Chapter 3

Useful Classes of Distributions

To adequately model the probabilistic behaviors of random variables, we must have access to a variety of theoretical probability distributions. The number of named distributions is large and there are distributions that have been created for specific purposes (see, e.g., the series of works edited by Johnson and Kotz). As discussed in most courses on mathematical statistics, a number of distributions have been derived as sampling distributions, used to construct tests and confidence intervals. There has also been extensive work to develop systems of distributions that include distributions that cover a range of characteristics such as skewness, including Pearson's system and Burr's system (reference needed?) but full coverage of these is beyond the scope of this book. Rather, we will present two *classes* of distributions that include many of the distributions used in applications. First, although this terminology is not followed by all statisticians, we distinguish between classes and families of distributions. Consider the probability mass or density function $f(x|\boldsymbol{\theta})$ with support $x \in \Omega_x$, and where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is a parameter such that $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. As $\boldsymbol{\theta}$ varies over its parameter space $\boldsymbol{\Theta}$ we say that $f(x|\boldsymbol{\theta})$ generates a family of distribu-

tions. Thus, what we usually refer to as the normal distribution constitutes a family of distributions, as does the Poisson distribution and many others. Collections of different families of distributions may be grouped into classes of distributions. The two classes of distributions we will discuss in this chapter are the classes of location-scale families and exponential families.

Much of this chapter will be presented in the context of a single random variable. 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. We will desire this set to also correspond to the support of a distribution assigned to the random variable.

- 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 nonrandom quantities such as

parameters or covariates. The other will be in the usual conditioning notation for two or more random variables. It is important to understand the context being used in a conditional statement so that, for example, the distinction between $E(Y|x)$ and $E(Y|X)$ is clear.

3.1 Location-Scale Families

This section briefly discusses a portion of the larger topic of *group* (Lehmann, 1983, e.g.,) or *transformation* (Lindsey, 1996, e.g.,) families of distributions. In particular, we will restrict attention to the class of families resulting from location and scale transformations.

Let U be a random variable with a fixed distribution F. If U is transformed into Y as,

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

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. 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

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

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,

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

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 can sometimes be difficult to grasp. For example, a location family can be generated starting with nearly any distribution but that does not mean the starting distribution constitutes a location family.

Example 3.1

Let X be a random variable following an exponential distribution with probability density function, for some $\beta > 0$,

$$f_x(x|\beta) = \beta \exp(-\beta x); \quad x > 0.$$

Now let $Y = \mu + X$ for $-\infty < \mu < \infty$. The density of Y is

$$f_Y(y|\beta,\mu) = \beta \exp[-\beta(y-\mu)]; \quad y > \mu$$

which is a legitimate distribution but is not an exponential distribution. To verify this, note that the moment generating function of X is, for $t < \beta$,

$$M_x(t) = \frac{\beta}{\beta - t}$$

while that of Y is, also for $t < \beta$,

$$M_Y(t) = \exp(\mu t)\frac{\beta}{\beta - t},$$

and these two moment generating functions are not of the same form. Any additional location transformations such as $Z = \psi + Y$ result in a distribution with moment generating function, for $t < \beta$,

$$M_Z(t) = \exp[(\psi + \mu)t]\frac{\beta}{\beta - t},$$

which has the same form as $M_Y(t)$. Thus, the exponential distribution can generate a location family of distributions, but is not itself a location family.

3.1.1 Properties of Location-Scale Families

Location-scale families have simple properties that stem directly from the transformations. For example, if Y is produced as a location-scale transformation of U, $Y = \mu + \sigma U$, then $E(Y) = \mu + \sigma E(U)$ and $\text{var}(Y) = \sigma^2 \text{var}(U)$. Traditionally, if $E(U) = 0$ and $\text{var}(U) = 1$ the the distribution of U is called the *parent* distribution for the family. This may not be the best terminology, since we must be able to arrive at any member from any other member through the same family of transformations (Lehmann, 1983, p. 25). What can be called the *standard* form of a distribution is the distribution that results from eliminating parameters. While this does often result from taking the distribution that has expected value 0 and variance 1, such as the standard normal, that is not always the case. One type of extreme value distribution, for example, has a standard form,

$$f(x) = \exp(-x)\, \exp[-\exp(-x)]; \quad -\infty < x < \infty, \tag{3.1}$$

which contains no parameters but has expected value given by Euler's constant (0.5772, to the first four decimal places) and variance $\pi^2/6$. This distribution does constitute a location-scale family but the location parameter is equal to the mode of the distribution and the scale parameter is $6/\pi^2$ times the variance. Nevertheless, location-scale families of distributions that have standard forms with expectation 0, variance 1, and support on the entire line are the traditional building blocks for models formulated in terms of what we will come to call a *signal plus noise* structure. A location scale transformation of a random variable X with probability density (3.1), $Y = \xi + \phi X$ has density, with

$-\infty < \xi < \infty$ and $0 < \phi$,

$$g(y|\xi,\phi) = \frac{1}{\phi}\exp\left[-\left(\frac{y-\xi}{\phi}\right)\right]\exp\left[-\exp\left\{-\left(\frac{y-\xi}{\phi}\right)\right\}\right]; \quad -\infty < y < \infty. \tag{3.2}$$

The mode of this distribution is ξ, the expected value is about $\xi + 0.5772$ and the variance is $\phi^2\,\pi^2/6$.

3.1.2 Location-Scale Error Distributions

Following directly from the end of the previous subsection, we are familiar with models such as linear regression and analysis of variance in which a set of independent response random variables $\{Y_i : i = 1, \ldots, n\}$ are modeled as, for some $n \times p$ design matrix $\boldsymbol{X}$ and parameter $\boldsymbol{\beta} \in \mathbb{R}^p$

$$Y_i = \boldsymbol{x}_i^T\,\boldsymbol{\beta} + \sigma\epsilon_i, \tag{3.3}$$

where the ϵ_i are independent and identically distributed according to a location-scale family with $E(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = 1$. The response variables then arise as the result of location-scale transformations of these error random variables. The location transformations are specific to individual response variables or perhaps groups of variables, depending on the structure of the covariate vectors $\boldsymbol{x}_i$. The scale transformations are all identical, at least in model (3.3). In this model we may think of response variables as being structured by the expectation function $\boldsymbol{x}_i^T\boldsymbol{\beta}$ or *signal* and additive errors $\sigma\epsilon_i$ or *noise*. We may attempt to avoid designating a specific distribution for the additive errors in developing estimators and some of their properties, such as ordinary least squares estimators of $\boldsymbol{\beta}$ to which we attach the Gauss-Markov theorem. But eventually when it comes time to produce inferential statements we nearly always resort to adding the assignment of normal distributions to the ϵ_i. The response variables

then have normal distributions with equal variances, least squares estimators of the elements of $\boldsymbol{\beta}$ are linear combinations of these normally distributed responses and thus also have normal distributions with variances that depend on only the one unknown parameter σ, and inferential statements follow from exact or small-sample theory. We could, of course, specify some other location-scale distributional family for the error terms such as a logistic distribution or the extreme value distribution of the previous subsection. We would then loose the ability to make inferences on the basis of exact theory, but we might improve our ability to model the spread of responses about their expectations. It is useful to contemplate when we might and when we might not want to consider using something other than a normal distribution for the ϵ_i in (3.3). If our only interest is in making inference about the expectation function $\boldsymbol{x}_i\boldsymbol{\beta}$ across values of the covariates, then there would seem to be little motivation to consider anything other than a normal error specification. With the normal distributional assignment we obtain a strong body of results on which to base inference. But in many problems, making inference only about expectations falls short of what is desired. We will see examples of simple regression settings (only one type of covariate) for which demonstrating that responses increase or decrease with covariate values is not even in question, such as tree volume as a response and tree diameter as the covariate. But we may desire estimation of certain quantiles for responses, or in the probability that a response will exceed some regulatory threshold at a given covariate value, and those questions depend on more than expectations alone. So if we have interest in distributional characteristics of response variables other than expectation (or location) alone, we may be well served by considering distributions for the ϵ_i of model (3.3) other than normal distributions.

3.2 Exponential Families

The class of exponential families of distributions constitutes an essential set of distributions for modeling purposes. There are various equivalent ways to write what is called the exponential family form. For a random variable Y having possible values in a set Ω and corresponding probability density function (pdf) or probability mass function (pmf) some of these representations are, all for $y \in \Omega$ and 0 otherwise:

$$\begin{aligned} f(y|\boldsymbol{\eta}) &= \exp\left\{\sum_{j=1}^{s} q_j(\boldsymbol{\eta}) T_j(y)\right\} c(\eta) h(y), \\ f(y|\boldsymbol{\theta}) &= a(\boldsymbol{\theta})\, g(y) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{t}(y)\right\}, \\ f(y|\boldsymbol{\eta}) &= \exp\left\{\sum_{j=1}^{s} q_j(\boldsymbol{\eta}) T_j(y) - B(\boldsymbol{\eta})\right\} c(y) \\ f(y|\boldsymbol{\theta}) &= \exp\left\{\sum_{j=1}^{s} \theta_j T_j(y) - B(\boldsymbol{\theta}) + c(y)\right\}. \end{aligned} \tag{3.4}$$

Note that, while $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^T$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_s)^T$ and $\boldsymbol{t}(y) = (t_1(y), \ldots, t_s(y))^T$ may be vectors, $B(\boldsymbol{\theta})$, $B(\boldsymbol{\eta})$, $h(y)$, $g(y)$, $c(y)$, and $a(\boldsymbol{\theta})$ are real-valued functions. The definition of functions such as $B(\cdot)$, $c(\cdot)$, $a(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are not exactly the same in these various expressions so that, for example, $c(y)$ is not the same function in the third and fourth lines of (3.4), but the equivalences are not difficult to work out.

Example 3.2

If Y is a random variable such that $Y \sim N(\mu, \sigma^2)$, the fourth version of the exponential family given in (3.4) 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,

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

We will use the fourth (last) expression in (3.4) as our basic form for exponential family representation. The term $\exp\{c(y)\}$ in the last expression of (3.4) 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. Other common densities or mass functions that can be written this way include,

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

3.2.1 Properties of Exponential Families

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

1. The parameter space $\boldsymbol{\Theta}$ (the set of points such that $f(y|\boldsymbol{\theta})$ is a density or mass function for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$) is a convex set. Recall that K is a convex set if, for $x, y \in K$, $\lambda x + (1-\lambda)y \in K$ for all $0 \leq \lambda \leq 1$, that is, the line segment joining x and y lies entirely within K.

2. To avoid difficulties, we will consider only members of exponential families such that neither the $T_j(y)$ nor the θ_j satisfy a linear constraint. In this case the representation is said to be *minimal* or sometimes *full*. If $\boldsymbol{\Theta}$ 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 often want exponential families to be written so that they are minimal and regular which is assumed in many theoretical results we wish to use for estimation and inference. For example, a multinomial with H categories will only be minimal if we write the pmf for $H-1$ random variables.

3. For a minimal, regular exponential family, the statistic $\boldsymbol{T} \equiv (T_1, \ldots, T_s)$ is minimal sufficient for $\boldsymbol{\theta}$. 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.

4. For an integrable function $h(\cdot)$, dominating measure ν, and any $\boldsymbol{\theta}$ in the interior of $\boldsymbol{\Theta}$, the integral
$$\int h(y) \exp\left\{\sum_{j=1}^{s} \theta_j T_j(y) - B(\boldsymbol{\theta}) + c(y)\right\} d\nu(y)$$
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., Lehmann, 1983, Theorem 4.1). This property does several things for us. First, it can be used to derive additional properties of

exponential families such as the form of the moment generating function in property 6. 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.

5. Property 4 can be used to show that (e.g., Lehmann, 1983),

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

These lead directly to $E(Y)$ and $var(Y)$ for what are called *natural exponential* families and *exponential dispersion* families, which will be discussed in the sequel. They also will provide an alternative parameterization of exponential families in general.

6. The moment generating function of an exponential family is defined to be that for the moments of the T_js and may be derived to be,

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

3.2.2 Parameterizations

In the final expression of (3.4) the parameters denoted as $\theta_j;\ \ j = 1, \ldots, s$ 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 those just given, it is not always the best parameterization for purposes of estimation, inference, or model interpretation. Parameter transformations are simple substitutions in density and mass functions, and it several parameterizations other than the canonical form are often useful.

We will describe two parameterizations here that have both been called *mean value* parameterizations, although they are not necessarily 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 values of response variables are usually of interest, and are often needed to quantify other characteristics in a concise manner. It may be the case that none of the canonical parameters θ_j in (3.4) correspond to the expected value of the random variable Y. 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, as defined previously. Note, however, that the reparametrized family may no longer be in canonical form.

Example 3.3

Consider a beta random variable Y with pdf,

$$f(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \tag{3.5}$$

where $\Omega = (0,1)$, $f(y|\alpha,\beta) = 0$ if $y \notin \Omega$, and $\alpha, \beta > 0$. We know for this density that $E(Y) = \alpha/(\alpha+\beta)$. First, write the density in canonical exponential family form as,

$$\begin{aligned} f(y|\theta) &= \exp\left[\theta_1 \log(y) + \theta_2 \log(1-y) + \log\{\Gamma(\theta_1+\theta_2)\}\right. \\ &- \left.\log\{\Gamma(\theta_1)\} - \log\{\Gamma(\theta_2)\} - \log(y) - \log(1-y)\right], \end{aligned} \tag{3.6}$$

where $\theta_1 = \alpha$ and $\theta_2 = \beta$. In terms of the last expression of (3.4), $T_1(y) = \log(y)$, $T_2(y) = 1 - \log(y)$, $B(\boldsymbol{\theta}) = \log\{\Gamma(\theta_1)\} + \log\{\Gamma(\theta_2)\} - \log\{\Gamma(\theta_1+\theta_2)\}$,

and $c(y) = -\log(y) - \log(1-y)$. We can achieve a mean value parameterization by taking,

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

We can then write the density in mean value parameterization by substituting into (3.6) the quantities

$$\theta_1 = \phi\mu; \quad \theta_2 = \phi(1-\mu).$$

Notice that we have not manipulated Y in any way, so that Ω remains unchanged throughout.

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 in property 5 of the previous section as first derivatives of the function $B(\theta)$. Thus, we transform $(\theta_1, \ldots, \theta_s) \to (\mu_1(\theta), \ldots, \mu_s(\theta))$ 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 an element of the complete sufficient statistic, namely $T_j(Y)$, which then immediately give us UMVU estimators for the parameters $\mu_j(\theta)$. The relevant question is whether such parameters represent quantities that are meaningful for inference.

Example 3.4

From example 3.1, 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 the expected values of $T_1(Y) = Y$ and $T_2(Y) = Y^2$. Notice for this example that the 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 $T_1(Y) = Y$. Families with this structure are among the more commonly used distributions in many types of models such as generalized linear models.

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 $\boldsymbol{\theta} = (\mu_1(\theta), \theta_2)$. One does not see such parameterizations used a great deal, but they apparently (Lindsey, 1996, p. 29) have the intriguing property of *variation independent* parameters. For a parameter $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \boldsymbol{\Theta}$, $\theta_1 \in \Theta_1$ and $\theta_2 \in_2$ are variation independent if $\boldsymbol{\Theta} = \theta_1 \times \theta_2$.

Uses of Various Parameterizations

As seen in Example 3.4, 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 (due to invariance of the likelihood function) or posterior simulation in Bayesian analysis. 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 sometimes 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 statisticians with extensive consulting experience will point out, most scientists do not think in terms of statistical models. It can be difficult 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 is to sometimes clearly identify how covariate information can appropriately be incorporated into a model. Our natural inclination is for covariate values to influence the marginal expectations of random variables. 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 3.2, 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 (14.6) or with canonical parameterization as in expression (3.6). 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 $\boldsymbol{\theta} \equiv (\theta_1, \theta_2)$ 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. While having variation independent parameters is probably typical of distributions we are familiar with, it is worth noting this property. In models that have multiple random components, such as hierarchical models, variation independent parameters in the data model translate into something called the *positivity* condition for modeling random parameter values. Formulating a proper model can become much more difficult when this condition does not hold.

3.2.3 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 one of the most common forms of exponential family distributions that appear in applications.

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* fol-

lowing the extensive investigation of their behavior by Morris (1982, 1983). If a family of distributions has only one canonical parameter, then both the expectations and variances of those distributions must be functions of the sole parameter.

Example 3.5

Consider the exponential form of a binomial random variable Y for a fixed number of associated binary trials n. Letting $p = Pr(Y = 1)$, the pmf of such a random variable is,

$$\begin{aligned} f(y|\theta) &= \exp\left[y\{\log(p) - \log(1-p)\} + n\{\log(1-p)\}\right. \\ &+ \left.\log\{n!\} - \log\{y!\} - \log\{(n-y)!\}\right] \\ &= \exp\{y\theta - b(\theta) + c(y)\};\ y = 0, 1, \ldots, n, \end{aligned}$$

where $\theta = \log\{p/(1-p)\}$, $b(\theta) = n\log\{1 + \exp(\theta)\}$ and $c(y) = log(n!) - \log(y!) - \log\{(n-y)!\}$. Note here that the parameter space of p is (0, 1) while that of θ is $(-\infty, \infty)$. Here, using the facts that $T(y) = y$ and $b(\theta)$ is a simple function, property 4 of canonical exponential families given previously implies that

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

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 3.6

Consider a random variable $Y \sim N(\mu, \sigma_*^2)$ for which σ_*^2 is considered a fixed, known value. In this case we can write, for $-\infty < \mu < \infty$ and $0 < \sigma^2$,

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

Letting $\theta = \mu$, $b(\theta) = (1/2)\theta^2$, $\phi = 1/\sigma_*^2$, and $c(y, \phi) = (1/2)[y/\sigma_*^2 - \log)2\pi\sigma_*^2)]$ this density may be written as what is called an *exponential dispersion family* which has the general form of,

$$f(y|\theta, \phi) = \exp\left[\phi\{y\theta - b(\theta)\} + c(y, \phi)\right]. \tag{3.7}$$

For a distribution with pdf or pmf of the form (3.7) the properties of $s-$parameter exponential families may be used to demonstrate that,

$$\begin{aligned} E(Y) &= \frac{d}{d\theta}b(\theta) = b'(\theta), \\ var(Y) &= \frac{1}{\phi}\frac{d^2}{d\theta^2}b(\theta) = \frac{1}{\phi}b''(\theta) = \frac{1}{\phi}V(\mu). \end{aligned} \tag{3.8}$$

The rightmost portion of the expression for $var(Y)$ in (3.8) follows from the fact that $\mu = b'(\theta)$ so that $b''(\theta)$ is a function of μ. The function $V(\cdot)$ in (3.8) is often called the *variance function*, which is not the variance except for a few cases in which $\phi = 1$. The variance function is important because it quantifies the relation between the mean and variance of the distribution.

Comments

1. What has happened in (3.7) is that we have coerced a two parameter exponential family to look almost like a natural exponential family (see

Example 3.5) but with the addition of an extra parameter ϕ called the *dispersion parameter*. This parameter is a scale factor for the variance (3.8).

2. Clearly, it will not be possible to write an exponential family in the form of expression (3.7) 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 (3.7), distributions for which one of the sufficient statistics is y and which can subsequently be written in exponential dispersion family form include the binomial, Poisson, normal, gamma, and inverse Gaussian. But it is not possible, for example, to write a beta pdf in this form.

3. Exponential dispersion families of the form (3.7) are the exponential families upon which *generalized linear models* are based (e.g., McCullagh and Nelder, 1989) but, as discussed in Chapter 1, 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.

3.2.4 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 (e.g., Arnold and Strauss, 1991; Kaiser and Cresie, 2000) here we will consider the situation only for sets of independent random variables, that is, random samples. Recall from Chapter 1 that a statistical

model must result in a joint distribution for the entire set of random variables involved in a problem.

One additional property of exponential families will be useful in this subsection. For Y distributed according to an s-parameter exponential family as in (3.4) with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$, the sufficient statistic $\boldsymbol{T}(y) = (T_1(Y), \ldots, T_s(Y))$ is distributed according to an exponential family with density or mass function

$$g(\boldsymbol{t}|\boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{s} \theta_j t_j - B(\boldsymbol{\theta}) + k(\boldsymbol{t})\right]. \tag{3.9}$$

Note that the dominating measure of the distributions of Y and $\boldsymbol{T}$ may differ, and that $k(\boldsymbol{t})$ may or may not be easily derived from the original $c(y)$, but $\boldsymbol{\theta}$ and $B(\boldsymbol{\theta})$ are the same as for the original distributions $f_Y(y|\boldsymbol{\theta})$.

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

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

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

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{s} \theta_j \sum_{i=1}^{n} T_j(y_i) - n\, B(\boldsymbol{\theta}) + \sum_{i=1}^{n} c(y_i)\right\}. \tag{3.10}$$

Notice that expression (3.10) 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 3.7

Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed (iid) fol-

lowing a common Poisson distribution with pmf, for some $\lambda > 0$,

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

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

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

where $b(\theta) = \exp(\theta)$ and $c(y) = -\log(y!)$. 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 that, using the properties of exponential families provided previously, we 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 $W = \sum_{i=1}^{n} Y_i$ 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 2.1.1 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.

Chapter 5

Basic Likelihood Estimation and Inference

This chapter is intended to be summary of fundamental likelihood-based estimation and inference. It will also serve as a convenient reference source, with information on basic likelihood collected in one place.

5.1 Likelihood Functions

Recall from Chapter 1 that a statistical model must lead to a joint probability distribution for the entire collection of random variables involved in a problem. We will assume here that this is accomplished in the form of a joint probability mass or probability density function. Although some models benefit from multiple subscripting, it is always possible to arrange for random variables to have a single subscript and then use indicator functions to define groups and so forth. Thus, we will assume here that we have a set of random variables $\{Y_i : i = 1, \ldots, n\}$ having possible values in a set Ω_Y and with parameterized

joint probability mass or probability density function, for $\boldsymbol{\theta} \in \Theta$,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = f(y_1, \ldots, y_n|\boldsymbol{\theta}); \quad \boldsymbol{y} \in \Omega_Y,$$

and where $f(\boldsymbol{y}|\boldsymbol{\theta}) = 0$ for any $\boldsymbol{y} \notin \Omega_Y$. Notice that we have made special note of both the support of $f(\cdot)$, Ω_Y, and the parameter space $\boldsymbol{\theta} \in \Theta$. Often, the set of possible values, matching the support of the joint distribution, will satisfy what is called the *positivity* condition. Let the possible values or support of the marginal distribution of Y_i be denoted as Ω_i for $i = 1, \ldots, n$. The positivity condition is that

$$\Omega_Y = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n. \tag{5.1}$$

The positivity condition states that, if a random variable can assume a given value, it can assume that value in combination with any other set of values for the other random variables involved in the problem, that is, there are no forbidden states in the possible joint configurations.

The likelihood function is defined as a function of the parameter $\boldsymbol{\theta}$ that is given by the same formula as the joint mass or density function, and the log likelihood is the logarithm of the likelihood,

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y}) &= f(\boldsymbol{y}|\boldsymbol{\theta}); \quad \boldsymbol{\theta} \in \Theta \\ \ell(\boldsymbol{\theta}|\boldsymbol{y}) &= \log[L(\boldsymbol{\theta}|\boldsymbol{y})]; \quad \boldsymbol{\theta} \in \Theta. \end{aligned} \tag{5.2}$$

If the random variables $Y_1, \ldots, Y_n$ are independent with probability mass or density functions $f_i(y_i|\boldsymbol{\theta})$; $y_i \in \Omega_i$, then the likelihood and log likelihood can be expressed as

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y}) &= \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}); \quad \boldsymbol{\theta} \in \Theta \\ \ell(\boldsymbol{\theta}|\boldsymbol{y}) &= \sum_{i=1}^{n} \log[f_i(y_i|\boldsymbol{\theta})]; \quad \boldsymbol{\theta} \in \Theta \end{aligned} \tag{5.3}$$

Expression (5.2) gives usual definitions of likelihood and log likelihood functions that we see in textbooks. When the context is that of a given set of observed data, we often write the likelihood without explicit conditioning on those data as simply $L(\boldsymbol{\theta})$ or $\ell(\boldsymbol{\theta})$. On the other hand, if the context involves probabilistic behavior of the likelihood we may write the random version as $L(\boldsymbol{\theta}|\boldsymbol{Y})$ or $\ell(\boldsymbol{\theta}|\boldsymbol{Y})$. This is analogous to notation for conditional expectations as $E(Y|x)$ or $E(Y|X)$, the former necessary for computation and the later for derivation of properties such as $E(Y) = E[E(Y|X)]$.

While likelihoods are not necessarily equal to probabilities, there is a connection between likelihood and probability. In the case of discrete random variables this is immediate in that probability mass functions do return probabilities,

$$Pr(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y}).$$

For continuous random variables, if observation 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. We could, and it is sometimes advocated that we should, write likelihoods so that they do always correspond to probabilities even in the continuous case. If we have assumed independence, then for a set of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, we could define the likelihood to be

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = Pr(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{F_i(y_i + \Delta_i|\boldsymbol{\theta}) - F_i(y_i - \Delta_i|\boldsymbol{\theta})\right\},$$

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

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = Pr(\boldsymbol{y}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}),$$

and the product of densities is sometimes called the density approximation to the likelihood. More often than not, however, we ignore this piece of mathematical technicality and define the likelihood and log likelihood as given in (5.3).

In many cases involving continuous random variables we assume that all $\Delta_i = \Delta$ and Δ is small enough to be ignored, but there are examples of where this is not the case (Lindsey, 1996) and it can be advantageous or even necessary to write the likelihood function in terms of probabilities rather than densities. This can occur, for example, if we are using continuous random variables to approximate a situation in which observable quantities are actually discrete, or if we are concerned about the precision with which data values have been recorded. We will assume unless otherwise noted that the density approximation of (5.3) is adequate for our purposes.

5.2 Maximum Likelihood Estimation

A *maximum likelihood* estimator or estimate of $\boldsymbol{\theta}$ is a value $\hat{\boldsymbol{\theta}} \in \Theta$ such that

$$L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}); \quad \text{for any } \boldsymbol{\theta} \in \Theta. \tag{5.4}$$

To avoid confusion, be aware that in discussing maximum likelihood we typically use the notation $\hat{\boldsymbol{\theta}}$ to denote either an estimate for a particular set of data or to denote an estimator defined by a procedure, assuming that the difference is clear by context. This is in contrast to, for example, the use of $\bar{y}$ to denote an estimate and $\bar{Y}$ to denote an estimator in introduction of unbiased

estimation. If the distinction between estimate and estimator is not clear or deserves special emphasis we can replace $L(\hat{\boldsymbol{\theta}})$ in (5.4) with $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ or $L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})$.

Now, given the preceding material, we have that $L(\boldsymbol{\theta}) \propto Pr(\boldsymbol{y}|\boldsymbol{\theta})$, 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 motivation for using 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 maximum likelihood is a good procedure for estimation 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 **?**, Theorem 2.1.

Result

Let P_θ represent the distribution 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 $f(y|\theta);\ y \in \Omega$

(iii) the true value of θ, say θ_0, lies in the interior of Θ

Then, as $n \to \infty$

$$P_{\theta_0}\left\{f(Y_1|\theta_0)\ldots f(Y_n|\theta_0) > f(Y_1|\theta)\ldots f(Y_n|\theta)\right\} \to 1,$$

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 hypothetical 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.

5.3 Asymptotic Normality and Efficiency

In this section we discuss some preliminaries for the development of asymptotic properties of likelihood-based estimators and inference and, in particular, maximum likelihood estimators. Although our context is likelihood analysis, the material of this section applies more generally and that will be reflected in the notation used. To emphasize the role of sample size we will now begin to index likelihoods and other quantities by the sample size n.

5.3.1 Asymptotic Normality

Asymptotic normality refers to the convergence in distribution of a suitably centered and scaled sequence of statistics to a standard normal distribution. The notion of asymptotic normality is even more general than what is presented here, but the slightly restricted context of estimation of parameters is appropriate in the context of applied statistical analysis.

If ψ is a scalar parameter and $\hat{\psi}_n$ denotes a consistent sequence of estimators of ψ, that sequence is asymptotically normal if there exists a sequence of

constants σ_n such that,

$$\frac{(\hat{\psi}_n - \psi)}{\sigma_n} \xrightarrow{d} \mathrm{N}(0, 1), \tag{5.5}$$

where $\xrightarrow{d}$ denotes convergence in distribution. If (5.5) holds we say $\hat{\psi}_n$ $\mathrm{AN}(\psi, \sigma_n^2)$. It is not uncommon that σ_n depends on n only as a factor, usually $\sigma_n = \sigma/\sqrt{n}$, but this is not always true. In any case, (5.5) implies that for a sufficiently large n we may behave as if $\hat{\theta}_n \sim \mathrm{N}(\psi, \sigma_n^2)$. In particular, we may compute an approximate interval estimate of ψ as,

$$\hat{\psi}_n \pm z_{1-\alpha/2}\sigma_n,$$

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

If $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^T$, a consistent estimator $\hat{\boldsymbol{\psi}}_n = (\hat{\psi}_{1,n}, \ldots, \hat{\psi}_{p,n})^T$ is asymptotically normal if there exists a sequence of matrices Σ_n such that,

$$\Sigma_n^{-1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathcal{I}_p), \tag{5.6}$$

where $\mathcal{I}_p$ is the $p \times p$ identity matrix with 1s on the diagonal and 0s elsewhere. Typically, Σ_n will be a function of the parameter $\boldsymbol{\psi}$. It may be that $\Sigma_n = \Sigma/n$ for a constant matrix Σ, in which case (5.6) may be written making the role of sample size explicit as,

$$n^{1/2}\Sigma^{-1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathcal{I}_p). \tag{5.7}$$

In this case, because Σ is non-stochastic and does not depend on n this may also be written as,

$$n^{1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \Sigma).$$

What (5.6) or its alternative forms mean is that for every real $p-$vector $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^T\boldsymbol{\psi}$ satisfies (5.5) with ψ replaced by $\boldsymbol{\lambda}^T\boldsymbol{\psi}$ and σ_n^2 replaced by $\boldsymbol{\lambda}^T\Sigma_n\boldsymbol{\lambda}$. The implications of (5.6) are that for a sufficiently large n we may behave as if

$\hat{\boldsymbol{\psi}}_n \sim \mathrm{N}(\boldsymbol{\psi}, \Sigma_n)$. Because the marginals of a multivariate normal distribution are also normal this means that we may behave as if $\hat{\psi}_j \sim \mathrm{N}(\psi_j, \sigma^2_{j,n})$ for $j = 1, \ldots, p$, where $\sigma^2_{j,n}$ is the j^{th} diagonal element of Σ_n.

A purely technical point is that despite the suggestive notation, σ^2_n in (5.5) may not correspond to variances of the sequence of estimators $\hat{\psi}_n$ and Σ_n in (5.6) may not correspond to covariance matrices of $\hat{\boldsymbol{\psi}}_n$. Nevertheless, we behave as if σ^2_n is the variance of $\hat{\psi}_n$ or as if Σ_n is the covariance matrix of $\hat{\boldsymbol{\psi}}_n$. If σ_n or Σ_n are not stochastic, typical assumptions are that $n^{1/2}\,\sigma_n \to \sigma > 0$ or $n\,\Sigma_n \to \Sigma$ for positive definite Σ with elementwise ordinary convergence. If σ_n or Σ_n are stochastic then the convergences are in probability.

5.3.2 Total Information

The expected or Fisher information plays an important role in the theory of estimation. In the case of iid random variables $Y_1, \ldots, Y_n$ with common distribution depending on a scalar parameter ψ, define the expected information in a single random variable as,

$$I(\psi) = E\left(\left[\frac{d}{d\psi}\log\{f(Y|\psi)\}\right]^2\right). \tag{5.8}$$

The total information in a sample of size n is $I_n(\psi) = nI(\psi)$. If the distribution of $Y_1, \ldots, Y_n$ depends on multiple parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^T$ the expected information is defined as the $p \times p$ matrix $I(\boldsymbol{\psi})$ with jk^{th} element,

$$I_{j,k}(\boldsymbol{\psi}) = E\left[\frac{\partial}{\partial\psi_j}\log\{f(Y|\boldsymbol{\psi})\}\,\frac{\partial}{\partial\psi_k}\log\{f(Y|\boldsymbol{\psi})\}\right], \tag{5.9}$$

and the total information is again $I_n(\boldsymbol{\psi}) = nI(\boldsymbol{\psi})$.

If the random variables $Y_1, \ldots, Y_n$ are independent but not identically distributed (inid) but with distributions that depend on a common parameter $\boldsymbol{\psi}$,

the total expected information is defined as the $p \times p$ matrix $I_n(\boldsymbol{\psi})$ with jk^{th} element,

$$I_{n,j,k}(\boldsymbol{\psi}) = \sum_{i=1}^{n} E\left[\frac{\partial}{\partial\psi_j}\log\{f_i(Y_i|\boldsymbol{\psi})\}\,\frac{\partial}{\partial\psi_k}\log\{f_i(Y_i|\boldsymbol{\psi})\}\right]. \tag{5.10}$$

If the random variables $Y_1, \ldots, Y_n$ are neither independent nor identically distributed, total information is the $p \times p$ matrix with jk^{th} element,

$$I_{n,j,k}(\boldsymbol{\psi}) = E\left[\frac{\partial}{\partial\psi_j}f(\boldsymbol{Y}|\boldsymbol{\psi})\,\frac{\partial}{\partial\psi_k}f(\boldsymbol{Y}|\boldsymbol{\psi})\right]. \tag{5.11}$$

Another way to express this same quantity is,

$$I_n(\boldsymbol{\psi}) = E[U_n(\boldsymbol{\psi})\,U_n^T(\boldsymbol{\psi})], \tag{5.12}$$

where $U_n(\boldsymbol{\psi}) = (U_{n,1}(\boldsymbol{\psi}), \ldots, U_{n,p}(\boldsymbol{\psi}))^T$ with, for $j = 1, \ldots, p$,

$$U_{n,j}(\boldsymbol{\psi}) = \frac{\partial}{\partial\psi_j}\log\{f(\boldsymbol{Y}|\boldsymbol{\psi})\}.$$

5.3.3 Efficiency

In the development of unbiased estimation, if $\hat{\psi}_n$ is an unbiased estimator of the scalar parameter ψ then the Information Inequality states that

$$\mathrm{var}(\hat{\psi}_n) \geq \frac{1}{nI}, \tag{5.13}$$

where I is the information in a single random variable (5.8) and will usually be a function of ψ, $I = I(\psi)$. An estimator for which there is equality in the information inequality is said to be *efficient*, which is a small sample or exact property. Asymptotic analogs of the information inequality assume that consistency and asymptotic normality holds. Suppose that observations are iid, ψ is a scalar, and $\hat{\psi}_n$ is $\mathrm{AN}(\psi, \sigma_n^2)$ such that $n\sigma_n^2 \to \sigma^2$, the asymptotic information inequality becomes,

$$\sigma^2 \geq \frac{1}{I}, \tag{5.14}$$

and again $I = I(\psi)$ will typically be a function of ψ. Note that (5.14) and (5.13) are quite distinct. The variance in (5.13) is an exact variance while σ^2 in (5.14) is a limiting value. Also, $\hat{\psi}_n$ in (5.13) is unbiased for all n, while $\hat{\psi}_n$ in (5.14) is not necessarily even asymptotically unbiased, although it is assumed to be consistent.

In the case of iid random variables depending on multiple parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^T$ suppose that $\hat{\boldsymbol{\psi}}_n$ is AN$(\boldsymbol{\psi}, \Sigma_n)$ and that $n\Sigma_n \to \Sigma$. The asymptotic information inequality takes a different form here with the result being that,

$$\Sigma - I^{-1} \text{ is nonnegative definite,}$$

where I is the expected information matrix for a single random variable and usually both $\Sigma = \Sigma(\boldsymbol{\psi})$ and $I = I(\boldsymbol{\psi})$ are functions of $\boldsymbol{\psi}$. This inequality has interpretation in terms of concentration ellipsoids Serfling (see 1980, Chapter 4.1.2) but is perhaps more easily understood in terms of its implication for individual elements of $\hat{\boldsymbol{\psi}}_n$. For $\sigma_{j,j}$ being the j^{th} diagonal element of $\Sigma(\boldsymbol{\psi})$, and $[I^{-1}]_{j,j}$ the j^{th} diagonal element of the inverse information matrix for a single random variable,

$$\sigma_{j,j} \geq [I^{-1}]_{j,j} \tag{5.15}$$

If there is equality in (5.15) then $\hat{\psi}_{n,j}$ is asymptotically efficient.

For inid or dependent random variables, assume asymptotic normality (5.6) holds such that $n\Sigma_n \to \Sigma$ and again let $\sigma_{j,j}$ denote the j^{th} diagonal element of Σ. Let the elements of the expected information matrix $I_{n,j,k}$ in (5.10) or (5.11) be such that $I_{n,j,k}/n \to I_{j,k}$. Asymptotic efficiency can then be defined for the elements of $\hat{\boldsymbol{\psi}}_n$ as equality in (5.15). Again, Σ_n, I_n, Σ and I are usually functions of $\boldsymbol{\psi}$.

Extension of the basic information inequality for unbiased estimators of

scalar parameters to asymptotic efficiency for estimators that are consistent and asymptotically normal allows us to say that if response variables $Y_1, \ldots, Y_n$ have joint pmf or pdf $f(\boldsymbol{y}|\boldsymbol{\psi})$ and there exists a sequence of consistent estimators of $\boldsymbol{\psi}$ such that $\hat{\boldsymbol{\psi}}_n$ AN$[\boldsymbol{\theta}, I_n^{-1}(\boldsymbol{\psi})]$ then the elements of $\hat{\boldsymbol{\theta}}_n$ are efficient.

It is important to understand that efficiency is a property of estimators, not estimates. So the fact that an estimator is efficient does not imply that an estimate produced from a set of data somehow has optimal properties. An estimate has no properties at all and cannot be claimed to be accurate or biased, precise or imprecise, it is simply a numerical value whose relation to the true parameter is unknown. Properties such as precision are properties of procedures or estimators only and statements of those properties such as (5.14) and (5.15) involve limiting quantities that can never be used or even approximated in practice. We can have some level of confidence or comfort about a particular estimate only because we know the tool used to produce it had good or optimal properties such as efficiency or asymptotic efficiency.

Asymptotic normality of likelihood-based estimators does have implications for analysis of a given set of data. Consider the limiting expression that corresponds to $\hat{\boldsymbol{\theta}}_n$ AN$[\boldsymbol{\theta}, I_n^{-1}(\boldsymbol{\theta})]$,

$$[I_n(\boldsymbol{\theta})]^{-1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathcal{I}_p).$$

To make use of the implication of this result we must have something that can replace $I_n(\boldsymbol{\theta})$ without changing the convergence in distribution. If $I_n(\boldsymbol{\theta})$ is available in closed form, that is, the expectation involved can be evaluated analytically, then we can use $I_n(\hat{\boldsymbol{\theta}}_n)$. As long as $\hat{\boldsymbol{\theta}}_n$ is consistent and the elements of $I_n(\boldsymbol{\theta})$ are smooth functions of $\boldsymbol{\theta}$, both of which are being assumed as part of the regularity conditions given previously in this chapter, then the limiting result continues to hold and we have $\hat{\boldsymbol{\theta}}_n$ AN$[\boldsymbol{\theta}, I_n^{-1}(\hat{\boldsymbol{\theta}}_n)]$. If the total

information matrix $I_n(\boldsymbol{\theta})$ is not available in closed form we could formulate a numerical approximation to $I(\hat{\boldsymbol{\theta}}_n)$ using the available data, typically through the use of a numerical integration method. Alternatively, and probably more common because of computational considerations, we may replace $I_n(\boldsymbol{\theta})$ with what is called the *observed information*. For the inid case under suitable regularity conditions, the total observed information is a $p \times p$ matrix with jk^{th} element,

$$\begin{aligned} I^{ob}_{n,j,k}(\hat{\boldsymbol{\theta}}_n) &= \sum_{i=1}^{n} \left[\frac{\partial}{\partial \theta_j} \log\{f(y_i|\boldsymbol{\theta})\} \, \frac{\partial}{\partial \theta_k} \log\{f(y_i|\boldsymbol{\theta})\} \right] \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \\ &= \sum_{i=1}^{n} \left[\frac{\partial^2}{\partial \theta_j \, \partial \theta_k} \log\{f(y_i|\boldsymbol{\theta})\} \right] \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} . \end{aligned} \tag{5.16}$$

5.4 Efficient Likelihood Estimators

In Chapter 5.3 we discussed asymptotic normality and efficiency in a somewhat general setting without reference to any particular type of estimator. In this section we describe what is necessary for these properties to hold for a class of likelihood-based estimators.

The estimators we consider are obtained as solutions to the *likelihood equations*. Quite generally, if the joint pmf or pdf of the response random variables $f(\boldsymbol{y}|\theta)$ depends on a single scalar parameter, then the likelihood equation is,

$$\frac{d}{d\,\theta} \log\{f(\boldsymbol{y}|\theta)\} = 0, \tag{5.17}$$

while if the joint $f(\boldsymbol{y}|\boldsymbol{\theta})$ depends on a vector-valued parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, then the likelihood equations are, for $j = 1, \ldots, p$,

$$\frac{\partial}{\partial\, \theta_j} \log\{f(\boldsymbol{y}|\theta)\} = 0. \tag{5.18}$$

If the response variables $Y_1, \ldots, Y_n$ are independent, then these equations may be written in terms of sums of derivatives of univariate densities or mass functions.

We can identify four issues connected with the development of asymptotic results for estimators obtained as solutions to the likelihood equations, (i) consistency of a sequence of likelihood equation solutions, (ii) demonstration that asymptotic normality holds for such a sequence, (iii) verification that the information inequality holds for such a sequence, and (iv) uniqueness of such a sequence. The first of these is concerned with locating a consistent sequence of estimators, regardless of whether such a consistent sequence is unique or is a global maximum of the likelihood function. Resolution of the second issue will provide an approximate sampling distribution from which to compute inferential quantities. The third issue will determine whether we can claim asymptotic efficiency if the variance of the limiting distribution is equal to the inverse information. Finally, uniqueness indicates that the solution found is the maximum likelihood estimator. Note here that the conditions and results to be presented directly do not exhaust the set of asymptotic results possible for maximum likelihood estimators in various situations. They do represent, however, what might be thought of as the typical or usual framework within which to consider asymptotic inference based on properties of likelihood-based estimators and asymptotic normality in particular.

Properties of estimators are developed under sets of conditions called *regularity conditions*. There are a variety of regularity conditions that have been developed, and different sets of conditions are needed to prove different results about likelihood-based estimators. It is not our intention to catalog all of these here. Rather, we will attempt to merge conditions that lead to various results into a basic package amenable to reference for practical use. For a more

detailed and piecewise presentation of this topic, see Lehman (1983, Chapter 6).

5.4.1 Observations iid and Scalar Parameter

We will list one set of conditions sufficient to produce desired results for random variables $\{Y_i : i = 1, \ldots, n\}$ that are independent and identically distributed with a scalar parameter. This will be extended to the multi-parameter situation in the sequel.

Scalar Parameter Regularity Conditions

R1. The distributions of the response variables are identifiable, meaning that different parameter values result in distinct distributions. We will assume these distributions have a common probability density or mass function $f(y|\theta)$; $y \in \Omega$ and that the support Ω does not depend on θ.

R2. The parameter space Θ is an open interval (not necessarily finite).

R3. The common density or mass function $f(y|\theta)$ has three continuous derivatives with respect to θ.

R4. With $\mu(y)$ denoting the dominating measure (Lebesque or counting) the first and second derivatives of the integral $\int f(y|\theta)\, d\mu(y)$ can be evaluated by passing the derivative under the integral operator, that is, for $k = 1, 2$,

$$\frac{d^k}{d\,\theta^k} \int f(y|\theta)\, d\mu(y) = \int \frac{d^k}{\theta^k} f(y|\theta)\, d\mu(y).$$

R5. The expected (or Fisher) information in a single random variable, $I(\theta) = E[\frac{d}{d\theta} log\{f(y|\theta)\}]^2$, is positive and finite, $0 < I(\theta) < \infty$.

R6. For all elements of the support $y \in \Omega$ and in an interval of the true parameter value, $\theta_0 - c < \theta < \theta_0 + c$, the third derivative of $\log\{f(y|\theta)\}$ satisfies

$$\left| \frac{d^3}{d\theta^3} \log\{f(y|\theta)\} \right| \leq M(y),$$

where,

$$E_{\theta_0}[M(y)] < \infty.$$

It should be noted that the regularity conditions just listed are typical but not unique in developing asymptotic results for likelihood estimation. For example, any number of authors replace condition R4 with
Alternative R4:

$$\left| \frac{df(y|\theta)}{d\theta} \right| \leq g(y) \qquad \text{and} \left| \frac{d^2 f(y|\theta)}{d\theta^2} \right| \leq h(y)$$

such that

$$\int g(y)\, dy < \infty \qquad \text{and} \int h(y)\, dy < \infty.$$

Also, not all of these conditions are needed for resolution of each of the individual issues listed previously. For example, existence of a consistent sequence of solutions to the likelihood equations can be demonstrated given only R1 and R2, and actually even with a slightly relaxed version of R2 that requires only that the true parameter lie in an open interval of the parameter space, regardless of whether the entire space is open or not. Understanding that various subsets of the regularity conditions given previously can be used to demonstrate some but not all of the results we wish to use in an analysis is not always vital in an application, but can be quite useful in developing research into new models.

Lemma 1

Suppose that $Y_1, \ldots, Y_n$ are iid with common pmf or pdf $f(y|\theta)$, and the conditions R1-R2 are satisfied. Then a consistent sequence of solutions to the likelihood equation $\frac{d}{d\theta} \log\{f(\boldsymbol{y}|\theta)\} = 0$ exists.

Likelihood Theorem 1

Suppose that conditions R1-R6 are satisfied. If the sequence of solutions to the likelihood equation given by Lemma 1 is unique for all n and $\boldsymbol{y}$, then it is a sequence of maximum likelihood estimators $\hat{\theta}_n$ and,

$$\sqrt{n}\,[I(\theta)]^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathrm{N}(0,1), \tag{5.19}$$

where

$$I(\theta) = -E\left[\frac{d^2}{d\,\theta^2} \log\{f(\boldsymbol{Y}|\theta)\}\right].$$

A minor point is that the assumption of unique solutions to the likelihood equation for all n and $\boldsymbol{y}$ in this theorem can be replaced with a condition that the probability of multiple solutions tends to 0 as $n \to \infty$.

Likelihood Theorem 1 results in asymptotic normality for maximum likelihood estimators in regular problems (i.e., under the suitable regularity conditions) which implies that in practice we may behave as if

$$\hat{\theta}_n \sim \mathrm{N}[0, I^{-1}(\theta)/n]. \tag{5.20}$$

That the asymptotic variance is equal to the inverse expected information indicates that in these cases maximum likelihood estimators are efficient, satisfying the information inequality.

Notice that in Likelihood Theorem 1 we have assumed the existence of unique solutions rather than arriving at it as a consequence of assumed regularity conditions. The possibility that the likelihood or log likelihood might

have multiple local maxima or even a saddle point is the most difficult of the four issues identified previously to verify in practice and this issue is, frankly, often ignored unless problems arise in numerical algorithms to locate maximum likelihood estimates or counter-intuitive results are obtained in estimation. As we will see, when such difficulties are encountered, the possibility that there are multiple likelihood modes or that the likelihood is unbounded present themselves as potential causes. There is, however, one large class of problems for which unique solutions to the likelihood equations are guaranteed.

Corollary 1.1

If $Y_1, \ldots, Y_n$ in Likelihood Theorem 1 follow a common distribution that constitutes an exponential family, then solutions to the likelihood equations, if they exist, are unique.

The result of Likelihood Theorem 1 contains the expected or Fisher information. Following the discussion of Chapter 3.3, we will usually need an alternative quantity to use in (5.19) that does not alter the result. Two such alternatives are given in the following result.

Corollary 1.2

The asymptotic normality of Theorem 1 continues to hold if $I(\theta)$ is replaced with $I(\hat{\theta}_n)$ or $I^{obs}(\hat{\theta}_n)$ in (5.19), where,

$$\begin{aligned} I(\hat{\theta}_n) &= -E\left[\frac{d^2}{d\,\theta^2}\log\{f(\boldsymbol{Y}|\theta)\}\right]\Bigg|_{\theta=\hat{\theta}_n}. \\ I^{obs}(\hat{\theta}_n) &= -\left[\frac{d^2}{d\,\theta^2}\log\{f(\boldsymbol{y}|\theta)\}\right]\Bigg|_{\theta=\hat{\theta}_n}. \end{aligned}$$

5.4.2 Observations iid With Multiple Parameters

The result of Likelihood Theorem 1 can be extended to situations in which we have iid random variables $Y_1, \ldots, Y_n$ with common pdf or pmf $f(y|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. In this case we have a set of likelihood equations, and we also extend the definition of expected information so that $I(\boldsymbol{\theta})$ becomes a $p \times p$ matrix with j, k^{th} element (5.9), using $\boldsymbol{\theta}$ instead of $\boldsymbol{\psi}$. Suitable regularity conditions for the multiple parameter case are similar to R1-R6 described previously.

Multiple Parameter Regularity Conditions

R1,R2. The first two regularity conditions remain identical to those given previously.

RM3. The common density or mass function $f(y|\theta)$ has continuous partial derivatives up to order three with respect to the elements of $\boldsymbol{\theta}$.

RM4. Typically, R4 is restated as the direct consequence of what that condition implies for the single parameter case, generalized to multiple parameters so that, for $j, k = 1, \ldots, p$,

$$E\left[\frac{\partial}{\partial\theta_j}\log\{f(\boldsymbol{y}|\boldsymbol{\theta})\}\right] = 0$$

and

$$I_{j,k}(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial\theta_j}\log\{f(\boldsymbol{y}|\boldsymbol{\theta})\}\ \frac{\partial}{\partial\theta_k}\log\{f(\boldsymbol{y}|\boldsymbol{\theta})\}\right] = -E\left[\frac{\partial^2}{\partial\theta_j\,\partial\theta_k}\log\{f(\boldsymbol{y}|\boldsymbol{\theta})\}\right].$$

RM5. Each element of the information matrix $I_{j,k}(\boldsymbol{\theta})$ is positive and finite and the matrix itself $I(\boldsymbol{\theta})$ is positive definite.

RM6. The smoothness condition R6 is generalized to hold for third partial derivatives as, for $j, k, \ell = 1, \ldots, p$,

$$\left| \frac{\partial^3}{\partial\theta_j \, \partial\theta_k \, \partial\theta_\ell} \log\{f(\boldsymbol{y}|\boldsymbol{\theta})\} \right| \leq M_{j,k,\ell}(\boldsymbol{y}),$$

such that for $j, k, \ell = 1, \ldots, p$,

$$E_{\boldsymbol{\theta}_0}[M_{j,k,\ell}(\boldsymbol{\theta})] < \infty.$$

Likelihood Theorem 2

Suppose that $Y_1, \ldots, Y_n$ are iid with common pmf or pdf $f(\boldsymbol{y}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, and that the regularity conditions R1, R2 and RM3-RM6 hold. Then with probability tending to 1 as $n \to \infty$ there exists one or more consistent sequences of solutions to the likelihood equations such that,

$$\sqrt{n}[I(\boldsymbol{\theta})]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}_p, \mathcal{I}_p), \tag{5.21}$$

where $\mathbf{0}_p$ is a $p-$vector of $0s$ and $\mathcal{I}_p$ is the $p \times p$ identity matrix with values of 1 on the diagonal and 0 elsewhere.

If the sequence of solutions to the likelihood equations is unique, then for each $j = 1, \ldots, p$, the elements of $\hat{\boldsymbol{\theta}}_n$ are asymptotically normal and efficient, that is,

$$\sqrt{n}[I^{-1}(\boldsymbol{\theta})]_{j,j}(\hat{\theta}_{j,n} - \theta_{j,n}) \xrightarrow{d} \mathrm{N}(0, 1), \tag{5.22}$$

where $[I^{-1}(\boldsymbol{\theta})]_{j,j}$ is the j^{th} diagonal element of $I^{-1}(\boldsymbol{\theta})$.

As for Likelihood Theorem 1, uniqueness of solutions to the likelihood equations are an additional assumption of the theorem and are not necessarily implied by the regularity conditions given. An interesting difference with Likelihood Theorem 1 here is that uniqueness of solutions to the likelihood equations does not guarantee that those solutions are maximum likelihood

estimators, or even that a maximum likelihood estimator exists. But from a practical standpoint this is largely a technical detail because the theorem provides consistency, asymptotic normality, and efficiency. These properties are sufficient for producing approximate inferential statements. The result of Corollary 1.1 in the case of a scalar parameter continues to hold for each element of a vector-valued parameter.

Corollary 2.1

If $Y_1, \ldots, Y_n$ in Theorem 1 follow a common distribution that constitutes an exponential family, then solutions to the likelihood equation$\hat{\theta}_{j,n}$, if they exist, are unique for $j = 1, \ldots, p$.

In a similar manner as for Likelihood Theorem 1 we can replace the expected information matrix, which contains unknown quantities, with alternatives that can be computed.

Corollary 2.2

The asymptotic normality of Likelihood Theorem 2 continues to hold if $I(\hat{\boldsymbol{\theta}})$ in (5.21) is replaced with $I(\hat{\boldsymbol{\theta}}_n)$ or $I^{obs}(\hat{\boldsymbol{\theta}}_n)$, which are $p \times p$ matrices with jk^{th} elements

$$\begin{aligned} I_{j,k}(\hat{\boldsymbol{\theta}}_n) &= -E\left[\frac{\partial^2}{\partial\theta_j\,\partial\theta_k}\log\{f(\boldsymbol{Y}|\boldsymbol{\theta})\}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}. \\ I^{obs}_{j,k}(\hat{\boldsymbol{\theta}}_n) &= -\left[\frac{\partial^2}{\partial\theta_j\,\partial\theta_k}\log\{f(\boldsymbol{y}|\boldsymbol{\theta})\}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}. \end{aligned}$$

5.4.3 Extensions to Non-iid settings

There are many situations in which we want to apply likelihood estimation but for which collections of response random variables are not identically dis-

tributed, not independent, or both. Regression models, both linear and nonlinear, fall into this category, as do problems involving likelihoods formulated with censored observations, random effects, or stochastic processes. For some models formulated for independent but not identically distributed random variables such as generalized linear models, asymptotic results for maximum likelihood estimators are readily available. This is largely the result of generalized linear models being formulated for random model components consisting of exponential dispersion families as described in Chapter 2.

In many non-iid cases, however, additional conditions that allow the development of asymptotic properties for likelihood-based estimators are necessary. Some of the conditions that are sufficient for consistency and asymptotic normality of maximum likelihood estimators become quite technical and their effects are not intuitive. While we will not attempt to identify any set of appropriate conditions explicitly, several general observations can be made pertaining to the character of what needs to result. First, most proofs of consistency for likelihood estimators make use of Taylor series expansions of the score functions followed by application of a central limit theorem. In the independent but not identically distributed case, conditions that allow a central limit theorem for non-identically distributed random variables is needed. With dependent observations a common tact is to rely on martingale central limit theorems. Secondly, the total information in a sample must grow without bound as the sample size increases. For independent but not identically distributed random variables the total information is a sum of contributions from individual random variables. The additional conditions specified in non-iid problems typically are sufficient to allow a appropriate central limit and law of large numbers to apply to score functions and to ensure that total information tends to infinity with increasing sample size. Results for independent but

not identically distributed problems may be found in Bradley and Gart (1962) and Hoadley (1971). Various settings involving dependent random variables are addressed by, among others, Bar-Shalom (1971), Bhat (1974), and Crowder (1976).

An issue that is important in non-iid settings is the manner in which sample size grows large, which is known as the *asymptotic context* for a model. Consider first random variables that are independent but not identically distributed. One setting involves groups of random variables $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})^T$ for $i = 1, \ldots, k$ such that the joint distributions within groups $f_i(\boldsymbol{y}_i|\boldsymbol{\theta}_i)$ all share one or more common parameters. That is, $\cap_{i=1}^k \boldsymbol{\theta}_i \neq \emptyset$; it is not necessary that each f_i depends on all the parameters although that is possible. A problem that would fit this scenario is a simple linear regression with separate intercept parameters for k groups of observations but a common slope. Bradley and Gart (1962) consider likelihood asymptotics when k remains fixed, $n_i \to \infty$ for each $i = 1, \ldots, k$ and $n_i/N = c$ where $N = \sum n_i$ and c is constant. Another possibility would be for $n_i = n$ for $i = 1, \ldots, k$ but $k \to \infty$ which would be the case, for example, if $n = 1$ in a linear or nonlinear regression.

The situation is similar but more complex for non-independent random variables. Consider, for example, a longitudinal linear model for random variables $Y_{i,j}$ with $j = 1, \ldots, n_i$ observations on $i = 1, \ldots, k$ individuals. Suppose the regression parameters giving the marginal expected values are common to all of the $Y_{i,j}$ but there is also an individual-specific random effect. In the joint marginal distribution $Y_{i,j}$ and $Y_{i,k}$ are correlated, for $j, k = 1, \ldots, n_i$. Here, one could take $n_i = n$ and allow $k \to \infty$ or take k to be fixed but $n_i \to \infty$ for each $i = 1, \ldots, k$ or allow $n_i \to \infty$ and $k \to \infty$, usually such that $k/n_i \to 0$ or $k/n_i = c$ for some constant c. Another situation would be presented by a spatial problem in which random variables are located on a regular lattice

or a spatio-temporal situation in which random variables are indexed in both space and time. Here, possible asymptotic contexts include what are called *repeating lattice* and *expanding lattice* scenarios. The former involves replicate observations, usually assumed to be independent, of a fixed spatial or spatio-temporal domain that contains some type of dependence structure. The joint distribution of the complete set of random variables involved in the problem then becomes a product of multivarite distributions. In contrast, an expanding lattice scenario presumes the set of spatial or spatio-temporal locations grows without bound, extending over a larger and larger region. In this case, the joint distribution of the entire set of random variables involved is a single multivariate distribution of ever increasing dimension. Full maximum likelihood estimation for problems involving spatial and spatio-temporal models is often difficult at best and alternative methods may be employed for estimation and inference, but likelihood-based estimation and inference are practical in a some cases involving Gaussian distributions. The concept of asymptotic context is valuable to understand when searching for a theoretical result that justifies methods of estimation and inference being applied to a particular model.

Example 5.1

Suppose $Y_1, \ldots, Y_n$ are iid following an $s-$parameter exponential family with density, for $y \in \Omega$ and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s$,

$$f(y|\boldsymbol{\theta}) = \exp\left[\sum_{q=1}^{s} \theta_q T_q(y) - B(\boldsymbol{\theta}) + c(y)\right].$$

The log likelihood for this sample is,

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{s}\left[\theta_j \sum_{i=1}^{n} T_j(y_i)\right] - nB(\boldsymbol{\theta}) + \sum_{i=1}^{n} c(y_i),$$

and the likelihood equations are, for $j = 1, \ldots, s$,

$$\frac{1}{n}\sum_{i=1}^{n} T_j(y_i) = E[T_j(y_1)].$$

Here, the expected information is, for a single random variable, the $p \times p$ matrix with jk^{th} element,

$$I_{j,k}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_j\,\partial\theta_k} B(\boldsymbol{\theta}) = \mathrm{cov}(T_j(y_1), T_k(y_1)).$$

From this and Corollary 2.1 we have that the mle of $\boldsymbol{\theta}$, if it exists, is (i) unique, (ii) uniform minimum variance unbiased (UMVU), (iii) asymptotically normal and (iv) asymptotically efficient. That the mle is UMVU follows from the fact that the $\{T_j : j = 1, \ldots, s\}$ are minimal sufficient for $\boldsymbol{\theta}$ and the mle is unbiased and a function of these statistics. That the mle is UMVU gives an optimal exact-theory property, but the sampling distribution is not immediately available, except in the case of normal distributions, for which we typically do not rely on likelihood asymptotics for inference. Likelihood Theorem 2, however, provides an approximate sampling distribution and asymptotic efficiency for exponential families in general.

Example 5.2

Suppose $Y_1, \ldots, Y_n$ are iid with common probability density function (3.2). The second derivative of the log likelihood corresponding to this density with respect to the location parameter xi is strictly negative for any set of observations $\boldsymbol{y}$, so the log likelihood is concave in the dimension of xi. This is not true, however, for the second derivative of the log likelihood with respect to the scale parameter ϕ, causing potential difficulties with asserting a solution to the likelihood equations is unique. Note that this does not mean the solution cannot be unique, just that we can not guarantee uniqueness based on the sufficient condition of a concave log likelihood. Also, we would like to have a

concave function in two-dimensions. A sufficient condition for this is that all of the eigenvalues of the Hessian matrix be negative, which is difficult to prove in general (for any $\boldsymbol{y}$) and may not be uniformly true in any case. In practice, then, we attempt to assure ourselves that the log likelihood is concave at least locally. Given numerical solutions to the likelihood equations for a given set of data, we may compute numerical values for the Hessian, the matrix of second derivatives of the log likelihood at that parameter value, and verify that at the eigenvalues are both negative. This seems to typically be the case, but one cannot guarantee such results for any set of data.

Example 5.3

Consider random variables $Y_1, \ldots, Y_n$ that are iid with common density

$$f(y|\theta) = \frac{1}{\theta}\mathcal{I}(0, y, \theta),$$

where $\mathcal{I}(A)$ is the indicator function that assumes a value of 1 if A is true and 0 otherwise. The log likelihood and its derivatives are then,

$$\begin{aligned} \ell_n(\theta) &= -n\log\{\theta\}, \\ \frac{\partial}{\partial\theta}\ell_n(\theta) &= \frac{-n}{\theta}, \\ \frac{\partial^2}{\partial\theta^2}\ell_n(\theta) &= \frac{n}{\theta^2}. \end{aligned}$$

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 follows from $L_n(\theta) = 1/\theta^n$, which gives

$$L_n(\max\{y_1, \ldots, y_n\}) \geq L_n(\theta); \quad \text{any } \theta \in (0, \infty).$$

The asymptotics of Likelihood Theorem 1 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. Note 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 does not apply in this case because the support of the distribution of response variables depends on the value of the parameter.

Two additional properties of likelihood-based estimators, and maximum likelihood estimators in particular, are worthy of mention to close out our discussion of this section.

1. If a given scalar parameter θ (which may be an element of the parameter vector $\boldsymbol{\theta}$) has a single sufficient statistic $T(\boldsymbol{y})$, then the maximum likelihood estimator must be a function of that sufficient statistic. If that sufficient statistic is minimal and complete, then the maximum likelihood estimator is unique. If the maximum likelihood estimator is unbiased then it is the UMVU (e.g., Stuart and Ord, 1994, 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. Likelihood-based estimators, determined as solutions to the likelihood equations, possess a property called *invariance* that is very useful but is not, in general, a property of other types of estimators, such as unbiased or least square estimators. The invariance property can be stated as, if $\hat{\boldsymbol{\theta}}_n$ is a consistent sequence of solutions to the likelihood equations, and $g(\boldsymbol{\theta})$ is a continuous, real-valued function of $\boldsymbol{\theta}$, then $g(\hat{\boldsymbol{\theta}}_n)$ is a consistent sequence of solutions to the likelihood equations when the likelihood is reparameterized in terms of $g(\boldsymbol{\theta})$. If $\hat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ then $g(\hat{\boldsymbol{\theta}}_n)$ is the maximum likelihood estimator of $g(\boldsymbol{\theta})$. Invariance is particularly useful in considering alternate parameterizations of random model components. It implies that, if using maximum likelihood, there is no need to explicitly reparameterize a likelihood and conduct an optimization procedure in order to move between several parametric forms for a model.

Example 5.4

Suppose that $Y_1, \ldots, Y_n$ are iid having a common beta distribution with parameters α and β, and let maximum likelihood estimators of these parameters be $\hat{\alpha}$ and $\hat{\beta}$ and note that these estimators are consistent and asymptotically normal. A maximum likelihood estimator of the common expected value of the response variables is then $\hat{\mu} = \hat{\alpha}/\hat{\beta}$. Because $\hat{\mu}$ is a maximum likelihood estimator and the maximum likelihood estimators of α and β are consistent, $\hat{\mu}$ is AN[μ, $V(\mu)$]. A method for determining $V(\mu)$ will be given in the next section.

5.5 Wald Theory Inference

The title of this section stems from the fact that the inferential methods presented are based on a test statistic introduced by Wald (1943) which is given in the main result. This statistic can be used to formulate tests of hypotheses and form confidence regions and intervals for parameters. We assume the conditions of Likelihood Theorem 2 hold for iid random variables $Y_1, \ldots, Y_n$.

5.5.1 Wald Theory Main Result

If $\{\hat{\boldsymbol{\theta}}_n\}$ is a sequence of consistent, asymptotically normal and efficient estimators of $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)^T$ then,

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T I_n(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \chi^2_p, \tag{5.23}$$

where $I_n(\hat{\boldsymbol{\theta}}_n)$ is the total expected information matrix evaluated at the estimate $\hat{\boldsymbol{\theta}}_n$ and χ^2_p is a Chi-squared random variable with p degrees of freedom For a proof of this result see, Serfling (e.g., 1980, Chapter 4.4), but note that the

result is often (usually) written with $I_n(\hat{\boldsymbol{\theta}}_n) = nI(\hat{\boldsymbol{\theta}}_n)$ where $I(\cdot)$ denotes the expected information matrix for a single random variable.

Following Serfling (1980, Chapter 4), we will consider one or more restrictions placed on the elements of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, specified as, for $r \leq p$,

$$R_j(\boldsymbol{\theta}) = 0; \quad j = 1, \ldots, r.$$

Example 5.5

A. Let $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \theta_3)^T$. With θ_1^0, θ_2^0 and θ_3^0 denoting specific values of these parameters, specify the restrictions, $R_1(\boldsymbol{\theta}) = \theta_1 - \theta_1^0 = 0$, $R_2(\boldsymbol{\theta}) = \theta_2 - \theta_2^0 = 0$, and $R_3(\boldsymbol{\theta}) = \theta_3 - \theta_3^0 = 0$. These restrictions correspond to the hypothesis that $\theta_1 = \theta_1^0$, $\theta_2 = \theta_2^0$ and $\theta_3 = \theta_3^0$.

B. Let $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \theta_3)^T$. Specify the single restriction $R_1(\boldsymbol{\theta}) = \theta_1 - \theta_1^0 = 0$. This restriction corresponds to the hypothesis that $\theta_1 = \theta_1^0$, but leaves θ_2 and θ_3 unrestricted.

C. Let $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \theta_3, \theta_4)^T$. Specify the restrictions, $R_1(\boldsymbol{\theta}) = \theta_1 - \theta_2 = 0$ and $R_2(\boldsymbol{\theta}) = \theta_3 - \theta_4 = 0$. These restrictions correspond to the hypothesis that $\theta_1 = \theta_2$ and $\theta_3 = \theta_4$.

In these examples, 5.5A would be called a *simple* hypothesis while 5.5B and 5.5C would be called *composite* hypotheses, the distinction resting 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) allows the development of both tests of hypotheses corresponding to the restrictions specified and confidence regions or intervals for subsets of the parameter vector $\boldsymbol{\theta}$.

5.5.2 Wald Theory Tests

As in Example 5.5, a set of r restrictions on the elements of $\boldsymbol{\theta}$ corresponds to a hypothesis about those parameters. Let $b(\boldsymbol{\theta}) = (R_1(\boldsymbol{\theta}), \ldots, R_r(\boldsymbol{\theta}))^T$ and 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 $H : R_1(\boldsymbol{\theta}) = 0, \ldots, R_r(\boldsymbol{\theta}) = 0$,

$$W_n = b^T(\hat{\boldsymbol{\theta}}_n) \left[C(\hat{\boldsymbol{\theta}}_n) I_n^{-1}(\hat{\boldsymbol{\theta}}_n) C^T(\hat{\boldsymbol{\theta}}_n) \right]^{-1} b(\hat{\boldsymbol{\theta}}_n) \xrightarrow{d} \chi^2_r, \tag{5.24}$$

Example 5.5 (cont.)

Revisiting the cases given in Example 5.5, the results play out as follows. Let the jk^{th} element of $I^{-1}(\hat{\boldsymbol{\theta}}_n)$ be denoted as i^{jk} and note that $i^{jk} = i^{kj}$.

A. Here, C in (5.24) is the 3×3 identity matrix 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,

$$\begin{aligned} W_n &= (\hat{\theta}_{n,1} - \theta_1^0)^2 i^{11} + (\hat{\theta}_{n,2} - \theta_2^0)^2 i^{22} + (\hat{\theta}_{n,3} - \theta_3^0)^2 i^{33} \\ &+ 2(\hat{\theta}_{n,1} - \theta_1^0)(\hat{\theta}_{n,2} - \theta_2^0) i^{12} + 2(\hat{\theta}_{n,1} - \theta_1^0)(\hat{\theta}_{n,3} - \theta_3^0) i^{13} \\ &+ 2(\hat{\theta}_{n,2} - \theta_2^0)(\hat{\theta}_{n,3} - \theta_3^0) i^{23} \end{aligned}$$

B. 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 $W_n = (\hat{\theta}_{n,1} - \theta_1^0)^2 \frac{1}{i^{11}}$.

C. Here,

$$C(\theta) = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

and,

$$\begin{aligned} W_n &= (\hat{\theta}_{n,1} - \hat{\theta}_{n,2})^2 \left(i^{11} + i^{22} - 2i^{12}\right) \\ &+ 2(\hat{\theta}_{n,1} - \hat{\theta}_{n,2})(\hat{\theta}_{n,3} - \hat{\theta}_{n,4})(i^{13} - i^{23} - i^{14} + i^{24}) \\ &+ (\hat{\theta}_{n,3} - \hat{\theta}_{n,4})^2 \left(i^{33} + i^{44} - 2i^{34}\right). \end{aligned}$$

5.5.3 Wald Theory Intervals

To develop confidence intervals and sets, let $\boldsymbol{\theta}^0$ denote the true parameter value and let $\mathcal{S} \subseteq \{1, \ldots, p\}$. Specify the set of restrictions $\{R_j : j \in \mathcal{S};\, \theta_j - \theta_j^0 = 0\}$. Then an approximate $100(1-\alpha)\%$ confidence region for $\{\theta_j : j \in \mathcal{S}\}$ is given by,

$$\left\{\theta_j^0 : j \in \mathcal{S};\; b^T(\hat{\boldsymbol{\theta}}_n)\left[C(\hat{\boldsymbol{\theta}}_n) I_n^{-1}(\hat{\boldsymbol{\theta}}_n) C^T(\hat{\boldsymbol{\theta}}_n)\right]^{-1} b(\hat{\boldsymbol{\theta}}_n) \leq \chi^2_{r,1-\alpha}\right\}. \tag{5.25}$$

If $\mathcal{S} = \{2\}$, then $b(\boldsymbol{\theta}) = (\theta_2 - \theta_2^0)$,

$$\left[C(\hat{\boldsymbol{\theta}}_n) I_n^{-1}(\hat{\boldsymbol{\theta}}_n) C^T(\hat{\boldsymbol{\theta}}_n)\right]^{-1} = \frac{1}{i^{22}},$$

and the confidence region becomes,

$$\{\theta_2^0 : (\hat{\theta}_{n,2} - \theta_2^0)\frac{1}{i^{22}}(\hat{\theta}_{n,2} - \theta_2^0) \leq \chi^2_{1,1-\alpha}\}.$$

Taking the square root of both sides of the inequality in this set,

$$(\hat{\theta}_{n,2} - z_{1-\alpha/2}\sqrt{i^{22}} \;\leq\; \theta_2^0 \;\leq\; \hat{\theta}_{n,2} + z_{1-\alpha/2}\sqrt{i^{22}}),$$

which is identical to what results from (5.22) with $I(\hat{\boldsymbol{\theta}}_n)$ in place of $I(\boldsymbol{\theta})$ as justified by Corollary 2.2.

5.5.4 The Delta Method

The method described in this section applies to any asymptotically normal estimator and to emphasize this we return to the notation used previously

in Chapter 5.3 with $\hat{\boldsymbol{\psi}}$ denoting a generic estimator of a parameter $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^T$. Suppose that $\hat{\boldsymbol{\psi}}_n$ AN($\boldsymbol{\psi}$, $a_n^2\Sigma$) such that $a_n \to 0$ as $n \to \infty$. Let $g(\boldsymbol{\psi}) = [g_1(\boldsymbol{\psi}), \ldots, g_r(\boldsymbol{\psi})]^T$; $r \leq p$, where each component function $g_k(\boldsymbol{\psi})$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}$. Then,

$$g(\hat{\boldsymbol{\psi}}_n)\text{AN}[g(\boldsymbol{\psi},\, a_n^2\boldsymbol{D}\Sigma\boldsymbol{D}^T], \tag{5.26}$$

where, $\boldsymbol{D}$ is an $r \times p$ matrix with k, j^{th} element,

$$\frac{\partial}{\partial\, \psi_j} g_k(\boldsymbol{\psi}).$$

In both the covariance Σ and in the matrix $\boldsymbol{D}$, $\hat{\boldsymbol{\psi}}_n$ may be used as a a plug-in estimator of $\boldsymbol{\psi}$. Consistency of $\hat{\boldsymbol{\psi}}$ allows the asymptotic result to be applied without modification.

In likelihood estimation and inference, $a_n^2\Sigma$ is typically one of the forms of inverse *total* information given in Chapter 5.3.2. The classical setting is the *iid* case for which $a_n = 1/\sqrt{n}$ and Σ is the inverse expected information for a single random variable. But it may also be, for example, that $a_n^2\Sigma = n[(1/n)I_n(\boldsymbol{\theta})]$ where $I_n(\boldsymbol{\theta})$ is $p \times p$ with j, k^{th} element (5.10).

5.6 Inference from Properties of the Log Likelihood

Wald theory inference proceeds largely from properties of likelihood-based estimators. It is also possible to develop inferential procedures based on asymptotic properties of the log likelihood function itself.

The name of this section is perhaps something of a misnomer, since everything that has been discussed in this chapter 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 section, 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

$$\ell_n(\boldsymbol{\theta}) = log\{f(\boldsymbol{y}|\boldsymbol{\theta}); \quad \theta \in \Theta,$$

and a *reduced model* of the form,

$$\ell_n(\boldsymbol{\theta}_0) = log\{f(\boldsymbol{y}|\boldsymbol{\theta}_0); \quad \theta_0 \in \Theta_0,$$

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. 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 Chapter 5.5.

Likelihood Ratio Main Result

Let $\dim\{\Theta\} = p$ and $\dim\{\Theta_0\} = r$, and,

$$\hat{\boldsymbol{\theta}}_n = \sup_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}) \qquad \tilde{\boldsymbol{\theta}}_n = \sup_{\boldsymbol{\theta}\in\Theta_0} \ell_n(\boldsymbol{\theta}).$$

Then, assuming that $\boldsymbol{\theta} \in \Theta_0$ (the reduced model),

$$T_n \equiv -2\left\{\ell_n(\tilde{\boldsymbol{\theta}}_n) - \ell_n(\hat{\boldsymbol{\theta}}_n)\right\} \xrightarrow{d} \chi^2_{p-r}. \tag{5.27}$$

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. This 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 (5.27) would not reject θ_0 at the α level. That is, $\boldsymbol{\theta}_0$ is any value of $\boldsymbol{\theta}$ such that,

$$-2\left\{\ell_n(\boldsymbol{\theta}_0) - \ell_n(\hat{\boldsymbol{\theta}}_n)\right\} \leq \chi^2_{p,1-\alpha}.$$

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 (5.27).

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 $\boldsymbol{\theta}$,

$$\left\{\boldsymbol{\theta}_0 : -2\left[\ell_n(\boldsymbol{\theta}_0) - \ell_n(\hat{\boldsymbol{\theta}}_n\right] \leq \chi^2_{p,1-\alpha}\right\}. \tag{5.28}$$

As a final comment, we will point out that the likelihood region (5.28) is invariant to parameter transformation, while the Wald theory region of (5.25) 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 $\ell_n(h(\boldsymbol{\theta})) = \ell_n(\boldsymbol{\theta})$. Thus, any $\boldsymbol{\theta}_0$ that is contained in the set (5.28) 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 (5.25) is not invariant under parameter transformation. Any number of simulation studies have been conducted that indicate the likelihood region is superior to the Wald region in maintaining nominal coverage when the two differ, which typically occurs 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 (5.28) tends to be more difficult to compute than the Wald region of (5.25), even if $\boldsymbol{\psi}$ only contains two elements. What are called normed profile likelihoods are introduced in a later chapter and they allow the computation of confidence intervals based on inversion of likelihood ratio tests for individual elements of a parameter vector.

5.7 Numerical Algorithms for Likelihood Estimation

In the vast majority of problems for which we might choose to conduct analysis based on likelihood theory, the likelihood or log likelihood cannot be maximized analytically. The same will be true for profile likelihoods, other modified likelihoods, and composite likelihoods to be discussed in later chapters. In all of these situations our fundamental objective will be to optimize some objective

function. If the objective function is a full likelihood or log likelihood, then optimization consists of locating the maximum value, and similarly for profile, marginal, conditional, or false likelihoods. This section describes several basic algorithms for maximizing an objective function $Q(\boldsymbol{x})$, for some argument $\boldsymbol{x} \equiv (x_1, \ldots, x_p)$. If required, maximization can be accomplished by minimizing the negative objective function and it might be noted that literature on numerical analysis usually takes the problem to be minimization rather than maximization.

5.7.1 Types of Basic Optimization Algorithms

Basic numerical algorithms for optimization can be divided into three broad categories of (1) direct search algorithms, (2) gradient-based algorithms, and (3) Newton-type algorithms. These categories differ in the type of information that the algorithm must be provided, the type of information provided as output from the algorithm, and properties of the objective function required for the algorithm to be appropriate, such as continuous derivatives or not.

1. Direct Search Algorithms.

   Direct search algorithms are characterized by requiring computation of only the relevant objective function and not any derivatives of that function. Thus, direct search algorithms are useful in problems for which derivatives of the objective function are difficult to compute, or in which we do not need derivatives for the purpose of calculating inferential quantities. This will be true, for example, of confidence intervals computed from inversion of likelihood ratio tests. Direct search algorithms usually assume that the objective function is unimodal with a unique maximum and typically return only the value of the argument that maximizes the

objective function and the maximum value.

2. Gradient-Based Algorithms.

   Algorithms that fall into this class make use of the gradient or first derivatives of the objective function to help guide the direction of search. The direction of the gradient will be the path of steepest increase in the objective function. Gradient algorithms provide the same information as direct search algorithms along with the value of the gradient, which should be zero at the maximum. They are generally more efficient than direct search algorithms in terms of the number of function evaluations needed, at the cost of requiring computer functions to be written for evaluation of the first derivatives of the objective function, and typically need the first derivatives to be continuous as well.

3. Newton-type Algorithms.

   Newton-type algorithms make use of information provided by not only the objective function and gradient, but also the second derivatives of the objective function. As such, they tend to be more efficient than either direct search or gradient-based algorithms, with the obvious cost in programing of computer functions and requirement for greater smoothness of the objective function. In the case that the objective function is a full log likelihood, they provide the benefit of including the observed information matrix as part of the output, which can make inference easier if an approach based on Wald theory is to be used.

The sections that follow describe one direct search algorithm and several Newton-type algorithms. While the number of optimization algorithms that have been developed far exceeds this, the algorithms presented here have proven useful in a wide range of problems, and it could be argued that they

form a fundamental or minimal necessary set of tools that can be modified or built on to attack the computation needed for likelihood analysis.

5.7.2 Equal Interval Search

An equal interval search is the most easily comprehended and programmed direct search algorithm in one dimension. While by itself it is rarely sufficient to solve an estimation problem, it is extremely useful to quickly deal with portions of a larger and more complex problem. In addition, this algorithm is useful for investigation of likelihoods in problems for which simultatneous maximization in multiple dimensions (e.g., with a Newton-type algorithm) seems to fail. If an initial attempt to use simultaneous maximization does not appear to be working properly, a series of direct searches in one dimension each can help locate the dimension in which problems are most severe, if that is the primary cause of failure, or can help locate better starting values if applied sequentially to difference dimensions, if that is the primary cause of failure. Equal interval search algorithms also have application in Bayesian analyses for which we need to quickly locate the mode of a (posterior) distribution in one dimension.

An equal interval search algorithm for maximization of an objective function $Q(x)$ for scalar x can be described as follows.

1. Begin with the endpoints of an interval $[a, b]$ known to be within the domain of Q and to bracket the maximum value. If this is not true of an arbitrary choice of a and b that will become clear from the results returned by the algorithm.

2. Define $x_1 = a+(1/4)(b-a)$, $x_2 = a+(1/2)(b-a)$ and $x_3 = a+(3/4)(b-a)$. Note that what we have done is use three equally spaced points on the

interval $[a, b]$ to divide the interval into four sub-intervals.

3. Evaluate $Q_1 = Q(x_1)$, $Q_2 = Q(x_2)$ and $Q_3 = Q(x_3)$. Let $M = \max\{Q_1, Q_2, Q_3\}$.

4. Adjust the interval $[a, b]$ as follows:

   (a) If $M = Q_1$ replace b with x_2.

   (b) If $M = Q_3$ replace a with x_2.

   (c) If $M = Q_2$ replace b with x_3 and replace a with x_1.

5. If $x_3 - x_1 \leq \delta$ for a specified small value δ (e.g., $\delta = 10^{-8}$) then declare convergence and return $x^* = (1/2)(x_1 + x_3)$ as the value at which $Q(x)$ is maximized. Otherwise, return to step 2.

Notice that if the original interval $[a, b]$ does not bracket the maximum the algorithm will return a value that differs from a or b by less than δ. It is easy to incorporate a check for this possibility in the algorithm and to indicate that a new starting bracket is needed if this occurs.

5.7.3 The Newton-Raphson Algorithm

The Newton-Raphson algorithm is most easily understood as a simple application of Newton's method for finding the roots of equations, where that method is applied to a function that is already a first derivative. This algorithm is useful in regular problems in which estimates may be determined by solving the likelihood equations.

Newton's Method for Finding Roots

Suppose we have some function $F(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x}))^T$ for which we would like to find the value $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_p^*)$ that gives $F(\boldsymbol{x}^*) = \mathbf{0}$. If

the component functions of F have continuous derivatives in a neighborhood of $\boldsymbol{x}^*$, $N(\boldsymbol{x}^*)$ say, then for $\boldsymbol{x}^(0) \in N(\boldsymbol{x}^*)$, a Taylor expansion of F is,

$$F(\boldsymbol{x}) = F(\boldsymbol{x}^{(0)}) + F'(\boldsymbol{x}^{(0)})(\boldsymbol{x} - \boldsymbol{x}^{(0)}) + R, \tag{5.29}$$

where

$$F'(\boldsymbol{x}^{(0)}) = \left(\begin{array}{ccc} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_p} \\ \vdots & \vdots & \\ \frac{\partial f_p(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_p(\boldsymbol{x})}{\partial x_p} \end{array} \right) \Bigg|_{\boldsymbol{x}=\boldsymbol{x}^{(0)}}$$

and

$$\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right) = \left\{(x_1 - x_1^{(0)}), \ldots, (x_p - x_p^{(0)})\right\}^T.$$

Dropping the remainder term R from expression (5.29) gives an approximation to $F(\boldsymbol{x})$ and we may solve this approximation for that value of $\boldsymbol{x}$, say $\boldsymbol{x}^{(1)}$ that makes the approximation zero as,

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \{F'(\boldsymbol{x}^{(0)})\}^{-1} F(\boldsymbol{x}^{(0)}).$$

Now $\boldsymbol{x}^{(1)}$ should be closer to the solution $\boldsymbol{x}^*$ than was $\boldsymbol{x}^{(0)}$, and we may repeat the operation using $\boldsymbol{x}^{(1)}$ in place of $\boldsymbol{x}^{(0)}$ and continue in this manner for a sequence of approximations and local solutions to those approximations. In general, at the k^{th} step, we update $\boldsymbol{x}^{(k)}$ as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \{F'(\boldsymbol{x}^{(k)})\}^{-1} F(\boldsymbol{x}^{(k)}).$$

The iterative process of updating $\boldsymbol{x}^{(k)}$ to $\boldsymbol{x}^{(k+1)}$ is continued until we obtain a value $\boldsymbol{x}^{(m)}$ for which $F(\boldsymbol{x}^{(m)}) \leq \delta$ for a specified delta (e.g., $\delta = 10^{-8}$) at which time convergence is declared and we take $\boldsymbol{x}^* = \boldsymbol{x}^{(m)}$ to be the desired solution.

Newton-Raphson

If we wish to maximize an objective function $Q(\boldsymbol{x})$ which has continuous first and second derivatives, then we may do so by finding that value $\boldsymbol{x}^*$ such that $\nabla Q(\boldsymbol{x}^*) = 0$, where,

$$\nabla Q(\boldsymbol{x}) = \left(\frac{\partial Q(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial Q(\boldsymbol{x})}{\partial x_p} \right)^T$$

is the gradient of Q. If we apply Newton's method for finding roots to ∇Q we arrive at an iterative algorithm called a Newton-Raphson algorithm, defined at the m^{th} iteration as

$$\boldsymbol{x}^{(m+1)} = \boldsymbol{x}^{(m)} - \{H(\boldsymbol{x}^{(m)})\}^{-1} \nabla Q(\boldsymbol{x}^{(m)}). \tag{5.30}$$

In (5.30), $H(\boldsymbol{x})$ is the $p \times p$ Hessian matrix, with jk^{th} element,

$$H_{j,k}(\boldsymbol{x}^{(k)}) = \left. \frac{\partial^2 Q(\boldsymbol{x})}{\partial x_j \, \partial x_k} \right|_{\boldsymbol{x}=\boldsymbol{x}^{(k)}}.$$

A number of comments about Newton-Raphson algorithms are in order at this point.

1. Newton-Raphson algorithms are designed to find the value of $\boldsymbol{x}^*$ at which the gradient $\nabla Q(\boldsymbol{x}^*) = 0$. In a typical application we also want convergence of the objective function and the argument. Thus, it is beneficial to use three convergence criteria, $\nabla Q(\boldsymbol{x}) \leq \delta_1$, $Q(\boldsymbol{x}^{(k+1)} - Q(\boldsymbol{x}^{(k)}) \leq \delta_2$ and $||\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}|| \leq \delta_3$, where δ_1, δ_2 and δ_3 need not all be the same value. In particular, if the dimension of $\boldsymbol{x}$ is large it may be difficult to specify δ_3 to be as small as might be possible for δ_2. Monitoring all of these modes of convergence allows us to verify that the algorithm is working as it was designed to.

2. A Newton-Raphson algorithm requires that starting values $\boldsymbol{x}^{(0)}$ be selected for the first iteration. The algorithm is not guaranteed to converge for all starting values. But, the algorithm should converge to the same value for all starting values that lead to convergence. Use of several starting values to ensure that solutions are all the same (within tolerance of convergence criterion) is one practical technique to provide assurance that the algorithm is working properly.

3. A common modification is called *step-halving* and can sometimes be quite useful. This modification replaces the update of expression (5.30) with the following embedded iterative procedure.

   (a) At current value $\boldsymbol{x}^{(m)}$ compute the update $\boldsymbol{x}^{(m+1)}$ as in expression (5.30).

   (b) If $Q(\boldsymbol{x}^{(m+1)}) \geq Q(\boldsymbol{x}^{(m)})$ proceed to the next iteration.

   (c) If $Q(\boldsymbol{x}^{(m+1)}) < Q(\boldsymbol{x}^{(m)})$ recompute $\boldsymbol{x}^{(m+1)}$ as

$$\boldsymbol{x}^{(m+1)} = \boldsymbol{x}^{(m)} - \frac{1}{2m} \{H(\boldsymbol{x}^{(m)})\}^{-1} \nabla Q(\boldsymbol{x}^{(m)},$$

   using $m = 1, 2, \ldots$ until $Q(\boldsymbol{x}^{(m+1)}) \geq Q(\boldsymbol{x}^{(m)})$ or m exceeds some specified threshold.

While step-halving can be effective in dealing with surfaces that have extreme curvature (essentially large second derivatives) it should also be monitored to avoid false convergence. This is one of the reasons to specify an upper limit for allowable values of m and to terminate the algorithm if that limit is exceeded in a step-halving procedure.

5.7.4 Fisher Scoring

Equal interval search and Newton-Raphson algorithms have many applications outside of statistical estimation. A Fisher Scoring algorithm, in contrast, is specific to the problem of estimation and, in particular, maximization of a log likelihood function. Suppose $\boldsymbol{\theta}$ is a parameter and we seek an estimate $\hat{\boldsymbol{\theta}}_n$ that maximizes the log likelihood $\ell(\boldsymbol{\theta}) = \log\{f(\boldsymbol{y}|\boldsymbol{\theta})\}$. Suppose also that conditions are satisfied that imply the Hessian matrix, the matrix of second derivatives of $\ell(\boldsymbol{\theta})$ is consistent for the expected information, $I(\boldsymbol{\theta})$. Then replace (5.30) with the update equation,

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \left[-E\{H(\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}\right]^{-1} \nabla\ell(\boldsymbol{\theta}^{(m)}). \tag{5.31}$$

An algorithm in which the update at each iteration is computed as in (5.31) is called a Fisher Scoring algorithm.

The choice between a Newton-Raphson algorithm and a Fisher Scoring algorithm is largely a matter of computational convenience. In some problems taking the expected value of the second derivatives of a log likelihood function can result in simplified expressions. In other problems, however, the expectation operation does not have a closed form solution and we prefer to use the original Newton-Raphson update of expression (5.30).

Chapter 6

Least Squares Estimation and Inference

Least squares is generally attributed to independent work by Gauss and Legendre in the late 1790's and early 1800's. Legendre published the first account of least squares in 1805, but there is some evidence that Gauss had previously used the method to solve an estimation problem in metrology (Stigler, 1981). Certainly, least squares estimation has a long history and can be motivated based on a number of perspectives including geometry, linear algebra, and minimum variance linear unbiased estimation. The fundamental domain of application for least squares estimation is linear models with constant variance additive errors, for which least squares estimators typically have optimal small sample properties, at least within a restricted class of estimators. Note that linear models and linear estimators refer to different things, models that are linear functions of additive errors versus estimators that are linear combinations of response variables, although the two often go hand in hand. Least squares can be applied to models that have nonlinear systematic components

or error variances that are not constant, but such estimators no longer have small sample properties, and inference generally relies on asymptotic results.

6.1 Linear Algebra and Least Squares

One way to approach least squares is to view the method as the solution of a minimization problem in linear algebra. This formulation of least squares makes the notion of weights explicit, which will prove useful in motivating the use of least squares with models beyond the classical linear model with constant error variance. To formulate the problem in a linear algebra context, consider a set of real numbers $\boldsymbol{y} \equiv \{y_i : i = 1, \ldots, n\}$ as a point in $\mathbb{R}^n$. Define the inner product of two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, both in $\mathbb{R}^n$, relative to an $n \times n$ positive definite matrix W as,

$$\langle \boldsymbol{u},\, \boldsymbol{v} \rangle_W = \boldsymbol{u}^T W\, \boldsymbol{v},$$

or,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} u_i\, v_j\, w_{i,j},$$

where $w_{i,j}$ is the ij^{th} element of the matrix W. Define the norm $||\boldsymbol{u}||_W = [\langle \boldsymbol{u}, \boldsymbol{u} \rangle_W]^{1/2}$ and the distance between $\boldsymbol{u}$ and $\boldsymbol{v}$ as $\text{dist}(\boldsymbol{u}, \boldsymbol{v})_W = ||\boldsymbol{u} - \boldsymbol{v}||_W$. Now, let M_L denote a linear manifold of $\mathbb{R}^n$, and $\boldsymbol{m} = (m_1, \ldots, m_n)$ an element of M_L and consider minimizing the squared distance between $\boldsymbol{y}$ and $\boldsymbol{m}$,

$$\min_{\boldsymbol{m} \in M_L} ||\boldsymbol{y} - \boldsymbol{m}||_W^2,$$

or,

$$\min_{\boldsymbol{m} \in M_L} (\boldsymbol{y} - \boldsymbol{m})^T W\, (\boldsymbol{y} - \boldsymbol{m}).$$

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

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} (\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta})^T W\,(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta}). \tag{6.1}$$

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 (6.1) as that value $\boldsymbol{\beta}^*$ such that

$$\begin{aligned} \langle (\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta}^*),\, (\boldsymbol{X}\,\boldsymbol{\beta}^*) \rangle_W &= 0, \\ \text{or} & \\ (\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta}^*)^T W\,\boldsymbol{X}\boldsymbol{\beta}^* &= 0 \\ \Rightarrow \boldsymbol{\beta}^{*T}\boldsymbol{X}^T W \boldsymbol{y} - \boldsymbol{\beta}^{*T}\boldsymbol{X}^T W \boldsymbol{X}\boldsymbol{\beta}^* &= 0 \\ \Rightarrow \boldsymbol{X}^T W \boldsymbol{X}\boldsymbol{\beta}^* &= \boldsymbol{X}^T W \boldsymbol{y} \\ \Rightarrow (\boldsymbol{X}^T W \boldsymbol{X})^{-1}\boldsymbol{X}^T W \boldsymbol{y} &= \boldsymbol{\beta}^*. \end{aligned} \tag{6.2}$$

To express the least squares problem (6.1) and its solution (6.2) 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 $\boldsymbol{X}$.

In the formulation of the least squares problem it is possible to replace the linear manifold M_L with a nonlinear manifold M_N. Let $g(\cdot)$ be a known nonlinear function and substitute an $n \times 1$ vector $g(\boldsymbol{X}, \boldsymbol{\beta})$ for the $n \times 1$ vector $\boldsymbol{X}\boldsymbol{\beta}$ in (6.1),

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} [\boldsymbol{y} - g(\boldsymbol{X},\, \boldsymbol{\beta})]^T W [\boldsymbol{y} - g(\boldsymbol{X},\, \boldsymbol{\beta})]. \tag{6.3}$$

Continuing to write $\boldsymbol{X}$ as a matrix in (6.3) 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 projection theorem continues to give a solution to this problem as,

$$\langle (\boldsymbol{y} - g(\boldsymbol{X}, \boldsymbol{\beta})), g(\boldsymbol{X}, \boldsymbol{\beta}) \rangle_W = 0,$$

although this solution cannot be determined in closed form similar to (6.2) for the case of a linear manifold M_L.

We can now state the projection theorem in a more general form. For the majority of statistical applications we can take the Hilbert space in this theorem to be $\mathbb{R}^n$ and the inner product to be the ordinary dot product with respect to a matrix of weights W.

Theorem: Let $\boldsymbol{y}$ be in a Hilbert space V with inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_W$, let M be a subspace of V such that $\boldsymbol{y} \notin M$ and let W 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 $\boldsymbol{m} \in M$

$$||\boldsymbol{y} - \boldsymbol{m}^*||_W^2 \leq ||\boldsymbol{y} - \boldsymbol{m}||_W^2,$$

with equality if and only if $\boldsymbol{m} = \boldsymbol{m}^*$.

This discussion of least squares from a linear algebra perspective leads to several conclusions. First, there is nothing necessarily inherently statistical about least squares, which can be thought of as a solution to a minimization problem in inner product spaces. If defined in terms of projecting a vector into a linear subspace there are explicit solutions to the least squares problem but this no longer remains the case with projections into nonlinear manifolds. Finally, and perhaps most importantly, we have defined the least squares problem relative to a known matrix of weights, W. This matrix will be determined

by the particular statistical model we would like to fit using least squares estimation.

6.2 Least Squares as Statistical Estimation

So far we have not attached least squares to a statistical model or procedure. As mentioned previously, least squares is well suited for application to additive error models, consisting of a systematic component or expectation function and additive error terms. Used as a statistical estimation procedure, least squares is concerned only with parameters in the systematic model component. In the projection theorem we usually take the space V to be $\mathbb{R}^n$ and the subspace of interest M is either a linear or nonlinear manifold of $\mathbb{R}^n$ which is defined by the systematic model component as the parameters of that component range across their set of possible values. That is, the vector of expectations $\{E(Y_1), \ldots, E(Y_n)\}$ is assumed to lie in the subspace M. The projection theorem decomposes the response vector $\boldsymbol{y}$ into two parts, one that is the expectation function lying within M for a particular set of parameters and another additive component that is orthogonal to M. Least squares then, is tailored for use with additive error models and, in particular, what was referred to in Chapter 2 as the signal plus noise statistical conceptualization of a problem. As indicated at the end of the previous section, determination of the weight matrix W in least squares is determined by the particular model under consideration. It turns out that what is needed for least squares estimators to have desirable statistical properties is for W to be chosen as proportional to the inverse covariance matrix of response random variables. We now consider three versions of least squares that are differentiated based on the choice of W and discuss the associated problems of finding numerical solutions to the least

squares problem and attaching statistical properties to those estimators.

6.2.1 Ordinary Least Squares

Let $\boldsymbol{x}_i^T$ denote a $p-$vector of nonrandom covariates, and consider a linear additive error model for $i = 1, \ldots, n$,

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \epsilon_i; \quad i = 1, \ldots, n, \tag{6.4}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma > 0$, and $\epsilon_i \sim$ iid F such that F is a location scale family of distributions with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$. Usually, F is taken to be normal but that is not required to attach at least some statistical properties to the ordinary least squares (ols) estimators of $\boldsymbol{\beta}$.

Here, it is beneficial to write the model (6.4) in matrix form as,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{\epsilon},$$

in which we have $cov(\boldsymbol{\epsilon}) = I_n$, the $n \times n$ identity matrix. Take the weight matrix W in the least squares problem (6.1) to be $W = I_n^{-1} = I_n$; the reason for the initial inverse will become clear shortly. The least squares problem may be formulated as minimization in $\boldsymbol{\beta}$ of the objective function,

$$Q = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2,$$

and by (6.2) the values of $\boldsymbol{\beta}$ that solve this problem are given by,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}, \tag{6.5}$$

which are the usual ols estimators. Statistical properties are attached to $\hat{\boldsymbol{\beta}}$ in (6.5) by means of the *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 estimators $\hat{\boldsymbol{\beta}}$ we make critical use of the fact that the estimators are linear functions of the response vector $\boldsymbol{Y}$. Combining this with the Gauss-Markov result of unbiasedness, and the fact that the model gives

$$cov(\boldsymbol{Y}) = E\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\right] = E[\sigma\boldsymbol{\epsilon}\,(\sigma\boldsymbol{\epsilon})^T] = \sigma^2 I_n$$

results in

$$\begin{aligned} cov(\hat{\boldsymbol{\beta}}) &= E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right] \\ &= E\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\right]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2 I_n\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned}$$

For estimation of this covariance we replace σ^2 with an unbiased estimator,

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2,$$

which is also UMVU if the error terms are normally distributed. This estimator can be developed as a moment-based estimator that turns out to have desirable small-sample properties. Specifically, from model (6.4) we have that $\epsilon_i = Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ are independent and identically distributed with expectation 0 and second moment (also variance) σ^2. Hence, $(1/n)\sum(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$ is a consistent moment estimator of σ^2. Since $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$, replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ in this estimator retains consistency (Mann-Wald Theorem). It turns out that a simple adjustment to the denominator produces an unbiased estimator, and with the addition of normality the estimator is a function of the complete sufficient statistic and is thus UMVU.

For inference, we generally strengthen model assumptions in (6.4) 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.

6.2.2 Weighted Least Squares

Now consider a model with $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ and ϵ_i as in (6.4) and, for $i = 1, \ldots, n$,

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + (\sigma/\sqrt{c_i})\, \epsilon_i; \quad i = 1, \ldots, n, \tag{6.6}$$

where the $\{c_i : i = 1, \ldots, n\}$ are assumed to be known constants. One type of problem in which model (6.6) might arise is if the response random variables $Y_i;\ \ i = 1, \ldots, n$ are averages of independent quantities with constant variance but in which the number of quantities averaged varies over the index i. The only difference in estimation and inference for this model from the constant variance model of (6.4) is that the covariance matrix for the vector $\boldsymbol{Y}$ becomes $cov(\boldsymbol{Y}) = \sigma^2 \boldsymbol{C}^{-1}$ where $\boldsymbol{C}^{-1}$ is a diagonal $n \times n$ matrix with elements $1/c_i^2$. Then the appropriate least squares weight matrix is $W = C$. The least squares problem is to minimize in $\boldsymbol{\beta}$ the objective function,

$$Q = \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2,$$

which results in,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}. \tag{6.7}$$

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 use a bias-corrected moment estimator of σ^2,

$$S_w^2 = \frac{1}{n-p} \sum_{i=1}^{n} w_i (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2,$$

which is unbiased, consistent, and is still a function of the complete sufficient statistic for a normal model. A studentization of elements of $\hat{\boldsymbol{\beta}}$ then again results in $t-$distributions which are used to produce intervals and other inferential quantities.

6.2.3 Generalized Least Squares

The models considered for estimation with ordinary or weighted least squares are constructed as linear functions of independent error terms in which there was no relation between expected values and variances and in which the variances depended on only one unknown parameter, 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. There are many additive error models that do not obey these restrictions but for which we may still consider least squares estimation. As mentioned previously, however, optimal small sample properties will not be available.

More involved additive error models are the subject of a later chapter and we do not presume these are familiar from previous statistical presentations the reader has been exposed to. These models will consist of two components, the systematic model component or expectation function, and a separate variance

model. The systematic component may be either linear or nonlinear and the variance model may depend on the expected values or not. In any case, additive error terms will be independent following location scale distributions, and the variance model will depend on no unknown parameters other than those from the expectation function and one additional scalar σ^2. Formulation of the least squares problem will depend on the combination of expectation function and variance model structure under consideration. For example, if the expectation function is linear with parameters $\boldsymbol{\beta}$ and the variance model depends on $\boldsymbol{\beta}$ as well as σ^2, then the least squares problem becomes similar to (6.1) except with a weight matrix $W(\boldsymbol{\beta})$ rather than a fixed W that is free of parameters. If the expectation function is nonlinear $g(\boldsymbol{X}, \boldsymbol{\beta})$ but the variance model is constant depending only on σ^2then the least squares problem is given by (6.3). If the expectation function is nonlinear and the variance model depends on $\boldsymbol{\beta}$ then the least squares problem is as in (6.3) except with $W(\boldsymbol{\beta})$ rather than W.

In any of the cases just mentioned, solution of the appropriate least squares problem will require an iterative numerical procedure with quantities that change from iteration to iteration. Let the expectation function for a model be given by a known function g as $E(Y_i) = g(\boldsymbol{x}_i, \boldsymbol{\beta})$ for $i = 1, \ldots, n$. Define $\boldsymbol{V}(\boldsymbol{\beta})$ as the $n \times p$ matrix with ik^{th} element,

$$v_{i,k}(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k} g(\boldsymbol{x}_i, \boldsymbol{\beta}). \tag{6.8}$$

Also define $\tilde{\boldsymbol{Y}}(\boldsymbol{\beta}) = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$ where $\tilde{Y}_i = Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})$. Finally, let $\boldsymbol{W}(\boldsymbol{\beta})$ denote a diagonal $n \times n$ matrix with i^{th} element $w_i \propto 1/\text{var}(Y_i)$. A generalized least squares algorithm produces a sequence of estimates $\{\boldsymbol{\beta}^{(j)} : j = 1, \ldots\}$.

Let

$$\boldsymbol{V}^{(j)} = \boldsymbol{V}(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(j)}},$$
$$\tilde{\boldsymbol{Y}}^{(j)} = \tilde{\boldsymbol{Y}}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(j)}}.$$
$$\boldsymbol{W}^{(j)} = \boldsymbol{W}(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(j)}}.$$

The generalized least squares algorithm is as follows.

Generalized Least Squares

1. Calculate initial estimates $\boldsymbol{\beta}^{(0)}$.

   For $j = 0, \ldots,$

2. Calculate the $\boldsymbol{W}^{(j)}$ matrix, the $\boldsymbol{V}^{(j)}$ matrix and the $\tilde{\boldsymbol{Y}}^{(j)}$ vector for the current value $\boldsymbol{\beta}^{(j)}$.

3. Calculate the step $\boldsymbol{\delta}^{(j)}$ as,

$$\boldsymbol{\delta}^{(j)} = \left(\boldsymbol{V}^{(j)T}\boldsymbol{W}^{(j)}\boldsymbol{V}^{(j)}\right)^{-1}\boldsymbol{V}^{(j)T}\boldsymbol{W}^{(j)}\tilde{\boldsymbol{Y}}^{(j)}.$$

4. Update estimates of the expectation function parameters $\boldsymbol{\beta}$ as,

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} + \boldsymbol{\delta}^{(j)}.$$

5. If $||\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}|| < c$ for some small value c (e.g., 10^{-6}) consider the algorithm to have converged and declare $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(j+1)}$. Otherwise, update $j = j + 1$ and return to step 2.

Differences in generalized least squares algorithms appropriate for use with different specific models are entirely contained in the identities of $\boldsymbol{W}^{(j)}$, $\boldsymbol{V}^{(j)}$, and $\tilde{\boldsymbol{Y}}^{(j)}$. The algorithm includes ordinary least squares for which $\boldsymbol{W}^{(j)} = I_n$, $\boldsymbol{V}^{(j)} = \boldsymbol{X}$ and $\tilde{\boldsymbol{Y}}^{(j)} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(j)}$, and weighted least squares with $\boldsymbol{W}^{(j)} = \boldsymbol{W}$ for a fixed W, and $\boldsymbol{V}^{(j)}$ and $\tilde{\boldsymbol{Y}}^{(j)}$ as for ordinary least squares. Taking $\boldsymbol{\beta}^{(0)} = \boldsymbol{0}$ results in convergence after one iteration to (6.5) or (6.7).

6.2.4 Inference from Generalized Least Squares

Inference based on least squares estimation is concerned with inference about parameters in the systematic model component or expectation function of an additive error model. While variances will need to be estimated, it is only the need to have such estimates to construct inferential quantities about parameters of the expectation function that motivates variance estimation.

Inference connected with ordinary and weighted least squares estimation was briefly described in Chapters 6.1.2 and 6.1.3 and it is assumed that the reader is familiar with these small sample procedures. In this section, then, we present an overview of inference connected with generalized least squares estimation applied to models with nonlinear expectation functions, variances that depend on parameters in the expectation function, or both. The Gauss Markov theorem does not apply to these models, and assuming normally distributed error terms does not necessarily result in explicitly identifiable sampling distributions for estimators. Thus, inference is approximate, based on a result that has been called the *Fundamental Theorem of Generalized Least Squares*. The context for this theorem is an additive error model such that $E(Y_i) = g(\boldsymbol{x}_i, \boldsymbol{\beta})$ and $\text{var}(Y_i) = w_i^{-1/2}\sigma^2$, where g is a known smooth function and w_i may be a constant or a function of $\boldsymbol{\beta}$. The quantities $\boldsymbol{V}(\boldsymbol{\beta})$ and $\boldsymbol{W}(\boldsymbol{\beta})$ are as previously defined, and we now include an index for sample size n to emphasize that we are concerned with the behavior of sequences of estimators as sample size grows without bound.

Fundamental Theorem of Generalized Least Squares

Under a set of fairly mild conditions, if the starting value $\boldsymbol{\beta}_n^{(0)}$ is $n^{1/2}-$consistent for $\boldsymbol{\beta}$, and for any number of iterations j in the generalized least squares

algorithm,

$$\boldsymbol{\beta}_n^{(j)} \text{ is } AN\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\Sigma_\beta^{-1}\right), \tag{6.9}$$

where

$$\Sigma_\beta = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{\beta})\boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{\beta})^T / w_i^2. \tag{6.10}$$

In (6.10) w_i will be determined by the particular model of concern and $\boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{\beta})$ is a $p \times 1$ column vector with k^{th} element

$$v_k(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k} g(\boldsymbol{x}_i, \boldsymbol{\beta}).$$

Note that this $\boldsymbol{v}_k(\boldsymbol{x}_i, \boldsymbol{\beta})$ is the ik^{th} element of $\boldsymbol{V}(\boldsymbol{\beta})$.

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. This means, for example, that if we take a starting value $\boldsymbol{\beta}_n^{(0)}$ and conduct $j = 1$ iterations of the algorithm we end up with the same asymptotic normality as if we iterate until $\boldsymbol{\beta}_n^{(j+1)} = \boldsymbol{\beta}_n^{(j)} + \delta$. This is not at all obvious, but recall that one of the conditions of this theorem is that $\beta_n^{(0)}$ constitute a root n consistent estimator for $\boldsymbol{\beta}$. An estimator $\hat{\theta}_n$ is $n^{1/2}-$consistent for a parameter θ if $n^{1/2}(\hat{\theta}_n - \theta)$ is bounded in probability (meaning that $\hat{\theta}_n$ converges to θ at least at the rate $1/n^{1/2}$). 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 **?**, 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 $\boldsymbol{\beta}_n^{(j+1)} = \boldsymbol{\beta}_n^{(j)}$ but in practice means $|\boldsymbol{\beta}_n^{(j+1)} - \boldsymbol{\beta}_n^{(j)}| < \delta$ 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 **?**.

Estimation of σ_n^2 is generally accomplished through the use of a moment-based estimator. Let $\hat{\boldsymbol{\beta}}_n$ denote a generalized least squares estimator of $\boldsymbol{\beta}$ based on n observations. Then σ^2 is typically estimated as,

$$\hat{\sigma}_n^2 = \frac{1}{n-p} \sum_{i=1}^{n} \left\{ \frac{Y_i - g(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_n)}{w_i} \right\}^2 . \tag{6.11}$$

Note that except in the special cases leading to models (6.4) or (6.6) this estimator no longer possesses any small sample properties, despite the divisor of $n-p$ which suggests it might be unbiased (it is not). It is, however, consistent as long as $\hat{\boldsymbol{\beta}}_n$ is consistent which was a condition of the theorem.

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 (6.9), with estimated variances produced by plug-in use of $\hat{\sigma}_n^2$ from (6.11) and $\hat{\boldsymbol{\beta}}_n$ from the generalized least squares algorithm. The asymptotic normality continues to hold, and we behave as if $\hat{\boldsymbol{\beta}}_n$ has a multivariate normal

distribution with expected value $\boldsymbol{\beta}$ and covariance matrix,

$$\hat{C}(\hat{\boldsymbol{\beta}}_n) = \frac{\hat{\sigma}_n^2}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}_n)\boldsymbol{v}(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}_n)^T/\hat{w}_i\right]^{-1}. \tag{6.12}$$

Notice that the term in square brackets on the right side of (6.12) is a matrix so that $[]^{-1}$ denotes the inverse of this matrix, not a simple scalar reciprocal. The hat notation for w_i is because these weights may be a function of the $\hat{\boldsymbol{\beta}}_n$. Interval estimates are then computed in the usual way. For an individual element β_k of $\boldsymbol{\beta}$ this is

$$\begin{aligned}\hat{\beta}_{n,k} &\pm t_{1-\alpha/2;n-p}\left\{\hat{C}(\hat{\boldsymbol{\beta}}_n)_{k,k}\right\}^{1/2}\\ \textit{or}&\\ \hat{\beta}_{n,k} &\pm z_{1-\alpha/2}\left\{\hat{C}(\hat{\boldsymbol{\beta}}_n)_{k,k}\right\}^{1/2}\end{aligned} \tag{6.13}$$

where $\hat{C}(\hat{\boldsymbol{\beta}}_n)_{k,k}$ is the k^{th} diagonal element of the estimated covariance matrix given in (6.12), $t_{1-\alpha/2;n-p}$ 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. Note that there is absolutely no justification for using the quantile of a $t-$distribution in these intervals rather than that of a standard normal distribution. These intervals are constructed on the basis of an asymptotic result, but you will unfortunately see such things done in the literature.

For inference concerning tests of hypotheses about parameter values 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 $\hat{\boldsymbol{\beta}}_n$ is a generalized least squares estimator. Based on the asymptotic normality of the generalized least squares estimator $\hat{\boldsymbol{\beta}}_n$, what was presented under the heading of Wald Theory in Chapter 5.5 applies, under slightly modified conditions from what was presented there in the context of likelihood estimation.

In particular, the delta method can be used to construct confidence intervals for functions of $\boldsymbol{\beta}$.

6.3 Summary of Least Squares Estimation and Inference

We conclude our consideration of least squares as a method of estimation and inference by summarizing a number of key points:

1. Least squares is used nearly exclusively with additive error models and is concerned primarily with parameters of the expectation function or systematic model component.

2. Least squares can be motivated in a number of ways, including as the solution to a general minimization problem in linear algebra.

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 so that estimated marginal sampling distributions of expectation function parameters are $t-$distributions, which are then used to construct intervals and tests.

4. For nonlinear models with constant variance, or for either linear or nonlinear models with variances that depend on parameters of the expectation function and σ^2, but no additional unknown parameters, generalized least squares estimators are asymptotically normal. 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. Note that the development of the approximate sampling distribution of expectation function parameters *does not* depend on the additional model assumption of normally distributed errors. 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 but only up to the point of verifying optimal behavior of *point* estimators. We then generally 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. But this concept of robustness is different than what is properly called *resistance*, which refers to the de-

gree 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 and thus have low resistance.

Chapter 7

Bayesian Fundamentals

This chapter introduces some basic concepts in Bayesian analysis of statistical problems and the fundamental distributions needed to put this type of an analysis into action. First note that we often talk about *the* Bayesian approach to analysis as if there was one standard conceptual basis that all Bayesian approaches to statistics employ. This is no more true than that there is one standard conceptual basis for all non-Bayesian analysis. There have been, and continue to be, a number of schools of thought about Bayesian analysis, just as there are a number of schools of thought about non-Bayesian approaches, such as design-based inference in contrast with model-based inference.

In the first section of this chapter one of those Bayesian schools of thought is outlined, drawing heavily on expositions by the statistician and physicist E.T. Jaynes, especially Jaynes (1986). Jaynes makes it clear that the distinction between Bayesian and what he calls orthodox approaches rests on interpretation of probability, not other areas of controversy.

7.1 Bayesian Concepts

Bayesian methods are sometimes characterized in a broad-brush manner as subjective, or involving personal reality, because of the use of a prior distribution. And there are indeed schools of thought in which probabilities are viewed as inherently subjective. But there was and is also at least one line of reasoning in the development of Bayesian methods that holds that there is, in fact, an absolute truth to the order of the universe, called the *true state of nature* in the historical literature. This view of Bayesian analysis is in total agreement with an extreme reductionist view of the world in which, if all forces in operation were known, observable quantities would be completely deterministic.

7.1.1 Parameters and Epistemic Probability

In the approach to Bayesian analysis presented here, the true state of nature is embodied in a fixed, but unknown parameter value that governs the distribution of observable quantities. Note that this sounds quite a bit like a typical frequentist idea, and that is the point. There is not necessarily anything different about the concept of controlling parameters between Bayesian and non-Bayesian approaches to statistical analysis. In fact, all of the issues addressed in Chapter 2 of these notes relative to scientific abstraction, statistical abstraction, and statistical modeling are just as pertinent for a Bayesian analysis as they are for a non-Bayesian analysis. As (Jaynes, 1986, p. 11) puts it,

> For decades, Bayesians have been accused of supposing that an unknown parameter is a random variable; and we have denied hundreds of times . . . that we are making any such assumption.

It may seem natural to suppose that the use of Bayes theorem must be what makes a Bayesian analysis Bayesian, but this is no more true than the fallacy that all Bayesians consider parameters to be random. Bayes theorem is a probability result, true for any legitimate concept of probability, and Bayes theorem has many uses other than in a Bayesian analysis.

Example 7.1

A classical non-Bayesian application of Bayes theorem is to problems in medical diagnostic testing. Jegerlehner, Suter-Riniker, Jent, Bittel, and Nagler (2021) studied the diagnostic accuracy of a rapid antigen test for COVID-19. Subjects were recruited from individuals visiting a COVID-19 testing facility in Sweden. Patients were tested simultaneously using the rapid antigen test and the gold standard PCR test. Let D be the event that an individual has COVID and N^c the event of no disease. Let P be the event of a positive result on the rapid antigen test and P^c the event of a negative result from that test. Of 1465 subjects, the PCR test was positive in 141 individuals and prevalence was considered to be $Pr(D) = 0.0962$. A total of 1462 individuals had enough sample material from nasal swabs to complete both the PCR and the rapid antigen test. Results for these individuals are given in the following table.

	D	D^c	Total
P	92	2	94
P^c	49	1319	1368
Total	141	1321	1462

From these values we can estimate sensitivity as $hat Pr(P|D) = 0.6525$, reported in the paper as 0.653, and specificity as $\hat{Pr}(P^c|D^c) = 0.9985$, reported in the paper as 0.999. Using Bayes theorem, the estimated probabilities that an individual who tests positive using the rapid antigen test does in fact have

COVID-19 and the probability that an individual who tests negative is in fact free of COVID-19 are then,

$$\begin{aligned}\hat{Pr}(D|P) &= \frac{\hat{Pr}(P|D)\,Pr(D)}{\hat{Pr}(P|D)\,Pr(D) + \hat{Pr}(P|D^c)\,Pr(D^c)} \\ &= \frac{0.6525(0.0962)}{0.6525(0.0962) + 0.0015(0.9038)} = 0.9788. \\ \hat{Pr}(D^c|P^c) &= \frac{\hat{Pr}(P^c|D^c)\,Pr(D^c)}{\hat{Pr}(P^c|D^c)\,Pr(D^c) + \hat{Pr}(P^c|D)\,Pr(D)} \\ &= \frac{0.9985(0.9038)}{0.9985(0.9038) + 0.3475(0.0926)} = 0.9643,\end{aligned}$$

which would suggest that the rapid antigen test is effective as a screening tool for COVID. A criticism of this study, however, is that the total number of individuals participating, 1465, were obtained as a convenience sample and may not be representative of the target population for inference. In fact, the authors report that of the total 1465 individuals visiting the medical facility to get tested for COVID-19, 1114 did so because they believed they were experiencing symptoms. This has an impact on the marginal probability $Pr(D)$ which is a driving factor in determining the conditional probabilities $Pr(D|P)$ and $Pr(D^c|P^c)$. It is hard to accept that a screening tool with sensitivity less than 0.75 could be effective in the general population, although it may be so in the population of individuals who seek a diagnostic test.

So taking parameters to be random variables does not distinguish a Bayesian analysis from a non-Bayesian one, and neither does the use of Bayes theorem. The defining characteristic of Bayesian methods is the use of an epistemic concept of probability for making inference (see Chapter 1). An epistemic concept of probability holds that probability is the language of belief or knowledge. Statements of probability for one time events, such as the probability that

your home team has a winning football season this year, the probability that some nation successfully develops a nuclear weapon within the next five years, or the probability that birds evolved from dinosaurs, are statements of epistemic probability (unless one believes in parallel universes). If one is willing to accept probability statements about events such as these as legitimate expressions of belief, then they accept an epistemic concept of probability. Again, Jaynes (1986, p. 11) makes this point plainly,

> In Bayesian parameter estimation, both the prior and posterior distributions represent, not any measurable property of the parameter, but only our own state of knowledge about it.

Suppose we have formulated a model on the basis of random variables connected with observable quantities, $Y_1, \ldots, Y_n$ in terms of a parametric probability mass function or probability density function $f(\boldsymbol{y}|\boldsymbol{\theta})$. Following notions about the process of developing useful models discussed in Chapter 2, we assume that the scientific mechanism of interest or, in the terminology of the current section, the true state of nature,is captured in the parameter $\boldsymbol{\theta}$, or perhaps some function of it. If we admit an epistemic concept of probability, we are free to represent our current knowledge or belief about the possible values of $\boldsymbol{\theta}$ as a probability distribution. When we do so before seeing data we will call such a distribution a *prior* distribution. When we do so after having seen data we obtain a *posterior* distribution. Note that, $f(\boldsymbol{y}|\boldsymbol{\theta})$ may be interpreted through a hypothetical limiting relative frequency concept of probability. It is distributions assigned to, or derived for, $\boldsymbol{\theta}$ that represent epistemic probability. Despite this, the mathematics of dealing with distributions on (what we believe about) $\boldsymbol{\theta}$ will be identical to what would be the case if $\boldsymbol{\theta}$ were considered a random variable subject to investigation via relative frequency.

This is because any legitimate concept of probability, including both relative frequency and epistemic probability, must obey the same mathematical rules of behavior.

7.1.2 Basic Distributions

There are a number of fundamental distributions inherent to a basic Bayesian analysis. The foremost of these are the data model, the prior distribution, and the posterior distribution. There are also what are known as predictive distributions, the principle of these being the posterior predictive distribution. We assume we are considering a problem that will be approached by formulating a model for random variables $Y_1, \ldots, Y_n$ that correspond to observable quantities. At this point we will make no other assumptions about these variables. They may be discrete or continuous, independent or not independent, and bounded or unbounded in possible values.

The *data model* in Bayesian analysis is the joint distribution for observable response variables $Y_1, \ldots, Y_n$, just as in any statistical model. Indeed, everything that has been discussed in this book concerning scientific abstraction, statistical abstraction, and model formulation applies equally to a Bayesian analysis as it does to any other model-based approach. We will assume that the joint distribution of response random variables depends on a parameter $\boldsymbol{\theta} \in \Theta$, which may be either a scalar or vector-valued, and we will write the data model as $f(\boldsymbol{y}|\boldsymbol{\theta})$. Also, as mentioned in the previous subsection, the data model may be interpreted according to relative frequency probability, the distribution representing frequencies that might occur in the limit with at least hypothetical repeated sampling.

A *prior distribution* is a distribution, usually in the form of a probability

density function, placed on what we believe about the possible values of the data model parameter before or *prior to* observing any data. With $\boldsymbol{\theta}$ denoting the data model parameter, the prior may be written as $\pi(\boldsymbol{\theta})$. The support of this prior must be contained in the data model parameter space Θ and the prior is often chosen so that these two sets are identical. If $\boldsymbol{\theta}$ is a vector, the prior distribution is a joint distribution. The prior distribution may itself depend on a parameter $\boldsymbol{\lambda}$, and priors are frequently written as either $\pi(\boldsymbol{\theta})$ or $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$. This is because if the prior involves a parameter $\boldsymbol{\lambda}$, specific numeric values will be chosen for $\boldsymbol{\lambda}$ before the analysis is conducted. The notation $\pi(\boldsymbol{\theta})$ makes it clear that the prior does not depend on any additional *unknown* parameters. The notation $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is sometimes convenient because $\boldsymbol{\lambda}$ may be involved in inferential quantities in the analysis. What is important to remember is that, if a prior is written as $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, in conducting an actual analysis specific values are chosen for the elements of $\boldsymbol{\lambda}$.

The *posterior distribution* is derived from the data model and the prior, and is a distribution for what we believe about the possible values of $\boldsymbol{\theta}$ after or *posterior to* seeing observed data. Assuming, as is usually the case, that the possible values of $\boldsymbol{\theta}$ are continuous in Θ, the posterior distribution is constructed as,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{m(\boldsymbol{\theta}, \boldsymbol{y})}{h(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}. \tag{7.1}$$

The posterior (7.1) is in the form of Bayes theorem which, given what is usually called the law of total probability (e.g., DeGroot and Schervish 2002 p. 67), follows directly from the definition of conditional probability density functions. It should be noted from the outset that the denominator of (7.1) is constant for the posterior distribution and thus $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto m(\boldsymbol{y}, \boldsymbol{\theta})$. This fact will often facilitate derivation of the posterior without the need for formal evaluation of

the integral that leads to $h(\boldsymbol{y})$.

If $\boldsymbol{\theta}$ is a vector, (7.1) is a joint distribution and as such it determines a set of additional posterior distributions that are often used in Bayesian analysis. For the purposes of inference we often want to use marginal posteriors such as $p(\theta_k|\boldsymbol{y})$ for the k^{th} element of $\boldsymbol{\theta}$. Also useful will often be conditional posteriors $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y})$ where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are disjoint subsets of $\boldsymbol{\theta}$. In particular, what are called *full conditional* posteriors, which occur if $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ partition $\boldsymbol{\theta}$. Often, in fact, $\boldsymbol{\theta}_1$ is an individual element of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_2$ are all of the remaining elements. Conditional posteriors play a role in certain algorithms used to simulate values from complex joint posterior distributions, a future topic.

There are also several *predictive* distributions involved in a Bayesian analysis. Consider a new set of observations $\boldsymbol{Y}^0$, assumed to follow the same data model as $\boldsymbol{Y}$, $f(\boldsymbol{y}|\boldsymbol{\theta})$. The *prior predictive* distribution is

$$p(\boldsymbol{y}^0) = \int f(\boldsymbol{y}^0|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{7.2}$$

and the *posterior predictive* distribution is,

$$p(\boldsymbol{y}^0|\boldsymbol{y}) = \int f(\boldsymbol{y}^0|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \tag{7.3}$$

Prior and posterior predictive distributions for elements of $\boldsymbol{Y}^0$ are the marginals determined by (7.2) and (7.3). Prior predictive distributions are used in problems that are sequential in nature that are beyond the scope of this work, but they can also play a role in model assessment. Posterior predictive distributions may also be used in those problems, but are important largely because they are central to the assessment of estimated models.

7.2 Basic Estimation and Inference

A strength of a Bayesian approach to analysis is that inference is wonderfully simple and intuitive. Given either a prior distribution $\pi(\boldsymbol{\theta})$ that represents our beliefs about the possible values of $\boldsymbol{\theta}$ before seeing the data, or a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ that represents our beliefs about the possible values of $\boldsymbol{\theta}$ after seeing the data, inference consists of making probability statements on the basis of those distributions. If the prior is used we are making *prior inference*, if the posterior is used we are making *posterior inference*. Except in sequential problems or with what are called dynamic models, inference is almost exclusively based on the posterior, and we will assume that is the case here. In addition we assume that, if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, the marginal posteriors corresponding to $p(\boldsymbol{\theta}|\boldsymbol{y})$ are available. These will be denoted as $p(\theta_k|\boldsymbol{y})$ for $k = 1, \ldots, p$.

7.2.1 Derivation of Posterior Distributions

The fundamental mathematical operation in a Bayesian analysis is combining the prior distribution and the likelihood to arrive at a posterior distribution via 7.1that represents our knowledge or belief about the possible values of the data model parameter. This topic will be revisited at several points in our overall presentation, becoming more complex when we discuss simulation of values from posteriors and, especially, when the topic becomes accomplishing such simulation based on the use of Markov Chain Monte Carlo methods. Here, we give some more basic and simpler examples to make the basic process concrete.

Example 7.2

Suppose we have a data model that consists of n independent and identically distributed random variables $Y_1, \ldots, Y_n$ with a common Poisson distribution having parameter $\lambda > 0$. Suppose further that we assign λ a prior in the form of an exponential distribution with parameter $\beta > 0$. Recall here that in an actual analysis β will be selected to have a particular numerical value, such as 0.5. The three basic distributions we need to deal with are then,

- Data Model: $f(\boldsymbol{y}|\lambda) = \left(\prod_{i=1}^{n} \frac{1}{y_i!}\right) \lambda^{\sum_{i=1}^{n} y_i} \exp(-n\lambda)$.
- Prior: $\pi(\lambda) = \beta \exp(-\beta\lambda)$.
- Posterior:

$$\begin{aligned} p(\lambda|\boldsymbol{y}) &= \frac{\beta \exp(-\beta\lambda) \left(\prod_{i=1}^{n} \frac{1}{y_i!}\right) \lambda^{\sum_{i=1}^{n} y_i} \exp(-n\lambda)}{\beta \left(\prod_{i=1}^{n} \frac{1}{y_i!}\right) \int_0^{\infty} \lambda^{\sum_{i=1}^{n} y_i} \exp(-n\beta\lambda)\, d\lambda} \\ &= \frac{\Gamma\left(1 + \sum_{i=1}^{n} y_i\right)}{(n+\beta)^{1+\sum_{i=1}^{n} y_i}}, \end{aligned}$$

  which is the probability density function of a gamma distribution with parameters $1 + \sum_{i=1}^{n} y_i$ and $n + \beta$.

As mentioned previously it is often convenient to use the fact that for a data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and prior $\pi(\theta)$, the posterior is $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta} 0\, \pi(\boldsymbol{\theta})$ and then simply recognize the right hand side as the kernel of some identifiable distribution. When we write this expression, only terms that contain the argument $\boldsymbol{\theta}$ need to be retained. This is true in this example, and we have

$$\begin{aligned} p(\lambda|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\lambda)\, \pi(\lambda) \\ &\propto \lambda^{\sum_{i=1}^{n} y_i} \exp[-(n+\beta)\lambda], \end{aligned}$$

which we recognize as the kernel of a gamma distribution with parameters $\sum_{i=1}^{n} y_i + 1$ and $n + \beta$.

A classic technique in the derivation of posterior distributions is completing the square in normal distributional forms. The following result is used many times in Bayesian analysis.

Completing the Square

Consider a random variable X that follows a probability distribution,

$$f(x|A, B) \propto \exp\left[-\frac{1}{2}(Ax^2 - 2Bx)\right].$$

Then X has a normal distribution with mean M and variance V where,

$$M = \frac{B}{A},$$
$$V = \frac{1}{A}$$

The proof of this result is left as an exercise. Completing the square is useful in many problems.

Example 7.3

Consider a one-sample normal model in which random variables $Y_1, \ldots, Y_n$ are taken as being independent and identically distributed with common probability density function, for some $-\infty < \mu < \infty$ and known $\sigma^2 > 0$,

$$f(y|\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]; \ -\infty < y < \infty.$$

Suppose that μ has been assigned a prior in the form of a normal distribution with expected value $-\infty < \lambda < \infty$ and known variance $\tau^2 > 0$,

$$\pi(\mu) = \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left[-\frac{1}{2\tau^2}(\mu - \lambda)^2\right]; \ -\infty < \mu < \infty.$$

The posterior of μ is then,

$$\begin{aligned} p(\mu|\boldsymbol{y}) &\propto \pi(\mu)\, f(\boldsymbol{y}|\mu) \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{1}{2\tau^2}(\mu-\lambda)^2)\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2\right) - \frac{1}{2\tau^2}\left(\mu^2 - 2\mu\lambda + \lambda^2\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) - 2\mu\left(\frac{n\bar{y}}{\sigma^2}+\frac{\lambda}{\tau^2}\right)\right\}\right], \end{aligned}$$

which, by the result, corresponds to a normal distribution with expected value M and variance V where,

$$\begin{aligned} M &= \frac{n\tau^2\bar{y} + \sigma^2\lambda}{n\tau^2+\sigma^2} \\ V &= \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}. \end{aligned}$$

Note that M is in the form of a weighted average of the prior mean λ and the data model mle $\bar{y}$. This weighted average can also be written as

$$M = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau^2}\lambda}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$$

which shows that the weights are inverse variances of $\bar{y}$ and λ.

Example 7.4

Consider a simple linear regression model for $i = 1, \ldots, n$, with $-\infty < \beta_0 < \infty$, $-\infty < \beta_1 < \infty$, and known error variance σ^2,

$$Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i,$$

where $\epsilon_i \sim$ iid N$(0,1)$. Suppose that β_0 and β_1 have been assigned prior distributions, for given values $-\infty < \lambda_0 < \infty$, $-\infty < \lambda_1 < \infty$, $\tau_0^2 > 0$ and

$\tau_1^2 > 0$,

$$\pi_0(\beta_0) = \frac{1}{(2\pi\tau_0^2)^{1/2}} \exp\left[-\frac{1}{2\tau_0^2}(\beta_0 - \lambda_0)^2\right]$$
$$\pi_1(\beta_1) = \frac{1}{(2\pi\tau_1^2)^{1/2}} \exp\left[-\frac{1}{2\tau_1^2}(\beta_1 - \lambda_1)^2\right]$$
$$\pi(\beta_0, \beta_1) = \pi_0(\beta_0)\, \pi_1(\beta_1).$$

The joint posterior distribution of β_0 and β_1 is given by,

$$p(\beta_0, \beta_1|\boldsymbol{y}) \propto \pi(\beta_0, \beta_1)\, f(\boldsymbol{y}|\beta_0, \beta_1).$$

This joint posterior cannot be derived in closed form, but we could derive the conditional posterior of β_0 given β_1 and the conditional posterior of β_1 given β_0. For the latter,

$$p_1(\beta_1|\beta_0, \boldsymbol{y}) \propto \pi_1(\beta_1) f(\boldsymbol{y}|\beta_0, \beta_1)$$
$$\propto \exp\left[-\frac{1}{2\tau_1^2}(\beta_1 - \lambda_1)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$
$$p_1(\beta_1|\beta_0, \boldsymbol{y}) \propto \exp\left[-\frac{1}{2}\left\{\beta_1^2\left(\frac{1}{\tau_1^2} + \frac{\sum_{i=1}^n x_i^2}{\sigma^2}\right) - 2\beta_1\left(\frac{\lambda_1}{\tau_1^2} + \frac{\sum_{i=1}^n x_i(y_i - \beta_0)}{\sigma^2}\right)\right\}\right],$$

and the result on completing the square gives that this conditional posterior is normal with mean M and variance V where,

$$M = \frac{\sigma^2\lambda_1 + \tau_1^2\sum_{i=1}^{n} x_i(y_i - \beta_0)}{\sigma^2 + \tau_1^2\sum_{i=1}^{n} x_i^2}$$
$$V = \frac{\sigma^2\tau_1^2}{\sigma^2 + \tau_1^2\sum_{i=1}^{n} x_i^2}$$

Such conditional posteriors will play an important role in determining posterior distributions in certain more complex models.

7.2.2 Point Estimation

Consider a univariate posterior for an element of $\boldsymbol{\theta}$. A point estimate of θ_k is usually one of the quantities that describe the location of the posterior, that is, the mean, median, or mode of $p(\theta_k|\boldsymbol{y})$. The posterior mode was, at one time, the most popular of the three because it can be located without finding the denominator of (7.1) which is the normalizing constant of the posterior. The posterior mean is probably the most used today, in part because of the extensive use of simulation to approximate posterior distributions, a topic that will be discussed in greater detail in what follows. The posterior mean or expected value can also be justified based on decision-theoretic grounds, if one considers squared error loss (e.g., Berger, 1985, p. 161). Moments of posterior distributions, such as means, variances, and covariances, are given by the usual definitions, again because probability distributions must obey the same mathematical rules regardless of the concept of probability they are interpreted under. Rules for expectations, variances and quantiles also are the same as for random variables, although we are not considering $\boldsymbol{\theta}$ or any of its elements to actually be random. The special status of posteriors as being expressions of epistemic probability impact interpretation not manipulation.

7.2.3 Interval Estimation

Although posterior variances and covariances can be, and often are, computed as summary quantities that describe posterior distributions, they are 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*. The basic

definition of a credible set for $\boldsymbol{\theta}$ is a set $\mathcal{C}$ such that

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in \mathcal{C}|\boldsymbol{y}) = \int_{\mathcal{C}} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \tag{7.4}$$

If $\boldsymbol{\theta}$ should happen to be discrete the integral in (7.4) is replaced with a summation, but this is unusual. Credible intervals for individual components of $\boldsymbol{\theta}$ are computed using the appropriate marginal posterior.

To emphasize the distinction between the interpretation of inferential statements made on the basis of epistemic probability and on the basis of relative frequency probability, consider the statement for a scalar parameter θ,

$$Pr(a < \theta < b) = 1 - \alpha. \tag{7.5}$$

If this probability is interpreted according to a relative frequency concept, the quantities a and b must be random variables, such as $a = \hat{\theta}(\boldsymbol{Y}) - z_{1-\alpha/2}\text{se}(\hat{\theta}(\boldsymbol{Y}))$ and $b = \hat{\theta}(\boldsymbol{Y}) - z_{\alpha/2}\text{se}(\hat{\theta}(\boldsymbol{Y}))$, where z_α is the α quantile of a standard normal distribution. The interpretation is based on repeated sampling and the limiting relative frequency with which the event inside the probability is true. In contrast, if (7.5) is interpreted under an epistemic concept of probability then a and b can be constants and the meaning of the interval is that it contains $(1 - \alpha)100\%$ of our belief about where the value of θ might lie.

For a given posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ there may be many sets that satisfy the original probability statement (7.4). One technique that has been used to help get around this difficulty is to define a *Highest Posterior Density* credible 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 $\boldsymbol{\theta} \notin \mathcal{C}^*$,

$$p(\boldsymbol{\theta}^*|\boldsymbol{y}) \geq p(\boldsymbol{\theta}|\boldsymbol{y}).$$

In other words, the posterior density for any value of $\boldsymbol{\theta}$ included in a highest posterior density credible set is at least as great as that for any value not in the 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 Berger (see 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 based on its marginal posterior $p(\theta_j|\boldsymbol{y})$, that interval is given by (L, U) where,

$$\begin{aligned} \alpha/2 &= \int_{-\infty}^{L} p(\theta_j|\boldsymbol{y})\, d\theta_j \\ \alpha/2 &= \int_{U}^{\infty} p(\theta_j|\boldsymbol{y})\, d\theta_j. \end{aligned} \tag{7.6}$$

Example 7.5

Using the data as given in the table of Example 7.1, point estimates and 95% confidence intervals are 0.6525 $(0.5739, 0.7311)$ for sensitivity and 0.9985 $(0.9964, 1.0006)$ for specificity. These values are computed based on an assumption that given the total number of diseased individuals, the number of positive rapid test results follows a binomial distribution and similarly for the number of negative results given the number of non-diseased individuals. These estimates and intervals are computed based on either unbiased estimation and the

central limit theorem, or maximum likelihood theory, which lead to the same values. Notice that the upper confidence limit for specificity extends beyond the parameter space for a binomial distribution. With a binomial data model $Y \sim \text{Binom}(\theta, p)$ we might choose a beta prior distribution, $\theta \sim \text{Beta}(\alpha, \beta)$, based on the agreement of support for a beta distribution and the parameter space of a binomial distribution. The posterior is then, for $0 < \theta < 1$,

$$\begin{aligned} p(\theta|y) &\propto f(y|\theta)\,\pi(\theta|\alpha,\beta) \\ &\propto \theta^{y}(1-\theta)^{n-y}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}, \end{aligned}$$

which can be recognized as the kernel of a beta density with parameters $\alpha + y$ and $\beta + n - y$. Notice that in the first line for $p(\theta|y)$ we have dropped any terms from the binomial and beta distributions that do not depend on the argument θ. Applied to the data of Example 7.1 with arbitrary choices of $\alpha = 3$ and $\beta = 2$, the posterior for sensitivity is a beta distribution with parameters 95 and 51 which leads to a posterior expected value of 0.6507 and a 95% hpd credible interval of (0.5718, 0.7257). For specificity the posterior is a beta distribution with parameters 1322 and 4 which gives expected value 0.9970 and interval (0.9934, 0.9992).

7.2.4 Hypothesis Testing

Consider testing a hypothesis about the value of an element of $\boldsymbol{\theta}$, $H_0 : \theta_k \in \Theta_0$ versus $H_1 : \theta_k \notin \Theta_0$. A Bayesian hypothesis test can be constructed as the ratio of posterior probabilities of the two hypotheses,

$$B(H_1, H_0) = \frac{Pr(H_1|\boldsymbol{y})}{Pr(H_0|\boldsymbol{y})}. \tag{7.7}$$

Interpreting the value of this test statistic depends somewhat on the approach one is taking toward the testing problem. If the testing problem is being approached as choosing between H_0 and H_1, then determining a decision rule for (7.7) is just as arbitrary as setting a rate for type I errors in frequentist hypothesis testing. One could make that choice depending on which has greater posterior probability or, equivalently, choose H_1 if $Pr(H_1|\boldsymbol{y}) > 0.50$ and choose H_0 otherwise. This is what is used in the presentation of **?**, p. 379 but who also point out that one could choose other values. Given that a posterior represents our belief about possible values of the parameter, the testing problem can also be approached as an assessment of the evidence provided by the data in favor of H_1. In this case, we don't really need explicit status for H_0 as a hypothesis *per se*. We simply want to assess the evidence for H_1, and do so by computing the posterior odds of H_1 as $Pr(H_1)/Pr(H_1^c)$.

Note that (7.7) is a special case of a more general quantity known as a Bayes Factor. Bayes Factors assess the change from prior odds to posterior odds of two models. For our purposes here, consider a model to consist of a data model, a prior distribution on the data model parameters, and the resulting posterior for those parameters. Consider two models M_1 and M_2, say. The posterior odds of model M_1 relative to model M_2 are,

$$\frac{Pr(M_1|\boldsymbol{y})}{Pr(M_2|\boldsymbol{y})} = \frac{Pr(M_1)}{Pr(M_2)} \frac{Pr(\boldsymbol{y}|M_1)}{Pr(\boldsymbol{y}|M_2)} \tag{7.8}$$

$$= \frac{Pr(M_1)}{Pr(M_2)} BF(M_1, M_2). \tag{7.9}$$

The quantity $BF(M_1, M_2)$ is called the Bayes Factor in favor of model M_1 relative to model M_2. From 7.8 we can arrive at a number of alternative ways

to represent a Bayes Factor,

$$BF(M_1, M_2) = \frac{Pr(\boldsymbol{y}|M_1)}{Pr(\boldsymbol{y}|M_2)} \tag{7.10}$$

$$= \frac{Pr(M_1|\boldsymbol{y})}{Pr(M_2|\boldsymbol{y})} \Big/ \frac{Pr(M_1)}{Pr(M_2)}. \tag{7.11}$$

The last expression in 7.10 is the ratio of posterior odds in favor of model M_1 to prior odds in favor of model M_1.

A number of scales of evidence for assessing Bayes Factors have been suggested in the literature. 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 categories of evidence for ranges of Bayes factors 3 to 20 (some evidence), 20 to 150 (strong evidence) and greater than 150 (decisive evidence).

Bayes Factors have a number of subtle and not entirely pleasing properties, which we will not go into here. Most of these deal with different aspects of what we take to be a model in formulating M_1 and M_2. Here, we will stick with a reasonably non-controversial setting, which was introduced at the beginning of this subsection in which models M_1 and M_2 correspond to hypotheses that partition the parameter space of the data model parameter $\boldsymbol{\theta}$.

Example 7.6

The posterior distribution for specificity in Example 7.5 was a beta distribution with parameters 1322 and 4. The expected value of this distribution is 0.9970. If we are willing to *a priori* declare the prior odds $P4(M_1)/Pr(M_2) = 1$, then to gauge the evidence in support of a hypothesis that specificity is greater than 0.9950 we would compute, for a quantity $q \sim \text{Beta}(1322, 4)$, the test

statistic $B = Pr(q > 0.9950)/Pr(q \leq 0.9950) = 8.7135$. This value would support a claim that the data provide some but not strong evidence that specificity exceeds 0.995. Alternatively, we could compute prior probabilities of $M_1 = H_0 : \theta > 0.995$ and $M_2 = H_1 : \theta \leq 0.995$ from the prior, which was a Beta$(3, 2)$ distribution. Then,

$$\begin{aligned} Pr(M_1) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_{0.9950}^{1} t^{\alpha-1} \exp(-\beta t), \\ Pr(M_2) &= 1 - Pr(M_1). \end{aligned}$$

With $\alpha = 3$ and $\beta = 2$ these integrals give $Pr(M_1) = 0.00015$ and $Pr(M_2) = 0.99985$. The Bayes Factor in favor of model M_1 over model M_2 is then,

$$\begin{aligned} BF(M_1, M_2) &= \frac{Pr(M_1|y)}{Pr(M_2|y)} \Big/ \frac{Pr(M_1)}{Pr(M_2)} \\ &= \frac{0.89705}{0.10295} \Big/ \frac{0.00015}{0.99985} = 58,081.43, \end{aligned}$$

which certainly provides overwhelming evidence that specificity is greater than 0.995. The disconnect between these two results is due to the rather dramatic discrepancy between the hypothesis being tested and the prior distribution chosen. If we truly believed that *a priori* $Pr(M_1) = Pr(M_2)$ then we have chosen a prior that does not reflect our true prior beliefs. If, on the other hand, the Beta$(3, 2)$ prior truly represents our prior beliefs, then we probably have no business testing a hypothesis that the binomial parameter is greater than 0.9950. That hypothesis could only have come from an examination of the results of posterior derivation, which is similar to what is sometimes called data dredging or data fishing in other statistical contexts.

7.3 Specification of Prior Distributions

The literature on prior formulation is vast, and the number of viewpoints contained in that literature is diverse. Prescriptions about how to construct prior distributions have failed to garner general acceptance, in part because it is difficult to agree on what qualities a prior distribution should possess. In other words, we have arrived at no generally accepted answer to the question of what makes a prior good. In this section we present a few of the more common methods for choosing prior distributions but, more importantly, we then offer a set of issues to consider that can help avoid selection of prior distributions that are clearly "not good" in a particular problem. Here, we consider the fundamental structure of data model-prior-posterior. There are differences between this structure and more complex frameworks such as hierarchical models and models in high dimensions, and we defer discussion of some issues pertinent for those models until later in the work. A final point in these introductory comments is that the basic methods for prior specification will be presented in terms of a scalar parameter. Arriving at a joint prior, which is necessary, is typically accomplished through a constructive process in which fundamental methods for scalar parameters are combined. We will discuss this later in the chapter.

7.3.1 The Gold Standard

Under a pure version of our conceptualization of Bayesian analysis a prior distribution represents a quantification of everything we know about the possible values of a data model parameter before the observation of data. This implies that the parameter corresponds to some quantity that has definition independent of the data model that will be used in an analysis, such as a

quantity with physical meaning for a problem or that has an unambiguous meaning for the subject matter. Examples might include the distribution of impurities in a crystal lattice (Jaynes 1968) or the proportion of registered voters that are in favor of a proposed municipal bond issue. Given that such a meaningful quantity will become a parameter in some data model, we do not need to know what that data model is to formulate a prior, only the set of possible values of the quantity. Of course, any additional knowledge about where the quantity (now a parameter) is likely to be located in its space of possible values should be incorporated into the prior. The point is that in such problems a gold standard of prior formulation is possible that does not depend on the planned measurement operation, definition of random variables for that operation, or specification of a probability distribution or model for those variables. Gelman et al. (2017) call this an *ideal* conceptual interpretation of prior distributions. But problems that allow priors to be formulated purely external to the remainder of the analysis are few and far between. It is more typical that prior distributions must be selected for parameters that are defined by a posited data model that will be used to analyze data from a given study. Thus, in practice, prior distributions are usually selected within the context of a known study design, data model, and associated likelihood. That priors are entwined with likelihoods will be elaborated on in what follows. Nevertheless, perhaps the least controversial assertion about choosing prior distributions is that if previous studies or experience are available, that information should be incorporated into prior formulation. The most concrete form of such information is data from previous, related studies. This is not dissimilar to construction of a meta-analysis in which a collection of medical studies are used to arrive at a combined estimate of some clinical effect. The studies included in a meta-analysis do not all need to involve exactly the same

study design, but they need to allow estimation of a common effect, such as the degree of improvement in a physical condition for patients treated with a given procedure. A properly conducted meta-analysis includes a set of criteria that must be met by individual clinical trials for inclusion in the analysis. It would be valuable for prior formulation based on previous data to follow this same type of protocol, but we have not seen this in the literature.

Example 7.7

The U.S. National Marine Fisheries Service (NMFS) conducts standardized surveys on research cruises off the Northeast coast of the U.S. from Cape Hatteras, North Carolina to the Canadian Scotian Shelf. On these surveys, trawls are conducted and the entire catch is enumerated, weighed and measured. The simplest indicator of the health of many fish stocks is just the probability that certain species are captured in a given tow. These surveys have been conducted since roughly the mid-1960s. Although the abundance and distribution of fish do change from year to year, which is the reason for the surveys, dramatic declines or increases in any one year are unlikely. Given previous data over a 50 to 60 year time span, one would be foolish not to allow this information to influence both the location and variance of a prior distribution for the probability a given species is captured in the current year.

7.3.2 Prescriptions for Prior Formulation

Any number of prescriptions for constructing prior distributions have been proposed in the literature. And there are a number of ways to attempt to categorize or organize these prescriptions in terms of categories such as informative versus non-informative priors, subjective versus objective priors, and structural versus regularizing priors. Here, we briefly review some of the prin-

cipal techniques that have been proposed for constructing prior distributions.

Priors from Expert Opinion

The desire to incorporate actual prior knowledge or belief into prior specification has led to an entire body of literature on what is called expert *prior elicitation*(e.g., Chaloner, 1996; Mikkola et al., 2024) The process of prior elicitation involves seeking the opinions of subject matter experts, with a key being the translation of opinion about the physical or biological system under investigation into distributions for data model parameters. It is difficult, except in very simple situations, to directly ask a subject matter expert about parameter values and there are a variety of techniques that have been used to try and translate the answers to questions that subject matter experts are comfortable with into components of statistical models (e.g., Garthwaite et al., 2005). At one time, prior elicitation was widely considered to be of concern mostly to extreme subjectivist Bayesians (a different school of Bayesian thought than has been espoused in these notes). But the importance of prior elicitation is becoming more widely recognized, at least in part because of its role in assessments of risk and reliability. For example, eliciting prior information is essential in problems for which observation of complete systems, such as the functioning of nuclear warheads, is impossible (Hamada, Wilson, Reese, and Martz, 2008).

Structural Priors

What are often referred to as *structural* priors attempt to incorporate known properties of a problem or measurement operation into the prior specification. Various problems in engineering, physics, and statistical mechanics may involve

certain symmetries or known constraints on the evolution of systems and the attempt is to encode such properties mathematically into prior distributions. One principle that has been used to do so with discrete sets of probabilities is maximum entropy, which has been extended and offered in combination with the concept of transformation groups as a prescription for prior formulation in more general situations (e.g., **?**).

Personal Priors

We much prefer the phrase *personal* to the broader term subjective when talking about prior distributions that are generated solely by the opinion of a single individual. Any prior that is not generated entirely by previously observed data could be thought of as subjective, as could most statistical models in their entirety. By personal prior we mean a distribution that is not necessarily motivated by any consideration of the problem or previously available information and no concern is attached to the possibility that no other individuals may see such a prior as reasonable; they may do so, but the question is simply of no concern. Personal probabilities are a hallmark of one school of Bayesian thought which rejects the notion that objectivity is a useful concept (e.g., Savage, 1954).

Conjugate Priors

Prior distributions do not function in a vacuum and are connected with the form of the data models to which they are attached. As such, in some cases the form of the likelihood can suggest families of prior distributions that are mathematically convenient in that they lead to posteriors with known closed forms. In particular, prior and likelihood pairs that result in posteriors that

belong to the same family of distributions as the prior are called *conjugate* pairs. Because the data model and its likelihood function are often considered given for the remainder of statistical analysis, priors may be referred to as conjugate for a parameter θ, but this phrase must be interpreted within the context of the data model to which θ belongs.

To understand conjugacy, consider a data model $f(\boldsymbol{y}|\theta)$ and a prior $\pi(\theta|\boldsymbol{\lambda}_0)$, 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,

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\theta)\,\pi(\theta|\boldsymbol{\lambda}_0)}{\displaystyle\int f(\boldsymbol{y}|\theta)\,\pi(\theta|\boldsymbol{\lambda}_0)\,d\theta} \\ &= \pi(\theta|h(\boldsymbol{y},\boldsymbol{\lambda}_0)), \end{aligned}$$

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

We have already seen examples of conjugate prior and likelihood pairs in Example 7.1 in which a gamma (exponential) prior was conjugate for the parameter of a one-sample Poisson data model, and Example 7.2 in which a normal prior was conjugate for the mean of a one-sample normal data model with known variance. Other conjugate prior and likelihood pairs include a gamma prior for the parameter of a one-sample exponential model, an inverse Gaussian prior for the variance of a one-sample normal data model with known mean, and a beta prior for the parameter of a binomial data model.

In our discussion of conjugate prior and likelihood pairs and, indeed, in the examples given earlier in this chapter, we have focused either on one param-

eter models or have taken portions of a parameter vector to be known (e.g., Examples 7.2, 7.3 and 7.4). While part of the reason for this was to reduce complexity in pedagogical presentation, these examples are also useful in their own right, because the posterior of a given parameter assuming other parameters are known is exactly the same as a conditional posterior. For example, the posterior $p(\mu|\boldsymbol{y})$ in Example 7.3 is precisely the same distribution as the conditional posterior $p(\mu|\sigma^2, \boldsymbol{y})$ in a model with unknown μ and σ^2.

Proper Uniform Priors

Uniform priors attach equal probability to any equal intervals inside a bounded parameter space or finite subset of an unbounded parameter space. The most obvious examples are uniform distributions on the unit interval for the parameters of binary, binomial, or beta data models. But uniform priors are also sometimes used for variances with parameter spaces consisting of the positive real line. Here, the uniform distribution will be specified on an interval $(0, A)$ for some positive constant A. Uniform distributions can also be assigned to location parameters using some large interval (A, B) for which knowledge of the problem indicates a mean (for example) will not lie outside of. If, for example, we are modeling departures of a temperature in degrees C from the long-term mean at a given location, we have little doubt that the expected departure will lie in the interval $(-40, 40)$.

Improper Priors

If the space of a given parameter is not bounded and we allow a uniform prior to extend over that entire space, we arrive at an improper prior which has the

form,

$$\pi(\theta) = 1; \quad \theta \in \Theta. \tag{7.12}$$

Priors of the form (7.12) are clearly not distributions as long as Θ is not a bounded set since they do not integrate to any finite value. Improper priors do not, however, necessarily imply improper posteriors. As long as the integral

$$\int f(\boldsymbol{y}|\theta)\, d\theta = K(\boldsymbol{y}) < \infty,$$

then the posterior distribution

$$p(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)}{\displaystyle\int f(\boldsymbol{y}|\theta)\, d\theta} = \frac{f(\boldsymbol{y}|\theta)}{K(\boldsymbol{y})},$$

will exist and will integrate to 1.

The posterior that corresponds to a data model with an improper prior is proportional to the likelihood, so that the posterior mode will be equal to the maximum likelihood estimate of the parameter in question.

Example 7.8

Consider a simple normal one sample data model, $Y_1, \ldots, Y_n \sim iid\, N(\mu, \sigma^2)$ 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,

$$\begin{aligned} p(\mu|\boldsymbol{y}) &= \frac{\exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right\}}{\displaystyle\int_{-\infty}^{\infty} \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right\}} \\ &\propto \exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right\}, \end{aligned}$$

which is the density of a normal distribution with mean $\bar{y}$ and variance σ^2/n, that is, $p(\mu|\boldsymbol{y})$ is $N(\bar{y},\, \sigma^2/n)$. In this case, not only is the mode of the posterior distribution equal to the maximum likelihood estimate, but so too is

the posterior expectation. Similarly, a reasonable 90% interval estimate would be the central 90% of the posterior density, namely $\bar{y} \pm 1.645\,(\sigma^2/n)^{1/2}$ which again agrees with what would be obtained from a non-Bayesian approach that relies on asymptotic results.

It is sometimes the case that we lack any prior information for one or more parameters in a complex model. We might then choose, for example, a conjugate prior for one parameter (which will then be conditionally conjugate) and an improper prior for another. An important point about improper priors is that they should not be used in a careless manner. Any time an improper prior is specified for one or more parameters, one must be able to demonstrate that the posterior is proper. This is frequently not a trivial matter. In fact, the use of improper prior distributions is the most attractive in situations in which we have little notion of what values a parameter might assume which, in turn, occurs in situations with complex models which, in turn, renders demonstration of posterior propriety difficult.

Jeffreys' Priors

Consider a uniform prior on the unit interval for some parameter $0 < \theta < 1$. While this prior gives equal probability to any equal interval, it is not invariant to transformation or change of scale. For example, if $\theta \sim U(0,\, 1)$ then $\eta = 1/\theta$ has density $h(\eta) = 1/\eta^2; \quad 1 < \eta < \infty$ which is far from uniform relative to η, although the data model may be equivalently expressed in terms of either θ or η.

Jeffreys proposed a method for ensuring that priors are invariant under transformation. Consider a data model $f(\boldsymbol{y}|\theta)$ and a transformation of θ into $\eta = h(\theta)$ for some one-to-one and monotone function $h(\cdot)$. The same data

model may now be written as $f(\boldsymbol{y}|\eta)$. Suppose we use some procedure to assign a prior $\pi_\theta(\theta)$ and the same procedure to assign a prior $\pi_\eta(\eta)$. For example, assigning both θ and η uniform distributions would qualify as a procedure for assigning these priors. Now, the prior $\pi_\theta(\theta)$ also implies a prior for η through transformation of variables, say $\pi'_\eta(\eta)$. Jeffreys goal was to arrive at a procedure for assigning priors such that the result would be that $\pi_\eta(\eta) = \pi'_\eta(\eta)$.

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,

$$\pi_\theta(\theta) = \{I(\theta)\}^{1/2}. \tag{7.13}$$

The form (7.13) is known as a *Jeffreys prior*.

Example 7.9

Suppose that we have a single observation corresponding to the data model $Y \sim Bin(\theta, n)$ where n is fixed. The expected information about θ in this model is $I(\theta) = n[\theta(1-\theta)]^{-1}$ so that the Jefferys prior would be $\pi_2(\theta) \propto \{\theta(1-\theta)\}]-1/2$.

Example 7.10

Consider again the normal one sample model of Example 7.8 with $Y_1, \ldots, Y_n \sim iidN(\mu, \sigma^2)$ with σ^2 considered known. In this case, we know that $I(\mu) = n/\sigma^2$

which, with σ^2 known, is a constant. Thus, the Jefferys prior for μ in this case is improper.

7.3.3 Non-Informative and Diffuse Priors

A substantial portion of the literature on prior specification concerns ways to arrive at prior distributions that are in some sense *non-informative* about data model parameters. Exactly what is meant by non-informative has been a subject of some debate, but the essential idea is that a prior distribution is non-informative if it has little influence on the posterior relative to the contribution of the likelihood. Attempts to arrive at priors that fulfill some notion of being non-informative have resulted in proposals to use uniform priors, Jeffreys' priors, reference and default priors, regularizing priors, and improper priors. Uniform and Jeffreys' priors have been introduced previously. The claim for being non-informative for uniform priors is that they are flat but, as we have seen, this property depends on the scale of expression. Jeffreys' priors attempt to rectify this deficiency with uniform priors, but clearly tie prior formulation to the likelihood chosen for analysis of a given problem. In certain constrained problems maximum entropy priors for discrete, finite parameter spaces may reduce to uniform priors (e.g., Zellner, 1991) or Jeffreys prior (e.g., Bernardo's so-called reference prior in the case of a single scalar parameter,(see **?**Berger and Bernardo, 1992). In an attempt to divorce prior formulation from specific likelihood functions, general reference priors depend on asymptotic approximations to likelihoods, the quality of which influences performance of the prior (e.g., Berger et al., 2009; Gelman et al., 2017). Regularizing priors attempt to lend some robustness or stability to derivation of a posterior distribution, and tend to produce smoother posteriors than the use of uniform priors. Improper

priors lay claim to being non-informative because the resultant posterior distributions are proportional to likelihoods and inferences tend to be quantitatively similar to non-Bayesian analyses, at least in some problems as seen in Example 7.8. Improper priors, however, require a demonstration that the associated posterior is proper and this can be a complex endeavor that tends to be specific to particular classes of models, such as linear mixed models (e.g., Sun, Tsutakawa, and He, 2001) or binomial regression (Roy and Kaiser, 2013).

It has become common, especially for parameters about which little is known or at least little can be coded into a mathematical expression, to take whatever form of prior is convenient and to reduce its influence on the posterior by making the prior variance large. These are often called *diffuse* priors and they can appear to be an effective way to avoid the need to check for posterior propriety while at the same time reaping the benefits of being nearly non-informative in the same way that an improper prior is non-informative. For example, in the analysis of a logistic regression for binary response variables one may choose to place normal priors on the regression coefficients β_0 and β_1 rather than attempt to prove that improper uniform priors result in a proper posterior. Then, those proper normal priors can be made diffuse by allowing the prior variances to be large. In this way, the effect of the prior distributions on the posterior can be made to mimic what might occur with improper priors, but without the need to check for posterior propriety. This device can be effective but is not without its potential pitfalls, as we will describe in the next section.

7.4 Prior Difficulties and Avoiding Them

Unwisely chosen prior distributions can have damaging consequences for an overall Bayesian analysis. Here, we review several of these in particular situations and then offer a general prescription for avoiding at least the most obvious problems that can occur.

7.4.1 Likelihood Incongruency

The consequences a prior has for the outcome of a Bayesian analysis are inherently entwined with the likelihood with which it is associated. The data model determines parameter spaces for the quantities that govern its characteristics, and it is those parameter spaces on which prior distributions exist. The support of a prior distribution cannot extend beyond the parameter space of the parameter to which it is assigned so that, for example, a prior distribution for a variance parameter cannot have support on the negative line. Beyond this, prior distributions should also not contain non-negligible probability mass on physically implausible values of a parameter. For example, in consideration of the proportion of male births of nearly any mammal species, the parameter space of any reasonable data model will be the interval (0, 1), but extreme values are clearly not within the realm of possibility. Thus, a uniform prior on this proportion, although sometimes thought of as somehow "fair" or "non-informative", is anything but, giving equal probability to values less than 0.10 as to values in the range 0.45 to 0.55.

It is quite common to formulate a data model that has support well beyond what might be reasonable for the quantity under consideration. Consider, for example, assigning a normal distribution to random variables connected with the weight of individual hummingbirds of a given species. This might be

entirely reasonable, but only because we know the tails of a normal distribution with anything but massive variance die off rapidly so that a normal distribution centered at something like 10 (grams) and variance anything less than 10 places probability of less than 0.003 outside the range of $1-20$ grams, which is roughly the range of weights for all known hummingbird species. So, even though the support of normal distributions is the entire real line, as a data model for observed hummingbird weights it could be quite adequate. The same flexibility is not enjoyed with prior specifications. In this same problem, it would not be unreasonable to specify a normal prior distribution for the mean of the normal data model based on, if nothing else, the convenience of conjugacy. If, however, we then attempt to make that prior something like non-informative by using a large variance of 100, say, then the prior distribution places probability of only about 0.657 on the interval 1 to 20 and, in fact, places probability of just over 0.18 on the negative line which, of course, is impossible even for a small hummingbird. Thus, we have assumed prior information that is known to be totally out of concert with any sensible view of reality.

7.4.2 The Effect of Priors Depend on Parameterization

We have seen in previous chapters that probability distributions are not influenced by the parameterization used. In fact, we sometimes use multiple parameterizations in the analysis of a given data model, one that facilitates estimation and another that leads to a more natural interpretation for the problem under investigation. This arises from the fact that while probabilities can be equivalently expressed under multiple parameterizations, the shape of likelihood functions does depend on the particular parameterization used to compute it. Similarly, how a prior distribution distributes probability across

a parameter space depends on the way that parameter space is expressed.

Example 7.11

In discussion of Jeffreys' priors we noted that a binomial probability mass function can be equivalently expressed in terms of the usual parameter as $f(y|\theta) = \theta^y (1-\theta)^{n-y} + c(y,n)$ or in terms of $\eta = 1/\theta$ as $f(y) = (\eta-1)^{n-y}\,\eta^{-n} + c(y,n)$. The parameter spaces associated with this model are $0 < \theta < 1$ and $1 < \eta < \infty$. Suppose we put uniform prior distributions on each of θ and η so that the prior on θ is proper but that on η is improper. The posterior distribution for θ is then a beta distribution with parameters $y+1$ and $n-y+1$. The posterior for η is $p(\eta|y) \propto (\eta-1)^{n-y}\,\eta^{-n};\ 1 < \eta < \infty$. Transforming this posterior using $\theta = 1/\eta$ results in a beta distribution with parameters n and $n-y$. Now suppose further that the observation turns out to be $y = 10$ for a binomial sample size of $n = 20$. The two posteriors that result for θ are shown in Figure 7.1. The solid curve is for the model with a uniform prior placed on θ and the dashed curve is for the model with a uniform prior placed on η. These are actually quite distinct posterior distributions. With a uniform prior on θ the posterior expectation is $E(\theta|y) = 0.50$ and a 90% credible interval of (0.33, 0.67) while the model with a uniform prior on η gives a posterior expectation of $E(\theta|y) = 0.67$ and a 90% credible interval of (0.52, 0.80).

7.4.3 Avoiding Poor Priors

There is no magic procedure that can be relied on to identify priors that are out of concert with scientific understanding or that are likely to have deleterious consequences for posterior inference. There are, however, several points that can be made to help avoid major problems with the formulation of prior distributions and the inference they eventually lead to.

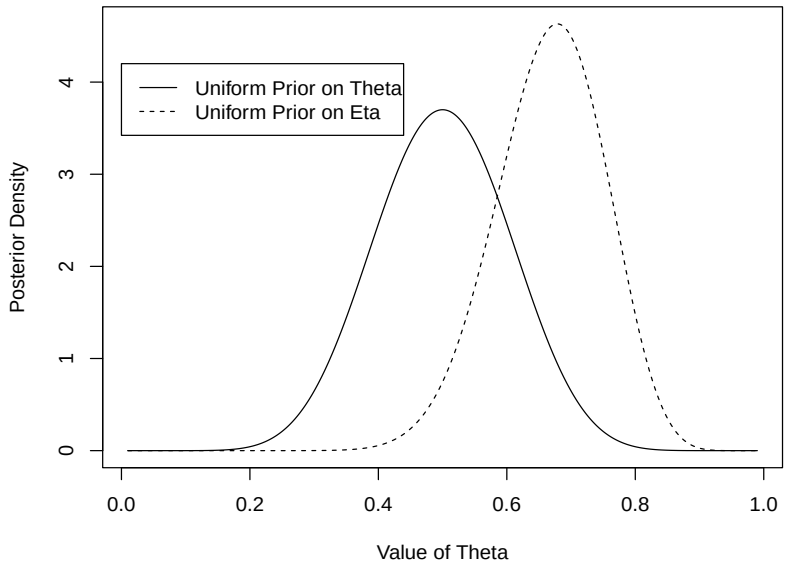


Figure 7.1: Posterior beta distributions for a binomial observation of $y = 10$, $n = 20$, for prior distributions on two parameterizations of the binomial data model.

Priors and Current Observations

That a prior distribution cannot be divorced entirely from its associated data model is not a license for complete disregard of the ideal situation in which a prior represents everything we know about a parameter before data are observed. There are clear ways in which one can produce misleading inferences from a Bayesian analysis by matching prior distributions to observed data values. In an analysis of data assumed to having arisen from a one-sample normal model, for example, choosing a prior for the mean that itself has a mean equal to the observed sample average clearly vitiates the Bayesian prescription for

using data to update existing knowledge or belief. Uncertainty in the posterior will be under-estimated, possibly drastically so. A reasonable rule of thumb is that initial examinations of data can influence the choice of data model distributional families and this may, in turn, influence the choice of distributional families for prior distributions (e.g., conjugate prior-likelihood pairs) but should not influence the choice of parameter values for those distributions. That is, examination of current data should influence prior specification only through the form chosen for the data model.

The point of the preceding paragraph notwithstanding, the data that are anticipated may have a proper impact on specification of prior distributions. This anticipation is prior information, not the result of examination of current data values. For example, if the problem involves estimation of the proportion of registered voters who favor a given candidate in a state-wide election, past experience indicates there is a small chance that proportion will be less than 40% or greater than 60%. Thus, selecting a prior that places substantial probability on these extreme regions is likely to influence results in a negative way.

It is often asserted that priors become "swamped" by current data so that poor choices of prior distributions are protected against if sample sizes are large enough. This can be true, if the increase in sample size offers replication of information about a particular parameter from a reasonably specified model. If, however, the data model is poorly specified so that the parameter we believe should be estimated is poorly identified by the data, then the same phenomenon can result in quite misleading inferences. In addition, some more complex models are formulated such that the number of parameters increases as the sample size increases so that additional observations do not necessarily represent an increase in information about a fixed set of parameters. In these

cases, increasing sample size may not alleviate problems caused by poor specification of prior distributions.

Example 7.12

Suppose we have a small sample of 15 observations from a normal distribution with mean $\mu = 10$ and known variance $\sigma^2 = 1$. We intend to fit a one sample normal model with mean μ and known variance 1 to these data. Examination of a stem plot, although of limited value for a small sample, exhibits no drastic contradiction to this intention. Due to some misinformation we determine an appropriate prior might be normal with mean $\lambda = -10$ and variance $\tau^2 = 0.2$ (which might come from the belief that our prior information is worth 5 current observations, see section 7.x). By Example 7.4, if $\bar{y} = 9.44$ the posterior distribution of μ will be normal with mean $M = 4.58$ and variance $V = 0.050$ with a 95% credible interval of (4.14, 5.02) which is quite far off. If, however, we have a sample of size 100 with the same sample mean of 9.44, the posterior mean and variance become $M = 8.51$ and $v = 0.0095$ with 95% credible interval (8.32, 8.70). If the sample size had been 500 these values would be $M = 9.25$, $V = 0.002$ and (9.16, 9.34).

Assessment Using Prior Predictive Distributions

Consider a problem in which we have solid prior information, perhaps from previous studies, but are unable to produce new or current data. Our only choice for making inference would be use of the prior distribution. We would consider potential data to be generated from the prior predictive distribution (7.2). We can use simulation to actually generate data from a prior predictive distribution and compare such data with scientific reasonableness. This can often be an effective check to flag prior distributions that will be incongruous with data

that will be gathered. To simulate data from the prior predictive distribution is a simple matter. First simulate a value θ^* from a prior distribution $\pi(\theta)$ and then use that value as a parameter in the data model to simulate a value y^* from $f(y|\theta^*)$. The result is a simulated value from $p(y^0) = \int f(y^0|\theta)\,\pi(\theta)\,d\theta$.

Example 7.13

Consider the problem of estimating the probability of a male birth in a given mammal species that generally gives birth to only one offspring at a time. We have previously used this scenario to caution against indiscriminate use of uniform priors, as biologically the probability of a male birth in a mammal species should not depart from 0.50 in a dramatic way. If we are able to observe a set of n births and assume unique parentage for each we might model the number of male offspring using a binomial data model with binomial sample size n and probability (parameter) θ. We know that a beta prior distribution for θ is conjugate for this model. An alternative would be to express the binomial in exponential family form for which the natural parameter is the logit of θ, $\eta = \log(\theta) - \log(1-\theta)$ so that the parameter space of η is now the entire real line. We could then attempt to make our prior diffuse by assigning η a normal prior with mean 0 and a large variance, say 100. A check on this potential prior consisted of simulating 100 binomial observations with $n = 25$ from the prior predictive distribution that results from assigning η a normal prior with mean 0 and variance 100. The results are summarized in the following stem plot:

```
The decimal point is 1 digit(s) to the right of the |

  0 | 00000000000000000000000000000001112222233444
  0 | 556799
```

```
1 | 33
1 | 9
2 | 33344444
2 | 55555555555555555555555555555555555555
```

The use of simulation from the prior predictive has clearly indicated that this prior does not reflect the biological reality connected with our problem. Additional investigation of this prior predictive distribution suggests that the prior variance needs to be made quite small before the distribution begins to behave like even a uniform distribution, which we have already flagged as inappropriate for this problem.

7.5 Constructing Joint Priors

The practical assignment of a joint prior to a p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ 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 7.14

Consider a one sample model using a beta distributon, that is, $Y_1, \ldots, Y_n \sim$ *iid* Beta(α, β). What is necessary to conduct a Bayesian analysis for this model is a prior for the parameter (α, β). Now, the parameter space is $\alpha > 0$, $\beta > 0$

and, to construct a joint prior through multiplication we could take

$$\pi(\alpha, \beta) = \pi_\alpha(\alpha)\, \pi_\beta(\beta),$$

for any density functions $\pi_\alpha(\cdot)$ and $\pi_\beta(\cdot)$ that each have support on the positive line. To assist in selection of these marginal priors it may be helpful to reparameterize the beta 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 $p(\alpha,\, \beta|\boldsymbol{y})$ or $p(\mu,\, \phi|\boldsymbol{y})$ would, in this example, require the use of simulation methods.

Example 7.15

Consider again the normal one-sample problem, 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 not jointly sufficient for μ and 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 a 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 $\pi_1(\mu|\sigma^2)$ in conjunction with the marginal prior $\pi_2(\sigma^2)$.

7.6 Choosing Prior Parameter Values

We have stated several times that, in an actual application, prior distributions should contain no unknown parameters. Proper prior distributions are, however, often in the form of parameterized probability density functions. There are a number of basic ideas that help in selecting the values for these parameters.

7.6.1 Previous Studies

The most scientifically defensible way to choose prior parameters is on the basis of previous studies that share some features with the problem under current

investigation. In some cases a study design may be repeated multiple times, such as in monitoring programs. In other cases, studies may share objectives for different, but related, topics, such as the prevalence of related diseases. In any situation for which there have been the same or similar investigations in the past we should carefully consider whether previous analyses can be used to inform our prior beliefs in the current situation.

Previous Posteriors and Discounting

In some situations the same study or investigation may be repeated in different situations, such as occurs in programs to monitor animal populations or in surveys to produce estimates of unemployment or other economic indicators. An important question is whether the situations produce data that can be considered to arise from independent sets of random variables. Different situations often correspond to different time periods, such as years or months, and the question becomes whether the time lag between periods of observation is long enough to allow an assumption of independence. If not, then models we have not yet discussed, such as dynamic models or models with autoregressive structure should be considered. Here, we consider only situations in which an assumption of independence among situations is reasonable.

The use of conjugate prior and likelihood pairs lends itself to sequential analysis in problems for which data accumulate over time. If we are able to assume the same data model, meaning with the same value of the parameter, applies to each wave of data, then we have a situation in which a cascade of prior-posterior pairs results from the use of conjugacy. The initial data model is $f(y|\theta)$ and the initial prior is $\pi(\theta|\lambda_0)$. Observation of data y_1 and conjugacy of f and π results in the posterior $p(\theta|y_1) = \pi(\theta|h(\lambda_0, y_1))$. Now

using this posterior as a prior, observation of data y_2 presumed to also be from $f(y|\theta)$ leads to the posterior $p(\theta|y_1, y_2) = \pi(\theta|h(\lambda_0, y_1, y_2)$. This progression continues, with the posterior from one stage of analysis becoming the prior for the next.

A difficulty with the scenario just presented is that posterior variance and, hence, the subsequent prior variance decreases with each addition of new data. For example, as in Example 7.4, the posterior variance of μ from a one sample normal (μ, σ^2) model combined with a conjugate normal (λ, τ^2) prior is $\tau^2\sigma^2/(n\tau^2 + \sigma^2)$. If waves of data having sizes $n_1, n_2, \ldots$ become available, at wave k we have $n = n_1 + \ldots + n_{k-1}$ so that, with tau^2 and σ^2 fixed, posterior variance is monotone decreasing as data accumulate. This will be true even if the data model parameter μ does not actually remain constant. Thus, even if our posterior mean changes at some point our quantification of uncertainty in that value will still get smaller and smaller. Thus, we want to apply this type of sequential analysis only in situations in which the data model parameter can be assumed constant over time. Justification of such an assumption must come from scientific understanding because results can be misleading and the fact that a data model parameter may not remain constant can be difficult to diagnose.

Example 7.16

To illustrate the points of the previous paragraph, data were simulated from two different scenarios. In the first, four values of binomial random variables all with parameter $\theta = 0.50$ and $n = 25$ were simulated, resulting in $y_1 = 10$, $y_2 = 15$, $y_3 = 13$, and $y_4 = 14$. These values were analyzed sequentially beginning with a uniform prior on the unit interval. For the second scenario, four values of binomial random variables were simulated with parameters $\theta_1 = 0.50$,

$\theta_2 = 0.45$, $\theta_3 = 0.60$ and $\theta_4 = 0.35$, each again with $n = 25$. These values were $x_1 = 15$, $x_2 = 11$, $x_3 = 16$ and $x_4 = 11$, and the same sequential model employed for the first set of values was used to determine posterior distributions. The sequences of posteriors distributions are presented in Figure 7.2, with the top two rows corresponding to the variables with equal binomial probability (arranged as 1 and 2 in the first row and 3 and 4 in the second row) and the bottom two rows corresponding to the variables with varying binomial probabilities. It is easy to see that one would be unable to distinguish between situations for which the binomial parameter was fixed (top two rows) or variable (bottom two rows) based on these sequential analyses. In both cases posterior variance decreases as the number of observations increases from one to four. If we had unknowingly applied this analysis in the situation with varying binomials we might very well conclude that the analysis is "closing in on the truth" (the true state of nature again) over the sequence of four values. There is no way to tell that is not the case from the sequence of posteriors. Justification for the use of one posterior as the exact prior for another observation must come from the substantive problem under investigation.

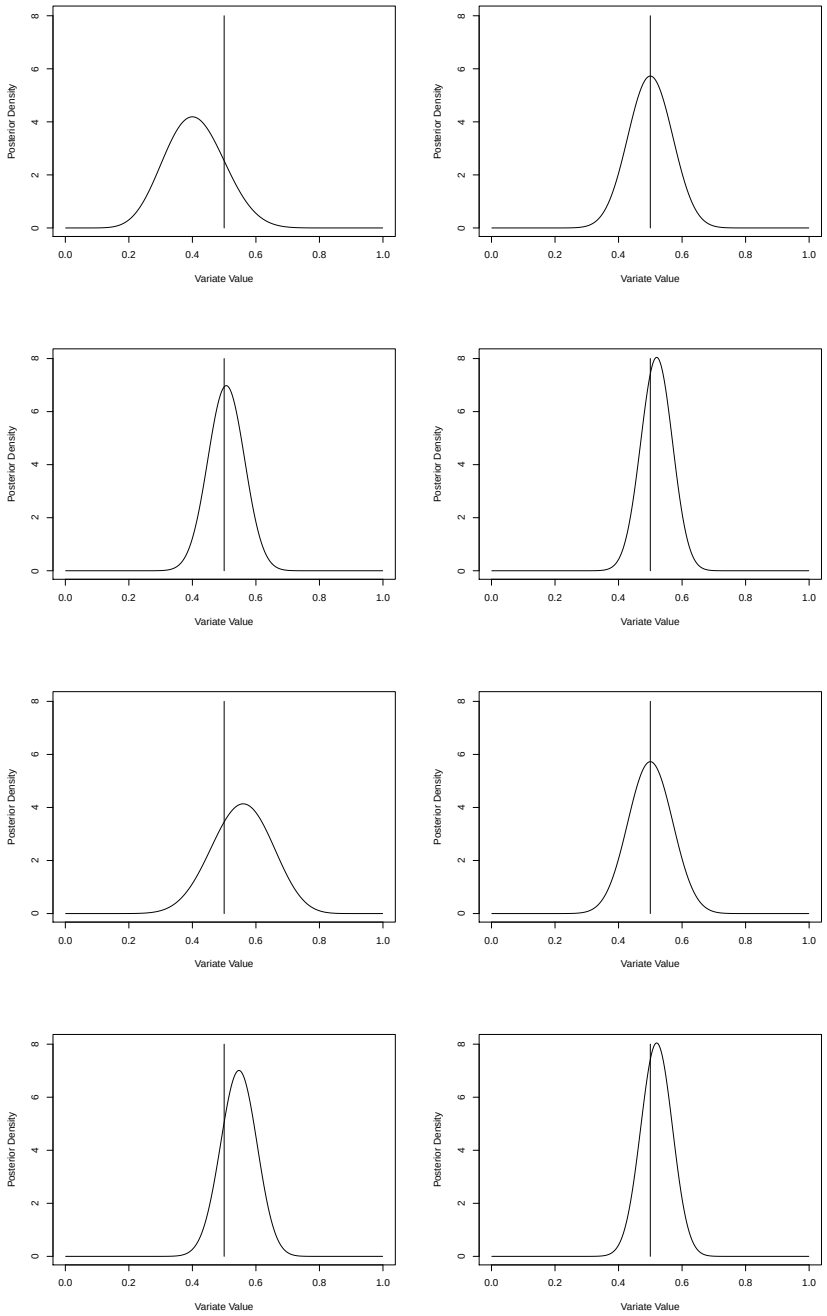


Figure 7.2: Sequential analysis of two sets of binomial observations. Top 2 rows have common parameter. Bottom 2 rows have varying parameters.

We may wish to make use of previous analyses even in cases for which we do not believe the data model parameter has remained fixed. For example, monitoring the presence/absence of frogs along survey routes in the Midwestern United States produces binomial-type data (number of suitable locations such as ponds for which a given species is present), but there is no reason to believe that the binomial parameter should be the same from year to year. In fact, detecting changes in the probability of the presence of frogs is the whole point, as the disappearance of frogs is an early indicator of environmental degradation. At the same time, it is reasonable to believe that the overall probability of presence in one year provides useful information about what to expect in the next year. It is unlikely, for example, that the probability of presence on a given route will be 0.10 one year and 0.85 the next. In situations that have this structure we may take a posterior distribution from one time point and *discount* it for use as a prior at the next time point. Discounting is the process of making a distribution (usually in the form of a density function) more variable or diffuse. There are a number of techniques for accomplishing this, one of which we will consider in the next subsection.

Similar Studies

The majority of scientific investigations are not repeated multiple times in exactly the same way. It is common, however, for scientific investigations to build on previous studies or to examine the generality of an idea that has been proposed in earlier work. Thus, studies may be similar but involve differences in physical environment such as different times, or different places. These situations can be treated in much the same way as a sequential analysis with posterior discounting.

Perhaps even more common than studies that vary only one or two aspects of the setting in which a design is used are studies conducted with related objects – animal species, diseases, classes of vaccines, and so forth. Because these types of objects tend to require greater differences in how they are handled than minor differences in physical environments, study designs may be quite distinct. But even in this type of situation it may still be possible to glean some information about a current study from the results of previous investigations. Diseases with similar eitiologies may serve as surrogates for one another in terms of population at risk or background prevalence. Studies on an organism with similar physiological responses to environmental stressors as humans (such as pigs) may provide information on what to expect *a priori* in a clinical study. In the rapidly developing field of "functional genomics" model organisms are defined as simple organisms (yeast, worms, flies) that allow rapid experimental evaluation of gene function. At the molecular level, the phenotype of an organism may be of only minor importance, and it may be possible to obtain useful prior information about one type of organism (e.g., humans) from studies on quite dissimilar organisms (e.g., *Drosophilia*).

7.6.2 The Current Worth of Prior Information

At least for several specific models, it can be useful to consider the number of current observations to which we feel our prior information is equivalent. This process is model specific and there are no general prescriptions available. Thus, we illustrate through the presentation of several examples.

Example 7.17

Consider again a binomial data model $Y \sim \text{Binomial}(\theta)$ with known binomial sample size n, and a beta prior $\theta \sim \text{Beta}(\alpha_0, \beta_0)$. We know for this conjugate

pair that the posterior will again be a beta distribution with parameters

$$\alpha = \alpha_0 + y \text{ and } \beta = \beta_0 + n - y.$$

The prior parameter α_0 has been updated by adding the number of binomial "successes", and the prior parameter β_0 has been updated by adding the number of binomial "failures". We can think of α_0 then as the number of prior successes out of a total of $\alpha_0 + \beta_0$ binary trials. This can put selection of α_0 and β_0 on more solid footing than defaulting to a uniform specification of $\alpha_0 = \beta_0 = 1$.

Example 7.18

Consider the one sample normal data model of Example 7.12 with a conditional prior on the expected value $\mu \sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0)$. This prior looks like the sampling distribution of a sample mean from a set of κ_0 independent and identically distributed random variables having $\mathrm{N}(\mu_0, \sigma^2)$ distributions. Since the posterior expectation is a weighted average of the prior mean and the sample mean, we can think of κ_0 as the number of observations the prior is worth.

Example 7.19

A Poisson data model combined with a gamma prior constitute a conjugate pair. For some $\lambda > 0$, let $Y_1, \ldots, Y_n \sim iid\text{Poisson}(\lambda)$ and let $\lambda \sim \text{Gamma}(\alpha_0, \beta_0)$ where $\alpha_0 > 0$ and $\beta_0 > 0$ will be specified numerical values. In this case, the posterior distribution of λ is gamma with parameters

$$\alpha = \alpha_0 + \sum_{i=1}^{n} y_i \text{ and } \beta = \beta_0 + n.$$

We can choose prior parameter values by thinking of α_0 as the sum of Poisson counts arising from a previous sample of size β_0.

7.7 Model Assessment

Assessing a model in a Bayesian analysis can involve both the prior distribution specified, and the associated data model which combine to determine the posterior.

7.7.1 Assessing Prior Specifications

We have previously discussed some ideas to avoid the selection of poor priors in the first place. Now, however, we are concerned with the effect that prior choice has had on the outcomes of an analysis. Investigation of this issue can involve the form of a prior distribution but often focuses on the choice of prior parameter values.

A basic tool in assessing how much influence a prior has on the outcome of an analysis is to simply vary the prior used within some class of priors, and observe the degree to which posterior inference is affected. This type of a procedure is called a sensitivity analysis. For example, if a proper uniform prior on the interval $(0, A)$ is selected for some parameter with an arbitrary choice of A, a simple sensitivity analysis might consist of examining posterior distributions for a set of possible choices $A_1 < A_2 < \ldots < A_k$. If the results obtained are sensitive to the choice of A that indicates that more attention should be paid to motivating its value, or perhaps the uniform specification should be changed to some other distributional form.

A sensitivity analysis is typically not conducted with the intent of modifying a prior based on what is learned. In fact, that might come close to using particular data values to determine a prior, a practice already cautioned against. If care is taken in the selection of prior distributions in the first place (see Chapter 7.4.3) then uncovering disasters in a sensitivity analysis should

be rare. Rather, a sensitivity analysis is conducted to learn about the overall behavior of a model. If posterior results differ for different prior distributions that is not necessarily surprising or a cause for concern. It is the degree to which results differ across a set of prior settings that is meaningful and perhaps uncovering how priors interact with the likelihood to produce those differences.

How one quantifies the degree to which posterior results are affected across a set of prior distributions remains an arbitrary aspect of sensitivity analysis, and it is not uncommon for the results to be presented in rather broad categories, such as minor, moderate, or high sensitivity of results to choice of prior. Visual inspection of posterior densities can be valuable, but also difficult to quantify. Measures that might be useful in examining results of sensitivity analyses include shifts in distributional characteristics such as means, variances, and skewness, and indicators of differences in inferential conclusions. The later might, for example, be simply a designation of whether a credible interval contains a particular parameter value (such as 0) for each of a set of priors.

For assessing shifts in distributional characteristics it is nice to have a benchmark from which to compute proportional shifts. In attempting to use diffuse prior specifications it is often beneficial to determine the limiting prior that corresponds to letting variance grow large, which will be an improper prior. For most simple non-hierarchical data-model/prior structures the posterior corresponding to this limiting prior will be proper, although one always needs to check. The posterior that corresponds to the improper limiting prior can make a convenient benchmark against which to assess shifts of proper priors that vary in the degree to which they are diffuse.

Example 7.20

Consider a problem in which we will employ a one sample model with a Poisson

distribution, $Y_1, \ldots, Y_n \sim$ iid Po(λ). Gamma prior distributions are conjugate for this model, so let $\lambda \sim$ Ga(α, β) parameterized so that $E(\lambda) = \alpha/\beta$ and $var(\lambda) = \alpha/\beta^2$. Suppose that the context of the problem (not the data) leads us to believe that the Poisson distribution will be centered somewhere within the single digits, from 1 to 10. We might then decide to assign λ a Gamma distribution with expected value 5. To make the variance of this prior large requires making β small, but subject to α/β not blowing up. This results from making both alpha and beta small. In the limit, as $\alpha \Rightarrow 0$ and $\beta \Rightarrow 0$ the prior becomes improper. The posterior that corresponds to this improper prior is a Gamma distribution with parameters $\sum_{i=1}^{n} y_i$ and $\beta = n$, so the posterior mean and variance are $E(\lambda|\boldsymbol{y}) = \bar{y}$ and $var(\lambda|y) = (1/n)\bar{y}$. A sensitivity analysis of the class of Gamma priors with prior expectation 5 and increasing variances could be conducted with a set of priors Ga(5, 1), Ga(2.5, 0.50), Ga(1.25, 0.25) and Ga(0.625, 0.125). Suppose the data result in a value $\bar{y} = 2.2$. We might compute posterior means, variances and proportional departures of these values from those obtained from the baseline Ga(0, 0) prior. Table 7.1 gives results of this sensitivity analysis.

Proportional departures of posterior expectations for this set of priors from the baseline improper prior are all less than 5% and, aside from the most concentrated prior of Ga(5, 1) are less than 2.5%. Posterior variances show even less departure from the baseline, all being less than 1% with most less than 0.5%. Overall, it appears that the results of this analysis are robust to selection of prior within the class examined.

		Posterior		
Prior	Mean	Departure	Variance	Departure
$\text{Ga}(0, 0)$	2.20	0	0.08800	0
$\text{Ga}(5, 1)$	2.3077	0.04895	0.08876	0.00861
$\text{Ga}(2.5, 0.50)$	2.2459	0.02496	0.08843	0.00486
$\text{Ga}(1.25, 0.25)$	2.2277	0.01260	0.08823	0.00257
$\text{Ga}(0.625, 0.125)$	2.2139	0.01393	0.08812	0.00132

Table 7.1: Results of sensitivity analysis for Poisson data model and Gamma priors.

7.7.2 Assessing the Overall Model

One fundamental notion connected with model aptness is that an adequate model should generate data that share important features in common with the actual data. To assess a fitted model, then, we can make use of the posterior predictive distribution (7.3) by simulating data from it and determining whether simulated realizations are sufficiently similar to the actual data in terms of certain key features. In order to put such a strategy into action we need to address two issues, (1) what key features of the actual data should be the focus of attention and (2) what is meant by sufficiently similar.

There are few definite guidelines relative to what characteristics of the actual data should be reproduced by simulations from the posterior predictive. Perhaps the only point that can be made without qualification is that any features of the data used to assess model adequacy *should not* be sufficient statistics or functions of sufficient statistics. This is because it is through sufficient statistics that the data model likelihood informs the posterior about the possible values of the data model parameter. Thus, any model that is

not completely and obviously inadequate for a given problem should be able to generate data that have values of a sufficient statistic similar to what is observed in the data.

Most problems have data features that are not in the form of sufficient statistics that can be used to assess whether data simulated from the posterior predictive distribution are similar to the actual observations. Extreme values, ranges, skewness or kurtosis are obvious examples for some problems. The frequency of zero values in counts for which the data model was taken to be Poisson. The level of serial correlation for data gathered over time. All of these types of characteristics of data sets might be reasonable choices to quantify data behavior, depending on the problem. Suppose we have selected a quantity through which to characterize some aspect of data behavior, which we will denote as $Q(\boldsymbol{y})$. A procedure for assessing a model fitted to actual data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ is as follows. For an index $m = 1, \ldots, M$,

1. Draw a value of the data model parameter $\boldsymbol{\theta}^*_m$ from its posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$.

2. Simulate a set of values $\boldsymbol{y}^*_m = \{y^*_1, \ldots, y^*_n\}_m$ from the data model $f(\boldsymbol{y}|\boldsymbol{\theta}^*_m)$, that is, using the parameter $\boldsymbol{\theta}^*_m$.

3. Compute the quantity $Q(\boldsymbol{y}^*_m)$.

4. Compute the quantity for the actual data, $Q(\boldsymbol{y})$.

5. A *posterior predictive p−value* is then computed as

$$p = \frac{1}{M} \sum_{m=1}^{M} I[\, Q(\boldsymbol{y}) \leq Q(\boldsymbol{y}^*_m)\,], \tag{7.14}$$

where $I[A]$ is the indicator function that assumes a value of 1 if A is true and a value of 0 otherwise.

The posterior predictive p−value in step 5 of the above algorithm can be used to assess the adequacy of the overall model. This $p-$value is similar to a $p-$value from a goodness of fit test, in which the null hypothesis is that the model is adequate. Exceptionally large or exceptionally small $p-$values indicate the model is lacking in ability to describe the data feature quantified in $Q(\boldsymbol{y})$. Non-extreme $p-$values function in favor of model adequacy. One note that can be important in some problems is that the $p-$value is computed using an inclusive inequality ($\geq$ rather than $>$). If the quantity $Q(\cdot)$ can assume only a discrete set of possible values this can change the $p-$value dramatically for some data sets. For example, suppose the assessment quantity Q we have chosen in a problem is the frequency of a particular value (discrete random variables) or the number of observations greater than a particular value. The possible values of these quantities in a finite set of data may be small and discrete. In these cases we should compute both upper and lower $p-$values as

$$p_{up} = \frac{1}{M}\sum_{m=1}^{M} I[\, Q(\boldsymbol{y}) \leq Q(\boldsymbol{y}^*_m)\,] \tag{7.15}$$

$$p_{low} = \frac{1}{M}\sum_{m=1}^{M} I[\, Q(\boldsymbol{y}) \geq Q(\boldsymbol{y}^*_m)\,] \tag{7.16}$$

In cases for which Q is highly discrete p_{up} and p_{low} may not sum to 1, and small values of either indicate a deficiency with the model.

7.8 Introduction to Simulating From Distributions

Many problems involve posterior distributions that are difficult to deal with analytically. We will still be able to make inference, however, if we can simulate

from such posteriors. If we are able to simulate a large number of independent values from a given distribution, then the Gilvenko-Cantelli theorem (e.g., Billingsley, 1986, p. 275) implies that the empirical distribution of those values converges to the true distribution. If the distribution being simulated is a posterior, then the empirical distribution of simulated values can be used to make probability statements based on that posterior, which is the essential form of Bayesian inference. Even if the values simulated from a posterior distribution are not independent, under certain conditions the Gilvenko-Cantelli result holds, and we again can use the empirical distribution of simulated values as an approximation to the true posterior.

7.8.1 Fundamental Principles of Simulation

Before discussing procedures for simulating values from (posterior) distributions it is helpful to review a few basic principles involved with simulation, which we do using generic notation for random variables X, Y, and Z and their density or mass functions as $f(x)$, $f(y)$ and $f(z)$. Within the context of Bayesian statistical analysis, the random variables X, Y, and Z will correspond to elements of a data model parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

1. Simulation From Joint Distributions Also Simulates From Marginal Distributions.

   If we simulate M values from a joint distribution with density or mass function $f(x, y, z)$, we obtain a set of values

Iteration	Value of X	Value of Y	Value of Z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
M	x_M	y_M	z_M

The joint empirical distribution of X, Y and Z based on this sample is,

$$F_M(x, y, z) = \frac{1}{M} \sum_{m=1}^{M} I[(x_m \leq x) \cap (y_m \leq y) \cap (z_m \leq z)].$$

The marginal empirical distribution function of X is,

$$F_M(x) = \frac{1}{M} \sum_{m=1}^{M} I(x_m \leq x),$$

and similarly for the marginal empirical distribution functions of Y and Z.

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 $j = 1, \ldots, p$.

2. Sequential Simulation from a Marginal and Then a Conditional Simulates Joint Distributions.

   Simulation of a value x^* from $f(x)$ followed by simulation of a value y^* from $g(y|x^*)$ results in a value (x^*, y^*) simulated from the joint $p(x, y)$. By principle number 1 this also gives values from the marginal $h(y)$. In this way, we can accomplish the integration,

   $$h(y) = \int g(y|x)\, f(x)\, dx.$$

   Similarly, if Y and X are conditionally indpendent given Z so that $m(y|z) = m(y|x, z)$, then simulation of one value z^* from $f(z)$ followed

by simulation of one value x^* from $g(x|z^*)$, followed in turn by simulation of one value y^* from $m(y|z^*)$ produces one value (x^*, y^*, z^*) from the joint $p(x, y, z)$. Repeated to produce M values, the marginal empirical distribution of the values $\{y_m^* : m = 1, \ldots, M\}$ again approximates the marginal distribution,

$$h(y) = \int \int m(y|z)g(x|z)f(z)\,dx\,dz. \tag{7.17}$$

3. Averaging over Simulations Approximates Expectations

   This principle embodies the fundamental idea of Monte Carlo approximation. Consider 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 approximation to the expected value $E(X)$ is,

$$\hat{E}_M(X) = \frac{1}{M}\sum_{j=1}^{M} x_j^*,$$

   and a Monte Carlo approximation to the expected value of any suitable function $q(X)$ is,

$$\hat{E}_M\{q(X)\} = \frac{1}{M}\sum_{j=1}^{M} q(x_j^*). \tag{7.18}$$

   That $\hat{E}_M\{q(X)\}$ is consistent for $E\{q(X)\}$ follows immediately from the law of large numbers.

   This principle of simulation or, more specifically, Monte Carlo simulation, applies for values that correspond to independent and identically distributed realizations of random variables. If the simulated values are not independent, as will be the case when we discuss simulation from Markov Chains, then additional conditions are needed to ensure that

Monte Carlo averages continue to approximate expected values. There is a large literature on this issue and detailed discussion is beyond the scope of the material presented here.

7.8.2 Basic Methods of Simulation

Assume in this section that our goal is to simulate one or more values x^* from a univariate distribution having probability density function $f(x)$ such that $\Omega \equiv \{x : f(x) > 0\}$ is the support of $f(x)$. Note that $f(x)$ may be a parameterized density and we are suppressing explicit representation of this. What follows are several fundamental methods to obtain this goal. These methods can be useful in their own right, but also serve as foundations for more elaborate procedures that we will not discuss. A basic question that impacts simulation is whether we know the form of $f(x)$ exactly or whether all we know is the form of a function $g(x) \propto f(X)$, but not $f(x)$ itself.

Inversion

Perhaps the simplest of method for simulating from $f(x)$ occurs if we are able to derive in closed form the distribution function $F(x) = \int_{-\infty}^{x} f(t)\, d\mu(t)$. If X is a continuous random variable, the probability integral transform then implies that if $\{x_j : j = 1, \ldots, M\}$ is a random sample from $F(x)$, the values $\{u_j : j = 1, \ldots, M\}$ where $u_j = F(x_j)$ are a random sample from a uniform distribution on the unit interval, $Unif(0, 1)$. Thus, for any continuous distribution with density function $f(x)$, we may sample values $\{x_j^* : j = 1, \ldots, M\}$ from that distribution in the following way.

1. Simulate M values $\{u_j : j = 1, \ldots, M$ from a uniform distribution on the interal (0, 1).

2. For each j, compute values $x_j^* = F^{-1}(u_j)$. Then $\{x_j^* : j = 1, \ldots, M\}$ are a simulated sample from the distribution $F(x)$.

This technique is also easily adapted to sample from discrete distributions. Suppose that $\Omega = \{v_1, v_2, \ldots\}$ is the support of a probability mass function for a discrete random variable X, with $v_1 < v_2, < \ldots$ being ordered values. Note that Ω may be either finite or infinite in size. Then a sample of values $\{x_j^* : j = 1, \ldots, M\}$ may be simulated from the distribution with probability mass function $f(x)$ as follows.

1. Simulate M values $\{u_j : j = 1, \ldots, M\}$ from a uniform distribution on the interval (0, 1).

2. For each j, let $x_j^* = \min v_k \in \Omega\{v_k : u_j \leq F(v_k)\}$. Then $\{x_j^* : j = 1, \ldots, M\}$ are a simulated sample from the distribution $F(x)$.

Composition

Simulating from a distribution using the method of composition is essentially an application of one or more of the relations among distributions you remember so fondly from your introductory probability class. For example, to simulate M values from a chi-squared distribution with n degrees of freedom, we could simulate M sets of values $\{z_{j,k} : k = 1, \ldots, n\}; \;\; j = 1, \ldots, M$ from a standard normal distribution and then take $x_j^* = \sum_{k=1}^{n} z_{j,k}^2$. To simulate M values from an inverse gamma distribution with parameters α and β we could simulate M values $\{w_j : j = 1, \ldots, M\}$ from a gamma distribution with parameters α and β and then take $x_j^* = 1/w_j$ for $j = 1, \ldots, M$. Most of the pre-packaged simulation functions contained in computational software (e.g., `rpois` or `rgamma` in R) use either inversion or some type of composition

starting with psuedo-random numbers turned into realizations from a uniform distribution on the interval (0, 1).

Example 7.21

Suppose we have a multinomial model with k categories or groups. A basic Bayesian analysis might take the multinomial parameters $p_1, \ldots, p_{k-1}$ to have a Dirichlet distribution, which is conjugate for a multinomial data model. If the Dirichlet prior had parameters given as $\alpha_1, \ldots, \alpha_{k-1}, \beta$ and out of n observations y_i fall in category $i = 1, \ldots, k-1$, then the posterior is Dirichlet with parameters $\alpha_1 + y_1, \ldots, \alpha_{k-1} + y_{k-1}$ and $\beta + n - \sum_i y_i$. Suppose we would now like to simulate M observations from the posterior predictive distribution of $y_1, \ldots, y_k$. To accomplish this, we could use the following algorithm.

1. Simulate values w_i^* from gamma distributions with parameters $\alpha_i' = \alpha_i + y_i$ and $\beta' = \beta + n - \sum_i y_i$ for $i = 1, \ldots, k-1$.

2. Let $p_i^* = w_i^* / \sum_i w_i^*$ for $i = 1, \ldots, k-1$.

3. Simulate n values from a uniform distribution on the interval (0, 1), say $\{u_j^* : j = 1, \ldots, n\}$. Construct values for $y_i^*; \quad i = 1, \ldots, k-1$ as,

$$y_i^* = \sum_{j=1}^{n} I(u_j^* \leq \sum_{h \leq i} p_h^*)$$

   Then $(y_1^*, \ldots, y_{k-1}^*)$ is an observation from the desired posterior predictive distribution.

4. Repeat the above steps M times.

Notice that in Example 7.21 we have used composition in steps 1 and 2, combined with inversion in step 3.

Basic Rejection Sampling

Rejection sampling is a technique that allows us to sample from a distribution for which we know the density function only up to some constant of proportionality that will typically depend on the values of parameters of the distribution. We can formulate the general problem as follows. Suppose that we would like to sample from a distribution with probability density function $f_x(x)$ having support $x \in \Omega_x$, but all we know is a function $g(x)$ such that $f_x(x) \propto g(x)$, in other words, $f_x(x) = g(x)/\int g(x)\,dx$. Suppose in addition that we do have available a distribution with density $f_y(y)$ with the same support as $f_x(x)$ and that we do know how to simulate values from this distribution. Consider the following algorithm.

1. Simulate y^* from $f_y(y)$.

2. Let $x^* = y^*$ with a specified probability $h(y^*)$ that may depend on the value of y^*, or else reject y^* as a value of x^*, and return to step 1.

Rejection Sampling Result 1

Repeating the steps above until a "candidate" value y^* is accepted as a value of x^* produces one sampled value from a distribution with density proportional to $f_y(x)\,h(x)$; $\quad x \in \Omega_x$.

Proof:

Directly, we have that for any real constant c, $Pr[y^* < c \text{ and } y^* \text{ is accepted}] = \int_{-\infty}^{c} f_y(t)\,h(t)\,d\mu(t)$. Also, $Pr[y^* \text{ is accepted}] = \int_{-\infty}^{\infty} f_y(t)\,h(t)\,d\mu(t)$, this from the fact that the probability y^* is accepted is the probability that $y^* < \infty$ and

y^* is accepted. Then,

$$Pr[y^* < c | y^* \text{ is accepted}] = \frac{\int_{-\infty}^{c} f_y(t)\, h(t)\, d\mu(t)}{\int_{-\infty}^{\infty} f_y(t)\, h(t)\, d\mu(t)}.$$

If y^* is accepted, then any probability statement that applies to y^* also applies to x^*. Thus,

$$Pr[x^* < c] = \frac{\int_{-\infty}^{c} f_y(t)\, h(t)\, d\mu(t)}{\int_{-\infty}^{\infty} f_y(t)\, h(t)\, d\mu(t)}.$$

Taking the derivative with respect to c,

$$\frac{d}{c} Pr[x^* < c] = \frac{f_y(c)\, h(c)}{\int_{-\infty}^{\infty} f_y(t)\, h(t)\, d\mu(t)}, \tag{7.19}$$

which proves the result.

Rejection Sampling Result 2

Provided that $f_x(x) \leq M\, f_y(x) < \infty$ for some constant M, if we take $h(y^*)$ in the rejection algorithm to be

$$h(y^*) = \frac{f_x(y^*)}{f_y(y^*)\, M}, \tag{7.20}$$

then the density of x^* is $f_x(x^*)$.

Proof

Substituting (7.20) into (7.19) gives,

$$\frac{d}{c} Pr[x^* < c] = \frac{f_y(c)\, f_x(c)}{f_y(c) M \int_{-\infty}^{\infty} f_y(t) \frac{f_x(t)}{f_y(t)\, M\, d\mu(t)}} = f_x(c).$$

Result 2 also holds if we use $h(y^*) = g(y^*)/(f_y(y^*)\tilde{M})$ for an $\tilde{M}$ such that $g(x) \leq \tilde{M}\, f_y(x)$, and this then provides the means to simulate from $f_x(x)$ even if we only know a function $g(x)$ that is proportional to it, as in the introduction to this subsection. A general basic rejection algorithm is then,

1. Simulate y^* from $f_y(y)$.

2. Simulate u from a uniform distribution on the interval (0, 1).

3. If $Mu \leq f_x(y^*)/f_y(y^*)$ or $\tilde{M}u \leq g(y^*)/f_y(y^*)$ then take $x^* = y^*$. Otherwise reject y^* and return to step 1.

Key components of rejection sampling are finding a candidate distributon $f_y(y)$ (with the same support as the target distribution $f_x(x)$ such that (i) $f_y(y)$ is easy to sample from, and (ii) the probability that a candidate y^* is accepted as a value of x^* from $f_x(x)$ is high. Often, however, the most difficult part of a basic rejection algorithm is determining the value of the bounding constant M or $\tilde{M}$. There are several modifications of basic rejection that can help overcome this problem. We describe one of those next.

Ratio of Uniforms

Consider first the problem of sampling from a univariate distribution with density $f_x(x)$ for which the support is a bounded interval, $\Omega = (A, B)$, and the density is itself bounded. This can be accomplished by sampling uniformly on a bounded region that "covers" the density function $f_x(x)$ and then rejecting or "throwing away" any values that do not fall under the density.

Example 7.22

Let $f_x(x) = 4\,[\pi(1 + x^2)]^{-1};\ \ 0 < x < 1$. A graph of this density function is shown in Figure 7.3. Suppose we were to sample uniformly over the "bounding box" given by the dashed lines in Figure 7.3, that is, uniform on the region $[0,\, 1]\times[0,\, 4/\pi]$, and accept those values falling under the density while rejecting those falling above the density. Given a sufficient number of samples produced in this way we would accept values of x (the horizontal axis in Figure 7.3) with the correct relative frequencies. A rejection algorithm to accomplish this

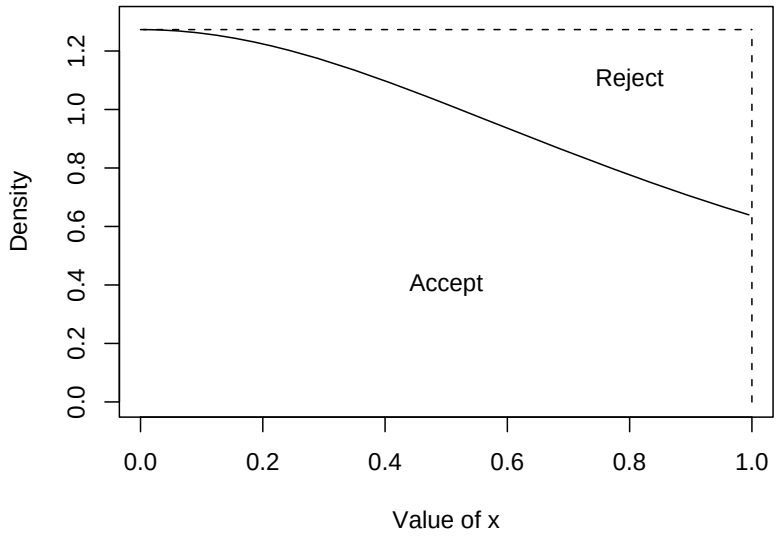


Figure 7.3: Density and sampling region for the distribution of Example 7.22.

would have the following form.

1. Simulate a value y^* from a uniform distribution on the interval (0, 1).

2. Simulate an independent value u from a uniform distribution on the interval (0, 1).

3. If $(4/\pi)u \leq 4[\pi(1+y^{*2})]^{-1}$ (equivalently, if $u \leq (1+y^{*2})^{-1}$) let $x^* = y^*$.

This is a basic rejection algorithm with $f_y(y) = 1; \;\; 0 < y < 1$, $f_x(x) = 4\,[\pi(1+x^2)]^{-1}; \;\; 0 < x < 1$, and $M = 4/\pi$.

The same idea can be extended to deal with other densities having non-

finite support through use of the following two results.

Ratio of Uniforms Result 1

Let $g(x)$ be any smooth function such that $g(x) \geq 0$; $-\infty < x < \infty$ and $\int_{-\infty}^{\infty} g(x)\,dx < \infty$. Let

$$C_g = \left\{(u, v) : 0 \leq u \leq [g(v/u)]^{1/2}\right\}. \tag{7.21}$$

If (u, v) are uniformly distributed on C_g then $x^* = v/u$ has density $f_x(x) = g(x)/\int g(t)\,dt$. The proof of this result will be presented below, but notice that, while this result does not detail the range of v, $v/0$ must be in the support of $f_x(x)$. Thus, if the range of v is strictly positive, the support of $f_x(x)$ must include ∞, if the range of v is strictly negative, the support of $f_x(x)$ must include $-\infty$ and, if v can be either positive or negative, the support of $f_x(x)$ must be $(-\infty, \infty)$. If $g(x)$ is proportional to a target density $f_x(x)$ from which we would like to sample, the next result indicates how to determine a rectangle that includes the region C_g.

Ratio of Uniforms Result 2

If $g(x)$ and $x^2 g(x)$ are both bounded, then a bounding box for the region C_g of Result 9.5 can be formed as $[0,\, a] \times [b_-,\, b_+]$, that is, $C_g \subset [0,\, a] \times [b_-,\, b_+]$, where

$$\begin{aligned} a &= [\sup\{g(x) : -\infty < x < \infty\}]^{1/2} \\ b_- &= -[\sup\{x^2 g(x) : x \leq 0\}]^{1/2} \\ b_+ &= [\sup\{x^2 g(x) : x \geq 0\}]^{1/2} \end{aligned} \tag{7.22}$$

Ratio of Uniforms Result 2 is purely geometric, and a proof may be found in Ripley (1987, p. 67). Ratio of Uniforms Result 1 is distributional, and we now

give a proof.

Proof of Ratio of Uniforms Result 1

Let $|C_g|$ be the area of the region C_g. If (u, v) are uniformly distributed on C_g, then the joint density is

$$f_{u,v}(u, v) = \frac{1}{|C_g|}; \quad 0 \leq u \leq [g(v/u)]^{1/2}.$$

Now, let $x = v/u$ and $y = u$. The Jacobian for this transformation is y and the joint density of x and y is $f_{x,y}(x, y) = y/|C_g|$, which leads to the marginal density of x as,

$$f_x(x) = \frac{1}{|C_g|} \int_0^{\sqrt{g(x)}} y \, dy = \frac{1}{2|C_g|} g(x).$$

Because $f_x(x)$ must be a density function, $|C_g| = (1/2) \int g(t) \, dt$, and then

$$f_x(x) = \frac{g(x)}{\int g(t) \, dt}.$$

A basic ratio of uniforms algorithm is then

1. Compute values of a, b_-, and b_+.

2. Simulate u_1 and u_2 as two independent values from a uniform distribution on the interval (0, 1).

3. Let $u = au_1$ and $v = b_- + (b_+ - b_-)u_2$.

4. If $(u, v) \in \{(u, v) : 0 \leq u \leq [g(v/u)]^{1/2}\}$ then let $x^* = v/u$.

Example 7.23

Let $g(x) = 1/(1 + x^2)$; $-\infty < x < \infty$. In this case we can determine the

region C_g of Ratio of Uniforms Result 1 exactly.

$$\begin{aligned} C_g &= \{(u,v) : 0 \le u \le [g(v/u)]^{1/2}\} \\ &= \left\{(u,v) : 0 \le u \le \left[\frac{1}{1+(v/u)^2}\right]^{1/2}\right\} \\ &= \left\{(u,v) : 0 \le u \text{ and } u^2 \le \frac{1}{1+(v/u)^2}\right\} \\ &= \{(u,v) : 0 \le u \text{ and } u^2 + v^2 \le 1\}, \end{aligned}$$

which is the right half of the unit circle. If we apply the basic ratio of uniforms algorithm, we have that $a = 1$ since $g(x) = 1/(1+x^2)$ is decreasing in $|x|$, $b_- = -1$, which is $\lim_{x\to-\infty} x^2/(1+x^2)$ and $b_+ = 1$ which is $\lim_{x\to\infty} x^2/(1+x^2)$.

Adaptive Ratio of Uniforms

Continue to consider the problem of sampling from a target density $f_x(x)$ with support Ω_x when all that is known is a function proportional to the target, $f_x(x) \propto g_x(x)$. The basic idea that any other function proportional to $g_x(x)$ is also proportional to $f_x(x)$ can be used to simplify a ratio of uniforms algorithm and render it useful even in cases for which the proportionality constant between $f_x(x)$ and $g_x(X)$ is enormously large or small, approaching or exceeding what can be evaluated by many computers.

If $g_x(x)$ is unimodal then the value x_a at which it attains its maximum value can be determined by any number of computational algorithms, for example an equal interval search in one dimension (see Chapter 5.7.2). If $g_x(x)$ is proportional to the target $f_x(x)$, then so also is $\tilde{g}_x(x) = g_x(x)/g_x(x_a)$. The

bounding box of a ratio of uniforms algorithm is then given by,

$$\begin{aligned} a &= [\sup\{\tilde{g}_x(x) : -\infty < x < \infty\}]^{1/2} = 1 \\ b_- &= -[\sup\{\tilde{g}_x(x) : x \leq 0\}]^{1/2} \\ b_+ &= [\sup\{\tilde{g}_x(x) : x \geq 0\}]^{1/2} \end{aligned}$$

If $g_x(x)$ is highly concentrated or otherwise differs from $f_x(x)$ by an extreme scaling factor, finding x_a is facilitated by working with the logarithm $\log\{g_x(x)\}$, and we will adopt this convention in presentation of the algorithm, which is as follows.

1. Determine the value x_a that maximizes $\log\{g_x(x)\}$ and let $g_m = \log\{g_x(x_a)\}$.

2. Determine the value x_{b+} that maximizes $\log\{x^2\, g_x(x)\} = 2\log(x) + \log\{g_x(x)\}$ for $x \geq 0$ and the value x_{b-} that maximizes $\log\{x^2\, g_x(x)\} = 2\log(x) + \log\{g_x(x)\}$ for $x \leq 0$.

3. Compute

$$\begin{aligned} b_+ &= \exp\left[2\log(x_{b+}) + \log\{g_x(x_{b+})\} - \log\{g_m\}\right] \\ b_- &= \exp\left[2\log(x_{b-}) + \log\{g_x(x_{b-})\} - \log\{g_m\}\right] \end{aligned}$$

4. Simulate u^* and v^* independently from uniform distributions on the interval (0, 1).

5. Let $u = u^*$ and $v = b_- + v^*(b_+ - b_-)$.

6. Simulate u_2 from a uniform distribution on the interval (0, 1).

7. If $2\log(u_2) \leq \log\{g_x(v/u)\} - \log(g_m)$ let $x^* = v/u$, otherwise return to step 4.

While the adaptive ratio of uniforms algorithm was designed to easily produce one sample from (conditional) posterior distributions in conjunction with a Gibbs Sampling algorithm (to come), we can also use it to produce a larger sample from a single posterior distribution in a manner similar to a basic rejection algorithm.

Example 7.24

Lawless (1982, p. 86) presents an example of failure times of airplane components, with the original data attributed to Mann and Fortig (1973). The sample size is small with $n = 10$ and the data are given as 0.22, 0.50, 0.88, 1.00, 1.32, 1.33, 1.54, 1.76, 2.50 and 3.00 (the units of time in this example are unknown to me). Suppose that, after examination of a stem-and-leaf plot, we decide to model these failure times as realizations from a model for random variables having independent and identical exponential distributions with parameter $\beta > 0$. The data model is then $Y_1, \ldots, Y_n \sim iid$ with common density $f(y|\beta) = \beta \exp(-\beta y); \quad y > 0$. If we make use of an improper prior for β the posterior is,

$$p(\beta|\boldsymbol{y}) \propto g(\beta|\boldsymbol{y}) = \beta^n \exp(-\beta \sum y_i); \quad \beta > 0,$$

which is shown for the data of this example in Figure 7.4. It can be shown this is an integrable function of β so that the posterior is proper. Using an equal interval algorithm with $\log\{g(\beta|\boldsymbol{y})\}$, the maximum of $g(\beta|\boldsymbol{y})$ is found to occur at $\beta = 0.71$. A sample of size $50,000$ from $p(\beta|\boldsymbol{y})$ produced from an adaptive ratio of uniforms algorithm produced a posterior mean and variance of 0.785 and 0.0559, respectively. A 95% central credible interval was (0.388, 1.308). For comparison, maximum likelihood produced a point estimate of $\hat{\beta} = 0.712$ with estimated variance $\hat{V}\hat{\beta} = 0.0506$ and a 95% Wald confidence interval of (0.271, 1.153). Although the Bayesian and likelihood results are similar, it

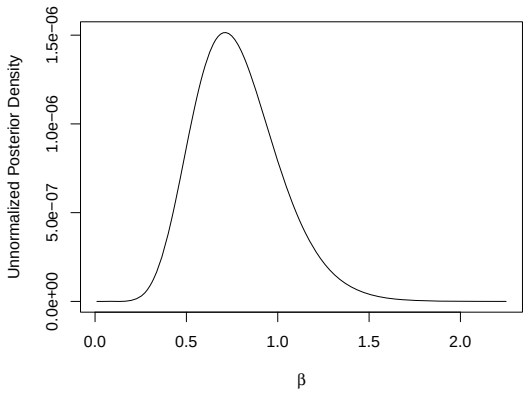


Figure 7.4: Unnormalized posterior density of β from exponential model with improper prior.

appears from this example that improper priors for scale parameters may not give the same results as likelihood, which they do for location parameters.

7.9 Simulation From Unnormalized Posteriors and MCMC

In the examples of the previous section we simulated from distributions *directly*, meaning that we knew each pair of values in the algorithm presented was a draw from the *target distribution*, the distribution from which we wanted to produce samples. In the majority of modern Bayesian the joint posterior is unavailable in closed form. If we are able to simulate from these posteriors, however, we can still make inference within a Bayesian framework. In this section we introduce the general problem and indicate that a set of procedures that fall under the title of *Markov Chain Monte Carlo* (MCMC) methods can

provide the tools we need to simulate from the posterior distribution for many models.

The basic problem to be addressed through the use of MCMC is not difficult to formulate. 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

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \end{aligned} \tag{7.23}$$

In simple examples we made use of the last line of (7.23) to recognize the kernel of known distributions, thus avoiding the need to formally evaluate the integral in the denominator of the first line of (7.23). In most applied problems we will find that the last line above cannot be matched with the kernel of a known distribution nor can the integral be evaluated analytically. Notice, however, that as functions of $\boldsymbol{\theta}$ the last line does give us a formula that is proportional to the posterior we desire to find. The integral is not a function of $\boldsymbol{\theta}$ and thus is a constant for the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. So the generic problem we are faced with is the desire to simulate from a distribution $f(x)$ when all we know is some other function that is not a distribution, $g(x) \propto f(x)$.

7.9.1 Markov Chain Samplers

The fundamental principles of simulation were presented in the context of simulated values (or *draws*) that were independent and we knew the distribution from which they were drawn. These results 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 Monte Carlo approximations $E_M\{q(X)\}$ and $F_M(x)$ converge to $E\{q(X)\}$ and $F(x)$ (as $M \to \infty$) just as for 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, meaning that values in the chain cover or visit all possible values in Ω and do so with relative frequencies dictated by F. Suppose we partition Ω into an 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 of adequate quality. Sequences $X(t)$ that have these behaviors are called ergodic.

If the sequence of variables $\{X(t) : t = 0, 1, \ldots\}$ involve dependencies among elements, it should be intuitive that the properties of those dependencies will determine whether the chain is ergodic or not. While not enough, by itself, to ensure ergodicity, dependencies that follow what is called a *Markov property* make it much easier to verify other conditions that are sufficient to

ensure ergodicity. A chain $\{X(t) : t = 0, 1, \ldots, \}$ is said to be a Markov chain if the conditional distribution of $X(t)$ given all previous values is the same as the conditional distribution of $X(t)$ given only $X(t-1)$. Using $[X]$ to denote "the distribution of X", this Markov property can be formally stated as, for any $t \geq 1$,

$$[X(t) \,|\, X(t-1), X(t-2), \ldots, X(0)] = [X(t) \,|\, X(t-1)]. \tag{7.24}$$

There are a couple of points that should be kept in mind about what this Markov property does and does not imply.

1. The Markov property (7.24) **does not** imply that $X(t)$ is independent of $X(t-2)$, or any other $X(t-k)$ for $k > 1$. The independence implied by (7.24) is *conditonal*.

2. The Markov property (7.24) **does** imply that, in an explicit formula for a conditional probability density function or probability mass function giving the left hand side of (7.24), the only other variable that will appear in that formula is $X(t-1) = x(t-1)$.

The last section of this chapter contains an introduction to the theory of Markov chains. It will be indicated there that a Markov chain will be ergodic if it is (1) irreducible, (2) positive recurrent, and (3) aperiodic. Typically, it is positive recurrence that is the most difficult to verify.

If we are attempting to produce a sample from a distribution $F(x)$, having a Markov chain is not enough. Having an ergodic Markov chain is not enough. We need an ergodic Markov chain that mimics the probabilistic behavior of $F(x)$. Constructing such a Markov chain will usually only be possible in the limit. That is, the probabilistic behavior of values in the chain $\{X(t) : t = 0, \ldots, \}$ will only agree with those dictated by $F(x)$ as $t \to \infty$.

The distribution $F(x)$ is known as the *target distribution* of the chain. Although probabilities reflected by relative frequencies of values assumed by the chain will only converge to those of $F(x)$ as $t \to \infty$, as with all asymptotic results we assume that at some point the approximation is close enough to prove useful in making inferences. Thus, in practice, we construct an appropriate Markov chain, run it (on the computer) for a certain number of iterations (or cycles) say $t = 1, 2, \ldots, B$, discard all values obtained to that point, and then start collecting subsequent values produced by the chain. The number of values discarded ,B, is called the *burn-in* period, and it is then assumed that all subsequent values can be treated as samples from the target distribution $F(x)$, although still not necessarily independent samples. In the coming sections we will discuss several algorithms for constructing ergodic Markov chains that have the target distributions we wish to sample from, which in Bayesian analyses will be joint posteriors $p(\boldsymbol{\theta}|\boldsymbol{y})$.

7.9.2 Metropolis-Hastings

We can now present the first of two MCMC algorithms that can be used to simulate from intractable posterior distributions. What are known as Metropolis-Hastings Algorithms were originally due to Metropolis et al. (1953) who studied the behavior of molecules in statistical physics, and were generalized by Hastings (1970). There are now any number of specific Monte Carlo sampling algorithms that fit under this general heading, distinguished by the manner in which *candidate values* for *jumps* of the chain are generated. These terms will become clear as we proceed. Metropolis-Hastings algorithms may be used to sample from multivariate distributions and/or from distributions that are known only up to some constant of proportionality.

General Form

The general form of Metropolis-Hastings algorithms (MH) is as follows. We take as our goal the simulation of a sample from some target distribution that has probability mass or density function $p(\cdot)$; we will assume in this section that p is a density. As previously mentioned, in our Bayesian applications this target will be $p(\boldsymbol{\theta}|\boldsymbol{y})$, a joint posterior. But because MH does not require the target distribution to be a posterior we will here simply consider some joint distribution $p(\boldsymbol{x})$ with support Ω.

Assume that a Markov chain $\{\boldsymbol{X}(t) : t = 0, 1, \ldots\}$ is in a state $\boldsymbol{X}(t) = \boldsymbol{x}_t \in \{\boldsymbol{x} : p(\boldsymbol{x}) > 0\}$ at time t. Assume further that we have available a density $q(\boldsymbol{y}|\boldsymbol{x})$ with support that is either the same as or larger than the support of $p(\boldsymbol{x})$. Given a simulated value $\boldsymbol{y}^*$ from $q(\boldsymbol{y}|\boldsymbol{x}_t)$, let $\boldsymbol{X}(t+1) = \boldsymbol{y}^*$ with probability α defined below, otherwise let $\boldsymbol{X}(t+1) = \boldsymbol{x}_t$. The acceptance probability for the candidate value $\boldsymbol{y}^*$ is defined as,

$$\alpha(\boldsymbol{x}_t, \boldsymbol{y}^*) = \min\left\{\frac{p(\boldsymbol{y}^*)\, q(\boldsymbol{x}_t|\boldsymbol{y}^*)}{p(\boldsymbol{x}_t)\, q(\boldsymbol{y}^*|\boldsymbol{x}_t)},\, 1\right\}. \tag{7.25}$$

Do not confuse this acceptance probability with the acceptance probability of a rejection algorithm. In rejection sampling a candidate value for the target distribution is either accepted or rejected and, if rejected, it is simply discarded and a new candidate produced. In contrast, the candidate $\boldsymbol{y}^*$ of a Metropolis-Hastings algorithm is a candidate for a *jump* or change in the state of a Markov chain. If it is accepted the chain makes a transition from $\boldsymbol{X}(t) = \boldsymbol{x}_t$ to $\boldsymbol{X}(t+1) = \boldsymbol{y}^*$. If it is not accepted the chain also makes a transition (but to the same value), namely $\boldsymbol{X}(t+1) = \boldsymbol{x}_t$. We can then say that the chain always moves from $\boldsymbol{X}(t) = \boldsymbol{x}_t$ to $\boldsymbol{X}(t+1) = \boldsymbol{y}$ and the possible values for $\boldsymbol{y}$ are $\boldsymbol{y} = \boldsymbol{y}^*$ or $\boldsymbol{y} = \boldsymbol{x}_t$. The form of $\alpha(\boldsymbol{x}, \boldsymbol{y})$ in (7.25) automatically ensures that the chain is reversible. For an irreducible chain this is sufficient to ensure an invariant

distribution of $p(\boldsymbol{x})$, the target distribution. Combined with aperiodicity, we have that the limit distribution is also $p(\boldsymbol{x})$.

Now, notice that the desired target distribution $p(\cdot)$ enters this progression only through the ratio $p(\boldsymbol{y}^*)/p(\boldsymbol{x}_t)$, which means it is enough to know $p(\cdot)$ only up to a constant. In other words, suppose that the target density $p(\cdot)$ is a posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$, but all we know about this posterior is that $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$ for data model f and prior distribution $\pi(\boldsymbol{\theta})$. Suppose that the current value of the chain is $\boldsymbol{\theta}_t$, which is $\boldsymbol{x}_t$ in (7.25) and the jump candidate is $\boldsymbol{\theta}^*$, which is $\boldsymbol{y}^*$ in (7.25). Let $p(\boldsymbol{\theta}|\boldsymbol{y}) = k(\boldsymbol{y})f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where

$$k^{-1}(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

In this case, the acceptance probability (7.25) becomes

$$\begin{aligned}
\alpha(\boldsymbol{\theta}_t, \boldsymbol{\theta}^*) &= \min\left\{\frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})\,q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}_t|\boldsymbol{y})\,q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_t)},\,1\right\} \\
&= \min\left\{\frac{f(\boldsymbol{y}|\boldsymbol{\theta}^*)\,\pi(\boldsymbol{\theta}^*)\,k(\boldsymbol{y})\,q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{f(\boldsymbol{y}|\boldsymbol{\theta}_t)\,\pi(\boldsymbol{\theta}_t)\,k(\boldsymbol{y})\,q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_t)},\,1\right\} \\
&= \min\left\{\frac{f(\boldsymbol{y}|\boldsymbol{\theta}^*)\,\pi(\boldsymbol{\theta}^*)\,q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{f(\boldsymbol{y}|\boldsymbol{\theta}_t)\,\pi(\boldsymbol{\theta}_t)\,q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_t)},\,1\right\}. \qquad (7.26)
\end{aligned}$$

The (rather dramatic) implication of (7.26) is that it is not actually necessary to know $p(\boldsymbol{\theta}|\boldsymbol{y})$ in close form in order to simulate from it. It is necessary only to know the data model f the prior distribution π and the proposal distribution q.

We now return to the convention in discussion of Markov chain samplers of using $p(\boldsymbol{x})$ to denote the target distribution.

Versions of Metropolis-Hastings

As mentioned previously, different versions of Metropolis-Hastings algorithms result from different choices of the candidate density $q(\boldsymbol{y}|\boldsymbol{x})$. We briefly list several of the possibilities here, drawing heavily on Tierney (1996).

1. Original Metropolis-Hastings.

   What is often referred to as the original version of the Metropolis-Hastings algorithm, we specify the candidate density such that $q(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{y})$. Then the acceptance probability for a value $\boldsymbol{y}^*$ simulated from $q(\boldsymbol{y}|\boldsymbol{x}_t)$ is

   $$\alpha(\boldsymbol{x}_t, \boldsymbol{y}^*) = \min\left\{\frac{p(\boldsymbol{y}^*)q(\boldsymbol{x}_t|\boldsymbol{y}^*)}{p(\boldsymbol{x}_t)q(\boldsymbol{y}^*|\boldsymbol{x}_t)}, 1\right\} = \min\left\{\frac{p(\boldsymbol{y}^*)}{p(\boldsymbol{x}_t)}, 1\right\}.$$

   Example 7.25

   Suppose that the support of a univariate target distribution $p(x)$ is the set of non-negative integers $x \in \{0, 1, \ldots, \}$. We might choose the proposal density as,

   $$\begin{aligned} q(y|x \neq 0) &= \begin{cases} \frac{1}{2} & \text{for } y = x - 1 \\ \frac{1}{2} & y = x + 1 \end{cases} \\ q(y|0) &= \begin{cases} \frac{1}{2} & y = 0 \\ \frac{1}{2} & y = 1 \end{cases} \end{aligned}$$

2. Independence Chains.

   Independence chains are so named because the proposal density $q(\boldsymbol{y}|\boldsymbol{x})$ is taken to be independent of the current state $\boldsymbol{x}_t$, so that $q(\boldsymbol{y}|\boldsymbol{x}_t) = h_y(\boldsymbol{y})$ for some density $h_y(\boldsymbol{y})$. The acceptance probability then becomes,

   $$\alpha(\boldsymbol{x}_t, \boldsymbol{y}^*) = \min\left\{\frac{p(\boldsymbol{y}^*)h_y(\boldsymbol{x}_t)}{p(\boldsymbol{x}_t)h_y(\boldsymbol{y}^*)}, 1\right\} = \min\left\{\frac{w(\boldsymbol{y}^*)}{w(\boldsymbol{x}_t)}, 1\right\},$$

   where $w(\boldsymbol{x}) = p(\boldsymbol{x})/h_y(\boldsymbol{x})$.

3. Random Walk Chains.

   To construct a Metropolis-Hastings algorithm using a random walk for candidate jumps, let $f_z(\cdot)$ be a density with the same support as the target $p(\boldsymbol{x})$. Simulate $\boldsymbol{z}^*$ from f_z independent of the current state $\boldsymbol{x}_t$, and let $\boldsymbol{y}^* = \boldsymbol{x}_t + \boldsymbol{z}^*$. Then the candidate density is

$$q(\boldsymbol{y}|\boldsymbol{x}) = f_z(\boldsymbol{y} - \boldsymbol{x}).$$

   For many problems f_z can be taken as Gaussian if one has some idea of the covariance matrix that should be used.

   Random walk chains have become popular because they lend themselves to "tuning" a Metropolis-Hastings algorithm to achieve a desired proportion of acceptance of jump proposals. The proportion of proposed jumps that are accepted is related to the idea of mixing described briefly in introducing ergodicity. Briefly, one wants a Metropolis-Hastings algorithm to accept proposed jumps often enough to ensure that the entire sample space of $\boldsymbol{x}$ is covered, but seldom enough so that regions of high probability are correctly reflected by the values of the chain. Rough rules of thumb have been suggested to the effect that the percentage of proposed jumps that are accepted should be in the range of 20% to 60%. Typically, one monitors the proportion of jump proposals that are being accepted in a chain. If that proportion is too high (say 70% or greater) then one can increase the variance of the random walk proposal distribution to decrease it. Conversely, if the acceptance of jump proposals is too low (say 15% or less) then one can decrease the variance of the proposal distribution.

7.9.3 Gibbs Sampling

Gibbs Sampling algorithms have become a common class of algorithms for simulating values from posterior distributions. They also are often useful in simulation of data sets from complex models. Widespread statistical awareness of the Gibbs sampler started with Geman and Geman (1984) and was accelerated by Gelfand and Smith (1990), although its roots go back much further (see Cassella and George, 1992; Gelfand, 2000).

The Basic Gibbs Algorithm

A Gibbs Sampling algorithm is based on the idea that, under suitable conditions, simulated values can be obtained from a joint distribution $p(\boldsymbol{x})$ by sequentially simulating values from a set of simpler conditional distributions. Specifically, let $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_p)^T$ have p scalar components, and let $\{\tilde{\boldsymbol{x}}_q : q = 1, \ldots, k\}$ denote some partition of $\boldsymbol{x}$. Any given component $\tilde{\boldsymbol{x}}_q$ may contain a single element such as $\tilde{x}_1 = x_1$ or multiple components such as $\tilde{\boldsymbol{x}}_1 = (x_1, x_2)$. We assume that we have available the conditional distributions $p(\tilde{\boldsymbol{x}}_q|\{\tilde{\boldsymbol{x}}_j : j \neq q\})$, and a means to simulate values from these conditionals. Often, we take $q = i$ and $\tilde{\boldsymbol{x}}_i = x_i$ so that the conditional distributions are univariate full conditionals $p(x_i|\{x_j : j \neq i\})$; $i = 1, \ldots, n$. When this is the case we need only simulate from univariate distributions, which can simplify the required computations.

A Gibbs Sampling algorithm is described as follows.

1. Choose a starting value $\boldsymbol{x}^{(0)}$ within the set of possible values for $\boldsymbol{x}$, and form the partition $\{\tilde{\boldsymbol{x}}_q : q = 1, \ldots, k\}$.

2. At iteration $t = 1, \ldots$, select an ordering of the indices $\phi(1, 2, \ldots, k)$. Here, ϕ may be either a random permutation operator or the identity

function, which produces what are called *random-scan* and *systematic-scan* algorithms, respectively. Re-index the $\tilde{\boldsymbol{x}}$ according to the selected ordering.

3. For $q = 1, \ldots, k$, simulate $\tilde{\boldsymbol{x}}_q^{(t)}$ from the conditional distribution with density

$$p\left(\tilde{\boldsymbol{x}}_q | \{\tilde{\boldsymbol{x}}_j^{(t)} : j < q\}, \{\tilde{\boldsymbol{x}}_j^{(t-1)} : j > q\}\right).$$

It sometimes appears as if there are nearly as many ways to understand the Gibbs Sampler as there are statisticians trying to understand it. Gelfand and Smith (1990) and Cassella and George (1992) emphasize connections between Gibbs sampling and what are sometimes called substitution algorithms, which have an exact correspondance in the bivariate case but not higher dimensions. Smith and Roberts (1993), Tierney (1994), and Liu (2001) all approach Gibbs sampling from the standpoint of general state space Markov chain theory, although in reading each of these authors one can get the impression that different aspects of the theory constitute the crucial component for success. Some of the confusion that easily results from reading literature on the Gibbs sampler can be avoided by understanding that Gibbs algorithms are very flexible. Gibbs algorithms can be applied to simulation of posteriors in Bayesian analysis of typical statistical models, which is the focus of this chapter, but can also be applied to simulation of data from models with complex dependence structures, or simulation of values from distributions being used as importance sampling distributions in Monte Carlo evaluation of integrals or problems that combine several of these features. Some of the differences in emphasis among authors describing Gibbs algorithms can be explained by either differences among the problems being considered or the level of generality the authors wish to convey.

Conventional Bayesian Problems

Assume that the full conditional distributions to be used in a Gibbs algorithm have densities and will be written for individual scalar elements of the variable for which a joint distribution is to be simulated. Consider a problem in which a data model has been formulated for a set of random variables $Y_1, \ldots, Y_n$ and leads to a joint probability mass or density function $f(\boldsymbol{y}|\boldsymbol{\theta})$ for some parameter $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)$ such that, for $\boldsymbol{\theta} \in \Theta$, $\Omega_{\boldsymbol{y}} = \{\boldsymbol{y} : f(\boldsymbol{y}|\boldsymbol{\theta}) > 0\}$. Suppose that $\boldsymbol{\theta}$ has been assigned the joint prior $\pi(\boldsymbol{\theta})$ that is either proper or in such a way that we know the posterior is proper. Then we know that $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$, and the right hand side of this expression is available in closed form.

Situations for which this is the case provide a huge advantage in verifying that a Gibbs algorithm to simulate from $p(\boldsymbol{\theta}|\boldsymbol{y})$ satisfies the conditions of being irreducible, positive (Harris) recurrent, and aperiodic. This is because we then know the target distribution in closed form, at least up to a constant of proportionality. And, we know that the conditional distributions $p(\theta_i|\boldsymbol{y}, \{\theta_j : j \neq i\})$ correspond to that joint posterior, and we know that the appropriate marginal distributions exist as well. What is needed is then to verify that the algorithm produces a Markov chain that is irreducible, aperiodic, and has $p(\boldsymbol{\theta}|\boldsymbol{y})$ as its invariant distribution so that the chain converges to $p(\boldsymbol{\theta}|\boldsymbol{y})$. The first two of these, irreducibility and aperiodicty, follow from conditions on the support of the marginal distributions for $\theta_1, \ldots, \theta_p$. Although stronger than what is necessary, the following condition, called the *positivity condition* will lead to irreducibility and aperiodicity.

Definition

For a variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$, let $\Omega_{\boldsymbol{x}}$ denote the support of the joint distribution and Ω_i denote the support of the marginal distribution of x_i; $i =$

$1, \ldots, p$, assuming all of these distributions exist. The positivity condition is satisfied if $\Omega_{\boldsymbol{x}} = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_p$.

The positivity condition essentially states that, if θ_i^* is a possible value of θ_i, then θ_i^* can occur in combination with any of the possible values for the remaining components of $\boldsymbol{\theta}$, $\{\theta_j : j \neq i\}$. The implication is then that any of the possible values of θ_i can be simulated with positive probability from the full conditional $p(\theta_i|\boldsymbol{y}, \{\theta_j : j \neq i\})$ for any set of conditioning values, and that this is true for all θ_i; $i = 1, \ldots, p$. Because this is true for all of the full conditional distributions being used in the Gibbs algorithm, one transition, which is a move from $(\theta_1^{(t-1)}, \ldots, \theta_p^{(t-1)})$ to $(\theta_1^{(t)}, \ldots, \theta_p^{(t)})$, can result in any of the possible values $\boldsymbol{\theta} \in \Theta$ with positive probability. This immediately gives irreducibility and aperiodicity.

It remains to show then, that $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the invariant distribution of the Gibbs algorithm. This can be demonstrated directly as a consequence of reversibility for what were previously called random-scan algorithms, and indirectly for what were called systematic-scan algorithms. Systematic-scan algorithms do not possess the property of being reversible, but random-scan algorithms do. This will be demonstrated for a finite state-space example, but the principle applies to general state-space chains as well.

Example 7.26

Suppose that the target distribution for a Gibbs algorithm is the joint distribution for two variables X and Y that have the same sets of possible values $\Omega = \{\omega_1, \omega_2, \ldots\}$ and satisfy the positivity condition. Let x_1, x_2, y_1 and y_2 be any values in Ω. The Markov chain defined by a systematic-scan algorithm makes the transition from (x_1, y_1) to (x_2, y_2) with probability $Pr(x = x_2, |y = y_1)\, Pr(y = y_2|x = x_2)$. To demonstrate reversibility we need

to show that, for any x_1, y_1, x_2, y_2,

$$\begin{aligned} & [Pr(x = x_1, y = y_1)\, Pr(x = x_2|y = y_1)\, Pr(y = y_2|x = x_2)] \\ = \; & [Pr(x = x_2, y = y_2)\, Pr(x = x_1|y = y_2)\, Pr(y = y_1|x = x_1)] \end{aligned}$$

which implies that

$$\begin{aligned} & Pr(x = x_1, y = y_1)\, \frac{Pr(x = x_2, y = y_1)}{Pr(y = y_1)}\, \frac{Pr(x = x_2, y = y_2)}{Pr(x = x_2)} \\ = \; & Pr(x = x_2, y = y_2)\, \frac{Pr(x = x_1, y = y_2)}{Pr(y = y_2)}\, \frac{Pr(x = x_1, y = y_1)}{Pr(x = x_1)}. \end{aligned}$$

Consider the case that $x_2 \neq x_1$ but $y_2 = y_1 = y^*$, so the transition under consideration is from (x_1, y^*) to (x_2, y^*). Then, if reversibility holds,

$$\begin{aligned} & Pr(x = x_1, y = y^*)\, \frac{Pr(x = x_2, y = y^*)}{Pr(y = y^*)}\, \frac{Pr(x = x_2, y = y^*)}{Pr(x = x_2)} \\ = \; & Pr(x = x_2, y = y^*)\, \frac{Pr(x = x_1, y = y^*)}{Pr(y = y^*)}\, \frac{Pr(x = x_1, y = y^*)}{Pr(x = x_1)}, \end{aligned}$$

which implies that,

$$Pr(y = y^*|x = x_1) = Pr(y = y^*|x = x_2),$$

and this cannot be true unless x and y are independent. Thus, the systematic-scan algorithm is not reversible. Now consider a random-scan algorithm in which the first conditional step of the transition is chosen at random, here with probability 0.5. Then a transition from (x_1, y_1) to (x_2, y_2) occurs with probability $0.5Pr(x = x_2|y = y_1)\, Pr(y = y_2|x = x_2) + 0.5\, Pr(y = y_2|x = x_1)\, Pr(x = x_2|y = y_2)$. Reversibility then requires that,

$$\begin{aligned} & Pr(x = x_1, y = y_1)\, \{0.5Pr(x = x_2|y = y_1)\, Pr(y = y_2|x = x_2) \quad + \\ & \qquad 0.5Pr(y = y_2|x = x_1)\, Pr(x = x_2|y = y_2)\} \\ = \; & Pr(x = x_2, y = y_2)\, \{0.5Pr(x = x_1|y = y_2)\, Pr(y = y_1|x = x_1) \quad + \\ & \qquad 0.5Pr(y = y_1|x = x_2)\, Pr(x = x_1|y = y_1)\} \end{aligned}$$

which implies that

$$\begin{aligned} & Pr(x = x_1, y = y_1) \left\{ 0.5\, Pr(x = x_2 | y = y_1) \frac{Pr(x = x_2, y = y_2)}{Pr(x = x_2)} \quad + \right. \\ & \qquad \left. 0.5\, Pr(y = y_2 | x = x_1) \frac{Pr(x = x_2, y = y_2)}{Pr(x = x_2)} \right\} \\ = \quad & Pr(x = x_2, y = y_2) \left\{ 0.5\, Pr(x = x_1 | y = y_2) \frac{Pr(x = x_1, y = y_1)}{Pr(x = x_1)} \right. \\ & \qquad \left. 0.5\, Pr(y = y_1 | x = x_2) \frac{Pr(x = x_1, y = y_1)}{Pr(y = y_1)} \right\} \end{aligned}$$

which in turn gives

$$\begin{aligned} & Pr(x = x_1, y = y_1)\, Pr(x = x_2, y = y_2) \left\{ 0.5 \frac{Pr(x = x_2 | y = y_1)}{Pr(x = x_2)} \quad + \right. \\ & \qquad \left. 0.5\, \frac{Pr(y = y_2 | x = x_1)}{Pr(y = y_2)} \right\} \\ = \quad & Pr(x = x_2, y = y_2)\, Pr(x = x_1, y = y_1) \left\{ 0.5 \frac{Pr(x = x_1 | y = y_2)}{Pr(x = x_1)} \quad + \right. \\ & \qquad \left. 0.5 \frac{Pr(y = y_1 | x = x_2)}{Pr(y = y_1)} \right\} \end{aligned}$$

Then canceling the leading factors and the definition of conditional probabilities gives that,

$$\begin{aligned} & \frac{Pr(x = x_2, y = y_1)}{Pr(x = x_2)\, Pr(y = y_1)} \frac{Pr(x = x_1, y = y_2)}{Pr(x = x_1)\, Pr(y = y_2)} \\ = \quad & \frac{Pr(x = x_1, y = y_2)}{Pr(x = x_1)\, Pr(y = y_2)} + \frac{Pr(x = x_2, y = y_1)}{Pr(x = x_2)\, Pr(y = y_1)} \end{aligned}$$

which holds for any $x_1, y_1, x_2, y_2 \in \Omega_{\boldsymbol{x}}$, verifying reversibility for the random-scan chain so that the joint distribution of x and y is an invariant distribution for the chain. Combined with irreducibility and aperiodicity (as established by the positivity condition) we have that this form of Gibbs sampler converges to the desired distribution.

The result of Example 7.26 generalizes to higher dimensions and to general state-space chains, so that random-scan chains are reversible and give the desired target distribution as the invariant distribution, if that target distribution is a joint that is known to exist and known to have the conditionals used to construct the Gibbs algorithm. It turns out that systematic-scan algorithms also typically have the desired joint as an invariant distribution and hence also converge to that distribution, although showing this becomes more involved. Roberts and Smith (1994) give general conditions under which Gibbs algorithms converge.

Unconventional Problems

It is worth re-iterating that the relative ease with which Gibbs algorithms for conventional problems can be shown to be irreducible, aperiodic and to have the desired invariant distribution depend heavily on the fact that we formulate such algorithms based on the conditional distributions that correspond to a known joint, which is the target distribution. This, combined with the positivity condition essentially provides enough to ensure that the algorithm behaves properly, at least under a random-scan formulation. But, as alluded to previously, Gibbs algorithms can also be useful in situations for which we do not know the joint target distribution and have available only a set of full conditionals.

The critical aspect of ensuring success in a Gibbs sampling algorithm then consists of showing that the set of full conditional distributions at hand determines a corresponding joint. The following result, due to Arnold and Press (1989) indicates what is needed in the case of a bivariate situation for two

variables x_1 and x_2.

Result (Arnold and Press, 1989)

Suppose variables x_1 and x_2 have specified conditional densities $p_1(x_1|x_2)$ and $p_2(x_2|x_1)$ with respect to measures μ_1 and μ_2, respectively. Let

$$\begin{aligned} N_1 &\equiv \{(x_1, x_2) : p_1(x_1|x_2) > 0\} \\ N_2 &\equiv \{(x_1, x_2) : p_2(x_2|x_1) > 0\} \end{aligned}$$

A joint density $\pi(x_1, x_2)$ exists and has p_1 and p_2 as its conditionals if and only if the following two conditions hold.

1. $N_1 = N_2 = N$
2. For all $(x_1, x_2) \in N$ there exist functions $u(x_1)$ and $v(x_2)$ such that

$$\frac{p_1(x_1|x_2)}{p_2(x_2|x_1)} = u(x_1)\, v(x_2) \text{ where } \int u(x)\, d\mu_1(x) < \infty.$$

Note that the positivity condition is sufficient to ensure that condition 1 of this result is met. Condition 2 of the result essentially ensures that appropriate marginals $f_1(x_1)$ and $f_2(x_2)$ exist such that $p(x_1|x_2) = \pi(x_1, x_2)/f_2(x_2)$ and $p_2(x_2|x_1) = \pi(x_1, x_2)/f_1(x_1)$ because $u(x_1) \propto f_1(x_1)$ and $v(x_2) \propto 1/f_2(x_2)$. The integrability condition indicates that these marginals integrate to 1.

The following example indicates that not all sets of full conditionals determine a corresponding joint, for a problem in which we do not really need this advanced theory.

Example 7.27

Consider the pair of conditional distributions where $p_1(x_1|x_2)$ is a normal density with mean $\gamma\, x_2$ and variance σ_1^2, while $p_2(x_2|x_1)$ is a normal density with

mean βx_1 and variance σ_2^2. Since the support of each of these conditional distributions is the entire real line, condition 1 of the Result of Arnold and Press is met. Now, after a bit of algebra, we have that

$$\frac{p_1(x_1|x_2)}{p_2(x_2|x_1)} = \frac{\sigma_2}{\sigma_1} \exp\left[-\frac{1}{2\sigma_1^2\sigma_2^2}\left\{x_1^2(\sigma_2^2 - \sigma_1^2\beta^2) + x_2^2(\sigma_2^2\gamma^2 - \sigma_1^2) - 2x_1x_2(\sigma_2^2\gamma - \sigma_1^2\beta)\right\}\right].$$

If condition 2 of the Result is to be met we must have that the cross-product term equals zero, that is,

$$2x_1x_2(\sigma_2^2\gamma - \sigma_1^2\beta) = 0 \Rightarrow \frac{\gamma}{\sigma_1^2} = \frac{\beta}{\sigma_2^2}.$$

We also need $\int u(x_1)\, d\mu_1(x_1) < \infty$ and here we can take this to be

$$\frac{\sigma_2}{\sigma_1} \int \exp\left\{-\frac{1}{\sigma_1^2\sigma_2^2} x_1^2(\sigma_2^2 - \sigma_1^2\beta^2)\right\} dx_1 < \infty.$$

This integral will be finite if and only if $\sigma_2^2 - \sigma_1^2\beta^2 > 0$, or with the already existing condition that $\gamma\,\sigma_2^2 = \beta\,\sigma_1^2$,

$$\beta^2\,\sigma_1^2 < \sigma_2^2 \Rightarrow |\beta|\,|\gamma| < 1.$$

Thus, in this example the two specified conditionals determine a joint if $\gamma\,\sigma_2^2 = \beta\,\sigma_1^2$ and $|\beta|\,|\gamma| < 1$, but not otherwise.

The parameter restrictions in this example match what is needed for a bivariate normal distribution to exist, the demonstration of which is a fairly simple but instructive exercise that is left to the reader. In example 7.27 parameter restrictions were sufficient to give conditionals that characterize a joint, but even this is not always the case. Arnold, Castillo, and Sarabia (1992) give a number of examples of impossible models in which conditionals cannot be made to satisfy the conditions of the Result.

Metropolis Within Gibbs

A hybrid algorithm known as Metropolis within Gibbs arises when we have a problem that seems well suited for use of a Gibbs sampler, but there are one or more of the conditional posteriors that we cannot derive in closed form, or sample from directly. In these cases, one common approach is to use a Metropolis-Hastings algorithm to sample from the unnormalized conditional, but embed this is an overall Gibbs algorithm to simulate from the entire joint posterior.

Example 7.28

Consider a one sample gamma model, $Y_1, \ldots, Y_n \sim$ iid Gamma(α, β). We will form a joint prior as a product form. Take the prior distribution of β to be a Gamma(α_0, β_0) distribution. Let the prior for α be denoted as $\pi_2(\alpha)$ without specifying a particular form as of yet. Then the joint posterior of α and β may be written as

$$p(\alpha, \beta|\boldsymbol{y}) \propto \pi(\alpha) \frac{\beta^{n\alpha}}{\{\Gamma(\alpha)\}^n} \left(\prod_{i=1}^{n} y_i^{\alpha-1}\right) \exp\left(-\beta \sum_{i=1}^{n} y_i\right) \beta^{\alpha_0-1} \exp(-\beta_0 \beta). \tag{7.27}$$

For a fixed value of α, (7.27) gives the conditional posterior $p(\beta|\boldsymbol{y}, \alpha)$ as a gamma distribution with parameters $\alpha_0 + n\alpha$ and $\beta_0 + \sum y_i$, and this will be easy to sample from using a prepackaged computational function. There is no form for $\pi_2(\alpha)$, however, that will give a recognizable form for the conditional posterior $p(\alpha|\boldsymbol{y}, \beta)$, so that all we can say about this conditional distribution is that

$$p(\alpha|\boldsymbol{y}, \beta) \propto \pi(\alpha) \frac{\beta^{n\alpha}}{\{\Gamma(\alpha)\}^n} \left(\prod_{i=1}^{n} y_i^{\alpha-1}\right).$$

We could, however, sample from this conditional posterior using one cycle of a Metropolis-Hastings algorithm. Note that one cycle here means one tran-

sition, regardless of whether that involves accepting or rejecting a proposed jump. The theoretical behavior of Metropolis within Gibbs is not well understood. Although the theory of MCMC is an active field of research and new results are being rapidly developed, at the time these notes were written no result sufficient to guarantee convergence of a hybird chain had been developed. Nevertheless, Metropolis within Gibbs seems to work in many cases and is certainly popular.

7.9.4 Monitoring Convergence

In an application we need to determine when a Markov chain has made a sufficient number of transitions for us to behave as if the values produced are a sample from the target distribution, at which point we say the chain has *converged*. Quite a few diagnostics have been developed for monitoring the output of Markov chain samplers for convergence. Cowles and Carlin (1996) compared thirteen convergence diagnostics and concluded that no single diagnostic can suffice in all problems and, as a result, "automated convergence monitoring (as by a machine) is unsafe and should be avoided" (Cowles and Carlin, 1996, p. 903). Thus, determining convergence in an application must remain a matter of judgment on the part of the investigator, and how the diagnostics currently available to guide this judgment can be improved is still a matter of active research.

In this section we present a practical diagnostic due to Gelman and Rubin (1992) that contains both a graphical component and a summary statistic as a global measure of convergence. We also indicate that the very simple technique of examining autocorrelations among values simulated at various lags can provide valuable information about the behavior of a Markov chain.

As a preliminary comment, note that convergence of a Markov chain can have more than one meaning. In the context of this chapter we are concerned with convergence of sampled values from a chain to the underlying invariant (and limiting) distribution, given irreducibility, positive recurrence and aperiodicity. This is sometimes called determination of the previously mentioned burn-in period. Once an appropriate burn-in has been determined (in terms of number of iterations of the chain), values prior to that point are discarded and values after that point are collected and considered to be samples from the target distribution. A topic we will not be able to cover in these notes is convergence of Monte Carlo approximations of quantities computed on the basis of values from a chain. The reason there might be a difference is that values from a Markov chain are not independent. Simulated values from a given distribution might have converged in the sense that they are drawn from the appropriate distribution, but different portions of the distribution might not be visited independently, even if they are ultimately visited with the correct frequencies, causing quantities such as the sample mean to be slower in convergence. This is a rather subtle point, but is of considerable importance in the Monte Carlo portion of Markov Chain Monte Carlo methods.

The Scale Reduction Factor

Gelman and Rubin (1992) introduced what they called the (estimated) scale reduction factor as a practical way to assess convergence of a Markov chain sampler. The basic idea is that, given positive recurrence, a Markov chain will converge to its invariant distribution regardless of starting value. Thus, to determine how many iterations are needed before a chain can be reliably considered to be generating values from its invariant distribution, one can

examine the behavior of multiple chains, each of which is started at a different point in the state space. When the variability of values within each chain is the same as the variability of values among different chains, we can consider the chains to have converged and to be sampling from the same distribution. This idea is formalized as follows.

Suppose we run m chains that have starting values "widely dispersed" in the appropriate state space. Let the values produced by these chains be denoted as $\boldsymbol{x}_j(t)$; $t = 1, \ldots, ;\ j = 1, \ldots, m$. Consider any one of the scalar components of $\boldsymbol{x}_j(t)$, $x_{j,k}(t)$, say for the kth component of $\boldsymbol{x}_j(t)$. The sample variance of a portion of a given chain j, $x_{j,k}(t)$; $t = 1 \ldots, n$, is

$$s^2_{j,k}(n) = \frac{1}{n-1} \sum_{t=1}^{n} \{x_{j,k}(t) - \bar{x}_{j,k}(n)\}^2,$$

where $\bar{x}_{j,k}(n) = (1/n) \sum_t x_{j,k}(t)$. The average of the sample variances across the m chains is called the within-sequence variance,

$$W_k(n) = \frac{1}{m} \sum_{j=1}^{m} s^2_{j,k}(n).$$

The between-sequence variance is also computed as,

$$B_k(n) = \frac{n}{m-1} \sum_{j=1}^{m} \{\bar{x}_{j,k}(n) - \bar{x}_k(n)\}^2,$$

where $\bar{x}_k(n) = (1/m) \sum_j \bar{x}_{j,k}(n)$.

Assuming that correlation between successive values in the chain is positive, the within-sequence variance W underestimates the variance of $x_k(t)$ in the target distribution, which is also the invariant and limit distribution of the chain, but does converge to $var\{x_k(t)\}$ as $n \to \infty$, where n is the length of the sequences used to compute W. Another estimate of $var\{x_k(t)\}$ in the target distribution can be constructed as

$$\hat{var}\{x_k(t)\} = \frac{n-1}{n} W_k(n) + \frac{1}{n} B_k(n).$$

Gelman and Rubin (1992) indicate that $\hat{var}\{x_k(t)\}$ should be an overestimate of $var\{x_k(t)\}$ because of the dispersion of starting values, but should also approach that quantitiy as $n \to \infty$. Thus, we have two estimates that approach the variance of $x_k(t)$ in the target distribution as length of the chains n increases, but do so from opposite directions. What is called the *estimated scale reduction factor* is then defined as,

$$R_k(n) = \left[\frac{\hat{var}\{x_k(t)\}}{W_k(n)}\right]^{1/2}, \tag{7.28}$$

and $R_n \to 1$ as $n \to \infty$.

As practical advice, Gelman and Rubin (1992) recommend running the chains until $R_k(n)$ is less than 1.2 or 1.1 for all components of $\boldsymbol{x}(t)$. Assuming this is the case at a given number of iterations n^* say, it has become common practice to run the chains for twice that long, $2n^*$ iterations in total, discard all values up to iteration n^* as burn-in, and use the values from iterations $n^* + 1$ to $2n^*$ for inference. If all of the m individual chains have "reached" the target distribution, then values from those chains may be combined (i.e., lumped together into one sample) providing a total of mn^* values.

As a closing remark to this subsection, it should be emphasized that the efficacy of the scale reduction factor in assessing convergence relies heavily on the concept of starting values for the individual chains that are widely dispersed in the state space. If this is not the case, then $\hat{var}\{x_k(t)\}$ will not sufficiently overestimate $var\{x_k(t)\}$ and $R(n)$ will approach 1 too rapidly as n increases. This becomes something of a sticking point for the method, because it is not clear how to pick widely dispersed starting values. For chains in which the components of $\boldsymbol{x}(t)$ are correlated with each other, some starting values within the total state space will cause numerical errors in sampling, such as numerical evaluations of densities that are below or above the internal machine

capabilities of most computers (e.g., acceptance probabilities in a Metropolis Hastings algorithm cannot be computed). When this occurs, sampling algorithms crash in much the same way that iterative algorithms for evaluating maximum likelihood estimates may crash for poor starting values. Thus, to get multiple chains to run, one may end up picking starting values that are not truly widely dispersed and obtain a misleading indication of the number of iterations needed for the chains to have converged.

Examination of Autocorrelations

The rate at which Markov chains convergence to their limit (and invariant) distributions is a function of autocorrelation among successive states. This, in turn, also determines the effect of starting values. As a result, the examination of (estimated) autocorrelations in a Markov chain can provide insight into how long a chain should be run in practice before one begins to accept values from the chain as representing values from the target distribution. Autocorrelation is essentially an (inverse) measure of the rate at which a chain visits all portions of the state space, which is guaranteed for an irreducible chain. Thus, a chain for which autocorrelations remain meaningful for a greater number of lagged values should be run longer than chains for which autocorrelations are meaningful for only a smaller number of lagged values. The basic tool in this examination is a plot of estimated autocorrelations versus number of lagged values used. Most software packages easily produce such plots.

One view is that when autocorrelation has declined to a negligible level the indication is that the effect of starting value has diminished to the point of being ignorable. Burn-in can then be taken as the number of iterations needed for this to occur or some small multiple of that number, to be conservative.

This outlook is paired with the viewpoint that it may be preferable to run one chain for a large number of iterations rather than combining values from multiple chains run for a shorter number of iterations. It seems generally accepted that the examination of autocorrelations can provide a valuable tool in diagnosing the behavior of a Markov chain sampler even if one is relying on the scale reduction factor as an indicator of convergence. For example, if the variance reduction factor appears to have initially declined to about 1 after a small number of iterations (e.g., 25) but autocorrelation is clearly not negligible until lagged vlaues of 50, then one might suspect that the results of the variance reduction factor have been caused by chance in that the particular sample paths taken by the chains examined just happen to give the results seen. Or, as noted at the end of the previous subsection, the starting values for individual chains may be too similar for the variance reduction factor to provide a meaningful diagnostic of convergence. In such instances one should re-run multiple chains with more widely dispersed starting values or, if that is not possible due to computational problems, run the individual chains for a longer period to see if the variance reduction factor remains near 1.

As a final note on the usefulness of autocorrelation in examining the behavior of Markov chain samplers, autocorrelations that do not decay rapidly often indicate that the variance reduction factor may decay to 1 more rapidly for some quantities that for others. In this chapter we have introduced the use of variance reduction factors only in examination of the individual elements of the variable under examination, which are simulated directly by the sampler being used. This may, indeed, occur fairly rapidly, with the variance reduction factor remaining near 1 for all subsequent iterations. But, as already mentioned, we may also wish to estimate scalar quantities constructed from simulated values (e.g., Monte Carlo approximations to means or certain quantiles of the target

distribution). If autocorrelation lasts for a substantial number of iterations relative to the point that the variance reduction factor (for individual components of the variable under examination) decreases to near 1, the indication is that it may take many more iterations for the variance reduction factor to decrease to 1 if it is computed for the Monte Carlo approximation constructed from simulated values.

7.9.5 A Touch of Markov Chain Theory

This section is primarily for those who wish a bit deeper understanding of how Markov chain samplers operate. It is not absolutely necessary to have mastered this material to be able to conduct analyses using MCMC methods, but it is also a good idea to know "what can go wrong" when conducting an application, so some attention to this topic is a good idea even if you do not wish to delve into convergence of Markov chains from a theoretical viewpoint.

To understand how, why, and when MCMC techniques are effective requires some background knowledge of Markov chains. It is not my intent in this section is to summarize available results, as an effort to do so would almost surely prove to be inadequate. Rather, my intention is to provide sufficient understanding of enough of the basic concepts and issues involved that the reader may approach the rather daunting volume of literature on the subject with some modest chance of success.

A stochastic process $\{\boldsymbol{X}(t) : t \geq 0\}$ with state space $\boldsymbol{x}(t) \in E \subset \mathbb{R}^p$ is called a Markov process if, for any countable set of indices $\{t_0, t_1, \ldots\}$, $\boldsymbol{X}(t)$ has the property that

$$Pr\{\boldsymbol{X}(t_{n+1} \leq \boldsymbol{x} | \boldsymbol{X}(t_n) = \boldsymbol{x}_n, \ldots, \boldsymbol{X}(t_0) = \boldsymbol{x}_0\} = Pr\{\boldsymbol{X}(t_{n+1}) \leq \boldsymbol{x} | \boldsymbol{X}(t_n) = \boldsymbol{x}_n\} \tag{7.29}$$

A Markov process is called a Markov chain if the index set is discrete (i.e., t can assume only values in the set $t \in \{t_0, t_1, \ldots\}$). Markov chains may have either discrete state space or a general state space. We will consider only chains with discrete state space, $E \equiv \{e_1, e_2, \ldots, \}$ since dealing mathematically with general state spaces requires some measure theoretic concepts. For a Markov chain $\{\boldsymbol{X}(t) : t = 0,\, 1, \ldots\}$ with discrete state space E, *transition probabilities* are defined as, for all $\boldsymbol{e}_i, \boldsymbol{e}_j \in E$,

$$p_{i,j} \equiv Pr\{\boldsymbol{X}(t+1) = \boldsymbol{e}_j | \boldsymbol{X}(t) = \boldsymbol{e}_i\}. \tag{7.30}$$

Notice here that we are assuming that these transition probabilities do not depend on t, in which case the chain is called *time homogeneous* and, given that these are probabilities, we must have that $\sum_j p_{i,j} = 1.0$. Now define *k-step transition probabilities* as,

$$p_{i,j}^{(k)} \equiv Pr\{\boldsymbol{X}(t+k) = \boldsymbol{e}_j | \boldsymbol{X}(t) = \boldsymbol{e}_i\} = Pr\{\boldsymbol{X}(k) = \boldsymbol{e}_j | \boldsymbol{x}(0) = \boldsymbol{e}_i\}. \tag{7.31}$$

The last equality in expression (7.31) follows from time homogeneity. Note also that $p_{i,j}^{(1)} = p_{i,j}$. A basic result in Markov chain theory, known as the *Chapman-Kolmogorov* equations, is that, for $p_{i,j}^{t+k} = Pr\{\boldsymbol{X}(t+k) = \boldsymbol{e}_j | \boldsymbol{X}(0) = \boldsymbol{e}_i\}$,

$$p_{i,j}^{t+k} = \sum_{m=0}^{\infty} p_{i,m}^{(t)}\, p_{m,j}^{(k)}, \tag{7.32}$$

for all $t, k > 0$ and all i, j.

A state $\boldsymbol{e}_j$ is said to be *accessible* from state $\boldsymbol{e}_i$ if $p_{i,j}^{(k)} > 0$ for some $k \geq 0$; some authors say states $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ *communicate*. Notice that this implies all states are accessible from themselves since $p_{i,i}^{(0)} = 1.0$ for all i. A Markov chain is said to be *irreducible* if all states are accessible from all other states. Thus, for example, counting processes in which the states of a univariate $X(t)$ represent the number of events that have occurred by time t are not irreducible,

since we can never get to state $e_j = 1$ from state $e_i = 2$. If a Markov chain is irreducible, then there exists a value $d > 0$ such that, for all states $\boldsymbol{e}_i \in E$, $p_{i,i}^{(d)} > 0$, that is, starting in state $\boldsymbol{e}_i$ the probability of returning to state $\boldsymbol{e}_i$ in d transitions is greater than zero. But note that we have not yet said anything about what the value of d might be, or even if it is finite. A great deal of what can be learned about the behavior of a chain depends on the behavior of d, but for the moment all we have is its existence for an irreducible chain. The state e_i of a Markov chain is called *periodic* with a period of ν if the only d for which $p_{i,i}^{(d)} > 0$ are $d = a\nu$; $a = 2, \ldots$. It can be shown that, if any one state of an irreducible Markov chain is periodic, then all states are periodic or, conversely, if any one state is not periodic then no states are periodic. A Markov chain that has no (implied by even one) periodic states is called *aperiodic*. Note that if a chain has one state that can be reached from itself in one transition then the chain is aperiodic. For a Markov chain $\{\boldsymbol{X}(t) : t = 0, 1, \ldots\}$, define the *time of first return* to any state $\boldsymbol{e}_i$ as,

$$\tau_{i,i} \equiv \min\{t > 0 : \boldsymbol{X}(t) = \boldsymbol{e}_i | \boldsymbol{X}(0 = \boldsymbol{e}_i\}.$$

Note that, for any state $\boldsymbol{e}_i$, the time of first return $\tau_{i,i}$ is a random variable. A state $\boldsymbol{e}_i$ for which $Pr(\tau_{i,i} < \infty) = 1.0$ is called *persistent* or *recurrent*. If a state $\boldsymbol{e}_i$ is not recurrent then it is said to be *transient*. It can be shown that, if one state of an irreducible Markov chain is recurrent, then all states are recurrent and we say that the chain is recurrent. A state $\boldsymbol{e}_i$ of a Markov chain is called *positive recurrent* if

$$E(\tau_{i,i}) < \infty,$$

and a state that is recurrent but not positive recurrent is called *null reccurrent*. It can be shown that if any state of an irreducible Markov chain is positive recurrent, then all states are positive recurrent. An equivalent condition for

positive recurrence is that there exists a probability distribution $\pi(\cdot)$ such that, for all $\boldsymbol{e}_j \in E$,

$$\sum_{i \in E} \pi(\boldsymbol{e}_i)\, p_{i,j} = \pi(\boldsymbol{e}_j), \tag{7.33}$$

and where $\sum_j \pi(\boldsymbol{e}_j) = 1.0$. A distribution $\pi(\cdot)$ such that (7.33) holds is called the *invariant* or *stationary* distribution of the Markov chain $\{\boldsymbol{X}(t) : t = 0, 1, \ldots\}$.

Notice at this point that the property of irreducibility implies that all states of a Markov chain have the same probabilistic behavior, in that all are either periodic or aperiodic, all are either recurrent or transient, and all are either positive recurrent or null recurrent assuming they are recurrent in the first place. The following result indicates how we can locate the stationary distribution of a Markov chain that possesses the properties of irreducibility, aperiodicicy, and positive recurrence.

Result (c.f., Feller, Vol. 1, 3^{rd} ed., p. 391)

For an irreducilbe, aperiodic, positive recurrent Markov chain with $k-$step transition probabilities $p_{i,j}^{(k)}$, for all $\boldsymbol{e}_j \in E$,

$$\pi(\boldsymbol{e}_j) = \lim_{k \to \infty} p_{i,j}^{(k)}, \tag{7.34}$$

and this $\pi(\boldsymbol{e}_j)$ is the stationary distribution of the chain given in expression (7.33).

This is a crucial result for simulation using Markov chain samplers becasue it states that the limiting distribution given by (7.34) exists and it is the same as the stationary distribution of (7.33). One additional concept will prove useful in assessing whether a chain is irreducible, positive recurrent and aperiodic, that being chains that have the property of being *reversible*.

Consider a time homogeneous Markov chain with finite state space E and transition probabilities $p_{i,j}$ that is irreducible and positive recurrent, denoted as,

$$\boldsymbol{X} = \{\boldsymbol{X}(t) : t = \ldots, -2, -1, 0, 1, 2, \ldots\}.$$

Suppose that $\boldsymbol{X}$ is at its stationary distribution, which is also its limiting distribution if the chain is aperiodic. Then,

$$Pr\{\boldsymbol{X}(t) = \boldsymbol{e}_j\} = \pi(\boldsymbol{e}_j); \quad \text{for all } \boldsymbol{e}_j \in E.$$

Define another Markov chain as $\boldsymbol{Y}(t) = \boldsymbol{X}(-t)$, so that $\boldsymbol{Y}(2) = \boldsymbol{X}(-2)$, $\boldsymbol{Y}(1) = \boldsymbol{X}(-1)$, $\boldsymbol{Y}(0) = \boldsymbol{X}(0)$, $\boldsymbol{Y}(-1) = \boldsymbol{X}(1)$, $\boldsymbol{Y}(-2) = \boldsymbol{X}(2)$, and so forth. The chain $\boldsymbol{Y}(t)$ is called the *time reversed* version of $\boldsymbol{X}(t)$. Let the transition probabilities for the chain $\boldsymbol{Y}(t)$ be denoted as $q_{i,j} = Pr\{\boldsymbol{Y}(t+1) = \boldsymbol{e}_j | \boldsymbol{Y}(t) = \boldsymbol{e}_i\}$. We wish to express the $q_{i,j}$ in terms of the $p_{i,j}$, the transition probabilities of the original chain $\boldsymbol{X}(t)$. Simply applying the definition of conditional probability,

$$\begin{aligned}
q_{i,j} &= Pr\{\boldsymbol{Y}(t) = \boldsymbol{e}_j | \boldsymbol{Y}(t-1) = \boldsymbol{e}_i\} \\
&= Pr\{\boldsymbol{X}(-t) = \boldsymbol{e}_j | \boldsymbol{X}(-t+1) = \boldsymbol{e}_i\} \\
&= \frac{Pr\{\boldsymbol{X}(-t) = \boldsymbol{e}_j \bigcap \boldsymbol{X}(-t+1) = \boldsymbol{e}_i\}}{Pr\{\boldsymbol{X}(-t+1) = \boldsymbol{e}_i\}} \\
&= \frac{Pr\{\boldsymbol{X}(-t+1) = \boldsymbol{e}_i | \boldsymbol{X}(-t) = \boldsymbol{e}_j\} \, Pr\{\boldsymbol{X}(-t) = \boldsymbol{e}_j\}}{Pr\{\boldsymbol{X}(-t+1) = \boldsymbol{e}_i\}} \\
&= \frac{p_{j,i} \, \pi(e_j)}{\pi(e_i)}.
\end{aligned}$$

A Markov chain for which $q_{i,j} = q_{j,i}$ for all $\boldsymbol{e}_i, \boldsymbol{e}_j \in E$ is said to be *time reversible* or often simply reversible. Notice that $q_{i,j} = q_{j,i}$ implies that

$p_{i,j}\pi(\boldsymbol{e}_i) = p_{j,i}\pi(\boldsymbol{e}_j)$ so that, if this holds for all $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, then,

$$\begin{aligned} & \sum_i p_{j,i}\pi(\boldsymbol{e}_j) = \sum_i p_{i,j}\pi(\boldsymbol{e}_i) \\ \iff \quad & \pi(\boldsymbol{e}_j)\sum_i p_{j,i} = \sum_i p_{i,j}\pi(\boldsymbol{e}_i) \\ \iff \quad & \pi(\boldsymbol{e}_j) = \sum_i p_{i,j}\pi(\boldsymbol{e}_i), \end{aligned} \tag{7.35}$$

which is the invariant distribution of expression (7.33) which, in turn, is equivalent to the condition of positive recurrence for an irreducible chain. What this gives us then is another way to verify positive recurrence. Suppose we have a set of target probabilities $\pi \equiv \{\pi(\boldsymbol{e}_i) : \boldsymbol{e}_i \in E\}$ and we specify a set of transition probabilities $\{p_{i,j} : \boldsymbol{e}_i, \boldsymbol{e}_j \in E\}$ such that the chain is reversible. Then, provided the chain is irreducible, (7.35) holds and π is the invariant distribution for the chain. If, in addition, the chain is aperiodic, then π is also the limiting distribution. To foreshadow, this is essentially what happens if we form a Metropolis-Hastings algorithm, which guarantees reversibility by its construction. To summarize the important results of this subsection for discrete state space Markov chains we have the following.

1. Irreducibility is a condition of "probabilistic connectedness", a phrase due to Roberts (1996), in that it implies all states of a chain share properties such as being recurrent, positive recurrent, and aperiodic.

2. Irreducibility plus positive recurrence implies the existence of an invariant (or stationary) distribution. Similarly, irreducibility plus the existence of an invariant distribution implies positive recurrence.

3. Irreducibility plus positive recurrence plus aperiodic implies the existence of a limiting distribution that is the same as the invariant distribution.

4. Reversibility implies the existence of an invariant distribution so that irreducibility plus reversibility implies positive recurrence.

Our context for the use of Markov chain samplers is simulation from the posterior distribution of a data model parameter. Most data models have parameters that can assume any value in some interval of the line, and thus posterior distributions for such parameters are usually in the form of probability density functions. The presentation of this subsection, in contrast, has been in terms of discrete state space chains. All of the concepts contained here apply to general state space chains, but require a shift from discrete to continuous mathematical settings.

Part III

Models With a Single Random Component

Chapter 8

Generalized Linear Models

What are known as *generalized linear models* (glms) are a class of (often, but not necessarily) nonlinear regresssion models that are typically formulated by considering an appropriate random model component first, and then pairing that with a specified systematic model component. As discussed in Chapter 2.4, 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. Historically, glms became popular because of a unified computational algorithm for estimation that covered the basic glm models. This is no longer a major motivational factor for the use of glms, but the basic ideas underlying the develop of these models are important to the development of regression models for more general situations, and glms remain a useful class of models in their own right for many problems.

The classic analysis of glms is typically presented from a frequentist viewpoint with maximum likelihood estimation of regression coefficients combined with a moment-based estimator of what will be called the dispersion parameter. This parallels what would be considered for an analysis of a normal linear

regression model where the dispersion parameter corresponds to the error variance. We will present estimation and inference for generalized linear models using both this traditional approach and Bayesian methods.

8.1 Structure of Generalized Linear Models

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 (3.7); note that this immediately implies the properties given in expression (3.8). We will not consider these random variables as *iid*, although we will allow them to differ only through their natural parameters (θ_i), and 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,

$$f(y_i|\theta_i) = \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i,\, \phi)\right], \tag{8.1}$$

and the properties of expression (3.8) as,

$$\begin{aligned} \mu_i \equiv E(Y_i) &= \frac{d}{d\theta_i} = b'(\theta_i), \\ var(Y_i) &= \frac{1}{\phi}\frac{d^2}{d\theta_i^2}b(\theta_i) = \frac{1}{\phi}\, b''(\theta_i) = \frac{1}{\phi}\, V(\mu_i). \end{aligned}$$

It is sometimes the case that the dispersion parameter ϕ is replaced by a function $a_i(\phi) = \phi m_i$ for some constant m_i. This is useful for binomial models and models formulated for weighted variables.

The random model component is given by expression (8.1). The systematic model component in glms 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

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{8.2}$$

where $\boldsymbol{x}_i^T = (x_{1,i}, x_{2,i}, \ldots, x_{p,i})$ is 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 $x_{1,i} = 1$, but this is neither necessary, nor does the term intercept parameter always have the same interpretation as for linear models.

The other portion of the systematic model component is the link function, which is defined as the relation,

$$g(\mu_i) = \eta_i. \tag{8.3}$$

The covariates may be quantities measured on a ratio/interval scale, or may be 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 an interval/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 (8.3) as $g(b'(\theta_i)) = \eta_i$. The link function $g(\cdot)$ is generally taken as a 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 that 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:

- Normal random component: $g(\mu_i) = \mu_i$
- Poisson random component: $g(\mu_i) = \log(\mu_i)$
- Binomial random component: $g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$
- Gamma random component: $g(\mu_i) = 1/\mu_i$
- 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 rather than observed counts. 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 $\eta_i =$

$\boldsymbol{x}_i\,\beta$. If this is not true we must 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 $\boldsymbol{\beta}$ may allow the use of an identity link with a Poisson random component. Other common link functions, without attaching them to any particular random components, include:

- Log link: $g(\mu_i) = \log(\mu_i)$ for $0 < \mu_i$.
- 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 $-\infty < \mu_i < \infty$.
- Complimentary Log-Log Link:
  $g(\mu_i) = \log\{-\log(1 - \mu_i)\}$ 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, but this is a more advanced topic that we will not cover here.

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. 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:

- Normal random component: $V(\mu_i) \equiv 1$.

- Poisson random component: $V(\mu_i) = \mu_i$.
- Binary random component: $V(\mu_i) = \mu_i(1 - \mu_i)$.
- Binomial* random component: $V(\mu_i) = \mu_i(1 - \mu_i)$.
- Gamma random component: $V(\mu_i) = \mu_i^2$.
- Inverse Gaussian random component: $V(\mu_i) = \mu_i^3$.

$*$ binomial random component written for observed proportions not observed counts

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 and plots of variances against means can often provide information about potentially useful random component specifications, which will be illustrated in the next section.

8.2 Choosing Random and Systematic Model Components

In this section we consider the practical issues of selecting random and systematic model components for a problem in which we are considering the use of a basic generalized linear model. As an introductory comment, we first point out that, in many problems, if one is not interested in the distribution of responses there is probably little motivation for considering generalized linear models. The exception to this might be if responses have extremely low variability and

there is a solid reason based on knowledge of the scientific problem for choosing a particular distributional form for the random model component. But, as we will see in the sequel, for problems in which responses exhibit considerable variability there is often little difference in estimated systematic model components for different choices of random components. If one is only interested in the systematic component, the choice of random component then seems relatively unimportant. A counter-argument is that choosing a more appropriate random component can lead to smaller standard errors for estimated regression coefficients (parameters in the systematic model component) than a less appropriate random component. This is true, but there are options other than generalized linear models for attaining precision in estimation of regression functions. Thus, there is nothing that suggests the use of a generalized linear model in particular if one has interest only in the systematic model component. In contrast, if one has interest in, for example, the 75^{th} percentile of the response distribution at one or more values of a covariate, then the choice of random model component has a major impact on the outcome of the analysis. Similarly, if one is interested in any aspect of the distribution of response variables other than expectations, then serious consideration of random model components is appropriate.

8.2.1 Random Model Components

The first consideration in choice of a random model component is the set of possible values for response variables. Situations in which a binomial distribution is appropriate are often fairly easy to detect, especially if those binomial distributions can be constructed from groups of independent and identically distributed binary trials. In other cases we might choose a binomial random

component not through a construction but simply as a reasonable representation, just like assignment of any other distributional form. Likewise, some problems indicate a natural choice of a Poisson random component because the responses are obtained in the form of counts and contain small values. In yet other problems the quantities represented by response random variables may be continuous but for which negative values are physically impossible, such as chemical concentrations. These situations might suggest one of gamma, inverse Gaussian, or lognormal random components.

A simple tool that can sometimes assist in choice of a random model component is a basic scatterplot, particularly if one has only a single covariate. Situations in which scatterplots may provide information about the distributional form of responses include problems involving a covariate that indicates group membership (i.e., an one-way ANOVA setting), or in which observations are fairly dense (numerous) along the covariate gradient. If one imagines a visual curve through the plot, it is sometime possible to determine the relative skewness of an appropriate random component.

Example 8.1

As part of a study on factors that influence the release of CO_2 from soils, soil temperature and soil respiration were measured at sites in several ecosystems, including temperate and tropical forests, and grasslands (Raich, Kaiser, Dornbusch, Martin, and Valverde-Barrantes, 2023). A scatterplot of soil temperature (°C) against soil respiration ($g\,C\,m^{-2}\,d^{-1}$) is shown in Figure 8.1. One might envision a straight line or perhaps a slightly increasing curve through these values. With either of those expectation functions, the distribution of data at given covariate values appears to be skew right with tails that fan out from the main body more in the upward direction than the downward

direction.

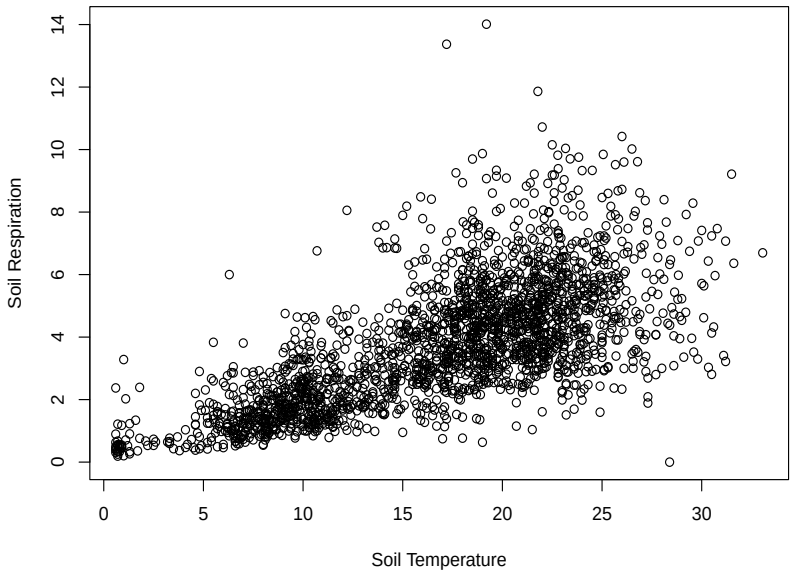


Figure 8.1: Scatterplot of soil respiration against soil temperature.

In addition to examination of scatterplots, the relation between means and variances exhibited in the data is often a useful tool in choosing a random model component. Recall from (3.8) that exponential dispersion family distributions dictate certain relations between expected values and variances through the relation $var(Y_i) = (1/\phi)V(\mu_i)$, where $\mu_i = E(Y_i)$. As a result, information the data provide about the variance function $V(\mu_i)$ is information about what an appropriate choice of random component might be. Also note from the list of variance functions give previously that $V(\mu_i)$ is a power of μ_i for continuous random components in basic generalized linear models. Now, if

$var(Y_i) = (1/\phi)\mu_i^\theta$ for some power θ, then

$$\log\{var(Y_i)^{1/2}\} = \frac{1}{2}[\log(1/\phi) + \theta\log(\mu_i)], \tag{8.4}$$

which suggests that a plot of the logarithm of standard deviations against the logarithm of expected values could help determine the appropriate value of θ and, hence, the appropriate random component.

If replicate response variables are not available at distinct covariate values, which typically occurs if covariates are measured on a ratio/interval scale, then some type of binning procedure can be employed to create groups across values of the covariate. Sample means and variances are computed for each group, and the logarithm of group standard deviation plotted against the logarithm of group mean. If this plot looks roughly like a straight line, then 2 times the slope of that line gives a rough idea of the value of θ in $var(Y_i) = (1/\phi)\mu_i^\theta$. You may have run into this same plot in previous courses, because it can also be used to help determine a reasonable power for a power transformation and it is then often called a Box-Cox plot. This is the same plot, but we are using it for a decidedly different purpose.

Example 8.1 (continued)

For the data of Figure 8.1, a Box-Cox plot is shown in Figure 8.2. These data were binned by dividing the range of the covariate into 30 bins of equal length. Only bins having at least 15 observations were retained, which resulted in 27 bins supplying values of sample means and sample variances. An ordinary least squares line fit to the values of Figure 8.2 gives a slope of 0.65, which suggests that variances are increasing roughly as a power 1.3 of the expected values.

In any application the choice of binning rule is arbitrary, and several plots should be produced to ensure that the choice of binning rule is not having too

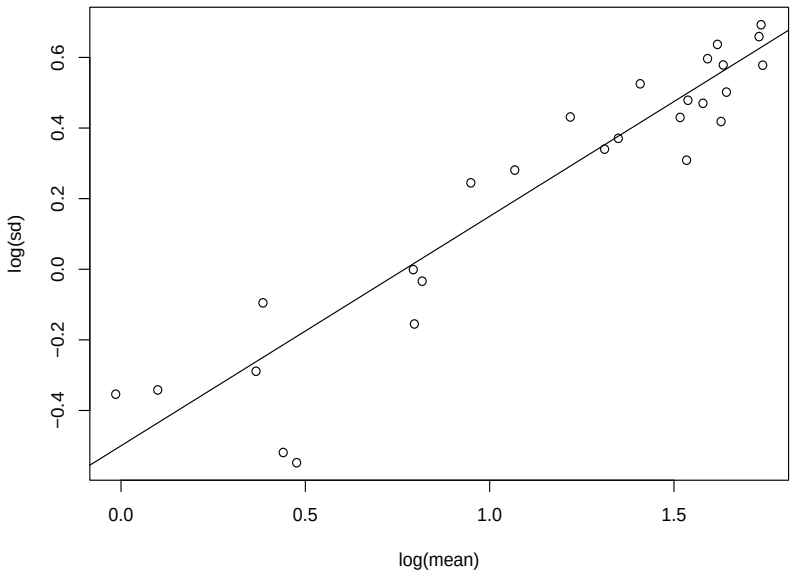


Figure 8.2: Box-Cox plot for the data of Figure 8.1.

great an influence on the diagnostic. Another Box-Cox plot was produced for these data by dividing the covariate into 30 bins such that each bin contained the same number of values, which was 73; not all bins are of equal length. The plot looked quite similar to Figure 8.2 and an ordinary least squares line had slope 0.66. The exploratory analyses for this example motivate the choice of a random component that is continuous and skew to the right, with variances that increase proportional to expected values raised to a power of about 1.3. Unfortunately, none of the standard random components for basic generalized linear models satisfy both of these characteristics. Understanding that the Box-Cox procedure serves only as a rough guide, we might contemplate a gamma random component, which has variances increasing proportional to the square of expected values. Alternatively, we might also consider abandoning a

generalized linear model framework to formulate an additive error model with power of the mean structure, and a power of about 0.65 coupled with a right skew location-scale response distribution such as an extreme value.

8.2.2 Systematic Model Component

For models with a single type of covariate, selection of a suitable systematic model component is often simply a matter of visual inspection of the scatterplot. Sometimes it can help to also examine plots of transformed responses against the covariate. If $T(y_i)$ is a transformation of the response data, and plotting $T(y_i)$ against the covariate x_i results in a straight line, then a suitable link function should be $g(\mu_i) = T(\mu_i)$.

Example 8.1 (continued)

The scatterplot of Figure 8.1 looks like a straight line might prove reasonable to describe the expected values, but there is also a hint of a gentle increasing curve. To examine this possibility, Figure 8.3 shows scatterplots of the logarithm of soil respiration versus soil temperature and the square root of soil respiration versus soil temperature.

Note that we do not want to transform responses and then fit a linear regression because we are interested in the distribution of reponses at given covariate values. Transformation of responses was used only as a device to identify an appropriate systematic model component. In Figure 8.3 we can see that a log transformation has had an overly dramatic effect on the relation between soil respiration and temperature. The scatterplot for the square root transformaton, on the other hand seems to produce a plot that could well be described with a straight line. We thus have two potential systematic model components which, in a generalized linear model would be the identity link,

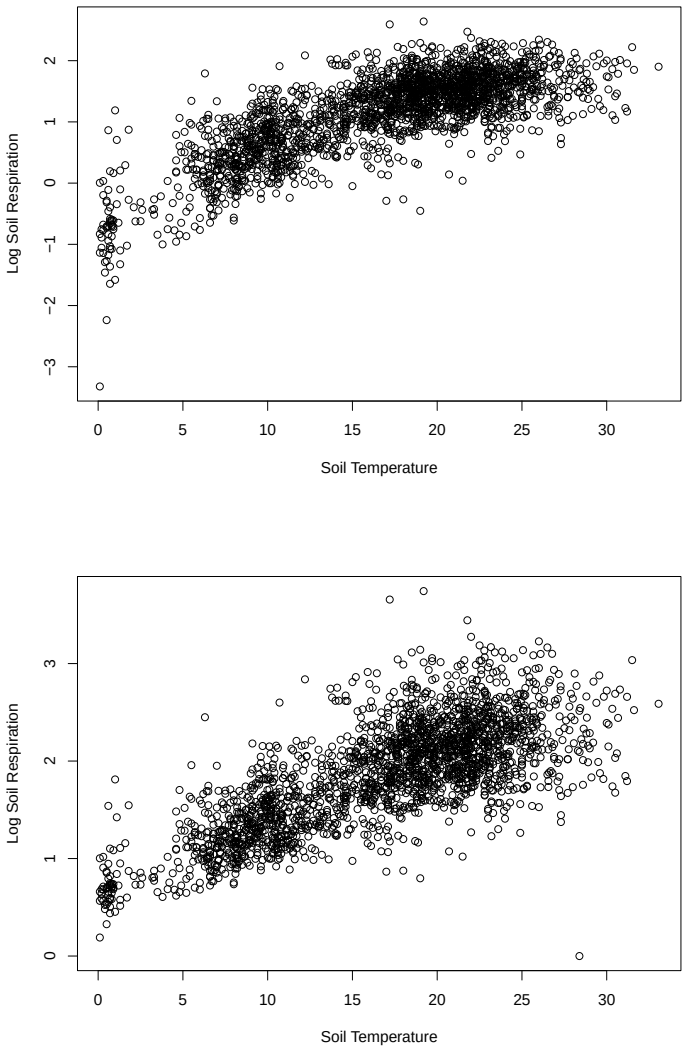


Figure 8.3: Scatterplots of log (upper) and square root (lower) transformation of soil respiration versus soil temperature.

$g(\mu_i) = \mu_i$, and the square root link, $g(\mu_i) = \sqrt{\mu_i}$. Either of these could be easily formulated as systematic components in an additive error model, if we would choose to go that route, as mentioned previously.

Choosing a suitable link function in situations that involve more than one type of covariate is considerably more difficult than in the case of a single type of covariate, because the relation specified in the model is between the response means and an unknown linear combination of the covariate types. Without knowing what linear combination to compute (i.e., without knowing the regression coefficients) it is not possible to construct diagnostic plots. In many cases, generalized linear models that involve a number of types of covariates seem to use canonical links as default choices. Whether this is a good idea or not is open to debate.

8.3 Likelihood Estimation and Inference

Non-Bayesian analysis of generalized linear models is accomplished through likelihood-based methods, at least as far as the regression parameters $\boldsymbol{\beta}$ are concerned. In models that contain an additional dispersion parameter ϕ, that parameter may also be dealt with based on likelihood methods, but is more commonly approached through moment-based methods. The reasons for this will be discussed after we have presented the usual treatment for regression parameters.

8.3.1 Maximum Likelihood Estimation of $\boldsymbol{\beta}$

. Consider estimation of the regression parameters $\boldsymbol{\beta}$. For the time being, consider ϕ (if there is one) to be a fixed constant; it will turn out that ϕ will not be involved in the maximum likelihood estimator of $\boldsymbol{\beta}$. Because we are dealing with exponential family response distributions that have common support not depending on the parameter, conditions are met that are needed to maximize

the likelihood by solving for roots of the score equations (see Chapter 5) but iterative numerical methods will be needed to locate the maximizing values. It turns out that a unified algorithm can be developed for any basic generalized linear model. That algorithm will be a Fisher scoring algorithm, as described in Chapter 5.7.4, and we will develop that algorithm carefully because it serves as a prototype for developing estimation procedures for models other than basic generalized linear models.

We assume that a basic generalized linear model has been formulated with a random component that forms an exponential dispersion family $f(y_i|\theta_i, \phi)$ as in (8.1) and a continuous link function $g(\mu_i)$. Given independence of the response variables $Y_1, \ldots, Y_n$, the log likelihood is,

$$\ell(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}, \phi), \tag{8.5}$$

where ℓ_i is the contribution of the i^{th} random variable,

$$\ell_i(\boldsymbol{\beta}, \phi) = \phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi). \tag{8.6}$$

Expression (8.6) 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 can be useful in deriving the derivatives of $\ell_i(\boldsymbol{\beta}, \phi)$ and thus also those of $\ell(\boldsymbol{\beta}, \phi)$ since the latter is just a sum over the former by (8.5). In particular, consider estimation of the components of $\boldsymbol{\beta}$ by deriving first the likelihood equations. We have that

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \tag{8.7}$$

Now, given the random component as an exponential dispersion family, and

the properties of such families, we have that,

$$\begin{aligned} \frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \theta_i} &= \phi\{y_i - b'(\theta_i)\} = \phi\{y_i - \mu_i\}, \\ \frac{d\theta_i}{d\mu_i} &= \frac{1}{V(\mu_i)}, \\ \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{i,j} \end{aligned} \tag{8.8}$$

The second line of expression (8.8) 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 (8.8) into (8.7) results in,

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \phi\{y_i - \mu_i\} \frac{1}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{i,j},$$

or, summing over observations,

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\phi\{y_i - \mu_i\} \frac{1}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{i,j} \right]. \tag{8.9}$$

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 (8.9) to arrive at,

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\phi\{y_i - \mu_i\} W_i \frac{d\eta_i}{d\mu_i} x_{i,j} \right]. \tag{8.10}$$

The set of likelihood equations are then given by setting (8.10) equal to zero for $j = 1, \ldots, p$.

To derive expressions for the second derivatives, make additional use of the

chain rule applied to (8.7), which results in,

$$\begin{aligned}
\frac{\partial^2 \ell_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k} &= \frac{\partial}{\partial\beta_k}\left[\frac{\partial\ell_i(\boldsymbol{\beta},\phi)}{\partial\theta_i}\,\frac{d\theta_i}{d\mu_i}\,\frac{d\mu_i}{d\eta_i}\,\frac{\partial\eta_i}{\partial\beta_j}\right] \\
&= \frac{\partial^2\ell_i(\boldsymbol{\beta},\phi)}{\partial\theta_i^2}\left(\frac{d\theta_i}{d\mu_i}\right)^2\left(\frac{d\mu_i}{d\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k} \\
&+ \frac{\partial\ell_i(\boldsymbol{\beta},\phi)}{\partial\theta_i}\,\frac{d^2\theta_i}{d\mu_i^2}\left(\frac{d\mu_i}{d\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k} \\
&+ \frac{\partial\ell_i(\boldsymbol{\beta},\phi)}{\partial\theta_i}\,\frac{d\theta_i}{d\mu_i}\,\frac{d^2\mu_i}{d\eta_i^2}\,\frac{\partial\eta_i}{\partial\beta_j}\,\frac{\partial\eta_i}{\partial\beta_k}. \qquad (8.11)
\end{aligned}$$

In (8.11) we would have

$$\begin{aligned}
\frac{\partial\ell_i(\boldsymbol{\beta},\phi)}{\partial\theta_i} &= \phi\{y_i - b'(\theta_i)\}, \\
\frac{\partial^2\ell_i(\boldsymbol{\beta},\phi)}{\partial\theta_i^2} &= \frac{\partial}{\partial\theta_i}\left[\phi\{y_i - b'(\theta_i)\}\right] \\
&= -\phi\, b''(\theta_i) = -\phi\, V(\mu_i).
\end{aligned} \qquad (8.12)$$

Substituting (8.12) into (8.11) we can see that the only terms in (8.11) that depend on the response value y_i are those that involve

$$\frac{\partial\ell_i(\boldsymbol{\beta},\phi)}{\partial\theta_i},$$

and, since $E(Y_i) = b'(\theta_i)$, the expected value of the random version of this first derivative is 0. This fact will render the expected second derivatives quite a bit easier to compute than the second derivatives themselves. That is, write

(8.11) as,

$$\begin{aligned}\frac{\partial^2 \ell_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k} &= -\phi V(\mu_i)\left(\frac{d\,\theta_i}{d\,\mu_i}\right)^2\left(\frac{d\,\mu_i}{d\,\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\frac{\partial\eta_i}{\partial\beta_k} \\ &+ \phi\{y_i - b'(\theta_i)\,\{\text{ terms without } y_i\}\end{aligned} \tag{8.13}$$

Taking the negative expectation of the random version of (8.13) results in,

$$-E\left\{\frac{\partial^2 \ell_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k}\right\} = \phi V(\mu_i)\left(\frac{d\,\theta_i}{d\,\mu_i}\right)^2\left(\frac{d\,\mu_i}{d\,\eta_i}\right)^2\frac{\partial\eta_i}{\partial\beta_j}\frac{\partial\eta_i}{\partial\beta_k} \tag{8.14}$$

Now, use the definition of W_i given just before expression (8.10) 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 (8.13) results in,

$$\begin{aligned}-E\left\{\frac{\partial^2 \ell_i(\boldsymbol{\beta},\phi)}{\partial\beta_j\partial\beta_k}\right\} &= \phi V(\mu_i)\frac{1}{\{V(\mu_i)\}^2}\left(\frac{d\,\mu_i}{d\,\eta_i}\right)^2 x_{i,j}x_{i,k} \\ &= \phi W_i\, x_{i,j}x_{i,k}.\end{aligned} \tag{8.15}$$

Summing (8.15) across observations (i) gives the total expected information. That is, let $I_n(\boldsymbol{\beta},\phi)$ be a $p \times p$ matrix with jk^{th} element

$$I_{j,k}(\boldsymbol{\beta},\phi) = \phi\sum_{i=1}^{n} W_i\, x_{i,j}\, x_{i,k}. \tag{8.16}$$

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,

$$\nabla\ell_n(\boldsymbol{\beta}^{(m)},\phi) = \left(\frac{\partial\ell_n(\boldsymbol{\beta},\phi)}{\partial\beta_1},\ldots,\frac{\partial\ell_n(\boldsymbol{\beta},\phi)}{\partial\beta_p}\right)^T\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}},$$

we can write the parameter update as,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + I_n^{-1}(\boldsymbol{\beta}^{(m)}, \phi)\, \nabla \ell_n(\boldsymbol{\beta}^{(m)}, \phi). \tag{8.17}$$

Now, expression (8.17) 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 (8.17) by $I_n(\boldsymbol{\beta}^{(m)}, \phi)$ to obtain,

$$I_n(\boldsymbol{\beta}^{(m)}, \phi)\boldsymbol{\beta}^{(m+1)} = I_n(\boldsymbol{\beta}^{(m)}, \phi)\boldsymbol{\beta}^{(m)} + \nabla \ell_n(\boldsymbol{\beta}^{(m)}, \phi),$$

or, using $\delta\boldsymbol{\beta} \equiv \boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}$,

$$I_n(\boldsymbol{\beta}^{(m)}, \phi)\, \delta\boldsymbol{\beta} = \nabla \ell_n(\boldsymbol{\beta}^{(m)}, \phi). \tag{8.18}$$

Note: expression (8.18) is what McCullagh and Nelder (1989) give on page 42 as $A\delta b = u$. Now, recall from expression (8.10) that,

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\phi\{y_i - \mu_i\}\, W_i\, \frac{d\eta_i}{d\mu_i}\, x_{i,j} \right].$$

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,

$$\nabla \ell_n(\boldsymbol{\beta}, \phi) = \phi \boldsymbol{X}^T \boldsymbol{W} \left((\boldsymbol{y} - \boldsymbol{\mu}) \frac{d\,\boldsymbol{\eta}}{d\,\boldsymbol{\mu}} \right),$$

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,

$$\nabla \ell_n(\boldsymbol{\beta}, \phi) = \phi \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}. \tag{8.19}$$

Similarly, inspection of (8.16) shows that the total expected information may be written in matrix form as,

$$I_n(\boldsymbol{\beta}, \phi) = \phi \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}. \tag{8.20}$$

Then, substitution of (8.19) and (8.20) into (8.18) gives the following equivalent statements (the first of these is just (8.18) repeated for ease of development):

$$\begin{aligned} I_n(\boldsymbol{\beta}^{(m)}, \phi)\, \delta\boldsymbol{\beta} &= \nabla \ell_n(\boldsymbol{\beta}^{(m)}, \phi), \\ \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} \delta\boldsymbol{\beta} &= \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \\ \delta\boldsymbol{\beta} &= \left[\left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}. \end{aligned} \tag{8.21}$$

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 (8.21) and updated estimates as,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \delta\boldsymbol{\beta}. \tag{8.22}$$

It is possible to make one further step, as in McCullagh and Nelder (1989; p. 43) to arrive at,

$$\boldsymbol{\beta}^{(m+1)} = \left[\left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{z}}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{8.23}$$

where,

$$\tilde{\boldsymbol{z}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}.$$

The use of (8.21) and (8.22) or (8.23) are entirely equivalent, and I don't really see much computational benefit one way or the other.

Comments

1. Although this derivation seems 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.

2. The dispersion parameter ϕ cancels in the progression leading to expression (8.21). Thus, just as for normal linear regression models, 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 already mentioned, a common method of estimation for ϕ is to use a moment estimator, and this will be developed in the next section.

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 (8.16) and the log likelihood given in expression (8.5) and (8.6) involve ϕ. Thus, inference from either Wald theory or more general likelihood-based procedures will require that an estimate of ϕ be available.

8.3.2 Estimation of ϕ

It is possible to use maximum likelihood to estimate the dispersion parameter ϕ along with the regression parameters β, and there may be good reasons to do so, such as inference based on likelihood-based procedures other than

asymptotic normality of estimators. Nevertheless, the traditional estimator of ϕ is based on moments. In basic generalized linear models we assume that the link function $g(\cdot)$ is continuous. We know that maximum likelihood estimators of $\boldsymbol{\beta}$ are consistent. An application of what is often called the Mann-Wald theorem then gives that

$$\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T \hat{\beta}) \xrightarrow{p} \mu_i. \tag{8.24}$$

From the structure of basic generalized linear models 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 not know 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,

$$\frac{1}{n}\sum_{i=1}^{n} W_i^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\{Y_i - \mu_i\}^2}{V(\mu_i)} \xrightarrow{p} \frac{1}{\phi}.$$

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

$$\hat{\phi} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\{Y_i - \hat{\mu}_i\}^2}{\hat{V(\mu_i)}}\right]^{-1} \xrightarrow{p} \phi. \tag{8.25}$$

8.3.3 Inference

Sufficient regularity conditions are satisfied by basic generalized linear models to allow inference about $\boldsymbol{\beta}$ based on the Wald Theory Main Result (or

Likelihood Theorem 2) of Chapter 5.5.1. In particular, we have that

$$\hat{\boldsymbol{\beta}} \quad AN(\boldsymbol{\beta}, I_n^{-1}(\boldsymbol{\beta}, \phi)), \tag{8.26}$$

where $I_n(\boldsymbol{\beta}, \phi)$ is given in (8.16) or (8.20). In practice, we can (by the Mann Wald Theorem again) replace $\boldsymbol{\beta}$ with the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ and ϕ with its consistent moment-based estimator $\hat{\phi}$ and use entries of

$$I_n^{-1}(\hat{\boldsymbol{\beta}}, \hat{\phi}) \tag{8.27}$$

to compute approximate confidence intervals for regression parameters. Similarly, the Wald Theory Tests of Chapter 5.5.2 could be used to compare some models such as those resulting from setting and not setting certain regression coefficients to zero. Note, however, that the use of likelihood ratio tests would require maximum likelihood estimation of the dispersion parameter ϕ along with the regression parameters.

Because expected values are functions of the regression parameters, intervals for these quantities can be obtained through use of the delta method. If there is only one type of covariate, pointwise confidence bands for the expectation function can be obtained through the use of the delta method of Chapter 5.5.4. Specifically, $\mu_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$ and the $1 \times p$ matrix (row vector) $\boldsymbol{D}$ is

$$\boldsymbol{D} = \left(\frac{\partial \mu_i}{\partial \beta_1}, \ldots, \frac{\partial \mu_i}{\partial \beta_p}\right).$$

The matrix $c_n^2\Sigma$ is $I_n^{-1}(\boldsymbol{\beta}, \phi)$ so a maximum likelihood estimator of μ_i is $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$, and the variance of the limit distribution of $\hat{\mu}_i$ is

$$v(\hat{\mu}_i) = \boldsymbol{D} I_n^{-1}(\boldsymbol{\beta}, \phi)\boldsymbol{D}^T. \tag{8.28}$$

In practice we use $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ as plug-ins to obtain

$$\tilde{v}(\mu_i) = \boldsymbol{D} I_n^{-1}(\hat{\boldsymbol{\beta}}, \hat{\phi})\boldsymbol{D}^T, \tag{8.29}$$

also using $\hat{\boldsymbol{\beta}}$ as a plug-in for elements of $\boldsymbol{D}$ if needed. An approximate $(1-\alpha)100\%$ interval for μ_i is then formed as

$$\hat{\mu}_i \pm z_{1-\alpha/2}\sqrt{\tilde{v}(\hat{\mu}_i)}. \tag{8.30}$$

As already mentioned, likelihood ratio tests cannot be validly conducted with standard results for basic generalized linear models because of the moment-based estimator used for ϕ. Were ϕ estimated using maximum likelihood, then likelihood ratio tests and intervals obtained by inverting likelihood ratio tests could also be used for inference. Historically, the reason this has not been more common in practice is that no unified algorithm for maximum likelihood estimation of ϕ is possible and computing for each model would need to be approached as a separate problem. This is perhaps no longer the burden it was when generalized linear models were being developed, but the use of the moment-based estimator of ϕ is still traditional. Note, however, that if a maximum likelihood estimator of ϕ is used, a number of other inferential avenues become available, such as interval estimation of quantiles of response distributions at given covariate values, which could be obtained through use of the delta method.

8.3.4 Deviance

What is called deviance is really connected with exponential dispersion families in general, not only generalized linear models, but it is most commonly encountered in basic generalized linear models and so is presented here.

Consider a set of independent random variables $Y_i;\quad i = 1,\ldots,n$ with density or mass functions of exponential dispersion family form,

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

Notice that we are allowing the distributions of the Y_i to vary only through the scalar natural parameter θ_i. Recall 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 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 $\hat{\boldsymbol{\mu}} \equiv \{\hat{\mu}_i : i = 1, \ldots, n\}$ and hence also a set of estimated natural parameters $\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}) \equiv \{\theta(\hat{\mu}_i) : i = 1, \ldots, n\}$.

Suppose, for the moment, that the dispersion parameter ϕ is known. Then, 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 full model that consists of 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 $\boldsymbol{\theta}(\boldsymbol{y}) = \{\theta(y_i) : i = 1, \ldots, n\}$. 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. With known ϕ, a likelihood ratio comparison of fitted and saturated models would then become,

$$D^* \equiv -2\{\ell(\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}), \phi) - \ell(\boldsymbol{\theta}(\boldsymbol{y}), \phi)\}, \tag{8.31}$$

where

$$\ell(\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}), \phi) = \sum_{i=1}^{n} \left[\phi\{y_i\theta(\hat{\mu}_i) - b(\theta(\hat{\mu}_i))\} + c(y_i, \phi)\right],$$

and

$$\ell(\boldsymbol{\theta}(\boldsymbol{y}), \phi) = \sum_{i=1}^{n} \left[\phi\{y_i\theta(y_i) - b(\theta(y_i))\} + c(y_i, \phi)\right].$$

Expression (8.31) defines the *scaled deviance* for a model based on independent exponential dispersion family random variables. Notice that it may also be written as

$$D^* = -2\phi \sum_{i=1}^{n} \left[y_i\{\theta(\hat{\mu}_i) - \theta(y_i)\} - b(\theta(\hat{\mu}_i)) + b(\theta(y_i)\right], \tag{8.32}$$

because, with ϕ considered known, the terms $c(y_i, \phi)$ cancel in the difference. The parameter ϕ may be seen in (8.32) to constitute a scaling factor, and the *unscaled deviance* is defined as $D = D^*/\phi$, or

$$D = -2 \sum_{i=1}^{n} \left[y_i\{\theta(\hat{\mu}_i) - \theta(y_i)\} - b(\theta(\hat{\mu}_i)) + b(\theta(y_i)\right]. \tag{8.33}$$

Scaled and unscaled deviance are measures of the departure of a fitted model from a saturated model, which intuitively 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 and 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 (8.32) 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 8.2

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 binomial 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) pa-

rameterization, $\theta_i = \mu_i$, $\phi = 1/\sigma^2$, 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 $\phi = \alpha$ we can arrive at an exponential dispersion family representation of the density with $b(\theta) = -\log(-\theta)$. 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 8.2 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, usually with the moment-based estimator of Chapter 8.3.2.

8.3.5 Deviance Residuals

Deviance residuals are simply component quantities in deviance and form the basic residuals used to assess generalized linear models. Each observation y_i contributes one term to (8.32) or (8.33), 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 $i = 1, \ldots, n$,

$$d_i \equiv \text{sign}(y_i - \hat{\mu}_i)\sqrt{D_i^*}. \tag{8.34}$$

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,

$$d_i' = \frac{d_i}{(1 - h_{i,i}^{(G)})^{1/2}}, \tag{8.35}$$

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 (8.35) is justified by results on the first two moments of "generalized residuals" 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 all of this is contained in Davison and Snell (1991), who also point out that (8.35) 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 except 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 (8.35) 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 (8.35), and $\boldsymbol{Q} = \boldsymbol{X}$.

8.4 Bayesian Estimation and Inference

To conduct a Bayesian analysis of a basic generalized linear model requires that we specify prior distributions for model parameters, derive a joint posterior, and consider how we might assess the performance of a fitted model. For completeness, we restate the overall structure of the data model for a basic glm. We denote independent response random variables as $\{Y_i : \ i = 1, \ldots, n\}$. Random Model Component: The probability density or mass function of Y_i is assumed to be of the form, for $y \in \Omega_i$, $\theta_i \in \Theta$ and $\phi \in \Phi$,

$$f(y|\theta_i, \phi) = \exp\left[a_i(\phi)\{y\theta_i - b(\theta_i)\} + c(y, \phi)\right]. \tag{8.36}$$

As before, in most cases $a_i(\phi) = \phi$ although if the Y_i are binomial proportions we may have $a_i(\phi) = m_i$ where m_i is the binomial sample size for Y_i. Given this form, we have that $E(Y_i) = \mu_i = b'(\theta_i)$ and $var(Y_i) = [1/a_i(\phi)]b''(\theta_i)$. Systematic Model Component: For a given monotone function $g(\cdot)$, known covariates $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})^T$ and unknown regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$,

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Typically, $x_{i,0} = 1$ for all i, β_1 is re-labeled as β_0 and the indices of $\boldsymbol{\beta}$ go from 0 to $p - 1$, but this is not necessary.

Bayesian analysis of a basic glm is driven by choice of prior distributions. There has been a good deal of discussion of the issue of how to choose priors for glms in the literature, and a variety of suggestions for general approaches have been put forward. Similarly to non-Bayesian analysis, the focus has been nearly entirely on the regression coefficients with little consideration given to the dispersion parameter.

8.4.1 Inducing a Prior on $\boldsymbol{\beta}$

One approach to prior specification for glms and, in fact, regression models in general, is to avoid formulating prior distributions for regression parameters directly and instead specify prior distributions on the expected values of response variables at various points in the covariate space. Two particular versions of this approach are data augmentation priors and conditional means priors (Bedrick, Christensen, and Johnson, 1996; Greenland and Christensen, 2001). A related procedure is suggested by (Gelman et al., 1995, p. 389). A description of conditional means priors will convey the basic ideas. If $\boldsymbol{\beta}$ contains p elements, select p vectors of covariates which will be denoted as $\tilde{\boldsymbol{x}}_k;\ k = 1, \ldots, p$. For a given link function g, let the expected responses at those covariates be denoted as $\tilde{m}_k = g^{-1}(\tilde{\boldsymbol{x}}_k);\ k = 1, \ldots, p$. Based on previous information or expert opinion, obtain prior distributions on the expected responses $\pi_{0,k}(\tilde{m}_k);\ k = 1, \ldots, p$. Assume these prior distributions are proper. Let $\boldsymbol{G}$ and $\boldsymbol{G}^{-1}$ be vector operators that apply g and g^{-1} to each element of their arguments, respectively. If $\tilde{\boldsymbol{X}}$ is the $p \times p$ matrix with $\tilde{\boldsymbol{x}}_k$ in the k^{th} row and $\tilde{\boldsymbol{X}}$ is nonsingular, we have

$$\tilde{\boldsymbol{m}} = G^{-1}(\tilde{\boldsymbol{X}}\boldsymbol{\beta}) \qquad\qquad \boldsymbol{\beta} = \tilde{\boldsymbol{X}}^{-1} G(\tilde{\boldsymbol{m}})$$

We typically assume the prior distributions $\pi_{0,k}(\tilde{m}_k)$ are independent so that the joint prior on $\tilde{\boldsymbol{m}}$ is

$$\pi_0(\tilde{\boldsymbol{m}}) = \prod_{k=1}^{p} \pi_{0,k}(\tilde{m}_k). \tag{8.37}$$

Let $DG^{-1}(\tilde{\boldsymbol{X}}\boldsymbol{\beta})$ be the $p \times p$ matrix with jk^{th} element

$$\frac{\partial}{\partial \beta_j} g^{-1}(\tilde{\boldsymbol{x}}_k \boldsymbol{\beta}),$$

and let $dG^{-1}(\tilde{\boldsymbol{x}}_k\boldsymbol{\beta})$ denote the $p-$vector with k^{th} element,

$$\frac{d}{d(\tilde{\boldsymbol{x}}_k\boldsymbol{\beta})}g^{-1}(\tilde{\boldsymbol{x}}_k\boldsymbol{\beta}) = \frac{d}{d\tilde{m}_k}g^{-1}(\tilde{m}_k).$$

Then $DG^{-1}(\tilde{\boldsymbol{X}}\boldsymbol{\beta}) = \tilde{\boldsymbol{X}}dG^{-1}(\tilde{\boldsymbol{x}}_k\boldsymbol{\beta})$ and the induced joint prior on $\boldsymbol{\beta}$ is,

$$\begin{aligned}\pi(\boldsymbol{\beta}) &= \pi_0[\boldsymbol{G}^{-1}(\tilde{\boldsymbol{X}}\boldsymbol{\beta})]\left|DG^{-1}(\tilde{\boldsymbol{X}})\right| \\ &= \prod_{k=1}^{p}\pi_{0,k}(\tilde{m}_k)\left|\tilde{\boldsymbol{X}}dG^{-1}(\tilde{m}_k)\right|. \end{aligned} \tag{8.38}$$

A related, but quite distinct procedure is proposed by Chen and Ibrahim (2003). These authors consider glms with $n \times p$ covariate matrices $\boldsymbol{X}$ of full rank and suggest that elicited values for the expected values of responses $\boldsymbol{y}_0 = (y_{0,1}, \ldots, y_{0,n})^T$ at the design points (i.e., the rows) of $\boldsymbol{X}$ be used as parameter values in a prior for the natural parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ of (8.36). Assuming independence of the elements of $\boldsymbol{y}_0$,

$$\pi(\boldsymbol{\theta}|a_0, \tau, \boldsymbol{y}_0) \propto \prod_{i=1}^{n}\exp\left[a_0\tau\{(y_{0,i}\theta_i - b(\theta_i)\}\right] = \exp\left[a_0\tau\sum_{i=1}^{n}\{(y_{0,i}\theta_i - b(\theta_i)\}\right], \tag{8.39}$$

where $\tau = \phi^{-1}$ and $b(\theta_i)$ are the same as in (8.36). In (8.39) $E(y_{0,i}) = \theta_i$ so the elicited values $y_{0,i}$ can be thought of as prior predictions for $E(Y_i)$ in (8.36). The quantity a_0 is a scalar parameter that can be viewed as a relative prior sample size n_0/n with n being the sample size of the actual data. Chen and Ibrahim (2003) point out that while (8.39) is conjugate for the θ_i of (8.36), the prior it induces on $\boldsymbol{\beta}$ is not conjugate for that parameter. But, by noting that $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\mu_i = b'(\theta_i)$ implies that

$$\theta_i = b'^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}) = h(\boldsymbol{x}_i\boldsymbol{\beta}),$$

substituting $h(\boldsymbol{x}_i, \boldsymbol{\beta})$ directly into (8.39) results in,

$$\pi(\boldsymbol{\beta}|a_0, \tau, \boldsymbol{y}_0) \propto \exp\left[a_0\tau\sum_{i=1}^{n}y_{0,i}h(\boldsymbol{x}_i^T\boldsymbol{\beta}) - a_0\tau\sum_{i=1}^{n}b\{h(\boldsymbol{x}_i^T\boldsymbol{\beta})\}\right]. \tag{8.40}$$

Note that the procedure of Chen and Ibrahim (2003) differs from those of Bedrick et al. (1996) in that $\boldsymbol{y}_0$ is dimension n (the same as $\boldsymbol{y}$), while the $\tilde{\boldsymbol{m}}$ in (8.38) is dimension p (the same as $\boldsymbol{\beta}$). This results from elicitation of a prediction for $E(Y_i)$ at each design point of $\boldsymbol{X}$ by Chen and Ibrahim, while the procedures of Bedrick *et al.* do so only for a set of $p < n$ points $m\tilde{b}X$ which may not even correspond to as subset of those in $\boldsymbol{X}$. Let $m_i = (a_0 y_{0,i} + y_i)/(a_0 + 1)$ and $s = a_0 + 1$. Chen and Ibrahim (2003) show that the prior (8.40) combined with the likelihood (8.36 results in a posterior for $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}|a_0, \tau, \boldsymbol{y}_0, \boldsymbol{y}) \propto \exp\left[s\tau \sum_{i=1}^{n} m_i h(\boldsymbol{x}_i^T\boldsymbol{\beta}) - s\tau \sum_{i=1}^{n} b\{h(\boldsymbol{x}_i^T\boldsymbol{\beta})\}\right]. \qquad (8.41)$$

Thus, the prior (8.40) is conjugate for $\boldsymbol{\beta}$ in basic glms or, more accurately, conditionally conjugate given the dispersion parameter ϕ.

8.4.2 Direct Priors for $\boldsymbol{\beta}$

The most straightforward way to assign a prior distribution to $\boldsymbol{\beta}$ is to recognize that a proper prior avoids the need to demonstrate posterior propriety and that these regression parameters control the location of response distributions, albeit usually through a nonlinear link function. Still, most link functions applied to expected values in glms have ranges on the entire line and normal prior distributions for coefficients in the linear predictor seem a natural choice. Dellaportas and Smith (1993) use a normal prior for regression parameters in a logistic regression. Assume a prior distribution for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is specified as,

$$\pi(\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right], \qquad (8.42)$$

where $\boldsymbol{\beta}_0$ is the prior mean and Σ is a positive definite covariance matrix. Ignoring, for the moment, that the model might involve a dispersion parameter

ϕ, the posterior for $\boldsymbol{\beta}$ (or more properly, $\boldsymbol{\beta}$ given ϕ) is,

$$p(\boldsymbol{\beta}|\boldsymbol{y},\phi) \propto \exp\left[\sum_{i=1}^{n} a_i(\phi)\{y_i\theta_i - b(\theta_i)\} - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right], \quad (8.43)$$

where $b'(\theta_i) = \mu_i = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})$. The posterior (8.43) will not be available in closed form for other than a model with normal random component and identity link and MCMC methods will be required to approximate it. As will be illustrated in sequel, the prior (8.42) can be combined with a proper uniform prior on the dispersion parameter ϕ such as $\pi(\phi) = 1/AI(0 < \phi < A)$ and then the joint posterior becomes $p(\boldsymbol{\beta},\phi|\boldsymbol{y}) = p(\boldsymbol{\beta}|\boldsymbol{y},\phi)\pi(\phi)$.

8.4.3 Improper Priors for $\boldsymbol{\beta}$

The use of improper priors for the regression coefficients in a basic glm is a complex topic because general results are illusive, leaving demonstration of posterior propriety largely a model-by-model exercise. Ibrahim and Laud (1991) discuss the use of Jeffreys' priors for the regression coefficients $\boldsymbol{\beta}$ in glms. Such priors can be either proper or improper and these authors give conditions necessary for posterior propriety for improper Jeffreys' priors. Along the way, **?** also present a simple example of a case in which an improper prior leads to an improper posterior. Consider a model with an exponential random component suppose we have only one observation y_1, which then has density, for some $\beta > 0$,

$$f(y_1|\beta) = \beta\exp(-\beta y_1); \quad y_1 > 0$$

Suppose further that our model uses the identity link function $g(\mu_1) = \mu_1 = \eta_1 = x_1\gamma$ and that $x_1 > 0$. If we specify an improper prior for γ, the posterior becomes, for $\gamma > 0$,

$$p(\gamma|y_1) \propto \frac{1}{x_1\gamma}\exp\left(-y_1/x_1\gamma\right).$$

Let $z = (x_1\gamma)^{-1}$ and note that for $0 < \gamma < \infty$ we have $0 < z < \infty$. Then the integral of the posterior is,

$$\int_0^\infty p(\gamma|y_1)\,d\gamma = \int_0^\infty \frac{1}{z}\exp(-y_1 z)\,dz$$

and it is not difficult to show that this integral is not finite. Thus, the posterior for this highly simplified situation is not proper. Now, the model of Ibrahim and Laud (1991) is a bit odd in that it pairs a gamma random component with an identity link, thus requiring a restriction on the values of the linear predictor $x_1\gamma > 0$. If we instead use the canonical link $g(\mu_1) = 1/\mu_1$ the posterior can be shown to be proper, and similarly for a log link $g(\mu_1) = \log(\mu_1)$. Similar to Chen and Ibrahim (2003), Ibrahim and Laud (1991) consider situations in which the $n \times p$ glm covariate matrix $\boldsymbol{X}$ is full rank. In constast, Gelfand and Sahu (1999) focus on situations in which the rank of $\boldsymbol{X} = r < p$, such as in ANOVA models written with an overall mean parameter (so-called effects models). These authors connect posterior propriety with a Bayesian notion of identifiable parameters and give conditions for posterior propriety in that context. The role of the link function in determining posterior propriety is also clear in this work and canonical link functions are relied on heavily by Gelfand and Sahu (1999).

Overall, propriety of posterior distributions for regression parameters in glms when prior distributions are taken to be improper depends on the combination of random model component and link function. Further, in models for discrete response variables with binary, binomial, and Poisson random components and canonical links, if all observed responses are 0, improper priors will lead to improper posteriors. Although extreme, the indication is that posterior propriety with improper priors can depend on observed data values, which should not be comforting in a world where the choice of prior distributions is

intended to be free of the data-generating mechanism.

8.4.4 Priors for the Dispersion Parameter

As previously noted, the majority of material on prior specification for glms dismisses the role of the dispersion parameter ϕ in a complete analysis. Gelfand and Ghosh (2000) consider posterior propriety for glms at some length, but do not even mention dispersion parameters. Ibrahim and Laud (1991) indicate that joint Jeffreys' priors for $(\boldsymbol{\beta}, \phi)$ are often not practical for numerical computations. They recommend assuming that $\boldsymbol{\beta}$ and ϕ are independent a priori, and forming a joint prior as the product of a Jeffreys' prior for $\boldsymbol{\beta}$ and some proper prior for ϕ. In a similar vein, Gelman et al. (1995, p. 388) indicate that joint priors for $\boldsymbol{\beta}$ and ϕ can be constructed as $\pi(\boldsymbol{\beta}, \phi) = \pi(\boldsymbol{\beta}|\phi)\,\pi(\phi)$ but otherwise offer no advice on formulation of $\pi(\phi)$, or how $\boldsymbol{\beta}$ might depend on ϕ in $\pi(\boldsymbol{\beta}|\phi)$. Chen and Ibrahim (2003) show that if $\pi(\boldsymbol{\beta}|\phi)$ is a conjugate conditional means prior and the likelihood is bounded, then for any $\pi(\phi)$, proper or improper, the joint prior $\pi(\boldsymbol{\beta}|\phi)\,\pi(\phi)$ is proper, so that the posterior will be as well.

8.4.5 Strategies for Prior Formulation

The preceding discussion of prior specification for basic glms should make it clear that this is not a trivial subject and that there is no generally accepted standard procedure available. Where does this leave us from a practical standpoint? The following points are relevant.

1. If one does not have previous assurance from the literature that improper uniform priors for $\boldsymbol{\beta}$ and ϕ will lead to a proper posterior for a particular

model, one can attempt to verify that

$$p(\boldsymbol{\beta}|\boldsymbol{y},\phi) \propto \int \exp\left[\sum a_i(\phi)\{y_i\theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))\}\right] d\boldsymbol{\beta} < \infty,$$

where $\theta_i(\boldsymbol{\beta}) = b'^{-1}[g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})]$. This could be an arduous task. Another possibility would be to verify the condition of Gelfand and Sahu (1999). Then a proper prior, such as a proper uniform on $(0, A)$, could be assigned to ϕ.

2. If one is able to elicit prior guesses for expected responses at either a subset or the entire set of design points one could attempt one of the strategies of Bedrick et al. (1996) *et al.* or Chen and Ibrahim (2003). Either of these can be combined with an independent prior on ϕ. For use with the conditional means prior of Bedrick *et al.* a proper prior on ϕ would most likely be called for as these authors assume ϕ is known in developing their prior for $\boldsymbol{\beta}$. For combination with the conjugate prior strategy of Chen and Ibrahim (2003), either proper or improper priors for ϕ would be possible under the condition of a bounded likelihood, as indicated at the end of the previous section.

3. A Jeffreys' prior for $\boldsymbol{\beta}$ can be derived and verified to lead to a proper posterior if ϕ is fixed as in binary, binomial, and Poisson models. If Jeffreys' prior is proper it can be combined with a proper prior on ϕ. But if a Jeffreys' prior is not proper and is not independent of ϕ then forming a joint prior in this same way is not guaranteed to result in a proper posterior.

4. The most straightforward strategy is to use proper priors for both $\boldsymbol{\beta}$ and ϕ, obviating the need to check conditions for posterior propriety. For many basic glms the parameter space of $\boldsymbol{\beta}$ is $\mathbb{R}^p$ and a normal prior on

$\boldsymbol{\beta}$ would seem to be a natural choice. In the absence of prior information to the contrary, one can specify independent normal priors on the elements of $\boldsymbol{\beta}$. The locations of these prior distributions can come from previous or related studies, if possible or, if we have little or no knowledge about how a covariate might influence responses, we might choose a non-zero value for the mean of an intercept parameter with means of zero for the remainder. Variances can be chosen as moderately large, but not extreme. This is because regression coefficients for nonlinear expectation functions generally operate within particular numerical windows of rationality for responses measured or observed on a given scale. For example, if responses are anticipated to be no more than 100 with a single covariate, the parameters of an exponential expectation function would not be anticipated to result in $\eta_i = \beta_0 + \beta_1 x_i > 4.6$ and we might choose prior variances for β_0 and β_1 so that this will be true with high probability. Note that, assuming that covariates are fixed and nonrandom, using their values does not violate the precept that priors should not depend on examination of the data. If there are multiple covariates, standardizing those quantities will be beneficial in keeping regression coefficients within their numerical ranges of rationality. Normal priors for regression coefficients can be paired with a proper uniform prior for ϕ if the model includes a dispersion parameter. This strategy is perhaps the easiest of those covered to implement, but should be combined with posterior checks on model adequacy. This topic will be covered in a later section of this chapter.

8.4.6 Analysis Using MCMC Methods

It is unlikely that posterior distributions for a basic glm will be available in closed form, and analysis generally requires the use of MCMC procedures. Note that there are normal approximations to posteriors that can be used (e.g., Gelman et al., 1995, p. 330-331) that make the use of Markov Chain sampling unnecessary, but here we will suppose that we wish to determine the actual posterior. The two basic MCMC algorithms of Metropolis-Hastings and Gibbs Sampling covered in Chapter 7.9 are usually suitable for use with glms. The joint posterior for a basic glm with $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$has the form,

$$p(\boldsymbol{\beta}, \phi|\boldsymbol{y}) \propto \exp\left[\sum_{i=1}^{n} a_i(\phi)\{y_i\theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))\} + c(y_i, \phi)\right] \pi(\boldsymbol{\beta}, \phi), \quad (8.44)$$

where $\theta_i(\boldsymbol{\beta}) = b'^{-1}[g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})]$. Even if we take $\pi(\boldsymbol{\beta}, \phi) = \pi(\boldsymbol{\beta})\pi(\phi)$ and $\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \pi_j(\beta_j)$, not much simplification is available for a Gibbs Sampling algorithm, because the full conditional posterior of each element of $\boldsymbol{\beta}$ depends on the full likelihood,

$$p(\beta_j|\boldsymbol{y}, \boldsymbol{\beta}_{-j}, \phi) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \phi)\, \pi_j(\beta_j),$$

as does the full conditional posterior of phi,

$$p(\phi|\boldsymbol{y}, \boldsymbol{\beta}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \phi)\, \pi(\phi).$$

A Metropolis-Hastings algorithm is attractive if the dimension p of $\boldsymbol{\beta}$ is small, but it can be difficult to mix over the entire space of $\boldsymbol{\beta}$ if p is large, say greater than 2 or 3. In such situations it may be possible to partition $\boldsymbol{\beta}$ into small pieces $\{\boldsymbol{\beta}_k : \ k = 1, \ldots, K\}$ and use a Metropolis-Hastings algorithm to sample each piece within the overall structure of a Gibbs Sampling algorithm. Splitting ϕ off into its own piece is also often useful, because tuning a Metropolis-

Hastings algorithm may be quite different for the dispersion parameter than for the regression coefficients.

If $\pi(\boldsymbol{\beta}, \phi) = \pi(\boldsymbol{\beta})\pi(\phi)$ with $\pi(\phi) = (1/A)I(0 < \phi < A)$ and

$$\pi(\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right],$$

then a Metropolis-Within-Gibbs agorithm has two steps, to sample from the two full conditionals,

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \phi) \propto \exp\left[\sum_{i=1}^{n} a_i(\phi)\{y_i\theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))\} - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right],$$

$$p(\phi|\boldsymbol{y}, \boldsymbol{\beta}) \propto \exp\left[\sum_{i=1}^{n} a_i(\phi)\{y_i\theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))\}\right] I(0 < \phi < A). \tag{8.45}$$

This algorithm will be illustrated with example of cadmium in yellow perch considered previously from a likelihood standpoint.

Inference for the regression parameters $\boldsymbol{\beta}$ and the dispersion parameter ϕ is straightforward based on the joint posterior distribution. Typically, one will report summary quantities and credible intervals for individual parameters based on the marginal posteriors contained in the MCMC output. It would seem natural to represent the fitted regression curve by evaluating the expectation function of the model at the posterior means of the regression coefficients, which is analogous to evaluating the expectation function at maximum likelihood estimates in a likelihood analysis. But this would not be the posterior mean of the expectation function except for models with identity link. What is needed is the posterior of the μ_i themselves, which we can produce in the following way. First, determine a set of covariate values at which to evaluate the regression, $\{\boldsymbol{x}_h : h = 1, \ldots, H\}$. These do not need to be the same as the set of observed covariate values and typically $n < H$, but they should be

contained in the convex hull of those values. Assuming we are using MCMC to approximate the joint posterior of model parameters $(\boldsymbol{\beta}, \phi)$, evaluate the expectation function at each of the chosen set of covariate values for each posterior draw $\tilde{\mu}_{h,m} = g^{-1}(\boldsymbol{x}_h \boldsymbol{\beta}_m)$, where $\boldsymbol{\beta}_m$ is the m^{th} draw from the joint posterior distribution, $m = 1, \ldots, M$ and $h = 1, \ldots, H$. The empirical distribution of the M values of $\mu_{h,m}$ at $\boldsymbol{x}_h$ approximate the posterior distribution of the expectation function at that covariate value. From here, we can obtain the posterior mean of the expectation function for $h = 1, \ldots, H$,

$$\hat{E}(\mu_h|\boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M} \tilde{\mu}_{h,m},$$

and these quantities are a pointwise approximation to the posterior mean of the regression function. Similarly, a pointwise credible band for the regression function is calculated based on quantiles of the empirical distributions of the $\mu_{h,m}$. To compute the q^{th} quantile of a set of values $\boldsymbol{Z} = \{Z_j : j = 1, \ldots, N\}$ let $\tilde{q} = \lfloor Nq \rfloor$ be the largest integer less than or equal to Nq, and let $Z_{[a]}$ be the a^{th} largest value of $\boldsymbol{Z}$ for any integer $1 \leq a \leq N$. Then the q^{th} quantile of $\boldsymbol{Z}$ is $Z_{[\tilde{q}]}$. Based on this, a $1 - \alpha$ pointwise credible band for the regression function is calculated by taking $q_\ell = \lfloor (M+1)\alpha \rfloor$ and $q_u = \lfloor (M+1)(1-\alpha/2) \rfloor$ and determining the set of lower and upper endpoints, for $h = 1, \ldots, H$,

$$L_h = \mu_{[q_\ell]}$$
$$U_h = \mu_{[q_u]},$$

8.5 Case Study: Cadmium in Yellow Perch

Contamination of fish by metals can be a serious problem due to both effects on ecosystem function and potential health effects for humans. Heavy metals

typically accumulate in fish over time because uptake from the environment is usually faster than depuration (elimination) from the body. As part of a much larger study investigating bio-accumulation of cadmium (Cd) and how it is affected by pH (Powell, 1993), data were collected on the whole-body concentration of Cd (in ng/g wet weight) and length of Yellow Perch in Little Rock Lake, Wisconsin. Because fish have indeterminate growth (they keep growing their entire lives, like trees) length is a surrogate measure of age or time of exposure to Cd in the lake. Thus, the relation between length and Cd concentration is of interest to aquatic scientists. A finding that concentration increases with time of exposure would not impress anyone, it is how concentration increases with time of exposure, including the distribution of concentrations in individual fish exposed to the same source (i.e., in the same lake) for the same amount of time or, roughly, fish of the same length. This is a problem for which we might naturally think of generalized linear models as a potential for analysis. A scatterplot of Cd concentration (in ng/g wet weight) versus length (in mm) is presented in Figure 8.4.

8.5.1 Selection of Random Model Component

In choosing a random component for this problem we first consider the possible values of the response variables, which are associated with a chemical concentration and thus consists of the positive line. Certainly we would like a continuous random component. This suggests gamma, inverse Gaussian and possibly normal, assuming normal distributions fitted to the data do not put more than negligible probability on the negative line. The scatterplot of Figure 8.4 indicates that there are small response values, but the scale of the vertical axis makes it difficult to determine exactly how small. Also, the variability of

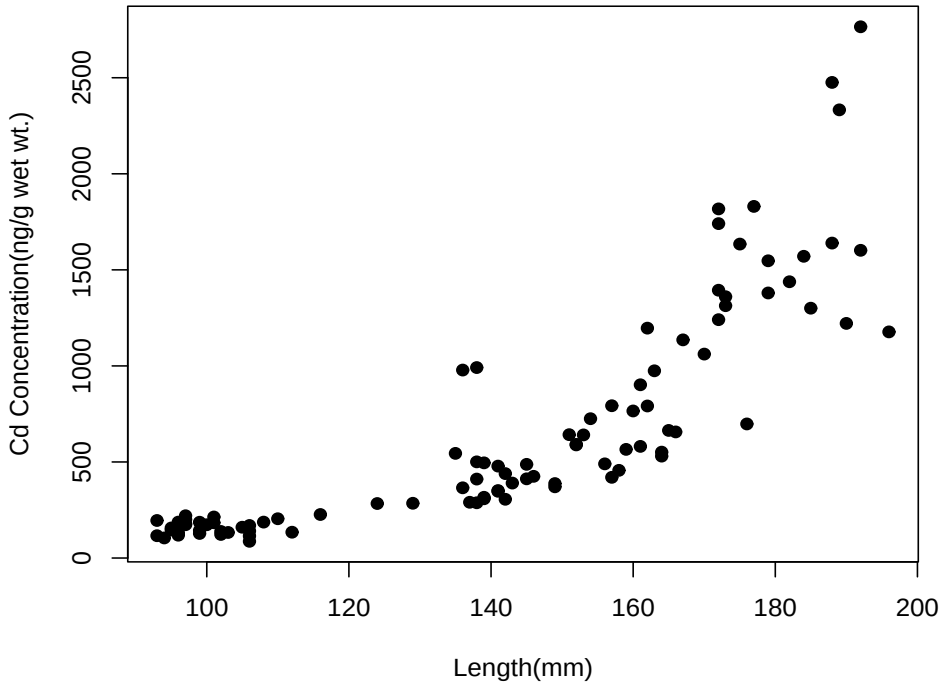


Figure 8.4: Scatterplot of Cd concentration against length in Yellow Perch from Little Rock Lake, Wisconsin.

responses in the covariate region where there are small values seems considerably less than the variability for larger covariate values. This might indicate that a normal would not put too much probability on the negative line, but it also indicates that the constant variance of a normal across the covariate range might not be appropriate. Finally, there may be some hint of right skew behavior in the responses as well, suggesting again gamma or inverse Gaussian random components. Just from consideration of the problem and examination of a simple scatterplot, we have arrived at a set of two candidate random com-

ponents. We might retain the possibility of a normal random component, but mostly to verify that our assessment has not been off target. That is, if we fit a normal model we would anticipate determining that it is not appropriate.

These data were binned by taking sets of 10 observations from the ordered values of Cd concentration. This is effective here because observations are scattered roughly uniformly along the covariate axis; that is, there are no big gaps. In any application the choice of binning rule is arbitrary, and several plots should be produced to ensure that the choice of binning rule is not having too great an influence on the diagnostic. It is also advisable to keep track of how many observations fall into each bin to ensure that one or two groups with very few observations are not having too great an influence. The plot of Figure 8.5 looks like it could be reasonably described by a straight line. The line shown was determined by ordinary least squares and has a slope of 1.14 suggesting a value of $\theta = 2$ in the variance function $V(\mu_i) = \mu_i^\theta$. This is the variance function of a gamma random component. At this point we would have a preference for a gamma random component over either a normal or an inverse Gaussian random component for this problem.

8.5.2 Selection of Systematic Model Component

The scatterplot in Figure 8.4 suggests a regression function that increases more rapidly than the covariate of length. A log link might be possible, or a power link. We can examine the possibility of a log link by plotting the logarighm of Cd concentrations against length, which is shown in Figure 8.6. The points of Figure 8.6 appear to be reasonably well described by a straight line, supporting the choice of a log link function. Similarly to Example 8.1, we are not interested in transforming response variables because we are interested in the distribution

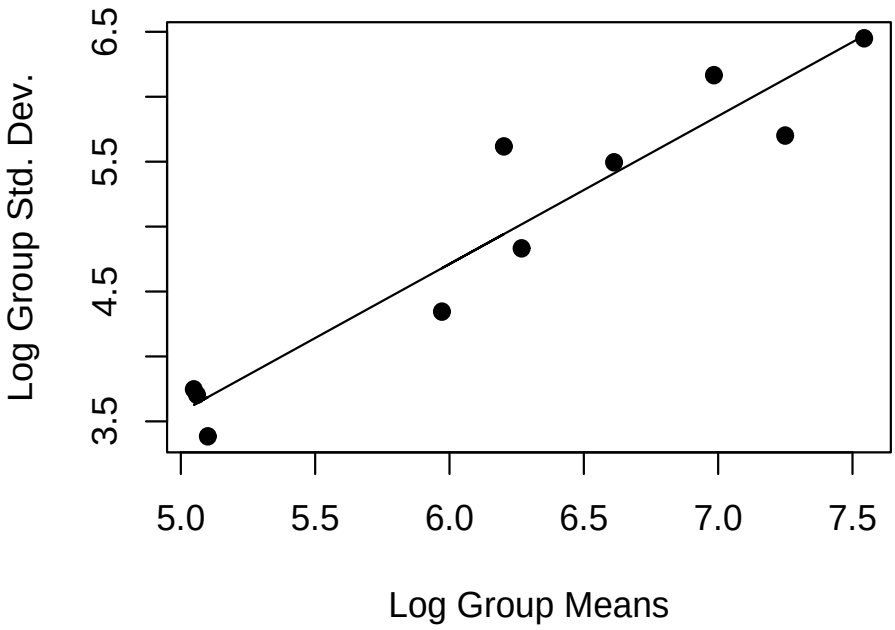


Figure 8.5: Box-Cox plot for the data of Figure 8.4.

of responses at given covariate values. Transformation of responses was used only as a device to identify an appropriate systematic model component.

8.5.3 Likelihood Analysis and Comparison of Models

Consideration of the problem of cadmium concentration in Yellow Perch in Little Rock Lake, Wisconsin has led to the likely candidate model consisting of a gamma random component and a log link. For purposes of illustration, in this section we also fit models with normal random component and log link and inverse Gaussian random component and log link. These three models differ

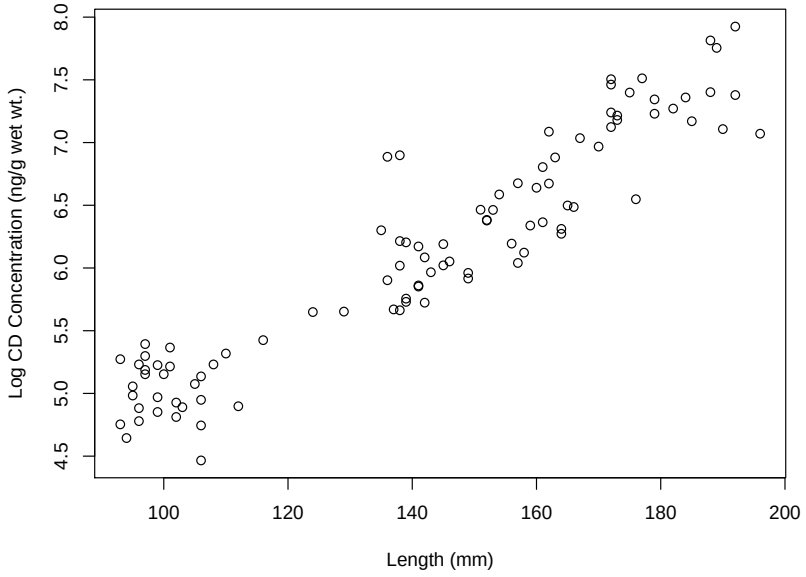


Figure 8.6: Scatterplot of log Cd concentration against length in Yellow Perch from Little Rock Lake, Wisconsin.

in the random model component but all use the same systematic component, namely $g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 x_i$, where x_i is length (mm) of fish i.

Maximum likelihood estimates of the regression parameters β_0 and β_1, Wald intervals and scaled deviances for these models are contained in Table 8.1. Estimated dispersion parameters were computed using the moment-based estimator (8.25) as $\hat{\phi} = 9.028$ for the Gamma model, $\hat{\phi} = 1.4 \times 10^{-5}$ for the normal model, and $\hat{\phi} = 3,531$ for the inverse Gaussian model. Clearly, the dispersion parameters in these models do not have the same meaning and so comparison of these estimates is not meaningful.

Notice from the estimates of Table 8.1 that $\hat{\beta}_0$ and $\hat{\beta}_1$ are quite similar among all three models, and even intervals do not differ dramatically, although

	β_0		β_1		Scaled
Model	Point Est.	Interval Est.	Point Est.	Interval Est.	Deviance
Gamma	2.327	$(2.019, 2.635)$	0.027	$(0.025, 0.029)$	84.4
Normal	2.234	$(1.587, 2.880)$	0.028	$(0.024, 0.032)$	95.0
Inv. Gaussian	2.477	$(2.196, 2.758)$	0.026	$(0.024, 0.028)$	86.4

Table 8.1: Maximum likelihood estimates and deviances for three models.

those for the model with a normal random component are a bit wider than for the other two models. This means that graphs showing fitted expectation functions for the three models will all look the same. The differences in these models, if there are any, are in terms of model structures other than expected values. It is interesting to determine the ways these models are representing variances. The variances of the gamma, normal, and inverse Gaussian models are $var(Y_i) = (1/\phi)V(\mu_i)$. Using estimates from Table 8.1 we can examine estimated variances for responses of different magnitudes. The first, second, and third quartiles of all lengths (the covariate) in these data are $Q_1 = 106$, $Q_2 = 143$ and $Q_3 = 165$, respectively. Estimated standard deviations for these values are given in Table 8.2

From Table 8.2 we can see that the differences among these models are in the representation given to variance. In particular, using the estimated values of ϕ given just before Table 8.1, the gamma model takes variances as $\hat{var}(Y_i) = (1/9.028)\mu_i^2$ while the inverse Gaussian model uses $\hat{var}(Y_i) = (1/3531)\mu_i^3$. Thus, the estimated variance of Y_i from the gamma model will be greater than that from the inverse Gaussian model for any $\mu_i < 391.12$ and less than that from the inverse Gaussian model for any $\mu_i > 391.12$.

We can represent the mean-variance relations for the three models graphically by plotting variance (or standard deviation) against mean, which is done

Model	Length(mm)	$\hat{\mu}_i$	$\{\hat{var}(Y_i)\}^{1/2}$
Gamma	106	182.73	60.82
	143	499.52	166.25
	165	908.33	302.31
Normal	106	178.67	269.02
	143	500.64	269.02
	165	923.82	269.02
Inverse Gaussian	106	187.64	43.26
	143	491.32	183.28
	165	870.81	432.48

Table 8.2: Estimated means and standard deviations at quartiles of the covariate.

in Figure 8.7 using standard deviation. Note that the gamma model produces a straight line since what is being plotted for that model is $\sqrt{(1/9.028)\mu_i^2}$ against μ_i. The group means and variances (as standard deviations) from Figure **??** are also plotted on this graph. The model with gamma random components seems to be a better reflection of the mean-variance relation exhibited by the data than either of the other models.

We can see these same model behaviors by looking at standardized deviance residuals plotted against estimated expectations or fitted values. These plots can be interpreted in a similar fashion to residual plots from linear regression models; no pattern is good, trend indicates a problem with the expectation function, and uneven spread indicates a problem with variance modeling. Residuals plots for our three models are presented in Figures 8.5.3, 8.8, and 8.9 for gamma, normal, and inverse Gaussian models, respectively. These residual

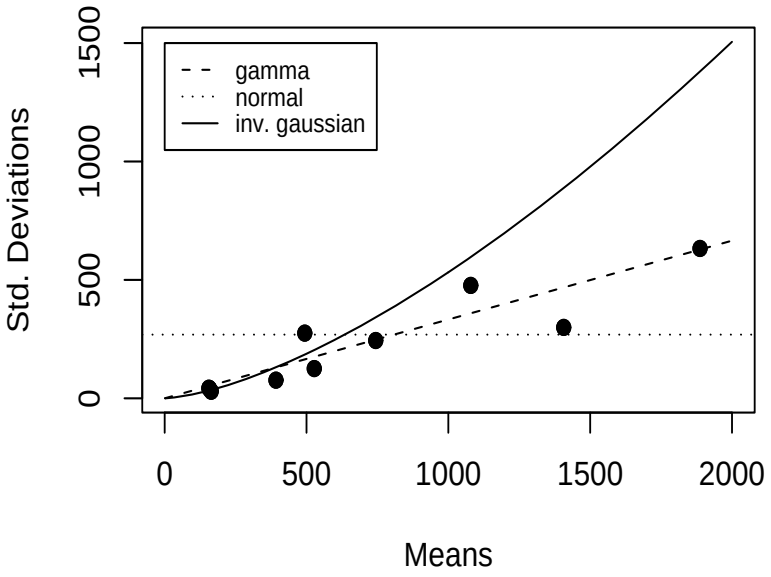
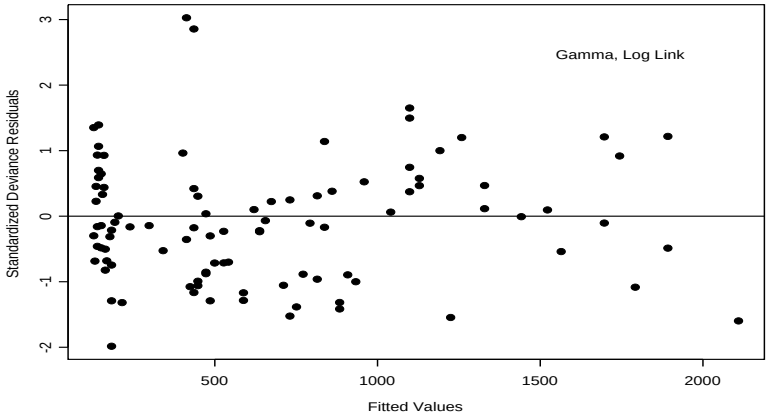


Figure 8.7: Estimated standard deviations as functions of expected value for the three models with data values overlaid.

plots confirm our impression that the model with gamma random component appears more appropriate to represent these data than either of the other models. Note that the residual plots reflect exactly the same behaviors as shown in Figure 8.7. The model with normal random component fails to account for the increase in variance as mean increases so that residuals have this pattern. The inverse Gaussian model "goes too far" in assuming that variance increases as a cubic function of mean. The resulting residuals then show the opposite behavior, having smaller variances for larger fitted values. While not perfect, the residual plot for the gamma model is much more well behaved and would



captionStandardized deviance residuals for model with gamma random component and log link.

be our 'Goldilocks' choice in this example.

Finally, we illustrate the production of a pointwise confidence band for the expectation function for the model with gamma random component. A pointwise band results from computing interval estimates for expected values across the range of the covariate and then simply connecting the endpoints. Since expectations μ_i are functions of the regression parameters β_0 and β_1, and we have maximum likelihood estimates of the regression parameters, we

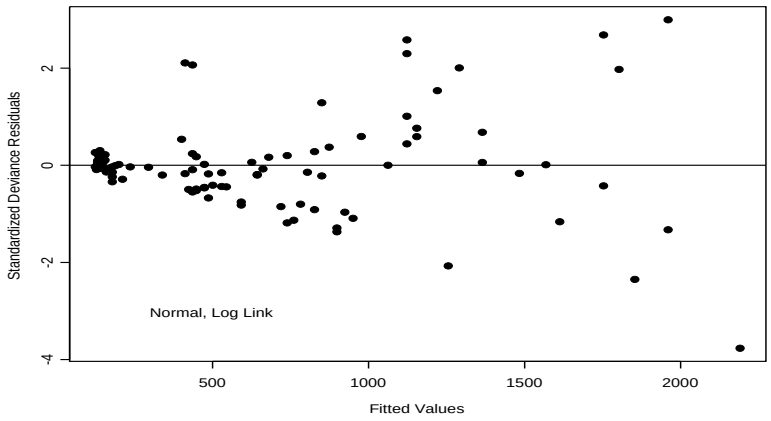


Figure 8.8: Standardized deviance residuals for model with normal random component and log link.

can compute variances for estimated expectations using the delta method. Specifically, what we need for any given covariate x_j (which may or may not

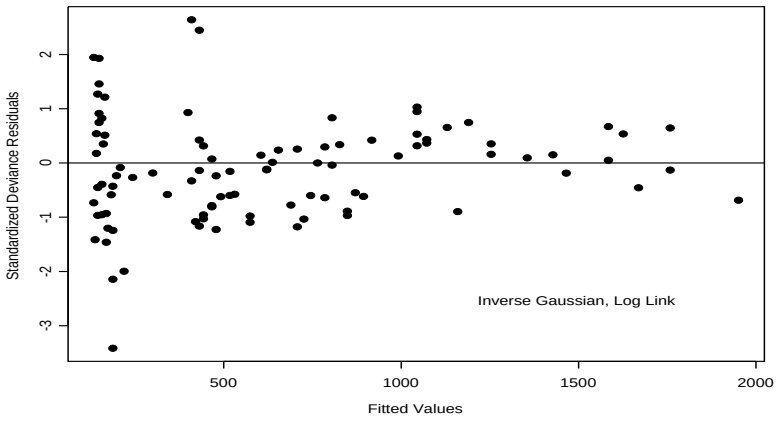


Figure 8.9: Standardized deviance residuals for model with Inverse Gaussian random component and log link.

be included in our set of data) is,

$$\begin{aligned} \mu_j &= \exp\{\beta_0 + \beta_1 x_j\} \\ \frac{\partial}{\partial \beta_0}\mu_j &= \mu_j \\ \frac{\partial}{\partial \beta_1}\mu_j &= \mu_j\, x_j. \end{aligned}$$

Then let $d_j^T = (\mu_j, \mu_j\, x_j)$ and

$$var\{\hat{\mu}_j\} = d_j^T I^{-1}(\beta_0, \beta_1, \phi)\, d_j,$$

where $I^{-1}(\beta_0, \beta_1, \phi)$ is the inverse information matrix for the regression parameters. We estimate this variance by substituting estimates for values of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ and ϕ, specifically,

$$\hat{var}\{\hat{\mu}_j\} = var\{\hat{\mu}_j\}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\phi=\hat{\phi}}$$

The result is Figure 8.10.

8.5.4 A Bayesian Analysis

We illustrate a Bayesian approach in analysis of a basic glm using the preferred model of a gamma random component and log link to relate cadmium concentration to length in Yellow Perch. We continue to use random variables $Y_1, \ldots, Y_n$ connected with the concentration of cadmium in individual fish. To formulate this model for a Bayesian analysis we take the data model as, for $\alpha > 0$ and $\beta_i > 0$, $i = 1, \ldots, n$,

$$f(\boldsymbol{y}|\alpha, \beta_i) = \prod_{i=1}^{n} \left(\frac{\beta_i^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} \right) \exp\left(\sum_{i=1}^{n} \beta_i y_i \right);\ \ y_i > 0, \tag{8.46}$$

$$\frac{\alpha}{\beta_i} = \exp(\gamma_0 + \gamma_1 x_i),$$

where x_i is the length of fish i in mm.

Notice that we have not written the data model in exponential dispersion family form nor the systematic model component with a link function in the usual way. This is because the explicit form of exponential dispersion families was used in a likelihood approach to analysis primarily to allow development of a unified algorithm for finding maximum likelihood estimates of the regression

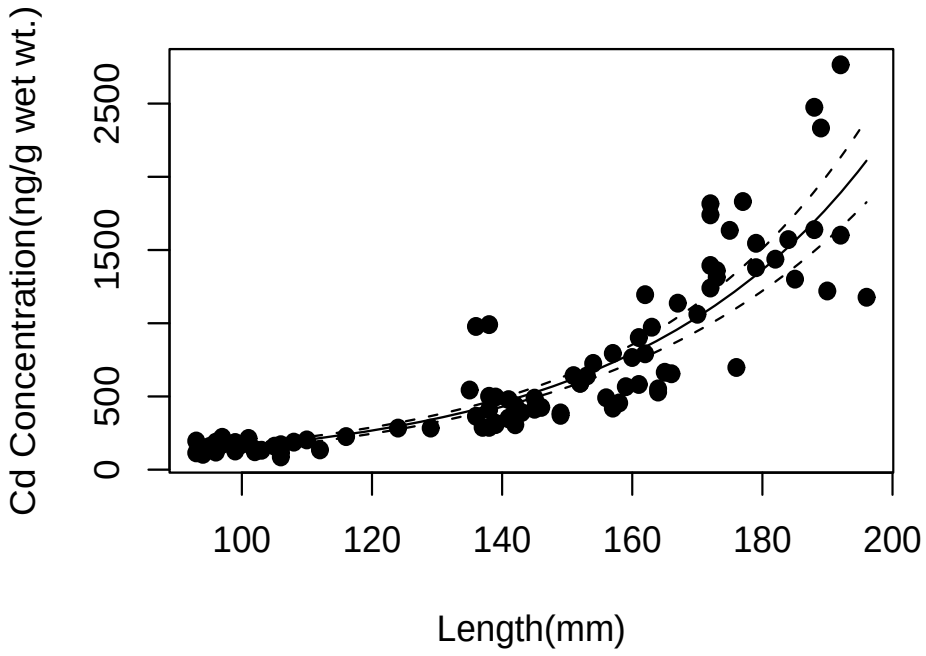


Figure 8.10: Scatterplot with estimated expectation function and pointwise 95% confidence band.

coefficients and to define deviance and deviance residuals. Here, we will find that writing computer code for conducting MCMC algorithms will be easier by keeping the original α, β parameterization of a gamma density. We have, however, drawn on a knowledge of exponential dispersion families to realize that keeping α fixed over observations and allowing β_i to vary results in variances proportional to the square of expected values, as previous analysis has indicated this is what we desire in our model.

To construct a joint prior on $(\gamma_0, \gamma_1, \alpha)$ we will assume our prior knowledge

about any one of these parameters does not depend on values of the others and use a product form, $\pi(\gamma_0, \gamma_1, \alpha) = \pi(\gamma_0)\pi(\gamma_1)\pi(\alpha)$. We assign both γ_0 and γ_1 normal priors as, for $p = 0, 1$, $V_p > 0$ and $-\infty < M_p < \infty$,

$$\pi(\gamma_p) \propto \exp\left[-\frac{1}{2V_p}(\gamma_p - M_p)^2\right]. \tag{8.47}$$

In these prior distributions, the V_p and M_p will be fixed numbers, which is why they are not given as parameters in the definition of the distributions. The prior for ϕ will be taken to be a proper uniform on the interval $(0, A)$. A sensitivity analysis will then be called for to determine the influence, if any, of the value A on the analysis.

An overall Gibbs Sampling algorithm with embedded Metropolis-Hastings steps for first $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ and then α was used to simulate values from the joint posterior. The conditional posterior of $\boldsymbol{\gamma}$ given α is,

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{y}, \alpha) &\propto \pi(\boldsymbol{\gamma}) f(\boldsymbol{y}|\boldsymbol{\gamma}, \alpha) \\ &\propto \exp\left[-\frac{1}{2V_p}(\gamma_p - M_p)^2\right] \prod_{i=1}^{n} \left(\frac{\beta_i^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1}\right) \exp\left(\sum_{i=1}^{n} \beta_i y_i\right); \;\; y_i > 0, \end{aligned}$$

where $\beta_i = \alpha / \exp(\gamma_0 + \gamma_1 x_i)$. The prior for α was taken as a proper uniform distribution on the interval $(0, A)$. The conditional posterior for α is then,

$$p(\alpha|\boldsymbol{y}, \boldsymbol{\gamma}) \propto \frac{1}{A} f(\boldsymbol{y}|\boldsymbol{\gamma}, \alpha).$$

Jump proposals for γ_0, γ_1, and α were all taken to be independent random walks with their own variances for tuning purposes.

Prior parameters were set to $M_0 = 5$, $M_1 = 0$, $V_0 = V_1 = 10$ and $A = 20$. Initial runs were used to tune the chains to achieve acceptance rates for jump proposals in reasonable ranges between about 0.20 and 0.50, and to ensure that the value of A was not influencing results. This resulted in random

walk variances of 0.01, 0.000001 and 4.0 for γ_0, γ_1 and α, respectively. Next, three chains were run, using starting values of $(2, 0.03, 10)$, $(0, 0.05, 5)$ and $(5, 0.01, 15)$. Trace plots of the first $2,000$ iterations are given in Figure 8.12, from which it appears that all three parameters are mixing by about 500 iterations. Autocorrelations for the parameters are shown in Figure **??** and it appears that the influence of the starting value dies off by about 100 to 150 iterations. To quantify this behavior, Gelman-Rubin scale reduction factors were computed at intervals of 100 iterations for γ_0 and γ_1. Values for the first 1500 iterations are given in Table 8.3. Scale reduction factors for both of these parameters become less than 1.1 by about iteration $1,100$, which suggests mixing is a bit slower than we would think from the visual assessment of trace plots and autocorrelations. Because the MC algorithm ran rapidly in real time, setting a burn-in period of $2,000$ iterations was not prohibitive. A final chain was run using a burn-in of $2,000$ and collection of the subsequent $25,000$ values. Summaries of the marginal posterior distributions for γ_0, γ_1 and α are given in Table 8.13. Credible interals (95%) were $(2.558, 3.185)$ for γ_0, $(0.021, 0.026)$ for γ_1 and $(6.691, 12.371)$ for α. Histograms of the marginal posteriors are shown in Figure **??**. Correlations between parameters in the Markov Chain were -0.978 for γ_0 and γ_1, -0.431 for γ_0 and α and 0.418 for γ_1 and α. Thus, the joint posterior for γ_0 and γ_1 will differ substantially from marginals.

MC Iteration	γ_0	γ_1
100	12.168	9.511
200	2.969	2.811
300	1.747	1.708
400	1.474	1.456
500	1.336	1.325
600	1.251	1.240
700	1.222	1.210
800	1.161	1.155
900	1.136	1.132
1000	1.119	1.115
1100	1.097	1.095
1200	1.089	1.087
1300	1.074	1.072
1400	1.057	1.056
1500	1.049	1.048

Table 8.3: Scale reduction factors for γ_0 and γ_1.

Parameter	Min	Q_1	Q_2	Mean	Q_3	Max
γ_0	2.310	2.743	2.852	2.856	2.964	3.401
γ_1	0.019	0.023	0.024	0.024	0.026	0.028
α	4.688	8.369	9.323	9.371	10.298	15.134

Table 8.4: Posterior summaries for regression of Cd concentration of fish length.

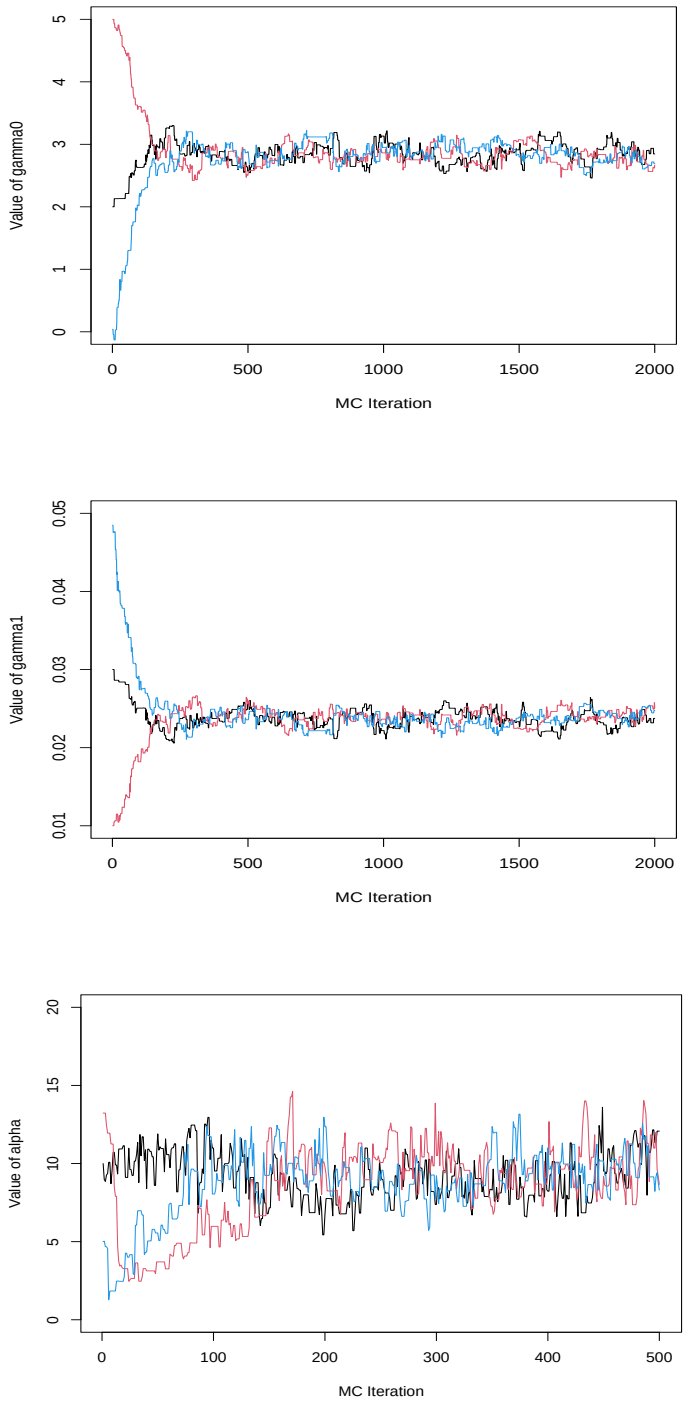


Figure 8.11: Trace plots for regression of Cd concentration on fish length.

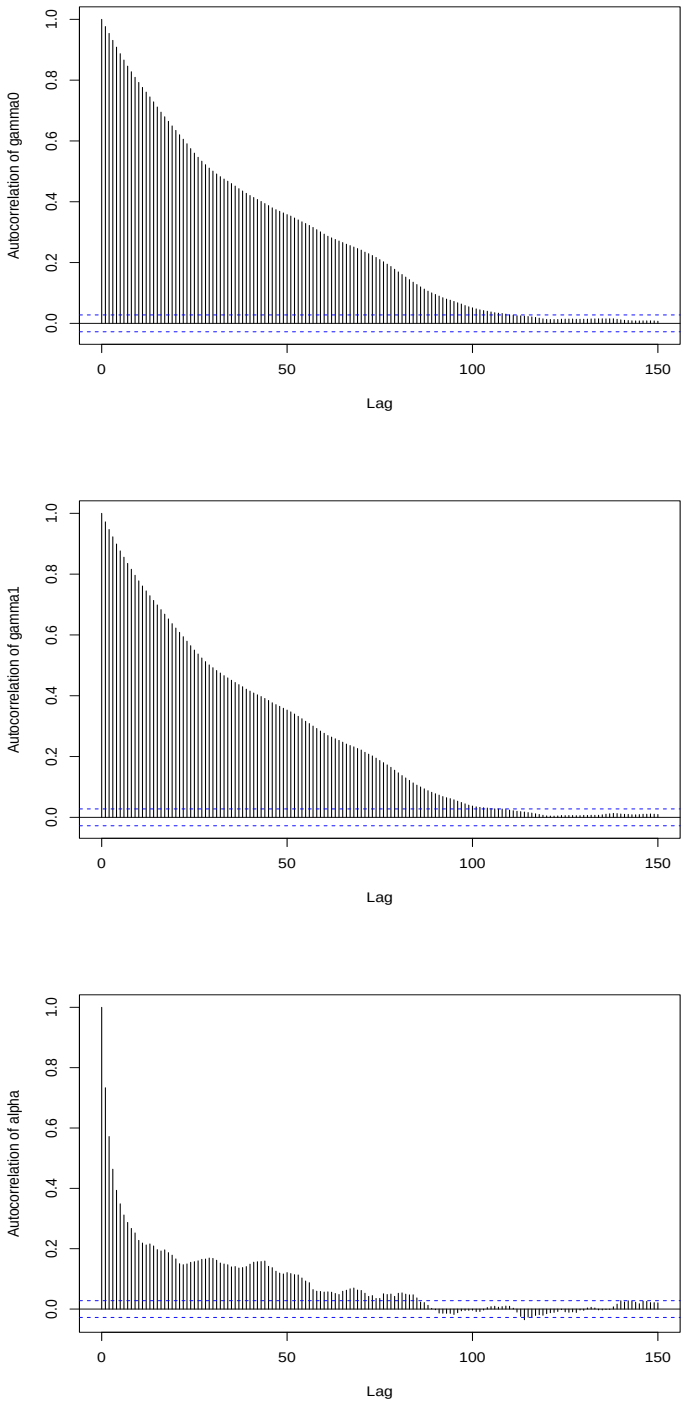


Figure 8.12: MC autocorrelation plots for regression of Cd concentration on fish length.

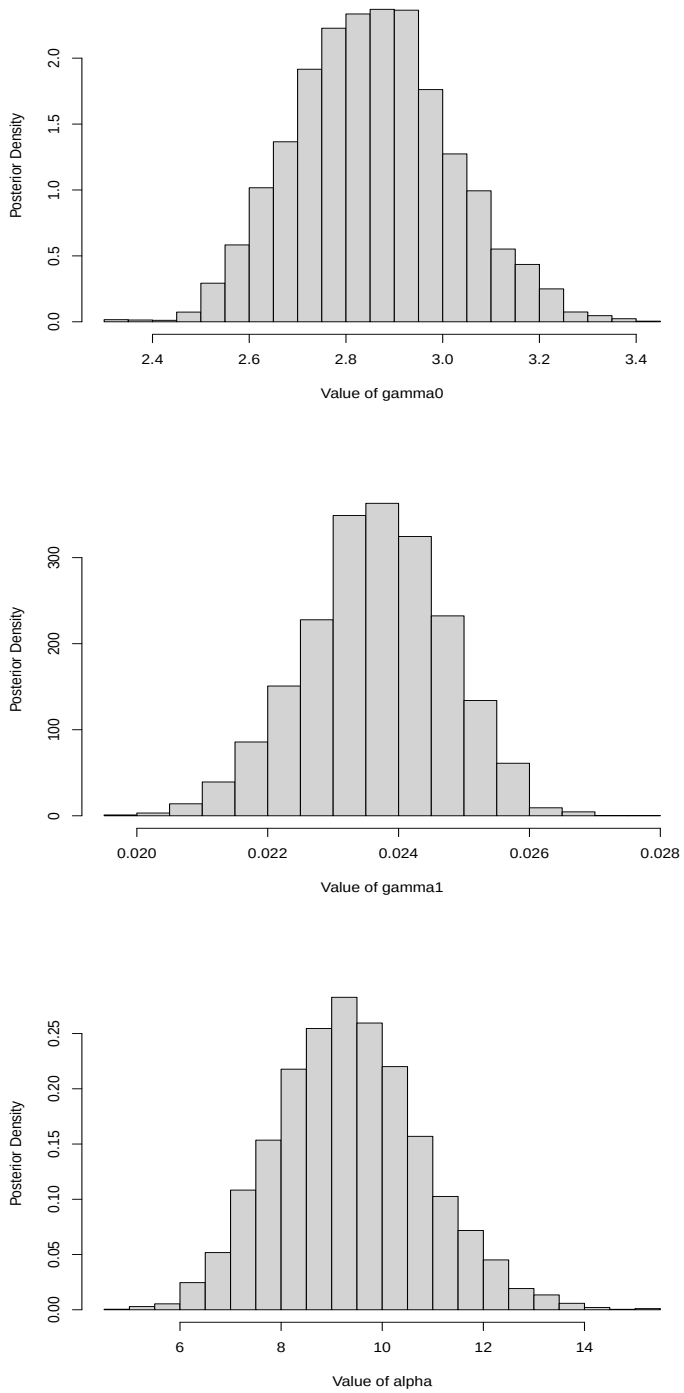


Figure 8.13: Posterior histograms from $2,500$ MCMC iterations.

Although there is not necessarily any technical or philosophical reason that results from the previous likelihood analysis and results from this Bayesian analysis should be in agreement, there is a general notion that for most problems any reasonable analysis should not be wildly different. In comparing the results of Table 8.13 with those of Table 8.1 we can see a reasonable amount of similarity. The posterior mean of γ_0, 2.86 is a bit larger than the mle of 2.33 while that of γ_1, 0.024 is a bit smaller than the mle of 0.027. The posterior mean of α, 9.37 is also roughly comparable to the moment-based estimate of ϕ, 9.03 used in the likelihood analysis; note that these two parameters are the same. The widths of credible and confidence intervals also compare favorably, being 0.627 (credible) and 0.616 (confidence) for γ_0 and 0.005 (credible) versus 0.004 (confidence) for γ_1. Because $\hat{\phi}$ was not estimated using maximum likelihood, no interval is available from the likelihood analysis. Approximate sampling distributions for γ_0 and γ_1 in the likelihood analysis are both normal. The marginal posterior distributions for these parameters in Figure 8.13 are unimodal, with that for γ_0 suggesting just a bit of right skewness and that for γ_1 a slight left skewness. The likelihood analysis does not allow estimation of a sampling distribution for ϕ. The posterior mean expectation function and 95% confidence bands were produced from the MCMC output as described in Chapter 8.6.4. A scatterplot of cadmium concentration versus length for Yellow Perch in Little Rock Lake is reproduced in Figure **??** with these posterior quantities overlaid, and can be compared to their likelihood counterparts as shown in Figure 8.10.

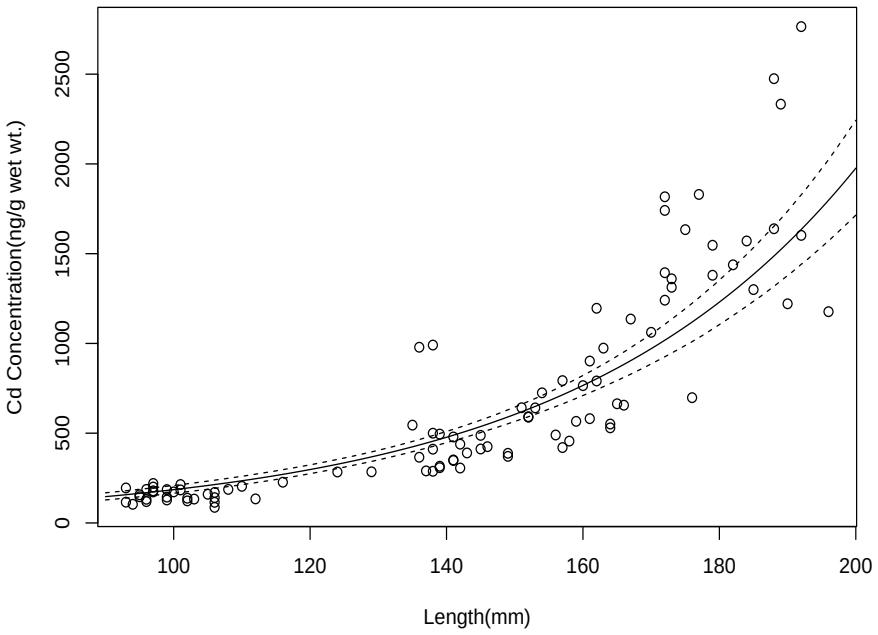


Figure 8.14: Posterior expectation function and 95% credible band for regression of cadmium concentration on length of fish.

Chapter 9

Additive Error Regression Models

9.1 Signal Plus Noise as a Statistical Model

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 $i = 1, \ldots, n$,

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i, \tag{9.1}$$

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 (9.1) 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, $E(Y_i) = g(\boldsymbol{x}_i, \boldsymbol{\beta})$.

The modeling task with an additive error specifications largely centers on two issues, appropriate specification of the function g, and 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 or through what is essentially an arbitrary selection of some function based on examination of the data.

Notice that the general form (9.1) 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 (9.1) 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,

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

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 what follows. 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 $g(\boldsymbol{x}_i, \boldsymbol{\beta})$ in (9.1) 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. It is this

portion of the model formulation that renders additive error models a viable option for many situations. By far-and-away, the most common location-scale family chosen for specification of the error distribution is the normal. We will briefly consider here four situations for modeling the variance of the error terms in (9.1); constant variance, variance models with known parameters, variance models with unknown parameters, and what are called *transform both sides* models.

9.2 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 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 testing scientific theory. Bates and Watts (1988) present any number of examples of such situations.

Example 9.1

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 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 9.1. 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

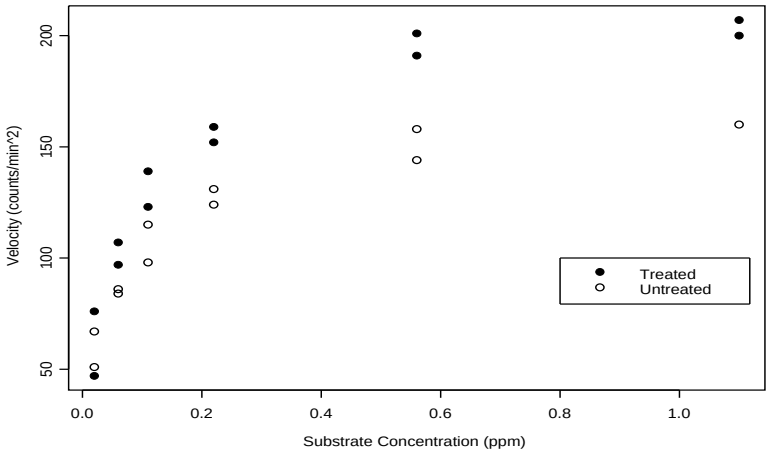


Figure 9.1: Scatterplot of data on the velocity of an enzyme reaction on substrate treated with Puromycin and untreated substrate.

or untreated) the Michaelis-Menten equation can be expressed as in model (9.1) with, for $\beta_1 > 0$ and $\beta_2 > 0$,

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

where x_i represents the substrate concentration for observation i.

Bates and Watts (1988, p. 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 9.1 if we use this transformation is evident from Figure 9.2.

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.

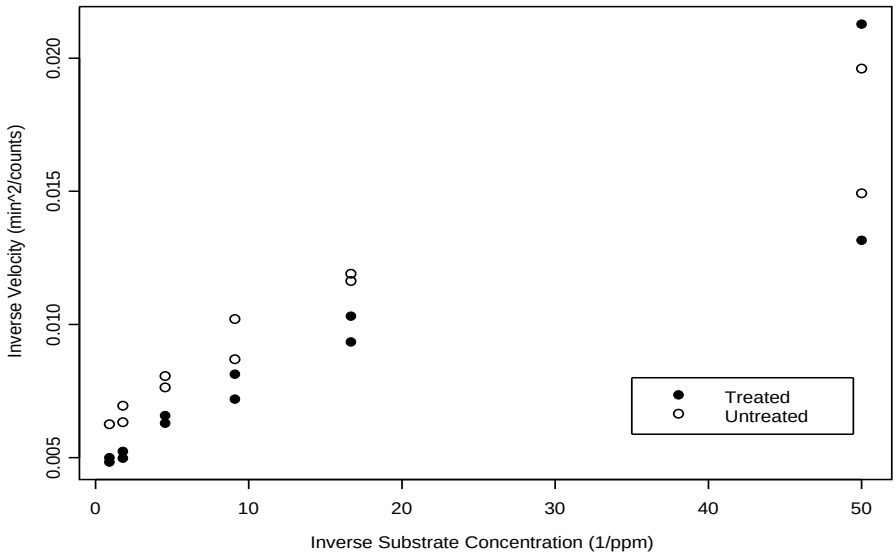


Figure 9.2: Scatterplot of transformed data for Puromycin example using reciprocal expression of both covariate and response variables.

9.3 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 (9.1) for which the error terms have zero expectation and constant variance.

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 $\boldsymbol{\beta}$ 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 although it is transformably linear.

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 (9.1) with a single type of covariate x_i measured on a ratio/interval scale. The quantities involved in this model, other than the parameters $\boldsymbol{\beta}$ 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 formed by $\beta_0 + \beta_1 x_i$ in the $2-$dimensional space of a scatterplot. 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 in $n-$dimensional space. 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 intrinsic and parameter effects curvatures can be gained by considering two aspects of a solution locus $\boldsymbol{g}$.

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

2. Secondly, $\boldsymbol{g}$ maps points from the $p-$dimensional space of $\boldsymbol{\beta}$ to the $n-$dimensional space of $(\boldsymbol{x}, \boldsymbol{y})$. 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.

9.4 Models with Known Variance Parameters

In both this section and the next we will consider additive error models for which the assumption of constant error variance is relaxed. The result is 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*. At first, this may seem an artificial device, similar to specifying a normal model with known variance, but that is not really the case. As will become clear shortly, 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 section 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 $\boldsymbol{\eta}$.

9.4.1 Known Weights

The simplest extension of model (9.1) 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,

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + (\sigma/\sqrt{w_i})\,\epsilon_i, \tag{9.2}$$

where, as in model (9.1), 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 (9.2) 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 9.2

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. Suppose 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 (9.2) 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 9.2, in which we have known weights, is a quite simple one. It is clear that model (9.2) could be easily re-expressed as

$$Y_i^* = g^*(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma\,\epsilon_i,$$

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 response as would, in fact, be true for any linear transformation applied to the random variables $\{Y_i : i = 1, \ldots, n\}$.

9.4.2 Weights as Specified Functions of Means

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 (9.2), except in which we replace $(\sigma/w_i^{1/2})$ by σ_j where j indexes distinct levels of the covariate. Carroll and Ruppert (1988, p. 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 Ruppert, 1988, p. 87). Why would we suggest something like model (9.2) 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 (9.2) for the hypothetical situation of Example 9.2, 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 Example 9.2, we used a model that specified variances differing only through the factor of sample sizes used to calculate averages. This amounts to modeling the variances in a manner analogous to modeling expected values in which variances are different at different covariate levels, but there are a small

number (in Example 9.2, one) of parameters to estimate. It is true that while we sometimes have scientific knowledge available to help with formulating a model for means, this is rarely the case 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 is that modeling variances as functions of means is often a useful technique. The type of model that results may be written as, for $i = 1, \ldots, n$,

$$Y_i = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta)\, \epsilon_i, \tag{9.3}$$

where, as before, $\epsilon_i \sim iid\, F$ with $E(\epsilon_i) = 0$ and, almost always, F is the standard normal distribution. If the $\boldsymbol{x}_i$; $i = 1, \ldots, n$ are considered known constants and we assume that the dependence of $g_2(\cdot)$ on $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ is only through the way these quantities are combined in the function $g_1(\cdot)$, then we can also write model (9.3) as,

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma\, g(\mu_i(\boldsymbol{\beta}), \theta)\, \epsilon_i, \tag{9.4}$$

with assumptions on $\{\epsilon_i : i = 1, \ldots, n\}$ as before. Here, $g_1(\boldsymbol{x}_i, \boldsymbol{\beta})$ in (9.3) has been replaced with $\mu_i(\boldsymbol{\eta})$ and $g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta)$ has been replaced with $g(\mu_i(\boldsymbol{\beta}), \theta)$. What renders model (9.4) appropriate under the topic of this section (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 *formulation* 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 formulation in linear regression analyses. If we take the $\sqrt{w_i}$ from model (9.2) to be given by $1/g(\mu_i(\boldsymbol{\beta}), \theta)$ in (9.4), then this model can be considered to be in the form of a weighted regression, but one in which

the weights must be estimated, since $\boldsymbol{\beta}$ is unknown. On the other hand, the situation is simplified by taking the additional parameter θ as a known value in the model.

Probably the most common model formulation of the type (9.4) is called a *power of the mean* model,

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma\{\mu_i(\boldsymbol{\beta})\}^{\theta}\,\epsilon_i. \tag{9.5}$$

In this case, we have that

$$var(Y_i) = \sigma^2\,\{\mu_i(\boldsymbol{\beta})\}^{2\theta},$$

or,

$$2\log\left[\{var(Y_i)\}^{1/2}\right] = 2\left[\log(\sigma) + \theta\,\log\{\mu_i(\boldsymbol{\beta})\}\right],$$

or,

$$\log\left[\{var(Y_i)\}^{1/2}\right] = \log(\sigma) + \theta\,\log\{\mu_i(\boldsymbol{\beta})\}, \tag{9.6}$$

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,

$$\left(\frac{d}{d\mu_i}h(\mu_i)\right) \propto \mu_i^{-\theta}, \tag{9.7}$$

any constant of proportionality being absorbed into σ^2. This relation will hold if

$$h(\mu_i) \propto \mu_i^{1-\theta}. \tag{9.8}$$

Now, when $h(\cdot)$ of (9.8) 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, (9.6) 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 have already made use of this Box-Cox plot in selecting a suitable random component for a basic generalized linear model.

We are not advocating here the indiscriminate use of power transformations to produce constant variance but, rather, the use of model (9.5) 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 (9.5) since this is where you actually started (whether that is made clear in courses on applied regression methods or not).

Example 9.3

This example is taken from Trumbo (2002). Major airlines 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 9.3. 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

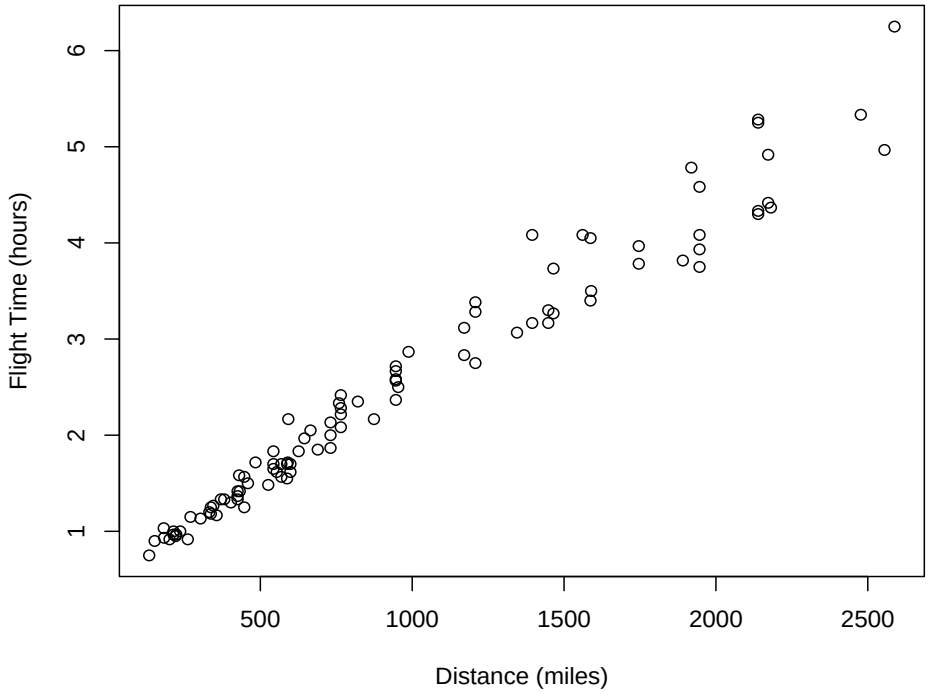


Figure 9.3: Scatterplot of travel time versus distance for a sample of 98 flights conducted by Delta airlines in 1994.

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 9.4, which exhibits the unequal variances noticed in the scatterplot. We might consider, for this situation, model (9.5) with $\mu_i(\beta) \equiv \beta_0 + \beta_1 x_i$, where 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_0, β_1 and σ^2.

Alternatively, we might consider using a power transformation to try and

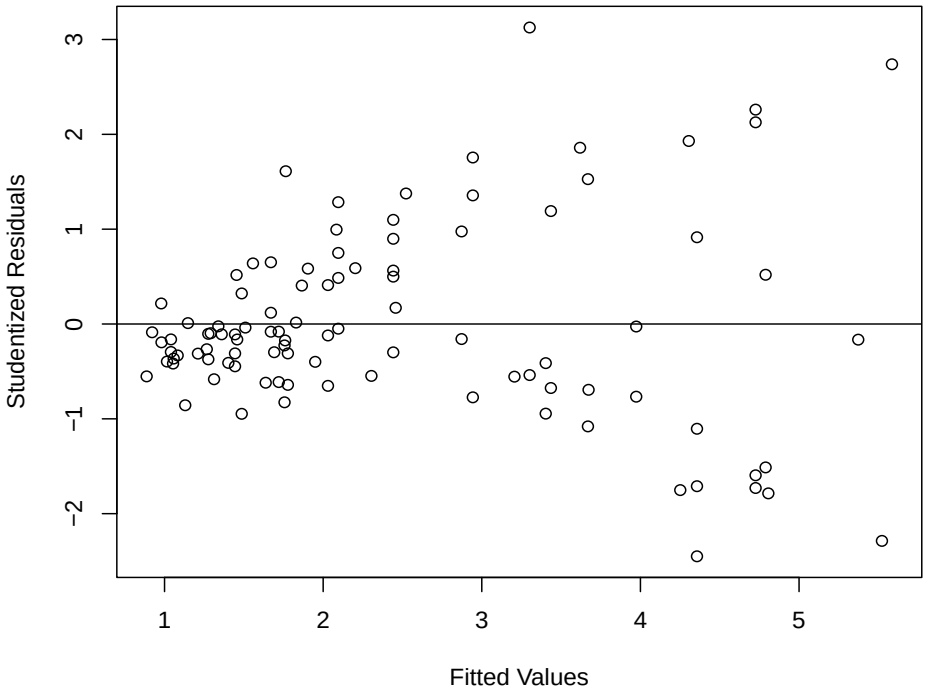


Figure 9.4: Plot of studentized residuals for an ordinary least squares fit to the data of Figure 9.3.

stabilize the variance. We are not advocating a transformation here, but it may be useful to see what occurs in this example if we take that approach, since it would correspond to advice in many courses on regression. We might consider a power transformation, which is sometimes also called a Box-Cox transormation. While there are a number of observations for some of the 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, and there are two obvious ways to bin those values, by making bins of equal length or by making bins with equal numbers of observations. Here, the density of points is somewhat greater for smaller values of distance than for the larger values and we will choose to create bins with equal numbers of observations; using 12 bins gives 8 observations per bin, dropping only 2 of the points. The resulting Box-Cox plot is given in Figure 9.5 This plot is a diagnostic or exploratory

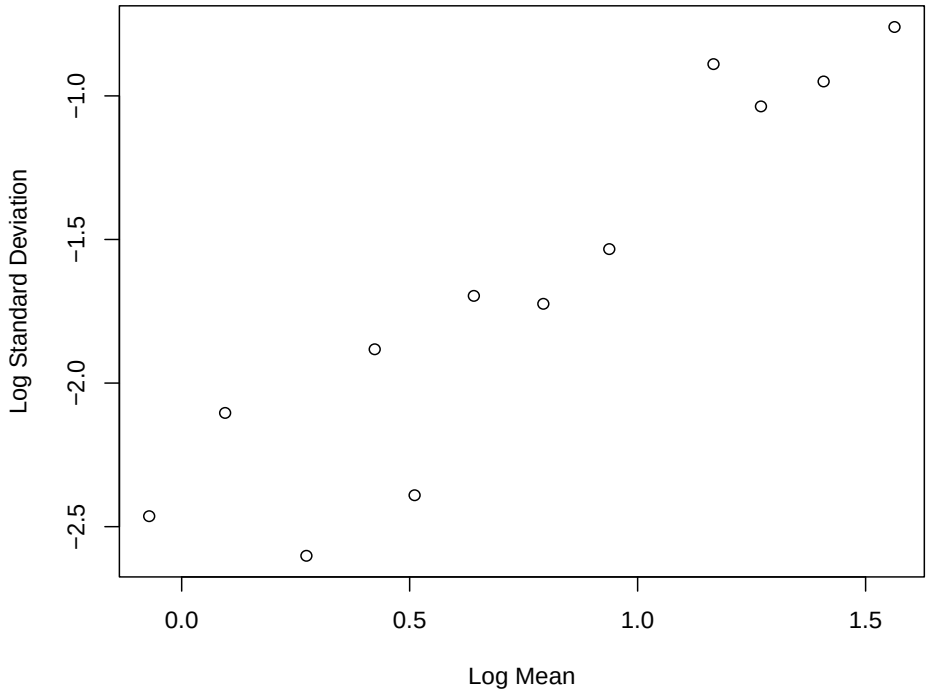


Figure 9.5: Box-Cox transformation plot from using binned data on flight times.

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 9.5 is 1.14, which would suggest a reciprocal square root transformation $Y_i^* = 1/\sqrt{(Y_i)}$. A scatterplot of the transformed responses versus distance in shown in Figure 9.6. Examination of this plot shows that there has been some stabilization of the variances, but the transformation has also taken what appeared to be a relatively straight line relation (Figure 9.3) and changed it into a nonlinear one. Mathematically, the reason for this is obvious. If Y is linear in x, then

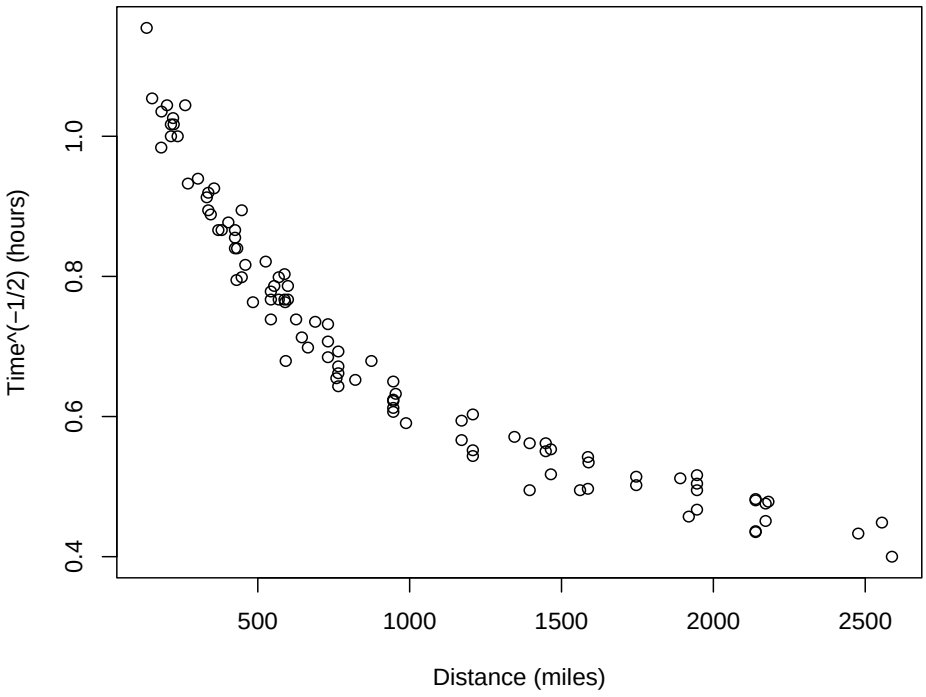


Figure 9.6: Scatterplot for reciprocal root transformed flight times.

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 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

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \sigma\, \epsilon_i; \qquad i = 1, \ldots, n,$$

where $\epsilon_i \sim iid\, N(0, 1)$, then,

$$E(Y_i) = \exp\left\{\beta_0 + \beta_1 x_i + \sigma^2/2\right\}.$$

The advantage of model (9.5) fit to the original responses over a constant variance model fit to the transformed responses 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 (9.5) with $\mu_i(\beta) = \beta_0 + \beta_1\, x_i$ and $\theta = 1.5$.

9.4.3 An Aside on Transformations

In an number of books on linear regression, transformation of response variables is presented as a standard approach to mitigating difficulties encountered by violation of assumptions, particularly that of constant variances. It is often assumed that desirable effects occur to both expectation functions and variances (e.g., Neter et al., 1989) and transformations do no harm to the fundamental model structure as additive error. As we have seen in both Example 9.1 and Example 9.3, this may not be the case. In Example 9.1, transformation to produce a linear expectation function created non-constant variances. In

Example 9.3, transformations to stabilize variances also turned a straight line expectation function into nonlinear ones.

Less obvious than the effects of transformation on expectation functions and variances, but equally important, 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). For example, suppose that $\log(Y_i) = \beta_0 + \beta_1 x_i + \sigma\epsilon_i$ with $\epsilon_i \sim iid\mathrm{N}(0,1)$. Then $Y_i = \exp(\beta_0)\exp(\beta_1 x_i)\exp(\sigma\epsilon_i)$ which is not an additive error model. Thus, the same model structure of a simple linear regression with additive error cannot hold for both $\log(Y_i)$ and Y_i.

Interpretation of a model relative to a set of response variables $\{Y_i : i = 1,\ldots,n\}$ is not necessarily straightforward if an additive error model was fit to transformed response variables $\{Y_i^* : i = 1,\ldots,n\}$. One of the simplest examples, is the same as that just given in which, for $i = 1,\ldots,n$, $\log(Y_i) = \beta_0 + \beta_1 x_i + \sigma\,\epsilon_i$ with the ϵ_i having standard normal distributions. Under this model,

$$\begin{aligned} E(Y_i) &= \exp\left\{\beta_0 + \beta_1 x_i + \sigma^2/2\right\} \\ var(Y_i) &= \exp\left\{2(\beta_0 + \beta_1 x_i) + \sigma^2\right\}\left\{\exp(\sigma^2) - 1\right\}. \end{aligned}$$

The naive mistake is to assume that $E(Y_i) = \exp[E\{\log(Y_i)\}] = \exp(\beta_0 + \beta_1 x_i)$ which can produce misleading conclusions. Interpretation of confidence intervals computed on the transformed scale is even more involved.

In general, scientists usually measure quantities using scales that have meaning to them. Interpreting results of an analysis performed on a transformed scale can produce quite misleading inferences if care is not taken to appropriately express those results on the original scale. The one situation in

which transformation of response variables does not raise any concerns is if the objective of an analysis is only comparison of the order of group means. In this case any monotone transformation will produce inferences that translate directly between original and transformed scales. Thus, taking logarithms or square roots of response variables to produce symmetry, for example, and conducting a test for equality of group means allows us to conclude that $\mu_2 > \mu_1$ without concern about unwanted effects of the transformation or which scale we are operating on. But making an inferential statement about how much greater than μ_1 μ_2 might be is another question and we return to the need to use caution about the effects of transformation.

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? The answer to this question probably depends at least in part on computational history and in part on institutional inertia. Estimation with nonlinear models, although not difficult today, does require iterative numerical techniques. Transformations have been one way to replace a model with a nonlinear expectation function or with non-constant variances with a model for which a linear expectation function and/or constant variance seems more reasonable that for responses in the original scale. Estimation by ordinary least squares is then possible and computations are easy. Once such methods made it into the standard set of material taught to scientists and applied statisticians there has been resistance to removing them.

There is also, however, more than just computation and history that underlies the tendency of statisticians to hang onto the basic ideas of additive error constant variance models with considerable fervor. 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 onto parts of this body of methodology even when it is clear that not all of it is appropriate for a given problem. What might be termed 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.

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.

A healthy reluctance to use transformations on response variables does not carry over to quantities that are not random. We have already discussed transformation of parameters and the potential benefits of reparameterization in some situations. Transformation of covariate values in a regression fall into a similar category. Covariates are considered to constitute given values that are fixed in an analysis. If the covariate or covarites are not fixed by design, which is probably the typical case, we conduct the entire analysis conditional on the values observed. If covariates are to be considered random, then their distribution enters into the analysis and different models are typically needed.

9.5 Unknown Variance Parameters

We turn now to models very similar to that of expression (9.3) but for which we generalize the variance model. Specifically, in this section we consider models of the form,

$$Y_i = g_1(\boldsymbol{x}_i\,\boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}_i, \boldsymbol{\theta})\,\epsilon_i, \tag{9.9}$$

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 (9.9) extends model (9.3) in that the function g_2 includes $\boldsymbol{z}_i$, which may be a part of $\boldsymbol{x}_i$ or may be other covariates that are believed to affect the variance but not the mean, and the now possibly vector-valued parameter $\boldsymbol{\theta}$ is no longer assumed known. Sometimes, we we can impose a restriction similar to that used in moving from expression (9.3) to expression (9.4) by taking $\mu_i(\boldsymbol{\beta}) = g(\boldsymbol{x}_i, \boldsymbol{\beta})$ and writing,

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma\, g(\mu_i(\boldsymbol{\beta}),\, \boldsymbol{z}_i,\, \boldsymbol{\theta})\,\epsilon_i, \tag{9.10}$$

with the same assumptions on the ϵ_i as in model (9.9). Models (9.9) and (9.10) allow the variance to depend on the covariates, possibly only through the mean, but no longer assume that $\boldsymbol{\theta}$ is a part of model formulation. Rather, $\boldsymbol{\theta}$ is to be estimated along with the other parameters $\boldsymbol{\beta}$ and σ^2. The inclusion of additional covariates $\boldsymbol{z}_i$ in the model for variances could also have been made to (9.3) and (9.4) but, as noted previously, the power of the mean model is dominant among those for which θ becomes a part of model specification; thus, there seemed little motivation to include $\boldsymbol{z}_i$ in the formulations of (9.3) or (9.4).

A number of possible forms (not meant to be exhaustive, by any means) for g are given in Carroll and Ruppert (1988), and I have extended the suggestions

below with a few additional 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 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 to either cover all of the possibilities or to constitute common models. The fact is that we are much less advanced in our modeling of response variances than in modeling response means.

Example 9.4

Foresters and environmental scientists are interested in estimating the volume of trees (obvious from a commercial standpoint, but also an indicator of biomass production in a forest ecosystem). 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 31 Black Cherry 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 9.7, 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.

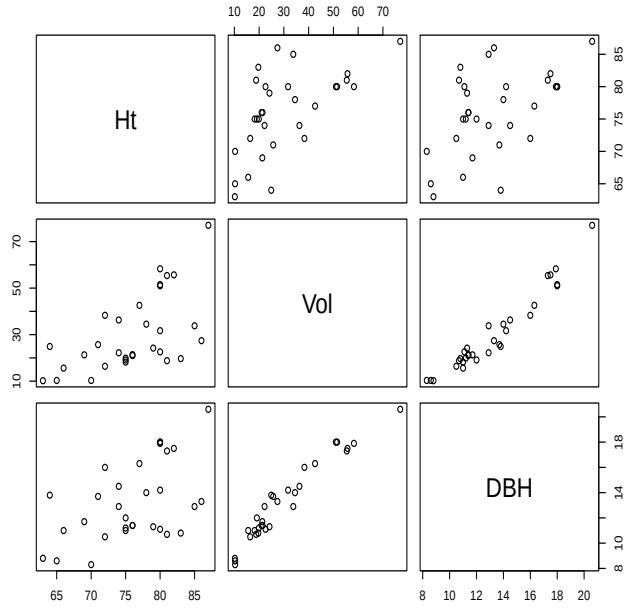


Figure 9.7: Scatterplot matrix of volume, height, and DBH for Black Cherry trees.

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 9.8, while the second (on height) results in the analogous Figure 9.9. An examination of these plots reveals the following:

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 9.8, and this appears to be due to more than the 3 small trees in the scatterplot.

3. The relation between volume and height is weak, as we already knew, and the variances of volume clearly increase with (estimated) volume in this regression.

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 given in Table 9.1 (which has been arranged so that parameter estimates in the same column are comparable).

Table 9.1 largely reflects what has already been seen in the plots of Figures 9.7 through 9.9. It is perhaps surprising that what is certainly a weak linear

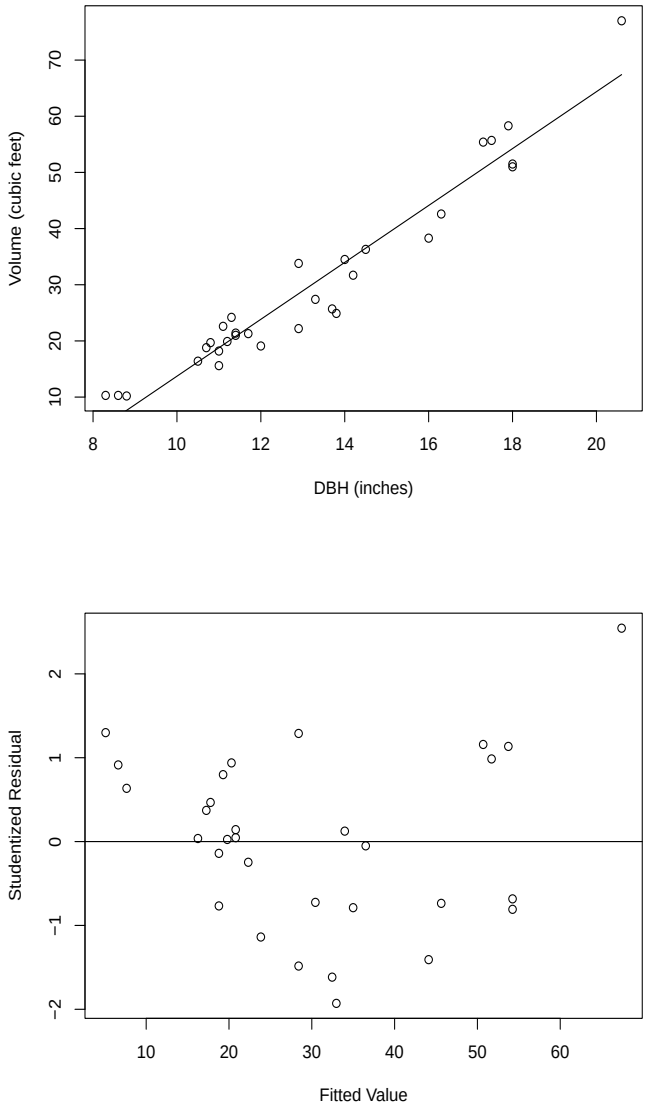


Figure 9.8: Regression of volume on DBH (upper) and studentized residuals (lower).

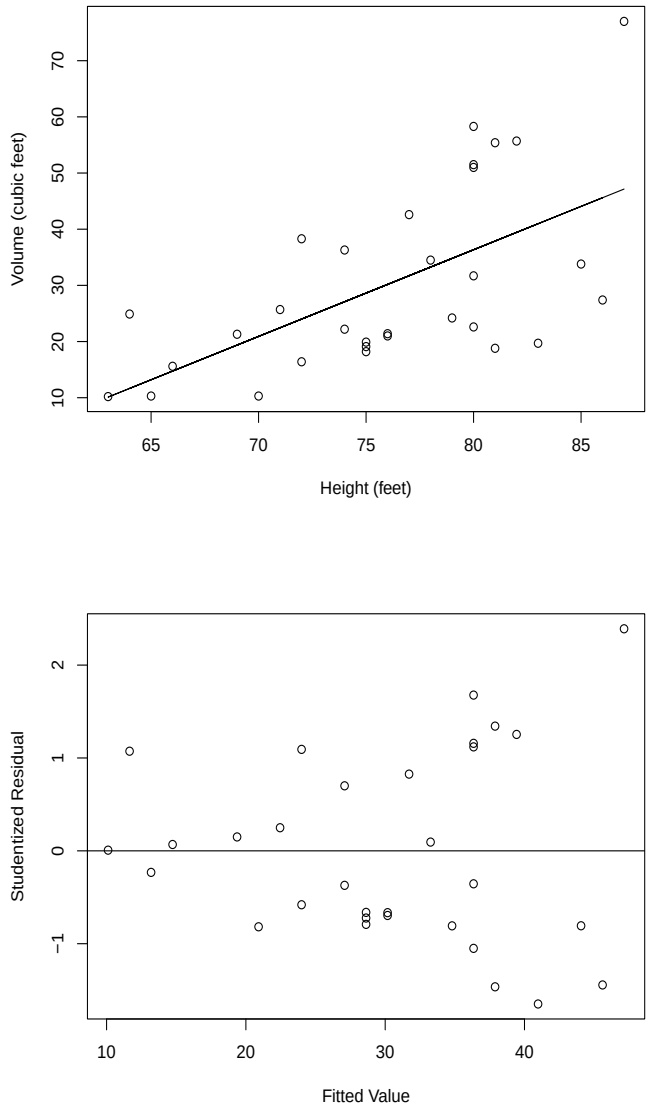


Figure 9.9: Regression of volume on height (upper) and studentized residuals (lower).

	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

Table 9.1: Estimated regression parameters for models of tree volume on DBH and height.

relation between height and DBH (see Figure 9.7) has such a great impact on the estimated value of the regression coefficient associated with height (β_2 in the table) and such a small impact on the coefficient of determination (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. A residual plot for the multiple regression is presented in Figure 9.10.

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 against fitted values 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 Figure 9.11. Figure 9.11 reinforces the suggestion that the mean function is not correctly specified in terms of DBH, and the same $U-$shaped residual pattern is hinted at for height, although in the absence of previous evidence one would be reluctant to see much in this plot.

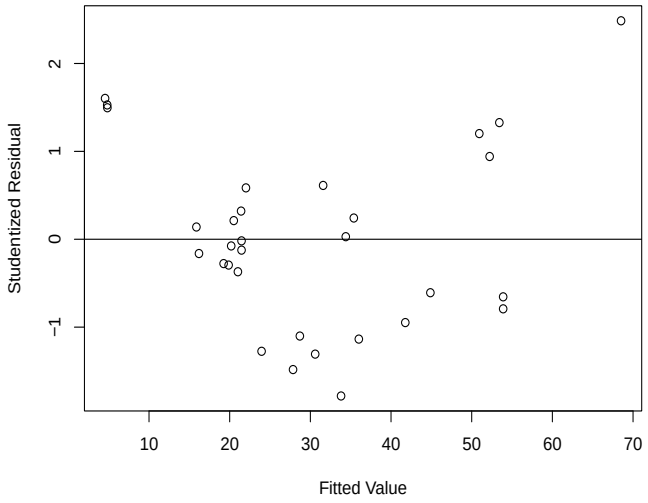


Figure 9.10: Studentized residuals for the regression of volume on DBH and height.

Where does this leave us? We have a linear multiple regression model that appears quite good for describing the pattern of data, with an R^2 value of nearly 0.95. 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 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

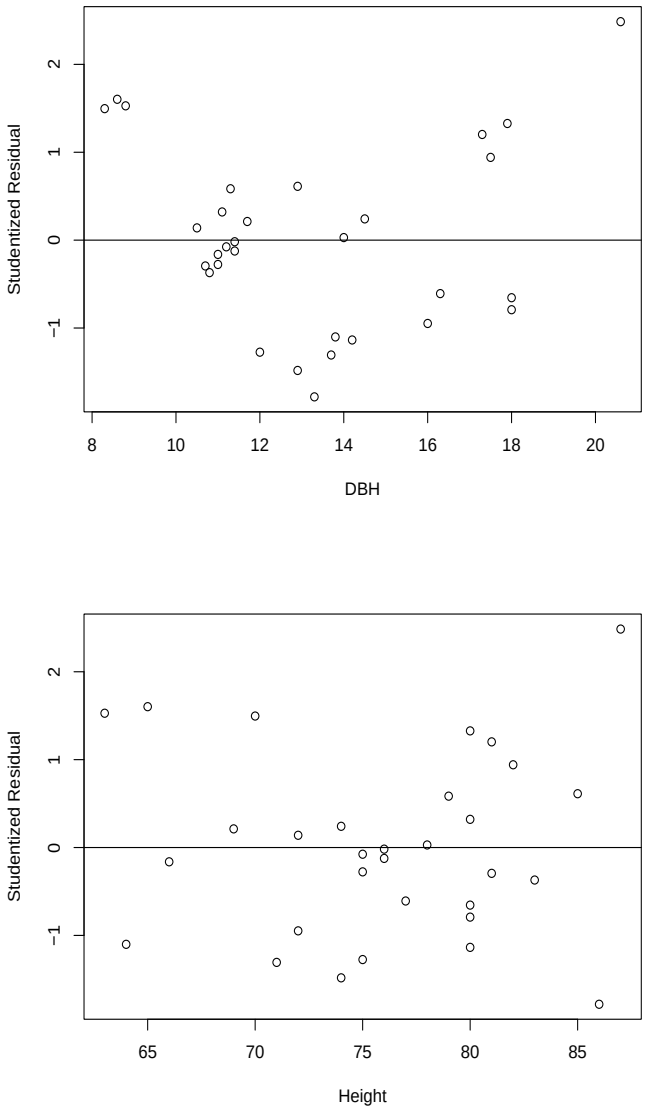


Figure 9.11: Studentized residuals from the regression of volume on DBH and height against DBH (upper) and height (lower).

volumes.

Is there a simple conceptualization of this problem other than "adding things together"? The problem essentially deals with quantities that reflect basic geometry relative to trees, a basic concept of which might be a cylinder, $V = \pi r^2 H$. To make use of this idea for an expectation function for 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,

$$E(Y_i) = \beta_0 + \beta_1 \left\{ 2\pi (x_{1,i}/24)^2 \, x_{2,i} \right\}, \tag{9.11}$$

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 (9.11) as a expectation function we can simply fit a constant variance regression using ordinary least squares,

$$Y_i = \beta_0 + \beta_1 \, \phi(\boldsymbol{x}_i) + \sigma \, \epsilon_i, \tag{9.12}$$

where, for $i = 1, \ldots, n$, $\epsilon_i \sim iid\, F$ with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$. The results are shown in Figure 9.12. Estimated values for the regression model (9.12) are $\hat{\beta}_0 = -0.298$, $\hat{\beta}_1 = 0.195$, $\hat{\sigma}^2 = 6.2150$, and $R^2 = 0.9778$. Relative to the regressions in Table 9.1, 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 9.12 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 (9.12) against height ($x_{2,i}$) results in the plot of Figure 9.13. This plot suggests that

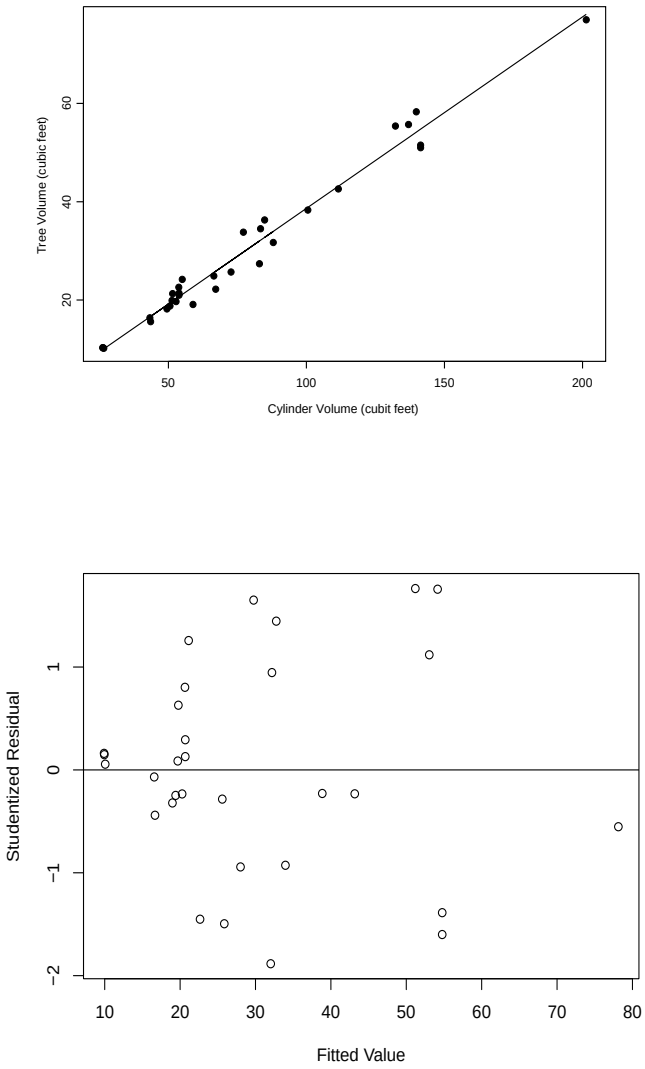


Figure 9.12: Results for regression of volume against cylinder.

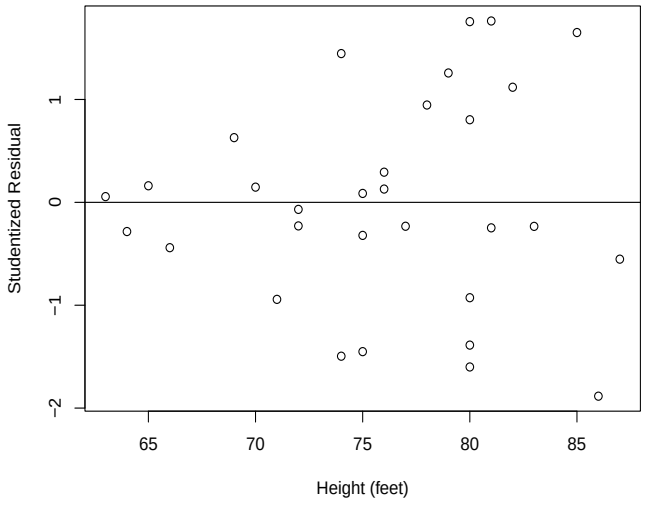


Figure 9.13: Studentized residuals for the regression of volume on cylinder plotted against values of height.

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 9.12 and is not presented.

Putting these preliminary analyses of tree geometry together, we should be willing to entertain a model of the general form (9.10), 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

$$\begin{aligned} \sigma \, g(z_i, \, \theta) &= \theta_0 + \theta_1 \, z_i \\ &\text{and} \\ \sigma \, g(z_i, \, \theta) &= \sigma \exp\{\theta_0 + \theta_1 \, z_i\} \end{aligned} \tag{9.13}$$

9.6 Transform Both Sides Models

We close discussion of additive error models with a brief mention of one additional modeling idea, promoted by **?**. 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 Ruppert, 1988, pp. 199-121) apparently **?** were the first to suggest this methodology as a general modeling strategy.

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 9.4 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)?

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 usually symmetric. The basic idea is that we begin with a model of the form,

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \text{error},$$

where $g(\cdot)$ has scientific meaning or is a pleasing empirical form, but for which the error term does not led itself to modeling through a location-scale specification. To mitigate 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,

$$h(Y_i, \lambda) = h\{g(\boldsymbol{x}_i, \boldsymbol{\beta}), \lambda\} + \sigma\, \epsilon_i, \tag{9.14}$$

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 (9.14) can be extended to include additional modeling of variance structure as,

$$h(Y_i, \lambda) = h\{g_1(\boldsymbol{x}_i, \boldsymbol{\beta}), \lambda\} + \sigma\, g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, z_i, \theta)\, \epsilon_i, \tag{9.15}$$

where assumptions on the error terms ϵ_i are the same as for (9.14). In the same way that we moved from model (9.3) to model (9.4) and model (9.9) to model (9.10), if the variance portion of (9.15) depends on β only through the expectation function g_1, we may write

$$h(Y_i, \lambda) = h\{\mu_i(\boldsymbol{\beta}), \lambda\} + \sigma\, g(\mu_i(\boldsymbol{\beta}), z_i, \theta)\, \epsilon_i. \tag{9.16}$$

Now, the models given in (9.15) and its reduced version in (9.16) 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 (9.15) or g in (9.16). Including both of these in a model of the form (9.16) would result in,

$$Y_i^{\lambda} = \{\mu_i(\boldsymbol{\beta})\}^{\lambda} + \sigma\, \{\mu_i(\boldsymbol{\beta})\}^{\theta}\, \epsilon_i.$$

This model would 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 $\mu_i(\boldsymbol{\beta})$

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 Ruppert (1988, p. 151) indicate, "More research is needed on predication intervals based on transformation models."

9.7 Non-Bayesian Analysis

9.7.1 Least Squares

The use of least squares estimation is traditional for additive error models. Least squares was covered from a general standpoint in Chapter 6. Here, we give additional details for some of the model forms covered previously in this chapter.

Consider first an additive error model such as (9.3) in which we take the expectation function to be linear,

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma g(\boldsymbol{x}_i^T\boldsymbol{\beta},\, \theta)\epsilon_i, \tag{9.17}$$

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 formulation). Now, model (9.17) is quite similar to model (6.6) 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 (6.6) 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 this suggests a weighted least squares solution of the form

$$\hat{\boldsymbol{\beta}}^{(1)} = (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(0)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(0)})\boldsymbol{Y}, \tag{9.18}$$

where $\boldsymbol{W}(\boldsymbol{\beta}^{(0)})$ is an $n \times n$ diagonal matrix with elements

$$w_i(\boldsymbol{\beta}^{(0)}) = \frac{1}{g^2(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(0)},\, \theta)}. \tag{9.19}$$

As suggested by the notation of (9.18) and (9.19), we might then iterate this process, taking new weights calculated as $w_i(\boldsymbol{\beta}^{(1)})$ from (9.19), then solving (9.18) 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,

$$\min_{\beta} \sum_{i=1}^{n} w_i(\boldsymbol{\beta})\{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\}^2. \tag{9.20}$$

Now consider an additive error model in which the expectation function is nonlinear but in which the variance model $g(\mu_i(\boldsymbol{\beta}), \theta) = 1$ for all $i = 1, \ldots, n$, namely,

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma\,\epsilon_i, \tag{9.21}$$

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 $k = 1, \ldots, p$,

$$V_{i,k}^{(0)} = \frac{\partial}{\partial \beta_k} g(\boldsymbol{x}_i, \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}} . \tag{9.22}$$

This approximation then allows us to write,

$$Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta}^{(0)}) \approx \sum_{i=1}^{n} V_{i,k}^{(0)} (\beta_k - \beta_k^{(0)}) + \sigma \, \epsilon_i, \tag{9.23}$$

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 (9.23) 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

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \boldsymbol{\delta}^{(0)}. \tag{9.24}$$

Replacing $\boldsymbol{\beta}^{(0)}$ with $\boldsymbol{\beta}^{(1)}$ in (9.22) and (9.23) allows expression of an updated model form in terms of $\boldsymbol{V}^{(1)}$ and $\tilde{\boldsymbol{Y}}^{(1)}$, and (9.24) 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

$$\min_{\beta} \sum_{i=1}^{n} \{y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\}^2. \tag{9.25}$$

Finally, consider a combination of the two models discussed above, namely,

$$Y_i = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta) \, \epsilon_i,$$

where we are still considering θ as known. Here, a combination of the thinking that resulted in (9.18) and (9.19) for linear models and (9.22) through (9.24) for nonlinear models results in a full-blown generalized least squares algorithm as laid out in Chapter 6.2.3. The least squares minimization problem this algorithm finds a solution to is

$$\min_{\beta} \sum_{i=1}^{n} w_i(\boldsymbol{\beta})\{y_i - g_1(\boldsymbol{x}_i, \boldsymbol{\beta})\}^2, \tag{9.26}$$

where $w_i(\boldsymbol{\beta})$ is now defined as (c.f. expression (9.19)),

$$w_i(\boldsymbol{\beta}) = \frac{1}{g_2^2(\boldsymbol{x}_i, \boldsymbol{\beta}, \theta)}.$$

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 (9.26). Thus, if the generalized least squares algorithm is, in fact, solving (9.26) 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).

Inference based on generalized least squares estimators typically consists of interval estimation of elements of $\boldsymbol{\beta}$ based on the Fundamental Theorem of Generalized Least Squares presented in Chapter 6.2.4 using the moment-based estimator of σ_n^2 also given in that chapter. Pointwise confidence bands for the expectation function or any other functions of $\boldsymbol{\beta}$ can be computed using the delta method of Chapter 5.5. Another approach for the construction of confidence intervals for functions of any of the model parameters will be the use of *parametric bootstrap* methods, discussed in later chapters of this material.

9.7.2 Pseudolikelihood

While some version of least squares can be applied to models with constant variance, variances proportional to known constants, or variances that depend on parameters in the expectation function but not other unknown parameters, least squares is not a viable estimation method for models with additional unknown parameters in the variance model, such as those of the general form (9.9). This is because estimation of the regression parameters $\boldsymbol{\beta}$ cannot be obtained without knowledge of θ, as they can be without knowledge of σ^2. There is no way to define a least squares problem that produces estimates of both $\boldsymbol{\beta}$ and θ.

The Pseudolikelihood of Carroll and Ruppert

One alternative was suggested by Carroll and Ruppert (1988) as a type of pseudo-likelihood estimation. Be aware that this is not the only procedure that is called pseudo-likelihood. There are other procedures for vastly different problems than estimation of additive error models that are sometimes called pseudo-likelihood (e.g., the pseudo-likelihood of Besag, 1975) for spatial Markov random field models). The procedure suggested by Carroll and Ruppert (1988) concerns the very general additive error model of (9.9), namely,

$$Y_i = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, z_i, \theta)\, \epsilon_i,$$

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

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma g(\mu_i(\boldsymbol{\beta}), z_i, \theta)\, \epsilon_i, \tag{9.27}$$

which was previously given as expression (9.10).

For model (9.27) not assuming that θ is known, the pseudo-likelihood strategy is an attempt to allow 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 $Y_i \sim$ *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,

$$\begin{aligned} L_*(\theta, \sigma^2 | \boldsymbol{\beta}^{(0)}) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n} \log\left[g^2\left\{ \mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta \right\}\right] \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{y_i - \mu_i(\boldsymbol{\beta}^{(0)})}{g\left\{ \mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta \right\}} \right]^2 . \end{aligned} \tag{9.28}$$

One way to maximize the pseudo-likelihood (9.28) in θ and σ^2, is to apply the idea of profiling for θ. That is, if we take the partial derivative of (9.28) with respect to σ^2 and set it equal to zero, the solution is,

$$\hat{\sigma}^2(\theta | \boldsymbol{\beta}^{(0)}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \mu_i(\boldsymbol{\beta}^{(0)})}{g\left\{ \mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta \right\}} \right]^2 . \tag{9.29}$$

So the maximum pseudo-likelihood estimate of σ^2 is a function of θ. Thus, to maximize (9.28) in θ and σ^2, we maximize in θ what could be called a

log-profile-pseudo-likelihood, formed by substituting the solution (9.29) into (9.28) to arrive at,

$$L_*^p(\theta|\boldsymbol{\beta}^{(0)}) \;=\; -\frac{n}{2}\log\left\{\hat{\sigma}^2(\theta|\boldsymbol{\beta}^{(0)})\right\} - \frac{1}{2}\sum_{i=1}^{n}\log\left[g^2\left\{\mu_i(\boldsymbol{\beta}^{(0)}), z_i, \theta\right\}\right]. \tag{9.30}$$

To find maximum pseudo-likelihood estimates of θ and σ^2, for a given fixed value of $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$, we would first maximize (9.30) in θ to give $\theta^{(0)}$ and then we would use both $\boldsymbol{\beta}^{(0)}$ and $\theta^{(0)}$ in (9.29) 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 $\boldsymbol{\beta}_n^{(0)}$, estimating θ by maximizing (9.30) with $\hat{\sigma}^2(\theta|\boldsymbol{\beta}^{(0)})$ given in (9.29), completing the steps of the generalized least squares algorithm of Chapter 6.2.4 to obtain updated estimates of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_n^{(1)}$, repeating estimation of θ as above with $\boldsymbol{\beta}_n^{(1)}$ replacing $\boldsymbol{\beta}_n^{(0)}$, returning to the generalized least squares algorithm with the new value of θ, and so forth until a given stopping rule is met. In essence, what has been done is to insert an estimation phase for θ between steps 1 and 2 of the generalized least squares algorithm.

Pseudolikelihood Inference

There is no one clear path for making inference about the parameters using the pseudo-likelihood procedure outlined in the previous subsection. 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) 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. A criticism of this approach is that uncertainty in the estimation of θ is not accounted for in making inference about $\boldsymbol{\beta}$.

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 at this point, 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 (9.27) 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 is beyond the scope of these notes.

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.

9.7.3 Likelihood Estimation and Inference

It is, of course, possible to specify a parametric distribution for additive error terms in any of the models discussed in this chapter, use maximum likelihood estimation, and base inference on either asymptotic normality of those estimates as summarized in Chapter 5.5, or on what was called inference from properties of the log likelihood in Chapter 5.6.

There is little unification that can be provided for likelihood analysis with additive error models, each model needing to be approached separately. The ease or difficulty of locating maximum likelihood estimates will depend on the specific forms chosen for expectation functions and variance models. The use of profiles, which we have already seen in conjunction with pseudolikelihood estimation is often a useful device, particularly for parameters that are part of the variance model but not included in the expectation function (e.g. θ in the models of Chapter 9.5).

Maximum likelihood is probably the default (non-Bayesian) method of estimation for transform both sides models. The use of profiling is again often helpful in maximization of the log likelihood function, at least in models with constant error variance. Specifically, consider model (9.14),

$$Z_i = h(Y_i, \lambda) = h\left\{g(\boldsymbol{x}_i, \boldsymbol{\beta}), \lambda\right\} + \sigma\epsilon_i,$$

where now we assume that $\epsilon_i \sim iidN(0,\, 1)$ for $i = 1, \ldots, n$. In this notation, the Z_i are independent and have normal distributions with expected values $h\{g(\boldsymbol{x}_i, \boldsymbol{\beta})\}$ and common variance σ^2.

The transformation from Z_i to Y_i is $Y_i = h^{-1}(Z_i, \lambda)$ and has Jacobian

$$J_i(\lambda) = \frac{\partial h\{Y_i, \lambda)}{\partial Y_i}.$$

The density of Y_i is then,

$$m(y_i|\boldsymbol{\beta}, \sigma^2, \lambda) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left[h(y_i, \lambda) - h\{g(\boldsymbol{x}_i, \boldsymbol{\beta}), \lambda\}\right]^2\right) J_i(\lambda). \tag{9.31}$$

The log likelihood for the set of responses $Y_1, \ldots, Y_n$ is, up to an additive constant,

$$\ell(\boldsymbol{\beta}, \sigma^2, \lambda) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[h(y_i, \lambda) - h\{g(\boldsymbol{x}_i, \boldsymbol{\beta}), \lambda\}\right]^2 + \sum_{i=1}^{n}\log\{J_i(\lambda)\}. \tag{9.32}$$

For fixed values of $\boldsymbol{\beta}$ and λ, (9.32) is maximized in σ^2 by

$$\hat{\sigma}^2(\boldsymbol{\beta}, \lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[h(y_i, \lambda) - h\{g(\boldsymbol{x}_i, \boldsymbol{\beta}), \lambda\}\right]^2. \tag{9.33}$$

Maximum likelihood estimates of $\boldsymbol{\beta}$ and λ can then be located by maximizing (9.32) after substitution of (9.33) for σ^2. This is one type of profiling. Another type of profiling may be useful to assist with locating the mle of λ, but this will not be covered until a later section of these notes.

Likelihood inference may proceed using either Wald theory, if the observed information has been computed, perhaps as part of an iterative algorithm for maximization of (9.32) in $\boldsymbol{\beta}$, or through likelihood ratio tests for model selection and the inversion of likelihood ratio tests to compute confidence regions (see Chapters 5.5 and c.6). Confidence intervals for individual elements of $\boldsymbol{\beta}$ through straight likelihood methods (i.e., not Wald theory) can be obtained through a profile procedure, which will be covered later in these notes.

9.8 Residuals and Residual Plots

We have already seen the use of residuals in a number of examples in this chapter, and it was assumed that the reader is familiar with the basic use of residuals from previous courses. In this section we present several types of residuals that can be useful with additive error models, and illustrate their development with a number of examples. Throughout this section we will assume that we have an additive error model of the form (9.9) which we will write using $\mu_i(\boldsymbol{\beta}) = g_1(\boldsymbol{x}_i, \boldsymbol{\beta})$ as

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma g(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}_i, \theta)\epsilon_i, \tag{9.34}$$

which is similar to (9.10) except that we retain the added generality of allowing $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$ to enter the variance in possibly ways other than through $\mu_i(\boldsymbol{\beta})$.

9.8.1 Types of Residuals

There are any number of quantities we might consider to be residuals in particular models. It would seem we may place the majority of such quantities into the broad categories, (1) raw and absolute residuals and (2) standardized and studentized residuals.

Raw and Absolute Residuals

The most basic form of residuals are what we can call *raw residuals*, defined as,

$$\begin{aligned} r_i &= y_i - \mu_i(\hat{\boldsymbol{\beta}}), \\ \text{or} & \\ r_i &= \mu_i(\hat{\boldsymbol{\beta}}) - y_i. \end{aligned} \tag{9.35}$$

We will use the two forms of (9.35) interchangeably although for interpretation of over or under estimation of $\mu_i(\boldsymbol{\beta})$ it is obviously important to keep track of which form is being used. Raw residuals can be useful in their own right in models such as 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 the basic building blocks for many other residual quantities as they clearly embodied what we intuitively think of as a residual.

A number of authors (Cook and Weisberg, 1982; Carroll and Ruppert, 1988) have advocated using either squared residuals or absolute residuals as more informative about variance structure than simple raw residuals. Squared residuals are $s_i = r_i^2$ and absolute residuals are $a_i = |r_i|$. 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 squared and absolute residuals have been suggested as useful in certain situations. We defer a discussion of such transformations until the portion of this section that discusses plotting residuals.

Standardized and 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 possess constant variance, even if the response variables Y_i or the error terms ϵ_i do. It is typically desirable then to use *studentized* residuals, which should have constant variance equal to about 1. Some statisticians distinguish between *standardization*, in which a random

variable is divided by its standard deviation, producing a quantity that has variance 1, and *studentization* in which the variable is divided by its estimated standard deviation and has variance that should be close to 1. Other statisticians use the two terms interchangably. If one is going to distinguish, then in practice we can only studentize, not standardize.

For ease of presentation, consider a model written as

$$Y_i = \mu_i(\boldsymbol{\beta}) + \sigma_i \, \epsilon_i,$$

where for $i = 1, \ldots, n$, $\epsilon_i \sim iidF$, $E\{\epsilon_i\} = 0$ and $var\{\epsilon_i\} = 1$. Of course, the variances σ_i^2 will be modeled in terms of reduced sets of parameters, as in any of the models considered in this chapter. Suppose, for the time being, that the variances σ_i^2 are known, but the expected values $\mu_i(\boldsymbol{\beta})$ are to be estimated. This model, along with the definition of raw residuals in (9.35), indicates that the random form of residuals is,

$$\begin{aligned} R_i &= \mu_i(\hat{\boldsymbol{\beta}})_i - Y_i \\ &= \mu_i(\hat{\boldsymbol{\beta}}) - \mu_i(\boldsymbol{\beta}) - \sigma_i \, \epsilon_i. \end{aligned}$$

Then,

$$var(R_i) = var\{\mu_i(\hat{\boldsymbol{\beta}})\} + \sigma_i^2 - 2\sigma_i cov\{\mu_i(\hat{\boldsymbol{\beta}}), \, \epsilon_i\},$$

and we can define studentized residuals as, for $i = 1, \ldots, n$,

$$b_i = \frac{r_i}{\left[var\{\mu_i(\hat{\boldsymbol{\beta}})\} + \sigma_i^2 - 2\sigma_i cov\{\mu_i(\hat{\boldsymbol{\beta}}), \, \epsilon_i\}\right]^{1/2}}. \tag{9.36}$$

Now, it will not be the case that the σ_i^2 will be known, and the typical approach is to use plug-in estimates of σ_i^2 in (9.36) giving

$$\tilde{b}_i = \frac{r_i}{\left[var\{\hat{\mu}_i\} + \hat{\sigma}_i^2 - 2\hat{\sigma}_i cov\{\hat{\mu}_i, \, \epsilon_i\}\right]^{1/2}}. \tag{9.37}$$

Note that by doing this plug-in procedure we have ignored any possible covariance of $\mu_i(\hat{\boldsymbol{\beta}})$ with $\hat{\sigma}_i^2$ since (9.36) was developed assuming that the σ_i^2 were known, that is, in deriving $var(R_i)$ just prior to (9.36). Thus, common practice is to worry about the covariance of $\mu_i(\hat{\boldsymbol{\beta}})$ with ϵ_i, but not covariance between $\mu_i(\hat{\boldsymbol{\beta}})$ and estimates of σ_i^2. 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 unequal variances. If both $\mu_i(\boldsymbol{\beta})$ and σ_i^2 depend on $\boldsymbol{\beta}$, as in a number of our models, ignoring the covariance between estimated values $\mu_i(\hat{\boldsymbol{\beta}})$ and $\hat{\sigma}_i^2$ may not be the best idea, but it is what is commonly done because dealing with that covariance is exceptionally difficult.

Example 9.5

If ordinary least squares is used to estimate $\boldsymbol{\beta}$ in the linear regression model (6.4) we have, from $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\,\boldsymbol{X})^{-1}$ and $\mu_i(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_i^T\,\hat{\boldsymbol{\beta}}$, that

$$\begin{aligned} var\{\mu_i(\hat{\boldsymbol{\beta}})\} &= \sigma^2 \boldsymbol{x}_i^T(\boldsymbol{X}^T\,\boldsymbol{X})^{-1}\boldsymbol{x}_i \\ &= \sigma^2 h_{i,i}, \end{aligned} \tag{9.38}$$

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, since $\hat{\boldsymbol{\mu}} = \boldsymbol{H}\,\boldsymbol{Y}$

$$\begin{aligned} cov\{\mu_i(\hat{\boldsymbol{\beta}}),\, \epsilon_i\} &= E\{\mu_i(\hat{\boldsymbol{\beta}})\epsilon_i\} - 0 \\ &= E\left\{\epsilon_i \sum_{j=1}^{n} Y_j h_{i,j}\right\} \\ &= \sum_{j=1}^{n} h_{i,j} E\{Y_j \epsilon_i\} \\ &= \sum_{j=1}^{n} h_{i,j} E\{(\mu_j(\boldsymbol{\beta}) + \sigma\epsilon_j)\epsilon_i\} = \sigma\, h_{i,i}. \end{aligned} \tag{9.39}$$

For this model, $\sigma_i^2 = \sigma^2$ for $i = 1, \ldots, n$. Replacing σ^2 in (9.39) with its usual moment-based estimator and then substituting into (9.37) gives

$$\tilde{b}_i = \frac{r_i}{[\hat{\sigma}^2(1 - h_{i,i})]^{1/2}}, \tag{9.40}$$

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 $\epsilon_i \sim iidF$ where F is assumed to be a location-scale family, $E(\epsilon_i) = 0$ 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 giving asymptotic normality of $\hat{\boldsymbol{\beta}}$. Hence, derivation of exact forms for the component quantities of (9.37) is difficult. One development of what is usually considered a 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,

$$[\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})] = [I - \boldsymbol{H}][\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^*)], \tag{9.41}$$

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 for a linear model this gives studentized residuals in the form of expression (9.40). 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,

$$\mu(\boldsymbol{\beta}) \approx \mu(\boldsymbol{\beta}^*) + V(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \tag{9.42}$$

Note that in (9.42) 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 (9.42), 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,

$$\min_{\boldsymbol{\beta}} \left[\{\boldsymbol{y} - \mu(\boldsymbol{\beta}^*)\} - V(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right]^T \left[\{\boldsymbol{y} - \mu(\boldsymbol{\beta}^*)\} - V(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right],$$

which has the ordinary least squares solution,

$$\tilde{\boldsymbol{\delta}} = (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = [V^T(\boldsymbol{\beta}^*)\, V(\boldsymbol{\beta}^*)]^{-1}\, V^T(\boldsymbol{\beta}^*) \{\boldsymbol{y} - \mu(\boldsymbol{\beta}^*)\}. \tag{9.43}$$

Now, we can't actually compute $\tilde{\delta}$ or $\tilde{\boldsymbol{\beta}}$. But, asymptotic results (see e.g., Seber and Wild, 1989, 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 (9.43) with $\hat{\boldsymbol{\beta}}$ in place of $\tilde{\boldsymbol{\beta}}$.

Now, consider the vector of raw residuals,

$$\begin{aligned}
\boldsymbol{r} &= \boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) \\
&\approx \boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^*) + V(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&\approx \boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^*) + V(\boldsymbol{\beta}^*)[V^T(\boldsymbol{\beta}^*)\, V(\boldsymbol{\beta}^*)]^{-1}\, V^T(\boldsymbol{\beta}^*)\, [y\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^*)] \\
&= \left[I - V(\boldsymbol{\beta}^*)(V^T(\boldsymbol{\beta}^*)\, V(\boldsymbol{\beta}^*))^{-1} V^T(\boldsymbol{\beta}^*)\right]\, [\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^*)] \\
&= [I - \boldsymbol{H}^{(N)}(\boldsymbol{\beta}^*)][\boldsymbol{Y} - \mu(\boldsymbol{\beta}^*)].
\end{aligned} \tag{9.44}$$

The second line of (9.44) follows from substitution of (9.42) evaluated at $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$, while the third line results from further use of (9.43) with $\hat{\boldsymbol{\beta}}$ in place of $\tilde{\boldsymbol{\beta}}$ as just discussed. The final line of (9.44) is analogous to the linear model result (9.41) 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 expression (9.44) being the parallel of the linear model result (9.41)) in hand, we appeal to analogy with linear model results and *define* studentized residuals to be

$$\tilde{b}_i = \frac{r_i}{[\hat{\sigma}^2\,\{1 - h_{i,i}^{(N)}(\hat{\boldsymbol{\beta}})\}]^{1/2}}. \tag{9.45}$$

Notice that in (9.45) 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 a more 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 formulation). 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 $\mu_i(\boldsymbol{\beta})/g(\mu_i(\boldsymbol{\beta}), z_i, \theta)$. 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 (9.45) as,

$$\tilde{b}_i = \frac{\tilde{r}_i}{[\hat{\sigma}^2 \{1 - \tilde{h}_{i,i}^{(N)}(\hat{\boldsymbol{\beta}})\}]^{1/2}}, \tag{9.46}$$

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,

$$\frac{1}{g(\mu_i(\hat{\boldsymbol{\beta}}), z_i, \theta)} \left[\frac{\partial}{\partial \beta_k} \mu_i(\boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \right].$$

9.8.2 Plotting Residuals

Any number of diagnostic plots can be constructed residuals, 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 abscissa. 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 perceived 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 understanding 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

What we might think of as a basic or standard residual plot results from plotting residuals against fitted values from a model. Fitted values are 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.

Example 9.8

In Example 9.1 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 9.14 presents the studentized residuals (9.45) for both groups. This residual plot does not reveal any serious problems with the model, although it is less than textbook 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 that variability appears 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 9.14 exhibit some degree of increasing variance as a function of increasing mean? To help in this assessment, we might plot the cube root of squared studentized residuals against the fitted values. 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 9.15. There does not appear to be a increasing wedge or fan of residuals in the plot of Figure 9.15, suggesting that there is little evidence of nonconstant variance for this model. Looking closely at the residual plot of Figure 9.14 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 9.16 verifies that this seems to be the case, at least for the treated preparations. In fact, there appear to be values at a substrate concentration of just over 0.2 ppm for which the expectation function misses for both treated and untreated groups. The cause of this

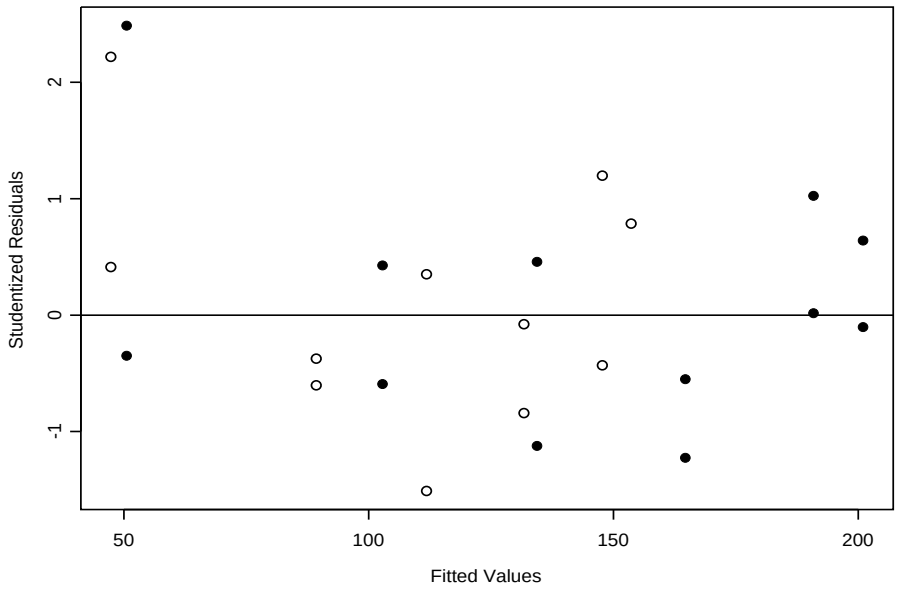


Figure 9.14: Studentized residuals from fitting a nonlinear regression based on the Michaelis-Menten equation to the enzyme reaction times of Example 5.1. Open circles are the untreated preparations while solid circles are the treated preparations.

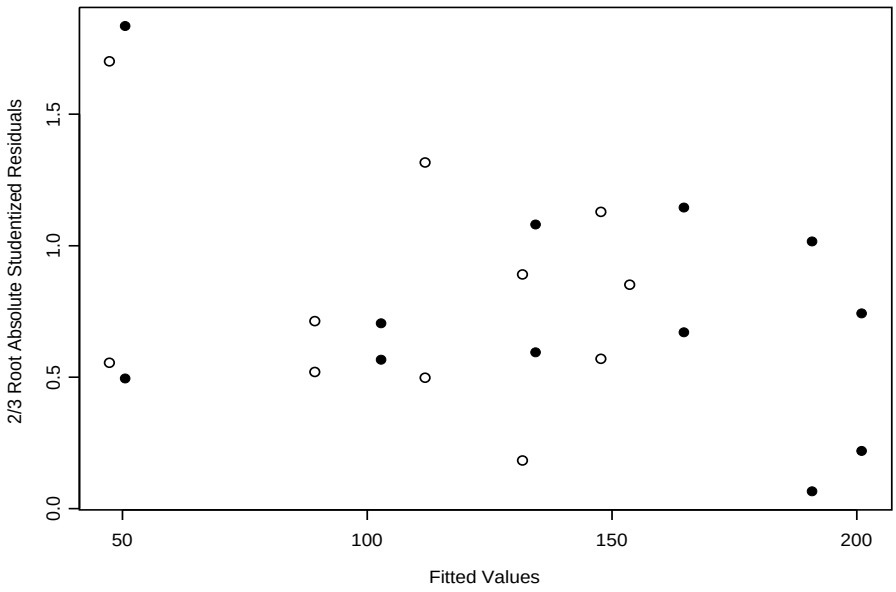


Figure 9.15: Cube root squared studentized residuals from fitting a nonlinear regression based on the Michaelis-Menten equation to the enzyme reaction times of Example 5.1. Open circles are the untreated preparations while solid circles are the treated preparations.

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, which transforms chi-squared variables into variables with normal distributions, but the basic value of such plots seems due to more practical than theoretical 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. **?** 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 alleviate this potential difficulty, but point out that they view the result essentially as a transformation of absolute residuals. From this standpoint, it would seem to make little difference if one used absolute residuals, the square root of absolute residuals or, as in Figures 9.14 and 9.15, a 2/3 power of absolute residuals.

A common difficulty with the basic plot of residuals against fitted values is that the density of points on the axis of fitted values (the horizontal axis of the plot) can differ across the range of values. This can easily lead to the visual impression of a pattern when there really is none, particularly in terms of nonconstant variance.

Example 9.9

Data were simulated from a nonlinear regression model with constant variance

$$Y_i = \alpha x_i^{\beta} \exp(-x_i) + \sigma\epsilon_i,$$

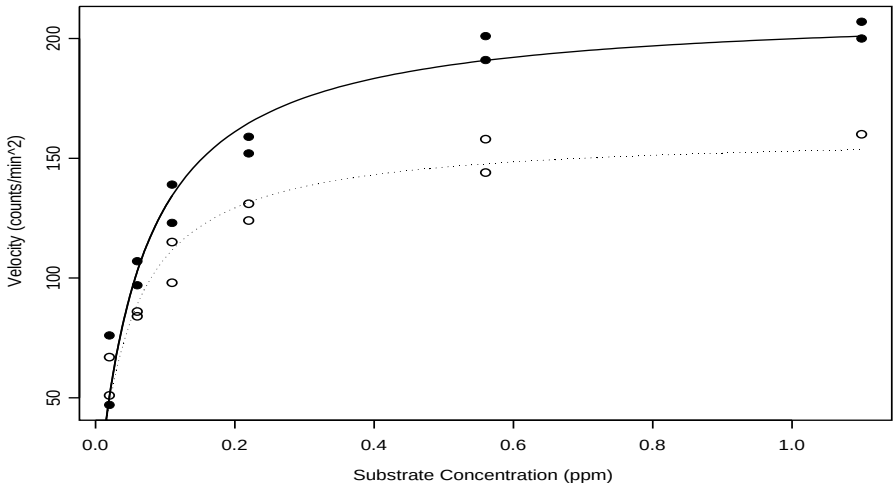


Figure 9.16: Fitted regressions based on the Michaelis-Menten equation to the enzyme reaction times of Example 9.1. Open circles are the untreated preparations while solid circles are the treated preparations.

where the covariates x_i were in the range $(0, 10)$, $\alpha = 1.5$, $\beta = 2.5$ and $\epsilon_i \sim iidN(0, 0.1225)$. A scatterplot of the simulated data is presented in Figure 9.17 along with true (as dashed curve) and estimated (as solid curve) expectation functions.

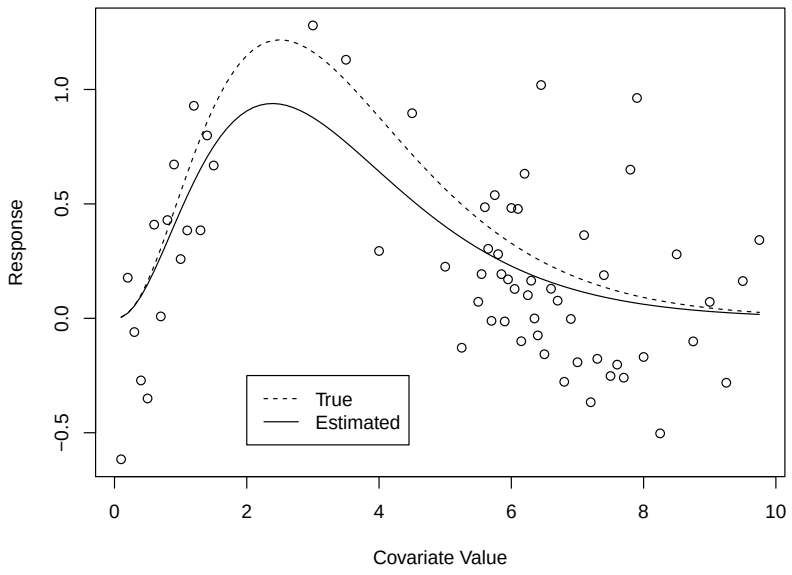


Figure 9.17: Scatterplot for simulated example showing true expectation function as dashed curve and estimated function as solid curve.

The model was fit using generalized least squares, giving $\hat{\alpha} = 1.28$ and $\hat{\beta} = 2.39$. A basic residual plot is shown in Figure 9.18, with studentized residuals (9.45) plotted against fitted values $\mu_i(\hat{\alpha}, \hat{\beta})$. It is easy to arrive at a visual impression from this residual plot that variances are decreasing as expected values increase. But this impression is enitrely an artifact of the changing density of points along the axis of fitted values. If we construct the residual plot using the logarithm of fitted values we arrive at what is shown in

Figure 9.19. Although using the logarithm of fitted values has resulted in one

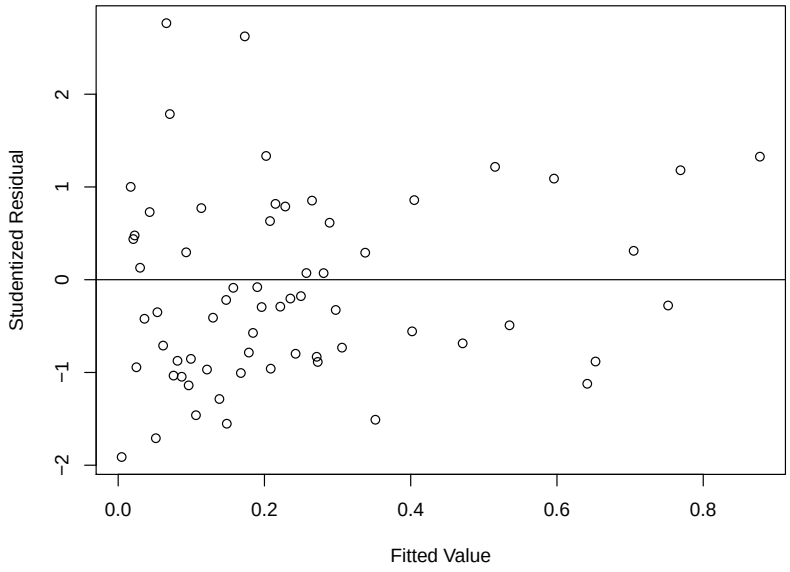


Figure 9.18: Studentized residuals against the fitted values for a model fit to the data of Figure 9.17.

data point in the extreme lower left hand corner, the visual impression that the amount of scatter in the residuals decreases as fitted values increase has largely disappeared.

There is no one approach by which to determine a good scale for the horizontal axis in a plot of residuals against fitted values. The logarithm of fitted values is often effective, but the square root and other powers of fitted values have also been suggested. **?**, p. 32 also consider a quantity for the model of Example 9.7 that they attribute to Cook and Weisberg for the horizontal axis

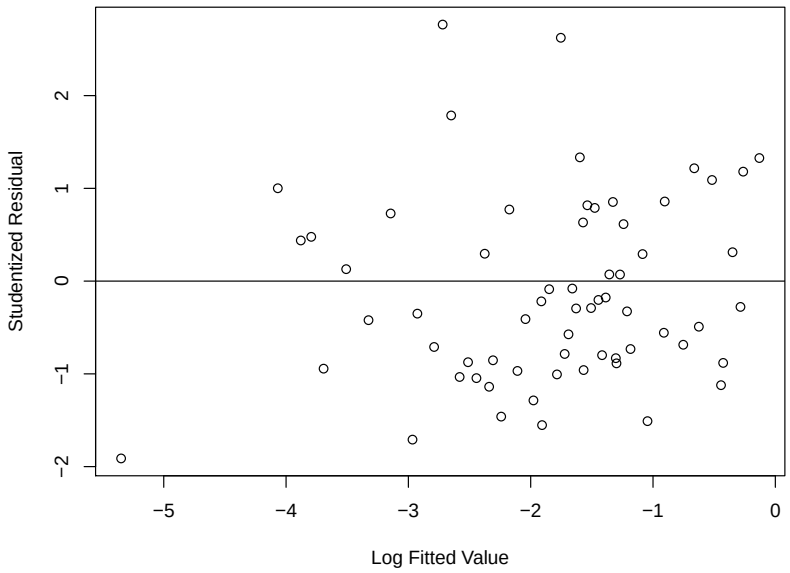


Figure 9.19: Studentized residuals against the logarithm of fitted values for a model fit to the data of Figure 9.17.

of a basic residual plot which is,

$$s_i = \frac{\partial}{\partial \theta} \left[g(\mu_i(\boldsymbol{\beta}), z_i, \theta) \right] \Big|_{\theta=0} .$$

Another possibility to avoid being visually misled by residual plots when using them to assess heterogeneity of variances is to plot absolute, log absolute, or cube root of squared residuals against fitted values and then apply a nonparametric smoother to estimated the expected values of the transformed residuals as a function of the response means. This same technique can be used with various transformations of absolute residuals. Silverman (1985) provides additional information on this idea.

Plotting against Covariates and Potential Covariates

Although plotting residuals against fitted values is often all that is needed to determine whether there are gross violations of a specified expectation function, it can sometimes be useful to also plot residuals against covariates used in a model. This has been illustrated in consideration of Example 9.4 relating tree volume to height and diameter. In particular, Figure 9.11 suggested that a multiple regression of volume on both height and diameter did not alleviate a potential problem with an assumption that expected volume was related to diameter as a straight line, and Figure 9.13 suggested that a model with a constructed covariate called cylinder still showed increasing variances as a function of height.

It can also be useful to plot residuals from a regression against potential covariates that were not included in the model that produced residuals. The basic idea is to assume that a regression has successfully accounted for the influence of a covariate or covariates used in the model (at least influence on expected values). If residuals are related in a systematic way to another possible covariate that is not currently in the model, that new covariate has something to contribute to the relation that has not already been accounted for.

Plotting Against Time or Space

Suppose that data are collected over a given time span, or gradient in space such as latitude, but time or space is not included in a regression model used to analyze the data. An effective way to examine whether time or space has an effect on response values beyond whatever covariate or covariates are included in the model is to plot residuals against time or the spatial gradient. This

is, essentially the same as plotting residuals against potential covariates not included in the model. A twist, however, is that data indexed in time and/or space are often modeled as nonindependent in those dimensions.

9.9 Bayesian Analysis

As for basic generalized linear models, conducting a Bayesian analysis of an additive error model is largely a matter of specifying prior distributions and determining what type of MCMC algorithm to use for approximation of the joint posterior distribution.

9.9.1 Assigning Prior Distributions

Using improper priors in nonlinear additive error models is difficult due to the need to demonstrate posterior propriety. Inducing priors on the regression parameters through the use of methods such as conditional means priors is considerably more difficult than for basic generalized linear models because the form of expectation functions contains no certain structure. That is, for a nonlinear regression we simply have $\mu_i = g(\boldsymbol{x}_i, \boldsymbol{\beta})$ while in glms there is always a linear component $\mu_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Connected with this is that the parameter space for $\boldsymbol{\beta}$ is typically $\mathbb{R}^p$ in a glm, while the parameter spaces for elements of $\boldsymbol{\beta}$ in an additive error nonlinear regression are often restricted to only a portion of the real line, and may differ for different components.

A common approach to assign prior distributions to the components of $\boldsymbol{\beta}$ then is to consider each element individually and assign proper prior distributions, often either chosen to match the relevant parameter space, or truncated to do so. It is also the case that many nonlinear functions are well-behaved only

for parameters in some *window* of the entire parameter space, often connected with the magnitude of covariates. For example, if $x \in (0, 10)$ the inverse of the logit function $\exp(-2+\alpha x)/[1+\exp(-2+\alpha x)]$ will have a sigmoidal shape between 0 and 1 only for values of α between about 0.4 and 1.2. Truncation can also be used to restrict parameters to lie in regions of the parameter space that produce certain behaviors in nonlinear expectation functions. Alternatively, it is sometimes the case that parameters are transformed to have support on the entire line, and then normal priors are a natural default, although the typical device of making prior variances large may cause difficulties.

Example 9.10

In the Michaelis-Menton expectation function of Example 9.1 is $\mu_i = \beta_1 x_i/(\beta_2 + x_i)$. The parameter spaces are $\beta_1 > 0$ and $\beta_2 > 0$, and in the application to enzyme reaction scientific understanding indicates that the curve should increase from 0 at $x_i = 0$ to a asymptote given by β_1 as x_i becomes large. Three possible strategies for assigning priors as $\pi(\beta_1, \beta_2) = \pi(\beta_1)\pi(\beta_2)$to these parameters follow.

1. Take $\pi(\beta_1)$ and $\pi(\beta_2)$ to both be gamma distributions but with the mean of β_1 substantially larger than that for β_2.

2. Take $\pi(\beta_1)$ and $\pi(\beta_2)$ to both have normal distributions truncated below at 0 again with a larger mean for β_1 than β_2 and moderate variances.

3. Let $\tilde{\beta}_1 = \log(\beta_1)$ and $\tilde{\beta}_2 = \log(\beta_2)$. Assign both $\tilde{\beta}_1$ and $\tilde{\beta}_2$ normal distributions.

In some ways, choosing parameter values for the prior distributions of regression parameters in an additive error nonlinear model is analogous to choosing starting values for iterative algorithms in maximum likelihood estimation.

It is difficult to achieve success if one is totally ignorant of how the model behaves relative to values of the parameters involved. But how the model behaves is typically related to the data and, in particular, the magnitude and range of values observed for both response values and covariates. In a Bayesian analysis, however, examination of the data for the purpose of assigning prior distributions is generally viewed as a violation of the principle that a prior distribution represents our beliefs before seeing the data. The question, then, is if adjustment of prior parameter values based on whether or not an MCMC algorithm converges or perhaps even numerically crashes constitutes examination of the data. We take the position that it does, and that failure of an MCMC procedure to converge because of prior specification indicates simply that one does not possess enough prior knowledge to enact a Bayesian analysis. This view is that having some prior belief about the value of data model parameters is a prerequisite for taking a Bayesian approach, in the same way that having a probability-based sample is a prerequisite for using survey sampling methodology.

The parameter σ^2 in any of the model forms presented in this chapter is a scale parameter in normal distributions. As such, we can exploit conditional conjugacy and assign σ^2 an inverse gamma prior. The full conditional posterior of σ^2 will then also be inverse gamma and sampling will be straightforward if an overall Gibbs Sampling algorithm is used to approximate the joint posterior. An alternative is to assign σ^2 a proper uniform prior or the half-t distribution of Gelman (2006). Assigning prior distributions to other parameters in the variance model faces the same difficulties as for parameters in the expectation function, and is highly dependent on the exact form of the variances. It is sometimes advisable to at least initially treat parameters in the variance model that are not also part of the expectation function as tuning parameters. They

then stay fixed for approximation of the joint posterior for $(\boldsymbol{\beta}, \sigma^2)$.

9.9.2 Derivation of Posterior Distributions

Identifying joint posterior distributions for additive error models is largely a matter of using the general form of likelihood times prior. There are few simplifications that are general in nature or that can be applied to a large number of situations. We present a few cases to illustrate.

Simple Linear Regression

One case in which considerable simplification of conditional posteriors is possible is a simple linear regression model with normal errors, $Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i$ where $\epsilon_i \sim$ iid N$(0,1)$. The parameter space for (β_0, β_1) is $\mathbb{R}^2$ and $\sigma^2 > 0$. Suppose we assign individual priors to the three parameters as $\beta_0 \sim N(\lambda_0, \tau_0^2)$, $\beta_1 \sim N(0, \tau^2)$, and $\sigma^2 \sim IG(\xi_1, \xi_2)$ and use $\pi(\beta_0, \beta_1, \sigma^2) = \pi(\beta_0)\pi(\beta_1)\pi(\sigma^2)$. In anticipation of using a Gibbs Sampling algorithm to simulate values from the joint posterior we can examine the full conditional posterior distributions as,

$$
\begin{aligned}
p(\beta_0|\cdot) &\propto f(\boldsymbol{y}|\beta_0, \beta_1, \sigma^2)\, \pi(\beta_0) \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - \frac{1}{2\tau_0^2}(\beta_0 - \lambda_0)^2\right] \\
p(\beta_1|\cdot) &\propto f(\boldsymbol{y}|\beta_0, \beta_1, \sigma^2)\, \pi(\beta_1) \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - \frac{1}{2\tau_1^2}(\beta_1 - \lambda_1)^2\right] \\
p(\sigma^2|\cdot) &\propto f(\boldsymbol{y}|\beta_0, \beta_1, \sigma^2)\, \pi(\sigma^2) \\
&\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right] \frac{1}{(\sigma^2)^{\xi_1+1}} \exp(-\xi_2/\sigma^2).
\end{aligned}
$$

Let $S_0 = \sum(y_i - \beta_1 x_i)^2$, $S_1 = \sum(y_i - \beta_0)$, $S_2 = (\sum x_i)^2$ and $S_3 = \sum(y_i - \beta_0 - \beta_1 x_i)^2$. Inspection of $p(\sigma^2|cdot)$ and combining powers gives that this conditional posterior is inverse gamma with parameters $\xi_1 + (n/2)$ and $\xi_2 + (1/2)S_3$. Completing the square in each of $p(\beta_0|\cdot)$ and $p(\beta_1|\cdot)$ shows that $p(\beta_0|\cdot)$ is normal with mean M_0 and variance V_0 while $p(\beta_1|\cdot)$ is normal with mean M_1 and variance V_1, where

$$M_0 = \frac{\tau_0^2 S_0 + \sigma^2 \lambda_0}{n\tau_0^2 + \sigma^2} \qquad V_0 = \frac{\sigma^2 \tau_0^2}{n\tau_0^2 + \sigma^2}$$
$$M_1 = \frac{\tau_1^2 S_1 + \sigma^2 \lambda_1}{\tau_1^2 S_2 + \sigma^2} \qquad V_0 = \frac{\sigma^2 \tau_1^2}{\tau_1^2 S_2 + \sigma^2}.$$

Nonlinear Model Constant Variance

Now consider $Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma\epsilon_i$ where $\epsilon_i \sim \mathrm{N}(0,1)$. Here, we might assign $\boldsymbol{\beta}$ a normal prior with expectation $\boldsymbol{\lambda}_0$ and covariance Σ, and σ^2 again an inverse gamma prior with parameters ξ_1 and ξ_2. The conditional posterior of $\boldsymbol{\beta}$ is,

$$\begin{aligned} p(\boldsymbol{\beta}|\cdot) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\, \pi(\boldsymbol{\beta}) \\ &\propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \{y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\}^2 - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\lambda}_0)^T \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\lambda}_0) \right]. \end{aligned}$$

No simplification of this posterior is readily available in general, although the form of $g(\boldsymbol{x}_i, \boldsymbol{\beta})$ may allow simplification in some cases. The conditional posterior of σ^2 is

$$\begin{aligned} p(\sigma^2|\cdot) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\, \pi(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \{y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\}^2 \right] \frac{1}{(\sigma^2)^{\xi_1 + 1}} \exp(-\xi_2/\sigma^2), \end{aligned}$$

and this may be recognized as an inverse gamma distribution with parameters

$$\xi_1 + (n/2) \qquad \text{and} \qquad \xi_2 + (1/2) \sum_{i=1}^{n} \{y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\}^2.$$

General Additive Error Model

Our most general model with additive errors was written as $Y_i = g_1(\boldsymbol{x}_i, \boldsymbol{\beta}) + \sigma g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}_i, \boldsymbol{\theta})\epsilon_i$ where $\epsilon_i \sim$ iid $\mathrm{N}(0,1)$. The variance parameter $\boldsymbol{\theta}$ may be considered as known (and selected as part of model formulation) or unknown. If $\boldsymbol{\theta}$ is considered known, and if we again assign $\boldsymbol{\beta}$ a normal prior with expectation $\boldsymbol{\lambda}_0$ and covariance Σ and assign σ^2 and inverse gamma prior with parameters ξ_1 and ξ_2, we have conditional posteriors

$$\begin{aligned} p(\boldsymbol{\beta}|\cdot) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)\pi(\boldsymbol{\beta}) \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\frac{y_i - g_1(\boldsymbol{x}_i, \boldsymbol{\beta})}{g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}_i, \boldsymbol{\theta})}\right\}^2 - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\lambda}_0)^T\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\lambda}_0)\right], \\ p(\sigma^2|\cdot) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)\pi(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\frac{y_i - g_1(\boldsymbol{x}_i, \boldsymbol{\beta})}{g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}_i, \boldsymbol{\theta})}\right\}^2\right]\frac{1}{(\sigma^2)^{\xi_1+1}}\exp(-\xi_2/\sigma^2). \end{aligned}$$

The conditional posterior of σ^2 may again be recognized as an inverse gamma, while that of $\boldsymbol{\beta}$ in general will defy simplification. If $\boldsymbol{\theta}$ is considered unknown, the conditional posteriors of $\boldsymbol{\beta}$ and σ^2 remain as above while that of $\boldsymbol{\theta}$, if it is added to the set, is,

$$\begin{aligned} p(\boldsymbol{\theta}|\cdot) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)\,\pi(\boldsymbol{\theta}) \\ &\propto \frac{1}{\prod_{i=1}^{n}\{g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}_i, \boldsymbol{\theta})\}^{1/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\frac{y_i - g_1(\boldsymbol{x}_i, \boldsymbol{\beta})}{g_2(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}_i, \boldsymbol{\theta})}\right\}^2\right]\pi(\boldsymbol{\theta}). \end{aligned}$$

Estimation and Inference

Bayesian estimation and inference for additive error models follows that previously described for basic glms. MCMC algorithms are typically necessary to approximate posterior distributions of the model parameters and an overall structure of Metropolis within Gibbs is often appropriate, which is why we

presented forms for conditional posteriors in the previous section. Depending on the dimension of $\boldsymbol{\beta}$, this parameter vector can be updated in its entirety with one Metropolis-Hastings step or partitioned into several pieces which are updated in sequence. If the conditionally conjugate inverse gamma prior was assigned to σ^2 then update of this parameter is straightforward. Alternatively, σ^2 and $\boldsymbol{\theta}$ might be updated together in one Metropolis step, particularly if $\boldsymbol{\theta}$ is a scalar or some prior other than an inverse gamma has been used for σ^2. In developing an overall algorithm it is often beneficial to consider some parameters fixed and verify that one can successfully deal with the remaining parameters. Then the pieces can be joined together. This type of a strategy helps with debugging algorithms and determining whether an algorithm is sensitive to one or a few particular parameters.

As with basic glms, inference on individual model parameters results from the estimation of moments and quantiles of marginal posterior distributions based on the MCMC output. Again, inferences relative to the expectation function are based on the posterior of that function, produced by computing the functions at each draw of the joint posterior. This is the distribution used to compute the fitted regression model and credible bands in the same manner as for glms in Chapter 8.4.6. Many nonlinear expectation functions contain maxima or minima, asymptotes, inflections points or other features that are meaningful to the scientific question of interest. These may, as in the enzyme reaction problem of Example 9.1, correspond to a particular parameter (β_1 in that example) or some function of the parameters.

Example 9.11

A function of a single covariate that has a maximum at $x = (\beta + 1)/\alpha$ and an inflection point at $x = (\beta + 2)/\alpha$ is given by $g(x, \alpha, \beta) = (\alpha x - \beta)\exp(-\alpha x)$.

To find the posterior distributions of these features of the expectation function we would compute $y_{max}^{(m)} = \exp[-(\beta^{(m)} + 1)]$ and $y_{inf}^{(m)} = 2\exp[-(\beta^{(m)} + 2)]$ at each draw from the joint posterior of $(\alpha^{(m)}, \beta^{(m)})$. The empirical distribution of $\{(y_{max}^{(m)}, y_{inf}^{(m)}) : m = 1, \ldots, M\}$ approximates the joint posterior of the maximum and inflection points of the model.

9.9.3 Model Assessment

Residuals form the backbone of model assessment for linear models. In non-Bayesian analyses, Studentized, squared studentized, or cube root squared studentized residuals play the same role in nonlinear regressions with additive errors. The exact form of these residuals will depend on the particular model under investigation, as illustrated in Examples 9.5 to 9.7. In Bayesian analysis of additive error models, using estimated standard deviations to studentize residuals is not relevant because posterior distributions of regression parameters are not sampling distributions. They are distributions of belief about the values of unknown but fixed parameters. As a result, quantities such as (9.36) and (9.37) are not really meaningful for a Bayesian analysis. We can, however, form residual quantities that are meaningful in a Bayesian approach by isolating the additive error terms in a model. This leads to quantities such as

$$r_i = \frac{y_i - \mu_i(\boldsymbol{\beta})}{\sigma} \tag{9.47}$$

in a constant variance model and

$$\tilde{r}_i = \frac{y_i - \mu_i(\boldsymbol{\beta})}{\sigma\, g(\mu_i(\boldsymbol{\beta}), z_i, \theta)}, \tag{9.48}$$

in a more general model. Quantities (9.47) and (9.48) are functions of fixed values $\{(y_i, z_i) : \ i = 1, \ldots, n\}$ and parameters $\boldsymbol{\beta}$, σ, and θ. As such, we can examine the posterior distributions of these quantities. As with the posterior of $\boldsymbol{\beta}$, a summary statistic such as the mean or median of these posterior distributions can be examined as a useful diagnostic. If the posterior distribution of model parameters is being approximated through the use of an MCMC algorithm, the posteriors of (9.47) or (9.48) can be approximated by computing them for each sample from $p(\boldsymbol{\beta}, \sigma^2, \theta|\boldsymbol{y})$. This results in sets of M empirical values sampled from posteriors $p(r_i|\boldsymbol{y})$ or $p(\tilde{r}_i|\boldsymbol{y})$ for each $i = 1, \ldots, n$. The Bayesian version of a basic residual plot results from plotting the median of each posterior against the median of the posterior distributions of $\mu_i(\boldsymbol{\beta})$. Other quantiles or summary values of the residual posterior distributions are also available if one desires to make use of them.

Additional Bayesian assessment of additive error models proceeds by simulating data sets from the posterior predictive distribution of the model, computing test quantities for each predictive data set, and constructing posterior predictive $p-$values as described in Chapter 7.7. The choice of test quantities is flexible, but should reflect issues of scientific importance and/or choices that were made as part of model formulation.

Example 9.12

Suppose we are working with a power-of-the-mean model $Y_i = \mu_i(\boldsymbol{\beta}) + \sigma\mu_i(\boldsymbol{\beta})^\theta \epsilon_i$ where $\epsilon_i \sim$ iid $\mathrm{N}(0, 1)$ in which we select θ as part of model formulation from a Box-Cox plot of log standard deviations against log means for bins of data created on the basis of a covariate. Let s_a denote the slope of an ordinary least square fit to points of the Box-Cox plot constructed from the actual observed data and let s_m denote the slope of for a fit to points of the Box-Cox

plot constructed from posterior predictive data set $m = 1, \ldots, M$. A posterior predicative $p-$value to assess reproduction of the observed mean-variance relation by the fitted model is computed as $p = (1/M) \sum_{m=1}^{M} I(s_m \geq s_a)$ where $I(A)$ is the indicator function that assumes a value of 1 if A is true and 0 otherwise. Extreme values of p either large or small are indicative of a problem with the model.

Test quantities for posterior predictive assessments should be quantities that can be computed based on the data alone, as opposed to functions of both the observed responses and estimated model parameters which, in a Bayesian analysis would correspond to a summary value of the posterior distribution, such as the mean, median, or mode. In the latter case, one would need to fit the model to each posterior predictive data set before computing the test statistic for that data set. While possible in principle, doing so would require a potentially prohibitive computational burden. In particular, test quantities based on residuals fall into this category. For example, although it might seem useful to examine the ratio of positive to negative residuals to assess distributional issues such as skewness, doing so would impose a heavy cost in terms of computation. Alternative test quantities related to the same issue but that depend only on responses can sometimes be identified. For example to examine skewness in response distributions we might bin the data based on covariate values and then compute the average (across bins) of the ratio of maximum minus median to median minus minimum.

9.10 Case Study: Walleye Length and Weight

In this case study we wish to develop a regression model to relate weight as a response to length as a covariate in a species of freshwater fish, Walleye. There

are a number of scientific reasons such a relation might be of interest. First, fisheries scientists are often interested in obtaining weights of fish for use with additional procedures. Weighing a fish in the field and returning it unharmed to the water is not a task lacking in complications, particularly on a small boat in rough water. The procedure needs to be done rapidly so that a whole net of fish can be weighed before any of them expire, with acceptable precision and without harming the fish. In contrast to weight, the length of fish is relatively easy to measure with good precision in the field. Thus, fisheries scientists have long relied on a strong relation between weight and length to allow length to be recorded in the field, and then weight obtained from a known *length-weight key* at a later time. Another reason to have interest in the relation between weight and length in many fish species is to understand something about the manner in which fish accumulate biomass (weight) relative to their morphological characteristics. Some fish species are long and skinny (e.g., pike), others are more blockly (e.g., carp), and yet others have blockly bodies with fairly extended tails (e.g., sturgeon). While some studies might include the measurement of a whole set of body form characteristics, most studies are not centered on this question. Yet, simply knowing weight and length can perhaps provide some information.

The data set we will use contains values of length and weight for 151 individual walleye captured in a certain set of lakes in the state of Minnesota in 1991, 1992 and 1993. These fish were being collected for the analysis of certain contaminants (e.g., mercury) and so were brought back to the lab, where precise measurements of length (in inches) and weight (in pounds) were recorded.

9.10.1 Exploratory Analysis

Our statistical objective is fairly simple, to develop the best regression model we can to relate weight to length in the available data. A scatterplot of weight versus length is presented in Figure 9.20. It seems fairly clear from the scatter-

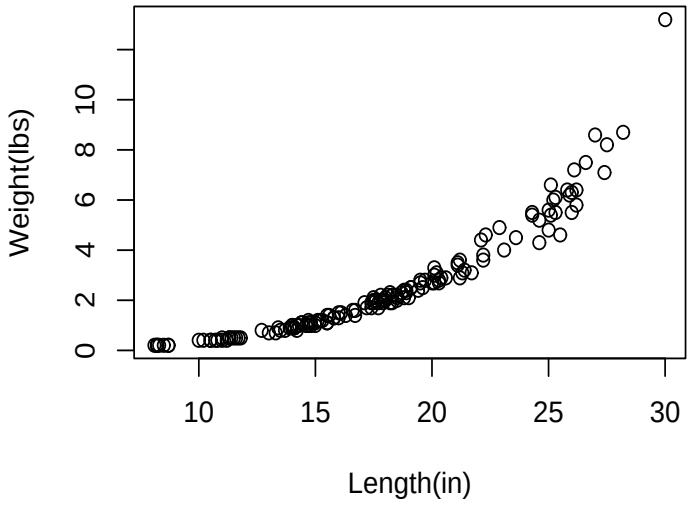


Figure 9.20: Scatterplot of weight versus length in walleye.

plot that (1) the relation between weight and length is not a straight line, (2) there is no evidence that the points would fail to be symmetrically distributed about a curve through the data, (3) the variance seems rather small and (4) despite this, the variances do seem to increase as the expected values increase. Because the variances of responses at given covariate values seems rather small relative to the overall spread of responses we might begin with consideration of determining a suitable systematic model component.

Finding a Good Empirical Fit

An expectation function appropriate for these data appears to be a convex curve, and we might at this point naturally think of a generalized linear model with log link. To examine this possibility we might plot log weight versus length as in Figure 9.21. This graph does not support the idea that a log

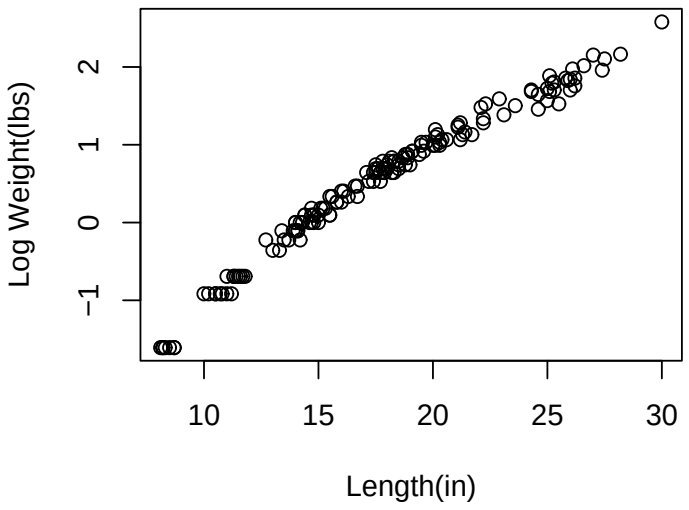


Figure 9.21: Log transformed weight versus length in walleye.

link function would be a good choice to relate $\mu_i \equiv E(Y_i)$ to x_i =length as $\log(\mu_i) = \beta_0 + \beta_1 x_i$.

Mean-Variance Relation

We are by now familiar with the use of Box-Cox plots to investigate possible associations between means and variances. A Box-Cox plot for these data is

presented in Figure 9.22, constructed from 7 groups formed from equal length classes. An ordinary least squares fit to the log group means and log group standard deviations in this figure has slope 0.716, which would suggest that variances increase proportional to means raised to the 1.5 or maybe 2.0 power.

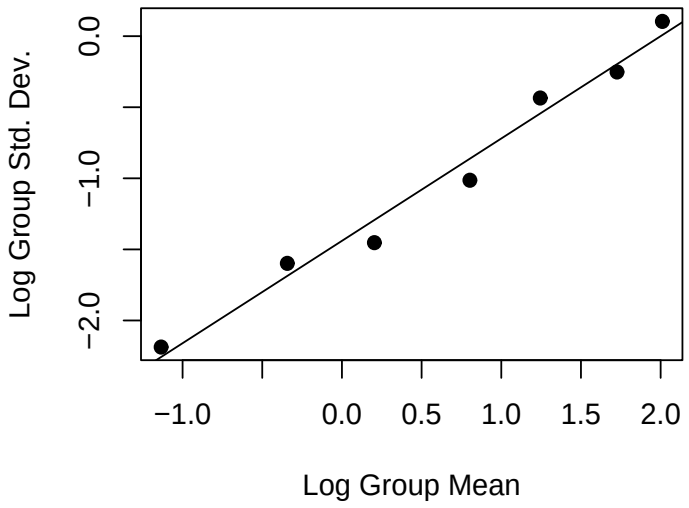


Figure 9.22: Box-Cox plot for length and weight in walleye.

Suppose that in this problem we have no solid scientific basis for choosing any particular form for the systematic model component or expectation function. This is actually not true as we will see later, but for now suppose that in an exploratory approach we are willing to accept this as our state of knowledge. We know that polynomials often provide an extremely flexible tool to fit curves through scatterplots of data. From Figure 9.20 it appears that a quadratic polynomial might provide an adequate fit to these data. A purely

empirical model might then be, for $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sigma\epsilon_i \tag{9.49}$$

where Y_i is connected with the weight of fish i, x_i is the length of fish i, and $\epsilon_i \sim iidF$ with F a location-scale family such that $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$. An ordinary least squares fit of model (9.49) to these data results in the estimated expectation function of Figure 9.23 and the parameter estimates given in Table 9.2. Note that we are producing these estimates as a part of our exploratory analysis, not as estimates for a model that we believe will be entirely reasonable for the problem.

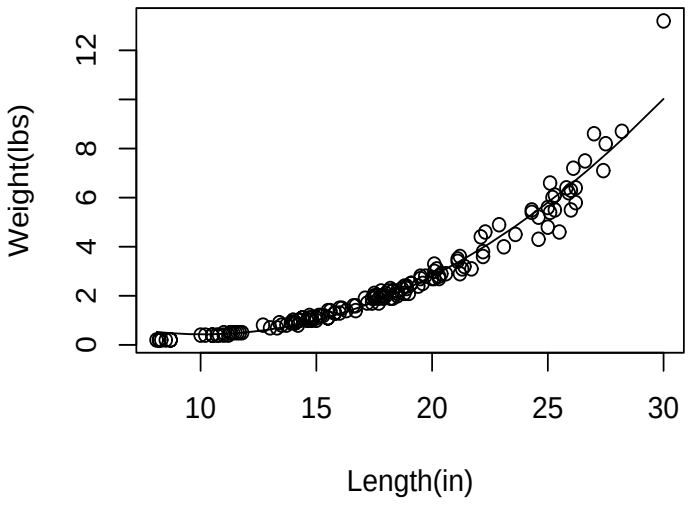


Figure 9.23: Fitted quadratic polynomial expectation function.

The estimated expectation function in Figure 9.23 appears quite reasonable. As we would expect from previous indications, standardized residuals

Parameter	Point Estimate	95% Interval
β_0	2.925	(2.179, 3.672)
β_1	-0.492	$(-0.577, -0.408)$
β_2	0.024	(0.022, 0.027)

Table 9.2: Ordinary least squares estimates for model (9.49).

presented in Figure 9.24 for this fitted model exhibit rather poor behavior, demonstrating non-constant variances.

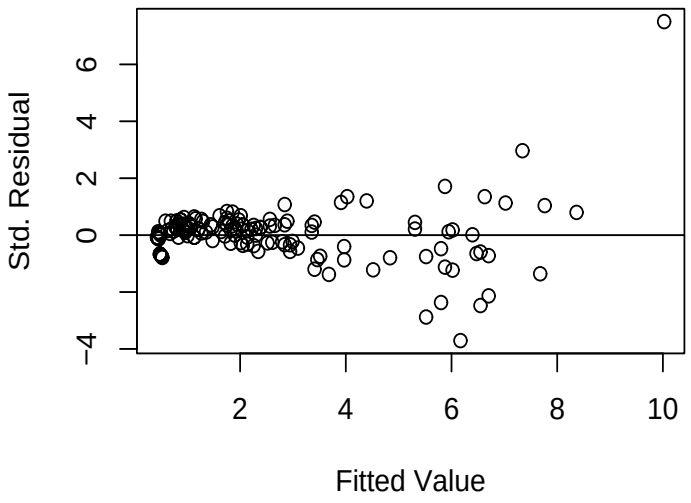


Figure 9.24: Residual plot for the ordinary least squares fit of a quadratic polynomial expectation function.

9.10.2 A Linear Model with Power of the Mean Variances

Our exploratory analysis suggests a model with a linear quadratic expectation function as in (9.49), but with non-constant variances. Given the Box-Cox plot of Figure 9.22, which has a slope of just over 0.70, a power of the mean model for variances might well prove adequate, with a power of 0.75 or perhaps 1.0. The model desired is, for $i = 1, \ldots, n$,

$$Y_i = \mu_i + \sigma\, \mu_i^{\theta}\, \epsilon_i, \tag{9.50}$$

where $\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ and $\theta = 0.75$ or perhaps $\theta = 1.0$.

Non-Bayesian Analysis

A generalized least squares fit of this model can be accomplished through use of a Gauss-Newton algorithm and inference could be based on the Fundamental Theorem of Generalized Least Squares. Fitting the model of expression (9.50) with $\theta = 0.75$ and with $\theta = 1.0$ results in the estimates of Table 9.3. Visually, the fitted expectation functions from either of these models differ little from that of Figure 9.23. Standardized residual plots for these two models are shown in Figure 9.25. Both plots contain one extreme value corresponding to the largest fitted value. Overall, residuals from the model with $\theta = 1.0$ appear a bit more well behaved than those from the model with $\theta = 0.75$, which still suggests variances that increase to some degree with (estimated) expected values. Our current model at this point would be (9.50) with $\theta = 1.0$.

Variance Parameter	Parameter	Point Estimate	95% Interval
$\theta = 0.75$	β_0	1.377	(1.176, 1.579)
	β_1	-0.290	$(-0.320, -0.260)$
	β_2	0.018	(0.017, 0.019)
	σ^2	0.023	
$\theta = 1.0$	β_0	1.130	(0.980, 1.280)
	β_1	-0.251	$(-0.275, -0.226)$
	β_2	0.017	(0.016, 0.018)
	σ^2	0.012	

Table 9.3: Generalized least squares estimates for model (9.50) using $\theta = 0.75$ and $\theta = 1.0$.

Bayesian Analysis

To enact a Bayesian analysis of model (9.50) we might assign prior distributions as $\beta_0 \sim N(\lambda_0, \tau_0^2)$, $\beta_1 \sim N(\lambda_1, \tau_1^2$, $\beta_2 \sim N(\lambda_2, \tau_2^2)$ and $\sigma^2 \sim IG(\xi_1, \xi_2)$ and then take $\pi(\beta_0, \beta_1, \beta_2, \sigma^2) = \pi(\beta_0)\pi(\beta_1)\pi(\beta_2)\pi(\sigma^2)$. For the constant variance model, full conditional posterior distributions of the regression parameters were all n=normal, $p(\beta_0|\cdot) \propto N(M_0, V_0)$, $p(\beta_1|\cdot) = N(M_1, V_1)$ and $p(\beta_2|\cdot) = N(M_2, V_3)$. Let $S_0 = \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_2 x_i^2)$, let $S_1 = \sum_{i=1}^n (y_i - \beta_0 - \beta_2 x_i^2)x_i$ and let $S_2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)x_i^2$. Also let $S_3 = \sum_{i=1}^n x_i^2$ and $S_4 = \sum_{i=1}^n x_i^4$. The means and variances of these conditional posteriors are, then,

$$
\begin{aligned}
M_0 &= \frac{\tau_0^2 S_0 + \sigma^2 \lambda_0}{\tau_0^2 n + \sigma^2} & V_0 &= \frac{1}{\tau_0^2 n + \sigma^2}, \\
M_1 &= \frac{\tau_1^2 S_1 + \sigma^2 \lambda_1}{\tau_1^2 S_3 + \sigma^2} & V_1 &= \frac{1}{\tau_1^2 S_3 + \sigma^2}, \\
M_2 &= \frac{\tau_2^2 S_2 + \sigma^2 \lambda_2}{\tau_2^2 S_4 + \sigma^2} & V_2 &= \frac{1}{\tau_2^2 S_4 + \sigma^2}.
\end{aligned}
$$

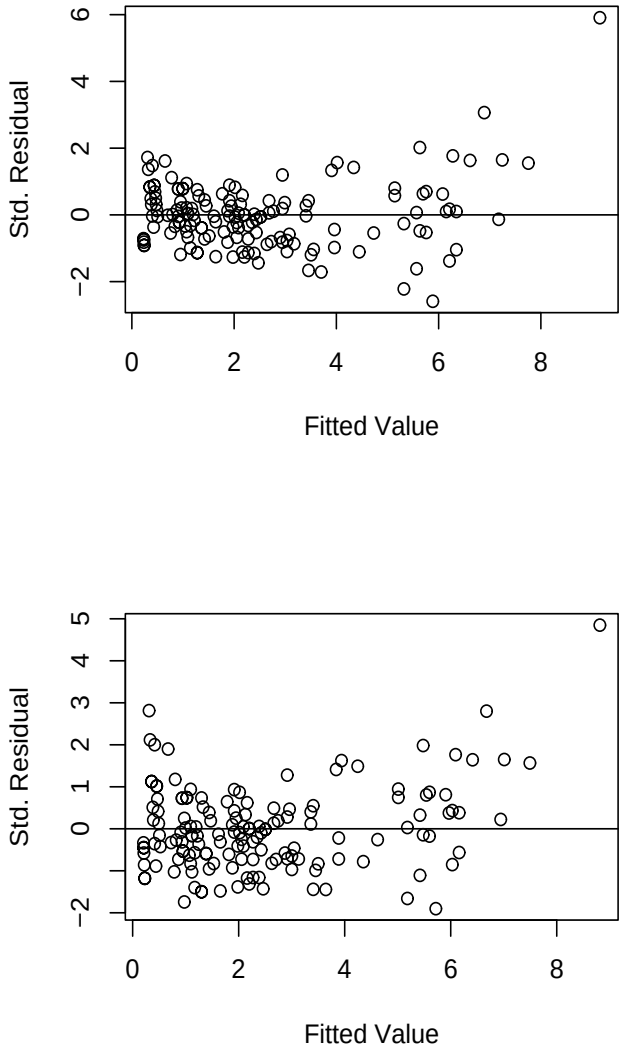


Figure 9.25: Standardized residuals from generalized least squares fits of the model of expression (9.50) with $\theta = 0.75$ (upper) and $\theta = 1.0$ (lower).

Finally, letting $S_5 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$, the full conditional posterior of σ^2 is inverse gamma with parameters,

$$\xi_1 + n/2 \qquad \text{and} \qquad \xi_2 + \frac{S_5}{2}.$$

We can simulate from the joint posterior of $(\beta_0, \beta_1, \beta_2, \sigma^2)$ for this constant variance model using a standard Gibbs Sampling algorithm, sampling sequentially from the above conditional posteriors. That algorithm was run with prior parameters $\lambda_0 = 5$, $\tau_0^2 = 10$, $\lambda_1 = 0$, $\tau_1^2 = 10$, $\lambda_2 = 0$, $\tau_2^2 = 10$, $xi_1 = 1$ and $xi_2 = 1$. The algorithm was run again using $\xi_1 = 0.5$ and $\xi_2 = 0.5$ with essentially no change in results. The algorithm turns out to have extremely slow mixing properties for the regression parameters but quite rapid mixing for σ^2. Ultimately, a burn-in of $100,000$ iterations followed by the collection of $M = 500,000$ values was used to approximate the posterior.

For the two power-of-the-mean models, one with $\theta = 0.75$ and one with $\theta = 1.0$, the full conditional posterior distributions of the β_j, for $j = 0, 1, 2$ are,

$$\begin{aligned}
p(\beta_0|\cdot) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \theta, \sigma^2)\, \pi(\beta_j) \\
&\propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\{ \frac{y_i - \mu_i(\boldsymbol{\beta})}{\mu_i^{\theta}(\boldsymbol{\beta})} \right\}^2 - \frac{1}{2\tau_j^2}(\beta_j - \lambda_j)^2 \right] \\
p(\sigma^2|\cdot) &\propto \frac{1}{(\sigma^2)^{\xi_1 + n/2 + 1}} \exp\left[-\left(\xi_2 + (1/2) \sum_{i=1}^{n} \left\{ \frac{y_i - \mu_i(\boldsymbol{\beta})}{\mu_i^{\theta}(\boldsymbol{\beta})} \right\}^2 \right) / \sigma^2 \right],
\end{aligned}$$

where $\mu_i(\boldsymbol{\beta}) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. These distributions are not easily simplified. The conditional posterior for σ^2 can be recognized as an inverse gamma distribution with parameters,

$$\xi_1 + n/2 \qquad \text{and} \qquad \xi_2 + \frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\mu_i^{\theta}} \right)^2.$$

Overall Gibbs Sampling algorithms with separate Metropolis steps for β_0, β_1, and β_2 were run for the two fixed values of $\theta = 0.75$ and $\theta = 1.0$ and with the same prior parameter values used for the constant variance model. Jump proposals for the three regression parameters were random walks.

With $\theta = 0.75$ chains seemed to mix more rapidly than for the constant variance model. Autocorrelation functions for the parameters all decreased to negligible levels after $4,500$ to $5,000$ iterations, resulting in the choice of a burn-in period of $10,000$ iterations, to be conservative. Tuning of the algorithm resulted in jump proposal variances of 0.001 for β_0, 0.00005 for β_1 and 0.0000005 for β_2. While these variances appear quite small, trace plots for the parameters showed considerable movement over sufficient iterations. Acceptance rates were 0.40, 0.16 and 0.11 for β_0, β_1, and β_2, respectively. Following burn-in, $M = 500,000$ values were collected to approximate the posterior.

With $\theta = 1.0$, mixing was even more rapid than for $\theta = 0.75$, with autocorrelation functions dying off by 50 to 100 iterations. The result was a burn-in of $2,500$ iterations followed by collection of $M = 500,000$ values. Jump proposal variances were the same as used with $\theta = 0.75$ and acceptance rates were 0.29, 0.12, and 0.10 for β_0, β_1 and β_2, respectively.

Posterior means and 95% credible intervals are given for all three models, constant variance, power-of-the-mean with $\theta = 0.75$, and power-of-the-mean with $\theta = 1.0$ in Table 9.4. Estimated regression parameters were similar for the two power-of-the-mean models, with these differing somewhat from values for the constant variance model, particularly for β_0 and β_1. The greatest difference between the constant variance model and the two with non-constant variances, however, was in the posterior values for σ^2.

Bayesian residual plots were constructed for the model with constant variance and the power of the mean models with $\theta = 0.75$ and $\theta = 1.0$. Each of

	Model		
Param.	Const. Variance	POM ($\theta = 0.75$)	POM ($\theta = 1.0$)
β_0	2.946 $(2.274, 3.623)$	1.438 $(1.187, 1.718)$	1.146 $(0.985, 1.337)$
β_1	-0.495 $(-0.572, -0.419)$	-0.299 $(-0.340, -0.263)$	-0.253 $(-0.283, -0.227)$
β_2	0.024 $(0.022, 0.026)$	0.019 $(0.017, 0.020)$	0.017 $(0.016, 0.018)$
σ^2	0.184 $(0.147, 0.231)$	0.025 $(0.020, 0.031)$	0.018 $(0.014, 0.022)$

Table 9.4: Posterior means and intervals for quadratic regressions fit to Walleye length and weight data.

these residual assessments required computation of $\mu_i(\boldsymbol{\beta}^{(m)})$ for $m = 1, \ldots, M$ and each $i = 1, \ldots, n$. Here, $M = 500,000$. For the constant variance model residuals (9.47) were computed for each of the $500,000$ Monte Carlo iterations, and the median of these residuals for each $i = 1, \ldots, n$ plotted against the median value of the $\{\mu_i(\boldsymbol{\beta}^{(m)} : m = 1, \ldots, M\})$. The same procedure was employed for the power of the mean models with and residual quantities (9.48). Figure 9.26 contains these residuals plots.

Consistent with previous results, the constant variance model is not supported. Plots for both power of the mean models contain one extreme residual corresponding to the largest fitted value. Aside from this one value, the plot for $\theta = 0.75$ appears more balanced between positive and negative residuals than that for $\theta = 1.0$, but there may remain a slight indication of increasing variance which is lacking for $\theta = 1.0$. Thus, the residual plot for $\theta = 0.75$ supports a random model component that is symmetric and with variances that increase just slightly faster than $\mu_i^{1.5}$, while the plot for $\theta = 1.0$ supports a skew right random component with variances proportional to μ_i^2. However, the major point to be gleaned from these plots is that both power of the mean models are vastly superior to the constant variance model.

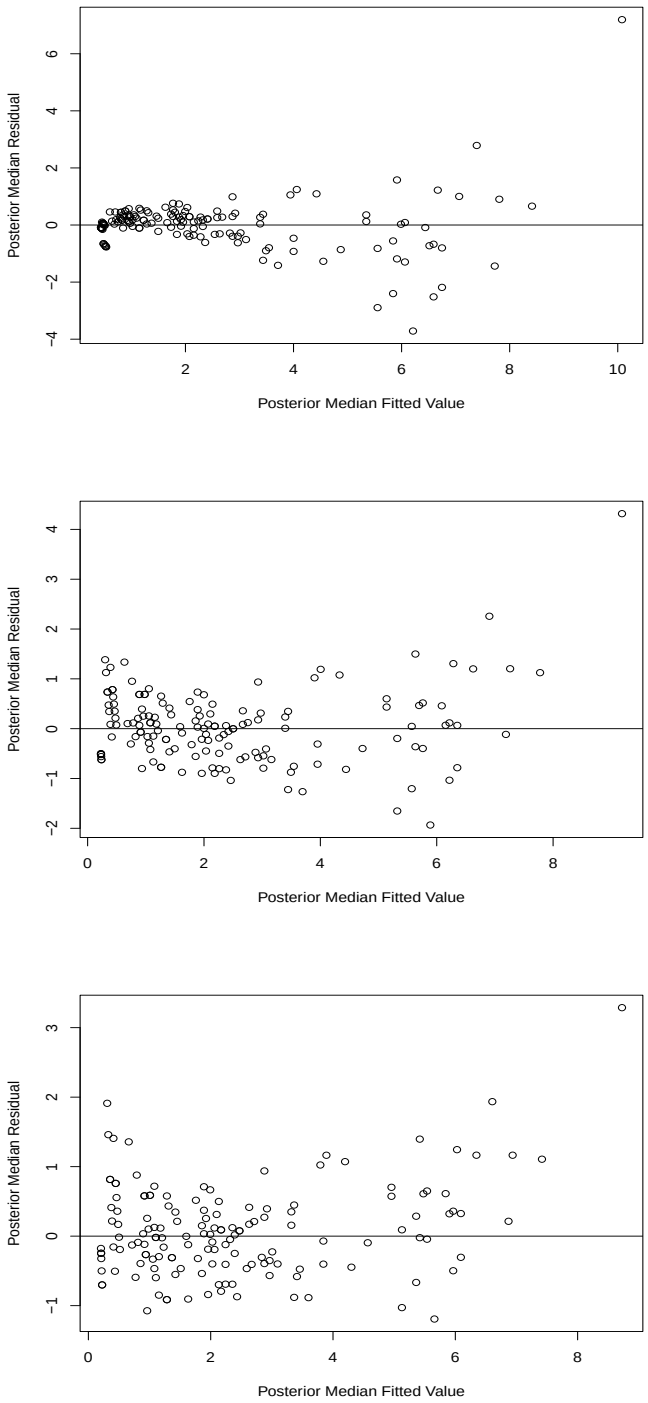


Figure 9.26: Bayesian residual plots for quadratic regression with constant variance (upper) and power of the mean ($\theta = 0.75$ middle and $\theta = 1.0$ lower) models fit to walleye data on weights versus lengths.

9.10.3 Developing a Model with Greater Scientific Potential

The polynomial models of expression (9.50) with θ somewhere in the range of 0.75 to 1.0 would appear entirely adequate for this problem if our only objective were to predict weight from length for fish from the given set of lakes in the time period of data collection. However, the regression coefficients lack interpretability within the context of the problem. For example, it would not be reasonable to conclude from these models that the linear part of the association between length and weight is negative ($\hat{\beta}_1 < 0$ for any of the polynomial models). If these parameters did have some scientific meaning, the previously noted property that different sets of values for β_0, β_1 and β_2 can lead to similar functions would be troubling. In addition, the description of this problem suggested that observation of length and weight might have potential for providing information about how the morphology (i.e., body shape) of fish species affects growth. In this section we wish to examine the possibility of developing a model with greater potential for scientific interpretation than possessed by the pure empirical descriptions provided by a quadratic expectation function.

We have seen in Figure 9.21 that a systematic model component or expectation function corresponding to a log link as $\log(\mu_i) = \beta_0 + \beta_1 x_i$ would likely prove inadequate for description of these data. However, if we plot log weight against log length we obtain the scatterplot of Figure 9.27, for which a straight line would seem to provide quite a good description. If, without any errors, $log(Y_i) = \beta + \alpha log(x_i)$, then $Y_i = \exp(\beta)\, x_i^{\alpha}$, and this suggests a nonlinear model as, for $i = 1, \ldots, n$,

$$Y_i = \mu_i + \sigma\, \mu_i^{\theta}\, \epsilon_i, \tag{9.51}$$

where $\mu_i = \beta\, x_i^{\alpha}$ and $\epsilon_i \sim iidF$ for F a location-scale family with $E(\epsilon_i) = 0$

and $var(\epsilon_i) = 1$. As before we might take $\theta = 0.75$ or $\theta = 1.0$.

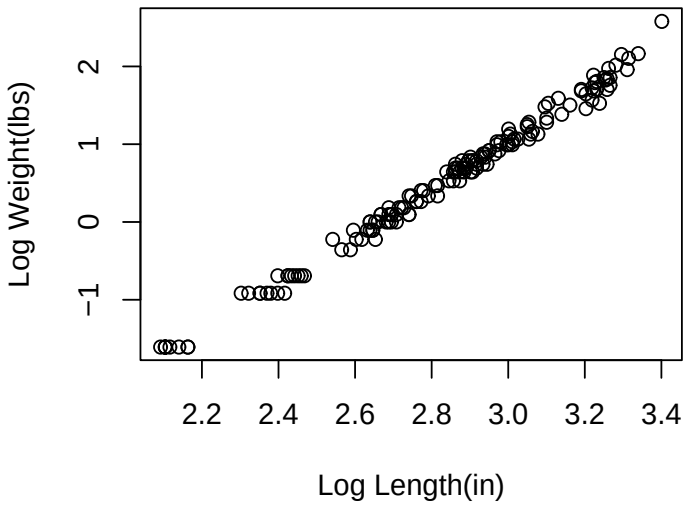


Figure 9.27: Scatterplot of log weight against log length.

This model suggests that we might represent weight as proportional to a power of length. If we would consider density of fish tissue as roughly constant in all regions of a fish body (not strictly true, but perhaps adequate for our purposes) the weight should be proportional to volume. And, although fish are not cuboids, we can envisage the "corresponding" cuboid with volume being the product of length, width, and height. If width and height increase proportionally to length, then using V for volume, L for length, W for width, and H for height we should have $W = k_1 L$, $H = k_2 L$, and $V = L\,W\,H = L\,k_1 L\,k_2 L = \tilde{k} L^3$. If, on the other hand, the geometry of a fish species is such

that increases in width and height are not proportional to increases in length as a fish grows, then we might well have $V \propto L^{\alpha}$ for some power $\alpha \neq 3$. The power α should then be characteristic of a given species, representing the biological reality of species-specific body morphology.

Non-Bayesian Analysis

An ordinary least squares fit of the data points in Figure 9.27 results in an intercept of -8.2 and a slope of 3.1, which gives what should be reasonable starting values for generalized least squares estimation as $\beta^{(0)} = 0.00027$ and $\alpha^{(0)} = 3.1$. Using these starting values and fitting the model of expression (9.51) using generalized least squares results in the estimated parameter values of Table 9.5 and the fitted regression function shown in Figure 9.28 which was visually identical for $\theta = 0.75$ and $\theta = 1.0$. The estimates of σ^2 were produced using the usual moment-based approach described in Chapter 6.2.4.

Variance Parameter	Parameter	Point Estimate	95% Interval
$\theta = 0.75$	β	0.00024	(0.00020, 0.00028)
	α	3.127	(3.069, 3.186)
	σ^2	0.0136	
$\theta = 1.0$	β	0.00027	(0.00023, 0.00030)
	α	3.084	(3.035, 3.132)
	σ^2	0.0086	

Table 9.5: Generalized least squares estimates for model (9.50) using $\theta = 0.75$ and $\theta = 1.0$.

Standardized residual plots for the fitted models with $\theta = 0.75$ and $\theta = 1.0$ are shown in Figure 9.29. Even ignoring the upper right most point, it again

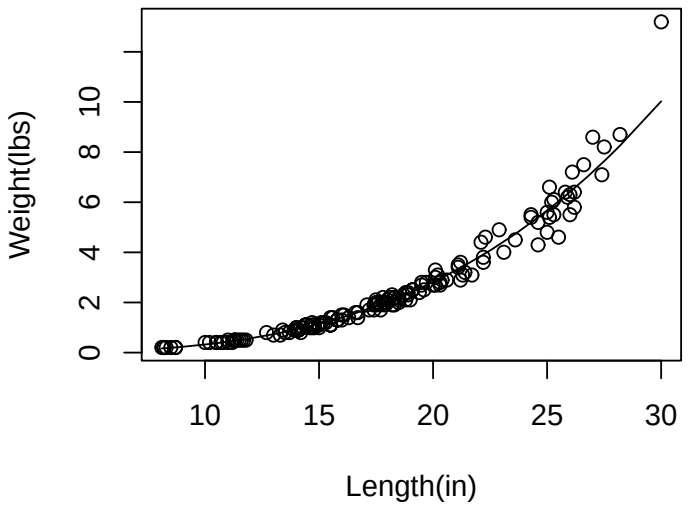


Figure 9.28: Estimated expectation function for the model of expression (9.51).

does not appear that a power of $\theta = 0.75$ is sufficiently strong to account for the unequal variances. In total, this analysis shows that model (9.51) with a power just a bit greater than 3 provides a good fit to the available data; neither confidence interval for α in Table 9.5 includes 3. It would also seem that a model with $\theta = 1.0$ is slightly preferable to one with $\theta = 0.75$, which implies that variances increase in proportion to the square of expected values in these data. Estimation of the regression function itself, however, appears to be robust to specification of the power θ in the variance model.

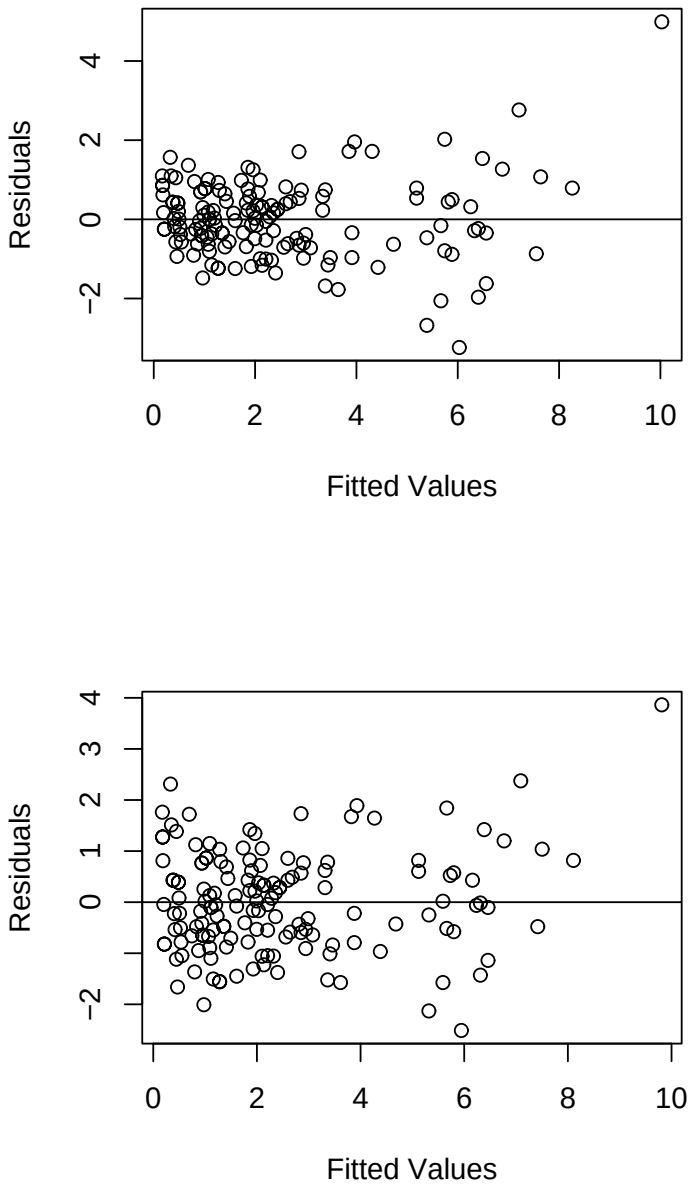


Figure 9.29: Standardized residuals for the model of expression (9.51) with $\theta = 0.75$ (upper) and $\theta = 1.0$ (lower).

Bayesian Analysis

To conduct a Bayesian analysis of model (9.51) we would like to specify prior distributions for α, β, and σ based only on information that is completely divorced from the current data. A web page operated by the Wisconsin Dept. of Natural Resources

`verb|https://dnr.wisconsin.gov/topic/Fishing/questions/estfishweight`

concerns the topic of Estimating Fish Weight. This web site gives the formula for estimating the weight of walleye (in pounds) based on length (in inches) as weight = length^3/2700. Other sources such as

`https://targetwalleye.com/whats-the-best-walleye-length-to-weight-formula`

claim that length alone is insufficient to estimate weight and what is needed is both length and girth, then weight=length $\times$ girth^2/750. Girth can apparently be quite variable but will almost certainly be less than length. The New York State Dept. of Environmental Conservation

`https://dec.ny.gov/things-to-do/freshwater-fishing/learn-to-fish/tips-skills/use-ruler-to-weigh`

does not give a formula, but contains a table in which weight increases more slowly than the formula given by the Wisconsin Dept. of Natural Resources. In terms of (9.51) the Wisconsin information suggests that $\alpha = 3$ and $\beta = 1/2700$. Considering the other sources, if girth is proportional to length as girth=δ length for $\delta < 1$, then we could still be successful with (9.51) but would expect a smaller leading factor, $\beta = \delta^2/2700$. If girth is not necessarily proportional to length but can vary considerably for the same length, then we would expect

that $\alpha \neq 3$ and the model may be difficult to fit at all. These considerations might motivate prior distributions of the following forms.

$$\begin{aligned}\beta &\sim \mathrm{N}(\lambda, \tau^2)\\ \alpha &\sim \mathrm{Gamma}(\psi_1, \psi_2)\\ \sigma^2 &\sim \mathrm{IG}(\xi_1, \xi_2)\end{aligned}$$

Based on the formula provided by the Wisconsin DNR we might choose parameters to center β somewhere in the neighborhood of 1/2700 with a value of τ^2 large enough to give a fairly diffuse prior. The values selected here were $\lambda = 0.0004$ and $\tau^2 = 0.01$, which results in a large coefficient of variation of 250. Based on the same information, parameters for the prior of α were selected as $\psi_1 = 3.0$ and $\psi_2 = 1.0$ giving a prior mean of 3.0 and prior variance also 3.0, which results in a probability of $1 < \alpha < 5$ of roughly 0.90. It seems unlikely that α would be anywhere outside of this range. The prior for σ^2 was selected to give conditional conjugacy and a sensitivity analysis was conducted with parameter values of $\xi_1 = \xi_2 \in \{0.01, 0.1, 1.0\}$ which had little to no effect on the results of analysis.

An overall Gibbs algorithm was employed for estimation and inference. The joint data model is,

$$f(\boldsymbol{y}|\beta, \alpha, \sigma^2) = \frac{1}{[2\pi\sigma^2\mu_i^{2\theta}]^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{\mu_i^{\theta}}\right)^2\right], \tag{9.52}$$

where $\mu_i = \beta x_i^{\alpha}$. The prior distributions for β, α and σ^2 are,

$$\begin{aligned}\pi(\beta) &\propto \exp\left[-\frac{1}{2\tau^2}(\beta - \lambda)^2\right],\\ \pi(\alpha) &\propto \alpha^{\psi_1 - 1}\exp(-\psi_2\alpha),\\ \pi(\sigma^2) &\propto \frac{1}{(\sigma^2)^{\xi_1+1}}\exp(-\xi_2/\sigma^2).\end{aligned} \tag{9.53}$$

The full conditional posterior distributions of β and α do not simplify and have the forms

$$p(\beta|\cdot) \propto f(\boldsymbol{y}|\beta, \alpha, \sigma^2)\, \pi(\beta)$$
$$p(\alpha|\cdot) \propto f(\boldsymbol{y}|\beta, \alpha, \sigma^2)\, \pi(\alpha).$$

These conditional posteriors will be sampled using a Metropolis-Hastings algorithm at each iteration of the overall Gibbs Sampler. As already mentioned, the conditional posterior of σ^2 is inverse gamma in form with parameters

$$\xi_1 + \frac{n}{2} \text{ and } \xi_2 + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{\mu_i^{\theta}}\right)^2.$$

An overall Gibbs Sampling algorithm using the full conditionals just presented proved to mix slowly in the dimensions of β and α, but very rapidly with respect to σ^2. Running a sequence of chains with longer and longer durations showed that autocorrelations became negligible at about $2,500$ iterations for chains of length at least $10,000$. To be conservative, burn-in was selected to be $B = 250,000$ and the subsequent $M = 1,0000,000$ values were sampled. After tuning random walk jump proposals to have variances 10^{-10} for β and 0.01 for α, acceptance rates were 0.269 for β and 0.339 for α. Posterior means and 95% credible intervals for a model with $\theta = 1.0$ are given in Table 9.6.

Parameter	Posterior Mean	95% Credible Interval
β	0.00027	$(0.00023, 0.0003)$
α	3.0880	$(3.0380, 3.1370)$
σ^2	0.0097	$(0.0077, 0.0122)$

Table 9.6: Posterior means and 95% credible intervals for model (9.51) fit to Walleye length and weight data with $\theta = 1.0$.

Posterior means of all three parameters in this model are nearly identical to the generalized least squares estimates of Table 9.5 and the posterior mean estimated regression function was visually identical to that of Figure 9.28. A plot of the median Bayesian residuals (9.48) against the posterior median fitted values is presented in Figure 9.30. This residual plot also looks quite similar to the standardized residual plot resulting from generalized least squares estimation for $\theta = 1.0$ in the lower panel of Figure 9.29.

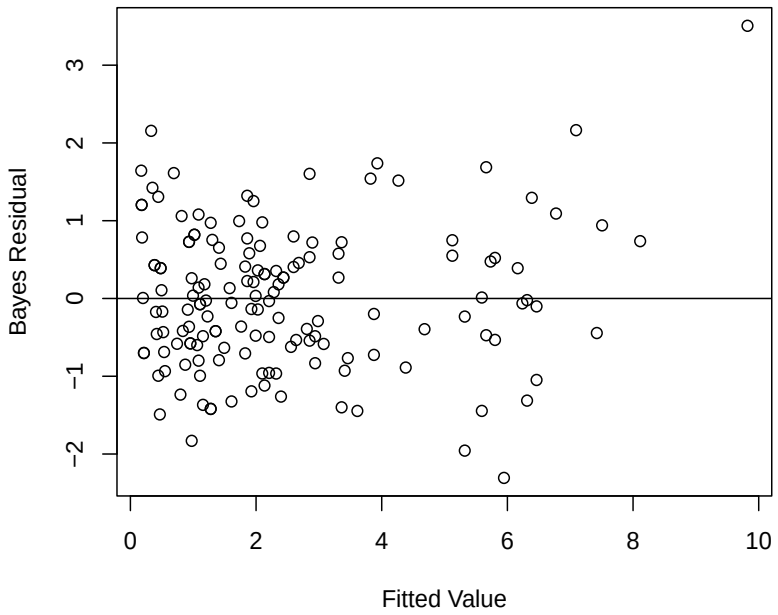


Figure 9.30: Bayesian residual plot for analysis of model (9.51).

Conclusions

There are a number of conclusions that can be reached as a result of careful analysis of this case study, about both the statistical methods used and the implications for walleye. First, results from a frequentist analysis based on generalized least squares estimation and those from a Bayesian analysis were quite similar across all of the models considered. Posterior expected values were quite close to generalized least squares point estimates, and even intervals were similar. Polynomial expectation functions are flexible tools for fitting curves to data, but lack interpretability within the context of the problem. A simpler model (in terms of number of parameters) formulated on the basis of just thinking about how weight might be related to length in fish and using crude geometry was able to fit the data at least as well as the polynomials, and with one less parameter. From a purely statistical viewpoint, this is an interesting data set. The relation between expected values and variances is clearly exhibited in the data, and yet it would also appear that a symmetric random component is appropriate to describe the data. In fact, it seems that variances are proportional to the square of the expected values, a pheonomenon we have used previously as an indication that perhaps a gamma or a lognormal random model component is called for. That is clearly not the case for these data. If we were to estimate quantiles of responses (or predict responses) based on these data, a normal random component or, equivalently, a normal distribution for additive error terms would likely be entirely appropriate. In terms of the relation between weight and length in walleye, it would appear that weight is roughly proportional to the cube of length in Walleye or perhaps just a bit higher power, as none of the intervals for α in any of the models fit by either least squares or Bayesian methods include 3.0. Based on searching

for information on walleye length and weight relations, this may be affected by geographic factors. The data for this analysis were collected in Minnesota and the formula promoted by the Wisconsin DNR appears quite a good approximation. There are indications from other regions (e.g., New York) that length alone is not adequate to reflect weight. Whether or not this is true would require additional data sets on which to examine the hypothesis. But, if weight is proportional to the cube of weight, either in a particular region or more generally, the implication is that the *condition* of walleye among various situations (lakes or sets of lakes in different regions, for example) could be compared through the parameter β in the model of expression (9.51).