STATISTICS 601

Advanced Statistical Methods

Mark S. Kaiser

Department of Statistics

Iowa State University

Fall 2005

Preface

The title given these notes, and the course numbered Statistics 601 at Iowa State University, is *Advanced Statistical Methods*. One might reasonably wonder, as did I in preparing these notes, what characteristics are needed for a statistical method to be considered advanced as opposed to elementary, introductory, or basic. Is a method advanced if it demands a certain level of mathematical sophistication to employ? Is a method advanced simply because it does not fall into the body of topics usually contained in courses taught at the undergraduate or beginning graduate level? Is a method advanced only if it requires extensive knowledge of computing to bring into practice? The Department of Statistics at Iowa State University now requires that all students intending to persue a PhD degree in statistics (including co-majors) take this course on advanced methods in preparation for the written preliminary examination. Thus, even more troubling than the question of what makes a statistical method advanced, is the question of what every PhD student should know beyond the topics contained in courses required of all MS students.

I have chosen to avoid addressing these questions directly because answers to them, even if such answers exist, fail to capture the intent of an upper-level course in statistical methods. I believe it is more profitable to ask what every PhD student should be able to *do*, rather than to ask what every PhD student should *know*. What every PhD student should be able to do is something our faculty refer to with the slippery phrase "demonstrate methodological maturity". I say this phrase is slippery because it falls into the category of something we can't define but can usually agree on when it is exhibited. That is, we can't tell you what it is but we know it when we see it, which is a most disturbing and unsatisfactory situation for graduate students who want

to know what is needed to be successful in their program of study. While not exhaustive of what everyone might consider methodological maturity, it seems that two characteristics are often involved in demonstrations of such maturity. First is the understanding that *methods* and *theory* are simply convenient titles used in our profession to refer to various topics, with theory comprising many topics that can profitably be considered without motivating examples or the existence of even hypothetical problems, and methods comprising many topics that generally do require at least a hypothetical situation in which they could be applied. The application of statistics often requires the use of both topics we would typically categorize as theory and those we would typically categorize as methods. Secondly, a demonstration of methodological maturity is often characterized by the avoidence of a modular organization of possible analyses. It is natural to organize our thinking along the lines of courses we have taken and, when faced with a problem requiring statistical treatment, to assign that problem to one of the modules formed by our formal educational experiences. I spent any number of years in which I would determine that a given problem was a "Stat 557 problem", or a "Stat 515 problem", or a "Stat 512 problem", although my numbering system was based on my own education rather than courses at Iowa State. This is not necessarily a bad thing, and may even be beneficial to our overall learning progress. And, many problems can be adequately dealt with using such an approach. But at some point a PhD statistician is expected to move beyond such a categorization. It is not the organization of topics into discrete courses that must be overcome (some type of organization is necessary to enable learning) but the use of this organization in considering how a scientific problem is to be approached from a statistical viewpoint. That is, methodological maturity is demonstrated when a statistician uses the knowledge and techniques at his or her disposal to con-

struct an appropriate analysis for a given problem rather than determining if the problem is sufficiently round to fit in the round hole or sufficiently square to fit in the square hole.

This course is intended to help you develop methodological maturity. It is organized along the lines of what I call *approaches* to statistical analysis. These notes are divided into three major parts, *Approaches Based on Randomization*, *Model Based Approaches*, and *The Bayesian Approach*. At some point I hope to add a fourth portion to the notes tentatively titled *Approaches to Inference*. Even within one of these major parts I have avoided a modular presentation. Thus, for example, general methods for constructing models are presented with little consideration of what estimation procedures might be used. Estimation procedures are presented without necessarily being attached to a particular type of model.

Following from the above considerations, the *advanced* in *Advanced Statistical Methods* refers to the way statistical procedures are used to build an analysis. You can expect to see some procedures and topics which are familiar to you, and you will also see many topics that are new. But this still leaves the issue of which particular topics should be included, and which passed over. These decisions were made largely on the basis of issues raised in the Introduction. The introductory portion of these notes begins with what will strike many as a somewhat obtuse philosophical discussion. Some might even describe the first section to follow as "fluffy" or a matter of irrelevant semantics alone. I believe, however, that consideration of what we mean by statistical methods and, even more generally, statistical analysis is important in understanding the structure of this course and, in particular, the course notes. Thus, I present this material not as a thorough discussion of philosophical considerations about what makes statistics a legitimate scientific discipline but, rather,

as an idication of what drove the necessary decisions about what to include and what to leave out of these notes. So read the Introduction, not from the viewpoint of how we make philosophical sense out of what we call the field of statistics, but from the viewpoint of how such considerations unfold into a way to organize our thinking about the topic we call statistical methods.

Mark S. Kaiser
August 2004

Contents

List of Figures

Chapter 1

INTRODUCTION

If you have not already done so, please read the preface to these notes now. The material in this introduction is presented to explain why many of the topics included in these notes are included, and why many other quite fine topics in their own right are not included. The broad context in which these considerations are made is that of scientific investigation, a bias that runs throughout these notes. While many of the procedures discussed in this course may be useful for the analysis of, for example, acceptance sampling data in a manufacturing process or data gathered from a stringent protocol for the licensing of a drug, our primary focus will be the analysis of problems from research in the applied sciences. Most scientific investigations are not repeatedly conducted in the same way making, for example, the concept of error rates in hypothesis testing of less importance than they would be in repeated decisions to accept or reject batches of manufactured goods. What is desired is an analysis of uncertainty and, by extension, a quantification of scientific evidence.

1.1 Statistical Analyses, Statistical Methods, and Statistical Techniques

The statistics profession seems to be in a continual process of attempting to define itself. This may be due, in part, from a long-perceived need to distinguish statistics from mathematics and to establish recognition not only in the academic community but also in society as a whole. In addition, however, the rapid increase in computational capability has provided new tools for both statisticians and workers in many other fields who pursue ways of examining data and making inferential statements on the basis of those analyses. This helps fuel what some statisticians see as an "identity crisis" for the profession (see, for example, the ASA Presidential Address published in the March 2004 issue of JASA). So what, if anything, defines statistics as a discipline? Given the diversity of activities that statisticians are involved in this question may be too broad to be given a satisfactory answer. But we may be able to make some progress by asking more specific questions about what constitutes statistical analyses, statistical methods, or statistical techniques.

1.1.1 Discussion Items on Statistical Analyses

What does it mean to say that a particular examination of a problem constitutes a *statistical analysis*? Some aspects of the procedures with which a problem can be investigated that might be mentioned as possible characteristics that qualify such procedures as statistical analyses include the following:

1. involves the use of observed data

2. involves learning from data

3. involves mathematical analysis

4. involves the examination of hypotheses

5. involves inferential statements

6. involves uncertainty

7. involves computation

The above list is not intended to include all the possibilities that one might think of, but these types of characteristics are often mentioned in attempts to "define" what constitutes a statistical analysis. In fact, phrases such as these are sometimes combined to produce an answer to the question "what is statistics?" such as several contained on the web page of the American Statistical Association,

> I like to think of statistics as the science of learning from data . . .
> Jon Kettenring, 1997 ASA President

> The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.
> American Heritage Dictionary

or

> The steps of statistical analysis involve collecting information, evaluating it, and drawing conclusions.
> Author of the ASA web page "What is Statistics?"

A difficulty with such attempts to give definition to statistical analyses is that they (almost necessarily) lack precision. Is solution of a set of partial differential equations, using observed data to set initial and/or boundary value conditions, necessarily a statistical analysis? Certainly this procedure involves data and mathematical analysis. Is solution of a stochastic differential equation necessarily a statistical analysis? It does involves uncertainty and, again, mathematical analysis. Are the use of machine learning or running a neural network necessarily statistical analyses? These activities involve computation and learning from data. Is examination of time recorded in flights around world in different directions to confirm or contradict Einstein's theory of special relativity necessarily a statistical analysis? Here we have both a hypothesis and data. I believe most of us would answer all of the above questions in the negative. While these types of activities and procedures *might* be included in something we would consider a statistical analysis they do not, in and of themselves, qualify as statistical analyses.

The quote from Dr. Kettenring given previously is an attempt to provide a simple indication of what statistical analysis is about in plain language that is readily interpretable, and I do not fault such efforts that appear in places such as the ASA web page. But they are not sufficient to provide guidance for what topics should be covered in an advanced course on statistical methods. The "learning from data" phrase has become popular, but really provides little distinction between procedures we would consider statistical in nature as opposed to other approaches. A few years ago I made a number of trips over a relatively short period of time. As I am slow to turn in travel expenses for reimbursement, I ended up with both high credit card balances and low cash availability. For several months I sent in the minimum monthly payment listed on several credit card statements. I noticed that, although I had made

regular payments in the minimum required amounts, the total credit card balances went up, not down! I learned from this experience, on the basis of data consisting of my account balances, that minimum monthly payments are not sufficient to reduce credit card debt. But I doubt that many of us would consider this a statistical analysis.

Just as it is perhaps too general to ask what statistics is, perhaps attempting to characterize statistical analyses remains too vague to address in a satisfactory manner. We might narrow the discussion by considering what we mean by a statistical *method* and then building up a statistical analysis as consisting of the application of one or more such methods.

1.1.2 Discussion Items on Statistical Methods

We turn our attention, then,to the question of what it means to say that some procedure is a *statistical method*. Consideration of this question might lead us to assert that one or more of the following characteristics apply to such methods:

1. involves a coherent (complete, logically consistent) process for the examination of data in all or a portion of a statistical analysis (even though we haven't entirely determined what we mean by analysis)

2. involves the expression or manipulation of data in a way that summarizes the information the data contain about a question or quantity of interest

3. involves mathematical expressions for estimation and/or testing of quantities in populations or theoretical probability distributions

Possible characterizations such as these are more well-focused than those given in the previous subsection for statistical analyses, but are also more tech-

nical in their implications. These more precise suggestions for characteristics of a statistical method do embody many of the possibilities given previously relative to the issue of a statistical analysis. The role of data comes through clearly, and learning from data is sharpened to the summarization of information contained in data about a particular question (hypothesis) or quantity (object of inference) of interest. Uncertainty appears at least in the form of probability distributions, and the role of mathematics in producing the appropriate summarization through estimation or testing is strongly implied. Several aspects of the phrases above have also restricted the range of procedures that might fall into one or more of these categories to a greater degree than what was considered for statistical analyses. The concept of population or distribution alone is sufficient to exclude possibilities such as the credit card story of Section 1.1; note here that the existence of such concepts does not imply that only indefinite probabilities are involved in inference. The insertion of a necessary logical basis for an overall procedure places a greater demand on what might be considered a statistical method than is implied by a mathematical solution to a well formulated problem, such as the solution of sets of differential equations.

It seems that general agreement concerning the status of various procedures as statistical methods is easier to attain than is true for statistical analyses. For example, I believe most statisticians would not have difficulty agreeing that maximum likelihood estimation, two-sample t-tests, and bootstrap estimation of standard errors qualify as statistical methods. But it is less clear that other common procedures employed by statisticians reach the level of a method. Whether the production of a scatterplot is a statistical method could certainly be questioned under criteria such as those listed above, and similarly for many other common data displays such as stem-and-leaf plots, boxplots or even

sophisticated rotating scatterplots in high dimension. Do procedures for the interpolation of data through the use of splines or kriging qualify as statistical methods? What if such procedures are combined with other assumptions or procedures to produce measures of uncertainty for predictions? Perhaps we should lower the demands another notch in our effort to define characteristics and consider something that might be called statistical *techniques*. We could then consider building up our concepts of statistical methods and eventually analyses as organized collections of such techniques.

1.1.3 Discussion Items on Statistical Techniques

We have now reached the point in our discussion at which we wish to consider what criteria might be used to categorize some procedure as a statistical technique. Possibilities might include that a technique:

1. forms a part of a statistical method
2. is anything that proves useful in a statistical analysis
3. is used primarily by statisticians but not other disciplines

At this point many of us, myself included, would conclude that we have simply progressed through the semantics of analysis, method, and technique to make the question of what qualifies under the headings less controversial and to allow nearly any procedure one wishes to consider to qualify as at least a statistical technique. But this recognition brings with it an important message. We typically use the words technique, method, and analysis with a sense of a progression that entails increasing demands on organized structure, completeness, and end product. A technique does not need to result in an inference or conclusion, and need not contain an overall logical structure. A method

must be more complete in attaining an narrow objective such as estimation or prediction of quantities and, in the view of many statisticians, also providing an associated measure of uncertainty. An analysis must combine one or more methods and techniques in an organized and logical fashion to culminate in conclusions, typically inferential in nature, about a question of interest.

I will close this brief subsection by pointing out that the development of the phrases statistical technique, method, and analysis is analogous to the way we use the words illustration, example, and application in referring to the presentation of manipulations of data or numbers. An illustration is constructed to display some aspect of the manner in which a mathematical manipulation of numbers functions. Illustrations may be built around real data but are also often constructed in a purposeful manner (e.g., with carefully chosen numerical values) for pedagogical purposes. Examples generally consist of at least portions of an actual data set to demonstrate the way some procedure functions in a real situation. Examples do, however, often simplify the setting through either culling of the data to be used (e.g., ignoring extreme values) or by putting aside justification of certain assumptions (e.g., taking "tuning" parameters as known). An application, in contrast, focuses on a particular scientific problem and must address all important issues involved in addressing that problem statistically (e.g., determining how one sets tuning parameter values in the context of the problem). A complete application must also culminate in an indication of what can be concluded about the problem based on the statistical analysis (yes, analysis) conducted. The connection with our present topic is that the same type of progression with increasing requirements in terms of completeness and logical organization is present in the progression of illustration, example, and application as has been developed for technique, method, and analysis in these notes.

1.1.4 Statistical Methods and Analyses Revisited

The primary thesis of this portion of the Introduction results from revisiting the concepts of methods and analyses in the light of the preceding discussion, and is illustrated by considering the statistical procedure of simple linear regression analysis. By far the most commonly employed manner of estimating the parameters of a simple linear regression model is ordinary least squares (ols). I would consider ols a statistical technique but not, in and of itself, a statistical method. This is because ols simply constitutes a solution to a problem of geometry, based on a mathematical projection theorem. What is statistical about ols estimators is that they possess certain properties (i.e., minimum variance among all unbiased linear estimators) as given by the Gauss-Markov theorem. It is when, in the usual progression in linear regression, we attach normal distributions to the independent and identically distributed error terms in the regression model that the procedure truly qualifies as a statistical method. In addition, at this point we are able to use the sampling distributions of estimators to develop inferential quantities (e.g., intervals for estimates or predictions, confidence bands for the regression line, etc.). When such quantities are used to reach conclusions about scientific aspects of the problem that led to the use of linear regression we can attach the label of analysis to the entire process. I would argue that the critical step in this development is the introduction of probability through the assignment of specific distributions to the error terms. Certainly, to realize the Gauss-Markov results we need to assume that these error terms are *iid* random variables with expectation zero and constant variance so, in a sense, probability has come into play at this point as well. But, we are unable to make use of this probability (except, perhaps, asymptotically) until we have a more clear description of the associated

distribution.

The more general conclusion suggested by this example is that it is the infusion of probability into a problem, and the subsequent derivation of results based on the probability structure developed, that forms the critical step from technique to statistical method. The addition of inference procedures based in a formal logical system then provides sufficient flesh for status as a statistical analysis. But, fundamentally, probability is the "glue" that unites various procedures and techniques that we use into what we call a statistical method and statistical methods into what we call a statistical analysis.

Probability is not a *thing*, or subject to easy definition. Probability is a concept or, rather, any of a number of concepts. As pointed out by Pollock (1990) in the preface to his book on nomic probability, "... concepts are characterized by their role in reasoning". All concepts of probability obey the same fundamental mathematical rules of behavior, which is why we can get so far in our statistical education without belaboring the distinctions among different concepts. But, the concept of probability that is utilized in a statistical analysis determines to a large extent the manner in which it is brought into play in the formulation of a problem, and also the interpretation of inferential statements that result from an analysis. Indeed, different probability concepts lead to different *approaches* to developing a statistical analysis for a given problem, and this course is organized broadly around several of the most common probability concepts employed by statisticians and the approaches to analysis attached to them.

1.2 Concepts of Probability

An argument has been made in Chapter 1.1 that probability is what we "hang our hats on" as statisticians. Probability is what allows a quantification of uncertainty in scientific investigation. But what is probability? There are any number of notions of probability, indicating that probability is not a *thing* but a *concept*. Concepts of probability include at least the following, taken from books on the topic by Edwards (1972), Kyburg (1974), Oakes (1986) and Pollock (1990):

1. Laplacian Probability
2. Relative Frequency Probability
3. Hypothetical Limiting Relative Frequency Probability
4. Nomic Probability
5. Logical Probability
6. Fiducial Probability
7. Propensity
8. Subjective Probability
9. Epistemic Probability

While comparing and contrasting these various notions of probability is a fascinating topic and has formed the basis for more than one or two book-length treatments, our concern with concepts of probability is the impact they might have on how we design a statistical analysis of a problem. As mentioned at the end of Chapter 1.1, statisticians do not often spend a great deal of time

worrying about the nuances of various probability concepts, because all legitimate concepts follow the same rules of mathematical behavior developed as the probability calculus. But we perhaps should be more concerned with concepts of probability than we generally appear to be. The concept of probability used in a statistical analysis influences first of all the way probability is brought into a statistical formulation of a problem (i.e., the approach to analysis). In addition, the concept of probability being employed in an analysis influences the meaning we should to inferential statements that result from the analysis. In short, concepts of probability are important to statisticians because they influence where probability "comes from" and where probability "goes to" in a statistical analysis.

Here, we will briefly cover four major concepts of probability: Laplacian Probability, Relative Frequency Probability, Hypothetical Limiting Relative Frequency, and Epistemic Probability. For each of these probability concepts we will list, in outline form, the basic notions involved, the fundamental characteristic, calculation, and a few brief comments.

1.2.1 Laplacian Probability

Sometimes also called *Classical Probability*, the Laplacian probability concept is well suited for problems involving fair coins, balanced dice or well-shuffled decks of cards, and so it could also be considered as *Gambling Probability*. This is the concept of probability we often see first in a presentation of set-theoretic probability operations and rules (e.g., Stat 101, Stat 104).

Basic Notions

1. Operation: observation, measurement, or selection

2. Sample Space: set of possible outcomes of an operation

3. Events: subsets of elements in the sample space

Fundamental Characteristic

Elements of the sample space (basic outcomes) are *equally likely*

Calculation

1. Let $\mathcal{S}$ denote the sample space, $E \subset \mathcal{S}$ denote an event, and $|A|$ denote the size of any set A.

2. $Pr(E) \equiv \frac{|E|}{|\mathcal{S}|}$.

Comments

1. It is easy to verify that the axioms of probability are all met by Laplacian probability (which is why we start with it in courses like 101 and 104).

2. The necessary fundamental characteristic of *equally likely* outcomes is typically the result of physical properties of the operation (e.g., flipping a coin, drawing a card).

3. Although essentially no one considers the Laplacian concept an acceptable general notion of probability, I believe it can be applicable in quite a number of situations and that statisticians use this probability concept more than we sometimes realize. In fact, Laplacian probability is used directly in some randomization based procedures.

1.2.2 Relative Frequency

When we talk about *relative frequency* probability we usually mean the topic of the next subsection, namely hypothetical limiting relative frequency. But direct relative frequency probability has some application in finite population problems.

Basic Notions

1. There exist a finite number of physically existing objects in a class $\mathcal{B}$.

2. An operation consists of observing whether a selected object also belongs to another class $\mathcal{A}$.

Fundamental Characteristic

Probability is a direct consequence of physical realities, that is, things that have actually happened.

Calculation

$Pr(\mathcal{A}|\mathcal{B}) = \frac{|\mathcal{A}|}{|\mathcal{B}|}$

Comments

1. This is a purely *material* concept of probability that is clearly inadequate for many problems that we would like to apply probability to. For example, a fair coin is to be tossed 3 times and then destroyed. What is the probability that an arbitrary toss is a H? Here, our three tosses of this coin are the class $\mathcal{B}$ and tosses that result in H are then our class $\mathcal{A}$. We might like to say 1/2 but, given there will be exactly 3 tosses, a H

on and arbitrary toss will have relative frequency of either 0, 1/3, 2/3,or 1 so that only these choices agree with physical reality.

2. Despite its clear inadequacy as a general notion of probability, I have included relative frequency here because it can apply in problems that involve finite populations (we will illustrate this later in the course).

1.2.3 Hypothetical Limiting Relative Frequency

The concept of probability we usually defer to in traditional analyses based on the theories of Fisher, Neyman, and Pearson (Egon, not Karl). This is what we usually mean when we refer to relative frequency or frequentist probability.

Basic Notions

1. Operations (as in Laplacian and Relative Frequency probability) but that can *at least hypothetically* be repeated an infinite number of times.
2. Sample Space (as in Laplacian probability)
3. Events (as in Laplacian probability)

Fundamental Characteristic

Operations that can be repeated *hypothetically* an infinite number of times.

Calculation

1. Let n denote the number of operations conducted, and let E_n denote the number of operations, out of the n operations conducted, that result in an outcome contained in an event E.

2. $Pr(E) \equiv \lim_{n\to\infty} \left(\frac{E_n}{n}\right)$.

Comments

1. It is not as easy to verify that the axioms of probability are all met by Hypothetical Limiting Relative Frequency Probability, but this concept agrees with Laplacian Probability *when both are applicable* (e.g., flipping a coin).

2. Outcomes need not be equally likely, but one-time or individual-specific events are problematic (e.g., evolutionary events)

1.2.4 Epistemic Probability

Any concept of probability that cannot be expressed in terms of *physical events* can be considered epistemic probability. In literature on theories of probability, epistemic probability is often equated with *subjective* or *personal* probability. These are somewhat "loaded" terms and there has been extensive debate about whether objective probabilities can truly exist or, conversely, whether subjective probability is legitimate as a vehicle for empirical investigation. We will take the more pragmatic view of many statisticians that non-physical probability concepts can be useful, and refer to such concepts as epistemic probability.

Basic Notions

1. Probability ≡ knowledge or belief.

2. Belief is updated or modified in the light of observed information.

3. Mathematical formalism is necessary for belief to be modified in a *coherent* manner.

Fundamental Characteristic

Probability $\equiv$ knowledge or belief

Calculation

1. Let $Pr(E)$ denote my belief about an event E. Let $Pr(\boldsymbol{y}|E)$ denote the probability of observations $\boldsymbol{y}$ under event E and $Pr(\boldsymbol{y}|E^c)$ the probability of observations $\boldsymbol{y}$ under the complement of E.
2. $$Pr(E|\boldsymbol{y}) = \frac{Pr(\boldsymbol{y}|E)Pr(E)}{Pr(\boldsymbol{y}|E)Pr(E) + Pr(\boldsymbol{y}|E^c)Pr(E^c)}$$

Comments

1. Does *not* necessarily contradict the notion of an absolute truth.
2. Does *not* necessarily minimize the importance of empirical evidence in scientific evaluation.
3. *Does* presume that scientific investigation rarely (if ever) takes place in a vacuum of knowledge or belief.

1.2.5 Transition to Approaches to Statistical Analysis

Recall from the beginning of this section that concepts of probability affect the way that probability is brought into a problem and the manner in which it gives meaning to inference that results from the analysis of a problem. Another way to say this is that concepts of probability are important in determining *where probability comes from* in an analysis and *where probability goes to* as the result of an analysis. One organization of this is in terms of *approaches to statistical analysis*, which are divided here along the following lines:

1. Analysis Through Randomization

   Approaches based on randomization make use primarily of Laplacian and Relative Frequency probability concepts as the basis for analysis.

2. Analysis Using Models

   Model based approaches are often presented in terms of Hypothetical Limiting Relative Frequency and this is largely adequate. There can be questions of whether such relative frequency is always adequate for these approaches, particularly for interpretation of inferential statements. Considerations related to this issue have, in fact, motivated many of the probability concepts listed at the beginning of this section but not discussed further (e.g., fiducial probability).

3. Bayesian Analysis

   The Bayesian approach may well make use of relative frequency probability, particularly in construction of the *data model*. But, the distinguishing characteristic of a Bayesian analysis is that it also makes use of Epistemic probability in the form of *prior* and *posterior* distributions.

Part I

APPROACHES BASED ON RANDOMIZATION

Chapter 2

Populations, Attributes, and Responses

Part 1 of these notes includes discussion of two approaches to statistical analysis that I call the *sampling approach* and the *experimental approach*. The sampling and experimental approaches to statistical analysis differ in ways that will be made explicit in the sequel but, in their most basic form, they have in common the underlying concept of a *population*. We will present these approaches largely under the assumption of a finite, physically existing population, but will indicate attempts that have been made to extend the the basic ideas involved to broader classes of problems. In this chapter, then, we consider first how we might define the concept of a population. Chapter 2.1 presents the strict concept of a finite, physically existing population, while alternative notions that relax the strictness of this definition are considered in Chapter 2.2. In large part, the distinctions involved between the ideas presented in these two sections impact delicate philosophical issues involved in determining a precise meaning for inferential statements, rather than the operational aspects of sam-

pling or experimental approaches. Much of the relevant discussion concerning such issues will be deferred until the (currently non-existing, but planned) part of the notes on Approaches to Inference. Nevertheless, I believe it is beneficial to understand that defining a population is not necessarily a trivial exercise even at this point in our presentation. At a gross level, some understanding of the importance of population definition can be obtained simply by recalling traditional presentations of what is meant by the concepts of error rates or coverage levels in introductory statistics courses. To communicate these concepts we generally refer to repeated sampling or repeated observation of portions of a population. Precisely what entities are to be sampled or observed, and what collection of such entities constitutes the total population are questions that are usually brushed over or taken as obvious (and examples are carefully chosen in which the answer to these questions are fairly obvious). If an inferential statement derives its meaning from repeated operations conducted on components of some population, it behooves us to expend some effort in determining precisely how we define a population and its associated characteristics.

Connected with the definition of a population are the concepts of *attributes* and *responses*. These refer to quantifiable and observable phenomena attached to the basic units of a population (once we have determined what might be meant by the phrase population unit). For the most part, the concept of attributes is involved with the sampling approach while the concept of responses is involved with the experimental approach. The notions of attribute and response will be discussed in Chapter 2.3.

2.1 Finite, Physically Existing Populations

The concept of a finite, physically existing population is a relatively "pure" or "strict" notion of what we would consider to constitute a population. It is also the easiest to comprehend, although too often we fail to consider the potential ramifications of departures from this concept of a population when we allow more relaxed definitions such as those introduced later in the Chapter.

By a finite, physically existing population we mean a finite collection of discrete entities we will call *population units*. Such units may be people, animals, or objects such as steel beams, buildings, or ships. The phrase *physically existing* implies that such units are manifested in the real world in which we live, not of hypothetical existence, and also are not subject to arbitrary definition by a scientific investigator or statistician.

Arbitrary means not governed by any principle, or totally capricious in nature. There is often no issue that requires extensive consideration in the case of populations that consist of living organisms, such as people or cows. Cows exist, and the definition of what constitutes a cow is not subject to a great deal of debate (comical considerations aside). But other common situations may demand more detailed examination. Consider, for example, the division of an agricultural field into plots. Suppose we have a 1 hectare square field; one hectare is 100 are and 1 are is 100 square meters. We want to divide this field into 25 plots. It might be natural to begin at one corner of the field and lay out square plots 20 meters on a side. Each plot would then consist of 4 are in a square shape. But, we could just as easily form 25 plots by taking 4 meter strips running the entire length of the field (100 m). In this configuration, each plot would also consist of 4 are, but in a much different configuration. Alternatively, we could discard a boarder strip of some given width around the

entire field and divide the remaining area into 25 plots in various ways. The point is that the plots are subject to arbitrary definition by an investigator, and cannot be considered units of a physically existing population as we have defined this concept. Contrast this with a situation in which we are given a map of 100 fields, which may differ in size and shape. While these fields may well have resulted from an arbitrary decision by someone at some point in time (a surveyor, a politician, a farmer, etc.) for our purposes they simply exist as given on the map, and are not subject to arbitrary definition *by us*. We might well consider such fields as constituting a physically existing population.

Assuming a finite population that consists of well-defined physical objects that are not subject to arbitrary definition, these fundamental population units are sometimes aggregated into larger units, and a population defined as consisting of the larger units. It is not entirely clear whether this procedure introduces any serious difficulty for the strict concept of a finite, physically existing population. Certainly, if the fundamental units are discrete, physically existing objects then so too are any aggregates of those units. But do the aggregate units exist without arbitrary definition? The answer may well depend on how the aggregation occurs, and in how flexible one is willing to be in what is considered arbitrary. If there are "natural" groupings, such as people into nuclear families (based on genetic relatedness) or functional families (based on living arrangements) one might argue that the aggregation is not arbitrary. What if the aggregation is based on random selection of fundamental units for aggregate units? Random selection is diametrically opposed to arbitrary selection in that it follows an exact and stringent principle (recall that arbitrary means not based on any principle). On the other hand, determination of the size of aggregate units may well be arbitrary in nature; you put 5 chickens in a group, I'll put 6. I would argue that truly random aggregation results in

a *conditional* population definition, the conditioning being on the size of the aggregate units.

2.2 More Flexible Notions of Population

As should be clear from the previous section, definition of what is meant by a population is not as clear cut as we sometimes would like to believe. A fundamental consequence of a finite, physically existing population as defined in Chapter 2.1 is that all units of a population may be individually identified and assigned discrete labels (e.g., 1, 2, ..., N). This automatically results in the classical *sampling frame* of survey sampling methodology or a clearly defined *reference class* for the experimental approach. But these approaches to statistical analysis would certainly be of limited use if the situations in which they could be applied was totally constrained by the need for a strictly defined finite, physically existing population. There are a number of violations of the population concept of Chapter 2.1 that are both common in applications and appear to pose less than insurmountable obstacles for these statistical approaches. Several of these difficulties are presented below and, for each, the extent to which one is willing to accept departures from a tight definition of population will depend on philosophical considerations and the degree to which a convincing argument can be constructed for the adequacy of the analysis undertaken.

1. Non-Static Populations

   Many, if not most, interesting situations in which we might consider making use of either the sampling or experimental approaches do not lend themselves to the definition of a static population. Consider a survey of

public opinion about construction of a new mega-mall on the outskirts of a small city. One desires a (random) sample of city residents, and it is that group of people about which conclusions are to be drawn (favor or oppose construction of the new mall). There do not appear to be major issues relative to defining a finite, physically existing population of people in this case. The context of the problem is likely sufficient to deal with issues such as transient versus permanent city residents or to restrict the relevant population based on age (although what to do with teenagers might be an interesting question). But, even in small cities, people move in, move away, die, and are born on a daily basis. Thus, the actual population about which we wish to draw a conclusion will necessarily be different from the actual population from which we can draw a sample. Nevertheless, few of us would consider the use of sampling methodology inappropriate in this situation. This is because we believe (expect, anticipate) that the population available for sampling will be sufficiently similar in all relevant respects (primarily attitude about a proposed mall) to the population about which a conclusion is desired.

The occurrence of a non-static population is arguably a reality in nearly all problems involving living organisms. In an experiment to determine the efficacy of a veterinary vaccine to prevent viral infections in pigs, conclusions are made about a certain type of pig (breed, age, health status) in general, not about the collection of pigs of that type that actually existed at the beginning of the study. In a study of the association of having taken advanced placement courses in high school with the chances of success (leaving this undefined at the moment) in college, the population to be sampled is college-bound high school students at

a particular point in time and region in space. But the population of interest is college-bound high school students in general.

One aspect of non-static population status can have important consequences for statistical analysis. A catastrophic change in the composition of a population during the time period of concern can vitiate the intent of a study. It is clearly not wise for a political candidate to survey residents of a given congressional district about a proposed tax increase if the state is in the process of re-districting, particularly if the new boundaries are likely in include or exclude a major population center (e.g., large city) from the district. With these types of population changes forming fairly obvious caveats, we are typically willing to apply the methods of sampling or experimentation to even non-static populations.

2. Populations of Unknown Size

It may be the case that a population of discrete physical objects can be identified, and that population must be logically finite in size, but the number of units comprised by the population is unknown. This may be due to the fact that individual units in the population cannot be uniquely identified for the entire population, but only for portions of the population chosen for observation. For example, consider a study of nitrate concentration in wells used for human water supply in a rural portion of a state. It is most likely that a list of all such wells is not available. If a particular areal unit (quadrat or subsection) is visited, it may be possible to enumerate all of the wells that occur in that small area. These situations are common, and survey sampling methodology has been developed to deal with many such situations.

Many, if not most, controlled experiments are not conducted with units

from a population of known size. Consider an experiment to assess the effect of a growth hormone given to chickens on the protein content of eggs produced by those chickens. Here, the fundamental population units of concern are chickens (perhaps of a particular type). They are discrete, physically existing entities not subject to arbitrary definition by a statistician. In this case, and for many experimental settings, the unknown size of a population and the non-static nature of the population of interest do not seem to be distinct issues. To object that we do not know the total number of chickens of the type of concern is largely irrelevant, since it will change over the course of the study anyway. It is mentioned here because the experimental approach in "pure" form demands random selection of population units to be included in a study. That this is almost never possible changes the inferential framework available, as will be discussed at greater length in later chapters.

Another type of situation that results in unknown population size occurs from populations defined in terms of units that "join" the population over time. Consider a study to determine the proportion of automobiles that exceed the speed limit by more than 5 miles per hour during the weekend on a given portion of interstate highway. Here, discrete, physically existing population units are readily available, but the size of the population is not only unknown but actually undefined prior to the time period of interest. In this example we would also have a non-static population, since it is unlikely that only one particular weekend would be of interest.

Of these types of departures from our basic concept of a physically existing population, the first and second seem to pose less difficulty for a

statistical approach based on randomization than the last, in which units join the population over time. This is because such situations generally result in population units that are available for sampling only at the time point in which they are identified as belonging to the population. To accomplish some type of randomization in sampling we must make use of surrogate units assumed to be unrelated to the object of investigation. In the example of automobile speeds we might, for instance, define *sampling units* as intervals of time, distinct from the populations units of true interest. A complete list of all time intervals in a weekend is both of known size and is easily obtainable. If time of day is unrelated to the object of investigation (proportion of speeding autos) then this poses little difficulty. On the other hand, if time is related to the phenomenon of interest, then this complicates the situation and our ability to apply sampling methodology.

3. Hypothetical and Constructed Populations

   This departure from our strict concept of a finite, physically existing population impacts primarily what is presented later in these notes as the experimental approach. Many controlled experiments are conducted on population units that are "constructed" by the investigator. Petri dishes containing a given amount of growth media and inoculated with a certain quantity of some micro-organism are constructed in the laboratory. Cell cultures to which are added various antibodies and antigens are similarly produced by the scientist conducting the investigation. Plantings of a certain horticultural variety are made by adding plant material and soil or other growth media, and may be subjected to various light regimens to determine the effect on floral productivity. In these cases, the defin-

ition of what constitutes a population unit may be less than arbitrary, yet those units do not exist outside the universe defined by the study in question. If the population is considered to be finite with a size determined by the number of units constructed, then the entire population has been used in the experiment and there is no need to make inferential statements about it. But, almost certainly, that collection of units is not the object of inference. If we are to make inference based on what would be expected in repetitions of the study protocol we are referring to repetitions conducted with hypothetical population units, since the exact study protocol is rarely repeated more than a few times, and generally not more than once.

The impact on the experimental approach to statistical analysis of populations defined on the basis of constructed units and, typically, of a hypothetical nature, is almost identical to that of experiments conducted with populations of unknown size or with units from non-static populations. This impact will be discussed at greater length in Chapter 4 and, particular, in Part IV of the course when we deal with inference.

4. Arbitrarily Defined Population Units

   The issue of arbitrary definition of population units has been raised previously in the discussion of Chapter 2.1. This issue may be considered relevant primarily for the sampling approach since under the experimental approach one could consider arbitrary population units to have been constructed. The example of Chapter 2.1 concerning agricultural plots in a field illustrates this nicely. If these plots are to become the objects of some type of treatment (e.g., fertilizer level) then they may easily be viewed in the same light as the petri dishes or horticultural plantings of

the previous item in this list.

Concerning the sampling approach, there is lack of agreement among statisticians about how serious the consequences of arbitrary definition of population units are. On one hand, it is often possible to directly apply the operational procedures of survey sampling methodology once population units have been defined, regardless of whether that definition was arbitrary or not. If we attach a caveat to inferential statements that such conclusions are conditional on the definition of population units employed, similar to what was discussed for aggregation of fundamental units based on random grouping in Chapter 2.1, then many statisticians see little reason the sampling approach cannot be used with confidence. On the other hand, arbitrary definition of population units may well interfere with the notion of units having inherent *attributes* of interest as they will be defined in the next section, and this is of fundamental importance for the sampling approach. If this is the case, then one may well prefer to formulate the statistical problem in the context of a model based or Bayesian approach, and this is the opinion of many other statisticians. How far one is willing to "bend the rules", so to speak, is a matter of personal choice by a statistician. A statistician who has considered the issue and made a deliberate decision is likely to be able to defend that decision, even if it does not meet with universal approval. A statistician who is not even aware there is an issue involved is on much less solid footing.

2.3 Attributes and Responses

Nearly all statistical applications involve the analysis of observed data, at least a portion of which typically constitute a quantification of some phenomenon of primary interest in a study. Even qualitative phenomena are typically assigned arbitrary numerical values, such as 0 and 1 for binary measures. In the sampling approach, we will refer to such values as *attributes* of population units, while in the experimental approach we will modify the concept of an attribute slightly to develop the notion of a *response*.

In its fundamental form, the sampling approach assumes that associated with each unit in the population is a fixed value of interest, called an attribute. These attributes are characteristics of the population units, so that observation of the same unit will always produce the same value of the attribute. In this framework, the values of attributes associated with population units are not realizations of some random mechanism (i.e., not realized values of random variables). They are, rather, fixed, immutable characteristics of the units. My car is a Honda civic. It will always be a Honda civic, unless it is transformed into a different car, in which case it is no longer truly the same unit in a population of cars. This notion of an attribute is quite stringent. My car is red, which is one of its attributes, unless I decide to get it painted a different color. Is it's color then no longer one of its attributes? Clearly, we cannot demand of an attribute that it remain characteristic of a population unit in perpetuity. We can only require that an attribute remain characteristic of a population unit for a certain time span appropriate within the context of any given study.

All studies are relevant for only a certain time span. Public opinions change based on events that cannot be anticipated at the time of a study; consider, as

an extreme example, the general attitude about figherfighters in New York city prior to and after the terrorist attack of September 11, 2002. The effect of an insecticide on yield of a given crop may change as the target organism(s) evolve resistance to the active ingredients of the insecticide. The average price of a 3 bedroom home in a given region will change over time. That an attribute is an unchanging characteristic of a population unit must be interpreted within the time span of relevance for a particular study. We take an attribute to be a characteristic of population units that has a fixed value for the course of a study. In fact, the time frame within which attributes can change for given population units helps define the time frame within which conclusions from a study are relevant.

Under the experimental approach, we will refer to measured or observed values of the primary quantity or quantities of interest as responses. Responses are essentially the same concept as attributes of population units in sampling and are considered characteristics of population units, rather than values associated with random variables. A difference, however, is that responses are allowed to be influenced by external factors within the time frame of a study. For example, while low density lipoprotein (so-called "bad"cholesterol) levels in humans may be considered characteristics of individuals for a reasonable period of time, the level of cholesterol in an individual can certainly be influenced by exercise, diet, and certain drugs. If we were interested in the average cholesterol level in male professors between the ages of 40 and 60 we could take a sample in which cholesterol level is considered the attribute of interest. If, however, we were interested in whether cholesterol level in male professors between 40 and 60 years of age could be influenced by consumption of one serving of oatmeal per day (for 30 days say), we would exercise control over at least a portion of the influence of diet and consider cholesterol level a response

of interest. Thus, the term *responses* refers to attributes under a given set of pertinent conditions and external influences. It is precisely the effect of certain of these external influences (i.e., treatments or factors) that the experimental approach is designed to assess.

One additional aspect of how attributes and responses are defined is important, particularly under the experimental approach. In introductory courses we often emphasize the concept of *experimental unit* and distinguish it from *sampling unit*. In an experimental situation, a sampling unit is the entity on which an observation or measurement is taken. Sampling units frequently correspond to the fundamental population units of Chapter 2.1. Experimental units may also correspond to these population units. But, if fundamental population units are aggregated into groups to which treatments are applied, experimental units correspond to these aggregates. For example, if plantings of a horticultural variety are constructed as described in Chapter 2.2, and groups of these plantings placed in controlled environmental chambers to expose them to different light regimens, then it is a group of plantings placed in one controlled environmental chamber that constitutes an experimental unit in the study. What is important is that the concept of a response applies to experimental units not sampling units, unless the two coincide. Responses, then, may consist of aggregated measurements made on individual sampling units just as experimental units may consist of aggregated groups of fundamental population units. In the horticultural example, the phenomenon of interest may be the number of blossoms produced in a given time span, which can be counted for each planting (sampling unit) but then must be totaled for all plantings in a given chamber to constitute the response for that experimental unit. The reason this is such a crucial point is that, as we will see in Chapter 4, probability enters a problem under the experimental approach only through

randomized treatment assignment. Treatments are assigned to experimental units and thus responses must be considered characteristics of experimental units as well. To fore-shadow a latter portion of the course, we will note that there are alternative structures in which to formulate this problem based on statistical modeling, in which individual plantings do have individual "responses". If such an alternative structure is used for analysis, however, one is no longer applying the experimental approach to statistical analysis, which is our main point here.

Chapter 3

The Survey Sampling Approach

References: Cochran (1977), Wolter (1985), Sarndal *et al.* (1992), Thompson (1992).

Consider a situation that involves a population of the type described in Chapter 2.1, for which we are interested in a given attribute of the population units; in nearly any real application we will have interest in multiple attributes but, for now, consider only one of these. Let the values of this attribute for any arbitrary ordering of the units in the population be denoted as $x_1, x_2, \ldots, x_N$.

Example 3.1

The National Marine Fisheries Service (NMFS) is responsible for estimating the total commercial catch of groundfish (a certain class of commercially valuable fish) in the West Coast fishery, which consists of waters off the west coast of the US from California up to southern Washington. The total commercial catch consists of fish that are caught and delivered to a processing plant (i.e., are sold) plus fish that are caught but dumped back into the ocean (called

"discard"). Fish may be discarded for any number of reasons, such as because they are too small to be economically valuable, but discard adds to the overall mortality caused by the fishing industry and is thus important to know. Records of kept catch are obtained by accessing records that must be provided by processors (how much they bought, who they bought it from, when they bought it, etc.) but getting an estimate of the discarded catch is more difficult. To facilitate this estimation, trained observers are placed aboard some (generally about 10%) of the trips made by fishing vessels. The observers are able to record the amount of discard (in weight) for hauls that are observed. Vessels, on the other hand, are required to keep "logbooks" containing information on trip dates, number of hauls made, time and geographic area of hauls, and other information regardless of whether the trip was officially observed or not.

Nearly all of the estimation strategies considered to date by NMFS have been based on survey sampling methodology. A fundamental question in the use of such estimators is what definition should be given to the population and, in particular, the basic population units. One option is to define population units as trips made by vessels. The total number of trips made over a given period (e.g., fishing season) is known. On the other hand, a population so defined cannot be enumerated until after the fact (i.e., after the fishing season is over) because there is no way to tell how many trips will be conducted before they actually occur. Alternatively, one might define the population to consist of vessels that purchase permits for the fishery (which they have to do before fishing); units then are vessels. In this case, a list of the entire population is available prior to any fishing. In the former case (population units defined as fishing trips) we may encounter difficulties with populations of unknown size, as discussed in Chapter 2.2, while in the latter case (population units defined as vessels) we may encounter difficulties in defining an observable attribute

for each unit. That is, if the attribute of interest is total catch, it may not be possible to obtain these values for each sampled unit since it is difficult to place an observer on board a vessel for all of the trips made by that vessel in such a way that all hauls made on each trip are observed.

This example also illustrates what may be a difficulty in defining the quantity of concern as an actual attribute of population units, regardless of whether those are considered to be trips or vessels. Observers cannot actual measure the total weight of discard for a given haul. Rather they estimate discard using several approved methods that may involve sampling of a portion of the fish in a haul or in a "discard pile". Thus, if two observers were to observe the same haul, it is exceedingly unlikely they would arrive at the same value for weight of discard. Similarly, processors may sometimes use estimation procedures rather than exact measurements for weight (particularly in large deliveries of fish).

3.1 The Sampling Frame

Of fundamental importance in putting the sampling approach to statistical analysis into action is the formation of what is known as a *sampling frame*. In its most basic form, the sampling frame for a given problem is simply a list, in arbitrary order, of all of the units in a population, and a unique identifier for each unit. Forming sampling frames in more complex situations, such as when the basic population units cannot be enumerated or identified in total, is addressed in Statistics 521 and Statistics 621. Here, we are attempting to communicate the fundamental ideas of the sampling approach, and we will assume that a complete sampling frame of all basic population units is able to be constructed.

In operation, an important aspect of a sampling frame is that, once formulated, the identifiers of population units remain inviolate. That is, if a given population unit is designated as unit 4, it is always unit 4, and x_4 refers to the value of its attribute, never the attribute of any other unit in the population, regardless of whether unit 4 is selected for observation (i.e., to be included in a sample) or not.

Example 3.1 (cont.)

NMFS identifies vessels and trips with unique numbers, and hauls within trips are identified sequentially as $1, 2, \ldots,$. The sampling frame for our population of vessel/trip/haul might be represented, for example, as a table of the following form:

Vessel	Trip	Haul	Unit Index	Total Catch(lbs)
115	81	1	1	x_1
115	81	2	2	x_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
131	51	5	N	x_N

3.2 Population Statistics as Parameters

Within the context of a basic sampling approach, statistics that are computed over the entire population are considered "parameters". For example, the population mean, total, and variance are defined as,

$$\mu \equiv \frac{1}{N}\sum_{i=1}^{N} x_i,$$

$$\begin{aligned} \tau &\equiv \sum_{i=1}^{N} x_i = N\,\mu, \\ \sigma^2 &\equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2. \end{aligned} \tag{3.1}$$

Similarly, the proportion of the population satisfying some condition or having an attribute that belongs to a particular class $\mathcal{A}$ is,

$$P_{\mathcal{A}} \equiv \frac{1}{N}\sum_{i=1}^{N} I(x_i \in \mathcal{A}), \tag{3.2}$$

where $I(\cdot)$ is the usual indicator function having the value 1 if the condition of the argument is true and 0 otherwise.

We will be concerned with the estimation of population parameters such as μ, τ, σ^2, and $P_{\mathcal{A}}$. Note, at this point, that we have discussed nothing involving randomness, probability, or uncertainty. If every unit in the population could be observed, we would be able to *compute* these quantities exactly, and there would be no need for statistics to enter the picture – this would be a census.

The need for statistics arises from the (usual) situation in which not every unit in the population can be observed. We must then estimate population parameters rather than simply compute them, and there is uncertainty involved in our estimation because we have less than complete observation. This is called *sampling error* in the survey sampling literature, and is the only source of uncertainty to be considered. What is called *nonsampling error* may arise from sources such as measurement error or nonresponse (a unit chosen for observation is impossible to actually observe). We will mention nonresponse in a later section, but for the most part in this development will consider only sampling errors.

3.3 Simple Random Sampling

The basic method for obtaining a portion of the population for observation (i.e., a sample) is to use simple random sampling. This method is often used in its own right to obtain a sample directly, and also forms the basic building blocks from which we may construct more complex random sampling designs.

3.3.1 Simple Random Sampling Defined

Definition:
A simple random sample of n units, selected from a population of N units, is any set of n units selected in such a way that all possible distinct sets of size n are equally likely to be selected.

How many sets (samples) of size n are there? That is, if $\mathcal{S}_n$ denotes the set of possible samples (the set consisting of distinct sets of size n), what is the size of $\mathcal{S}_n$?

$$|\mathcal{S}_n| = \left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{(N-n)!\,n!}.$$

Denote (with an arbitrary ordering) the set of possible samples as $\mathcal{S}_n \equiv \{S_{n,1}, \ldots, S_{n,M}\}$, where $M = |\mathcal{S}_n|$. Then using Laplacian probability, we can calculate the probability of any given sample $S_{n,k}$ say, as, for $k = 1, \ldots, M$,

$$Pr(S_{n,k}) = \frac{1}{M} = \frac{1}{|\mathcal{S}_n|} = \frac{(N-n)!\,n!}{N!}.$$

Comments

1. What is of fundamental importance here is that, given that we will estimate a population parameter on the basis of a sample, we have just introduced probability into the overall problem formulation.

2. It is worth emphasizing that, if one looses control of the sampling process, then there is in fact no basis for statistical treatment of the problem. That is, with a non-random sample, it is more than an issue of answers that are likely to be "less than perfect", it is an issue of having lost any logical basis for conducting an analysis based on the use of probability.

3.3.2 Obtaining a Simple Random Sample

There are several ways to actually obtain a simple random sample, although the sequential selection procedure described below is by far the easiest to program. We assume that a complete sampling frame (list of population units with unique identifiers) is available, such as the table of vessels, trips, and hauls given previously for the groundfish example. For both of the procedures to obtain a simple random sample given in what follows, let $\{U_i : i = 1, \ldots, N\}$ denote the population units (in any arbitrary, but fixed order).

Group Selection Procedure

Directly from the definition of a simple random sample at the beginning of Chapter 3.3.1, it is clear that one procedure to select such a sample from a population of N units would be to enumerate all of the possible samples of size n, and select one of them at random. For example, if $N = 6$ and $n = 2$, the possible samples could be enumerated as follows:

Sample	Composition	Sample	Composition
1	$\{U_1, U_2\}$	9	$\{U_2, U_6\}$
2	$\{U_1, U_3\}$	10	$\{U_3, U_4\}$
3	$\{U_1, U_4\}$	11	$\{U_3, U_5\}$
4	$\{U_1, U_5\}$	12	$\{U_3, U_6\}$
5	$\{U_1, U_6\}$	13	$\{U_4, U_5\}$
6	$\{U_2, U_3\}$	14	$\{U_4, U_6\}$
7	$\{U_2, U_4\}$	15	$\{U_5, U_6\}$
8	$\{U_2, U_5\}$		

Something worth mentioning at this point, because we will most likely see it repeatedly, is how one of these samples would be randomly selected in a computer algorithm. There are $M = 15$ possibilities. A computer algorithm to select one at random would be:

1. Generate one value u^* from a uniform distribution on the interval (0, 1). This is easily accomplished using almost any statistical software or language (e.g., Splus, R, SAS, etc.)

2. If $u^* \leq (1/15)$ select sample number 1, which would consist of population units 1 and 2, $\{U_1, U_2\}$.

3. If $(1/15) < u^* \leq (2/15)$ select sample number 2.

4. In general, if $((k-1)/15) < u^* \leq (k/15)$ select sample k; $k = 1, \ldots, 15$.

While programming such a group sampling method is not difficult for problems such as this illustration (with small N and n), it can become much more cumbersome for most real problems, such as that of Example 3.1 in which $N = 6,312$ and we may want a sample of size $n = 100$ say. This motivates

the next procedure which is equivalent to the group selection procedure, but is easier to accomplish (i.e., program) in practice.

Sequential Selection Procedure

The sequential selection procedure is simply a computational version of exactly what you would do if asked to draw n chips from a bowl containing a total of N numbered chips. That is,

1. Select 1 unit at random from a population of size N.
2. Select 1 unit at random from the remaining population of size $N-1$.
3. Select 1 unit at random from the remaining population of size $N-2$.

$\vdots$

n. Select 1 unit at random from the remaining population of size $N-(n-1)$.

It is easy to prove that the sequential procedure is equivalent to the group procedure. Restricting attention to our small example with $N = 6$ and $n = 2$, suppose that the group selection procedure resulted in selection of sample number 5, consisting of $\{U_1, U_6\}$, which had probability $(1/15)$ of being chosen. What would be the probability of this sample under the sequential procedure?

There are two mutually exclusive ways to select the sample $\{U_1, U_6\}$ in the sequential procedure: (1) select U_1 at step one and U_6 at step two (call this the event F_1), and, (2) select U_6 at step one and U_1 at step two (call this the event F_2). Now,

$$\begin{aligned} Pr(S_{2,5}) &= Pr(F_1 \cup F_2) = Pr(F_1) + Pr(F_2) \\ &= (1/6)(1/5) + (1/6)(1/5) \\ &= (1/15), \end{aligned}$$

which is the same as under the group selection procedure. Since this does not depend at all (i.e., is WLOG) on which particular sample the demonstration is for, the sampling procedures are equivalent for this example.

3.4 Estimation For Simple Random Samples

3.4.1 The Basic Estimators

It is natural to consider estimating population parameters such as μ, τ, σ^2, or $P_{\mathcal{A}}$ using the same formulae as in (3.1) and (3.2) with only units selected for the sample, rather than the entire population. Letting S_n^* denote the sample (of size n) selected, we can represent such estimates without re-indexing sampled and other population units as follows:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{N} x_i\, I(U_i \in S_n^*) \\
\hat{\tau} &= N\,\hat{\mu} = \frac{N}{n}\sum_{i=1}^{N} x_i\, I(U_i \in S_n^*) \\
\hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{N} (x_i - \hat{\mu})^2\, I(U_i \in S_n^*) \\
\hat{P}_{\mathcal{A}} &= \frac{1}{n}\sum_{i=1}^{N} I(x_i \in \mathcal{A})\, I(U_i \in S_n^*)
\end{aligned} \tag{3.3}$$

3.4.2 Properties of the Estimators

There are two major avenues by which to approach elucidation of properties of the basic estimators of Section 3.4.1. One, which is perhaps the more straightforward, is to introduce the concept of random variables and consider the indicator functions contained in the expressions of equation (3.3) to be binary random variables. One may then use the operations of mathematical

expectation and higher moments to derive properties of the estimators (which, of course, then being functions of random variables are themselves random variables). But, we have not used the concepts of mathematical expectation, moments, or distributions up to this point, and it is possible to derive properties of the estimators without those mechanisms, which is what we will attempt here, in an effort to keep the sampling approach "pure" with respect to its origins. Thus, we will rely on the basic operator of averaging, namely that the average of a function $h(\cdot)$ applied to each of a set of numbers $\{x_i : i = 1, \ldots, N\}$ is defined as

$$avg(h) \equiv \frac{1}{N} \sum_{i=1}^{N} h(x_i).$$

Note that, under this convention, the population parameter μ of equation (3.1) is the average of the attribute values x_i in the population. Similarly, the proportion $P_{\mathcal{A}}$ of equation (3.2) is the average of the indicator variables $I(x_i \in \mathcal{A})$.

Averaging and Probability of Sample Inclusion

Now, notice that the population average of the indicator that units are included in a particular sample S_n^* is

$$avg\{I(U \in S_n^*)\} = \frac{1}{N} \sum_{i=1}^{N} I(U_i \in S_n^*) = \frac{n}{N}, \tag{3.4}$$

which turns out to be equal to the probability that a particular unit U_i is included in an arbitrary sample $S_n \in \mathcal{S}_n = \{S_{n,k} : k = 1, \ldots, M\}$. This can be seen to be true because, still relying on the Laplacian concept of probability,

$$Pr(U_i \in S_n) \quad = \quad \frac{\sum_{k=1}^{M} I(U_i \in S_{n,k})}{M}$$

$$= \frac{\left[\frac{(N-1)!}{((N-1)-(n-1))!\,(n-1)!}\right]}{\left[\frac{N!}{(N-n)!\,n!}\right]}$$

$$= \frac{(N-1)!\,(N-n)!\,n!}{(N-n)!\,(n-1)!\,N!}$$

$$= \frac{n}{N}, \tag{3.5}$$

where the first step in this progression is a direct reflection of the definition of Laplacian probability, and the second step follows because the number of samples of which a particular unit U_i is a member is equal to the number of samples of size $n-1$ that can be formed from the $N-1$ population units excluding U_i, while the total number of samples of size n is $M = N!/((N-n)!\,n!)$, as defined in Chapter 3.3.1. Thus, we have that

$$avg\{I(U \in S_n^*)\} = Pr(U_i \in S_n).$$

Now, the remarkable thing is that the probability that unit U_i is included in some arbitrary sample of size n, namely $Pr(U_i \in S_n)$, and the average over population units that those units are included in a particular sample S_n^*, namely $avg\{I(U \in S_n^*)\}$, are also equal to the average over possible samples $\{S_{n,k} : k = 1, \ldots, M\}$ of the events $U_i \in S_{n,k}$. That is,

$$\begin{aligned} avg_{\mathcal{S}}\{I(U_i \in S_n)\} &= \frac{1}{M}\sum_{k=1}^{M} I(U_i \in S_{n,k}) \\ &= \frac{\left[\frac{(N-1)!}{(N-n)!\,(n-1)!}\right]}{\left[\frac{N!}{(N-n)!\,n!}\right]} \\ &= \frac{n}{N} \end{aligned} \tag{3.6}$$

So, from (3.4), (3.5), and (3.6), we have that

$$avg\{I(U \in S_n^*)\} = Pr(U_i \in S_n) = avg_{\mathcal{S}}\{I(U_i \in S_n)\}.$$

That is, the average over population units that those units are included in a particular sample is equal to the probability that a particular unit is included in an arbitrary sample, is equal to the average over possible samples that a particular unit is included in those samples.

Design Unbiasedness

Estimator of Population Mean

Denote the estimator $\hat{\mu}$ in expression (3.3) for a particular sample k as $\hat{\mu}_k$, and consider the average of $\hat{\mu}_k$ over all possible samples of size n $\{S_{n,k} : k = 1, \ldots, M\}$,

$$\begin{aligned}
avg_{\mathcal{S}}(\hat{\mu}) &= \frac{1}{M}\sum_{k=1}^{M}\hat{\mu}_k \\
&= \frac{1}{M}\sum_{k=1}^{M}\frac{1}{n}\sum_{i=1}^{N}x_i I(U_i \in S_{n,k}) \\
&= \frac{1}{n}\sum_{i=1}^{N}x_i\frac{1}{M}\sum_{k=1}^{M}I(U_i \in S_{n,k}) \\
&= \frac{1}{n}\sum_{i=1}^{N}x_i\frac{n}{N} \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i \\
&= \mu
\end{aligned}$$

Any estimator of a population quantity θ that satisfies $avg_{\mathcal{S}}(\hat{\theta}) = \theta$ when averaged over all possible samples (under a given sampling design) is called *design unbiased*. Thus, $\hat{\mu}$ as defined in expression (3.3) is design unbiased for μ under

simple random sampling.

Estimator of Population Variance

Similar to the notation $\hat{\mu}_k$, let the estimator of population variance in expression (3.3) for a given sample k be denoted $\hat{\sigma}_k^2$. We would like to show that $\hat{\sigma}^2$ is design unbiased (under simple random sampling) for the population variance σ^2, that is,

$$avg_{\mathcal{S}}(\hat{\sigma}_k^2) = \sigma^2.$$

To accomplish this requires several preliminary results. The first two of these results you will verify as an assignment, namely,

$$\frac{1}{N}\sum_{i=1}^{N} x_i^2 = \frac{N-1}{N}\sigma^2 + \mu^2, \tag{3.7}$$

$$\frac{2}{N}\sum_{1\le i<j\le N}\sum x_i x_j = \frac{(N-1)^2}{N}\sigma^2 + (N-1)(\mu^2 - \sigma^2). \tag{3.8}$$

In addition, it will prove useful to consider the average over possible samples of the squared estimator of the mean,

$$\begin{aligned}
\frac{1}{M}\sum_{k=1}^{M}\{\hat{\mu}_k\}^2 &= \frac{1}{M}\sum_{k=1}^{M}\left\{\frac{1}{n}\sum_{i=1}^{N} x_i\, I(U_i \in S_{n,k})\right\}^2 \\
&= \frac{1}{M}\sum_{k=1}^{M}\frac{1}{n^2}\left\{\sum_{i=1}^{N} x_i^2\, I(U_i \in S_{n,k}) \right. \\
&\quad \left. + \; 2\sum_{1\le i<j\le N}\sum x_i x_j\, I(U_i \in S_{n,k}) I(U_j \in S_{n,k})\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{N}\frac{1}{M}\sum_{k=1}^{M} x_i^2 I(U_i \in S_{n,k}) \\
&\quad + \frac{2}{n^2}\sum_{1\le i<j\le N}\sum\frac{1}{M}\sum_{i=1}^{M} x_i x_i I(U_i \in S_{n,k}) I(U_j \in S_{n,k})
\end{aligned}$$

$$= \frac{1}{nN}\sum_{i=1}^{N} x_i^2 + \frac{2(n-1)}{nN(N-1)}\sum\sum_{1\le i<j\le N} x_i x_j \tag{3.9}$$

Finally, it will also be useful to re-express the estimator $\hat{\sigma}_k^2$ in the following manner,

$$\begin{aligned}
(n-1)\hat{\sigma}_k^2 &= \sum_{i=1}^{N}\left(x_i - \hat{\mu}_k\right)^2 I(U_i \in S_{n,k}) \\
&= \sum_{i=1}^{N}\left(x_i^2 - 2x_i\hat{\mu}_k + \hat{\mu}_k^2\right) I(U_i \in S_{n,k}) \\
&= \sum_{i=1}^{N} x_i^2 I(U_i \in S_{n,k}) - 2\hat{\mu}_k \sum_{i=1}^{N} x_i I(U_i \in S_{n,k}) \\
&\quad + \hat{\mu}_k^2 \sum_{i=1}^{N} I(U_i \in S_{n,k}) \\
&= \sum_{i=1}^{N} x_i^2 I(U_i \in S_{n,k}) - n\hat{\mu}_k^2. \qquad (3.10)
\end{aligned}$$

We are now prepared to demonstrate that the estimator $\hat{\sigma}^2$ of expression (3.3)is design unbiased for σ^2 under simple random sampling. To do so, we begin with the result of equation (3.10),

$$\begin{aligned}
\frac{1}{M}\sum_{k=1}^{M}\hat{\sigma}_k^2 &= \frac{1}{M}\sum_{k=1}^{M}\frac{1}{n-1}\left[\sum_{i=1}^{N} x_i^2 I(U_i \in S_{n,k}) - n\hat{\mu}_k^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{N} x_i^2 \frac{1}{M}\sum_{k=1}^{M} I(U_i \in S_{n,k}) - \frac{1}{M}\sum_{k=1}^{M}\hat{\mu}_k^2\right]
\end{aligned}$$

which, on substitution from (3.9) becomes

$$\begin{aligned}
&= \frac{n}{N(n-1)}\sum_{i=1}^{N} x_i^2 - \frac{n}{n-1}\left[\frac{1}{nN}\sum_{i=1}^{N} x_i^2 + \frac{2(n-1)}{nN(N-1)}\sum\sum_{1\le i<j\le N} x_i x_j\right] \\
&= \frac{n}{n-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \frac{1}{n-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \frac{1}{N-1}\left(\frac{2}{N}\sum\sum_{1\le i<j\le N} x_i x_j\right)
\end{aligned}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N-1}\left(\frac{2}{N}\sum\sum_{1\leq i<j\leq N} x_i x_j\right)$$

Finally, using (3.8) and (3.7) we arrive at,

$$\begin{aligned}\frac{1}{M}\sum_{k=1}^{M}\hat{\sigma}_k^2 &= \left(\frac{N-1}{N}\sigma^2+\mu^2\right)\\ &\quad -\frac{1}{N-1}\left(\frac{(N-1)^2}{N}\sigma^2+(N-1)(\mu^2-\sigma^2)\right)\\ &= \frac{N-1}{N}\sigma^2+\mu^2-\frac{N-1}{N}\sigma^2-\mu^2+\sigma^2\\ &= \sigma^2. \end{aligned} \tag{3.11}$$

Thus, the estimator $\hat{\sigma}^2$ of expression (3.3) is design unbiased for σ^2 under simple random sampling.

Estimators of Population Totals and Proportions

That the basic estimator of a population total, $\hat{\tau}$ in expression (3.3) is design unbiased, under simple random sampling, for τ is immediate from design unbiasedness of $\hat{\mu}$ and the fact that $\hat{\tau} = N\hat{\mu}$. Design unbiasedness of $\hat{P}_{\mathcal{A}}$ for population proportions under simple random sampling is simple to show, and is left as an exercise.

Variances of the Estimators

Notice that the population variance in (3.1) is (essentially) the average over basic units of observation of the squared differences of an attribute (x) and the average of the attribute (μ). Using this same notion of variance, variances for the basic estimators of (3.3) are averages over basic units of observation for those estimators (i.e., samples) of the squared differences of their values with

their average value. That is, for an estimator $\hat{\theta}$ which can be computed for individual samples $\{S_{n,k} : k = 1, \ldots, M\}$, the variance of $\hat{\theta}$ is

$$var(\hat{\theta}) \equiv \frac{1}{M} \sum_{k=1}^{M} \{\hat{\theta}_k - avg_{\mathcal{S}}(\hat{\theta})\}^2,$$

where

$$avg_{\mathcal{S}}(\hat{\theta}) \equiv \frac{1}{M} \sum_{k=1}^{M} \hat{\theta}_k.$$

Using this convention, and the same techniques of derivation employed in demonstrating design unbiasedness of the basic estimators, we can easily derive variances for those estimators.

Notice that, in particular, if $\hat{\theta}$ is design unbiased for the corresponding population quantity θ,

$$\begin{aligned} var(\hat{\theta}) &= \frac{1}{M} \sum_{k=1}^{M} \left\{\hat{\theta}_k^2 - 2\hat{\theta}_k\theta + \theta^2\right\} \\ &= \frac{1}{M} \sum_{k=1}^{M} \hat{\theta}_k^2 - \theta^2. \end{aligned} \tag{3.12}$$

Population Mean

Substituting (3.7) and (3.8) directly into (3.9) yields,

$$\frac{1}{M} \sum_{k=1}^{M} \hat{\mu}_k^2 = \frac{\sigma^2}{n} \frac{N-n}{N} + \mu^2,$$

so that, design unbiasedness of $\hat{\mu}$ and application of (3.12) gives,

$$var(\hat{\mu}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N}\right). \tag{3.13}$$

Population Total and Proportions

The simple relation between population mean and total again immediately yields

$$var(\hat{\tau}) = N(N-n)\frac{\sigma^2}{n}. \tag{3.14}$$

Notice from (3.3) that the estimator of a population proportion is of the same form as the estimator of a population mean with the attributes $\{x_i : i = 1, \ldots, N\}$ replaced by the indicator that x_i is in some class $\mathcal{A}$. We could then, in deriving properties of the estimator of a proportion, simply define a new attribute $x_i' \equiv I(x_i \in \mathcal{A})$ and then use all of the results for mean estimation. This is typically what is done, with one change in notation. In the case of a numerical attribute, the variance σ^2 depends on functions of the attributes other than the average (which is μ); in particular, the function $\sum x_i^2$. In the case that we replace attributes x_i with indicator attributes x_i' we get,

$$\begin{aligned} \sigma_p^2 &\equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i' - P)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i' - 2x_i' P + P^2) \\ &= \frac{N}{N-1}P(1-P), \end{aligned} \tag{3.15}$$

which depends on the attributes x_i' only through their mean, P.

From results for the estimator $\hat{\mu}$ then, we immediately have that,

$$\hat{\sigma}_p^2 = \frac{n}{n-1}\hat{P}(1-\hat{P}), \tag{3.16}$$

and

$$var(\hat{P}) = \left(\frac{N-n}{N-1}\right)\frac{P(1-P)}{n}. \tag{3.17}$$

Estimated Variances

Estimation of the variances of the basic estimators consists of using "plug-in" estimates of the population quantities involved in expressions for variances

presented previously. That is, given that $\hat{\sigma}^2$ is an unbiased estimator of the population parameter σ^2 and $\hat{P}$ is an unbiased estimator of the population parameter P, substitution of $\hat{\sigma}^2$ for σ^2 in expressions (3.13) and (3.14) yields unbiased estimators for $var(\hat{\mu})$ and $var(\hat{\tau})$. Similarly, substitution of $\hat{P}$ for P in expression (3.15) yields an unbiased estimator for σ_p^2. Unbiasedness of these variance estimators follows immediately from the fact that the variances are all constants (functions of N and n) multiplied by the respective population parameters (σ^2 in the cases of μ and τ, and P in the case of P).

3.5 Unequal Probability Samples

Thus far, all of our probability calculations can be envisaged as applications of Laplacian probability. Recall from section 3.3.1 that, under Laplcian probability, any given sample of size n, namely $S_{n,k}$, has probability of being selected

$$Pr(S_{n,k}) = \frac{1}{M} = \frac{(N-n)!\,n!}{N!}.$$

Thus, averages of estimators $\hat{\theta}_k$ over possible samples as $(1/M)\sum_k \hat{\theta}_k$ could also be represented as,

$$avg_{\mathcal{S}}(\hat{\theta}) = \sum_{k=1}^{M} \hat{\theta}_k \, Pr(S_{n,k}). \tag{3.18}$$

Notice that expression (3.18) is in agreement with the usual definition of expectation for a random variable (which here would be $\hat{\theta}$) since, given that some particular sample will be chosen,

$$\sum_{k=1}^{M} Pr(S_{n,k}) = \sum_{k=1}^{M} \frac{1}{M} = 1.$$

We will use the usual notion of expectation for discrete random variables to extend the idea of averaging over possible samples. Note here that any $\hat{\theta}$ computed from possible samples in a finite population of *fixed* attributes x_i *must*

have a finite set of discrete values. Another way to see this is that, interpreting expression (3.18) as the definition of expectation for discrete random variables, the quantities $\{\hat{\theta}_k : k = 1, \ldots, M\}$ are *possible values* of the random variable, not the random variable itself. Now, also note that the key operation in derivation of properties of estimators under simple random sampling was interchanging summations over possible samples (sum over k from 1 to M) and over population units (sum over i from 1 to N). This was key because, after interchanging summations, summations over samples became constants and then entire expressions reduced to sums over only population units; see, for example, the demonstration of design unbiasedness for the basic estimator of population mean in section 3.4.2.

If we have a set of possible samples of size n, $\mathcal{S}_n = \{S_{n,1}, \ldots, S_{n,M}\}$ such that not each of these samples has the same probability of being chosen, there are a number of modifications we must make to our basic development.

1. First, we must determine what is meant that not each sample has the same probability of being chosen, and how those probabilities for various possible samples are to be computed.

2. Second, we replace simple averaging over possible samples with expectation for discrete random variables (which is weighted averaging) to determine properties of estimators.

3. Finally, we must determine whether any relations exist between probabilities of samples and quantities attached to population units that allows reduction of sums over possible samples to sums over population units.

3.5.1 Appropriate Probability Concept

The first of the issues listed above, determination of an appropriate concept of probability within which to consider sets of samples that may not all be equally likely, is more difficult than it may at first appear. For one thing, we need to distinguish between designs in which not all *possible samples* have the same chance of being selected, and designs in which all possible samples are equally likely but not all *population units* have the same chance of being selected for the sample. The first of these possibilities (not all samples equally likely) is probably much more rare than the second (equally likely samples constructed from unequal chances of unit selection) but is at least hypothetically possible. This situation also presents the greater difficulty for determining an appropriate probability context. While we will not go into this topic in detail, the following comments seem pertinent:

1. It would seem difficult to envisage a situation in which one might construct unequally likely samples that could not be better handled by constructing equally likely samples from selection of units with unequal probability.

2. It does not seem possible to attach any material concept of probability to the selection of unequally likely samples. At fist glance it appears that one should be able to list all samples to be considered (all possible samples) and then simply add "copies" of samples to be given higher probability to the list, finally arriving at a set of possible outcomes which could be considered equally likely (and hence to which Laplacian probability would still apply). The difficulty here is that only a discrete set of varying probabilities would be possible. For example, if I start with 5 possible samples, of which I would like to increase the probability of

> selecting sample S_3, I could make the probability of choosing sample S_3 have values of 2/6, 3/7, 4/8, 5/9, etc., but nothing else. It would appear that samples of unequal chances of being selected can be given no general probability framework without recourse to hypothetical limiting frequency concepts. And, what would this do, for example, to the interpretation of an estimated population proportion in a finite population?

For samples constructed in such a way so that all possible samples are equally likely of selection, but for which the chances that individual population units are included differ, the situation becomes less complex. In this situation we are still able to apply the concept of Laplacian probability because the samples form a set of equally likely basic outcomes. The probability that individual units are included in the sample may then be computed as the probability of events under this concept of probability. Such samples are sometimes called *unequal probability samples* but, because of the discussion presented here and what appears in the next subsection, I prefer an often-used alternative and refer to such samples as samples that result from *restricted randomization.*

3.5.2 Obtaining Samples Through the Use of Restricted Randomization

Obtaining random samples of equal probability, but with unequal probabilities of selection for individual population units, is based on methods for obtaining simple random samples, as eluded to at the beginning of Chapter 3.3. Two related sampling designs that are common will be used to illustrate this point, *stratified random sampling* and *multistage sampling.*

Stratified Sampling

In a stratified sampling design the overall population is divided into subsets or groups called *strata*, based (it is presumed) on some external (or prior) knowledge that the groups differ in some systematic manner relative to the attribute of interest. For example, it may be known (or generally accepted or believed) that age groups in the voting population will demonstrate a systematic difference in support for a ballot initiative on decriminalization of marijuana (e.g., old voters more opposed, young voters more opposed, voters of middle age more in favor). In essence, we delineate a partition of the overall population into subpopulations. For an overall population consisting of units $\{U_i : i = 1, \ldots, N\}$, let such a partition be denoted as $\{U_{h,i} : h = 1, \ldots, H, i = 1, \ldots, N\}$. In this notation, h indexes the strata while i continues to index population units from 1 to N; an alternative notation would be to allow i to run from 1 to N_h within each stratum. The sampling frame is then also partitioned into frames for each stratum according to the same system of indices.

Example 3.1 (cont.)

In the example of the West Coast groundfish fishery introduced at the beginning of Chapter 3.1, it is known that vessels tend to fish in either the northern portion of the fishery or the southern portion. A part of the reason for this is that NMFS regulates the fishery in two "zones", north of $40°10'$ latitude, and south of $40°10'$ latitude, which separates the very northern coast of California, Oregon, and Washington from the bulk of California. Also, fishermen operating in these different zones catch a somewhat different mix of species and thus use different gear (net types, mesh sizes, etc.). If we believe these difference likely affect the way that vessels produce total catch values, we might decide

to stratify the populations by zone and the sampling frame becomes,

Vessel	Trip	Haul	Zone	Unit Index	Total Catch (lbs)
64	78	1	S	1	x_1
64	78	2	S	2	x_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
131	51	5	N	N	x_N

Since the ordering of units in a sampling frame is arbitrary, it is not necessary that all of the units in the same stratum have consecutive unit indices (although it is natural to arrange things so that this does occur). A desired overall sample size n is then also divided into sample sizes for each stratum $\{n_h : h = 1, \ldots, H\}$, such that $n = \sum_h n_h$, and a simple random sample of size n_h is selected from each stratum; $h = 1, \ldots, H$.

Stratification basically consists of dividing one larger population into a set of smaller populations and applying all of the methods of simple random sampling to each of the small populations. Estimates for the total population are then obtained by simply summing across each of the small populations or strata (e.g., Thompson, 1992). We will not consider the details further in lecture, although we might have an example in lab.

Multistage Sampling

Multistage sampling designs are similar to stratified designs, except that, rather than taking a sample from each group (stratum) we first select a sample of groups at random from which to obtain observations. This is called

multistage sampling because it can be extended to additional nested levels of grouping, but we will deal only with two-stage designs.

It is typical in this setting to refer to the groups as *primary sampling units* rather than strata. A fairly typical indexing system for multistage designs is to now take N as the number of primary sampling units in the population, n as the number of sampled primary units, M_i as the number of fundamental population units in the ith primary unit, and m_i as the number of population units in the ith primary unit that are sampled (e.g., Thompson, 1992, Chapter 13). The population units are then indexed as $\{U_{i,j} : i = 1, \ldots, N;\ j = 1, \ldots, M_i\}$. We will not make use of this, or any other alternative indexing system, but be aware that you are likely to encounter such systems in the literature.

3.5.3 Inclusion Probabilities and Linear Estimators

We are now prepared to address the second and third items needed to extend the fundamental ideas of the survey sampling approach beyond simple random sampling as identified in the list at the beginning of Chapter 3.5. In particular, we want to make use of mathematical expectation to represent averaging over samples (this was item 2), and to connect such expectation over samples with expectation over population units (this was item 3).

Linear Estimators

Given attributes of population units $\{x_i : i = 1 \ldots, N\}$ and a population parameter θ to be estimated, a *linear* estimator of θ for a given sample $S_{n,k}$ has the form

$$\hat{\theta}_k = \sum_{i=1}^{N} \beta_i x_i I(x_i \in S_{n,k}), \tag{3.19}$$

for a set of pre-specified (i.e., fixed) values $\{\beta_i : i = 1 \ldots, N\}$. Linear estimators play a central role in survey sampling methods, and many estimators may be written in linear form. For example, the basic estimator of the mean under simple random sampling is a linear estimator with $\beta_i \equiv (1/n)$ for a sample of size n. Similarly, the basic estimator of the total under simple random sampling is a linear estimator with $\beta_i \equiv N/n$.

More Advanced Note: One of the reasons for saying linear estimators are central to survey sampling methods is that, not only are many of the standard estimators linear, but variances for nonlinear estimators are often derived by forming a Taylor series expansion of the nonlinear estimator and then deriving the variance for the linear approximation (e.g., Wolter, 1985).
Consider, for illustration, estimation of a population total τ with a linear estimator

$$\hat{\tau}_k = \sum_{i=1}^{N} \beta_i x_i I(U_i \in S_{n,k}). \tag{3.20}$$

Now, suppose we have a set of possible samples of size n $\mathcal{S}_n \equiv \{S_{n,1}, \ldots, S_{n,M}\}$. Considering $\hat{\tau}$ as a random variable, with possible values $\{\hat{\tau}_1, \ldots, \hat{\tau}_M\}$ in one-to-one correspondence with the possible samples, the expected value of $\hat{\tau}$ is

$$E\{\hat{\tau}\} = \sum_{k=1}^{M} \hat{\tau}_k \, Pr(S_{n,k}). \tag{3.21}$$

Expression (3.21) applies to any set of possible samples, even if they are not equally likely, but recall that the interpretation of this type of expectation can become murky if the samples are not equally likely.

Now, combining (3.20) and (3.21),

$$E\{\hat{\tau}\} \quad = \quad \sum_{k=1}^{M} \sum_{i=1}^{N} \beta_i x_i I(U_i \in S_{n,k}) \, Pr(S_{n,k})$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \beta_i x_i \sum_{k=1}^{M} I(U_i \in S_{n,k})\, Pr(S_{n,k}) \\
&= \sum_{i=1}^{N} \beta_i x_i E\{I(U_i \in S_n^*)\} \\
&= \sum_{i=1}^{N} \beta_i x_i Pr(U_i \in S_n^*), \qquad (3.22)
\end{aligned}
$$

where, as before, S_n^* denotes the sample of size n to be (or which will be) selected. Note that, if the possible samples are all equally likely, the second line of (3.22) becomes the number of samples of which U_i is a member divided by the number of (equally likely) possible samples, and we may go directly to the concluding line based on Laplacian probability.

Now, (3.22) implies that $\hat{\tau}$ will be design unbiased for τ if

$$\beta_i \equiv \frac{1}{Pr(U_i \in S_n^*)}; \quad i = 1, \ldots, N.$$

Inclusion Probabilities and the Horvitz-Thompson Estimator

Building on the result of expression (3.22) define the *inclusion probability* for population unit U_i as

$$\pi_i \equiv Pr(U_i \in S_n^*); \quad i = 1, \ldots, N. \qquad (3.23)$$

The linear estimator eluded to just above, with β_i given by the reciprocal of the probability unit i is included in the sample is called the *Horvitz-Thompson* estimator of a population total,

$$\hat{\tau} = \sum_{i=1}^{N} x_i I(U_i \in S_n^*) \frac{1}{\pi_i} = \sum_{i \in S_n^*} \frac{x_i}{\pi_i}, \qquad (3.24)$$

and expression (3.21) implies that this estimator will be design unbiased for *any* sampling design for which the inclusion probabilities can be computed for each unit in the population (since we do not know which particular units will

in fact end up in the selected sample S_n^*.)
Now define what are called the *second order* inclusion probabilities for $i, j = 1, \ldots, N;\ i \neq j$,

$$\pi_{i,j} \equiv Pr\{(U_i \in S_n^*) \cap (U_j \in S_n^*)\}. \tag{3.25}$$

In sampling without replacement (which is all we are considering in this class) the events $U_i \in s_n^*$ and $U_j \in S_n^*$ will not be independent, so that the $\pi_{i,j}$ are not merely products of π_i and π_j (although this *will* be true, for example, for units belonging to different strata in a stratified design with independent selection for each stratum). Consider, for example, simple random sampling. The probability that two units U_i and U_j are both included in the sample is $\{n(n-1)\}/\{N(N-1)\}$, which we have actually seen before in expression (3.9). Second-order inclusion probabilities are important in deriving variances for estimators that can be placed into the Horvitz-Thompson framework, which includes many of the typical estimators used in the survey sampling approach.

3.5.4 The Overall Generalization

The generalization of simple random sampling that has now been achieved may be outlined as follows.

1. For a finite population of units $\{U_i : i = 1, \ldots, N\}$, define the binary random variables $Z_i \equiv I(U_i \in S_n^*)$.

2. Define inclusion probabilities as
   $\pi_i \equiv Pr(Z_i = 1)$ and

   $\pi_{i,j} \equiv Pr\{(Z_i = 1) \cap (Z_j = 1)\}$.

3. For a given population quantity θ consider linear estimators of the form

$$\begin{aligned}\hat{\theta} &= \sum_{i=1}^{N} \beta_i(\pi_i) x_i I(U_i \in S_n^*) \\ &= \sum_{i=1}^{N} \beta_i(\pi_i) x_i Z_i.\end{aligned}$$

Notice that we have taken the values $\{\beta_i : i = 1, \ldots, N\}$ to be, in general, functions of the inclusion probabilities $\{\pi_i : i = 1, \ldots, N\}$, but they may or may not depend explicitly on these inclusion probabilities.

We may derive properties of the linear estimators considered in this generalization as,

$$E\{\hat{\theta}\} = \sum_{i=1}^{N} \beta_i(\pi_i) x_i E\{Z_i\}, \tag{3.26}$$

and,

$$var\{\hat{\theta}\} = \sum_{i=1}^{N} \beta_i^2(\pi_i) x_i^2 var\{Z_i\} + \sum\sum_{1 \le i < j \le N} x_i x_j cov\{Z_i, Z_j\}. \tag{3.27}$$

Now, under this formulation of estimators, the first and second moments of the Z_i are,

$$\begin{aligned}E\{Z_i\} &= \pi_i, \\ var\{Z_i\} &= \pi_i - \pi_i^2 = \pi_i(1 - \pi_i), \\ cov\{Z_i, Z_j\} &= E\{Z_i Z_j\} - \pi_i \pi_j = \pi_{i,j} - \pi_i \pi_j.\end{aligned} \tag{3.28}$$

As an example, consider the Horvitz-Thompson estimator $\hat{\tau}$ of expression (3.24), for which $\beta_i(\pi_i) = 1/\pi_i$. Then,

$$\begin{aligned}var\{\hat{\tau}\} &= var\left\{\sum_{i=1}^{N} \frac{x_i}{\pi_i} Z_i\right\} \\ &= \sum_{i=1}^{N} \frac{x_i^2}{\pi_i^2} var\{Z_i\} + 2 \sum\sum_{1 \le i < j \le N} \frac{x_i x_j}{\pi_i \pi_j} cov\{Z_i, Z_j\}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \frac{x_i^2}{\pi_i^2}\pi_i(1-\pi_i) + 2\sum\sum_{1\le i<j\le N} \frac{x_i x_j}{\pi_i\pi_j}(\pi_{i,j}-\pi_i\pi_j) \\
&= \sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i}\right) x_i^2 + 2\sum\sum_{1\le i<j\le N} \left(\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\right) x_i x_j.
\end{aligned}
\tag{3.29}
$$

Now, for estimation of variances developed from expression (3.27), such as given in (3.29) for the standard Horvitz-Thompson estimator of population total, we need "plug-in" estimators that are unbiased for portions of the expression, such that the result is an unbiased estimator for the overall variance. This is not always possible. Properties of expectations tell us that this possible when the variance is a linear combination of quantities that can be estimated in an unbiased manner.

Continuing with the example of the Horvitz-Thompson estimator of population total given in (3.24), notice that the variance is composed of two additive terms, each of which consists of a sum over all population units. Notice that, for any simple function of attributes in the population $h(x_i)$, an unbiased estimator of the linear combination

$$\theta \equiv \sum_{i=1}^{N} a_i h(x_i)$$

is given by

$$
\begin{aligned}
\hat{\theta} &= \sum_{i\in S_n^*} a_i \frac{h(x_i)}{\pi_i} \\
&= \sum_{i=1}^{N} a_i h(x_i) I(U_i \in S_n^*) \\
&= \sum_{i=1}^{N} a_i h(x_i) Z_i.
\end{aligned}
$$

Unbiasedness of $\hat{\theta}$ follows from,

$$E\{\hat{\theta}\} = \sum_{i=1}^{N} a_i \frac{h(x_i)}{\pi_i} E\{Z_i\}$$

$$= \sum_{i=1}^{N} a_i \frac{h(x_i)}{\pi_i}\pi_i$$
$$= \sum_{i=1}^{N} a_i h(x_i).$$

Thus, in the variance of expression (3.29), an unbiased estimator of the first additive term is

$$\sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i}\right)\frac{x_i^2}{\pi_i} I(U_i \in S_n^*) = \sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i^2}\right) x_i^2 Z_i,$$

while an unbiased estimator of the second term is,

$$2\sum_{1\le i<j\le N}\sum\left(\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{x_i x_j}{\pi_{i,j}} I\{(U_i \in S_n^*) \cap I(U_j \in S_n^*)\}$$

$$= 2\sum_{1\le i<j\le N}\sum\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{i,j}}\right) x_i x_j\, Z_i Z_j.$$

Substituting these expressions into (3.29) gives an unbiased estimator of that variance,

$$\hat{var}(\hat{\tau}) = \sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i^2}\right) x_i^2 Z_i + 2\sum_{1\le i<j\le N}\sum\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{i,j}}\right) x_i x_j\, Z_i Z_j. \qquad (3.30)$$

A number of important estimators may be formulated as functions of Horvitz-Thompson estimators of population totals. For example, an estimator of the population mean is,

$$\hat{\mu} = \frac{\hat{\tau}}{N}.$$

An estimator of the ratio of two attributes in the population is

$$\hat{\left(\frac{\tau_x}{\tau_y}\right)} = \frac{\hat{\tau}_x}{\hat{\tau}_y}.$$

Thus, the use of inclusion probabilities forms an important methodology in the development of estimators for finite population problems.

Example 3.2

As an illustration what is now possible, consider the following (hypothetical) example. A federal resource agency has completed a wolf re-introduction program in a major national park (e.g., Yellowstone National Park in the western United States). Several members of congress from districts within which the park lies are concerned with the economic impact this effort has had on sheep ranchers near the park due to depredations from wolves that wander out of the park (wolves are smart, but they can't read boundary signs). You work as a statistician for the resource agency involved, and are asked to design a survey of sheep ranchers in the potentially affected areas to determine the total economic loss to the local sheep industry due to wolf depredations. After much discussion, it is decided that individual sheep ranching operations will constitute the basic units in the population, and that the total dollar loss due to wolf depredations will be the attribute of interest. Assuming that there are N ranches in the region of interest, and with x_i representing dollar loss for ranch i, the quantity to be estimated is

$$\tau = \sum_{i=1}^{N} x_i.$$

Now, suppose that licensing of sheep ranching operations is administered through county offices (I don't know if this is true or not, but suppose it is). It is not difficult to obtain the number of licenses issued from each office in the region of interest, but no records are kept of economic loss due to wolf depredations. Obtaining this information requires visiting the individual ranches and going over detailed records. This would be, of course, impossible to do for every ranch in the region. In addition, there are 32 counties in the region and obtaining the locations, name of the responsible owner or man-

ager, contact information, etc. for the individual ranches in counties is a time consuming activity. It is decided to use a multi-stage sampling design with counties as the primary sampling units and ranches as the secondary sampling units. The sampling plan may be described as:

1. A simple random sample of 7 of the 32 counties will be selected.

2. From each of the sampled counties, a simple random sample of 10 individual ranches will be selected to be visited.

3. For each ranch visited, the attribute of interest will be determined.

For our purposes in this example, assume there are no refusals to cooperate from the ranchers, and the attribute of interest can be determined without question (no lying, false records, etc.).

Everything from here on is largely a matter of notation and keeping track of the units and quantities involved in an organized fashion. Denote the primary sampling units as $\{U_h^{(1)} : h = 1, \ldots H\}$ from which a simple random sample of $n^{(1)}$ is to be selected; the superscript (1) denotes *first stage sample*, and in this example we would have $H = 32$ possible primary sampling units for the stage one samples and $n^{(1)} = 7$. Let $S^{(1)}$ denote the first stage sample selected, let $\{n_h : h = 1, \ldots, H\}$ denote the number of population units to be sampled from each primary sampling unit if those units are selected in the stage one sample, and denote the number of population units in each of the primary sampling units as $\{N_h : h = 1, \ldots, H\}$. Then $N = \sum_{h=1}^{H} N_h$ is the population size and $n = \sum_{h=1}^{H} n_h I(U_h^{(1)} \in S^{(1)})$ is the total sample size. Finally, as before, let S_n^* denote the sample of population units selected from the two-stage sampling design. For estimation of the population total, all that is necessary is to

determine the inclusion probabilities for individual population units. Let

$$\begin{aligned}
\pi_h^{(1)} &= Pr(U_h^{(1)} \in S^{(1)}), \\
\pi_{h,k}^{(1)} &= Pr\{(U_h^{(1)} \in S^{(1)}) \cap (U_k^{(1)} \in S^{(1)})\}, \\
\pi_{i|h}^{(2)} &= Pr\{(U_i \in S_n^*)|(U_h^{(1)} \in S^{(1)})\}; \quad U_i \in U_h^{(1)}, \\
\pi_{i,j|h}^{(2)} &= Pr\{(U_i \in S_n^*) \cap (U_j \in S_n^*)|U_h^{(1)} \in S^{(1)}\}; \quad U_i, U_j \in U_h^{(1)}.
\end{aligned}$$

Here, we would have

$$\begin{aligned}
\pi_h^{(1)} &= \frac{n^{(1)}}{H}, \\
\pi_{h,k}^{(1)} &= \frac{n^{(1)}(n^{(1)}-1)}{H(H-1)}, \\
\pi_{i|h}^{(2)} &= \frac{n_h}{N_h}, \\
\pi_{i,j|h}^{(2)} &= \frac{n_h(n_h-1)}{N_h(N_h-1)}.
\end{aligned}$$

From these we can calculate the inclusion probabilities for individual population units as,

$$\begin{aligned}
\pi_i &= \pi_h^{(1)}\,\pi_{i|h}^{(2)}, \\
\pi_{i,j} &= \pi_h^{(1)}\pi_{i,j|h}^{(2)}; \quad U_i, U_j \in U_h^{(1)} \\
\pi_{i,j} &= \pi_{h,k}^{(1)}\pi_{i|h}^{(2)}\pi_{j|k}^{(2)}; \quad U_i \in U_h^{(1)},\, U_j \in U_k^{(1)}.
\end{aligned}$$

Relying on the particular forms of these probabilities for simple random sampling at each stage given previously, we arrive at the following inclusion probabilities for individual units in the population:

$$\begin{aligned}
\pi_i &= \frac{n^{(1)} n_h}{H N_h}; \quad U_i \in U_h^{(1)}, \\
\pi_{i,j} &= \frac{n^{(1)} n_h (n_h - 1)}{H N_h (N_h - 1)}; \quad U_i, U_j \in U_h^{(1)} \\
\pi_{i,j} &= \frac{n^{(1)} (n^{(1)} - 1) n_h n_k}{H(H-1) N_h N_k}; \quad U_i \in U_h^{(1)}, \, U_j \in U_k^{(1)}.
\end{aligned}$$

With these inclusion probabilities in hand, the population total, its variance, and estimated variance are available from equations (3.24), (3.29) and (3.30), respectively.

Now, one might question the use of going through all of this if what we have arrived at is a standard design and associated estimators, which are available in what is arguably simpler form in standard texts (e.g., Thompson (1992), Chapter 13). Still within the context of this example, consider the following scenario:

Congressman Higgenbottom (one of the congressmen from the districts in which the park is located) contacts your agency with the names of two particular ranchers in his district who would very much like to be included in the survey (they may, for example, be major campaign contributors for the congressman). The good Congressman expresses the opinion that he would be "mightily disappointed" if these two fine citizens, who are willing to go out of their way to help the government in its efforts, could not contribute information to the "fancy survey" planned by the agency. Given that Congressman Higgenbottom is a member of the appropriations committee that passes budgets for your agency, your supervisor translates the message to you as follows. "Rancher A_1 and rancher A_2 will be included as sampled units in the survey.

Supposing that the sample list has not yet been drawn, how do you maintain the scientific integrity of the survey while at the same time managing to retain your job?

3.6 Extensions to Ill-Defined Populations

Our entire development of the survey sampling approach has assumed that there is available a population of discrete physical units which are not subject to question, that is, are not ambiguous in definition, and which may be individually identified (see Chapter 2). In addition, we have assumed that each unit has one or more attributes which are characteristic of those units. But, sampling methods are commonly employed in situations for which one or more of these assumptions are not readily verified. There are three concerns in such applications, two of which are fundamentally more troublesome than the other. Specifically, the survey sampling approach is often applied in the following situations.

1. Not all units in the population can be uniquely identified, and perhaps only those units selected for observation can in fact be uniquely identified.

2. The population of interest does not consist of naturally occurring discrete units. Rather, units must be defined in some arbitrary manner.

3. The attribute associated with each population unit is really an *estimate* rather than an unambiguous characteristic of the unit.

The first of these difficulties is actually more easily (not easily, but more easily than the others) overcome. Statistics 521 and Statistics 621 are courses that cover this topic. I believe that the second and third difficulties present more

fundamental (inter-related) problems for the survey sampling approach. The second potential difficulty, arbitrary definition of population units, is not insurmountable in and of itself, as noted in Chapter 2.2. That is, given any specific delineation of population units, Laplacian probability may be applied under that definition, and everything developed to this point applies *conditionally* on the population unit definition. However, the issue of unit definition is also connected with the concept of an attribute (fixed, immutable characteristic of a unit), and it is the third potential difficulty that constitutes the crux of the matter.

Consider, for example, the problem of determining (estimating) the proportion of acres (or hectares) planted in soybeans in Iowa that have "serious" infections of soybean cyst nematode, a pest that attacks the roots of soybeans and decreases yield. First, how does one unambiguously define units that consist of an acre of planted soybeans? You draw your lines, but my lines might be a translation consisting of a shift 1000 meters south and 500 meters east. Is it possible that such a shift in unit definition changes the value of the attribute of concern for at least some units in the population? Is it obvious that such effects should "average out" over different definitions of population units (i.e., increases are about the same as decreases?). Even without the potential problem of unit definition, how is a "serious" infection of soybean cyst nematodes defined? And, how is it "observed" for a given population unit? Recall that, in all of the statistical development for the sampling approach, the attributes $\{x_i : i = 1, \ldots, N\}$ have been taken as fixed values for population units $\{U_i : i = 1, \ldots, N\}$.

These concerns do not invalidate the survey sampling approach to many problems, but they do indicate that not every problem can be forced into the confines necessary for the approach to be applicable. It is difficult, for example,

to see how the problem of non-characteristic attributes for population units can be overcome without recourse to the notion of random variables associated with the observation or measurement process.

3.7 Interval Estimation

To this point, we have said nothing about inference from sampling finite populations. First, we indicate the standard approach. Given an estimator $\hat{\theta}$ of a population quantity θ, a derived variance $var(\hat{\theta})$ and an estimate of that variance $\hat{var}(\hat{\theta})$, a typical approach is to form an $(1-\alpha)$ confidence interval for θ as either,

$$\hat{\theta} \pm t_{1-(\alpha/2);n-1} \left\{ \hat{var}(\hat{\theta}) \right\}^{1/2}, \tag{3.31}$$

where $t_{1-(\alpha/2);n-1}$ is the $1-(\alpha/2)$ quantile of a t-distribution with $n-1$ degrees of freedom, or as,

$$\hat{\theta} \pm z_{1-(\alpha/2)} \left\{ \hat{var}(\hat{\theta}) \right\}^{1/2}, \tag{3.32}$$

where $z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of a standard normal distribution.

Now, (3.31) comes directly from the elementary result for sample means of normally distributed random variables. Its applicability seems to rest on what might be called the *ubiquitous statistical appeal to normality*, stated quite simply by Cochran (1977) as

> "It is usually assumed that the estimates $\bar{y}$ and $\hat{Y}$ are normally distributed about the corresponding population values."

Chapter 2.15 in Cochran (1977) gives some discussion of the normality assumption, and it is clear support for the assumption rests almost entirely on asymptotic arguments for means of random variables (i.e., central limit theorem results). In particular, a finite population central limit theorem for the

basic estimator of the mean ($\hat{\mu}$) under simple random sampling requires that $N \to \infty$ and $n \to \infty$ such that $n/N \to r$ for some $r < 1$. Thus, while some support exists for the use of (3.32), (3.31) has no theoretical basis. It is important to note here, that the random variables for which means are calculated are estimators such as $\hat{\mu}$, not the population attributes $\{x_i : i = 1, \ldots, N\}$, which are considered fixed.

In a more complex setting involving averages of estimators from "random groups" selected from the population, Wolter (1985) comments that

> "Notwithstanding these failures of the stated assumptions [normality assumptions] Theorem 2.2 [standardization of normal random variables to t-distributions] has historically formed the basis for inference in complex surveys, largely because of the various asymptotic results"
> (Wolter, 1985, p. 23).

Chapter 4

The Experimental Approach

References: Kempthorne and Folks (1971), Edgington (1980), Milliken and Johnson (1992), Good (1994).

The second approach we will cover that relies on randomization to bring probability to bear on a problem is what will be called the *Experimental Approach*. Recall the concept of a physically existent population discussed in Chapter 2. Assume that such a population exists in a given problem, and that a simple random sample of units from the population has been obtained; this assumption is rarely met in the experimental approach, but the approach is most easily described under this assumption. Our objective is to determine the *effect* of a given treatment or set of treatments on a response of interest (i.e., an attribute) among units of the population. For most of the discussion of this chapter we will assume not only that we have obtained a random sample of units in a population, but also that those units correspond to both experimental units and sampling units. But see Chapter 2.3 for a discussion of the distinction and the importance of connecting responses with experimental

units. In its most basic form, the experimental approach shares with sampling and survey methods two characteristics:

1. The existence of a well-define population of discrete objects (e.g., humans, pigs, plants, cows, horses, etc.). We will extend this notion in a later section to what were called *constructed* populations in Chapter 2.2, but for now we will retain the concept of an existent population.

2. Responses, similar to what were called attributes of population units in Chapter 3, are considered characteristics of population units, not associated with random variables. See Chapter 2.3 for a more extensive discussion of attributes and responses.

Recall that responses are allowed to be influenced by external factors within the time frame of a study, as illustrated with the discussion of influences of exercise and diet on cholesterol level in Chapter 2.2. Thus, the term *responses* refers to attributes under a given set of pertinent conditions and external influences. It is precisely the effect of certain of these external influences (i.e., treatments) that the experimental approach is designed to assess.

4.1 Scientific Abstraction and Experiments

The word *abstract* can have a number of different meanings. We often use abstract in its meaning of *abstruse*, or difficult to understand. But a fundamental meaning of abstract is to separate, to express a quality apart from an object, or to consider a part as divorced from the whole. This is, in many ways, the essence of scientific experimentation. Consider an experiment in which green leaves are brought into the laboratory and it is discovered that, in the presence of radiant energy, certain cells (chloroplasts) can produce carbohydrates

from water and carbon monoxide. Has this experiment explained how plants grow? Of course not, but it has examined a particular aspect of that problem, divorced from the whole. In the experimental approach, the *whole* consists of all of the external conditions to which a population unit is subject. The *part* is to examine fluctuations in a small number of those conditions while holding all others constant. The key element is *control* of all relevant factors. That is, the external conditions, or factors, to which population units are subject must be determined by the investigator, or under the control of the investigator. This brings us to an important point, that experimentation involves *invasive actions* on the part of the investigator (i.e., the assignment of treatment groups).

Now, it is physically impossible to exercise perfect control over all factors that may influence a response of interest among population units. This would require absolute control over, for example, both genetic and all environmental conditions in a study on biological organisms. Differences that exist among population units that are not subject to control by an investigator must be considered *inherent* differences among units. Inherent differences produce differences in responses and, hence, a certain level of uncertainty in response values among population units (enter statistics). This is more than a trivial matter, as the assessment of the effect of a treatment depends on quantification of the amount of variability among units subject to the same conditions (intra-treatment or inherent variability) to the amount of variability among units subject to different conditions (inter-treatment variability).

4.2 The Nested Syllogism of Experimentation

The experimental approach also has close ties to fundamental logical arguments known as syllogisms. A syllogism is a valid logical argument concerning

propositions and *conclusions* such as the following, which is known as a disjunctive syllogism (disjunctive because of the first proposition):

Either A or B.
Not A.
Therefore B.

Consider the following propositions within the context of the disjunctive syllogism immediately above.

Proposition A: All parrots are green.
Proposition B: Some parrots are not green.

If a red parrot is observed, then we have verified "Not A" in the syllogism, and the conclusion B that some parrots are not green has been proved.

Now consider the following syllogism, which is an argument known as *modus tollens* in logic:

If A then C.
Not C.
Therefore, not A.

As a side note, you should recognize the similarity of these logical syllogisms to some methods of mathematical proof. Finally, a nesting of these two valid syllogisms yields the following nested syllogism of experimentation:

Either A or B.
If A then C.
Not C.
Therefore not A.
Therefore B.

Lines 1, 4, and 5 constitute a valid disjunctive syllogism. Lines 2, 3 , and 4 constitute modus tollens. In the traditional experimental approach, we connect the propositions with a disjunction between "chance" and "design" as follows:

A: chance alone is in effect
B: design (a systematic force) is in effect
C: observable implication of chance alone

What happens statistically is that C is replaced with observable results "expected" if chance alone is in operation. The statement "Not C" in the nested syllogism is then replaced with "exceedingly low probability of C", and the statement "Therefore not A" is replaced with "either an exceptionally rare event has occurred, or not A". This should all look quite familiar to you, in the following form:

H_0: $\mu = \mu_0$ or $H_1 : \mu \neq \mu_0$ (Either A or B).
If H_0 then t^* has a t-distribution with $n-1$ degrees of freedom (If A then C).
$Pr(t_{n-1} \geq t^*) < \alpha$ (C has low probability).
Reject H_0 (Therefore not A).
Accept H_1 (Therefore B).

4.3 Randomized Treatment Assignment

The demonstration immediately above pulls us too quickly into the use of theoretical probability distributions as an approximation to randomization probability, but is instructive for understanding the progression of the experimental argument. We need to retreat, however, to the manner in which probability is introduced in the experimental approach, without resort to t-distributions and the like.

Consider a set of n population units that are to be divided into groups exposed to k different sets of external factors (i.e., treatments). All other conditions (subject to inherent differences among units) are to be controlled at the same levels. Randomized treatment assignment is somewhat analogous to simple random sampling in that each possible assignment of the k treatments to the n units should be equally likely. The number of ways that n units can be assigned to k treatments of sizes $n_1, \ldots, n_k$ is

$$\frac{n!}{n_1!\, n_2! \ldots n_k!}.$$

Example 4.1

Consider a small experiment in which 5 population units are to be assigned to 2 treatments of sizes 3 and 2. Denote the population units as U_1, U_2, U_3, U_4, and U_5, and the treatments as T_1 and T_2. The possible treatment assignments are:

Assignment	Treatment	Units
1	1	U_1, U_2, U_3
	2	U_4, U_5
2	1	U_1, U_2, U_4
	2	U_3, U_5
3	1	U_1, U_2, U_5
	2	U_3, U_4
4	1	U_1, U_3, U_4
	2	U_2, U_5
5	1	U_1, U_3, U_5
	2	U_2, U_4
6	1	U_1, U_4, U_5
	2	U_2, U_3
7	1	U_2, U_3, U_4
	2	U_1, U_5
8	1	U_2, U_3, U_5
	2	U_1, U_4
9	1	U_2, U_4, U_5
	2	U_1, U_3
10	1	U_3, U_4, U_5
	2	U_1, U_2

In a manner similar to the group versus sequential selection procedures in sampling, we may either enumerate all possible arrangements (as in the previous table) and choose one at random, or choose n_1 units sequentially for assignment to treatment 1, then choose n_2 of the remaining units sequentially for assignment to treatment 2, and so forth, until the remaining n_k units

are assigned to treatment k. In practice, only one of the possible treatment assignments will be used. We will refer to that particular arrangement as the "actual assignment", and all others as "possible assignments".

In the above presentation, we have assumed that *experimental units* are the same as *sampling units* and that these correspond to the fundamental units of a defined population. It is worth reiterating the message of Chapter 2.3 that these need not be the same, in which case the "units" of concern are experimental units (why?).

4.4 Quantifying Differences Among Treatments

Given a particular treatment assignment, we need to quantify the difference among units subject to the various treatments. We are familiar with typical test statistics such as the $t-$statistic and the $F-$statistic. These quantities may certainly be used to quantify the differences among treatment groups, but they are by no means necessary; recall that we are not relying on random variables or theoretical probability distributions at this point.

Any potential test statistics that are perfectly correlated in rank are called *equivalent test statistics*. For example, consider a situation with two treatment groups, denoted T_1 and T_2. Let the observed responses from units assigned to T_1 be denoted as $\{x_{1,j} : j = 1, \ldots, n_1\}$ and those from units assigned to treatment T_2 be denoted $\{x_{2,j} : j = 1, \ldots, n_2\}$. Let

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}; \quad i = 1, 2$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2 ; \quad i = 1, 2,$$

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}.$$

It can be shown that the typical $t-$statistic

$$t^* \equiv \frac{\bar{x}_1 - \bar{x}_2}{\left\{(s_p^2/n_1) + (s_p^2/n_2)\right\}^{1/2}},$$

is perfectly monotonically correlated in absolute value with the simpler statistic

$$D^* = |\bar{x}_1 - \bar{x}_2|. \tag{4.1}$$

Note that this does not depend on the denominator of t^* being invariant to data permutations (it is, in fact, not). In a similar way, it can be shown that the traditional $F-$statistic in a one-way analysis of variance is equivalent to the simpler statistic

$$T^* = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} x_{i,j}\right)^2}{n_i}. \tag{4.2}$$

Notice that the two examples we have given center on the magnitude of response values in treatment groups, which is typical of the experimental approach. The appeal to equivalence with traditional $t-$statistics and F-statistics is something of a false justification for the use of D^* and T^*. That is, $t-$statistics and $F-$statistics are justified based on the concepts of random variables, normal distributions, and statistical independence. None of those concepts are necessary to support the logical basis of the experimental approach. Thus, an appeal to equivalence with these statistics to support use of simpler forms such as D^* and T^* given above is without logical force. What *is* necessary, however, is that a test statistic is chosen that meets two requirements:

1. The test statistic must reflect the systematic effect that might be anticipated under whatever physical meaning is attached to treatment groups.

At least hypothetically, it may be possible for a given treatment to decrease small responses and increase large responses, leading to an increase in variability among units given that treatment. That such a situation is not amenable to examination through the experimental approach is indicated by the next requirement.

2. The treatment effect that might be anticipated must be realized in all, or least the majority, of units which are exposed to it, and this must be reflected in the statistic chosen.

A reasonable supposition is that the type of effect that distinguishes A from B in the nested syllogism of experimentation of Section 4.2 should apply more or less uniformly to each unit in the population. This makes it difficult to envisage anticipated treatment effects other than a change in magnitude of response.

4.5 Permutation Tests

The fundamental procedure connected with the experimental approach is that of permutation tests. Recall from the discussion of Chapter 2.3 that responses are considered characteristic of units under a given set of pertinent external influences (i.e., factors). In addition, recall from the discussion of Chapter 4.1 that all such factors are controlled to be constant for all units in the experiment *other than those that define the treatments of interest*. Recall from Chapter 4.2 that a disjunction has been defined between chance alone and a systematic effect of the (small number of) factors that define treatment groups. Finally, recall from Chapter 4.3 that one of a number of equally likely assignments of units to treatment groups has been chosen as the actual treatment assignment

used in the experiment. Putting these together yields the following:

1. If chance alone is in operation, then the treatment groups have no effect on responses, and all responses observed are subject only to inherent variability among units chosen for the experiment. The response attached to a given unit would thus be the same no matter which treatment group it happened to be assigned to.

2. If, further, random treatment assignment was conducted, each of the possible treatment assignments was equally likely. Thus, the statistic calculated to reflect between treatment differences for the actual treatment assignment used is simply one value chosen with equal probability from a set of values that could have arisen from the various possible treatment assignments.

3. Given 1 and 2, the rank of the actual test statistic among those calculated from arbitrary re-assignment of units to treatment groups can be used to calculate the probability of obtaining a statistic as extreme as that actually observed, under the hypothesis of chance alone.

This progression motivates a permutation test procedure, which is conducted as follows:

1. For a given experimental design, consisting of n units assigned to k treatment groups of sizes $n_1, \ldots, n_k$, list the set of all possible treatment assignments (as in Example 4.1).

2. For a given test statistic D, let D^* represent the value calculated for the actual treatment assignment, and compute the value of D for all other possible assignments under the assumption that chance alone is in operation.

3. Rank all values of D computed in step 2, and define the $p-$value of the test as (assuming that a treatment effect is associated with large values of D) as,

$$Pr(D \geq D^*) = \frac{\text{No. } D \geq D^*}{(n!/(n_1!n_2!\ldots n_k!))}.$$

Comments

1. The $p-$value in item 3 above has exactly the same definition as that you are familiar with from previous courses (the probability of obtaining a test statistic at least as extreme as that observed under the hypothesis to be discredited).

2. This probability is calculated as that of an event under the Laplacian concept of probability (number of outcomes for which the condition of an event is satisfied divided by the total number of equally likely outcomes).

3. While the $p-$value has the definition you are familiar with, it may not have the same interpretation you are accustomed to, as we are about to see.

Example 4.1 (cont.)

Consider again the experiment of Example 4.1 with 5 units assigned to 2 treatments of sizes 3 and 2. There were 10 possible treatment assignments in this experiment. Suppose the responses attached to these units were as follows:

Unit:	U_1	U_2	U_3	U_4	U_5
Value:	4.446	4.882	3.094	11.887	5.034

Suppose that the actual treatment assignment was assignment 1 in the table of Chapter 4.3, namely units U_1, U_2 and U_3 to treatment 1 and units U_4 and U_5

to treatment 2. Values of the test statistic D corresponding to the assignments of that table are:

Assignment	D
1	4.319
2	3.008
3	2.703
4	1.152
5	4.194
6	3.134
7	1.881
8	3.830
9	3.498
10	2.007

The value of the actual assignment was $D^* = 4.319$, giving a $p-$value of

$$p = \frac{1}{10} = 0.10$$

Should this $p-$value of 0.10 be considered a "significant" value? That is, does $p = 0.10$ provide evidence against the proposition of chance alone? On one level, we are generally trained that a $p-$value of 0.10 is not particularly small (although it is on the boundary of what many scientists would accept). If we do not consider 0.10 "small", we would reach a conclusion that chance alone cannot be discounted. But, $p = 0.10$ is the smallest $p-$value that could be obtained from this experiment, no matter how extreme the effect of the treatment. Thus, we should either be willing to accept it as an indication of significance or admit that the experiment was designed in such a way that no treatment effect could be detected (i.e., was a worthless exercise).

Now, you may already be thinking that the use of a t-test could remove the need to take the size of an experiment into account in the assessment of $p-$values. This is true and, although in this situation the use of a t-test would need to be considered an application of a modeling approach, here's what would happen:

T_i	$\bar{x}_i$	s_i^2	n_i
T_1	4.141	0.8694	3
T_2	8.460	23.4873	2

These values lead to:

s_p^2	t^*	df	p
12.0334	-1.2453	3	0.8493

Clearly, a $p-$value of 0.8493 would not lead to rejection of the null hypothesis (here, equality of treatment means). What happened here, and why such a difference in the results of the permutation test and the t-test?

Relative to the outcome of the t-test, attention is immediately drawn to the difference in sample variances for the two treatment groups (0.8694 versus 23.4873), and we might reasonably feel that the assumptions of the standard t-test (i.e., equal variance) have been violated, thus nullifying the result of this test for these data. Although one may always question the assumption of normality, it would not be possible to assess this assumption with the amount of data included in the experiment. In addition, we may understand that a 3 degree of freedom t-test is not very powerful for any but the most extreme displacements (i.e., differences in distribution means).

On the other hand, it is also true that, in the actual assignment, all (i.e., both) values for treatment T_2 are greater than all values for treatment T_1. The permutation test is formulated on the basis of "chance alone" – the use

of group averages in the test statistic was simply to capture the characteristic of interest, not as estimates of any population quantity. Thus, "population means" do not enter into the formulation, and the distance between group means is irrelevant, only the ordering of those means.

Comments

1. Combining the above observations gives insight into the fact that the use of theoretical probability distributions (e.g., normal distributions for responses, leading to typical test statistics distributed as t or F) is not merely as an approximation to the $p-$value that would result from a permutation test, at least in small experiments; we will discuss such approximation for large experiments in Chapter 4.9. Until only recently I used to believe that theoretical distributions were motivated by such approximation, and to remove the dependence of $p-$values on experiment size, and that these motivations did not depend on the size of the experiment; at my age, "recently" includes any time interval up to about 4 years in the past.

2. In actual fact, the units U_1, U_2 and U_3 (treatment group T_1 in the actual assignment) were randomly generated from a $N(5, 2)$ distribution, using notation $N(\mu, \sigma^2)$, while units U_3 and U_4 were randomly selected from a $N(7.8, 2)$ distribution. Since we know that the assumptions of the t-test are satisfied, this emphasizes again that a parametric test such as the t-test may not be a good approximation to the actual permutation procedure in small experiments. Thus, in a true experimental approach, it may well be impossible to divorce assessment of $p-$values from size of experiment.

3. It is certainly true that normal distributions might be used in a model-

ing approach for this example, leading here to a t-test which would be perfectly valid. Thus, the important point is not that either the permutation $p-$value or the t-test $p-$value are incorrect. They are, in fact, both correct in that no assumptions under which they have been formulated have been violated in either case. The important point is that the experimental approach and the use of normal distributions as a model are, in fact, different approaches.

4. Finally, one might question whether the values reported for this example were somewhat contrived, and this would, in a way, be correct. The values were in fact generated from $N(5, 2)$ and $N(7.8, 2)$ distributions, although such simulation was continued until a realization appropriate for the example was obtained. A pertinent point is that this was easy to do. In fact, only 4 data sets were generated from these two distributions before obtaining the one used. This might give pause to any eager to designate certain data values as "outliers" (such as the value of 11.887 in this example) in small sets of data.

4.6 Toward Inductive Inference

At the beginning of this chapter, it was assumed that a simple random sample of n units had been selected (from a population of N units) for use in a given experiment. The comment was made that this is rarely true in experimental procedures, but that the experimental approach is most easily understood under this assumption. In fact, obtaining a simple random sample of population units to use in an experiment plays only one role, to allow construction of an inductive argument, which is something of a elusive gold standard in statistical

inference. Put simply, the question of induction is how the results of a particular experiment or study (using only n units) might be logically extended to the entire population (consisting of N units). Under an assumption that a simple random sample of size n from a population of size N has been obtained for use in an experiment, an informal inductive argument can be constructed as follows:

1. Under simple random sampling, component estimates of a test statistic (such as means or variances for individual treatment groups) are unbiased estimates of the corresponding population quantities under those treatments. Unbiased here is in the sense of population averages as for survey sampling methodology (see Chapter 3).

2. Test statistics constructed as linear functions of the component estimates are thus unbiased estimates of those test statistics in the population, this latter being an average over all possible samples of size n from a population of size N.

3. Significance values ($p-$values) computed under permutation of treatment assignments are thus unbiased estimates of a "population-level $p-$value", to test the hypothesis of chance alone.

Comments

1. This informal argument is difficult to make mathematically precise, primarily because definition of the "population $p-$value" is elusive. Such a quantity must depend on not only attributes that are characteristic of population units (as in survey sampling) but on quantities that are characteristic of population units under all relevant external factors. Since

the set of relevant factors includes treatments that are not actually applied, this population $p-$value is a quantity that is hypothetical in nature.

2. Despite the difficulty raised in comment 1, the notion that permutation tests constitute a procedure that allows extension of inference beyond the n units actually used in an experiment seems to have some rational force. Since the n units selected for the experiment are chosen in a manner such that all such groups are equally likely, conclusions from the n units used should extend in a natural manner to the entire population of units.

3. It would be possible to devote a great deal of time to the issue of this subsection, but it is doubtful that doing so would be of much value. The primary reason for this is that it is extremely rare that experiments are conducted with n units randomly sampled from a larger population. In fact, the strongest proponents of procedures based on randomization probability dismiss this possibility from the outset (e.g., Edgington, 1980; Good, 1994).

4.7 Randomization Tests

The term *randomization test* is generally used to refer to a permutation test procedure applied to an experiment for which the participating units have not been selected by a random sample from an existing population. In this context, the term *randomization* stems from the fact that the appropriate permutation of data among groups is dictated by the (randomized) manner in which treatments are assigned. Indeed, if a random sample of population units was available there would be no need for further random assignment

of treatments. We would simply assign the first n_1 units sampled from the population to treatment T_1, the next n_2 units sampled to treatment T_2 and so forth.

4.7.1 Experiments Lacking Random Samples

In designed experiments, a random sample from a population is rarely, if ever, available. Scientists use units that are available, perhaps taking a random sample of available units if that is a large group, but more often it is a struggle simply to obtain enough units in the first place.

Example 4.2

Consider an experiment which is designed to assess the effect of two different diet regimens on weight gain in pigs. External factors in such a situation may include breed, initial age at the start of experimentation, previous housing conditions, previous diet, sex, and birth condition (weak or normal say). The experiment must control as best possible for all of these factors, and the simplest structure is to select pigs for which all of these factors have been essentially the same up until the time of the experiment (this issue is called *uniformity* of units). From where does a scientist obtain pigs that satisfy a reasonable condition of uniformity in all of these characteristics?

Certainly, it is not possible to obtain a list (i.e., sampling frame) of all such pigs in existence. Even more, the intent of the experiment is clearly meant to apply to *future* pigs, as well as current pigs, a "population" from which it is most definitely impossible to draw a random sample.

In fact, the need to select units which are *uniform* in external factors that

may influence the response of interest generally outweighs the desire to obtain a random sample from some population, leading to a typical situation in experiments in which units to be included are carefully selected (not at all randomly selected from some existent population).

4.7.2 Experiments With Constructed Units

It is also common, at least in laboratory studies, to have units that are *constructed* or *manufactured* to certain specifications. For example, petri dishes constructed to contain a given medium for growth of a particular cell type would fit this situation, as would microarray plates prepared with particular gene segments (sequences of nucleotides) in a series of "wells".

Example 4.3

A simple experiment was conducted to investigate whether one of the primary plant nutrients (phosphorus and nitrogen) limits algal growth in a reservoir in Thailand (Sinagarind Reservoir). This reservoir, located northwest of Bangkok provides water supply, irrigation and hydropower to the surrounding area. Excessive algal growth in this region with a 12 month growing season can clog pumps, turbines, and filters, causing severe problems with these functions. The experimental design was as follows:

1. A total of 12 translucent plastic containers (pre-treated so that the plastic material would not affect the outcome of interest) were filled with surface water from the reservoir.

2. Three of the containers received addition of $7.5\mu g/L$ phosphorus (as K_2HPO_4), three containers received $112.5\mu g/L$ of nitrogen (as NH_4NO_3),

three containers received both phosphorus and nitrogen, and three containers received no nutrient addition. Which treatment was applied to which container was determined by a random treatment assignment (this was really true, I was there).

3. All containers were suspended at a depth of 1/2 the photic zone for 2 days.

4. Containers were collected, samples were filtered and algal content determined as the concentration of chlorophyll using standard procedures in limnology (e.g., Jones et al., 1990).

4.8 Random Selection of Permutations

The use of randomly selected permutations of treatment assignments, from the set of all possible permutations, applies to any permutation or randomization test procedure. The fundamental idea is that one may obtain an unbiased estimate of the "true" $p-$value by randomly sampling only a portion of the possible treatment assignments. The "true" $p-$value here refers to the $p-$value that would be obtained from computing the test statistics for all possible treatment assignments. Random sampling of possible treatment assignments is often called the use of *random data permutations*, while the use of all possible treatment assignments is often called *systematic data permutation*.

One motivation for the use of random data permutation is certainly the fact that the number of possible treatment assignments increases rapidly with the size of an experiment and computing a test statistic for all of these can be time consuming and difficult. For example, the table below presents the number of possible treatment assignments for some experiments of differing

types with k treatments and sample sizes of $n_1 \ldots, n_k$, up to $k = 4$.

k	n_1	n_2	n_3	n_4	No. Assignments
2	2	2			6
2	3	2			10
2	3	3			20
2	5	5			252
2	10	10			184,756
3	3	3	3		1,680
3	5	5	5		756,756
4	3	3	3	3	369,600
4	5	5	5	5	11,732,745,024

Note that, first of all, none of the above experimental designs contains more than 20 units (i.e., they are all small experiments). Also, note that the effect on number of treatment arrangements of an increase in group sizes becomes much more extreme as the number of groups increases. For two groups ($k = 2$), an increase from 3 to 5 units per group increases the number of possible assignments by $252/20 = 12.6$. For $k = 3$ an increase from 3 to 5 units per group results in a 450.45 times increase, while for 4 groups this value becomes $31,744.44$. The implications for computation should be obvious, although with modern computing speeds, the only one of the experiments in this table that might give one pause is the last experiment with $k = 4$ treatments and 5 units per treatment.

Equally important to the sheer number of possible assignments in practice, however, is the difficulty of programming data permutations. For two groups (treatments) of equal size, writing a computer function to identify all possible

assignments is not difficult. But the task increases noticeably for groups of unequal size, and becomes even more onerous as the number of groups increases. On the other hand, as we will see in lab, computing random data permutations is not difficult, even for multiple groups of unequal sizes. Thus, the use of random data permutations remains probably the fundamental computational method by which to conduct randomization (or permutation) tests in all but the simplest of situations.

Example 4.3 (cont.)

Consider the nutrient enrichment experiment introduced in Example 4.3. The data obtained from this experiment were as follows:

Treatment	Rep	Chl(μg/L)
Control (C)	1	2.375
	2	2.350
	3	2.500
Nitrogen (N)	1	3.325
	2	3.175
	3	3.525
Phosphorus (P)	1	2.450
	2	2.575
	3	2.400
N Plus P (NP)	1	4.950
	2	4.900
	3	4.875

We will use data from this entire experiment in lab, but for now suppose we consider only the C, N, and P treatments. There were a total of 9 units (bags

of water) assigned to these treatments with sample sizes of $n_c = n_n = n_p = 3$. Under an assumption that treatment has no effect on the response of interest (i.e., chance alone) there are a total of

$$\frac{9!}{3!\,3!\,3!} = 1,680$$

possible treatment assignments, which of which has equal probability. Using a random selection of $S = 100$ of those possible assignments with the test statistic T of expression (4.2) resulted in $p = 0.03$ which should, by any standard, be judged as evidence against the hypothesis of chance alone.

Ecological theory holds that there should be only one "limiting factor" in operation for algal growth, and here we are considering nitrogen (N) and phosphorus (P) as possibilities. We can use the available data as a test of this scientific theory. If we examine only the C and P treatments we have 2 groups of size 3 each and thus 20 possible treatment assignments. If we list out these assignments we will discover that there are "symmetric" or "mirror image" pairs of assignments. For example, the assignment of U_1, U_5, and U_6 to treatment C and U_2, U_3 and U_4 to treatment P gives the same test statistic as the assignment of U_2, U_3 and U_4 to treatment C and U_1, U_5 and U_6 to treatment P. Thus, there are only 10 assignments that need to be examined in a systematic permutation procedure (i.e., considering all possible assignments). The test statistic D of expression (4.1) yields the actual value $D^* = 0.067$ and values for the 10 unique possible assignments of the data are:

Assignment	D	Assignment	D
1	0.067*	6	0.080
2	0.100	7	0.033
3	0.017	8	0.050
4	0.133	9	0.067
5	0.000	10	0.017

Here, the $*$ superscript denotes the actual treatment assignment used. The calculated $p-$value for this treatment comparison is then

$$p = \frac{1}{10} I(D \geq 0.067) = \frac{5}{10} = 0.50$$

Keeping in mind that the smallest $p-$value possible from this experiment would be $2/20 = 1/10 = 0.10$, we must conclude that the data provide no evidence against a hypothesis of chance alone concerning the responses to the C and P treatments.

A similar procedure applied to the C and N treatments yields:

Assignment	D	Assignment	D
1	0.930*	6	0.380
2	0.380	7	0.150
3	0.480	8	0.170
4	0.250	9	0.400
5	0.280	10	0.300

and an associated $p-$value of 0.10, the smallest possible value. Overall, then, we would conclude that there is no evidence for a systematic difference between C and P treatments but there is evidence for a systematic difference between C

and N treatments. Thus, based on the evidence provided by this experiment, we would conclude that N is the limiting factor in Sinagarind Reservoir, and that our results are consistent with (i.e., are confirmatory for) ecological theory.

The overall test in this experiment was of the ANOVA type with $k = 3$ treatment groups. With $S = 100$, the possible values of the calculated $p-$value are contained in a set of values that have an increment of 0.01. That is, possible values for the calculated p are in the set $\{0.01, 0.02, \ldots, 1.00\}$. Values for p with all permutations would have increments of only $1/1,680 = 0.000595$. The $p-$value reported from a set of $S = 100$ randomly selected data permutations was $p = 0.03$ which, as we have already noted, is an unbiased estimate of the $p-$value that would be obtained from a systematic use of all $1,680$ possible permutations (i.e., treatment assignments). This suggests the possibility of conducting a Mote Carlo procedure to more precisely estimate the true $p-$value, by conducting our procedure with $S = 100$ repeatedly. We will call one procedure a "trial". For example, if we conduct the randomization test with $S = 100$ for a total of M independent trials, the average of the M $p-$values should be a more precise estimator of the true $p-$value than that of only one trial. The following table presents results for several different values of the number of random permutations S and the number of trials M.

S	M	Mean	Variance
100	20	0.025	0.000110
	50	0.023	0.000135
	100	0.022	0.000093
	500	0.024	0.000137
200	20	0.018	0.000048
	50	0.019	0.000065
	100	0.017	0.000063
	500	0.018	0.000075
500	20	0.015	0.000025
	50	0.015	0.000028
	100	0.015	0.000029
	500	0.015	0.000026

It should be clear from the above table that increasing the number of permutations used lends stability to mean $p-$value and the variance of those values. What is curious is that the mean $p-$value appears to decrease as the number of random data permutations used increases from $S = 100$ to $S = 500$. The computer function I used in this example samples possible permutations *with replacement*. At first glance, this may seem contradictory to the intention of sampling a subset of permutations; if we were able to use all possible permutations we would do so, each permutation appearing once and only once in the set of possible arrangements. But this is not what we actually want, unless we can indeed compute D for all possible arrangements. By embedding the computation of a $p-$value in what is essentially a Monte Carlo procedure, we have brought relative frequency probability into play. That is, we have turned estimation of the true $p-$value into a problem of sampling

from a multinomial distribution (in which the categories are defined by the set of possible permutations). This is an interesting turn of events – we are using randomization and Laplacian probability for the basis of inference, but relying on relative frequency in computation of the $p-$value. A similar phenomenon occurs in the computation of posterior distributions via simulation in some Bayesian analyses. It is important in such situations to keep the roles of various probability concepts clearly in mind. The *analysis* itself may not rely on relative frequency probability but, when computations are conducted through simulation or Monte Carlo methods, relative frequency becomes relevant for the computational aspects of the analysis.

4.9 Theoretical Probability Approximations

As we have seen in consideration of Example 4.1, the experimental approach, based on the application of Laplacian probability to possible treatment assignments, is fundamentally distinct from the use of distributional assumptions of normality, equal variance, and independence to derive t and F distributions for comparison of group means. In fact, these procedures depend on the concept of hypothetical limiting relative frequency discussed in Chapter 1.2. Nevertheless, there is a connection between typical procedures based on what is often called "normal sampling theory" (t-tests, F-tests, and the like) and the experimental approach. This connection is part historical and part asymptotic in nature.

4.9.1 The Historical Connection

Before the advent of high speed computers, the use of randomization procedures in situations even as small as $k = 3$ treatments of sizes $n_1 = n_2 = n_3 = 3$ was somewhat prohibitive (there are $1,680$ possible treatment assignments in this setting). Random selection of the possible assignments using a computational algorithm was not possible either. And, then as now, this illustration with a total of 9 units would have constituted a quite small experiment. With the mathematical work of Fisher, Neyman, Pearson, Pitman, and others, there was widespread use of theory based on the concepts of random variables having normal distributions to compute $p-$values for tests. Such methods actually fall into what we will call the modeling based approach, but this was (as far as I can tell) not in the minds of these early workers. It does appear, however, that a fair number of these eminent statisticians were willing to use theoretical distributions (e.g., independent normal random variables and the associated sampling distributions for t and F statistics) to approximate what were otherwise incalculable randomization $p-$values. Several quotes to this effect, drawn from the works of Edgington (1980) and Good (1994) are:

1. In the context of permutation tests, Fisher (1936, p. 59):

   > Actually the statistician does not carry out this very tedious process but his conclusions have no justification beyond the fact they could have been arrived at by this very elementary method.

2. From Kempthorne (1955, p. 947):

   > Tests of significance in the randomized experiment have frequently been presented by way of normal law theory, whereas

their validity stems from randomization theory.

3. In summarizing a number of sources, Bradley (1968, p. 85):

> Eminent statisticians have stated that the randomization test is the truly correct one and that the corresponding parametric test is valid only to the extent that it results in the same statistical decision.

4.9.2 The Asymptotic Connection

Good (1994) provides perhaps the most extensive summarization to date of asymptotic connections between randomization (permutation) tests and typical parametric tests. These connections center on the concept of power of a test under the usual parametric assumptions. The general conclusion, in terms of t-tests and F-tests is that, asymptotically, randomization and parametric tests "make equally efficient use of the data" (Good, 1994, p. 177).

The idea of assessing permutation or randomization test procedures relative to their asymptotic properties under the assumptions of typical parametric models seems to clash with the underlying motivations of at least randomization test procedures in the first place. That is, randomization tests are based on the principles of Laplacian probability applied to a small, finite set of units manipulated in an experimental procedure. It would seem difficult to apply asymptotic concepts to this setting. It would appear that the intent of Good (1994), and references therein, is to demonstrate that little or nothing is lost by using a randomization procedure even when large sample theory connected with parametric tests is available and, while this may be indeed be an argument in favor of randomization tests, it does not seem to justify the randomization

theory in the first instance.

4.9.3 Where Does This Leave Us?

It is legitimate to question at this point, where we stand with regard to the experimental approach and, in particular, the use of randomization test procedures, be they applied through either systematic or random data permutation. The following issues represent my own opinion and may not be (are probably not, in fact) accepted by a majority of statisticians. At the same time, I would claim that the majority of statisticians have not given these issues the serious thought they deserve.

1. As clearly asserted by Edgington (1980) and eluded to by Good (1994), randomization test procedures do not share the same conceptual basis as parametric tests such as t-tests and F-tests. I trace the difference back to the fundamental reliance of the experimental approach (and its manifestation in permutation and randomization tests) on Laplacian probability.

2. It does appear to be the case that parametric test procedures and randomization procedures tend to agree for large experiments (leaving the definition of large vague). The primary differences seem to occur in experiments of limited size. Such experiments do occur quite frequently.

3. Any justification for approximating randomization $p-$values with those from tests based on theoretical probability concepts seems to have vanished with advances in computational ability. Particularly with random data permutation, programming randomization tests is not a major difficulty, and computational time has ceased to be an issue at all. The logical distinction seems to be a victim of history in that the developers

of what is called here the experimental approach embraced parametric test procedures so that they could accomplish computations.

4. It is true that randomization procedures appear most applicable in situations for which the treatment effect is believed to be an effect on the magnitude of response that applies to all units in a more or less uniform manner. This is acknowledged by Good (1994, page 2) in pointing out that one of the few assumptions of a randomization test is that ". . . the alternatives are simple shifts in value".

Putting all of this together, I would suggest that the experimental approach, while often eluded to as a justification for random sampling (if possible) and randomized treatment assignment (as a basic tenet of experimentation), has not been faithfully adhered to by statisticians. The experimental approach is limited in application, but represents a forceful methodology. When appealed to as justification for an analysis, it should be more than just "window dressing". It should be carried out in a manner consistent with the manner that it introduces probability into a problem (i.e., through randomization founded on the concept of Laplacian probability).

Part II

STATISTICAL MODELING

Chapter 5

Statistical Abstraction, Random Variables, and Distributions

We turn now to an approach that arguably constitutes (along with the Bayesian approach, which can be cast as an extension of modeling) the majority of all statistical analyses conducted today. This approach differs in a radical way from those of sampling and what we have called the experimental approach. The modeling approach rests on the mathematical concepts of random variables and theoretical probability distributions. While a population of physically existing (or constructed) units is allowed, it is not required for these concepts to be valid. The focus of analysis becomes the values of parameters in a model, not the value of some attribute or response in a population or collection of experimental units. A fundamental characteristic of statistical modeling is that the model represents a statistical *conceptualization* of the scientific mechanism or phenomenon of interest, which could "produce" the observed data as possible values of the random variables involved.

Recall the discussion of scientific abstraction in Chapter 4.1, where ab-

straction was taken to mean consideration of a part divorced from the whole. We now wish to translate this same idea into a statistical formulation called a *model*. Any number of statisticians have indicated that the advent of *generalized linear models* (Nelder and Wedderburn, 1972) as an important landmark in statistical modeling. Lindsey (1996, p. 21) indicates this paper of Nelder and Wedderburn as seminal for the modeling approach. Why is this? After all, models of one type or another have been around for far longer than the mid-1900s, and there is little explicit discussion of the general topic of modeling in the Nelder and Wedderburn paper.

The answer lies in the impetus of this paper for consideration of observable phenomena as something more than simply *signal plus noise*. Clearly, the signal plus noise concept for modeling has produced many useful results. Additionally, from the viewpoint of extreme reductionism, signal plus noise is "true" in that all events in the world are the result of some complex set of deterministic processes; this is true of even human behavior if we understood all of the chemical and physiological processes in the brain and spinal cord. Under this reductionist view, and a perfect understanding of the subject under study, the only role for uncertainty (to statisticians as represented in probability structures) is through *measurement error*. There has long been a tendency for statistics to mimic this thinking in models that consist of an expected value component and an error distribution component. Many traditional linear models, such as the simple linear regression models, take this form.

$$Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i; \quad \epsilon_i \sim iidN(0,1).$$

This model is a direct reflection of signal (as the expectation $\beta_0 + \beta_1 x_i$) plus noise (as the additive error ϵ_i). The standard interpretation we can find in

many (probably most) texts on applied regression models is that the error terms represent measurement error and other uncontrolled influences. The fundamental assumption, then, is that other uncontrolled influences can be adequately combined with measurement error into a single error term for the model.

The major impact of generalized linear models was to promote consideration of *random* and *systematic* model components rather than signal plus noise, although Nelder and Wedderburn (1972) did not present their work in this context. As we will see in the sequel, the random model component consists of a description of the basic distributional form of response random variables, *not* necessarily an error distribution, while the systematic model component consists of a description of the expected values of the random component. I would argue that this is more than a matter of semantics. The encouragement is to consider the random model component first, rather than a form for the expectation function first. In so doing, the stochastic model (i.e., distributional portion of the model) becomes much more than a description of the manner in which observations are dispersed around their expectations. Having made this step, we are then poised to consider models with, for example, multiple stochastic elements or even models based on nonstationary stochastic processes.

One additional point should be made at this introductory stage in our discussion of statistical models. We will not consider in detail either *exploratory* phases of an analysis (although some use of exploratory techniques may appear in lab) or purely *data descriptive* techniques such as nonparametric "models" (e.g., spline or kernel smoothers). Our focus will be nearly entirely on parametric models. As we will indicate in greater detail later in this section, this focus is tied to the role of a parametric model as a statistical *conceptualiza-*

tion of a scientific mechanism or phenomenon of interest. As such, we desire more than to detect possible patterns in data (exploratory approaches) or to describe the way in which data fluctuate over time or in relation to other data (e.g., nonparametric smoothers).

5.1 The Concept of Random Variables

The basic building blocks of a statistical model are random variables, but what *are* random variables? First, random variables are not necessarily (in fact, usually are not) something that can be physically realized, they are a mathematical concept. Although we often refer to data as "observed values of random variables" this is not, strictly speaking, a valid notion. What we actual mean is that data represent possible values that might be assumed by random variables. This is a subtle, but important, difference. Consider a collection of n boards all milled to the same nominal thickness; the thicknesses will, of course, vary due to any number of factors. Suppose that we have attached random variables $Y_1, \ldots, Y_n$ to those thicknesses, and have specified that, for $i = 1, \ldots, n$, $Y_i \sim iidN(\mu, \sigma^2)$. Now, the thickness of board 1 (under constant environmental conditions) is what it is; in the terminology of Part 1 of these notes, an attribute of the board. To presume that it "could be" any value $-\infty < y_1 < \infty$ with probabilities given by a normal distribution is, frankly, ludicrous. To say that the thickness of board 1 is a random variable before observation because we don't know what it is, and the particular value of the random variable after observation because we then do know what it is, does not entirely solve this dilemma. The thickness of board 1 is *not* a random variable. We may, however, use a random variable that is connected with the thickness of board 1 to allow a mathematical conceptualization of the values

and uncertainties in thickness of the collection of boards. To put it a slightly different way, random variables and theoretical probability distributions are not simply an extension of finite populations to infinite collections of physical units (if that were even a possibility). Random variables are mathematical beasts, and they do not exist outside of the world of mathematics and statistics. As we will reiterate in Section 5.3, our goal is to use the concepts of random variables and associated probability distributions to formulate a meaningful conceptualization of a real situation called a model and, through statistical analysis of such a model, increase our knowledge of the real situation.

Formally, random variables are mathematical functions that map a set Ω onto the real line $\Re$. You have likely by this time seen the use of the triple

$$(\Omega, \mathcal{F}, P)$$

to represent a probability space, and understand that a random variable Y is a real-valued function $Y : \Omega \rightarrow \Re$ such that $Y^{-1} : \mathcal{B} \rightarrow \mathcal{F}$, where $\mathcal{B}$ is the $\sigma-$algebra of Borel sets on the real line and $\mathcal{F}$ is a $\sigma-$algebra of the set Ω (typically the $\sigma-$algebra generated by Y, the smallest $\sigma-$algebra for which Y is $\mathcal{F}-$measurable). Here, Ω is often unfortunately called a *sample space* containing the possible outcomes of a *random experiment*. Note that this use of the term experiment is not necessarily connected with our description of experiments in Part 1 of this course. It is preferable to consider Ω an arbitrary set of any objects or elements of your choosing. These elements will be denoted as ω, and we will assume that $\omega \in \Omega$. We will consider such elements to be values of a scientific *constructs*. In the case of observable constructs the phrase sample space for Ω may seem fairly obvious, as Ω then consists of the set of possible outcomes of an observation or measurement operation. Even for observable constructs, however, the concept of Ω as a set of possible outcomes

is not entirely unambiguous.

Example 5.1

Consider a study of the composition of a forest bird community that involves setting what are called mist nets (nets of fine mesh strung between two poles, much like a very fragile volleyball net) in a forest. Birds are unable to detect such nets and those that fly into the net are ensnared. Nets are typically set, abandoned for a specified period of time (e.g., 2 hours), and then re-visited. Birds are gently removed from the net, various characteristics recorded, and released; this is, by the way, no small test of respect for bird life when it comes to removing woodpeckers from a mist net without harming them. What sets of outcomes Ω might arise from such an observational operation?

1. Species.

   Here, Ω would consist of a list of all bird species that occur in the study area such as

   $$\Omega \equiv \{\text{wood thrush, black and white warbler, yellowthroat, etc.}\}.$$

2. Sex.

   Here, $\Omega \equiv \{\text{Male, Female}\}$.

3. Weight.

   In this case, we might take

   $$\Omega \equiv \{\omega : 0 < \omega < \infty\}.$$

Notice that, for the third observational characteristic of weight, we have already departed from physical reality. That is, the set of possible outcomes of

the actual measurement operation is determined by the inherent precision of the measurement instrument (e.g., 0.5 grams for a scale marked in grams). In addition, it is physically impossible that a hummingbird, for example, could weigh 3 metric tons (which is still considerably less than ∞). But the construct of *weight* does not depend on the particular measurement tool used to observe it, nor does it depend on a set of physically real "outcomes" of an observation process. For unobservable scientific constructs the situation becomes even more obscure.

Example 5.2

A social scientist is interested in studying the effect of violent cartoons on "aggressive behavior" in children ages 5 to 7. The construct of "aggressive behavior" is ill-defined in terms of observable quantities. Rather, aggression is assessed relative to a set of indicators assumed to be indicative of aggression. In this situation, would it be possible to define a random variable that is a direct reflection of the notion of aggression?

The point of the two examples above is that, while it is sometimes possible to define Ω as a sample space in the traditional textbook sense (e.g., item 2 of example 5.1), and then proceed to a random variable that maps this set to the real line, it is perhaps not the sample space that is fundamental, but the concept of a random variable itself. In fact, in applications, Ω is often determined relative to the random variable Y rather than the other way around. It is important that a legitimate set Ω exist, of course, and it is common that we take $(\Omega, \mathcal{F}) = (\Re, \mathcal{B})$.

5.2 Probability Distributions

In the probability space $(\Omega, \mathcal{F}, P)$, the set function P, defined on $\mathcal{F}$, constitutes a mapping from $\mathcal{F}$ to [0, 1], and obeys the axioms of probability. Briefly, we can derive from $Y : \Omega \to \Re$ and $P : \mathcal{F} \to [0, 1]$ the *probability law* P_Y as,

$$P_Y(B) \equiv P(Y^{-1}(B)) = P(Y \in B); \quad B \in \mathcal{B},$$

the *distribution function* F as,

$$F(y) \equiv P(\omega : Y(\omega) \leq y) = P_Y(Y \leq y),$$

and, with the addition of the Radon-Nikodym theorem, the *density function* f as,

$$f \equiv \frac{dP}{dP_0} \text{ a.s. on } \mathcal{F},$$

where P is dominated by P_0. For our purposes, P_0 will be either Lebesgue (for continuous Y) or counting (for discrete Y) measure.

What is the value of all of this for a consideration of statistical *methods*? The answer is that, in the formulation of models, we often construct functions of random variables, specify either marginal, conditional, or joint distributions, and either aggregate or disaggregate basic random variables. It is essential that, whatever we end up with, a joint distribution appropriate for the definition of a likelihood function exists. This is true regardless of the approach taken to estimation and inference, be it exact theory, maximum likelihood, likelihood approximations (e.g., quasi- and pseudo-likelihoods), or Bayesian. While we may often "work in reverse" in that we may proceed to identify probability laws and measures P_Y and P based on formulated densities f and distribution functions F, we must arrive at a situation in which all of the above holds. This is, in some ways, similar to the previously mentioned

tendency to describe Ω based on the concept of a random variable, rather than vice versa.

Throughout the remainder of this part of the course we will use a number of theoretical probability distributions, many of which you will already be familiar with, at least in passing. In the next full section we will summarize useful results for several *families* of distributions, most notably the exponential family. Distributions within a family share various statistical properties (which properties depends on which family). This makes it possible to use families of distributions in basic model frameworks, arriving at classes of models that inherit certain behaviors from the families on which they are founded.

5.3 Statistical Abstraction

We come now to one of the main events in the modeling approach, which is the idea of statistical abstraction. If scientific abstraction consists of considering a part of a problem divorced from the whole, statistical abstraction consists of capturing the key elements of a problem in a small set of parameters of a probabilistic model. What do we mean by the *key elements*?

The majority of scientific investigation is based on the concept that there exists a *mechanism* that underlies the production of observable quantities. A mechanism is the set of physical, chemical, and biological forces that govern the manner in which some process functions. Certainly, discovery of a mechanism is a key component of medical research; if the mechanism by which a disease affects the body is known, the chances of developing an effective treatment or vaccine are vastly increased. In about 1999, the US Environmental Protection Agency (EPA) released regulations for "fine particulate matter", defined as particles of mean aerodynamic diameter less than 2.5 microns (a smoke par-

ticle, for comparison is about 8 to 10 microns in diameter). The impetus for these regulations was largely a number of large-scale studies that indicated a relation between the ambient concentration of fine particulates in cities and health outcomes such as the number of hospital admissions for asthma in those cities. These new regulations prompted a number of legal battles and a study commissioned to the National Research Council (NRC). In its report, a major difficulty the NRC saw with the evidence used by EPA to issue new regulations was that the mechanism by which fine particulates might produce respiratory problems in humans had not been identified. I believe that enforcement of the new regulations was suspended until additional studies could be conducted.

In many (most?) areas of science, mechanisms are not fully understood; if they were, we would be moving closer to the type of perfect understanding of the "signal" discussed at the beginning of this chapter. This is perhaps less true of physics than many areas of biology, but even there understanding of a basic mechanism under highly controlled conditions in a laboratory does not necessarily indicate the exact physical behavior of the world in an uncontrolled setting. Nevertheless, in nearly all scientific disciplines a finding of relation among various quantities or constructs, or differences among groups in those quantities or constructs, will likely not meet with acceptance among workers in the discipline unless a plausible mechanism can be suggested (this is, of course, not the same as having all of the details worked out). The one exception to this is the occurrence of a phenomenon that is highly repeatable, but not in the least understood. In such situations intense study is generally conducted to determine why the phenomenon persists, that is, to suggest mechanisms. Some of you may be old enough to remember a movie called *The Awakening*, which was based on a true story. In it, a young psychiatrist (played by Robin Williams) treated a group of nonfunctional mental patients with some type of

a drug "cocktail" and an entire group of patients began functioning normally for a time. They eventually relapsed. To me, the most poignant moment in the movie was at the end, where they presented little descriptions of the real life fates of the various characters (right before the credits). The psychiatrist, it was noted, spent the remainder of a frustrated career attempting to duplicate the effect of his treatment, to no avail. In short, something had occurred for which the good doctor could suggest no plausible mechanism, and which could not be repeated. While the event was wonderful, it provided no useful results for the treatment of patients with the given affliction.

It does occur that a study will be conducted in which everything that might possibly be related to the objective of the study is measured or observed. Nearly every statistical consultant of any tenure has been presented with a mountain of data gathered on a cornucopia of variables and asked to determine "what is important". Despite such horror stories, and despite the emergence of what is called "data mining" (and other exploratory approaches for extremely large collections of data) the norm in scientific investigation is still that data are collected in a more focused manner.

The upshot of the above discussion for the purpose of statistical modeling is that scientific mechanisms or repeatable phenomena represent the *key elements* of a problem to be captured in a small set of model parameters, which is the question we have been attempting to address. By this we do not mean a direct translation of a mechanism into mathematical terms (this would be a deterministic process model, which are generally based on sets of partial differential equations) but rather that we be able to "locate" the mechanism in a small set of model parameters, and determine the relation of other model parameters to those that represent the mechanism.

Example 5.3

Suppose we have interest in the effect of altitude on the boiling point of some substance; there is no need to get fancy, water works fine for this example. The plausible (actually at this point in time more than merely plausible) mechanism is the effect of atmospheric pressure on the amount of energy needed to produce vaporization. To investigate this phenomenon we may use various study designs, and these may, in turn, lead to different models. To simplify the situation somewhat, suppose we have located a relatively small geographic region (e.g., near Seattle in the western US, or perhaps in Nepal) where altitude covers a wide range in a short distance. To begin our modeling exercise, we define random variables associated with the boiling temperature of water (in degrees Kelvin, say) $\{Y_i : i = 1, \ldots, n\}$ and covariates as altitude (in meters above sea level, say) $\{x_i : i = 1, \ldots, n\}$.

One possibility is to place observers at n various altitudes along our chosen gradient with portions of a common stock of water (such as deionized water), identical (to within manufacturers specifications) thermometers and other equipment, and have them measure the temperature at which this water boils at the same time (starting heating to within one minute, say). We might then begin a modeling exercise through examination of a linear model,

$$Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i; \quad \epsilon_i \sim iidN(0, 1).$$

In this model, the phenomenon of interest is embodied in the systematic model component $\beta_0 + \beta_1 x_i$.

What is the relation of the other model parameter σ to those that represent the phenomenon of interest? This dispersion (or variance, or precision) parameter quantifies the degree to which observed values of the boiling point

of water differ from what is "explained" by the modeled version of the effect of altitude. In the study described, this would incorporate measurement error and microclimate effects. But what, exactly, have we modeled through the systematic component? The effect of altitude on the boiling point of deionized water on a given day. If ordinary least squares is used to produce parameter estimates for β_0 and β_1, and the usual bias-corrected moment estimator of σ^2 is used, we might quantify our uncertainty about the modeled effect of altitude on boiling point of deionized water on the given day through a joint confidence region for β_0 and β_1. Consider, as a part of this uncertainty, only the variance of the marginal distribution of $\hat{\beta}_1$,

$$var\{\hat{\beta}_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Here we have the typical case that our uncertainty about the phenomenon of interest is directly related to the variance of the observable random variables.

Another possible study design would be to place observers at n various altitudes with (possibly randomly assigned) water from n different sources, on n different days (again, perhaps randomly selected from days within some specified time frame). We could fit the exact same linear regression model to the resultant data from this study. The effect of altitude on boiling point of water would be modeled through exactly the same portion of the model as before, namely $\beta_0 + \beta_1 x_i$. But what would be the relation of σ to this systematic model component? Now, σ would reflect not only measurement errors and microclimate effects, but also varying water compositions and factors associated with the days of observation (e.g., humidity, pressure due to atmospheric cells of air that vary over time, etc.).

One likely effect of this design would be to increase the (true) value of the variance of observable random variables, namely σ^2. This would, in turn,

increase our uncertainty about the phenomenon under study. But, we have used exactly the same model form as under the first study design, and we end up with exactly the same portion of that model reflecting the phenomenon of interest. How can we then have greater uncertainty (*a priori*, in fact) about that phenomenon? Is, in fact, the same model term $\beta_0 + \beta_1 x_i$ modeling the same phenomenon in these two situations?

Questions

1. Why might we want to randomly assign water types to observers, and altitudes to days of observation?

2. Why might we *not* want to have each altitude observed on the same day?

3. Might might we *not* want to use this same simple linear regression model if we provide each observer with several water types to be used at their observation altitude on the same day?

4. Under what assumptions does the systematic model component $\beta_0 + \beta_1 x_i$ represent the same scientific phenomenon under these two study designs? Is this reasonable given that the first study design used deionized water?

Example 5.3 (cont.)

Now, consider a study design which is an extension of the type suggested in question 2 above. That is, suppose we provide multiple water types to observers on each day of observation. The water types might be randomly chosen from a list of locations across the world. It might not be necessary to actually visit these locations if the compositions of water at those locations

were known (water can be completely deionized and then "reconstituted" to contain various ionic compositions). This could easily lead to a larger study if, for example, we choose K water types to be evaluated at our original n altitudes, so that we may have the index i going from 1 to nK, or i going from 1 to n_k for each value of k from 1 to K (to allow for possible differences in the number of altitudes that water types are successfully observed). Consider now fitting a model of the following form, for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$,

$$\begin{aligned} Y_{i,k} &= \beta_{k,0} + \beta_{k,1} x_{i,k} + \sigma \epsilon_{i,k}, \\ \epsilon_{i,k} &\sim iid\, N(0,\, 1) \\ (\beta_{k,0},\, \beta_{k,1}) &\sim iid\, N\left(b,\, \Sigma_b\right), \end{aligned}$$

for some $b \equiv (b_0,\, b_1)$ and 2×2 covariance matrix Σ_b with m, h entry $\sigma_{(m,h)}$ for $m, h \in \{0, 1\}$.

For this model, it is clear that

$$E(Y_{i,k}) = b_0 + b_1 x_{i,k},$$

and

$$var(Y_{i,k}) = \sigma_{(0,0)} + \sigma_{(1,1)} x_{i,k}^2 + 2\sigma_{(0,1)} x_{i,k} + \sigma^2.$$

Where now has our phenomenon of interest gone (i.e., where is it in this model)? The effect of altitude on the boiling point of water is now captured in the distribution of $(\beta_{k,0},\, \beta_{k,1})$; we will later call such a distribution a *mixing distribution*. Notice that a major change has suddenly occurred in our representation of the effect of altitude on the boiling point of water. That effect is no longer a constant term but is, rather, an entire (bivariate) distribution. Does this imply we no longer believe that there is a "true" effect of altitude? No, it means that we no longer believe the the effect of altitude is manifested in the

same way under all conditions. But, would it not be possible to determine the effect of these other factors (ionic composition of the water, atmospheric conditions) and incorporate them into a model? In principle, certainly it would. In principle, we could carry such an exercise to the extreme in which the only uncertainty involved would be a (in this case presumably very small) measurement error. That we have been able to make modeling the simple process of boiling water as complex as we have should convince you that this deterministic approach, while appealing in some cases, is not generally applicable to the range of complex scientific problems under consideration in the world.

What now are the relations of the other model parameters σ^2 and Σ_b to this mathematical formulation of the phenomenon of interest? The variance σ^2 is back to its interpretation as in the first study design using only deionized water (measurement error plus micro-scale effects). The elements of Σ_b are now indicative of the variability in how the effect of altitude is realized or manifested in various situations. The exact meaning attached to this concept depends fundamentally on the manner in which the situations to be observed were chosen. It would, clearly, be an entirely different matter to assign $(\beta_{k,0}, \beta_{k,1})$ a probability distribution across purposefully chosen situations than across situations chosen at random from some "pool" of those possible.

Returning from this example to the overall topic of this subsection, it should be clear that what we are calling statistical abstraction is the process by which a problem from the real world is brought into a conceptual world of random variables, theoretical probability distributions, and hence is subject to the methods of mathematical statistical analysis. A criticism that is sometimes leveled at the modeling approach is that a model "doesn't care where the data come from", or "how the data were obtained". In one way this is true – given the assumptions inherent in a model formulation, analysis will proceed in the

same manner regardless of how the data used in its analysis were obtained. What is missing from this criticism, however, is that no model can properly operate beyond the context given it by the process of statistical abstraction which (in a proper application) must have been given careful consideration. We perhaps do not spend enough time discussing this aspect of statistical analysis in the educational process, but it is my hope that, having been introduced to the topic, you will see it woven into the material presented in the remainder of this part of the course.

5.4 Summary of Key Points

It might be useful, at this point, to summarize many of the key ideas presented in this section (the word model below implies a statistical model).

1. Modeling represents a fundamentally different approach to bringing probability concepts to bear on a problem than that of approaches based on randomization.

2. Models represent a probabilistic conceptualization of a scientific mechanism or phenomenon of interest, and the situation(s) under which that mechanism or phenomenon leads to observable quantities.

3. The basic building blocks of a model are random variables. For application in an empirical scientific investigation, models must include at some point random variables associated with observable quantities.

4. The process of statistical abstraction, by which a mechanism or phenomenon of interest is captured in a model formulation, involves the objectives of study, the data collection design, and the choice of model.

Finally, we will end this introduction to the modeling effort with a preliminary comment about the *modeling process*. While certainly connected with the material of this chapter, theoretical probability distributions have not been the focus of discussion. This is a direct reflection of point 3 immediately above. Distributions are meaningless without random variables that (are assumed to) follow them. In an application, the first step is *not* to select distributions, to write down forms for systematic model components, or to begin considering likelihoods. The *first* and, in many ways most important, step in developing a statistical model is to define random variables appropriate for the problem at hand. The fundamental properties of such random variables (i.e., set of possible values, dependence or independence structure) will largely determine at least a set of possible theoretical distributions that may be used to describe their probabilistic behaviors.

Chapter 6

Families of Distributions Useful in Modeling

We now begin consideration of the tools that are needed for successful statistical modeling. These will include distributions, model structures, estimation methods, inference methods, and model assessment methods. Certainly, to adequately model the probabilistic behaviors of random variables, we must have access to a variety of theoretical probability distributions. We will organize the presentation of such distributions around the concept of *families* which, as mentioned previously, often provide us a means of formulating classes of models that share important characteristics. It is important to note, however, that we are involved in an *introduction* of useful distributions, not an exhaustive effort at developing a catalog (see, e.g., the series of works edited by Johnson and Kotz for such an effort).

Much of this section will be presented in the context of a single random variable. When groups of random variables are necessary they will be indexed by the subscript i. It is important to note, however, that models always deal

with groups or collections of random variables. Notation that will be used throughout this section is as follows:

- Upper case letters such as X, Y, Z, W will be used to denote random variables. The corresponding lower case letters will denote values that could be assumed by these variables.

- The symbol Ω will be used to denote the set of possible values of a random variable, subscripted with the variable symbol if needed for clarity, such as Ω_Y.

- Parameters will be denoted with Greek letters such as θ, ϕ, and λ. The *parameter space*, defined as the set of possible values of a parameter, will be denoted as the corresponding upper case Greek letters, except as noted.

- All parameters may be either scalars or vectors, the difference should be clear from the context. When a generic symbol for a parameter is needed it will be denoted as θ.

- Conditioning notation, $y|x$, will be used in two contexts. One is in which the conditioning value(s) represent fixed quantities such as parameters or covariates not considered random. The other will be in the usual conditioning notation for two or more random variables. It is important that you understand the context being used in a conditional statement (so ask if it is not clear).

6.1 Exponential Families

You have been introduced to exponential families of distributions in previous courses (e.g., Statistics 542). These families constitute an essential class of distributions for modeling purposes. There are various, equivalent, ways to write what is called the exponential family form. For a random variable Y and corresponding probability density function (pdf) or probability mass function (pmf) some of these representations are, all for $y \in \Omega$:

$$\begin{aligned} f(y|\eta) &= \exp\left\{\sum_{j=1}^{s} q_j(\eta)T_j(y)\right\} c(\eta)h(y), \\ f(y|\theta) &= a(\theta)\, t(y) \exp\left\{\theta^T\, t(y)\right\}, \\ f(y|\eta) &= \exp\left\{\sum_{j=1}^{s} q_j(\eta)T_j(y) - B(\eta)\right\} c(y) \\ f(y|\theta) &= \exp\left\{\sum_{j=1}^{s} \theta_j T_j(y) - B(\theta) + c(y)\right\}. \end{aligned} \tag{6.1}$$

Note that, while $\theta = (\theta_1, \ldots, \theta_s)$ (or η) may be a vector, $B(\cdot)$ is a real-valued function. Clearly, the definition of functions such as $B(\cdot)$, $c(\cdot)$, $a(\cdot)$, and $h(\cdot)$ are not exactly the same in these various expressions, but you should be able to easily work out the equivalence.

Example 6.1

If Y is a random variable such that $Y \sim N(\mu, \sigma^2)$, the fourth version of the exponential family given in (1) can be used to write the density of Y with,

$$T_1(y) = y \quad \theta_1 = \frac{\mu}{\sigma^2},$$

$$T_2(y) = y^2 \quad \theta_2 = \frac{-1}{2\sigma^2},$$

and,

$$\begin{aligned} B(\theta) &= \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\{2\pi\sigma^2\} \\ &= \frac{-\theta_1^2}{4\theta_2} + \frac{1}{2}\log\left\{\frac{-\pi}{\theta_2}\right\} \end{aligned}$$

We will use the fourth (last) expression in (6.1) as our basic form for exponential family representation. Other common densities or mass functions that can be written this way include

- The Poisson pmf with $\Omega \equiv \{0, 1, \ldots\}$
- The binomial pmf with $\Omega \equiv \{0, 1, \ldots, n\}$
- The negative binomial pmf with $\Omega \equiv \{0, 1, \ldots\}$
- The gamma pdf with $\Omega \equiv (0, \infty)$
- The beta pdf with $\Omega \equiv (0,\ 1)$
- The log-normal pdf with $\Omega \equiv (0,\ \infty)$
- The inverse Gaussian pdf with $\Omega \equiv (0,\ \infty)$

6.1.1 Properties of Exponential Families

Recall we are using the fourth (last) form for the expression of the exponential family given in (6.1). Note, first, that the term $\exp\{c(y)\}$ in this expression could be absorbed into the relevant measure. This is typically not done so that integrals can be written with respect to dominating Lebesgue (for continuous Y) or counting (for discrete Y) measures.

Exponential families possess a number of useful properties for modeling, some of which we review here in a brief manner.

1. The parameter space Θ (the set of points such that $f(y|\theta) > 0$ for $\theta \in \Theta$) is a convex set. To avoid difficulties, we will consider only members of the exponential family such that neither the $T_j(y)$ nor the θ_j satisfy a linear constraint (in which case the representation is said to be "minimal" or sometimes "full"). If Θ contains an open $s-$dimensional rectangle, then the exponential family is said to be of "full rank", or "regular". These items affect us in model specification because we want exponential families to be written so that they are minimal and regular. For example, a multinomial with H categories will only be minimal if we write the pmf for $H-1$ random variables.

2. For a minimal, regular exponential family, the statistic $T \equiv (T_1, \ldots, T_s)$ is minimal sufficient for θ. This property is often useful because, as we will see, the joint distribution of *iid* random variables belonging to an exponential family are also of the exponential family form.

3. For an integrable function $h(\cdot)$, dominating measure ν, and any θ in the interior of Θ, the integral

$$\int h(y) \exp\left\{\sum_{j=1}^{s} \theta_j T_j(y) + c(y)\right\} d\nu(y)$$

is continuous, has derivatives of all orders with respect to the θ_js, and these derivatives can be obtained by interchanging differentiation and integration (e.g., Theorem 4.1 in Lehmann, 1983). This property does several things for us. First, it can be used to derive additional properties of exponential families (such as the next property given for the form of the moment generating function). In addition, it allows us to evaluate expressions needed for estimation and variance evaluation through numerical integration of derivatives, which can be important to actually

conduct an analysis with real data.

4. The property in item 3 can be used directly (e.g., Lehmann, 1983, p.29) to show that

$$\begin{aligned} E\{T_j(Y)\} &= \frac{\partial}{\partial \theta_j} B(\theta), \\ cov\{T_j(Y), T_k(Y)\} &= \frac{\partial^2}{\partial \theta_j \partial \theta_k} B(\theta) \end{aligned}$$

   These lead directly to $E(Y)$ and $var(Y)$ for what are called *natural exponential* families and *exponential dispersion* families (coming soon). They also will provide an alternative parameterization of exponential families in general (coming even sooner).

5. The moment generating function of an exponential family is defined to be that for the moments of the T_js, as,

$$M_T(u) = \frac{\exp\{B(\theta + u)\}}{\exp\{B(\theta)\}}.$$

6.1.2 Parameterizations

In the final expression of (6.1) the parameters denote $\theta_j;\ j = 1, \ldots, s$ are called *canonical* or sometimes *natural* parameters for the exponential family. While the canonical parameterization usually leads to the easiest derivation of properties (such as given above) it is not always the best parameterization for purposes of estimation, inference, or model interpretation. While parameter transformations can be used in a quite flexible manner (they are simple substitutions in density and mass functions), it is helpful to know several other parameterizations that are fairly standard, and are often useful. We will describe two parameterizations here that have both been called "mean value" parameterizations, although they are not the same.

Mean Value Parameterization 1

While we certainly wish to dispel the notion that *location* is the only distributional characteristic of concern in a model, it is true that the expected value is usually of interest (and is often needed to quantify other characteristics in a concise manner). It is nearly always the case that none of the canonical parameters θ_j in (6.1) correspond to the expected value of the random variable Y. Thus, a mean value parameterization can be accomplished by a transformation $(\theta_1, \ldots, \theta_s) \rightarrow (\mu, \phi_1, \ldots, \phi_{s-1})$, where $\mu \equiv E(Y)$ and $\phi_1, \ldots, \phi_{s-1}$ are arbitrarily defined; we will still need s parameters because we are assuming the canonical representation is minimal (see Section 6.1.1).

Example 6.2

Consider a beta random variable Y with pdf

$$f(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \tag{6.2}$$

where $\Omega = (0,1)$ and $\alpha,\ \beta > 0$. As we know for this density, $E(Y) = \alpha/(\alpha+\beta)$. First, let's write this density in canonical exponential family form as,

$$\begin{aligned} f(y|\alpha,\beta) \;=\; & \exp\left[(\alpha-1)\log(y) + (\beta-1)\log(1-y)\right. \\ & +\; \left.\log\{\Gamma(\alpha+\beta)\} - \log\{\Gamma(\alpha)\} - \log\{\Gamma(\beta)\}\right], \end{aligned} \tag{6.3}$$

or,

$$\begin{aligned} f(y|\theta) \;=\; & \exp\left[\theta_1\log(y) + \theta_2\log(1-y)\right. \\ & +\; \log\{\Gamma(\theta_1+1)\} - \log\{\Gamma(\theta_2+1)\} \\ & -\; \left.\log\{\Gamma(\theta_1+\theta_2+2)\}\right]. \end{aligned} \tag{6.4}$$

The equality of (6.2) and (6.3) is immediate, and (6.4) is obtained from (6.3) by taking

$$\theta_1 = \alpha - 1 \quad T_1(y) = \log(y)$$
$$\theta_2 = \beta - 1 \quad T_2(y) = \log(1-y)$$

and

$B(\theta) =$

$$\log\{\Gamma(\theta_1+1)\} - \log\{\Gamma(\theta_2+1)\} - \log\{\Gamma(\theta_1+\theta_2+2)\}.$$

In terms of the canonical exponential representation, $E(Y) = (\theta_1+1)/(\theta_1+\theta_2+2)$. We can then achieve a mean value parameterization by taking,

$$\mu = \frac{\theta_1+1}{\theta_1+\theta_2+2}; \quad \phi = \frac{1}{\theta_1+\theta_2+2}$$

We can then write the density in mean value parameterization by substituting into (6.4) the quantities

$$\theta_1 = \frac{\mu-\phi}{\phi}; \quad \theta_2 = \frac{1-\mu-\phi}{\phi}.$$

Notice, in this example, that while we have not manipulated Y in any way, so that Ω remains unchanged throughout, we have gone from $\alpha, > 0$ and $\beta > 0$ in (6.2) and (6.3) to $\theta_1 > -1$ and $\theta_2 > -1$ in (6.4) to $0 < \mu < 1$ and $\phi > 0$ in the mean value parameterization.

Mean Value Parameterization 2

In the canonical parameterization for exponential families there is a clear association between parameters θ_j and sufficient statistics T_j. It is perhaps natural then to attempt to parameterize families using the expected values of the T_j,

which are given by first derivatives of the function $B(\theta)$. Thus, we transform $(\theta_1, \ldots, \theta_s) \rightarrow (\mu_1(\theta), \ldots, \mu_s(\theta))$ where

$$\mu_j(\theta) = E\{T_j(Y)\} = \frac{\partial}{\partial \theta_j} B(\theta).$$

This parameterization has the potential advantage that each parameter of the density is then the expected value of a random variable associated with an observable quantity, namely $T_j(Y)$.

Example 6.3

From example 6.1 we have that, for a normal density, $T_1(Y) = Y$, $T_2(Y) = Y^2$, and,

$$\begin{aligned} \frac{\partial}{\partial \theta_1} B(\theta) &= \frac{-\theta_1}{2\theta_2}, \\ \frac{\partial}{\partial \theta_2} B(\theta) &= \frac{\theta_1^2 - 2\theta_2}{4\theta_2^2}. \end{aligned}$$

Given that $\theta_1 = \mu/\sigma^2$ and $\theta_2 = -1/(2\sigma^2)$, we then have that,

$$\begin{aligned} \mu_1(\theta) &= \frac{\partial}{\partial \theta_1} B(\theta) = \mu, \\ \mu_2(\theta) &= \frac{\partial}{\partial \theta_2} B(\theta) = \mu^2 + \sigma^2, \end{aligned}$$

and these are easily seen to be the expected values of $T_1(Y) = Y$ and $T_2(Y) = Y^2$. Notice for this example that mean in mean value parameterization 1 and the first parameter under mean value parameterization 2 are the same, namely the expected value of Y. This is, rather obviously, because the first sufficient statistic is Y. Families with this structure are among the more commonly used modeling distributions (see Section 6.1.3 on exponential dispersion families).

Mixed Parameterizations

It is also possible to write an exponential family in terms of a parameterization that is part mean value and part canonical, for example, with parameter $(\mu_1(\theta), \theta_2)$. I have not seen such parameterizations used much, but they apparently (Lindsey, 1996, p. 29) have the intriguing property of *variation independent* parameters (see below).

Comments on Parameterizations

As should be clear from Example 6.2, parameterizations other than the canonical one are generally *not* chosen to make the expression of the density shorter or less complex. There are a number of other reasons one might choose one parameterization over another, some at the modeling stage, some at the estimation (and/or inferential) stage, and some at the interpretational stage.

1. Parameter transformations made for the purposes of interpretation are frequently conducted after estimation has been completed. This is often not too difficult, at least for estimation using maximum likelihood or posterior simulation (as we will see later in the course). It is possible, however, that with estimation by exact theory or least squares one might need to conduct a transformation before estimation to allow inference to be made on the transformed parameters.

2. Parameter transformations are not infrequently conducted to produce increased stability in numerical estimation procedures. Parameter transformations can affect the shape of a likelihood function, and what is called *parameter effects* curvature in nonlinear models. Numerical optimization algorithms, for example, tend to perform with greater stability when applied to log likelihoods that are relatively quadratic near the

maximum for a given set of data. For an extensive treatment of this topic, see the book by Ross (1990).

3. Recall that, in model formulation, a primary goal is to connect the key elements of a scientific problem with parameters of a probabilistic model. It can occur that one parameterization makes this more clearly the case than does an alternative. This assertion comes dangerously close to being something of a platitude, however. As any number of my colleagues with extensive consulting experience will point out, most scientists do not think in terms of statistical models. It can be difficult enough to determine the basic objectives in model form, and seeking scientific advice on the appropriate parameterization is a step or two beyond that. Nevertheless, this is an aspect of parameterization that should not be dismissed out of hand.

4. A more easily comprehended goal of parameterization can be to provide an indication of how covariate information can appropriately be incorporated into a model. Our natural inclination is for covariate values to influence the marginal expectation of a random variable. Mean value parameterizations can then aid in the manner that covariates are incorporated into a model structure. For example, suppose we have a beta random variable as in Example 6.2, used to model the proportion of a river sediment sample that consists of particles larger than what would be considered "sand" (defined by particle size). A covariate of water flow (call it x) is believed to influence this (i.e., the faster water moves the more energy it has, the larger the size of the particles it can transport downstream). It is not clear how such a covariate would be incorporated into a distribution written with standard parameterization as expression

(6.2) or with canonical parameterization as in expression (6.4). But, using a mean value parameterization (version 1) we might take

$$\mu = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)},$$

which would give the expected value of the random variable as a monotonically increasing function of x that has the appropriate range in the interval $(0, 1)$.

5. In the investigation of different parameterizations it is essential that one keep track of possible restrictions on the parameter space, both in terms of allowable values and in terms of restrictions that may be imposed on one parameter component (e.g., θ_2) by the value of another (e.g., θ_1). Such restrictions (including possibly the lack of such restrictions) can render a parameterization either more or less appropriate to describe a given situation. From a purely statistical viewpoint, it seems pleasing to have parameter elements that are *variation independent*. A generic vector-valued parameter $\theta \equiv (\theta_1, \theta_2)$ has variation independent components if the parameter space can be written as the Cartesian product $\Theta = \Theta_1 \times \Theta_2$, where Θ_1 and Θ_2 are sets of possible values for θ_1 and θ_2, respectively. It is, at first blush, easy to attribute the usefulness of the normal distribution to the fact that it can be so easily parameterized in this manner (either the standard (μ, σ^2) or mixed mean value and canonical $(\mu, \theta_2) = (\mu, -1/(2\sigma^2))$ parameterizations meet this criterion). But other distributions share this property without the same broad applicability (a gamma with canonical parameters, for example). One is led to the conclusion that the wide applicability of a normal distribution stems not only from the ease with which it is expressed in terms of variation independent parameter elements, but also the fact that those elements

quantify discrete characteristics of the distribution (location by μ and spread by σ^2). This does not hold for many other distributions that can be given variation independent parameters. Nevertheless, the notion is worth taking note of.

6.1.3 Exponential Dispersion Families

The name of this particular subsection is somewhat larger than its true content. We will not discuss exponential dispersion families in their full generality, but rather a certain subclass of families that are essentially one parameter families extended to include an additional dispersion parameter. This particular subclass of exponential dispersion families is, however, arguably the most common form of exponential family distributions that appear in applications at the current time.

An important role is played in both the theory and application of exponential family distributions by one-parameter families for which the sufficient statistic is $T(y) = y$. These are often called "natural exponential families" following the extensive investigation of their behavior by Morris (1982, 1983). If a family of distributions has only one canonical parameter, it is clear that both the expectation and variance of those distributions must be functions of the sole parameter.

Example 6.4

Consider the exponential form of a binomial random variable Y for a fixed number of associated binary trials n. The pmf of such a random variable is,

$$\begin{aligned} f(y|\theta) &= \exp\left[y\{\log(p) - \log(1-p)\} + n\{\log(1-p)\}\right. \\ &+ \left.\log\{n!\} - \log\{y!\} - \log\{(n-y)!\}\right] \\ &= \exp\{y\theta - b(\theta) + c(y)\}, \end{aligned}$$

where $\theta = \log\{p/(1-p)\}$ and $b(\theta) = n\log\{1+\exp(\theta)\}$. Here, using the facts that $T(y) = y$ and $b(\cdot)$ is a simple function, property 4 of canonical exponential families (given in Section 6.1.1) implies that

$$\begin{aligned} E(Y) &= n\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right) = np \\ var(Y) &= n\left(\frac{\{1+\exp(\theta)\}\exp(\theta) - \exp(2\theta)}{\{1+\exp(\theta)\}^2}\right) \\ &= np(1-p). \end{aligned}$$

Thus, both mean and variance are simple functions of the canonical parameter θ. Also notice that the variance can be written as $var(Y) = np - np^2 = \mu - \mu^2/n$, where $\mu = np$. This is the type of "quadratic variance function" referred to in the papers by Morris.

Example 6.5

Consider again a random variable $Y \sim N(\mu, \sigma_*^2)$ except for which σ_*^2 is now considered a fixed, known value. In this case we can write,

$$\begin{aligned} f(y|\mu) &= \exp\left[\frac{-1}{2\sigma_*^2}(y-\mu)^2 - \frac{1}{2}\log(2\pi\sigma_*^2)\right] \\ &= \exp\left[\frac{1}{\sigma_*^2}\left(y\mu - \frac{1}{2}\mu^2\right) - \frac{1}{2}\left\{\frac{y^2}{\sigma_*^2} - \log(2\pi\sigma_*^2)\right\}\right], \end{aligned}$$

which can be written as

$$f(y|\theta) = \exp\left[\phi\{y\theta - b(\theta)\} + c(y,\phi)\right],$$

for $\theta = \mu$, $b(\theta) = (1/2)\theta^2$, $\phi = 1/\sigma_*^2$, and $c(y,\phi)$ contains the remaining terms which involve only y and ϕ.

This latest expression, namely,

$$f(y|\theta) = \exp\left[\phi\{y\theta - b(\theta)\} + c(y,\phi)\right] \tag{6.5}$$

is the restricted version of an exponential dispersion family we will consider. For a distribution with pdf or pmf of the form (6.5) the same techniques as presented in Section 6.1.1 for general $s-$parameter exponential families may be used to demonstrate that,

$$\begin{aligned} E(Y) &= \frac{d}{d\theta}b(\theta) = b'(\theta), \\ var(Y) &= \frac{1}{\phi}\frac{d^2}{d\theta^2}b(\theta) = \frac{1}{\phi}b''(\theta) = \frac{1}{\phi}V(\mu). \end{aligned} \tag{6.6}$$

The rightmost portion of the expression for $var(Y)$ in (6.6) follows from the fact that $\mu = b'(\theta)$ so that $b''(\theta)$ is a function of μ. The function $V(\cdot)$ in (6.6) is often called the "variance function" (which is *not* the variance except for a few cases in which $\phi \equiv 1$) and is actually quite important since it quantifies the relation between the mean and variance of the distribution.

Comments

1. What has essentially happened in (6.5) is that we have "coerced" a two parameter exponential family to look "almost" like a natural exponential family (see Example 6.4) but with the addition of an extra parameter ϕ. The resultant form is the same as what we would get from a normal with known variance (see Example 6.5), except that we typically do not assume ϕ is known. It is sometimes relegated to the role of "nuisance parameter" which is a scale factor for the variance (see expression (6.6)) but it can also be of considerable interest; ϕ is often called a "dispersion parameter".

2. Clearly, it will not be possible to write an exponential family in the form of expression (6.5) unless one of the sufficient statistics is given by the identity function (i.e., $T_j(y) = y$ for some j). While this is not, in itself, sufficient for representation of a pdf or pmf as in (6.5), distributions for which one of the sufficient statistics is y and which can subsequently be written in this form include the binomial, Poisson, normal, gamma, and inverse Gaussian. But it is not possible, for example, with a beta pdf.

3. Exponential dispersion families of the form (6.5) are the exponential families upon which *generalized linear models* are based (e.g., McCullagh and Nelder, 1989) but, as already noted in the introduction to this part of the course notes, the impetus provided by generalized linear models to consider random model components in a more serious light than mere "error distributions" has much wider applicability than just these families.

4. The machinations required to render some distributions amenable to expression as in (6.5) provide an opportunity to reiterate the importance of Comment 5 of Section 6.1.2 about the possible effects of parameter transformations. A prime example is that of a gamma distribution. The standard approach for writing a gamma pdf as in (6.5) takes two parameters that are variation independent and maps them into parameters θ and ϕ. Now, these parameters are are also variation independent, but one should not be misled into believing that the parameter transformation involved has resulted in parameters that have separate effects on moments of the distribution. Even in exponential dispersion family form, both θ and ϕ effect all moments of a gamma distribution. This can, in fact, restrict the types of situations that can be represented by a model that represents a number of distributions as a gamma written

in the form (6.5) but assumes the dispersion parameter ϕ has the same value for each distribution.

6.1.4 Exponential Families for Samples

Thus far, we have dealt only with exponential family distributions for a single random variable Y. While there are a number of results that make exponential families a potentially useful vehicle for the construction of multivariate distributions in general (see e.g., Arnold and Strauss, 1991; Kaiser and Cressie, 2000) here we will consider the situation only for sets of independent random variables.

One additional property of exponential families will be useful in this subsection. We could have covered it in Section 6.1.1, but it really doesn't come into play until now. For Y distributed according to an exponential family as in (6.1) with $\theta = (\theta_1, \ldots, \theta_s)$, the sufficient statistic $(T_1, \ldots, T_s)$ is distributed according to an exponential family with density or mass function

$$f(\boldsymbol{t}|\theta) = \exp\left[\sum_{j=1}^{s} \theta_j t_j - B(\theta) + k(\boldsymbol{t})\right]. \tag{6.7}$$

Note that the dominating measure of the distributions of Y and $\boldsymbol{T}$ may differ, and $k(\boldsymbol{t})$ may or may not be easily derived from the original $c(y)$.

Consider now the case of n independent and identically distributed random variables $Y_1, \ldots, Y_n$, with each variable having a pdf or pmf of the form

$$f(y|\theta) = \exp\left\{\sum_{j=1}^{s} \theta_j T_j(y) - B(\theta) + c(y)\right\}.$$

Under the *iid* assumption, the joint distribution of $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)^T$ is,

$$f(\boldsymbol{y}|\theta) = \exp\left\{\sum_{j=1}^{s} \theta_j \sum_{i=1}^{n} T_j(y_i) - n\, B(\theta) + \sum_{i=1}^{n} c(y_i)\right\}. \tag{6.8}$$

Notice that expression (6.8) is still in the form of an exponential family, with sufficient statistics given by the sums of the $T_j(\cdot)$. In particular, let $Y_1, \ldots, Y_n$ be distributed according to a one-parameter exponential family. Then the joint distribution is again a one-parameter exponential family with the same canonical parameter and sufficient statistic given as the sum $\sum_{i=1}^{n} T(Y_i)$.

Example 6.6

Suppose that $Y_1, \ldots, Y_n$ are distributed *iid* following a Poisson distribution with $E(Y_i) = \lambda$, that is,

$$f(y_i|\lambda) = \frac{1}{y_i!}\,\lambda^{y_i}\,\exp(-\lambda); \qquad y_i = 0, 1, \ldots, ; \quad \lambda > 0,$$

which is a one-parameter family, and can be written for $\theta \equiv \log(\lambda)$ as,

$$f(y_i|\theta) = \exp\left[y_i\,\theta - b(\theta) + c(y_i)\right],$$

where $b(\theta) = \exp(\theta)$ and $c(y_i) = -\log(y_i!)$. Then the joint distribution of $Y_1, \ldots, Y_n$ is,

$$f(y_1, \ldots, y_n|\theta) = \exp\left[\theta \sum_{i=1}^{n} y_i - nb(\theta) + \sum_{i=1}^{n} c(y_i)\right],$$

Notice here that, using the property of exponential families listed as property 4 in Section 6.1.1, which is the same here as that of expression (6.6), we immediately have that,

$$E\left\{\sum_{i=1}^{n} Y_i\right\} = n\,b'(\theta) = \frac{d}{d\theta} n \exp(\theta) = n\exp(\theta),$$

so that $E(\bar{Y}) = \exp(\theta) = \lambda$ which we already know. What may not be so obvious is that the distribution of $W \equiv \sum_{i=1}^{n} Y_i$ is also now available as ,

$$f(w|\theta) = \exp\left[w\theta - b^*(\theta) + c^*(w)\right],$$

which is in the basic form of a one-parameter exponential family with canonical parameter θ, and we know that $b^*(\cdot) = nb(\cdot)$. We do not know $c^*(\cdot)$ directly from knowledge of $c(\cdot)$, but in this case property 5 of exponential families from Section 6.1.1 indicates that

$$\begin{aligned} M_W(u) &= \frac{\exp\{nb(\theta+u)\}}{\exp\{nb(\theta)\}} \\ &= \frac{\exp\{n\exp(\theta+u)\}}{\exp\{n\exp(\theta)\}} \\ &= \frac{\exp\{\exp(\log(n)+\theta+u)\}}{\exp\{\exp(\log(n)+\theta)\}}, \end{aligned}$$

which is the form of a Poisson moment generating function for canonical parameter $\log(n)+\theta$. Thus, the distribution of W is also Poisson.

6.2 Location-Scale Families

A larger topic of which this subsection is only a portion is that of *group* (e.g., Lehmann, 1983) or *transformation* (e.g., Lindsey, 1996) families of distributions. While families of distributions formed from classes of transformations holds potential for greater use in applications than has been the case to date, we will restrict attention here to what is certainly the most important case, that of *location-scale* transformations.

Let U be a continuous random variable with a fixed distribution F (typically we will assume U has pdf or pmf f). If U is transformed into Y as

$$Y = U + \mu,$$

then Y has distribution $F(y-\mu)$ since $Pr(Y \leq y) = Pr(U \leq y-\mu)$. The set of distributions generated for a fixed F, as μ varies from $-\infty$ to ∞, is called

a *location family* of distributions generated by F. If the resultant distribution is of the same form as F only with modified parameter values, then F forms a location family. A similar definition of a distribution F forming a *scale family* is if F is unchanged other than parameter values under transformations

$$Y = \sigma U; \quad \sigma > 0,$$

in which case the distribution of Y is $F(y/\sigma)$ since $Pr(Y \leq y) = Pr(U \leq y/\sigma)$.

The composition of location and scale transformations results in,

$$Y = \mu + \sigma U; \quad -\infty < \mu < \infty; \quad \sigma > 0,$$

and Y has distribution $F((y-\mu)/\sigma)$. If F has a density f, then the density of Y is given by

$$g(y|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right).$$

Location-scale families include the double exponential, uniform, and logistic, but by far the most frequently employed member of this class of distributions is the normal. As we will soon see, location-scale families are well suited for modeling true error processes.

6.2.1 Properties of Location-Scale Families

Location-scale families have beautifully simple properties that stem directly from the transformations. For example, if Y is produced as a location transformation of U then

$$E(Y) = E(U + \mu) = E(U) + \mu.$$

While the notion of a "parent" distribution is somewhat misleading for such families (since we must be able to arrive at any member from any other member

through the same family of transformations, see Lehmann, 1983, p.25) we often begin in the development of models with a random variable U for which $E(U) = 0$ so that $E(Y) = \mu$. Similarly, for a scale transformation, the variance of Y is,

$$var(Y) = var(\sigma\, U) = \sigma^2 var(U),$$

and we often begin with a random variable U such that $var(U) = 1$.

Although we have not emphasized it, we are assuming here that the location and scale transformations to be made are closed under composition and inversion. This is, in fact, important in the definition of a group family (e.g., Lehmann, 1996, Chapter 1.3). Caution is needed, for example, in situations for which location transformations are defined only for positive shifts, since then the class of transformations $Y = U + \mu; \quad \mu > 0$ is not closed under inversion. It is clear that location-scale transformations "work" if the location transformation is defined for any $-\infty < \mu < \infty$ and $\sigma > 0$.

6.2.2 The Prominence and Limitations of the Normal Distribution

Without question, the most commonly used distribution in statistical modeling is the normal. Why is this? Is it due, as is sometimes asserted, to some mystical "ubiquitous occurrence" of the normal distribution in nature? Is it due, as is sometimes asserted, to the "mathematical tractability" of its properties? Is it due, as is sometimes asserted, to the fact that so many distributions seem to result in weak convergence to a normal law? The answer seems to be, at the same time, all and none of these explanations. While the

normal is frequently useful in describing the relative frequencies of observable quantities, the first notion has been soundly discredited. While the second point is certainly true, this does not really explain why the normal is so useful for modeling purposes (unless we can give greater detail to which properties we refer to). Mathematical properties are important (a model is no good unless you can do something with it) but a useful model must be an adequate vehicle for statistical abstraction, as discussed in Chapter 5.3. There must be more to this than merely mathematical nicety. The third notion goes a long way toward explaining the prominence of the normal distribution in the production of *inferential quantities*, but not its use in modeling.

What properties does a normal distribution possess that renders it attractive for modeling? The following seem pertinent:

1. A normal density may be easily expressed in terms of variation independent parameters. This is certainly an important property but, as also noted previously, does not distinguish the normal from many other distributions.

2. In a $N(\mu, \sigma^2)$ parameterization, the individual parameter values quantify basic characteristics of a distribution, namely location and spread. But this is also true of other distributions and, in particular, many location-scale families such as the logistic distributions (although the scale parameter here is proportional to the variance rather than equal to the variance).

3. In samples, the normal allows a reduction through sufficiency. That is, the dimension of the minimal sufficient statistic is less than that of the number of random variables. This is generally true of exponential families but not location-scale families (e.g., the minimal sufficient statistic

for a set of *iid* logistic random variables is the set of order statistics, see Lehmann, 1983, p.43). This is something that comes close to being unique for the normal among location-scale families. In fact, the only two location families that are also exponential families are the normal (with fixed variance) and the distribution of any real constant times the log of a gamma random variable (e.g., Dynkin, 1951; Ferguson, 1963).

In summary, what the normal distribution possesses in terms of statistical properties that, to my knowledge make it unique, is a combination of all of the above; variation independent parameterizations in which parameters independently represent fundamental distributional characteristics, minimal sufficient statistics of small dimension, and membership in exponential families, exponential dispersion families, and location or location-scale families.

Despite its undeniable appeal and usefulness, the normal distribution also possesses some limitations, three general types of which may be given as:

1. Since the normal is a distribution for continuous random variables, the most obvious limitation may occur in situations which involve discrete random variables. Sometimes, this is a matter of little concern, since all data are discrete regardless of the statistical conceptualization. Consider, for example, a model for the number of stars in various portions of the galaxy. Clearly, this situation should be conceptualized by discrete random variables. On the other hand, the numbers are large, and when one is dealing with a range of integer data in the thousands of values, the discrete nature of the actual situation may become relatively unimportant. On the other hand, consider the number of deaths due to SARS. Here, the situation is also one of counts, but the discrete nature of the problem may be more important to accurately reflect in a model. The

issue concerns the ratio of the interval between observable values to the total range of data likely to be observed; in the star example this ratio might be 10^{-5}, while in the SARS example it is more likely to be 10^{-1} or less. It should also be noted that this is not a "potential limitation" unique to the normal distribution, but is shared by any distribution for continuous random variables.

2. Since the normal is characterized by its first two moments (expectation and variance), its usefulness in situations for which the general *shape* of a distribution is important is limited. That is, all normal distributions are unimodal and symmetric. Consider the implications, for example, of conducting a two sample $t-$test under the assumption of normally distributed groups of random variables. The model that corresponds to this situation may be written as, for $k = 1, 2$,

$$Y_{k,i} = \mu_k + \sigma\, \epsilon_{k,i}; \qquad \epsilon_{k,i} \sim \; iid\, N(0, 1).$$

What are the implications of this model? First, all of the random variables involved have unimodal and symmetric distributions with tails that die off at the same rate. Secondly, all of the random variables involved have distributions with the same spread or variability. Thus, this simple model has restricted the two groups of random variables to have identical distributions with the possible exception of location (one may wish, at this point, to revisit comment 2 of Chapter 4.4 in Part I of these notes).

3. Since the set of possible values of a normal random variable is the entire real line, a model that relies on normal distributions may place positive probability on sets that are physically impossible for some problems. Some of the ramifications of applying a model based on normal distrib-

utions in such a situation are illustrated in the following example.

Example 6.7

While the data used in this example are simulated, the setting is hypothetically real; that is, the problem is real. There are certain types of algae that produce toxins. Ingestion of these toxins can lead to adverse health effects, including death in rare instances (and yes, there have been deaths in the US attributed to this cause, as well as elsewhere in the world). In North America, the most predominant such toxin-producing algae is a genus called *Microcystin*. It is of interest to limnologists (scientists who study water quality and water chemistry), as well as public health officials, to determine factors that may be related to the concentration of these algae in lakes and reservoirs. A study was conducted in which the concentration of *Microcystin spp.* and various water chemistry variables were measured. Exploratory analyses suggested that a model of *Microcystin* concentration versus the nitrogen concentration of waterbodies might be useful in describing the situation in the Midwestern US, and could potentially lead to prediction of possible problem waters. Little is known about the manner in which various water chemistry variables may be related to *Microcystin* abundance (i.e., concentration) in lakes and reservoirs. Consider the (hypothetical) data presented in Figure 6.1, a scatterplot of *Microcystin* concentration versus nitrogen concentration in a collection of lakes and reservoirs in the Midwestern United States. In the absence of scientific knowledge about the way that nitrogen concentration should be related to *Microcystin* abundance, a linear regression model would seem to be a logical choice to describe these data. Let $Y_i;\ i = 1, \ldots, n$ be random variables associated with the concentration of *Microcystin* in waterbody i. Let $x_i;\ i = 1, \ldots, n$ denote

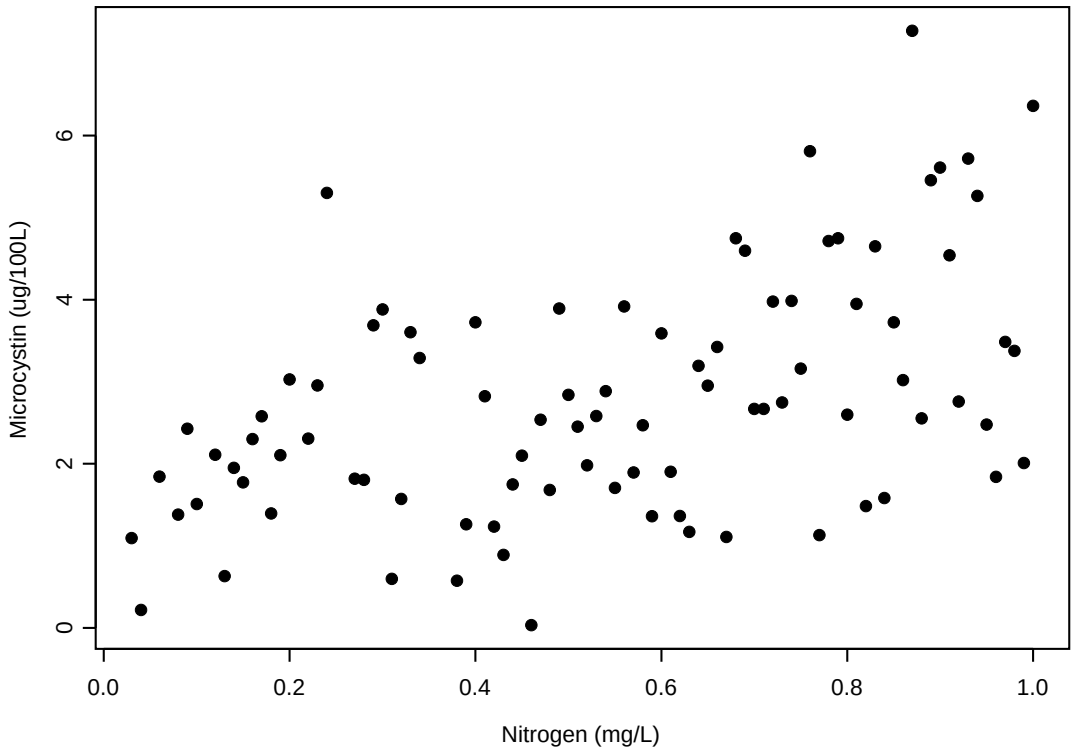

Figure 6.1: Scatterplot of simulated data for concentration of *Microcystin* versus concentration of nitrogen in lakes and reservoirs.

the corresponding nitrogen concentrations in those same waterbodies. Now we know that *Microcystin* concentration cannot assume negative values, and it is true that some nitrogen values in Figure 6.1 are quite small, so that a normal linear regression model would specify some conditional distributions (conditional on x_i) that likely would place positive probability on the negative portion of the real line (likely given the spread of data exhibited in Figure 6.1). But, under the assumption that a normal distribution provides a reasonable approximation even in these cases, and knowing also that the tails of normal distributions "die out" fairly quickly, we might proceed to fit a normal linear model of the form,

$$Y_i = \beta_0 + \beta_1 \, x_i + \sigma \, \epsilon_i, \tag{6.9}$$

where $\epsilon_i \sim iid \, N(0,1)$ for $i = 1, \ldots, n$. Although we have not discussed estimation and inference yet, you know from Statistics 500 and Statistics 511 that ordinary least squares estimators for this model are minimum variance among linear unbiased estimators (i.e., are BLUE), and have normal sampling distributions which may be used for inference. The ols estimates of model (6.9), estimate of σ^2 by the usual bias-corrected moment estimator, associated standard errors and 95% intervals based on the data of Figure 6.1 are:

Parameter	Estimate	Std. Error	95% Interval
β_0	1.381	0.2999	(0.785, 1.977)
β_1	2.564	0.4907	(1.589, 3.539)
σ^2	1.658		

So far, so good, but let's examine residuals to determine whether the assumptions of normal error terms having constant variance appears to have caused any problems. Figure 6.2 presents a plot of studentized residuals against fitted

values. Here, fitted values are given by:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \, x_i,$$

raw residuals are defined as $r_i = Y_i - \hat{Y}$, and these values are studentized as, for $i = 1, \ldots, n$,

$$b_i = \frac{r_i}{\{\hat{\sigma}^2 \, (1 - h_{ii})\}^{1/2}},$$

where h_{ii} denotes the ith diagonal element of the "hat matrix",

$$H = X \, (X^T \, X)^{-1} X^T,$$

where X is the $n \times 2$ matrix with ith row $(1, \, x_i)$. The hat matrix is also sometimes called the "projection matrix", and notice that $\hat{Y} = H \, Y$. The diagonal elements of this matrix are also called the "leverages" of the observations, and for this simple linear regression model may be computed as,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $\bar{x}$ is the sample mean of the $\{x_i : \, i = 1, \ldots, n\}$. The standardization (studentization) of residuals adjusts for the fact that the raw residuals do not have constant variance, even if model (6.9) holds. These residuals are often plotted against covariates and/or fitted values as diagnostics for departures from model assumptions. In simple linear regression, plots of studentized residuals against the x_is and the $\hat{Y}_i$s are equivalent (since the $\hat{Y}$s are linear transformations of the x_is). The residual plot of Figure 6.2 does not indicate any drastic problems with the model. One might be tempted to "read too much" into this plot, an urge that should be resisted; in the vast majority of applications, this residual plot would be welcomed as indicating no problems.

What about the assumption of normal distribution for the error terms? Figure 6.3 presents a histogram of the studentized residuals.

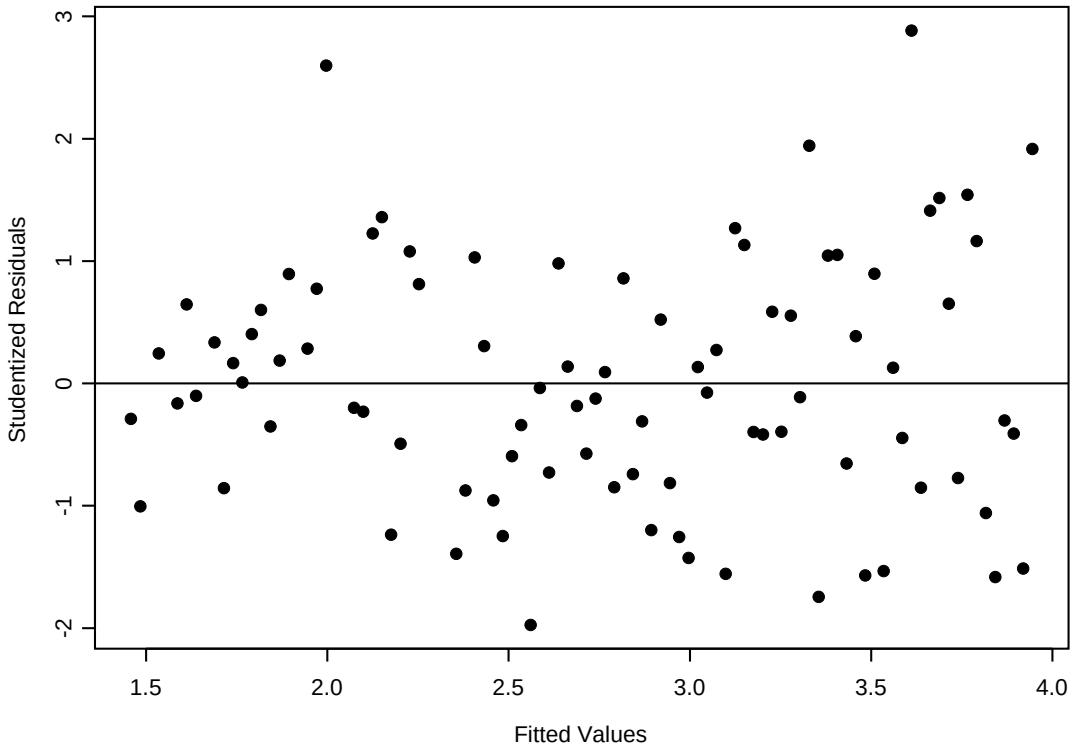

Figure 6.2: Studentized residuals from ols fit to the data of Figure 6.1.

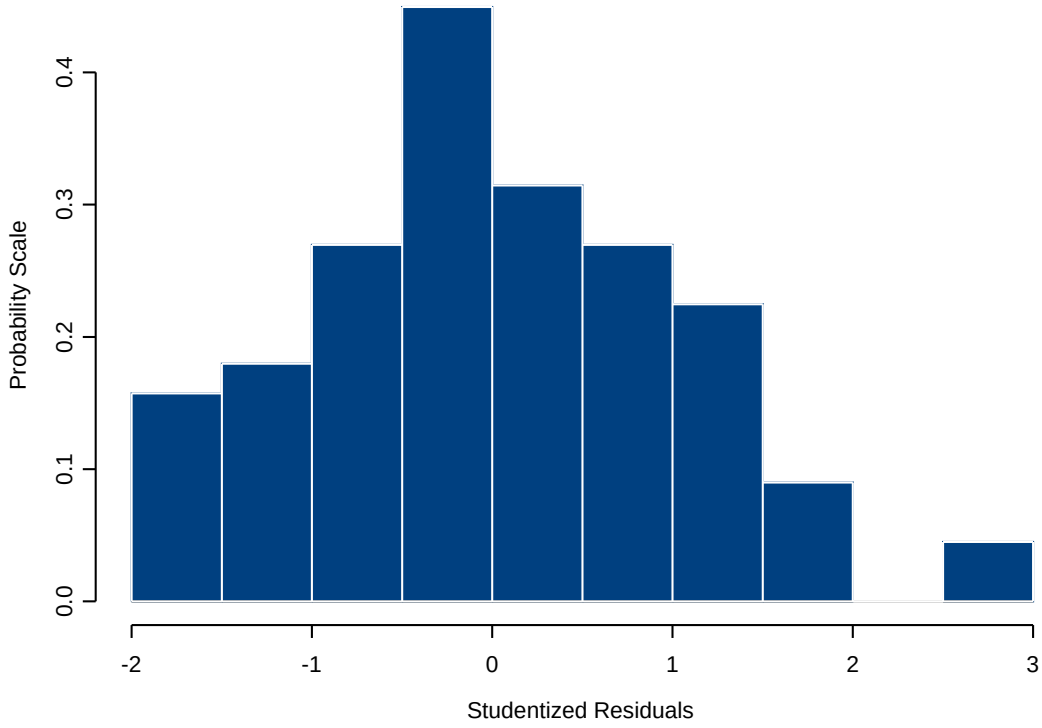

Figure 6.3: Histogram of studentized residuals from ols fit to the data of Figure 6.1.

While other graphical displays often present more information than histograms (e.g., normal probability plots) here the visual impression of the histogram matches what can also be seen in the scatterplot of Figure 6.1. The left tail of the histogram is rather "bunched up" relative to the right tail and the residual quantities appear almost "truncated" on the left. A re-examination of the scatterplot of Figure 6.1 indicates the absence of negative response values for small covariate values, a sort of "missing piece" of what should be described by the model (6.9); of course in this example we know such values are physically impossible.

Is this, however, indicative of a problem for model (6.9) in describing the data of Figure 6.1? The histogram of residuals in Figure 6.3 is, after all, unimodal and relatively symmetric, and has roughly the correct range of values for standardized normal quantities; we must always keep in mind that any such plot will exhibit effects of sampling variability (recall the ease with which the data in Example 4.1 of Part I were simulated from a normal model, see comment 4 of Chapter 4.5). A Kolmogorov goodness of fit test for normality conducted with the studentized residuals results in a $p-$value of 0.924, hardly indicative of any problem with an assumption of normality. All in all, our linear normal regression model (6.9) appears to have done a quite adequate job in describing the relation between *Microcystin* concentration and nitrogen concentration in the data at hand.

Our real concern with model (6.9) in this situation is (or should be) not totally centered on whether it provides an adequate description of the observed data. The degree to which a model provides an adequate "fit" to a set of data is only one indicator of the adequacy of the model in conceptualizing the underlying *problem*; it is important (necessary) but is not sufficient for a model to adequate. Statistical abstraction, as discussed in Chapter 5.3, involves more

than "fitting data". Our concern with model (6.9) for this example is the use of normal distributions, having infinite support, to conceptualize a process that is strictly non-negative. This is an inadequacy of the normal model. The question is whether that inadequacy is important in the current problem, or whether it can be reasonably ignored. Consider estimation of the probabilities with which the response is negative, $Pr(Y_i < 0|x_i, \beta_0, \beta_1)$. For model (6.9) these probabilities can be estimated from the predictive distribution with mean $\hat{Y}_i$ and estimated variance

$$\hat{\sigma}^2 \left\{ 1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}.$$

A plot of these estimated probabilities against the values of the x_is (nitrogen concentration) is presented in Figure 6.4. For the data observed, points at which these probabilities become just less than the 10%, 5%, and 1% levels are $x_i = 0.12$, $x_i = 0.30$ and $x_i = 0.64$, respectively. The range of covariate values is 0 to 1 in these data. Thus, we must move to greater than roughly a third of the entire range of covariate values before the estimated probability of a negative response becomes smaller than 5%. This is not a pleasing aspect of the model.

One might argue that, despite this difficulty with a normal model, this should not affect the prediction of *exceeding* a given response value. That is, estimating probabilities of small *Microcystin* values is not as much of a concern as estimating large values. To assess this requires knowledge of the "truth" which is possible here since the data were simulated. In fact, the model used to simulate the data of Figure 6.1 was,

$$Y_i = \{0.25 + 3.75\, x_i + \sigma\, \epsilon_i\}\, I\{\sigma\epsilon_i > -(0.25 + 3.75 x_i)\}, \qquad (6.10)$$

where $\epsilon_i \sim iid\, N(0, 1)$ for $i = 1, \ldots, n$. The relation between nitrogen concentration (x_i) and *Microcystin* concentration (Y_i) is embodied by the linear

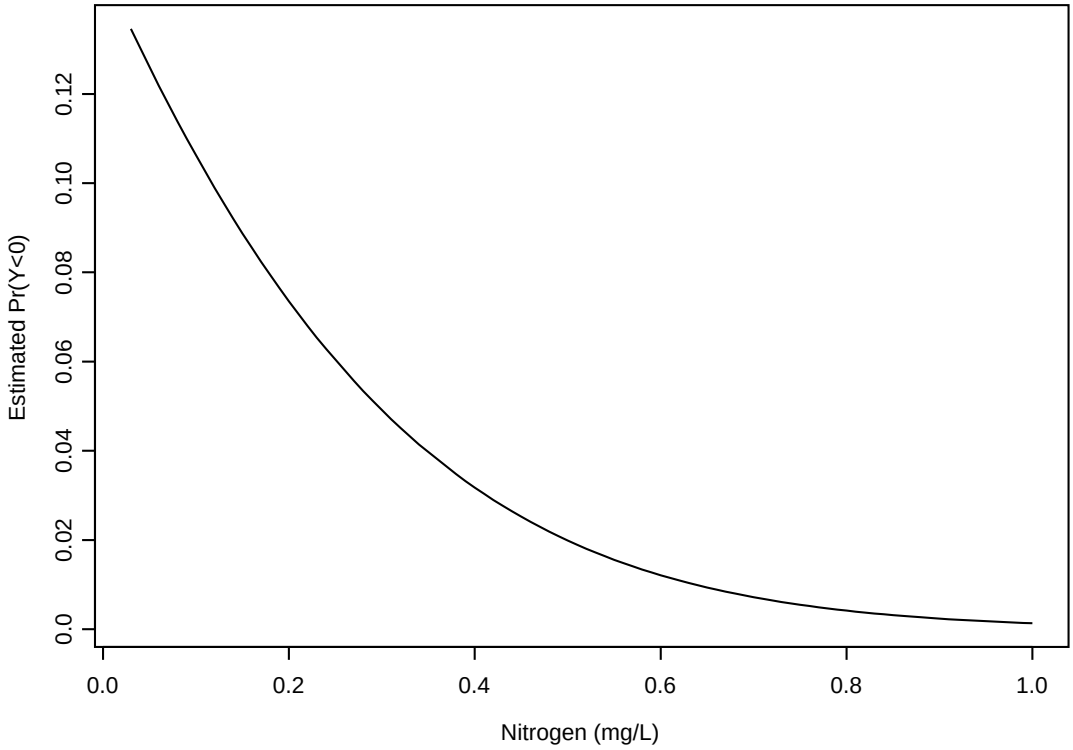

Figure 6.4: Estimated probabilities that *Microcystin* concentration is less than zero, calculated from a normal linear model and plotted against nitrogen concentration.

equation $0.25 + 3.75x_i$, although this is no longer the expected value of Y_i. It is possible with this model, however, to compute the probabilities that the Y_i exceed a given value such as 3.0, for example. Suppose that $Y_i > 3.0$ is indicative of conditions in which *Microcystin* pose a potential health hazard. Estimating these probabilities under model (6.9) and under the true model (6.10) shows that the probabilities are uniformly over-estimated under model (6.9) (i.e., positive bias). Figure 6.5 describes the magnitudes of over-estimation as a function of nitrogen concentrations.

Consider a hypothetical situation in which some action would be taken (e.g., public access to a waterbody restricted) if the probability of *Microcystin* exceeding a value of 3.0 were estimated to be greater than 0.5. Measuring nitrogen concentration in a lake or reservoir is relatively easy and inexpensive compared to sampling for and measurement of *Microcystin*, so this decision will be based on observed values of nitrogen and the associated probabilities that *Microcystin* exceeds 3.0. These probabilities would be greater than 0.50 (the decision rule) under a fit of model (6.9) at $x_i = 0.64$, while the actual value should be $x_i = 0.74$ (a difference of 10% of the total range in the x_is).

Interestingly, the estimation errors shown in Figure 6.5 are greater for larger values of x_i than for smaller ones. This illustrates that model inadequacies in one portion of the data (here for small x_i values) can have effects that are not restricted to these portions of the "data space". It would clearly be a mistake to assume that, because the model inadequacy should occur primarily at small values of the covariate, it is only if we have interest in some aspect of the problem connected with small covariate values that we need to worry about the model inadequacy.

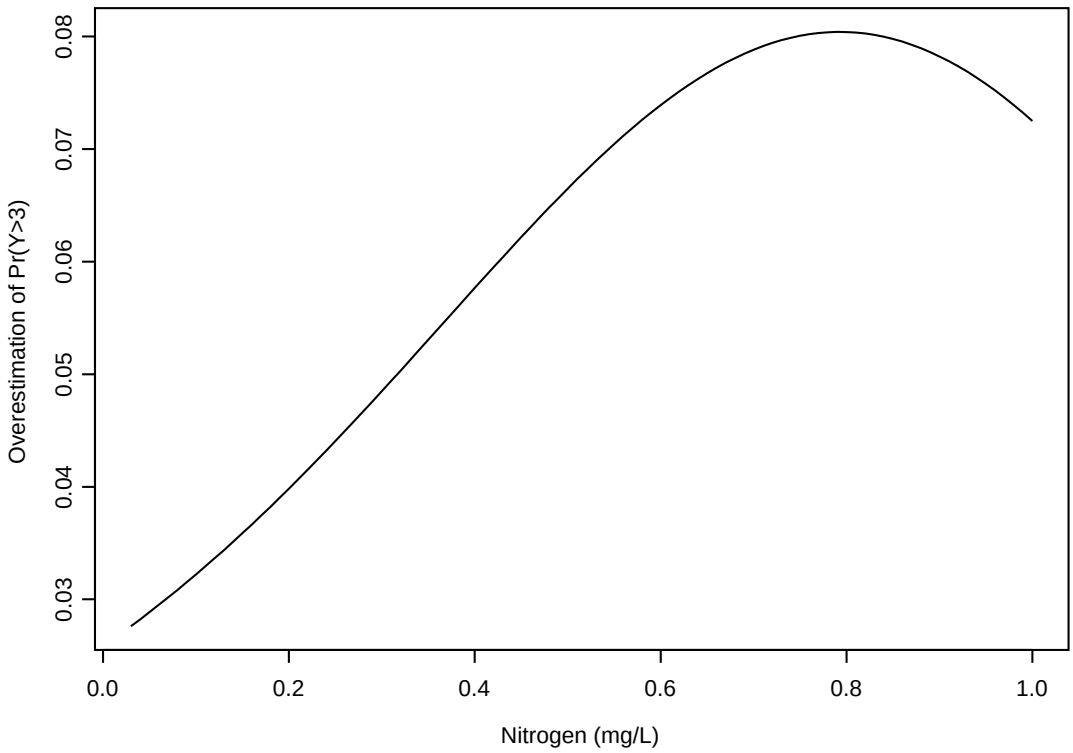

Figure 6.5: Overestimation of probabilities that *Microcystin* concentration is greater than 3 from normal linear regression model fit to the data of Figure 6.1.

Chapter 7

Basic Methods of Model Specification

In this chapter we will discuss a number of fundamental ways that models can be specified. Although various of these methods of model formulation do result in particular "types" or "classes" of models (e.g., generalized linear models) we will attempt to maintain a focus on modeling concepts and methods rather than models of certain "brand names". This is, in part, to help prevent our thinking from becoming too modular and, in part, because certain estimation and inference methods have become closely associated with various classes of models. We want to avoid the idea that various models are "always" estimated with a certain method or that a certain type of estimation is "correct" for a given model type.

7.1 Modeling and the Objectives of Analysis

Before launching into the subject of methods by which models are formulated, we want to briefly consider the role that the objectives of an analysis may play in modeling considerations. Although not meant to be exhaustive, the following list covers many of the objectives, and the broad implications that these might have for model formulation, that are frequently addressed by modeling efforts:

1. Data Description.

   I almost left this one off because it is so incomplete, but many statisticians would accuse me of heretical behavior if it was not included. It is almost a no-brainer, the idea that we may use a model to describe patterns in observed data. But this is incomplete in that if this is *all* that is desired we are probably better served by using purely descriptive approaches such as nonparametric smoothers or exploratory approaches rather than parametric models, through which we try to achieve statistical abstraction.

2. Problem Conceptualization.

   This is what I would replace data description with. It has been discussed at some length in Chapter 5.3 under the title of *Statistical Abstraction*. One point that has not been made previously is that it is often useful to focus on the very basic or very fundamental tenets of a scientific discipline when formulating a model for the purpose of problem conceptualization.

3. Examination of Scientific Theory.

   It is sometimes the case that a body of scientific knowledge has produced a particular theory. Some such theories may be in the form of a "law",

such as Newton's law of cooling which states that for the temperature of a material T, ambient (surrounding) temperature T_0, and time t,

$$T = f(t, \beta) \text{ such that } \frac{df}{dt} = -\beta(f - T_0).$$

We may be faced with an empirical study designed to "test" such a theory.

In other situations, scientific theory may lead to more complex representations, such as sets of differential equations meant to describe atmospheric processes (these are, for example, commonly used in models to predict climate or, on a finer scale, weather). Such theory often leads to what are called *process models* in the geophysical sciences (the term "model" here is typically a totally deterministic model, not a probabilistic model). In complex situations, no one claims such models are complete, but we may be faced with a situation in which the objective is to construct a statistical model based on a process model and examine its usefulness from a set of observed data. This is not always as straightforward as it may sound. A large class of problems in the geophysical sciences are known as "inverse problems" in which a response of interest, such as permeability of a geological formation, is to be estimated based on observations of a few direct measurements of permeability (difficult to obtain), a greater collection of connected quantities such as seismic readings (essentially travel times of sound or radar waves between sensors at fixed locations) and a body of scientific theory that relates the two.

4. Estimation and Inference About a Specific Quantity.

   It may be that, particularly if the process of problem conceptualization

or statistical abstraction has been well addressed, interest may focus on a particular parameter, function of parameters, or other functional of the distributional model (e.g., cumulative probabilities). In these situations, the focus of model specification is clear – primary interest centers on the quantity of interest, and other model aspects that may affect the precision with which it can be estimated.

5. Prediction or Forecasting.

   We often utter the words prediction and forecasting in the same breath (or the same sentence or the same item title as above). In this course we will distinguish between the two as they have seriously different ramifications for both model formulation and analysis. We will refer to "prediction" as the prediction of an unobserved random variable or functional of a distribution that is given (conceptualized) existence within the spatial and/or temporal extent of a set of data. The prediction of the level of an air pollutant at a spatial location in the interior of a region in which observations are made at a given time would be an example. The prediction of the number of occupants of a rental property within a 3 mile radius of the Iowa State Campus center in 2003, based on a (possibly) random sample of such properties, would be another. If prediction is the primary objective, we may choose to model data patterns that have no ready explanation (e.g., trend over time) with mathematical structures that have no ready interpretation (e.g., polynomial regression). Forecasting, on the other hand, we will take to mean the prediction of random quantities that are given (conceptual) existence outside the spatial and/or temporal extent of the available data. This is a fundamentally different task, as the description of patterns in data that have no ready

explanation form dangerous structures on which to base a forecast.

We end this brief discussion of objectives with an indication that the categories given above, while probably not exhaustive, are certainly not mutually exclusive. A paper by Raftery, Givens, and Zeh (1995) for example, concerns a modeling problem in which a major element was the (statistical) conceptualization of a problem in a manner that necessarily incorporated deterministic models from scientific theory, with the ultimate goals being estimation and forecast of a particular meaningful quantity.

7.2 Additive Error Models

A basic concept in statistical modeling, and one with which you are familiar from previous courses, is the use of additive error models. The basic concept is epitomized by the following quote from a book on (both linear and nonlinear) regression analysis by Carroll and Ruppert (1988):

> When modeling data it is often assumed that, in the absence of randomness or error, one can predict a response y from a predictor x through the deterministic relationship
>
> $$y = f(x, \beta)$$
>
> where β is a regression parameter. The [above] equation is often a theoretical (biological or physical) model, but it may also be an empirical model that seems to work well in practice, e.g., a linear regression model. In either case, once we have determined β then the system will be completely specified.

These authors proceed to discuss reasons why the deterministic relation between y and x may not hold in practice, including measurement error (potentially in both y and x), slight model misspecification, and omission of important covariates.

The model form that results is that of an additive error model, in our notation, for $i = 1, \ldots, n$,

$$Y_i = g(x_i, \beta) + \epsilon_i, \tag{7.1}$$

where g is a specified function, $\epsilon_i \sim iid\, F$ with F an absolutely continuous distribution function with density f and, typically, $E(\epsilon_i) = 0$.

The model (7.1) is a direct mathematical expression of the concept that observable quantities arise from scientific mechanisms or phenomena that can be represented as "signal plus noise". The typical assumption that "noise" has expectation 0 renders "signal" the expected value of responses, that is, $E(Y_i) = g(x_i, \beta)$.

The modeling task with an additive error specifications largely centers on two issues:

1. Appropriate specification of the function g.

2. Modeling of the variance of the additive errors ϵ_i; $i = 1, \ldots, n$.

The first of these, specification of the expectation function g can be approached either through scientific knowledge (re-read the brief description of the objective of examining scientific theory of Chapter 7.1) or through what is essentially an arbitrary selection of some function that "looks right" based on examination of the data.

Notice that the general form (7.1) encompasses situations involving the comparisons of groups. For example, we may define x_i to be an indicator of

group membership as $x_i \equiv j$ if $Y_i \in$ group j, and

$$g(x_i, \beta) = \beta_j \quad \text{if } x_i = j \ ,$$

which could then constitute a one-way ANOVA model, depending on how the distribution of the ϵ_i are specified. Also, model (7.1) includes group regression equations if, for example, we define $x_i \equiv (j, z_i)$, where j is an indicator of group membership as before, z_i is a continuous covariate associated with the random variable Y_i, and, for example,

$$g(x_i, \beta) = g(j, z_i, \beta) = \beta_0^j \exp\{-\beta_1^j z_i\}.$$

Comment

Notice that we have, to a large extent, avoided using multiple subscripting (e.g., $Y_{i,j}$ for response variable i in group j) and have also written expressions for individual (univariate) random variables. This is a convention we will try to adhere to throughout the semester. Multivariate random variables are simply collections of univariate variables, and vector and matrix notation are simply convenient ways of reducing notation (primarily in the case of linear models). There is no notion, for example, of a vector expectation operator; the expectation of a vector is merely the vector of expectations for the individual random variables included. Expectation and other properties of random variables are only *defined* for scalar quantities. Everything else is just notation.

The digression of the preceding comment aside, the fundamental concept involved in the specification of additive error models is that of signal plus noise, with noise consisting of sources of error that combine in a simple manner with a correctly specified signal given as an expectation function. Additive error models are clearly well suited for use with location-scale families of distributions for modeling the error terms. That is, the expectation function $g(x_i, \beta)$

in (7.1) constitutes a location transformation of the error random variables $\{\epsilon_i : i = 1, \ldots, n\}$. What remains in model formulation is to specify a model for scale transformations (or the variances) of these error terms, as indicated in the second item in the list of modeling tasks that follow expression (7.1). It is this portion of the model formulation that renders additive error models a viable option for many situations. It is nearly ubiquitous, however, that the location-scale family chosen for specification of the error distribution is the normal, the reasons for which are touched on in Section 6.2.2. We will briefly consider here four situations for modeling the variance of the error terms in (7.1); constant variance, variance models with known parameters, variance models with unknown parameters, and what are called "transform both sides" models.

7.2.1 Constant Variance Models

Models that specify a constant variance for the error terms $\{\epsilon_i : i = 1, \ldots, n\}$ perhaps form the "backbone" of statistical modeling as applied to much of scientific investigation; it is worthy of note, however, that this "backbone" is becoming more "cartilaginous" as computational power increases. The reason for the historical (at least) prominence of constant variance models may be the fact that "exact" or "small sample" theory can be developed for linear models with additive normal errors that have constant variance, but for few other situations.

Curiously, statisticians have had the tendency to hang on to this "gold standard" idealization despite the fact that it is of limited application. What do we (we meaning statisticians) typically teach individuals learning basic regression methods in situations for which a linear, constant variance model does not

appear to be appropriate? Why, *transformation* of course. Transform (usually) the response variables so that they more nearly meet the assumptions of a linear expectation function and normally (or at least symmetrically) distributed error terms with constant variance. No matter that the transformed scale of measurement may be totally inappropriate for scientific inference, the statistical gold standard has been achieved.

The above assessment is unnecessarily harsh. Constant variance models and, in particular, linear constant variance models, are highly useful, both in their own right and as "baseline" formulations that allow modification to more complex structures. The intention of the negative comment relative to linear constant variance models is to help us escape from the idea that this is what statistical modeling is all about. Under what situations, then, does one naturally turn to a constant variance model as the *a priori* choice for model formulation? Fundamentally, in situations for which the assumption of a deterministic relation between a response quantity and a covariate is plausible in the absence of measurement error. These situations are common in studies for which the objective is essentially that of testing scientific theory (see the first portion of item 3 of Chapter 7.1). Bates and Watts (1988) present any number of examples of such situations.

Example 7.1

One of the examples presented in Bates and Watts involves enzyme kinetics for a given enzyme treated with Puromycin (see Bates and Watts, 1988, Figure 2.1 and Appendix A1.3). In this example, response random variables were associated with the "velocity" of a chemical reaction (measured in counts of a radioactive substance per squared minute), and a covariate of substrate

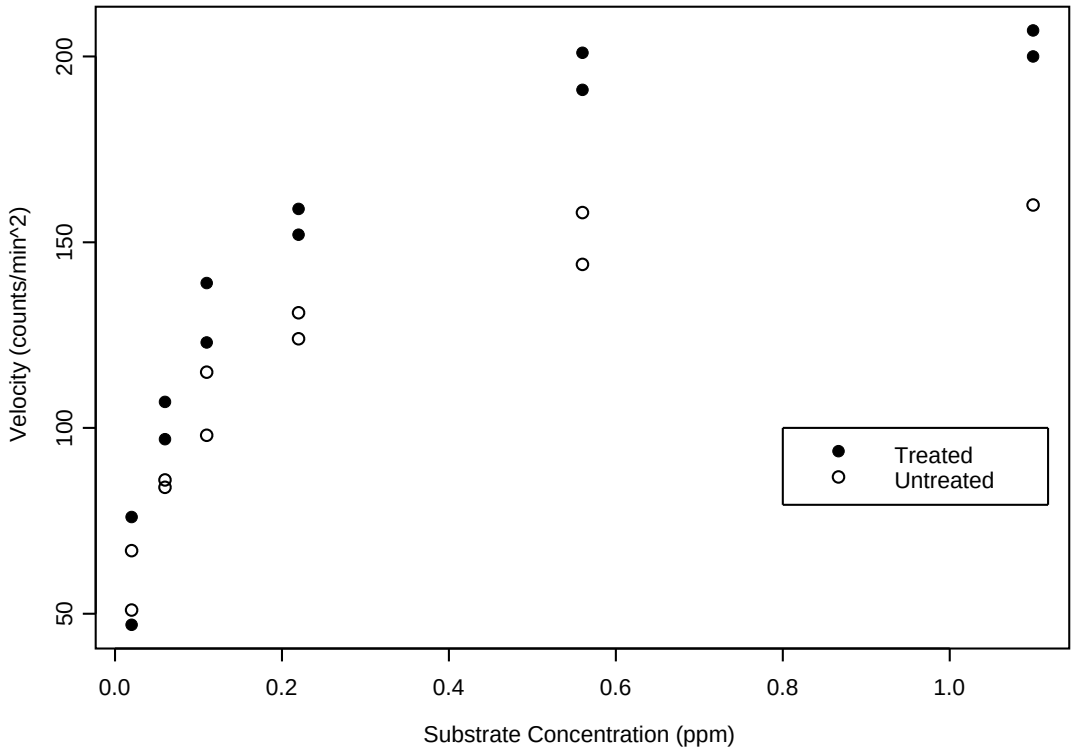

Figure 7.1: Scatterplot of data on the "velocity" of an enzyme reaction on substrate treated with Puromycin and untreated substrate.

concentration (what the substrate was is not identified by Bates and Watts). These data, from an original source cited in Bates and Watts, are reproduced in Figure 7.1. It can be seen that the variability of these data about a reasonable expectation function should be small. It was hypothesized in this example that the data could be described by a Michaelis-Menten equation, which relates the theoretical velocity of an enzyme reaction to the associated substrate concentration. For one group of random variables from this example (treated or untreated) the Michaelis-Menten equation can be expressed as in model

(7.1) with,

$$g(x_i, \beta) = \frac{\beta_1 x_i}{\beta_2 + x_i},$$

where x_i represents the substrate concentration for observation i.

Bates and Watts (1988, page 35) point out that it is possible to transform the Michaelis-Menten equation to have a linear form by taking the reciprocals of both sides of the equation.

$$\begin{aligned} \frac{1}{g(x_i, \beta)} &= \frac{\beta_2 + x_i}{\beta_1 x_i} \\ &= \frac{1}{\beta_1} + \frac{\beta_2}{\beta_1}\frac{1}{x_i}, \end{aligned}$$

and this is in the form of a linear model $y' = \beta_0' + \beta_1' x_i'$ say. What happens to the data plot of Figure 7.1 if we use this transformation?

It is clear from Figure 7.2 that the transformation has indeed made the relation between the (transformed) response and the (transformed) covariate linear. It has also, however, produced a situation in which the variance of an additive error term could not be reasonably assumed constant, and has also produced a situation in which observations at the highest (transformed) covariate value would have exceedingly great leverage on a fitted equation. Bates and Watts demonstrate that fitting a linear, constant variance model and back-transforming parameter estimates to reflect the values of the Michaelis-Menten equation results in a poor fit to the data in the region of the asymptote (which is of primary scientific interest in this problem).

This example has involved a specific model of scientific interest (the Michaelis-Menten equation). Many other situations can be initially approached with a constant variance model, the estimation and assessment of which provide information for model modifications.

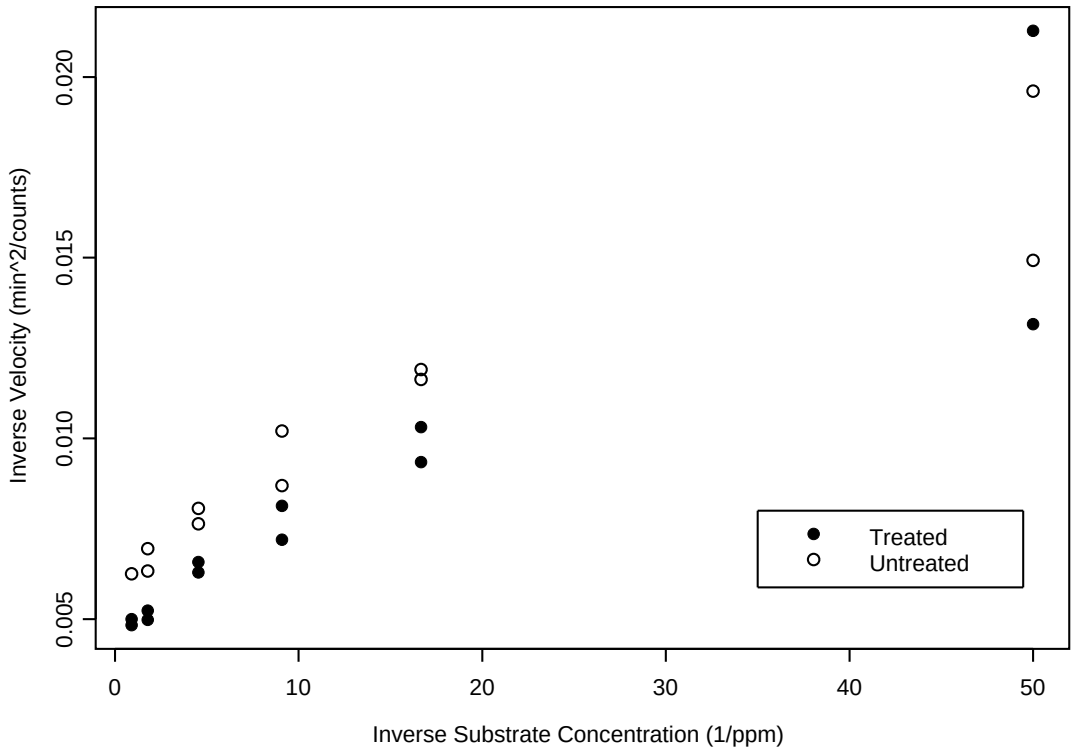

Figure 7.2: Scatterplot of transformed data for Puromycin example using reciprocal expression of both covariate and response variables.

7.2.2 Linear and Nonlinear Models

Before moving on to additive error models that have nonconstant variance, we pause to briefly indicate the meaning of *linear* and *nonlinear* models. Consider an additive error model of the form (7.1) for which

$$E(Y_i) = g(x_i,\, \beta); \qquad i = 1, \ldots, n, \tag{7.2}$$

under only the assumption that the error terms have zero expectation.

We define a model of this type to be *nonlinear* if at least one of the derivatives of $g(\cdot)$ with respect to elements of β depends on one or more elements of that parameter; note that this is obviously not the case for a linear expectation function. One point of clarification is in order. Some authors of applied linear regression texts use the phrase "intrinsically linear" to refer to models that we will consider intrinsically nonlinear, but "transformably linear". For example, Draper and Smith (1981) consider the following model to be intrinsically linear.

$$Y_i = g(x_i, \beta) = \exp(\beta_0)\, \exp(-\beta_1 x_i),$$

because it may be transformed to

$$\log(Y_i) = \beta_0 - \beta_1\, x_i.$$

Since the derivatives of $g(x_i, \beta)$ with respect to either β_0 or β_1 depend on β, we will consider this an intrinsically nonlinear model.

The topic of nonlinearity results in two notions of the way in which an additive error model can be nonlinear, and these are called *intrinsic* curvature and *parameter effects* curvature. While there are techniques for quantifying the relative contributions of these types of nonlinearity for specific models, for now we confine our efforts to gaining a more intuitive understanding of just what these types of nonlinearity are.

To have a basic understanding of intrinsic and parameter effects curvatures we must first introduce the concept of an *expectation surface*, which is also frequently called a *solution locus* (some authors use both terms interchangeably, e.g., Seber and Wild, 1989). Consider a model of the form (7.1) in which the covariates x_i consist of a single numerical value. The quantities involved in this model, other than the parameters β and σ, may be viewed as the vectors $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{x} \equiv (x_1, \ldots, x_n)^T$. Think of $\boldsymbol{Y}$ and $\boldsymbol{x}$ not as vectors of length n, but rather as individual points in $n-$dimensional real space. Similarly, think of $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)^T$ as a point in $p-$dimensional real space, with $p < n$. The expectation function, which we will momentarily write as $\boldsymbol{g} \equiv (g(x_1, \boldsymbol{\beta}), \ldots, g(x_n, \boldsymbol{\beta}))^T$, defines a relation between the $p-$dimensional space of $\boldsymbol{\beta}$ and the $n-$dimensional space of $\boldsymbol{x}$ and $\boldsymbol{Y}$. Now, for a fixed $\boldsymbol{x}$, $\boldsymbol{g}$ is a $p-$dimensional surface in $n-$space, that is, a $p-$dimensional *manifold* (recall $p < n$). This manifold is what is called the solution locus (or expectation surface).

To avoid confusion here, note that we are *not* describing the straight line in the $2-$dimensional space of a scatterplot that is formed by $\beta_0 + \beta_1 x_i$ as the x_i vary. Rather, for fixed $\boldsymbol{x}$ of any dimension (> 2) the solution locus of a simple linear regression model is a $2-$dimensional plane formed as β_0 and β_1 vary. For a multiple regression model the solution locus is a $p-$dimensional plane, assuming $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.

All we will say about the quantification of intrinsic and parameter effects curvatures are that such quantification depends on arrays of first and second derivatives of $g(x_i, \beta)$ with respect to the elements of β. Note that, for any function linear in the elements of β, the first derivatives are constants and the second derivatives are all 0. Curvature is thus exhibited by any surface that has non-zero second derivatives. This is where the geometry and algebra of

vector spaces becomes more complex than what we desire to get into at this point, but an intuitive understanding can be gained by considering two aspects of a solution locus $\boldsymbol{g}$ (both of which have already been mentioned above):

1. First, $\boldsymbol{g}$ forms a $p-$dimensional manifold in $n-$dimensional space. The degree to which this manifold differs from a $p-$dimensional plane is reflected in intrinsic curvature.

2. Secondly, $\boldsymbol{g}$ maps points from $p-$dimensional space (i.e., $\boldsymbol{\beta}$s) to $n-$dimensional space. If equally spaced points in $p-$space are mapped into unequally spaced points in $n-$space, then the model exhibits parameter effects curvature; note that for a linear manifold equally spaced points in the parameter space are mapped into equally spaced points in the data space.

We mention these two types of curvature because one of them, intrinsic curvature, cannot be changed by re-expression of the model through parameter transformations while the other, parameter effects curvature can be changed by this means. This can sometimes be desirable for purposes of estimation, inference, and interpretation (see Ross, 1990, for an extensive discussion). Note that transformation of parameters is an entirely different matter than transformation of random variables. A distribution is invariant to the former but obviously not the latter.

7.2.3 Models with Known Variance Parameters

In both this subsection and the next we will consider additive error models for which the assumption of constant error variance is relaxed. The result is, essentially, that we need to form a model for the variance structure, similar to forming a model for the mean structure. At present, we will consider

models for the variance that contain no unknown parameters other than those also involved in the model for mean structure (the β of expression (7.1)). At first, this may seem an artificial device, similar to specifying a normal model with known variance, but that is not really the case. There are two realistic situations in which this approach to model formulation is quite viable. At the same time, the reason for separating the models of this subsection from those of the next does depend on methods of estimation that may be applied, thus fore-shadowing topics to come. What ties the two situations discussed in this subsection together is that they may both be considered as producing "regression weights" for individual random variables (and the associated observations). In the first case the resultant weights are fixed and known, while in the second they must be estimated, but only as functions of the parameters β that also appear in the model for expectations.

Known Weights

The simplest extension of model (7.1) occurs in situations for which the variances of the response variables $\{Y_i : i = 1, \ldots, n\}$ are not equal, but differ by only known constants of proportionality. A model appropriate for this situation is,

$$Y_i = g(x_i, \beta) + (\sigma/\sqrt{w_i})\,\epsilon_i, \tag{7.3}$$

where, as in model (7.1), the ϵ_i are assumed to be *iid* random variables following a location-scale family F such that $E(\epsilon_i) = 0$ and (usually) $var(\epsilon_i) = 1$. As for constant variance models, the nearly ubiquitous choice for F is the normal distribution.

The most obvious situations in which we might want to consider model (7.3) with known weights $\{w_i : i = 1, \ldots, n\}$ are those for which the data used

as a realization of the model are composed of sample means.

Example 7.2

Consider a situation in which the phenomenon of interest is the infestation of a commercial crop by an insect pest, to be compared among groups of insecticide treatments (e.g., passive control, standard chemical insecticide, natural biological insecticide). Since our primary concern in this example is a large-scale phenomenon (we don't really care about individual plants, only the average effects when the treatments are applied to fields), the observations may be in the form of the average number of pests found on plants in experimental plots given the various treatments (note that this also corresponds to the notion from the experimental approach that observations should be made on "experimental units" to which treatments are independently applied, not necessarily "sampling units" on which individual measurements are made). In this hypothetical example, suppose that recorded observations are the average number of insects of concern on plants in a number of experimental plots under each treatment. Say there are 5 plots per treatment, but that the number of plants actually sampled per plot varies from 12 to 30, depending on the number of field assistants available to visit the various plots on the day of observation. We could also imagine an experimental design in which averages are taken over a number of days. Regardless, if we would believe that a constant variance model is appropriate for random variables associated with the sampling units (plants), then this would not be true for plot (or plot/time) averages. Model (7.3) would likely be more reasonable, with the w_i given as n_i, the number of observed plants in plot (or plot by time unit) i.

The situation of Example 7.2, in which we have known weights, is a quite

simple one. It is clear that model (7.3) could be easily re-expressed as

$$Y_i^* = g^*(x_i,\, \beta) + \sigma\, \epsilon_i,$$

where $Y_i^* \equiv (w^{1/2})Y_i$, $g^*(\cdot) \equiv (w^{1/2})g(\cdot)$ and, for example, $\epsilon_i \sim iid\, N(0,1)$; $\quad i = 1, \ldots, n$. In this case, we have done nothing untoward to the model by transformation of the responses (this would, in fact, be true for any linear transformation applied to the random variables $\{Y_i : i = 1, \ldots, n\}$ – why?)

Finally, it is also true in this example that we would need to be mindful of the potentially deleterious effect of assuming normal distributions that was seen in Example 6.7. If plot averages for one or more of the treatment groups constituted sets of small numbers, something other than a model with additive normal errors might be suggested, regardless of how variances were dealt with.

Weights as Specified Functions of Means With Unknown Parameters

Consider applying the basic concept of weights based on variances in a simple linear regression model for which we have available replicate values of the response for a given level of covariate. In this situation it may be tempting to apply something similar to model (7.3), except in which we replace $(\sigma/w_i^{1/2})$ by σ_j where j indexes distinct levels of the covariate. Carroll and Rupert (1988, section 3.3.6, page 86) caution against this type of model in situations for which the number of replicates is small; apparently even from 6 to 10 replicates per value of the covariate can lead to poor estimates of the weights and subsequent overestimation of the variance of the estimated regression parameters (see references given in Carroll and Rupert, page 87). Why would we suggest something like model (7.3) and then turn around and caution against what appears to be a straightforward extension of the same idea? What's the difference?

The difference between what we have considered in model (7.3) with the hypothetical situation of Example 7.2, and the notion of the preceding paragraph is that, in the former case but not the latter, we have assigned a "reduced structure" to the variances in a manner similar to that used for the mean structure. That is, in the hypothetical Example 7.2, we used a model that specified variances differing only through the "mechanism" of sample sizes used to calculate averages (ok, ok, I won't say go read Chapter 5.3 on statistical abstraction again, but you should). This amounts to modeling the variances in a manner analogous to modeling expected values. A difference is that, while we sometimes have scientific knowledge available to help with a model for means, this is rarely true for modeling variances. We must, for the most part, rely on what we know about the behavior of statistical models, and the experiences of previous analyses.

One basic idea that has emerged from these sources is that modeling variances as functions of the mean is often a useful technique. The type of model that results may be written as, for $i = 1, \ldots, n$,

$$Y_i = g_1(x_i, \beta) + \sigma g_2(x_i, \beta, \theta)\, \epsilon_i, \tag{7.4}$$

where, as before, $\epsilon_i \sim iid\, F$ with $E(\epsilon_i) = 0$ and, almost always, F is the standard normal distribution. If the x_i; $i = 1, \ldots, n$ are considered known constants and we assume that the dependence of $g_2(\cdot)$ on the combination of x_i and β is only through the way these quantities are combined in the function $g_1(\cdot)$, then we can also write model (7.4) as,

$$Y_i = \mu_i(\beta) + \sigma\, g(\mu_i(\beta), \theta)\, \epsilon_i, \tag{7.5}$$

with assumptions on $\{\epsilon_i : i = 1, \ldots, n\}$ as before. Here, $g_1(x_i, \beta)$ in (7.4) has been replaced with $\mu_i(\beta)$ and $g_2(x_i, \beta, \theta)$ has been replaced with $g(\mu_i(\beta), \theta)$.

What renders model (7.5) appropriate under the topic of this subsection (known variance model parameters) is that we assume the value of the parameter θ is known. Specification of this value is generally considered a part of model *selection* rather than an issue of estimation, in the same manner that selection of an appropriate power for a Box-Cox transformation is considered a part of model selection in linear regression analyses. If we take the $\sqrt{w_i}$ from model (7.3) to be given by $1/g(\mu_i(\beta), \theta)$ in (7.5), then this model can be considered to be in the form of a "weighted" regression. The result, however, is that the weights for model (7.5) must be estimated, since β is unknown, rather than specified as known constants. On the other hand, the situation is simplified by taking the additional parameter θ as a known value in the model.

By far, the most common model formulation of the type (7.5) is the "power of the mean" model, in which,

$$g(\mu_i(\beta), \theta) = \{\mu_i(\beta)\}^{\theta}.$$

In this case, we have, from model (7.5), that

$$var(Y_i) = \sigma^2 \{\mu_i(\beta)\}^{2\theta},$$

or,

$$2 \log \left[\{var(Y_i)\}^{1/2}\right] = 2 \left[\log(\sigma) + \theta \log\{\mu_i(\beta)\}\right],$$

or,

$$\log \left[\{var(Y_i)\}^{1/2}\right] = \log(\sigma) + \theta \log\{\mu_i(\beta)\}, \tag{7.6}$$

that is, the logarithm of the standard deviation of Y_i should be linearly related to the logarithm of its expectation.

Now, a result due to Bartlett (1947) is that, if Y_i, having mean μ_i and variance $\sigma^2 g^2\{\mu_i\}$, is transformed to $h(Y_i)$, then a Taylor series expansion

results in,

$$var\{h(Y_i)\} \approx \left(\frac{d}{d\mu_i}h(\mu_i)\right)^2 \{\sigma g(\mu_i)\}^2.$$

Thus, if $g(\mu_i, \theta) = \mu_i^\theta$, the transformed variable $h(Y_i)$ has approximately constant variance if,

$$\left(\frac{d}{d\mu_i}h(\mu_i)\right) \propto \mu_i^{-\theta}, \tag{7.7}$$

any constant of proportionality being absorbed into σ^2. The relation of expression (7.7) will hold if

$$h(\mu_i) \propto \mu_i^{1-\theta}. \tag{7.8}$$

Now, when $h(\cdot)$ of (7.8) is applied to response random variables Y_i, we have obtained a power (Box-Cox) transformation of the Y_i that will stabilize variance. Also, (7.6) indicates a practical manner by which the power parameter θ may be easily estimated (plotting the logarithm of standard deviations against the logarithms of means for groups of data), and looking at the slope to estimate θ.

We are not advocating here the indiscriminant use of power transformations to produce constant variance but, rather, the use of model (7.5) to reflect the phenomenon of interest. The point is, simply, that this is the exact same theory that leads to power transformations. In effect, if you are willing to accept the latter as potentially useful, you should be equally willing to accept model (7.5) since this is where you actually started (whether that is made clear in courses on applied regression methods or not).

Example 7.3

This example is taken from Trumbo (2002). Major airlines must schedule flights based on any number of factors, one of which is the necessary "flight

time" (time in the air) to complete a given trip. A data set presented in Trumbo (2002) contains data from 100 non-randomly chosen flights made by Delta airlines in 1994. The data set contains several variables, of which we will use the distance of the flights (recorded in miles) and the flight time (recorded in hours). For our purposes we also ignore two flights of much greater distance than the others; inclusion of these two flights would change nothing in this example, but excluding them makes it easier to look at plots. A scatter plot of the 98 observations used here is presented in Figure 7.3. It is clear from this display that time appears linearly related to distance, and it also seems that the variability among times increases as distance increases. An ordinary least squares fit to these data results in $\hat{\beta}_0 = 0.63064$, $\hat{\beta}_1 = 0.00192$ and $\hat{\sigma}^2 = 0.06339$. A plot of the studentized residuals from this regression are presented in Figure 7.4, which exhibits the unequal variances noticed in the scatterplot. We might consider, for this situation, model (7.5) with $\mu_i(\beta) \equiv \beta_0 + \beta_1 x_i$, where for flight i, $i = 1, \ldots, n$, Y_i corresponds to time, x_i is distance, $\epsilon_i \sim iid\, N(0, 1)$, and θ is to be determined prior to estimation of β and σ^2.

Alternatively, we might consider using a power transformation to try and stabilize the variance. What occurs in this example if we take this approach and proceed to employ data transformation in a typical manner (i.e., power or Box-Cox transformations)? While there are a number of observations for the same distances in the data, this is not true for many other distances. But we might bin or group the data by values of distance, compute sample means and variances within each group, and examine a plot of log standard deviations against log means. Travel distances range from 134 miles to 2588 miles. If our 98 observations were evenly spaced in this range there would be one observation for each 25 mile increase in distance. If we want about 4 observations per group we might choose bin sizes of 100 miles, $\{100-200,\ 200-300,\ \ldots, 2500-2600\}$.

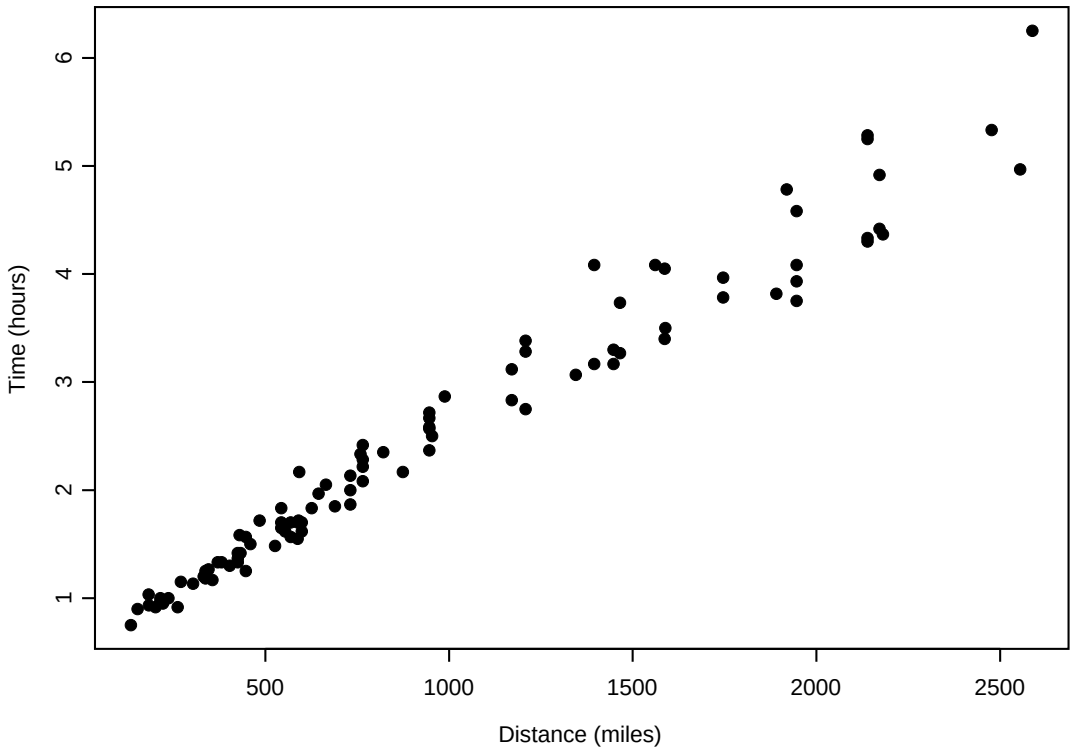

Figure 7.3: Scatterplot of travel time versus distance for a sample of flights conducted by Delta airlines in 1994.

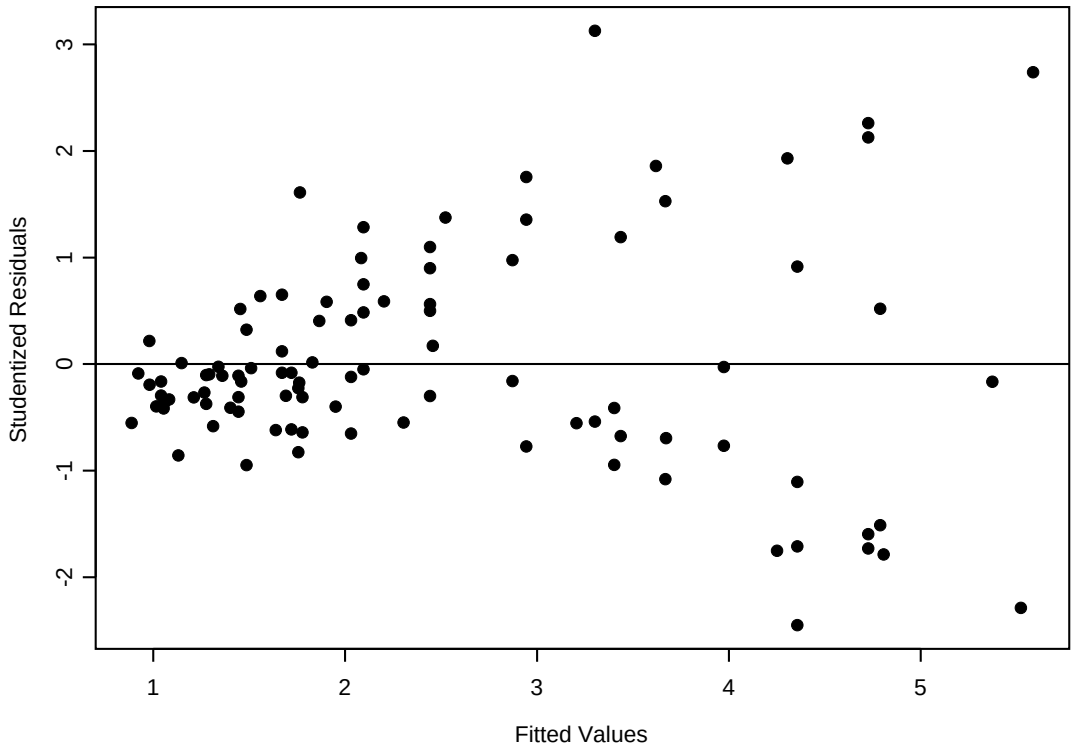

Figure 7.4: Plot of studentized residuals for an ordinary least squares fit to the data of Figure 7.3.

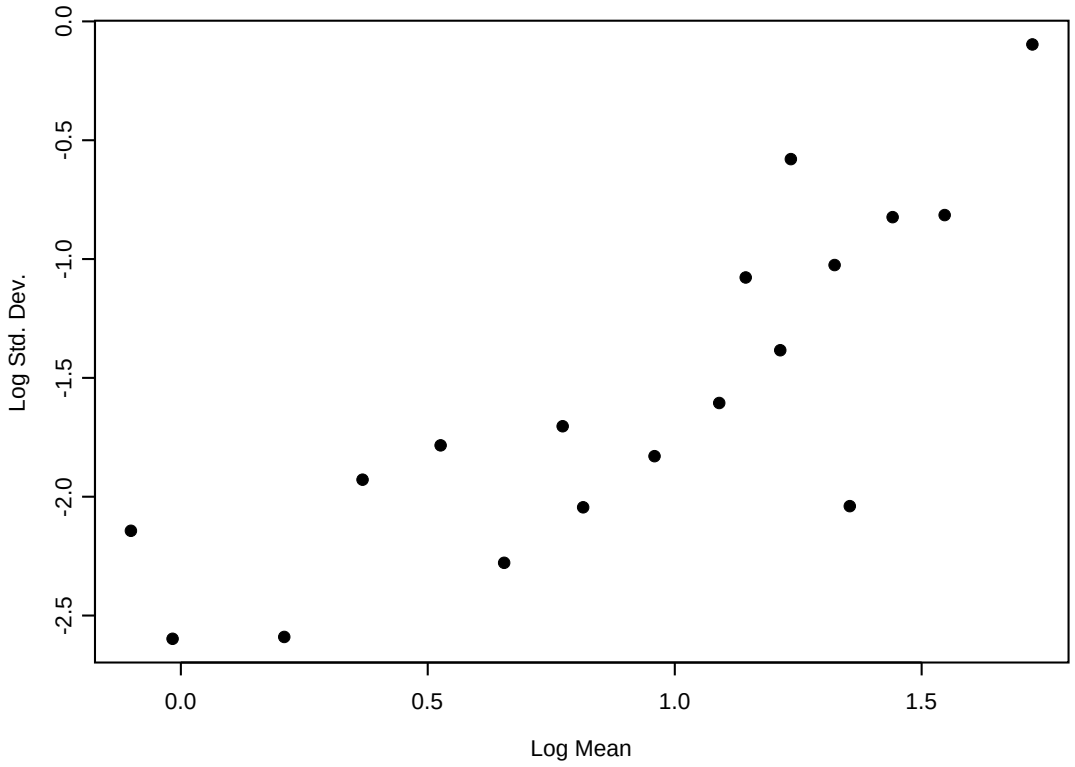

Figure 7.5: Box-Cox transformation plot from using binned data from Figure 7.3.

This is a crude and *ad hoc* manner for forming bins, but it is effective here since the observations are fairly well spread out over the range of distances. The resulting plot is given in Figure 7.5. This plot is playing the role of a diagnostic or exploratory tool, and we want to avoid getting too "fine-grained" or picky in its assessment. The slope of an ordinary least squares fit to the values of Figure 7.5 is 1.09. If one uses an eyeball method, not paying too much attention to the first two values on the left side of the figure, a value of around 1.5 for the slope also may be reasonable. The first of these values

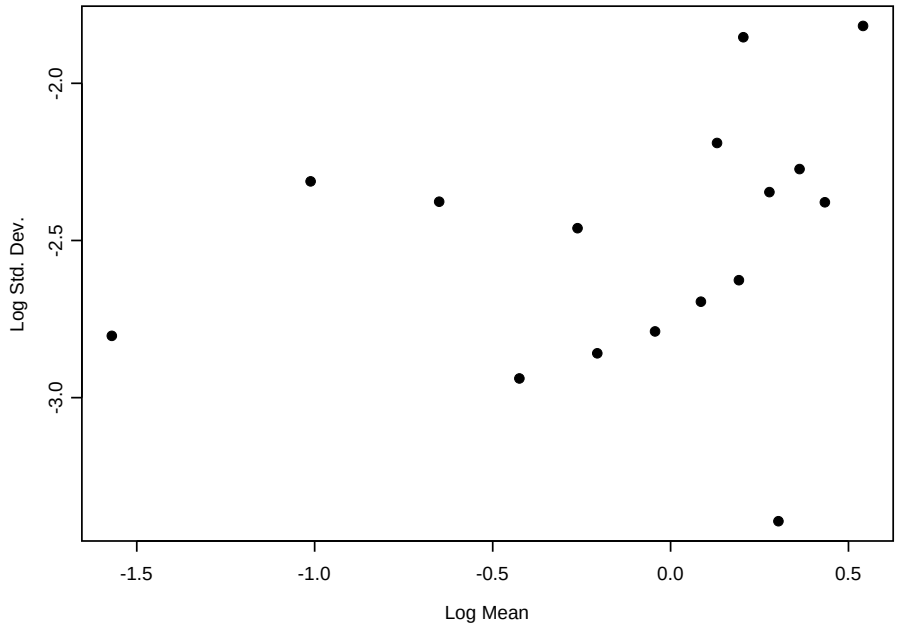

Figure 7.6: Box-Cox diagnostic plot for log transformed data.

would suggest a logarithmic transformation $Y_i^* = \log(Y_i)$ and the second a reciprocal square root transformation $Y_i^* = 1/\sqrt{(Y_i)}$. If we conduct these transformations and then reconstruct Box-Cox diagnostic plots beginning with the transformed values, we obtain what is presented in Figure 7.6 (for the log transform) and Figure 7.7 (for the reciprocal root transform). Neither of these plots are "textbook" nice, but they both look as if there has been some improvement over the pattern of Figure 7.5. We can now try to fit regressions using one or the other of these data transformations. The first indications of unequal variances came from the scatterplot of Figure 7.3 and the residual plot of Figure 7.4 (the Box-Cox plots were constructed primarily to help choose a power for the transformation, not to indicate whether one was

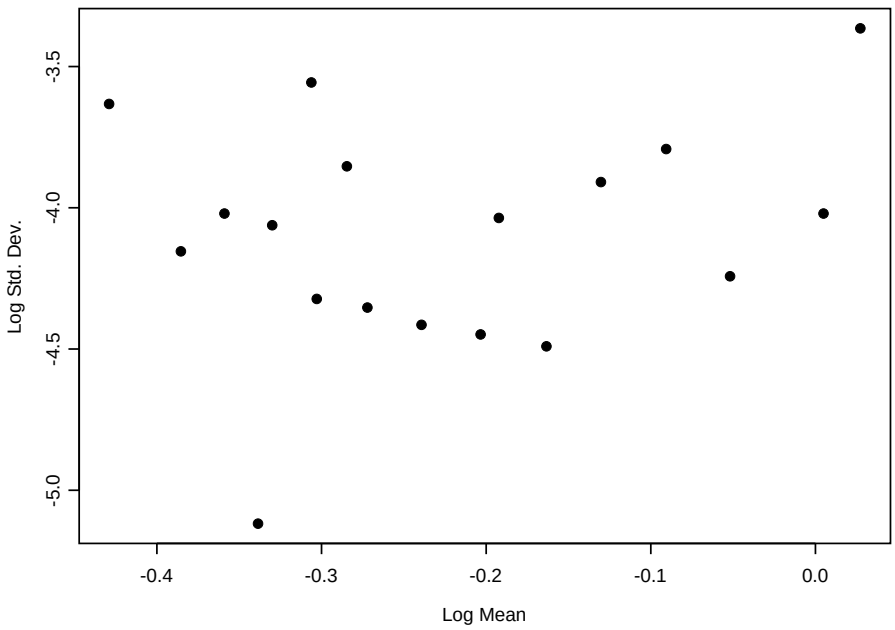

Figure 7.7: Box-Cox diagnostic plot for reciprocal root transformed data.

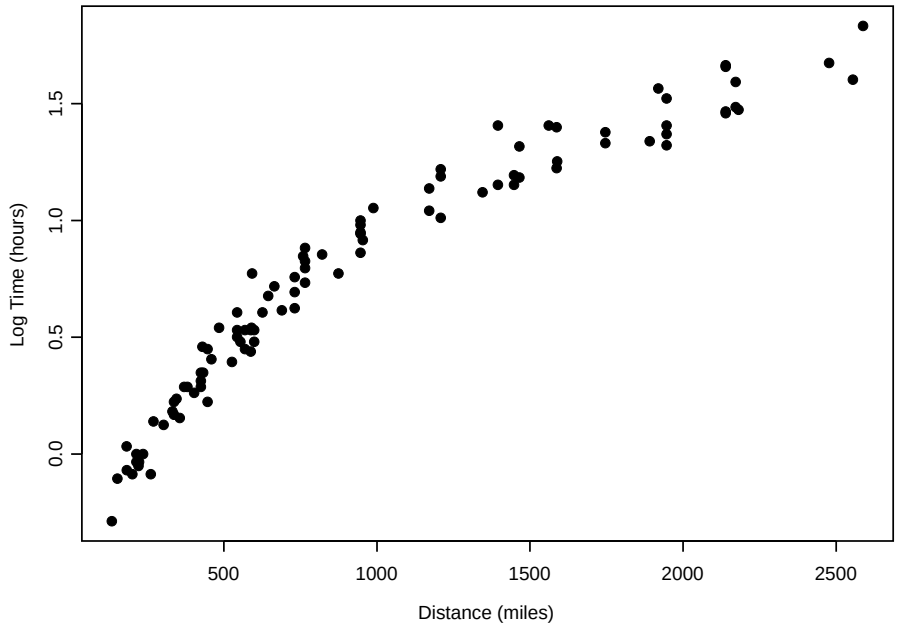

Figure 7.8: Scatterplot for log transformed time.

needed). Scatterplots for $\log(Y_i)$ and $1/\sqrt{(Y_i)}$, both against distance (x_i) are presented in Figure 7.8 and Figure 7.9. What has happened? First, the power transformations conducted did seem to help with the problem of nonconstant variance (which is what they were intended to do). But, they also changed what was originally a fairly linear relation (in Figure 7.3) into nonlinear relations. Mathematically, the reason for this is obvious. If Y is linear in x, then Y^z will not be linear in x. Less obvious, but equally true, is that if additive error terms ϵ_i in a model for response variables Y_i are normally distributed, then additive error terms in a model for transformed Y_i cannot be normally distributed (and may not even have identical location-scale distributions). Interpretation of a model relative to a set of response variables $\{Y_i : i = 1, \ldots, n\}$ is not

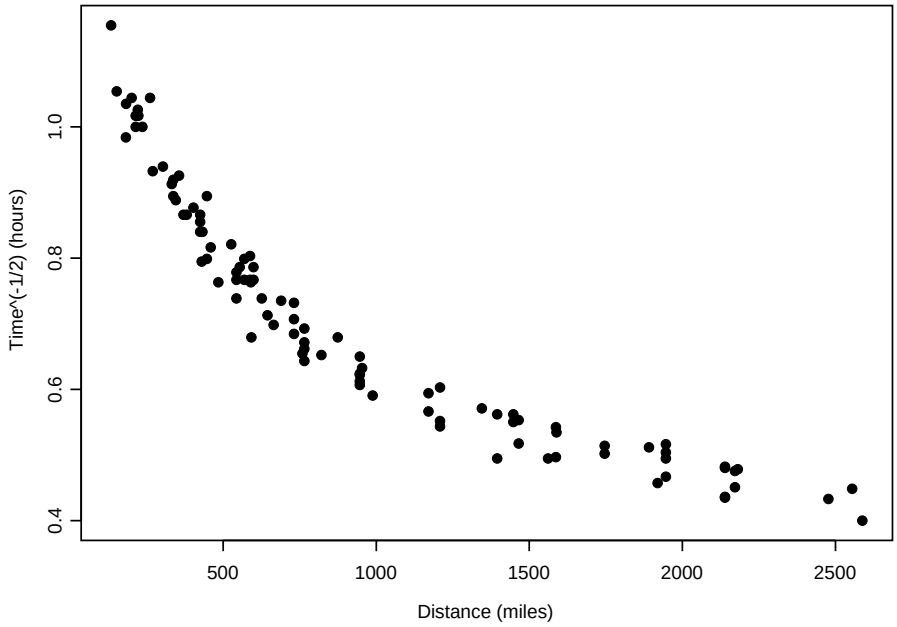

Figure 7.9: Scatterplot for reciprocal root transformed time.

necessarily straightforward if an additive error model was fit to transformed response variables $\{Y_i^* : i = 1, \ldots, n\}$. One of the simplest examples, of which you may already be aware, is that if

$$\log(Y_i) = \beta_0 + \beta_1 + \sigma\,\epsilon_i; \qquad i = 1, \ldots, n,$$

where $\epsilon_i \sim iid\, N(0, 1)$, then,

$$E(Y_i) = \exp\left\{\beta_0 + \beta_1\, x_i + \sigma^2/2\right\},$$

not,

$$E(Y_i) = \exp\left\{\beta_0 + \beta_1\, x_i\right\}.$$

The advantage of model (7.5) over a constant variance model fitted to the transformed responses $\{Y_i^{1-\theta} : i = 1, \ldots, n\}$ is that all inferences or predictions concerning the responses are maintained in the original scale of measurement or observation. For the problem of this example, a reasonable model would be of the form (7.5) with $\mu_i(\beta) = \beta_0 + \beta_1\, x_i$ and $\theta = -0.5$. A fit of this (or any other model) should certainly be subject to model assessment procedures which will come later in the course.

Comments

1. If transformations (of response variables) are so fraught with dangers, why have they been so popular, even making their way into courses for applied scientists (e.g., Statistics 401 at ISU)?

   The answer to this question probably depends at least in part on computational history and institutional inertia. That is, estimation with model (7.5), although not difficult today, does require iterative numerical techniques. Transformations have always been one way to replace a model

for which constant variance is clearly not appropriate with a model for which constant variance is more appropriate, and for which estimation by ordinary least squares is possible. Once such methods made it into the "standard" set of material taught to scientists and applied statisticians there has always been resistance to removing them.

2. But why have statisticians hung onto the basic ideas of additive error constant variance models with such fervor? Computation of estimates for models such as (7.5) is not something that has only recently become available in, say, the past 15 years. Surely there must be more to it than just computation and history.

   Yes, there is. Recall the opening paragraph to Section 7.2.1. It is typically the case that exact theory is only available for constant variance models (and, in fact, linear constant variance models). This is certainly a mature and beautiful set of theory, and one that has proven to be of great applicability and value in practice. But there has been a tendency for statisticians to hang on to parts of this body of methodology even when it is clear that not all of it is appropriate for a given problem. This "statistical denial" leads to such things as, for example, computing interval estimates with quantiles from $t-$distributions even in cases for which the only distributional result available is asymptotic normality (see, e.g., Section 3.7 of Part 1 of this course).

   A counter-point to the above assertion is that it is not really exact theory that is the goal, but having estimation methods that are robust, and these are the most easily developed for constant variance models (linear models are also helpful, but nonlinearity is not the same roadblock to achieving robustness that it is for exact theory). Note that the term *robust* is used

here in a distributional context, not relative to extreme observations. Methods that are not greatly affected by extreme observations are called *resistant*; ordinary least squares, for example, is robust but not resistant.

7.2.4 Models with Unknown Variance Parameters

We turn now to models very similar to that of expression (7.4) but for which we generalize the variance model. Specifically, in this subsection we consider models of the form,

$$Y_i = g_1(x_i\,\beta) + \sigma g_2(x_i, \beta, z_i, \theta)\,\epsilon_i, \tag{7.9}$$

with, for $i = 1, \ldots, n$, $\epsilon_i \sim iid\, F$ such that $E(\epsilon_i) = 0$ and (usually) $var(\epsilon_i) = 1$. As for all additive error models, F is taken to be in a location-scale family and is usually specified to be $N(0, 1)$. Model (7.9) extends model (7.4) in that the function g_2 includes z_i, which may be a part of x_i or may be other covariates that are believed to affect the variance but not the mean, and the parameter θ is no longer assumed known. Sometimes, we we can impose a restriction similar to that used in moving from expression (7.4) to expression (7.5) by taking $\mu_i(\beta) \equiv g(x_i,\, \beta)$ and writing,

$$Y_i = \mu_i(\beta) + \sigma\, g(\mu_i(\beta),\, z_i,\, \theta)\,\epsilon_i, \tag{7.10}$$

with the same assumptions on the ϵ_i as in model (7.9). Models (7.9) and (7.10) are the basic structures we will consider in this subsection. They allow the variance to depend on the covariates (possibly only through the mean), but no longer assume that θ is a part of model selection. Rather, θ is to be estimated along with the other parameters β and σ^2. The inclusion of additional covariates z_i in (7.9) and (7.10) could also have been made to (7.4) and (7.5) but, as noted previously, the power of the mean model is dominant

among those for which θ becomes a part of model selection; thus, there seemed little motivation to include z_i in the formulations of (7.4) or (7.5).

A number of possible forms (not meant to be exhaustive, by any means) for g are given in Carroll and Rupert (1988), and I have extended the suggestions below with a few of my own possibilities. These include:

$$\begin{aligned}
\sigma\, g(\mu_i(\beta),\, z_i,\, \theta) &= \sigma\, \{\mu_i(\beta)\}^{\theta} \\
\sigma\, g(\mu_i(\beta),\, z_i,\, \theta) &= \sigma \exp\{\theta \mu_i(\beta)\} \\
\sigma\, g(\mu_i(\beta),\, z_i,\, \theta) &= \sigma \exp\{\theta_1 x_i + \theta_2 x_i^{-1}\} \\
\sigma\, g(x_i,\, \beta,\, z_i,\, \theta) &= \sigma(1 + \theta_1 x_i + \theta_2 x_i^2) \\
\sigma\, g(x_i,\, \beta,\, z_i,\, \theta) &= \theta_0 + \theta_1 x_i + \theta_2 x_i^2 \\
\sigma\, g(x_i,\, \beta,\, z_i,\, \theta) &= \theta_0 + \theta_1 z_i + \theta_2 z_i^2
\end{aligned}$$

Notice that the first of these is the power of the mean model discussed in the previous subsection. We may certainly specify this model without setting θ to a known (or selected) value. Note also that the first three models in this list take the logarithm of the standard deviations of the response variables Y_i as linear in either the mean or covariates, while the last three take the standard deviations of the responses as linear in covariate values. By no means should you consider the above list either all of the possibilities or even to constitute "tried and true" suggestions. The fact is that we are much less advanced in our modeling of response variances than in modeling response means.

Example 7.4

Foresters and environmental scientists are interested in estimating the volume of trees (obvious from a commercial standpoint, but also an indicator of

biomass production). Measuring the volume of a tree is a difficult and destructive process. On the other hand, field workers can easily measure the height of trees and what is known as "diameter at breast height" (DBH) in an efficient and non-destructive manner. The question is how these variables are related to the characteristic of interest, which is volume. Data for this example come from a study conducted in the Allegheny National Forest in Pennsylvania in which height and DBH were recorded for a number (here 31) trees which were subsequently cut and the volume measured in a more elaborate process. Our goal is to develop a statistical model that relates DBH and height to volume in a manner that would allow prediction for trees left standing, and may be applicable (with different parameter values) to species other than Black Cherry. The data used here are given by Ryan, Joiner, and Ryan (1985), where they are used to illustrate multiple linear regression.

A scatterplot matrix of the three variables of concern is presented in Figure 7.10, from which we see that volume and DBH are strongly linearly related, volume and height are weakly linearly related, and height and DBH are also weakly linearly related. To develop an additive error model for these data we begin with definition of variables involved. Let $\{Y_i : i = 1, \ldots, n\}$ be random variables associated with the actual volume of trees. Let $\{x_{1,i} : i = 1, \ldots, n\}$ be fixed variables that represent the measured DBH of trees (at 4.5 ft above ground level), and let $\{x_{2,i} : i = 1, \ldots, n\}$ be fixed variables that represent the measured height of trees. As a first step in developing a model we might conduct simple linear regressions of the Y_i (volumes) on each of $x_{1,i}$ (DBHs) and $x_{2,i}$ (heights). The first of these regressions (on DBH) yields results depicted in Figure 7.11 with studentized residuals presented in Figure 7.12, while the second (on height) results in the analogous Figures 7.13 and 7.14. An examination of these plots reveals the following:

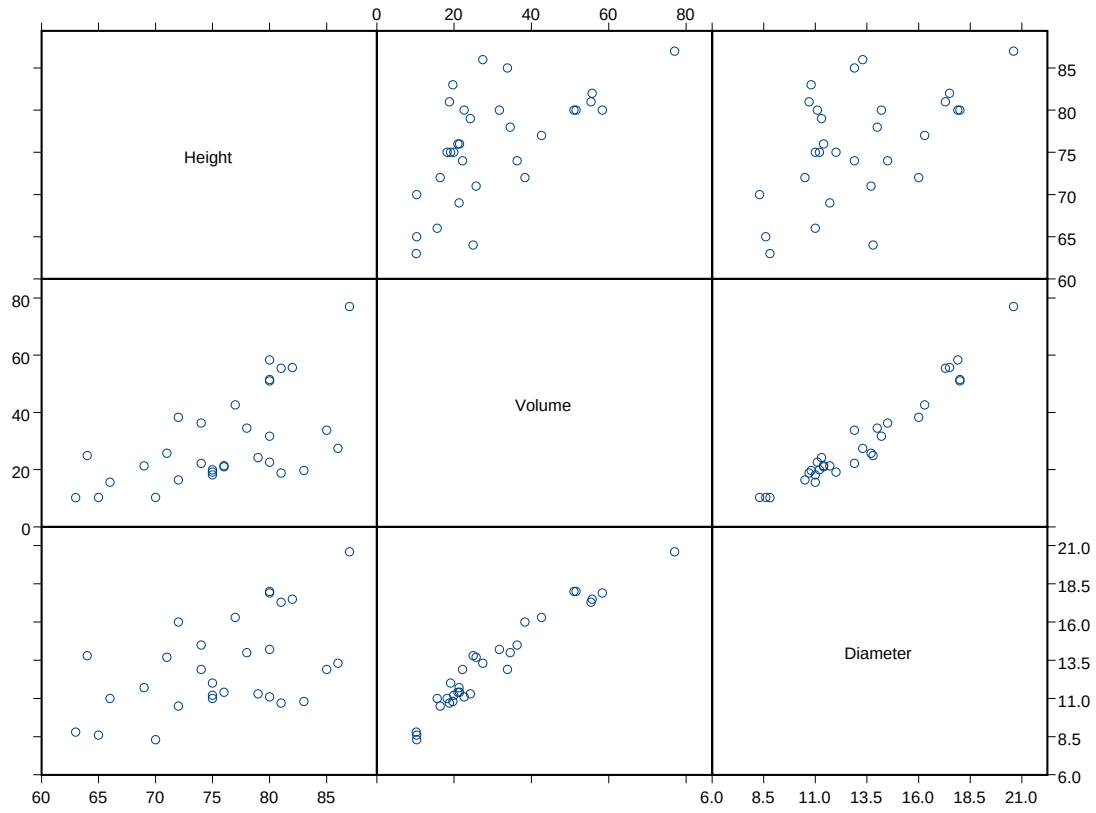

Figure 7.10: Scatterplot matrix of volume, height, and DBH for Black Cherry trees.

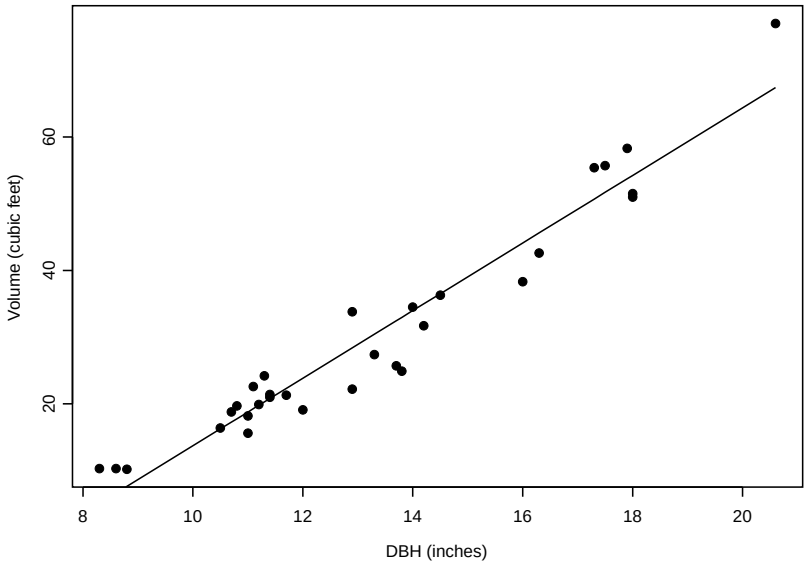

Figure 7.11: Regression of volume on DBH.

1. While the regression of volume on DBH is fairly nice, there are a few "small" trees that are not well described by the regression line.

2. More disturbing is the $U-$shaped pattern in residuals for this model, seen in Figure 7.12, and this appears to be due to more than the 3 small trees of Figure 7.11.

3. The relation between volume and height is weak, as we already knew (Figures 7.10 and 7.13), and the variances of volume clearly increase with (estimated) volume in this regression (Figure 7.14).

The natural next step is to fit a multiple linear regression model using both DBH and height as covariates. Estimated parameters for this multiple regression, as well as the two simple linear regressions using only one covariate are

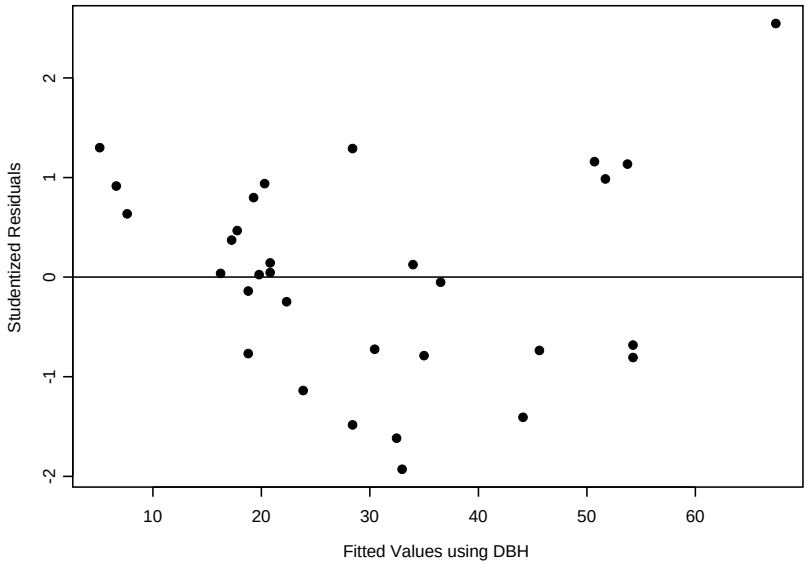

Figure 7.12: Studentized residuals for the regression of Figure 7.11.

presented in the following table (which has been arranged so that parameter estimates in the same column are comparable):

	Estimated Values				
Model	β_0	β_1	β_2	σ^2	R^2
DBH	-36.94	5.06		18.079	0.9353
Ht	-87.12		1.54	179.48	0.3579
DBH, Ht	-57.99	4.71	0.34	15.069	0.9479

This table largely reflects what has already been seen in the plots of Figures 7.10 through 7.14. It is perhaps surprising that what is certainly a weak linear relation between height and DBH (see Figure 7.10) has such a great impact on the estimated value of the regression coefficient associated with this covariate (β_2 in the table) and such a small impact on the coefficient of determination

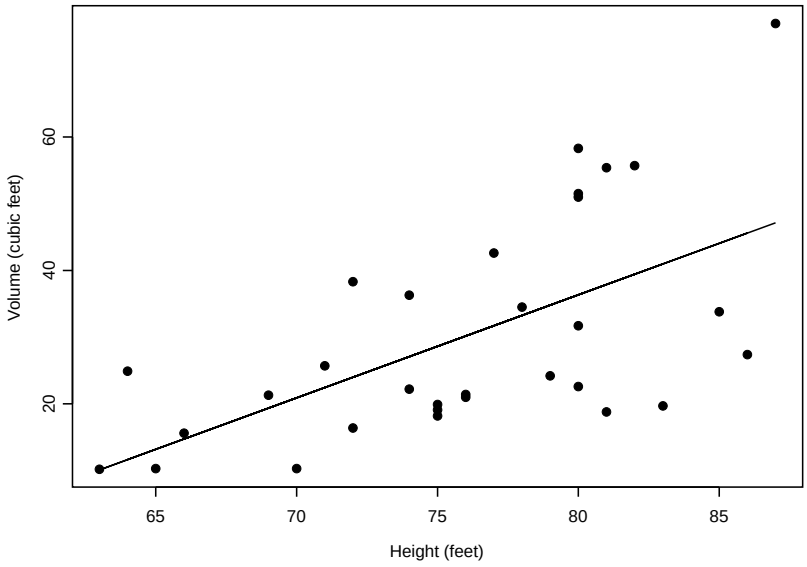

Figure 7.13: Regression of volume on height.

(R^2 in the table). Nonetheless, we might choose to retain both covariates in the model based on the reality that height must certainly be important in modeling the volume of trees.

As good statisticians we should certainly examine a residual plot for the multiple regression, which is presented in Figure 7.15. The curious $U-$shaped pattern of residuals seen in the regression of volume on DBH is repeated in this residual plot, even ignoring the three leftmost and one rightmost points of the plot (which may not be a good idea here as with 31 data values this represents about 15% of the total data).

In a multiple regression, a plot of residuals may not reveal everything shown in plots of residuals against the individual covariates. Plotting the studentized residuals against both DBH and height individually results in Figures 7.16 and 7.17. Figure 7.16 reinforces the suggestion that the mean function is not

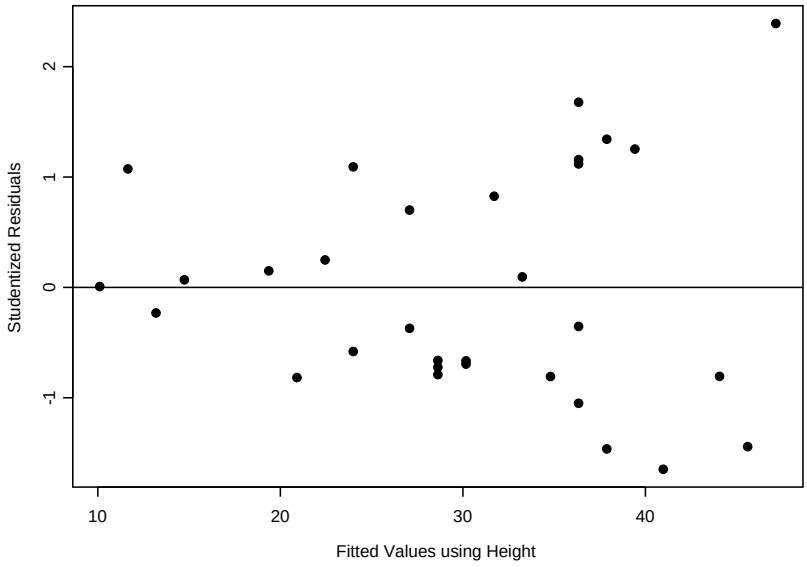

Figure 7.14: Studentized residuals for the regression of Figure 7.13.

correctly specified in terms of DBH. The same $U-$shaped residual pattern is hinted at for height in Figure 7.17, although in the absence of previous evidence one would be reluctant to see much in this plot.

Where does this leave us? We have a linear multiple regression model that appears quite good for describing the pattern of data (an R^2 value of nearly 0.95 is, in general, nothing to sneeze at). On the other hand, we certainly want to accomplish more than describing the data pattern. The finding that volume is greater for taller, fatter trees than it is for shorter, thinner trees is not likely to set the world of forest mensuration on fire. We would like to develop a model that can predict well, and the general form of which might be amenable to use for other tree species. This means that we would like to determine a pleasing statistical conceptualization for the problem that can hopefully take into account the anomalies seen in the residual plots of the

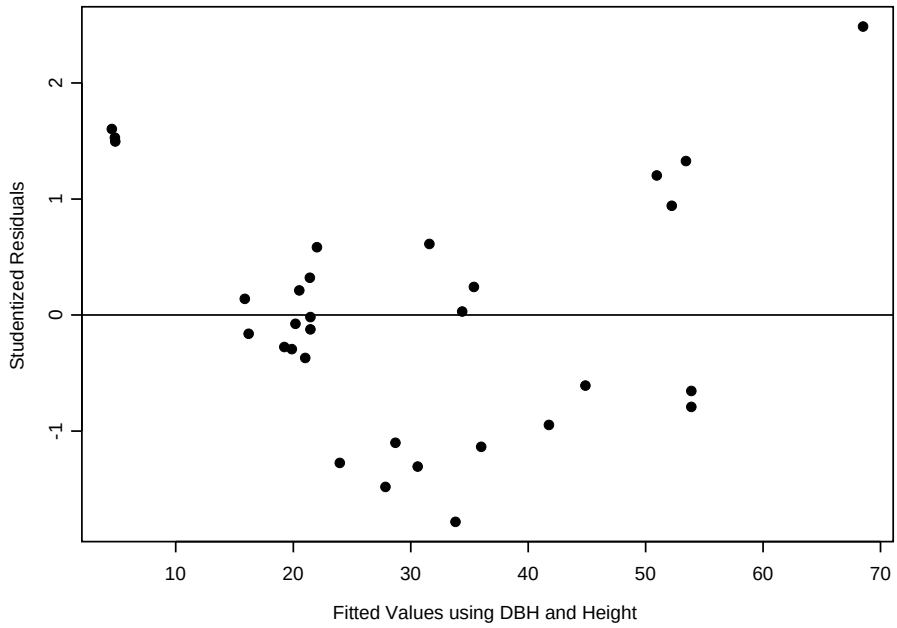

Figure 7.15: Studentized residuals for the regression of volume on DBH and height.

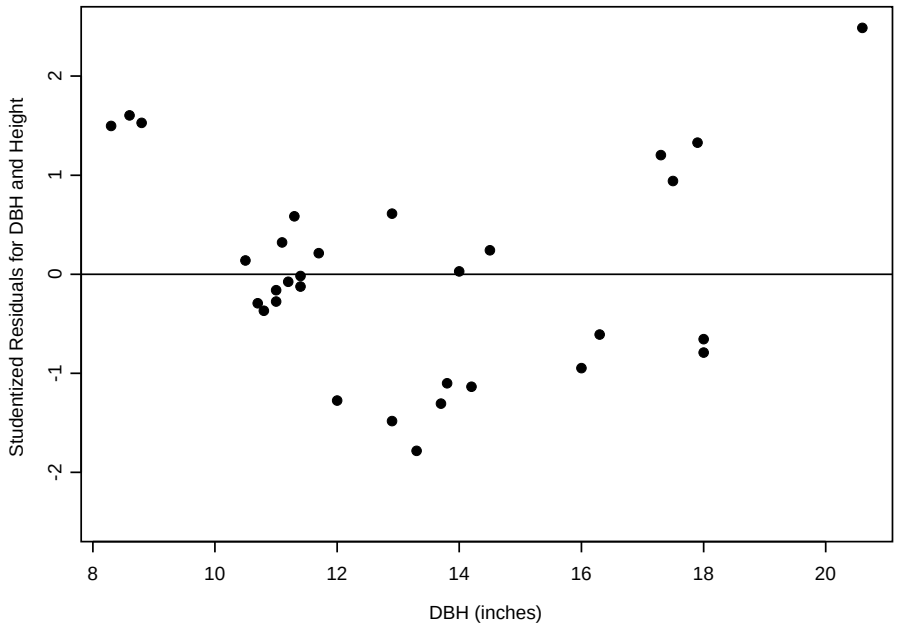

Figure 7.16: Studentized residuals from the regression of volume on DBH and height against DBH.

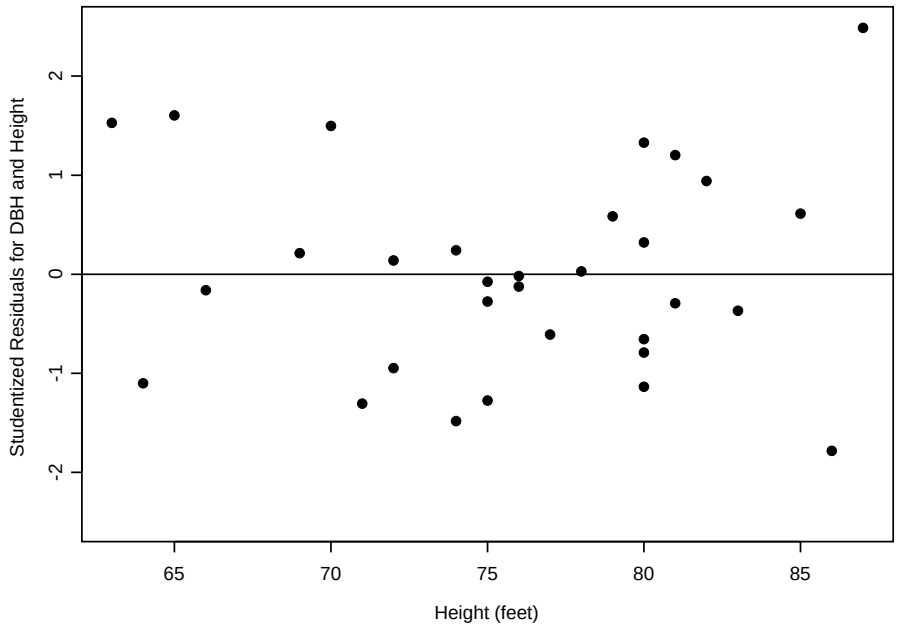

Figure 7.17: Studentized residuals from the regression of volume on DBH and height against height.

linear regression models. These plots have suggested that the relation between DBH and volume is not exactly a straight line, and that height may have some connection with variability in volumes. Residual plots from the multiple regression point toward a strategy of addressing the expectation function first (why?).

Is there a simple conceptualization of this problem other than "adding things together"? Well, the problem essentially deals with quantities that reflect basic geometry relative to trees. What is a simple geometric concept of a tree (no need to get fancy – close your eyes and think of a wooden telephone pole). It would seem that a very basic connection between volume and the two measurements of height and diameter (twice the radius) would be the volume of a cylinder, $V = \pi r^2 H$. To make use of this idea for an expectation function in this example we must bring the units of measurement into agreement. Volume (Y_i) is in cubic feet, height ($x_{2,i}$) is in feet, DBH ($x_{1,i}$) is in inches and is also 2 times the radius. A possible model for the expectation function is then,

$$E(Y_i) = \beta_0 + \beta_1 \left\{ 2\pi (x_{1,i}/24)^2 \, x_{2,i} \right\}, \tag{7.11}$$

which, if we define $\phi(\boldsymbol{x}_i) = \{2\pi(x_{1,i}/24)^2 \, x_{2,i}\}$ is just a simple linear regression of volume (Y_i) on $\phi(\boldsymbol{x}_i)$), which we might call "cylinder". To investigate the possibility of using (7.11) as a linear expectation function we might simply fit a constant variance regression using ordinary least squares,

$$Y_i = \beta_0 + \beta_1 \, \phi(\boldsymbol{x}_i) + \sigma \, \epsilon_i, \tag{7.12}$$

where, for $i = 1, \ldots, n$, $\epsilon_i \sim iid \, F$ with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$. The result is shown in Figure 7.18, with a studentized residual plot presented in Figure 7.19. Estimated values for the regression model (7.12) are $\hat{\beta}_0 = -0.298$, $\hat{\beta}_1 = 0.389$, $\hat{\sigma}^2 = 6.2150$, and $R^2 = 0.9778$. Relative to the regressions in

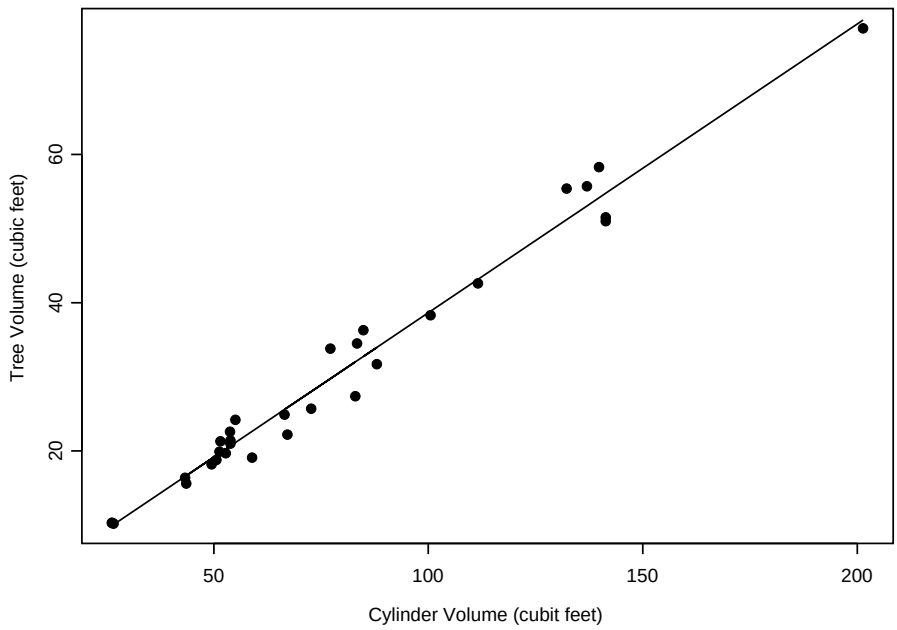

Figure 7.18: Scatterplot and least squares fit for volume against cylinder.

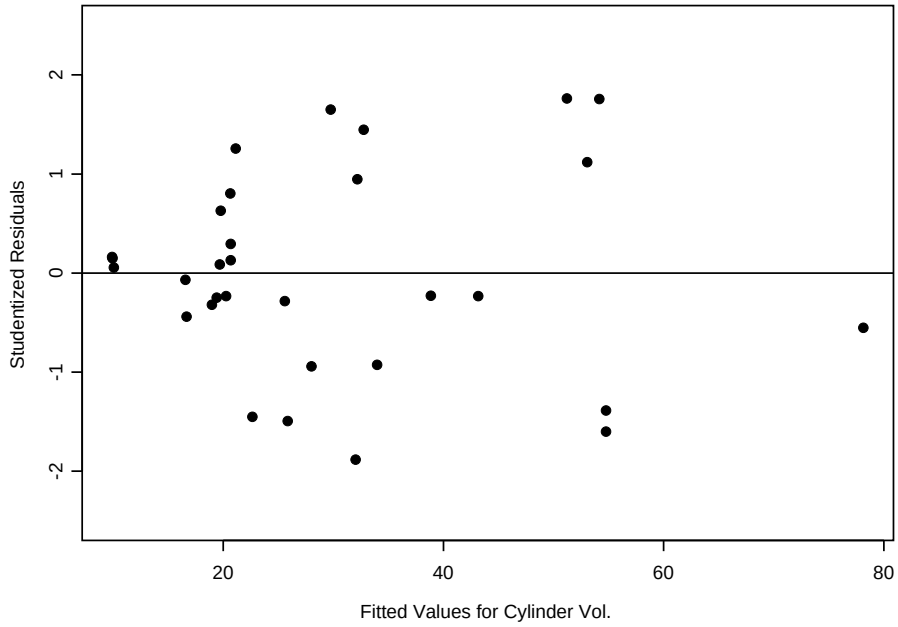

Figure 7.19: Studentized residuals for the regression of volume on cylinder.

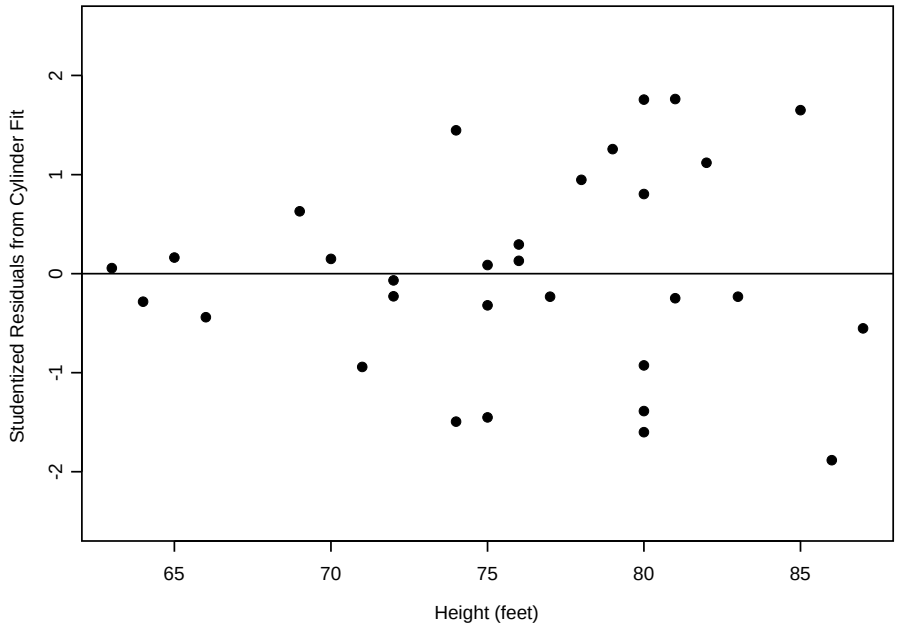

Figure 7.20: Studentized residuals for the regression of volume on cylinder plotted against values of height.

the table immediately following Figure 7.12, we have reduced the estimate of σ^2 by more than half, and increased R^2 over the regression with only DBH by more than twice the increase resulting from the multiple regression model. Perhaps more importantly, there is nothing in the residual plot of Figure 7.19 to indicate that our expectation function is lacking in form.

We might wonder if there remains a relation between variance and height for this regression. Plotting studentized residuals from the fit of model (7.12) against height ($x_{2,i}$) results in the plot of Figure 7.20. This plot suggests that there is still a relation of the variability in tree volumes, after adjusting for the effects of height and DBH through use of the variable cylinder, to the

measurement of tree height. A similar plot of residuals against DBH looks nearly identical to the plot of Figure 7.19 and is not presented.

Putting these preliminary analyses of tree geometry together, we should be willing to entertain a model of the general form (7.10), with $z_i \equiv x_{2,i}$ and $\mu_i(\beta) = \beta_0 + \beta_1 \, \phi(\boldsymbol{x})$. Possible forms for $\sigma g(z_i, \, \theta)$ would include

$$\begin{aligned} \sigma \, g(z_i, \, \theta) &= \theta_0 + \theta_1 \, z_i \\ &\text{and} \\ \sigma \, g(z_i, \, \theta) &= \sigma \exp\{\theta_0 + \theta_1 \, z_i\} \end{aligned} \tag{7.13}$$

We will (hopefully) discuss additional plots that are useful in selecting a variance model later in the course.

7.2.5 Transform Both Sides Models

We close discussion of additive error models with a brief mention of one additional modeling idea, promoted by Carroll and Rupert (1988). While this idea, transforming both sides of a theoretical relation between a response variable and a set of covariates (including the case in which covariates are group membership indicators) has been used in a number of particular situations over the years (see Carroll and Rupert, 1988, pages 119-121) apparently Carroll and Rupert (1984) were the first to suggest this methodology as a general modeling strategy.

Consider again the quote from Carroll and Rupert (1988) presented at the beginning of Chapter 7.2 of these notes, which reflects the concept that a response y may sometimes (i.e., in some, but not all, problems) be viewed as a deterministic function of a covariate x *if there were no sources of uncertainty.*

This concept feeds nicely into the use of additive constant variance models (Section 7.2.1) at least as an initial step. Evidence of nonconstant variance may be exhibited (e.g., Example 7.3), leading to the desire to either (1) transform the responses, or (2) model the heteroscedastic variances. In Example 7.3 it was seen that modeling the variances may be preferable if transformation "destroys" what was a pleasing expectation function (in that example this was an empirical linear function, in other situations it may be a theoretical model). In addition, transformation of responses affects distributional form, a topic we have not discussed explicitly but certainly underlies, for example, our discussion of potential limitations to the use of a normal distribution (see Example 6.7).

Three fundamental departures from an additive error model with constant variance are:

1. Incorrect specification of the expectation function.

2. Nonconstant (heteroscedastic) error variances.

3. Nonsymmetric error distributions (usually non-normal error distributions).

It can, in fact, be difficult to separate these three types of departures from an additive error model with constant variance. For example, is the pattern of residuals in Figure 7.4 really due to heteroscedastic error variances (the focus of that example), or might there be evidence of either a nonlinear expectation function (there is some hint of an inverted U pattern), or an error distribution that is skew left (count points above and below the zero line)?

Modeling of variance structures (what Carroll and Rupert call "weighting", in deference to the general form of equation (7.3) as discussed in Section 7.2.3)

addresses heteroscedastic error variance without changing either the expectation function or the assumed distribution of (location-scale family) errors. Transformation affects all three properties listed above. In elementary courses it is generally assumed that all three aspects of expectation function, variance structure, and error distribution "go together", and that transformation to "fix" one of these problems also "fixes" the others, or at least does no damage. Examples 7.1, 7.3, and 7.4 are all indications that this is not, in general, the case.

The *transform both sides* methodology was developed in response to situations in which there exists a fundamental expectation function for the original variables that we do not wish to change, and yet there is evidence of either nonsymmetry or nonconstant variance for additive error terms. In particular, nonsymmetric error distributions may indicate that, in the original scale of observation, an additive error model is not really appropriate (since additive error models essentially imply location-scale distributions which are typically symmetric). The basic idea is that we begin with a model of the form,

$$Y_i = g(x_i, \beta) + \text{error},$$

where $g(\cdot)$ has scientific meaning or is a pleasing empirical form (e.g., linear), but for which the error term does not led itself to modeling through a location-scale specification. To "fix" the problem with error specification, but without changing the expectation function beyond hope, we might transform the responses Y_i to produce "nice" error terms but also transform g to maintain the basic relation between responses and covariates. This leads to the transform both sides (TBS) model,

$$h(Y_i, \lambda) = h\{g(x_i, \beta), \lambda\} + \sigma\, \epsilon_i, \tag{7.14}$$

where, for $i = 1, \ldots, n$, $\epsilon_i \sim iid\, F$ with $E(\epsilon_i) = 0$, usually $var(\epsilon_i) = 1$, and frequently F is $N(0, 1)$.

Because it is not a certainty that a transformation $h(\cdot, \lambda)$ will have appropriate effects on *both* symmetry and constancy of error variance, model (7.14) can be extended to include additional modeling of variance structure as,

$$h(Y_i, \lambda) = h\{g_1(x_i, \beta), \lambda\} + \sigma\, g_2(x_i, \beta, z_i, \theta)\, \epsilon_i, \tag{7.15}$$

where assumptions on the error terms ϵ_i are the same as for (7.14). In the same way that we moved from model (7.4) to model (7.5) and model (7.9) to model (7.10), if the variance portion of (7.15) depends on β only through the expectation function g_1, we may write

$$h(Y_i, \lambda) = h\{\mu_i(\beta), \lambda\} + \sigma\, g(\mu_i(\beta), z_i, \theta)\, \epsilon_i. \tag{7.16}$$

Now, the models given in (7.15) and its reduced version in (7.16) are very general structures indeed. A word of caution is needed, however, in that one can easily use these models to produce the statistical version of "painting oneself into the corner". This stems from the fact that it is not merely diagnosing differences among the three effects listed previously that is difficult, but also modeling them separately. For example, probably the most common form of the transformation h is a power transformation $h(Y_i, \lambda) = Y_i^{\lambda}$, but this is also a common form for the variance model g_2 in (7.15) or g in (7.16). Including both of these in a model such as (7.16) would result in,

$$Y_i^{\lambda} = \{\mu_i(\beta)\}^{\lambda} + \sigma\, \{\mu_i(\beta)\}^{\theta}\, \epsilon_i.$$

This model will prove difficult if one wishes to estimate both λ and θ simultaneously. In principal, such problems can be avoided (e.g., a power transformation is often used to remove dependence of variance on mean so that $\mu_i(\beta)$

can probably be eliminated from the variance model), but they are certainly a consideration in model formulation. There are, also, difficulties in deriving predictions (and the associated intervals) from such a model with anything other than a "plug-in" use of parameter estimates. That is, uncertainty in parameter estimates are not reflected in predication intervals. As Carroll and Rupert (1988, page 151) indicate, "More research is needed on predication intervals based on transformation models."

7.3 Models Based on Response Distributions

We turn now to an approach to model formulation that differs in a radical way from the concept of additive error. It is this approach to modeling that I referred to in Chapter 5 of these notes as being greatly influenced by what are called *generalized linear models*. While this is true, we should avoid thinking that generalized linear models encompass all that is available in this approach; in fact they are but a small, and fairly restricted, subset.

7.3.1 Specifying Random Model Components

All of the model formulations in Chapter 7.2 on additive error models can be thought of as mathematical representations of *signal plus noise*. Throughout that section we increased the complexity with which we were willing to model the noise (error) component, and attempted to deal with lack of independence between signal (i.e., expectation function) and noise (i.e., error) resulting from modifications to the model by transformations and weighting.

A radical departure from that strategy of formulating models as signal plus noise is to consider models as consisting of *random* and *systematic* components

which may combine in a nonadditive fashion. While the systematic model component is essentially the same thing as the expectation function in additive error models, the random component refers to the basic distribution of response variables, not additive errors.

While for additive error models we generally begin model formulation by specifying a form for the expectation function or systematic model component, the models under discussion in this section are usually formulated by first determining an appropriate distribution for response variables if any other factors involved in the model are held constant. The last portion of the preceding sentence is important. One cannot examine the appropriateness of an assumption of normality for additive errors in a simple linear regression model by examining a histogram (or conducting a test) of response observations across levels of the covariate. Similarly, one cannot determine an appropriate random model component (normal or otherwise) by looking at the empirical distribution of responses across all levels of other factors that may be involved in the problem.

Considering the random model component first has sometimes been criticized because, due to the fact that one often needs some type of "residual" or "conditional quantity" to assess distributional assumptions, attempting to specify a distribution before a mean structure results in something of a "catch 22"; you need residuals to examine conditional distributions, but you need means to define residuals, but you need conditional distributions before specifying the structure of means. This argument is essentially a red herring (something that needlessly confuses an issue). An assumption of normality is just as much a distributional assumption as any other specification. A full model is arrived at regardless of the order in which one considers systematic and random model components. All initial models should be assessed relative to the specification of both systematic and random components, and adjusted

accordingly. What then, becomes the difficulty with giving first consideration to the random rather than systematic model component? I'm left with the conclusion that for many statisticians who have been raised on additive error models the answer is that it is simply not something that they are accustomed to doing. This said, there does remain a valid question about how one is to start the process of considering distributional form first.

In the development of an additive error model we may have the first step handed to us as a consequence of scientific theory (this is the happy situation). If not, we generally look first at a graphical displays such as histograms and side-by-side boxplots (for a problem with groups) or a scatterplot or scatterplot matrix (for a regression problem). The same type of displays are useful in considering random components as well as systematic components. If one constructs a scatterplot of responses on one covariate quantity and sees a increasing curve with a "fan-shaped" scatter, attention is immediately focused on distributions for which the variance increases as a function of the mean (e.g., gamma or inverse Gaussian). If one constructs boxplots for which the "tails" extend about the same amount from both ends of the box, attention is immediately drawn to symmetric distributions (e.g., normal or logistic). If one constructs a histogram in which there appear to be two or even three local modes, attention is drawn to finite mixture distributions (more on this to come). This is really nothing different than what we always do, except the focus of attention is shifted from looking at means first to looking at distributional characteristics.

Even prior to the examination of basic data displays, however, it is often possible to form a preliminary idea of the kinds of distributions that might be appropriate. This is tied in with the definition of random variables and the sets of possible values Ω that are attached to them (see Chapter 5.1 and,

in particular, Example 5.1). If Ω consists of the non-negative integers, we may begin thinking of Poisson-like distributions. If Ω is the positive line, we may begin thinking of distributions with this support such as gamma, inverse Gaussian, or lognormal distributions or, if the observable quantities to be modeled are of great enough magnitude we might be thinking of a normal approximation.

Sometimes it is even possible to construct random variables that have particular sets Ω. This occurs, for example, in situations for which the observation or measurement process is not conducted on a well-defined physical quantity.

Example 5.2 (cont.)

Consider again Example 5.2 from Chapter 5.1, which was about the effect of violent cartoons on aggression in young children. Suppose that, in this study, children were (randomly would be a good idea) divided into four groups. One group were read "happy" children's stories. One group were read "violent" children's stories (e.g., Grimm's Fairy Tales or Little Red Riding Hood or the Three Pigs). One group viewed "happy" cartoons (e.g., Barney, or Winne the Pooh). Finally, one group viewed "violent cartoons" (most of the Looney Tunes). The intent of these groups, of course, is to investigate possible interaction between violent content *per se* and the media of communication; this possible interaction is confounded with the reality of the violence – in Little Red Riding Hood, for example, both Grandma and the Wolf actually die of non-natural causes, while the Coyote always recovers from being crushed or blown up in the Roadrunner cartoons. Suppose that the observation process to assess aggression is to present the children with a series of (maybe 10 to 20) pictures judged by psychologists to appeal to aggressive instincts. The

children are asked whether they "like", "do not like", or "don't care about" each picture.

How might we go about defining random variables by which to compare the "treatment" groups in this example, and what type of a distribution might be considered for such random variables? There is no one answer to this question. Some possibilities include the following.

Binomial Formulation

One could choose to ignore the "do not like" and "don't care about" responses in the study and focus only on the number of "like" responses, which supposedly indicate aggressive tendencies; the more "like" responses, the more aggression. For each child and picture, then, define

$$X_{i,j} = \begin{cases} 0 & \text{if response was not "like" to picture } j \\ 1 & \text{if response was "like" to picture } j \end{cases},$$

where $i = 1, \ldots, n$ indexes child and $j = 1, \ldots, m$ indexes picture. Combine these variables for each child as

$$Y_i = \frac{1}{m} \sum_{j=1}^{m} X_{i,j},$$

where now $\Omega_Y \equiv \{0, (1/m), \ldots, 1\}$. It is natural to consider a binomial specification for the probability mass functions of Y_i as, for $y_i = 0, (1/m), \ldots, 1$,

$$\begin{aligned} f(y_i|\theta) &= \frac{m!}{(my_i)!\,(m - my_i)!}\, p^{my_i}\,(1-p)^{m-my_i} \\ &= \exp\left[m\{y_i\theta - b(\theta)\} + c(y_i, m)\right], \end{aligned} \tag{7.17}$$

where $\theta = \log(p) - \log(1-p)$, $b(\theta) = \exp(p)/(1+\exp(p))$, and $c(y_i, m) = \log\{m!\} - \log\{(my_i!)\} - \log\{(m - my_i)!\}$.

Our modeling exercise would now consist of determining how many values of θ are necessary; same θ for everyone; one θ for each group; one θ for each child (more on this later under mixture models). One advantage of expressing this binomial distribution (for observed proportions) as an exponential dispersion family in (7.17) is that we immediately have available expressions for the mean and variance of Y_i, and also an unbiased estimator of $b'(\theta)$ from any joint distribution for a collection of Y_is assumed to have the same θ.

Multinomial Formulation

A multinomial formulation is very similar to the binomial model, and its development will be left as an exercise. Note here, however, that each Y_i will be a vector, as will each p, but we will only include two components in these vectors due to a "bounded sum" condition on their elements.

Beta Formulation

An alternative to the binomial and multinomial models would be to define random variables having a conceptual set of possible values $\Omega_Y \equiv (0,\ 1)$ as follows. First, for each child (indexed by i) and picture (indexed by j) combination, define

$$X_{i,j} = \begin{cases} 0 & \text{if response was "do not like"} \\ 1 & \text{if response was "don't care"} \\ 2 & \text{if response was "like"} \end{cases}$$

Note that this assumes some sort of ratio scale for these categories, which might be a question. Now define, for each child, the random variable

$$Y_i = \frac{1}{2m} \sum_{j=1}^{m} X_{i,j}.$$

The value $2m$ is the "maximum" aggression score possible, so that Y_i represents

the proportion of maximal aggression that is exhibited by a child. This variable is more easily handled than binomial or multinomial formulations in the case that not all children provide recorded responses for all pictures (children of this age group are unpredictable, may refuse to reply to some questions, get picked up early by their parents, etc.); we simply replace m by m_i in the above definition of Y_i. On the other hand, our assumption that this theoretical construct has possible values in the interval $(0, 1)$ may cause difficulties if most m (or m_i) values are not large enough to prevent many "ties" in the data; a potential way around this difficulty will be discussed in the portion of the course on likelihood estimation.

We might now assign the random variables Y_i beta distributions, based on the fact that a beta distribution has support that matches Ω_Y and is quite flexible in shape; a beta density may be unimodal symmetric, unimodal skew (both left and right), $J-$shaped, or $U-$shaped. We can then assign distributions as,

$$f(y_i|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, y_i^{\alpha-1}\,(1 - y_i)^{\beta-1}, \tag{7.18}$$

for $y_i \in (0, 1)$. Model (7.18) may again be expressed in exponential family form, although for this a two-parameter family is required. Analysis of this hypothetical example would then proceed in a manner similar to that mentioned under the binomial formulation in determining how many distinct values of the parameters (α, β) are necessary to adequately account for the entire set of children.

One advantage (a major one in my opinion) of the beta formulation is that it does not assume that individual "trials" (corresponding to decisions on individual pictures by individual children) are independent and identically distributed. With the same set of pictures shown to each child, this is the

only (other than totally artificial) way to have the records for children within a group result in identically distributed replicates of the same multinomial. Consider, for example, 3 children from each of two groups, "happy readers" and "violent readers". Suppose the following responses were obtained:

		Category			Beta
Group	Obs.	0	1	2	Score
Happy					
Readers	1	4	8	8	0.60
	2	6	5	9	0.58
	3	4	9	7	0.58
Violent					
Readers	1	10	5	5	0.38
	2	3	12	5	0.55
	3	1	5	14	0.82

Consider the multinomial formulation for this problem. Without going into a full discussion of estimation, if we just wanted unbiased estimates of the multinomial probabilities (probabilities that individual observations fall in each category) these are immediate from (1) expressing the multinomial in exponential family form, and (2) the result of expression (6.8) of Section 6.1.4 for joint distributions.

The result is that, for *iid* random variables $\{X_{i,j} : i = 1, \ldots, n; j = 1, \ldots, m\}$ defined as indicated under the heading Beta Formulation, and a corresponding multinomial formulated using $Y_{i,1} = \sum_j I(X_{i,j} = 0)$ and $Y_{i,2} = \sum_j I(X_{i,j} = 1)$, parameterized with $p_0 \equiv Pr(X_{i,j} = 0)$, $p_1 \equiv Pr(X_{i,j} = 1)$ and $m = 20$, unbiased estimators are,

$$\hat{p}_0 = \frac{1}{mn} \sum_{i=1}^{n} Y_{i,1},$$

$$\hat{p}_1 = \frac{1}{mn}\sum_{i=1}^{n} Y_{i,2}.$$

Thus, for the data in the table above, $\hat{p}_0 = 0.233$ and $\hat{p}_1 = 0.367$ for both the happy reader and violent reader groups. It is impossible, then, that any assessment would indicate the distributions differ for these two groups. But it seems clear from the numbers, and is reflected in the "Beta Score" values, that there is much greater uniformity in responses for the happy reader group than for the violent reader group. The difficulty for the multinomial formulation illustrated in this example is that the assumption of *iid* individual trials (and, hence, replicate observations from the same multinomial) may not be reasonable.

7.3.2 Generalized Linear Models

What are known as *generalized linear models* are a class of (typically, but not necessarily) nonlinear models that begin by specifying response random variables $\{Y_i : i = 1, \ldots, n\}$ as following probability density or mass functions that belong to exponential dispersion families of the form of expression (6.5); note that this immediately implies the properties given in expression (6.6). We do not, however, need to consider these random variables as *iid*, although we will allow them to differ only through their natural parameters (θ_i), but not in what is assumed to be a constant dispersion parameter (ϕ). For the set of response variables, then, we can write the pdf or pmf functions as,

$$f(y_i|\theta_i) = \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i,\, \phi)\right], \tag{7.19}$$

and the properties of expression (6.6) as,

$$\mu_i \equiv E(Y_i) = \frac{d}{d\theta_i} = b'(\theta_i),$$

$$var(Y_i) \quad = \quad \frac{1}{\phi}\frac{d^2}{d\theta_i^2}b(\theta_i) = b''(\theta_i) = \frac{1}{\phi}V(\mu_i).$$

As mentioned way back in the introduction to this part of the course on modeling, it was the advent of generalized linear models (Nelder and Wedderburn 1972; McCullagh and Nelder 1989) that gave rise to the terminology of *systematic* and *random* model components. The random model component is given by expression (7.19). The systematic model component (in generalized linear models) consists itself of two parts, the *linear predictor* and the *link function.*

The linear predictor is exactly what it sounds like, and is usually represented as a typical linear model. For random variable Y_i this is

$$\eta_i = \boldsymbol{x}_i^T \beta, \tag{7.20}$$

for $\boldsymbol{x}_i^T = (x_{1,i}, x_{2,i}, \ldots, x_{p,i})$ a vector of covariates associated with Y_i. Often, as in linear models, the first of these covariates plays the role of an intercept term as $x_{1,i} \equiv 1$, but this is neither necessary, nor does "intercept" always have the same interpretation as for linear models.

The other portion of the systematic model component is the link function, and is defined as the relation,

$$g(\mu_i) = \eta_i. \tag{7.21}$$

As for additive error models, the covariates may be simply group indicators, assigning a separate fixed value of expectations to random variables that are members of certain groups. In the case that the covariate vectors $\boldsymbol{x}_i$ contain one or more quantities that function on a ratio scale of measurement (e.g., continuous covariates) the link function $g(\cdot)$ is a monotonic function of the linear predictors η_i.

Note at the outset that there exists a duplicity of notation in generalized linear models. Since $\mu_i = b'(\theta_i)$ for a simple function $b(\cdot)$, there is a one-to-one relation between the expected value of Y_i and the exponential dispersion family natural parameter θ_i. So, we could equally well write expression (7.21) as $g(b'(\theta_i)) = \eta_i$. The link function $g(\cdot)$ is generally taken as a (often nonlinear) smooth function and is given its name because it "links" the expected values (and hence also the natural parameters) of response pdfs or pmfs to the linear predictors.

There is a special set of link functions called *canonical* links which are defined as $g(\cdot) = b'^{-1}(\cdot)$. The name stems from the fact that what I have usually called *natural* parameters are also known as *canonical* parameters in exponential families. Canonical link functions have the property that, if $g(\cdot)$ is a canonical link for the specified random model component, then,

$$g(\mu_i) = b'^{-1}(\mu_i) = b'^{-1}(b'(\theta_i)) = \theta_i.$$

For particular common random components, the corresponding canonical link functions may be seen to be:

1. Normal random component: $g(\mu_i) = \mu_i$

2. Poisson random component: $g(\mu_i) = \log(\mu_i)$

3. Binomial random component:
   $g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$

4. Gamma random component: $g(\mu_i) = 1/\mu_i$

5. Inverse Gaussian random component:
   $g(\mu_i) = 1/\mu_i^2$

Note here that, in particular, the binomial random component is assumed to be written in terms of random variables associated with observed proportions as in expression (7.17). Now, since, for independent exponential dispersion family random variables, the joint distribution is of exponential family form with sufficient statistic $\sum_i Y_i$, canonical links lead to $\theta_i = \eta_i = \boldsymbol{x}_i^T \beta$ and thus sufficient statistics for each of the of the β_j that consist of $\sum Y_i \, x_{j,i}$. While this is a "nice" property, there is nothing particularly special about what it allows in practice, and we should avoid attaching any "magical" properties to canonical link functions.

What link functions must, under most situations, be able to do is map the set of possible expected values (i.e., the possible values of the μ_i) onto the entire real line, which is the fundamental range of linear predictors $\eta_i = \boldsymbol{x}_i \, \beta$ (unless we restrict both $\boldsymbol{x}_i$ and β). For example, any link function appropriate for use with binomial random variables must map the interval (0, 1) onto the real line. This makes, for example, the use of an identity link function $g(\mu_i) = \mu_i$ potentially dangerous with a binomial random component. A similar situation exists for Poisson random components, although constraints on the allowable values of the covariates and the "regression" parameters in β may allow, for example, the use of an identity link with a Poisson random component. Other common link functions, without attaching them to any particular random components, include:

1. Log link: $g(\mu_i) = \log(\mu_i)$ for $0 < \mu_i$.

2. Power link: $g(\mu_i) = \begin{cases} \mu_i^\lambda & \lambda \neq 0 \\ \log(\mu_i) & \lambda = 0 \end{cases}$

   for any fixed λ and $-\infty < \mu_i < \infty$.

3. Complimentary Log-Log Link:
   $g(\mu_i) = \log\{1 - \log(1 - \mu_i)\}$ for $0 < \mu_i < 1$.

It is also possible to embed link functions into parameterized families of functions, without specifying the value of the parameter. One example is the power family, for which an equivalent specification to that given in item 2 immediately above is,

$$g(\mu_i, \lambda) = \frac{(\mu_i^\lambda - 1)}{\lambda}.$$

We then would need to estimate the parameter λ along with all of the other parameters of the model (we may or may not get to this, but see Kaiser 1997).

One additional aspect of the generalized linear model formulation is of fundamental importance, that being the variance function $V(\mu_i)$. This function is proportional to the variance of the response variables Y_i (equal up to the scaling factor $1/\phi$, see expression (7.19) and what follows immediately after). The variance function is not something that is open to specification in the model, but is determined by the choice of random component. For some of the more common random components, the variance function takes the forms:

1. Normal random component: $V(\mu_i) \equiv 1$.

2. Poisson random component: $V(\mu_i) = \mu_i$.

3. Binomial random component:
   $V(\mu_i) = \mu_i(1 - \mu_i)$.

4. Gamma random component: $V(\mu_i) = \mu_i^2$.

5. Inverse Gaussian random component:
   $V(\mu_i) = \mu_i^3$.

Keeping in mind that specific random components imply specific variance functions, which dictates the relation between means and variances, and combining this with knowledge of the set of possible values for response variables Ω, the examination of scatterplots can often provide *hints* about potentially useful random component specifications (although this can almost never distinguish among several possibilities).

Example 7.6

Contamination of fish by metals can be a serious problem, due to both effects on ecosystem function and potential health effects for humans. Metals can be accumulated by fish through a number of mechanisms, including primary routes of uptake from water through the gill tissue and dietary uptake (by eating other contaminated organisms). Particularly for uptake from direct contact with contaminated water (i.e., absorption in the respiratory system through gills) many metals exhibit a change in what is called "bioavailability" with changes in other water chemistry variables, primarily as reflected in PH or acidity of the water. That is, in more acidic waters, many metals become dominated by ionic forms that are more easily bound to biological tissues than the ionic forms that predominate in less acidic waters. Thus, the more acid the water, the more accumulation of a metal occurs, even with constant (total) concentrations of the metal in the water itself. This is perhaps exemplified by aluminum (Al) which, in water that is neutral pH (pH= 7.0), is relatively inert and is not a problem in fish or other aquatic organisms. But, in water that is low pH (pH < 5.5), Al assumes an ionic form that is readily accumulated by aquatic organisms and becomes highly toxic. On the other hand, for metals with a primary route of uptake from gill contact with water, lower pH can

(at least initially) work in the opposite manner, since H ions can compete for binding sites on gill tissue with bioavailable forms of metals. If the pH of a lake were to decrease relatively rapidly, the issue becomes partially one of whether H ions or bioavailable forms of metals are becoming abundant more rapidly.

This phenomenon of metals becoming more available and toxic at low water pH values has been a concern for states with large amounts of sport fishing. such as Minnesota, Wisconsin, Michigan and, to a lesser extent, Iowa.

A great deal of the materials that lower the pH of a lake come from atmospheric deposition (e.g., what is called "acid rain"), a major source being power plants that burn "high-sulfur" coal. But because of the complexity of the chemical and physiological processes involved, there has been a good deal of controversy about whether decreasing pH values in lakes of these states constitutes a serious environmental problem (at least as regards metal concentrations in fish).

A study was conducted beginning in the mid-1980s to help resolve some of these issues (see Powell, 1993). Based on the concept of the experimental approach (see Part 1 of the course notes) that there are no such things as "natural" experiments, a lake in north-central Wisconsin was used for an experimental acidification study. This lake, called Little Rock Lake, has no natural inflow or outflow, but receives all of its water from rainfall and runoff from the surrounding watershed. The lake consists of two natural basins, which are connected by a thin stretch of water (think of an hourglass). In the early fall of 1984 a special vinyl curtain was used to separate the two basins and prevent the exchange of water between them. One of the basins, called the "reference" basin, was not tampered with. The other, called the "treatment" basin, was artificially acidified over a period of years. The basins were quite similar in morphology; the reference basin was 8.1 ha in size with a mean depth

of 3.1 m, while the treatment basin was 9.8 ha in size with a mean depth of 3.9 m. The beginning pH of both basins was 6.1, but the pH of the treatment basin was lowered to 5.6 by the fall of 1986, to 5.2 by the fall of 1988, and to 4.9 by the fall of 1990.

The data used in this example comprise measurements of the whole body concentration of the metal Cadmium (Cd) and length of fish taken on one of the principal forage fishes (fish that are eaten by other fish) in Little Rock Lake, namely Yellow Perch. Length is an important covariate for the concentration of nearly all pollutants in fish; growth in fish is indeterminate, making length an indicator of the length of exposure for one thing. The data were collected in the summer of 1989 after the pH of the treatment basin had been reduced to 5.2 from its original value of 6.1. Observations verified that the pH of the reference basin remained at a mean of 6.1 at this time.

The objectives of an analysis with these data centers on what was called *problem conceptualization* on page 231 of Section 3.1. Certainly, we would like to know if acidification makes the concentration of Cd in Yellow Perch greater or lesser, but the problem is clearly not that simple, particularly given the interaction of changing H ion concentrations relative to biologically available forms of metals, as mentioned above. What we would really like to achieve is a statistical model that seems to "capture' the important features of the relation between Cd concentration and length in this fish. We must admit that what we mean by the important features of the relation is, at this point, based primarily on empirical patterns in the data. On the other hand, we would like a parsimonious description of the distribution of Cd concentrations for given lengths (since, for example it may the the upper quantiles rather than the mean that are of primary concern in the protection of human health).

With all of this in mind, we may examine scatterplots of Cd concentration

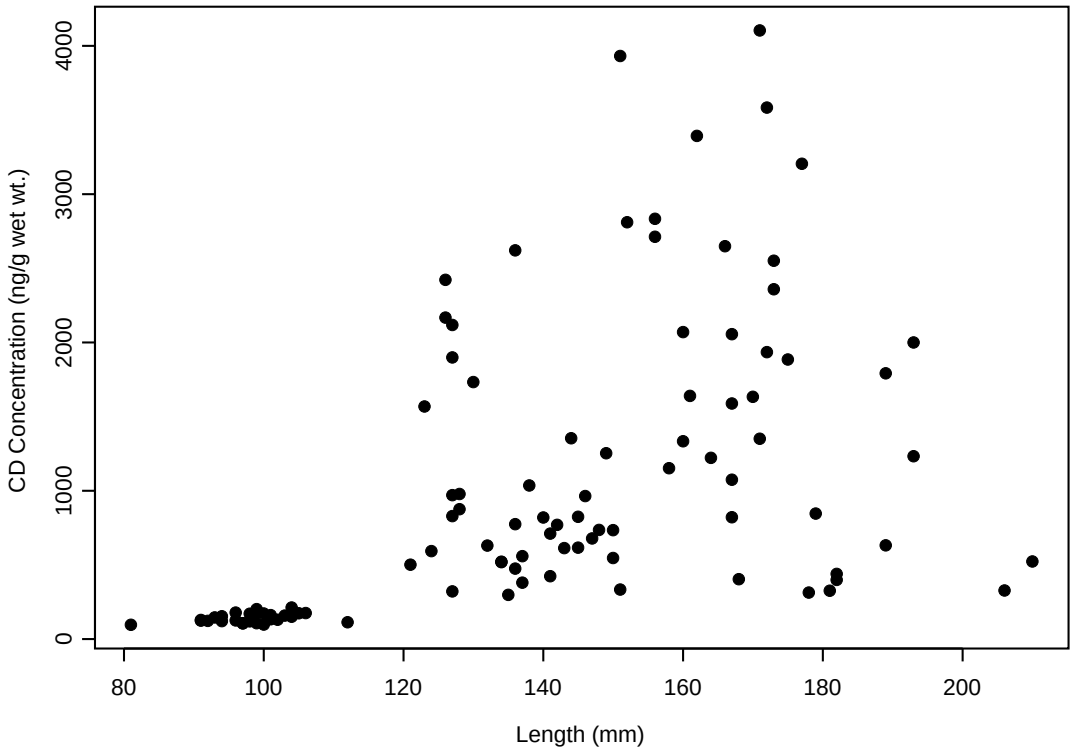

Figure 7.21: Scatterplot of Cd concentration against length in Yellow Perch from Little Rock Lake, WI for the reference basin.

(which was measured in ng/g) against fish length (which was measured in mm). A plot for the reference basin is presented in Figure 7.21, and one for the treatment basin is presented in Figure 7.22. These two plots, which have the same plotting scales, are rather dramatically different in what they suggest about the relation between Cd concentration and length in Yellow Perch in this lake. Clearly, whatever the effect of acidification, it involves more than the expectation function of responses. Also, it appears that, in both basins, the variability of responses increases as the mean increases. What is the effect

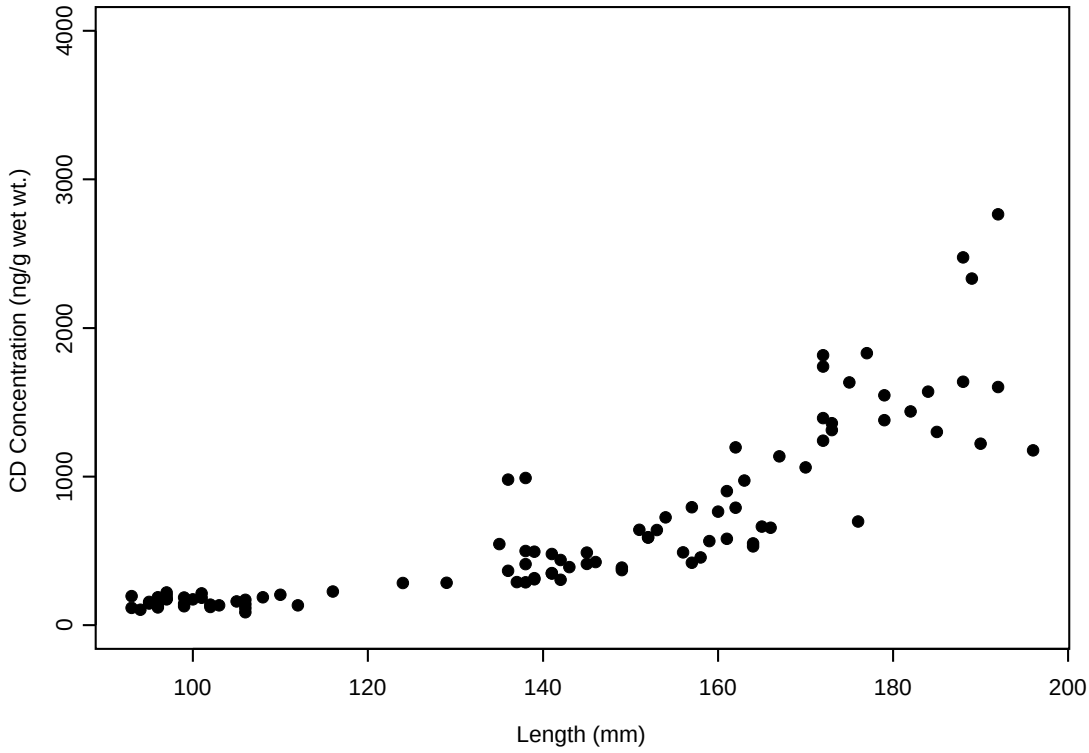

Figure 7.22: Scatterplot of Cd concentration against length in Yellow Perch from Little Rock Lake, WI for the treatment basin.

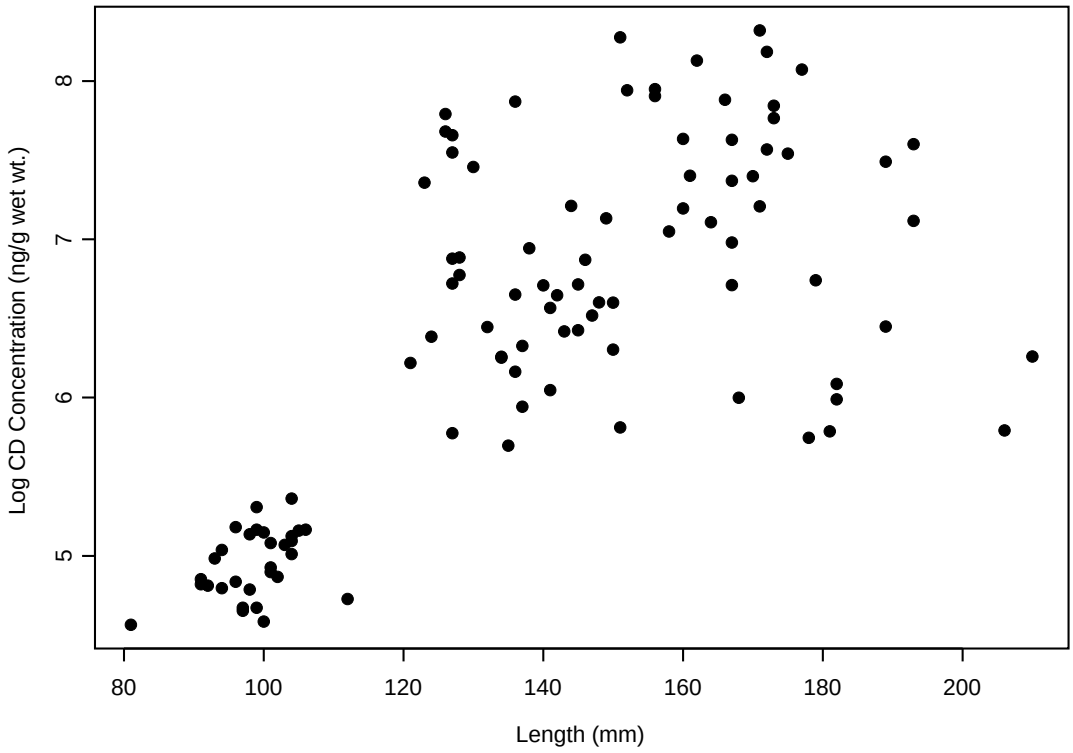

Figure 7.23: Scatterplot of log Cd concentration against length in Yellow Perch from Little Rock Lake, WI for the reference basin.

of a log transformation, for example, on these data? Plots of the logarithm of Cd concentration against length for the reference and treatment basins are presented in Figure 7.23 and 7.24, respectively. The log transformation appears to have done a fairly nice job of rendering the expectation function for the treatment basin both linear and nearly constant variance (Figure 7.24), but the same cannot be said for the reference basin (Figure 7.23).

Perhaps the log transformation was simply not "strong" enough. What happens if we try a more severe transformation, say a reciprocal? The results

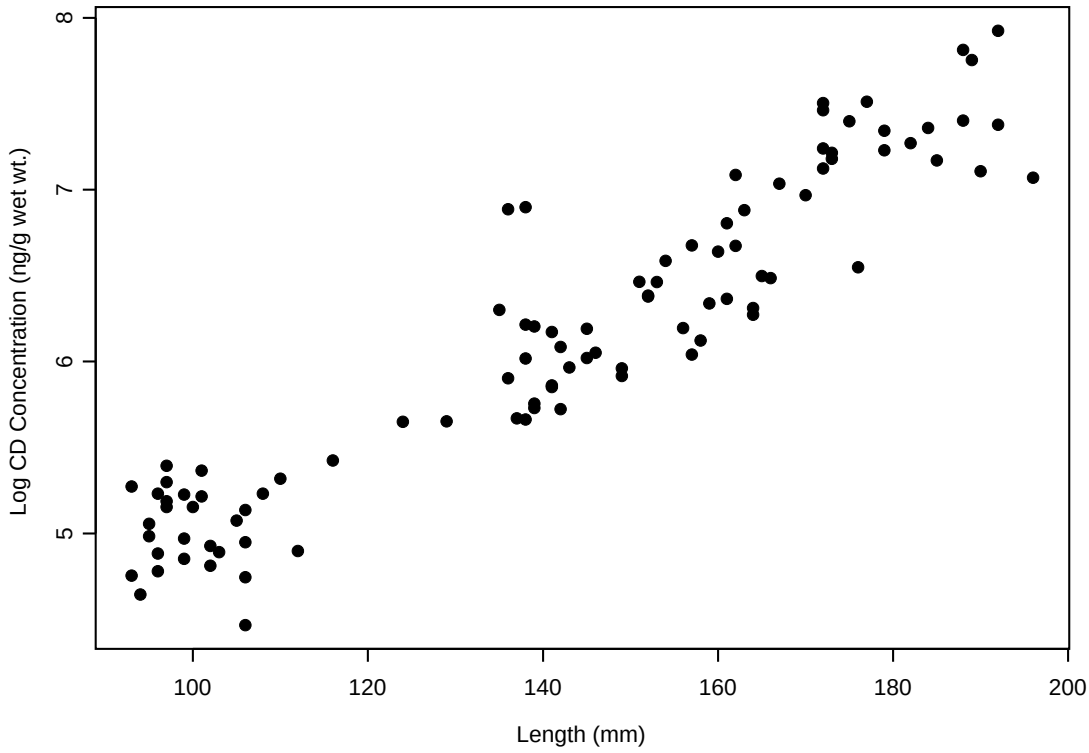

Figure 7.24: Scatterplot of log Cd concentration against length in Yellow Perch from Little Rock Lake, WI for the treatment basin.

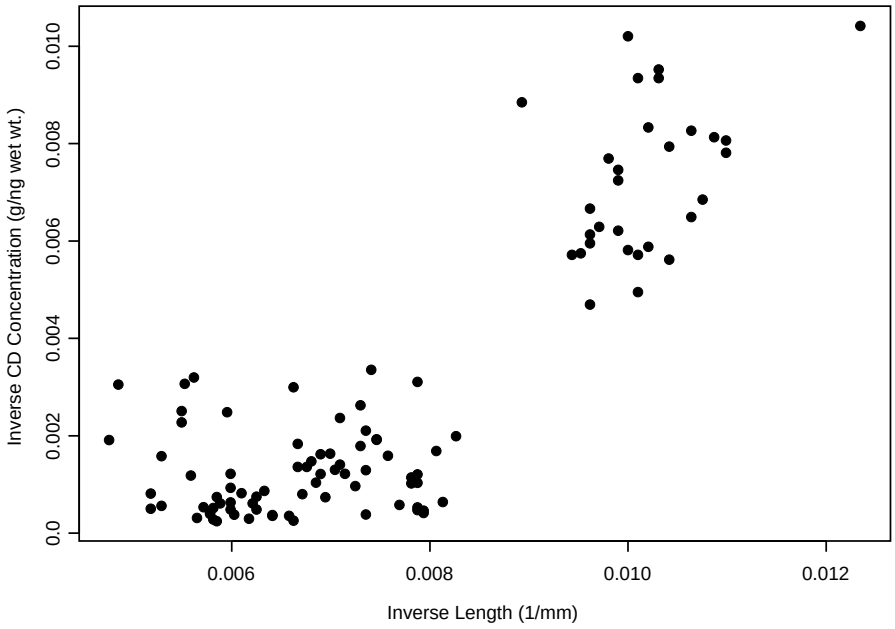

Figure 7.25: Scatterplot of reciprocal Cd concentration against reciprocal length in Yellow Perch from Little Rock Lake, WI for the treatment basin.

are shown in Figure 7.25, in which we have also used the same transformation on length to prevent the relation from becoming "reversed"; this is essentially the "transform both sides" idea of Section 7.2.5. Certainly, Figure 7.25 seems to exhibit more constant variance than do either Figures 7.21 or 7.23. It is also clear, however, that an additive normal error model fit to these data would suffer from the difficulties of Example 6.7 of Section 6.2.2 with normal distributions putting positive probability on regions that are outside of the physical reality of the response variables.

Where does this leave us in model development? First of all, it indicates that the distributions of original responses (Cd concentration in ng/g) seems to

differ between the reference and treatment basins. Thus, we strongly suspect that the effect of acidification (in the treatment basin) has affected more than the expected value of responses. If, as suggested in the description of Example 7.6, quantiles of the (conditional on length) distributions of Cd concentration in these two basins is of interest, this is a major issue for modeling. Consider, for example, fitting linear regression models to the log transformed Cd concentrations against length for the treatment basin data of Figure 7.24, and similarly for the reciprocal transformed Cd concentrations against reciprocal length for the reference basin data of Figure 7.25 (ignore, for the moment the difficultly with a normal error assumption in the latter). How would we compare these models with regard to the relation of Cd concentration to length in these two situations? What would it mean if our two regressions had the same or different slopes? Would it even be possible (without a great deal of pain, at least) to estimate the 85% or the 90% level of Cd (not transformed Cd against possibly transformed length) for fish of length 150 mm for these two basins?

Let's go back to the scatterplots of the measured variables in Figures 7.21 and 7.22. What might we be able to discern from these figures that could help us form a combination of random and systematic model components for a potential analysis using generalized linear models? First, it is clear that, for both basins, the expectation function is nonlinear in length. In your "minds eye" envision a curve through the points of Figure 7.21. Most reasonable such curves would be such that more points lie above the curve than below (based on the "density" of points on the plot). Thus, our attention is drawn to random model components that are skew right. Possibilities include gamma and inverse Gaussian among the basic exponential dispersion families. Now, conduct the same exercise for the data of Figure 7.22. Here, it is less clear. It could be that a skew random component is appropriate (especially given the

two points at length of about 140 mm) but this is by no means certain to be better than a symmetric random component, most likely a normal. On the other hand, the variance in Figure 7.22 does appear to increase as a function of the mean, which casts additional doubt on the appropriateness of a straight normal specification (for which $V(\mu_i) \equiv 1$). Thus, for the reference basin we might initially consider gamma and inverse Gaussian random components, and for the treatment basin, we might consider gamma, perhaps inverse Gaussian, and perhaps normal random components.

In both Figure 7.21 and Figure 7.22 the mean seems to be a concave curve that increases with length. Our initial attention is then drawn to an exponential-like function, or perhaps a positive power function. These correspond to the (inverse form of) link functions $g(\mu_i) = \log(\mu_i)$ or $g(\mu_i) = \mu_i^\lambda$ with $\lambda < 1$. Of these two, initial modeling attempts might be conducted with the log link, particularly in light of the evidence of Figure 7.24 that, at least for the treatment basin, this may well be an adequate link function.

Overall, it seems clear that the random components appropriate for the data of Figures 7.21 and 7.22 will differ. Whether this means we must use different distributional forms or only different parameters (e.g., in a gamma or inverse Gaussian random component) remains an open question. It also seems clear that the expectation functions will differ for these two data sets, but again, whether this requires different functional forms or only differences in parameter values remains to be seen.

The point of this example is not that we have, through consideration of models formulated as generalized linear models, "solved" a difficult modeling problem (we don't know that yet). And it is true that, as yet, we have no solid evidence that we can "do better" with this problem than we could with models formulated as either weighted or transformed additive error models. The point

of this example is that we have been able to suggest plausible models that may well be better statistical abstractions of the problem under study than are additive error (and, in particular, linear additive error) models. And, importantly, we have done so without modification of the original scales of measurement for the quantities observed.

We end this introduction to generalized linear models with a very brief indication of the flexibility that has been gained through consideration of models as composed of random and systematic components that have more "equal" stature than in additive error models, in which the distributional attributes of the model play a decidedly subservient role to specification of the expectation function. It was not that long ago (I would say less than 25 years) that nonlinear models were thought applicable almost exclusively in situations for which scientific theory indicated a particular (nonlinear) expectation function, and all that remained was to formulate an appropriate additive error distribution for the observational process. This view of statistical modeling was still fixated on what I have called the "signal plus noise" approach. If "errors" in the original model formulation did not appear to follow an "appropriate" (i.e, location-scale and primarily normal) distribution, the solution was to employ transformations of one type or another to produce such appropriate behavior in the resultant quantities. Certainly, advances have been made to these ideas even within the context of additive errors (e.g., see Sections 7.2.4, 7.2.5), but the idea that, perhaps, one should consider the distribution of responses on a more equal footing with the expectation function has been a major advancement. The demonstration that what are now called generalized linear models could unify this idea for at least a particular class of models (with distributions given by exponential dispersion families) was a huge influence in promoting the general concept of paying attention to distributions as more than mere "error"

models.

It should also be noted that, although the specification of link functions is definitely facilitated with models that contain only one type of covariate (because then you can see the inverse link exhibited in scatterplots), the ideas underlying generalized linear models are not limited to single covariate situations. The book by McCullagh and Nelder (1990) contains a number of examples, and Kaiser and Finger (1997) presents an example that uses not only multiple covariates but also lagged values of those covariates; it is not entirely clear whether the model used in this reference is *really* a generalized linear model or not, but it certainly makes use of the general structure of random and systematic model components.

7.4 Models With Multiple Random Components

The topic of this section is models that contain not a single random component, but several random components. I believe that there are three fundamental schools of thought from which the development of models with more than one stochastic component can be approached.

The first school of thought draws on an extension of additive error modeling, under which multiple sources of variation function at different levels of a data structure. This leads, for example, to the types of random effects and mixed models you have seen in Statistics 500 and 511 with linear systematic model components. Usually, the division of errors into "variance components" under this approach are determined from the observational structure of the study (e.g., the entire book by Bryk and Raudenbush (1992) is founded on

this idea). Estimation and inference can proceed in a manner that attempts to maintain as much of the exact theory (i.e., derivable sampling distributions for estimators under all sample sizes) of ordinary least squares as possible, although this doesn't really work; consider, for example "approximate $t-$intervals", which is a complete oxymoron, that result from the Cochran-Satterhwaite approximation. I say that approximate $t-$intervals is an oxymoron because $t-$distributions depend exactly on the sample size n, while approximations are inherently asymptotic in nature. I'm not claiming that the Cochran-Satterhwaite approximation is not a useful statistical result. It is. But to believe that we can use it to hang on to Gauss-Markov-like properties for estimators is simple self-delusion.

The second school of thought for formulation of models with multiple stochastic elements rests on the concept of "random parameter" specifications for models of random variables associated with observable quantities. This leads to multi-level or what are called *hierarchical* models. As we will see, there is nothing inherently Bayesian about such model formulations; it is true that Bayesian analysis is quite natural for many such models, but analysis of a model is a separate issue from formulation of a model. In the case of linear models it turns out that mixed model formulations are usually nothing more than special cases of random parameter formulations. Except in situations for which the investigator is able to exercise a great deal of control over factors that may influence a response of interest, this second approach has more scientific appeal than does the "extension of additive error" approach briefly described above. We will illustrate this in the following subsections.

The third school of thought is not so much a different view of the same situation as the first two (i.e., a true school of thought), but is more of a "catch-all" for situations that cannot be addressed by those approaches di-

rectly. We might call this the "latent variable" approach. There are few, if any, conventions for how one might add to a model random variables that represent unobservable phenomena (which is the latent variable idea). We will briefly discuss this approach after covering the first two.

7.4.1 Mixed Models

The entire presentation of Chapter 7.2 on additive error models was designed to communicate the fact that, for models with additive error terms, nonlinear models are not an *extension* of linear models but, rather, linear models are *special cases* of general additive error models. Linear models are of particular interest because of the results they allow for estimation and inference procedures. But as *models* they form only a class of restricted expectation functions for additive error formulations. My view on what are called *mixed models* is different. Mixed models *are* fundamentally an extension of linear models. In this subsection, then, we will discuss mostly linear mixed models. Toward the end, I will provide a few comments indicating what I view as the extreme limitations of trying to use the same formulation with nonlinear models.

You have already had an introduction to linear mixed models in Statistics 500 and 511, primarily from the viewpoints of estimation (e.g., REML) and structures for controlled experimental studies (e.g., balanced nested designs). We will not dwell on these same topics here, but will instead focus on the underlying scientific mechanisms or phenomenon being represented in these models; see Chapter 5.3. It is generally useful in this context to distinguish between what are often called *random effects* models and *random coefficient* models (e.g., Longford, 1993).

Linear Random Effects Models

Random effects models refer to situations in which responses of interest are subject to a number of identifiable sources of variability, *and in which we have information that will allow us to estimate the magnitudes of those sources as distinct from one another*. The latter portion of the preceding sentence is italicized because it is fundamental to the statistical conceptualization of a "source of variability". In practical terms, what this means is that we must have *replicate observations* of a source of variability in order to identify it as such a source in the first place.

Example 7.7

This example was adapted from one given in Longford (1992). Consider the formulation of a linear model for the relation between log liabilities and log assets for companies of various types in various parts of the world. The basic idea is that, if Y_i represents the log liabilities of a company and x_i the log assets, then we might consider the simple linear regression model,

$$Y_i = \beta_0 + \beta_1 x_i + \sigma\, \epsilon_i.$$

This model implies that, if $\beta_1 = 1$, the logarithm of the ratio of liabilities to assets, namely $\log(\text{liabilities})/\log(\text{assets})$ is constant, and the position of a company in the distribution of this quantity might be used by financial institutions in making loan decisions (e.g., a company in the far right tail might not be a good "bet"). In fitting this model to observed data, it is presumed that the estimated value of β_1 will be near 1 or perhaps just slightly below, at least for companies that survive. Primary interest then centers on the constant β_0, which then represents the mean of the log ratio of liabilities

to assets; the only reason for writing the model with liabilities as a function of assets is to ensure that an estimate of β_1 is, in fact, near a value of 1. Otherwise, we could write, with $Y_i = \log(\text{liabilities})/\log(\text{assets})$,

$$Y_i = \mu + \sigma\,\epsilon_i, \tag{7.22}$$

Now, it is generally felt (i.e., known or accepted) among financial analysts that the sector of the economy a company operates in may influence the value of μ; sectors include categories such as construction, food, transportation, conglomerates, clothing, and so forth. It is also believed that the area of the world in which a company operates is an important factor; Global, North America, Central America, South America, Southeast Asia, Central Asia, Western Europe, Eastern Europe, Southern Africa, and so forth are possibilities. It would certainly make sense to model these additional sources of variability, which we could do by defining $Y_{i,j,k}$ as the logarithm of liabilities to asset ratio associated with company i in economic sector j in world region k, and specifying a random effects model of the type

$$Y_{i,j,k} = \mu + \delta_j + \lambda_k + \epsilon_{i,j,k}, \tag{7.23}$$

where

$$\begin{aligned} \delta_j &\sim iid\, N(0,\, \tau^2) \\ \lambda_k &\sim iid\, N(0,\, \psi^2) \\ \epsilon_{i,j,k} &\sim iid\, N(0,\, \sigma^2) \end{aligned}$$

But suppose that the only data available are a hodgepodge of information from companies in different economic sectors from different regions around the world, with few, if any, replicate observations of companies in the same economic sector and region. Or, data were available for only one economic

sector in all the regions. Or, data were available for most economic sectors in all regions but for only one economic sector in one region. The result is that we would be unable to distinguish model (7.23) from model (7.22).

The point of Example 7.7 is not that random effects models are ineffectual, but that there is an intimate connection between useful model formulations and the data collection scheme. In fact, when one has control over the observational structure, random effects models are both powerful and elegant. The connection between linear random effects models and situations that allow carefully designed data collection plans is indicated by the fact that much (if not, in fact, most) early development of these models was motivated by the analysis of problems stemming from experiments in agricultural and animal breeding. The connection between control over observational structure and the appropriateness of linear models will also appear in consideration of linear random coefficient models.

A major hurdle to understanding the structure of linear random effects (and, later mixed effects) models is the variety of notations that can and have been used to represent such models. Consider, first, the representation of a simple linear regression model with $\{Y_i : i = 1, \ldots, n\}$ as response variables, $\{x_i : i = 1, \ldots, n\}$ as covariates, and $\{\epsilon_i : i = 1, \ldots, n\}$ as iid $N(0, 1)$ errors. This model can be written in various forms as,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \, x_i + \sigma \, \epsilon_i, \\ Y_i &= \boldsymbol{x}_i^T \, \boldsymbol{\beta} + \sigma \, \epsilon_i; \quad \boldsymbol{x}_i^T = (1, x_i), \\ \boldsymbol{Y} &= \boldsymbol{X} \, \boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ E(\boldsymbol{Y}) &= \boldsymbol{X} \, \boldsymbol{\beta}; \quad var(\boldsymbol{Y}) = \sigma^2 \, I_n, \end{aligned}$$

where, in the last expression, I_n stands for the identity matrix of dimension $n \times n$.

Now, these various model forms typically cause us no difficulties, as we are able to move from one to the other quite easily. The situation can become less clear rapidly, however, when one begins adding in additional random variables on the right-hand-side (rhs) of these expressions. For example, a random effects model for responses that are observed in "clusters" or "groups" could be written in at least the following ways.

$$\begin{aligned} Y_{i,j} &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \delta_j + \sigma \epsilon_{i,j}, \\ Y_i &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{\gamma} + \sigma \boldsymbol{\epsilon}_i \\ \boldsymbol{Y} &= \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{Z} \boldsymbol{\gamma} + \sigma \boldsymbol{\epsilon} \\ E(\boldsymbol{Y}) &= \boldsymbol{X} \boldsymbol{\beta}; \quad var(\boldsymbol{Y}) = V + \sigma^2 I_n \end{aligned}$$

It is much less clear than in the simple linear regression case, how to consider indexes, dimension of matrices, and whether those matrices have unique only or repeated rows and/or columns, in order to make these formulations equivalent. Now, imagine what is possible if one adds more than one random variable to the rhs of such linear models, variables that might correspond to clusters or groups of varying sizes or nestings. Consider what might be possible if one allows nonlinear models into the picture. Even reasonably competent statisticians (if I may flatter myself for the moment) can become easily confused. Is there a way to "unwind" the potential confusion caused by various notations, or are we doomed to have to learn multiple equivalent forms for which the equivalence is not immediately obvious?

One possibility is to draw on the previously expressed reluctance to use multiple subscripting in cases that would require a large number of subscripts, and recall that statistical operators such as expectation are inherently univariate in nature. While this technique seems, at first, overly cumbersome, it can help us keep track of "what is truly going on" in a model. Under this

convention, the simple linear model is perhaps best presented in one of the first two forms given (the use of $\boldsymbol{x}_i^T \boldsymbol{\beta}$ seems to cause no great difficulties here, since responses are still indexed by a single variable i).

The key notational convention being suggested is that a collection of response variables be indexed individually as, perhaps, $Y_1, Y_2, \ldots, Y_n$, and that auxiliary information is available, such as sets $\mathcal{C}_j$ that contain the elemental indices i belonging to different "clusters" indexed by j. We could then write a linear random effects model as

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \tau \sum_{j=1}^{J} \delta_j \, I(i \in \mathcal{C}_j) + \sigma \, \epsilon_i, \tag{7.24}$$

where, for $j = 1, \ldots, J < n$, $\delta_j \sim iid\, N(0,1)$, and, for $i = 1, \ldots, n$, $\epsilon_i \sim iid\, N(0,1)$ and with independence among all δ_j and ϵ_i. What is accomplished by this "univariate" view of the world? For one thing, it facilitates the derivation of expected values and covariances for the joint distribution of $Y_1, \ldots, Y_n$ on which estimation must be based. That is, given normal distributions for both δ_j and ϵ_i, all i and j, it is a simple matter to determine for model (7.24) that the joint distribution of $Y_1, \ldots, Y_n$ is multivariate normal such that,

$$E(Y_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}; \quad var(Y_i) = \tau^2 + \sigma^2,$$

and,

$$cov(Y_i, Y_k) = \begin{cases} \tau^2 & \text{if } i \in \mathcal{C}_j \text{ and } k \in \mathcal{C}_j, \\ 0 & \text{o.w.} \end{cases}$$

Now, for a situation in which we might only need double subscripting (using i and j, say) this convention presents no huge advance. But consider the following example.

Example 7.8

The classical example of a situation in which a random effects model is a useful conceptualization of the problem is that of sub-sampling and extensions of sub-sampling to contain both nested (sub-sampling structure) and crossed (factorial structure) elements. Consider a study in which the objective is to estimate the magnitude of some soil constituent (e.g., nitrogen content or contamination with polychlorinated biphenyls known as PCBs) in a particular region (e.g., field). A number of soil samples are taken, and each sample is homogenized (i.e., physically mixed). Each homogenized sample is divided into subsamples, and three subsamples of each sample are sent to each of a number of different chemical laboratories. Each laboratory again divides each of its three subsamples into further subsamples sometimes called aliquots; suppose for simplicity that three aliquots of each subsample are used. Each aliquot undergoes a separate extraction procedure in which the compound of interest is "extracted" from the soil, often by running a solvent (such as dimethylsulfoxide, known as DMSO, or some chloric substance such as chloroform) through a column that contains the material and collecting the elutriant (what comes out the bottom). Each processed aliquot is then subject to an appropriate measurement procedure from analytical chemistry (e.g., liquid chromatography).

Using a multiple subscripting scheme, basic random variables would need to be defined as $Y_{i,j,k,s}$ where i indexes sample, j indexes subsample, k indexes laboratory, and s indexes aliquot. If, as indicated in this example, the observational structure was completely balanced, this could certainly be done in matrix notation, but doing so would, in my mind, not be easy. For example, is it clear what $\boldsymbol{Z}$ would look like, and what $\boldsymbol{\gamma}$ would be if we wrote the model

as follows?

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{Z}\,\boldsymbol{\gamma} + \sigma\,\boldsymbol{\epsilon}.$$

Other situations are impossible to write in clean matrix notation, such as illustrated by the following.

Example 7.9

Suppose we have a study that results in longitudinal data, such as the observation of low density lipoproteins (LDL) over time for patients on a given treatment (e.g., the drug with brand name Zocor?) Each patient is to be observed (have LDL measured) every 4 weeks, except when they forget their appointment, or except when their son or daughter has an important school event but they can make it in the next week, or except when they move, or except when they die, or except when too many medical staff are unavailable to take the readings, or . . .

Clearly, a model for this situation will be in the form of random variables for elemental measurements (LDL for a given patient at a given time) with a covariate of time, and clusters of observations taken within patients. Suppose that we assume (not unquestionable, but suppose that we do) that the "trajectories" of LDL over time are the same for each patient and linear (i.e., equal slopes in linear regressions on time). Then we may wish to use a mixed model with fixed regression slope but random effect (which then affects the model intercept for each patient). It is unclear in this setting how to use a double index system (i could be patient, but how would time be indexed by j when the values of j are not the same among patients). It would be difficult to construct a clean matrix representation for a model in this situation(vectors of observa-

tions are of different lengths for different patients). On the other hand, we may define response variables for the total number of observations made (across all patients, which we will denote as n) in the form $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ in which the (non-random) variables $\boldsymbol{s}_i$ are defined as

$$\boldsymbol{s}_i \equiv (m, j),$$

where m denotes patient ($m = 1, \ldots, k$) and j denotes observation number for patient m, with $j = 1, \ldots, n_m$. Note that $n = \sum_m n_m$. Associated with each $Y(\boldsymbol{s}_i)$ is a covariate vector $x(\boldsymbol{s}_i) \equiv (1, t_i)^T$ in which t_i is the time of observation. Then, we can easily write

$$Y(\boldsymbol{s}_i) = x(\boldsymbol{s}_i)^T \boldsymbol{\beta} + \tau\, \delta(\boldsymbol{s}_i) + \sigma\, \epsilon(\boldsymbol{s}_i), \tag{7.25}$$

where $\delta(\boldsymbol{s}_i) \equiv \delta_p$ if $\boldsymbol{s}_i = (p, j)$ for any j and $p = 1, \ldots, k$. In (7.25) we would most likely assume that $\delta_p \sim iid\, N(0, 1)$ for $p = 1, \ldots, k$, that $\epsilon(\boldsymbol{s}_i) \sim iid\, N(0, 1)$ for $i = 1, \ldots, n$, and that there is independence between these terms.

A final issue that should be addressed here is the question of how one decides whether to model an effect as fixed or random. This is, first of all, an old question that eludes a simple answer. A basic idea, given in one form or another in most books on linear models, is whether one "cares" about the particular levels of a given factor. Longford (1993, p.16) casts this in terms of "exchangeability" and whether the inferences we wish to make depend on the particular levels of a factor included in a study. As noted in Longford (1993, p.24), the view presented by Searle (1971) is to consider a study a result of a random process. If another realization of this process would lead to the same levels for a given factor, then that factor should be considered fixed, and random otherwise. As also noted by Longford (1993, p.24) this idea

seems quite applicable for designed experiments, but may not be as effective for observational situations. One point that is often overlooked is that random effects models can be dangerous when there are a small number of levels of the random term; of course, fixed effect models can also be dangerous if there is, in fact, a random term with a small number of levels but this is ignored.

Example 7.10

Consider a situation in which a mixed model with simple linear fixed effect and a clustering or group structure for the random effect is appropriate. We could write a model for this situation as

$$\begin{aligned} Y_{i,j} &= \beta_0 + \beta_1 x_{i,j} + \tau \delta_j + \sigma \epsilon_{i,j}, \\ Y_i &= \beta_0 + \beta_1 x_i + \tau \sum_{j=1}^{J} \delta_j I(Y_i \in \mathcal{C}_j) + \sigma \epsilon_i, \\ Y(\boldsymbol{s}_i) &= x(\boldsymbol{s}_i)^T \boldsymbol{\beta} + \tau \delta(\boldsymbol{s}_i) + \sigma \epsilon(\boldsymbol{s}_i). \end{aligned}$$

Either of the first two formulations may be the easiest in this example, the third is valuable primarily because it generalizes more easily than either of the others. Suppose then that in the first model formulation given above $j = 1, 2, 3$ and $x_{i,j} = 1, 2, \ldots, 25$ for each j, $\delta_j \sim iid\, N(0,1)$, $\epsilon_{i,j} \sim iid\, N(0,1)$, $\tau = 0.5$, $\sigma = 0.5$, $\beta_0 = 1.2$ and $\beta_1 = 0.5$.

From this model, the conditional expected value of $Y_{i,j}$, given the value of δ_j is,

$$E(Y_{i,j}|\delta_j) = (\beta_0 + \tau \delta_j) + \beta_1 x_{i,j},$$

and the conditional variance is,

$$var(Y_{i,j}|\delta_j) = \sigma^2.$$

At the same time, the unconditional (i.e., marginal) expectation and variance of $Y_{i,j}$ are,

$$E(Y_{i,j}) = \beta_0 + \beta_1 x_{i,j}; \quad var(Y_{i,j}) = \tau^2 + \sigma^2.$$

A random realization of this model is presented in Figure 7.26, where group membership is not indicated, and the marginal expectations $\beta_0 + \beta_1 x_{i,j} = 1.2 + 0.5 x_{i,j}$ are given as the solid line.

While at first glance this looks like a fairly "good regression", a closer look will reveal that there are substantially (in fact, twice as many) points above the line as there are below the line. Recall that the line in Figure 7.26 is the "true" line, not a estimated line. Any estimated line would almost certainly be "higher" than (but possibly parallel to) the true line. Recall that this example data set was simulated using 3 values of δ_j, assumed to be normal random variables with mean 0 and variance 0.25. It turned out that the values obtained were $\delta_1 = 1.349$, $\delta_2 = 2.248$ and $\delta_3 = -1.271$. That is, two of the three values of δ_j were greater than 0 and one was less. The same data, but this time identified by "group", along with the conditional means and same overall marginal mean as in Figure 7.26 are given in Figure 7.27. What can be gained from this example?

1. When data are obtained from clusters or groups, it can be misleading to base estimates of marginal structure on a small number of groups. The collection of data from an *odd* or *even* number of groups is of little import; it could easily have occurred that all three values of δ_j were positive (or negative).

2. It may be that, in situations that involve a small number of groups, *conditional* analyses are more meaningful than attempting to model the

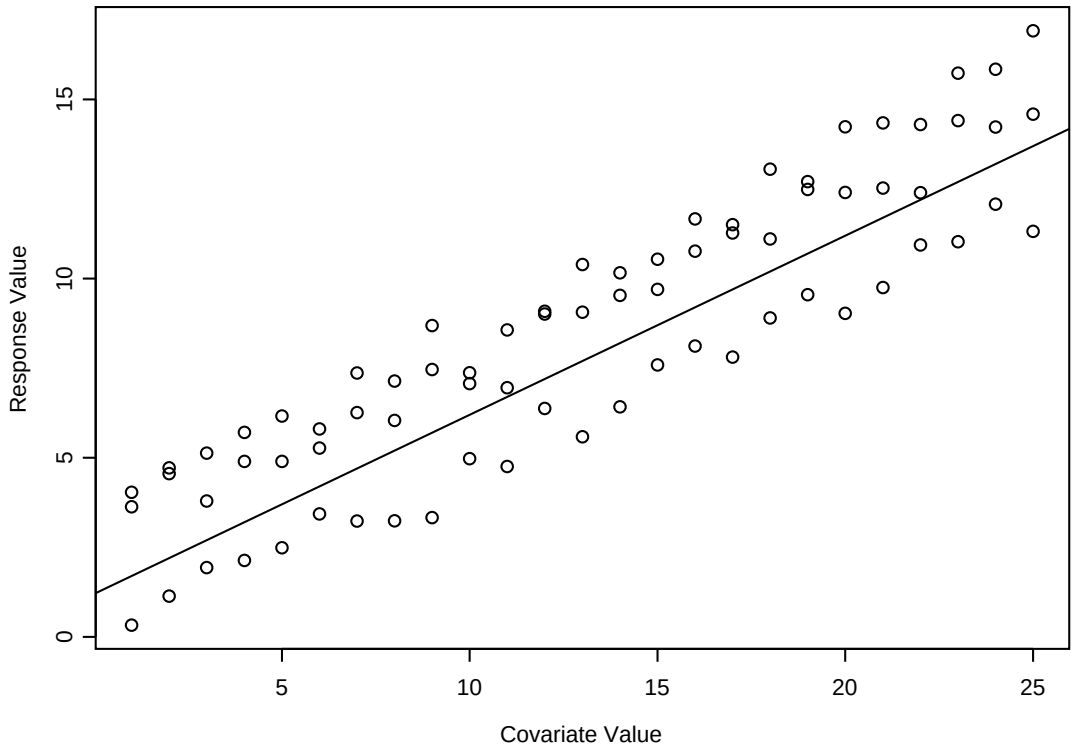

Figure 7.26: Scatterplot of simulated data from a random effects model with three clusters or groups.

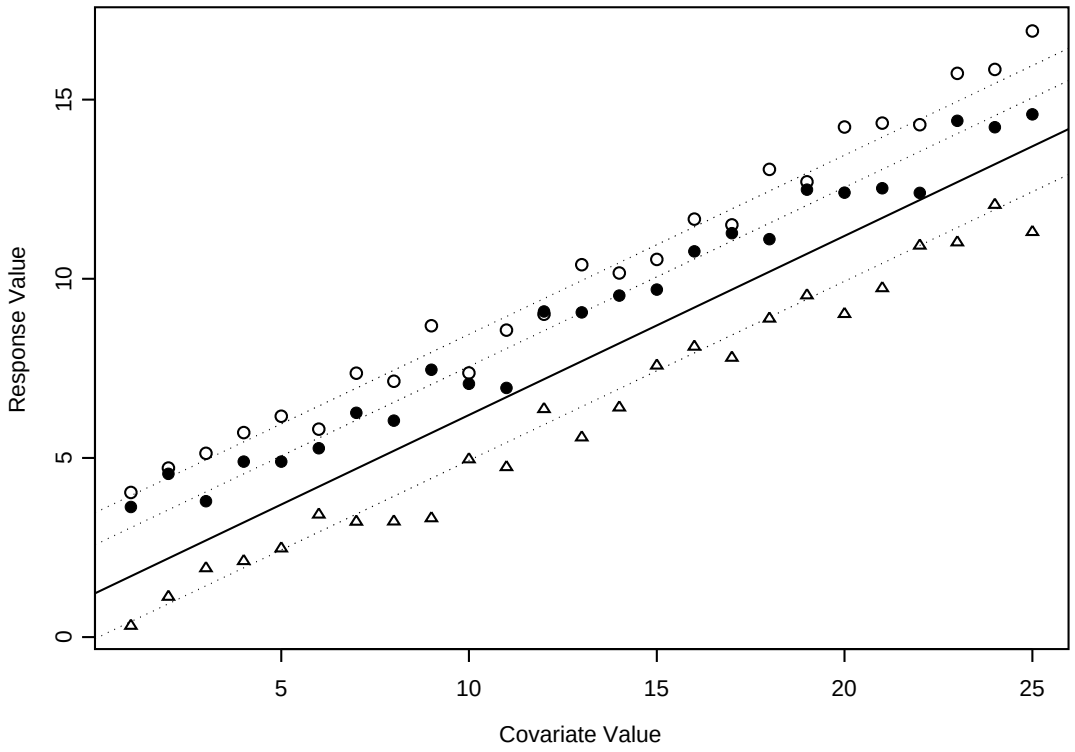

Figure 7.27: Scatterplot of simulated data as in Figure 7.26, but with group identification and conditional means added.

marginal structure. For example, it is clear from Figure 7.27 that conditional regressions fit to each group individually would give reasonably good estimates of the conditional mean structures (relation between responses and covariate). It would be difficult, if not impossible, however, to obtain a "good" estimate of the marginal mean structure from these data (although an unbiased estimate could be obtained).

Linear Random Coefficient Models

In at least Statistics 500 you were introduced to mixed models that included an interaction term between the random and fixed effects, probably along the lines of

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \sigma\,\epsilon_{i,j,k}, \tag{7.26}$$

where $i = 1, \ldots, a$ indexed fixed effects, $j = 1, \ldots, b$ indexed random effects, and the interaction $(\alpha\beta)_{i,j}$ was hence also a random term in the model, usually with an assumption that $(\alpha\beta)_{i,j} \sim iid\, N(0, \sigma_{\alpha\beta})$ or something similar. This formulation draws on the usual interpretation of models for only fixed effects; the interpretation of a finding that $\sigma^2_{\alpha\beta}$ differs from zero implies that fixed effects are not constant across levels of the random effect or vice versa. When the fixed effects in model (7.26) involve one or more covariates that function on a ratio scale, this implies that each group (as defined by the random effect terms β_j) may have its own "regression line". Compare this to the model of Example 7.10 in which each group had its own intercept, but there was common slope across groups (see, e.g., Figure 7.27).

This situation is often modeled by what we are calling *random coefficient* models after Longford (1993). A standard formulation for this is to take groups

indexed by j and assign a linear structure as

$$\boldsymbol{Y}_j = \boldsymbol{X}_j\,\boldsymbol{\beta} + \boldsymbol{Z}_j\,\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_j; \quad j = 1, \ldots, k, \tag{7.27}$$

where it is assumed that the columns of $\boldsymbol{Z}_j$ consist of a subset of the columns of $\boldsymbol{X}_j$. It is often assumed that the first column of $\boldsymbol{X}_j$ is a column of 1s and that this is the same as the first column of $\boldsymbol{Z}_j$. In (7.27) we generally have that $\boldsymbol{\gamma}_j \sim iid\, N(0,\, V)$ for some covariance matrix V, and $\boldsymbol{\epsilon}_j = (\epsilon_{1,j}, \ldots, \epsilon_{n_j,j})^T$ is $N(0,\, \sigma^2 I_{n_j})$.

Notice what is implied by model (7.27), which is that the "fixed" parameters $\boldsymbol{\beta}$ do not depend on the group j, while the "random" parameters $\boldsymbol{\gamma}$ do depend on the group j. It would, of course, be possible to "stack" vectors and matrices in (7.27) to arrive at the general but uninformative version $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ given in Example 7.8 for random effects models.

Example 7.11

Example 7.10 resulted in a model of parallel regression lines that differed in (random) intercept values. We may expand that situation to one that includes random slopes by taking

$$Y_{i,j} = \beta_0 + \beta_1\, x_{i,j} + \tau_1 \delta_j + \tau_2 \gamma_j\, x_{i,j} + \sigma\, \epsilon_{i,j}, \tag{7.28}$$

where $\delta_j \sim iid\, N(0,\, 1)$, $\gamma_j \sim iid\, N(0,\, 1)$, $\epsilon_{i,j} \sim iid\, N(0, 1)$ and all of these variables are independent; note that we could assign (δ_j, γ_j) a joint distribution, but for our purposes at the present independence is sufficient. Model (7.28) describes a situation in which each group (index j) has its own conditional regression line (conditional on the random variables δ_j and γ_j). In fact, the

conditional model takes $Y_{i,j}$ to be normally distributed with,

$$E(Y_{i,j}|\delta_j, \gamma_j) = (\beta_0 + \tau_1\delta_j) + (\beta_1 + \tau_2\gamma_j)\, x_{i,j},$$

and,

$$var(Y_{i,j}|\delta_j, \gamma_j) = \sigma^2.$$

In the conditional model, $Y_{i,j}$ is independent of all other response variables.

The marginal model, on the other hand, gives,

$$E(Y_{i,j}) = \beta_0 + \beta_1\, x_{i,j},$$

$$var(Y_{i,j}) = \tau_1^2 + \tau_2^2\, x_{i,j}^2 + \sigma^2,$$

and,

$$cov(Y_{i,j}, Y_{k,j}) = \tau_1^2 + \tau_2^2 x_{i,j} x_{k,j},$$

for $i, k = 1, \ldots, n_j$, $j = 1, \ldots, k$. All other covariances are zero.

Notice immediately that the marginal model indicates the variances of responses becomes large as a function of the magnitude (or absolute value) of the covariates $x_{i,j}$. This is not quite so obvious if, for example, we write the model in the (semi) matrix form of expression (7.27), although that expression is certainly legitimate. Consider again a situation in which $j = 1, 2, 3$ and $x_{i,j} = 1, 2, \ldots, 25$ for each j.

Suppose now that model (7.28) applies with $\sigma^2 = 4$, $\tau_1^2 = 0.25$ (same as τ^2 in Example 7.10) and $\tau_2^2 = 0.25$ as well. Using the same values of δ_j; $j = 1, 2, 3$ as in Example 7.10, but generating independent values for γ_j; $j = 1, 2, 3$ and $\epsilon_{i,j}$; $i = 1, \ldots, 25$; $j = 1, 2, 3$ produces an overall scatterplot presented in Figure 7.28. These data used the same values δ_j; $j = 1, 2, 3$ as Example 7.10. The values of γ_j used were $\gamma_1 = -0.181$, $\gamma_2 = 0.311$ and $\gamma_3 = -0.420$. The scatterplot of Figure 7.29 shows the data with conditional regression lines

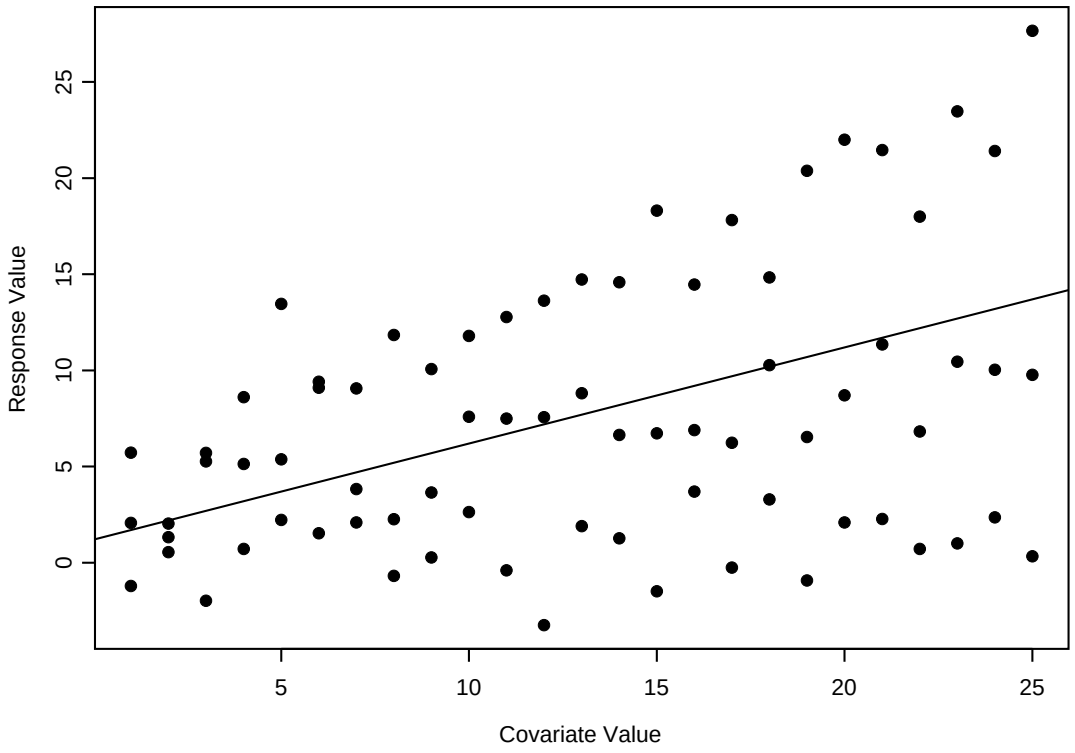

Figure 7.28: Scatterplot of simulated data from a random coefficient model with three clusters or groups.

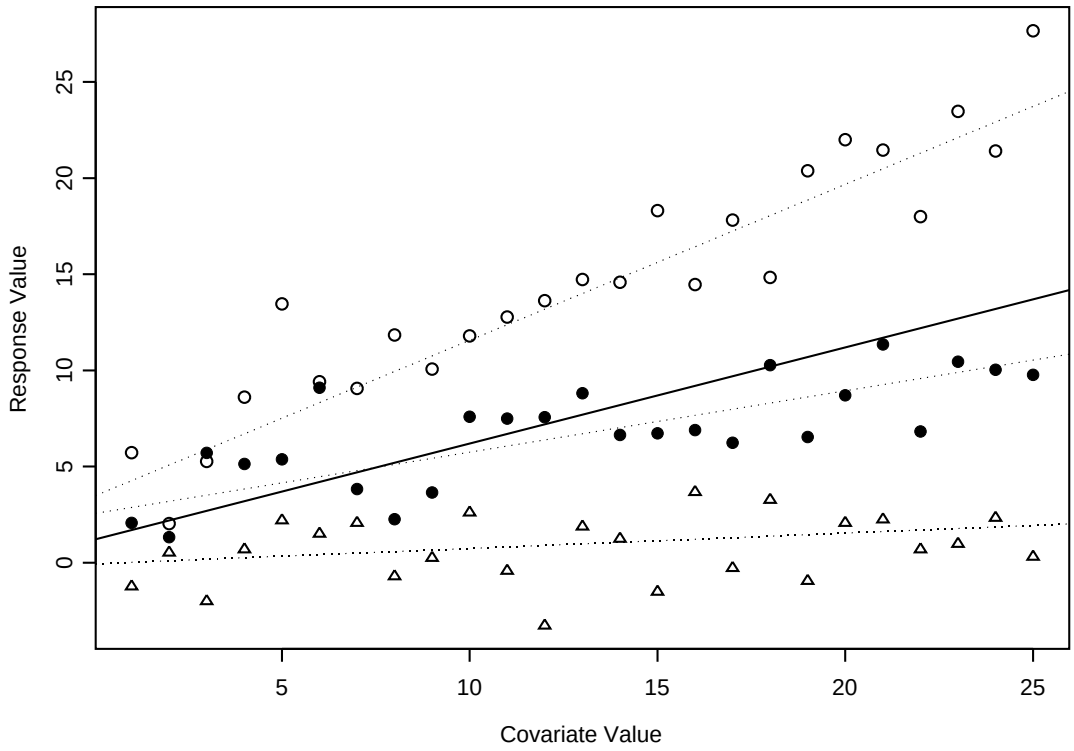

Figure 7.29: Scatterplot of simulated data as in Figure 7.28, but with conditional regressions added.

(dashed lines), as well as the marginal expectation function (solid line). Notice that, in example 7.10 the variance of the random terms was similar to that of the independent error terms ($\tau^2 = 0.25$ and $\sigma^2 = 0.25$). In example 7.11, however, the variance of, in particular, the random term for slopes (γ_j in model (7.28)) was much smaller than the error variance ($\tau_2 = 0.25$ while $\sigma^2 = 4$). These differences were used to produce "nice" pictures, meaning simulated data that could be displayed on one page. This is, however, the same pattern we would expect in situations for which linear random effects or linear random coefficient models are appropriate; note that the relative magnitudes of error and random effect terms for "nice" data realizations is also influenced by the range of covariate values.

We will seize the opportunity to use this example for reinforcement of a point made awhile back that appropriate random components (e.g., symmetric errors) cannot be determined by examination of histograms that represent marginal distributions. For the data of Example 7.11, a histogram of all the responses in presented Figure 7.30. This graph clearly shows a distribution with long right tail. Yet this is no indication at all that normal error terms are inappropriate in models for these data. All of the random terms used to generate these data were, in fact, simulated from normal distributions.

Nonlinear Mixed Effects Models

We will give only a brief overview to efforts at extending the linear mixed model strategy to formulations involving nonlinear response functions, but will revisit the issue in a latter section in a discussion of mixed versus mixture modeling strategies. There have been two primary vehicles used to extend the mixed modeling idea to nonlinear models based, naturally enough, on nonlinear

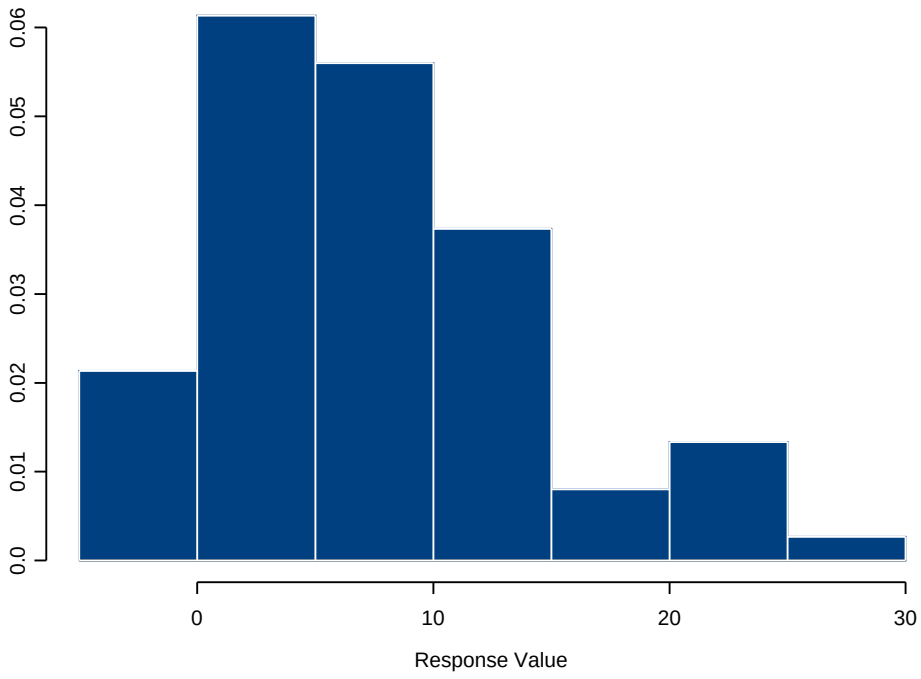

Figure 7.30: Histogram of response variables for Example 7.11.

additive error models on the one hand, and generalized linear models on the other.

In the presentation of this subsection we will use the double subscripting notation $Y_{i,j}$ to facilitate comparison of these notes with most of the published references cited. This is the context of clusters or groups of random variables in which $Y_{i,j}$ denotes the i^{th} random variable for the j^{th} cluster. The classic example is that of a longitudinal study on $j = 1, \ldots, k$ individuals with $i = 1, \ldots, n_j$ observations for the j^{th} individual. Be alert in reading the literature, however, as some authors use other notations; see, for example Breslow and Clayton (1993) for notation more similar to my inclination for modeling response variables indexed sequentially with $i = 1, \ldots, n$, while Lindstrom and Bates (1990) reverse the use of i and j subscripts.

What are typically called *generalized linear mixed models* have the following form. Let $\{Y_{i,j}i : i = 1, \ldots, n_j;\ \ j = 1, \ldots, k\}$ denote independent response variables with exponential dispersion family distributions of the form of expression (7.19),

$$f(y_{i,j}|\theta_{i,j}, \phi) = \exp\left[\phi\{y_{i,j}\theta_{i,j} - b(\theta_{i,j})\} + c(y_{i,j}, \phi)\right],$$

from which we have $\mu_{i,j} \equiv E(Y_{i,j}) = b'(\theta_{i,j})$ as before. A basic generalized linear model is completed by taking a known smooth link function $g(\cdot)$ to be such that $g(\mu_{i,j}) = \eta_{i,j}$, where $\eta_{i,j}$ is given as the linear predictor in expression (7.20), $\eta_{i,j} = \boldsymbol{x}_{i,j}^T\beta$. To formulate a generalized linear mixed model, we simply extend the linear predictor in the same manner as the linear response function in expression (7.27),

$$\eta_{i,j} = \boldsymbol{x}_{i,j}^T\,\beta + \boldsymbol{z}_{i,j}^T\,\gamma_j,$$

which then gives the systematic component of the mixed model as

$$g(\mu_{i,j}) = \eta_{i,j} = \boldsymbol{x}_{i,j}^T\,\beta + \boldsymbol{z}_{i,j}^T\,\gamma_j, \tag{7.29}$$

for $\gamma_j \sim F_\gamma$ with $E(\gamma_j) = 0$. Almost always, F_γ is taken to be multivariate normal with a covariance matrix of known parameterized form, $\Gamma(\psi)$ say. For examples of this type of model see Zeger, Liang, and Albert (1988), Zeger and Karim (1991), Breslow and Clayton (1993), McGilchrist (1994), and McCulloch (1997).

The other form of nonlinear mixed models is to take the basic form of a nonlinear additive error model, given in expression (7.1) as, for $i = 1, \ldots, n_j$, $j = 1, \ldots, k$,

$$Y_{i,j} = g(\boldsymbol{x}_{i,j}, \beta) + \sigma\, \epsilon_{i,j},$$

where, usually, $\epsilon_{i,j} \sim iid\, N(0, 1)$. This model is extended in a straightforward manner to have the mixed effects formulation as,

$$Y_{i,j} = g(\boldsymbol{x}_{i,j}, \lambda_j) + \sigma\, \epsilon_{i,j}, \tag{7.30}$$

where

$$\lambda_j = \beta + \gamma_j,$$

where, as for generalized linear mixed models, $\gamma_j \sim F_\gamma$ with $E(\gamma_j) = 0$ and, almost always, F_γ is taken to be multivariate normal with a covariance matrix $\Gamma(\psi)$. Note that Lindstrom and Bates, for example, use a more complex version in which $\lambda_j = A_j\, \beta + B_j\, \gamma_j$, where A_j and B_j are "design matrices" for parameters. I have kept the specification of (40) simple, which corresponds to $A_j = I$ and $B_j = I$, the identity matrix (this is the most common situation).

The vast majority of attention in the literature has centered on estimation of parameters in these models (e.g., Zeger, Liang, and Albert, 1988; Schall, 1991; Zeger and Karim, 1991; Breslow and Clayton, 1993; McGilchrist, 1994; Kuk, 1995; McCulloch, 1997). While this is not inappropriate, given the difficulties involved (we hope to touch on this in the section on estimation methods), it is true that much less attention has been paid to the modeling aspects

of how such nonlinear mixed structures conceptualize scientific mechanisms or phenomena of interest (i.e., Section 5.3 again). Without going into great detail at this point we will indicate the basic difficulty in interpretation.

Consider first a linear mixed model of the form

$$Y_{i,j} = \boldsymbol{x}_{i,j}^T \beta + \boldsymbol{z}_{i,j}^T \gamma_j + \sigma \epsilon_{i,j}, \tag{7.31}$$

with $\epsilon_{i,j} \sim iid\, N(0,1)$ and $\gamma_j \sim indep\, N(0, \Gamma(\psi))$. Here, in a manner similar to what we saw in Example 7.11, conditional and marginal systematic model components (expectation functions) are,

$$E(Y_{i,j}|\gamma_j) = \boldsymbol{x}_{i,j}\beta + \boldsymbol{z}_{i,j}^T \gamma_j,$$

$$E(Y_{i,j}) = \boldsymbol{x}_{i,j}^T \beta.$$

That is, the fixed parameters β have exactly the same effect on responses in a marginal model as they do in a conditional model. Thus, if we estimated the model from a set of longitudinal observations on individuals $\{Y_{i,j} : i = 1, \ldots, n_j;\ j = 1, \ldots, k\}$ using model (7.31), or estimated the model from a set of independent observations $\{Y_{i,j} : i \equiv 1;\ j = 1, \ldots, k\}$, we would be estimating the same β. The difference, relative to the fixed parameter β is only in the marginal covariance structure that results from considering "commonalities" among groups of responses in (7.31). This is exemplified in the comment of Lindstrom and Bates (1990, p. 686) that "In fact, the SS linear mixed effects model [our conditional model] can be viewed as just one way to generate a parameterization for the marginal covariance in a PA model [our marginal model]."; the acronyms SS and PA stand for "subject specific" and "population average" models and were used initially by Zeger, Liang and Albert (1988). The implication is that it is parameters in the marginal model that are truly

of interest, and consideration of the conditional model primarily provides information by which can make estimation of those parameters more efficient. This idea is common in the literature on nonlinear mixed effects models which has, in large part, centered on population-level models in biostatistics.

Now consider a nonlinear mixed model of the form of expression (40), which can also be written as,

$$Y_{i,j} = g(\boldsymbol{x}_{i,j},\, \beta + \gamma_j) + \sigma\, \epsilon_{i,j}. \tag{7.32}$$

Here, β is still the fixed parameter. The expected values for conditional and marginal models under the formulation (7.32) are,

$$E(Y_{i,j}|\gamma_j) = g(\boldsymbol{x}_{i,j},\, \beta, +\gamma_j),$$

$$\begin{aligned} E(Y_{i,j}) &= E\left\{g(\boldsymbol{x}_{i,j},\, \beta, +\gamma_j)\right\} \\ &= \int g(\boldsymbol{x}_{i,j},\, \beta, +\gamma_j)\, d\, F_\gamma(\gamma_j). \end{aligned}$$

If $g(\cdot)$ is a nonlinear function, it will *not* be true that $E(Y_{i,j}) = g(\boldsymbol{x}_{i,j},\, \beta)$ even though we do have $E(\gamma_j) = 0$; this is a consequence of Jensen's Inequality. So, the basic difficulty with nonlinear mixed models (the same phenomenon is true for generalized linear mixed models) is what interpretation should be attached to the fixed parameter. Clearly, the parameter beta in a model with $E(Y_{i,j}) = g(\boldsymbol{x}_i,\, \beta)$ is not the same as the β in model (7.32); there is an argument that it can be "approximately" the same under certain conditions (see Section 7.4.2).

In the introduction to Chapter 7.4 I offered some relatively harsh words regarding nonlinear mixed models. This is not a consequence that they suffer from the fact that expectations of nonlinear functions of random variables are not equal to those same functions of the expectations (this is simply Jensen's Inequality). Models formulated in the next subsection as hierarchical models

will not be different in this regard. A stumbling block for writing nonlinear models in a mixed model form occurs if one attempts to write random parameters as $\theta = \theta_{\text{fixed}} + \theta_{\text{random}}$, but yet to interpret θ_{fixed} as something meaningful other than simply the expected value of θ. Why has there been so much effort devoted to writing models in this way? I have no convincing answer. In the next subsection we will consider an alternative approach to the formulation of models with multiple stochastic components that seems to have a more pleasing scientific context in general than do mixed models, and focuses to a much greater extent on the conditional model form as what is meaningful.

7.4.2 Models With Parameter Hierarchies

Consider a set of random variables $\{Y_i : i = 1, \ldots, n\}$ that are associated with observable quantities. We will call these "observable random variables", even though we know that "observable" is not quite an accurate adjective to apply to the concept of random variables (see Chapter 5.1). A basic statistical model for such random variables represents a conceptualization of a scientific mechanism or phenomenon of interest, through some (possibly function of) the parameter vector θ. Probability distributions assigned to the random variables Y_i represent the variability or uncertainty associated with observable quantities, given our conceptualization of the underlying mechanism through θ. The specification of a model such as $f(y_i|\theta)$ thus is a model for an observable process, and we will call it the *observation process* or *data* model.

Now, consider the collection of observations corresponding to the conceptualization $\{Y_i : i = 1, \ldots, n\}$ with distributions $f(y_i|\theta);\ i = 1, \ldots, n$ under a number of different sets of circumstances. The setting of Example 7.6 can

be used to illustrate this. Suppose, for example, that some generalized linear model (e.g., inverse Gaussian random component with log link) was deemed appropriate to represent the relation between Cd concentration and length in Yellow Perch from the reference basin in Little Rock Lake (see Figure 7.21). Would we expect this same regression equation (that is, with the same parameter values) to also describe the relation of Cd concentration to length in Yellow Perch from Lake Puckaway (a lake somewhat larger than Little Rock Lake, and located in south-central rather than north-central Wisconsin)? Most likely we would not be that naive. We might believe, hope, or wish to investigate whether the same model *structure* (i.e., inverse Gaussian random component with log link) is adequate in both situations, but it would be unrealistic to assume that the same parameter values would apply. In effect, given that that the relation between responses (Cd concentration) and covariate (length) does reflect a meaningful mechanism (bioaccumulation of Cd), the two lakes, Little Rock in the north and Puckaway in the south, represent two different *manifestations* of that mechanism. If our model form is adequate to describe the mechanism over a range of situations, differences in the parameter values reflect the variability in the way the mechanism is manifested under different circumstances.

Now suppose that we were able to obtain observations from a variety of particular manifestations of the mechanism in, for example, k different lakes (a random sample of lakes would be good here). Then, we might take each lake as having its own regression, and model the parameters of those regressions as coming from some distribution. The mechanism we are trying to model is then embodied in the distribution of parameters, not necessarily a marginal model. It will likely be true that we need the joint marginal distribution of all our observable random variables in order to estimate that distribution, but

the form of the marginal model itself may be of little concern otherwise.

Basic Mixture Models

We will first consider situations that involve groups of independent response variables, for which each variable has its own distribution. The simplest of these are the *beta-binomial* and *gamma-Poisson* mixture models. Such models are useful in comparison of groups. These models have sometimes been consider ways to cope with what is called *overdispersion*, but we present them here in the context of the introductory comments to this section. In addition, we will describe these models as they would be formulated for a single group; this is adequate for the comparison of groups as will become clear when we discuss estimation and inference.

Consider a set response variables $\{Y_i : i = 1, \ldots, n\}$, assumed to be independent given a corresponding set of parameters $\{\theta_i : i = 1, \ldots, n\}$. Let the density or mass functions of the Y_i be denoted as $\{f(y_i|\theta_i) : i = 1, \ldots, n\}$. This set of distributions then constitutes the data model or observation process. Now, let the parameters θ_i be *iid* random variables following a common density or mass function $g(\theta_i|\lambda)$. This is then the *random parameter* model or what we will call the *mixing* distribution. Following the argument in the introductory comments, the scientific mechanism or phenomenon of interest is now conceptualized through the parameter λ, or some function of λ. We can write the joint data model as

$$f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\theta_i),$$

and the joint random parameter model as,

$$g(\theta_1, \ldots, \theta_n|\lambda) = g(\boldsymbol{\theta}|\lambda) = \prod_{i=1}^{n} g(\theta_i|\lambda).$$

The joint "marginal" distribution of the response variables is then derived as,

$$h(\boldsymbol{y}|\lambda) = \int \ldots, \int f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\lambda)\, d\theta_1, \ldots, d\theta_n.$$

Now, because of independence throughout this model formulation, it is generally simpler to derive $h(\boldsymbol{y}|\lambda)$ as

$$h(\boldsymbol{y}|\lambda) = \prod_{i=1}^{n} h(y_i|\lambda),$$

where,

$$h(y_i|\lambda) = \int f(y_i|\theta_i)\, g(\theta_i|\lambda)\, d\theta_i. \tag{7.33}$$

In the above general notation, we use the following nomenclature:

- $f(y_i|\theta)$ is the data model for Y_i
- $g(\theta|\lambda)$ is the *mixing distribution*
- $h(y_i|\lambda)$ is the resultant *mixture* of f over g
- $\log\{h(\boldsymbol{y}|\lambda) = \sum_i \log\{h(y_i|\lambda)\}$ is the marginal or mixture log likelihood, a figures prominently in most useful methods for estimation of λ (i.e., maximum likelihood or Bayesian analysis)

Example 7.12

The Central Valley of California is a large agricultural region, but largely because of extensive irrigation. The Central Valley was originally quite arid, and the underlying geology is that of an ancient sea bed. It is also a historical stopping grounds for migratory waterfowl in what is called the "Pacific Flyway", essentially a broad corridor for waterfowl that breed in the north (e.g., Alaska and British Columbia) but winter in Mexico and Central America. When one

irrigates an area heavily, over a period of years the water table rises. If that area was formed on the sedimentary material of an ancient sea bed, the underlying bedrock contains a large amount of minerals and salts, which become dissolved as excess irrigation water percolates down through the soil. When the water table rises to the level of the root zone of plants, the salinity kills the plants. The engineering solution to this problem is to tile agricultural fields, and drain excess irrigation water from the ground. There of course needs to be a depository for this "irrigation return flow" which, in the Central Valley was accomplished by construction of a great number of "evaporation ponds". The original thought was that such ponds would also be ideal habitat for migrating waterfowl as they moved through the area, as well as holding the potential for benefits from human recreation. But, when salt and mineral-laden irrigation return water evaporates it leaves behind much of the salt and mineral burden, which can become toxic in high concentrations. When the evaporation ponds in the Central Valley began to yield deformed frogs (e.g., six legs, two heads but no legs) and other aquatic life, concern was raised for both the health of the ecosystem and potential implications for humans using the ponds for fishing, boating, and other recreational activities.

To make a long story somewhat shorter, attention eventually focused on Selenium (Se), a necessary trace element for life to exist, but teratogenic in high concentrations. A contentious issue, however, was whether Se was in fact causing problems "in the real world", or whether it could only be shown to have an effect in controlled laboratory studies using unrealistically high exposures. A large number of field studies of the Central Valley region ensued. One such study was conducted by the U.S. Fish and Wildlife Service to examine the potential teratogenic effect of irrigation return water to aquatic life by looking at reproductive success in Mosquitofish *Gambusia spp.*. *Gambusia* are a small

fish that form the basis of many aquatic food chains in this region, and they are also one of the few fish taxa that are viviparous (give live birth).

Now, for irrigation return water to be delivered to evaporation ponds requires the construction of what are called "irrigation return flow canals". One of the larger of these in the Central Valley is called the *San Luis Drain*. In 1983 a large fish kill was observed in the San Luis Drain and Se levels at that time were extremely high. From 1983 to 1985, *Gambusia* was the only fish observed in the San Luis Drain, although previously the canal had supported populations of largemouth bass, striped bass, catfish, and bluegill, among others. A nearby area, the *Volta National Wildlife Refuge* receives no irrigation return water, and did not experience a similar fish kill.

In June, 1985, gravid female *Gambusia* were collected from both the San Luis Drain and the Volta NWR. Fish of similar length, weight, and stage of pregnancy were held in the laboratory until parturition and the number of live and stillborn young counted for each female. Now, the problem was, of course, not so simple. One concern was that fish collected from one type of water but held in clean laboratory water could undergo shock, thus affecting their reproductive success. This is particularly true if the salinity level of the collection water and holding water differs. The average salinity of water in the San Luis Drain (SLD) around the time of the study was judged to be about 16 ppm, while that of the Volta area was 10 ppm. As a result, laboratory water was clean of all ionic content ($R - 0$, pronounced "R-oh") and then reconstituted to 10 and 16 ppm salinities ($R - 10$ and $R - 16$). There were actually four treatment groups used in the study. SLD fish held in $R - 10$ and $R - 16$ water, and Volta fish held in $R - 10$ and $R - 16$ water. The data used here consisted of observations of the total number of young for each female and the number of young born live. Thus, high proportions are an indication

of good reproductive success while low proportions are an indication of poor reproductive success.

We will hopefully have access to the data in lab, but for now suffice it that the observed proportions indicate the presence of "overdispersion", that is, more variability among females than if all individuals within a treatment group were conceptualized as generating identical binomial outcomes (we can find a test for this, e.g., in Snedecor and Cochran, 1967). As a result, a beta-binomial model was fit to each group.

For one treatment group (e.g., SLD with R-16 holding water) let $Y_i;\ i = 1, \ldots, m$ be random variables associated with the number of live young produced by female i. Let $n_i;\ i = 1, \ldots, m$ be the total number of young for female i; we will consider the n_i fixed constants, although it would certainly be reasonable to also model them as random variables in a more complex structure. Given parameters $\theta_i;\ \imath = 1, \ldots, m$, assume that the Y_i are conditionally independent with probability mass functions

$$f_i(y_i|\theta_i) \propto \theta_i^{y_i} \, (1-\theta_i)^{n_i - y_i}, \tag{7.34}$$

for $y_i = 0, 1, \ldots, n_i$ and where $0 < \theta_i < 1$. Further, assume that $\theta_i;\ i = 1, \ldots, m$ are *iid* with probability density functions,

$$g(\theta_i|\alpha,\, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \, \theta_i^{\alpha-1} \, (1-\theta_i)^{\beta-1}, \tag{7.35}$$

for $0 < \theta_i < 1$ and where $0 < \alpha$ and $0 < \beta$.

Combining the data model (7.34) and the mixture (or random parameter model) (7.35) yields the marginal pmf,

$$h(y_i|\alpha,\beta) \;\; \propto \;\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^1 \theta_i^{\alpha+y_i-1} \, (1-\theta_i)^{\beta+n_i-y_i-1} \, d\,\theta_i$$

$$
\begin{aligned}
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\;\frac{\Gamma(\alpha+y_i)\,\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)} \\
&= \frac{\prod_{j=0}^{y_i-1}(\alpha+j)\prod_{j=0}^{n_i-y_i-1}(\beta+j)}{\prod_{j=0}^{n_i-1}(\alpha+\beta+j)} \qquad (7.36)
\end{aligned}
$$

Comment

It is important here to keep track of the sets of possible values for all of the various quantities involved in these derivations. We have,

1. In $f(y_i|\theta_i)$, $y_i \in \Omega_Y = \{0, 1, \ldots, n_i\}$ and $\theta_i \in \Theta = (0,\, 1)$.
2. In $g(\theta_i|\alpha, \beta)$, $\theta_i \in \Theta = (0,\, 1)$ and $\alpha > 0$, $\beta > 0$.
3. In $h(y_i|\alpha, \beta)$, $y_i \in \Omega_Y = \{0, 1, \ldots, n_i\}$ and $\alpha > 0$, $\beta > 0$.

It is crucial that these sets of possible values (for y_i, θ_i, α, β) all match throughout the progression. Thus, the function $h(\cdot)$ is a probability mass function for the discrete random variable Y_i, and the derivation of $h(y_i|\alpha, \beta)$ has not changed the set of possible values from $f(y_i|\theta_i)$. If it had, our model would not make sense.

Now, in any estimation method we use, the log likelihood formed from the pmfs in (7.36) will be important (e.g., method of moments, maximum likelihood or Bayesian estimation). Using independence (Y_is conditionally independent given the θ_is, and the θ_is *iid* implies that marginally the Y_is are *iid*) we have that the log likelihood is,

$$
\begin{aligned}
L(\alpha, \beta) \;\propto\; & \sum_{i=1}^{m}\left[\sum_{j=0}^{y_i-1}\log(\alpha+j) + \sum_{j=0}^{n_i-y_i-1}\log(\beta+j)\right. \\
& \left. -\; \sum_{j=0}^{n_i-1}\log(\alpha+\beta+j)\right]. \qquad (7.37)
\end{aligned}
$$

To illustrate what can be obtained from this model, consider the *Gambusia* data. The scientific mechanism or phenomenon of interest is embodied in the parameters α and β of the mixture model with log likelihood (7.37) which is written for one group (i.e., treatment group). In this example we have 4 treatment groups, $SLD\,R-10$, $SLD\,R-16$, $Volta\,R-10$ and $Volta\,R-16$. An initial question of interest might be whether there is evidence of a difference between the holding water ionic contents $R-10$ and $R-16$ within each location (SLD and Volta). The log likelihood (7.37) was maximized (how to do this in the estimation section) in α and β separately for each of the 4 treatment groups. The resulting maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ were put into beta pdfs, and those functions graphed. The two resulting estimated densities for $R-10$ and $R-16$ holding waters are shown for the SLD area in Figure 7.31 and for the Volta area in Figure 7.32. Recall that the observed quantities were number of *live* young, so that high values of the beta variates are "good". It would appear from Figure 7.31 that a holding water of lower ionic content $R-10$ than the SLD environment from which these fish were collected (16 ppm) has shifted probability mass to the right ("good") in these densities. The same effect seems to be even more pronounced in Figure 7.32, computed for data from the Volta area. Thus, it appears that fish in $R-10$ holding water may have had higher reproductive success than their counterparts (i.e., fish from the same area) that were held in $R-16$ water and that this could be true for both the SLD and Volta areas. We might proceed then to compare fish between the two areas at a constant level of holding water ionic strength. If, for example, it is only the difference between $R-10$ and $R-16$ water that is important, then there should be no difference between SLD fish at $R-16$ and Volta fish at $R-16$.

Estimated beta mixing densities are presented for fish from the two areas

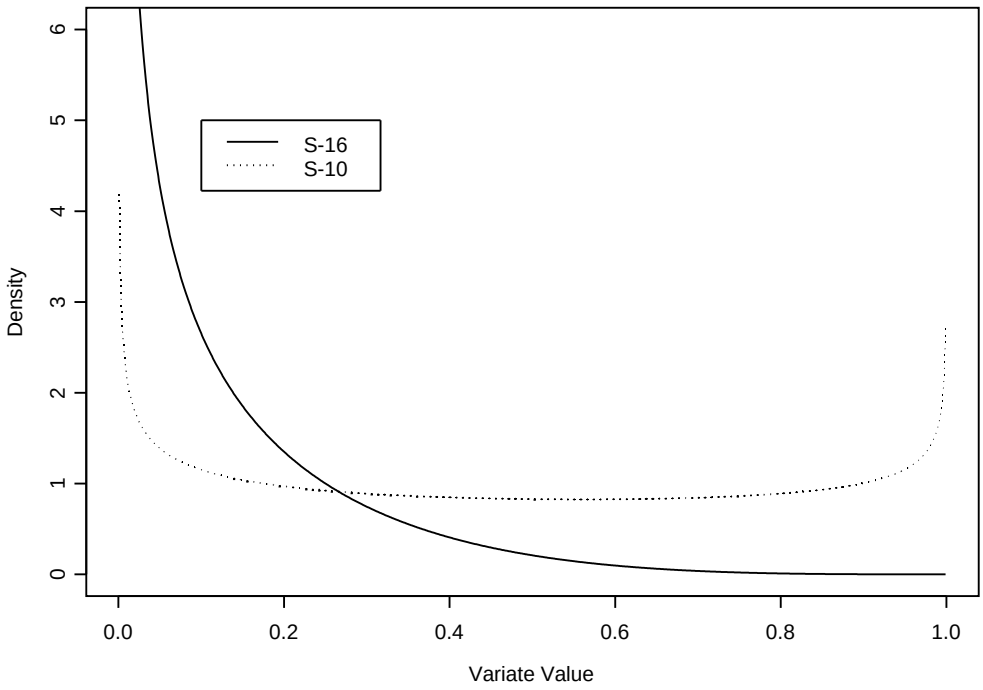

Figure 7.31: Estimated beta mixing pdfs for the SLD area with $R-10$ and $R-16$ holding waters.

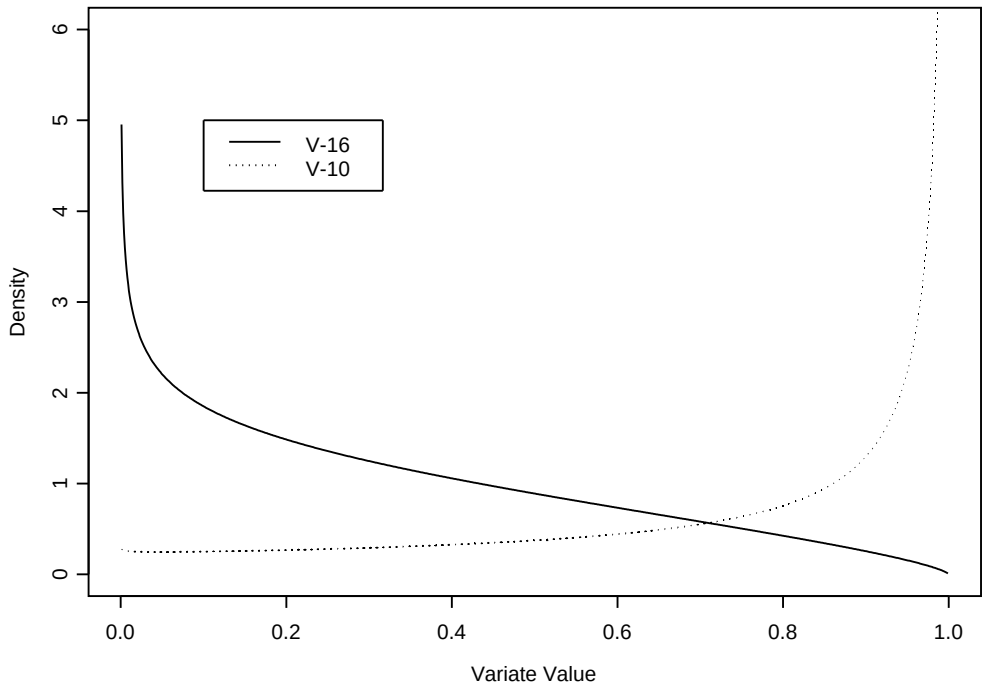

Figure 7.32: Estimated beta mixing pdfs for the Volta area with $R-10$ and $R-16$ holding waters.

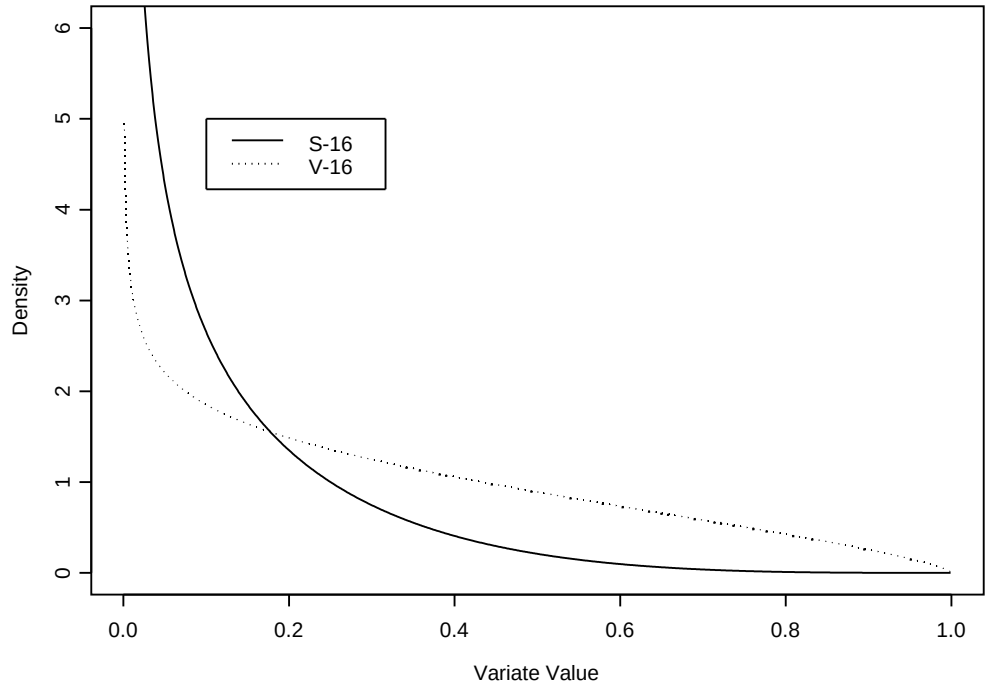

Figure 7.33: Estimated beta mixing pdfs for the SLD and Volta areas with $R-16$ holding water.

in $R-16$ holding water in Figure 7.33. Here, both densities are J shaped, although it appears that the density for the Volta area "picks up probability" more slowly than that for the SLD area. Often, in comparison of J and U shaped densities it is easier to interpret cumulative densities, and the cdfs that correspond to the densities of Figure 7.33 are presented in Figure 7.34. It is clear from this plot that the distribution for the SLD fish does place higher probability on smaller values than the distribution for the Volta fish. We have not conducted a formal testing procedure but for now assume that the difference in these distributions is meaningful (i.e., *significant*). The conclusion

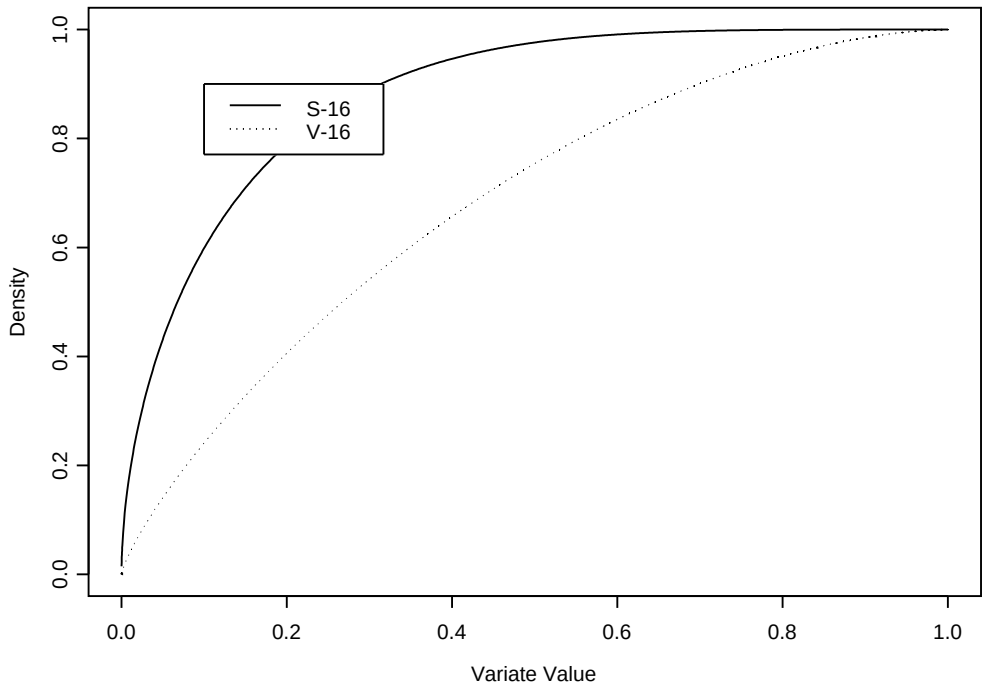

Figure 7.34: Estimated beta mixing cdfs for the SLD and Volta areas with $R-16$ holding water.

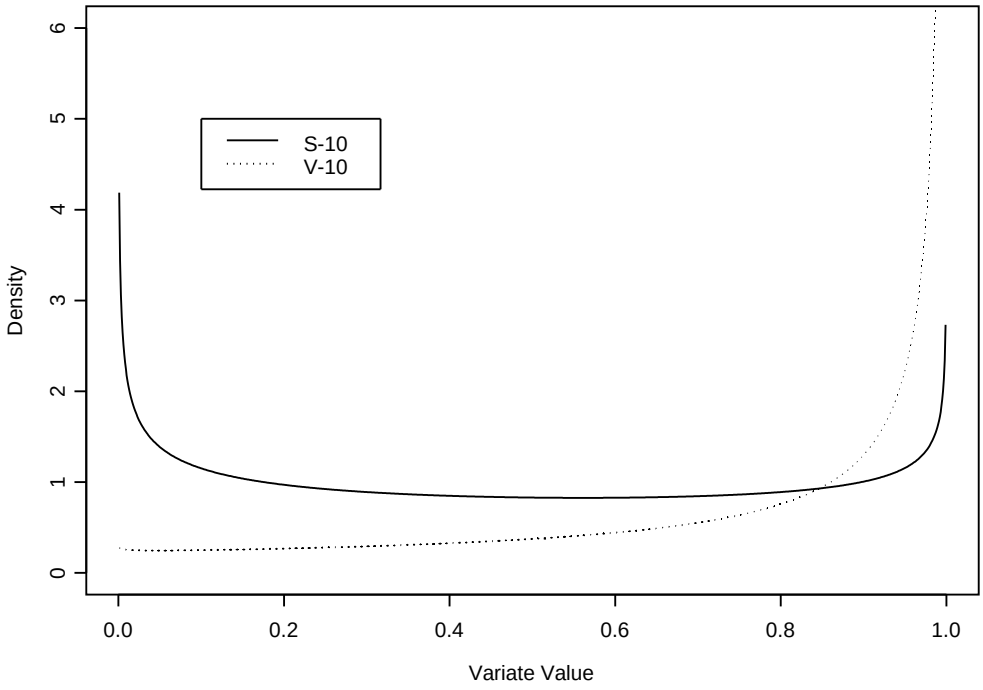

Figure 7.35: Estimated beta mixing pdfs for the SLD and Volta areas with $R-10$ holding water.

we would reach is that the fish from SLD tend to have fewer live young than the fish from Volta.

A similar comparison can be made for fish held in $R-10$ water, and the resulting densities and cumulative densities are presented in Figure 7.35 and 7.36, respectively. The pdf for the SLD area indicates the possibility of positive probability at both ends of the range (0, 1) with a uniform accumulation of probability over the remainder of the possible values, although it is difficult to tell whether or not there is substantial mass at the ends (the upward bends of the U may be so sharp as to have little probability mass beneath them).

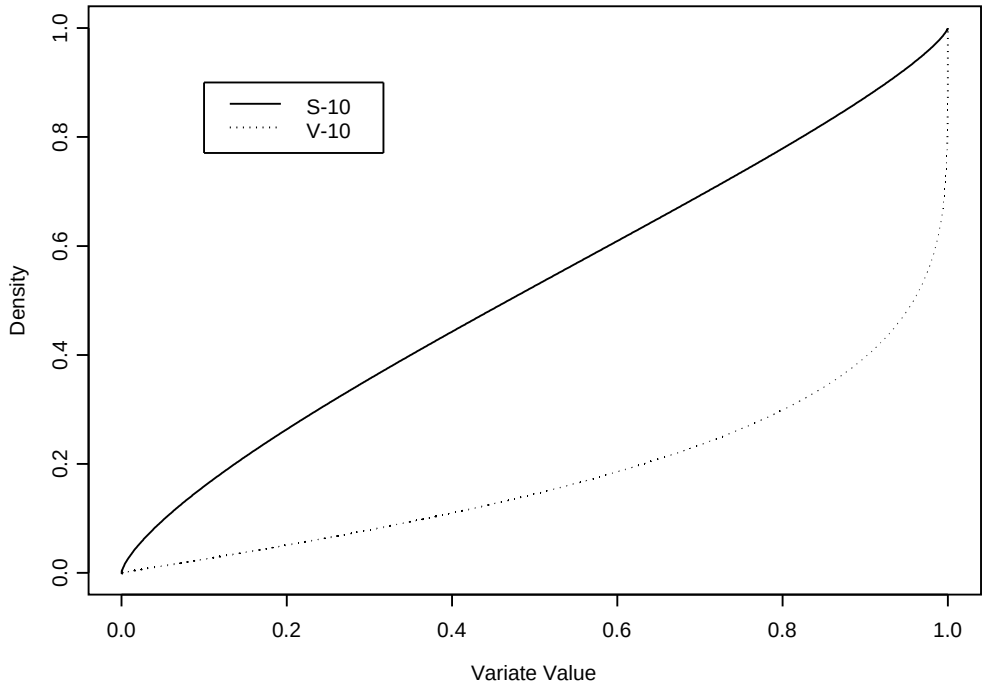

Figure 7.36: Estimate beta mixing pdfs for the SLD and Volta areas with $R - 10$ holding water.

The density for the Volta area again indicates the majority of probability is attained for larger values. The cdfs make the situation with the SLD area more clear. In Figure 7.36 it is clear that the distribution is nearly uniform, with little additional probability contained at the two extremes.

Thus far, we would seem to have an indication (albeit only speculative at this point without a formal probabilistic comparison) that fish from the SLD area tend to have broods of fish with a smaller proportion of live young than do fish from the Volta region, and this is true across ionic strengths of holding water. This provides at least indirect evidence that there is something about

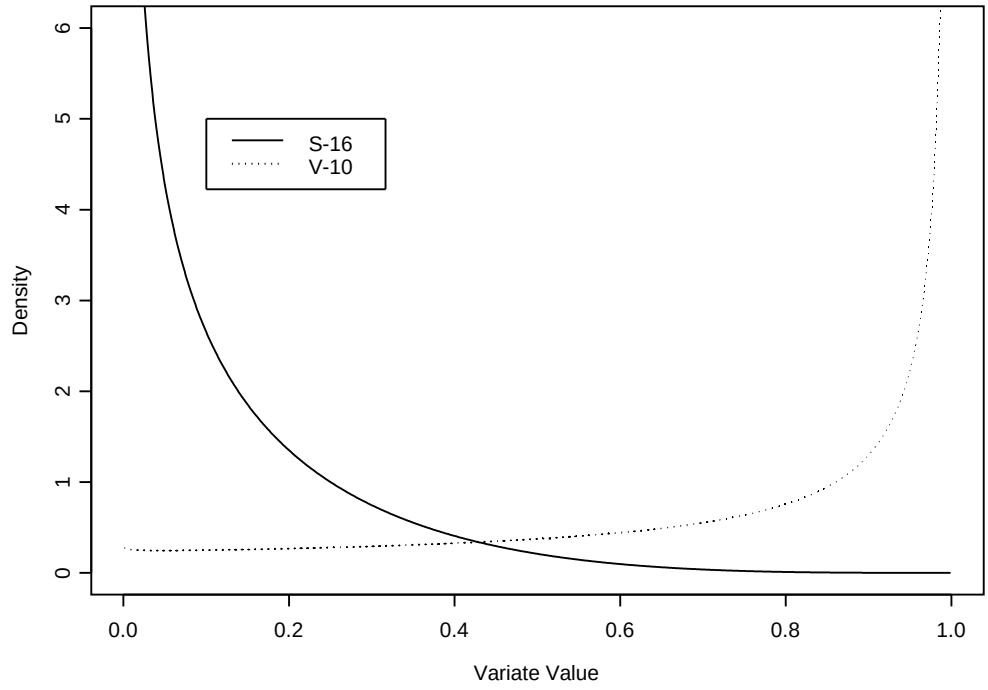

Figure 7.37: Estimate beta mixing pdfs for the Volta area with $R-10$ and $R-16$ holding waters.

the SLD area that decreases reproductive success other than simple that fact that the water is of higher ionic strength. In nature, the SLD area has ionic strength of about 16ppm, while the value for the Volta area is about 10ppm, and fish from these areas are not moved to different water before giving birth. What difference might we expect to see in the actual situations of SLD and Volta? This is represented by a comparison of the SLD $R-16$ and Volta $R-10$ groups, the pdfs for which are presented in Figure 7.37. The distinction between these pdfs is clear, the J shaped densities appearing as almost "mirror images" relative to regions of positive probability. An interesting summary of

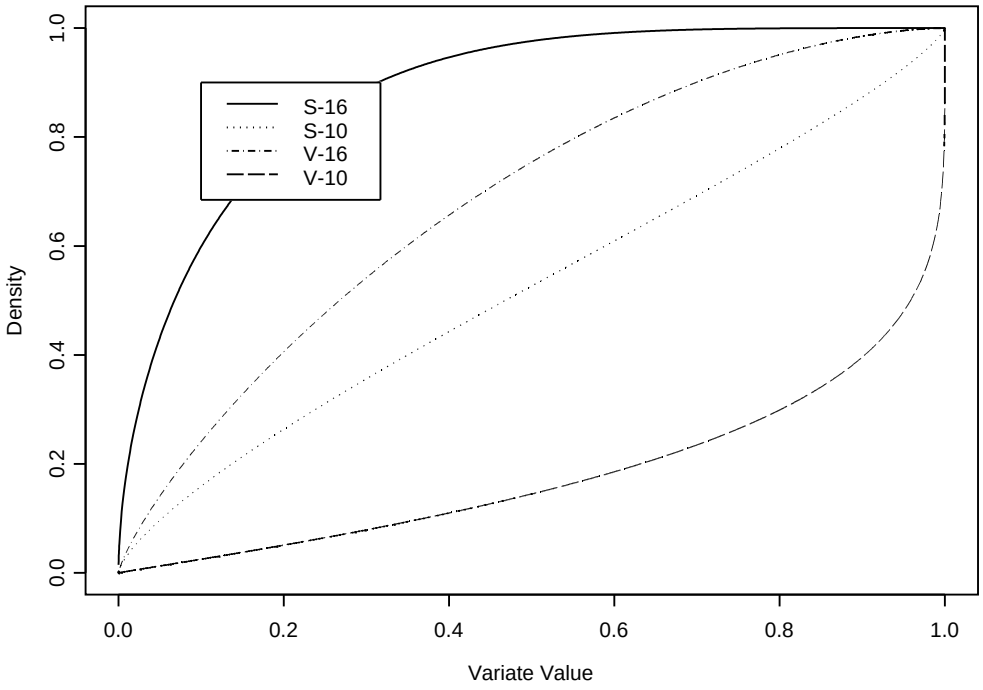

Figure 7.38: Estimated beta mixing cdfs for all four groups of fish.

the four groups used in this study is presented by the plot containing all four cdfs in Figure 7.38. Here, the effect of higher ionic strength of holding water is seen as a shift of probability to smaller values of the beta random variables θ_i for both SLD and Volta areas, while the same difference is seen between distributions for the two areas when compared within ionic strength of holding waters. A complication to reaching a totally unambiguous conclusion is that the cdf for SLD $R-10$ puts more probability at higher values than does the cdf for Volta at $R-16$, although we certainly do not know at this point whether that apparent difference is meaningful.

Our general notation for a mixture distribution given by expression (7.33)

applies to situations in which the dominating measure ν is either Lebesgue measure or counting measure. When ν is counting measure, the resultant $h(\cdot)$ is called a *finite mixture* distribution. In this case the mixture distribution $g(\cdot)$ is a probability mass function that is typically assumed to put positive probability at a small set of points given by $\{\theta_i : i = 1, \ldots, k\}$. If, in this context, we have

$$g(\theta_i|\lambda) = \begin{cases} \pi_1 & \text{if } \theta_i = \theta_1 \\ \pi_2 & \text{if } \theta_i = \theta_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \pi_k & \text{if } \theta_i = \theta_k \end{cases}$$

then (7.33) becomes

$$h(y_i|\pi_1, \ldots, \pi_k) = \sum_{j=1}^{k} \pi_j \, f(y_i|\theta_j). \tag{7.38}$$

The density or mass function given by (7.39), depending on whether f is a pdf or pmf, is that of a finite mixture. The term *mixture model* historically implied a finite mixture and, to some statisticians, still does; the term *compound distribution* used to be used to refer to what we are calling mixtures, although this has nearly died in standard nomenclature.

Finite mixtures are useful in a number of circumstances. We may have a scientific hypothesis that the observed data have arisen from several "groups" that differ in distributional characteristics, but it is impossible to assign sampling units to these groups *a priori*.

Example 7.13

Consider a hypothetical example in which the goal is to determine whether two different "substrates" have an effect on the length of time blooms last for cultured orchids. It is believed (or suspected) that the orchids in question consist of two sub-populations of "short-timers" and "long-timers" for genetic reasons that we have not yet discovered (based, for example, on the observation of bimodal histograms for bloom length in previous observations). Since plants are relatively "plastic" in terms of genetic expression, we would like to know whether the growth substrates might affect the proportion of a given seed stock that develop into short-timers and long-timers, as well as determining whether the substrates affect length of viable bloom within each sub-population. We have not been able to establish a relation between any observable characteristics and short versus long bloom times and, in addition, it is likely this is a artificial division with considerable overlap in the tails of bloom time distributions. In this situation it might be appropriate to consider fitting a finite mixture model with $k = 2$ "components"and component distributions $f(y_i|\theta_1)$ and $f(y_i|\theta_2)$. We might, based on either previous experience or simply in hopes that it would prove appropriate, take the component distributions to be normal with parameters $\theta_1 \equiv (\mu_1, \sigma_1^2)$ and $\theta_2 \equiv (\mu_2, \sigma_2^2)$. This would result in a model (for each treatment group) containing 6 parameters to be estimated, namely π_1, π_2, μ_1, μ_2, σ_1^2, and σ_2^2.

It is also the case that many observed data patterns may be described by finite mixtures of normal distributions. This application is perhaps less of a true "modeling" exercise than it is a matter of pure "data description". Note that, in both of these potential applications of finite mixture distributions, the number of components k is assumed known. The estimation of a finite mixture with an unknown number of component distributions is a difficult

problem; there are several ideas for how to approach the problem of estimating k along with other model parameters, which are beyond the scope of this course.

To close this subsection, we list a number of data model/random parameter model combinations that are frequently used. There is nothing "magic" about these combinations, and our thinking should not be constrained by the following list. Nevertheless, the following data model/random parameter model combinations do have nice mathematical properties and often seem reasonable combinations to use in practical situations.

1. Beta-Binomial. This mixture model, illustrated in example 7.11, was introduced by Williams (1982) in the context of studies of teratogenic effects. Despite that fact that Williams presents both the model and analysis in purely frequentist terms, (too) many statisticians still associate mixture models with Bayesian formulations.

2. Gamma-Poisson. Here, the data model consists of conditionally independent Poisson distributions (e.g., for counts) with parameters that follow *iid* gamma distributions. This model is sometimes referred to as a "negative binomial" distribution, but under the development of a negative binomial as the number of binary trials preceding a given number of successes, this is only true for integer-valued gamma parameters.

3. Normal-Normal. If the data model consists of conditionally independent normals with either known variances or variances considered as uninteresting "nuisance" parameters, a natural choice for the distributions of data distribution means is again the normal.

4. Normal-Inverse Gamma-Normal. If the data model consists of condi-

tionally independent normals with unknown variances that are also to be modeled as random variables (as well as the means), then a common structure is to assign the random data model means a normal distribution and the random data model variances an inverse gamma distribution; an inverse gamma distribution is the distribution of $1/X$ for a random variable X that follows a gamma distribution.

5. Multinomial-Dirichlet. The multinomial is an extension of the binomial. The Dirichlet is an extension of the beta. In the same way that a beta makes an attractive mixing distribution for binomial data models, the Dirichlet makes an attractive mixing distribution for multinomial data models. It is important to note, however, that what are called *bounded sum* conditions become important in such formulations. That is, if $\boldsymbol{Y}_i \equiv (Y_{1,i}, \ldots, Y_{k,1})$ is multinomial with m polychotomous trials , then $\sum Y_{h,i} = m$ and, if the parameters of a multinomial are taken to be $p_1, \ldots, p_k$, then $\sum p_h = 1$. Such bounded sum conditions automatically introduce negative dependence among the component quantities.

While we certainly want to avoid thinking that the combinations given above are the "best" combinations of data models with random parameter models, one thing the all share in common is that the sets of possible values for the various quantities involved match nicely.

Mixture Models in Regression

It should be clear that for linear regression models, assigning the regression parameters a normal distribution gives equivalent models to using random effects and/or random coefficient linear models as discussed under the heading of Mixed Models in Section 7.4.1.

For example, a linear random effects model for grouped random variables was given in expression (7.24) as

$$Y_i = \boldsymbol{x}_i^T \beta + \tau \sum_{j=1}^{J} \delta_j I(i \in \mathcal{C}_j) + \sigma \epsilon_i,$$

and was also written for double subscripting and a single covariate in Example 7.10 as,

$$Y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \tau \delta_j + \sigma \epsilon_{i,j}.$$

In these models we took $\delta_j \sim iid\, N(0,\, 1)$. The conditional models may be written as,

$$\begin{aligned} Y_i &= \left\{ \beta_0 + \tau \sum_{j=1}^{J} \delta_j I(i \in \mathcal{C}_j) \right\} + \boldsymbol{x}_i^T \beta + \sigma \epsilon_i \\ \textit{or} & \\ Y_{i,j} &= (\beta_0 + \tau \delta_j) + \beta_1 x_{i,j} + \sigma \epsilon_{i,j}, \end{aligned}$$

still with $\delta_j \sim iid\, N(0,\, 1)$. But, these models could also be written with random intercept parameters as,

$$\begin{aligned} Y_i &= \sum_{j=1}^{J} \beta_{0,j} I(i \in \mathcal{C}_j) + \boldsymbol{x}_i^T \beta + \sigma \epsilon_i, \\ \textit{or} & \\ Y_{i,j} &= \beta_{0,j} + \beta_1 x_{i,j} + \sigma \epsilon_{i,j}, \end{aligned} \tag{7.39}$$

where now $\beta_{0,j} \sim iid\, N(B_0,\, \tau^2)$. The models of (7.39) are in the form of mixture models, but are equivalent to the previous random effects models.

Similarly, linear random coefficients models, such as those of (7.27),

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{\gamma}_j + \sigma \boldsymbol{\epsilon}_j,$$

or (7.28),

$$Y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \tau_1 \delta_j + \tau_2 \gamma_j x_{i,j} + \sigma \epsilon_{i,j}$$

may be written as a random parameter models. Let $\boldsymbol{W}_j$ denote a the matrix of columns of $\boldsymbol{X}_j$ that are not included in $\boldsymbol{Z}_j$. Then the first model above can be written as, with $\boldsymbol{\beta}_f$ a fixed constant,

$$\boldsymbol{Y}_j = \boldsymbol{W}\boldsymbol{\beta}_f + \boldsymbol{Z}_j\beta_j + \sigma\boldsymbol{\epsilon}_j.$$

In situations for which $\boldsymbol{Z}_j = \boldsymbol{X}_j$ (which are probably the most common) or in which there is only one covariate measured on a ratio or interval scale we can define $\boldsymbol{x}_i \equiv (1, x_{1,i}, \ldots, x_{p,i})^T$, $\boldsymbol{\beta}_i \equiv (\beta_{0,i}, \ldots, \beta_{p,i})^T$, and write the linear random parameter regression model as,

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}_i + \sigma\epsilon_i, \tag{7.40}$$

where $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_n^T)^T$ has a joint $(n \times (p+1))$ Gaussian distribution. If the random variables Y_i are grouped, we can use model (7.40) with $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ if i is in group j, $j = 1, \ldots, k$, and take $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_k^T)^T$ to have a joint $(k \times (p+1))$ Gaussian distribution.

Notice that the notation becomes far more involved than the concept, here. For linear regressions in which we want one or more of the regression parameters to have a random component, we can write those coefficients as either a mean plus random portion (having mean 0), which is the mixed model formulation, or as random variables having a location scale distribution (almost always normal), which is the random parameter formulation. The models are equivalent, as we have shown for the conditional models. This is also true for the marginal models since, if $h(x, \beta)$ is a linear function of fixed x and random β, $E\{h(x, \beta)\} = h(x, E\{\beta\})$.

The situation changes little, relative to conditional model specification, for nonlinear additive error models. That is, if a model is formulated as in expression (7.1), it makes little difference whether we take $\beta = \beta_f + \beta_r$, with

f denoting fixed and r denoting random, or take β as random with some location-scale distribution. This all stems from the fact that, for *any* random variable ϵ following a location scale distribution F with mean 0 and variance 1, the following are equivalent:

$$\begin{aligned} Q &= \mu_Q + \tau\,\epsilon; \quad \epsilon \sim F \\ \text{and} & \\ Q &\sim F(\mu_Q, \tau^2) \end{aligned}$$

The modeling decision between mixed and mixture formulations will depend on (1) whether the mean μ_Q has any viable interpretation in a marginal model, and (2) whether a location-scale distribution is appropriate for modeling β.

The situation is slightly different in the case of models formulated along the lines of generalized linear models. Here, of course, the emphasis may well be on a (other than location-scale) distribution for the response variables, as discussed at some length in Section 7.3.2. There are two alternative model formulations that are commonly used in this situation. The first is similar to what has already been covered in this subsection, except that the systematic model component is not linear. Specifically, let $\{Y_i : i = 1, \ldots, n\}$ denote response random variables that follow a typical generalized linear model, with pdfs or pmfs of the Y_i as in expression (7.19) and,

$$g(\mu_i) = \boldsymbol{x}_i^T\,\beta,$$

which combines expressions (7.20) and (7.21). If the link function $g(\cdot)$ has range over the entire line, then it is common to assign β a Gaussian distribution (e.g., Gelfand and Ghosh, 2000). For example, if the Y_i are Poisson with means λ_i, in the exponential dispersion family form of (7.19) the natural parameter

is

$$\theta_i = \log(\lambda_i),$$

which has values $-\infty < \theta_i < \infty$. If $g(\cdot)$ is the canonical link $g(\mu_i) \equiv g(\lambda_i) = \theta_i$, then we have $-\infty < \boldsymbol{x}_i^T \beta_i < \infty$, and a Gaussian distribution for β_i may well be adequate.

But, the exponential dispersion family form of (7.19) presents an alternative for model formulation. This is to take the natural parameters θ_i as random, and then assign the systematic model component at the level of these random parameters. Specifically, let Y_i have the form of an exponential dispersion family (7.19), and take, the θ_i, for $i = 1, \ldots, n$ to have density or mass functions given by,

$$h(\theta_i|\lambda_i,\, \psi) = \exp\left[\psi\{\lambda_i\, \theta_i - b(\theta_i)\} + k(\lambda_i,\, \psi)\right], \tag{7.41}$$

where $b(\cdot)$ is the same function that appears in the exponential form for the response variables Y_i. For a more complete discussion of such models see Albert (1988), although that paper focuses on a Bayesian analysis, which is not (at this point) our emphasis. To complete the model, then, take a known smooth function $g(\cdot)$ such that,

$$g(\lambda_i) = \boldsymbol{x}_i^T \beta,$$

for fixed parameters β. Now, it turns out that (e.g., Morris, 1982) that $E(\theta_i) = \lambda_i$ so that assigning a systematic model component through the link function $g(\cdot)$ to λ_i from (7.41) is assigning a systematic model component to the *expectation* of the natural parameters rather than the random parameters themselves. It is not well understood what the difference is in what is being modeled (i.e., the statistical conceptualization given to a problem) between formulations in which a generalized linear model is assigned to observable random variables Y_i with random regression parameters β and formulations

in which natural parameters θ_i are random with a (fixed) generalized linear model assigned to their expected values λ_i.

Conditional and Marginal Models

We have touched briefly on the issue of conditional versus marginal models in a comparison of nonlinear mixed versus mixture models in Section 7.4.1. The basic issue was presented there as follows:

1. In linear models, regardless of whether they are specified in mixed model or random parameter (i.e., mixture model) form, the fixed parameters of the conditional model (or the means of the random parameter distribution) are parameters in the systematic component of a marginal model that has the same form (linear) as the conditional model. That is, if $g(\boldsymbol{x}_i, \beta_f, \beta_r)$ is the systematic component of the conditional model specification in a mixed model structure, with β_f fixed and β_r random such that $E\{\beta_r\} = 0$, then

$$E\{g(\boldsymbol{x}_i, \beta_f, \beta_r)\} = g(\boldsymbol{x}_i, \beta_f).$$

   Alternatively, if $g(\boldsymbol{x}_i,\ \beta_r)$ is the systematic model component of the conditional model such that $E\{\beta_r\} = \beta_f$, then

$$E\{g(\boldsymbol{x}_i,\ \beta_r\} = g(\boldsymbol{x}_i,\ E\{\beta_r\}) = g(\boldsymbol{x}_i,\ \beta_f).$$

2. Due to Jensen's Inequality, this same relation does not hold if $g(\cdot)$ is not a linear function. That is, the expected value of a conditional model with random parameters (or some fixed and some random parameters) is not equal to the conditional model evaluated at the expected values of the random parameters.

This discrepancy motivated Zeger, Liang and Albert (1988) to introduce the terms *population average* and *subject specific* models to refer to marginal and conditional model specifications, respectively.

Unfortunately, these terms have sometimes (and are, in fact) easily misinterpreted to mean that, in a nonlinear mixed model parameterization, the fixed parameters represent "typical parameters" in the conditional model while they represent parameters that produce a "typical response vector" in the marginal model (e.g., Lindstrom and Bates, 1990, p.686). This may be true for linear models, but is not at all clear thinking when it comes to nonlinear models. How in the world can we, as statisticians, explain to scientists that the "typical response in the population" is not the same as the response for "typical parameter values". Zeger, Liang and Albert (1988) were very careful in their original wording, indicating that the population averaged (i.e., marginal) model was a model for responses averaged over the population, while the subject specific (i.e., conditional) model was a model for individual responses. This is entirely correct, and is related to what is often called the *aggregation effect* in regression, which is also related to *Simpson's Paradox*, often presented in terms of categorical data. Put simply, there is no reason to believe that a model for aggregated data should be of the same form as models for unaggregated data.

The question is whether one wishes to formulate a model in which the focus is on population-level responses or on responses at the level of individual sampling units. If the former, a mixed model formulation may have some justification in terms of interpretation of the fixed model parameters (if the model is linear, but not if it is nonlinear). This is the motivation for considering conditional (or subject-specific) models as a mechanism by which to formulate a covariance matrix for the marginal distribution of responses (see the material following expression 7.31 in these notes). If, however, one makes use of the

mixture concept that a scientific mechanism is leading to different realizations in different circumstances, then interest centers on conditional model specification. That is, the mechanism of interest is not necessarily reflected in a model for population averages. The "average" mechanism in the population is the conditional model evaluated at the expected parameter values, $h(\boldsymbol{x}_i, E\{\beta\})$, *not* the marginal model $E\{h(\boldsymbol{x}, \beta)\}$.

There has been considerable effort to collapse the emphasis on conditional or marginal models. Zeger, Liang and Albert (1988) and Breslow and Clayton (1993) both explore the possibility that marginal models constitute first-order approximations of conditional models, sometimes by adding what is known as an "offset" in generalized linear models, and sometimes by altering values of either the covariates or the regression parameters in a systematic fashion. While such attempts are interesting, and certainly add to our overall knowledge of what such models are doing, they are fundamentally misplaced effort. That is, if one is primarily interested in modeling population-level responses, consideration of conditional models serves only to help formulate the marginal model, primarily in terms of the covariance matrix. If, on the other hand, one is primarily interested in modeling responses of individual sampling units, the marginal model is really not a "model" at all, but merely a joint distribution of response variables that must be used for estimation.

Consider the situation of Example 7.6, dealing with the concentration of cadmium in fish. Given observations from a number of lakes (i.e., groups or clusters) our concern is with the manner in which the lengths of fishes are related to the corresponding cadmium concentrations. Thus, the scientific mechanism or phenomenon of interest is embodied in the conditional model. There really is no marginal model in the sense that we care little what relation (if any) might exist between average length and average cadmium concentra-

tion for fish from a variety of lakes (which differ in all kinds of factors that may play a role in determining the relation of interest – lengths and cadmium concentrations of individual fish).

7.4.3 Latent Variable Models

We turn now to our third manner of formulating models that results in mixture distributions for marginal response models. This is the use of random variables to represent "unobservable" effects in a model. The title of *latent variables* is, in this context, a broad concept. In many applications connected with the social sciences, the term latent variable model implies what are known as "structural equations" models. This is not our implication. Rather, we will use the term latent variable to refer to the portion of a model that corresponds to a phenomenon (or a collection of phenomena) that cannot be measured or observed.

As indicated in the introductory comments of this section on models with multiple stochastic elements, there are few guiding principles as to how such models are formulated. Thus, our presentation will consist of an example intended to illustrate the flexibility of such models. In addition, this example is intended to convey one other aspect of modeling, namely that useful models are often formulated on the basis of very fundamental scientific concepts. We will, yet again, bring up the basic message of Chapter 5.3, which is the encapsulation of a scientific mechanism or phenomenon of interest in a small set of model parameters. Latent variables are often a useful manner of modeling at least some of the things left unexplained by that mechanism in various circumstances.

Example 7.14

We have already introduced an example (albeit hypothetical in nature) that concerned the level of a particular toxic algal genus known as *Microcystin* (Example 6.7). This example concerns the same problem, but with real data. What is pertinent from the previous example is that the concentration of nitrogen in lakes and reservoirs may be related to the concentration of *Microcystin* in those waters. This idea is based on a very basic ecological concept known as the *Law of Limiting Factors*. The fundamental ideas are captured in the two portions of this theory, called *Leibig's Law of the Minimum* and *Shelford's Law of Tolerance*. Very briefly (and in reduced technical form) these two laws are as follows:

1. Leibig's Law of the Minimum.

   When a biological process (such as growth) depends on a number of necessary inputs (such as nutrients) that are consumed during the process, the process is limited (i.e., stopped) by the input that is used up the most quickly. For example, if the growth of a plant depends on the primary plant nutrients of phosphorus and nitrogen, radiant energy (as sunlight), and carbon, whichever of these factors are in shortest supply will stop the growth process when it runs out. This concept is employed, by the way, in agriculture when a farmer decides what type of fertilizer to apply to a crop.

2. Shelford's Law of Tolerance.

   The ecological fitness of an organism is often reflected in the abundance of that organism (e.g., species or genus). The environment in which organisms exist consist largely of a set of environmental *gradients* such

as gradients in altitude, temperature, salinity, oxygen availability, etc. Along a given gradient, a type of organism will have a range of tolerance outside of which it cannot exist. But, even within that range, there will be a preferred level along the gradient at which the organism is most "comfortable". This is often represented as a unimodal curve for which the vertical axis is abundance and the horizontal axis is the environmental gradient. The mode of the curve is interpreted as the "optimal" level of the gradient for the given organism.

How do we reflect these ideas in statistical models? Leibig's law of the minimum is reflected, for a single factor, in the type of model introduced by Kaiser, Speckman and Jones (1994), a simple version of which has the form

$$Y_i = x_i \gamma \, U_i + \sigma \, \epsilon_i, \tag{7.42}$$

in which Y_i is the response variable, x_i is the factor of concern, and U_i is an unobservable (i.e., latent) random variable with possible values on the interval (0, 1). If $U_i < 1$, then some factor other than x_i must be limiting for the observation connected with the response Y_i. Kaiser, Speckman and Jones (1994) took $U_i \sim iid$ beta (α, β) and $\epsilon_i \sim iid\, N(0, 1)$ for $i = 1, \ldots, n$, and U_i independent of ϵ_i. This model leads to an expected data pattern of a triangular array (or "wedge") of data scattered below a straight line through the origin, as illustrated by a simulated data set in Figure 7.39. These data were simulated using a beta $(2, 3)$ distribution for the latent U_i variables, which gives an expected value of 0.40 times the line $x_i \, \gamma$. There are a host of questions regarding response function form (e.g., other than linear through the origin) that we are brushing aside at the moment. The point is that a latent random variable U_i has been used to model departures from the mechanism (limiting

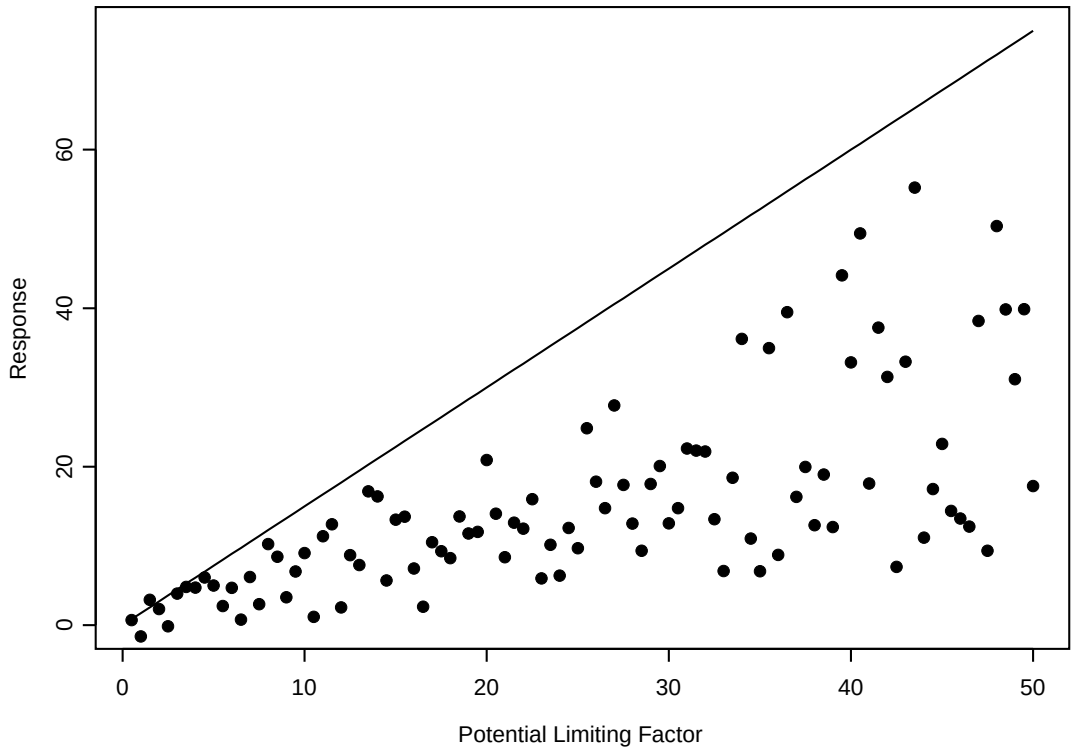

Figure 7.39: Simulated data showing model for limiting factors based on Leibig's law of the minimum.

factor) of interest. Various data patterns can be generated by this model, depending on the values of the beta parameters and σ that are used.

Now, combine this modeling idea with Shelford's law of tolerance. Consider a unimodal curve that describes the tolerance or affinity of an organism for various levels of a single environmental gradient (this is Shelford's law of tolerance). Certainly, the abundance of that organism depends on a number of factors besides the one gradient in question, and abundance may not reach the potential level it could for a given level of the gradient under study (the limiting factor model concept). The inescapable conclusion is that the entire unimodal curve in question represents an "optimum", given that all other factors are at their "most favorable" levels. Thus, in observed data what should we expect? We should expect observed values of the abundance of an organism to be scattered below a unimodal curve, because it is not always true that all other factors are at their most favorable levels.

What is needed to formulate a model for this situation? Essentially, replacing the linear limit function $x_i\,\gamma$ of model (7.42) with a unimodal curve. There are any number of possible choices for unimodal curves, many of which have restrictions on shape other than unimodality (e.g., a normal pdf is a unimodal curve, but must always be symmetric). One possibility is the function

$$f(x_i,\,\boldsymbol{\theta}) = \frac{\theta_1}{\Gamma(\theta_2 + \theta_3 x_i + \theta_4 x_i^2)}. \tag{7.43}$$

The function in expression (7.43) is quite flexible, with fairly nice connections between its behavior and the values of θ_1, θ_2, θ_3 and θ_4; θ_1 governs height, θ_2 governs whether both or only one tail is seen, θ_3 governs rate of increase, and θ_4 governs rate of decrease. A model for the situation we are trying to capture may then be formulated as,

$$Y_i = f(x_i,\,\boldsymbol{\theta})U_i + \sigma\epsilon_i, \tag{7.44}$$

where $f(\cdot)$ is given in (7.43), $U_i \sim iid\, G$ for some distribution on the interval $(0,\ 1)$, and $\epsilon \sim iid\, F_\epsilon$ for a location-scale family F_ϵ with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$.

Model (7.44) has been applied to data on Microcystin abundance (Y_i) with the covariate or potential limiting factor of nitrogen concentration (x_i). In this application, F_ϵ was taken to be logistic, and G was a "histogram model" formed by dividing the unit interval into a partition (e.g., 0 to 0.25, 0.25 to 0.5, 0.5 to 0.75, and 0.75 to 1.0). A simulated data set is presented in Figure 3.41, along with the true limit function $f(x_i,\ \boldsymbol{\theta})$ (as the solid curve) and an estimated limit function based on maximum likelihood estimates of the components of $\boldsymbol{\theta}$ (as the dashed curve).

To fully specify models (7.42) or (7.44) we need to write down forms of the various distributions involved so that we can derive the marginal distribution of the Y_i given parameters involved in the distributions of the U_i and ϵ_i.

Assuming continuous Y_i and continuous U_i, the general form implied by model (7.44) is as follows. First, for given U_i, the conditional density of Y_i is a location-scale transformation of F_ϵ,

$$f(y_i|u_i, \boldsymbol{\theta}, \sigma) = \frac{1}{\sigma} f_\epsilon \left(\frac{y_i - f(x_i,\ \boldsymbol{\theta})\, u_i}{\sigma} \right). \tag{7.45}$$

The marginal distribution of the U_i are (iid) with pdf $g(u_i, \boldsymbol{\eta})$ which may depend on the parameter (vector) $\boldsymbol{\eta}$. Then the joint of Y_i and U_i is,

$$p(y_i,\ u_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = f(y_i|u_i, \boldsymbol{\theta}, \sigma)\, g(u_i, \boldsymbol{\eta}), \tag{7.46}$$

and the marginal of Y_i is given by the mixture distribution,

$$h(y_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = \int f(y_i|u_i, \boldsymbol{\theta}, \sigma)\, g(u_i, \boldsymbol{\eta})\, du_i. \tag{7.47}$$

For model (7.42), which is a special case of (7.44), $f(x_i,\ \boldsymbol{\theta}) = x_i\gamma$, $\boldsymbol{\theta} \equiv \gamma$, $f_\epsilon(\cdot)$ is standard normal and $g(\cdot)$ is beta with $\boldsymbol{\eta} \equiv (\alpha,\ \beta)$ so that

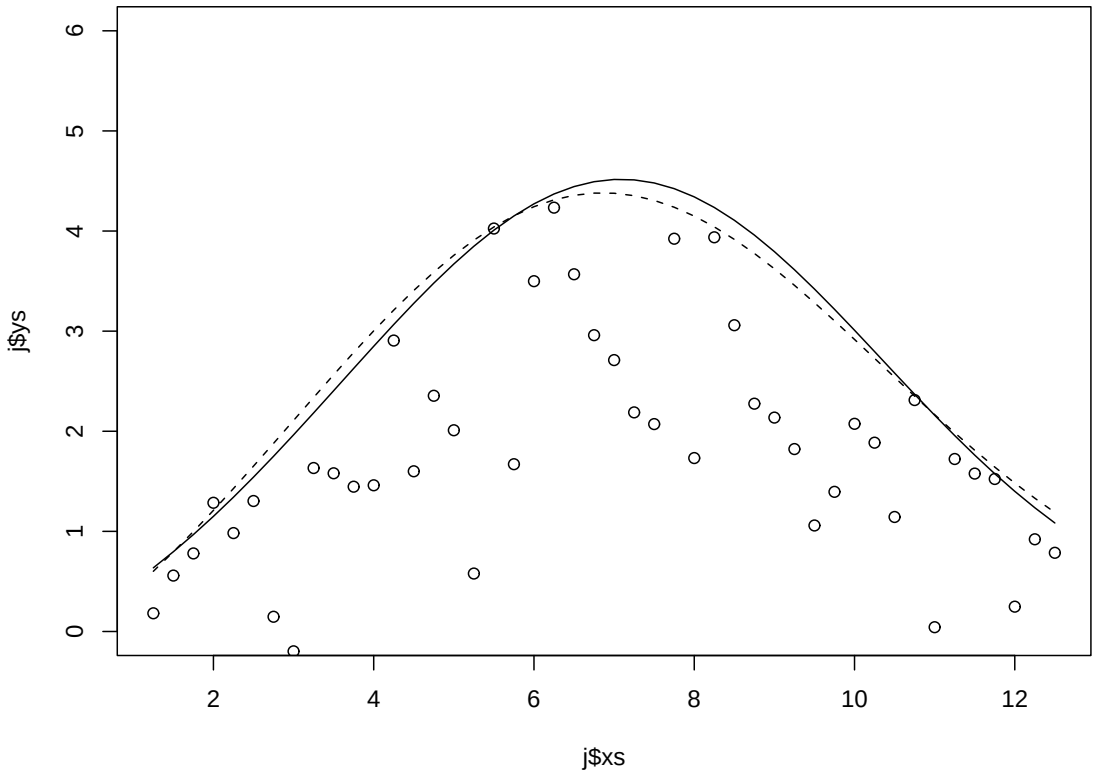

Figure 7.40: Data simulated from model (7.44) with true limit function given as the solid curve and estimated limit function as the dashed curve.

$$p(y_i|\gamma,\sigma,\alpha,\beta) =$$

$$\frac{\Gamma(\alpha+\beta)}{(2\pi\sigma^2)^{1/2}\Gamma(\alpha)\Gamma(\beta)}$$

$$\times \quad \int_0^1 \exp\left\{\frac{-1}{2\sigma^2}(y_i-\gamma x_i u_i)^2\right\} u_i^{\alpha-1}(1-u_i)^{\beta-1}\,d\,u_i.$$

In the application to Microcystin data, we defined $\lambda_j \equiv 0.25\,j$, for $j = 0,1,2,3,4$, defined $\boldsymbol{\eta} \equiv (\eta_1,\eta_2,\eta_3,\eta_4)$, and took

$$g(u_i|\boldsymbol{\eta}) = \eta_j\, I(\lambda_{j-1} < u_i < \lambda_j); \quad j = 1,2,3,4.$$

Note this is a "histogram" model with probabilities $0.25\eta_j$ so that we have imposed the restriction

$$\sum_{j=1}^{4} \eta_j = 4.$$

For the moment, leave $f_\epsilon(\cdot)$ unspecified. Then (7.46) becomes, for $j = 1,2,3,4$,

$$p(y_i,\, u_i|\boldsymbol{\theta},\sigma,\boldsymbol{\eta}) = f(y_i|u_i,\boldsymbol{\theta},\sigma)\eta_j\, I(\lambda_{j-1} < u_i < \lambda_j),$$

which leads to the mixture distribution of (7.47) as,

$$h(y_i|\boldsymbol{\theta},\sigma,\boldsymbol{\eta}) = \sum_{j=1}^{4} \eta_j \int_{\lambda_{j-1}}^{\lambda_j} f(y_i|u_i,\boldsymbol{\theta},\sigma). \tag{7.48}$$

Now, let

$$w_i \equiv \frac{1}{\sigma}\left\{y_i - f(x_i,\boldsymbol{\theta})u_i\right\},$$

or,

$$u_i = \frac{1}{f(x_i,\boldsymbol{\theta})}(y_i - \sigma w_i); \quad \frac{d\,u_i}{d\,w_i} = \frac{-\sigma}{f(x_i,\boldsymbol{\theta})}.$$

Then from (7.45) and (7.48),

$$h(y_i|\boldsymbol{\theta},\sigma,\boldsymbol{\eta}) \;=\; \sum_{j=1} 4\eta_j \int_{\xi_{j-1}}^{\xi_j} \frac{f_\epsilon(w_i)}{f(x_i,\boldsymbol{\theta})}\,d\,w_i$$

$$= \sum_{j=1}^{4} \frac{\eta_j}{f(x_i, \boldsymbol{\theta})} \left[F_\epsilon(\xi_j) - F_\epsilon(\xi_{j-1}) \right], \tag{7.49}$$

where

$$\xi_j \equiv \frac{1}{\sigma} \left\{ y_i - f(x_i, \boldsymbol{\theta}) \lambda_{j-1} \right\}; \quad \xi_{j-1} \equiv \frac{1}{\sigma} \left\{ y_i - f(x_i, \boldsymbol{\theta}) \lambda_j \right\}$$

The fact that ξ_j is a function of λ_{j-1} and ξ_{j-1} is a function of λ_j comes from $(d\, u_i / d\, w_i) < 0$. In this particular example, F_ϵ was taken to be a logistic distribution, $F_\epsilon(x) = (1 + \exp(x))^{-}1$.

Now, all of this effort has been expended to render the mixture (7.59) in a form that does not involve an unevaluated integral as is present in the mixture formulated from model (7.42). It would, of course, have been possible to use the same type of distributions for both U_i and ϵ_i in (7.44) as was done in (7.42) or, to use the histogram model for U_i and logistic F_ϵ in model (7.42).

In any case, returning to the general notation of (7.45) through (7.47), the log likelihood for a set of observed $y_1, \ldots, y_n$ is,

$$L(\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = \sum_{i=1}^{n} h(y_i | \boldsymbol{\theta}, \sigma, \boldsymbol{\eta}). \tag{7.50}$$

An application of model (7.44) with $f(x_i, \boldsymbol{\theta})$ as in (7.43) and the resulting mixture in the form of (7.49) to the actual Microcystin data resulted in estimates presented in Figure 7.41. In this figure, the unimodal "tolerance" curve (7.43), evaluated at maximum likelihood estimates of the components of $\boldsymbol{\theta}$ is shown as the solid line, with point-wise 90% interval estimates given as dashed lines. Actual inference in this problem depends, of course, not only on the estimated value of $\boldsymbol{\theta}$ but also σ and, importantly, $\boldsymbol{\eta}$; these latter values determine the probability that responses reach various proportions of the tolerance curve. The models presented in this subsection are intended to *illustrate* the use of latent variables in the formulation of models, not be an exhaustive survey of

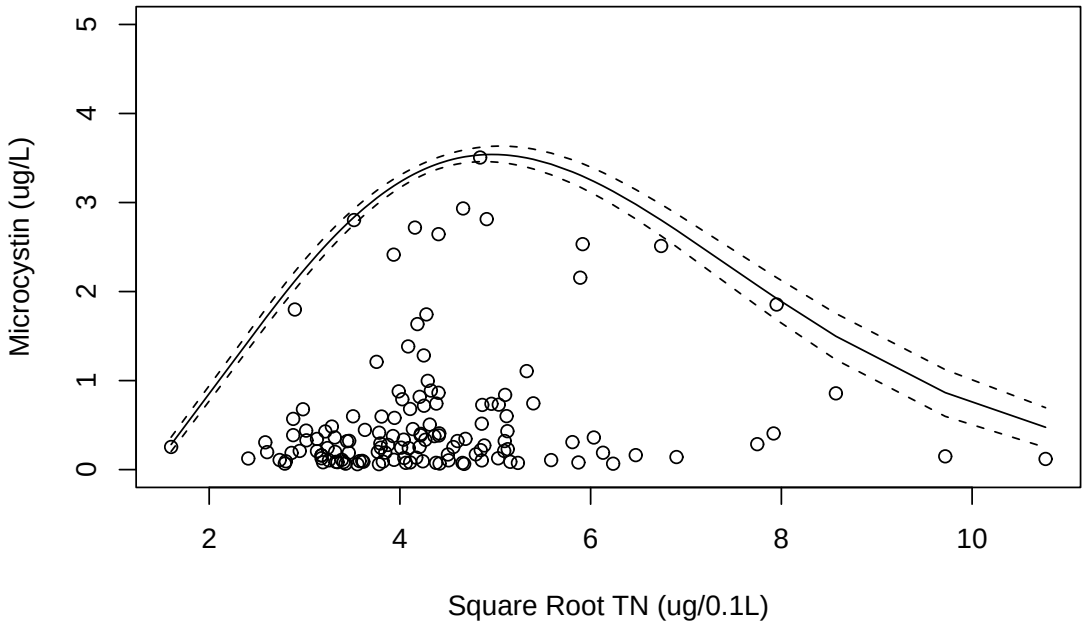

Figure 7.41: Actual data on abundance of Microcystin as a function of nitrogen concentration in midwestern lakes and reservoirs.

the ways that latent variables can be incorporated into models. As we will (hopefully) see in consideration of methods of estimation, latent variables may often be considered as "missing information", which leads to the need for either numerical methods for dealing with intractable integrals or estimation strategies such as the EM algorithm, or both.

7.5 Models Based on Stochastic Processes

The final topic in our tour of basic methods of model specification (Chapter 7) is models that are based on what are typically called *stochastic processes*. The most common models included in this category are time-series models, models of spatial processes, and models used in what is known as queuing theory. Each of these types of models are topics unto themselves (our department offers 3 courses on time series at different levels, 2 courses on spatial analysis at different levels, and queuing theory is covered in various places, notably in a service course offered to computer scientists). Thus, our objective is not to cover this topic in all of its variety but, rather, to communicate the basic concepts of stochastic processes and the manner in which they lead to certain statistical model formulations.

7.5.1 Restrictions in Statistical Models

Consider the simple model, for $i = 1, \ldots, n$,

$$Y_i = \mu + \sigma\,\epsilon_i; \quad \epsilon_i \sim iid\, N(0,1).$$

This model offers several restrictions on the distribution of the variables Y_i. Certainly there is a restriction to normal distributions, but even more we have specified that each random variable has the same expected value (μ) and the same variance (σ^2). Such restrictions serve a fundamental purpose, which is to give us multiple variables from the same *statistical population*. This is necessary for progress to be made, first of all in statistical abstraction and also if we have any hope of producing estimators with known properties. For example, contrast the above mode with the alternative

$$Y_i = \mu_i + \sigma_i \, \epsilon_i; \quad \epsilon_i \sim iid \, N(0,1).$$

What would we do with such a model without some restrictions or further modeling for the μ_i and/or σ_i? (not much would be a good answer).

We often place restrictions on distributions through the first two moments (mean and variance) by:

1. Specifying a constancy among a group of random variables, such as $\mu_i = \mu$ or $\sigma_i^2 = \sigma^2$.

2. Modeling unequal quantities as a function of covariates that depends on a small number of parameters, such as regression models for means or variance models such as those described in Section 7.2.4.

3. Modeling of means and/or variances as random variables that follow distributions with a small number of parameters, such as the hierarchical or mixture models described in Section 3.4.2.

7.5.2 Stochastic Processes and Random Fields

The world we encounter on the scale of typical human experience (i.e., forget special relativity for the moment) is 4−dimensional in nature. Three of these are spatial dimensions and the fourth is temporal. A stochastic process is a collection of random variables indexed by one or more of these dimensions, in which restrictions on distributional characteristics (usually means and variances) are also functions of the indices. Such collections of random variables are often called *stochastic processes* for indices in one dimension, and *random fields* for indices in more than one dimension, although technically these terms may be used interchangeably.

We will present a few of the basic models for processes in time and space, but first list two examples that lie outside of these realms to indicate that we are not covering all possibilities in detail.

1. Let $t = 1, 2, \ldots$ index discrete points in time, and define a random variable for each of these points as $Y(t)$. Suppose that $Y(t) \in \Omega_Y \equiv \{0, 1, 2, 3\}$. Here, the set of possible values Ω_Y is often called the *state space* of the process $\boldsymbol{Y} \equiv \{Y(t) : t = 1, 2, \ldots, \}$ since it is the set (i.e., space) of the possible states that the process can assume. Now, for $j, k \in \Omega_Y$, define values $t_{j,k}$ as,

$$t_{j,k} \equiv Pr[Y(t) = k \,|\, Y(t-1) = j].$$

The values $t_{j,k}$ may be collected in a matrix

$$T \equiv \begin{pmatrix} t_{0,0} & t_{0,1} & t_{0,2} & t_{0,3} \\ t_{1,0} & t_{1,1} & t_{1,2} & t_{1,3} \\ t_{2,0} & t_{2,1} & t_{2,2} & t_{2,3} \\ t_{3,0} & t_{3,1} & t_{3,2} & t_{3,3} \end{pmatrix},$$

such that $t_{j,k} \geq 0$ for all j, k, and $\sum_k t_{j,k} = 1$ for all j. The stochastic process $\boldsymbol{Y}$ is called a discrete Markov process (discrete because it is indexed at discrete points in time) that also has a discrete state space Ω_Y This *transition matrix* can be estimated from a finite sequence of observed values if suitable restrictions are placed on its component quantities (the $t_{j,k}$), generally as properties of the matrix T. Suitable restrictions include properties of T known as *irreducible*, *positive recurrent*, and *aperiodic*, which we will not go into further here.

2. Let $\{N(t) : t \geq 0\}$ represent the number of some type of events that have occurred between time 0 and time t. Suppose that,

   - $N(0) = 0$.
   - $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are independent for any disjoint intervals (s_1, t_1) and (s_2, t_2).
   - For any $s, t \geq 0$ and any $x \in \{0, 1, \ldots\}$,

   $$Pr[N(t+s) - N(s) = x] = \frac{1}{x!}(\lambda t)^x \exp(-\lambda t).$$

   Then $\{N(t) : t \geq 0\}$ is a continuous-time Poisson process (which has discrete state space). In this instance, all necessary restrictions are built into the definition of the process.

7.5.3 Stationarity

For stochastic processes, the types of restrictions made on means and variances of the process random variables (which are model assumptions) are often those necessary to produce *stationary* behavior in the model. There are actually

several types of stationarity. To introduce these concepts of stationarity, we first set forth some general notation appropriate for random fields.

Let $\boldsymbol{s}$ denote a (non-random) variable that contains information on a "location" in a system of spatial/temporal indices. For example, in a two-dimensional spatial process $\boldsymbol{s}$ might be defined as $\boldsymbol{s} \equiv (u, v)$ for longitude u and latitude v, or some transformation of latitude and longitude that makes (u, v) amenable to Euclidean geometry (e.g., universal transverse-mercator coordinates, or UTMs). In a one-dimensional time series, we may have $\boldsymbol{s} \equiv t$, where t is a point in time. Notice that we have taken this location variable $\boldsymbol{s}$ to be continuous in its time/space domain, which leads to a continuous process $\boldsymbol{Y} \equiv \{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D} \subset \Re^d\}$, where $\mathcal{D}$ is the domain of the process. We assume from the outset that $var\{Y(\boldsymbol{s}_i)\} < \infty$ for all $\boldsymbol{s}_i \in \mathcal{D}$.

1. First-Order Stationarity.

   The process $\boldsymbol{Y}$ is said to be *first-order* stationary if,

   $$E\{Y(\boldsymbol{s}_i)\} = \mu \qquad \forall \boldsymbol{s}_i \in \mathcal{D}$$

2. Intrinsic Stationarity.

   The process $\boldsymbol{Y}$ is said to be *intrinsically* stationary if it is first-order stationary and

   $$var\{Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)\} = V(\boldsymbol{s}_i - \boldsymbol{s}_j) \qquad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{D},$$

   for some function $V(\cdot)$. This concept of stationarity appears most often in spatial problems, but there is no reason it does not apply equally to temporal processes.

3. Second-Order Stationarity.

   The process $\boldsymbol{Y}$ is said to be *second-order* stationary if it is first-order

stationary and

$$cov\{Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\} = C(\boldsymbol{s}_i - \boldsymbol{s}_j) \quad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{D},$$

for some function $C(\cdot)$. While second-order and intrinsic stationarity are clearly related, they are not the same. In fact, it is possible to demonstrate the result that second-order stationarity implies intrinsic stationarity, but not the converse. Thus, second-order stationarity is a stronger condition than is intrinsic stationarity.

4. Strict Stationarity. Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$ be any finite collection of locations in $\mathcal{D}$. Define
$F_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m}(y_1, \ldots, y_m) \equiv$

$$Pr[Y(\boldsymbol{s}_1) \leq y_1, \ldots, Y(\boldsymbol{s}_m) \leq y_m].$$

The process $\boldsymbol{Y}$ is said to be *strictly* stationary if, for all $h \in \Re^d$ and all $m \geq 1$,

$$F_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m}(y_1, \ldots, y_m) = F_{\boldsymbol{s}_1+h,\ldots,\boldsymbol{s}_m+h}(y_1, \ldots, y_m)$$

Now, what do these various types of stationarity mean? Begin by considering a process in Re^1 such as a time series (could be a spatial transect, but time is adequate).

If a process in time is strictly stationary, then random variables for any set of times separated by a fixed distance have the same joint distribution. For example, $\{Y(1), Y(3), Y(10)\}$, $\{Y(6), Y(8), Y(15)\}$, and $\{Y(20), Y(22), Y(29)\}$ all have the same joint $3-$dimensional distribution. Similarly, $\{Y(4), Y(5)\}$ and $\{Y(150), Y(151)\}$ have the same $2-$dimensional joint distribution. This is, in fact, true for any number of random variables and any fixed difference in time. Clearly, this is a strong property. If a process in time is second-order

stationary, then random variables for any set of times separated by a fixed distance have the same first two moments, but not necessarily the same distribution. Thus, strict stationarity implies second-order stationarity, but not the converse. Note, however, that if we specify normal distributions then second-order stationarity does implies strict stationarity since normal distributions are characterized by the first two moments.

If a process in time is intrinsically stationary, then random variables for any set of times separated by a fixed distance have variances of their differences that are the same. For example, $var\{Y(1) - Y(3)\} = var\{Y(6) - Y(8)\} = var\{Y(20) - Y(22)\}$. What is the relation between this and second-order stationarity? For any two random variables X and Y, $var(X - Y) = var(X) + var(Y) - 2cov(X, Y)$, or, $cov(X, Y) = (1/2)\{var(X) + var(Y) - var(X - Y)\}$. It is then entirely possible for $var\{Y(\boldsymbol{s}_i), -Y(\boldsymbol{s}_j)\}$ to be a function of $\boldsymbol{s}_i - \boldsymbol{s}_j$ alone (intrinsic stationarity) without $cov\{Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\}$ also being such a function (second-order stationarity); in fact, this is guaranteed only if $var\{Y(\boldsymbol{s}_i)\} = var\{Y(\boldsymbol{s}_j)\}$. Thus, intrinsic stationarity is a weaker condition than is second-order stationarity. To generalize the above interpretations from $1-$dimensional real space to $d-$dimensional real space requires only replacing time differences with higher dimension displacement.

7.5.4 Two Fundamental Time Series Models

In this subsection we will consider stochastic processes for which we have a discrete set of indices indicating equally spaced points in time, $\boldsymbol{s}_i \equiv t$ for $t = 0, \pm 1, \pm 2, \ldots$. Two basic structures for temporal models, from which many more elaborate structures can be constructed (e.g., Newton, 1988; Shumway, 1988; Brockwell and Davis, 1991) are *moving average* and *autoregressive* mod-

els. In most texts on time series, these processes are presented in terms of a "backshift operator" $Bx_t = x_t - 1$. While this leads to compact notation and is useful in writing down results about these models, it does not necessarily promote understanding of the *structures* involved. Thus, we will avoid its use here, and present moving average and autoregressive models in a very simple form.

Moving Average Models

A basic moving average model formulation can be written as,

$$Y(t) = \sum_{k=0}^{q} \beta_k \, \epsilon(t-k), \tag{7.51}$$

in which we take $\beta_0 = 1$, and $\epsilon(t) \sim iid\, N(0, \sigma^2)$. Expression (7.51) would be said to represent a "qth order moving average model. To see the manner in which moving average models conceptualize a temporal process, consider a $2nd$ order process written as

$$Y(t) = \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2}. \tag{7.52}$$

In (7.52) we see clearly that the process at time t is composed of a linear combination of independent "error" or "innovation" terms. Clearly, the process has mean 0 for all t and is thus automatically first-order stationary. Although we will not write down an explicit form here, it should also be clear that the variances of $Y(t)$ and covariances of $Y(t)$, $Y(t+s)$, $s \leq q$, will be the variance of the innovation terms (σ^2) times a polynomial in the coefficients, the βs. For "lags" of greater than q, the covariance will be 0. Thus, the process is second-order stationary.

A major concern with time series models of this type (this will be true also for autoregressive models) is whether parameters of the model (7.51) can be "identified" based on the covariance function of the process. The answer, in general, is no we cannot do so. Consider an example taken from Newton (1988, p. 96), for a first order moving average process,

$$Y(t) = \epsilon_t + \beta\,\epsilon_{t-1},$$

which has first and second moments,

$$\begin{aligned} E\{Y(t)\} &= 0, \\ var\{Y(t)\} &= \sigma^2(1+\beta^2), \\ cov\{Y(t), Y(t+1)\} &= \beta\,\sigma^2, \\ cov\{Y(t), Y(t-1)\} &= \beta\,\sigma^2. \end{aligned}$$

Suppose that $var\{Y(t)\} = 5$ and $cov\{Y(t), Y(t+1)\} = 2$. Combinations of $\sigma^2 = 1$ with $\beta = 2$ or $\sigma^2 = 4$ with $\beta = 0.5$ both lead to these same moments. In general, for a moving average process of order q, there are 2^q sets of parameters that lead to the same variance and covariances. There is only one set of parameters, however, that lead to what is called an *invertible* model. We will leave this topic until after we have introduced autoregressive models.

Autoregressive Models

The basic form of an autoregressive model is,

$$Y(t) = \sum_{k=1}^{p} \alpha_k\, Y(t-k) + \epsilon(t), \tag{7.53}$$

where, as before, $\epsilon(t) \sim iid\, N(0, \sigma^2)$. Expression (7.53) is said to represent a p^{th} order autoregressive process. Consider, analogous to our $2nd$ order moving

average model (7.52) a $2nd$ order autoregressive model,

$$Y(t) = \alpha_1 Y(t-1) + \alpha_2 Y(t-2) + \epsilon(t). \tag{7.54}$$

Models (7.54) and (7.53) look like a linear regression of values of $Y(t)$ on previous values (which it is). Hence the name *autoregressive* which is more intuitive than the previous name of moving average for (7.51) or (7.52). The process of model (7.53) does not necessarily have constant mean, so it is usually assumed that the $Y(t)$ have constant mean zero. In practice, data are generally "detrended" through regression (over time) or the process of "first-differencing" to remove at least linear trends in means of the $Y(t)$.

Neither is it the case that an autoregressive process is necessarily second-order stationary unless some conditions are placed on the coefficients $\alpha_1, \ldots, \alpha_p$. This is also related to invertibility and will be covered shortly. The covariance of an autoregressive model can be determined using what are called the *Yule-Walker* equations (e.g., Shumway 1988, p.135). To illustrate the ideas involved, consider a first order autoregressive model,

$$Y(t) = \alpha\, Y(t-1) + \epsilon(t),$$

with $\epsilon(t) \sim iid\, N(0, \sigma^2)$. Directly from this model we have the inequalities

$$\begin{aligned}
Y(t)Y(t) &= \alpha Y(t-1)Y(t) + \epsilon(t)Y(t) \\
Y(t)Y(t-1) &= \alpha Y(t-1)Y(t-1) + \epsilon(t)Y(t-1) \\
&\vdots \quad\quad \vdots \\
Y(t)Y(t-k) &= Y(t-1)Y(t-k) + \epsilon(t)Y(t-k)
\end{aligned}$$

Since all $Y(t)$ are assumed to have expectation 0, taking expectations in these expressions yields,

$$\begin{aligned} var\{Y(t)\} &= \alpha cov\{Y(t-1),\, Y(t)\} + \sigma^2 \\ cov\{Y(t),\, Y(t-1)\} &= \alpha var\{Y(t-1)\} + 0 \\ &\cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \\ cov\{Y(t),\, Y(t-k)\} &= cov\{Y(t-1),\, Y(t-k)\} + 0 \end{aligned} \tag{7.55}$$

The final term on the rhs of the first equality comes from the fact that

$$E\{Y(t)\epsilon(t)\} = E\left[\alpha Y(t-1)\epsilon(t) + \{\epsilon(t)\}^2\right] = \sigma^2,$$

since $Y(t-1)$ contains only random variables that are independent of $\epsilon(t)$. This also immediately explains why the remaining right most terms are all 0.

Substituting the second line of (7.55) into the first, and assuming that we have an equal variance process gives an equality that allows derivation of the variance of the $Y(t)$,

$$\begin{aligned} var\{Y(t)\} &= \alpha^2 var\{Y(t-1)\} + \sigma^2 \\ var\{Y(t)\} &= \frac{\sigma^2}{1-\alpha^2} \end{aligned} \tag{7.56}$$

Substituting the first line of (7.55) into the second gives,

$$\begin{aligned} cov\{Y(t),\, Y(t-1)\} &= \alpha^2 cov\{Y(t),\, Y(t-1)\} + \alpha\sigma^2 \\ &= \frac{\alpha\,\sigma^2}{1-\alpha^2}. \end{aligned}$$

In fact, continuing in this manner shows that

$$cov\{Y(t),\, Y(t-k)\} = \frac{\alpha^k\,\sigma^2}{1-\alpha^2}. \tag{7.57}$$

Now, equation (7.57) indicates that, for increasing lag k and positive variances for sums, we must have $|\alpha| < 1$ in order for a legitimate model to exist (i.e., to have positive covariances). Given that this is true, the correlations between values of the process separated by a distance k is, (here allowing k to take values $0, \pm 1, \pm 2, \ldots$),

$$corr\{Y(t),\, Y(t-k)\} = \alpha^{|k|}$$

An important point here is the distinction between a moving average model of order one and an autoregressive model of order one. In the moving average model, covariance (and hence correlation) between values separated by more than one lag are 0, while in the autoregressive model the covariance (and hence the correlation) decays more slowly over time as a power function of the coefficient α (which must be smaller than 1 in absolute value). This same distinction extends to moving average and autoregressive models of higher orders.

To gain some intuition regarding this difference, consider an moving average model of order one and an autoregressive model of order one that are generated by the same innovation process. I know we are suppose to consider time point 0 as arbitrary, but suppose we have an actual starting point for the processes, and that the moving average parameter β is equal to the autoregressive parameter α; call this common parameter θ. As these moving average (MA) and autoregressive (AR) processes progress through time we obtain:

$$\begin{aligned} MA: Y(0) &= \epsilon(0) \\ AR: Y(0) &= \epsilon(0) \end{aligned}$$

$$MA: Y(1) = \epsilon(1) + \theta\,\epsilon(0)$$

$$\begin{aligned} AR: Y(1) &= \epsilon(1) + \theta\,\epsilon(0) \end{aligned}$$

$$\begin{aligned} MA: Y(2) &= \epsilon(2) + \theta\,\epsilon(1) \\ AR: Y(2) &= \theta Y(1) + \epsilon(2) \\ &= \epsilon(2) + \theta\epsilon(1) + \theta^2\,\epsilon(0) \end{aligned}$$

$$\begin{aligned} MA: Y(3) &= \epsilon(3) + \theta\,\epsilon(2) \\ AR: Y(3) &= \theta Y(2) + \epsilon(3) \\ &= \theta\{\theta Y(1) + \epsilon(2)\} + \epsilon(3) \\ &= \theta[\theta\{\theta\,\epsilon(0) + \epsilon(1)\} + \epsilon(2)] + \epsilon(3) \\ &= \epsilon(3) + \theta\,\epsilon(2) + \theta^2\,\epsilon(1) + \theta^3\,\epsilon(0) \end{aligned}$$

Deleting the intermediate steps for the AR processes makes the result more clear as:

$$\begin{aligned} MA: Y(0) &= \epsilon(0) \\ AR: Y(0) &= \epsilon(0) \end{aligned}$$

$$\begin{aligned} MA: Y(1) &= \epsilon(1) + \theta\,\epsilon(0) \\ AR: Y(1) &= \epsilon(1) + \theta\,\epsilon(0) \end{aligned}$$

$$\begin{aligned} MA: Y(2) &= \epsilon(2) + \theta\,\epsilon(1) \\ AR: Y(2) &= \epsilon(2) + \theta\,\epsilon(1) + \theta^2\,\epsilon(0) \end{aligned}$$

$$\begin{aligned} MA: Y(3) &= \epsilon(3) + \theta\,\epsilon(2) \end{aligned}$$

$$AR : Y(3) \quad = \quad \epsilon(3) + \theta\,\epsilon(2) + \theta^2\,\epsilon(1) + \theta^3\,\epsilon(0)$$

It is clear from this progression why the pairwise dependence of an autoregressive process lasts longer through time than does that of a moving average process; it is also more clear why the coefficient of an autoregressive process of order one needs to satisfy the condition of being less than 1 in absolute value.

Inversion

We have mentioned the issue called *inversion* previously, which deals with conditions under which there exists a duality between moving average and autoregressive processes, our primary interest begin situations in which both processes are stationary. Note that, in the progression used to contrast MA and AR models of order 1 in this section, we had the same coefficients of the MA and AR process, but the processes themselves were not the same (equivalent). Inversion concerns conditions under which the processes are the same. Box and Jenkins (1970, p. 50) point out that invertibility and stationarity are different properties. This is true, but it turns out that conditions needed to produce invertibility of stationary moving average processes are similar to those needed to produce stationarity in invertible autoregressive processes. We need to explain this remark, and do so in the following, although without the requisite derivations. For those see any of the references cited in this subsection (7.5.4).

A key quantity in determination of both invertibility and stationarity is what is called the *characteristic polynomial* or *characteristic equation* which, for finite moving average (order q) and finite autoregressive (order p) processes, and a possibly complex argument z are,

$$h(z) \quad = \quad 1 + \sum_{k=1}^{q} \beta_k\, z^k,$$

$$g(z) \;=\; 1 - \sum_{k=1}^{p} \alpha_k \, z^k. \tag{7.58}$$

Comments

1. In these notes, I have written both moving average (expression 7.51) and autoregressive (expression 7.53) models as sums rather than differences. It is also common (e.g., Box and Jenkins 1970; Shumway 1988) to write moving average processes with coefficients being the negative of those in (7.51), in which case the characteristic polynomials of (7.58) have the same form (same as given in $g(z)$); note also that Newton (1988) does neither of these, writing autoregressive models to isolate the error term $\epsilon(t)$ on the rhs of the model.

2. In consideration of a first order autoregression process we arrived at the need to have $|\alpha| < 1$ in order for covariances to remain finite and, in fact, then also stationary. For a general finite autoregressive process of order p, this condition is translated into conditions on the roots (zeros) of the characteristic polynomial $g(z)$, since these will be determined by values z^0 that are functions of the coefficients $\{\alpha_1, \ldots, \alpha_p\}$. Similarly, conditions on the coefficients of a finite moving average process of order q can be specified in terms of conditions on the roots of the characteristic polynomial $h(z)$.

3. The conditions that produce desired behaviors in finite moving average and autoregressive processes turn out to be conditions on whether the roots of the characteristic polynomials in (68) lie *inside*, *on*, or *outside* of the unit circle (i.e., less than, equal, or greater than 1 in modulus).

We can now summarize, without proof, what can be a confusing array of results regarding moving average and autoregressive time series models.

1. Finite moving average processes are always second order stationary.

2. Finite autoregressive processes are second order stationary if all of the zeros of $g(z)$ in (7.58) are greater than 1 in modulus.

3. Finite moving average processes can be written as infinite autoregressive processes if the zeros of $h(z)$ in (7.58) are all greater than 1 in modulus.

4. Finite autoregressive processes can be written as (doubly) infinite moving average processes as long as none of the zeros of $g(z)$ in (7.58) is equal to one in modulus. *Note: a doubly infinite moving average processes is as in (7.51) but with the summation going from $-\infty$ to ∞, (see, e.g., Fuller 1996).* In addition, finite autoregressive processes can be written as (singly) infinite moving average processes if all of the zeros of $g(z)$ in (7.58) are greater than 1 in modulus.

Given invertible models (moving average and autoregressive) the question in an application is which representation is more parsimonious (adequate, with as few parameters as possible). Clearly, if a process is truly a moving average model of order one, using an autoregressive representation is not desirable as the moving average model would have 2 parameters, β and σ^2, while the autoregressive model would have an infinite number of parameters. The reverse would be true for a process that was in fact an autoregressive process of order one. This leads, in time series analysis, to the representation of processes as a combination of moving average and autoregressive models. These are called autoregressive-moving average (ARMA) models, and have the general form,

$$\sum_{j=0}^{p} \alpha_j Y(t-j) = \sum_{k=0}^{q} \beta_k \epsilon(t-k),$$

with $\alpha_0 = \beta_0 = 1$. This looks more familiar if we write it as,

$$Y(t) = \sum_{j=1}^{p} \alpha_j Y(t-j) + \sum_{k=1}^{q} \beta_k \epsilon(t-k) + \epsilon(t),$$

Note that the coefficients α_j in these two expressions are negatives of each other.

Dependence on the Past

One final point relative to moving average and autoregressive time series models is worthy of note (it is, perhaps, even more important from a modeling standpoint than the topic of inversion). This concerns the manner in which moving average and autoregressive models represent the dependence of the current $Y(t)$ on past values. We have already seen, through consideration of first order processes, that moving average models have pairwise correlation of zero for lags greater than 1 (in general this will be for lags greater than the order of the model q), while the pairwise correlation for autoregressive models dies off more slowly (e.g., expression 7.57).

Now, consider the conditional distribution of $Y(t)$ given $Y(t-1) = y(t-1)$, for simplicity in the case of first order autoregressive and first order moving average models. For the autoregressive model we have,

$$Y(t) = \alpha\, y(t-1) + \epsilon(t),$$

which clearly demonstrates that, given a value $Y(t-1) = y(t-1)$, the distribution of $Y(t)$ does not depend on any previous values of the process. On the other hand, for a first order moving average model of the form $Y(t) = \beta\epsilon(t-1) + \epsilon(t)$,

$$\begin{aligned} Y(t) &= \beta\, \epsilon(t-1) + \epsilon(t) \\ &= \beta\, y(t-1) - \beta^2\, \epsilon(t-2) + \epsilon(t), \end{aligned}$$

so that, given $Y(t-1)$, $Y(t)$ is not independent of $Y(t-2)$, which is a function of $\epsilon(t-2)$ as $\epsilon(t-2) = Y(t-2) - \beta\epsilon(t-3)$.

How can this be? After all, we have shown that the covariance of $Y(t)$ and $Y(t-2)$ is 0 for a first order moving average process. How can they then be dependent? The answer lies in what a covariance matrix represents; we have been assuming normality throughout this presentation and will continue to do so. A covariance matrix represents *pairwise dependence* in a *joint distribution* of variables. In the case of normality, this also represents dependence in the *marginal* distributions of any subset of component variables. But, even in the case of normality, the covariance matrix *does not* represent *conditional dependence*. For normal distributions conditional dependence is given by the *inverse* covariance matrix.

One way to think of this is that is, in a Gaussian (or joint normal) distribution, the elements of the covariance matrix represents dependence when all other variables than the pair involved are *averaged* over. This is marginal dependence. On the other hand, the inverse covariance matrix represents dependence when all other variables than the pair involved are conditioned on. In a first order moving average model, *given* $Y(t-1)$, $Y(t)$ is dependent on $Y(t-2)$ because they are both related to (dependent on) $Y(t-1)$. But *marginally*, all of that shared dependence of $Y(t)$ and $Y(t-2)$ has been averaged out by integrating over possible values of $Y(t-1)$. In a first order autoregressive model, on the other hand, $Y(t)$ and $Y(t-2)$ are marginally dependent (under normality) but conditionally independent.

Another way to form some intuition about all of this is to consider the process of *generation through time* by which values of time series are produced (statistically). In a first order moving average model, it is innovation terms that are "propagated" through time to directly influence future variables. If we

would condition on these innovation terms, then there would be no conditional dependence beyond lag one, that is, $Y(t)$ given $\epsilon(t-1)$ would be independent of all previous terms (either Ys or ϵs). But, because these innovations correspond to latent (unobservable) variables, we condition on the actual value of the series at lag one $Y(t-1)$, which is composed partly of $\epsilon(t-1)$ but also partly of $\epsilon(t-2)$. Thus, conditioned on $Y(t-1)$ the current value $Y(t)$ is still dependent on $\epsilon(t-2)$ and, hence also $Y(t-2)$, and this extends to all previous terms $Y(t-3)$, $Y(t-4)$, etc. In a first order autoregressive model, it is the actual process variables (the Ys) that are directly propagated through time. Thus, all of the effect of previous ϵs is embodied in $Y(t-1)$ and, since this value directly affects $Y(t)$, conditioning on $Y(t-1)$ is the same as conditioning on all previous terms. That is, first order autoregressive models possess a first order Markov property in time (while p^{th} order autoregressive models are p^{th} order Markov in time).

Goals and Limitations of Traditional Time Series Models

We close our discussion of basic time series structures with a few thoughts on modeling issues related to these formulations. It should be noted, however, that the theory and application of time series forms a rich body of statistical knowledge and investigation, and there are many extensions of the basic ideas we have presented. Nevertheless, the following seem pertinent:

1. Everything we have considered has assumed zero mean processes (or constant mean processes in which the constant is easily removed by subtraction). In applications for which the mean is not constant, time series models are usually applied to series of residual quantities from which an estimated mean structure has been removed, or to a series of differenced

values (this results in what are often called autoregressive-integrated-moving average models (ARIMA models). The point is that time series modeling is focused on a signal plus noise conceptualization of scientific phenomena, in which the signal is removed and time series are used to describe remaining structure in the noise, or error, process.

2. Building on comment 1, the goals of time series analyses are generally those of data description, prediction, and forecasting (see Chapter 7.1) without an overly great concern for conceptualization of the scientific problem under study, other than what might be incorporated in modeling of the mean structure.

3. A practical limitation of the models presented in this section is the assumption of equally-spaced time points $t = 0, \pm 1, \pm 2, \ldots$. While there are certainly a good number of applications in which this assumption is met (the references given above contain a large number of examples and applications) there are also many problems that logically involve temporal dependence in which observations cannot be taken at equally spaced points in time, have a few missing values, or cannot be observed for even entire time periods (e.g., many environmental studies).

4. We have also, in our presentation of moving average and autoregressive models, assumed that innovation terms followed normal distributions. Following the same tradition as additive error models, time series models are most often presented without specific distributional assumptions, taking those terms to be simply random variables with expectation zero and constant variance. However, again in a manner similar to additive error models, there is an *implicit* appeal to normal distributions. We

have simply made this explicit in the preceding presentation.

7.5.5 Random Field Models

It is tempting, at this point, to discuss models for spatial processes as a distinct topic, but doing so would restrict the more general nature of models for *random fields*. In fact, everything covered in the previous subsection on time series models would fit under the heading of models for (discrete index) random fields and, although we did not cover this, there are models for processes with continuous time index that would fit under the heading of continuous index random fields.

Thus, time series models of the type discussed in Section 7.5.4 (which are called models in the *time domain*) are simply one class of models for random fields. It is also true that problems other than those involving only temporal and/or spatial dependence can be considered within the context of models for random fields. Despite this, however, spatial problems present a natural situation which which to illustrate random fields in more than one dimension, and this subsection will have a decidedly "spatial" flavor. In addition, much of the discussion of the topics included in this subsection will be found in the literature on spatial statistics (e.g., Haining 1990; Cressie 1993; Griffith and Layne 1999; Lawson 2001).

Recall from section 7.5.3 the notion of a (often non-random) "location" index $\boldsymbol{s}$. While such indices often contain variables that indicate position in a space/time structure, this is not entirely necessary. Consider, for example, a longitudinal study in which n_i observations are taken in time for each of k situations or "subjects". We can place response random variables in a random field structure by defining $\boldsymbol{s}_i \equiv (t_i, j)$ where t_i is the time of an observation

on a given subject j, where $j = 1, \ldots, k$. It is sufficient for this discussion of models for random fields, however, to generally take $\boldsymbol{s}_i \equiv (u_i, v_i)$ to denote a location in physical space, such as longitude u_i and latitude v_i.

We will, very briefly, present two random field structures which we will call *continuous index* random fields and *discrete index* random fields. A third type of structure, *point processes*, will not be discussed here. The primary distinction among these types of random field models rests on the type of location indices $\boldsymbol{s}$ assumed for each. A general structure is given in the following:

Let $\mathcal{D} \subset \Re^d$ be a subset of $\Re^d$ that has positive volume. We define a general random field process as,

$$\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}, \tag{7.59}$$

where, at this point, we are making no assumptions about the quantities $Y(\boldsymbol{s})$ and $\boldsymbol{s}$ (i.e., random or non-random).

Continuous Index Random Fields

In a continuous index random field, we take, in the general process (7.59), $\mathcal{D}$ to be a fixed subset of $\Re^d$, and allow $\boldsymbol{s}$ to vary continuously over $\mathcal{D}$. The response variables $Y(\boldsymbol{s})$ are taken as random. In general, $Y(\boldsymbol{s})$ may be multivariate, but we will consider only univariate processes. It is worthy of mention that, although the process itself is considered continuous, data will be available at only a discrete and finite number of locations in the form

$$\{y(\boldsymbol{s}_1), y(\boldsymbol{s}_2), \ldots, y(\boldsymbol{s}_n)\}.$$

In essence this means that, for a random field with no restrictions on means, variances, or distributional form, we have available only a partial realization

of a unique stochastic process (i.e., a sample of size less than one). This means then, from a modeling viewpoint, that restrictions on means, variances, or distributional forms are crucial in order to make progress in terms of estimation, inference and, even more fundamentally, statistical conceptualization or abstraction.

Consideration of dependence in a random field leads to a quantification of dependence that is an alternative to that of covariance – values of quantities that are closer together (in a random field distance) should be more similar than values of quantities that are farther apart. This concept of dependence is most easily understood if distance means geographic (i.e., spatial) distance, but a spatial setting is not necessary if a suitable metric can be defined for locations in a random field. The idea that things that are "closer" together should be more similar than things that are "farther" apart is directly represented in a quantity called the *variogram*, defined as a function which describes the variances of the differences of random variables at two locations,

$$2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \equiv var\{Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)\}; \quad \text{all } \boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{D}. \tag{7.60}$$

In (7.60) the function $2\gamma(\cdot)$ is called the variogram. Just as functions must satisfy certain conditions to be density or mass functions (≥ 0 and integrate to one), and matrices must satisfy certain conditions to be covariance matrices (positive or at least non-negative definite), so too must a function $2\gamma(\cdot)$ satisfy certain conditions to be a variogram. This condition is called *conditional negative definiteness*, and is defined as follows.

Definition

A variogram $2\gamma(\cdot)$ is conditionally negative definite if, for any finite number of locations $\{\boldsymbol{s}_i : i = 1, \ldots, m\}$ and real numbers $\{a_i : i = 1, \ldots, m\}$ such that

$\sum a_i = 0$,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j 2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0. \tag{7.61}$$

That is, any function $2\gamma(\cdot)$ that satisfies (7.60) must have the property (7.61). Note, at this point, that we have made no assumptions regarding $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ in terms of mean (i.e., first order stationarity) or variance (i.e., constant variance). Note also, however, that by writing a variogram as a function of displacement between locations (i.e., the argument to $s\gamma(\cdot)$ in 7.60 is the displacement $\boldsymbol{s}_i - \boldsymbol{s}_j$) we have, in fact, assumed intrinsic stationarity if it has constant mean.

An additional assumption is often made about the variogram, which may be checked by a data-driven model assessment, that the value of the variogram for two locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ depends only on the distance between, $d_{i,j} \equiv ||\boldsymbol{s}_i - \boldsymbol{s}_j||$. For example, if $\boldsymbol{s}_i \equiv (u_i, v_i)$ for a horizontal position u_i and vertical position v_i, Euclidean distance would be $d_{i,j} = \{(u_i - u_j)^2 + (v_i - v_j)^2\}^{1/2}$. In this case, we may write the variogram (70) as a function of only a distance $h \in \Re^1$ as,

$$2\gamma(h) = var\{Y(\boldsymbol{s} + \boldsymbol{w}) - Y(\boldsymbol{s})\}; \quad \boldsymbol{s} \in \mathcal{D}, \tag{7.62}$$

for any $\boldsymbol{w}$ such that $||\boldsymbol{s} - \boldsymbol{w}|| = h$. If a continuous index random field process has a variogram that satisfies (7.62) then we say the process is *isotropic*.

While we will not go into details here, the goal in application of a continuous index random field model is often prediction. Forecasting is possible, but becomes more tenuous than in time series because assumptions are being made that the form of a process remains similar for observations beyond the scope of the data in more than one dimension. Forecasting based on data description but not understanding is, in my opinion, dangerous even in one dimension (such as time series) but this danger is compounded when a lack of understanding is extended to more than one dimension of our physical world.

The general progression of developing a predictor for unobserved locations is as follows:

1. The variogram $2\gamma(\cdot)$ is estimated from a given set of data (this is most often a method of moments estimator) for a finite set of displacements (usually distances, under an assumption of isotropy).

2. A theoretical model that ensures conditional negative definiteness is fit to the estimated variogram values (much like a nonlinear regression).

3. A linear predictor to minimize prediction mean squared error is developed as a function of observed data and values of the variogram, which are now taken from the model fitted in step 2.

4. Uncertainty in predictions may be derived by substituting the linear predictor into the prediction mean squared which it minimizes and assuming normality.

In the area called *Geostatistics* such a prediction system is known as *kriging*. For a full development of kriging under various assumptions on the process (e.g., constant versus nonconstant mean, variances, etc.) see Cressie (1993).

As a final comment on continuous index random field models, note that nothing in the above development has assumed constant variance for the process $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$. A variogram model, combined with a model for variances yields a covariance model since,

$$\begin{aligned} cov\{Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\} &= var\{Y(\boldsymbol{s}_i)\} + var\{Y(\boldsymbol{s}_j)\} \\ &- var\{Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)\}. \end{aligned}$$

For example, under an assumptions of constant variance σ^2, and an isotropic variogram $2\gamma(h)$, a covariance model for an isotropic, second order stationary

process is,

$$2\,C(h) = 2\,C(0) - 2\,\gamma(h).$$

Thus, modeling the variogram plus the variance leads to a model for the covariance, but not necessarily the other way around.

Discrete Index Random Fields

To formulate a discrete index random field we take, in the general process of (7.59), $\mathcal{D}$ to be a fixed set of countable (usually finite) points, so that we have random field locations $\boldsymbol{s}_i \in \mathcal{D}$ for a discrete set of locations $\{\boldsymbol{s}_i : i = 1, \ldots\}$. As for continuous index processes, $Y(\boldsymbol{s}_i)$ are considered random variables.

In our (very) brief consideration of time series models, dependence was represented through the manner in which past values (autoregressive models) or past innovations (moving average models) where propagated through a process in a forward manner (there, through time). Models for continuous index random fields allowed an alternative formulation of dependence as a variogram model. Discrete index random fields present yet another vehicle by which dependence can be incorporated into a model, through what is called *neighborhood* structure in a *Markov random field*. It has become common to equate models for Markov random fields with what are called *conditionally specified* models. But it should be noted that models for Markov random fields are not a necessary consequence of discrete index random fields, nor are conditionally specified models a necessary consequence of Markov random fields. In fact, discrete index time series are discrete index random fields, some models for which (e.g., first order autoregressive models) possess a Markov property. Thus, one could consider a first order autoregressive model a model for a Markov random field, although it would not typically be considered a

conditionally specified model (it would, at best, constitute a highly restricted conditional specification).

First, we define what is meant by a *neighborhood* structure. The *neighborhood* of a location $\boldsymbol{s}_i$ is that set of locations N_i such that the random variable $Y(\boldsymbol{s}_i)$ is believed to "depend on" the variables at locations contained in N_i. For the moment we leave what is meant by "depend on" vague. This will be determined by the manner in which dependence is allowed to enter a particular model. Neighborhoods are usually determined by other than statistical considerations and are often, in fact, rather arbitrary in nature. Determination of appropriate neighborhoods is a difficult problem in the application of the types of models to be considered below. Examples of neighborhoods for spatial problems with $\boldsymbol{s}_i \equiv (u_i, v_i)$ include:

1. If $\boldsymbol{s}_i$ denotes the center of a county or other political subdivision, we might define N_i to include those counties that share a common border with the county of which $\boldsymbol{s}_i$ is the centroid.

2. If the values $u_i : i = 1, \ldots, C$ and $v_i : i = 1, \ldots, R$ denote equally spaced vertices on a regular grid, we might specify that $N_i \equiv \{(u_j, v_j) : u_j = u_i \pm \delta,\ v_j = v_i \pm \delta\}$, where δ is the grid spacing. This is known as a "four nearest neighbors" structure.

3. For locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ that are either uniformly or non-uniformly distributed in space, we might define $N_i \equiv \{\boldsymbol{s}_j : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| \leq \kappa\}$ for some predetermined distance κ.

Simultaneous Autoregressive Model

If specified neighborhoods are used to define dependence parameters $\{b_{i,j} :$

$i, j = 1, \ldots, n\}$, the following model has sometimes been called a *simultaneous autoregressive model* (SAR):

$$Y(\boldsymbol{s}_i) = \mu_i + \sum_{j=1}^{n} b_{i,j}\{Y(\boldsymbol{s}_j) - \mu_j\} + \epsilon_i, \tag{7.63}$$

where (usually), $\epsilon_i \sim iid\, N(0,\, \sigma^2)$, and $b_{i,i} = 0$, for all i. Model (73) is called an autoregressive model because of the structure of $Y(\boldsymbol{s}_i)$ "regressed" on values of $Y(\boldsymbol{s}_j)$, but it does *not* share properties with the time series autoregressive model (63). For one thing, as shown in Cressie (1993, p. 406), the error ϵ_i is not independent of the $\{Y(\boldsymbol{s}_j) : j \neq i\}$.

Markov Random Fields

We are familiar with the standard Markov assumption in time (given the entire past, the present depends on only the most immediate past). What does this imply for a joint distribution of variables on a one-dimensional random field (time or otherwise)? Let $\{Y_1, Y_2, \ldots, Y_n\}$ denote ordered random variables (e.g., time or a spatial transect). Let $p(\cdot)$ denote a generic probability density or mass function corresponding to its arguments. That is, $p(y)$ is the marginal density of Y, $p(y_1, y_2)$ is the joint density of Y_1 and Y_2, $p(y_1|y_2)$ is the conditional density of Y_1 given Y_2, and so forth. Then, as is always true,

$$p(y_1, \ldots, y_n) = p(y_1)p(y_2|y_1) \ldots, p(y_n|y_1, \ldots, y_{n-1})$$

which, by the Markov property, becomes

$$p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2|y_1) \ldots, p(y_n|y_{n-1}).$$

Note also that the Markov property implies that

$$p(y_i|\{y_j : j = 1, \ldots, i-2\}) = p(y_i|y_{i-2})$$

Now,

$p(y_i|\{y_j : j \neq i\})$

$$
\begin{aligned}
&= \frac{p(y_1, \ldots, y_n)}{p(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)} \\
&= \frac{p(y_1)p(y_2|y_1)\ldots p(y_{i-1}|y_{i-2})p(y_i|y_{i-1})}{p(y_1)p(y_2|y_1)\ldots p(xyi-1|y_{i-2})} \\
&\times \frac{p(y_{i+1}|y_i)p(y_{i+2}|y_{i+1})\ldots p(y_n|y_{n-1})}{p(y_{i+1}|y_{i-1})p(y_{i+2}|y_{i+1})\ldots p(y_n|y_{n-1})} \\
&= \frac{p(y_i|y_{i-1})p(y_{i+1}|y_i)}{p(y_{i+1}|y_{i-1})} \\
&= \frac{p(y_i|y_{i-1})p(y_{i+1}|y_i, y_{i-1})}{p(y_{i+1}|y_{i-1})} \\
&= \frac{p(y_i, y_{i-1})p(y_{i-1}, y_i, y_{i+1})p(y_{i-1})}{p(y_{i-1})p(y_{i-1}, y_i)p(y_{i-1}, y_{i+1})} \\
&= \frac{p(y_{i-1}, y_i, y_{i+1})}{p(y_{i-1}, y_{i+1})} \\
&= p(y_i|y_{i-1}, y_{i+1})
\end{aligned}
$$

Thus, the typical Markov property in one dimension (e.g., time) implies that the conditional distribution of Y_i given all other variables depends on only the adjacent values Y_{i-1} and Y_{i+1}.

It is the structure of such *full conditional* distributions that are the concern of Markov random fields. A collection of random variables $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ are said to constitute a Markov random field if, for each $i = 1, \ldots, n$,

the conditional density or mass functions satisfy,

$$p(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}) = p(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in N_i\}), \tag{7.64}$$

where $\{N_i : i = 1, \ldots, n\}$ are neighborhood structures.

Conditional Model Specification

While the definition of Markov random fields just given indicates that neighborhoods are determined by conditional distributions, in model formulation we typically want to reverse this progression by starting with defined neighborhoods and then using these to write forms for conditional density or mass functions. This method of formulating models is called *conditional model specification* and owes a great deal to early work by Besag (1974). In this method of modeling, we specify forms for the n full conditionals making use of neighborhoods as in (7.64) to reduce the complexity of these distributions. The key for modeling is to ensure that a joint distribution exists that has the corresponding set of specified conditionals. The key for statistical analysis is to identify this joint and make use of it in statistical estimation and inference procedures. This becomes a long and involved topic, beyond the scope of our class. See Besag (1974) for models based on one-parameter exponential families. Kaiser and Cressie (2000) relax some of the assumptions of Besag, and extend this modeling approach to multiple parameter exponential family conditionals. Kaiser (2001) gives an introductory overview of this approach to modeling. Arnold, Castillo and Sarabia (1992) provide some characterization results, especially for bivariate settings. Probably the most common conditional model in applications is formed from conditional normal distributions and is often called simply the *conditional autoregressive model*.

Let $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ be a set of n random variables with (spatial) locations $\boldsymbol{s}_i$; $i = 1, \ldots, n$, and let the full conditional densities of these random variables be given by,

$$f_i(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}) =$$

$$\frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left[\frac{-1}{2\tau_i^2}\left\{y(\boldsymbol{s}_i) - \mu(\{y(\boldsymbol{s}_j) : j \neq i\})\right\}^2\right], \tag{7.65}$$

where

$$\begin{aligned} \mu(\{y(\boldsymbol{s}_j) : j \neq i\} &\equiv E\left[Y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}\right] \\ \tau_i^2 &\equiv var\left[Y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}\right] \end{aligned}$$

Now, further model

$$\mu(\{y(\boldsymbol{s}_j) : j \neq i\}) = \alpha_i + \sum_{j=1}^{n} c_{i,j}\left\{y(\boldsymbol{s}_j) - \alpha_j\right\}, \tag{7.66}$$

subject to the conditions that $c_{i,j}\tau_j^2 = c_{j,i}\tau_i^2$, $c_{i,j} = 0$; $i, j = 1, \ldots, n$, and $c_{i,j} = 0$ unless $\boldsymbol{s}_j \in N_i$, the neighborhood of $\boldsymbol{s}_i$. It is common to take $\tau_i = \tau$ in this model so that the condition on the $c_{i,j}$ reduces to $c_{i,j} = c_{j,i}$.

Let C denote the $n \times n$ matrix with ij^th element $c_{i,j}$ and M the $n \times n$ matrix with diagonal elements τ_i^2; for constant conditional variance $M = \tau^2 I_n$ where I_n is the $n \times n$ identity matrix. As shown by Besag (1974) and Cressie (1993), the joint distribution of $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$ is then

$$\boldsymbol{Y} \sim Gau(\boldsymbol{\alpha}, (I_n - C)^{-1}M), \tag{7.67}$$

provided that the $n \times n$ matrix $(I_n - C)$ is invertible and the $n \times n$ matrix $(I_n - C)^{-1}M$ is positive definite. In (54) $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_n)^T$.

7.5.6 An Application to Modeling Nitrates in the Des Moines River

We present in this subsection an application of a model based on stochastic processes. This application illustrates, in particular, the flexibility allowed by conditional model specification in representing dependence structures. The problem involves modeling nitrate concentrations in a river monitoring network. What is known as the Des Moines River Water Quality Network was instituted for the purpose of monitoring water quality in the Des Moines River prior to impoundment of Saylorville Reservoir. The upper map of Figure 7.42 shows the portion of the State of Iowa involved, which is the Des Moines River from Boone to Pella. Since the original creation of the network, the number and location of monitoring stations has changed, but for the period of time involved in this example the network consisted of seven stations along about a 116 mile stretch of the river. These stations are indicated in the schematic map in the lower panel of Figure 7.42. The triangles in this picture represent two reservoirs, Saylorville Reservoir upstream (upper left) and Red Rock Reservoir downstream (lower right). Two of the monitoring stations (stations 2 and 6) are located in these reservoirs just above the dam sites. Samples from the monitoring stations are assessed for over 100 water quality variables, but our focus will be on the analysis of nitrate/nitrite concentrations. The data available for our analysis consist of $2,954$ observations between late November 1981 and December 1996. The time between observations is roughly two weeks, but the data record contains irregular spacings in time as well as more than a negligible number of missing observaitons at particular stations for various sampling dates. Given the large number of observations, a smaller data set consisting of 938 observations from November 1981 through December 1985 was created

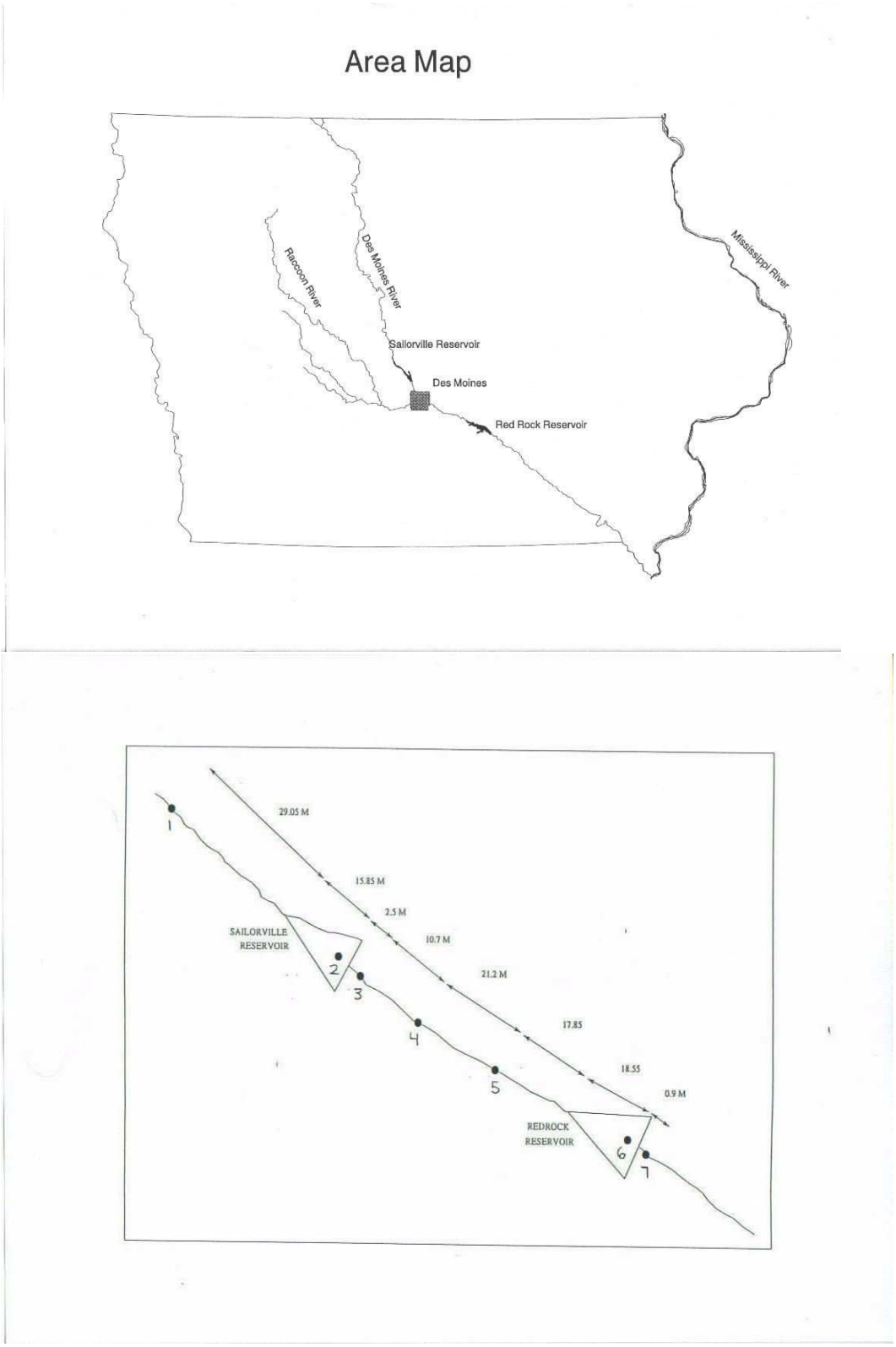

Figure 7.42: Maps of Des Moines River Quality Network

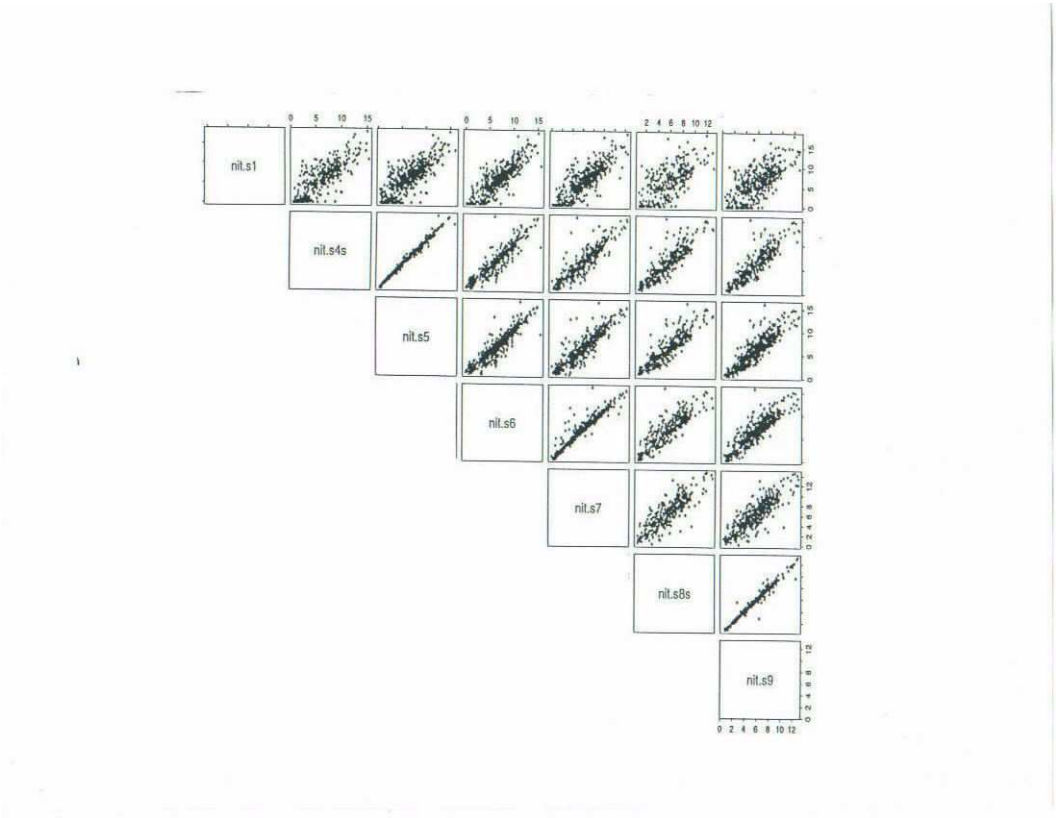

Figure 7.43: Scatterplot matrix for the Des Moines River nitrate data.

for model development. The final model structure chosen was then used to fit the entire data set of almost $3,000$ values.

Data Exploration and Modeling Concepts

We begin by making use of some simple exploratory data analysis conducted with the reduced data set, including examination of scatterplots and basic sample statistics. Figure 7.43 presents a scatterplot matrix of values for the seven monitoring stations. There appear to be fairly strong pairwise correlations among values from the stations. This is verified by examination of the

sample partial and marginal correlations, which are presented in the following table: (marginal below diagonal, partial above diagonal):

Site	1	2	3	4	5	6	7
1	1	-.32	.34	.48	.20	-.12	.05
2	.75	1	.94	.18	.19	-.07	.15
3	.77	.99	1	.00	-.15	.14	-.16
4	.84	.93	.93	1	.33	-.05	.14
5	.79	.88	.87	.91	1	.12	-.05
6	.63	.85	.84	.82	.80	1	.88
7	.64	.85	.84	.83	.80	.97	1

The marginal correlations are substantial, and appear to become weaker for stations separated by greater distances (e.g., correlation between stations 4 and 5 is 0.91 but this drops to 0.82 between stations 4 and 6). Note also that the correlations between reservoir stations and the corresponding nearest upstream station are weaker than most of the other correlations between adjacent stations. The correlation bewteen station 1 and station 2 is 0.75, while that between stations 5 and 6 is 0.80, the two weakest values for adjacent stations. This does not appear to be true for correlations of reservoir stations with the nearest downstream station, which is not surprising given the configuration in the schematic diagram of Figure 7.42. The partial correlations given in the above table do not appear to be entirely consistent. This would make application of, for example, a covariance selection model (Dempster, 1972) difficult to formulate. The strong marginal correlations suggest, of course, that a full multivariate model (Gaussian if that appears appropriate) would be possible to fit to these data (this is, in fact done later in the example). But, our objective is to model the network as a whole making use of structure to describe the

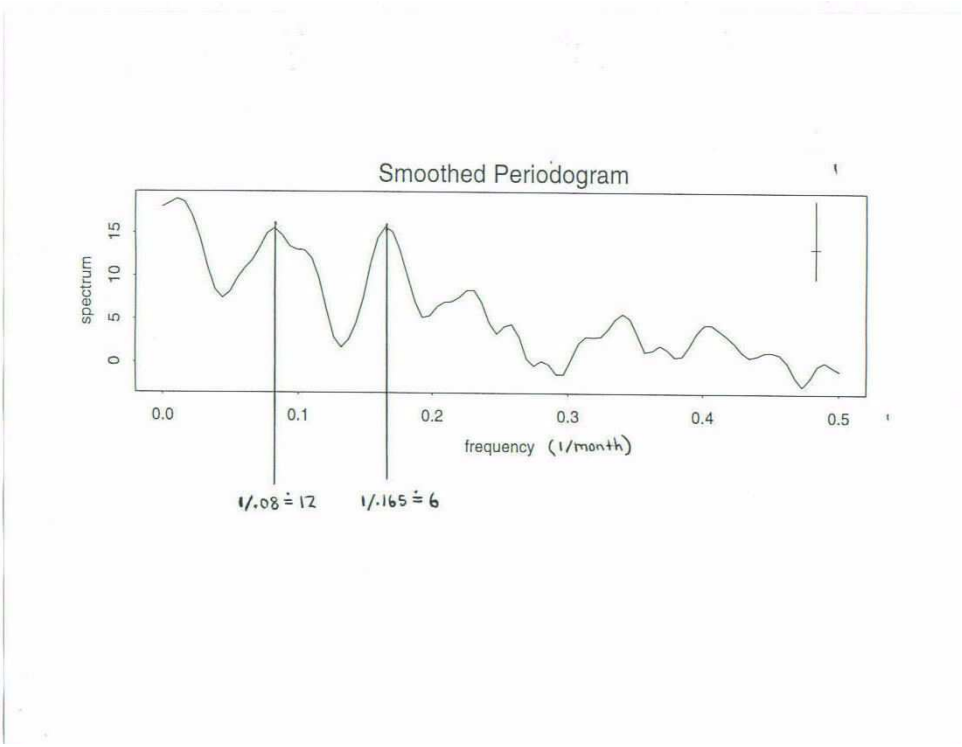

Figure 7.44: Sample periodogram for the Des Moines River nitrate data.

relations among values for the different stations. A full multivariate Gaussian model would not accomplish this goal, although we might expect that it would describe the data quite well.

Given that observations are available over a long time span (3 years even in the reduced data set), another exploratory tool is examination of a periodogram, presented in Figure 7.44 for one of the riverine stations, station number 4. This periodogram suggests the presence of both 12 and 6 month cycles in the values. The investigators who collected these data have indicated that the six month cycle is expected based on the hydrology of the region, but the same was not true of the 12 month cycle. Similar periodograms resulted for the other stations, although the 12 month cycle was not always identified

as a major feature for every station.

We might consider a number of modeling structures to represent these data. Some of our options include:

1. A regression model over time in which all values (among or within stations) are taken to be independent.

2. Time series structures for each station individually.

3. A spatial structure for all stations, with no explicit structure for temporal dependence.

4. A multivariate structure across stations with no explicit structure for temporal dependence.

5. A stochastic structure including terms for both spatial and temporal dependencies.

The first option listed above constitutes something of a "straw man" model, given the correlations among stations and periodic structure over time identified in initial examination of the data. The fourth option results in a large model with many unconstrained parameters, and fails to provide much in the way of a parsimonious conceptualization of the structure that might underlie the processes involved in producing the observed data patterns. As mentioned previously, however, the initial examination of of the data suggests that such a model might be expected to represent the data quite well, and will be used in the sequel as a "brick man" model (in contrast to the straw man model of opiton 1).

All of the other options listed (numbers 2, 3, and 5) are viable modeling approaches. Option 2, consisting of individual time series for different monitoring stations, fails to provide an overall structure for the entire monitoring

network, unless the series at the various stations appear to be identical. The unequal spacing of observation and presence of numerous missing values might also result in difficulties for this type of model. Thus, we will not consider the time series approach further. Of the two remaining options, the last results in the most complex model structure and introduces problems in combining spatial and temporal dependencies that are beyond the scope of this example. In addition, the roughly two week separation of observations represents what might be considered a substantial time gap for there to remain temporal dependence in the data. Nevertheless, it would, in future consideration of this problem, be interesting to see if a model with both spatial and temporal structures offers improvement to the models presented in what follows.

Having determined that a reasonable approach is to fit a model containing spatial structure but lacking explicit terms for temporal dependencies, we must consider how we will formulate the various portions of the model. In particular, questions center on what effects will be included in an expectation function (or systematic model component), what effects will be included in a dependence structure, and what distributional form will be used. A typical approach, common in time series applications, would be to remove all structure possible as systematic trend, even if that structure is not well understood from a scientific viewpoint. Any remaining structure is then modeled through dependence. This approach is essentially a "signal plus noise" formulation, with signal taken as any patterns in the data that can be modeled through an expectation function, and noise taken to be correlated error terms. We will make use of an alternative approach here, to illustrate the flexibility of conditional model specification. Under this alternative approach, we consider the underlying scientific process of interest to consist of factors that are *external* to the observed process (i.e., the data record of nitrate concentrations) and fac-

tors that are *internal* to that process. Overall, we seem to simply have replace the words "trend" and "dependence" with "external" and "internal" and to some extent this is true, but with an important impact on the development of our model. The basic guidelines that emerge may be summarized as follows.

1. External factors should be modeled as systematic trend, but only external causes which are understood or at least scientifically plausible should be included. Here, this implies that a six month cycle should be included in the systematic model component but not a twelve month cycle.

2. Internal factors and remaining influences that are not understood should be modeled through dependence structure. Dependence structure should be modeled according to basic properties of the situation under investigation.

These prescriptions, while still somewhat vague, are intended to provide some guidance for making necessary modeling decisions. It is well understood that, in stochastic process models, there is no unique way to decompose trend and dependence (or large-scale and small-scale structure, or external and internal processes). As Cressie (1993, page 114) states "... one person's deterministic mean structure may be another person's correlated error structure". If the sole or primary objective of an analysis is prediction of unobserved values this may cause no serious difficulties. But if, as is the case here, a primary goal is a meaningful conceptualization of the scientific processes under study, this issue should be given careful consideration.

What are the "basic properties" of this situation, which are meant to guide our modeling of dependence in this problem? To re-phrase this question, what are the basic reasons we might expect spatial dependence, as exhibited in the data? The presence of systematic trends that are common to all monitoring

stations should produce correlations among values. Beyond this, however, values might be correlated essentially because of physical transport. Put very plainly, nitrate concentrations at stations along a river might be correlated because water flows downhill. While there certainly must be other processes occurring, such as chemical reactions, binding of particles to the sediment, and so forth, these other (which I would consider internal) factors of the overall process contribute uncertainty and would be difficult to give mathematical form in a model. The upshot of this discussion is that we will attempt to model the six month cycle of hydrology through systematic trend, and other influences through spatial dependence based on the concept that it is transport that is the underlying cause of dependence.

Definition of Random Field

To define random variables appropriate for use in this problem, let $\boldsymbol{s}_i \equiv (\ell, t)$ where ℓ is station no. (1-7) and t is Julian date, beginning with $t = 1$ as 24 November, 1981. The $\boldsymbol{s}_i$ are thus nonrandom variables that indicate the "location" of each random variable, where location is taken to mean a position in a random field, not necessarily a physical location. Here, location consists of a combination of geographic location (station) and a time of observation. Let

$$\begin{aligned} \boldsymbol{Y} &\equiv \{Y(\boldsymbol{s}_i) : i = 1, \ldots, N\} \\ &= \{Y(\ell, t) : \ell = 1, \ldots, 7;\ t = 1 \ldots, T\} \end{aligned}$$

Drawing on the fundamental idea that these random variables may fail to be independent because of a transport process in the river, we can define neigh-

borhoods by making a standard Markov assumption, but in terms of water flowing down a river rather than time. That is, we will assume that the distribution of nitrates at a station, conditional on all stations upstream depends only on the closest upstream station. The implication of this assumption is that the neighbors of location $\boldsymbol{s}_i$ are, for $i = 1, \ldots, 7$,

$$N_i \equiv \{\boldsymbol{s}_j : \boldsymbol{s}_j \in \{(\ell - 1, t), (\ell + 1, t)\}; \;\; i = 1, \ldots, N$$

Then,

$$[Y(\boldsymbol{s}_i) \mid \{Y(\boldsymbol{s}_j) : j \neq i\}] = [Y(\boldsymbol{s}_i) \mid \boldsymbol{Y}(N_i)]$$

Conditional Distributions

To formulate our model we need to specify functional forms for the full conditional distributions $[Y(\boldsymbol{s}_i) \mid \boldsymbol{Y}(N_i)]$. It is, in general, difficult to determine appropriate data-driven procedures on which to base distributional choice. In this case, we first fit a regression containing a six month cycle to data from each station, and then produced conditional histograms of residuals, as presented in Figure 7.45 for Station 7. The conditioning used was to bin neighboring residuals into categories. For example, the upper left histogram of Figure 7.45 contains residuals for Station 7 given that its only neighbor, Station 6, had residual values between -7 and -2. The reamining histograms of Figure 7.45 were similary produced for various (non-mutually exclusive) bins. Overall, the histograms of Figure 7.45 appear fairly symmetric, although they are not all centered at zero, which is expected since these are conditional distributions. The histograms of Figure 7.45 and similar graphs for other stations provides at least some evidence to support a choice of conditional Gaussian distributions on which to base a model. Thus, for $i = 1, \ldots, N$ let $Y(\boldsymbol{s}_i)$ have conditional

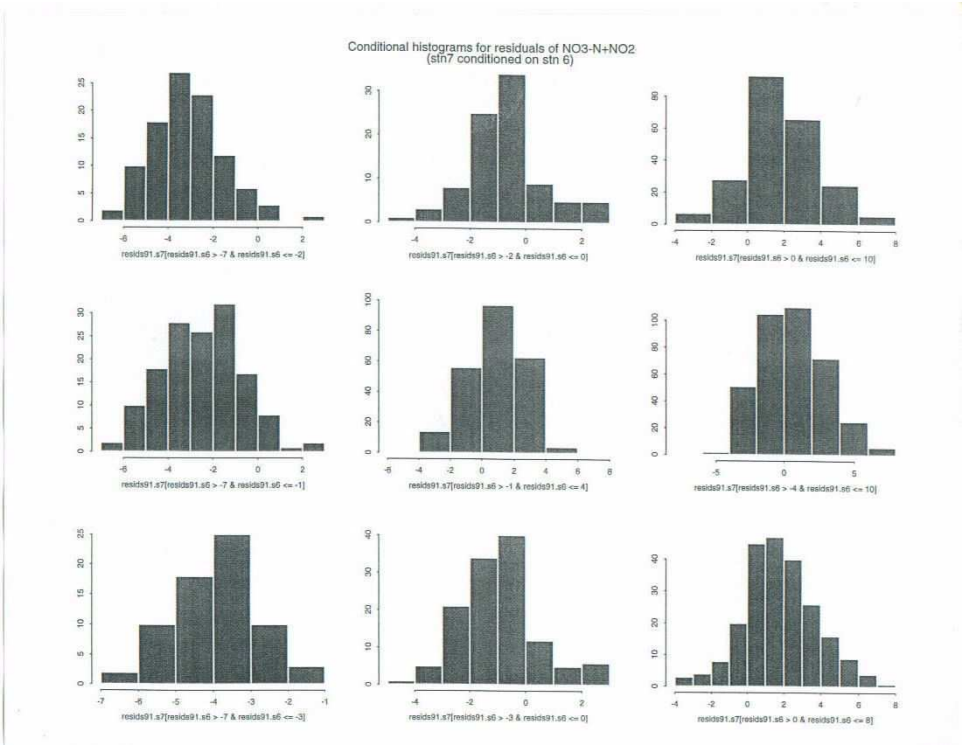

Figure 7.45: Conditional histograms of regression residuals

density

$$f(y(\boldsymbol{s}_i) \mid \boldsymbol{y}(N_i)) = \text{ Gau } (\mu_i, \tau^2)$$

where

$$\mu_i = \theta_i + \sum_{j \in N_i} c_{i,j} \ (y(\boldsymbol{s}_j) - \theta_j) \tag{7.68}$$

subject to $c_{i,j} = c_{j,i}$. The above distributions specify what has been called a conditional autoregressive model in the portion of Section 7.5.5 on Discrete Index Random Fields; see, in particular, expressions (7.65) through (7.67). We will formulate several models having this general form by specifying different forms for the θ_i and $c_{i,j}$.

Joint Distribution

As indicated in Section 7.5.5 this conditional Gaussians model also has a Gaussian joint distribution which may be written as,

$$\boldsymbol{Y} \sim \text{ Gau } \left(\boldsymbol{\theta} ; \ (I - C)^{-1} I\tau^2 \right) \tag{7.69}$$

where

$\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_N)^T$

$C \equiv [c_{i,j}]_{N \times N}$ and $c_{i,j} = 0$ if $j \notin N_i$. Note that the leading constants in the conditional means of expression (7.68) are also the marginal means. The conditional means then become an additive combination of the marginal means and departures of neighboring values from their marginal means. The $c_{i,j}$ in (7.68) are incorporated in the matrix C of the joint distribution which is, essentially, the inverse covariance matrix of the distribution. That the joint distribution for a conditional Gaussians model is available in closed form is, while not unique to Gaussian conditionals, the exception rather than the rule in conditionally specified models. Having this joint distribution in closed form

greatly simplifies estimation, and least for the maximum likelihood approach we will take here.

What remains in model formulation is to give explicit form to the leading θ_i and $c_{i,j}$ of (7.68). In doing so we wish to reduce the number of free parameters in the model and to provide a mathematical conceptualization (i.e., model) that adheres to the modeling guidelines developed earlier. To accomplish this, we return to the ideas of factors that are external and internal to the process under investigation.

Model for External Factors

As mentioned above, it is the leading θ_i in (7.68) that become marginal expectations, and thus it is in these model terms that we wish to incorporate identified external influences. In this application we have identified only one factor for which sufficient scientific knowledge is available to incorporate into the systematic model component, that being the six month hydrological cycle. Therefore, model the θ_i in (7.68) as,

$$\theta_i = \beta_0 + \beta_1 \sin\left(\frac{t\,\pi}{91}\right) + \beta_2 \cos\left(\frac{t\,\pi}{91}\right) \tag{7.70}$$

In (7.70) the terms $(t\pi/91)$ determine a six month cycle, and both sine and cosine functions are included to deal with phase shifts at the beginning of the data record. The regression conducted to produce residuals for construction of conditional histograms used (7.70) as the (linear) expectation function.

Models for Dependence Structure

All models fit to the data made use of (7.70) to describe the marginal expectations over time. Several different models were then formulated through various

specifications for the dependence terms $c_{i,j}$ in the conditional means of (7.68).

1. Distance Model
$$c_{i,j} \equiv \eta \left\{ \frac{\min\{d_{i,j}\}}{d_{i,j}} \right\}^k$$
where $d_{i,j}$ is distance (river miles) between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. In the distance model, dependence is modeled as a function of physical or geographic distance between monitoring stations. This is a simple formulation corresponding to the previous notion that dependence is related to the transport of materials along a river.

2. Flow Model
$$c_{i,j} \equiv \eta \left\{ \frac{\min\{f_{i,j}(t)\}}{f_{i,j}(t)} \right\}^k$$
where $f_{i,j}(t)$ is time (hours) for water to travel between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ (both downstream and upstream) at time (Julian date) t

The flow model is an extension of the basic concept embodied in the distance model. That is, if the distance model represents the fact that water flows downhill, then the flow model represents the fact that water flows downhill but not always at the same speed.

Calculations of the $f_{i,j}(t)$ in the flow model is a complicated process that will not be explained in detail here. Suffice it to say that these variables depend on measurements of river discharge at time t, the cross-sectional profiles of the river at the different monitoring stations, and reservoir inflow, outflow and pool elevation, which are also time-dependent factors. Computation of these flow calcualtions involved three distinct cases:

1. Two river stations

   Station 3 to Station 4 and Station 4 to Station 5

2. Upstream river station, downstream reservoir station Station 1 to Station 2 and Station 5 to Station 6

3. Upstream reservoir station, downstream river station Station 2 to Station 3 and Station 6 to Station 7

Straw and Brick Man Models

Recall from our previous discussion of potential model structures the two options identified as straw man and brick man models. The straw man model was specified as,

$$\boldsymbol{Y} \sim \text{ Gau } (\boldsymbol{\theta}, \tau^2 I) \tag{7.71}$$

$$\theta_i = \beta_0 + \beta_1 \sin\left(\frac{t\,\pi}{91}\right) + \beta_2 \cos\left(\frac{t\,\pi}{91}\right),$$

which contains 4 parameters, β_0, β_1, β_2, and τ^2. Note that, while the regression function parameters have the same interpretation in this model as in the conditionally specified distance and flow models, τ^2 here is the marginal variance of each random variable as well as the conditional variance (which are the same here). To specify the brick man model, let $Y_t \equiv (Y(l,t) : l = 1, \ldots, 7)^T$ with Y_t independent for $t = 1, \ldots, T$. Then take,

$$\boldsymbol{Y} \sim \text{ Gau } (\boldsymbol{\mu}, \Sigma) \tag{7.72}$$

with $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_7)^T$, and unconstrained Σ other than the requirement that it be a legitimate covariance matrix (symmetric, positive definite). The brick man model contains 35 parameters, the 3 regression coefficients β_0, β_1 and β_2 as in all of the models, and 31 parameters in the covariance matrix Σ.

The names of straw and brick man models are attached to (7.71) and (7.72) because of the ease with which their abilities to fit the data should be

surpassed. That is, the straw man model with 4 parameters and assumed independence should be relatively easy to improve upon (the straw man is easy to knock over) if there is any dependence exhibited in the data (and we know there is). On the other hand, the brick man model with 35 parameters should be difficult to improve upon (the brick man is difficult to knock over). Model performance will be assessed using the mean squared difference between predicted (or fitted) values and observed values for each location in the random field. One might (correctly, as it turns out) assume that the straw man model will have the worst performance by this measure, and the brick man model the best performance. An assessment of the spatial distance and flow models is how close to the brick man model the mean squared error can be "moved" from the straw man model while adding as few parameters as possible.

Results – 1982 to 1985 Data

N = 938

Maximum likelihood estimates for parameters in the straw man, distance, and flow models are presented in the table below.

	Estimated Value					
Model	β_0	β_1	β_2	τ^2	η	k
Straw Man	6.83	1.40	2.15	6.68		
Distance	6.77	1.69	2.81	2.49	0.42	0.25
Flow	6.75	1.70	2.78	2.50	0.40	0.01

Notice that the parameters of the systematic model component (7.70) are quite similar for all three models. While this is "proof" of nothing, and our

models have not overcome the problem of non-unique decomposition into trend and dependence that all models share, this is reassuring for our modeling concept. That is, the systematic model component (7.70) which we developed by considering external influences that are reasonably understood does seem to represent effects that can be consistently estimated from the data, even under different structures for dependence. This is not necessarily the case in many situations, particularly with models that attempt to model as much structure as possible through the expectation. Maximized log likelihood values for these models are given in the following table.

Model	No. Parameters	Log Likelihood
Straw Man	4	−2022.8035
Distance	5	−1730.7696
Flow	5	−1730.6467

While we should hesitate to think of any type of likelihood ratio test in this situation, these values seem to indicate a substantial improvement of the distance and flow models over the independence (straw man) model, as expected.

We turn now to our assessment based on mean squared errors for the various models. Note that dates of observation containing missing values had to be eliminated from the available data to allow the brick man model to be estimated. This is why that model was not included in the previous model comparisons, but average square error should be at least roughly comparable.

Station	Straw Man	Distance	Flow	Brick Man*
1	9.37	6.59	5.67	3.86
2	6.35	2.23	1.79	0.12
3	8.10	2.41	1.36	0.14
4	7.32	3.19	1.37	0.85
5	6.54	2.78	2.03	1.61
6	3.08	0.88	0.95	0.35
7	4.94	1.83	1.51	0.36
TOTAL	6.53	2.90	2.10	1.04

* Only 623 observations used

If we use mean squared errors as a scalar quantification of model performance, we can determine the "distance" between the straw man and brick man models as $6.53 - 1.04 = 5.49$. For the distance and flow models, the percentages of this distance "moved" with the addition of only one parameter to the model were:

Model	Percent of Distance Moved	Added Parameters
Distance	66.1	1
Flow	80.7	1

Plots of predicted versus actual data for two monitoring stations for both the independence and flow models are presented in Figure 4.46. Station 2 is located near the dam of Saylorville Reservoir, while station 4 has two neighbors that are both riverine locations (see Figure 7.42). The most noticeable feature of these plots is that the spatial flow model has managed to pick up a great deal more detail in the series than the straw man model.

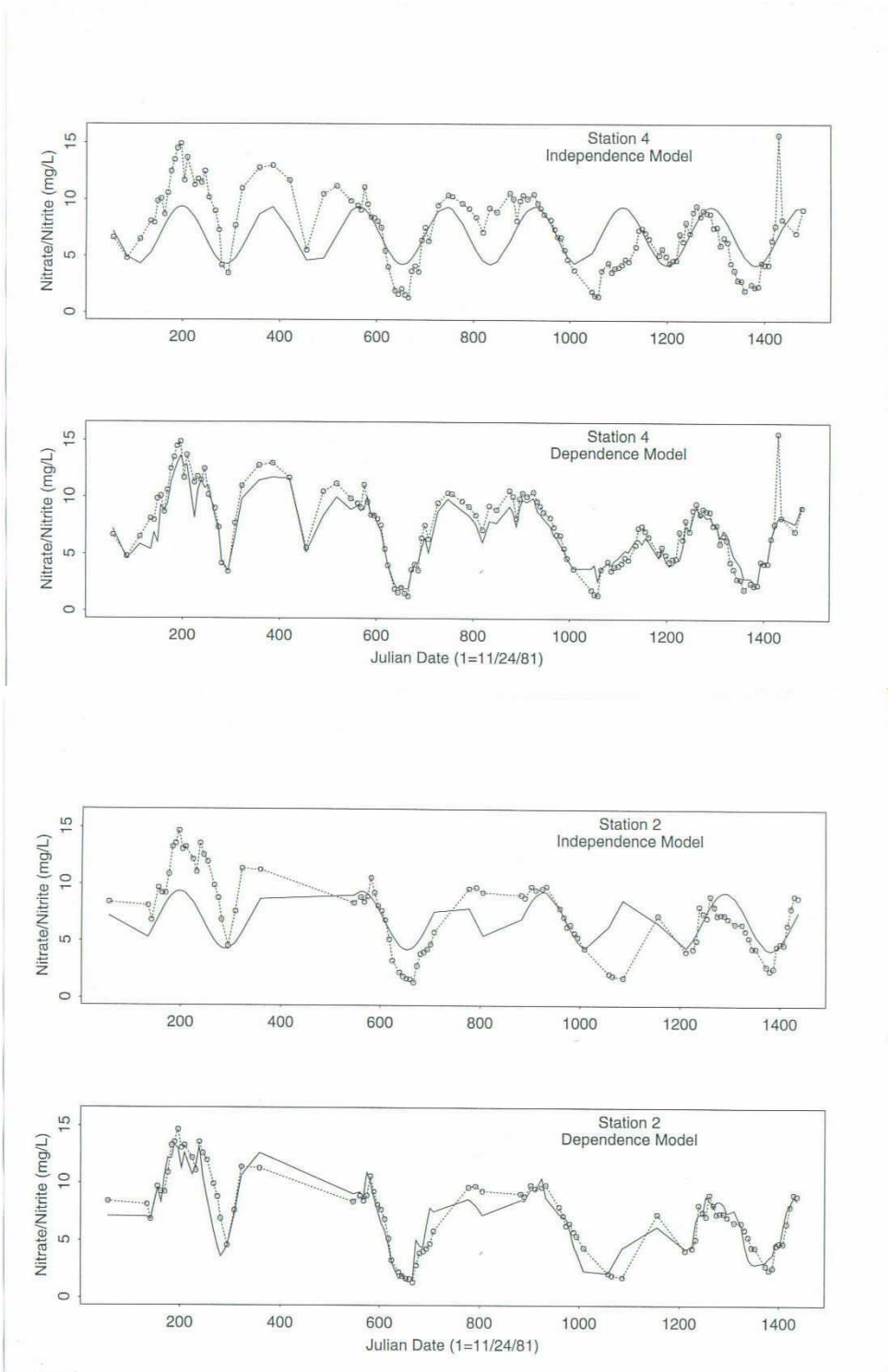

Figure 7.46: Fitted values for straw man and flow models for Station 2 and Station 4 using the reduced data set.

Results – 1982 to 1996 Data

The straw man model and the two spatial models (distance and flow) were also fit to the entire record of 2,954 available data values. Maximum likelihood estimates of model parameters are given in the table below.

	Estimated Value					
Model	β_0	β_1	β_2	τ^2	η	k
Straw Man	6.20	1.20	1.26	8.71		
Distance	6.16	1.23	1.36	4.38	0.35	0.05
Flow	6.24	1.23	1.37	4.00	0.48	-0.04

The same consistency in estimation of expectation function parameters observed in the reduced data set is also exhibited here, although the values have changed somewhat between the two data sets, most notably the values estimated for β_2. The values of the dependence parameter η are remarkably simlar for the two data sets, although note that the fixed "tuning parameter" k, which influences interpretation of η differs. We have not said anything about this k parameter, and will delay discussion of how these values were chosen until Chapter 8 (one form of profile likelihood).

The maximized log likelihood values that resulted from the model fits are given in the table below.

Model	No. Parameters	Log Likelihood
Straw Man	4	-6756.10
Distance	5	-6010.94
Flow	5	-5936.71

Here, in contrast to results for the reduced data set, the flow model appears to have increased the log likelihood substantially over the distance model as well

as both spatial models over the straw man model (we are not yet prepared to consider what might be meant by a "substantial" increase at this point in the course).

Finally, plots of fitted values for straw man and flow models for Station 3 are presented in Figure 7.47, similar to those of Figure 7.46 using the reduced data set. The traces for Station 3 were chosen because this station exhibited what would appear to be an "intervention", resulting in a major dramatic shock to the series of values at about Julian date 3000. The dependence contained in the flow model was able to pick up this event, without a need to include an arbitrary shift parameter in an expectation function (arbitrary because the cause is unknown).

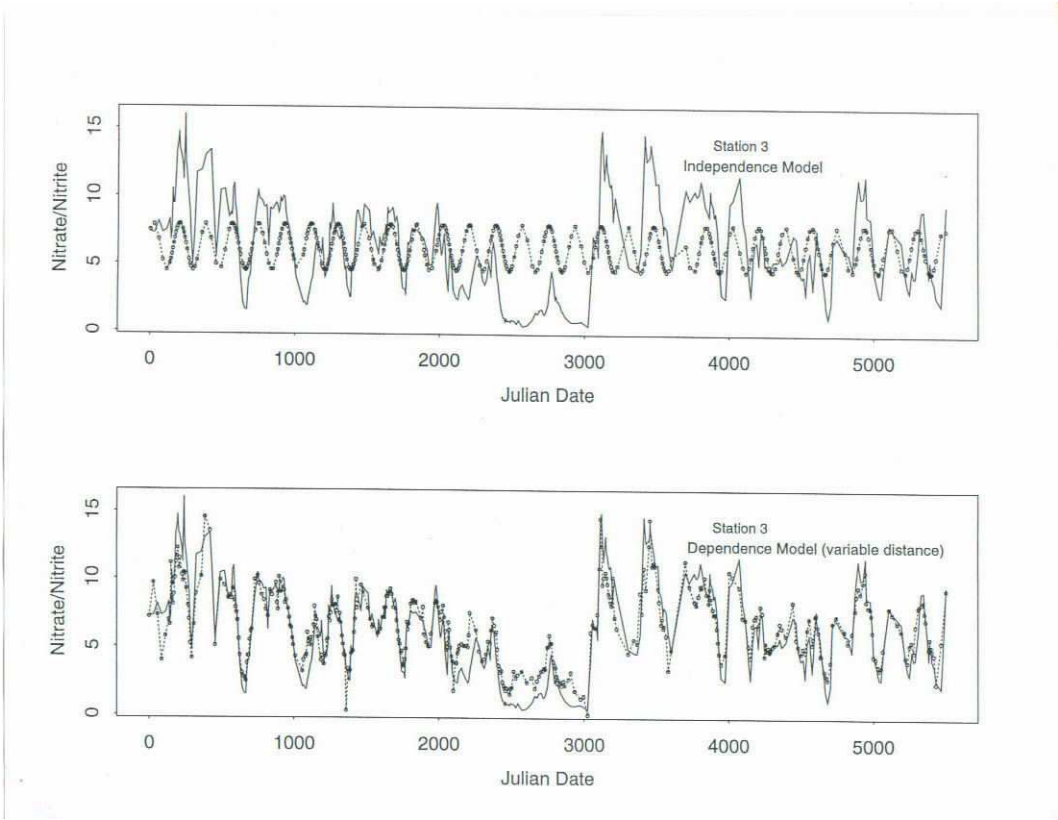

Figure 7.47: Fitted values from the flow model for Station 3 over the entire data record

Chapter 8

Methods for Estimation and Inference

We have discussed (at some length) basic methods of specifying statistical models, and a variety of implications of these methods for statistical abstraction. We now turn our attention to methods used in estimation of parameters of various models, along with procedures for quantification and estimation of uncertainty in those estimates, as well as the issue of model selection. Pertinent introductory comments include the manner in which we will approach *exact* or *small sample* theory versus *approximate* or *large sample* theory, and the basic modes used for making inferential statements about parameters or models.

First, let us recognize that, in parametric modeling, inferential statements can only be made on the basis of properties of statistical quantities, often the estimators of parameters, although sometimes the likelihood function itself. In small sample theory the primary properties of interest are bias (or the lack thereof), efficiency or minimum variance, and distributional form. In

large sample theory we have consistency, asymptotic bias (or lack thereof), asymptotic efficiency, and asymptotic distribution. Other properties that are less frequently considered include equivariance and general minimum risk (e.g., Lehmann 1983). Small sample properties are extremely desirable as they dictate statistical behavior of estimators for any sample size. The availability of small sample properties is also the exception, rather than the rule, for statistical estimators (despite the fact that we often spend a great deal of time on such properties in theory courses). Thus, while we will point out situations in which the methods presented in this section possess small sample properties, we will not dwell on such properties. You have already seen in previous courses most of the situations under which small sample properties are available, which consist primarily of linear models with constant variance additive errors and *iid* samples from exponential families.

Secondly, our approach to inference will consist primarily of two modes, interval estimation of parameters and model selection procedures. Thus, we will deal little with traditional formulations of "hypothesis" tests or "significance" tests, although we may revisit both of these in Part 4 of the course which will deal with issues in inference. Thus, we are by design (although not necessarily choice – we simply cannot cover everything in one course) largely excluding an entire body of statistical theory called *decision theory* from consideration, at least as relevant to issues of statistical testing. Bits and pieces of this theory may, however, find their way into our discussion.

8.1 Estimators Based on Sample Moments

In our discussion of Sampling and Survey Methodology in Part 1 of the course we defined population quantities (for physically existing populations of discrete

units) as averages over population units of functions of attributes. In defining basic estimators under this approach we then simply replaced averages over population units with averages over sample units. Estimation based on sample moments embodies this same basic concept. The population average and variance have, in particular, been replaced with expectation and variance for theoretical distributions, but these quantities turn out to be functions of model parameters. It is natural, then, to consider estimating such functions of parameters using sample moments and, sometimes, solving the resultant expressions for the individual parameter values themselves.

Estimators based on sample moments are used in (at least) the following circumstances:

1. To obtain initial "guesses" for estimators that will be adjusted to produce small sample properties (primarily unbiasedness and/or minimum variance).

2. To obtain estimators of model parameters in situations for which efficiency is not a crucial concern.

3. To obtain starting values for iterative estimation procedures that produce maximum likelihood estimates.

4. To obtain consistent estimators of variance parameters for use with asymptotic properties of parameters involved in expectation functions.

Note that these uses of estimators based on sample moments are not mutually exclusive and that the third and fourth of these circumstances are actually contained in the second. We will thus deal separately with the situation of item 1 above, and the remaining situations.

8.1.1 Sample Moments as Launching Pads for Optimal Estimators

You either are, or will soon become through other courses, aware of the derivation of estimators possessing some type of optimality property. Although not exhaustive of the possibilities, by far the most common criterion for a definition of "optimal" is minimization of expected squared error loss. For a parameter θ, collection of random variables $\boldsymbol{Y}$ having a distribution that depends on θ, and estimator of θ $\delta(\boldsymbol{Y})$, we define the loss function $L\{\theta, \delta(\boldsymbol{Y})\}$. It is important for deriving optimal properties of estimators that this loss function is restricted to have the two properties that,

$$\begin{aligned} L\{\theta, \delta(\boldsymbol{y})\} &\geq 0 \text{ for all } \theta \text{ and possible values } \delta(\boldsymbol{y}) \\ L\{\theta, \theta\} &= 0 \text{ for all } \theta. \end{aligned}$$

The squared error loss function is,

$$L\{\theta, \delta(\boldsymbol{Y})\} = \{\theta - \delta(\boldsymbol{Y})\}^2,$$

and the corresponding expected loss or *risk* becomes,

$$E[L\{\theta, \delta(\boldsymbol{Y})\}] = E[\{\theta - \delta(\boldsymbol{Y})\}^2]. \tag{8.1}$$

A fair amount of statistical theory is devoted to finding estimators $\delta(\boldsymbol{Y})$ that minimize (8.1) for all possible values of θ, so-called *minimum risk* estimators. The usual progression is to point out that such uniform minimum risk estimators do not, in general, exist unless we restrict attention to particular classes of estimators. The usual class considered is that of unbiased estimators for which

$$E\{\delta(\boldsymbol{Y})\} = \theta,$$

in which case minimization of (8.1) becomes the same as minimization of the variance of $\delta(\boldsymbol{Y})$. If such estimators exist for a given model they are then called *uniform minimum variance unbiased* (UMVU) estimators. Other situations in which minimum risk estimators may sometimes exist include the class of *equivariant* estimators and we sometimes refer to *minimum risk equivaraint* (MRE) estimators (see, e.g., Lehmann 1983).

Among the most important messages from consideration of optimal estimators for statistical modeling are the following:

1. Restricting the classes of estimators considered (e.g., to unbiased estimators) is typically not sufficient to guarantee that a uniform minimum risk, or minimum variance, estimator exists, yet alone indicate how we might find one if it does.

2. In situations for which both unbiased estimators and sufficient statistics are available for a parameter, the Rao-Blackwell theorem (e.g., Lehmann 1983, p.50) provides the key for determination of an optimal estimator. Under the additional restriction that the model under consideration provides a *complete* sufficient statistic, an application of the Rao-Blackwell theorem provides the unique UMVU estimator (this is actually true under any strictly convex loss, not just squared error).

3. The set of situations under which we have parameters for which unbiased estimators exist and such that the model provides complete sufficient statistics are relatively limited. Complete sufficient statistics are mostly constrained to exponential family distributions (at least if we include "curved" exponential families).

4. Given that a UMVU or MRE estimator can be shown to exist and can

be found (i.e., derived) we must still face the problem of determining what its variance or risk is if it is to be useful in inferential procedures. It seems to be the case that the calculation of exact variances is often quite complex, if it can even be achieved (e.g., Lehmann 1983, p. 106).

5. Given the historical emphasis in statistical modeling on additive error models, the connection of such models with location scale families of distributions, the fact that the normal distribution is one of only two location families (i.e., normal with known variance) that are also exponential families (the other being the distribution of the log of a gamma random variable), the fact that the normal may also be written as an exponential dispersion family, and the fact that the variance of normal theory estimators may often be expressed in explicit formulae, all of the above go a long way toward explaining why the normal distribution has achieved its high status in both statistical theory and methods.

6. Two caveats to the above indication that optimal small sample theory is often applicable to models based on normal distributions are that this also usually requires constant variance, and that functions of UMVU estimators are *not*, in general, UMVU for the same function of parameters. Linear transformations of UMVU estimators are UMVU since the expected value of a linear function of random variables is the same linear function of the expectations, but this does not hold for nonlinear transformations (Jensen's Inequality).

Before moving on we give two simple examples based on models with normal error distributions and constant variance to illustrate the use of sample moments in developing estimators with optimal small sample properties.

Example 8.1

Consider the simple one-sample normal model,

$$Y_i = \mu + \sigma\,\epsilon_i,$$

where, for $i = 1, \ldots, n$, $\epsilon_i \sim iid\, N(0,1)$. Consider estimation of μ and σ^2 using the sample moments $\bar{Y} = (1/n)\sum Y_i$ and $S_*^2 = (1/n)\sum(Y_i - \bar{Y})^2$ In this model Y_i follows a two parameter exponential family with complete sufficient statistic $(Y_i,\, Y_i^2)^T$. Using the property of exponential families that the joint distribution of a sample is an exponential family with the same natural parameters and complete sufficient statistics given by sums of the corresponding statistics for individual random variables (see Section 6.1.4), the joint distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is also normal with complete sufficient statistic $(\sum Y_i,\, \sum Y_i^2)^T$. Now, as we well know, the sample mean $\bar{Y}$ is unbiased for μ. Since this estimator is unbiased and a function of the complete sufficient statistic, it is UMVU for μ. The second (central) sample moment S_*^2 is not unbiased for σ^2, but a simple correction yields the usual sample variance $S^2 = (1/(n-1))\sum(Y_i - \bar{Y})^2$ which is unbiased for σ^2 and is a function of the complete sufficient statistic and, hence, is also UMVU. As an illustration of the point that functions of UMVU estimators are not generally also UMVU, note that $S = \{S^2\}^{1/2}$ is not UMVU for $\sigma = \{\sigma^2\}^{1/2}$.

Example 8.2

Consider now a linear regression model with constant variance,

$$Y_i = \boldsymbol{x}_i^T\,\beta + \sigma\,\epsilon_i,$$

where the $\epsilon_i \sim iid\, N(0,1)$ for $i = 1, \ldots, n$, and $\boldsymbol{x}_i^T = (x_{1,i}, \ldots, x_{p,i})$. The distribution of the response variables $\{Y_i : i = 1, \ldots, n\}$ in this model, $Y_i \sim indep\, N(\boldsymbol{x}_i^T\beta, \sigma^2)$ may be written in exponential family form with complete sufficient statistics $\sum Y_i^2$ and, for $k = 1, \ldots, p$, $\sum Y_i x_{k,i}$. The ordinary least squares estimators for β are unbiased and functions of the complete sufficient statistics and are thus UMVU. Similarly, $S^2 = \sum\{(Y_i - \boldsymbol{x}_i^T\hat{\beta})^2\}/(n-p)$ is unbiased for σ^2 and thus also UMVU. This example perhaps extends a bit beyond the use of sample moments as initial estimators from which to develop optimal properties, although least squares estimators are sometimes referred to as moment-based estimators (e.g., Lindsey 1996, p. 123).

8.1.2 Method of Moments Estimators

What is called the *method of moments*, usually attributed to Pearson (1948), is often presented as a "preliminary" or "crude" method for the derivation of estimators. This is because, as a method of estimation *per se*, method of moment estimators are not efficient in either exact or asymptotic senses. Nevertheless, we still employ the fundamental concept of method of moments estimation to obtain either starting values for iterative numerical methods for maximum likelihood estimation or, perhaps more importantly, to obtain consistent estimators of variances in a model as discussed in the previous subsection.

The basic concept behind the method of moments is that, in a situation involving *iid* random variables, estimators of a set of model parameters $\{\theta_k : k = 1, \ldots, s\}$ may be obtained by equating the theoretical moments with the first s sample moments $(1/n)\sum(Y_i^k)$; $k = 1, \ldots, s$, and solving the resultant set of equations for the parameter values. This concept is most easily grasped through an example.

Example 8.3 A One Sample Gamma Model

Suppose that we have a gamma model for a set of *iid* random variables. That is, for $i = 1, \ldots, n$, the density of Y_i is,

$$f(y_i|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} \exp\{-\beta\, y_i\}; \quad 0 < y_i.$$

From this model, we know that the first two moments of each Y_i are

$$\begin{aligned} E(Y_i) &= \frac{\alpha}{\beta}, \\ E(Y_i^2) &= \frac{\alpha(\alpha+1)}{\beta^2}. \end{aligned}$$

Let the first two sample moments be denoted as,

$$\begin{aligned} a_1 &= \frac{1}{n}\sum_{i=1}^{n} Y_i, \\ a_2 &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2. \end{aligned}$$

Then method of moment estimators for α and β are values that solve,

$$\begin{aligned} a_1 &= \frac{\alpha}{\beta}, \\ a_2 &= \frac{\alpha(\alpha+1)}{\beta^2}, \end{aligned}$$

which results in,

$$\hat{\alpha} = \frac{a_1^2}{a_2 - a_1},$$

$$\hat{\beta} = \frac{a_1}{a_2 - a_1}.$$

Now, we could make use of asymptotic results for sample moments under *iid* conditions (e.g., Serfling 1980, p. 67) to derive consistent estimators for the

variances of $\hat{\alpha}$ and $\hat{\beta}$ from the result that,

$$var(a_k) = \frac{1}{n}[E(Y_i^{2k}) - \{E(Y_i^k)\}^2],$$

and an application of what is often called the Mann-Wald theorem, but there is little cause for doing so, since these moment estimators have greater variance than do maximum likelihood estimators which are no longer prohibitive to compute. Nevertheless, iterative algorithms for numerically approximating maximum likelihood estimates are often sensitive to starting values, and moment estimators may provide good starting values since they do constitute consistent estimators in their own right.

Estimators that make use of the basic idea of method of moments estimators, but for which we have no pretense of deriving variances or expected squared errors are usually not called method of moments estimators. Nevertheless, I include them in this category because the concept of "matching" sample and theoretical moments seems the basic justification for their development. Situations to which moment estimation is applied in this context are often those of estimating parameters involved in the variance of response variables in a model. Often, but not always (as we will illustrate below) these are variances of additive error models. We will again illustrate this type of estimation through several examples.

Example 8.4 Linear Regression Revisited

Consider again the linear regression model of Example 8.2,

$$Y_i = \boldsymbol{x}_i^T \beta + \sigma \epsilon_i,$$

with $\epsilon_i \sim iid\, N(0,1)$ and $\boldsymbol{x}_i^T = (x_{1,i}, \ldots, x_{p,i})$. Suppose that consistent estimators are available for the expectation function parameters β, but we are given

nothing else. Directly from this model, we have that

$$\sigma\,\epsilon_i = Y_i - \boldsymbol{x}_i^T\,\beta,$$

so that the random variables $W_i = Y_i - \boldsymbol{x}_i^T\,\beta$ are iid with $N(0,\,\sigma^2)$ distributions. The second sample moment of the W_i is then,

$$\begin{aligned} a_2 &= \frac{1}{n}\sum_{i=1}^{n} w_i^2, \\ &= \frac{1}{n}\sum_{i=1}^{n}\{y_i - \boldsymbol{x}_i^T\,\beta\}^2. \end{aligned} \tag{8.2}$$

The sample moment a_2 in (8.2) is consistent for $E(\sigma^2\,\epsilon^2) = \sigma^2$, from basic properties of sample moments (e.g., Serfling 1980, p. 67).

Now, we were given in this problem that an estimator $\hat{\beta}$ was available such that

$$\hat{\beta} \xrightarrow{p} \beta.$$

If we define (8.2) as a function of β, we may write $a_2(\beta)$ on the left hand side of (8.2). Then, given a $\hat{\beta} \xrightarrow{p} \beta$, and understanding that $a_2(\cdot)$ is continuous, a portion of the Mann-Wald theorem (e.g., Serfling 1980, p. 24) gives that

$$a_2(\hat{\beta}) \xrightarrow{p} a_2(\beta) \xrightarrow{p} \sigma^2,$$

or

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \boldsymbol{x}_i^T\,\hat{\beta}\}^2, \tag{8.3}$$

is consistent for σ^2. It is not true that $E(\tilde{\sigma}^2) = \sigma^2$ and it is typical to adjust (8.3) as $\hat{\sigma}^2 = (n/(n-p))\tilde{\sigma}^2$ to produce an unbiased estimator. Note, however, that the production of unbiasedness through this adjustment depends on having a linear model with variance that does not depend on the mean.

As we have seen in Section 8.1.1, in this normal-based linear model a much stronger result is available for the estimator $\hat{\sigma}^2$. In many other situations,

however, we may retain the argument immediately above for consistency even though the conditions needed for the stronger result will not be met.

Example 8.5 Constant Variance Nonlinear Regression

Consider a model in the form of expression (7.1) in Chapter 7.2,

$$Y_i = g(x_i, \beta) + \sigma \epsilon_i,$$

where $g(\cdot)$ is a specified function, $\beta = (\beta_1, \ldots, \beta_p)$ and, for $i = 1, \ldots, n$, we take $\epsilon_i \sim iid\, N(0, 1)$. This model implies that

$$W_i = Y_i - g(x_i, \beta),$$

are iid random variables with expectation 0 and variance σ^2. Suppose that consistent estimators $\hat{\beta}$ are a available for β. Proceeding in a manner directly analogous to that of Example 8.4, we can develop the consistent estimator of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \{y_i - g(x_i, \hat{\beta})\}^2. \tag{8.4}$$

Here, although it is typically the case that the denominator $n - p$ is used in (8.4), there is really no justification other than analogy with producing unbiasedness in normal linear models. That is, although $n - p$ is typically used, n provides the same results (consistency), and there is no formal justification for adjusting the "degrees of freedom" in this case.

Example 8.6 Nonlinear Regression with Known Variance Parameters

The argument we have developed in Examples 8.4 and 8.5 also applies to

models of the form (7.4), in which the resultant estimator of σ^2 becomes,

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \left(\frac{y_i - g_1(x_i, \hat{\beta})}{g_2(x_i, \hat{\beta}, \theta)} \right)^2 .$$

Recall that for this type of model we are assuming the parameter θ is known or selected as a part of model specification and, again, there is no theoretical justification for the choice of $n - p$ over n in the denominator.

Example 8.7 Nonlinear Regression with Unknown Variance Parameters

If we use the same form of model as in Example 8.6, but now considering the parameter θ of the variance model unknown, can the same development be used to estimate σ^2? The answer is no, it generally cannot. Why is this? Recall that the consistency of our estimators $\hat{\sigma}^2$ in the above examples (8.4, 8.5, 8.6) depended on the availability of a consistent estimator of parameters β in the expectation function. When the variance model depends on additional unknown parameters, such as for the model in expression (7.9), the development of estimators for β and θ cannot be undertaken in a separate manner. That is, there is no consistent estimator of β available without the same for θ, and *vice versa*. Thus, we must consider estimation of the entire parameter set $\{\beta, \theta\}$ in a simultaneous manner.

Example 8.8 Dispersion Parameters in Generalized Linear Models

Consider now, estimating the dispersion parameter ϕ in a generalized linear model as given by expressions (7.19), (7.20) and (7.21) in Section 7.3.2. Suppose that the link function $g(\cdot)$ is continuous and that consistent estimators of β are available for the systematic model component $g(\mu_i) = \boldsymbol{x}_i^T \beta$. Then, an

initial application of the Mann-Wald theorem gives that

$$\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T \hat{\beta}) \xrightarrow{p} \mu_i. \tag{8.5}$$

From (7.19) we have that $E(Y_i) = \mu_i = g^{-1}(\boldsymbol{x}_i^T \beta)$ and $var(Y_i) = (1/\phi)V(\mu_i)$. Thus, the random variables

$$W_i = \frac{Y_i - \mu_i}{\{V(\mu_i)\}^{1/2}}$$

are independent with distributions that have expectation 0 and variance $1/\phi$, a constant. While we may have no idea what these distributions actually are, we do have that the W_i are independent with $E(W_i^2) = (1/\phi)$. A basic moment estimator for $(1/\phi)$ is then,

$$\frac{1}{n}\sum_{i=1}^{n} W_i^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\{y_i - \mu_i\}^2}{V(\mu_i)}.$$

An additional application of the Mann-Wald theorem then results in

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\{y_i - \hat{\mu}_i\}^2}{\hat{V(\mu_i)}} \xrightarrow{p} \frac{1}{\phi}.$$

Since we are concerned with asymptotic behavior rather than expectation, we also have that

$$\hat{\phi} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\{y_i - \hat{\mu}_i\}^2}{\hat{V(\mu_i)}}\right]^{-1} \xrightarrow{p} \phi.$$

Example 8.9 Empirical Variograms

As a final example, consider estimation of the function defined as a *variogram* in our discussion of continuous index random fields; see expression (7.60). Suppose that $E\{Y(\boldsymbol{s})\} = \mu$ for all $\boldsymbol{s} \in \mathcal{D}$. Then we would have that,

$$2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) = E[Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)]^2; \quad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{D}.$$

The above expression suggests using the average of the squared differences for pairs of data values to estimate $2\gamma(\cdot)$. Suppose that, for a given displacement $\boldsymbol{h}$, we have a number of pairs of locations such that $\boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}$. In a set of data from locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ we might then form the set $n(\boldsymbol{h}) \equiv \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h};\ i, j = 1, \ldots, n\}$ and estimate $2\gamma(\cdot)$ as,

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} \{Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)\}^2; \quad \boldsymbol{h} \in \Re^d, \tag{8.6}$$

where $|N(\boldsymbol{h})|$ is the number of pairs of locations in the set $N(\boldsymbol{h})$. The estimator (8.6) is called "Matheron's estimator" and is generally considered as a moment-based estimator for obvious reasons. Note that we typically do not have multiple pairs of locations with the same displacements so that $N(\boldsymbol{h})$ is replaced with a "tolerance class" $N(\boldsymbol{h}(l)) \equiv \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j \in T(\boldsymbol{h}(l))\}; \quad l = 1, \ldots, k$, where $T(\boldsymbol{h}(l))$ is defined as some region around the displacement $\boldsymbol{h}$. In fact, it is not infrequent to make an (at least initial) assumption that the process is *isotropic* in which displacement $\boldsymbol{h}$ in $\Re^d$ is replaced with distance h in $\Re^+$.

A fundamental point to be gained through the above examples is that, although we typically do not resort to the "method of moments" as a self-contained estimation procedure in its own right, many of the estimators that we typically employ, particularly for estimation of variances, make use of the basic intuition of the method of moments which is that estimators may sometimes be developed by equating theoretical expectations with sample averages of the corresponding quantities.

8.2 Least Squares Estimation

Notice that most of the examples of moment-based estimation discussed in Section 8.1 made no use of distributional assumptions. In the case of estimat-

ing all model parameters by matching theoretical and sample moments, as in Example 8.3, we made use of explicit distributional assumptions (i.e., gamma). In Example 8.8 on estimation of dispersion parameters in glms we made use of general distributional structure for exponential dispersion families, but not specific distributional form. Other than these cases we made no use of distributional information, dealing only with basic definitions of moments in terms of the expectation operator (and, in particular, variances). The least squares approach to estimation shares this characteristic of avoiding specific distributional assumptions, and this is often mentioned as a "robustness" property of the method. On the other hand, it is also true that, when it comes time to consider inferential quantities such as interval estimates for parameters, we usually revert to an assumption of normality. As a final introductory comment note that we have also had an indication that least squares estimators are sometimes considered to be moment-based estimators (e.g., Lindsey 1996, p. 123). While this may be a legitimate connotation for least squares, I believe it is not quite true to the origins of the method and prefer to think of least squares estimators as distinct from moment estimators.

8.2.1 The Concept of Least Squares

One way to approach least squares is to view the method as the solution of a geometric minimization problem. To formulate the problem in this manner consider a set of real numbers $\boldsymbol{y} \equiv \{y_i : i = 1, \ldots, n\}$ as a point in $\Re^n$. Define the inner product of two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, both in $\Re^n$, relative to an $n \times n$ positive definite matrix A as,

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_A = \boldsymbol{u}^T A \, \boldsymbol{v},$$

or,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} u_i\, v_j\, a_{i,j},$$

where $a_{i,j}$ is the ij^{th} element of the matrix A. Now, let M_L denote a linear manifold of $\Re^n$, and $\boldsymbol{m} \equiv (m_1, \ldots, m_n)$ an element of M_L. Further, define the metric $||\boldsymbol{u}||_A \equiv <\boldsymbol{u}, \boldsymbol{u}>_A^{1/2}$. Now, consider minimizing the squared distance (metric) between $\boldsymbol{y}$ and $\boldsymbol{m}$,

$$\min_{\boldsymbol{m}\in M_L} ||\boldsymbol{y}-\boldsymbol{m}||_A^2,$$

or,

$$\min_{\boldsymbol{m}\in M_L} (\boldsymbol{y}-\boldsymbol{m})^T A\,(\boldsymbol{y}-\boldsymbol{m}).$$

As a final step to get this in familiar form, let $\boldsymbol{X}$ be an $n \times p$ matrix whose columns span the linear manifold M_L as $\boldsymbol{X}\,\boldsymbol{\beta} = M_L$. The problem then becomes

$$\min_{\boldsymbol{\beta}\in\Re^p} (\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta})^T A\,(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}). \tag{8.7}$$

As a side note, we have restricted the general problem from $\boldsymbol{y}$ being in a Hilbert space and $\langle\,\rangle$ a generic inner product to the particular instance of this problem that is usually the one of statistical interest. Now, what is known as the *Projection Theorem* gives the solution of (8.7) as that value $\boldsymbol{\beta}^*$ such that

$$\begin{aligned}
\langle(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}^*),\,(\boldsymbol{X}\,\boldsymbol{\beta}^*)\rangle_A &= 0, \\
&\text{or} \\
(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}^*)^T A\,\boldsymbol{X}\boldsymbol{\beta}^* &= 0 \\
\Rightarrow \boldsymbol{\beta}^{*T}\boldsymbol{X}^T A\boldsymbol{y} - \boldsymbol{\beta}^{*T}\boldsymbol{X}^T A\boldsymbol{X}\boldsymbol{\beta}^* &= 0 \\
\Rightarrow \boldsymbol{X}^T A\boldsymbol{X}\boldsymbol{\beta}^* &= \boldsymbol{X}^T A\boldsymbol{y} \\
\Rightarrow (\boldsymbol{X}^T A\boldsymbol{X})^{-1}\boldsymbol{X}^T A\boldsymbol{y} &= \boldsymbol{\beta}^*.
\end{aligned} \tag{8.8}$$

To express the least squares problem (8.7) and its solution (8.8) in a form that is statistically familiar, we made use of the restriction that M_L constituted a linear manifold spanned by the columns of a known matrix $\boldsymbol{X}$. If we replace M_L with a nonlinear manifold M_N what changes? Suppose we replace the $n \times 1$ vector $\boldsymbol{X}\boldsymbol{\beta}$ with an $n \times 1$ vector $g(\boldsymbol{X}, \boldsymbol{\beta})$ for some known nonlinear function $g(\cdot)$. Continuing to write $\boldsymbol{X}$ as a matrix implies nothing in particular to do with a linear vector space; $\boldsymbol{X}$ here is simply a convenient notation for a collection of vectors $\{\boldsymbol{x}_i : i = 1, \ldots, n\}$ where $\boldsymbol{x}_i \equiv (\boldsymbol{x}_{1,i}, \ldots, \boldsymbol{x}_{p,i})^T$. The least squares minimization problem analogous to (8.7) then becomes,

$$\min_{\boldsymbol{\beta} \in \Re^p} (\boldsymbol{y} - g(\boldsymbol{X}, \boldsymbol{\beta}))^T A (\boldsymbol{y} - g(\boldsymbol{X}, \boldsymbol{\beta})). \tag{8.9}$$

The projection theorem continues to give a solution to this problem as,

$$\langle (\boldsymbol{y} - g(\boldsymbol{X}, \boldsymbol{\beta})), g(\boldsymbol{X}, \boldsymbol{\beta}) \rangle_A = 0,$$

although this solution cannot be determined in closed form similar to (8.8) for the case of a linear manifold M_L.

What should we take away from this discussion of the concept of least squares?

1. There is nothing fundamentally statistical about the concept of least squares – least squares is the solution to a minimization problem in vector spaces.

2. In problems that are fundamentally linear, the least squares problem allows an explicit solution.

3. The least squares problem is defined with respect to a known, positive definite, "weight" matrix (the matrix A). This will be critical in terms of whether an exact or only approximate solution to the minimization

problem can be determined, and whether even an approximate solution is unique, even in the case of a linear problem.

One additional point that will become crucial in determining for which types of statistical models estimation by least squares might be appropriate comes from a more carefully examination of what was called the projection theorem, which offers a (unique, as it turns out, for a known weight matrix A) solution to the basic least squares problem.

The projection theorem may be stated as:

Theorem: Let $\boldsymbol{y}$ be in a Hilbert space V and let M be a subspace of V such that $\boldsymbol{y} \notin M$. Further, let A be a known positive definite matrix. Then $\boldsymbol{y}$ can be uniquely represented in the form $\boldsymbol{y} = \boldsymbol{m} + \boldsymbol{v}$ for some $\boldsymbol{m} \in M$ and $\boldsymbol{v} \perp M$ such that, for any $\boldsymbol{w} \in M$

$$||\boldsymbol{y} - \boldsymbol{w}||_A^2 \geq ||\boldsymbol{y} - \boldsymbol{m}||_A^2,$$

with equality if and only if $\boldsymbol{w} = \boldsymbol{m}$.

As already indicated, the subspace M corresponds to either a linear manifold M_L or a nonlinear manifold M_N as described by the vector of expectation functions $\{E(Y_1), \ldots, E(Y_n)\}$. The essence of the projection theorem is to decompose $\boldsymbol{y}$ into two parts, one that lies within this manifold, and the other *additive* component that is orthogonal to M. This points directly to the use of least squares for estimation of parameters involved in the expectation function of additive error models.

8.2.2 Least Squares as Statistical Estimation

So far we have attached no statistical properties to either the formulation of the least squares problem, or to the solution of this problem given by the pro-

jection theorem. As hinted at in the comments at the end of Section 8.2.1, there are some differences that arise depending on whether an additive error model to which least squares is applied contains a linear or nonlinear expectation function, and whether the appropriate "weight matrix" A is considered known or unknown. In fact, we have left the exact identity of this matrix somewhat vague. Consider that solution of a least squares problem leads to estimators of the expectation function parameters $\boldsymbol{\beta}$. In the same spirit of developing properties for moment estimators in Section 8.1, we wish to determine properties of these least squares estimators. It turns out that whether such properties exist and, if so, what they are, depends on the choice made in definition of the weight matrix A that helps define the least squares problem.

We will consider the issues of choosing an appropriate matrix A to formulate the least squares problem, finding numerical solutions to that problem for particular data sets, and attaching statistical properties to the resultant estimators for a number of general cases.

Ordinary Least Squares

Suppose that we have a linear additive error model with constant variance,

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \epsilon_i; \quad i = 1, \ldots, n, \tag{8.10}$$

where $\epsilon_i \sim iid\, F$ such that $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$; usually, F is taken as normal but (as you already know) that is not required to attach statistical properties to the ols estimators of $\boldsymbol{\beta}$.

Here, it is beneficial to write the model in matrix form as,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\, \boldsymbol{\epsilon},$$

in which we have $cov(\boldsymbol{\epsilon}) = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix. Take the weight matrix A in the least squares problem (8.7) to be $A = I_n^{-1} = I_n$;

the reason for the initial inverse will become clear shortly. Then by (8.8) the values of $\boldsymbol{\beta}$ that solve the least squares problem are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{8.11}$$

which are the usual ordinary least squares estimators. Statistical properties are attached to $\hat{\boldsymbol{\beta}}$ in (8.11) by means of the standard *Gauss-Markov* theorem, which states that $\hat{\boldsymbol{\beta}}$ is UMVU among all estimators that are linear functions of the random vector $\boldsymbol{Y}$.

To derive the variance of the ols estimator $\hat{\boldsymbol{\beta}}$ we make critical use of the fact that the estimator is a linear function of the response vector $\boldsymbol{Y}$. Combining this with the Gauss-Markov result of unbiasedness, and the fact that the model gives $cov(\boldsymbol{Y}) = E(\boldsymbol{Y}\boldsymbol{Y}^T) = \sigma^2 I_n$ we have,

$$\begin{aligned} cov(\hat{\boldsymbol{\beta}}) &= E\left[\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\}\boldsymbol{Y}\boldsymbol{Y}^T\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\}^T\right] \\ &= \{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\}E(\boldsymbol{Y}\boldsymbol{Y}^T)\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\}^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T I_n \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}^{-1})^T\sigma^2 \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2. \end{aligned}$$

For estimation of this covariance we replace σ^2 with an unbiased estimator,

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2,$$

which is UMVU if the error terms are normally distributed, see Example 8.2. For inference, we generally strengthen model assumptions in (8.10) to include that the error distribution F is normal, which then leads to a joint normal distribution for the elements of $\boldsymbol{\beta}$, the concomitant normal marginal distributions as normal, and the standardized elements of $\hat{\beta}$ using estimated variances as $t-$distributions from which intervals are formed. Take note of the fact that,

the exact theory results in this case lead to $t-$distributions as *results* so that it is entirely appropriate and correct to use quantiles of these distributions for interval estimation.

Weighted Least Squares

Now consider a model of the form of expression (7.3) with $g(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, namely,

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + (\sigma/\sqrt{w_i})\,\epsilon_i; \quad i = 1, \ldots, n, \tag{8.12}$$

where the $\{w_i : i = 1, \ldots, n\}$ are assumed to be known constants, and the same assumptions are made about the additive errors ϵ_i as for model (8.10). The only difference in estimation and inference for this model from the constant variance model of (8.10) is that the covariance matrix for the vector $\boldsymbol{Y}$ becomes $cov(\boldsymbol{Y}) = \sigma^2 \boldsymbol{W}^{-1}$ where $\boldsymbol{W}^{-1}$ is a diagonal $n \times n$ matrix with elements $1/w_i$. It turns out that the implication is that the appropriate weight matrix A in the least squares problem (8.7) and solution (8.8) is $A = \boldsymbol{W}$. Least squares estimators of the elements of $\boldsymbol{\beta}$ are then given as,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}. \tag{8.13}$$

The Gauss-Markov theorem continues to hold, and the derivation of the covariance for $\hat{\boldsymbol{\beta}}$ in a manner similar to that presented for ordinary least squares results in

$$cov(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \sigma^2.$$

To estimate this covariance we now use as an estimator of σ^2 the estimator

$$S_w^2 = \frac{1}{n-p} \sum_{i=1}^{n} w_i (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2,$$

which is developed as a bias corrected moment estimator in exactly the same way as S^2 for the constant variance model, is still a function of the complete

sufficient statistic for a normal model (and thus is UMVU) and may be used in a standardization of elements of $\hat{\boldsymbol{\beta}}$ that results in $t-$distributions.

Take note that, in the models considered for estimation with ordinary or weighted least squares, we have considered only linear models for which there was no dependence between the mean and variance models and for which the variance model has included no additional unknown parameters. In other words, first of all linear models, and secondly models for which the only unknown quantity involved in the variances of responses is a constant σ^2. Both of these restrictions turn out to be critical for application of the Gauss-Markov theorem that gives exact (or small sample) results for least squares estimators.

Generalized Least Squares

We have discussed in Chapter 7 a number of additive error models that do not fall into the categories described above for which the Gauss-Markov theorem provides small sample properties for estimators. There are several of these for which Gauss-Markov does not apply but for which we may, however, still consider the basic concept of least squares estimation. First are linear models with variances that may depend on the mean but not additional unknown parameters (i.e., models of the form of expression (7.5) of Section 7.2.3, for which we assume θ is known and $\mu_i(\beta)$ is linear in covariates x_i). Second are nonlinear models with constant variance (i.e., models of the form (7.1) in which g is a nonlinear function and the variance of error terms ϵ_i are constant). Finally, are nonlinear models of the form of model (7.5) in which variances may depend on the mean but not other unknown parameters.

There are two modifications to solution of the least squares problems in these situations that both result in iterative computational algorithms to solve

a least squares problem. Estimates that result from any algorithm depending on one or both of these modifications are typically called *generalized least squares* estimates.

Consider first an additive error model such as (7.5) in which we take the expectation function to be linear,

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma g(\boldsymbol{x}_i^T\boldsymbol{\beta},\, \theta)\epsilon_i, \tag{8.14}$$

with the usual additive error model assumptions on the ϵ_i and where θ is considered known (e.g., chosen prior to estimation as a part of model selection). Now, model (8.14) is quite similar to model (8.12) if we write

$$\sqrt{w_i(\boldsymbol{\beta})} = \frac{1}{g(\boldsymbol{x}_i^T\boldsymbol{\beta},\, \theta)},$$

the distinction being that here we have written the "weights" as functions of $\boldsymbol{\beta}$ whereas in (8.12) they were assumed to be known constants.

Consider taking preliminary estimates of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(0)}$ for use as fixed values in the weights but not the expectation function. Then our model could be written as,

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \frac{\sigma}{\sqrt{w_i(\boldsymbol{\beta}^{(0)})}}\,\epsilon_i,$$

and, in analogy with (8.12) and (8.13) this suggests a weighted least squares solution of the form

$$\hat{\boldsymbol{\beta}}^{(1)} = (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(0)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(0)})\boldsymbol{Y}, \tag{8.15}$$

where $\boldsymbol{W}(\boldsymbol{\beta}^{(0)})$ is an $n \times n$ diagonal matrix with elements

$$w_i(\boldsymbol{\beta}^{(0)}) = g^2(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(0)},\, \theta). \tag{8.16}$$

As suggested by the notation of (8.15) and (8.16), we might then iterate this process, taking new weights calculated as $w_i(\boldsymbol{\beta}^{(1)})$ from (8.16), then solving

(8.15) with these weights to produce $\hat{\boldsymbol{\beta}}^{(2)}$ and so forth until $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)}$ at which time we say the iterative procedure has "converged". Just to keep everything straight, note that the least squares minimization problem we are attempting to solve here is,

$$\min_{\beta} \sum_{i=1}^{n} w_i(\boldsymbol{\beta})\{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\}^2. \tag{8.17}$$

Now consider an additive error model such as (7.5) in which the expectation function is nonlinear but in which the variance model $g(\mu_i(\boldsymbol{\beta}), \theta) \equiv 1$ for all $i = 1, \ldots, n$,

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma\, \epsilon_i, \tag{8.18}$$

where $\epsilon_i \sim iid\, F$ with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$. Suppose here we also have a preliminary estimate $\boldsymbol{\beta}^{(0)}$ and we approximate the expectation function with a first-order Taylor expansion,

$$E(Y_i) = g(\boldsymbol{x}_i, \boldsymbol{\beta}) \approx g(\boldsymbol{x}_i, \boldsymbol{\beta}^{(0)}) + \sum_{k=1}^{p} V_{i,k}^{(0)}(\beta_k - \beta_k^{(0)}),$$

where, for $k = 1, \ldots, p$,

$$V_{i,k}^{(0)} = \frac{\partial}{\partial \beta_k}\, g(\boldsymbol{x}_i, \boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}}. \tag{8.19}$$

This approximation then allows us to write,

$$Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta}^{(0)}) \approx \sum_{i=1}^{n} V_{i,k}^{(0)}(\beta_k - \beta_k^{(0)}) + \sigma\, \epsilon_i, \tag{8.20}$$

which is in the form of a linear regression model with the "usual Y_i" replaced by $Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})$, the "usual $x_{i,k}$" replaced by $V_{i,k}^{(0)}$, and the "usual β_k" replaced by $(\beta_k - \beta_k^{(0)})$.

Equation (8.20) suggests the use of ordinary least squares to obtain an estimate of

$$\boldsymbol{\delta}^{(0)} \equiv (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^T$$

as,

$$\boldsymbol{\delta}^{(0)} = (\boldsymbol{V}^{(0)T}\boldsymbol{V}^{(0)})^{-1}\boldsymbol{V}^{(0)T}\tilde{\boldsymbol{Y}}^{(0)},$$

where $\boldsymbol{V}^{(0)}$ is an $n \times p$ matrix with ik^{th} element $V_{i,k}^{(0)}$ and $\tilde{\boldsymbol{Y}}^{(0)}$ is a vector of length n with elements $Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta}^{(0)})$. An updated estimate of $\boldsymbol{\beta}$ may then be obtained as

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \boldsymbol{\delta}^{(0)}. \tag{8.21}$$

Replacing $\boldsymbol{\beta}^{(0)}$ with $\boldsymbol{\beta}^{(1)}$ in (8.19) and (8.20) allows expression of an updated model form in terms of $\boldsymbol{V}^{(1)}$ and $\tilde{\boldsymbol{Y}}^{(1)}$, and (8.21) allows this to be updated to $\boldsymbol{\beta}^{(2)}$ and so on in an iterative manner. As before, when $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)}$ we would say the iterative procedure has converged. The least squares minimization problem we are attempting to solve with this model is

$$\min_{\beta} \sum_{i=1}^{n} \{y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})\}^2. \tag{8.22}$$

Finally, consider a combination of the two models discussed above, namely model of the form (7.4),

$$Y_i = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta)\, \epsilon_i,$$

where we are still considering θ as known. Here, a combination of the thinking that resulted in (8.15) and (8.16) for linear models and (8.19) through (8.21) for nonlinear models results in a full-blown generalized least squares algorithm of the following form.

Generalized Least Squares Algorithm

1. Calculate initial estimates $\boldsymbol{\beta}^{(0)}$.

   For $j = 0, \ldots,$

2. Calculate the $\boldsymbol{W}^{(j)}$ matrix as an $n \times n$ diagonal matrix with elements

$$w_i(\boldsymbol{\beta}^{(j)}) = g_2^2(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(j)}, \theta).$$

3. Calculate the $\boldsymbol{V}^{(j)}-$ matrix as an $n \times p$ matrix with ik^{th} element

$$V_{i,k}^{(j)} = \frac{\partial}{\partial \beta_k} g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(j)}}.$$

4. Calculate elements of $\tilde{\boldsymbol{Y}}^{(j)}$, the vector of "$j-$step response variables" as,

$$\tilde{Y}_i^{(j)} = Y_i - g_1(\boldsymbol{x}_i, \boldsymbol{\beta}^{(j)}).$$

5. Calculate the "step" $\boldsymbol{\delta}^{(j)}$ as,

$$\boldsymbol{\delta}^{(j)} = \left(\boldsymbol{V}^{(j)T} \boldsymbol{W}^{(j)} \boldsymbol{V}^{(j)}\right)^{-1} \boldsymbol{V}^{(j)T} \boldsymbol{W}^{(j)} \tilde{\boldsymbol{Y}}^{(j)}.$$

6. Update estimates of the expectation function parameters $\boldsymbol{\beta}$ as,

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} + \boldsymbol{\delta}^{(j)}.$$

7. Update $j = j + 1$ and return to step 2.

The least squares minimization problem this algorithm finds a solution to is

$$\min_{\beta} \sum_{i=1}^{n} w_i(\boldsymbol{\beta}) \{y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})\}^2, \tag{8.23}$$

where $w_i(\boldsymbol{\beta})$ is now defined as (c.f. expression (8.16)),

$$w_i(\boldsymbol{\beta}) = g_2^2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta).$$

Although for linear models with constant variance or linear models with variances that are functions of known weights we usually employ the much

simplified algorithms of ordinary least squares or weighted least squares, the minimization problems attached to those models fit the general form of (8.23). Thus, if the generalized least squares algorithm is, in fact, solving (8.23) it should work with any of the additive error models considered thus far (i.e., linear or nonlinear models with constant variance, variances that are functions of known weights, or variances that are functions of expectations with any additional parameters known).

As a fairly meaningless exercise relative to estimation, but to demonstrate that the generalized least squares algorithm is not out of concert with the "direct" (i.e., non-iterative) least squares algorithms, consider estimating expectation function parameters $\boldsymbol{\beta}$ for a constant variance linear model using the ordinary least squares estimates as starting values, $\beta^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$; if we did this we would already be done, which is why this is meaningless for actual estimation. The linear model form, however, would result in $\boldsymbol{V}^{(j)} = \boldsymbol{X}$ for all j, and the constant variance specification would give $\boldsymbol{W}^{(j)} = I_n$ for all j. Then on the first iteration of the generalized least squares algorithm we would have, in steps 4, 5 and 6,

Step 4

$$\tilde{Y}_i^{(1)} = Y_i - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

Step 5

$$\boldsymbol{\delta}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\}.$$

Step 6

$$\begin{aligned}\boldsymbol{\beta}^{(1)} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &+ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\}\end{aligned}$$

$$
\begin{aligned}
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\
&- (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\
&- (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \\
&= \boldsymbol{\beta}^{(0)}.
\end{aligned}
$$

Thus, the generalized least squares algorithm does not "contradict" ordinary least squares. A similar demonstration is possible for a linear model with variances that depend on known weights. Beginning with weighted least squares estimates as starting values, the generalized least squares algorithm returns those same values after any number of iterations.

The reason we separate ordinary and weighted least squares from generalized least squares in discussing methods of estimation is that, while the Gauss-Markov theorem provides exact theory results for the former cases, this is no longer true for the more complex situations in which generalized least squares is applicable. That is, when we have an additive error model for which either the expectation function is nonlinear or the variances depend on the parameters of that expectation function (but no other unknown parameters) or both, then the Gauss-Markov theorem no longer applies. Rather, we attach statistical properties to generalized least squares estimators through what has been called the *Fundamental Theorem of Generalized Least Squares* which may be stated as follows:

Fundamental Theorem of Generalized Least Squares

Consider any model of the form of expression (7.4),

$$Y_i = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta)\, \epsilon_i,$$

in which the value of θ is known and, for $i = 1, \ldots, n$, $\epsilon_i \sim iid\, F$ such that $E(\epsilon_i) = 0$ and $var(\epsilon_i) = c$ for a known constant c (usually $c \equiv 1$). Then, under mild smoothness conditions on $g_1(\cdot)$ and $g_2(\cdot)$, for any starting estimate $\boldsymbol{\beta}^{(0)}$ such that $\boldsymbol{\beta}^{(0)}$ is $n^{1/2}-$consistent, and for any j in the generalized least squares algorithm (i.e., any number of iterations),

$$\boldsymbol{\beta}^{(j)} \text{ is } AN\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\Sigma_\beta^{-1}\right), \tag{8.24}$$

where

$$\Sigma_\beta = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{\beta})\boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{\beta})^T / g_2^2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta). \tag{8.25}$$

In (8.25), $\boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{\beta})$ is a $p \times 1$ column vector with k^{th} element

$$v_k(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k} g_1(\boldsymbol{x}_i, \boldsymbol{\beta}).$$

Note that this $\boldsymbol{v}_k(\boldsymbol{x}_i, \boldsymbol{\beta})$ is essentially the same quantity given as $V_{i,k}^{(j)}$ in Step 3 of the generalized least squares algorithm except that here we are not (yet) evaluating these derivatives at estimated values of $\boldsymbol{\eta}$ since Σ_β in (8.25) gives the variance of the asymptotic normal distribution (up to the scale factor σ^2/n), not the estimated variance.

Before we discuss estimating the variance parameters σ^2 and Σ_β we should say a few words about the "for any number of iterations" part of the fundamental theorem of generalized least squares, since this is not an intuitive portion of the result. Does this mean, for example, that if we take a starting value

$\boldsymbol{\beta}^{(0)}$ and conduct $j = 0$ iterations of the algorithm we end up with the same asymptotic normality as if we iterate until $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)}$? The answer is yes, it does. How can this be? Recall one of the other conditions of this theorem, that $\beta^{(0)}$ constitute a "root n consistent" estimator for $\boldsymbol{\beta}$. Given this, the stated asymptotic normality holds for estimators that result from *any number* of iterations of the algorithm, and there are proponents for various choices. Some references, taken from the discussion by Carroll and Rupert (1988, Section 2.3) are given in the table below:

Iterations	Proponents
1	Goldberger (1964)
	Matloff, Rose and Tai (1984)
2	Williams (1959)
	Seber (1977)
2 or 3	Carroll and Ruppert (1988)
∞	McCullagh and Nelder (1989)

In this table, ∞ means iteration until convergence which is technically $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)}$ but in practice means $\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)} < \delta$ for some suitably small δ such as 10^{-6} or 10^{-8}. For further discussion of generalized least squares and connected asymptotic results, see also Jobson and Fuller (1980) and Carroll and Rupert (1982).

Estimation of σ^2 is generally accomplished through the use of a moment estimator. Let $\hat{\boldsymbol{\beta}}$ denote a generalized least squares estimator of $\boldsymbol{\beta}$.

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \left\{ \frac{Y_i - g_1(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})}{g_2(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}, \theta)} \right\}^2 . \tag{8.26}$$

Note that this estimator no longer possesses any small sample properties (despite what is suggested by the divisor of $n - p$). It is, however, consistent as

long as $\hat{\boldsymbol{\beta}}$ is consistent which was a starting point in the theorem (actually, root n consistency is stronger than consistency alone).

For inference connected with generalized least squares estimators then, we make use of the result of the fundamental theorem of generalized least squares given in (8.24), with estimated variances produced by plug-in use of $\hat{\sigma}^2$ from (8.26) and $\hat{\boldsymbol{\beta}}$ from the generalized least squares algorithm, giving

$$\hat{cov}(\hat{\boldsymbol{\beta}}) = \frac{\hat{\sigma}^2}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})\boldsymbol{v}(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})^T/g_2^2(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}},\theta)\right]^{-1}. \tag{8.27}$$

Interval estimates are then computed in the usual way. For an individual element β_k of $\boldsymbol{\beta}$ this is

$$\begin{aligned} \hat{\beta}_k &\pm t_{1-\alpha/2;n-p}\left\{\hat{cov}(\hat{\boldsymbol{\beta}})_{k,k}\right\}^{1/2} \\ or& \\ \hat{\beta}_k &\pm z_{1-\alpha/2}\left\{\hat{cov}(\hat{\boldsymbol{\beta}})_{k,k}\right\}^{1/2} \end{aligned} \tag{8.28}$$

where $\hat{cov}(\hat{\boldsymbol{\beta}})_{k,k}$ is the k^{th} diagonal element of the estimated covariance matrix given in (8.27), $t_{1-\alpha/2;n-p}$ is the $1-\alpha/2$ quantile of a $t-$distribution with $n-p$ degrees of freedom and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution.

For inference concerning tests of hypotheses about parameter values (i.e., model selection) and the development of joint confidence regions for sets of the elements of $\boldsymbol{\beta}$ there are a number of approaches, none of which depend explicitly on the fact that $\hat{\boldsymbol{\beta}}$ is a generalized least squares estimator. These inferential methods can be based on what is called "Wald theory", which applies to any asymptotically normal estimator, likelihood theory which we will cover in the next section, approximate likelihood theory which we will cover after that, or sample reuse (specifically, either nonparametric or parametric bootstrap) which we cover after the other methods. While we have, in fact, used a

portion of Wald theory in the intervals of expression (8.28), we will delay full discussion of this topic until we encounter it again in likelihood estimation.

8.2.3 Summary of Least Squares Estimation

We conclude our consideration of least squares as a method of estimation by summarizing most of the key points:

1. Least squares is used nearly exclusively with additive error models.

2. Least squares is, fundamentally, not a statistical estimation procedure. It is motivated by solving a general minimization problem.

3. For linear models with either constant variance or variances that are proportional to known weights, least squares estimators have exact theory properties. In particular, they are UMVU estimators, and may be used in conjunction with UMVU estimators of the variance parameter σ^2. Inferential procedures in these situations are typically developed under the additional assumption of normally distributed errors to result in the types of intervals and tests you are familiar with from Stat 500 and Stat 511.

4. For nonlinear models with constant variance, or for either linear or nonlinear models with variances that depend on parameters of the expectation function but no additional unknown parameters, generalized least squares estimators are asymptotically normal as given by the fundamental theorem of generalized least squares. Generalized least squares estimators are typically used in conjunction with consistent estimators of σ^2 developed from a moment based approach. Intervals for individual parameter elements may be based on this asymptotic normality (which

can be considered a part of Wald theory). Note that, in all of the cases considered under generalized least squares estimation, the development of intervals depending on normal distributions *does not* depend on the additional model assumption of normally distributed errors (as it does in the ordinary and weighted least squares cases). It *does*, however, require reliance on an asymptotic result.

5. Putting together the information in items 3 and 4 immediately above, we arrive at the "no free lunch" conclusion for estimation by least squares methods. Avoiding strong distributional assumptions on model terms is often considered a "good" thing. Being able to develop exact properties for estimators that do not depend on asymptotic arguments is often considered a "good" thing. Under models for which we can apply ordinary or weighted least squares we can accomplish both for *point* estimation, but then must rely on strong distributional assumptions for *inference*. Under models for which we turn to generalized least squares we can avoid strong distributional assumptions on the model entirely, but must rely on asymptotic results for both properties of point estimators and inferential procedures.

6. The ability to develop properties for least squares estimators, either exact theory for point estimation or asymptotic theory for both point estimation and inference, without assuming a specific parametric form for model distributions is often considered a "robustness" property or aspect of least squares, and this is true insomuch as robustness refers to small departures from an assumed distributional form. This concept of robustness is different than what is properly called *resistance*, which refers to the degree to which an estimator is affected by extreme observations. It

is well known that, while least squares estimators have a certain amount of robustness, they are extremely sensitive to the effect of extreme and high leverage observations.

8.3 Basic Likelihood Estimation and Inference

We turn now to what, aside from the Bayesian approach to estimation and inference discussed in Part III of the course, is probably the most generally applicable estimation method for statistical models. This method, called *maximum likelihood* estimation, depends on having a specified parametric distribution of random variables connected with observable quantities. Although properties of maximum likelihood have been examined and any number of results are available for cases involving dependent random variables, we will focus here on situations involving independent response variables.

8.3.1 Maximum Likelihood for Independent Random Variables

Let $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)^T$ be a vector of independent random variables with possible values in the sets $\Omega_1 \ldots, \Omega_n$ and assume that the set of possible values for $\boldsymbol{Y}$ is $\Omega \equiv \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n$, which is sometimes called the *positivity* condition. Like independence, the positivity condition is not necessarily needed for the types of results we will present, but it does simplify the general treatment (i.e., without the positivity condition we must usually approach the proof of various properties of maximum likelihood estimates on a case-by-case basis for indi-

vidual models). Further, let $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)^T$ denote a vector of parameters such that $\boldsymbol{\theta} \in \Theta \subset \Re^p$ with $p < n$.

We will also assume that the random variables contained in $\boldsymbol{Y}$ have probability density or probability mass functions $f_i(y_i|\boldsymbol{\theta})$; $i = 1, \ldots, n$, such that,

$$Pr(Y_i = y_i|\boldsymbol{\theta}) = f_i(y_i|\boldsymbol{\theta}) \text{ if } Y_i \text{ is discrete,}$$

$$Pr(a < Y_i < b|\boldsymbol{\theta}) = \int_a^b f_i(y_i|\boldsymbol{\theta}) \text{ if } Y_i \text{ is continuous.}$$

Maximum likelihood is very much a "data dependent" method of estimation. Any observation or measurement process has a finite precision. If observation of a given quantity results in a value y_i we will take this to mean that the associated random variable Y_i has a value in the range $y_i - \Delta_i < Y_i < y_i + \Delta_i$ for some Δ_i. If Y_i is continuous for $i = 1, \ldots, n$ and we assume independence, then for a set of observations $\boldsymbol{y} \equiv (y_1, \ldots, y_n)^T$, define

$$Pr(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ F_i(y_i + \Delta_i|\boldsymbol{\theta}) - F_i(y_i - \Delta_i|\boldsymbol{\theta}) \right\},$$

where $F_i(\cdot)$ is the distribution function corresponding to $f_i(\cdot)$. If Y_i has density $f_i(\cdot)$, the intermediate value theorem of calculus gives that,

$$F_i(y_i + \Delta_i|\boldsymbol{\theta}) - F_i(y_i - \Delta_i|\boldsymbol{\theta}) = \int_{y_i - \Delta_i}^{y_i + \Delta_i} f_i(t|\boldsymbol{\theta})dt \approx 2\Delta_i f_i(y_i|\boldsymbol{\theta}),$$

and then

$$Pr(\boldsymbol{y}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}).$$

For discrete situations we will assume that $\Delta_i = 0$ for all i and

$$Pr(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}),$$

In many, if not most, cases involving continuous random variables we assume that all Δ_i are small enough to be ignored, but there are numerous examples of where this is not the case (e.g., Lindsey 1996) and it can be advantageous (even necessary) to write the likelihood function in terms of the above integrals rather than densities. For our purposes here, however, we will assume that the "density approximation" is adequate. We then define the *likelihood function* for a set of observations $\boldsymbol{y}$ as,

$$\ell_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}). \tag{8.29}$$

The important point, which you have heard before but is worth emphasizing again, is that (8.29) is considered a function of an argument $\boldsymbol{\theta}$ for a fixed, known set of observations $\boldsymbol{y}$.

Quite simply, then, the *maximum likelihood* estimator of $\boldsymbol{\theta}$ is that value $\hat{\boldsymbol{\theta}} \in \Theta$ such that

$$\ell_n(\hat{\boldsymbol{\theta}}) \geq \ell_n(\boldsymbol{\theta}); \quad \text{for any } \boldsymbol{\theta} \in \Theta.$$

Now, given the preceeding material we have that $\ell(\boldsymbol{\theta}) \propto Pr(\boldsymbol{y}|\boldsymbol{\theta})$, which leads to the intuitive interpretation and justification of a maximum likelihood estimate as that value of the parameter that "makes the probability of the data as great as it can be under the assumed model". This is actually very nice as both an intuitive understanding and justification for maximum likelihood, but it leaves us a little short of what we might desire as a statistical justification. That is, having the value of the parameter that maximizes the probability of seeing what we saw certainly justifies the maximum likelihood estimate (mle) as a summarization of the available data, but it does not necessarily indicate that the mle is a good estimate of the parameter of interest $\boldsymbol{\theta}$. This is provided by the following result at least for the *iid* case with scalar parameter θ, adapted here from Lehmann (1983, section 6.2, Theorem 2.1).

Result

Let P_θ represent the distribution (probability measure) of a random variable indexed by the parameter θ. Suppose that, for $\theta \in \Theta$,

(i) the distributions P_θ have common support Ω

(ii) the random variables Y_i are *iid* with common density or mass function $f(y_i|\theta)$

(iii) the true value of θ, say θ_0 lies in the interior of Θ

Then, as $n \to \infty$

$$P_{\theta_0}\{f(Y_1|\theta_0)\dots f(Y_n|\theta_0) > f(Y_1|\theta)\dots f(Y_n|\theta)\} \to 1,$$

for any fixed $\theta \neq \theta_0$. In other words,

$$Pr\{f(\boldsymbol{Y}_n|\theta_0) > f(\boldsymbol{Y}_n|\theta)\} \to 1,$$

as $n \to \infty$. This indicates that, for large samples (at least large *iid* samples) the density of $\boldsymbol{Y}$ at the true parameter value exceeds the density of $\boldsymbol{Y}$ for any other parameter value. This provides a connection between a maximum likelihood estimate and the "true" parameter value in a model. That is, as the sample size increases, the parameter value that maximizes the joint distribution not only provides a good value for describing the observations at hand, but also must become close to the true value under a given model.

8.3.2 Notation and Settings

In the development leading to expression (8.29) independence was used only to allow the derivation of a joint density or mass function for $Y_1, \dots, Y_n$ by

multiplication of the individual density or mass functions. Importantly, however, these functions were assumed to depend on a common parameter $\boldsymbol{\theta}$ of finite dimension. We can, in a completely general setting, define a maximum likelihood estimator in the following way.

Suppose that $Y_1, \ldots, Y_n$ are random variables with possible values $\boldsymbol{y} \in \Omega_Y$, and that, for $\boldsymbol{\theta} \in \Theta \subset \Re^p$, these variables have joint probability density or mass function $f(\boldsymbol{y}|\boldsymbol{\theta})$. Then the likelihood is,

$$\ell_n(\boldsymbol{\theta}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta}), \tag{8.30}$$

and a maximum likelihood estimator of $\boldsymbol{\theta}$ is any value $\hat{\boldsymbol{\theta}}$ such that

$$\ell_n(\hat{\boldsymbol{\theta}}) \geq \ell_n(\boldsymbol{\theta}) \quad \text{for any} \quad \boldsymbol{\theta} \in \Theta. \tag{8.31}$$

Maximum likelihood estimators are often found by maximizing the *log likelihood* function,

$$L_n(\boldsymbol{\theta}) = \log\{f(\boldsymbol{y}|\boldsymbol{\theta})\}, \tag{8.32}$$

since the logarithmic function is monotone. If a value $\hat{\boldsymbol{\theta}} \in \Theta$ maximizes the likelihood (8.30) as in (8.31) then it also maximizes the log likelihood (8.32) and *vice versa*.

Still in a completely general setting, define what is often called the *score function* as the $p-$vector $U_n(\boldsymbol{\theta}) = (U_{n,1}(\boldsymbol{\theta}), \ldots, U_{n,p}(\boldsymbol{\theta})^T$, where,

$$U_{n,k}(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \theta_k} L_n(\theta); \quad k = 1, \ldots, p. \tag{8.33}$$

The expected or Fisher information plays an important role in likelihood estimation. Define the expected information as the $p \times p$ matrix,

$$I_n(\boldsymbol{\theta}) \equiv E\{U_n^T(\boldsymbol{\theta})\, U_n(\boldsymbol{\theta})\}. \tag{8.34}$$

Now while (8.30) through (8.34) apply in any situation for which a joint distribution depends on a fixed parameter vector of lower dimension, the cornerstone property of maximum likelihood estimation is *asymptotic efficiency*. Efficiency depends, through the *information inequality* or *Cramer-Rao Inequality*, on the expected information. And, asymptotically, a crucial aspect of likelihood theory is that the total information in a sample of size n tends to ∞ as n tends to ∞. It can be extremely difficult to work out the way this might (or might not) take place in a model with dependent random variables; this is not to say that likelihood theory cannot provide results in dependent cases, only that it becomes more involved. We will thus restrict attention to situations for which $Y_1, \ldots, Y_n$ are independent which are also the cases for which the expected information in (8.34) can be written as a sum.

In the case that $Y_1, \ldots, Y_n$ are *iid* the likelihood, log likelihood, and elements of the score function and expected information matrix may be written as,

$$\begin{aligned} \ell_n(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}), \\ L_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\{f(y_i|\boldsymbol{\theta})\}, \\ U_{n,k}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \frac{1}{f(y_i|\boldsymbol{\theta})} \left\{ \frac{\partial}{\partial \theta_k} f(y_i|\boldsymbol{\theta}) \right\}, \\ I_{n,j,k}(\boldsymbol{\theta}) &= n\, E\left[\frac{\partial}{\partial \theta_k} \log\{f(y_i|\boldsymbol{\theta})\} \frac{\partial}{\partial \theta_j} \log\{f(y_i|\boldsymbol{\theta})\} \right]. \end{aligned} \tag{8.35}$$

If $Y_1, \ldots, Y_n$ are independent but not *iid*,

$$\ell_n(\theta) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}),$$

$$\begin{aligned} L_n(\theta) &= \sum_{i=1}^{n} \log\{f_i(y_i|\boldsymbol{\theta})\}, \\ U_{n,k}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \frac{1}{f_i(y_i|\boldsymbol{\theta})} \left\{ \frac{\partial}{\partial \theta_k} f_i(y_i|\boldsymbol{\theta}) \right\}, \\ I_{n,j,k}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta_k} \log\{f_i(y_i|\boldsymbol{\theta})\} \frac{\partial}{\partial \theta_j} \log\{f_i(y_i|\boldsymbol{\theta})\} \right]. \end{aligned} \tag{8.36}$$

Note that (8.36) is very similar to (8.35), but the change in form for the elements of the expected information matrix $I_n(\boldsymbol{\theta})$ is important. Also note here that we are writing the expected information with the index n to emphasize that what we have is the *total information* for a sample. It is not unusual for authors to write the expected information $I(\boldsymbol{\theta})$ in terms of a single random variable (and then, since nearly everyone presents the *iid* case, $I_n(\boldsymbol{\theta}) = nI(\boldsymbol{\theta})$). Terms depending on n then just appear or disappear in the stated asymptotic results relative to the information, which I have always found a source of confusion in moving from material that presents proofs of asymptotic properties to making use of those results in practice. To clarify, focus on the total information in a sample from the outset, and make note of the fact that for the *iid* case this turns out to be n times a finite constant while, in the independent but not *iid* case this must be represented as a sum over n terms. Now, what is necessary for the variance matrix of an asymptotic distribution to be given by the inverse (total) information matrix is that, for some function of n that tends to infinity with n, $h(n)$ say, for $j, k = 1, \ldots, p$, and for some constant $\tilde{I}_{j,k}(\boldsymbol{\theta})$,

$$\frac{1}{h(n)} I_{n,j,k}(\boldsymbol{\theta}) \xrightarrow{p} \tilde{I}_{j,k}(\boldsymbol{\theta}),$$

as $n \to \infty$ and such that the $p \times p$ matrix $\tilde{I}(\boldsymbol{\theta})$ with j, k^{th} element $\tilde{I}_{j,k}(\boldsymbol{\theta})$ is a covariance matrix. Note that, since $h(n) \to \infty$, this implies that $I_{n,j,k} \to \infty$

for $j,\, k = 1, \ldots, p$. That is, the total information in a sequence of increasing sample sizes tends to infinity. In proving theoretical results, such as asymptotic normality of maximum likelihood estimators, the sequence of values $h(n)$ then typically appears as part of the "normalizing constant" or "standardization term" and the fixed matrix $\tilde{I}(\boldsymbol{\theta})$ as the covariance of the limiting distribution, for example,

$$\frac{\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}}{\sqrt{h(n)}} \xrightarrow{d} N(\mathbf{0},\, \tilde{I}^{-1}(\boldsymbol{\theta})).$$

In the *iid* case, and often for independent but not *iid* situations, the appropriate function is $h(n) = n$. This is not always the case, however, as shown in Example 6.5 of Lehmann (1983).

For inference, the covariance matrix we would like to use is then $I_n^{-1}(\boldsymbol{\theta}) = [1/h(n)]\tilde{I}^{-1}(\boldsymbol{\theta})$, but this must itself usually be estimated. An appropriate estimator is generally determined by a combination of two factors.

1. Whether the random variables $Y_1, \ldots, Y_n$ are *iid* or are independent but not identically distributed.

2. Whether or not a closed form is available for terms in the total information. That is, whether or not a functional expression, depending only on a variable y_i, is available for the contribution of the i^{th} random variable to the total information, namely,

$$I_{i,j,k}(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial\theta_j}\log\{f_i(y_i|\boldsymbol{\theta})\}\,\frac{\partial}{\partial\theta_k}\log\{f_i(y_i|\boldsymbol{\theta})\}\right],$$

   for $j,\, k = 1, \ldots, p$. In what follows we will assume that $I_{n,j,k}(\boldsymbol{\theta}) = \sum_i I_{i,j,k}(\boldsymbol{\theta})$, and that $I_n(\boldsymbol{\theta})$ is the $p \times p$ matrix with j, k^{th} element $I_{n,j,k}(\boldsymbol{\theta})$.

Depending on the resolution of these two factors (i.e., what is available), the following possible estimators for $I_n(\boldsymbol{\theta})$ may be used in application.

1. Random variables *iid* and closed form available for $I_{j,k} = I_{i,j,k}$.
   Here, $I_{i,j,k}(\boldsymbol{\theta})$ does not depend on i, $h(n) = n$, and $\tilde{I}(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})$ is the expected information for a single random variable. In this situation one would typically estimate $I_n(\boldsymbol{\theta}) = nI(\boldsymbol{\theta})$ with $nI(\hat{\boldsymbol{\theta}}_n)$.

2. Random variables are independent but not identically distributed and a closed form is available for $I_{i,j,k}$.
   Here, $h(n)$ is often n, although this need not always be the case, and the existence of $\tilde{I}(\boldsymbol{\theta})$ is needed for theoretical results, but is typically never determined in application. Here, a common estimator of $I_n(\boldsymbol{\theta})$ would be $I_n(\hat{\boldsymbol{\theta}})$, the $p \times p$ matrix with j, k^{th} element $I_{n,j,k}(\hat{\boldsymbol{\theta}}_n) = \sum_i I_{i,j,k}(\hat{\boldsymbol{\theta}})$.

3. Random variables *iid* but closed form unavailable for $I_{i,j,k}(\boldsymbol{\theta})$.
   Here, a common estimator of $I_n(\boldsymbol{\theta})$ is the so-called *observed* information $I_n^{ob}(\hat{\boldsymbol{\theta}}_n)$, in which
   $$I_{i,j,k}^{ob}(\hat{\boldsymbol{\theta}}_n) = \frac{\partial}{\partial\theta_j}\log\{f(y_i|\boldsymbol{\theta})\}\frac{\partial}{\partial\theta_k}\log\{f(y_i|\boldsymbol{\theta})\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n},$$
   $I_{n,j,k}^{ob}(\hat{\boldsymbol{\theta}}_n) = \sum_i I_{i,j,k}^{ob}(\hat{\boldsymbol{\theta}}_n)$, and $I_n^{ob}(\hat{\boldsymbol{\theta}}_n)$ is the $p \times p$ matrix with j, k^{th} element $I_{n,j,k}^{ob}(\hat{\theta}_n)$.

4. Random variables are independent but not identically distributed and no closed form is available for $I_{i,j,k}$.
   In this case, a typical estimator of $I_n(\boldsymbol{\theta})$ is the observed information matrix defined as in item 3 immediately above, except that,
   $$I_{i,j,k}^{ob}(\hat{\boldsymbol{\theta}}_n) = \frac{\partial}{\partial\theta_j}\log\{f_i(y_i|\boldsymbol{\theta})\}\frac{\partial}{\partial\theta_k}\log\{f_i(y_i|\boldsymbol{\theta})\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}.$$

Finally, also note that we have been using θ as a generic fixed parameter in this sub-section, while we have also used it as the natural parameter for

exponential and exponential dispersion families in previous portions of these notes. To avoid confusion, consider an exponential dispersion family

$$f(y_i|\theta_i) = \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right],$$

from which we have formed a generalized linear model with link function $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. For this model, the θ of the current section would be $\boldsymbol{\beta}$ and the notation $f_i(\cdot)$ would incorporate the function of $\boldsymbol{\beta}$ that gives the exponential family θ_i as $\theta_i = b'^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$.

8.3.3 Properties of Maximum Likelihood Estimators

Properties of maximum likelihood estimators are developed under sets of technical conditions called "regularity conditions". There are a wide variety of regularity conditions that have been developed, and different sets of conditions are needed to prove different results about maximum likelihood estimators under various models. It is not our intention to catalog all of these here. But from the standpoint of *modeling* there are several of these conditions that are important because they are directly connected to the properties dictated by model structure. We will break these into two sets. The first are conditions that are sufficient to allow us to *find* a maximum likelihood estimator and guarantee that this estimator is consistent for θ. The second set of conditions are additional restrictions that allow us to derive asymptotic results for either a sequence of consistent estimators, or for the random version of the likelihood or loglikelihood functions themselves. In particular, the regularity conditions given are sufficient to demonstrate normal limit distributions for a consistent sequence of estimators, thus providing *inferential* quantities.

Regularity Conditions Set 1:

1. The distributions of $Y_1, \ldots, Y_n$ are distinct in that different values of the parameter θ necessarily lead to different distributions. This can be important, for example in the mixture distributions of Section 7.4.2. There, it is important that different values of λ in the mixing distributions $g(\boldsymbol{\theta}|\lambda)$ lead to different mixture distributions $h(\boldsymbol{y}|\lambda)$. This general issue is called *identifiability* of parameters in a model formulation.

2. The distributions of $Y_1, \ldots, Y_n$ have common support independent of θ. Note that, under independence, this implies the positivity condition mentioned earlier.

3. The true value of the parameter, θ_0 lies in the interior of an open interval contained in the parameter space Θ. Note that this does not necessarily imply that Θ *is* an open interval, only that it contains an open interval for which θ_0 is an interior point.

4. For almost all $\boldsymbol{y}$ the density or mass function $f(\boldsymbol{y}|\theta)$ is differentiable with respect to all elements of θ. As an advanced note, our statement of this condition is overly restrictive for the following results to apply to at least *some* of the elements of θ.

We may now present a result, which is somewhat of a conglomeration of various theorems and corollaries from Lehmann (1983, Chapter 6).

Likelihood Theorem 1

If $Y_1, \ldots, Y_n$ are independent random variables for which the four conditions

listed above hold, then a sequence of values $\{\hat{\theta}_n\}$ exists which solve the likelihood equations,

$$\frac{\partial}{\partial \theta_k} \ell_n(\theta) = 0; \quad k = 1, \ldots, p, \tag{8.37}$$

or, equivalently,

$$\frac{\partial}{\partial \theta_k} L_n(\theta) = 0; \quad k = 1, \ldots, p, \tag{8.38}$$

and is a consistent sequence of estimators for θ.

This result is essentially that of Theorem 2.2 in Chapter 6 of Lehmann (1983, p. 413) which is presented for the *iid* case with a scalar parameter. Notice that this theorem provides for a consistent set of solutions to the likelihood equations only. It does not indicate that such solutions are either unique or are a maximum likelihood estimator. We give two corollaries, the first of which is really a modification of the basic result, to demonstrate the complexity of relating specific regularity conditions with specific results.

Corollary 1: (cf. Lehmann, 1983, Corollary 2.1, p. 410)
Under conditions 1 through 2, if the parameter space Θ is finite (meaning that θ can take on only a finite number of values), then a sequence of unique maximum likelihood estimates exists and is consistent for θ.

Notice here that we have not said these estimates are solutions to the likelihood equations, and have used the rather strong restriction that Θ is finite. I have included this result to indicate the role of smoothness conditions such as 4, and potential difficulties caused by "parameters on the boundary of the parameter space", which are eliminated from consideration by condition 3. The assumption of a finite parameter space in Corollary 1 means neither condition 3 nor condition 4 are needed for the result.

Corollary 2: (cf. Lehmann, 1983, Corollary 2.2, p. 414)
Under conditions 1 through 4, if the likelihood equation has a unique root for each n, then that sequence of estimators is consistent. If, in addition, the parameter space Θ *is* an open interval (rather than only containing an open interval) then that sequence of roots is the sequence of maximum likelihood estimators (i.e., the sequence maximizes the likelihood).

Note here that we have assumed uniqueness rather than giving it as a consequence. For models with open parameter spaces, when the likelihood equations have a unique root, then that root provides the unique maximum likelihood estimator which is consistent. When the likelihood equations have multiple roots, the game becomes to determine which sequence of roots is consistent and efficient (and this may not always be the sequence of maximum likelihood estimators, even if they exist). In practice, when there are multiple solutions to the likelihood equations (implying local maxima and, hence also local minima) it is often time to examine the behavior of the likelihood or log likelihood in more detail. Note however, that a result analogous to the fundamental theorem of generalized least squares (Section 8.2.2) is available (e.g., Lehmann 1983, Chapter 6.3; Kendall and Stuart 1960, Chapter 18.21) given a $\sqrt{n}$ consistent starting value. Since situations in which the likelihood equations contain multiple roots are typically complex models for which even consistent starting values are difficult to determine, I have found this of little practical use.

To try and provide some order to this array of conditions and results, consider that there are four distinct issues involved:

1. The existence of a maximum likelihood estimator (or sequence of estimators).

2. The existence of roots of the likelihood equations (or sequence of roots).

3. Uniqueness of estimators (or sequences of estimators).

4. Consistency of sequences of estimators.

Now, none of these necessarily implies any of the others, except under various conditions. Kraft and Lecam (1956) provide an example of a multinomial with certain specification for the parameters for which a unique maximum likelihood estimator exists but is not consistent, but also for which a consistent root of the likelihood equations does exist. In short, the happy situation is provided by Corollary 2 in which unique roots correspond to unique maximum likelihood estimators which are consistent.

We turn now to a summarization of the conditions that lead to inference based on properties of the mle, which is a form of what is known as *Wald Theory*. The number of particular technical conditions for particular situations (*iid* single parameter, *iid* multiple parameters, indep. but not *iid*, etc.) can become quite confusing. We will give appropriate conditions only for the *iid* case with multiple parameters here, to communicate what needs to be achieved by such conditions.

Regularity Conditions Set 2:

For $Y_1, \ldots, Y_n$ *iid* with density or mass functions $f_i(y_i|\theta)$, and $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)^T$,

1.

$$E\left[\frac{\partial}{\partial \theta_k} log\{f(Y|\boldsymbol{\theta})\}\right] = 0; \quad k = 1, \ldots, p.$$

2.

$$\begin{aligned} I_{j,k}(\boldsymbol{\theta}) &\equiv E\left[\frac{\partial}{\partial\theta_k}\log\{f(Y|\boldsymbol{\theta})\}\,\frac{\partial}{\partial\theta_j}\log\{f(Y|\boldsymbol{\theta})\}\right] \\ &= -E\left[\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log\{f(Y|\boldsymbol{\theta})\}\right]. \end{aligned}$$

3. The $p \times p$ matrix $I(\boldsymbol{\theta})$ with kj^{th} element $I_{j,k}$ is positive definite.

4. There exist functions $M_{k,j,s}(\cdot)$ such that

$$\left|\frac{\partial^3}{\partial\theta_k\partial\theta_j\partial\theta_s}\log\{f(y|\boldsymbol{\theta})\}\right| \leq M_{k,j,s}(y) \quad \text{for all } \boldsymbol{\theta} \in \Theta,$$

and

$$E\left[M_{k,j,s}(Y)\right] < \infty.$$

Now, conditions 1, 2, and 3 above will often be written in terms of other conditions that lead to these as results, generally expressed in terms of the first two derivatives of $\log\{f(Y|\boldsymbol{\theta})\}$ in a way similar to the expression of condition 4 in the above list. For example, in the *iid* case with a single scalar parameter, conditions 1 and 2 above are often replaced with,

$$\left|\frac{\partial f(y|\theta)}{\partial\theta}\right| \leq g(y)$$

$$\left|\frac{\partial^2 f(y|\theta)}{\partial\theta^2}\right| \leq h(y)$$

such that $\int g(y)\,dy < \infty$ and $\int h(y)\,dy < \infty$. Then, in this case, conditions 1 and 2 follow as a result of being able to differentiate under the integral. Conditions 3 and 4 in the above list ensure that the derivatives of the log likelihood, considered as functions of the random variables $Y_1, \ldots, Y_n$, have finite but positive variances.

Likelihood Theorem 2

Let $Y_1, \ldots, Y_n$ be iid with densities $f(y_i|\boldsymbol{\theta})$ such that the conditions given in Regularity Conditions Sets 1 and 2 all hold. Then there exists a sequence of solutions to the likelihood equations $\{\hat{\boldsymbol{\theta}}_n\}$ such that

(i) $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$.

(ii) $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is asymptotically normal with mean $\mathbf{0}$ and covariance $n\{I(\boldsymbol{\theta})\}^{-1}$.

(iii) For $k = 1, \ldots, p$, $\hat{\theta}_{n,k}$ is asymptotically efficient in that,

$$\sqrt{n\{I(\boldsymbol{\theta})\}^{-1}_{k,k}}(\hat{\theta}_{n,k} - \theta_k) \xrightarrow{\mathcal{L}} N(0,\, 1),$$

where $\{I(\boldsymbol{\theta})\}^{-1}_{k,k}$ is the k, k^{th} element of the matrix $I(\boldsymbol{\theta})^{-1}$.

This result then provides an immediate means for constructing approximate intervals for individual elements of $\boldsymbol{\theta}$. A generalization that allows tests of hypotheses and the construction of confidence regions will be provided in Section 8.3.4 where we discuss Wald theory.

An example will serve to illustrate some of what can "go wrong" in asymptotic normality for maximum likelihood estimators. First, consider random variables $Y_1, \ldots, Y_n$ such that $Y_i \sim iid\, U(0,\, \theta)$. Here, $f(y_i|\theta) = (1/\theta)I(0 < y_i < \theta)$, where $I(\cdot)$ is the indicator function, not information. The log likelihood and its derivatives are then,

$$\begin{aligned} L_n(\theta) &= -n\log\{\theta\}, \\ \frac{\partial}{\partial\theta}L_n(\theta) &= \frac{-n}{\theta}, \\ \frac{\partial^2}{\partial\theta^2}L_n(\theta) &= \frac{n}{\theta^2}. \end{aligned}$$

Now, the likelihood equation (first derivative of L_n) clearly has no root. Thus, the maximum likelihood estimator, if it exists, cannot be obtained as a solution to the likelihood equation. That a maximum likelihood estimator does indeed exist is immediate from $\ell(\theta) = 1/\theta^n$, which gives

$$\ell(\max\{y_1, \ldots, y_n\}) \geq \ell(\theta); \quad \text{any } \theta \in (0, \infty).$$

The asymptotics of our Likelihood Theorem 2 certainly do not apply in this case. That does not mean, however, that asymptotics are not available. Only that they are not available from theorems on "regular" problems. To see what can be done, note first that, if $Y_{[n]}$ denotes the largest order statistic from a $U(0, \theta)$ distribution, then

$$Pr(Y_{[n]} \leq y) = Pr(Y_1, \ldots, Y_n \leq y) = \frac{y^n}{\theta^n}.$$

Thus,

$$\begin{aligned} Pr\left[n\{\theta - Y_{[n]}\} \leq y\right] &= Pr\left[Y_{[n]} > \theta - y/n\right] \\ &= 1 - Pr\left[Y_{[n]} \leq \theta - y/n\right] \\ &= 1 - \left(\frac{\theta - y/n}{\theta}\right)^n. \end{aligned}$$

Taking the limit as $n \to \infty$,

$$\begin{aligned} \lim_{n\to\infty} 1 - \left(\frac{\theta - y/n}{\theta}\right)^n &= 1 - \lim_{n\to\infty} \left(\frac{\theta - y/n}{\theta}\right)^n \\ &= 1 - \lim_{n\to\infty} \left(1 - \frac{y}{n\theta}\right)^n \\ &= 1 - \lim_{n\to\infty} \left(1 + \frac{-y/\theta}{n}\right)^n \\ &= 1 - \exp\{-y/\theta\}, \end{aligned}$$

the last line following from $\lim\{1 + (x/n)\}^n = \exp(x)$ for all x.

Thus, the maximum likelihood estimator for this problem is $\hat{\theta}_n = Y_{[n]}$ and this estimator has an asymptotic distribution given as,

$$n\{\theta - \hat{\theta}_n\} \xrightarrow{\mathcal{L}} E(0, \theta),$$

where $E(0, \theta)$ denotes a exponential $(0, \theta)$ distribution. The regular theory did not apply in this case because of condition 2 in Regularity Conditions Set 1 of Section 8.3.3.

Two additional properties of maximum likelihood estimators are worthy of mention to close out our discussion of this subsection.

1. If a given scalar parameter θ (which may be an element of the parameter vector $\boldsymbol{\theta}$) has a single sufficient statistic $T(\boldsymbol{y})$ say, then the maximum likelihood estimator must be a function of that sufficient statistic. If that sufficient statistic is minimal and complete, the the maximum likelihood estimator is unique. If the maximum likelihood estimator is unbiased then it is the UMVU (e.g., Kendall and Stuart 1960, Chapters 18.4-18.7). This property could have implications, for example, in mean value parameterization 2 for exponential families (e.g., Lindsey 1996, p. 307).

2. Maximum likelihood estimators possess a property called *invariance* that is very useful but is not, in general, a property of UMVU estimators (unless, of course, they happen to also be maximum likelihood estimators). The invariance property of maximum likelihood estimators can be stated as, if $\hat{\boldsymbol{\theta}}_n$ is an mle of $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)^T$, and $g(\cdot)$ is a real-valued function of $\boldsymbol{\theta}$, then

$$g(\hat{\boldsymbol{\theta}}_n) \text{ is an mle of } g(\boldsymbol{\theta}).$$

Invariance is often combined with what we will present as the *delta method* in the next subsection to derive the limit distribution of asymptotically normal estimators.

8.3.4 Wald Theory Inference

What we will present here in terms of inference from properties (essentially the asymptotic normality) of maximum likelihood estimators applies equally well to other asymptotically normal estimators, such as the generalized least squares estimators of Section 8.2.2. To make the results in what follows here applicable to generalized least squares estimators, simply replace the inverse information matrix with with the quantity $(\sigma^2/n)\Sigma_\beta^{-1}$ from the Fundamental Theorem of Generalized Least Squares.

Wald Theory Main Result

If $\{\hat{\boldsymbol{\theta}}_n\}$ is a sequence of maximum likelihood estimators of $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)^T$ for which the conditions of Likelihood Theorem 2 apply, then,

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T I_n(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \chi^2_p, \tag{8.39}$$

where χ^2_p is a Chi-squared random variable with p degrees of freedom. For proof of this result see, e.g., Serfling (1980, Chapter 4.4), but note that it is often (usually) written with $I_n(\hat{\boldsymbol{\theta}}_n) = nI(\hat{\boldsymbol{\theta}}_n)$ which is appropriate in the *iid* setting.

We are interested in using this result to obtain tests and interval estimates for sets of elements of $\boldsymbol{\theta}$. We first will consider hypotheses about the $p-$dimensional parameter vector $\boldsymbol{\theta}$. To formulate a hypothesis about $\boldsymbol{\theta}$ we

specify a set of $r \leq p$ restrictions of the form

$$R_j(\boldsymbol{\theta}) = 0; \quad j = 1, \ldots, r.$$

Example 8.10

1. Let $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \theta_3)^T$. Specify the hypothesis, $H_0 : \theta_1 = \theta_1^0,\ \theta_2 = \theta_2^0,\ \theta_3 = \theta_3^0$ for particular values θ_1^0, θ_2^0, θ_3^0. This hypothesis corresponds to the restrictions,

$$\begin{aligned} R_1(\boldsymbol{\theta}) &= \theta_1 - \theta_1^0 = 0 \\ R_2(\boldsymbol{\theta}) &= \theta_2 - \theta_2^0 = 0 \\ R_3(\boldsymbol{\theta}) &= \theta_3 - \theta_3^0 = 0. \end{aligned}$$

2. Let $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \theta_3)^T$. Specify the hypothesis, $H_0 : \theta_1 = \theta_1^0$. This corresponds to the single restriction,

$$R_1(\boldsymbol{\theta}) = \theta_1 - \theta_1^0 = 0,$$

with unrestricted parameters θ_2 and θ_3.

3. Let $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \theta_3, \theta_4)^T$. Specify the hypothesis $H_0 : \theta_1 = \theta_2,\ \theta_3 = \theta_4$. This corresponds to the restrictions,

$$\begin{aligned} R_1(\boldsymbol{\theta}) &= \theta_1 - \theta_2 = 0 \\ R_2(\boldsymbol{\theta}) &= \theta_3 - \theta_4 = 0. \end{aligned}$$

In these examples, 1 would be called a *simple hypothesis* while 2 and 3 would be called *composite* hypotheses (the distinction rests on whether the number of restrictions is $r = p$ or $r < p$).

The Wald Theory Main Result combined with results for quadratic transformations of normally distributed random variables (e.g., Serfling 1980, Chapter 3.5) leads to the following result for forming a test statistic. This result will also be used to form joint confidence regions for subsets of the parameter vector $\boldsymbol{\theta}$.

Wald Theory Tests

Let

$$b(\boldsymbol{\theta}) \equiv (R_1(\boldsymbol{\theta}), \ldots, R_r(\boldsymbol{\theta}))^T,$$

be an $r \times 1$ vector of defined restrictions on model parameters. Let $C(\boldsymbol{\theta})$ be an $r \times p$ matrix with jk^{th} element

$$C_{k,j} = \frac{\partial}{\partial \theta_k} R_j(\boldsymbol{\theta}).$$

Then, under $R_1(\boldsymbol{\theta}) = \ldots, R_r(\boldsymbol{\theta}) = 0$ (i.e., under H_0),

$$W_n = b^T(\hat{\boldsymbol{\theta}}_n) \left[C(\hat{\boldsymbol{\theta}}_n) I_n^{-1}(\hat{\boldsymbol{\theta}}_n) C^T(\hat{\boldsymbol{\theta}}_n) \right]^{-1} b(\hat{\boldsymbol{\theta}}_n) \xrightarrow{\mathcal{L}} \chi^2_r, \tag{8.40}$$

Revisiting the cases given in Example 8.10, this result plays out as follows. In these examples, let

$$I_n^{-1}(\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} i^{11} & i^{12} & \ldots & i^{1p} \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ i^{p1} & i^{p2} & \ldots & i^{pp} \end{pmatrix},$$

but note that this will be a symmetric matrix so that $i^{kj} = i^{jk}$.

1. Here,

$$C(\boldsymbol{\theta}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

so that

$$C(\hat{\boldsymbol{\theta}}_n) I^{-1}(\hat{\boldsymbol{\theta}}_n) C^T(\hat{\boldsymbol{\theta}}_n) = I^{-1}(\hat{\boldsymbol{\theta}}_n).$$

Then, also using the fact that $I^{-1}(\hat{\boldsymbol{\theta}}_n)$ is symmetric,

$$\begin{aligned} W_n &= (\hat{\theta}_{n,1} - \theta_1^0)^2 I_{n,11} + (\hat{\theta}_{n,2} - \theta_2^0) I_{n,22} \\ &+ (\hat{\theta}_{n,3} - \theta_3^0) I_{n,33} \\ &+ 2(\hat{\theta}_{n,1} - \theta_1^0)(\hat{\theta}_{n,2} - \theta_2^0) I_{n,12} \\ &+ 2(\hat{\theta}_{n,1} - \theta_1^0)(\hat{\theta}_{n,3} - \theta_3^0) I_{n,13} \\ &+ 2(\hat{\theta}_{n,2} - \theta_2^0)(\hat{\theta}_{n,3} - \theta_3^0) I_{n,23} \end{aligned}$$

2. Here, $C(\boldsymbol{\theta}) = (1,\, 0,\, 0)$ so that,

$$C(\hat{\boldsymbol{\theta}}_n) I^{-1}(\hat{\boldsymbol{\theta}}_n) C^T(\hat{\boldsymbol{\theta}}_n) = i^{11},$$

and,

$$W_n = (\hat{\theta}_{n,1} - \theta_1^0)^2 \frac{1}{i^{11}}.$$

Note: Compare this to the square of a normal-theory test statistic.

3. Here,

$$C(\theta) = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

and,

$$\begin{aligned} W_n &= (\hat{\theta}_{n,1} - \hat{\theta}_{n,2})^2 \left(i^{11} + i^{22} - 2i^{12} \right) \\ &+ 2(\hat{\theta}_{n,1} - \hat{\theta}_{n,2})(\hat{\theta}_{n,3} - \hat{\theta}_{n,4})(i^{13} - i^{23} - i^{14} + i^{24}) \\ &+ (\hat{\theta}_{n,3} - \hat{\theta}_{n,4})^2 \left(i^{33} + i^{44} - 2i^{34} \right). \end{aligned}$$

Wald Theory Intervals

To develop intervals along the same lines as tests, let $\boldsymbol{\theta}^0$ denote the true parameter value, and take

$$\begin{aligned} R_1(\boldsymbol{\theta}) &= \theta_1 - \theta_1^0 = 0 \\ &\cdot \\ &\cdot \\ &\cdot \\ R_r(\boldsymbol{\theta}) &= \theta_r - \theta_r^0 = 0, \end{aligned}$$

for $r \leq p$.

Then $b(\boldsymbol{\theta})$ from the test result is a subset of $\boldsymbol{\theta} - \boldsymbol{\theta}^0$, and an approximate $100(1-\alpha)\%$ confidence region (for $\boldsymbol{\theta}^0$) is given by

$$\left\{\boldsymbol{\theta}^0 : b^T(\hat{\boldsymbol{\theta}}_n)\left[C(\hat{\boldsymbol{\theta}}_n)I_n^{-1}(\hat{\boldsymbol{\theta}}_n)C^T(\hat{\boldsymbol{\theta}}_n)\right]^{-1} b(\hat{\boldsymbol{\theta}}_n) \leq \chi^2_{r,1-\alpha}\right\}. \tag{8.41}$$

What happens if $r = 1$ and, for example, $R_1(\boldsymbol{\theta}) = \theta_1 - \theta_1^0 = 0$? Then,

$$\left[C(\hat{\boldsymbol{\theta}}_n)I_n^{-1}(\hat{\boldsymbol{\theta}}_n)C^T(\hat{\boldsymbol{\theta}}_n)\right]^{-1} = \frac{1}{i^{11}},$$

and the confidence region becomes,

$$\{\theta^0 : (\hat{\theta}_{n,1} - \theta_1^0)\frac{1}{i^{11}}(\hat{\theta}_{n,1} - \theta_1^0) \leq \chi^2_{1,1-\alpha}\},$$

or, taking square roots on both sides of the inequality in this set,

$$\begin{aligned} \frac{(\hat{\theta}_{n,1} - \theta_1^0)}{\sqrt{i^{11}}} &\leq z_{1-\alpha/2} \\ \frac{(\hat{\theta}_{n,1} - \theta_1^0)}{\sqrt{i^{11}}} &\geq -z_{1-\alpha/2} \end{aligned}$$

which implies that

$$\begin{aligned} \theta_1^0 &\geq \hat{\theta}_{n,1} - z_{1-\alpha/2}\sqrt{i^{11}} \\ \theta_1^0 &\leq \hat{\theta}_{n,1} + z_{1-\alpha/2}\sqrt{i^{11}} \end{aligned}$$

The Delta Method

As we have seen, Wald theory is based on asymptotic normality of (in particular) maximum likelihood estimators. Now, likelihoods are invariant to parameter transformation. What was introduced in Chapter 8.3.3 as the invariance property indicates that functions of maximum likelihood estimates are also maximum likelihood estimates of the same functions of parameters, and are also asymptotically normal. But derivatives of likelihood functions (or log likelihood functions) are clearly not invariant to parameter transformation, and plugging a function of parameter values into the inverse information matrix does not provide the covariance matrix of the limit distribution of a transformed sequence of maximum likelihood estimators. That is, if $\hat{\boldsymbol{\theta}}_n$ is a maximum likelihood estimate of a parameter $\boldsymbol{\theta}$ with inverse information matrix $I^{-1}(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ a real-valued function of $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\psi}}_n = g(\hat{\boldsymbol{\theta}}_n)$ is a maximum likelihood estimate of $\boldsymbol{\psi}$ and is asymptotically normal, but $I^{-1}(g^{-1}(\boldsymbol{\psi}))$ is not the asymptotic variance of $\hat{\boldsymbol{\psi}}$.

What is often called the *delta method* provides a method by which the asymptotic covariance matrix of a function of an asymptotically normal quantity can be eaisly obtained. Note that the delta method applies more generally than only in the case of maximum likelihood estimation and could be used, for example, in conjunction with a generalized least squares estimator or any estimator shown to be asymptotically normal. It fits so nicely with the invariance property of maximum likelihood esitmates, however, that it seems natural to present it in that context. We now state the result in a manner similar to a combination of what is given in Serfling (1980, page 122) and Lehmann (1983, page 344).

Let $\hat{\boldsymbol{\theta}}_n$ be a sequence of asymptotically normal estimators of a parameter

$\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)$ with mean $\boldsymbol{\theta}$ and "covariance" matrix $c_n^2\Sigma$ (i.e., $(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})/c_n$ converges in law to the multivariate normal distribution with mean $\mathbf{0}$ and covariance Σ). Let $g_1(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta})$ be a set of real-valued functions of $\boldsymbol{\theta}$ with $r \leq p$ and such that each function $g_k; k = 1, \ldots, r$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}$. Let $\boldsymbol{D}$ be an $r \times p$ matrix with k, j^{th} element $\partial g_k/\partial\theta_j$. Then the vector $g_1(\hat{\boldsymbol{\theta}}), \ldots, g_r(\hat{\boldsymbol{\theta}})$ is asymptotically normal with mean $g_1(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta})$ and "covariance" matrix $c_n^2\boldsymbol{D}\Sigma\boldsymbol{D}^T$.

In likelihood estimation and inference, $c_n^2\Sigma$ is typically $I_n^{-1}(\boldsymbol{\theta})$, and in practice this is estimated using one of the forms described in Chapter 8.3.2. Similarly, the matrix $\boldsymbol{D}$ is estimated using $\hat{\boldsymbol{\theta}}_n$ as a plug-in estimator of $\boldsymbol{\theta}$. Consistency of these estimators allows the asymptotic result to be applied without modification.

8.3.5 Likelihood Inference

The name of this subsection is perhaps something of a misnomer, since everything that has been discussed in Chapter 8.3 could be considered a part of likelihood estimation and inference. The title is given, however, to distinguish inference based on the asymptotic normality of maximum likelihood estimates (i.e., Wald Theory) from the topic of this subsection, which is inference based on asymptotic properties of the log likelihood function itself. The basis of this type of inference is the asymptotic distribution of the likelihood ratio statistic.

To set the stage, consider two models of the same form (i.e., the same random component) but of differing parameter spaces. Specifically, suppose we have a "full model" of the form

$$L_n(\boldsymbol{\theta}) = log\{f(\boldsymbol{y}|\boldsymbol{\theta}); \quad \theta \in \Theta,$$

and a "reduced model" of the form,

$$L_n(\boldsymbol{\theta}_0) = log\{f(\boldsymbol{y}|\boldsymbol{\theta}_0); \quad \theta_0 \in \Theta_0,$$

where $\Theta_0 \subset \Theta$. This last condition is crucial, and is called the condition of "nested parameter spaces". For example, if we have two independent groups of random variables $\{Y_{1,i} : i = 1, \ldots, n_1\}$ and $\{Y_{2,i} : i = 1, \ldots, n_2\}$ such that within each group we assume an *iid* normal distribution, then we might have the following possible model structures.

1. Model 1.
   $Y_{1,i} \sim iid\, N(\mu_1, \sigma_1^2)$ and $Y_{2,i} \sim iid\, N(\mu_2, \sigma_2^2)$

2. Model 2.
   $Y_{1,i} \sim iid\, N(\mu_1, \sigma^2)$ and $Y_{2,i} \sim iid\, N(\mu_2, \sigma^2)$

3. Model 3.
   $Y_{1,i} \sim iid\, N(\mu, \sigma_1^2)$ and $Y_{2,i} \sim iid\, N(\mu, \sigma_2^2)$

4. Model 4.
   $Y_{1,i} \sim iid\, N(\mu, \sigma^2)$ and $Y_{2,i} \sim iid\, N(\mu, \sigma^2)$

Here, all other models would be "nested" within Model 1. Model 4 would be nested within either Model 2 or Model 3. But Model 2 would not be nested within Model 3, nor *vice versa*. The procedure we are about to discuss only applies to the comparison of nested models. What results in nested parameter spaces is not simply $\Theta_0 \subset \Theta$, but that the parameter $\boldsymbol{\theta}$ is the same for both full and reduced models. In particular, models with different random components or response distributions are not amenable to comparison using the procedures of this subsection.

Assume regularity conditions similar to those given previously in Section 8.3.3. Given models for independent random variables that differ only through nested parameter spaces $\Theta_0 \subset \Theta$, we have a result that will form the basis for both tests and intervals, in a manner similar to the Wald Theory Main Result for the inference of Section 8.3.4.

Likelihood Ratio Main Result

Let $\dim\{\Theta\} = p$ and $\dim\{\Theta_0\} = r$, and,

$$\hat{\boldsymbol{\theta}}_n = \sup_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta}) \qquad \tilde{\boldsymbol{\theta}}_n = \sup_{\boldsymbol{\theta}\in\Theta_0} L_n(\boldsymbol{\theta}).$$

Then, assuming that $\boldsymbol{\theta} \in \Theta_0$ (the reduced model),

$$T_n \equiv -2\left\{L_n(\tilde{\boldsymbol{\theta}}_n) - L_n(\hat{\boldsymbol{\theta}}_n)\right\} \xrightarrow{\mathcal{L}} \chi^2_{p-r}. \tag{8.42}$$

It is worthy of note here that, while this result is closely related to what were given as Likelihood Theorem 2 and the Main Wald Theory result, it is a distinct result that is not a direct consequence of those previous theorems. The proof the Main Likelihood Ratio Result depends on the ability to expand the log likelihood function itself as a Taylor series, while the proof of asymptotic normality of maximum likelihood estimators (Likelihood Theorem 2) and resulting Chi-squared limiting distribution for quadratic forms of asymptotically normal estimators (the Wald Theory Main Result) depend on expanding the score function, that is, the derivative of the log likelihood.

Given the Main Likelihood Ratio Result, we have an immediate test statistic for the comparison of full, $\boldsymbol{\theta} \in \Theta$, and reduced, $\boldsymbol{\theta} \in \Theta_0 \subset \Theta$, models.

Example 8.11

Consider a situation in which we have two groups of *iid* random variables. These may be of any given distributional form. Suppose that we have a beta-binomial model such as used in Example 7.12 for the number of live young born to *Gambusia* from the San Luis Drain and Volta regions in the Central Valley of California. Consider, for this example, the two groups designated as $SLD\,R-10$ and $Volta\,R-16$ from that example. Within each group, our model results in a log likelihood function of the form of expression (7.37). Now, denote the parameters for these groups as (α_s, β_s) and α_v, β_v) for the San Luis Drain and Volta areas, respectively.

The full model consists of separate beta mixing distributions for the two groups, so that $\boldsymbol{\theta} = (\alpha_s, \beta_s, \alpha_v, \beta_v)^T$ and $\Theta \equiv \Re^+ \times \Re^+ \times \Re^+ \times \Re^+$. The reduced model consists of a single beta mixing distribution that applies to both groups with parameters α_c and β_c, say; the subscript c is for "combined". Here, $\Theta_0 \equiv \Re^+ \times \Re^+$ and we have the necessary nested parameter spaces. We could also set restrictions of the type used in case 3 of Example 8.10 and formulate the comparison of full and reduced models as a hypothesis test, but it seems preferable to approach the entire issue as one of model comparison. That is, we are comparing the full and reduced models, but are allowing no other possibilities.

Given independence among observations from the San Luis Drain and Volta areas, the log likelihood for the full model may be written as,

$$\begin{aligned} L_{n,f}(\alpha_{n,s}, \beta_{n,s}, \alpha_{n,v}, \beta_{n,v}) &= L_{n,s}(\alpha_{n,s}, \beta_{n,s}) \\ &+ L_{n,v}(\alpha_{n,v}, \beta_{n,v}), \end{aligned}$$

where $L_{n,s}$ is the log likelihood for the San Luis Drain area, and $L_{n,v}$ is that

for the Volta area. Given maximum likelihood estimates of these parameters, the maximized log likelihood for the full model becomes,

$$\begin{aligned} L_{n,f}(\hat{\alpha}_{n,s}, \hat{\beta}_{n,s}, \hat{\alpha}_{n,v}, \hat{\beta}_{n,v}) &= L_s(\hat{\alpha}_{n,s}, \hat{\beta}_{n,s}) \\ &+ L_v(\hat{\alpha}_{n,v}, \hat{\beta}_{n,v}), \end{aligned}$$

The log likelihood for the reduced model may be written as $L_{n,r}(\alpha_{n,c}, \beta_{n,c})$, with maximized value, $L_{n,r}(\hat{\alpha}_{n,c}, \hat{\beta}_{n,c})$. In practice, we obtain $\hat{\alpha}_{n,c}$ and $\hat{\beta}_{n,c}$ by combining data across groups, and fitting the same model to these combined data that we did to each group separately. The test statistic of expression (8.42) then becomes,

$$T_n = -2\left\{L_{n,r} - L_{n,f}\right\},$$

which, in this example, is compared to a χ^2 random variable with $p - r = 4 - 2 = 2$ degrees of freedom.

The Main Likelihood Ratio Result also provides a method for forming confidence regions, which is sometimes referred to as "inverting" the likelihood ratio test statistic (e.g., Hahn and Meeker 1991, pp.240-241). The concept is straightforward and based on the relation between tests and intervals. Let $\boldsymbol{\theta}_0$ be any value of $\boldsymbol{\theta}$ such that a likelihood ratio test of the form (8.42) would not reject θ_0 at the α level. That is, $\boldsymbol{\theta}_0$ is any value of $\boldsymbol{\theta}$ such that,

$$-2\left\{L_n(\boldsymbol{\theta}_0) - L_n(\hat{\boldsymbol{\theta}}_n)\right\} \leq \chi^2_{p,1-\alpha}.$$

The reason for p degrees of freedom in this expression is as follows. In the main result, we took p as the dimension of the full model parameter space Θ and r as the dimension of the reduced model parameter space Θ_0 and the likelihood ratio statistic was asymptotically χ^2 with $p - r$ degrees of freedom. Here, we have a completely specified parameter $\boldsymbol{\theta}_0$. Now, while $\boldsymbol{\theta}_0$ is a $p-$dimensional vector, it consists of only one point in $p-$dimensional space. In other words,

the dimension of Θ_0 is zero. Thus, the degrees of freedom above are $p - r = p - 0 = p$, entirely in agreement with the main result of expression (8.42).

The set of all $\boldsymbol{\theta}_0$ such that a likelihood ratio test would not reject this value (or reduced model) at the α level of significance is then a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$,

$$\left\{\boldsymbol{\theta}_0 : -2\left[L_n(\boldsymbol{\theta}_0) - L_n(\hat{\boldsymbol{\theta}}_n\right] \leq \chi^2_{p,1-\alpha}\right\}. \tag{8.43}$$

As a final comment, we will point out that the likelihood region (8.43) is invariant to parameter transformation, while the Wald theory region of (8.41) is not. This is because the likelihood and log likelihood functions are invariant to parameter transformation. That is, if $h(\boldsymbol{\theta})$ is a transformation of $\boldsymbol{\theta}$ for some continuous function $h(\cdot)$, then $L_n(h(\boldsymbol{\theta})) = L_n(\boldsymbol{\theta})$. Thus, any $\boldsymbol{\theta}_0$ that is contained in the set (8.43) corresponds to an $h(\boldsymbol{\theta}_0)$ that is also within the set. On the other hand this same property does not hold for variances, so that (8.41) is not invariant under parameter transformation. Any number of simulation studies have been conducted that indicate the likelihood region of is superior to the Wald region in maintaining nominal coverage when the two differ. When will they differ? When the likelihood surface near its maximum is not well approximated by a quadratic surface. It is also true, however, that the likelihood region of (8.43) tends to be more difficult to compute than the Wald region of (8.41), even in two dimensions.

8.3.6 Example - Generalized Linear Models

As we have discussed in lab, one typical approach for finding maximum likelihood estimates in specific problems is to form the likelihood or log likelihood, derive the likelihood equations (i.e., set the score function equal to zero) and use numerical algorithms for solving these equations. The most commonly

used algorithms are probably those we categorized as being "Newton-type" algorithms such as the Newton-Raphson and Fisher Scoring algorithms. In a few instances, some unification is possible in this procedure for an entire class of models, and we illustrate this situation with the generalized linear models introduced in Section 7.3.2. Here, we have independent random variables $Y_1, \ldots, Y_n$ with probability mass or density functions that may be written as exponential dispersion families as,

$$f(y_i|\theta_i, \phi) = \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right].$$

This constitutes the random model component. The systematic model component is specified as,

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

for a known smooth function $g(\cdot)$.

From the above model, given independence of the response variables $Y_1, \ldots, Y_n$, the log likelihood is,

$$L(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} L_i(\boldsymbol{\beta}, \phi), \tag{8.44}$$

where L_i is the contribution of the i^{th} random variable,

$$L_i(\boldsymbol{\beta}, \phi) = \phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi). \tag{8.45}$$

Expression (8.45) makes sense as a function of $\boldsymbol{\beta}$ since $E(Y_i) = \mu_i = b'(\theta_i)$ from the random component, and $g(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ from the systematic model component. That is, we have a "cascade" of simple functions connecting θ_i to μ_i to $\boldsymbol{\beta}$. This suggests that the standard chain rule of elementary calculus should be useful in deriving the derivatives of $L_i(\boldsymbol{\beta}, \phi)$ and thus also those of $L(\boldsymbol{\beta}, \phi)$ since the latter is just a sum over the former by (8.44). In particular, consider estimation of the components of $\boldsymbol{\beta}$ by deriving first the likelihood

equations. We have that

$$\frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial \beta_j} = \frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial \theta_i}\,\frac{d\theta_i}{d\mu_i}\,\frac{d\mu_i}{d\eta_i}\,\frac{\partial \eta_i}{\partial \beta_j}. \tag{8.46}$$

Now, given the random component as an exponential dispersion family, and the properties of such families, we have that,

$$\begin{aligned}
\frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial \theta_i} &= \phi\{y_i - b'(\theta_i)\} = \phi\{y_i - \mu_i\}, \\
\frac{d\theta_i}{d\mu_i} &= \frac{1}{V(\mu_i)}, \\
\\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{i,j}
\end{aligned} \tag{8.47}$$

The second line of expression (8.47) follows because $\mu_i = b'(\theta_i)$ so that $d\mu_i/d\theta_i = b''(\theta_i) = V(\mu_i)$, and the third line follows from the linear form of $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

Substituting (8.47) into (8.46) results in,

$$\frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial \beta_j} = \phi\{y_i - \mu_i\}\,\frac{1}{V(\mu_i)}\,\frac{d\mu_i}{d\eta_i}\,x_{i,j},$$

or, summing over observations,

$$\frac{\partial L(\boldsymbol{\beta},\phi)}{\partial \beta_j} = \sum_{i=1}^{n}\left[\phi\{y_i - \mu_i\}\,\frac{1}{V(\mu_i)}\,\frac{d\mu_i}{d\eta_i}\,x_{i,j}\right]. \tag{8.48}$$

At this point, although there is no clear reason to do so in the above derivations, let

$$W_i \equiv \left\{\left(\frac{d\eta_i}{d\mu_i}\right)^2 V(\mu_i)\right\}^{-1},$$

and substitute into expression (8.48) to arrive at,

$$\frac{\partial L(\boldsymbol{\beta},\phi)}{\partial \beta_j} = \sum_{i=1}^{n}\left[\phi\{y_i - \mu_i\}\,W_i\,\frac{d\eta_i}{d\mu_i}\,x_{i,j}\right]. \tag{8.49}$$

The set of likelihood equations are then given by setting (8.49) equal to zero for $j = 1, \ldots, p$.

To derive expressions for the second derivatives, make additional use of the chain rule applied to (8.46), which results in,

$$\begin{aligned}
\frac{\partial^2 L_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k} &= \frac{\partial}{\partial\beta_k}\left[\frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial\theta_i}\,\frac{d\theta_i}{d\mu_i}\,\frac{d\mu_i}{d\eta_i}\,\frac{\partial\eta_i}{\partial\beta_j}\right] \\
&= \frac{\partial^2 L_i(\boldsymbol{\beta},\phi)}{\partial\theta_i^2}\left(\frac{d\theta_i}{d\mu_i}\right)^2\left(\frac{d\mu_i}{d\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k} \\
&+ \frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial\theta_i}\,\frac{d^2\theta_i}{d\mu_i^2}\left(\frac{d\mu_i}{d\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k} \\
&+ \frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial\theta_i}\,\frac{d\theta_i}{d\mu_i}\,\frac{d^2\mu_i}{d\eta_i^2}\,\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k}. \qquad (8.50)
\end{aligned}$$

In (8.50) we would have

$$\begin{aligned}
\frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial\theta_i} &= \phi\{y_i - b'(\theta_i)\}, \\
\frac{\partial^2 L_i(\boldsymbol{\beta},\phi)}{\partial\theta_i^2} &= \frac{\partial}{\partial\theta_i}\left[\phi\{y_i - b'(\theta_i)\}\right] \\
&= -\phi\, b''(\theta_i) = -\phi\, V(\mu_i).
\end{aligned} \qquad (8.51)$$

Substituting (8.51) into (8.50) we can see that the only terms in (8.50) that depend on the response value y_i are those that involve

$$\frac{\partial L_i(\boldsymbol{\beta},\phi)}{\partial\theta_i},$$

and, since $E(Y_i) = b'(\theta_i)$, the expected value of the random version of this first derivative is 0.

Now, recall for what purpose we are finding second derivatives – for use in an iterative numerical algorithm (such as Newton-Raphson) for approximation of maximum likelihood estimates of the elements of $\boldsymbol{\beta}$. As we have seen, one of

the family of Newton-like algorithms is Fisher Scoring, in which the matrix of second derivatives (i.e., the Hessian matrix) is replaced by its negative expected value. This suggests that we write (8.50) as,

$$\begin{aligned}\frac{\partial^2 L_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k} &= -\phi V(\mu_i)\left(\frac{d\,\theta_i}{d\,\mu_i}\right)^2\left(\frac{d\,\mu_i}{d\,\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k} \\ &+ \phi\{y_i - b'(\theta_i)\,\{\text{ terms without } y_i\}\end{aligned} \tag{8.52}$$

As a result, taking the negative expectation of the random version of (8.52) results in,

$$-E\left\{\frac{\partial^2 L_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k}\right\} = \phi V(\mu_i)\left(\frac{d\,\theta_i}{d\,\mu_i}\right)^2\left(\frac{d\,\mu_i}{d\,\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k} \tag{8.53}$$

Now, use the definition of W_i given just before expression (8.49) as,

$$W_i \equiv \left\{\left(\frac{d\eta_i}{d\mu_i}\right)^2 V(\mu_i)\right\}^{-1},$$

and,

$$\frac{\partial\eta_i}{\partial\beta_j} = x_{i,j}; \quad \text{and} \quad \frac{d\,\theta_i}{d\,\mu_i} = \frac{1}{V(\mu_i)}.$$

Using these in expression (8.53) results in,

$$\begin{aligned}-E\left\{\frac{\partial^2 L_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k}\right\} &= \phi V(\mu_i)\frac{1}{\{V(\mu_i)\}^2}\left(\frac{d\,\mu_i}{d\,\eta_i}\right)^2 x_{i,j}x_{i,k} \\ &= \phi W_i\, x_{i,j}x_{i,k}.\end{aligned} \tag{8.54}$$

Summing (8.54) across observations (i) gives the total expected information. That is, let $I_n(\boldsymbol{\beta})$ be a $p \times p$ matrix with jk^{th} element

$$I_{j,k}(\boldsymbol{\beta}) = \phi\sum_{i=1}^{n} W_i\, x_{i,j}\, x_{i,k}. \tag{8.55}$$

Then, at iteration m of a Fisher Scoring algorithm, and using the notation

$$\boldsymbol{\beta}^{(m)} = (\beta_1^{(m)}, \ldots, \beta_p^{(m)})^T$$

and,

$$\nabla L_n(\boldsymbol{\beta}^{(m)}) = \left(\frac{\partial L_n(\boldsymbol{\beta}, \phi)}{\partial \beta_1}, \ldots, \frac{\partial L_n(\boldsymbol{\beta}, \phi)}{\partial \beta_p} \right)^T \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}},$$

we can write the parameter update as,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + I_n^{-1}(\boldsymbol{\beta}^{(m)})\, \nabla L_n(\boldsymbol{\beta}^{(m)}). \tag{8.56}$$

Now, expression (8.56) is entirely sufficient to program a Fisher Scoring algorithm for generalized linear models. From the standpoint of computation, however, additional simplifications are possible. In particular, pre-multiply expression (8.56) by $I_n(\boldsymbol{\beta}^{(m)})$ to obtain,

$$I_n(\boldsymbol{\beta}^{(m)})\boldsymbol{\beta}^{(m+1)} = I_n(\boldsymbol{\beta}^{(m)})\boldsymbol{\beta}^{(m)} + \nabla L_n(\boldsymbol{\beta}^{(m)}),$$

or, using $\delta\boldsymbol{\beta} \equiv \boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}$,

$$I_n(\boldsymbol{\beta}^{(m)})\, \delta\boldsymbol{\beta} = \nabla L_n(\boldsymbol{\beta}^{(m)}). \tag{8.57}$$

Note: expression (8.57) is what McCullagh and Nelder (1989) give on page 42 as $A\delta b = u$.

Now, recall from expression (8.49) that,

$$\frac{\partial L(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\phi\{y_i - \mu_i\}\, W_i\, \frac{d\eta_i}{d\mu_i}\, x_{i,j} \right].$$

Then, with $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$, and $\boldsymbol{W}$ a diagonal $n \times n$ matrix with elements W_i,

$$\nabla L_n(\boldsymbol{\beta}, \phi) = \boldsymbol{X}^T \boldsymbol{W} \phi \left((\boldsymbol{y} - \boldsymbol{\mu}) \frac{d\,\boldsymbol{\eta}}{d\,\boldsymbol{\mu}} \right),$$

or, by writing $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ where

$$z_i = (y_i - \mu_i) \frac{d\,\eta_i}{d\,\mu_i},$$

we can express the gradient as,

$$\nabla L_n(\boldsymbol{\beta}, \phi) = \phi \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}. \tag{8.58}$$

Similarly, inspection of (8.55) shows that the total expected information may be written in matrix form as,

$$I_n(\boldsymbol{\beta}) = \phi \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}. \tag{8.59}$$

Then, substitution of (8.58) and (8.59) into (8.57) gives the following equivalent statements (the first of these is just (8.57) repeated for ease of development):

$$\begin{aligned} I_n(\boldsymbol{\beta}^{(m)})\, \delta\boldsymbol{\beta} &= \nabla L_n(\boldsymbol{\beta}^{(m)}), \\ \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} \delta\boldsymbol{\beta} &= \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \\ \delta\boldsymbol{\beta} &= \left[\left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}. \end{aligned} \tag{8.60}$$

The right hand side of this last expression is in the form of a weighted least squares equation. The left hand side is the change in estimates at iteration m, $\delta\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}$. Thus, at iteration m of a Fisher Scoring algorithm for numerical computation of maximum likelihood estimates of $\boldsymbol{\beta}$ we could compute $\delta\boldsymbol{\beta}$ as in (8.60) and updated estimates as,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \delta\boldsymbol{\beta}. \tag{8.61}$$

It is possible to make one further step, as in McCullagh and Nelder (1989; p. 43) to arrive at,

$$\boldsymbol{\beta}^{(m+1)} = \left[\left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{z}}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{8.62}$$

where,

$$\tilde{\boldsymbol{z}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}.$$

The use of (8.60) and (8.61) or (8.62) are entirely equivalent, and I don't really see much computational benefit one way or the other.

Comments

1. Although this derivation sees like a long haul (and perhaps it is) what we have arrived at is a simple algorithm for maximum likelihood estimation of the regression parameters ($\boldsymbol{\beta}$) in any standard generalized linear model. This is rather remarkable.

2. The dispersion parameter ϕ cancels in the progression leading to expression (8.60). Thus, just as for normal linear regression models (which are actually a particular type of generalized linear model), parameters of the systematic model component can be estimated independently of additional parameters involved in the variances.

3. It is possible to develop maximum likelihood estimates of the dispersion parameter ϕ, although there is no longer a general algorithm, and such estimation must be developed on a case-by-case basis for each particular model. As discussed in example 8.8, a common method of estimation for ϕ is to use a moment estimator, which is consistent

4. It is important in practice to realize that, while $\boldsymbol{\beta}$ can be estimated without knowledge of ϕ, an estimate of ϕ is needed for inference. That is, both the expected information matrix with components given by expression (8.55) and the log likelihood given in expression (8.44) and (8.45) involve ϕ. Thus, inference from either Wald theory or more general likelihood-based procedures will require that an estimate of ϕ be available.

8.4 Inference from Modified Likelihood Functions

We approach in this section a rather daunting set of topics on which there have been written a vast number of papers from a huge number of perspectives. Our goal is to provide an indication of the major lines of thought that have emerged from these works; true unification does not seem possible at this point. We will divide the topic of modified likelihood functions into three parts, *profile likelihoods*, *marginal likelihoods* and *conditional likelihoods*. The second two of these, in particular, have strong connections with the concepts of sufficiency and ancillarity, and so we include a subsection on these topics prior to discussion of marginal and conditional likelihoods as distinct topics.

What we are calling here *modified* likelihoods come into play in practice primarily in situations in which dealing with the full likelihood function is difficult. For example, simultaneous maximization, in all parameter values, of the log likelihood formed from the joint distribution implied by a conditional autoregressive model is computationally difficult. In other cases, we may be able to maximize the likelihood using, for example, gradient or direct search methods, but unable to compute the information for purposes of forming inferential quantities.

8.4.1 Profile Likelihoods

Profile likelihoods are discussed by a number of authors (e.g., Lindsey 1996, Meeker and Escobar 1998) primarily as a way to assess uncertainty in a portion of the parameter (i.e., some elements of a parameter vector) while essentially ignoring others. This type of profile likelihood extends the basic idea of what

is sometimes called a "normed likelihood" from the case of a scalar parameter, in which it is easily interpreted, to a multi-parameter situation. Normed likelihoods are closely related to the approximate likelihood confidence region of expression (8.43) and, in the case of a scalar parameter, intervals formed from a normed likelihood are identical to likelihood intervals; in this case, normed likelihood is also called "relative likelihood", and the interval result is simply a re-statement of (8.43) (e.g., Meeker and Escobar 1998, Appendix B.6.6). Normed likelihoods can also be interpreted from the viewpoint of likelihoods as proportional to the probability of the data, rather than from the standpoint of asymptotic distribution theory (e.g., Lindsey 1996). In this context, a normed likelihood represents the "strength of evidence" offered by the data for any $\theta \in \Theta$, relative to that offered for the most likely θ (i.e., the maximum likelihood estimate). Normed *profile* likelihoods, also called "maximized relative likelihoods" make use of a partition of the parameter vector, maximizing a normed likelihood over the portion of the parameter that is of lesser interest. Interpretation again can be based on either asymptotic results or strength of evidence concepts, as we will discuss presently.

Other authors (e.g., Barndorff-Nielsen and Cox 1994) present profile likelihoods from the perspective of partitioning the parameter vector and focusing on a likelihood that is considered a function of only the portion of the parameter that is of "interest". These profile likelihoods often have "likelihood-like" behavior, at least up to first-order approximation. We should note that, although we are separating this form of profile likelihood from those discussed above, they are really the same thing, being simply the numerator of a normed profile likelihood or maximized relative likelihood. It sometimes helps, however, to recognize these as a different "type" of profile likelihood because they typically are used in point estimation as well as forming intervals.

Normed Likelihoods and Normed Profile Likelihoods

We presented what was called the Likelihood Ratio Main Result using the difference in log likelihoods for "full" and "reduced" models. The factor -2 in expression (8.42) was the appropriate scaling factor to give the limiting distribution of this difference. Consider now the exponentiation of this expression, which is where the name "likelihood ratio" comes from. We will consider this ratio at present, not in the context of models with different parameter dimensions, but rather for a single model with parameter $\boldsymbol{\theta} \in \Theta$. The *normed likelihood* is defined as,

$$R_n(\boldsymbol{\theta}) \equiv \frac{\ell_n(\boldsymbol{\theta})}{\ell_n(\hat{\boldsymbol{\theta}}_n)}, \tag{8.63}$$

where $\hat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ for a given set of observed data.

Interpretation of $R_n(\boldsymbol{\theta})$ is not restricted to the case of a scalar parameter, but is certainly the most useful in this situation (or, at most, a two-dimensional parameter). This is true for either the asymptotic theory interpretation or the strength of evidence interpretation. As regards asymptotic theory, if $dim(\Theta) = 1$, the normed likelihood is just

$$R_n(\theta) = \exp\{L_n(\theta) - L_n(\hat{\theta}_n)\}.$$

For interval estimation, note that, if

$$-2\{L_n(\theta) - L_n(\hat{\theta}_n)\} < \chi^2_{1,1-\alpha},$$

where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$ quantile of a Chi-squared distribution with 1 degree of freedom, then

$$R_n(\theta) > \exp\{-\chi^2_{1,1-\alpha}/2\}.$$

The strength of evidence interpretation of $R_n(\theta)$ is based on the first result presented in Section 8.3.2 that the likelihood is equal to (in the discrete case)

or proportional to (in the continuous case) the probability of the observed data given the value of θ. Interpreting the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n$ as the parameter value that "maximizes the probability of the data", the normed likelihood gives the ratio of probability of the data for any $\theta \in \Theta$ to the probability of the data for the "most likely" value of θ. The normed likelihood $R_n(\theta)$ is bounded above by 1 by virtue of definition of the maximum likelihood estimate. Thus, if a given parameter value θ^* results in $R_n(\theta^*) = 0.5$, we would say that the data are twice as likely under $\hat{\theta}_n$ as under θ^*. Note that for this interpretation to be useful, we must have a given model form with a parameter of fixed dimension; the above illustration for $\dim(\Theta) = 1$ certainly meets this stipulation.

Example 8.12

To make the notion of a normed likelihood clear, consider a situation with a set of *iid* random variables and a scalar parameter θ. Suppose $Y_1, \ldots, Y_n \sim iid\, Po(\theta)$. The likelihood function for a set of observations $y_1, \ldots, y_n$ is,

$$\ell_n(\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \frac{1}{\{\prod_{i=1}^{n} y_i!\}}\ \theta^{\sum_{i=1}^{n} y_i}\ \exp\{-n\theta\}.$$

The normed likelihood function for any value $\theta > 0$ is then,

$$R_n(\theta) = \left\{\frac{\theta}{\hat{\theta}_n}\right\}^{\sum_{i=1}^{n} y_i}\ \exp\{-n(\theta - \hat{\theta}_n)\},$$

where $\hat{\theta}_n = (1/n)\sum_{i=1}^{n} y_i$, the maximum likelihood estimate of θ.

A set of 25 observations were generated from a Poisson distribution with $\theta = 6$, giving values:

6	3	5	7	5	7	6	0	3	4	3	6	5
8	9	4	2	4	8	4	2	3	5	10	11	

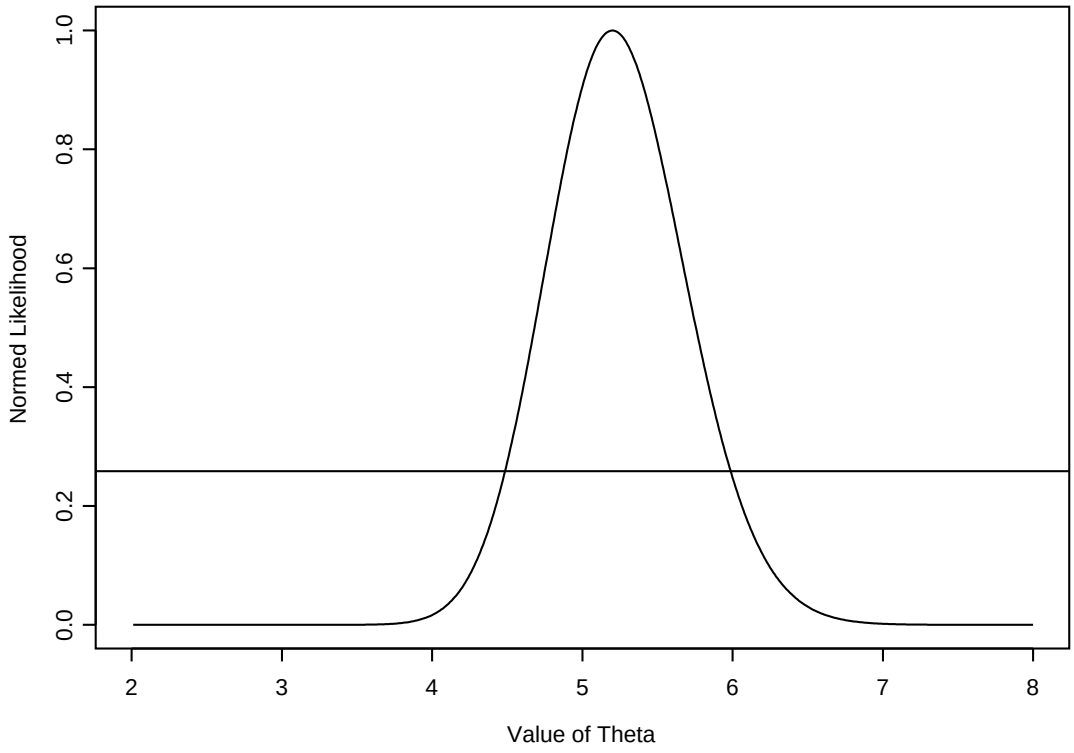

Figure 8.1: Normed likelihood for a random sample of size 25 from a $Po(6)$ distribution.

A graph of the normed likelihood function $R_n(\theta)$ is shown for these data in Figure 8.1. The horizontal line in this plot is drawn at a value of $\exp\{-\chi^2_{1,0.90}/2\} = 0.2585$, and this line intersects the normed likelihood at values of 4.49 and 5.98, giving in this case an exact likelihood interval for θ. The above likelihood interval for θ may be compared to a Wald theory 90% interval, and what might be called a "0.2 likelihood region", which could be depicted as a horizontal line at a value of 0.20 for the normed likelihood in Figure 8.1.

Interval Type	Lower Point	Upper Point
Likelihood	4.49	5.98
Wald	4.48	5.92
0.2 Likelihood	4.43	6.06

Interpretation of the Likelihood and Wald intervals is exactly what you are used to. The interval labeled 0.2 Likelihood is that the given data are less than 1/5 as likely under any parameter value outside the interval as they are for the maximum likelihood estimate. In this example, likelihood and Wald theory intervals are very similar, indicating that the likelihood surface (here curve) is well approximated by a quadratic near the maximum likelihood estimate.

Using Wald theory, confidence intervals for individual parameters (i.e., components of the parameter vector) may be constructed using the square root of the appropriate diagonal element of the inverse information matrix (expected or observed). We have given no analogous method of interval construction for portions of a parameter vector using basic likelihood intervals. Such a method is provided by normed profile likelihoods, also known as maximized relative likelihoods (e.g., Kalbfleisch and Sprott 1970). Suppose that a parameter may be partitioned into two parts as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$. Generally, the dimension of $\boldsymbol{\theta}_1$ is small (e.g., 1 or 2). We will consider the case in which $\boldsymbol{\theta}_1$ is a scalar θ and $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)^T$. Then, if the dimension of $\boldsymbol{\theta}$ is p, the dimension of $\boldsymbol{\theta}_2$ is $p - 1$. The *normed profile likelihood* for θ_1 is defined as,

$$R_n^p(\theta_1) \equiv \max_{\boldsymbol{\theta}_2} \left[\frac{\ell_n(\theta_1, \boldsymbol{\theta}_2)}{\ell_n(\hat{\boldsymbol{\theta}}_n)} \right]. \tag{8.64}$$

An approximate $(1 - \alpha)100\%$ interval for θ_1 may be formed in the same way as for a normed likelihood function with a scalar parameter, namely,

$$\{\theta_1 : \, R_n^p(\theta_1) > \exp\{-\chi^2_{1,1-\alpha}/2\}. \tag{8.65}$$

Example 8.13

Consider again the situation of Example 7.12, concerning the potential effect of selenium on reproductive success in *Gambusia*. Earlier, we looked at fitting beta-binomial mixture models to sets of fish that were held in water of varying ionic strength after being removed from their habitat. Here, we will consider the same model form for groups of fish held in the same water from which they were collected. That is, the holding water was from the same source as the fish, either the Volta or San Luis Drain areas, not water reconstituted to have the same ionic strength but without any contaminants. The actual data are given below for completeness.

The beta-binomial model with marginal log likelihood given in expression (7.37) could certainly be fit to these data. In fact, doing so results in maximized log likelihood values of $L_n(\hat{\alpha}_v, \hat{\beta}_v) = -53.4091$ for the Volta area and $L_n(\hat{\alpha}_s, \hat{\beta}_s) = -136.8254$ for the San Luis Drain area. For the combined data, $L_n(\hat{\alpha}_c, \hat{\beta}_c) = -195.9984$ which leads to the likelihood ratio test,

$$\begin{aligned} T_n &= -2\{-195.9984 - (-53.4091 + -136.8254)\} \\ &= 11.5277, \end{aligned}$$

which has an associated $p-$value of 0.00314 when compared to a Chi-squared distribution with 2 degrees of freedom. We would conclude that the full model with parameters α_v, β_v, α_s, β_s is to be preferred to the reduced model with parameters α_c, β_c. If we look at the estimated expectations of the beta mixing distributions, however, we find that for the Volta area, $\hat{\alpha}_v/(\hat{\alpha}_v + \hat{\beta}_v) = 0.89$, while for the San Luis Drain area, $\hat{\alpha}_s/(\hat{\alpha}_s + \hat{\beta}_s) = 0.82$. These values, which

are maximum likelihood estimates of the expected values, do not appear much different. We might wonder then whether the significant difference detected by the likelihood ratio test was due to something other than mean values.

Data For Study on Teratogenic Effect of Selenium in *Gambusia*

Volta Area		San Luis Drain	
No. Live	Total	No. Live	Total
28	28	36	40
31	31	33	34
9	11	27	28
68	68	4	18
32	32	13	18
37	37	22	26
19	19	20	24
17	17	20	22
26	26	38	41
52	52	21	21
30	30	20	25
46	46	26	27
0	9	7	16
47	51	18	18
22	22	23	25
18	19		
62	64		
4	5		

To investigate this question formally (i.e., by more than just visual assessment of graphs of density functions) we may proceed as follows. First, recall the log likelihood given in expression (7.37),

$$L(\alpha,\beta) \;\propto\; \sum_{i=1}^{m}\left[\sum_{j=0}^{y_i-1}\log(\alpha+j)+\sum_{j=0}^{n_i-y_i-1}\log(\beta+j)-\sum_{j=0}^{n_i-1}\log(\alpha+\beta+j)\right].$$

Now, let

$$\mu \equiv \frac{\alpha}{\alpha+\beta}\,; \qquad \theta \equiv \frac{1}{\alpha+\beta}.$$

Then $\alpha = \mu/\theta$ and $\beta = (1-\mu)/\theta$. Substituting these values into the log likelihood results in,

$$\begin{aligned} L(\mu,\theta) \;&\propto\; \sum_{i=1}^{m}\left[\sum_{j=0}^{y_i-1}\log(\mu/\theta+j)+\sum_{j=0}^{n_i-y_i-1}\log((1-\mu)/\theta+j)-\sum_{j=0}^{n_i-1}\log(1/\theta+j)\right]. \\ &=\; \sum_{i=1}^{m}\left[\sum_{j=0}^{y_i-1}\log(\mu+\theta j)+\sum_{j=0}^{n_i-y_i-1}\log((1-\mu)+\theta j)-\sum_{j=0}^{n_i-1}\log(1+\theta j)\right]. \end{aligned} \tag{8.66}$$

Maximization of (8.66) in μ and θ for Volta and San Luis Drain data values separately gives the following table, where the column labeled 90% Interval contains values from a Wald theory approach, and the subscripts v and s denote Volta and San Luis Drain, respectively.

Parameter	Estimate	Variance	$90\%Interval$
μ_s	0.822	0.00229	(0.744, 0.901)
θ_s	0.244	0.01404	(0.049, 0.439)
μ_v	0.889	0.00349	(0.792, 0.986)
θ_v	1.341	0.90867	(−0.227, 2.91)

There is, of course, no need to give values for maximized likelihoods since these remain the same under the new parameterization as they were previously.

Now, notice that the Wald interval for θ_v extends outside of the parameter space. Clearly, this parameter is not estimated as well as the others, having an (estimated) variance that is an order of magnitude greater than that for $\hat{\theta}_s$ and two orders of magnitude greater than either of the mean parameters. We might well question whether the quadratic approximation to the log likelihood on which the Wald interval is based is entirely appropriate for these data. It could be that a likelihood region might be preferable to the Wald interval in this case, and a one-dimensional (i.e., interval) form of such regions are the normed profile likelihood intervals presented in this section. For the parameter θ_v this profile likelihood is,

$$\begin{aligned} R_n^p(\theta_v) &= \max_{\mu_v} \left[\frac{\ell_n(\mu_v, \theta_v,)}{\ell_n(\hat{\mu}_v, \hat{\theta}_v)} \right] \\ &= \exp\left\{ \max_{\mu_v} L(\mu_v, \theta_v) - L(\hat{\mu}_v, \hat{\theta}_v) \right\}, \end{aligned} \tag{8.67}$$

where $L(\mu, \theta)$ is given in (8.66). Computing $R_n^p(\theta_v)$ for values of θ_v between 0 and 6, results in the normed profile likelihood shown in Figure 8.2, in which the horizontal line is drawn at $\exp(-\chi^2_{0.90,1}/2) = 0.2585227$. An approximate 90% profile likelihood interval for θ_v is then,

$$\{\theta_v : R_n^p(\theta_v) < \exp(-\chi^2_{0.90,1}/2)\},$$

which in this case is the interval (0.423, 4.370). This interval differs substantially from the Wald interval computed earlier. Repeating this procedure for the parameter μ_v results in the normed profile likelihood

$$R_n^p(\mu_v) = \max_{\theta_v} \left[\frac{\ell_n(\mu_v, \theta_v,)}{\ell_n(\hat{\mu}_v, \hat{\theta}_v)} \right] = \exp\left\{ \max_{\theta_v} L(\mu_v, \theta_v) - L(\hat{\mu}_v, \hat{\theta}_v) \right\}, \tag{8.68}$$

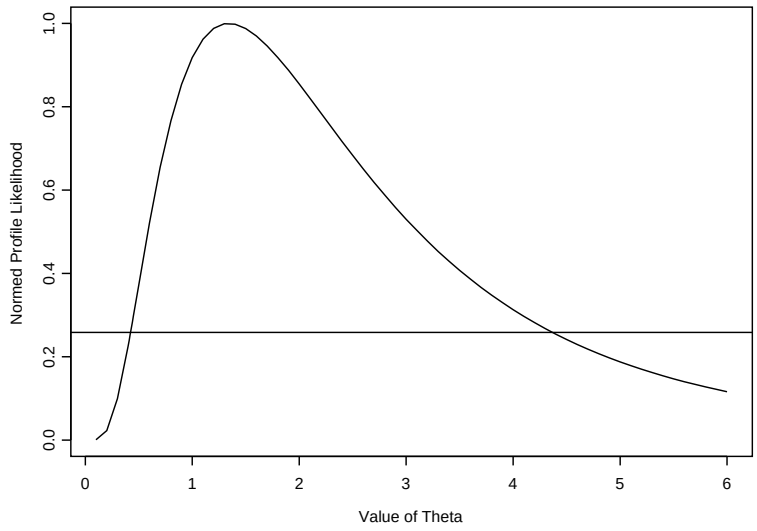

Figure 8.2: Normed profile likelihood for θ_v in analysis of *Gambusia* reproductive data.

which is shown graphically in Figure 8.3. The corresponding 90% approximate profile likelihood interval is (0.755, 0.960), which is somewhat different than the Wald interval, but not nearly so dissimilar as was the case for θ_v.

Now, although not strictly a part of the main topic of this subsection, we will use this same example to illustrate a likelihood ratio test of a full model with parameters μ_v, θ_v, μ_s, θ_s against a reduced model with parameters μ, θ_v, θ_s that is, a model in mixing distributions for both groups have a common mean, but are allowed to otherwise differ. The log likelihood for the full model may be written in the same way as previously,

$$L_n(\mu_v, \theta_v, \mu_s, \theta_s) = L_n(\mu_v, \theta_v) + L_n(\mu_s, \theta_s),$$

while the log likelihood for the reduced model may be written as,

$$L_n(\mu, \theta_v, \theta_s) = L_n(\mu, \theta_v) + L(\mu, \theta_s).$$

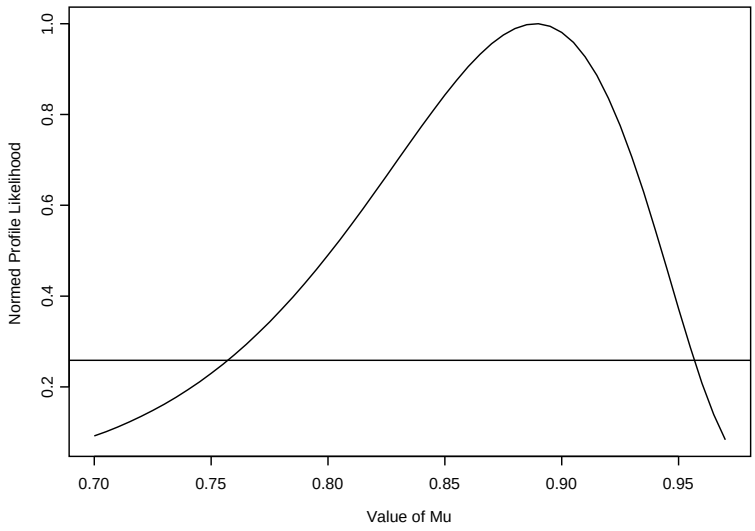

Figure 8.3: Normed profile likelihood for μ_v in analysis of *Gambusia* reproductive data.

maximizing the reduced model log likelihood results in $L_n(\hat{\mu},\, \hat{\theta}_v,\, \hat{\theta}_s) = -190.5655$, while the full model maximized log likelihood remains as before $L_n(\hat{\mu}_v,\, \hat{\theta}_v,\, \hat{\mu}_s,\, \hat{\theta}_s) = -53.4091 - 136.8254 = -190.2345$, leading to the likelihood ratio test,

$$T_n = -2\{-190.5655 + 190.2345\} = 0.6619.$$

Comparing to a Chi-squared distribution with 1 degree of freedom results in a $p-$value of 0.4159. Thus, we would not reject the reduced model in this case in favor of the full model.

Unscaled Profile Likelihoods

We now consider another function that is often called a *profile likelihood*. Although these profile likelihoods are not truly distinct from normed profile likelihoods, their use is typically that of an objective function for point estimation

(i.e., an alternative to the full likelihood) as well as a vehicle by which to obtain intervals which is the primary use for normed profile likelihoods. That is, normed profile likelihoods require that the full maximum likelihood estimate of a parameter $\boldsymbol{\theta}$ is available for use in the denominator of expression (8.64). Unscaled profile likelihoods, on the other hand, are often useful in finding maximum likelihood estimates of certain elements of the parameter vector in the first place and, like regular likelihoods, are often expressed in logarithmic form. There are two situations in which unscaled profile likelihoods appear to be the most useful, both of which operate from a partition of the parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$.

Situation 1.
The maximum likelihood estimates of one portion of the partition, say $\hat{\boldsymbol{\theta}}_1$ can be expressed as a function of any fixed value of the other as $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$.

Example 8.14

Consider a situation in which a small-scale disease epidemic has been observed, with individuals exposed to the disease agent (e.g., virus) at a common place and time. We assume that a time interval is known for exposure, but not the exact time. For example, passengers of a cruise ship come down with salmonella, possibly due to exposure at some port visited, or due to contamination in food supplies used on the ship (the difference is economically important to a cruise line). The available data consist of time to onset of disease for individuals, with time 0 defined as the start of the known interval in which exposure occurred. This might be, for example, the last resupply time of food stores on the vessel, or the last port of call the vessel made. Connect these observations

with random variables $Y_1, \ldots, Y_n$, which will be assumed to be *iid* following some distribution. Assume that the time from the point of common exposure to the onset of disease (symptoms) follows a log-normal distribution across individuals (this is not an unreasonable assumption given what is known about the incubation time for many diseases). One model that has been used in such situations is to take the random variables Y_i; $i = 1, \ldots, n$ to be *iid* following a "three-parameter log-normal" distribution which has density function, $f(y_i|\alpha, \mu, \sigma) =$

$$\begin{cases} \frac{1}{(y_i-\alpha)\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(\log(y_i-\alpha)-\mu\right)^2\right]; & \alpha < y_i \\ \\ 0\,; & \text{o.w.} \end{cases}$$

Here, the parameter α represents the time that exposure took place (measured from 0 as defined by the start of the known interval of exposure). The log likelihood formed from n observations under this model is,

$$\begin{aligned} L_n(\alpha, \mu, \sigma^2) &= -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\log(y_i-\alpha) \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\log(y_i-\alpha)-\mu\right\}^2, \end{aligned}$$

if $y_i > \alpha$ for $i = 1, \ldots, n$. Now, suppose α is fixed. Then,

$$\begin{aligned} \frac{\partial L_n}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left\{\log(y_i-\alpha)-\mu\right\}, \\ \frac{\partial L_n}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left\{\log(y_i-\alpha)-\mu\right\}^2, \end{aligned}$$

and we can write

$$\begin{aligned} \hat{\mu}(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\log(y_i-\alpha) \\ \hat{\sigma}^2(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\log(y_i-\alpha)-\hat{\mu}(\alpha)\right\}^2. \end{aligned}$$

In this situation, we can write the log likelihood $L_n(\alpha, \mu, \sigma^2)$ as a function of α alone, and this is one form of a profile log likelihood. Here, we would have,

$$L_n^p(\alpha) \propto -\frac{n}{2}\left[\log\{\hat{\sigma}^2(\alpha)\} + 2\hat{\mu}(\alpha)\right].$$

Advanced Note: this is a situation in which the log likelihood becomes unbounded as the parameter α approaches a finite bound (minimum y_i) as discussed in lab, and the density form of the likelihood may become unbounded as the parameter approaches that bound.

Situation 2.
The likelihood or log likelihood can be maximized over one portion of the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, say $\boldsymbol{\theta}_1$ for any fixed value of the other, say $\boldsymbol{\theta}_2$, although that maximizing value cannot be expressed as an explicit function. This is probably the most common situation for application of unscaled profile likelihoods.

Example 8.15

Consider a general power of the mean model,

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma\{\mu_i(\boldsymbol{\beta})\}^{\theta}\epsilon_i,$$

in which we assume $\epsilon_i \sim iid\, N(0,1)$ for $i = 1, \ldots, n$, but for which we wish to follow the prescription of Section 7.2.4 in assuming that θ is not known prior to estimation. In this case, for a fixed value of θ, the log likelihood could be maximized in $\boldsymbol{\beta}$ using any Newton-type algorithm such as Newton-Raphson or Gauss-Newton.

In either of the two situations discussed above, we may define a profile likelihood as,

$$\ell_n^p(\boldsymbol{\theta}_2) \equiv \max_{\boldsymbol{\theta}_1} \ell_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \tag{8.69}$$

for any fixed value of $\boldsymbol{\theta}_2 \in \Theta$. The logarithm of this profile likelihood is,

$$L_n^p(\boldsymbol{\theta}_2) = \max_{\boldsymbol{\theta}_1} \log\{\ell_n^p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}. \tag{8.70}$$

Notice that the profile likelihood is simply the numerator of the normed profile likelihood of expression (8.64). In the first situation, in which $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ we could write (8.69) and (8.70) as,

$$\ell_n^p(\boldsymbol{\theta}_2) = \ell_n\{\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2\},$$

$$L_n^p(\boldsymbol{\theta}_2) = \log[\ell_n^p\{\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2\}],$$

since $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ gives the maximized value over $\boldsymbol{\theta}_1$. This is, in fact, the form of the log profile likelihood given for the three parameter log-normal model.

The value of the profile likelihood (8.69) and log profile likelihood (8.70) functions is that they behave in many ways like true likelihood functions. In particular (see, e.g., Barndorff-Nielsen and Cox 1994, p.90):

1. The estimate of $\boldsymbol{\theta}_2$ found by maximizing the profile likelihood (8.69) is the maximum likelihood estimate of $\boldsymbol{\theta}_2$. This should be clear because

$$\begin{aligned} \max_{\boldsymbol{\theta}_2} \ell_n^p(\boldsymbol{\theta}_2) &= \max_{\boldsymbol{\theta}_2} \max_{\boldsymbol{\theta}_1} \ell_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= \max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \ell_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \end{aligned}$$

2. A likelihood ratio test statistic formed from the profile log likelihood (8.70) has a limiting Chi-squared distribution. That is, with $\dim(\boldsymbol{\theta}_2) =$

$p - r$ and $\dim(\boldsymbol{\theta}_1) = r$,

$$T_n(\boldsymbol{\theta}_2) = -2[L_n^p(\boldsymbol{\theta}_2^0) - L_n^p(\hat{\boldsymbol{\theta}}_2)] \xrightarrow{\mathcal{L}} \chi^2_{p-r},$$

for any fixed value $\boldsymbol{\theta}_2^0 \in \Theta_2$. Compare this with expression (8.42).

3. A profile likelihood confidence region,

$$\{\boldsymbol{\theta}_2^0 : -2[L_n^p(\boldsymbol{\theta}_2^0) - L_n^p(\hat{\boldsymbol{\theta}}_2)] \leq \chi^2_{p-r,1-\alpha}\},$$

is a valid approximate confidence region for $\boldsymbol{\theta}_2$. This follows from item 2, just as the likelihood region (8.43) follows from (8.42).

Despite these properties it should be noted that unscaled profile likelihoods or log likelihoods are *not*, in general, full likelihood functions. Although $\ell_n^p(\cdot)$ and $L_n^p(\cdot)$ behave asymptotically in the same way as $\ell_n(\cdot)$ and $L_n(\cdot)$, which is what leads to the above properties, their derivatives do not necessarily behave in the same way as those of the full likelihood functions. In particular, the expected value of the first derivative of $L_n^p(\boldsymbol{\theta}_2)$ is not necessarily equal to zero. If the portion of the parameter vector $\boldsymbol{\theta}_1$ being excluded from consideration in a profile likelihood is a "substantial fraction of n" (McCullagh and Nelder 1989, p. 255) then the difference of this expectation from zero is not negligible in asymptotics. The end result is that Wald theory can be difficult to adapt for unscaled profile likelihoods.

On the other hand, it is apparently true that the (negative) inverse second derivative matrix of the log profile likelihood is equal to the corresponding portion of the observed information matrix from the full likelihood (Barndorff-Nielsen and Cox 1994, p. 90). The apparent contradiction of this with the above assertion about Wald theory can be resolved at an intuitive level by examination of the log profile likelihood given in expression (8.70). In essence,

the log profile likelihood is simply considering $\boldsymbol{\theta}_1$ to be fixed, although the value at which it is fixed can depend of the value of $\boldsymbol{\theta}_2$ at which the log profile likelihood is being evaluated (that is, $\boldsymbol{\theta}_1$ is fixed at its maximum for the given value of $\boldsymbol{\theta}_2$). As a result, any effects of uncertainty about $\boldsymbol{\theta}_1$ are being ignored in quantification of uncertainty about an estimate of $\boldsymbol{\theta}_2$. To the degree to which uncertainty about $\boldsymbol{\theta}_1$ influences uncertainty about $\boldsymbol{\theta}_2$ (and our ability to estimate that uncertainty), inference based on profile likelihoods will be unreliable. If that degree of influence is small or negligible, then profile likelihood inference will be a good approximation to full likelihood inference.

Example 8.16

A model for which both situations that motivate unscaled profile likelihood are in force is the conditional autoregressive model (or conditionally specified Gaussian model) used in the analysis of nitrates in the Des Moines River presented in Section 7.5.6. Recall that this model resulted in a joint Gaussian distribution of the form,

$$\boldsymbol{Y} \sim Gau(\boldsymbol{\theta},\, (I_n - C)^{-1}M).$$

Consider what was called the "distance model" in that example. Then, for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$,

$$\theta_i = \beta_0 + \beta_1 \sin\left(\frac{t\pi}{91}\right) + \beta_2 \cos\left(\frac{t\pi}{91}\right),$$

which can be written as a linear model, $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. The matrix C was formed with elements $c_{i,j}$ given by

$$c_{i,j} \equiv \eta \left\{ \frac{\min\{d_{i,j}\}}{d_{i,j}} \right\}^k,$$

where $d_{i,j}$ is the distance (river miles) between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. This model then gives a joint Gaussian density function which may be written as,

$$\boldsymbol{Y} \sim Gau(\boldsymbol{X}\boldsymbol{\beta},\, (I-C)^{-1}\tau^2 I),$$

with I the $n \times n$ identity matrix. The parameters to be estimated from this model are $\boldsymbol{\beta}$, τ^2, the dependence parameter η, and k. From the joint density, the log likelihood function becomes,

$$\begin{aligned} L_n(\boldsymbol{\beta}, \tau^2, \eta, k) &= -(n/2)\log(2\pi\tau^2) + (1/2)\log(|(I-C)|) \\ & \\ & -(1/2\tau^2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(I-C)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{8.71}$$

Maximization of (8.71) in $\boldsymbol{\beta}$ and τ^2 gives,

$$\hat{\boldsymbol{\beta}} = \{\boldsymbol{X}^T(I-C)\boldsymbol{X}\}^{-1}\boldsymbol{X}^T(I-C)\boldsymbol{y}$$

$$\hat{\tau}^2 = (1/n)(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(I-C)(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

which are both functions of the parameters η and k involved in the matrix C. Thus, we have a situation that falls into the type given previously as Situation 1 in which the maximum likelihood estimate of a portion of the parameter vector can be expressed as a function of the other portion.

If we then substitute the solutions for $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}^2$ back into the log likelihood (8.71) we have the profile log likelihood function,

$$\begin{aligned} L_n^p(\eta, k) &= -(n/2)\{\log(2\pi)+1\} + (1/2)\log(|I-C|) \\ & \quad -(n/2)\log\left[\boldsymbol{y}^T(I-C)\{I-\boldsymbol{X}(\boldsymbol{X}^T(I-C)\boldsymbol{X})^{-1} \right. \\ & \quad \left. \boldsymbol{X}^T(I-C)\}\boldsymbol{y}\right]. \end{aligned} \tag{8.72}$$

Now, the profile log likelihood (8.72) can be maximized in η by any number of algorithms, but maximization in k proves difficult. This then fits the second situation in which profile likelihoods are useful. For any fixed value of k, maximization of (8.72) in η leads to estimates of η, $\boldsymbol{\beta}$, and τ^2, and we can then write another profile log likelihood in the form of expression (8.70) as,

$$L_n^p(k) = \max_{\eta}\{L_n^p(\eta, k)\} = \max_{\eta,\beta,\tau^2}\{L(\boldsymbol{\beta}, \tau^2, \eta, k\}. \tag{8.73}$$

The maximum likelihood estimate of k then results from the first property of unscaled profile likelihoods given immediately following expression (8.70) as,

$$\hat{k} = \max_{k} L_n^p(k),$$

with $L_n^p(k)$ as given in (8.73).

This procedure was in fact what was used to produce the parameter estimates given in the tables of example of Section 7.5.6. This example also involves several issues relative to the estimation of uncertainty in parameter estimates.

1. Although is is possible, as shown above, to obtain a maximum likelihood estimate of k, should this value be considered as a part of the overall model parameter to be estimated, or a part of model selection? The value of k "scales" the distance measures used to construct the model matrix C. This scaling is important as it determines the the appropriate relation between ratios of distances and ratios of spatial dependence; if the distance from a location is cut in 1/2 will the spatial dependence also be cut by 1/2 ($k = 1$), or should a change in distance of 1/4 give a change in dependence of 1/2 ($k = 2$). But, should uncertainty in k be allowed to affect uncertainty in other parameter estimates any more than a change from $c_{i,j} \propto \{1/d_{i,j}\}$ to $c_{i,j} \propto \{1/\log(d_{i,j})\}$?

2. Computation of any type of information matrix (observed or expected) would be difficult in this model. This may be somewhat less detrimental than it at first seems, since the lack of independence in the model makes justification of asymptotic likelihood inference less than a trivial matter, if it would even be possible or desirable. That is, even if computation of the information were possible, or even if normed profile likelihoods were computed, the application of normal and/or Chi-squared limit results would be questionable (at least without additional theoretical justification). A parametric bootstrap procedure, to be discussed in a future section, may provide a reasonable alternative.

8.4.2 Sufficiency and Ancillarity

Recall a typical formulation of sufficiency and ancillarity. We suppose a set of random variables $\boldsymbol{Y}$ with a possible value $\boldsymbol{y}$, but will now write the likelihood for a parameter $\boldsymbol{\theta}$ as $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ to make explicit the dependence on the observations $\boldsymbol{y}$, but momentarily dropping the subscript n that has been used previously to emphasize the role of sample size in asymptotic arguments. A statistic $\boldsymbol{T} \equiv T(\boldsymbol{Y})$ is *sufficient* for $\boldsymbol{\theta}_1$ if,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_1(\boldsymbol{\theta}; \boldsymbol{t})\, \ell_2(\boldsymbol{y}|\boldsymbol{t}), \tag{8.74}$$

where $\boldsymbol{t} \equiv T(\boldsymbol{y})$, the "observed" value of $\boldsymbol{T}$. Similarly, a statistic $\boldsymbol{U} \equiv \boldsymbol{U}(\boldsymbol{Y})$ is *ancillary* for $\boldsymbol{\theta}$ if,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_1(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{u})\, \ell(\boldsymbol{u}). \tag{8.75}$$

If estimation is to be based on the likelihood function, then $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ contains everything the data $\boldsymbol{y}$ can tell us about the value of the parameter $\boldsymbol{\theta}$. Examination of (8.74) indicates why we say that, if a sufficient statistics is available,

we can "reduce" the information content of $\boldsymbol{y}$ about $\boldsymbol{\theta}$ to consideration of $\boldsymbol{t}$ alone. In this case, maximization of $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ is the same as maximization of $\ell_1(\boldsymbol{\theta}; \boldsymbol{t})$. We also sometimes say that sufficiency implies marginalization (to the marginal distribution of $\boldsymbol{t}$). Examination of (8.75) likewise indicates why ancillarity implies conditioning, since maximization of $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ is equivalent to maximization of $\ell_1(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{u})$.

Throughout the remainder of this subsection we will again consider a partitioned parameter vector $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. We will use the terms "parameter of interest" and "nuisance parameter". These terms are in the context of a particular portion of the estimation procedure, and we may consider $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ to both be nuisance parameters in turn as we consider estimation of the other. For example, in an additive error model with variance that depends on an unknown parameter θ and an expectation function that depends on a set of parameters $\boldsymbol{\beta}$, we might consider θ a nuisance parameter for estimation of $\boldsymbol{\beta}$, but then turn around and consider $\boldsymbol{\beta}$ a nuisance parameter for estimation of θ. There are many situations in which simultaneous maximization of a likelihood in all elements of the parameter vector is difficult, and we make use of this type of multi-stage estimation, some of which we have already seen. Such estimation appears to be the most effective when we have some type of sufficiency or ancillarity to draw on; what we mean by "some type" here is the topic of this subsection.

Likelihood Orthogonality

To help understand the effects of sufficiency and ancillarity on estimation, note first that we have two general goals regarding either or both of the portions of $\boldsymbol{\theta}$, considered as consisting of two components $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. For ease of

presentation, we will consider $\boldsymbol{\theta}_1$ to be the parameter of interest and $\boldsymbol{\theta}_2$ to be a nuisance parameter, but keep in mind that the roles can be interchanged for parameters that have particular identities in a given model. The first goal is to use the maximal amount of "information" in the observation $\boldsymbol{y}$ for estimation of $\boldsymbol{\theta}_1$; look again at (9.74) and think sufficiency. The second goal, largely ignored in our discussion of normed profile likelihoods, is to take into account the effect of the nuisance parameter $\boldsymbol{\theta}_2$ on the quantification of uncertainty in estimation of $\boldsymbol{\theta}_1$. A rare but ideal situation is one in which there are two statistics $\boldsymbol{T}_1 \equiv T_1(\boldsymbol{Y})$ and $\boldsymbol{T}_2 \equiv T_2(\boldsymbol{Y})$ such that the likelihood may be written as,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_1(\boldsymbol{\theta}_1; \boldsymbol{t}_1)\, \ell_2(\boldsymbol{\theta}_2; \boldsymbol{t}_2). \tag{8.76}$$

If, in addition to (8.76) the parameter components $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are variation independent, we can estimate and make inferences about $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ totally independently of the other. The condition (8.76), along with variation independent parameters, is called *likelihood orthogonality* or sometimes *likelihood independence* (e.g., Lindsey 1996, p. 239).

Example 8.17

We have had a number of examples of beta-binomial models in which we have always considered the binomial sample sizes $n_1, \ldots, n_m$ to be fixed quantities. But clearly, in any of these examples, we could take the binomial sample sizes to be "observed" values of random variables $N_1, \ldots, N_m$. Two models immediately suggest themselves:

1. Let $Y_1, \ldots, Y_m$, conditional on the values $N_1 = n_1, \ldots, N_m = n_m$ and $\theta_1, \ldots, \theta_m$ have conditionally independent distributions that follow bino-

mial probability mass functions, $bin(n_i, \theta_i)$. Let $N_1, \ldots, N_m$ *iid* random variables following a common Poisson probability mass function $Po(\lambda)$. Let $\theta_1, \ldots, \theta_m$ be *iid* random variables following a common beta probability density function $Be(\alpha, \beta)$.

2. Let $Y_1, \ldots, Y_m$, conditional on the values $N_1 = n_1, \ldots, N_m = n_m$ and $\theta_1, \ldots, \theta_m$ have conditionally independent distributions that follow binomial probability mass functions, $bin(n_i, \theta_i)$. Let $N_1, \ldots, N_m$ given $\lambda_1, \ldots, \lambda_m$ be conditionally independent random variables following Poisson probability mass functions $Po(\lambda_i)$. Let $\theta_1, \ldots, \theta_m$ be *iid* random variables following a common beta probability density function $Be(\alpha, \beta)$, and let $\lambda_1, \ldots, \lambda_m$ be *iid* random variables following a common gamma probability density function $Ga(\gamma, \nu)$.

For either of these models, the likelihood has the property (8.76), for the first model with $\boldsymbol{\theta}_1 = (\alpha, \beta)^T$ and $\boldsymbol{\theta}_2 = \lambda$, and for the second model with $\boldsymbol{\theta}_1 = (\alpha, \beta)^T$ and $\boldsymbol{\theta}_2 = (\gamma, \nu)^T$.

Partial Sufficiency and Ancillarity

The happy occurrence of likelihood orthogonality is rare. In many other cases, however, we can come close to the complete factorization of expression (8.76). A statistic $\boldsymbol{T} \equiv T(\boldsymbol{Y})$ is said to be *partially sufficient* for $\boldsymbol{\theta}_1$ if,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_1(\boldsymbol{\theta}_1; \boldsymbol{t})\, \ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t}), \tag{8.77}$$

and $\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t})$ is "noninformative" about $\boldsymbol{\theta}_1$. What is meant by "noninformative" will be discussed shortly. A statistic $\boldsymbol{U} \equiv U(\boldsymbol{Y})$ is said to be *partially ancillary* for $\boldsymbol{\theta}_1$ if,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_1(\boldsymbol{\theta}_1; \boldsymbol{y}|\boldsymbol{u})\, \ell_2(\boldsymbol{\theta}; \boldsymbol{u}), \tag{8.78}$$

and $\ell_2(\boldsymbol{\theta}; \boldsymbol{u})$ is "noninformative" about $\boldsymbol{\theta}_1$. In comparing (8.77) and (8.78) to the ordinary definitions of sufficiency and ancillarity, notice that the terms $\ell_1(\cdot)$ in (8.77) and (8.78) are analogous to (8.74) and (8.75) with $\boldsymbol{\theta}_1$ replacing the full parameter $\boldsymbol{\theta}$. But the terms $\ell_2(\cdot)$ in (8.77) and (8.78) are now markedly different from those in (8.74) and (8.75), being allowed to depend not only on $\boldsymbol{\theta}_2$ but even $\boldsymbol{\theta}_1$.

The definitions of partial sufficiency and ancillarity, however, contained the qualifier that these $\ell_2(\cdot)$ terms be "noninformative" about the value of $\boldsymbol{\theta}_1$. We now must make this concept more explicit. Barndorff-Nielson and Cox (1994, p. 38) give three conditions under which the $\ell_2(\cdot)$ terms in (8.77) or (8.78) can be considered noninformative about the parameter of interest $\boldsymbol{\theta}_1$ Their presentation is in the context of a partially ancillary statistic, but we generalize the first two of their conditions to either of the $\ell_2(\cdot)$ terms; we also change the order of presentation.

1. The term $\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t}) = \ell_2(\boldsymbol{\theta}_2; \boldsymbol{y}|\boldsymbol{t})$ in (8.77) or $\ell_2(\boldsymbol{\theta}; \boldsymbol{u}) = \ell_2(\boldsymbol{\theta}_2; \boldsymbol{u})$ in (8.78). Notice that either of these "collapse" the conditions. That is, if $\boldsymbol{T}$ is a statistic such that (8.77) holds as,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_1(\boldsymbol{\theta}_1; \boldsymbol{t})\, \ell_2(\boldsymbol{\theta}_2; \boldsymbol{y}|\boldsymbol{t}),$$

   then (8.78) also holds. Then $\boldsymbol{T}$ is partially sufficient for $\boldsymbol{\theta}_1$ and partially ancillary for $\boldsymbol{\theta}_2$. Similarly, if (8.78) holds as,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_1(\boldsymbol{\theta}_1; \boldsymbol{y}|\boldsymbol{u})\, \ell_2(\boldsymbol{\theta}_2; \boldsymbol{u}),$$

   then (8.77) also holds, and $\boldsymbol{U}$ is partially ancillary for $\boldsymbol{\theta}_1$ and partially sufficient for $\boldsymbol{\theta}_2$.

2. The condition given in (1) does not always hold, but does hold for a particular value $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$.

3. The observation of $\boldsymbol{U}$ alone in (8.78) would render estimation of $\boldsymbol{\theta}_1$ difficult or impossible without knowledge of $\boldsymbol{\theta}_2$. While certainly a much more vague "condition" than (1) or (2), this situation seems to occur if, for example, $\dim(\boldsymbol{U}) \leq \dim(\boldsymbol{\theta}_2)$ or $E(\boldsymbol{U}) = \boldsymbol{\theta}_2$. In a colloquial sense, $\boldsymbol{U}$ is "used up" estimating $\boldsymbol{\theta}_2$, and has little or nothing left to contribute to the estimation of $\boldsymbol{\theta}_1$.

As a word of caution, do not start to believe that the expressions (8.77) and (8.78) will always apply to the same model. That is, they are separate conditions, not a pair of conditions that go together. It is true, however, that in any number of cases they do seem to both hold for different statistics $\boldsymbol{T}$ and $\boldsymbol{U}$.

8.4.3 Marginal and Conditional Likelihoods

Finding functions such as $\ell_1(\cdot)$ and $\ell_2(\cdot)$ in (8.77) or (8.78) is not necessarily an easy matter. First, in order to define marginal or conditional likelihoods, we need these functions to be proportional to probability mass or density functions. This is necessary so that the component functions $\ell_1(\cdot)$ in either (8.77) or (8.78) correspond to actual likelihood functions. Secondly, while we may be able to show relations such as (8.77) or (8.78) for probabilities or distributions in the general sense, these may not always be easily written in terms of density functions, which is also something we would like in practice (i.e., so we can write computer programs to compute these functions). Thus, to apply (8.77) or (8.78) in a given problem, we must be able to find component functions $\ell_1(\cdot)$ and $\ell_2(\cdot)$ that are proportional to density (or mass) functions which we can compute directly. It is not enough to show that these factorizations exist for given statistics $T(\boldsymbol{Y})$ or $U(\boldsymbol{Y})$ without being able to explicitly give the forms

of component functions.

A fundamental difficulty, which occurs in the case of continuous random variables $Y_1, \ldots, Y_n$, is that the sigma fields generated by $T \equiv T(\boldsymbol{Y})$ and by $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)^T$ are, in general, different. That is, the sigma field generated by the transformation T is a subfield of that generated by $\boldsymbol{Y}$. If the probability space of $\boldsymbol{Y}$ is $(\Omega_Y, \mathcal{B})$ and T is a mapping into $(\Omega_T, \mathcal{A})$, then the sigma field induced by T is,

$$\mathcal{B}_0 = T^{-1}(\mathcal{A}) = \{T^{-1}(A) : A \in \mathcal{A}\}.$$

This is typically not a problem for defining probabilities, since integrals of real-valued measurable functions with respect to the probability measure μ defined over $(\Omega_Y, \mathcal{B})$ and the measure $\mu^* = \mu[T^{-1}(A)]$; $A \in \mathcal{A}$ that is subsequently induced over $(\Omega_T, \mathcal{A})$ can be related via $T^{-1}(\mathcal{A})$ as (e.g., Lehmann 1986, Lemma 2, p. 43),

$$\int_{T^{-1}(A)} g[T(y)]d\mu(y) = \int_A g(t)d\mu^*(t).$$

Dissimilarity in the sigma fields generated by $\boldsymbol{Y}$ and $T(\boldsymbol{Y})$ does, however, cause problems for the derivation of a conditional density for $\boldsymbol{Y}$ given T using the simple method of dividing a joint density by a conditional density; the problem lies in defining a joint density.

There appear to be two avenues by which this difficult can be approached in application. The first is to find a statistic $T(\boldsymbol{Y})$ and verify two conditions which can be given as,

1. The sigma field generated by $T(\boldsymbol{Y})$ is contained in (usually equivalent to) that generated by the (either normed or unscaled) profile likelihood for $\boldsymbol{\theta}_1$,

$$\ell_n^p(\boldsymbol{\theta}_1) = \max_{\boldsymbol{\theta}_2} \ell_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

2. The marginal distribution of $T(\boldsymbol{Y})$ depends on $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}_1$.

This is the approach adopted by Lindsey (1996), who presents these conditions as one way to define what are sometimes called *L-sufficient* statistics. If $T(\boldsymbol{Y})$ meets these conditions and its marginal density is available (which would probably be true from verification of the second condition above), and the factorization (8.77) holds by making $\ell_1(\boldsymbol{\theta}_1; \boldsymbol{t})$ proportional to the marginal density of $T(\boldsymbol{Y})$, then what is left must necessarily correspond to $\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t})$.

In this case, we can define the *marginal likelihood* as

$$\ell_M(\boldsymbol{\theta}_1) \equiv \ell_1(\boldsymbol{\theta}_1; \boldsymbol{t}), \tag{8.79}$$

where $\ell_1(\boldsymbol{\theta}_1; \boldsymbol{t})$ is the first term in expression (8.77). This assumes that the second term in (8.77), $\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t})$ can be shown to provide "essentially" no information about $\boldsymbol{\theta}_1$, according to one of the three conditions given on pages 680-681. Application of this approach appears to be more problematic to derive, directly from density representations, a *conditional likelihood* as $\ell_C(\boldsymbol{\theta}_1) \equiv \ell_2(\boldsymbol{\theta}_1; \boldsymbol{y}|\boldsymbol{u})$ from (8.78).

It is instructive for understanding marginal and conditional likelihoods to consider simple examples where we may not necessarily apply these concepts, but which will indicate the underlying concepts. We introduce the simplest of such examples here to illustrate marginal likelihood.

Example 8.18

Let $Y_1, \ldots, Y_n \sim iid\, N(\mu, \sigma^2)$. Suppose that our interest focuses on estimation of σ^2; μ is then, for this estimation, considered the nuisance parameter. In the context of our presentation of partial sufficiency, $\boldsymbol{\theta} = (\mu, \sigma^2)$, $\boldsymbol{\theta}_1 = \sigma^2$ and $\boldsymbol{\theta}_2 = \mu$. The full likelihood function may be written as,

$$
\begin{aligned}
\ell_n(\mu, \sigma^2) = & \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2 - \frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right] \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right],
\end{aligned}
$$

which we have written in the form of (8.77) if we take,

$$
\begin{aligned}
\ell_1(\boldsymbol{\theta}_1; \boldsymbol{t}) &\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \\
\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t}) &\propto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right],
\end{aligned}
$$

where in these expressions we have taken $\boldsymbol{T} \equiv T(\boldsymbol{Y}) = \sum(y_i - \bar{y})^2$.

We know (from previous courses) that $\boldsymbol{T}/\sigma^2 \sim \chi^2_{n-1}$, which has density,

$$
\begin{aligned}
g_1\left(\frac{\boldsymbol{t}}{\sigma^2}\right) &= \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} \left(\frac{\boldsymbol{t}}{\sigma^2}\right)^{\{(n-1)/2\}-1} \exp\left\{-\frac{\boldsymbol{t}}{2\sigma^2}\right\} \\
&\propto \left(\frac{\boldsymbol{t}}{\sigma^2}\right)^{\{(n-1)/2\}-1} \exp\left\{-\frac{\boldsymbol{t}}{2\sigma^2}\right\}.
\end{aligned}
$$

This then implies that the density of $\boldsymbol{T}$ is,

$$
g_2(\boldsymbol{t}) \propto \left(\frac{1}{\sigma^2}\right)^{(n-1)/2} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{t}\right\},
$$

from a straightforward transformation of random variables (the Jacobian is σ^{-2}). The proposal for $\ell_1(\boldsymbol{\theta}_1; \boldsymbol{t}) = \ell_1(\sigma^2; \boldsymbol{t})$ is thus verified as proportional to

the marginal density of $T(\boldsymbol{Y})$. Now, the profile likelihood of σ^2 in this model is,

$$\begin{aligned}\ell_n^p(\sigma^2) &= \max_{\mu} \ell_n(\mu, \sigma^2) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \bar{y})^2\right],\end{aligned}$$

the random version of which clearly generates the same sigma algebra as $T(\boldsymbol{Y}) = \sum(Y_i - \bar{Y})^2$. It remains then only to verify that the proposed expression for $\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t})$ is essentially noninformative about σ^2. The first condition is not fulfilled, since the proposed $\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t})$ involves both $\boldsymbol{\theta}_1 = \sigma^2$ and $\boldsymbol{\theta}_2 = \mu$. The second condition is certainly met for any $\sigma^2 \propto \sum(y_i - \bar{y})^2$. The third condition would be met if, by defining $U(\boldsymbol{y}) \equiv \bar{Y}$, the factorization of (8.78) holds (this is, in fact, true for this example). Thus, we have justified, to the extent possible under this "less than exact" theory, the use of the marginal likelihood $\ell_M(\sigma^2)$ for estimation of σ^2. Here, $\ell_M(\sigma^2)$ is given equivalently as,

$$\ell_1(\boldsymbol{\theta}_1; \boldsymbol{t}) = \ell_1(\sigma^2; \boldsymbol{t}) = \left(\frac{1}{\sigma^2}\right)^{(n-1)/2} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{t}\right\}.$$

The maximum marginal likelihood estimate of σ^2 is then,

$$\begin{aligned}\hat{\sigma}^2 &= \max_{\sigma^2} \ell_M(\sigma^2) = \max_{\sigma^2} \ell_1(\sigma^2; \boldsymbol{t}) \\ &= \max_{\sigma^2} \left\{\log\left[\ell_1(\sigma^2; \boldsymbol{t})\right]\right\} = \max_{\sigma^2} L_1(\sigma^2; \boldsymbol{t}) \\ &= \max_{\sigma^2} \left\{-\frac{n-1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{t}\right\} \\ \Rightarrow\ \hat{\sigma}^2 &= \frac{1}{n-1}\boldsymbol{t} \\ &= \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2,\end{aligned}$$

which we recognize as the unbiased (in fact, UMVU) estimator of σ^2 in this problem. The use of this marginal likelihood is one way to motivate what is called *Restricted* or *Residual* maximum likelihood (REML) which you saw in Stat 511 connected with the estimation of variance terms in linear mixed models. For a relatively detailed derivation of REML as a marginal likelihood estimator for use with linear mixed models see McCullagh and Nelder (1989), Exercises 7.10-7.13.

The second manner in which researchers have dealt with the difficulty of defining a density for the numerator of $\ell_2(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{t})$ in (8.77) or $\ell_1(\boldsymbol{\theta}_1; \boldsymbol{y}|\boldsymbol{t})$ in (8.78) is to make use of one-to-one transformations of $\boldsymbol{Y}$ to variables $(\boldsymbol{Z}, \boldsymbol{V})$, say (e.g., Pawitan 2001, Chapter 10.3; Kalbfleisch and Sprott 1970). The idea is that one of $\boldsymbol{Z}$ or $\boldsymbol{V}$ corresponds to one of $\boldsymbol{T}$ or $\boldsymbol{U}$ in (8.77) or (8.78). This then allows the usual methods of transformations to derive the appropriate densities. We will first re-consider Example 8.18 from a different viewpoint that makes use of this technique.

Example 8.18 (cont.)

Consider again the one-sample normal problem of Example 8.18, but now define $V \equiv \sum Y_i$ and, for $i = 1, \ldots, n-1$, $Z_i \equiv Y_i$. Then V is dimension 1, $\boldsymbol{Z}$ is dimension $n-1$, the transformation from $\boldsymbol{Y}$ to $(V, \boldsymbol{Z})$ is one-to-one, and $Y_n = V - \sum_{i=1}^{n-1} Z_i$ so that the Jacobian of the transformation is 1. Then, $g(\boldsymbol{z}, v|\mu, \sigma^2) = f(\boldsymbol{y}|\mu, \sigma^2)$, the marginal of V is $N(n\mu, n\sigma^2)$, and in (8.78) (dropping constants of proportionality) $\ell_1(\boldsymbol{\theta}_1; \boldsymbol{y}|\boldsymbol{u})$ becomes,

$$\ell_1(\sigma^2; \boldsymbol{z}|v) = \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n-1} z_i^2 + \left(v - \sum_{i=1}^{n-1} z_i\right)^2\right\} - \frac{n-1}{2}\log(\sigma^2)\right].$$

We then use $\ell_1(\sigma^2; \boldsymbol{z}|v)$ as a *conditional* likelihood for estimation of σ^2, that

is, $\ell_C(\sigma^2) \equiv \ell_1(\sigma^2; \boldsymbol{z}|v)$. Notice that this conditional likelihood, if written in terms of the original $\boldsymbol{y}$, is identical to $\ell_M(\sigma^2)$ given in the first portion of Example 8.18. We have left $\ell_C(\sigma^2)$ in terms of the transformed variables V and $\boldsymbol{Z}$ because one cannot always count on the transformation from the original $\boldsymbol{Y}$ to $(\boldsymbol{Z}, \boldsymbol{V})$ to be linear.

To summarize this approach for derivation of marginal and conditional likelihoods, let there exist a one-to-one transformation of $\boldsymbol{Y}$ to $(\boldsymbol{Z}, \boldsymbol{V})$ such that,

$$\begin{aligned} \ell(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{v}) &= \ell_1(\boldsymbol{\theta}_1; \boldsymbol{v})\ell_2(\boldsymbol{\theta}; \boldsymbol{z}|\boldsymbol{v}) \\ \text{or} & \\ ell(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{v}) &= \ell_1(\boldsymbol{\theta}_1; \boldsymbol{v}|\boldsymbol{z})\ell_2(\boldsymbol{\theta}; \boldsymbol{z}). \end{aligned}$$

Then we take a marginal likelihood to be $\ell_M(\boldsymbol{\theta}_1) = \ell_1(\boldsymbol{\theta}_1; \boldsymbol{v})$ in the first instance or a conditional likelihood to be $\ell_C(\boldsymbol{\theta}_1) = \ell_1(\boldsymbol{\theta}_1; \boldsymbol{v}|\boldsymbol{z})$ in the second instance. The provision about the other terms in these likelihood factorizations being essentially noninformative about $\boldsymbol{\theta}_1$ continue to be needed for such marginal or conditional likelihoods to be useful. If $\boldsymbol{Z}$ and $\boldsymbol{V}$ in this progression are independent, then marginal and conditional likelihoods coincide, as in example 8.18, for which $\sum Y_i$ and $\sum(Y_i - \bar{Y})^2$ are known to be independent.

We will give one additional example to illustrate the potential uses of marginal and conditional likelihoods. This example, introduced originally by Neyman and Scott (1948), illustrates the use of marginal and conditional likelihoods (they will again be the same in this example) in a situation for which the number of nuisance parameters increases as a proportion of the sample size. This is a classic setting for marginal and/or conditional likelihood, and discussion of these likelihoods is often contained in sections of texts that discuss dealing with large numbers of nuisance parameters (e.g., Pawitan

2001). This example is used in Lindsey (1996), Pawitan (2001), and the entry by Kalbfleisch in the Encyclopedia of Statistical Science under the entry on "Pseudo-Likelihood".

Example 8.19

Consider a study designed to investigate the precision of an measurement instrument across a range of possible values of the quantity being measured. Suppose that observations are gathered in pairs corresponding to random variables $\{(Y_{i,1}, Y_{i,2}) : i = 1, \ldots, n\}$ for which the model is

$$Y_{i,1}, Y_{i,2} \sim iid\, N(\mu_i,\, \sigma^2),$$

where we also assume independence across i. The intention of the study is measurement precision so the parameter of interest is σ^2. The values of the means, $\{\mu_i : i = 1, \ldots, n\}$ are considered nuisance parameters, and the number of these nuisance parameters present would increase as n increases; in fact, for any fixed sample size, the model contains $n + 1$ parameters and $2n$ observations. Let $\boldsymbol{\theta} \equiv (\mu_1, \ldots, \mu_n,\, \sigma^2)^T$, $\theta_1 \equiv \sigma^2$ the parameter of interest, and $\boldsymbol{\theta}_2 \equiv (\mu_1, \ldots, \mu_n)^T$ the nuisance parameters. Then the full likelihood can be written as,

$$\begin{aligned} \ell(\boldsymbol{\theta}; \boldsymbol{y}) &= \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{2}(y_{i,j} - \mu_i)^2\right] \\ &= \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{2} y_{i,j}^2\right) - 2\mu_i \sum_{j=1}^{2} y_{i,j} + \mu_i^2\right]. \end{aligned}$$

Now, let

$$z_i = \frac{y_{i,1} - y_{i,2}}{\sqrt{2}}$$

$$v_i = \frac{y_{i,1} + y_{i,2}}{\sqrt{2}}.$$

Then the $2n$ values in $\{(Y_{i,1}, Y_{i,2}) : i = 1, \ldots, n\}$ can be transformed into the $2n$ values $\{Z_i, V_i : i = 1, \ldots, n\}$ with Jacobian 1, and, in particular,

$$\begin{aligned} z_i^2 + v_i^2 &= \sum_{j=1}^{2} y_{i,j}^2, \\ \sqrt{2}\, v_i &= \sum_{j=1}^{2} y_{i,j}. \end{aligned}$$

Then the likelihood becomes,

$$\begin{aligned} \ell(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{v}) &= \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i^2 + v_i^2 - 2\mu_i\sqrt{2}\, v_i + 2\mu_i^2\right] \\ &= \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{v_i^2 - 2\sqrt{2}\mu_i + 2\mu_i^2\right\}\right], \end{aligned}$$

and the marginal likelihood for $\theta_1 \equiv \sigma^2$ based on $\boldsymbol{z}$ would be

$$\ell_M(\sigma^2) = \ell_1(\theta_1; \boldsymbol{v}) = \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i^2\right],$$

because this is, in fact, the marginal distribution of $\{z_i : i = 1, \ldots, n\}$. It is a simple matter to show that the remaining term in the likelihood (involving only $\boldsymbol{v}$) is the conditional density of $\boldsymbol{v}$ given $\boldsymbol{z}$ since $z_i \sim iid\, N(0, \sigma^2)$ and $v_i \sim iid\, N(\sqrt{2}\mu_i, \sigma^2)$. The maximum marginal likelihood estimate of σ^2 is then obtained by, for example, maximizing the log of $\ell_M(\sigma^2)$ which gives,

$$\begin{aligned} \hat{\sigma}_M^2 &= \frac{1}{n}\sum_{i=1}^{n} z_i^2 \\ &= \frac{1}{2n}\sum_{i=1}^{n}(y_{i,1} - y_{i,2})^2. \end{aligned}$$

This example is generally used to contrast the estimate $\hat{\sigma}_M^2$ with that obtained by maximizing the profile likelihood $\ell_n^p(\sigma^2)$, which gives $\hat{\sigma}^2 = (1/4n)\sum(y_{i,1} - y_{i,2})^2$, only one-half the value of $\hat{\sigma}_M^2$.

8.4.4 Concluding Remarks

To conclude our discussion of estimation and inference from modified likelihood functions, the following are pertinent:

1. There is a good deal of overlap among the topics of profile likelihoods, marginal likelihoods, and conditional likelihoods. In fact, it is not infrequent that marginal or conditional likelihoods turn out to also be profile likelihoods. Pawitan (2001, Chapter 10) gives several examples for which conditional or marginal likelihoods are profile likelihoods, and several examples for which this is not true.

2. The problem of Example 8.19 could, of course, be approached through the use of a mixture (or random parameter) model, in which the $\mu_i;\ i = 1, \ldots, n$ are taken as random variables following some distribution, which would then be integrated out of the marginal distribution of the $(Y_{i,1},\ Y_{i,2})$ pairs. The use of marginal (or conditional in this case) likelihood may be viewed as an attempt to eliminate nuisance parameters without using random parameter models. This, in fact, is the essential focus of the discussion paper by Kalbfleisch and Sprott (1970) in which much of the following discussion centered on this issue. At that time, random parameter models were thought to imply a Bayesian approach. As we have seen, this is in fact not the case, although some statisticians still hold to this notion.

3. As is probably clear from the examples presented, marginal and conditional likelihoods do not necessarily offer a *constructive* methodology by which to develop estimators. Typically, once a "reasonable" approach for estimation has become the focus of investigation, we may try and

determine whether that approach corresponds to some type of marginal or conditional likelihood. But, it is usually difficult to divine the correct likelihood factorization to produce useful marginal or conditional likelihoods before we already have a pretty good idea of where we are headed in the development of an estimation strategy.

8.5 Quasi-Likelihood, Estimating Functions, and Pseudo-Likelihoods

In this section we consider estimation, and the associated inference procedures, based on a number of functions that might be considered as "likelihoods that aren't really likelihoods", or "functions that are sort-of likelihoods", or "functions that we might pretend are likelihoods". While the marginal and conditional likelihoods of the previous subsection are sometimes lumped with the types of functions we consider here, I have separated them because, while marginal and conditional likelihoods may not correspond to *full* likelihoods for the entire parameter vector $\boldsymbol{\theta}$, they are proportional to probability mass or density functions (perhaps marginal or conditional) and, thus, the probability of at least a portion of the obtained data. That is, they are "true" likelihood functions. The functions considered here do not share that property, are thus not "true" likelihoods.

We begin with a consideration of what are called *quasi-likelihoods* in the context of response distributions that are (but aren't really, are sort-of, or we might pretend are) in exponential dispersion family form. We then briefly indicate that the derivatives of such quasi-likelihoods are a special case of a more general structure called *estimating functions*. The term *pseudo-likelihood* can

technically apply to almost any function that is treated as if it were a likelihood function; we give two versions of pseudo-likelihoods that have proven useful for certain classes of models, nonlinear regressions with unknown parameters in the variance model, and conditionally specified models.

8.5.1 Quasi-Likelihood

To motivate quasi-likelihoods, consider again maximum likelihood estimation of generalized linear models as detailed in Section 8.3.6. In deriving derivatives of the log likelihood for these models we made use first of independence among response variables to write the likelihood and its derivatives in terms of sums of contributions from individual variables Y_i, and secondly of the chain rule to arrive at expression (8.49). Consider using these same techniques, but taking the derivation only to the point of obtaining a derivative of the log likelihood with respect to the expectation μ_i,

$$\frac{\partial L_i(\mu_i,\phi)}{\partial \mu_i} = \frac{\partial L_i(\mu_i,\phi)}{\partial \theta_i}\frac{d\theta_i}{d\mu_i},$$

$$\frac{\partial L(\mu_i,\phi)}{\partial \mu_i} = \sum_{i=1}^{n}\frac{\partial L_i(\mu_i,\phi)}{\partial \mu_i}.$$

Focusing on an individual response variable, and given an exponential dispersion family form for the density or mass functions of the Y_i; $i = 1,\ldots,n$, we have,

$$\frac{\partial L_i(\mu_i,\phi)}{\partial \mu_i} = \frac{\phi\{y_i - b'(\theta_i)\}}{V(\mu_i)} = \frac{\phi\{y_i-\mu_i\}}{V(\mu_i)}. \tag{8.80}$$

Notice that expression (8.80) (and thus also the sum across i) depends only on the first two moments of Y_i, $E(Y_i) = \mu_i$ and $var(Y_i) = (1/\phi)V(\mu_i)$. We might then consider trying to "recover" the log likelihood for μ_i (for fixed ϕ)

through integration as,

$$\begin{aligned} Q_i(\mu_i|\phi) &= \int_{y_i}^{\mu_i} \frac{\phi\{y_i - t\}}{V(t)}\, dt, \\ Q(\boldsymbol{\mu}|\phi) &= \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{\phi\{y_i - t\}}{V(t)}\, dt. \end{aligned} \tag{8.81}$$

Example 8.20

Suppose $Y_1, \ldots, Y_n$ have Poisson distributions with expected values $\mu_1, \ldots, \mu_n$. Then $\phi \equiv 1$, $V(\mu_i) = \mu_i$, and, up to an additive constant depending only on y_i,

$$Q_i(\mu) = \int_{y_i}^{\mu_i} \frac{y_i - t}{t}\, dt = y_i \log(\mu_i) - \mu_i,$$

so that,

$$Q(\boldsymbol{\mu}) = \sum_{i=1}^{n} \{y_i \log(\mu_i) - \mu_i\},$$

and, up to an additive constant, $Q(\boldsymbol{\mu}) = L(\boldsymbol{\mu})$.

Note that, in general, the additive constant that distinguishes $Q(\boldsymbol{\mu})$ from $L(\boldsymbol{\mu})$ in this situation (which is $Y_1, \ldots, Y_n$ independent with exponential dispersion families of the glm type) depends on both $\boldsymbol{y}$ and ϕ; this will be important if the value of ϕ is to be estimated.

Basic Quasi-Likelihood

The fundamental idea underlying basic quasi-likelihood is that, even in situations that do not correspond to independent random variables with fully specified exponential dispersion family distributions, the function $Q(\boldsymbol{\mu}|\phi)$ in (8.81) should behave in a manner that resembles a log likelihood function for $\boldsymbol{\mu}$.

Consider independent random variables $Y_1, \ldots, Y_n$ that have expectations $\mu_1, \ldots, \mu_n$ and variances $\phi V_1(\mu_1), \ldots, \phi V_n(\mu_n)$ for set of specified functions $V_1(\cdot), \ldots, V_n(\cdot)$, and assume that $\mu_i = h(\boldsymbol{x}_i, \boldsymbol{\beta})$ for some known function $h(\cdot)$ and unknown parameters $\boldsymbol{\beta}$ with $\dim(\boldsymbol{\beta}) = p < n$, but specify nothing additional about the model. Notice that we have allowed the functions $V_i(\cdot)$ to vary across observations. In the majority of situations it will be reasonable to take these to be the same function, but that is not necessary. What is necessary, however, is that $var(Y_i) = \phi V_i(\mu_i)$ where ϕ is constant and $V_i(\cdot)$ does not depend on elements of $\boldsymbol{\mu}$ other than μ_i (McCullagh and Nelder 1989, p. 324).

In this independence situation, define the quasi-likelihood to be as in (8.81),

$$\begin{aligned} Q_i(\mu_i|\phi) &= \int_{y_i}^{\mu_i} \frac{\phi\{y_i - t\}}{V_i(t)}\, dt, \\ Q(\boldsymbol{\mu}|\phi) &= \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{\phi\{y_i - t\}}{V_i(t)}\, dt. \end{aligned}$$

The quasi-score function is then,

$$\begin{aligned} U_i(\mu_i|\phi) &= \frac{\phi\{y_i - \mu_i\}}{V_i(\mu_i)} \\ U(\boldsymbol{\mu}|\phi) &= \sum_{i=1}^{n} \frac{\phi\{y_i - \mu_i\}}{V_i(\mu_i)}. \end{aligned} \tag{8.82}$$

It is easy to show that the elements of $U(\boldsymbol{\mu}|\phi)$, which are first derivatives of $Q(\boldsymbol{\mu}|\phi)$, have the following properties:

$$\begin{aligned} E\{U_i(\mu_i|\phi)\} &= 0, \\ var\{U_i(\boldsymbol{\mu}_i|\phi)\} &= \frac{\phi}{V_i(\mu_i)} \\ -E\left\{\frac{\partial}{\partial \mu_i} U_i(\mu_i|\phi)\right\} &= \frac{\phi}{V_i(\mu_i)} \end{aligned}$$

Notice that, given the first, the second and third of these properties constitute the analogous condition in regular likelihood problems that the expected information is equal to the negative expectation of second derivatives. In addition, the first property given above implies that

$$\begin{aligned} E\left\{\frac{\partial Q_i(\mu_i|\phi)}{\partial \beta_k}\right\} &= E\left\{\frac{\partial Q_i(\mu_i|\phi)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_k}\right\} \\ &= E\left\{U_i(\mu_i|\phi)\frac{\partial \mu_i}{\partial \beta_k}\right\} = 0. \end{aligned} \tag{8.83}$$

Quasi-likelihood functions share the property given in expression (8.83) with true likelihood functions, and suggest that maximum quasi-likelihood estimates of $\boldsymbol{\beta}$ might be found by solving these equations absent the expectation operator for $k = 1, \ldots, p$ (i.e., for the elements of $\boldsymbol{\beta}$).

In a similar way, the second and third properties imply that,

$$\begin{aligned} E\left\{\frac{\partial Q_i(\mu_i|\phi)}{\partial \beta_k}\frac{\partial Q_i(\mu_i|\phi)}{\partial \beta_j}\right\} &= E\left\{\{U_i(\mu_i|\phi)\}^2\frac{\partial \mu_i}{\partial \beta_k}\frac{\partial \mu_i}{\partial \beta_j}\right\} \\ &= -E\left\{\frac{\partial U_i(\mu_i|\phi)}{\partial \mu_i}\partial \beta_k\frac{\partial \mu_i}{\partial \beta_j}\right\} \\ &= \frac{\phi}{V_i(\mu_i)}\frac{\partial \mu_i}{\partial \beta_k}\frac{\partial \mu_i}{\partial \beta_j}. \end{aligned}$$

Given these results, we may derive a Fisher scoring algorithm in an entirely analogous manner to that used in Section 8.3.6 for maximum likelihood estimation of $\boldsymbol{\beta}$ in standard generalized linear models. Now, however, we are maximizing the quasi-likelihood function rather than the log likelihood function. The result is a Fisher scoring algorithm,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \delta\boldsymbol{\beta},$$

where,

$$\delta\boldsymbol{\beta} = \left(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}. \tag{8.84}$$

In (8.84) $\boldsymbol{V}$ is the $n \times n$ diagonal matrix with entries $V_i(\mu_i)$ and $\boldsymbol{D}$ is the $n \times p$ matrix with ik^{th} element $\partial\mu_i/\partial\beta_k$; recall that p is the dimension of $\boldsymbol{\beta}$. Notice that, in this algorithm, the parameter ϕ has canceled, in the same way that it did for development of the Fisher scoring algorithm for standard generalized linear models.

Inference for maximum quasi-likelihood is analogous to Wald theory inference for maximum likelihood. In particular, we can obtain a result analogous to a portion of that presented as Likelihood Theorem 2 in Section 8.3.3, namely that, if $\tilde{\boldsymbol{\beta}}$ denotes the maximum quasi-likelihood estimator of $\boldsymbol{\beta}$, then:

(i) $\tilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

(ii) $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $(n/\phi)\left(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}\right)^{-1}$.

Given the asymptotic normality of $\tilde{\boldsymbol{\beta}}$, Wald theory suggests that inference for $\boldsymbol{\beta}$ be based on normal theory forms with the asymptotic covariance matrix

$$cov(\tilde{\boldsymbol{\beta}}) = \frac{1}{\phi}\left(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}\right)^{-1}. \tag{8.85}$$

Use of (8.85) in practice requires and estimate of ϕ. A common estimator is the moment-based estimator of Example 8.8 in Section 8.1.2,

$$\hat{\phi} = \left[\frac{1}{n-p}\sum_{i=1}^{n}\frac{\{y_i-\hat{\mu}_i\}^2}{V(\hat{\mu}_i)}\right]^{-1}.$$

When might one consider the use of quasi-likelihood in estimation and inference? If the variance functions $V_1(\cdot), \ldots, V_n(\cdot)$ are all the same, and if $V(\cdot)$ corresponds to the variance function from an exponential dispersion family, then quasi-likelihood returns likelihood, as in Example 8.20.

One situation in which quasi-likelihood presents a viable option for estimation and inference is if a standard generalized linear model has been used, but the resulting fit exhibits a less than totally adequate description of the observed variances. That is, we formulate a model with $Y_1, \ldots, Y_n$ distributed according to an exponential dispersion family, which then dictates the variance function $V(\mu_i)$. Diagnostics, such as described in McCullagh and Nelder (1989) may reveal that this assumed variance function does not describe the data in a completely satisfactory manner. We may then choose to envisage a model with "essentially" the same type of random component only with a modified variance function.

Example 8.21

McCullagh and Nelder (1989, Chapter 9.2.4) present an example in which the initial model was a (sort of) binomial random component with logistic link, and hence variance function $V(\mu_i) = \mu_i(1-\mu_i)$ (the response variables Y_i were taken as proportions rather than counts). Diagnostic plots (McCullagh and Nelder 1989, p. 331) showed that the variances appeared to decrease too rapidly (as a function of μ_i) at the extremes (for μ_i close to 0 and 1). They then suggested a variance function $V(\mu_i) = \mu_i^2(1-\mu_i)^2$ as an alternative. This variance function does not correspond to any of the standard exponential dispersion family distributions and, hence, a fully specified random component is no longer available. A quasi-likelihood analysis was then conducted of these data with

$$E(Y_i) = \mu_i; \qquad var(Y_i) = \frac{1}{\phi}V(\mu_i) = \frac{1}{\phi}\mu_i^2(1-\mu_i)^2,$$

and with $\log\{\mu_i/(1-\mu_i)\} = \boldsymbol{x}_i^T\boldsymbol{\beta}$. This then leads to the quasi-likelihood (in

μ_i),

$$Q_i(\mu_i|\phi) = (2y_i - 1)\log\left(\frac{\mu_i}{1-\mu_i}\right) - \frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}.$$

The potential drawbacks to quasi-likelihood are relatively obvious, and include:

1. The asymptotic result given for maximum quasi-likelihood estimators in this section was presented as analogous to *part* of that of Likelihood Theorem 2. The part that is missing in the quasi-likelihood result is the statement of asymptotic efficiency (part (iii) of Likelihood Theorem 2).

2. While inferential procedures are available for parameters in the systematic model component (represented in this section as $\boldsymbol{\beta}$), quasi-likelihood no longer offers a vehicle by which to make inferences about any other portion of the model such as quantiles or other functionals of the distribution.

3. Related to item 2, interpretation of results relative to the underlying scientific phenomenon or mechanism of interest becomes less well-defined. Consider Example 8.21 given above. By replacing the variance function of a binomial with something else, we have admitted that we do not understand the observation process, since we no longer have an actual model. This should (in my opinion) cast substantial doubt on whether we have much of a grasp on modeling the scientific phenomenon of interest through the systematic model component.

Drawing on the third comment above, quasi-likelihood appears in many cases to be an attempt to account for the additional variance in an observable process without having to go to the trouble of modeling it in an adequate manner; effort is diverted to estimation rather than modeling.

Extended Quasi-Likelihood

It was noted immediately following Example 8.20 that, in exponential dispersion family situations, the additive constant that separates a quasi-likelihood and the true log likelihood will depend on both $\boldsymbol{y}$ and ϕ, and we have been careful to write the quasi-likelihood and quasi-score functions as conditional on the dispersion parameter ϕ, as in expression (8.82). With this conditioning, the quasi-likelihood method produces estimates for systematic model component parameters that behave in a manner similar to that of maximum likelihood estimates (minus asymptotic efficiency). If, however, we would like a quasi-likelihood function that is a "sort-of" likelihood in terms of ϕ as well as in terms of $\mu_i; \quad i = 1, \ldots, n$, something else is needed than what has been developed to this point. That development is the intent of what is typically called *extended quasi-likelihood* (Nelder and Pregibon 1987).

Extended quasi-likelihood functions are "derived" in McCullagh and Nelder (1989) by essentially pre-supposing the end product. Barndorff-Nielsen and Cox (1989) give a more convincing progression in which extended quasi-likelihood is derived as a "tilted Edgeworth" expansion (also often called a "saddlepoint approximation"). In either case, what results is the extended quasi-likelihood of Nelder and Pregibon (1987), written for a single random variable Y_i as,

$$\begin{aligned} Q_i^+(\mu_i, \phi) &= \int_{y_i}^{\mu_i} \frac{\phi\{y_i - t\}}{V(t)}\, dt - \frac{1}{2}\log\{2\pi(1/\phi)V(y_i)\} \\ &= Q_i(\mu_i, \phi) - \frac{1}{2}\log\{2\pi(1/\phi)V(y_i)\}. \end{aligned} \tag{8.86}$$

The only difference between $Q_i(\mu_i|\phi)$ in (8.81) and what has been written as $Q_i(\mu_i, \phi)$ in (8.86) is whether ϕ is considered a fixed value in the function, or

part of the argument. As pointed out by McCullagh and Nelder (1989, p. 350) the justification of extended quasi-likelihood as a saddlepoint approximation depends on some type of assumption that renders the contribution of higher-order cumulants to the Edgeworth expansion used negligible. Typically, in such an expansion for a density, cumulants higher than order 4 are dropped (e.g., Stuart and Ord 1987); in derivation of extended quasi-likelihood we make this assumption for cumulants of greater than order 2.

The use of extended quasi-likelihood is perhaps epitomized by models of the generalized linear model form in which we attempt to model both the expectation and variance as functions of covariates. Note that this is the generalized linear model counterpart of additive error models with variance that depends on unknown parameters (see Section 7.2.4). One way to specify such models is given in McCullagh and Nelder (1989, Chapter 10.2) as,

$$E(Y_i) = \mu_i; \qquad g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta};$$

$$var(Y_i) = \frac{1}{\phi_i} V(\mu_i),$$

and,

$$\phi_i = E\{d_i(Y_i, \mu_i)\}; \qquad h(\phi_i) = \boldsymbol{u}_i^T \boldsymbol{\gamma};$$

$$var\{d_i(Y_i, \mu_i)\} = \frac{1}{\tau}\phi^2.$$

In the above model formulation, $d_i(Y_i, \mu_i)$ is chosen as a measure of dispersion, typically,

$$d_i(Y_i, \mu_i) = \frac{Y_i - \mu_i}{V(\mu_i)},$$

$h(\cdot)$ is a "dispersion link function" that connects the expected value of $d_i(Y_i, \mu_i)$, namely ϕ_i, with covariates $\boldsymbol{u}_i$ (often some or all of $\boldsymbol{x}_i$, just as in the additive error models of Section 7.2.4), and the relation between ϕ and $var\{d_i(Y_i, \mu_i)\}$

is dictated if one uses the extended quasi-likelihood (164) for estimation. McCullagh and Nelder (1989, Chapter 10.5) give several additional adjustments to the extended quasi-likelihood procedure that may be desirable.

Dependent Random Variables

Thus far in this subsection our discussion has been entirely in the context of independent response variables Y_i; $i = 1, \ldots, n$. One of the primary areas of application for quasi-likelihood and similar ideas, however, has certainly been longitudinal studies in which random variables correspond to repeated observation of particular sampling units or "individuals" (e.g., Zeger and Liang 1986; Liang and Zeger 1986; Zeger, Liang, and Albert 1988). This type of situation clearly implies correlated random variables within individuals, as we have seen for marginal models in Chapter 7.4. A fundamental property of this setting is that we do, in fact, have independent realizations (across individuals) of some type of (marginal) model depending on the same parameters.

In the dependence setting, quasi-likelihood blends into what we will call "Estimating Functions" in Chapter 8.6, because the starting point is really the quasi-score rather than the quasi-likelihood. In fact, the quasi-likelihood itself is rarely mentioned or defined (but see McCullagh and Nelder 1989, Chapter 9.3.2). The term "Generalized Estimating Equations" has also been used, notably by Zeger and Liang (1986). The fundamental concept follows from noting two things about what was presented previously as the quasi-score function, which may be written, for $k = 1, \ldots, p$, as,

$$\sum_{i=1}^{n} U_i(\mu_i|\phi)\frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{n} \frac{\phi\{y_i - \mu_i\}}{V_i(\mu_i)}\frac{\partial \mu_i}{\partial \beta_k}.$$

Almost trivially, note that the roots of these equations in $\boldsymbol{\beta}$ will not involve ϕ. Secondly, note that we could use the same form of these equations with

y_i and μ_i replaced by vectors $\boldsymbol{y}_i \equiv (y_{i,1}, \ldots, y_{i,n_i})^T$ and $\boldsymbol{\mu}_i \equiv (\mu_{i,1}, \ldots, \mu_{i,n_i})^T$, and $V_i(\mu_i)$ replaced by an $n_i \times n_i$ matrix $\boldsymbol{V}$ which gives, up to the scalar multiple ϕ, the covariance matrix of $\boldsymbol{Y}$. Let k now denote the number of independent vectors $\boldsymbol{y}$ (e.g., number of individuals in a longitudinal study). The above equations may then be written as what Zeger and Liang (1986) called *generalized estimating equations*,

$$\sum_{i=1}^{k} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = 0, \tag{8.87}$$

where $\boldsymbol{S}_i \equiv (\boldsymbol{y}_i - \boldsymbol{\mu}_i)$ is $n_i \times 1$, $\boldsymbol{D}_i$ is $n_i \times p$ with jr^{th} element $\partial \mu_{i,j} / \partial \beta_r$ and $\boldsymbol{V}_i$ is $n_i \times n_i$ with structure to be given directly. For a particular model, an appropriate structure must be chosen for $\boldsymbol{V}_i$; $i = 1, \ldots, k$. Liang and Zeger (1986) suggested using a "working correlation matrix" $\boldsymbol{R}_i(\alpha)$ to help define $\boldsymbol{V}_i$ as,

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\alpha) \boldsymbol{A}_i^{1/2}, \tag{8.88}$$

where $\boldsymbol{A}_i$ is diagonal with elements given by what, in the independence case were the variance functions $V(\mu_i)$ and are now functions $w(\mu_{i,j})$ such that $var(Y_{i,j}) \propto w(\mu_{i,j})$.

Estimates of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are then found by solving the estimating equations (8.87) with the definition of $\boldsymbol{V}_i$ as in (8.88). In order to do this, a value is needed for the correlation matrix parameter α. We will discuss this, and estimation of ϕ shortly. But first, it can be helpful to describe the types of matrices $R(\alpha)$ that might be used in some structures.

1. Independence.

   If $R(\alpha) = I_{n_i}$, then the estimating equations (8.87) reduce to those of quasi-score functions in the independence case. Nevertheless, this choice of $R(\alpha)$ is sometimes used to obtain starting values for iterative algorithms to solve the full equations (8.87) with a different choice of $R(\alpha)$.

2. Random Effects Structure.

$$R(\alpha) = \begin{pmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & \dots & \alpha \\ \vdots & \vdots & \dots & \vdots \\ \alpha & \alpha & \dots & 1 \end{pmatrix}.$$

This is the correlation structure of a linear random effects model in which all variables in the same "group" (here the variables contained in $\boldsymbol{Y}_i$) have the same correlation.

3. Autoregressive Process of Order m.

Here, we would take the uv^{th} element of the matrix $\boldsymbol{R}_i$ to be

$$[\boldsymbol{R}_i]_{u,v} = \begin{cases} \alpha^{|t_{i,u} - t_{i,v}|}, & |t_{i,u} - t_{i,v}| \leq m \\ 0 & |t_{i,u} - t_{i,v}| > m, \end{cases}$$

where $t_{i,u}$ and $t_{i,v}$ are the u^{th} and v^{th} observation times of the individual (subject, group) indexed by i.

For estimation of $\boldsymbol{\beta}$, α, and ϕ, the general prescription is to utilize a Gauss-Newton type of algorithm in which one iteration for $\boldsymbol{\beta}$ is conducted using current estimates of α as fixed, then estimating new α for the current β and so forth. At each iteration, α is updated via a moment-based estimator, and ϕ is also estimated based on a moment-type estimator. We will not present the details here (since we've covered the basic ideas already), but see Zeger and Liang (1986) or Pawitan (2001) for all the details.

Inference about $\boldsymbol{\beta}$ is made using Wald theory asymptotics based on asymptotic normality of the GEE (generalized estimating equations) estimates. The form of the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ is given in Zeger and Liang

(1986) and Liang and Zeger (1986), and is estimated by using moment-based estimators $\hat{\alpha}$ and $\hat{\phi}$ as "plug-in" values.

8.5.2 Estimating Functions

The generalized estimating equations of the previous subsection are a special case of what are known as *estimating functions*. Given random variables $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)^T$ with a density $f_Y(\boldsymbol{y}|\boldsymbol{\theta})$, an unbiased estimating function for $\boldsymbol{\theta}$ can be defined quite simply as any function $v(\boldsymbol{Y}, \boldsymbol{\theta})$ such that, for all values of $\boldsymbol{\theta} \in \Theta$,

$$E\{v(\boldsymbol{Y}, \boldsymbol{\theta})\} = 0. \tag{8.89}$$

If we are interested in only a part of $\boldsymbol{\theta}$, then the estimating function is defined relative to only that component or components.

Example 8.22

Let $Y_1, \ldots, Y_n$ be $iid\, N(\mu, \sigma^2)$. Possible estimating functions for μ include,

$$\begin{aligned} v(\boldsymbol{y}, \mu) &= \sum_{i=1}^{n} (y_i - \mu) = 0, \\ v(\boldsymbol{y}, \mu) &= \sum_{i=1}^{n} sgn(y_i - \mu) = 0, \\ v(\boldsymbol{y}, \mu) &= \sum_{i=1}^{n} \psi(y_i - \mu) = 0, \end{aligned}$$

where in this third possibility, for some real number r,

$$\psi(x) = x(r^2 - x^2)^2 I(-r < x < r).$$

Estimates resulting from solving these three possible estimating functions for μ are the sample mean, the sample median, and a "robust" estimator that

corresponds to use of the function $\psi(\cdot)$ which in the above is known as the *biweight* or *bisquare* function, attributed to Tukey (e.g., Hampel *et al.* 1986).

Asymptotic inference for estimating functions is developed in much the same way as for maximum likelihood estimators and (although we did not cover it in any detail) maximum quasi-likelihood estimators. Specifically, in the case of a scalar parameter θ, suppose conditions sufficient to show the following are assumed:

1. The solution to an unbiased estimating function $\tilde{\theta}$ may be expanded as

$$v(\boldsymbol{y}, \tilde{\theta}) \approx v(\boldsymbol{y}, \theta) + \left. \frac{\partial v(\boldsymbol{y}, \tilde{\theta})}{\partial \tilde{\theta}} \right|_{\tilde{\theta}=\theta} (\tilde{\theta} - \theta),$$

   from which we have, since $v(\boldsymbol{y}, \tilde{\theta}) = 0$ by definition,

$$\tilde{\theta} - \theta = -v(\boldsymbol{y}, \theta) \left\{ \left. \frac{\partial v(\boldsymbol{y}, \tilde{\theta})}{\partial \tilde{\theta}} \right|_{\tilde{\theta}=\theta} \right\}^{-1}.$$

2. The random version of $v(\cdot)$ satisfies,

$$n^{-1/2} v(\boldsymbol{Y}, \theta) \;\text{ is }\; AN(0, q(\theta)),$$

   for some function $q(\cdot)$.

3. The random version of the derivative of $v(\cdot)$ satisfies,

$$\frac{1}{n} \frac{\partial v(\boldsymbol{Y}, \theta)}{\partial \theta} \xrightarrow{p} s(\theta),$$

   for some function $s(\cdot)$.

Then the result is (Barndorff-Nielsen and Cox 1994, p.303),

$$\tilde{\theta} \;\text{ is }\; AN\left(\theta, \frac{q(\theta)}{n\, s^2(\theta)}\right).$$

There appear to be two major areas in which estimating functions surface. The first is an an alternative "theory of estimation", in which estimating functions are presented as an alternative to both least squares (viewed as an exact theory procedure) and maximum likelihood. This is, for example, the view offered by Godambe and Kale (1991) in which the authors claim to demonstrate that estimating functions "unifies the method of maximum likelihood and the method of minimum variance unbiased estimation in the case of parametric models" (Godambe and Kale 1991, p.17). Under this viewpoint we must deal with the development of criteria under which one of a variety of possible estimating functions can be deemed "optimal". Clearly, there will be any number of unbiased estimating functions that can be developed in most situations (as illustrated in Example 8.22). A variety of possible optimality criteria are offered by Godambe and Heyde (1987). A slightly different view is offered by Barndorff-Nielsen and Cox (1994) who point out that, in the case of $Y_1, \ldots, Y_n \sim iid$ with common density $f(y|\theta)$ for a scalar θ, the two matrices $q(\cdot)$ and $s(\cdot)$ in conditions 2 and 3 listed above for asymptotic normality of $\tilde{\theta}$ become

$$q(\theta) = var\left\{\frac{\partial \log\{f(y_i|\theta)}{\partial\theta}\right\},$$

$$s(\theta) = E\left\{\frac{\partial^2 \log\{f(y_i|\theta)}{\partial\theta^2}\right\},$$

so that $q(\theta) = -s(\theta)$ and the asymptotic result leads to the solution to the likelihood equation (an estimating function) being asymptotically normal with variance given by the inverse information. This then motivates them to suggest a criterion

$$\rho_v(\theta) = \frac{[E\{\partial v(\boldsymbol{Y},\theta)/\partial\theta\}]^2}{var\{v(\boldsymbol{Y},\theta)\}},$$

as a measure of the lost efficiency of the solution to any estimating function $v(\boldsymbol{Y}, \theta)$; this is because $\rho_v(\theta)$ is maximized by the likelihood score function.

The second major area in which estimating functions play a role is in the development of *robust* estimators, as indicated by the third possible estimating function given in Example 8.22 for the one-sample normal problem. This is the flavor of the presentation by, for example, Pawitan (2001, Chapter 14). In this context, estimating functions are also often called *M-estimators*. See, for example, Hampel *et. al.* (1986), and Carroll and Ruppert (1988, Chapter 7).

One aspect of estimating functions that deserves mention is that, if $v_1(\cdot), \ldots, v_p(\cdot)$ are unbiased estimating functions, then any linear combination, $\sum a_j v_j$ is also an unbiased estimating function. This property can sometimes be utilized to form simple estimating functions in even complex situations. For example, in the autoregressive process

$$Y(t) = \theta Y(t-1) + \epsilon(t),$$

where $\epsilon(t) \sim iid\, N(0,1)$, we could take estimating functions,

$$v_t = y(t) - \theta y(t-1),$$

and form the combination,

$$v(\boldsymbol{y}, \theta) = \sum_t y(t-1)\, v_t.$$

An example of such development is given in McCullagh and Nelder (1989, p.341).

8.5.3 Pseudo-Likelihood

In one sense, all of the discussion so far in Chapter 8.5 could be categorized under the heading of *pseudo-likelihood* in that the prefix "pseudo-" implies

false or spurious. On the other hand, the distinction between the functions we have discussed under the heading *quasi-likelihood* and those to be considered in this subsection is perhaps worth maintaining. This is because the functions considered as *quasi-likelihoods* were developed in an attempt to maintain likelihood-like behavior.

Even the generalized estimating equations of Section 8.5.1 and the estimating functions of Section 8.5.2 mimic at least the derivatives of log likelihoods in asymptotic behavior. In one sense, those methods "abandoned a fully specified model" in order to maintain something akin to a likelihood function. In this subsection, however, we consider functions that are decidedly *not* likelihoods, but which are used "as if they were" likelihoods, in order to make estimation possible (or easier). In the first of these we discuss, formulated for estimation of unknown variance parameters in additive error models, the model does not necessarily make a full distributional assumption. In the second form of pseudo-likelihood we consider, the model is maintained, but the likelihood is not. Thus, one way to think of (at least this second type of) pseudo-likelihood is that we are willing to "abandon the complete likelihood" in order to maintain the model.

While many pseudo-likelihood functions are possible in varying situations (technically, anything that is not a likelihood could be considered a pseudo-likelihood) we will consider two examples of pseudo-likelihoods that have proven useful.

The Pseudo-Likelihood of Carroll and Ruppert

This version of what is sometimes called pseudo-likelihood was suggested by Carroll and Ruppert (1988) for estimation of unknown parameters in the vari-

ance functions of regression models. Consider, then, the general form of such a model discussed in Section 7.2.4,

$$Y_i = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, z_i, \theta)\, \epsilon_i,$$

where, for $i = 1, \ldots, n$, $\epsilon_i \sim iid\, F$ such that $E(\epsilon_i) = 0$, and $var(\epsilon_i) = 1$.

The functions $g_1(\cdot)$ and $g_2(\cdot)$ are assumed to be known, smooth functions, $\boldsymbol{x}_i;\ i = 1, \ldots, n$ are known covariates involved in the expectation function, $\boldsymbol{\beta}$ are unknown regression parameters, and z_i are covariates that may be involved in the variance model but not the expectation model. To simplify presentation, we will assume that $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ enter the variance function $g_2(\cdot)$ only through the expectation, which we will now denote as $\mu_i(\boldsymbol{\beta}) \equiv g_1(\boldsymbol{x}_i, \boldsymbol{\beta})$; we must keep in mind, with this notation, that $\mu_i(\boldsymbol{\beta})$ is a function of the covariates $\boldsymbol{x}_i$ as well as $\boldsymbol{\beta}$. Then, the model becomes

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma g(\mu_i(\boldsymbol{\beta}), z_i, \theta)\, \epsilon_i, \tag{8.90}$$

which was previously given as expression (7.10) in Section 7.2.4.

Now, we have seen that, with θ considered known in (8.90) the regression parameters $\boldsymbol{\beta}$ may be estimated using the generalized least squares algorithm outlined in Section 8.2.2, and inferences about the values of $\boldsymbol{\beta}$ may be accomplished through the Fundamental Theorem of Generalized Least Squares, which provides asymptotic normality for $\hat{\boldsymbol{\beta}}$. Under this strategy, the parameter σ^2 is usually obtained from a moment-based estimator given as expression (8.26) (if we make the substitution $\mu_i(\boldsymbol{\beta}) \equiv g_1(\boldsymbol{x}_i, \boldsymbol{\beta})$.

For model (8.90), however, we are not assuming that θ is known, and the pseudo-likelihood strategy of Carroll and Ruppert (1988) is an attempt to allow its estimation without making full distributional assumptions on the model. Suppose then, for the moment, that $\boldsymbol{\beta}$ is known to be equal to a particular value

$\boldsymbol{\beta}^{(0)}$, say. As Carroll and Ruppert (1988, p.71) put it, "pretend" that the ϵ_i have normal distributions so that $Y_i \sim \mathit{indep}\, N(\mu_i(\boldsymbol{\beta}^{(0)}),\, \sigma^2 g^2(\mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta))$. Then a log *pseudo-likelihood* for θ and σ^2 could be written as,

$$\begin{aligned} L_*(\theta, \sigma^2 | \boldsymbol{\beta}^{(0)}) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n} \log\left[g^2\left\{\mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta\right\}\right] \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{y_i - \mu_i(\boldsymbol{\beta}^{(0)})}{g\left\{\mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta\right\}}\right]^2 . \end{aligned} \tag{8.91}$$

One way to maximize the pseudo-likelihood (8.91) in θ and σ^2, is to apply the idea of profiling for θ. That is, if take the partial derivative of (8.91) with respect to σ^2 and set equal to zero, the solution is,

$$\hat{\sigma}^2(\theta | \boldsymbol{\beta}^{(0)}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \mu_i(\boldsymbol{\beta}^{(0)})}{g\left\{\mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta\right\}}\right]^2 . \tag{8.92}$$

That is, the maximum pseudo-likelihood estimate of σ^2 is a function of θ. Thus, to maximize (8.91) in θ and σ^2, we maximize in θ what could be called a log-profile-pseudo-likelihood, formed by substituting the solution (8.92) into (8.91) to arrive at,

$$L_*^p(\theta | \boldsymbol{\beta}^{(0)}) = -\frac{n}{2}\log\left\{\hat{\sigma}^2(\theta | \boldsymbol{\beta}^{(0)})\right\} - \frac{1}{2}\sum_{i=1}^{n}\log\left[g^2\left\{\mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta\right\}\right]. \tag{8.93}$$

Thus, to find maximum pseudo-likelihood estimates of θ and σ^2, for a fixed value $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$, we would maximize (8.93) in θ, and substitute the maximizing value into (8.92) to estimate σ^2.

Estimation of the full set of parameters $\{\boldsymbol{\beta}, \theta, \sigma^2\}$ by this strategy consists of beginning with an initial value $\boldsymbol{\beta}^{(0)}$, estimating θ from (8.93) with $\hat{\sigma}^2(\theta | \boldsymbol{\beta}^{(0)})$ given in (8.92), completing the steps of the generalized least squares algorithm

of Section 8.2.2 to obtain updated estimates of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^{(1)}$, repeating estimation of θ as above with $\boldsymbol{\beta}^{(1)}$ replacing $\boldsymbol{\beta}^{(0)}$, returning to the generalized least squares algorithms with the new value of θ, and so forth until a given stopping rule is met (recall the discussion of stopping rules in generalized least squares). In essence, all that has been done is to insert and estimation phase for θ between steps 1 and 2 of the generalized least squares algorithm.

There is no one clear path for making inference about the parameters using this pseudo-likelihood procedure. A common approach for making inferential statements about the regression parameters $\boldsymbol{\beta}$ is to fix θ at its estimated value (from the pseudo-likelihood procedure above) and then use the results of the Fundamental Theorem of Generalized Least Squares, usually with the moment-based estimator of σ^2 rather than the pseudo-likelihood estimator (8.92). A criticism of this approach is that uncertainty in the estimation of θ is not accounted for in the estimation of $\boldsymbol{\beta}$. A number of possible ways to make inference about θ are discussed in Carroll and Ruppert (1988, Chapter 3.4). Rather than go into detail about these possible methods, we will simply conclude this discussion of Carroll and Ruppert's pseudo-likelihood with a few comments about what might motivate its use, and connections with other estimation approaches we have discussed.

1. The entire concept of using a pseudo-likelihood for models such as (8.90) is based on the desire to maintain the "distribution-free" flavor of generalized least squares. An obvious alternative is to just assume normality in the first place, and apply full maximum likelihood estimation to all parameters in the model (possibly making use of profiling methods if needed). One motivation for making use of the pseudo-likelihood strategy then is to keep the potential robustness properties of least squares in

effect for estimation of $\boldsymbol{\beta}$, although whether this is truly possible without further restrictions on the variance (e.g., σ^2 is "small") remains less clear.

2. Following the point of comment 1, Carroll and Ruppert (1988, Chapter 6.4) extend the pseudo-likelihood estimation of θ to be instead based on an estimating function within the context of *M-estimators*. The connection between estimating functions and the development of robust estimators was briefly mentioned in these notes toward the end of Section 8.5.3.

3. Although robustness may motivate, to some extent, the use of pseudo-likelihood, we should be careful not to interpret robustness here to also imply *resistance*. Pseudo-likelihood, similar to full maximum likelihood based on an assumption of normal distributions, is typically sensitive to extreme observations. If such extreme values do, in fact, correspond to errors in data collection or recording, pseudo-likelihood has provided no additional protection against their effects over that given by full maximum likelihood.

The Pseudo-Likelihood of Besag

The second type of Pseudo-likelihood we will briefly mention is constructed in an entirely different context that that of the preceding sub-section. This pseudo-likelihood was suggested by Besag (1974) for use with conditionally specified models and, in particular, conditionally specified models having conditional distributions in one-parameter exponential families (or exponential dispersion families). We have discussed only one of these models in any detail, that being a Gaussian conditionals model, sometimes called a conditional au-

toregressive model (see Section 7.5.5). In that case, the joint distribution was available in closed form as a Gaussian density, and the full likelihood could be used (most easily through profiling methods). In the general case, however, the joint density or mass function is available only up to an unknown normalizing constant which involves the parameters of interest. While estimation in such cases can be approached in a number of ways (e.g., Monte Carlo maximum likelihood), most of these methods are either complex in theory or computationally intensive, or both. Besag (1974) suggested a type of pseudo-likelihood to deal with these difficulties.

The general situation is as follows. For a set of random variables $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, suppose that a conditionally specified model has been formulated through the set of full conditional density or mass functions, for $i = 1, \ldots, n$,

$$f_i(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : y \neq i\}) =$$

$$\exp\left[A_i(\{y(\boldsymbol{s}_j) : j \neq i\})T_i(y(\boldsymbol{s}_i)) - B_i(\{y(\boldsymbol{s}_j) : j \neq i\}) + C(y(\boldsymbol{s}_i))\right]; \qquad y(\boldsymbol{s}_i) \in \Omega_i. \tag{8.94}$$

A density or mass function in the form of (8.94) is written as a one-parameter exponential family, with the function $A_i(\{y(\boldsymbol{s}_j) : j \neq i\})$ playing the role of the natural parameter. Now, for a model such as (172) Besag (1974) showed that a *necessary* parameterization for $A_i(\cdot)$ is,

$$A_i(\{y(\boldsymbol{s}_j) : j \neq i\}) = \alpha_i + \sum_{j=1}^{n} c_{i,j} T_j(y(\boldsymbol{s}_j)), \tag{8.95}$$

where $c_{i,j} = c_{j,i}$, $c_{i,i} = 0$, and $c_{i,j} = 0$ unless locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are neighbors (i.e., $c_{i,j} = 0$ for $\boldsymbol{s}_j \notin N_i$).

Given difficulties in deriving the joint density (or mass function) of $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$ in closed form, to use the joint in estimation becomes a complex matter. Let

$\boldsymbol{\theta}$ denote the collection of parameters in the model. For example, in the conditional Gaussian model used in the spatial analysis of monitoring nitrates in the Des Moines River (Section 7.5.5) $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \eta, \tau^2, k)$. Besag (1974) suggested the following pseudo-likelihood to accomplish estimation of $\boldsymbol{\theta}$.

$$\ell_*(\boldsymbol{\theta}) \equiv \prod_{i=1}^{n} f_i(y(\boldsymbol{s}_i)|\boldsymbol{\theta}, \{y(\boldsymbol{s}_j) : y \neq i\}), \tag{8.96}$$

or, in log form,

$$L_*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left\{f_i(y(\boldsymbol{s}_i)|\boldsymbol{\theta}, \{y(\boldsymbol{s}_j) : y \neq i\})\right\}. \tag{8.97}$$

In (8.96) and (8.97) we have made explicit the dependence of the conditional density functions on the parameter vector $\boldsymbol{\theta}$.

The pseudo-likelihood (8.96) or log pseudo-likelihood (8.97) are used exactly as likelihood functions for estimation of $\boldsymbol{\theta}$, although the usual asymptotic inference does not apply. For a sketch of some of the asymptotic results available, see Cressie (1993, p. 487).

8.6 Parametric Bootstrap

What is known as a *parametric bootstrap* is a widely applicable method for assessing uncertainty in parameter estimates, often in the form of interval estimation, although the formation of prediction regions or intervals is also a clear area of application. We will present the parametric bootstrap in the case of a scalar parameter θ, although it appears that the same essential ideas could be easily extended to confidence regions for vector-valued parameters. In particular, since the parametric bootstrap is a version of simulation-based inference, using a model with $\boldsymbol{\theta} \equiv (\theta_1, \boldsymbol{\theta}_2)$ and simulating in the manner to be described below results in an assessment of the marginal distribution of an estimator $\hat{\theta}_1$;

often, but not necessarily, $\hat{\theta}_1$ is a maximum likelihood estimator.

8.6.1 Notation and Basic Simulation Estimators

To set the basic notation for this section, let $Y_1, \ldots, Y_n$ be independent random variables that follow a model giving joint density or mass function $f(y_1, \ldots, y_n|\theta)$. Suppose that an estimator of θ, $\hat{\theta}$ say, is available by some means (maximum likelihood, least squares, quasi-likelihood, an estimating function, or maximum pseudo-likelihood). We assume that $\hat{\theta}$ is a function of the observations, $\hat{\theta} \equiv \hat{\theta}(\boldsymbol{y})$, whether or not that function can be expressed in closed form; this dependence will be assumed in writing $\hat{\theta}$. Substitution of $\hat{\theta}$ for θ in the model gives the *fitted model* with density or mass function $f(y_1, \ldots, y_n|\hat{\theta})$ and distribution function $F(y_1, \ldots, y_n|\hat{\theta})$. The basic simulation process is to generate observations $y_1^*, \ldots, y_n^*$ from the fitted model and then calculate the estimator of θ from these simulated values, which we will denote θ^*. If this process is repeated a number of times, denoted by M, we obtain the values $\theta_1^*, \ldots, \theta_M^*$. The underlying idea is that the distribution of a function of the true parameter and the actual estimate, $h(\theta, \hat{\theta})$, can be approximated by the empirical distribution of that same function using $\hat{\theta}$ as the "true" parameter value and the values $\theta_1^*, \ldots, \theta_M^*$ as estimates. Thus, for example, the bias associated with $\hat{\theta}$, namely $B(\hat{\theta}) \equiv E\{\hat{\theta} - \theta\}$, may be approximated by,

$$B_M(\hat{\theta}) = \frac{1}{M} \sum_{m=1}^{M} (\theta_m^* - \hat{\theta}) = \bar{\theta}^* - \hat{\theta}. \tag{8.98}$$

Similarly, the variance of $\hat{\theta}$ is approximated by

$$V_M(\hat{\theta}) = \frac{1}{M-1} \sum_{m=1}^{M} (\theta_m^* - \bar{\theta}^*)^2. \tag{8.99}$$

In the same way, quantiles of the distribution of $h(\theta, \hat{\theta})$ are approximated by using the empirical distribution of $h(\hat{\theta}, \theta^*)$. That is, let $h_1, \ldots, h_M$ denote the values of $h(\hat{\theta}, \theta^*_m)$; $m = 1, \ldots, M$. If $G(u) = Pr\{h(\theta, \hat{\theta}) \leq u\}$, then $G(\cdot)$ is approximated by

$$G_M(u) = \frac{1}{M} \sum_{m=1}^{M} I\{h(\hat{\theta}, \theta^*_m) \leq u\}, \tag{8.100}$$

where $I(\cdot)$ is the usual indicator function. Now, if we are interested in quantiles of the distribution of $h(\theta, \hat{\theta})$, we use the general result that, if $X_1, \ldots, X_N$ are independently distributed with distribution function F_X, and if $X_{[1]}, \ldots, X_{[n]}$ denote the ordered values, then

$$E\{X_{[k]}\} \approx F^{-1}\left(\frac{k}{n+1}\right),$$

leading to the estimate of the q^{th} quantile of F_X, which is $x_q = F_X^{-1}(q)$ as,

$$x_q = X_{[(n+1)q]}.$$

Thus, the estimated value of the q^{th} quantile of $h(\theta, \hat{\theta})$ is the $(n+1)q^{th}$ largest value of $\{h(\hat{\theta}, \theta^*_m) : m = 1, \ldots, M\}$.

8.6.2 Normal Approximation Intervals

Consider a situation in which we are willing to accept (approximate or asymptotic) normality for $\hat{\theta}$, but the variance or standard error of this estimator is unavailable. If $\hat{\theta}$ is AN with mean $\theta + B(\hat{\theta})$ and variance $V(\hat{\theta})$ then an approximate normal interval can be derived in the usual way from,

$$Pr\{L_\alpha \leq \theta \leq U_\alpha\} = 1 - \alpha,$$

which leads to

$$\begin{aligned} L_\alpha &= \hat{\theta} - B(\hat{\theta}) - V^{1/2}(\hat{\theta}) z_{1-\alpha/2} \\ U_\alpha &= \hat{\theta} - B(\hat{\theta}) + V^{1/2}(\hat{\theta}) z_{1-\alpha/2}, \end{aligned} \tag{8.101}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. To see this, note that it is the quantity

$$\frac{\hat{\theta} - B(\hat{\theta})}{V^{1/2}(\hat{\theta})},$$

which is approximately $N(0,1)$. The derivation then proceeds exactly as in elementary statistics courses using $\hat{\theta} - B(\hat{\theta})$ in place of an unbiased estimator of θ.

To use the interval (8.101) in practice requires only estimation of $B(\hat{\theta})$ and $V(\hat{\theta})$ as given by $B_M(\hat{\theta})$ in (176) and $V_M(\hat{\theta})$ as given in (8.99). A bootstrap normal approximation interval for θ is then,

$$\left(\hat{\theta} - B_M(\hat{\theta}) - V_M^{1/2}(\hat{\theta}) z_{1-\alpha/2},\ \ \hat{\theta} - B_M(\hat{\theta}) + V_M^{1/2}(\hat{\theta}) z_{1-\alpha/2}\right). \tag{8.102}$$

8.6.3 Basic Bootstrap Intervals

Now suppose we are in a situation in which we are reluctant to assume approximate normality for our estimator $\hat{\theta}$. The distributional form of $\hat{\theta}$ can, by the way, be assessed for many functions such as $h(\theta, \hat{\theta}) = \hat{\theta} - \theta$ through examination of the empirical distribution of the values $\theta_1^*, \ldots, \theta_M^*$ for M sufficiently large (e.g., $M > 1000$).

We may desire to compute an interval with equal probability in each tail, that is, equal left-tail and right-tail errors as,

$$Pr\{\theta \leq L_\alpha\} = Pr\{U_\alpha \leq \theta\} = \alpha/2.$$

Then we desire L_α and U_α such that

$$\begin{aligned} Pr\{\theta - \hat{\theta} \leq L_\alpha - \hat{\theta}\} &= \alpha/2 \\ \\ Pr\{U_\alpha - \hat{\theta} \leq \theta - \hat{\theta}\} &= \alpha/2, \end{aligned}$$

which implies that,

$$Pr\{\hat{\theta} - L_\alpha \leq \hat{\theta} - \theta\} = \alpha/2$$

$$Pr\{\hat{\theta} - \theta \leq \hat{\theta} - U_\alpha\} = \alpha/2,$$

or,

$$Pr\{\hat{\theta} - \theta \leq \hat{\theta} - L_\alpha\} = 1 - \alpha/2$$

$$Pr\{\hat{\theta} - \theta \leq \hat{\theta} - U_\alpha\} = \alpha/2.$$

Using the notation of Section 8.6.1, if we knew the distribution of $h(\theta, \hat{\theta}) = (\hat{\theta} - \theta)$ as $G(\cdot)$, then we would have,

$$\hat{\theta} - L_\alpha = v_{1-\alpha/2}$$

$$\hat{\theta} - U_\alpha = v_{\alpha/2}, \tag{8.103}$$

where v_α denotes the α quantile of the distribution $G(\cdot)$. Now, we do not know $G(\cdot)$, but we can approximate it using a parametric bootstrap as in expression (8.100). Following the development presented immediately following expression (8.100) we approximate the values $v_{\alpha/2}$ and $v_{1-\alpha/2}$ with

$$v_{1-\alpha/2} \approx (\theta^* - \hat{\theta})_{[(M+1)(1-\alpha/2)]}$$

$$v_{\alpha/2} \approx (\theta^* - \hat{\theta})_{[(M+1)(\alpha/2)]}$$

which, from expression (8.103) leads to,

$$L_\alpha = \hat{\theta} - (\theta^* - \hat{\theta})_{[(M+1)(1-\alpha/2)]}$$

$$\begin{aligned} &= \hat{\theta} - (\theta^*_{[(M+1)(1-\alpha/2)]} - \hat{\theta}) \\ &= 2\hat{\theta} - \theta^*_{[(M+1)(1-\alpha/2)]}, \end{aligned}$$

and,

$$\begin{aligned} U_\alpha &= \hat{\theta} - (\theta^* - \hat{\theta})_{[(M+1)(\alpha/2)]} \\ &= \hat{\theta} - (\theta^*_{[(M+1)(\alpha/2)]} - \hat{\theta}) \\ &= 2\hat{\theta} - \theta^*_{[(M+1)(\alpha/2)]}. \end{aligned}$$

A $(1-\alpha)100\%$ interval estimate for θ is then,

$$\left(2\hat{\theta} - \theta^*_{[(M+1)(1-\alpha/2)]},\, 2\hat{\theta} - \theta^*_{[(M+1)(\alpha/2)]}\right). \tag{8.104}$$

The interval (8.104) is called the *basic bootstrap confidence interval* for θ by Davison and Hinkley (1997). Note that this interval estimate assumes that things have been arranged so that $(M+1)(\alpha/2)$ and $(M+1)(1-\alpha/2)$ are integers. This is not difficult if M is large; for example, take $\alpha/2 = 0.05$ and $M = 9999$ for a 90% interval.

The normal approximation interval given in expression (8.102) has the typical property of being symmetric, in this case about $\hat{\theta} - B(\hat{\theta})$; it will be symmetric about $\hat{\theta}$ if that estimator is known to be unbiased, in which case we take $B(\hat{\theta}) = 0$ rather than estimating it through expression (8.98). The basic bootstrap interval (8.104) however, is not necessarily symmetric. It will be symmetric or close to symmetric as the empirical distribution of $\{\theta^*_m : m = 1, \ldots, M\}$ is symmetric or close to symmetric.

8.6.4 Percentile Bootstrap Intervals

We will briefly mention one other approach to the formulation of bootstrap confidence intervals, called *percentile methods* by Davison and Hinkley (1997, Chapter 5.3). The origin of the name is not intuitively obvious but presumably comes from the fact that we end up using the $(\alpha/2)$ percentile and $(1-\alpha/2)$ percentile of the empirical distribution of bootstrap estimates θ^* directly as interval endpoints, as will be shown below. Percentile methods have been modified (or adjusted) in a number of ways that seem to offer some improvement over normal approximation or basic bootstrap intervals in meeting nominal coverage goals, although assessments have been made primarily through the use of nonparametric bootstrap sampling (see, e.g., Davison and Hinkley 1997, Chapter 5.4 and Chapter 5.7). The performance of percentile methods in parametric bootstrap, which is our concern here, is less well understood.

Suppose that, in application of the basic bootstrap method of Section 8.6.3, there exists some transformation of the estimator $\hat{\theta}$, say $\hat{\phi} \equiv W(\hat{\theta})$, such that the distribution of $\hat{\phi}$ is known to be symmetric; for the moment the existence of such a function $W(\cdot)$ is all that matters, not its identity. Consider, then, applying this transformation to $\hat{\theta}$ and then using the basic bootstrap method to form an interval for $\phi = W(\theta)$, with the following modifications. First, notice that we now have estimates of functionals of the distribution of $h(\phi, \hat{\phi})$ as in Section 8.6.1 through the bootstrap simulations $\{h(\hat{\phi}, \phi^*_m) : m = 1, \ldots, M\}$. In the basic bootstrap method we take $h(\phi, \hat{\phi}) = \phi - \hat{\phi}$ and $h(\hat{\phi}, \phi^*_m) = \hat{\phi} - \phi^*_m$. The development of the basic bootstrap interval for ϕ proceeds up to expression (8.103) exactly as given in Section 8.6.3 with ϕ replacing θ, $\hat{\phi}$ replacing $\hat{\theta}$ and ϕ^* replacing θ^*. Now, however, the symmetry of the distribution of $\hat{\phi} - \phi$ indicates that $v_{1-\alpha/2} = -v_{\alpha/2}$ so that expression (8.103) may be changed to

(in terms of ϕ),

$$\begin{aligned} \hat{\phi} - L_\alpha &= -v_{\alpha/2} \\ \\ \hat{\phi} - U_\alpha &= -v_{1-\alpha/2}, \end{aligned}$$

where v_α is now a quantile from the distribution of $\hat{\phi} - \phi$ rather than $\hat{\theta} - \theta$. Then,

$$\begin{aligned} v_{\alpha/2} &\approx (\phi^* - \hat{\phi})_{[(M+1)(\alpha/2)]}, \\ v_{1-\alpha/2} &\approx (\phi^* - \hat{\phi})_{[(M+1)(1-\alpha/2)]}, \end{aligned}$$

which leads to,

$$\begin{aligned} L_\alpha &= \hat{\phi} + (\phi^* - \hat{\phi})_{[(M+1)(\alpha/2)]} \\ &= \phi^*_{[(M+1)(\alpha/2)]} \end{aligned}$$

and,

$$\begin{aligned} U_\alpha &= \hat{\phi} + (\phi^* - \hat{\phi})_{[(M+1)(1-\alpha/2)]} \\ &= \phi^*_{[(M+1)(1-\alpha/2)]}. \end{aligned}$$

A $(1-\alpha)100\%$ basic bootstrap estimate for ϕ is then

$$\left(\phi^*_{[(M+1)(\alpha/2)]},\ \phi^*_{[(M+1)(1-\alpha/2)]}\right). \tag{8.105}$$

Now, if the transformation $W(\cdot)$ that produced ϕ from θ, $\hat{\phi}$ from $\hat{\theta}$ and ϕ^*_m from θ^*_m was monotone, then $\phi^*_{[k]}$ corresponds to $\theta^*_{[k]}$ for any integer $k \in \{1, \ldots, M\}$. Transforming the interval (8.105) back to the θ scale then results in the bootstrap percentile interval for θ of,

$$\left(\theta^*_{[(M+1)(\alpha/2)]},\ \theta^*_{[(M+1)(1-\alpha/2)]}\right). \tag{8.106}$$

What is surprising, then, is that such an interval can be formulated (and computed) for θ without ever determining what the transformation $\phi = W(\theta)$ might be.

8.6.5 Predication Intervals

Another use of parametric bootstrap is in forming prediction intervals for a new random variable Y^0, assumed to follow the same model as $Y_1, \ldots, Y_n$. Let the model evaluated at a possible value y^0 be denoted as $F(y^0|\theta)$. Given an estimated parameter $\hat{\theta}$ the natural starting point is an interval with endpoints given by the $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of the estimated model $F(y^0|\hat{\theta})$. Denote these values as

$$\begin{aligned} q(\hat{\theta})_{\alpha/2} &\equiv F^{-1}(\alpha/2 \,|\, \hat{\theta}) \\ q(\hat{\theta})_{1-\alpha/2} &\equiv F^{-1}(1-\alpha/2 \,|\, \hat{\theta}). \end{aligned} \tag{8.107}$$

The interval $(q_(\hat{\theta})\alpha/2,\, q(\hat{\theta})_{1-\alpha/2})$ will be overly optimistic (i.e., too short) because it does not take into account uncertainty in the estimation of θ by $\hat{\theta}$. That is, if we knew the true value θ, it would be the case that

$$Pr\left[q(\theta)_{\alpha/2} \leq Y^0 < q(\theta)_{1-\alpha/2} \,|\, \theta\right] = 1-\alpha.$$

Since we do not know θ but are estimating it with $\hat{\theta}$ we need to assess the actual coverage rate,

$$Pr\left[q(\hat{\theta})_{\alpha/2} \leq Y^0 < q(\hat{\theta})_{1-\alpha/2} \,|\, \theta\right] = 1-c(\alpha). \tag{8.108}$$

If there is a functional relation between $c(\alpha)$ and α, then we could "adjust" the procedure to use $q(\hat{\theta})_{\alpha'/2}$ and $q(\hat{\theta})_{1-\alpha'/2}$, where α' is chosen such that $c(\alpha') = \alpha$. The essential problem, then, is estimation of $c(\alpha)$ in expression (8.108). A parametric bootstrap may be used for this estimation in the following way.

In the notation of Section 8.6.1, the function of θ and $\hat{\theta}$ to be estimated is,

$$\begin{aligned} h(\theta, \hat{\theta}) &= Pr\left[q(\hat{\theta})_{\alpha/2} \leq Y^0 < q(\hat{\theta})_{1-\alpha/2} \,|\, \theta\right] \\ &= E\left\{I\left[q(\hat{\theta})_{\alpha/2} \leq Y^0 < q(\hat{\theta})_{1-\alpha/2} \,|\, \theta\right]\right\}. \end{aligned}$$

Given a fitted model through $\hat{\theta}$, simulate bootstrap data sets $\boldsymbol{y}^*_m \equiv (y^*_1, \ldots, y^*_n)^T_m$ in the usual way from $F(y_1, \ldots, y_n|\hat{\theta})$ to obtain bootstrap estimates θ^*_m; $m = 1, \ldots, M$. Also simulate values of the predictand y^0_m; $m = 1, \ldots, M$ from the fitted model $F(y^0|\hat{\theta})$. Compute the intervals $(q^*_{(\alpha/2),m},\ q^*_{(1-\alpha/2),m})$ with nominal coverage $1 - \alpha$ for each bootstrap data set as,

$$\begin{aligned} q^*_{(\alpha/2),m} &\equiv F^{-1}(\alpha/2 \,|\, \theta^*_m) \\ q^*_{(1-\alpha/2),m} &\equiv F^{-1}(1 - \alpha/2 \,|\, \theta^*_m). \end{aligned} \tag{8.109}$$

Estimate the actual coverage of the interval as,

$$1 - c_M(\alpha) = \frac{1}{M} \sum_{m=1}^{M} I\left(q^*_{(\alpha/2),m} \leq y^0_m < q^*_{(1-\alpha/2),m}\right). \tag{8.110}$$

Expression (8.110) is then a bootstrap estimate of the probability (8.108). There are then two options. One might simply report $1 - \hat{c}(\alpha)$ as the actual coverage, or one might relate $\hat{c}(\alpha)$ to α through some empirical model (e.g., a quadratic regression of $\hat{c}(\alpha)$ on α might provide a good description of the relation). In the latter case, we can attempt to select an appropriate value α' to use in expression (8.107) in calculating $q(\hat{\theta})_{\alpha'/2}$ and $q(\hat{\theta})_{1-\alpha'/2}$ to provide an actual coverage at level $1 - \alpha$.

8.6.6 Dependence and Other Complications

The usefulness of parametric bootstrap is perhaps the greatest in situations for which we have an estimator $\hat{\theta}$ but it is difficult to derive the variance or

distribution of $\hat{\theta}$. At the same time, we have presented parametric bootstrap methods for sets of independent random variables $Y_1, \ldots, Y_n$. This does seem somewhat incongruous, since it is situations in which we fail to have independence among response variables that most often leads to the inability to make use of distributional results for the purposes of inference. As pointed out by Davison and Hinkley (1997) the theoretical underpinnings of using bootstrap methods with models that contain complex dependence structures (e.g., spatial models) are both unresolved and an area of intensive research. This still remains true today, although any number of advances have been made since the late 1990s, largely in the area of nonparametric bootstrap. Nevertheless, the use of simulation from fitted parametric models seems to hold great potential for a large number of problems.

Consider problems which might be amenable to asymptotic inference. Underlying the development of theoretical properties of bootstrap estimators (either parametric or nonparametric) are two levels of asymptotics, and we will now index $\hat{\theta}$ by the available sample size as $\hat{\theta}_n$ to illustrate this.

At one level is the convergence of the distribution of $h(\hat{\theta}_n, \theta^*_m)$ computed from bootstrap estimates $\{\theta^*_m : m = 1, \ldots, M\}$ to the distribution of $h(\theta, \hat{\theta}_n)$. Here, we must recall that θ^*_m is a function of the bootstrap sample, $\boldsymbol{y}^*_m \equiv (y^*_1, \ldots, y^*_n)^T_m$, so that suppressed in this notation is the fact that each $\boldsymbol{y}^*_m$ is of dimension n and each θ^*_m is based on $\boldsymbol{y}^*_m$. The fact that bootstrap samples $\boldsymbol{y}^*_m$ have been independently generated certainly helps in demonstrating this convergence as M increases. A difficulty is if the distribution of $h(\theta, \hat{\theta}_n)$ depends heavily on the value of θ, since we are using $\hat{\theta}_n$ as the "true" value of θ for bootstrap simulations. The ideal situation is if $h(\theta, \hat{\theta}_n)$ is a *pivotal* quantity; recall this means that the distribution of $h(\theta, \hat{\theta}_n)$ is independent of θ. Unfortunately, we can often only demonstrate this in a definite manner

for fairly simple models, in which case we may have alternative procedures than bootstrap for computing inferential quantities. We may, however, always examine the dependence of the distribution of $h(\theta, \hat{\theta}_n)$ on θ in the following way. Let $Q_M(h|\theta^{(k)})$ denote the estimated q^{th} quantile of $h(\theta, \hat{\theta}_n)$ based on a bootstrap simulation of size M with data generated from the model with parameter value $\theta^{(k)}$. That is, from Section 8.6.1,

$$Q_M(h|\theta^{(k)}) = h(\theta^{(k)}, \theta^*)_{[(M+1)q]}.$$

If values $Q_M(h|\theta^{(k)})$ are computed for a range of values, $\theta^{(k)} \in \{\hat{\theta}_n \pm k\delta : k = 1, \ldots, K\}$ for some δ, then a simple plot of $Q_M(h|\theta^{(k)})$ against $\theta^{(k)}$ may demonstrate the degree of dependence of $h(\theta, \hat{\theta}_n)$ on θ, or at least the relative dependence for several possible choices of $h(\cdot)$.

The second level of convergence needed (at least in an asymptotic setting) arises because, in order for inference about θ based on $\hat{\theta}_n$ to have discernible properties, it is necessary that the distribution of $h(\theta, \hat{\theta}_n)$ allows some description of its probabilistic behavior as n grows large. This issue involves the proverbial "not loosing sight of the forest for the trees" as it applies to bootstrap methods, even outside of an asymptotic context. Consider, for example, estimating μ from a model that gives $Y_1, \ldots, Y_n \sim iid\, N(\mu, \sigma^2)$. My estimator of choice will be $\hat{\mu}_n = 0.1Y_{[2]} + 0.9Y_{[n-3]}$, where $Y_{[k]}$ denotes the k^{th} largest value of the set $\{Y_i : i = 1, \ldots, n\}$. Note, among other things, that here $E\{\hat{\mu}_n\} = \mu$ so $\hat{\mu}_n$ is an unbiased estimator. By taking $h(\mu, \hat{\mu}_n) = \hat{\mu}_n - \mu$ I will be perfectly capable of estimating the distribution of $h(\mu, \hat{\mu}_n)$ through a parametric bootstrap, forming bootstrap intervals, and so forth. This clearly does not, however, offer any justification for my choice of $\hat{\mu}_n$ in the first place.

The combined issues of what are typically called *simulation error* and *statistical error* present a number of perplexing problems, not the least of which

is *where* our concern about uncertainty should be focused. We present one hypothetical example to illustrate this.

Example 8.23

An important environmental characteristic of riverine ecosystems is the distribution of "sediment types" over the bottom of the river. Sediment types are often placed into fairly broad categories such as sand, silt, clay, and gravel. These categories have to do with a combination of particle size and organic matter content of the substrate. Sediment type is one of the factors that determine the abundances of many types of aquatic invertebrates and plants and, consequently, things that depend on them such as fish. Sediment type is also related to various characteristics of water quality, such as clarity and the availability of dissolved oxygen to aquatic life. Ecologists are interested in these relations both from the viewpoint of scientific understanding and to improve prediction of productivity. For example, models have been developed that relate the abundance of mayfly larvae to sediment types; mayflies are one of the "rabbits" of the aquatic world – everything likes to eat them. Now, sampling for sediment type is relatively easy compared to sampling for number of mayflies, and a spatial model that allows prediction of sediment types over a stretch of river might then also allow prediction of mayfly abundance. Observations are available for sediment samples on a portion of the Upper Mississippi River called "Pool 13". The data consist of a number of samples, each categorized into one of k sediment types, for the cells of a grid placed over the river.

A number of spatial models might be possible (are, in fact, possible) for use in this problem. Perhaps the simplest of these is a conditionally spec-

ified multinomial model. Let $\boldsymbol{s}_i \equiv (u_i, v_i)$ where u_i denotes horizontal index and v_i vertical index of a grid. The location variable $\boldsymbol{s}_i$ thus denotes a particular grid cell over the region of interest. Define random variables $\boldsymbol{Y}(\boldsymbol{s}_i) \equiv (Y_1(\boldsymbol{s}_i), Y_2(\boldsymbol{s}_i), \ldots, Y_{k-1}(\boldsymbol{s}_i))^T$ to denote the number of samples categorized as sediment types $1, \ldots, k$ for grid cell $\boldsymbol{s}_i$. We have defined only $k-1$ response variables for each location because, given these variables and the number of samples taken within this location, n_i, the number of samples placed into sediment type k is fixed. The exact structure used to build spatial dependence into this model is not vital for this example, but can be developed from the results of Kaiser and Cressie (2000). What is important is that it is not clear how to compute inferential quantities such as intervals for parameters.

One possible approach toward estimation of intervals for parameters and/or predictions is that of the parametric bootstrap. Given a fitted model, it is not difficult to simulate realizations from the conditional probability mass functions for each location (from the conditionally specified model) using a Gibbs sampler. Each simulated data set could then be used for estimation or, in a "leave-one-out" approach, the formation of prediction intervals. A difficulty, however, is that for categories (sediment types) with low frequency of occurrence, many of the locations $\boldsymbol{s}_i;\ i = 1, \ldots, n$ may not have any simulated values in those categories. That is, for a subset of locations, one or more of the elements of $\boldsymbol{Y}(\boldsymbol{s}_i)$ are zero. This makes estimation for these simulated realizations of the model difficult or impossible. It is not clear how this problem should be dealt with. One approach, which has been used in any number of similar situations, is to condition the simulation estimates of uncertainty on estimable data sets by simply discarding any simulated data sets that do not allow estimation. This must result in an underestimation of uncertainty, but uncertainty about *what*; the values of parameters, the model itself, or the

error of simulation? Put another way, should this phenomenon (the simulation of unestimable data sets from a perfectly acceptable model) impact error of simulation, statistical error in estimation, or error in model selection?

Chapter 9

Model Assessment

In statistical modeling, once one has formulated a model and produced estimates and inferential quantities, the question remains of whether the model is adequate for its intended purpose. This may well involve issues other than whether the model seems to describe the available data in a satisfactory manner, depending on the objectives of the analysis conducted (see Chapter 7.1). Nevertheless, the two conerstones of data-driven model assessment are examination of how well the fitted model describes the observed data, and how well the model predicts observations, and these issues will be the focus of our presentation in this chapter. Even here, however, there are questions regarding which components of a model should be the focus of assessment that depend on the objectives of analysis. Does interest center on a description of the systematic model component? On the modeled distribution more completely (e.g., quantiles of the distribution of responses)? On the ability of the model to predict unobserved random variables (within the extent of the available data)? Or, perhaps the degree and form of departures from a theoretical relation among variables is of central concern. Certainly, there is much overlap in

the manner one might approach these questions, but there may well be unique issues involved as well which simply indicates that model assessment is not a "one size fits all" activity.

Our goal in this chapter is to organize some of the main procedures used in assessment of statisical models, not to present a catalog of all (or even most) of the types of plots, tests, and assessment criteria that have been developed. Many useful procedures are fairly model specific, having been developed for certain types of models that become popular in application. In addition, the number of diagnostics and other assessment methods developed for linear models far exceeds the number and sophistication of methods developed for most other types of models, and has resulted in any number of book-length treatments of the subject (e.g., Belsley, Kuh, and Welsch, 1980; Cook and Weisberg, 1982). Such detailed procedures should be sought out and utilized when appropriate in a particular problem of data analysis. But what might one think of when faced with a model that has been formulated for a specific problem rather than drawn from a standard list of existing model types? In keeping with the theme of these notes, what are ways one might approach model assessment? Three major approaches to model assessment are the use of residuals, cross-validation, and simulation-based assessment.

9.1 Analysis of Residuals

Every student who has completed a basic course in regression is aware of the usefulness of residuals in assessing linear regression models; indeed, it has been assumed in previous chapters that readers were familiar with basic residual plots. Intuitively, residuals are a direct gauge of how far we have "missed" the target in a signal plus noise formulation of a model. Here, basic residuals

have the form of $\{y_i - \hat{y}_i : i = 1, \ldots, n\}$, where the combination of all of the influences (i.e., signals) that produce an observed value y_i is contasted with the (estimated) signals incorporated in a model that produce the fitted or predicted value $\hat{y}_i$. If all of the primary signals that are important in producing y_i have been (nearly) correctly modeled, then these residuals should reflect primarily measurement error. But, as we will see, there are other types of residuals as well, that may be useful in assessing aspects of a proposed model other than how well it reflects signal as modeled through expected values.

9.1.1 A General Notational Framework

Throughout this section we will rely on a general notation framework built around the concept of a random field. Let $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ denote a set of random variables connected with observable quantities, with $\boldsymbol{s}_i$ a non-random "location variable". Several possibilities for the location variables $\boldsymbol{s}_i$ are:

1. Independent random variables.
   Here, we would natrually take $\boldsymbol{s}_i = i$ and simplify notation by referring to $Y(\boldsymbol{s}_i)$ as just Y_i.

2. Groups of random variables.
   Here, we might define $\boldsymbol{s}_i = (k, j)$ where k indexes group and j indexes observation within group; $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$.

3. Geographic random variables.
   Here, we might take $\boldsymbol{s}_i = (u_i, v_i)$, where u_i denotes latitude and v_i longitude, or u_i denotes horizontal coordinate on a grid and v_i denotes vertical coordinate on a grid.

4. Time series of random variables.

   Here we might take $\boldsymbol{s}_i = t$ where t is time, if each $Y(\boldsymbol{s}_i)$ occurs at a unique time, or $\boldsymbol{s}_i = (t, j)$ where t is time and j is observation number at time t; $t = 1, \ldots, T$ and $j = 1, \ldots, n_t$.

We assume that each $Y(\boldsymbol{s}_i)$ is modeled through a parametric distribution having a density (or mass function) f_i, depending on parameter $\psi(\boldsymbol{s}_i)$ through the data model,

$$f_i(y(\boldsymbol{s}_i)|\psi(\boldsymbol{s}_i)); \quad y(\boldsymbol{s}_i) \in \Omega_i. \tag{9.1}$$

Here, the densities f_i are indexed by i to allow for the possibility of differing covariates $\boldsymbol{x}_i$ or auxilliary information (e.g., binomial sample size).

We will assume that the parameters $\{\psi(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ represent "minimal" parameters in the sense that any other parameters used in writing the densities f_i; $i = 1, \ldots, n$ are functions of the $\psi(\boldsymbol{s}_i)$, and also that we may write,

$$\psi(\boldsymbol{s}_i) = (\psi_f(\boldsymbol{s}_i),\, \psi_r(\boldsymbol{s}_i)), \tag{9.2}$$

where $\psi_f(\boldsymbol{s}_i)$ represents parameters that are fixed in the data model and $\psi_r(\boldsymbol{s}_i)$ denotes parameters that are random in the data model. We take $\psi_r(\boldsymbol{s}_i)$ to have a distribution with parameterized density $g_i(\psi_r(\boldsymbol{s}_i)|\lambda)$, where this density may result from marginalization over any additional levels of random terms in the model. For examle, if $\psi_r(\boldsymbol{s}_i)$ is modeled directly in terms of a distribution $g_{1,i}(\psi_r(\boldsymbol{s}_i)|\lambda_1(\boldsymbol{s}_i))$ with $\lambda_1(\boldsymbol{s}_i)$ having a distribution with density $g_2(\lambda_1(\boldsymbol{s}_i)|\lambda)$, then,

$$g_i(\psi_r(\boldsymbol{s}_i)|\lambda) = \int g_{1,i}(\psi_r(\boldsymbol{s}_i)|\lambda_1(\boldsymbol{s}_i)) g_2(\lambda_1(\boldsymbol{s}_i)|\lambda)\, d\lambda_1(\boldsymbol{s}_i). \tag{9.3}$$

Finally, we then take the marginal density of $Y(\boldsymbol{s}_i)$ to be given by

$$h_i(y(\boldsymbol{s}_i)|\psi_f(\boldsymbol{s}_i),\, \lambda) = \int f_i(y(\boldsymbol{s}_i)|\psi_f(\boldsymbol{s}_i),\, \psi_r(\boldsymbol{s}_i))\; g(\psi_r(\boldsymbol{s}_i)|\lambda)\; d\psi_r(\boldsymbol{s}_i). \tag{9.4}$$

This notation is sufficent to cover most of the models we have discussed. In particular, we have not considered any model that contains both fixed and random parameter components in the second (mixing) or higher levels.

Example 9.1

Consider a typical linear regression model with independent response variables, written as

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \epsilon_i; \quad i = 1, \ldots, n.$$

This model fits into our general notation by defining $\boldsymbol{s}_i \equiv i$ and $\psi_f(\boldsymbol{s}_i) \equiv (\beta, \sigma^2)$ and dropping remaining elements of the structure; there is no $\psi_r(\boldsymbol{s}_i)$ or density g.

Example 9.2

We have written a standard generalized linear model as in expressions (7.19) through (7.21) in Section 7.3.2, namely with responses independent and,

$$\begin{aligned} f(y_i|\theta_i, \phi) &= \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right], \\ \mu_i &= b'(\theta_i) \\ \eta_i &= \boldsymbol{x}_i^T \boldsymbol{\beta} \\ g(\mu_i) &= \eta_i \end{aligned}$$

which fits into our general notation with $\boldsymbol{s}_i \equiv i$ and $\psi_f(\boldsymbol{s}_i) \equiv (\boldsymbol{\beta}, \phi)$. Note here that all intermediate parameters can be written in terms of these fixed values as

$$\eta_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^T \boldsymbol{\beta}; \quad \mu_i(\boldsymbol{\beta}) = g^{-1}(\eta_i(\boldsymbol{\beta})); \quad \theta_i(\boldsymbol{\beta}) = b'^{-1}(\mu_i(\boldsymbol{\beta})).$$

Example 9.3

A beta-binomial mixture model was presented in expressions (7.34) and (7.35) for a set of independent random variables as,

$$f_i(y_i|\theta_i) \propto \theta_i^{y_i}\,(1-\theta_i)^{n_i-y_i},$$

$$g(\theta_i|\alpha,\,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1},$$

$$h_i(y_i|\alpha,\beta) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}.$$

This fits into our general notation with $\boldsymbol{s}_i \equiv i$, $\psi_r(\boldsymbol{s}_i) = \theta_i$ and $\lambda \equiv (\alpha, \beta)$.

Example 9.4

A model euqivalent to one presented in expression (7.24) for data with random cluster or group effects is,

$$Y_{i,j} = \boldsymbol{x}_i^T\,\boldsymbol{\beta} + \delta_j I(i \in \mathcal{C}_j) + \sigma\epsilon_{i,j}; \quad \delta_j \sim iidN(0,\tau^2); \quad \epsilon_{i,j} \sim iidN(0,1),$$

where j indexes group and i indexes observation within group. To put this into our general notation, define $\boldsymbol{s}_i \equiv (i,j)$, $\psi_f(\boldsymbol{s}_i) \equiv (\boldsymbol{\beta}, \sigma^2)$, $\psi_r(\boldsymbol{s}_i) \equiv \delta_j$, and $\lambda \equiv \tau^2$.

Corresponding to a model written as in expressions (9.1) through (9.4), we assume parameter estimates are available for the components of $\psi_f(\boldsymbol{s}_i)$ and λ and, where applicable, predictors are available for the components of $\psi_r(\boldsymbol{s}_i)$. Also, we assume that these estimates and predictors lead to estimated expected values $\hat{\mu}(\boldsymbol{s}_i) \equiv \hat{E}\{Y(\boldsymbol{s}_i)\}$ for any $\boldsymbol{s}_i$ in our set of observed locations and predicted values $\hat{p}_Y(\boldsymbol{s}_0)$ for any $\boldsymbol{s}_0$ not in our set of observed locations. If appropriate for a given model, we also assume that estimators of $E\{\psi_r(\boldsymbol{s}_i)\}$ are

available which will be considered in this section as predictors of the random data model parameters as $\hat{p}_\psi(\boldsymbol{s}_i) \equiv \hat{E}\{\psi_r(\boldsymbol{s}_i)\}$.

9.1.2 Types of Residuals

As eluded to in the introductory comments to this chapter, there are any number of quantities we might label as "residuals" in particular models. It would seem we may place the majority of such quantities into the broad categories discussed in this subsection.

Raw and Absolute Residuals

The most basic form of residuals are what we can call *raw residuals*, defined as,

$$r(\boldsymbol{s}_i) = y(\boldsymbol{s}_i) - \hat{\mu}(\boldsymbol{s}_i) \tag{9.5}$$

Raw residuals can be useful in their own right in simple models (e.g., simple linear regression) in which they reflect the same behaviors as more sophisticated residual quantities, and in extremely complex models where we have not yet developed the ability to make use of more refined values. In addition, raw residuals are basic building block for many other residual quantities as they clearly embodied what we intuitively think of as a "residual". Absolute residuals $a(\boldsymbol{s}_i) = |r(\boldsymbol{s}_i)|$ are often useful in detecting patterns of unequal variances and Carroll and Ruppert (1988, p.30) call absolute residuals "the basic building blocks in the analysis of heteroscedasticity" in regression. Any number of transformations of raw and absolute residuals are also useful in certain situations. We defer a discussion of such transformations until the section of this chapter that deals with residual plots since such transformations do not seem to represent truly different types of residuals than the basic underlying

unadjusted quantities.

Studentized Residuals

The use of raw residuals would seem to be well suited for examination of many additive error models, since they represent our "estimates" of the noise component in a model conceptualized as signal plus noise. But in most additive error models, raw residuals do not posess constant variance (even if the model error terms ϵ_i do). It is typically desirable then to use *studentized* residuals, which should have (at least approximately) constant variance.

In general, consider an additive error model of the form

$$Y(\boldsymbol{s}_i) = \mu(\boldsymbol{s}_i) + \sigma(\boldsymbol{s}_i)\,\epsilon(\boldsymbol{s}_i),$$

where $\epsilon(\boldsymbol{s}_i) \sim iidF$, $E\{\epsilon(\boldsymbol{s}_i)\} = 0$ and $var\{\epsilon(\boldsymbol{s}_i)\} = 1$ for $i = 1, \ldots, n$. Consider, for the time being, that the $\sigma(\boldsymbol{s}_i)$ are known, but that the $\mu(\boldsymbol{s}_i)$ are to be estimated. This model, along with the definition of raw residuals in (9.5), indicates that the random form of residuals is,

$$\begin{aligned} R(\boldsymbol{s}_i) &= Y(\boldsymbol{s}_i) - \hat{\mu}(\boldsymbol{s}_i) \\ &= \mu(\boldsymbol{s}_i) - \hat{\mu}(\boldsymbol{s}_i) + \sigma(\boldsymbol{s}_i)\,\epsilon(\boldsymbol{s}_i). \end{aligned}$$

Then,

$$var\{R(\boldsymbol{s}_i)\} = var\{\hat{\mu}(\boldsymbol{s}_i)\} + \sigma^2(\boldsymbol{s}_i) - 2\sigma(\boldsymbol{s}_i)cov\{\hat{\mu}(\boldsymbol{s}_i),\, \epsilon(\boldsymbol{s}_i)\},$$

and we can define studentized residuals as, for $i = 1, \ldots, n$,

$$b(\boldsymbol{s}_i) = \frac{r(\boldsymbol{s}_i)}{[var\{\hat{\mu}(\boldsymbol{s}_i)\} + \sigma^2(\boldsymbol{s}_i) - 2\sigma(\boldsymbol{s}_i)cov\{\hat{\mu}(\boldsymbol{s}_i),\, \epsilon(\boldsymbol{s}_i)\}]^{1/2}}. \tag{9.6}$$

In (9.6) we usually have means $\mu(\boldsymbol{s}_i)$ modeled in terms of a $p-$dimensional parameter $\boldsymbol{\beta}$ with $p < n$, and the first term in the denominator becomes a

function of the variance of $\hat{\boldsymbol{\beta}}$. Of course, it is not the case that the data model variances $\sigma^2(\boldsymbol{s}_i)$ will be known, and the typical approach is to use plug-in estimates of $\sigma^2(\boldsymbol{s}_i)$ in (9.6), ignoring any possible covariance with the estimator of $\mu(\boldsymbol{s}_i)$. That is, common practice is to worry about the covariance of $\mu(\hat{\boldsymbol{s}}_i)$ with $\epsilon(\boldsymbol{s}_i)$, but not covariance between $\mu(\hat{\boldsymbol{s}}_i)$ and estimates of $\sigma^2(\boldsymbol{s}_i)$. Carroll and Ruppert (1988, pp. 33-34) give a limited treatment of the effect of this common practice in terms of a nonlinear model with heteroscedastic errors that we discussed in Chapter 7 as additive error models with known variance parameters.

Example 9.5

If ordinary least squares is used to estimate $\boldsymbol{\beta}$ in the linear regression model of Example 9.1 we have, from $var\boldsymbol{\beta} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\mu}_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, that

$$\begin{aligned} var\{\hat{\mu}_i(\boldsymbol{\beta})\} &= \sigma^2\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \\ &= \sigma^2 h_{i,i}, \end{aligned} \tag{9.7}$$

where $h_{i,i}$ is the i^{th} diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Now,

$$\begin{aligned} cov\{\hat{\mu}_i(\boldsymbol{\beta}),\, \epsilon_i\} &= E\{\hat{\mu}_i(\boldsymbol{\beta})\epsilon_i\} - 0 \\ &= E\left\{\epsilon_i \sum_{j=1}^{n} y_j h_{i,j}\right\} \\ &= \sum_{j=1}^{n} h_{i,j} E\{y_j\epsilon_i\} \\ &= \sum_{j=1}^{n} h_{i,j} E\{(\mu_j + \sigma\epsilon_j)\epsilon_i\} = \sigma\, h_{i,i}. \end{aligned}$$

Substituting into the denominator of (9.6) and replacing σ^2 with the usual

moment estimator $\hat{\sigma}^2$ gives,

$$b_i = \frac{r_i}{[\hat{\sigma}^2(1-h_{i,i})]^{1/2}}, \tag{9.8}$$

the usual studentized residual for linear regression with constant variance.

Example 9.6

Consider a nonlinear regression model with constant variance,

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma\,\epsilon_i,$$

where $\epsilon_i \sim iidF$, $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$. With either generalized least squares or, under the additional assumption that F is $N(0,1)$, maximum likelihood estimation of $\boldsymbol{\beta}$, inference is based on asymptotic results, as discussed in Chapter 8. Hence, derivation of exact forms for the component quantities of (9.6) is difficult. One development of the usual studentized residual follows. For a linear model (i.e., $\mu_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^T\boldsymbol{\beta}$) with constant variance it is easy to show that, in matrix notation,

$$[\boldsymbol{Y} - \mu(\hat{\boldsymbol{\beta}})] = [I - \boldsymbol{H}][\boldsymbol{Y} - \mu(\boldsymbol{\beta}^*)], \tag{9.9}$$

where $\boldsymbol{\beta}^*$ is the true value of $\boldsymbol{\beta}$, and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the usual hat matrix. Recall that this gives studentized residuals in the form of expression (9.8). Now, in a nonlinear model with constant variance we can develop two approximations. First, by expanding the expectation function $\mu_i(\boldsymbol{\beta})$ about the true value $\boldsymbol{\beta}^*$, we have that for any $\boldsymbol{\beta}$ in a small neighborhood of $\boldsymbol{\beta}^*$,

$$\mu_i(\boldsymbol{\beta}) \approx \mu_i(\boldsymbol{\beta}^*) + \sum_{k=1}^{p} \frac{\partial}{\partial\beta_k}\mu_i(\boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\beta_k - \beta_k^*),$$

or, in matrix notation,

$$\mu(\boldsymbol{\beta}) \approx \mu(\boldsymbol{\beta}^*) + V(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \tag{9.10}$$

Note that in (9.10) the matrix of derivatives V is evaluated at the true value $\boldsymbol{\beta}^*$. Now, the minimization problem being solved by a generalized least squares estimation procedure (or maximum likelihood under normality) is,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \{y_i - \mu_i(\boldsymbol{\beta})\}^2,$$

which, after substitution of (9.10), becomes

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\{y_i - \mu_i(\boldsymbol{\beta}^*)\} - \sum_{k=1}^{p} \frac{\partial}{\partial \beta_k} \mu_i(\boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\beta_k - \beta_k^*) \right]^2,$$

or, in matrix notation,

$$\min_{\boldsymbol{\beta}} [\{\boldsymbol{y} - \mu(\boldsymbol{\beta}^*)\} - V(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)]^T [\{\boldsymbol{y} - \mu(\boldsymbol{\beta}^*)\} - V(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)],$$

which has the ordinary least squares solution,

$$\tilde{\boldsymbol{\delta}} = (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = [V^T(\boldsymbol{\beta}^*)\, V(\boldsymbol{\beta}^*)]^{-1}\, V^T(\boldsymbol{\beta}^*)\{\boldsymbol{y} - \mu(\boldsymbol{\beta}^*)\}. \tag{9.11}$$

Now, we can't actually compute $\tilde{\delta}$ or $\tilde{\boldsymbol{\beta}}$. But, asymptotoic results (see e.g., Seber and Wild Chapter 12.2.3) give that, for large enough n,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \approx (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

so that we can make use of (9.11) with $\hat{\boldsymbol{\beta}}$ in place of $\tilde{\boldsymbol{\beta}}$.

Now, consider the vector of residuals,

$$\begin{aligned}
r_i &= \boldsymbol{Y} - \mu(\hat{\boldsymbol{\beta}}) \\
&\approx \boldsymbol{Y} - \mu(\boldsymbol{\beta}^*) + V(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&\approx \boldsymbol{Y} - \mu(\boldsymbol{\beta}^*) + V(\boldsymbol{\beta}^*)[V^T(\boldsymbol{\beta}^*)\, V(\boldsymbol{\beta}^*)]^{-1}\, V^T(\boldsymbol{\beta}^*)\, [y\boldsymbol{Y} - \mu(\boldsymbol{\beta}^*)] \\
&= \left[I - V(\boldsymbol{\beta}^*)(V^T(\boldsymbol{\beta}^*)\, V(\boldsymbol{\beta}^*))^{-1} V^T(\boldsymbol{\beta}^*)\right]\, [\boldsymbol{Y} - \mu(\boldsymbol{\beta}^*)] \\
&= [I - \boldsymbol{H}^{(N)}(\boldsymbol{\beta}^*)][\boldsymbol{Y} - \mu(\boldsymbol{\beta}^*)]. \qquad (9.12)
\end{aligned}$$

The second line of (9.12) follows from substitution of (9.10) evaluated at $\mu(\hat{\boldsymbol{\beta}})$, while the third line results from further use of (9.11) with $\hat{\boldsymbol{\beta}}$ in place of $\tilde{\boldsymbol{\beta}}$ as just discussed. The final line of (9.12) is analogous to the linear model result (9.9) with the hat matrix $\boldsymbol{H}$ replaced by a matrix of the same form but with $V(\boldsymbol{\beta}^*)$ in place of $\boldsymbol{X}$ and denoted as $\boldsymbol{H}^{(N)}(\boldsymbol{\beta}^*)$. That is,

$$\boldsymbol{H}^{(N)}(\boldsymbol{\beta}^*) = V(\boldsymbol{\beta}^*)\,[V^T(\boldsymbol{\beta}^*)V(\boldsymbol{\beta}^*)]^{-1}V^T(\boldsymbol{\beta}^*),$$

where $V(\boldsymbol{\beta}^*)$ is $n \times p$ with i, k^{th} element,

$$\left.\frac{\partial}{\partial\beta_k}\mu_i(\boldsymbol{\beta})\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

With the parallel of expressions (9.12) and (9.9) in hand, we appeal to analogy with linear model results and *define* studentized residuals to be

$$b_i = \frac{r_i}{[\hat{\sigma}^2\,\{1 - h_{i,i}^{(N)}(\hat{\boldsymbol{\beta}})\}]^{1/2}}. \tag{9.13}$$

Notice that in (9.13) we have both replaced σ^2 with an estimator, and have also replaced $\boldsymbol{\beta}^*$ in the nonlinear "hat" matrix $\boldsymbol{H}^{(N)}(\boldsymbol{\beta}^*)$ with its generalized least squares estimator $\hat{\boldsymbol{\beta}}$.

Example 9.7

Now consider the general case of a nonlinear model with nonconstant variance,

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma\, g(\mu_i(\boldsymbol{\beta}), z_i, \theta)\,\epsilon_i,$$

where, as usual, $\epsilon_i \sim iidF$, $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$ but where θ is considered known (or chosen as part of model selection). The usual strategy to develop studentized residuals in this case is to note that this model could also be written as

$$\frac{Y_i}{g(\mu_i(\boldsymbol{\beta}), z_i, \theta)} = \frac{\mu_i(\boldsymbol{\beta})}{g(\mu_i(\boldsymbol{\beta}), z_i, \theta)} + \sigma\,\epsilon_i,$$

which is in the form of a constant variance nonlinear model with modified response $Y_i/g(\mu_i(\boldsymbol{\beta}), z_i, \theta)$ and modified expectation function $\mu_i(\boldsymbol{\beta})/g(\mu_i(\boldsymbol{\beta}), z_i, \theta)$. As indicated by Carroll and Ruppert (1988, p. 33) the standard approach is to ignore all effects of estimation of $g(\mu_i(\boldsymbol{\beta}), z_i, \theta)$ and define studentized residuals in the form of (19.3) as,

$$\tilde{b}_i = \frac{\tilde{r}_i}{[\hat{\sigma}^2 \{1 - \tilde{h}_{i,i}^{(N)}(\hat{\boldsymbol{\beta}})\}]^{1/2}}, \tag{9.14}$$

where

$$\tilde{r}_i = \frac{y_i - \mu_i(\hat{\boldsymbol{\beta}})}{g(\mu_i(\hat{\boldsymbol{\beta}}), z_i, \theta)},$$

and $\tilde{h}_{i,i}^{(N)}(\hat{\boldsymbol{\beta}})$ is the i^{th} diagonal element of the $n \times n$ matrix

$$\tilde{\boldsymbol{H}}^{(N)}(\hat{\boldsymbol{\beta}}) = \tilde{V}(\hat{\boldsymbol{\beta}})[\tilde{V}^T(\hat{\boldsymbol{\beta}})\tilde{V}(\hat{\boldsymbol{\beta}})]^{-1}\tilde{V}^T(\hat{\boldsymbol{\beta}}),$$

where $\tilde{V}(\hat{\boldsymbol{\beta}})$ is $n \times p$ with i, k^{th} element,

$$\frac{1}{g(\mu_i(\hat{\boldsymbol{\beta}}), z_i, \theta)} \left[\frac{\partial}{\partial \beta_k} \mu_i(\boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \right].$$

Deviance Residuals

Deviance residuals are closely connected with exponential families and, in particular, exponential dispersion families. They also represent a somewhat different approach to the conceptual question of what we mean by "residual" than does the fundamental notion of a raw (or studentized) residual. As we have seen, raw residuals are developed first and foremost by considering the deviation of individual response values from their (estimated) expected values. In contrast, deviance residuals are most easily developed as the contributions of individual response values to a quantity that reflects overall model fit. To develop this idea, assume we have a set of independent response variables $Y_1, \ldots, Y_n$ with density or mass functions of exponential dispersion family

form,

$$f_i(y_i|\theta_i, \phi) = \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right].$$

Notice that we are allowing the distributions of the Y_i to vary only through the scalar natural parameter θ_i. Recall from Section 6.1.3 this implies that $\mu_i \equiv E(Y_i) = b'(\theta_i)$, or $\theta_i = b'^{-1}(\mu_i)$ so that we can write the natural parameters as functions of the expected values, $\theta(\mu_i)$. Now, in almost all models formulated on the basis of exponential dispersion family distributions, we further model μ_i as a function of other parameters and, perhaps, covariates. Generalized linear models are the obvious example, but the concept of deviance being developed depends on exponential dispersion family properties not the specific form of generalized linear models. In any case, fitting a model will produce a set of estimated expectations $\{\hat{\mu}_i : i = 1, \ldots, n\}$ and hence also a set of estimated natural parameters $\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}) \equiv \{\theta(\hat{\mu}_i) : i = 1, \ldots, n\}$.

We have also seen that, given maximum likelihood estimates, full and reduced models with nested parameter spaces can be compared through likelihood ratio tests. Consider, then, comparison of a fitted model considered as a reduced model to a "saturated" model (or a "maximal model"); these labels are meant to evoke the notions of "fullest model possible" or "model with the highest likelihood value possible". Such a model will result from estimating μ_i as the observed value y_i, for $i = 1, \ldots, n$, which leads to another set of estimated natural parameters $\boldsymbol{\theta}(\boldsymbol{y}) \equiv \{\theta(y_i) : i = 1, \ldots, n\}$. Note that such a saturated or maximal model is not a viable or useful model in practice since it contains as many parameters as observations, and this is assuming that the dispersion parameter ϕ is known. With known ϕ, a likelihood ratio comparison of fitted and saturated models would then become,

$$D^* \equiv -2\{L(\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}), \phi) - L(\boldsymbol{\theta}(\boldsymbol{y}), \phi)\}, \tag{9.15}$$

where

$$L(\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}), \phi) = \sum_{i=1}^{n} \left[\phi\{y_i\theta(\hat{\mu}_i) - b(\theta(\hat{\mu}_i))\} + c(y_i, \phi)\right],$$

and

$$L(\boldsymbol{\theta}(\boldsymbol{y}), \phi) = \sum_{i=1}^{n} \left[\phi\{y_i\theta(y_i) - b(\theta(y_i))\} + c(y_i, \phi)\right].$$

Expression (9.15) defines the *scaled deviance* for a model based on independent exponential dispersion family random variables. Notice that it may also be written as

$$D^* = -2\phi \sum_{i=1}^{n} \left[y_i\{\theta(\hat{\mu}_i) - \theta(y_i)\} - b(\theta(\hat{\mu}_i)) + b(\theta(y_i)\right], \tag{9.16}$$

because, with ϕ considered known, the terms $c(y_i, \phi)$ cancel in the difference. The parameter ϕ may be seen in (9.16) to constitute a scaling factor, and the *unscaled deviance* is defined as $D \equiv D^*/\phi$, or

$$D = -2 \sum_{i=1}^{n} \left[y_i\{\theta(\hat{\mu}_i) - \theta(y_i)\} - b(\theta(\hat{\mu}_i)) + b(\theta(y_i)\right]. \tag{9.17}$$

Scaled and unscaled deviances are measures of the departure of a fitted model from a saturated model, which intutitively captures the concept of goodness of fit. Given the assumed distributional form and with a known value of ϕ (more on this in the sequel), nothing could fit the data better than the saturated model, which has the greatest log likelihood value possible (this explains my use of the phrase maximal model). If we would not prefer this maximal model to our reduced fitted model, then the fitted model provides an adequate representation of the observed data. In this sense, expression (9.16) constitutes a likelihood ratio goodness of fit test, and D could be compared to a χ^2 distribution with $n - p$ degrees of freedom. Unfortunately, when ϕ is not known this no longer is the case and, in fact, it is not even possible to estimate

ϕ under the saturated or maximal model.

Example 9.8

It is instructive to examine the forms taken by deviance for some of the more common exponential dispersion family distributions.

1. Poisson

   Here, $\phi \equiv 1$ and $\theta_i = \log(\mu_i)$ so that, for a fitted model with estimated expected values $\{\hat{\mu}_i : i = 1, \ldots, n\}$, $\theta(\hat{\mu}_i) = log(\hat{\mu}_i)$ and $\theta(y_i) = \log(y_i)$. Also, $b(\theta_i) = \exp(\theta_i)$ so that $D^* = D$, and

   $$\begin{aligned} D &= -2\sum_{i=1}^{n}[y_i\{\log(\hat{\mu}_i) - \log(y_i)\} - \hat{\mu}_i + y_i] \\ &= 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]. \end{aligned}$$

2. Binomial

   For a set of independent binomial random variables taken to represent proportions rather than counts, let $E(Y_i) = p_i$. In exponential dispersion family form, $\phi \equiv 1$, $\theta_i = \log\{p_i/(1-p_i)\}$, and $b(\theta_i) = \log\{1 + \exp(\theta_i)\}$. Then, $\theta(\hat{\mu}_i) = \log\{\hat{\mu}_i/(1-\hat{\mu}_i)\}$ and $\theta(y_i) = \log\{y_i/(1-y_i)\}$. It is convention to simply absorb the known biniomial sample sizes n_i into all formulas as weights, and then again $D^* = D$ where,

   $$\begin{aligned} D &= -2\sum_{i=1}^{n} n_i \left[y_i \left\{\log\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) - \log\left(\frac{y_i}{1-y_i}\right)\right\} - \log(1-\hat{\mu}_i) + \log(1-y_i)\right] \\ &= 2\sum_{i=1}^{n} n_i \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right]. \end{aligned}$$

3. Normal

   For normal distributions with the usual mean (μ) and variance (σ^2) parameterization, $\theta_i = \mu_i$, $\phi = 1/\sigma^2$, and $b(\theta_i) = (1/2)\theta_i^2$. Then scaled

deviance is,

$$\begin{aligned} D^* &= \frac{-2}{\sigma^2}\sum_{i=1}^{n}[y_i\{\hat{\mu}_i - y_i\} - (1/2)\hat{\mu}_i^2 + (1/2)y_i^2] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2. \end{aligned}$$

Notice that for this situation unscaled deviance is $D = \sigma^2 D^*$, the usual residual sum of squares.

4. Gamma

Since there are several versions of the "usual" parameterization of a gamma density function we need to be careful of our initial formulation for a problem involving independent gamma random variables. For an individual random variable Y, let the probability density function be

$$f(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, y^{\alpha-1}(1-y)^{\beta-1}; \quad y > 0.$$

With this form, $\mu \equiv E(Y) = \alpha/\beta$, and by writing $\nu = \alpha$ we can arrive at an exponential dispersion family representation of the density with $\theta = -1/\mu$, $\phi = 1/\nu$, and $b(\theta) = -\log(-\theta)$. Let $\{Y_i : i = 1,\ldots,n\}$ be a set of independent random variables have such densities with parameters $\{\theta_i : i = 1,\ldots,n\}$ and common ϕ. Then $\theta(\hat{\mu}_i) = -1/\hat{\mu}_i$ and $\theta(y_i) = -1/y_i$, and the scaled deviance becomes,

$$\begin{aligned} D^* &= -2\phi\sum_{i=1}^{n}\left[y_i\left\{\frac{-1}{\hat{\mu}_i} - \frac{-1}{y_i}\right\} + \log(-\hat{\mu}_i) - \log(-y_i)\right] \\ &= 2\phi\sum_{i=1}^{n}\left[\frac{y_i}{\hat{\mu}_i} - 1 + \log(\hat{\mu}_i) - \log(y_i)\right] \\ &= 2\phi\sum_{i=1}^{n}\left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\left(\frac{y_i}{\hat{\mu}_i}\right)\right]. \end{aligned}$$

For the Poisson and binomial portions of Example 9.8 we could use deviance as a likelihood ratio goodness of fit test statistic, but not for the normal and gamma. In these latter cases, deviance is generally calculated using an estimated value $\hat{\phi}$ from the fitted model.

Each observation y_i contributes one term to (9.16) or (9.17), and it is these terms that are used to define basic deviance residuals. Let,

$$d_i^* = -2\hat{\phi}\left[y_i\{\theta(\hat{\mu}_i) - \theta(y_i)\} - b(\theta(\hat{\mu}_i)) + b(\theta(y_i)\right],$$

and define deviance residuals as, for $i = 1, \ldots, n$,

$$d_i \equiv \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i^*}. \tag{9.18}$$

While, as mentioned, the ideas of deviance and deviance residuals have their genesis in results for exponential dispersion families, their use is most closely connected with generalized linear models. In this case, it is common to standardize deviance residuals as,

$$d_i' = \frac{d_i}{(1 - h_{i,i}^{(G)})^{1/2}}, \tag{9.19}$$

where $h_{i,i}^{(G)}$ is the i^{th} diagonal element of the matrix

$$\boldsymbol{H}^{(G)} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2},$$

in which $\boldsymbol{X}$ is the $n \times p$ matrix of covariate values of the linear predictor $\boldsymbol{\eta} \equiv (\eta_1, \ldots, \eta_n)^T$ and $\boldsymbol{W}$ is the $n \times n$ diagonal matrix with elements given in Section 8.3.6 as,

$$W_i \equiv \left\{\left(\frac{d\eta_i}{d\mu_i}\right)^2 V(\mu_i)\right\}^{-1}.$$

The standardization of (9.19) is justified by results on the first two moments of "generalized residuals", a topic we will cover briefly later in this section, and

conditions that make higher derivatives of the log likelihood negligible. As a result, $E(d_i) \approx 0$ and $var(d_i) \approx 1 - h_{i,i}^{(G)}$. A readable presentation of this is contained in Davison and Snell (1991), who also point out that (9.19) is a special case of a result that applies more generally to exponential dispersion families. In particular, consider a model formulated in the same manner as a generalized linear model expect that, rather than using a link to a linear prediction as $g(\mu_i) = \boldsymbol{x}_i^T \beta$, we simply take the expectations to be a given function of parameters and covariates as

$$\mu_i = \eta(\boldsymbol{x}_i, \beta),$$

denoted as η_i for brevity.

Then, define the matrix $\boldsymbol{W}$ as the diagonal matrix with i^{th} element

$$w_i = E\left[-\frac{\partial^2 \log\{f(y_i|\theta_i, \phi)\}}{\partial \eta_i^2}\right],$$

and the $n \times p$ matrix $\boldsymbol{Q}$ to have i, k^{th} element,

$$q_{i,k} = \frac{\partial \eta_i}{\partial \beta_k}.$$

Then, take

$$\tilde{\boldsymbol{H}}^{(G)} = \boldsymbol{W}^{1/2}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{W}\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{W}^{1/2},$$

and standardized deviance residuals are then given by (9.19) with $\tilde{\boldsymbol{H}}^{(G)}$ in place of $\boldsymbol{H}^{(G)}$. Note that, in the case of a generalized linear model, w_i has the same form as given following expression (9.19), and $\boldsymbol{Q} = \boldsymbol{X}$.

Many of the general ideas described in the past few pages are also applied in formulating a number of other residual quantities that seem to be less commonly used in practice. These include what are called *score residuals*, *likelihood residuals*, and *Anscombe residuals*. See, for example, Lindsey (1996, p. 168) for mention of the first two of these, and Davison and Snell (1991) for

the latter. Somewhat more common are *Pearson residuals* particularly in the case of models with discrete response varaibles such as Poisson or binomial.

Generalized Residuals

In what remains an important paper on the construction and analysis of residuals, Cox and Snell (1968) gave a general definition of a residual, as follows. The situation considered is one in which the location variable of our general notation from Section 9.1.1 is taken to be $\boldsymbol{s}_i = i$ and the random variables $\{Y_i : i = 1, \ldots, n\}$ are assumed independent. Consider a model in which

$$Y_i = g_i(\boldsymbol{\theta},\, \epsilon_i); \quad i = 1, \ldots, n, \tag{9.20}$$

where $\boldsymbol{\theta}$ is a p-dimensional parameter, the ϵ_i are independent and identically distributed random variables, and the model is indexed by i (i.e., $g_i(\cdot)$) to allow for covariates or other known factors. Expression (9.20) is sufficiently general to cover nearly all models formulated for independent random variables, a key aspect being the each Y_i depends on only one ϵ_i.

Example 9.9

1. Additive error models, either linear or nonlinear, are easily put into the form of (9.20) by taking $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \theta)$ and, using our previous notation for such models,

$$g_i(\boldsymbol{\theta}, \epsilon_i) = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, z_i, \theta)\epsilon_i,$$

where $\epsilon_i \sim iidF$ with $E(\epsilon_i) = 0$.

2. Multiplicative error models are also easy to put in the form of (9.20). Cox and Snell consider as an example a model

$$Y_i = \beta_1 \exp\{\beta_2(x_i - \bar{x})\}\, \epsilon_i,$$

where $\epsilon_i \sim iid$ exponential, with $E(\epsilon_i) = 1$.

3. Let $Y_1, \ldots, Y_n$ be independently and identically distibuted following an extreme value distribution with density written as, for $-\infty < \theta_1 < \infty$ and $\theta_2 > 0$,

$$f(y_i|\theta_1, \theta_2) = \frac{1}{\theta_2} \exp\left[\frac{y_i - \theta_1}{\theta_2} - \exp\left\{\frac{y_i - \theta_1}{\theta_2}\right\}\right]; \quad -\infty < y_i < \infty.$$

The distribution function that corresponds to this density is,

$$F_Y(y_i|\theta_1, \theta_2) = 1 - \exp\left[-\exp\left\{\frac{y_i - \theta_1}{\theta_2}\right\}\right],$$

and the inverse distribution function becomes,

$$F_Y^{-1}(a|\theta_1, \theta_2) = \theta_1 + \theta_2 \log\left[\log\left\{\frac{1}{1-a}\right\}\right]; \quad 0 < a < 1.$$

To formulate a model in the form of (9.20) for this situation, let

$$Y_i = g_i(\theta_1, \theta_2, U_i) = F_Y^{-1}(u_i|\theta_1, \theta_2),$$

where $U_i \sim iid\, U(0, 1)$.

4. The device of item 3 immediately above may be used directly for any set of continuous random variables for which a model leads to a parameterized distribution function. This is true even if the inverse distribution function is not available in closed form, such as for gamma, beta, or even normal distributions; for a normal distribution, however, we would probably use the easier additive model formulation. It is also not necessary

to have identical distributions. Specifically, if $Y_1, \ldots, Y_n$ are independent random variables with modeled probability density functions $f_i(y_i|\boldsymbol{\theta})$, then the model may be written as,

$$Y_i = g_i(\boldsymbol{\theta}, U_i) = F_i^{-1}(U_i|\boldsymbol{\theta}); \quad i = 1, \ldots, n,$$

where $U_i \sim iid\, U(0,1)$ and

$$F_i(y_i|\boldsymbol{\theta}) = \int_{-\infty}^{y_i} f_i(t|\boldsymbol{\theta})\, dt.$$

5. Although the prescription of items 3 and 4 applies only to random variables with continuous distributions, a similar device may be used for discrete random variables. Let $Y_1, \ldots, Y_n$ be independent with a common set of possible values Ω and suppose the elements of this set have been ordered as $\Omega \equiv \{y_{[1]}, y_{[2]}, \ldots, y_{[m]}\}$ so that $y_{[k-1]} < y_{[k]}$ for $k = 2, \ldots, m$. If we desire to assign probability mass functions $f_i(y_i|\boldsymbol{\theta})$ to these random variables, let

$$F_i(y_{[k]}|\boldsymbol{\theta}) = \sum_{j=1}^{k} f_i(y_{[j]}|\boldsymbol{\theta}),$$

and take, for $i = 1, \ldots, n$,

$$Y_i = \min\left\{y_{[k]} : F_i(y_{[k]}|\boldsymbol{\theta}) > U_i\right\},$$

where $U_i \sim iid\, U(0,1)$. Define $y_{[0]}$ to be any value such that $F_i(y_{[0]}|\boldsymbol{\theta}) = 0$. Then, for $k = 1, \ldots, m$,

$$\begin{aligned} Pr(Y_i = y_{[k]}) &= Pr\{F_i(y_{[k-1]}\boldsymbol{\theta}) < U_i < F_i(y_{[k]}|\boldsymbol{\theta})\} \\ &= F_i(y_{[k]}|\boldsymbol{\theta}) - F_i(y_{[k-1]}\boldsymbol{\theta}) \\ &= f_i(y_{[k]}|\boldsymbol{\theta}) \end{aligned}$$

as desired.

Now, writing a model in the form of expression (9.20) does not necessarily define a residual quantity. Cox and Snell (1968) do so as follows. Suppose that equation (9.20) has a unique solution for ϵ_i, say

$$\epsilon_i = h_i(Y_i, \boldsymbol{\theta}).$$

If this is the case, and if $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimate of $\boldsymbol{\theta}$, then define generalized residuals as, for $i = 1, \ldots, n$,

$$r_i = h_i(y_i, \hat{\boldsymbol{\theta}}). \tag{9.21}$$

The random version of (9.21) may be written as $R_i = h_i(Y_i, \hat{\boldsymbol{\theta}})$. The remainder of the treatment by Cox and Snell (1968) involved deriving approximations to the means, variances, and covariances of the random version of (9.21) by expanding $R_i - \epsilon_i$ as a Taylor series in terms of the components of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, which is where they use the condition that $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimate of $\boldsymbol{\theta}$. These approximations provide a means for modifying the residuals R_i so that the modified residuals have expectations and variances approximately equal to those of the ϵ_i. Adjustment of the residuals in (9.21) may be important, but the form of resultant equations is quite model specific. Here, we simply indicate that if the number of observations is large relative to the number of parameters, then the residual quantities (9.21) should have behavior similar to the ϵ_i in the model (9.20). This is because it is the observations $y_1, \ldots, y_n$ that are used to estimate $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}$ so that the R_i are not independent, and generally will not have expectations or variances equal to those of the ϵ_i. But as the number of observations increases relative to the dimension of $\boldsymbol{\theta}$ these effects diminish.

We will use these ideas to develop the notion of *generalized residual* that will be used in these notes, drawing on cases 4 and 5 of Example 9.9. Corre-

sponding to case 4, given continuous independent random variables $Y_1, \ldots, Y_n$ with model $Y_i = F_i^{-1}(U_i|\boldsymbol{\theta})$ for $U_i \sim iid\, U(0,1)$, define the residual quantities,

$$r_i \equiv \int_{-\infty}^{y_i} F_i(t|\hat{\boldsymbol{\theta}})\, dt; \quad i = 1, \ldots, n. \tag{9.22}$$

If the model is representative of the data, the $\{r_i : i = 1, \ldots, n\}$ should behave in a manner similar to a sample from the uniform distribution on $(0, 1)$. The probability integral transform would hold if the parameter $\boldsymbol{\theta}$ were used in (9.22) rather than an estimate $\hat{\boldsymbol{\theta}}$. While this result does not hold, we expect that the residuals of (9.22) should provide diagnostic quantities useful to detect gross discrepancies between the model and observed responses.

To define similar residuals corresponding to discrete random variables as in case 5 of Example 9.9 requires an extension of the definition of Cox and Snell (1968). In the develoment of (9.21) it was assumed that the model (9.20) allows a *unique* solution for the ϵ_i. Here, we define random residuals even for fixed values of observations $\{y_i : i = 1, \ldots, n\}$. Using the same notation as in case 5 of Example 9.9, let $Y_1, \ldots, Y_n$ be independent random variables with a common set of possible values, ordered as $\Omega \equiv \{y_{[1]}, y_{[2]}, \ldots, y_{[m]}\}$. Take the ordered value of observation i to be the k^{th} ordered value, that is, $y_i = y_{[k]}$. Define the (random) generalized residual r_i' to be the realized value of a random variable with distribution uniform on the interval $(F_i(y_{[q-1]}),\, F_i(y_{[q]}))$, that is,

$$r_i' \equiv u_i; \quad \text{where } U_i \sim iid\; U\left(F_i(y_{[q-1]}),\, F_i(y_{[q]}\right). \tag{9.23}$$

Similar to the residuals of expression (9.22), these residuals should behave in the manner of a sample of *iid* uniform variables on the interval (0, 1). A set of residuals $\{r_i' : i = 1, \ldots, n\}$ will not, however, be unique for a given set of observations $\{y_i : i = 1, \ldots, n\}$.

Other Types of Residuals

The classification of residuals into categories in this subsection is, of course, not exhaustive of the various quantities proposed for use as residuals. While many such quantities are rather model specific, there are some that are more general in nature. A few of these are listed here.

1. Pearson Residuals.

   Pearson residuals are motivated by considering individual contributions to a Pearson Chi-squared goodness of fit test for discrete random variables. Consider, for example, a set of independent Possion random variables $\{Y_i : i = 1, \ldots, n\}$ with expected values $\{\mu_i : i = 1, \ldots, n\}$ and variances $\{V(\mu_i) : i = 1, \ldots, n\}$, where $V(\mu_i) = \mu_i$. The Pearson χ^2 goodness of fit test statistic is,

   $$\chi^2_* = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

   The square root of the individual contributions are known as Pearson residuals,

   $$r_{p,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \tag{9.24}$$

   The residual quantities in (9.24) are more general that just Poisson, or even just discrete random variables, and are still called Pearson residuals even when applied to random variables with continuous distributons.

2. Score Residuals.

   For independent random variables, the score function for any given parameter component consists of a sum of quantities, the random versions of which each have expected value 0. With respect to a given component of a generic parameter, θ_j, say, the score residuals are then based on

standardized contributions to the total score function,

$$r_{s,i} = \left(\frac{1}{I_{i,i}(\hat{\theta})}\right) \frac{\partial}{\partial \theta_j} \log\{f_i(y_i|\theta)\}, \tag{9.25}$$

where $I_{i,i}(\hat{\theta})$ is the i^{th} diagonal element of the estimated information matrix.

3. Anscombe Residuals.

   Ansombe (1961) proposed a residual quantity of the essential form of the Pearson residual (9.24) but with y_i replaced by a transformed value that more nearly produces normal distributions than (9.24). The appropriate transformation depends on the model chosen, although it can be expressed in a general form for standard generalized linear models (e.g., McCullagh and Nelder, 1989).

9.1.3 Plotting Residuals

Any number of diagnostic plots can be constructed using the residuals quantities discussed in Section 9.1.2, with the intent of detecting departures from the model structure assumed in an analysis. We mention here some of the more common of these, along with the types of modeling inadequacies they are intended to detect. In general, residual plots involve plotting residuals (or some transformation of residuals) on the vertical axis or ordinate against corresponding quantities of some type on the horizontal axis or abcissa. Typically, any type of pattern exhibited by the points on such a plot indicates some type of model inadequacy. Gleaning useful information from residual plots then involves determination of whether a percieved pattern is due to more than random variability in a finite set of observed data, and the type of model inadequacy suggested by a pattern. The first of these is often a matter of judgment,

a process that is often made easier by comparison of plots for several models; the strength or degree of departures from model structure is typically more easily assessed on a relative scale than an absolute scale. The second requires undertanding of the expected behavior of residuals under a correctly specified model, as well as the types of behaviors that would be produced by departures from the assumed model structure.

Plotting Against Fitted Values

Perhaps the most common type of residual plot results from plotting residuals against fitted values from a model. Fitted values are generally taken as estimated expected values of the random variables associated with observed responses, that is, the estimated systematic model component. We have already seen a number of examples of this type of residual plot, at least for linear regression models using residuals as in expression (9.8). Figure 6.2 presents such a plot for the simulated data relating Microcystin concentrations to nitrogen. Figure 7.4 gives this residual plot for airplane flight times. Figures 7.12, 7.14, 7.15, and 7.19 show these plots for various models employed in the example of tree volume modeled as a function of diameter and height.

Example 9.10

In Example 7.1 a nonlinear regression model with additive constant variance errors was fitted to the reaction times of an enzyme as a function of substrate concentration for preparations treated with Puromycin and also for untreated preparations. The model was

$$Y_i = \frac{\beta_1 \, x_i}{\beta_2 + x_i} + \sigma \epsilon_i,$$

where x_i denoted substrate concentration and we took $\epsilon_i \sim iidF$ for some distribution F with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$ for $i = 1, \ldots, n$. This model was fit to each group (treated and untreated) separately. Figure 9.1 presents the studentized residuals of expression (9.13) for both groups. This residual plot does not reveal any serious problems with the model, although it is less than "picture perfect" in terms of what we might hope to see. Given that this model was formulated on the basis of a theoretical equation for enzyme reaction times (the Michaelis-Menten equation) and variability is anticipated (and appears in the data) to be small, we would be justified in assessing this residual plot with a fairly high level of scrutiny (relative to, say, a residual plot for a purely observational study with many potential sources of variability). Does the residual plot of Figure 9.1 exhibit some degree of increasing variance as a function in increasing mean? To help in this assessment, we might plot the cube root of squared studentized residuals against the fitted values (e.g., Carroll and Ruppert, 1988, p. 30). In this type of residual plot, nonconstant variance is exhibited by a wedge-shaped pattern of residuals. A plot of the cube root squared studentized residuals for these data is presented in Figure 9.2. There does not appear to be a increasing wedge or fan of residuals in the plot of Figure 9.2, suggesting that there is little evidence of nonconstant variance for this model. Looking closely at the residual plot of Figure 9.1 we can see a suggestion of a "U-shaped" pattern in residuals from both treated and untreated groups. This would indicate that the fitted expectation function from the Michaelis-Menten equation fails to bend correctly to fit data values at across the entire range of substrate concentrations. A close examination of the fitted curves, presented in Figure 9.3 verifies that this seems to be the case for at least the treated preparations. In fact, there appear to be values at a substrate concentration of just over 0./2 ppm for which the expectation

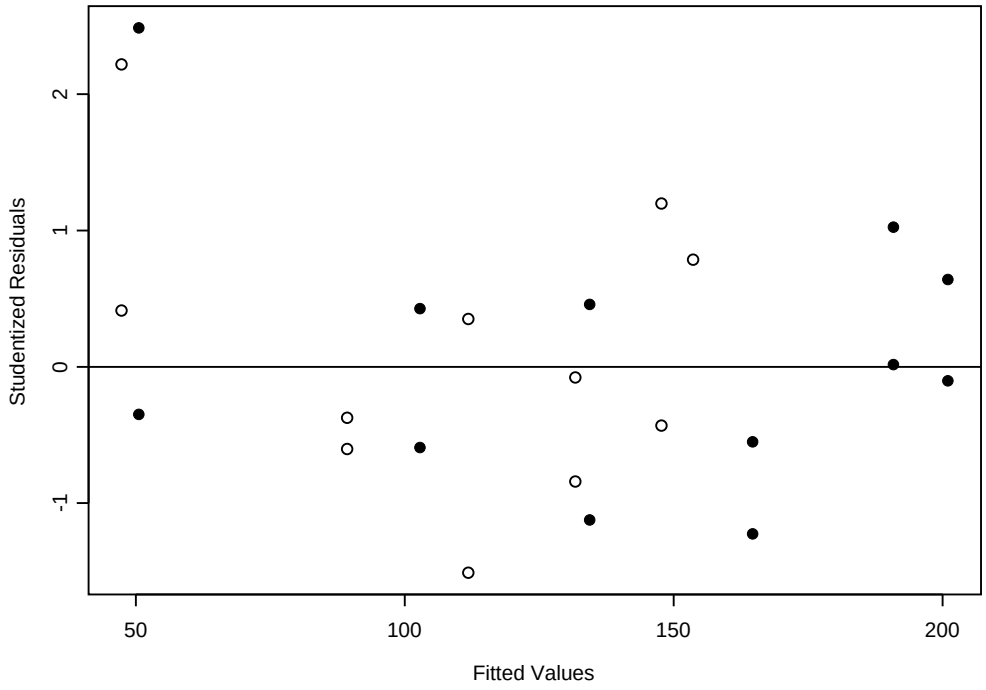

Figure 9.1: Studentized residuals from fitting a nonlinear regression based on the Michaelis-Menten equation to the enzyme reaction times of Example 7.1. Open circles are the untreated preparations while solid circles are the treated preparations.

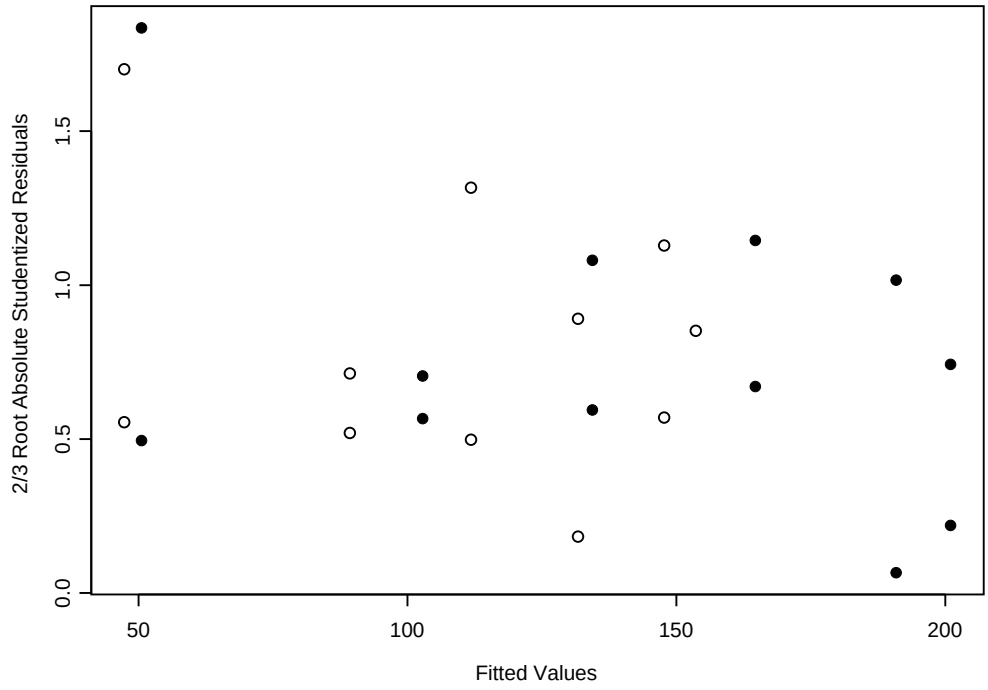

Figure 9.2: Cube root squared studentized residuals from fitting a nonlinear regression based on the Michaelis-Menten equation to the enzyme reaction times of Example 7.1. Open circles are the untreated preparations while solid circles are the treated preparations.

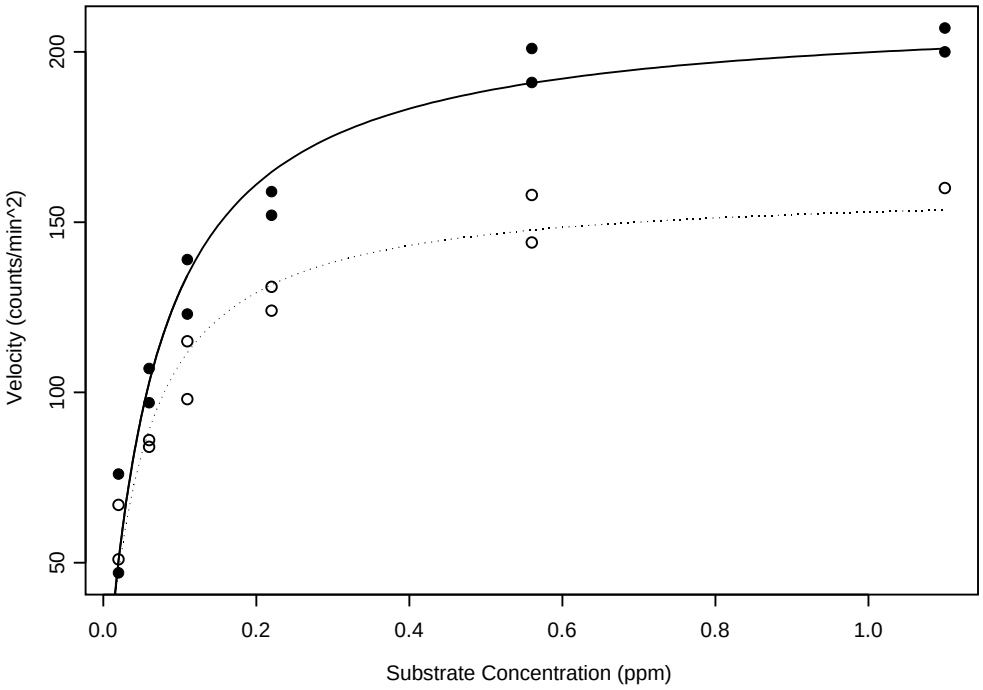

Figure 9.3: Fitted regressions based on the Michaelis-Menten equation to the enzyme reaction times of Example 7.1. Open circles are the untreated preparations while solid circles are the treated preparations.

function "misses" for both treated and untreated groups. The cause of this phenomenon is unknown to us as is, indeed, the degree of scientific import for what it suggests. It may well be the case, however, that there exists some evidence that the theoretical Michaelis-Menten equation does not adequately describe the enzyme reaction in this experiment.

In general, the use of cube root squared studentized residuals might be justified based on what is known as the *Wilson-Hilferty* transformation to normalize chi-squared variables, but the basic value of such plots seems due to

more practical than theorectical considerations. Cook and Weisberg (1982) suggested plotting squared residuals to help overcome sparse data patterns, particularly when it is not clear that positive and negative residuals have patterns symmetric about zero. Carroll and Ruppert (1988) echo this sentiment, but indicate that squaring residuals can create extreme values if the original residuals are moderately large in absolute value to begin with. They then suggest taking the cube root to aleviate this potential difficulty, but point out that they view the result essentially as a transformation of absolute residuals, which are taken as "the basic building blocks in the analysis of heteroscedasticity" (Carroll and Ruppert, 1998, p. 30). From this standpoint, it would seem to make little difference if one used absolute residuals, the squre root of absolute residuals or, as in Figures 9.1 and 9.2, a 2/3 power of absolute residuals.

Plotting Against Covariates

Plotting Against Time or Space

More on Scaling Residual Plots

9.1.4 Tests With Residuals

9.2 Cross Validation

9.2.1 Fundamental Concepts

9.2.2 Types of Cross Validation

Group Deletion

Leave One Out Deletion

Pseudo-Cross Validation

9.2.3 Discrepancy Measures

Loss Functions and Decision Theory

Global Versus Local Discrepancies

9.3 Assessment Through Simulation

9.3.1 Fundamental Concepts

9.3.2 Discrepancy Measures Revisited

9.3.3 Simulation of Reference Distributions

9.3.4 Sensitivity Analysis

Part III

BAYESIAN ANALYSIS

Chapter 10

Basic Bayesian Paradigms

In this part of the course we consider the topic of Bayesian analysis of statistical models. Note that I have phrased this as Bayesian *analysis* of *models*, as opposed to analysis of Bayesian models. In some ways this is simply a matter of semantics, but there is an issue involved that goes slightly beyond the mere use of words, which we will attempt to elucidate in what follows. The effects of the two viewpoints on Bayesian analysis presented in this chapter do not influence mathematical or statistical techniques in the derivation of distributions or inferential quantities, but they may have an impact on which quantities (i.e., which *distributions*) are deemed appropriate for inference. The issues involved come into play primarily in the case of *hierarchical models*, in which there is more than one level of quantities that play the role of parameters in probability density or mass functions. In this regard, there might be a distinction between thinking in terms of analysis of a Bayesian model as opposed to a Bayesian analysis of a mixture (i.e., random parameter) model.

In this chapter we will attempt to make clear both the fundamental nature of the Bayesian argument, and the potential impact of the distinction eluded

to in the previous paragraph. Since this issue is not even necessarily recognized as an issue by many statisticians, the subsection headings given below are my own device and should not be taken as standard terminology in the world of statistics.

10.1 Strict Bayesian Analysis

The heading of "strict" Bayesian analysis comes from a reading of the history of Bayesian thought. Although I will not give a list of references, I believe this line of reasoning to be faithful to that history, in which there was reference to what was called the "true state of nature". That is, at least one line of reasoning in the development of Bayes methods held that there is, in fact, an absolute truth to the order of the universe. This thought is in direct conflict with the frequent (at least in the past) criticism of Bayesian methods that they depend on totally subjective interpretations of probability (there were, however, other schools of thought in the development of Bayesian methods in which probabilities *were* viewed as essentially subjective values, but we will not discuss these). The fundamental point is that this view of Bayesian analysis is in total agreement with a reductionist view of the world in which, if all forces in operation were known, observable quantities would be completely deterministic. The true state of nature in this school of thought is embodied in a fixed, but unknown parameter value that governs the distribution of observable quantities. Note that this is starting to sound quite a bit like a typical frequentist argument, and that is the point.

There is not necessarily anything in a Bayesian approach to statistical analysis that contradicts the view that, if we knew everything about all physical relations in the world, we would know the values that would be assumed by

observable quantities with certainty. We are, of course, not capable of such exquisite knowledge, and thus the field of statistics has meaning in modeling the things we do not understand through the use of probability.

The essential departure of Bayesian thought from its frequentist counterpart, under this interpretation of Bayesian analysis, is that an epistemic concept of probability is legitimate. That is, given a fixed but unknown parameter θ that represents the true state of nature, it is legitimate to write $Pr(t_1 < \theta < t_2) = c$, with θ, t_1, t_2 and c all being constants, not as a statement about a random quantity (there are no random variables in the above probability statement) but rather as a statement about our *knowledge* that θ lies in the interval (t_1, t_2). Thus, under this interpretation of Bayesian thought, there must be somewhere in a model a fixed quantity or parameter that represents the "true state of nature".

Now, if we admit an epistemic concept of probability, then we are free to represent our current knowledge about θ as a probability distribution. And, given the possibility of modeling observable quantities as connected with random variables (the modeling concept), we can easily formulate the basic structure of a strict Bayesian approach to the analysis of data; note here that everything contained in Part II of the course prior to the discussion of estimation and inference (i.e., everything before Chapter 8) applies to Bayesian analysis as much as it applies to what was presented under the heading of Statistical Modeling. In particular, suppose that we have a situation in which we can formulate random variables $Y_1, \ldots, Y_n$ in connection with observable quantities. Suppose that the joint probability distribution of these variables can be written in terms of a density or mass function $f(y_1, \ldots, y_n|\boldsymbol{\theta})$ depending on an unknown (but fixed) parameter $\boldsymbol{\theta}$ that represents the true state of nature (i.e., the phenomenon or scientific mechanism of interest). Now, even in the

absence of data, we are almost never totally devoid of knowledge about $\boldsymbol{\theta}$. If an observable quantity Y is a count with a finite range, and the mass function assigned to Y is a binomial distribution with parameter θ, then we know that θ must lie between 0 and 1. If a set of observable quantities are connected with *iid* random variables having gamma distributions with common parameters α and β (here, $\boldsymbol{\theta} \equiv (\alpha, \beta)^T$), then we know that α and β must both be positive and that any distribution assigned to our knowledge of their values should give low probability to extremely large values (i.e., both the mean and variance of the gamma distribution should exist).

Given this background, consider a single random variable Y, to which we have assigned a probability distribution through density or mass function $f(y|\boldsymbol{\theta})$ with fixed but unknown parameter $\boldsymbol{\theta}$. Our knowledge about $\boldsymbol{\theta}$ before (i.e., *prior*) to observing a value connected with Y is embodied in a prior distribution $\pi(\boldsymbol{\theta})$. Note that, while $f(y|\boldsymbol{\theta})$ may be interpreted through a hypothetical limiting relative frequency concept of probability, the prior distribution $\pi(\boldsymbol{\theta})$ is an expression of epistemic probability, since $\boldsymbol{\theta}$ is considered a fixed, but unknown, quantity. The mathematics of dealing with $\pi(\boldsymbol{\theta})$ will, however, be identical to what would be the case if $\boldsymbol{\theta}$ were considered a random variable. Suppose that $\boldsymbol{\theta}$ can assume values in a region (i.e., $\boldsymbol{\theta}$ is continuous) with dominating Lebesgue measure. Then, given observation of a quantity connected with the random variable Y as having the particular value y, we may derive the *posterior* distribution of $\boldsymbol{\theta}$ as,

$$p(\boldsymbol{\theta}|y) = \frac{f(y|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int f(y|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}. \tag{10.1}$$

If $\boldsymbol{\theta}$ can assume only values in a discrete set and $\pi(\cdot)$ is dominated by counting measure, then (10.1) would become,

$$p(\boldsymbol{\theta}|y) = \frac{f(y|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} f(y|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}. \tag{10.2}$$

Expressions (10.1) and (10.2) are generally presented as applications of Bayes Theorem, the form of which they certainly represent. But Bayes Theorem is a probability result, which holds for random variables regardless of whether one is using it in the context of a Bayesian analysis or not. The only reason one may not accept the expressions (10.1) or (10.2) as valid is if one will not allow an epistemic concept of probability for the function $\pi(\boldsymbol{\theta})$ to describe our knowledge of the fixed but unknown value of $\boldsymbol{\theta}$. Thus, we are led to the conclusion that the fundamental characteristic of Bayesian analysis does not rely on Bayes Theorem, but on an epistemic concept of probability to describe what we know about the true state of nature $\boldsymbol{\theta}$.

In the above development we call the distribution for the observable quantities $f(y_1, \ldots, y_n|\boldsymbol{\theta})$ the *observation* or *data* model, and the distribution $\pi(\boldsymbol{\theta})$ (interpreted under an epistemic concept of probability) the *prior* distribution. The resultant distribution of our knowledge about $\boldsymbol{\theta}$ conditioned on the observations, namely $p(\boldsymbol{\theta}|y)$, is the *posterior* distribution of $\boldsymbol{\theta}$.

Now, notice that nothing above changes if we replace the data model for a single random variable Y with the joint distribution of a set of variables $Y_1, \ldots, Y_n$ all with the same parameter $\boldsymbol{\theta}$. That is, the data model becomes $f(\boldsymbol{y}|\boldsymbol{\theta})$, the prior remains $\pi(\boldsymbol{\theta})$, and the posterior would be $p(\boldsymbol{\theta}|\boldsymbol{y})$.

In this development we have taken the prior $\pi(\boldsymbol{\theta})$ to be *completely specified* which means that $\pi(\cdot)$ depends on no additional *unknown* parameters. That does not necessarily mean that $\pi(\cdot)$ depends on no controlling parameters, only that, if it does, those parameter values are considered known.

Example 10.1

Suppose that we have a random variable Y such that the data model is given

by $Y \sim Bin(\theta)$. That is, Y has probability mass function

$$f(y|\theta) = \frac{n!}{(n-y)!\,y!}\,\theta^y\,(1-\theta)^{n-y}; \quad y = 0, 1, \ldots, n.$$

Now, we know that $\theta \in (0, 1)$ and, if we wish to express no additional knowledge about what the value of θ might be, we could take the prior to be $\theta \sim U(0,\, 1)$, so that,

$$\pi(\theta) = 1; \qquad 0 < \theta < 1.$$

The posterior that would result from an observation of of the data model as y would then be,

$$\begin{aligned} p(\theta|y) &= \frac{\theta^y\,(1-\theta)^{n-y}}{\int_0^1 \theta^y\,(1-\theta)^{n-y}\,d\theta} \\ &= \frac{\Gamma(n+2)}{\Gamma(1+y)\,\Gamma(1+n-y)}\,\theta^y\,(1-\theta)^{n-y}, \end{aligned}$$

which is the density function of a beta random variable with parameters $1+y$ and $1+n-y$.

Now, suppose we have additional information about θ before observation of y which is represented by a beta distribution with parameters $\alpha = 2$ and $\beta = 2$; this would give expected value $\alpha/(\alpha+\beta) = 0.5$ and variance 0.05. Then, the posterior would be derived in exactly the same way as above, except that it would result in a beta density with parameters $2+y$ and $2+n-y$. In general, for any specific choices of α and β in the prior $\pi(\theta) = Beta(\alpha,\, \beta)$ the posterior will be a beta distribution with parameters $\alpha+y$ and $\beta+n-y$.

Now consider a problem in which we have a data model corresponding to

$$Y_1, \ldots, Y_m \sim \; iid\, Bin(\theta),$$

in which the "binomial sample sizes" $n_1, \ldots, n_m$ are considered fixed, and $\pi(\theta)$ is taken to be beta$(\alpha_0,\, \beta_0)$ with α_0 and β_0 specified (i.e., considered

known). Then a derivation entirely parallel to that given above, except using a joint data model $f(\boldsymbol{y}|\theta) = f(y_1|\theta)f(y_2|\theta)\ldots, f(y_m|\theta)$ results in a posterior $p(\theta|y_1, \ldots, y_m)$ which is a beta (α_p, β_p) density with parameters

$$\begin{aligned} \alpha_p &= \alpha_0 + \sum_{i=1}^{m} y_i, \\ \beta_p &= \beta_0 + \sum_{i=1}^{m} n_i - \sum_{i=1}^{m} y_i. \end{aligned} \tag{10.3}$$

The essential nature of what I am calling here the strict Bayesian approach to data analysis follows from the modeling idea that the scientific mechanism or phenomenon of interest is represented somewhere in a statistical model by a fixed parameter value. The additional component under a Bayesian approach is to assign our knowledge of that parameter a prior distribution under an epistemic concept of probability. This then allows derivation of a posterior distribution that represents our knowledge about the parameter value after having incorporated what can be learned from taking observations in the form of data. Thus, there are really no such things as "Bayesian models", only Bayesian analyses of statistical models. This view of Bayesian analysis extends from simple situations such as that of Example 10.1 to more complex situations in a natural manner.

Example 10.2

Suppose that we have a data model corresponding to

$$Y_1, \ldots, Y_m \sim \ indep\, Bin(\theta_i),$$

in which the binomial sample sizes $n_1, \ldots, n_m$ are considered fixed. The only difference between this data model that that of the second portion of Example 10.1 is that now each response variable $Y_i;\ i = 1, \ldots, m$ is taken to have its

own binomial parameter. If we interpret the values $\theta_1, \ldots, \theta_m$ as representing different "manifestations" of some scientific mechanism or phenomenon, then we might assign these values a mixing distribution (as in Part II of the course),

$$\theta_1, \ldots, \theta_m \sim iid\, Beta(\alpha, \beta),$$

which would result in a beta-binomial mixture model as discussed previously. A Bayesian analysis of this mixture model would then consist of assigning (our knowledge about) the parameter (α, β) a prior distribution $\pi(\alpha, \beta)$ and deriving the posterior distribution $p(\alpha, \beta|\boldsymbol{y})$.

Example 10.2 is an illustration of the general structure for a strict Bayesian analysis of a mixture model. In general we have,

1. Data Model:

   $Y_1, \ldots, Y_n \in \Omega$ have joint density or mass function $f(\boldsymbol{y}|\boldsymbol{\theta}) = f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n)$.

   If these response variables are independent, or conditionally independent given $\theta_1, \ldots, \theta_n$, then,

   $$f(\boldsymbol{y}|\boldsymbol{\theta}) = f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} f_i(y_i|\theta_i).$$

2. Mixing Distribution:

   $\theta_1, \ldots, \theta_n \in \Theta$ have joint density or mass function $g(\boldsymbol{\theta}|\boldsymbol{\lambda}) = g(\theta_1, \ldots, \theta_n|\boldsymbol{\lambda})$.

   If these random variables are independent, then,

   $$g(\boldsymbol{\theta}|\boldsymbol{\lambda}) = g(\theta_1, \ldots, \theta_n|\boldsymbol{\lambda}) = \prod_{i=1}^{n} g_i(\theta_i|\boldsymbol{\lambda}).$$

3. Prior:

   Given $f(\boldsymbol{y}|\boldsymbol{\theta})$ and $g(\boldsymbol{\theta}\ \boldsymbol{\lambda})$, the parameters $\boldsymbol{\lambda} \in \Lambda$ are assigned a prior distribution $\pi(\boldsymbol{\lambda})$.

4. Posterior:

   Using the above formulation, the posterior density or mass function of $\boldsymbol{\lambda}$ may be derived as,

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})}{\int_{\Lambda}\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}}. \tag{10.4}$$

   If the Y_i are conditionally independent and the θ_i are independent, then,

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{\prod_{i=1}^{n} f_i(y_i|\theta_i)\, g_i(\theta_i|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})}{\int_{\Lambda} \left\{ \prod_{i=1}^{n} \int_{\Theta_i} f_i(y_i|\theta_i)\, g_i(\theta_i|\boldsymbol{\lambda})\, d\theta_i \right\} \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}.$$

Now notice that, in the above progression, we could have equivalently combined items 1 and 2 into the overall mixture model

$$h(\boldsymbol{y}|\boldsymbol{\lambda}) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta},$$

or, in the case of conditional independence of $\{Y_i : i = 1, \ldots, n\}$ and independence of $\{\theta_i : i = 1, \ldots, n\}$,

$$h(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} \int_{\Theta_i} f_i(y_i|\theta_i)\, g_i(\theta_i|\boldsymbol{\lambda})\, d\theta_i.$$

The posterior of $\boldsymbol{\lambda}$ could then be expressed, in a manner entirely analogous with expression (10.1), as,

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{h(\boldsymbol{y}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})}{\int_{\Lambda} h(\boldsymbol{y}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}, \tag{10.5}$$

or, in the case of independence,

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{h_i(y_i|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})}{\int_{\Omega_i} h_i(y_i|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, dy_i}.$$

The preceding expressions are a direct expression of a Bayesian analysis of a mixture model. The mixture model is given by $h(\boldsymbol{y}|\boldsymbol{\lambda})$, and we simply assign the parameters of this model, namely $\boldsymbol{\lambda}$, a prior distribution $\pi(\boldsymbol{\lambda})$ and derive the corresponding posterior $p(\boldsymbol{\lambda}|\boldsymbol{y})$.

We are not yet entirely prepared to address the possible issue of the introduction, but notice that, in principle, other distributions are available to us, such as

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\int_{\Lambda} f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}{\int_{\Lambda}\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\lambda})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}}, \\ \\ p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})}{\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\lambda})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\theta}}. \end{aligned} \tag{10.6}$$

The first of these expressions is sometimes called the *marginal* posterior of $\boldsymbol{\theta}$ and the second the *conditional* posterior of $\boldsymbol{\theta}$. This last expression for the conditional posterior $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$ is of particular interest, since it can also be written as,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y}) &= \frac{m(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{\lambda})}{\int_{\Theta} m(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}} \\ &= \frac{m(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{\lambda})}{h(\boldsymbol{y}|\boldsymbol{\lambda})}, \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})}{h(\boldsymbol{y}|\boldsymbol{\lambda})}. \end{aligned} \tag{10.7}$$

where $h(\boldsymbol{y}|\boldsymbol{\lambda})$ is the same as the "marginal" density of $\boldsymbol{Y}$ discussed in marginal maximum likelihood analysis of mixture models. Notice that, to use $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$

directly for inference, we would need to have an estimate of $\boldsymbol{\lambda}$ to use in this density function as $p(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}}, \boldsymbol{y})$.

Finally, we also have the modeled distribution of $\boldsymbol{\theta}$ as $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and, given an estimate of $\boldsymbol{\lambda}$ we could focus attention on this distribution with estimated $\boldsymbol{\lambda}$ as $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$; this, in fact, is what we have done in Part II of the course in presenting plots of, e.g., beta mixing densities, using $\hat{\boldsymbol{\lambda}}$ as given by marginal maximum likelihood estimates. It would certainly be possible, however, to take $\hat{\boldsymbol{\lambda}}$ to be the mode or expectation of the posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$ in a Bayesian approach. A portion of the issue introduced in the introduction to this discussion of Bayesian analysis concerns whether inference about $\boldsymbol{\theta}$ in a Bayesian analysis of a mixture model (or hierarchical model) should be made based on $p(\boldsymbol{\theta}|\boldsymbol{y})$, $p(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}}, \boldsymbol{y})$ or $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$.

10.2 Bayesian Analysis of Unknowns

What I call here the Bayesian analysis of unknowns is not so much a different approach to that of the strict Bayesian analysis of Chapter 10.1 as it is just a slightly different angle on what is being accomplished. The philosophy of this viewpoint is that statistical models contain unknown quantities. Whether some might be considered random variables and others parameters under the strict Bayesian interpretation is not material; everything we do not know is simply an unknown quantity. Probability and, in particular, epistemic probability, is the way that we describe uncertainty. Thus, anything we do not know the value of can be assigned a "prior" distribution to represent our current knowledge and, given a subsequent set of observed data $\{y_1, \ldots, y_n\}$, these can be updated to be "posterior" distributions. This seems to be the view taken in several recent texts on Bayesian data analysis (Carlin and Louis, 2000; Gelman, Carlin, Stern

and Rubin, 1995). In fact, Carlin and Louis (2000, p.17) state that

> In the Bayesian approach, in addition to specifying the model for the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta}$, usually in the form of a probability distribution $f(\boldsymbol{y}|\boldsymbol{\theta})$, we suppose that $\boldsymbol{\theta}$ is a random quantity as well, . . .

It is unclear whether Carlin and Louis are asserting that $\boldsymbol{\theta}$ is a parameter or a random variable (in the sense we have used these words in this class). My own interpretation of their words is that they just don't really care. If one wishes to think of $\boldsymbol{\theta}$ as a parameter, that's fine, or if one wishes to think of $\boldsymbol{\theta}$ as a random variable that's fine too. Either way, $\boldsymbol{\theta}$ is unknown, and so we use probability to describe our uncertainty about its value.

Gelman, Carlin (a different Carlin from Carlin and Louis), Stern and Rubin (1995) are quite circumspect in talking about random variables *per se*, more often using phrases such as "random observables", "sampling models", or simply the unadorned "variable" (reference to random variables does pop up occasionally, e.g., p. 20, but that phrase is nearly absent from the entire text). These authors present (Chapter 1.5 of that text) a clear exposition of the basic notion of probability as a representation of uncertainty for *any* quantity about which we are uncertain. The implication is again that we need not be overly concerned about assigning some mathematical notion of "type" (e.g., random variable or parameter) to quantities in order to legitimately use probability as a measure of uncertainty.

This notion of using probability as a measure of uncertainty without getting all tangled up in whether the object of our uncertainty should be considered a parameter, random variable, or some other type of quantity is seductively simple, perhaps too much so, as I will argue below. It does have the attractive

flavor that all quantities involved in an analysis are put on a equal footing. The distinction is only that some of those quantities will be observable (i.e., the data) while others will not. Simplicity in analytical approach (but not necessarily practice) is then achieved through the assertion that the entire process of statistical inference is just a matter of conditioning our uncertainty (in the form of probability) about unobservable quantities on the values obtained for observable quantities.

This view of Bayesian analysis does not result in any differences with that of the strict Bayesian view in terms of manipulations of distributions, derivation of posteriors and so forth. In simple problems that involve only a data or observational model and a prior, such as those described in Example 10.1, there is only one possible way to condition the distribution of (our knowledge or uncertainty about) the unobservable $\boldsymbol{\theta}$ on that of the data. In these situations, I see no possible conflicts between the strict Bayesian view and that considered in this section.

In a hierarchical situation involving data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and additional distributions $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $\pi(\boldsymbol{\lambda})$, there may be some slightly different implications for inference of the strict Bayesian view and that of this section. In particular, as noted above, Bayesian analysis of unknowns would indicate that the proper distribution on which to base inference is the joint posterior $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y})$, that is, the conditional distribution of unobservables on observables. If one is uninterested in $\boldsymbol{\lambda}$, for example, the corresponding marginal $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior of concern. The possibilities of making use of the conditional posterior $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$ or (especially) the model distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ are not in total concert with Bayesian analysis of unknowns.

In fact, the motivation of specifying a distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ can be considered somewhat differently under this view of Bayesian analysis. In a strict

Bayes view, $g(\boldsymbol{\theta}|\lambda)$ is thought of as a mixing distribution that represents, in the model, a scientific mechanism or phenomenon of interest. The fixed parameter $\boldsymbol{\lambda}$ then embodies the immutable mechanism, while $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ describes the ways that the mechanism is manifested in different situations. Here, we are hesitant to assign such responsibilities to $\boldsymbol{\lambda}$ and $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$. The value of $\boldsymbol{\theta}$ is an unobservable quantity that controls the description of our uncertainty about the observable quantity y (or quantities $y_1, \ldots, y_n$). Our uncertainty about $\boldsymbol{\theta}$ is therefore assigned a distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, which may depend on an additional unknown quantity $\boldsymbol{\lambda}$. Our uncertainty about $\boldsymbol{\lambda}$ is then also assigned a distribution $\pi(\boldsymbol{\lambda})$. It is quite natural,then, to refer to *prior distributions* for both $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. The distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ would be the conditional prior for $\boldsymbol{\theta}$, while the marginal prior for $\boldsymbol{\theta}$ would be,

$$\pi_\theta(\boldsymbol{\theta}) = \int_\Lambda g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\lambda)\, d\boldsymbol{\lambda}.$$

One could then derive a posterior as in expression (10.1) using data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and prior $\pi_\theta(\boldsymbol{\theta})$, resulting in the posterior

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\, \pi_\theta(\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})\, \pi_\theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &= \frac{\int_\Lambda f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}{\int_\Lambda \int_\Theta f(\boldsymbol{y}|\boldsymbol{\lambda})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}} \end{aligned} \tag{10.8}$$

Notice that (10.8) is exactly the same marginal posterior for $\boldsymbol{\theta}$ given in expression (10.6), illustrating the point that all of the mathematics of Chapter 10.1 also apply here.

What I want to contrast (10.8) with, however, is the posterior for $\boldsymbol{\lambda}$ in expression (10.5). Expression (10.5) was developed by combining $f(\boldsymbol{y}|\boldsymbol{\theta})$ and

$g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ into the mixture model $h(\boldsymbol{y}|\boldsymbol{\lambda})$ which was then used with the prior $\pi(\boldsymbol{\lambda})$ to derive the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y})$. Expression (10.8) has been developed by combining $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $\pi(\boldsymbol{\lambda})$ into the prior $\pi_\theta(\boldsymbol{\theta})$ which was then used with the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ to derive the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. That is, to get to (10.5) most directly, $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ was considered part of the model $h(\boldsymbol{y}|\boldsymbol{\lambda}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}$, to which was assigned the prior $\pi(\boldsymbol{\lambda})$. To get to (10.8) most directly, $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ was considered part of the prior $\pi_\theta(\boldsymbol{\theta}) = \int g(\boldsymbol{\theta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$, which was applied to the model $f(\boldsymbol{y}|\boldsymbol{\theta})$.

While these progressions are not inherent parts of the two viewpoints that I have called here *strict Bayesian analysis* and *Bayesian analysis of unknowns* they do indicate a somewhat different slant to the way in which hierarchical models and, in particular, the roles of $p(\boldsymbol{\theta}|\boldsymbol{y})$, $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\lambda})$ and $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, are interpreted.

10.3 Summary of the Viewpoints

To encapsulate what has been presented in the two previous sections, there appear to be several angles from which Bayesian analysis of hierarchical models can be approached. Both depend on the derivation of posterior distributions for unknown quantities in a statistical model. Both collapse to the same viewpoint in the case of simple models with a given data or observational model and a fixed parameter that controls that model. The potential differences arise in consideration of multi-level or hierarchical models. The following points seem relevant.

1. Under either viewpoint, all of the same distributions are available. Distributions of $f(\boldsymbol{y}|\boldsymbol{\theta})$, $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $\pi(\boldsymbol{\lambda})$ constitute the statistical model.

Distributions $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y})$ and the associated marginals $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $p(\boldsymbol{\lambda}|\boldsymbol{y})$, the conditional form $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$, and the model distribution $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$ are available and identical under both viewpoints. For the purposes of inference, these last two would require a plug-in estimator of $\boldsymbol{\lambda}$ as $p(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}}, \boldsymbol{y})$ and $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$.

2. Under what I have called strict Bayesian analysis there is really only one prior distribution, that being for whatever quantities are considered fixed parameters in a model. If we have a data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ for which $\boldsymbol{\theta}$ is fixed, and to which we wish to assign the prior $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ but have uncertainty about an appropriate value for $\boldsymbol{\lambda}$, then we might combine $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $\pi(\boldsymbol{\lambda})$ into the prior $\pi_\theta(\boldsymbol{\theta})$ just as in the progression given for the viewpoint titled Bayesian analysis of unknowns. This seems unlikely, however. If I know enough about the possible values of $\boldsymbol{\theta}$ to assign them (my knowledge of them) a prior $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ that takes a particular form, surely I must also have some idea what the value of $\boldsymbol{\lambda}$ might be. Thus, it would be somewhat unnatural (not necessarily wrong or inappropriate, but perhaps a bit odd), given a strict Bayesian view of the world, to make use of the argument of Carlin and Louis (2000, p.19-20) that hierarchical models result from uncertainty about the appropriate values of the parameter $\boldsymbol{\lambda}$ to use in the "first-stage" prior $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$. A counter-argument to my assertion is possibly offered by a theorem due to de Finetti (1974) which we will discuss under the heading of *exchangeability* in Chapter 12. This theorem may be used to justify the formulation of a prior for $\boldsymbol{\theta}$ in the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ through the use of a mixture such as $\pi_\theta(\boldsymbol{\theta})$ developed just before expression (10.8).

   On the other hand, if we have a data model that consists of the obser-

vation model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and a mixing distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, then we are in one sense assigning a prior only to $\boldsymbol{\lambda}$, and considering $\boldsymbol{\theta}$ to be a random variable in the usual sense of the term, although typically a "latent" random variable that is not connected to observable quantities. In this latter setting, $p(\boldsymbol{\lambda}|\boldsymbol{y})$ is certainly meaningful. A question arises, however, concerning $p(\boldsymbol{\theta}|\boldsymbol{y})$ versus $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$ versus $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ where, for the purpose of inference, a plug-in estimate of $\boldsymbol{\lambda}$ would be needed in the latter two distributions.

3. Under what I have called Bayesian analysis of unknowns there is no distinction made between the status of quantities as random variables or parameters, or the mixing model idea of "random variables that play the role of parameters". Probability is epistemic in nature wherever it is applied; even to observable quantities. If my knowledge about the values of observable quantities, represented through the observation model $f(\boldsymbol{y}|\boldsymbol{\theta})$ happens to agree with what would result from repeated observation of the same situation (i.e., relative frequency) then so much the better; I then just have some empirical justification for my belief. Given this viewpoint, the appropriate quantities for inference would seem to be $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y})$ and the associated marginals $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $p(\boldsymbol{\lambda}|\boldsymbol{y})$. There is probably little role for $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$ and almost certainly little use for $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where these latter two quantities would again be used with a plug-in estimator of $\boldsymbol{\lambda}$.

4. What can be said about relations between $p(\boldsymbol{\theta}|\boldsymbol{y})$, $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$ and $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ in the mathematical sense? Such relations must apply regardless of the

viewpoint toward analysis that is being taken. First, we have that,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\int_{\Lambda} m(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}{\int_{\Lambda}\int_{\Theta} m(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}} \\ &= \frac{\int_{\Lambda} p(\boldsymbol{\theta}|\boldsymbol{\lambda},\boldsymbol{y})\, p(\boldsymbol{\lambda}|\boldsymbol{y})\, h(\boldsymbol{y})\, d\boldsymbol{\lambda}}{h(\boldsymbol{y})} \\ &= \int_{\Lambda} p(\boldsymbol{\theta}|\boldsymbol{\lambda},\boldsymbol{y})\, p(\boldsymbol{\lambda}|\boldsymbol{y})\, d\boldsymbol{\lambda}. \end{aligned}$$

Thus, as it should be, the marginal posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the expected value of the conditional posterior $p(\boldsymbol{\theta}|\boldsymbol{\lambda},\boldsymbol{y})$ taken over the distribution of $\boldsymbol{\lambda}$ given $\boldsymbol{y}$. The variance of $p(\boldsymbol{\theta}|\boldsymbol{y})$ will then be greater than that of $p(\boldsymbol{\theta}|\boldsymbol{\lambda},\boldsymbol{y})$.

Secondly, directly from (10.7) we have that,

$$g(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \frac{h(\boldsymbol{y}|\boldsymbol{\lambda})}{f(\boldsymbol{y}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{\lambda},\boldsymbol{y}).$$

Since $h(\boldsymbol{y}|\boldsymbol{\lambda})$ is more diffuse (variable) than $f(\boldsymbol{y}|\boldsymbol{\theta})$, so $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ will be more diffuse than $p(\boldsymbol{\theta}|\boldsymbol{\lambda},\boldsymbol{y})$. Thus, both $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ represent greater uncertainty about $\boldsymbol{\theta}$ than does $p(\boldsymbol{\theta}|\boldsymbol{\lambda},\boldsymbol{y})$, which is intuitive. The unknown relation, at this time, is between $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ when the latter is evaluated at a value $\hat{\boldsymbol{\lambda}}$, which will be either the expectation or mode of the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y})$.

5. Under what I have called a strict Bayesian viewpoint, inference about $\boldsymbol{\lambda}$ is most naturally based on $p(\boldsymbol{\lambda}|\boldsymbol{y})$. Inference about $\boldsymbol{\theta}$ is most naturally based on either $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$ or possibly $p(\boldsymbol{\theta}|\boldsymbol{y})$. The role, if any, for $p(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}},\boldsymbol{y})$ is not clear. Under what I have called the Bayesian analysis of unknowns, inference for $\boldsymbol{\lambda}$ is most naturally based on $p(\boldsymbol{\lambda}|\boldsymbol{y})$, although this will

typically not be of much interest. Inference for $\boldsymbol{\theta}$ should probably be based on $p(\boldsymbol{\theta}|\boldsymbol{y})$. There is little, if any, role for $p(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}}, \boldsymbol{y})$ and almost certainly no use for $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$. Note here, however, that using $p(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}}, \boldsymbol{y})$ is one formulation of methods that are called *empirical Bayes* (e.g., Carlin and Louis, 2000, Chapter 3).

Chapter 11

Sequential Bayes

Perhaps one of the strongest arguments in favor of a Bayesian approach to analysis is provided by the similarity between the view of science as a progressive "building up" of knowledge, and a sequential use of Bayesian analysis. To see this clearly, adopt for the moment the strict Bayesian viewpoint that our concern is learning about the value of a fixed "state of nature" represented by a parameter $\boldsymbol{\theta}$. That is, suppose we have a problem in which data are "generated" by a observational model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and we have specified a prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$. Now, $\pi(\boldsymbol{\theta})$ represents our knowledge about $\boldsymbol{\theta}$ before any observations are available. Given observations $\boldsymbol{y}$, we update that knowledge in the form of the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. Now, suppose that we (or someone else) are able to repeat the study that led to the observations $\boldsymbol{y}$, or at least conduct a similar study in which $\boldsymbol{\theta}$ is a controlling parameter in an observational model and has the same scientific meaning it did in the first study. Then it would be natural to take $p(\boldsymbol{\theta}|\boldsymbol{y})$ from the first study as representing our current knowledge about $\boldsymbol{\theta}$ (i.e., the posterior from the first study becomes the prior for the second study). In a sequence of k such studies, then, we could conduct an overall

analysis using data models $f_1(\boldsymbol{y}_1), \ldots, f_k(\boldsymbol{y}_k)$ and a "cascade" of prior and posterior distributions, in the following way:

Study	Prior	Data	Posterior
1	$\pi_1(\boldsymbol{\theta})$	$f_1(\boldsymbol{y}_1\|\boldsymbol{\theta})$	$p_1(\boldsymbol{\theta}\|\boldsymbol{y}_1)$
2	$\pi_2(\boldsymbol{\theta}) = p_1(\boldsymbol{\theta}\|\boldsymbol{y}_1)$	$f_2(\boldsymbol{y}_2\|\boldsymbol{\theta})$	$p_2(\boldsymbol{\theta}\|\boldsymbol{y}_2)$
3	$\pi_3(\boldsymbol{\theta}) = p_2(\boldsymbol{\theta}\|\boldsymbol{y}_2)$	$f_3(\boldsymbol{y}_3\|\boldsymbol{\theta})$	$p_3(\boldsymbol{\theta}\|\boldsymbol{y}_3)$
.	.	.	.
.	.	.	.
.	.	.	.
k	$\pi_k(\boldsymbol{\theta}) = p_{k-1}(\boldsymbol{\theta}\|\boldsymbol{y}_{k-1})$	$f_k(\boldsymbol{y}_k\|\boldsymbol{\theta})$	$p_k(\boldsymbol{\theta}\|\boldsymbol{y}_k)$

At each step in this progression we would have the basic Bayesian update of expression (10.1), namely,

$$p_j(\boldsymbol{\theta}|\boldsymbol{y}_j) = \frac{f_j(\boldsymbol{y}_j|\boldsymbol{\theta})\,\pi_j(\boldsymbol{\theta})}{\displaystyle\int_\Theta f_j(\boldsymbol{y}_j|\boldsymbol{\theta})\pi_j(\boldsymbol{\theta})\,d\boldsymbol{\theta}}; \quad j = 1, \ldots, k.$$

Implementation of this progression is made much easier if all of the data models have the same form, $f(\cdot) = f_1(\cdot) = f_2(\cdot) = \ldots = f_k(\cdot)$, and if the common data model $f(\cdot)$ and prior $\pi_1(\cdot)$ have a property called "conjugacy" which means that the posterior $p_1(\boldsymbol{\theta}|\boldsymbol{y})$ has the same form as the prior $\pi_1(\boldsymbol{\theta})$ (more on this in Chapter 12.2). This then implies that all of the priors $\pi_1(\cdot), \pi_2(\cdot), \ldots, \pi_k(\cdot)$ all have the same form as well.

Example 11.1

The sex ratio at birth of various species is of interest to ecologists in understanding evolutionary pressures and the way in which organisms have adapted in response to those pressures. A study was conducted over several years on the South American Guanaco (one of the South American "camels", the others

being the Llama, Alpaca, and Vicuna) to determine the proportion of males at birth. It is fairly well established that mortality is higher during the first year of life for males than for females, because males tend to venture farther from the protection of the adult "herd" than do females, and are thus more often attacked by predators (your own interpretation of this being because males are "adventurous", "developing skills to protect the herd when adults", or "just stupid" may well depend on your own sex). At any rate, this has led ecologists and geneticists to believe that the ratio of males to females at birth should be greater than 0.50. The basic idea has nothing to do with the "good of the species" but rather the genetic pay-off to adult females who produce male or female offspring. Under random assortment, each female contributes one X chromosome and each male either an X or Y chromosome with equal probability. Skewed sex ratio at birth could be due to viability of implantation, development during gestation, and so forth if there is a "physiological edge" to being a male with XY chromosomes. That is, if the sex ratio at conception is 50/50 but more males die before reaching breeding age, then a female who "decides" (in an evolutionary sense) to produce a male is playing the "genetic lottery" (less chance of having any grandchildren, but more of them if it happens). Females should "decide" to take this gamble up to a point, but no more. Given the pace of evolution in mammals, the ratio of male to female births should represent an "Evolutionary Stable Strategy", that is, a "true state of nature" at least for the duration of multiple human generations. The study under discussion was conducted out of what was then the Department of Animal Ecology at Iowa State University (under the direction of Dr. William Franklin) to determine if there is evidence that such a strategy has developed in Guanacos so that more males than females are born. I have been told that geneticists have, for reasons unclear to me, predicted that the proportion of

males at birth should be 0.524.

The study design was quite simple. Field workers located and observed adult female Guanacos over a period of 4 years, recording the number of male and female offspring in each year. Each female Guanaco that breeds in a given year produces one offspring. Since we are assuming random assortment in the genetic process, which males were involved is irrelevant (at least to us, maybe not to them). The number of male and female births recorded in this study are given in the following table.

Year	Males	Females
1987	33	30
1988	47	42
1989	51	42
1990	53	46

Since each female in a given year produces one offspring, and we are assuming that the male parent is not a factor in sex of that offspring, in a given year i it is not unreasonable to formulate an observation model for the number of male offspring as $Y_i \sim Bin(\theta, n_i)$ where n_i is taken as fixed by the number of births observed. Although some females may have been observed several times over the 4 year period, individual identification was either not possible or not recorded. Thus, from a statistical perspective, we will assume that Y_1, Y_2, Y_3 and Y_4 are *exchangeable* (more on exchangability in the next chapter). Our *absolute* prior knowledge about the value of θ before any data are observed is that $0 < \theta < 1$, and we will choose a beta distribution to represent our knowledge of this parameter, which represents the "true state of nature", based on arguments from the ecological and evolutionary sciences (as sketched above). We might, in an effort to appear "objective", choose to take the prior

for 1987 to have parameters $\alpha_0 = 1$ and $\beta_0 = 1$, resulting in a uniform distribution on $(0, 1)$ for θ. From the discussion of Example 10.1, the posterior after observation in 1987 is a beta distribution with parameters $\alpha_1 = \alpha_0 + y_1$ and $\beta_1 = \beta_0 + n_1 - y_1$. Proceeding in a sequential manner as illustrated in the table of page 801, we obtain the following posterior parameters:

Year	$\hat{\alpha}$	$\hat{\beta}$	Mean	Variance
1987	34	31	0.5231	0.0038
1988	81	73	0.5260	0.0016
1989	132	115	0.5344	0.0010
1990	185	161	0.5347	0.0007

Graphs of the posterior densities (which by conjugacy will all be beta densities) are shown in Figure 11.1; the prior for the 1987 sample is a uniform and not shown in this figure.

Credible intervals for θ (we will get to these in Chapter 13.2) were,

Year	95% Interval	90% Interval
1987	(0.402, 0.642)	(0.421, 0.624)
1988	(0.447, 0.604)	(0.460, 0.592)
1989	(0.472, 0.596)	(0.482, 0.586)
1990	(0.482, 0.587)	(0.490, 0.579)

As can be seen from the above tables and Figure 11.1, the posterior distributions appear to be "narrowing in" on a value of θ that is above the value of 0.50, although after 4 years of data the value 0.50 is still included in both 95% and 90% credible intervals.

While the results of Example 11.1 are a pleasing illustration of a sequential process of gaining more and more knowledge about a parameter θ, as evidenced by the decreasing variances of the posterior distributions in moving from 1987

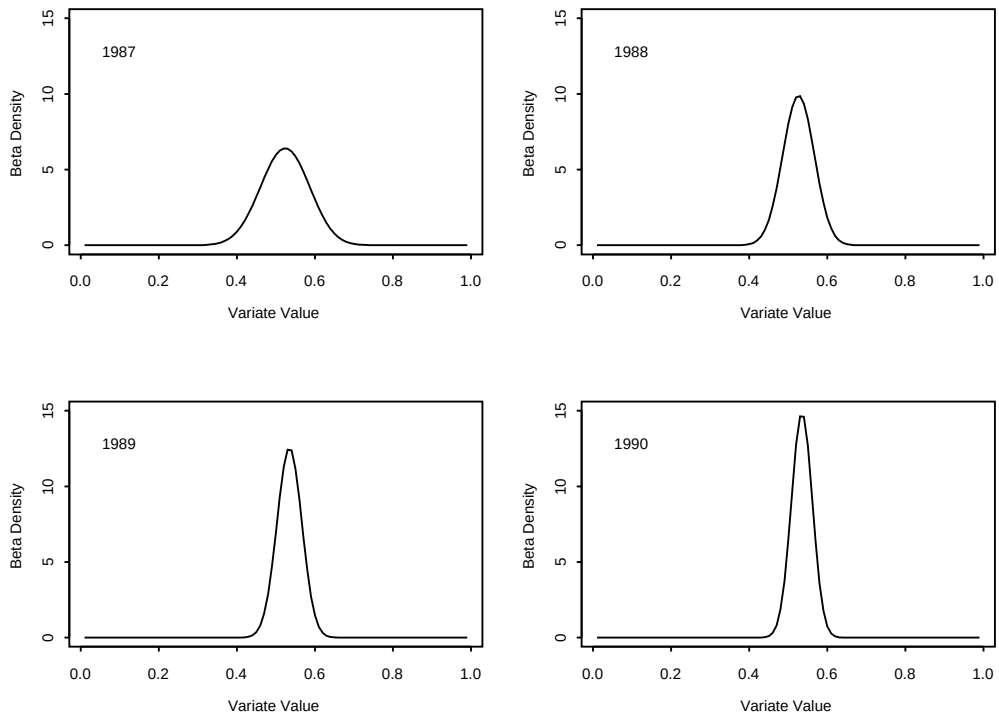

Figure 11.1: Sequential posterior densities for the analysis of sex ratio at birth in Guanacos. The initial prior was a uniform distribution on the interval (0, 1).

to 1990, one has to wonder if this truly represents some kind of scientific support for the concept that θ represents an evolutionarily stable strategy on which we are "zeroing in" as more data are accumulated. An examination of the model (and evidenced by the values of $\hat{\alpha}$ and $\hat{\beta}$ in the previous table), shows that the beta distribution parameters α and β will both increase as more data are collected. Does this essentially imply that the posterior variance will decrease as a function of amount of data, regardless of what those data indicate about the value of θ? This question is difficult to answer analytically, but an example will suffice to illustrate the point.

Example 11.2

First consider a simulated version of the situation of Example 11.1, in which 5 successive values $y_1, \ldots, y_5$ were independently simulated from a binomial distribution with parameter $\theta = 0.55$ and binomial sample size fixed at $n = 30$ for each value. Beginning with a uniform (0, 1) prior, analysis of these data in the same manner as that of Example 11.1 produced the following posterior values:

i	y_i	Mean	Variance
1	14	0.469	0.0075
2	19	0.548	0.0039
3	18	0.565	0.0026
4	16	0.557	0.0020
5	14	0.539	0.0016

We see in this table essentially the same behavior as that of the table of posterior values for Example 11.1, which is good, although in that example we do not know the actual value of θ, while here we know the true value is $\theta = 0.55$. Nevertheless, what we are doing in this example is essentially trying

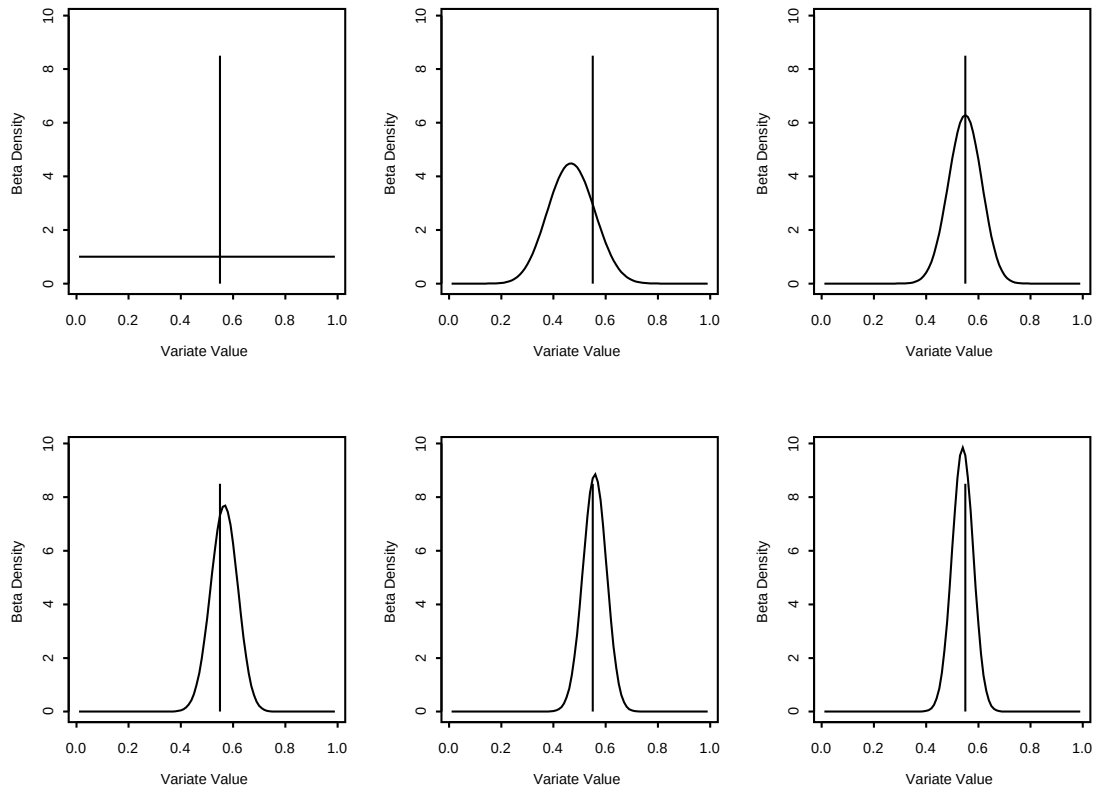

Figure 11.2: Posterior densities for the first set of simulated data in Example 11.2. The true value of $\theta = 0.55$ is shown by the solid vertical line.

to mimic the behavior of data and estimators from that situation with real data. Graphs of the sequence of prior/posterior densities for the simulated data are shown in Figure 11.2, which again looks quite similar to Figure 11.1.

Now consider values $y_1, \ldots, y_5$ simulated from a beta-binomial model in which the observation or data model is taken to be $Bin(\theta_i, n)$, again with $n = 30$, and $\theta_1, \ldots, \theta_5$ are values from a beta distribution with parameters $\alpha = 4$ and $\beta = 3.2727$. These values are the $\boldsymbol{\lambda}$ in the mixing distribution

$g(\theta|\boldsymbol{\lambda})$ and produce $E(\theta) = 0.55$ and $var(\theta) = 0.0299$. Suppose, however, that we apply the same model used previously in which each y_i; $i = 1, \ldots, 5$ is taken to be from the same binomial with parameter θ and $n = 30$. In this case, a sequential analysis, proceeding exactly as above yields the following results:

			Posterior	
i	y_i	θ_i	Mean	Variance
1	11	0.407	0.375	0.0071
2	16	0.412	0.452	0.0039
3	18	0.532	0.500	0.0027
4	25	0.855	0.582	0.0020
5	20	0.704	0.599	0.0016

While the posterior expectations in the above table are perhaps somewhat more variable than those from the previous table (although this would be exceedingly difficult to detect without having the previous values available), what I want to draw your attention to are the posterior variances, which are so similar to the first portion of this example as to be discomforting. Graphs of the posterior densities for these data from fitting the (incorrect) sequential model are shown in Figure 11.3.

Figure 11.3 is, again, quite similar to Figure 11.2, particularly if one did not have the solid vertical line showing, in this case, the true expected value of θ_i; $i = 1, \ldots, 5$. In fact, without knowledge of the true data generating mechanism (in this case the simulation distributions) we would not consider the results of this sequential analysis applied to the two sets of data different at all, and that is exactly the point. Although the data were simulated from quite distinct models, an analysis supposing a single binomial parameter gave quite similar results.

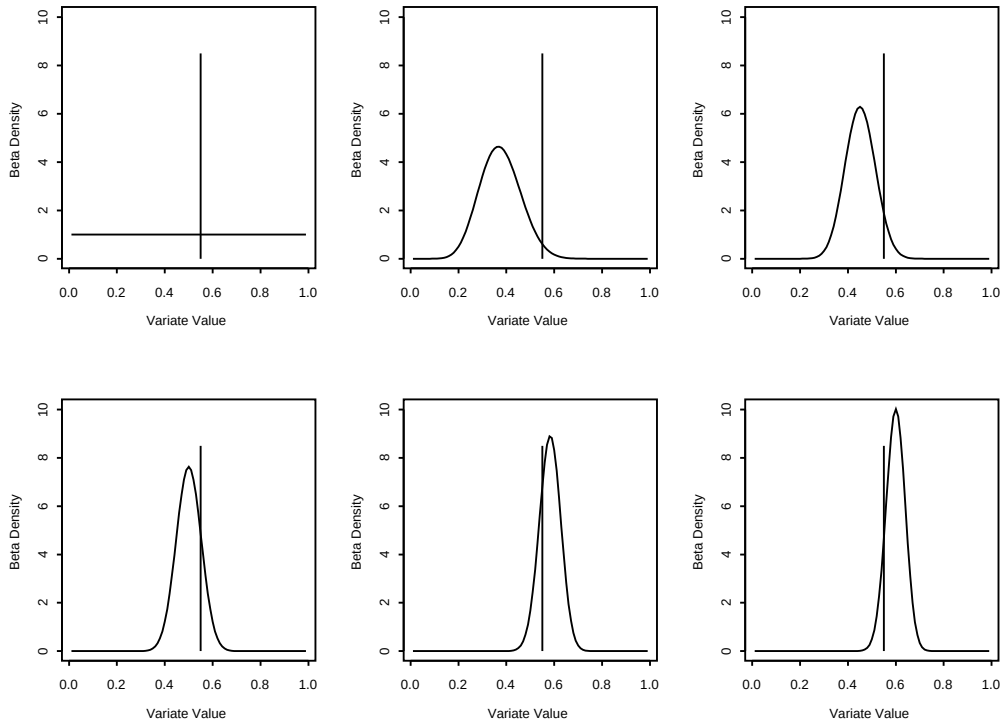

Figure 11.3: Posterior densities for the second set of simulated data in Example 11.2. The true value of $E(\theta) = 0.55$ is shown by the solid vertical line.

If we superimpose on the graphs of Figure 11.3 the true distribution $g(\theta|\boldsymbol{\lambda}) = g(\theta|\alpha, \beta)$ we obtain what is shown in Figure 11.4. If inferences were made about a value of θ that was (incorrectly) assumed to be constant for these data, we would base that inference on the densities shown by the solid curves of Figure 11.4, which would clearly be misleading. Relative to the analysis of Example 11.1 about sex ratio at birth in Guanacos, the implications of this exercise is that we should take no comfort from the progression of posterior densities that appear to be closing in on a given value of θ. Rather, support for that conclusion must come from the scientific argument that θ does in fact represent an evolutionarily stable strategy, and hence should be considered the same for each year of observation.

The message of this example is *not* that Bayesian analysis can result in misleading inferences. The same difficulty illustrated above could easily be encountered in a non-Bayesian analysis of this problem. The message is that *uncritical* Bayesian analysis can lead to such difficulties, just as uncritical non-Bayesian analysis. That is, there is no special protection against the deleterious effects of model misspecification offered by taking a Bayesian approach. No matter what approach is taken toward estimation and inference of parametric statistical models, the modeling process itself is critical. Gelman, Carlin, Stern and Rubin (1995, p. 161) make this same point in emphasizing that the use of sensitivity analysis in a Bayesian approach should consider not only the effects of prior specification, but also the likelihood that results from a given data model.

Now consider a Bayesian analysis of the actual model used to simulate the second data set of Example 11.2, which was a beta-binomial model. We would take the model to consist of binomial specifications for independent random variables $Y_1, \ldots, Y_5$, each with its own data model parameter $\theta_1, \ldots, \theta_5$, which

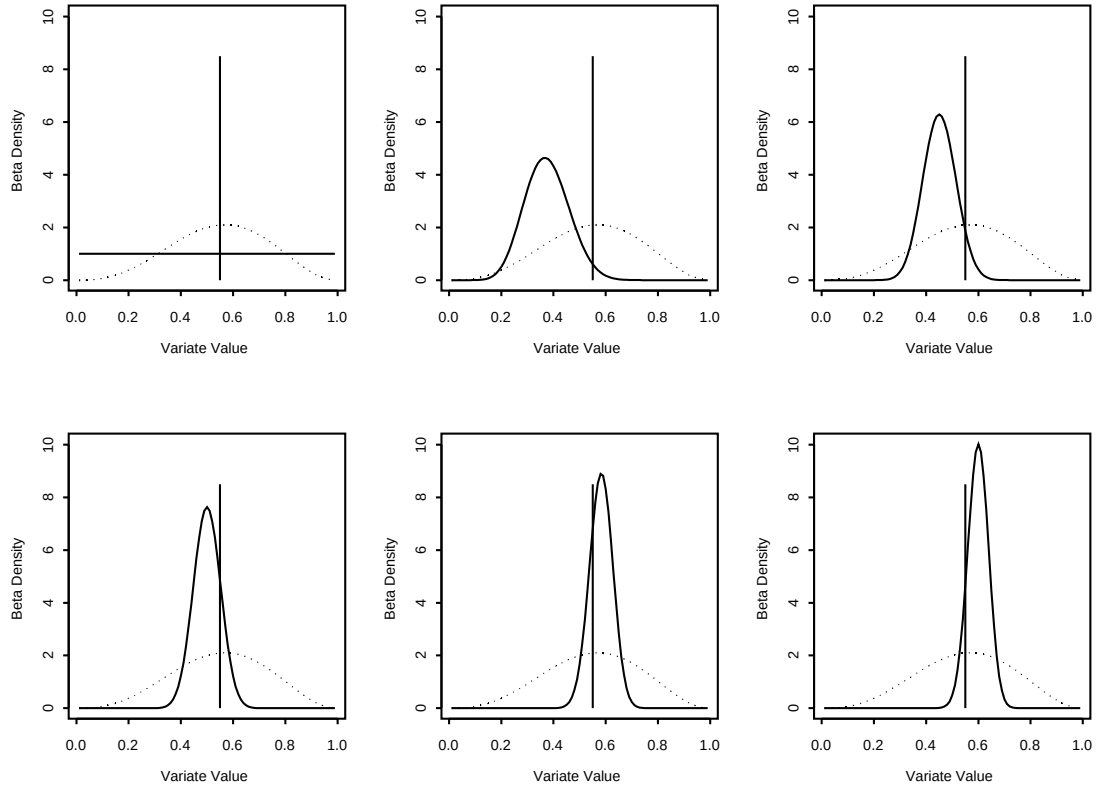

Figure 11.4: Posterior densities for the second set of simulated data in Example 1.2 with the true mixing density for values of θ overlaid. The true value of $E(\theta) = 0.55$ is shown by the solid vertical line.

follow the same beta distribution with parameters α and β. A prior would be specified for the distribution of $\boldsymbol{\lambda} \equiv (\alpha,\ \beta)^T$ as $\pi(\boldsymbol{\lambda})$. The posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$ would represent our knowledge about $\boldsymbol{\lambda}$ in the light of the observed data $y_1, \ldots, y_5$. But what about the "posterior" $p(\boldsymbol{\theta}|\boldsymbol{y})$? Here, given independence of the Y_is and exchangeability or independence of the θ_is, this distribution would have the form,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \prod_{i=1}^{5} p_i(\theta_i|\boldsymbol{y}) \\ &= \prod_{i=1}^{5} p_i(\theta_i|y_i). \end{aligned}$$

What interpretation should be given to the $p_i(\theta_i|y_i)$ in this case? This is the conditional distribution of θ_i given the observed value y_i, but, if each θ_i is representation of a unique random variable, what does this conditional distribution (i.e., this "posterior") tell us? In the model, the common distribution of the θ_is, namely $g(\theta|\boldsymbol{\lambda})$ represented the (distribution of the) ways in which the scientific mechanism or phenomenon of interest is expressed in observable situations. But here, each $p_i(\theta_i|y_i)$ will be a different distribution. Does $p_i(\theta_i|y_i)$ then represent what we know about how the mechanism was manifested in the particular situation that led to the observed value y_i? Perhaps. Does $p_i(\theta_i|y_i)$ represent what we know about the "true state of nature" in the situation that led to the observation y_i? Perhaps. But, extending the pure reductionist view that everything would be deterministic if we understood all of the forces at work in particular situations, the "true state of nature" should represent forces or factors (i.e., scientific mechanisms or phenomena) that are larger in scope than can be seen in those particular instances, that is, commonalities or fundamental laws about nature. And, from this perspective, the conditional distributions $p_i(\theta_i|y_i)$ seem to be less of any type of "posterior" distributions

than simply mathematical expressions of questionable value.

A number of points are relevant at this juncture:

1. It is difficult to fit the above situation into the framework offered by Carlin and Louis (2000) that in hierarchical models $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is a true prior on $\boldsymbol{\theta}$ that is controlled by a parameter $\boldsymbol{\lambda}$ about which we have additional uncertainty and thus use a "second-stage" prior $\pi(\boldsymbol{\lambda})$.

2. There does seem to be a role for $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ in inference, since this distribution is thought (from the modeling exercise) to apply to a broader set of situations than those under observation, while $p(\boldsymbol{\theta}|\boldsymbol{y})$ concerns only the particular values of $\boldsymbol{\theta}$ that may have been associated with what was observed.

3. Clearly, whatever view we have of Bayesian (or even non-Bayesain) analysis, a fundamental concern is model formulation. This is one of the reasons I indicated previously that the viewpoint of "simply putting distributions on unknown quantities" and then making inference by conditioning unobservable quantities on observable quantities was perhaps a bit overly simplistic. Without careful consideration of how a model is connected to a problem under study, which seems most natural by considering the concepts of random variables, random parameters, and fixed parameters, no approach to analysis will be guaranteed to provide reasonable scientific inferences.

4. Coming full circle, I now repeat the claim made in the first sentence of this part of the course (Chapter 10) that what we are concerned with is the Bayesian analysis of models, not the analysis of Bayesian models.

Chapter 12

Prior Distributions

Regardless of the overall view one might take toward Bayesian analysis, there is a need to specify prior distributions for one or more quantities in a model. In this chapter we consider several topics that are involved in this process of assigning priors.

12.1 Exchangeability

A common assumption in Bayesian analyses is that the "observable" random variables $Y_1, \ldots, Y_n$ are *exchangeable*. The meaning of exchangeable is given in the following definition.

Definition:

1. $Y_1, \ldots, Y_n$ are marginally exchangeable if, for a probability density or mass function $m(\cdot)$ and permutation operator $\mathcal{P}$,

$$m(y_1, \ldots, y_n) = m\left(\mathcal{P}(y_1, \ldots, y_n)\right).$$

2. $Y_1, \ldots, Y_n$ are conditionally exchangeable given z if, for a probability density or mass function $m(\cdot)$ and permutation operator $\mathcal{P}$,

$$m(y_1, \ldots, y_n|z) = m\left(\mathcal{P}(y_1, \ldots, y_n)|z\right).$$

The interpretation of these definitions needs clarification. For any valid joint distribution it is always true that the indices of variables may be permuted. That is, for random variables X_1 and X_2 is is always true that

$$Pr(X_1 = x_1,\, X_2 = x_2) = Pr(X_2 = x_2,\, X_1 = x_1).$$

This is trivial, not exchangeability. What exchangeability implies is that

$$Pr(X_1 = x_1,\, X_2 = x_2) = Pr(X_1 = x_2,\, X_2 = x_1),$$

which is a quite different condition. The implication of exchangeability is that the probability with which random variables assume various values does not depend on the "identity" of the random variables involved; this is essentially a "symmetry"condition.

It is true that *iid* random variables are exchangeable, and we often assume the condition of independent and identical distribution, but we should realize that exchangeability is not the same as either independence or identical distribution, as shown by the following example.

Example 12.1

1. Exchangeable but not Independent Random Variables. Let the pair of random variables (X, Y) be bivariate with possible values

$$(X, Y) \in \{(0, 1),\, (0, -1),\, (1, 0),\, (-1, 0)\},$$

such that each possible value has probability 0.25.

Exchangeability:

Clearly, $Pr(X = x, Y = y) = Pr(X = y, Y = x)$, since each possible value has the same probability.

Lack of Independence:

$Pr(X = 1) = 0.25$ and $Pr(Y = 0) = 0.5$, but $Pr(X = 1, Y = 0) = 0.25 \neq 0.25(0.5)$

2. Independent but not Exchangeable Random Variables.

Let X and Y be any two independent random variables with X being discrete with probability mass function $f_X(x);\ x \in \Omega_X$ and Y being continuous with probability density function $f_Y(y);\ y \in \Omega_Y$.

Independence: Independence is by assumption, so that the joint (mixed) density (and mass function) is

$$m(x, y) = f_X(x) f_Y(y);\ (x, y) \in \Omega_X \times \Omega_Y.$$

Lack of Exchangeability:

For any $y \notin \Omega_X$ we would have

$$f_X(y) f_Y(x) = 0 \neq f_X(x) f_Y(y),$$

so that X and Y cannot be exchangeable.

In the first portion of this example, what "messes up" independence is the lack of what has previously been called the positivity condition. That is, while $\Omega_X \equiv \{-1, 0, 1\}$ and $\Omega_Y \equiv \{-1, 0, 1\}$, and the probability distributions of X and Y on these sets is the same (i.e., X and Y are identically distributed), it is not true that $\Omega_{X,Y} = \Omega_X \times \Omega_Y$. In the second portion of the example, what

"messes up" exchangeability is that the sets of possible values Ω_X and Ω_Y are not the same, although the positivity condition does hold in this case.

In general, random variables that are not identically distributed cannot be exchangeable, random variables that are independent and identically distributed are exchangeable, but exchangeability is not the same property as independence.

The role of exchangeability in formulation of prior distributions follows from a famous theorem of de Finetti (1974) which essentially states that, if quantities $\theta_1, \ldots, \theta_n$ are exchangeable following the same distribution $m(\theta_i|\boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ is unknown, then any suitably "well-behaved" joint distribution for these quantities can be written in the mixture form,

$$m(\boldsymbol{\theta}) = \int_{\Lambda} \left\{ \prod_{i=1}^{n} m(\theta_i|\boldsymbol{\lambda}) \right\} \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda},$$

as $n \to \infty$.

What this theorem basically justifies is the use of a prior as a mixing distribution in distributions that have the form of an *iid* mixture. We relied on exchangeability of $Y_1, \ldots, Y_4$ in Example 12.1 for the sequential analysis of sex ratio at birth in Guanacos. In hierarchical models, exchangeability of the data model parameters $\{\boldsymbol{\theta}_i : i = 1, \ldots, n\}$ and de Finetti's theorem does lend some credence to thinking of a distribution,

$$\pi_{\theta}(\boldsymbol{\theta}) = \int_{\Lambda} \left\{ \prod_{i=1}^{n} g(\theta_i|\boldsymbol{\lambda}) \right\} \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda},$$

as a prior to be applied to the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$.

12.2 Conjugate Priors

We have already seen the use of conjugate priors in Example 10.1 and Examples 11.1 and 11.2 on sequential Bayes. To expand on conjugacy, consider a simple setting with observation model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_0)$, where we have written the prior as a parameterized distribution, but are considering $\boldsymbol{\lambda}_0$ to be a known (or specified) value. The prior $\pi(\cdot)$ is conjugate for the data model $f(\cdot|\cdot)$ if the resultant posterior has the form,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_0)}{\displaystyle\int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_0)\,d\boldsymbol{\theta}} \\ &= \pi(\boldsymbol{\theta}|h(\boldsymbol{y},\boldsymbol{\lambda}_0)), \end{aligned}$$

where $h(\boldsymbol{y},\boldsymbol{\lambda}_0)$ is some function of $\boldsymbol{y}$ and $\boldsymbol{\lambda}_0$. That is, if in the transition from prior to posterior, the effect of the data $\boldsymbol{y}$ is only to modify the parameter values of the prior, not its functional form, then the prior $\pi(\cdot)$ is said to be conjugate for the given data model $f(\cdot|\cdot)$.

Example 12.2

Consider a data model consisting of $Y_1,\ldots,Y_n \sim iid\, N(\mu,\sigma^2)$ where σ^2 is considered known, and our interest is in the fixed parameter μ. Let the prior for μ be specified as $\mu \sim N(\lambda,\tau^2)$, where both λ and τ^2 are specified values. The posterior of μ is easily shown to be a normal distribution with parameters

$$\frac{\sigma^2\lambda + \tau^2 n\bar{y}}{n\tau^2+\sigma^2};\quad \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}.$$

Note also in this example that the posterior mean may be written as a weighted average of the prior mean λ and the usual data estimator $\bar{y}$ as

$$\frac{\frac{n}{\sigma^2}\,\bar{y} + \frac{1}{\tau^2}\,\lambda}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}.$$

This form also indicates why many statisticians write model formulations to which a Bayesian analysis is to be applied in terms of "precision" parameters (here $1/\sigma^2$ and $1/\tau^2$) rather than variances.

12.3 Noninformative Priors

In some ways the phrase "noninformative prior" is an unfortunate title for several of the methods for prior formulation discussed in this section. In the sense that noninformative implies providing no information, not all "noninformative" priors should be considered as such. Nevertheless, this has become the standard heading under which to consider the methods for formulating priors that we consider here.

12.3.1 Proper Uniform Priors

The parameter space of some data or observation models is bounded both above and below. For example, a binomial model with parameter θ implies that $0 < \theta < 1$. A multinomial model with $k + 1$ categories and parameters $\theta_1, \ldots, \theta_k$ implies that both $0 < \theta_j < 1$ for $j = 1, \ldots, k$ and that $\sum_j \theta_j \leq 1$. In these situations, placing a uniform prior on the allowable interval allows us to express a prior belief that gives no preference to any of the possible values of the parameter $\boldsymbol{\theta}$. We have already seen examples of a uniform priors used in this manner in previous examples that combine binomial data models with priors that take $\theta \sim Unif(0, 1)$.

In other models, even when the data model parameter space is unbounded in one or both directions, it may be possible to determine an interval (a, b) such that it is either physically impossible, or scientifically implausible that

$\theta < a$ or $\theta > b$. In these cases, it may be reasonable to assign θ a prior distribution that is uniform on the given interval.

Uniform priors sometimes (but not always) simplify calculation of the integral that appears in the denominator of expression (10.1) since then we have,

$$\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \frac{1}{b-a}\int_{a}^{b} f(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

An objection to uniform priors as "noninformative" is that uniform distributions are not invariant to transformation. For example, if $\theta \sim U(0,\, 1)$ then $\eta = 1/\theta$ has density $h(\eta) = 1/\eta^2$; $\quad 1 < \eta < \infty$. Thus, the indifference that would seem to be expressed by the uniform prior on θ does not translate into indifference about values of η, although the data model may be equivalently expressed as either $f(\boldsymbol{y}|\theta)$ or $f(\boldsymbol{y}|\eta)$. Note that this is an objection to uniform priors being thought of as noninformative in nature, not as an objection to uniform priors *per se*.

12.3.2 Improper Priors

For data models that have parameters with unbounded parameter space from one or both directions, an extension of the idea of giving all possible values equal prior weight results in improper priors of the form

$$\pi(\boldsymbol{\theta}) = 1; \qquad \boldsymbol{\theta} \in \Theta,$$

which are clearly not distributions since they do not integrate to any finite value as long as Θ is not a bounded set. Improper priors do not, however, necessarily imply improper posteriors. As long as the integral

$$\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta} = K(\boldsymbol{\theta}) < \infty,$$

then the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{K(\boldsymbol{\theta})},$$

will exist and will integrate to 1. It has actually become quite popular to use improper priors, particularly for some elements of the parameter vector $\boldsymbol{\theta}$ in complex models that involve a parameter $\boldsymbol{\theta}$ of high dimension.

Improper priors perhaps deserve the label "noninformative" in that posterior distributions in simple cases with improper priors often result in essentially the same conclusions that would be reached under the sampling distribution of non-Bayesian estimators.

Example 12.3

Consider again the situation of Example 12.2 in which $Y_1, \ldots, Y_n \sim iid\, N(\mu, \sigma^2)$ with σ^2 known. By sufficiency, we may reduce this data model to consideration of $\bar{Y} \sim N(\mu, \sigma^2/n)$. Suppose that we place an improper prior on μ as $\pi(\mu) = 1; \quad -\infty < \mu < \infty$. The resulting posterior is,

$$\begin{aligned} p(\mu|\boldsymbol{y}) &= \frac{\exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right\}}{\int_{-\infty}^{\infty} \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right\}} \\ &\propto \exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right\}, \end{aligned}$$

which is the density of a normal distribution with mean $\bar{y}$ and variance σ^2/n, that is, $p(\mu|\boldsymbol{y})$ is $N(\bar{y},\, \sigma^2/n)$. While we have not yet discussed Bayesian estimation and inference, it should be intuitive that a reasonable point estimate of μ is the expectation of the posterior distribution which would be, in this case, $\bar{y}$ which agrees with any sensible non-Bayesian estimation. Similarly, a reasonable 90% interval estimate would be the central 90% of the posterior

density or distribution, namely $\bar{y} \pm 1.645\,(\sigma^2/n)$ which again agrees with what would be obtained from a non-Bayesian approach.

As illustrated by Example 12.3, improper priors often lead to situations in which the prior actually plays no role at all or is, in truth, "noninformative". It may be the case that we wish to take this approach for certain elements of $\boldsymbol{\theta}$, while placing informative proper prior distributions on our knowledge of other elements of $\boldsymbol{\theta}$. The standard caution in use of improper priors is that one must be certain that the resulting posterior is in fact a distribution. This is not always a trivial matter.

12.3.3 Jeffreys' Priors

Consider again the objection to using uniform prior distribution that they are not invariant to transformation in expressing lack of knowledge or indifference about values of a parameter. Jeffreys (1961) introduced a procedure for prior formulation based on the idea that any noninformative prior should be equivalent, in terms of expression of prior knowledge, on different scales.

To understand this, consider a data model with scalar parameter, and a procedure for assigning a prior, such as priors that are uniform on the range of the parameter space. If applied to the data model $f(\boldsymbol{y}|\theta)$ with parameter space $\theta \in \Theta \equiv (\theta_1, \theta_2)$, this procedure results in the prior $\pi_\theta(\theta)$. Now consider an alternative parameterization using $\eta \equiv h(\theta)$ for some one-to-one transformation $h(\cdot)$. An equivalent model is now $f(\boldsymbol{y}|\eta)$ with parameter space $\eta \in (h(\theta_1), h(\theta_2))$. Applying the same procedure for prior formulation as used under the model $f(\boldsymbol{y}|\theta)$ results in a prior $\pi_\eta(\eta)$. But, the original prior π_θ also implies a distribution for $\eta \equiv h(\theta)$ as,

$$\pi'_\eta(\eta) \;=\; \pi_\theta(h^{-1}(\eta)) \left| \frac{d\,h^{-1}(\eta)}{d\,\eta} \right|$$

$$= \pi_\theta(\theta) \left| \frac{d\,\theta}{d\,\eta} \right|. \tag{12.1}$$

Jeffreys idea was that the procedure for assigning priors $\pi_\theta(\theta)$ and $\pi_\eta(\eta)$ is invariant under transformation if

$$\pi_\eta(\eta) = \pi'_\eta(\eta),$$

that is, if the prior assigned under the model $f(\boldsymbol{y}|\eta)$ is the same as the distribution that results from the prior assigned under model $f(\boldsymbol{y}|\theta)$ by transforming θ into η. As we have seen in Section 12.3.1, assigning uniform priors on the ranges of parameter spaces does not result in this property.

The suggestion Jeffreys gave for a procedure to assign priors that would result in this property was to take, under a model $f(\boldsymbol{y}|\theta)$,

$$\begin{aligned} [\pi_\theta(\theta)]^2 &\propto E\left[\left(\frac{d\,\log f(\boldsymbol{y}|\theta)}{d\,\theta}\right)^2\right] \\ &= -E\left[\frac{d^2\,\log f(\boldsymbol{y}|\theta)}{d\,\theta^2}\right] \\ &= I(\theta), \end{aligned}$$

or,

$$\pi_\theta(\theta) = \{I(\theta)\}^{1/2}. \tag{12.2}$$

The form (12.2) is thus known as *Jeffreys prior*.

To verify that the procedure (12.2) does result in the desired property for priors, apply this procedure to the model $f(\boldsymbol{y}|\eta)$, which gives,

$$[\pi_\eta(\eta)]^2 \propto E\left[\left(\frac{d\,\log f(\boldsymbol{y}|\eta)}{d\,\eta}\right)^2\right]$$

$$
\begin{aligned}
&= E\left[\left(\frac{d\,\log f(\boldsymbol{y}|\theta)}{d\,\theta}\,\left|\frac{d\,\theta}{d\,\eta}\right|\right)^2\right] \\
&= E\left[\left(\frac{d\,\log f(\boldsymbol{y}|\theta)}{d\,\theta}\right)^2\right]\,\left|\frac{d\,\theta}{d\,\eta}\right|^2 \\
&= I(\theta)\left|\frac{d\,\theta}{d\,\eta}\right|^2,
\end{aligned}
$$

or,

$$\pi_\eta(\eta) = \{I(\theta)\}^{1/2}\,\left|\frac{d\,\theta}{d\eta}\right|.$$

Now, using (12.1), the distribution for η implied by (12.2) is

$$\pi'(\eta) = \{I(\theta)\}^{1/2}\,\left|\frac{d\,\theta}{d\,\eta}\right|.$$

Thus, $\pi'_\eta(\eta) = \pi_\eta(\eta)$ and the approach suggested by Jeffreys does result in the property desired.

Example 12.4

Suppose that we have a single observation corresponding to the data model $Y \sim Bin(\theta, n)$ where n is fixed. We now have two choices for assigning θ a so-called noninformative prior distribution. The first would be to take $\pi_1(\theta) = 1; \quad 0 < \theta < 1$, while the second would be to use the procedure of Jeffreys. In this case,

$$
\begin{aligned}
I(\theta) &= -E\left[\frac{d^2\,\log f(\boldsymbol{y}|\theta)}{d\theta^2}\right] \\
&= \frac{n}{\theta(1-\theta)},
\end{aligned}
$$

so that Jeffreys prior would be $\pi_2(\theta) \propto \{\theta(1-\theta)\}]-1/2$. As we have already seen in Example 10.1, the uniform prior results in a beta posterior with para-

meters $1+y$ and $1+n-y$. A parallel derivation gives the posterior associated with the Jeffreys prior as beta with parameters $(1/2)+y$ and $(1/2)+n-y$.

In principle, Jeffreys' method for forming noninformative priors can be extended to the case of vector valued $\boldsymbol{\theta}$ by taking

$$\pi_\theta(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2},$$

where $|I(\boldsymbol{\theta})|$ here denotes the determinant of the expected information matrix $I(\boldsymbol{\theta})$. This type of multidimensional prior can be difficult to achieve in practice, however, and Jeffreys priors are usually seen in simpler cases with scalar θ.

12.4 Priors for Vector Parameters

Although much of what we have presented applies in principle to parameter vectors $\boldsymbol{\theta}$ (e.g., Jeffreys priors), the practical assignment of a joint prior to a p-dimensional parameter $\boldsymbol{\theta}$ can become less than an easy matter. Two techniques that are often used in such situations are to take joint priors as products of individual (marginal) priors in the same way we would form a joint distribution for independent random variables, and to specify priors for some of the components of $\boldsymbol{\theta}$ as conditional on the others, and specify a marginal prior for those other components. We illustrate these two techniques here for a model with two-dimensional parameter $\boldsymbol{\theta} \equiv (\theta_1, \theta_2)^T$.

Example 12.5

Consider again the beta-binomial mixture model of Example 10.2. There we had

$$\begin{aligned} Y_1, \ldots, Y_m &\sim \mathit{indep}\, Bin(\theta_i) \\ \theta_1, \ldots, \theta_m &\sim \mathit{iid}\, Beta(\alpha, \beta) \end{aligned}$$

What is necessary to conduct a Bayesian analysis for this model is a prior for the parameter $\boldsymbol{\lambda} \equiv (\alpha, \beta)^T$. Now, the parameter space is $\alpha > 0$, $\beta > 0$, but if we first reparameterize the beta mixing distribution in terms of parameters,

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \eta = \frac{1}{\alpha + \beta + 1},$$

then $0 < \mu < 1$ and $0 < \eta < 1$. We might then assign the joint prior as

$$\pi(\mu, \eta) = \pi_\mu(\mu)\, \pi_\eta(\eta),$$

where both $\pi_\mu(\cdot)$ and $\pi_\eta(\cdot)$ are uniform distributions on the interval $(0, 1)$. Derivation of the posterior $p(\alpha, \beta|\boldsymbol{y})$ would, in this example, require the use of simulation methods.

Example 12.6

Consider again the normal one-sample problem of Example 12.3, but now not assuming that the variance σ^2 is known. Here, it would not be possible to consider only the distribution of $\bar{Y}$ in the likelihood, since $\bar{Y}$ is sufficient for μ but not σ^2. Thus, we must work with the full joint distribution of $Y_1, \ldots, Y_n$, which can be written as,

$$\begin{aligned} f(\boldsymbol{y}|\mu, \sigma^2) &= \{2\pi\sigma^2\}^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \\ &= \{2\pi\sigma^2\}^{n/2} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(y_i - \bar{y})^2\right\}\right. \\ &- \left.\frac{1}{2\sigma^2}\, n\,(\bar{y} - \mu)^2\right]. \end{aligned}$$

One way to assign the joint prior $\pi(\mu, \sigma^2)$ to this model is to use the conditional prior $\pi_1(\mu|\sigma^2)$ and the marginal prior $\pi_2(\sigma^2)$ as,

$$\pi_1(\mu|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\kappa_0}{2\sigma^2}\{\mu - \mu_0\}^2\right]$$

$$\pi_2(\sigma^2) \;=\; \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\{\sigma^2\}^{-(\alpha_0+1)}\exp\{-\beta_0/\sigma^2\}.$$

Here, $\pi_1(\cdot)$ is normal with parameters μ_0 and σ^2/κ_0, while $\pi_2(\cdot)$ is inverse gamma with parameters α_0 and β_0, which are conjugate for μ in a model with σ^2 assumed known and σ^2 with μ assumed known, respectively.

It can be shown, using this model with prior $\pi(\mu,\sigma^2) = \pi_1(\mu|\sigma^2)\,\pi_2(\sigma^2)$ that the marginal posterior $p(\mu|\boldsymbol{y})$ is a $t-$ distribution, the marginal posterior $p(\sigma^2|\boldsymbol{y})$ is an inverse gamma distribution, and the conditional posterior $p(\mu|\sigma^2,\boldsymbol{y})$ is a normal distribution (e.g., Gelman, Carlin, Stern and Rubin, 1995, pp. 72-73). What is important for us at this point is the use of the conditional prior $\pi_1(\mu|\sigma^2)$. First, the fact that σ^2 appears in this prior indicates that μ and σ^2 are not independent in the joint prior $\pi(\mu,\sigma^2)$.

Secondly, the fixed constant κ_0 plays the role of the "number of observations" we believe our prior information about μ is "worth" in the overall problem; note that $\pi_1(\cdot)$ has the form of the distribution of the sample mean of κ_0 observations taken from a $N(\mu,\sigma^2)$ distribution.

Gelman, Carlin, Stern and Rubin (1995) use this example also as a simple case in which simulation-based derivation of the posterior is valuable. While the marginal posteriors are of relative "nice" forms, and the joint posterior can be derived in closed form, this joint posterior is not necessarily easily manipulated to find, for example, expectations, quantiles, and so forth. But the fact that the conditional posterior $p(\mu|\sigma^2,\boldsymbol{y})$ is normal and the marginal $p(\sigma^2|\boldsymbol{y})$ is inverse gamma indicates that these two univariate distributions may be easily simulated from. An algorithm to simulate values from the joint posterior is then also easily constructed as follows:

1. Generate (or draw or simulate) a value σ^{2*} from the marginal posterior

$p(\sigma^2|\boldsymbol{y})$.

2. Using this value, generate a value μ^* from the conditional posterior $p(\mu|\sigma^{2*}, \boldsymbol{y})$.

3. The resulting pair $(\mu^*,\ \sigma^{2*})$ is one value from the joint posterior $P(\mu, \sigma^2|\boldsymbol{y})$.

4. Repeat this process a large number M times, and make inference based on the empirical distribution of the set of values $\{(\mu_j^*,\ \sigma_j^{2*}) : j = 1, \ldots, M\}$.

It is also instructive to contrast Example 12.6 to what would happen if, instead of specifying $\pi_1(\mu|\sigma^2)$ as $N(\mu_0,\ \sigma^2/\kappa_0)$ we would use independent priors and simply take $p(\mu)$ as $N(\mu_0, \tau_0^2)$) for example. In this case (e.g., Gelman, Carlin, Stern and Rubin, 1995, Chapter 3.4) μ and σ^2 are still dependent in the joint posterior, the conditional posterior of μ given σ^2 is again normal, but the marginal posterior of σ^2 cannot be derived in closed form. The simulation algorithm given above can also be used in this situation, but the initial step of generating σ^{2*} becomes more difficult, and must make use of an indirect method of simulation (e.g., inversion, rejection sampling, etc.).

Chapter 13

Basic Estimation and Inference

In some ways, discussing Bayesian methods of inference is a very short topic. Under the concept of epistemic probability, a posterior distribution represents our knowledge about a parameter of interest. That is, given a posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$, we are free to make probability statements about $\boldsymbol{\theta}$ with the understanding that our statements actually refer to our knowledge about the value of $\boldsymbol{\theta}$ which, under the strict Bayesian viewpoint is a fixed parameter and under the viewpoint of Bayesian analysis of uncertainty is just some unknown quantity. That is, in this context it is perfectly acceptable to write a statement such as

$$Pr(a \leq \theta \leq b) = \alpha,$$

where a, b, and α are all particular real numbers. Nevertheless, there are a few particular issues relative to Bayesian inference that merit brief consideration. Before moving on to these, we will mention that, although inference is to be based on a posterior, there are several broad methods for obtaining those posteriors, several of which have been eluded to in the previous sections. We may find a posterior distribution through one of three avenues:

1. Analytical derivation.

2. Approximation.

3. Simulation.

The first of these, analytical derivation, we have seen in a number of the examples presented. Here, a posterior is derived in closed mathematical form, usually as a probability density function. Simple models with conjugate priors are a prime example of situations in which this approach is useful. We will not discuss approximation in this course, but will only mention that the most popular such approximation is known as the *Laplace Approximation*. For a very accessible introduction to this approximation see Carlin and Louis (2000) Chapter 5.2.2. Finally, simulation of posteriors has become the most widely used method for finding posteriors in use today. This is because it may be applied to a huge variety of models, many of which could not be approached by any other technique. Our department offers a course (the old Stat 601, the number of which was usurped for this current course, and has yet to receive a new designation) in which simulation of posteriors using methods known collectively as Markov Chain Monte Carlo (MCMC) are the bulk of what is discussed.

13.1 Point Estimation

While it is well and fine to indicate that, given a posterior distribution, one is free to make any probability statement desired, there is typically still a desire for concise summarization of a posterior. For example, we may desire as a summarization of the location of a posterior distribution a single value or "point estimate". Here, I am about to start sounding like I'm teaching Statistics 101 because the values that are used are the posterior mean, median,

or mode. The median is not used that frequently, the mode is still popular, but less than it was before the advent of MCMC simulation, and the mean is often the quantity of choice in Bayesian analyses.

The posterior mode is often the easiest to find in an analytical approach, because it does not depend on finding the posterior "normalizing constant", the denominator of

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}.$$

As shown for one case in Example 12.3, if $\pi(\boldsymbol{\theta}) \propto 1$ for $\boldsymbol{\theta} \in \Theta$, then the posterior mode is equal to the maximum likelihood estimate, since then,

$$\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y}) = \max_{\boldsymbol{\theta}} f(\boldsymbol{y}|\boldsymbol{\theta}).$$

Use of the posterior mean or expected value can be justified based on decision-theoretic grounds, if one considers squared error loss (e.g., Berger, 1985, p. 161). For scalar θ, the posterior mean is given in the obvious way as,

$$E\{\theta|\boldsymbol{y}\} = \int_{\Theta} \theta\, p(\theta|\boldsymbol{y})\,d\theta, \tag{13.1}$$

and is sometimes reported along with the posterior variance,

$$var(\theta|\boldsymbol{y}) = \int_{\Theta} (\theta - E\{\theta|\boldsymbol{y}\})^2\; d\theta. \tag{13.2}$$

For vector-valued parameters $\boldsymbol{\theta}$,

$$E\{\boldsymbol{\theta}|\boldsymbol{y}\} = (E\{\theta_1|\boldsymbol{y}\}, \ldots, E\{\theta_p|\boldsymbol{y}\}),$$

and expression (13.1) continues to hold for the component quantities with θ replaced by θ_j and $p(\theta|\boldsymbol{y})$ replaced by $p_j(\theta_j|\boldsymbol{y})$, the marginal posterior of θ_j; $j = 1, \ldots, p$. The same is true for the variances and expression (13.2), with

the additional covariances given as,

$cov(\theta_j, \theta_k) =$

$$\int_{\Theta}\int_{\Theta}(\theta_j - E\{\theta_j|\boldsymbol{y}\})(\theta_k - E\{\theta_k|\boldsymbol{y}\})p_{j,k}(\theta_j, \theta_k|\boldsymbol{y})\; d\theta_j\, d\theta_k, \tag{13.3}$$

where $p_{j,k}(\theta_j, \theta_k|\boldsymbol{y})$ is the joint marginal posterior of θ_j and θ_k.

13.2 Interval Estimation

Although posterior variances and covariances can be, and often are, computed as given in expressions (13.2) and (13.3) they are typically not used to form interval estimates of $\boldsymbol{\theta}$ or its components. This is because we are not dealing with sampling distributions of estimators, and because we have at hand the entire posterior distribution of $\boldsymbol{\theta}$. The Bayesian analog of confidence sets or intervals are typically called "credible sets" or "credible intervals". All that is needed is a sensible way to make probability statements such as $Pr(a \leq \theta \leq b) = 1 - \alpha$ and find the appropriate values of a and b. The basic definition of a credible set for $\boldsymbol{\theta}$ is a set $\mathcal{C}$ such that

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in \mathcal{C}|\boldsymbol{y}) = \int_{\mathcal{C}} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \tag{13.4}$$

If $\boldsymbol{\theta}$ should happen to be discrete the integral in (13.4) is replaced with a summation.

For a given posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ there may be many sets $\mathcal{C}$ that satisfy (13.4). One technique that has been used to help get around this difficulty is to define a *Highest Posterior Density* credible set as a set

$$\mathcal{C}^* = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}|\boldsymbol{y}) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant such that $\mathcal{C}^*$ is a credible set. What this

means is that, for any $\boldsymbol{\theta}^* \in \mathcal{C}^*$ and any other $\boldsymbol{\theta} \in \mathcal{C}^{*\prime}$, where $\mathcal{C}^{*\prime}$ is the complement of $\mathcal{C}$,

$$p(\boldsymbol{\theta}^*|\boldsymbol{y}) \geq p(\boldsymbol{\theta}|\boldsymbol{y}).$$

In other words, the posterior density for any value of $\boldsymbol{\theta}$ included in the credible set is at least as great as that for any value not in the credible set.

While highest posterior density (HPD) credible sets are not hard to find for scalar θ, they can be quite difficult to determine in higher dimensions. In addition, HPD credible sets are not invariant to transformation of $\boldsymbol{\theta}$. For a more complete discussion of issues involved with credible sets, HPD credible sets and their extension to "optimal" credible sets see Berger (1985).

In many applications and, in particular, those in which the posterior is found through the use of simulation, a common practice is to use the "central" $1-\alpha$ interval for any component of $\boldsymbol{\theta}$, regardless of whether it would qualify as an HPD interval or not. That is, if we wish a $(1-\alpha)100\%$ credible interval for θ_j, that interval is given by (L, U) where

$$\begin{aligned} 1-\alpha/2 &= \int_{-\infty}^{L} p(\theta_j|\boldsymbol{y})\, d\theta_j \\ 1-\alpha/2 &= \int_{U}^{\infty} p(\theta_j|\boldsymbol{y})\, d\theta_j \end{aligned} \tag{13.5}$$

and where $p(\theta_j|\boldsymbol{y})$ is the marginal posterior of θ_j.

13.3 Model Comparison

Suppose that we have two competing models denoted as M_1 and M_2 that we would like to compare in light of a set of observations $\boldsymbol{y}$. These models may differ in the number of parameters associated with covariates (e.g., a typical "variable selection" problem in regression), by one model having fixed values

for a portion of a vector-valued parameter (e.g., $\mu = 0$ in a normal data model), by having parameter values that are restricted to different regions of a partitioned parameter space (e.g., $p < 0.5$ versus $p > 0.5$ in a binomial model), by having different priors, or by having different data models $f_1(\boldsymbol{y}|\theta_1)$ and $f_2(\boldsymbol{y}|\theta_2)$ (e.g., gamma versus lognormal). Notice that, in particular, we do not require nested parameter spaces for our competing models.

Now, suppose we represent our beliefs about the possible models M_1 and M_2 in terms of a prior distribution, which will necessarily place distinct probabilities on the possible models M_1 and M_2 (e.g., $\pi(M_1) = \gamma$ and $\pi(M_2) = 1-\gamma$). Let the values of this prior be represented as $Pr(M_1)$ and $Pr(M_2)$. The two models may then be compared by taking the ratio of posterior probabilities of the models as the "posterior odds ratio".

$$\begin{aligned} \frac{Pr(M_1|\boldsymbol{y})}{Pr(M_2|\boldsymbol{y})} &= \frac{Pr(M_1)}{Pr(M_2)} \frac{Pr(\boldsymbol{y}|M_1)}{Pr(\boldsymbol{y}|M_2)} \\ &= \frac{Pr(M_1)}{Pr(M_2)} BF(M_1,\, M_2) \end{aligned} \tag{13.6}$$

where $BF(M_1,\, M_2)$ denotes the "Bayes Factor" of model M_1 relative to model M_2, namely,

$$BF(M_1,\, M_2) = \frac{Pr(\boldsymbol{y}|M_1)}{Pr(\boldsymbol{y}|M_2)}. \tag{13.7}$$

Ways to interpret the concept quantified in a Bayes factor include:

1. BF is ratio of posterior odds in favor of model M_1 to the prior odds in favor of model M_1, as,

$$BF(M_1,\, M_2) = \frac{Pr(M_1|\boldsymbol{y})/Pr(M_2|\boldsymbol{y})}{Pr(M_1)/Pr(M_2)}.$$

2. BF is a likelihood ratio, which is a direct interpretation of expression (13.7), that is, the ratio of the likelihood of the data $\boldsymbol{y}$ under model M_1

to the likelihood of the data $\boldsymbol{y}$ under model M_2. Note, however, that the Bayes Factor of (13.7) is written in terms of probabilities rather than densities.

If the Bayes factor $BF(M_1, M_2)$ is greater than 1, then the posterior odds in favor of model M_1 are increased from the prior odds. If the prior odds ratio is taken to be 1 (i.e., $Pr(M_1) = Pr(M_2) = 0.5$) the the Bayes factor is equal to the posterior odds ratio in favor of M_1. How big should $BF(M_1, M_2)$ be in order for us to have substantially greater belief in M_1 than M_2? Kass and Raftery (1995) give a slightly modified version of a scale suggested by Jeffreys (1961) which suggests that values from 3.2 to 10 provide some evidence in favor of M_1, values from 10 to 100 provide strong evidence, and values greater than 100 provide "decisive" evidence. These authors also suggest their own scale which results in the same categories of evidence for ranges of Bayes factors 3 to 20 (some evidence), 20 to 150 (strong evidence) and greater than 150 (decisive evidence).

Now, in a situation in which a model is formulated through density functions,a given model M_i is embodied through its data model $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$ and its prior $\pi_i(\boldsymbol{\theta}_i)$ so that $Pr(\boldsymbol{y}|M_i)$ is associated with the density,

$$h_i(\boldsymbol{y}|M_i) = \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\, \pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i,$$

or, in the case of a hierarchical model,

$$h_i(\boldsymbol{y}|M_i) = \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\, g(\boldsymbol{\theta}_i|\boldsymbol{\lambda}_i)\, \pi_i(\boldsymbol{\lambda}_i)\, d\boldsymbol{\theta}_i\, d\boldsymbol{\lambda}_i.$$

The Bayes factor for models M_1 and M_2 is then often written in terms of these densities as,

$$BF(M_1, M_2) = \frac{h_1(\boldsymbol{y}|M_1)}{h_2(\boldsymbol{y}|M_2)}, \tag{13.8}$$

where the notation $h_i(\cdot);\ i = 1, 2$ has been used to emphasize that these densities are not necessarily of the same form.

Now, the Bayes factor of (13.8) is based on densities, not probabilities (cf. 13.7), but to have interpretation as developed from the posterior odds ratio of expression (13.6) we still need it to be true that,

$$Pr(M_1|\boldsymbol{y}) \quad = \quad \frac{Pr(M_1)h_1(\boldsymbol{y}|M_1)}{Pr(M_1)h_1(\boldsymbol{y}|M_1) + Pr(M_2)h_2(\boldsymbol{y}|M_2)}, \tag{13.9}$$

and similarly for $Pr(M_2|\boldsymbol{y})$. What is needed in order for (13.9) to hold? Essentially the same type of conditions as were needed in order for likelihoods to be proportional to the probability of data $\boldsymbol{y}$ for given parameters θ in Section 8.3.1 of Part II of these notes. Here, we extend the basic idea of using a linear approximation to the integral mean value theorem (what we actually depended on for likelihood is usually called the intermediate value theorem) from the case of independent variables to that of a joint density $h_i(\boldsymbol{y}|M_i)$.

Consider a joint density $h_i(\boldsymbol{y}|M_i)$ evaluated at the observed values $\boldsymbol{y}$. In order to apply the mean value or intermediate value theorems for multiple integrals, we can use the following:

(i) The value $\boldsymbol{y}$ implies that there exists an $n-$dimensional ball δ_y, centered at $\boldsymbol{y}$, such that for the random variable $\boldsymbol{Y}$ associated with the situation that led to $\boldsymbol{y}$,

$$Pr(\boldsymbol{Y} \in \delta_y) = \int_{\delta_y} h_i(\boldsymbol{t}|M_i)\, d\boldsymbol{t}.$$

(ii) The set δ_y is *connected* in the sense that any two points of δ_y can be joined by an arc that lies entirely in δ_y.

(iii) The density function $h_i(\cdot)$ is continuous on δ_y.

Under these conditions, and the assumption that δ_y is of sufficiently small volume, a similar argument as used to connect likelihood with probability gives that

$$\int_{\delta_y} h_i(\boldsymbol{t}|M_i)\, d\boldsymbol{t} \approx |\delta_y| h_i(\boldsymbol{y}|M_i), \tag{13.10}$$

or $Pr(\boldsymbol{y}|M_i) \propto h_i(\boldsymbol{y}|M_i)$. Now, the last thing needed in order to make (13.9) hold is then that the volume $|\delta_y|$ be the same for models M_1 and M_2. If this is the case, then (13.9) may be used on the right hand side of (13.6) to give,

$$\begin{aligned} \frac{Pr(M_1|\boldsymbol{y})}{Pr(M_2|\boldsymbol{y})} &= \frac{Pr(M_1)}{Pr(M_2)} \frac{h_1(\boldsymbol{y}|M_1)}{h_2(\boldsymbol{y}|M_2)} \\ &= \frac{Pr(M_1)}{Pr(M_2)} BF(M_1,\, M_2), \end{aligned}$$

with the Bayes factor BF defined as in expression (13.8). For the majority of models we deal with, the conditions needed to take the Bayes factor as a ratio of densities is not likely to raise any concern. But, if competing data models would happen to differ in functional form, then greater caution is needed.

Example 13.1

In the continuation of Example 5.2 contained in Section 7.3.1 of these notes we considered a number of possible models for analysis of a hypothetical study of the effect of violent cartoons and stories on aggression in children. The measured quantities were responses of "like", "do not like", or "don't care about" made to pictures shown to children who had watched or been read "violent" or "happy" cartoons or stories. In considering possible random variables that could be constructed for this setting, and possible distributions that might be assigned to those random variables, we discussed possible models based on binomial, multinomial, and beta distributions. Any of these models could be

analyzed through a Bayesian procedure. If we were to attempt a comparison of two of these models, say the beta model and the multinomial model, through a Bayes factor, the issue described immediately above would become important. Suppose, for example, that we took $\boldsymbol{y}$ to be the proportion of the total "score" possible over 20 pictures by making "don't like" correspond to 0, "don't care" correspond to 1 and "like" correspond to 2, then summing these scores across pictures for each child and dividing by 40 (the maximum "aggression score" possible).

We might then take $f_1(\boldsymbol{y}|\theta_1)$ to be beta with a (joint) prior for θ_1 formulated as a product of uniform (0, 1) densities as described in Example 12.5 in Chapter 12.4. Also suppose that we took $f_2(\boldsymbol{y}|\theta_2)$ to be multinomial with a (joint) prior for θ_2 given as a Dirichlet. It would be possible, of course, to derive $h_1(\boldsymbol{y}|M_1)$ from the beta model and $h_2(\boldsymbol{y}|M_2)$ from the multinomial model, and then just slap these distributions into expression (13.8) calling the result a Bayes factor. This would clearly be a mistake, and the formal reason for this is the above discussion.

Another potential difficulty with the use of Bayes factors occurs with the use of improper prior distributions. In this case it may be true that numerical values may be computed for the integrals preceding expression (13.8) (this will be the case whenever improper priors lead to proper posteriors) but the Bayes factor is nevertheless undefined, since these integrals cannot be considered proportional to $Pr(\boldsymbol{y}|M_i)$ as in expression (13.7). The distinction is that, in derivation of a posterior, we consider

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta}),$$

where the proportionality constant is the integral

$$\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Thus, so long as this integral is finite, the posterior exists as a proper distribution. For use in a Bayes factor, however, we need this same integral to result in a density for the argument $\boldsymbol{y}$ as,

$$h(\boldsymbol{y}) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

such that

$$\int_{-\infty}^{\infty} h(\boldsymbol{y})\, d\boldsymbol{y} = 1.$$

If $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a proper density with support Ω and $\pi(\boldsymbol{\theta})$ is an improper prior over Θ, the function $h(\cdot)$ cannot be a density with argument $\boldsymbol{y}$ because then,

$$\begin{aligned}
\int_{-\infty}^{\infty} h(\boldsymbol{y})\, d\boldsymbol{y} &= \int_{\Omega}\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\boldsymbol{y} \\
&= \int_{\Theta}\int_{\Omega} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{y}\, d\boldsymbol{\theta} \\
&= \int_{\Theta} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta},
\end{aligned}$$

which is not finite.

Despite the potential difficulties with Bayes Factors they can often be a useful method for comparing among models. In addition, for cases in which Bayes factors are defined, they may also be used to conduct what might be thought of as Bayesian tests of hypotheses.

Example 13.2

Consider again the analysis of sex ratio at birth in Guanacos of Example 11.1 discussed at some length in Chapter 11. With θ being defined as the probability of a male birth, we may formulate a hypothesis in this example of $H_0 : \theta \leq 0.5$ and an alternative hypothesis of $H_1 : \theta > 0.5$. Let these hypotheses correspond to the models M_2 and M_1, respectively. That is, model M_2 corresponds

to $\theta \leq 0.5$ while model M_1 corresponds to $\theta > 0.5$. Suppose that, *a priori* we give these two models equal weight, so that $Pr(M_1) = Pr(M_2) = 0.5$. Since, in this case, models correspond to values of θ, these hypotheses are naturally reflected in a uniform prior for θ on the interval (0, 1); there are other priors that could also reflect these hypotheses (anything with a density symmetric about 0.5 would suffice), but the uniform serves nicely in this case. The prior odds of M_1 to M_2 is then $0.5/0.5 = 1.0$. In this case, then, expression (13.6) indicates that the posterior odds of M_1 to M_2 become

$$\frac{Pr(M_1|\boldsymbol{y})}{Pr(M_2|\boldsymbol{y})} = BF(M_1, M_2).$$

Now,

$$Pr(M_1|\boldsymbol{y}) = Pr(\theta > 0.5|\boldsymbol{y}),$$

and,

$$Pr(M_2|\boldsymbol{y}) = Pr(\theta \leq 0.5|\boldsymbol{y}).$$

At the end of four years of data collection (1990) we had that, beginning with a uniform prior, the posterior distribution of θ was beta with parameters $\alpha = 185$ and $\beta = 161$ (see Chapter 11). Thus, the posterior odds, or Bayes factor, in favor of model M_1 are

$$\frac{Pr(\theta > 0.5|\boldsymbol{y})}{Pr(\theta \leq 0.5|\boldsymbol{y})} = \frac{Pr(\theta > 0.5|\alpha = 185, \beta = 161)}{Pr(\theta \leq 0.5|\alpha = 185, \beta = 161)}.$$

These values are easily computed for a beta distribution to be,

$$\frac{Pr(\theta > 0.5|\boldsymbol{y})}{Pr(\theta \leq 0.5|\boldsymbol{y})} = \frac{0.90188}{0.09812} = 9.1913.$$

Using typical scales for "strength of evidence" as discussed earlier in this Section we would conclude that there is some, but not strong, evidence against M_2 in favor of M_1, which agrees with our assessment from the credible intervals of Chapter 11.

13.4 Predictive Inference

It is often the case that we desire a predictive distribution for a "new" observation y^*, presumed to follow the same model as the components of $\boldsymbol{y}$. Such predictive distributions may be useful in their own right for the purposes of forecasting if y^* lies outside of the extent (or spatial and temporal window) of the available observations $\boldsymbol{y}$, or prediction if y^* lies within the extent of the available data but corresponds to a location or time or general "position" that is not observed in the data. In addition, predictive distributions may be useful in model assessment from the viewpoint that a good model predicts well. Quantifying the agreement of "replicated" data with actual data, where the replicated data are simulated from a predictive distribution is one way to accomplish this type of model assessment (e.g., Gelman, Carlin, Stern and Rubin 1995, Chapter 6.3).

The fundamental distribution used in predictive inference is the *posterior predictive* distribution $p(\boldsymbol{y}^*|\boldsymbol{y})$, which is in concert with the viewpoint of Bayesian analysis of unknowns in that inference about unobserved quantities are made on the basis of the conditional distributions of those quantities given the observed data. Consider either a simple model consisting of $f(\boldsymbol{y}|\boldsymbol{\theta})$ and the prior $\pi(\boldsymbol{\theta})$, or a hierarchical model in which we view the mixture $\int g(\boldsymbol{\theta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})\,d\boldsymbol{\lambda}$ to constitute a prior for $\boldsymbol{\theta}$ as $\pi_\theta(\boldsymbol{\theta})$. Using the notation $p(\cdot)$ as a generic probability density or mass function, and assuming that y^* is conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{\theta}$,

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}^*, \boldsymbol{y})}{p(\boldsymbol{y})} \\ &= \frac{1}{p(\boldsymbol{y})} \int_\Theta p(\boldsymbol{y}^*, \boldsymbol{y}, \boldsymbol{\theta})\,d\boldsymbol{\theta} \\ &= \frac{1}{p(\boldsymbol{y})} \int_\Theta p(\boldsymbol{y}^*|\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \end{aligned}$$

$$\begin{aligned} &= \frac{1}{p(\boldsymbol{y})} \int_{\Theta} p(\boldsymbol{y}^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}, \boldsymbol{y})\, d\boldsymbol{\theta} \\ &= \int_{\Theta} p(\boldsymbol{y}^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \end{aligned} \tag{13.11}$$

Conditional independence of $\boldsymbol{y}^*$ and $\boldsymbol{y}$ is used in the transition from line 2 to line 3 of expression (13.11). Notice that (13.11) indicates the posterior predictive density of $\boldsymbol{y}^*$ is the expected value of $p(\boldsymbol{y}^*|\boldsymbol{\theta})$, which will be of the same form as the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$, taken with respect to our knowledge about $\boldsymbol{\theta}$ as represented by the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. Now, reverting to the use of $f(\cdot)$ for the data model density and $p(\cdot|\boldsymbol{y})$ for a posterior density, the posterior predictive density of $\boldsymbol{y}^*$ may be written as,

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = \int_{\Theta} f(\boldsymbol{y}^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \tag{13.12}$$

In a hierarchical model it is also possible to derive a conditional posterior predictive density for $\boldsymbol{y}^*$ as,

$$p(\boldsymbol{y}^*|\boldsymbol{y}, \boldsymbol{\lambda}) = \int_{\Theta} f(\boldsymbol{y}^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})\, d\boldsymbol{\theta}. \tag{13.13}$$

Expression (13.12) is the same as expression (13.11) with notation to make the roles of the data model $f(\boldsymbol{y}^*|\boldsymbol{\theta})$ and posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ explicit. Expression (13.13) can be derived along the lines of (13.11), leading to the conditional posterior predictive distribution in notation parallel to that of the ordinary or marginal posterior predictive density (13.12). While the predictive density (13.13) is certainly less common in typical Bayesian analysis, it does appear on occasion, for example in the dynamic models of West and Harrison (1989). To make use of the conditional distribution (13.13) one would need a plug-in value for $\boldsymbol{\lambda}$ in the same way that this is required for use of the conditional posterior $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$ of expression (10.7).

Chapter 14

Simulation of Posterior Distributions

The basic idea of simulation of a posterior distribution is quite simple. Suppose that we have a model that consists of a data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and a prior $\pi(\boldsymbol{\theta})$. The posterior is

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \end{aligned} \tag{14.1}$$

Note that both the left and right hand sides of this expression must be considered as functions of $\boldsymbol{\theta}$. Suppose that the integral in the denominator of the first line of (14.1) is intractable so that the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ cannot be derived in closed form. A typical goal in simulation of posteriors is then to use the last line of expression (14.1) to allow simulation of values from $p(\boldsymbol{\theta}|\boldsymbol{y})$ even in such cases. In a hierarchical model we have a data model $f(\boldsymbol{y}|\boldsymbol{\theta})$, an intermediate distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, which can be considered either a part of

the model as a mixing distribution or a part of the prior, depending on the viewpoint taken, (see Chapter 10) and a prior or final stage prior $\pi(\boldsymbol{\lambda})$. For this type of model with $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\lambda} \in \Lambda$, we may have interest in the joint posterior,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})}{\int_\Lambda \int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda}). \end{aligned} \tag{14.2}$$

Again, the integrals in the first line of (14.2) may be intractable, and a common goal is to use the last line of (14.2) to allow simulation from $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y})$ without explicitly evaluating these integrals. There are several ways that simulation from the posteriors (14.1) and (14.2) can be accomplished, but before going into these methods it is useful to identify several basic principles of simulation.

14.1 Fundamental Principles of Simulation

It is useful to set forth several basic truths of simulation procedures, which we do in this section, using generic notation for random variables X, Y, and Z and their density functions as $f(x)$, $f(y)$ and $f(z)$.

1. Averaging over Simulations Estimates Expectations

   This first principle embodies the fundamental idea of Monte Carlo estimation, which we have already encountered in Chapter 8.6 in discussing Parametric Bootstrap methods. Consider first the case of a univariate random variable X with distribution function $F(x)$. If we obtain simulated values $\{x_j^* : j = 1, \ldots, M\}$ as independent and identical realizations

from $F(x)$, a Monte Carlo estimate of the expected value $E(X)$ is,

$$\hat{E}_M(X) = \frac{1}{M}\sum_{j=1}^{M} x_j^*.$$

Indeed, for any suitable function $q(X)$,

$$\hat{E}_M\{q(X)\} = \frac{1}{M}\sum_{j=1}^{M} q(x_j^*). \tag{14.3}$$

Thus, simulation and averaging accomplishes estimation of the integral

$$E\{q(X)\} = \int q(x)\, dF(x),$$

where it is assumed that the dominating measure is Lebesgue or counting measure. That $\hat{E}_M\{q(X)\}$ is consistent for $E\{q(X)\}$ follows immediately from the law of large numbers. Similarly, the empirical distribution function of the values $\{x_j^* : j = 1, \ldots, M\}$,

$$F_M(x) = \frac{1}{M}\sum_{j=1}^{M} I(x_j^* \leq x),$$

converges to $F(x)$ for each fixed x as $M \to \infty$. The Gilvenko-Cantelli Theorem gives that this convergence is uniform in x (e.g., Billingsley, 1986, p. 275).

These results, which form the basis for Monte Carlo estimation, may be extended to sequences of random variables that are not independent, if such sequences have a property called *ergodicity*. A complete coverage of ergodicity is beyond the scope of these notes, but an intuitive understanding of the fundamental idea can be gained as follows. Consider a distribution $F(x)$; $x \in \Omega$ from which we would like to simulate a sample $\{x_j^* : j = 1, \ldots, M\}$ so that the above Monte Carlo estimators $\hat{E}_M\{q(X)\}$ and $F_M(x)$ converge to $E\{q(X)\}$ and $F(x)$ as in the case of independent

realizations. Now, suppose we are unable to simulate values from $F(x)$ directly, but we are able to construct a sequence of random variables $\boldsymbol{X}(t) \equiv \{X(t) : t = 0, 1, \ldots, \}$ called a *chain* in such a way that the above results continue to hold using values simulated from $\boldsymbol{X}(t)$ rather than $F(x)$. This can only occur, for dependent $X(t)$, if the sequence "mixes" over the set Ω in the proper manner. Suppose we partition Ω into and arbitrary number of k subsets, $\Omega_1, \ldots, \Omega_k$. Suppose further that $\{X(t) : t = 0, 1, \ldots, \}$ has the property that for some value B and $t > B$, the relative frequencies with which $X(t) \in \Omega_k$ for each k converge to the probabilities dictated by F (as $t \to \infty$). If this is true for all arbitrary partitions of Ω, then the results desired will continue to hold using $\{x^*(t) : t = B, B + 1, \ldots B + M\}$ in place of $\{x_j^* : j = 1, \ldots, M\}$. What is needed, then, is for the sequence $X(t)$ to "visit" or "mix over" each of the subsets $\Omega_1, \ldots, \Omega_k$ with the correct frequencies, and with sufficient rapidity that we don't have to wait until M becomes too large for the approximations to be adequate. Sequences $X(t)$ have have these behaviors are called ergodic.

The construction of what are called *Markov chain samplers* is concerned with developing chains that are ergodic and, importantly, mimic the probabilistic behavior of the distribution $F(x)$, simulation from which was goal in the first place. In this context, $F(x)$ is often called the "target" distribution. We will encounter several Markov chain samplers in Chapters 14.4 and 14.5.

2. Expectations wrt to F Can be Estimated by Simulating from G.
   Consider basic Monte Carlo estimation of $E\{q(X)\}$ as described in the first part of item 1 above. There, it was assumed that values $\{x_j^* : j =$

$1, \ldots, M\}$ had been simulated from $F(x)$, the distribution of X. Assume that $F(x)$ has a corresponding density function $f(x)$. The expected value of $q(X)$ may than also be written as,

$$E\{q(X)\} = \int q(x)\frac{f(x)}{g(x)}g(x)\,d\mu(x),$$

for some function $g(\cdot)$ with domain matching the support Ω of $f(\cdot)$, and μ either Lebesgue or counting measure. If $g(\cdot)$ corresponds to a density over Ω, then a Monte Carlo estimate of the expectation can be computed as

$$\hat{E}_M\{q(X)\} = \frac{1}{M}\sum_{j=1}^{M} q(x_j^*)\,\frac{f(x_j^*)}{g(x_j^*)}, \tag{14.4}$$

where $\{x_j^* : j = 1, \ldots, M\}$ have been simulated from the distribution with density $g(x)$; $x \in \Omega$. Expression (14.4) is called an *importance sampling* estimator of $E\{q(X)\}$, and $g(x)$ is the importance (or importance sampling) distribution. While importance sampling had its origins in techniques useful to reduce the variance of Monte Carlo estimators, in most statistical applications it appears in one of two situations. First, notice that importance sampling can be used to formulate a Monte Carlo estimator of the integral of nearly any function, regardless of whether that integral is originally expressed as an expectation or not. That is, the integral of a function $h(x)$ can be expressed as the expectation of $h(x)/g(x)$ with respect to a distribution having density $g(\cdot)$ as long as the support of g matches the limits of the original integration. Secondly, and perhaps most commonly, importance sampling is useful in situations where the distribution with density $f(\cdot)$ in (14.4) is difficult to simulate from, but for which we can find an importance distribution with density $g(\cdot)$, with the same support, but from which it is easy to simulate. A key to the successful application of importance sampling, however, is not only

finding an importance distribution from which it is easy to sample, but also one that is "good" in the sense that it results in rapid convergence of $\hat{E}_M\{q(X)\}$ to $E\{q(X)\}$.

3. Successive Simulation of Marginal and Conditional Distributions Accomplishes Integration.

   For two random variables X and Y with marginal and conditional densities $f(x)$ and $f(y|x)$, simulation of a value x^* from $f(x)$ followed by simulation of a value y^* from $f(y|x^*)$, gives one value y^* from the marginal density

   $$f(y) = \int f(y|x)f(x)\,dx. \tag{14.5}$$

   Thus, if we want a set of simulated values $\{y_j^* : j = 1, \ldots, M\}$ from the distribution with density $f(y)$, we simulate M values $\{x_j^* : j = 1, \ldots, M\}$ independently from $f(x)$ and, for each of these values, simulate a value y_j^* from $f(y|x_j^*)$.

   Similarly, simulation of one value z^* from $f(z^*)$, followed by simulation of one value x^* from $f(x|z^*)$, followed in turn by one value y^* simulated from $f(y|x^*)$ produces one value y^* from the distribution,

   $$f(y) = \int\int f(y|x)f(x|z)f(z)\,dx\,dz, \tag{14.6}$$

   where a crucial assumption is that $f(y|x, z) = f(y|x)$, that is, y depends on z only through its effect on x. To simulate a set of M values we simulate $\{z_j^* : j = 1, \ldots, M\}$ from $f(z)$, simulate $\{x_j^* : j = 1, \ldots, M\}$ from $f(x|z_j^*)$, and then simulate $\{y_j^* : j = 1, \ldots, M\}$ from $f(y|x_j^*)$.

   This principle of simulation perhaps comes into play directly most often in simulating values from a given model, but it can also be useful in situations for which it is possible to derive in closed form marginal and

conditional posteriors $p(\boldsymbol{\lambda}|\boldsymbol{y})$ and $p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{y})$. In particular, for scalar θ and λ, the above prescription can be used to simulate values from $p(\theta|\boldsymbol{y})$ by simulating λ^* from $p(\lambda|\boldsymbol{y})$ and then θ^* from $p(\theta|\lambda^*, \boldsymbol{y})$. Here, $p(\lambda|\boldsymbol{y})$ plays the role of $f(x)$ in (14.5) while $p(\theta|\lambda, \boldsymbol{y})$ plays the role of $f(y|x)$.

4. Simulation of Joint Distributions Accomplishes Simulation of Marginal Distributions.

   Consider a joint distribution $F(x, y, z)$ for random variables X, Y, and Z. Suppose we are able to simulate the values $\{(x_j^*, y_j^*, z_j^*) : j = 1, \ldots, M\}$ from this joint distribution. Then, an estimator of $E\{q(X)\}$, for example, would be the same as given in expression (14.3) using only the values $\{x_j^* : j = 1, \ldots, M\}$ and ignoring the y_j^* and z_j^* values. Estimators for the expected values of functions of Y and Z would be formed in a similar manner. In fact, the empirical distribution of the values $\{x_j^* : j = 1, \ldots, M\}$ would approximate the true marginal distribution $F_X(x)$ say, and the same would be true for the empirical distributions of the values $\{y_j^* : j = 1, \ldots, M\}$ and $\{z_j^* : j = 1, \ldots, M\}$ as estimators of the marginals $F_Y(y)$ and $F_Z(z)$. Thus, if we are able to simulate from a joint distribution we have also accomplished simulation from any of the marginals.

   This principle of simulation comes into play in that if, for a model with multiple parameter elements, say $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)^T$, simulation from $p(\boldsymbol{\theta}|\boldsymbol{y})$ also provides simulated values from $p(\theta_j|\boldsymbol{y})$ for $j = 1, \ldots, p$. Similarly, if we are able to simulate from the joint posterior $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y})$ from a hierarchical model, then we have also simulated from the marginal posteriors $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $p(\boldsymbol{\lambda}|\boldsymbol{y})$ and, by what is immediately above, also the marginals of any of the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$.

14.2 Basic Methods of Simulation

14.2.1 Inversion

14.2.2 Composition

14.2.3 Basic Rejection Sampling

14.2.4 Ratio of Uniforms

14.2.5 Adaptive Rejection Sampling

14.3 The Method of Successive Substitution

14.4 The Gibbs Sampler

14.5 Metropolis Hastings