Chapter 9

Model Assessment

In statistical modeling, once one has formulated a model and produced estimates and inferential quantities, the question remains of whether the model is adequate for its intended purpose. This may well involve issues other than whether the model seems to describe the available data in a satisfactory manner, depending on the objectives of the analysis conducted (see Chapter 7.1). Nevertheless, the two cornerstones of data-driven model assessment are examination of how well the fitted model describes the observed data, and how well the model predicts observations, and these issues will be the focus of our presentation in this chapter. Even here, however, there are questions regarding which components of a model should be the focus of assessment that depend on the objectives of analysis. Does interest center on a description of the systematic model component? On the modeled distribution more completely (e.g., quantiles of the distribution of responses)? On the ability of the model to predict unobserved random variables (within the extent of the available data)? Or, perhaps the degree and form of departures from a theoretical relation among variables is of central concern. Certainly, there is much overlap in
the manner one might approach these questions, but there may well be unique issues involved as well which simply indicates that model assessment is not a “one size fits all” activity.

Our goal in this chapter is to organize some of the main procedures used in assessment of statistical models, not to present a catalog of all (or even most) of the types of plots, tests, and assessment criteria that have been developed. Many useful procedures are fairly model specific, having been developed for certain types of models that become popular in application. In addition, the number of diagnostics and other assessment methods developed for linear models far exceeds the number and sophistication of methods developed for most other types of models, and has resulted in any number of book-length treatments of the subject (e.g., Belsley, Kuh, and Welsch, 1980; Cook and Weisberg, 1982). Such detailed procedures should be sought out and utilized when appropriate in a particular problem of data analysis. But what might one think of when faced with a model that has been formulated for a specific problem rather than drawn from a standard list of existing model types? In keeping with the theme of these notes, what are ways one might approach model assessment? Three major approaches to model assessment are the use of residuals, cross-validation, and simulation-based assessment.

9.1 Analysis of Residuals

Every student who has completed a basic course in regression is aware of the usefulness of residuals in assessing linear regression models; indeed, it has been assumed in previous chapters that readers were familiar with basic residual plots. Intuitively, residuals are a direct gauge of how far we have “missed” the target in a signal plus noise formulation of a model. Here, basic residuals
have the form of \( \{ y_i - \hat{y}_i : i = 1, \ldots, n \} \), where the combination of all of the influences (i.e., signals) that produce an observed value \( y_i \) is contrasted with the (estimated) signals incorporated in a model that produce the fitted or predicted value \( \hat{y}_i \). If all of the primary signals that are important in producing \( y_i \) have been (nearly) correctly modeled, then these residuals should reflect primarily measurement error. But, as we will see, there are other types of residuals as well, that may be useful in assessing aspects of a proposed model other than how well it reflects signal as modeled through expected values.

### 9.1.1 A General Notational Framework

Throughout this section we will rely on a general notation framework built around the concept of a random field. Let \( \{ Y(s_i) : i = 1, \ldots, n \} \) denote a set of random variables connected with observable quantities, with \( s_i \) a non-random “location variable”. Several possibilities for the location variables \( s_i \) are:

1. Independent random variables.
   
   Here, we would naturally take \( s_i = i \) and simplify notation by referring to \( Y(s_i) \) as just \( Y_i \).

2. Groups of random variables.
   
   Here, we might define \( s_i = (k, j) \) where \( k \) indexes group and \( j \) indexes observation within group; \( k = 1, \ldots, K \) and \( j = 1, \ldots, n_k \).

3. Geographic random variables.
   
   Here, we might take \( s_i = (u_i, v_i) \), where \( u_i \) denotes latitude and \( v_i \) longitude, or \( u_i \) denotes horizontal coordinate on a grid and \( v_i \) denotes vertical coordinate on a grid.
4. Time series of random variables.

Here we might take $s_i = t$ where $t$ is time, if each $Y(s_i)$ occurs at a unique time, or $s_i = (t, j)$ where $t$ is time and $j$ is observation number at time $t$; $t = 1, \ldots, T$ and $j = 1, \ldots, n_t$.

We assume that each $Y(s_i)$ is modeled through a parametric distribution having a density (or mass function) $f_i$, depending on parameter $\psi(s_i)$ through the data model,

$$f_i(y(s_i)|\psi(s_i)); \ y(s_i) \in \Omega_i.$$  \hfill (9.1)

Here, the densities $f_i$ are indexed by $i$ to allow for the possibility of differing covariates $x_i$ or auxiliary information (e.g., binomial sample size).

We will assume that the parameters $\{\psi(s_i) : i = 1, \ldots, n\}$ represent “minimal” parameters in the sense that any other parameters used in writing the densities $f_i; i = 1, \ldots, n$ are functions of the $\psi(s_i)$, and also that we may write,

$$\psi(s_i) = (\psi_f(s_i), \psi_r(s_i)),$$  \hfill (9.2)

where $\psi_f(s_i)$ represents parameters that are fixed in the data model and $\psi_r(s_i)$ denotes parameters that are random in the data model. We take $\psi_r(s_i)$ to have a distribution with parameterized density $g_i(\psi_r(s_i)|\lambda)$, where this density may result from marginalization over any additional levels of random terms in the model. For example, if $\psi_r(s_i)$ is modeled directly in terms of a distribution $g_{1,i}(\psi_r(s_i)|\lambda_1(s_i))$ with $\lambda_1(s_i)$ having a distribution with density $g_2(\lambda_1(s_i)|\lambda)$, then,

$$g_i(\psi_r(s_i)|\lambda) = \int g_{1,i}(\psi_r(s_i)|\lambda_1(s_i))g_2(\lambda_1(s_i)|\lambda) \ d\lambda_1(s_i).$$  \hfill (9.3)

Finally, we then take the marginal density of $Y(s_i)$ to be given by

$$h_i(y(s_i)|\psi_f(s_i), \lambda) = \int f_i(y(s_i)|\psi_f(s_i), \psi_r(s_i)) \ g(\psi_r(s_i)|\lambda) \ d\psi_r(s_i).$$  \hfill (9.4)
This notation is sufficient to cover most of the models we have discussed. In particular, we have not considered any model that contains both fixed and random parameter components in the second (mixing) or higher levels.

Example 9.1
Consider a typical linear regression model with independent response variables, written as

\[ Y_i = x_i^T \beta + \sigma \epsilon_i; \quad i = 1, \ldots, n. \]

This model fits into our general notation by defining \( s_i \equiv i \) and \( \psi_f(s_i) \equiv (\beta, \sigma^2) \) and dropping remaining elements of the structure; there is no \( \psi_r(s_i) \) or density \( g \).

Example 9.2
We have written a standard generalized linear model as in expressions (7.19) through (7.21) in Section 7.3.2, namely with responses independent and

\[
\begin{align*}
  f(y_i|\theta_i, \phi) &= \exp \left[ \phi \{ y_i \theta_i - b(\theta_i) \} + c(y_i, \phi) \right], \\
  \mu_i &= b'(\theta_i) \\
  \eta_i &= x_i^T \beta \\
  g(\mu_i) &= \eta_i
\end{align*}
\]

which fits into our general notation with \( s_i \equiv i \) and \( \psi_f(s_i) \equiv (\beta, \phi) \). Note here that all intermediate parameters can be written in terms of these fixed values as

\[
\eta_i(\beta) = x_i^T \beta; \quad \mu_i(\beta) = g^{-1}(\eta_i(\beta)); \quad \theta_i(\beta) = b'^{-1}(\mu_i(\beta)).
\]
Example 9.3
A beta-binomial mixture model was presented in expressions (7.34) and (7.35) for a set of independent random variables as,

\[ f_i(y_i | \theta_i) \propto \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i}, \]

\[ g(\theta_i | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \theta_i^{\alpha - 1} (1 - \theta_i)^{\beta - 1}, \]

\[ h_i(y_i | \alpha, \beta) \propto \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \theta_i^{\alpha + y_i - 1} (1 - \theta_i)^{\beta + n_i - y_i - 1}. \]

This fits into our general notation with \( s_i \equiv i, \psi_r(s_i) = \theta_i \) and \( \lambda \equiv (\alpha, \beta). \)

Example 9.4
A model equivalent to one presented in expression (7.24) for data with random cluster or group effects is,

\[ Y_{i,j} = \mathbf{x}_{i,j}^T \beta + \delta_j I(i \in C_j) + \sigma \epsilon_{i,j}; \quad \delta_j \sim iidN(0, \tau^2); \quad \epsilon_{i,j} \sim iidN(0, 1), \]

where \( j \) indexes group and \( i \) indexes observation within group. To put this into our general notation, define \( s_i \equiv (i, j), \psi_f(s_i) \equiv (\beta, \sigma^2), \psi_r(s_i) \equiv \delta_j, \) and \( \lambda \equiv \tau^2. \)

Corresponding to a model written as in expressions (9.1) through (9.4), we assume parameter estimates are available for the components of \( \psi_f(s_i) \) and \( \lambda \) and, where applicable, predictors are available for the components of \( \psi_r(s_i). \) Also, we assume that these estimates and predictors lead to estimated expected values \( \hat{\mu}(s_i) \equiv \hat{E}\{Y(s_i)\} \) for any \( s_i \) in our set of observed locations and predicted values \( \hat{p}_Y(s_0) \) for any \( s_0 \) not in our set of observed locations. If appropriate for a given model, we also assume that estimators of \( E\{\psi_r(s_i)\} \) are
available which will be considered in this section as predictors of the random data model parameters as $\hat{p}_\psi(s_i) \equiv \hat{E}\{\psi_r(s_i)\}$.

### 9.1.2 Types of Residuals

As eluded to in the introductory comments to this chapter, there are any number of quantities we might label as "residuals" in particular models. It would seem we may place the majority of such quantities into the broad categories discussed in this subsection.

**Raw and Absolute Residuals**

The most basic form of residuals are what we can call *raw residuals*, defined as,

$$r(s_i) = y(s_i) - \hat{\mu}(s_i)$$  \hspace{1cm} (9.5)

Raw residuals can be useful in their own right in simple models (e.g., simple linear regression) in which they reflect the same behaviors as more sophisticated residual quantities, and in extremely complex models where we have not yet developed the ability to make use of more refined values. In addition, raw residuals are basic building block for many other residual quantities as they clearly embodied what we intuitively think of as a "residual". Absolute residuals $a(s_i) = |r(s_i)|$ are often useful in detecting patterns of unequal variances and Carroll and Ruppert (1988, p.30) call absolute residuals "the basic building blocks in the analysis of heteroscedasticity" in regression. Any number of transformations of raw and absolute residuals are also useful in certain situations. We defer a discussion of such transformations until the section of this chapter that deals with residual plots since such transformations do not seem to represent truly different types of residuals than the basic underlying
unadjusted quantities.

**Studentized Residuals**

The use of raw residuals would seem to be well suited for examination of many additive error models, since they represent our “estimates” of the noise component in a model conceptualized as signal plus noise. But in most additive error models, raw residuals do not possess constant variance (even if the model error terms $\epsilon_i$ do). It is typically desirable then to use studentized residuals, which should have (at least approximately) constant variance.

In general, consider an additive error model of the form

$$Y(s_i) = \mu(s_i) + \sigma(s_i) \epsilon(s_i),$$

where $\epsilon(s_i) \sim iidF$, $E\{\epsilon(s_i)\} = 0$ and $var\{\epsilon(s_i)\} = 1$ for $i = 1, \ldots, n$. Consider, for the time being, that the $\sigma(s_i)$ are known, but that the $\mu(s_i)$ are to be estimated. This model, along with the definition of raw residuals in (9.5), indicates that the random form of residuals is,

$$R(s_i) = Y(s_i) - \hat{\mu}(s_i)$$

$$= \mu(s_i) - \hat{\mu}(s_i) + \sigma(s_i) \epsilon(s_i).$$

Then,

$$var\{R(s_i)\} = var\{\hat{\mu}(s_i)\} + \sigma^2(s_i) - 2\sigma(s_i) cov\{\hat{\mu}(s_i), \epsilon(s_i)\},$$

and we can define studentized residuals as, for $i = 1, \ldots, n$,

$$b(s_i) = \frac{r(s_i)}{\left[var\{\hat{\mu}(s_i)\} + \sigma^2(s_i) - 2\sigma(s_i) cov\{\hat{\mu}(s_i), \epsilon(s_i)\}\right]^{1/2}}. \quad (9.6)$$

In (9.6) we usually have means $\mu(s_i)$ modeled in terms of a $p$–dimensional parameter $\beta$ with $p < n$, and the first term in the denominator becomes a
9.1. ANALYSIS OF RESIDUALS

function of the variance of $\hat{\beta}$. Of course, it is not the case that the data model variances $\sigma^2(s_i)$ will be known, and the typical approach is to use plug-in estimates of $\sigma^2(s_i)$ in (9.6), ignoring any possible covariance with the estimator of $\mu(s_i)$. That is, common practice is to worry about the covariance of $\mu(\hat{s}_i)$ with $\epsilon(s_i)$, but not covariance between $\mu(\hat{s}_i)$ and estimates of $\sigma^2(s_i)$. Carroll and Ruppert (1988, pp. 33-34) give a limited treatment of the effect of this common practice in terms of a nonlinear model with heteroscedastic errors that we discussed in Chapter 7 as additive error models with known variance parameters.

Example 9.5

If ordinary least squares is used to estimate $\beta$ in the linear regression model of Example 9.1 we have, from \( \text{var}\beta = \sigma^2 (X^T X)^{-1} \) and $\hat{\mu}_i = x_i^T \beta$, that

\[
\text{var}\{\hat{\mu}_i(\beta)\} = \sigma^2 x_i^T (X^T X)^{-1} x_i \\
= \sigma^2 h_{i,i}, \tag{9.7}
\]

where $h_{i,i}$ is the $i^{th}$ diagonal element of the hat matrix $H = X(X^T X)^{-1}X^T$.

Now,

\[
cov\{\hat{\mu}_i(\beta), \epsilon_i\} = E\{\hat{\mu}_i(\beta)\epsilon_i\} - 0 \\
= E\left\{\epsilon_i \sum_{j=1}^{n} y_j h_{i,j}\right\} \\
= \sum_{j=1}^{n} h_{i,j} E\{y_j \epsilon_i\} \\
= \sum_{j=1}^{n} h_{i,j} E\{(\mu_j + \sigma \epsilon_j) \epsilon_i\} = \sigma \ h_{i,i}.
\]

Substituting into the denominator of (9.6) and replacing $\sigma^2$ with the usual
moment estimator \( \hat{\sigma}^2 \) gives,

\[
b_i = \frac{r_i}{\hat{\sigma}^2(1 - h_{ii})^{1/2}}, \tag{9.8}
\]

the usual studentized residual for linear regression with constant variance.

**Example 9.6**

Consider a nonlinear regression model with constant variance,

\[
Y_i = \mu_i(\beta) + \sigma \epsilon_i,
\]

where \( \epsilon_i \sim iid F, E(\epsilon_i) = 0 \) 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 \( \beta \), inference is based on asymptotic results, as discussed in Chapter 8. Hence, derivation of exact forms for the component quantities of (9.6) is difficult. One development of the usual studentized residual follows.

For a linear model (i.e., \( \mu_i(\beta) = x_i^T \beta \)) with constant variance it is easy to show that, in matrix notation,

\[
[Y - \mu(\hat{\beta})] = [I - H][Y - \mu(\beta^*)], \tag{9.9}
\]

where \( \beta^* \) is the true value of \( \beta \), and \( H = X(X^T X)^{-1}X^T \) is the usual hat matrix. Recall that this gives studentized residuals in the form of expression (9.8). Now, in a nonlinear model with constant variance we can develop two approximations. First, by expanding the expectation function \( \mu_i(\beta) \) about the true value \( \beta^* \), we have that for any \( \beta \) in a small neighborhood of \( \beta^* \),

\[
\mu_i(\beta) \approx \mu_i(\beta^*) + \sum_{k=1}^{p} \frac{\partial}{\partial \beta_k} \mu_i(\beta) \bigg|_{\beta=\beta^*} (\beta_k - \beta_k^*),
\]

or, in matrix notation,

\[
\mu(\beta) \approx \mu(\beta^*) + V(\beta^*)(\beta - \beta^*). \tag{9.10}
\]
Note that in (9.10) the matrix of derivatives $V$ is evaluated at the true value $\beta^*$. Now, the minimization problem being solved by a generalized least squares estimation procedure (or maximum likelihood under normality) is,

$$
\min_{\beta} \sum_{i=1}^{n} \{y_i - \mu_i(\beta)\}^2;
$$

which, after substitution of (9.10), becomes

$$
\min_{\beta} \sum_{i=1}^{n} \left[\{y_i - \mu_i(\beta^*)\} - \sum_{k=1}^{p} \frac{\partial}{\partial \beta_k} \mu_i(\beta) \bigg|_{\beta=\beta^*} (\beta_k - \beta_k^*)\right]^2,
$$

or, in matrix notation,

$$
\min_{\beta} [\{\mathbf{y} - \mu(\beta^*)\} - \mathbf{V}(\beta^*) (\beta - \beta^*)]^T [\{\mathbf{y} - \mu(\beta^*)\} - \mathbf{V}(\beta^*) (\beta - \beta^*)],
$$

which has the ordinary least squares solution,

$$
\hat{\delta} = (\hat{\beta} - \beta^*) = \left[\mathbf{V}^T(\beta^*) \mathbf{V}(\beta^*)\right]^{-1} \mathbf{V}^T(\beta^*) \{\mathbf{y} - \mu(\beta^*)\}. \quad (9.11)
$$

Now, we can’t actually compute $\hat{\delta}$ or $\hat{\beta}$. But, asymptotic results (see e.g., Seber and Wild Chapter 12.2.3) give that, for large enough $n$,

$$(\hat{\beta} - \beta^*) \approx (\tilde{\beta} - \beta^*),$$

so that we can make use of (9.11) with $\tilde{\beta}$ in place of $\hat{\beta}$.

Now, consider the vector of residuals,

$$
r_i = \mathbf{Y} - \mu(\tilde{\beta})
\approx \mathbf{Y} - \mu(\beta^*) + \mathbf{V}(\beta^*) (\hat{\beta} - \beta^*)
\approx \mathbf{Y} - \mu(\beta^*) + \mathbf{V}(\beta^*) \left[\mathbf{V}^T(\beta^*) \mathbf{V}(\beta^*)\right]^{-1} \mathbf{V}^T(\beta^*) [\mathbf{y} - \mu(\beta^*)]
= \left[\mathbf{I} - \mathbf{V}(\beta^*) \left[\mathbf{V}^T(\beta^*) \mathbf{V}(\beta^*)\right]^{-1} \mathbf{V}^T(\beta^*)\right] [\mathbf{Y} - \mu(\beta^*)]
= \left[\mathbf{I} - \mathbf{H}^{(N)}(\beta^*)\right] [\mathbf{Y} - \mu(\beta^*)]. \quad (9.12)
The second line of (9.12) follows from substitution of (9.10) evaluated at \( \mu(\hat{\beta}) \), while the third line results from further use of (9.11) with \( \hat{\beta} \) in place of \( \tilde{\beta} \) as just discussed. The final line of (9.12) is analogous to the linear model result (9.9) with the hat matrix \( H \) replaced by a matrix of the same form but with \( V(\beta^*) \) in place of \( X \) and denoted as \( H^{(N)}(\beta^*) \). That is,

\[
H^{(N)}(\beta^*) = V(\beta^*) [V^T(\beta^*)V(\beta^*)]^{-1}V^T(\beta^*),
\]

where \( V(\beta^*) \) is \( n \times p \) with \( i,k^{th} \) element,

\[
\frac{\partial}{\partial \beta_k} \mu_i(\beta) \bigg|_{\beta=\beta^*}.
\]

With the parallel of expressions (9.12) and (9.9) in hand, we appeal to analogy with linear model results and define studentized residuals to be

\[
b_i = \frac{r_i}{\sqrt{\hat{\sigma}^2 \{1 - h_{i,i}^{(N)}(\hat{\beta})\}^{1/2}}}.
\]

(9.13)

Notice that in (9.13) we have both replaced \( \sigma^2 \) with an estimator, and have also replaced \( \beta^* \) in the nonlinear “hat” matrix \( H^{(N)}(\beta^*) \) with its generalized least squares estimator \( \hat{\beta} \).

Example 9.7

Now consider the general case of a nonlinear model with nonconstant variance,

\[
Y_i = \mu_i(\beta) + \sigma g(\mu_i(\beta), z_i, \theta) \epsilon_i,
\]

where, as usual, \( \epsilon_i \sim iidF \), \( E(\epsilon_i) = 0 \) and \( var(\epsilon_i) = 1 \) but where \( \theta \) is considered known (or chosen as part of model selection). The usual strategy to develop studentized residuals in this case is to note that this model could also be written as

\[
\frac{Y_i}{g(\mu_i(\beta), z_i, \theta)} = \frac{\mu_i(\beta)}{g(\mu_i(\beta), z_i, \theta)} + \sigma \epsilon_i.
\]
which is in the form of a constant variance nonlinear model with modified response \( Y_i / g(\mu_i(\beta), z_i, \theta) \) and modified expectation function \( \mu_i(\beta) / g(\mu_i(\beta), z_i, \theta) \).

As indicated by Carroll and Ruppert (1988, p. 33) the standard approach is to ignore all effects of estimation of \( g(\mu_i(\beta), z_i, \theta) \) and define studentized residuals in the form of (19.3) as,

\[
\tilde{b}_i = \frac{\tilde{r}_i}{\tilde{\sigma}^2 \{1 - \tilde{h}_{i,i}^{(N)}(\hat{\beta})\}]^{1/2},
\]

where

\[
\tilde{r}_i = \frac{y_i - \hat{\mu}_i(\hat{\beta})}{g(\hat{\mu}_i(\hat{\beta}), z_i, \theta)},
\]

and \( \tilde{h}_{i,i}^{(N)}(\hat{\beta}) \) is the \( i^{th} \) diagonal element of the \( n \times n \) matrix

\[
\tilde{H}^{(N)}(\hat{\beta}) = \tilde{V}(\hat{\beta})[\tilde{V}^T(\hat{\beta})\tilde{V}(\hat{\beta})]^{-1}\tilde{V}^T(\hat{\beta}),
\]

where \( \tilde{V}(\hat{\beta}) \) is \( n \times p \) with \( i, k^{th} \) element,

\[
\frac{1}{g(\mu_i(\hat{\beta}), z_i, \theta)} \left[ \frac{\partial}{\partial \beta_k} \mu_i(\beta) \bigg|_{\beta=\hat{\beta}} \right].
\]

**Deviance Residuals**

Deviance residuals are closely connected with exponential families and, in particular, exponential dispersion families. They also represent a somewhat different approach to the conceptual question of what we mean by “residual” than does the fundamental notion of a raw (or studentized) residual. As we have seen, raw residuals are developed first and foremost by considering the deviation of individual response values from their (estimated) expected values. In contrast, deviance residuals are most easily developed as the contributions of individual response values to a quantity that reflects overall model fit. To develop this idea, assume we have a set of independent response variables \( Y_1, \ldots, Y_n \) with density or mass functions of exponential dispersion family
form,

\[ f_i(y_i|\theta_i, \phi) = \exp \left[ \phi \{y_i - b(\theta_i)\} + c(y_i, \phi) \right]. \]

Notice that we are allowing the distributions of the \( Y_i \) to vary only through the scalar natural parameter \( \theta_i \). Recall from Section 6.1.3 this implies that \( \mu_i \equiv E(Y_i) = b'(\theta_i) \), or \( \theta_i = b^{-1}(\mu_i) \) so that we can write the natural parameters as functions of the expected values, \( \theta(\mu_i) \). Now, in almost all models formulated on the basis of exponential dispersion family distributions, we further model \( \mu_i \) as a function of other parameters and, perhaps, covariates. Generalized linear models are the obvious example, but the concept of deviance being developed depends on exponential dispersion family properties not the specific form of generalized linear models. In any case, fitting a model will produce a set of estimated expectations \( \{\hat{\mu}_i : i = 1, \ldots, n\} \) and hence also a set of estimated natural parameters \( \theta(\hat{\mu}) \equiv \{\theta(\hat{\mu}_i) : i = 1, \ldots, n\} \).

We have also seen that, given maximum likelihood estimates, full and reduced models with nested parameter spaces can be compared through likelihood ratio tests. Consider, then, comparison of a fitted model considered as a reduced model to a “saturated” model (or a “maximal model”); these labels are meant to evoke the notions of “fullest model possible” or “model with the highest likelihood value possible”. Such a model will result from estimating \( \mu_i \) as the observed value \( y_i \), for \( i = 1, \ldots, n \), which leads to another set of estimated natural parameters \( \theta(y) \equiv \{\theta(y_i) : i = 1, \ldots, n\} \). Note that such a saturated or maximal model is not a viable or useful model in practice since it contains as many parameters as observations, and this is assuming that the dispersion parameter \( \phi \) is known. With known \( \phi \), a likelihood ratio comparison of fitted and saturated models would then become,

\[ D^* \equiv -2\{L(\theta(\hat{\mu}), \phi) - L(\theta(y), \phi)\}, \quad (9.15) \]
where
\[
L(\theta(\hat{\mu}), \phi) = \sum_{i=1}^{n} \left[ \phi \{ y_i \theta(\hat{\mu}_i) - b(\theta(\hat{\mu}_i)) \} + c(y_i, \phi) \right],
\]
and
\[
L(\theta(y), \phi) = \sum_{i=1}^{n} \left[ \phi \{ y_i \theta(y_i) - b(\theta(y_i)) \} + c(y_i, \phi) \right].
\]
Expression (9.15) defines the scaled deviance for a model based on independent exponential dispersion family random variables. Notice that it may also be written as
\[
D^* = -2 \phi \sum_{i=1}^{n} \left[ y_i \{ \theta(\hat{\mu}_i) - \theta(y_i) \} - b(\theta(\hat{\mu}_i)) + b(\theta(y_i)) \right],
\]
(9.16)
because, with \( \phi \) considered known, the terms \( c(y_i, \phi) \) cancel in the difference. The parameter \( \phi \) may be seen in (9.16) to constitute a scaling factor, and the unscaled deviance is defined as \( D \equiv D^*/\phi \), or
\[
D = -2 \sum_{i=1}^{n} \left[ y_i \{ \theta(\hat{\mu}_i) - \theta(y_i) \} - b(\theta(\hat{\mu}_i)) + b(\theta(y_i)) \right].
\]
(9.17)

Scaled and unscaled deviances 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 \( \phi \) (more on this in the sequel), nothing could fit the data better than the saturated model, which has the greatest log likelihood value possible (this explains my use of the phrase maximal model). If we would not prefer this maximal model to our reduced fitted model, then the fitted model provides an adequate representation of the observed data. In this sense, expression (9.16) constitutes a likelihood ratio goodness of fit test, and \( D \) could be compared to a \( \chi^2 \) distribution with \( n - p \) degrees of freedom. Unfortunately, when \( \phi \) is not known this no longer is the case and, in fact, it is not even possible to estimate
\[ \phi \] under the saturated or maximal model.

**Example 9.8**

It is instructive to examine the forms taken by deviance for some of the more
common exponential dispersion family distributions.

1. **Poisson**

Here, \( \phi \equiv 1 \) and \( \theta_i = \log(\mu_i) \) so that, for a fitted model with estimated
expected values \( \{ \hat{\mu}_i : i = 1, \ldots, n \} \), \( \theta(\hat{\mu}_i) = \log(\hat{\mu}_i) \) and \( \theta(y_i) = \log(y_i) \).

Also, \( b(\theta_i) = \exp(\theta_i) \) so that \( D^\ast = D \), and

\[
D = -2 \sum_{i=1}^{n} [y_i \{ \log(\hat{\mu}_i) - \log(y_i) \} - \hat{\mu}_i + y_i]
= 2 \sum_{i=1}^{n} \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right].
\]

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 con-
vention to simply absorb the known binomial sample sizes \( n_i \) into all
formulas as weights, and then again \( D^\ast = D \) where,

\[
D = -2 \sum_{i=1}^{n} n_i \left[ y_i \left\{ \log \left( \frac{\hat{\mu}_i}{1 - \hat{\mu}_i} \right) - \log \left( \frac{y_i}{1 - y_i} \right) \right\} \right] - \log(1 - \hat{\mu}_i) + \log(1 - y_i)
= 2 \sum_{i=1}^{n} n_i \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (1 - y_i) \log \left( \frac{1 - y_i}{1 - \hat{\mu}_i} \right) \right].
\]

3. **Normal**

For normal distributions with the usual mean \( (\mu) \) and variance \( (\sigma^2) \) pa-
parameterization, \( \theta_i = \mu_i \), \( \phi = 1/\sigma^2 \), and \( b(\theta_i) = (1/2)\theta_i^2 \). Then scaled
deviance is,

\[ D^* = \frac{-2}{\sigma^2} \sum_{i=1}^{n} [y_i(\hat{\mu}_i - y_i) - (1/2)\hat{\mu}_i^2 + (1/2)y_i^2] \]
\[ = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2. \]

Notice that for this situation unscaled deviance is \( D = \sigma^2 D^* \), the usual residual sum of squares.

4. Gamma

Since there are several versions of the “usual” parameterization of a gamma density function we need to be careful of our initial formulation for a problem involving independent gamma random variables. For an individual random variable \( Y \), let the probability density function be

\[ f(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}(1-y)^{\beta-1}; \quad y > 0. \]

With this form, \( \mu \equiv E(Y) = \alpha/\beta \), and by writing \( \nu = \alpha \) we can arrive at an exponential dispersion family representation of the density with \( \theta = -1/\mu, \phi = 1/\nu \), and \( b(\theta) = -\log(-\theta) \). Let \( \{Y_i : i = 1, \ldots, n\} \) be a set of independent random variables have such densities with parameters \( \{\theta_i : i = 1, \ldots, n\} \) and common \( \phi \). Then \( \theta(\hat{\mu}_i) = -1/\hat{\mu}_i \) and \( \theta(y_i) = -1/y_i \), and the scaled deviance becomes,

\[ D^* = -2\phi \sum_{i=1}^{n} \left[ y_i \left( \frac{-1}{\hat{\mu}_i} - \frac{-1}{y_i} \right) + \log(-\hat{\mu}_i) - \log(-y_i) \right] \]
\[ = 2\phi \sum_{i=1}^{n} \left[ \frac{y_i}{\hat{\mu}_i} - 1 + \log(\hat{\mu}_i) - \log(y_i) \right] \]
\[ = 2\phi \sum_{i=1}^{n} \left[ \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log \left( \frac{y_i}{\hat{\mu}_i} \right) \right]. \]
For the Poisson and binomial portions of Example 9.8 we could use deviance as a likelihood ratio goodness of fit test statistic, but not for the normal and gamma. In these latter cases, deviance is generally calculated using an estimated value \( \hat{\phi} \) from the fitted model.

Each observation \( y_i \) contributes one term to (9.16) or (9.17), and it is these terms that are used to define basic deviance residuals. Let,

\[
d_i^* = -2 \hat{\phi} \left[ y_i \{ \theta(\hat{\mu}_i) - \theta(y_i) \} - b(\theta(y_i)) + b(\theta(\hat{\mu}_i)) \right],
\]

and define deviance residuals as, for \( i = 1, \ldots, n \),

\[
d_i \equiv \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i^*}.
\]

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,

\[
d'_i = \frac{d_i}{(1 - h_i^{(G)})^{1/2}},
\]

where \( h_i^{(G)} \) is the \( i^{th} \) diagonal element of the matrix

\[
H^{(G)} = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2},
\]

in which \( X \) is the \( n \times p \) matrix of covariate values of the linear predictor \( \eta \equiv (\eta_1, \ldots, \eta_n)^T \) and \( W \) is the \( n \times n \) diagonal matrix with elements given in Section 8.3.6 as,

\[
W_i \equiv \left\{ \left( \frac{d\eta_i}{d\mu_i} \right)^2 V(\mu_i) \right\}^{-1}.
\]

The standardization of (9.19) is justified by results on the first two moments of “generalized residuals”, a topic we will cover briefly later in this section, and
9.1. ANALYSIS OF RESIDUALS

conditions that make higher derivatives of the log likelihood negligible. As a result, \( E(d_i) \approx 0 \) and \( \text{var}(d_i) \approx 1 - h^{(G)}_{i,i} \). A readable presentation of this is contained in Davison and Snell (1991), who also point out that (9.19) is a special case of a result that applies more generally to exponential dispersion families. In particular, consider a model formulated in the same manner as a generalized linear model expect that, rather than using a link to a linear prediction as \( g(\mu_i) = \mathbf{x}_i^T \beta \), we simply take the expectations to be a given function of parameters and covariates as

\[
\mu_i = \eta(x_i, \beta),
\]

denoted as \( \eta_i \) for brevity.

Then, define the matrix \( \mathbf{W} \) 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 \( \mathbf{Q} \) to have \( i,k^{th} \) element,

\[
q_{i,k} = \frac{\partial \eta_i}{\partial \beta_k}.
\]

Then, take

\[
\tilde{H}^{(G)} = \mathbf{W}^{1/2} \mathbf{Q} (\mathbf{Q}^T \mathbf{W} \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{W}^{1/2},
\]

and standardized deviance residuals are then given by (9.19) with \( \tilde{H}^{(G)} \) in place of \( H^{(G)} \). Note that, in the case of a generalized linear model, \( w_i \) has the same form as given following expression (9.19), and \( \mathbf{Q} = \mathbf{X} \).

Many of the general ideas described in the past few pages are also applied in formulating a number of other residual quantities that seem to be less commonly used in practice. These include what are called score residuals, likelihood residuals, and Anscombe residuals. See, for example, Lindsey (1996, p. 168) for mention of the first two of these, and Davison and Snell (1991) for
the latter. Somewhat more common are *Pearson residuals* particularly in the case of models with discrete response variables such as Poisson or binomial.

**Generalized Residuals**

In what remains an important paper on the construction and analysis of residuals, Cox and Snell (1968) gave a general definition of a residual, as follows. The situation considered is one in which the location variable of our general notation from Section 9.1.1 is taken to be $s_i = i$ and the random variables \( \{Y_i : i = 1, \ldots, n\} \) are assumed independent. Consider a model in which

\[
Y_i = g_i(\theta, \epsilon_i); \quad i = 1, \ldots, n, \tag{9.20}
\]

where $\theta$ is a $p$-dimensional parameter, the $\epsilon_i$ are independent and identically distributed random variables, and the model is indexed by $i$ (i.e., $g_i(\cdot)$) to allow for covariates or other known factors. Expression (9.20) is sufficiently general to cover nearly all models formulated for independent random variables, a key aspect being the each $Y_i$ depends on only one $\epsilon_i$.

**Example 9.9**

1. Additive error models, either linear or nonlinear, are easily put into the form of (9.20) by taking $\theta = (\beta, \sigma^2, \theta)$ and, using our previous notation for such models,

\[
g_i(\theta, \epsilon_i) = g_1(x_i, \beta) + g_2(x_i, \beta, z_i, \theta)\epsilon_i,
\]

where $\epsilon_i \sim iidF$ with $E(\epsilon_i) = 0$. 
2. Multiplicative error models are also easy to put in the form of (9.20). Cox and Snell consider as an example a model

\[ Y_i = \beta_1 \exp\{\beta_2 (x_i - \bar{x})\} \varepsilon_i, \]

where \( \varepsilon_i \sim iid \) exponential, with \( E(\varepsilon_i) = 1 \).

3. Let \( Y_1, \ldots, Y_n \) be independently and identically distributed following an extreme value distribution with density written as, for \(-\infty < \theta_1 < \infty\) and \( \theta_2 > 0 \),

\[ f(y_i|\theta_1, \theta_2) = \frac{1}{\theta_2} \exp \left[ \frac{y_i - \theta_1}{\theta_2} - \exp \left\{ \frac{y_i - \theta_1}{\theta_2} \right\} \right]; \quad -\infty < y_i < \infty. \]

The distribution function that corresponds to this density is,

\[ F_Y(y_i|\theta_1, \theta_2) = 1 - \exp \left[ -\exp \left\{ \frac{y_i - \theta_1}{\theta_2} \right\} \right], \]

and the inverse distribution function becomes,

\[ F_Y^{-1}(a|\theta_1, \theta_2) = \theta_1 + \theta_2 \log \left[ \log \left\{ \frac{1}{1-a} \right\} \right]; \quad 0 < a < 1. \]

To formulate a model in the form of (9.20) for this situation, let

\[ Y_i = g_i(\theta_1, \theta_2, U_i) = F_Y^{-1}(u_i|\theta_1, \theta_2), \]

where \( U_i \sim iid U(0,1) \).

4. The device of item 3 immediately above may be used directly for any set of continuous random variables for which a model leads to a parameterized distribution function. This is true even if the inverse distribution function is not available in closed form, such as for gamma, beta, or even normal distributions; for a normal distribution, however, we would probably use the easier additive model formulation. It is also not necessary
to have identical distributions. Specifically, if \( Y_1, \ldots, Y_n \) are independent random variables with modeled probability density functions \( f_i(y_i|\theta) \), then the model may be written as,

\[
Y_i = g_i(\theta, U_i) = F_i^{-1}(U_i|\theta); \quad i = 1, \ldots, n,
\]

where \( U_i \sim iid U(0,1) \) and

\[
F_i(y_i|\theta) = \int_{-\infty}^{y_i} f_i(t|\theta) \, dt.
\]

5. Although the prescription of items 3 and 4 applies only to random variables with continuous distributions, a similar device may be used for discrete random variables. Let \( Y_1, \ldots, Y_n \) be independent with a common set of possible values \( \Omega \) and suppose the elements of this set have been ordered as \( \Omega \equiv \{y_1, y_2, \ldots, y_m\} \) so that \( y_{k-1} < y_k \) for \( k = 2, \ldots, m \). If we desire to assign probability mass functions \( f_i(y_i|\theta) \) to these random variables, let

\[
F_i(y_k|\theta) = \sum_{j=1}^{k} f_i(y_j|\theta),
\]

and take, for \( i = 1, \ldots, n \),

\[
Y_i = \min \left\{ y_k : F_i(y_k|\theta) > U_i \right\},
\]

where \( U_i \sim iid U(0,1) \). Define \( y_{00} \) to be any value such that \( F_i(y_{00}|\theta) = 0 \). Then, for \( k = 1, \ldots, m \),

\[
Pr(Y_i = y_k) = Pr\{F_i(y_{k-1}|\theta) < U_i < F_i(y_k|\theta)\} = F_i(y_k|\theta) - F_i(y_{k-1}|\theta) = f_i(y_k|\theta)
\]

as desired.
Now, writing a model in the form of expression (9.20) does not necessarily define a residual quantity. Cox and Snell (1968) do so as follows. Suppose that equation (9.20) has a unique solution for \( \epsilon_i \), say

\[
\epsilon_i = h_i(Y_i, \theta).
\]

If this is the case, and if \( \hat{\theta} \) is a maximum likelihood estimate of \( \theta \), then define generalized residuals as, for \( i = 1, \ldots, n \),

\[
r_i = h_i(y_i, \hat{\theta}).
\]  

(9.21)

The random version of (9.21) may be written as \( R_i = h_i(Y_i, \hat{\theta}) \). The remainder of the treatment by Cox and Snell (1968) involved deriving approximations to the means, variances, and covariances of the random version of (9.21) by expanding \( R_i - \epsilon_i \) as a Taylor series in terms of the components of \( \hat{\theta} - \theta \), which is where they use the condition that \( \hat{\theta} \) is a maximum likelihood estimate of \( \theta \).

These approximations provide a means for modifying the residuals \( R_i \) so that the modified residuals have expectations and variances approximately equal to those of the \( \epsilon_i \). Adjustment of the residuals in (9.21) may be important, but the form of resultant equations is quite model specific. Here, we simply indicate that if the number of observations is large relative to the number of parameters, then the residual quantities (9.21) should have behavior similar to the \( \epsilon_i \) in the model (9.20). This is because it is the observations \( y_1, \ldots, y_n \) that are used to estimate \( \theta \) as \( \hat{\theta} \) so that the \( R_i \) are not independent, and generally will not have expectations or variances equal to those of the \( \epsilon_i \). But as the number of observations increases relative to the dimension of \( \theta \) these effects diminish.

We will use these ideas to develop the notion of generalized residual that will be used in these notes, drawing on cases 4 and 5 of Example 9.9.
Corresponding to case 4, given continuous independent random variables $Y_1, \ldots, Y_n$ with model $Y_i = F_i^{-1}(U_i|\theta)$ for $U_i \sim iid U(0, 1)$, define the residual quantities,

$$r_i \equiv \int_{-\infty}^{y_i} F_i(t|\hat{\theta}) \, dt; \quad i = 1, \ldots, n. \quad (9.22)$$

If the model is representative of the data, the \{r_i : i = 1, \ldots, n\} should behave in a manner similar to a sample from the uniform distribution on $(0, 1)$. The probability integral transform would hold if the parameter $\theta$ were used in (9.22) rather than an estimate $\hat{\theta}$. While this result does not hold, we expect that the residuals of (9.22) should provide diagnostic quantities useful to detect gross discrepancies between the model and observed responses.

To define similar residuals corresponding to discrete random variables as in case 5 of Example 9.9 requires an extension of the definition of Cox and Snell (1968). In the development of (9.21) it was assumed that the model (9.20) allows a unique solution for the $\epsilon_i$. Here, we define random residuals even for fixed values of observations \{yi : i = 1, \ldots, n\}. Using the same notation as in case 5 of Example 9.9, let $Y_1, \ldots, Y_n$ be independent random variables with a common set of possible values, ordered as $\Omega \equiv \{y_1, y_2, \ldots, y_m\}$. Take the ordered value of observation $i$ to be the $k^{th}$ ordered value, that is, $y_i = y_k$.

Define the (random) generalized residual $r'_i$ to be the realized value of a random variable with distribution uniform on the interval $(F_i(y_{q-1}), F_i(y_q))$, that is,

$$r'_i \equiv u_i; \quad \text{where } U_i \sim iid U \left( F_i(y_{q-1}), F_i(y_q) \right). \quad (9.23)$$

Similar to the residuals of expression (9.22), these residuals should behave in the manner of a sample of iid uniform variables on the interval $(0, 1)$. A set of residuals \{r'_i : i = 1, \ldots, n\} will not, however, be unique for a given set of observations \{yi : i = 1, \ldots, n\}. 
9.1. ANALYSIS OF RESIDUALS

Other Types of Residuals

The classification of residuals into categories in this subsection is, of course, not exhaustive of the various quantities proposed for use as residuals. While many such quantities are rather model specific, there are some that are more general in nature. A few of these are listed here.

1. Pearson Residuals.

Pearson residuals are motivated by considering individual contributions to a Pearson Chi-squared goodness of fit test for discrete random variables. Consider, for example, a set of independent Possion random variables \( \{Y_i : i = 1, \ldots, n\} \) with expected values \( \{\mu_i : i = 1, \ldots, n\} \) and variances \( \{V(\mu_i) : i = 1, \ldots, n\} \), where \( V(\mu_i) = \mu_i \). The Pearson \( \chi^2 \) goodness of fit test statistic is,

\[
\chi^2 = n \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.
\]

The square root of the individual contributions are known as Pearson residuals,

\[
r_{p,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \tag{9.24}
\]

The residual quantities in (9.24) are more general that just Poisson, or even just discrete random variables, and are still called Pearson residuals even when applied to random variables with continuous distributions.

2. Score Residuals.

For independent random variables, the score function for any given parameter component consists of a sum of quantities, the random versions of which each have expected value 0. With respect to a given component of a generic parameter, \( \theta_j \), say, the score residuals are then based on
standardized contributions to the total score function,
\[
r_{s,i} = \left( \frac{1}{I_{i,i}(\hat{\theta})} \right) \frac{\partial}{\partial \theta_j} \log\{f_i(y_i|\theta)\},
\]
where \(I_{i,i}(\hat{\theta})\) is the \(i^{th}\) diagonal element of the estimated information matrix.

3. Anscombe Residuals.
Ansombe (1961) proposed a residual quantity of the essential form of the Pearson residual (9.24) but with \(y_i\) replaced by a transformed value that more nearly produces normal distributions than (9.24). The appropriate transformation depends on the model chosen, although it can be expressed in a general form for standard generalized linear models (e.g., McCullagh and Nelder, 1989).

### 9.1.3 Plotting Residuals

Any number of diagnostic plots can be constructed using the residuals quantities discussed in Section 9.1.2, with the intent of detecting departures from the model structure assumed in an analysis. We mention here some of the more common of these, along with the types of modeling inadequacies they are intended to detect. In general, residual plots involve plotting residuals (or some transformation of residuals) on the vertical axis or ordinate against corresponding quantities of some type on the horizontal axis or 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**

Perhaps the most common type of residual plot results from plotting residuals against fitted values from a model. Fitted values are generally taken as estimated expected values of the random variables associated with observed responses, that is, the estimated systematic model component. We have already seen a number of examples of this type of residual plot, at least for linear regression models using residuals as in expression (9.8). Figure 6.2 presents such a plot for the simulated data relating Microcystin concentrations to nitrogen. Figure 7.4 gives this residual plot for airplane flight times. Figures 7.12, 7.14, 7.15, and 7.19 show these plots for various models employed in the example of tree volume modeled as a function of diameter and height.

**Example 9.10**

In Example 7.1 a nonlinear regression model with additive constant variance errors was fitted to the reaction times of an enzyme as a function of substrate concentration for preparations treated with Puromycin and also for untreated preparations. The model was

\[ Y_i = \frac{\beta_1 x_i}{\beta_2 + x_i} + \sigma \epsilon_i, \]
where \( x_i \) denoted substrate concentration and we took \( \epsilon_i \sim iidF \) for some
distribution \( F \) with \( E(\epsilon_i) = 0 \) and \( \text{var}(\epsilon_i) = 1 \) for \( i = 1, \ldots, n \). This model
was fit to each group (treated and untreated) separately. Figure 9.1 presents
the studentized residuals of expression (9.13) for both groups. This residual
plot does not reveal any serious problems with the model, although it is less
than “picture perfect” in terms of what we might hope to see. Given that
this model was formulated on the basis of a theoretical equation for enzyme
reaction times (the Michaelis-Menten equation) and variability is anticipated
(and appears in the data) to be small, we would be justified in assessing this
residual plot with a fairly high level of scrutiny (relative to, say, a residual plot
for a purely observational study with many potential sources of variability).

Does the residual plot of Figure 9.1 exhibit some degree of increasing variance
as a function in increasing mean? To help in this assessment, we might plot
the cube root of squared studentized residuals against the fitted values (e.g.,
Carroll and Ruppert, 1988, p. 30). In this type of residual plot, nonconstant
variance is exhibited by a wedge-shaped pattern of residuals. A plot of the
cube root squared studentized residuals for these data is presented in Figure
9.2. There does not appear to be a increasing wedge or fan of residuals in
the plot of Figure 9.2, suggesting that there is little evidence of nonconstant
variance for this model. Looking closely at the residual plot of Figure 9.1 we
can see a suggestion of a “U-shaped” pattern in residuals from both treated
and untreated groups. This would indicate that the fitted expectation function
from the Michaelis-Menten equation fails to bend correctly to fit data values
at across the entire range of substrate concentrations. A close examination
of the fitted curves, presented in Figure 9.3 verifies that this seems to be the
case for at least the treated preparations. In fact, there appear to be values
at a substrate concentration of just over 0.\( \sqrt{2} \) ppm for which the expectation
Figure 9.1: Studentized residuals from fitting a nonlinear regression based on the Michaelis-Menten equation to the enzyme reaction times of Example 7.1. Open circles are the untreated preparations while solid circles are the treated preparations.
Figure 9.2: Cube root squared studentized residuals from fitting a nonlinear regression based on the Michaelis-Menten equation to the enzyme reaction times of Example 7.1. Open circles are the untreated preparations while solid circles are the treated preparations.
Figure 9.3: Fitted regressions based on the Michaelis-Menten equation to the enzyme reaction times of Example 7.1. Open circles are the untreated preparations while solid circles are the treated preparations.

function “misses” for both treated and untreated groups. The cause of this phenomenon is unknown to us as is, indeed, the degree of scientific import for what it suggests. It may well be the case, however, that there exists some evidence that the theoretical Michaelis-Menten equation does not adequately describe the enzyme reaction in this experiment.

In general, the use of cube root squared studentized residuals might be justified based on what is known as the *Wilson-Hilferty* transformation to normalize chi-squared variables, but the basic value of such plots seems due to
more practical than 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. Carroll and Ruppert (1988) echo this sentiment, but indicate that squaring residuals can create extreme values if the original residuals are moderately large in absolute value to begin with. They then suggest taking the cube root to alleviate this potential difficulty, but point out that they view the result essentially as a transformation of absolute residuals, which are taken as “the basic building blocks in the analysis of heteroscedasticity” (Carroll and Ruppert, 1998, p. 30). From this standpoint, it would seem to make little difference if one used absolute residuals, the square root of absolute residuals or, as in Figures 9.1 and 9.2, a 2/3 power of absolute residuals.
Plotting Against Covariates

Plotting Against Time or Space

More on Scaling Residual Plots

9.1.4 Tests With Residuals

9.2 Cross Validation

9.2.1 Fundamental Concepts

9.2.2 Types of Cross Validation

Group Deletion

Leave One Out Deletion

Pseudo-Cross Validation

9.2.3 Discrepancy Measures

Loss Functions and Decision Theory

Global Versus Local Discrepancies

9.3 Assessment Through Simulation

9.3.1 Fundamental Concepts

9.3.2 Discrepancy Measures Revisited

9.3.3 Simulation of Reference Distributions

9.3.4 Sensitivity Analysis
Part III

BAYESIAN ANALYSIS
Chapter 10

Basic Bayesian Paradigms

In this part of the course we consider the topic of Bayesian analysis of statistical models. Note that I have phrased this as Bayesian analysis of models, as opposed to analysis of Bayesian models. In some ways this is simply a matter of semantics, but there is an issue involved that goes slightly beyond the mere use of words, which we will attempt to elucidate in what follows. The effects of the two viewpoints on Bayesian analysis presented in this chapter do not influence mathematical or statistical techniques in the derivation of distributions or inferential quantities, but they may have an impact on which quantities (i.e., which distributions) are deemed appropriate for inference. The issues involved come into play primarily in the case of hierarchical models, in which there is more than one level of quantities that play the role of parameters in probability density or mass functions. In this regard, there might be a distinction between thinking in terms of analysis of a Bayesian model as opposed to a Bayesian analysis of a mixture (i.e., random parameter) model.

In this chapter we will attempt to make clear both the fundamental nature of the Bayesian argument, and the potential impact of the distinction eluded
to in the previous paragraph. Since this issue is not even necessarily recognized as an issue by many statisticians, the subsection headings given below are my own device and should not be taken as standard terminology in the world of statistics.

10.1 Strict Bayesian Analysis

The heading of “strict” Bayesian analysis comes from a reading of the history of Bayesian thought. Although I will not give a list of references, I believe this line of reasoning to be faithful to that history, in which there was reference to what was called the “true state of nature”. That is, at least one line of reasoning in the development of Bayes methods held that there is, in fact, an absolute truth to the order of the universe. This thought is in direct conflict with the frequent (at least in the past) criticism of Bayesian methods that they depend on totally subjective interpretations of probability (there were, however, other schools of thought in the development of Bayesian methods in which probabilities were viewed as essentially subjective values, but we will not discuss these). The fundamental point is that this view of Bayesian analysis is in total agreement with a reductionist view of the world in which, if all forces in operation were known, observable quantities would be completely deterministic. The true state of nature in this school of thought is embodied in a fixed, but unknown parameter value that governs the distribution of observable quantities. Note that this is starting to sound quite a bit like a typical frequentist argument, and that is the point.

There is not necessarily anything in a Bayesian approach to statistical analysis that contradicts the view that, if we knew everything about all physical relations in the world, we would know the values that would be assumed by
observable quantities with certainty. We are, of course, not capable of such exquisite knowledge, and thus the field of statistics has meaning in modeling the things we do not understand through the use of probability.

The essential departure of Bayesian thought from its frequentist counterpart, under this interpretation of Bayesian analysis, is that an epistemic concept of probability is legitimate. That is, given a fixed but unknown parameter \( \theta \) that represents the true state of nature, it is legitimate to write
\[
Pr(t_1 < \theta < t_2) = c,
\]
with \( \theta, t_1, t_2 \) and \( c \) all being constants, not as a statement about a random quantity (there are no random variables in the above probability statement) but rather as a statement about our knowledge that \( \theta \) lies in the interval \((t_1, t_2)\). Thus, under this interpretation of Bayesian thought, there must be somewhere in a model a fixed quantity or parameter that represents the “true state of nature”.

Now, if we admit an epistemic concept of probability, then we are free to represent our current knowledge about \( \theta \) as a probability distribution. And, given the possibility of modeling observable quantities as connected with random variables (the modeling concept), we can easily formulate the basic structure of a strict Bayesian approach to the analysis of data; note here that everything contained in Part II of the course prior to the discussion of estimation and inference (i.e., everything before Chapter 8) applies to Bayesian analysis as much as it applies to what was presented under the heading of Statistical Modeling. In particular, suppose that we have a situation in which we can formulate random variables \( Y_1, \ldots, Y_n \) in connection with observable quantities. Suppose that the joint probability distribution of these variables can be written in terms of a density or mass function \( f(y_1, \ldots, y_n|\theta) \) depending on an unknown (but fixed) parameter \( \theta \) that represents the true state of nature (i.e., the phenomenon or scientific mechanism of interest). Now, even in the
absence of data, we are almost never totally devoid of knowledge about $\theta$. If an observable quantity $Y$ is a count with a finite range, and the mass function assigned to $Y$ is a binomial distribution with parameter $\theta$, then we know that $\theta$ must lie between 0 and 1. If a set of observable quantities are connected with iid random variables having gamma distributions with common parameters $\alpha$ and $\beta$ (here, $\theta \equiv (\alpha, \beta)^T$), then we know that $\alpha$ and $\beta$ must both be positive and that any distribution assigned to our knowledge of their values should give low probability to extremely large values (i.e., both the mean and variance of the gamma distribution should exist).

Given this background, consider a single random variable $Y$, to which we have assigned a probability distribution through density or mass function $f(y|\theta)$ with fixed but unknown parameter $\theta$. Our knowledge about $\theta$ before (i.e., prior) to observing a value connected with $Y$ is embodied in a prior distribution $\pi(\theta)$. Note that, while $f(y|\theta)$ may be interpreted through a hypothetical limiting relative frequency concept of probability, the prior distribution $\pi(\theta)$ is an expression of epistemic probability, since $\theta$ is considered a fixed, but unknown, quantity. The mathematics of dealing with $\pi(\theta)$ will, however, be identical to what would be the case if $\theta$ were considered a random variable. Suppose that $\theta$ can assume values in a region (i.e., $\theta$ is continuous) with dominating Lebesgue measure. Then, given observation of a quantity connected with the random variable $Y$ as having the particular value $y$, we may derive the posterior distribution of $\theta$ as,

$$p(\theta|y) = \frac{f(y|\theta) \pi(\theta)}{\int f(y|\theta) \pi(\theta) d\theta}. \quad (10.1)$$

If $\theta$ can assume only values in a discrete set and $\pi(\cdot)$ is dominated by counting measure, then (10.1) would become,

$$p(\theta|y) = \frac{f(y|\theta) \pi(\theta)}{\sum_{\theta} f(y|\theta) \pi(\theta)}. \quad (10.2)$$
Expressions (10.1) and (10.2) are generally presented as applications of Bayes Theorem, the form of which they certainly represent. But Bayes Theorem is a probability result, which holds for random variables regardless of whether one is using it in the context of a Bayesian analysis or not. The only reason one may not accept the expressions (10.1) or (10.2) as valid is if one will not allow an epistemic concept of probability for the function \( \pi(\theta) \) to describe our knowledge of the fixed but unknown value of \( \theta \). Thus, we are led to the conclusion that the fundamental characteristic of Bayesian analysis does not rely on Bayes Theorem, but on an epistemic concept of probability to describe what we know about the true state of nature \( \theta \).

In the above development we call the distribution for the observable quantities \( f(y_1, \ldots, y_n|\theta) \) the observation or data model, and the distribution \( \pi(\theta) \) (interpreted under an epistemic concept of probability) the prior distribution. The resultant distribution of our knowledge about \( \theta \) conditioned on the observations, namely \( p(\theta|y) \), is the posterior distribution of \( \theta \).

Now, notice that nothing above changes if we replace the data model for a single random variable \( Y \) with the joint distribution of a set of variables \( Y_1, \ldots, Y_n \) all with the same parameter \( \theta \). That is, the data model becomes \( f(y|\theta) \), the prior remains \( \pi(\theta) \), and the posterior would be \( p(\theta|y) \).

In this development we have taken the prior \( \pi(\theta) \) to be completely specified which means that \( \pi(\cdot) \) depends on no additional unknown parameters. That does not necessarily mean that \( \pi(\cdot) \) depends on no controlling parameters, only that, if it does, those parameter values are considered known.

**Example 10.1**
Suppose that we have a random variable \( Y \) such that the data model is given
by $Y \sim Bin(\theta)$. That is, $Y$ has probability mass function

$$f(y|\theta) = \frac{n!}{(n-y)!y!} \theta^y (1-\theta)^{n-y}; \quad y = 0, 1, \ldots, n.$$  

Now, we know that $\theta \in (0, 1)$ and, if we wish to express no additional knowledge about what the value of $\theta$ might be, we could take the prior to be $\theta \sim U(0, 1)$, so that,

$$\pi(\theta) = 1; \quad 0 < \theta < 1.$$  

The posterior that would result from an observation of the data model as $y$ would then be,

$$p(\theta|y) = \frac{\theta^y (1-\theta)^{n-y}}{\int_0^1 \theta^y (1-\theta)^{n-y} d\theta} = \frac{\Gamma(n+2)}{\Gamma(1+y) \Gamma(1+n-y)} \theta^y (1-\theta)^{n-y},$$

which is the density function of a beta random variable with parameters $1 + y$ and $1 + n - y$.

Now, suppose we have additional information about $\theta$ before observation of $y$ which is represented by a beta distribution with parameters $\alpha = 2$ and $\beta = 2$; this would give expected value $\alpha / (\alpha + \beta) = 0.5$ and variance 0.05. Then, the posterior would be derived in exactly the same way as above, except that it would result in a beta density with parameters $2 + y$ and $2 + n - y$. In general, for any specific choices of $\alpha$ and $\beta$ in the prior $\pi(\theta) = Beta(\alpha, \beta)$ the posterior will be a beta distribution with parameters $\alpha + y$ and $\beta + n - y$.

Now consider a problem in which we have a data model corresponding to

$$Y_1, \ldots, Y_m \sim iid Bin(\theta),$$

in which the “binomial sample sizes” $n_1, \ldots, n_m$ are considered fixed, and $\pi(\theta)$ is taken to be $\text{beta}(\alpha_0, \beta_0)$ with $\alpha_0$ and $\beta_0$ specified (i.e., considered
known). Then a derivation entirely parallel to that given above, except using a joint data model \( f(y|\theta) = f(y_1|\theta)f(y_2|\theta) \ldots , f(y_m|\theta) \) results in a posterior \( p(\theta|y_1, \ldots, y_m) \) which is a beta \((\alpha_p, \beta_p)\) density with parameters

\[
\begin{align*}
\alpha_p &= \alpha_0 + \sum_{i=1}^{m} y_i, \\
\beta_p &= \beta_0 + \sum_{i=1}^{m} n_i - \sum_{i=1}^{m} y_i.
\end{align*}
\tag{10.3}
\]

The essential nature of what I am calling here the strict Bayesian approach to data analysis follows from the modeling idea that the scientific mechanism or phenomenon of interest is represented somewhere in a statistical model by a fixed parameter value. The additional component under a Bayesian approach is to assign our knowledge of that parameter a prior distribution under an epistemic concept of probability. This then allows derivation of a posterior distribution that represents our knowledge about the parameter value after having incorporated what can be learned from taking observations in the form of data. Thus, there are really no such things as “Bayesian models”, only Bayesian analyses of statistical models. This view of Bayesian analysis extends from simple situations such as that of Example 10.1 to more complex situations in a natural manner.

**Example 10.2**

Suppose that we have a data model corresponding to

\[ Y_1, \ldots, Y_m \sim \text{indep Bin}(\theta_i), \]

in which the binomial sample sizes \( n_1, \ldots, n_m \) are considered fixed. The only difference between this data model that that of the second portion of Example 10.1 is that now each response variable \( Y_i; \ i = 1, \ldots, m \) is taken to have its
own binomial parameter. If we interpret the values $\theta_1, \ldots, \theta_m$ as representing different “manifestations” of some scientific mechanism or phenomenon, then we might assign these values a mixing distribution (as in Part II of the course),

$$\theta_1, \ldots, \theta_m \sim iid \text{Beta}(\alpha, \beta),$$

which would result in a beta-binomial mixture model as discussed previously. A Bayesian analysis of this mixture model would then consist of assigning (our knowledge about) the parameter $(\alpha, \beta)$ a prior distribution $\pi(\alpha, \beta)$ and deriving the posterior distribution $p(\alpha, \beta|y)$.

Example 10.2 is an illustration of the general structure for a strict Bayesian analysis of a mixture model. In general we have,

1. Data Model:
   
   $Y_1, \ldots, Y_n \in \Omega$ have joint density or mass function $f(y|\theta) = f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n)$.
   
   If these response variables are independent, or conditionally independent given $\theta_1, \ldots, \theta_n$, then,

   $$f(y|\theta) = f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} f_i(y_i|\theta_i).$$

2. Mixing Distribution:
   
   $\theta_1, \ldots, \theta_n \in \Theta$ have joint density or mass function $g(\theta|\lambda) = g(\theta_1, \ldots, \theta_n|\lambda)$.
   
   If these random variables are independent, then,

   $$g(\theta|\lambda) = g(\theta_1, \ldots, \theta_n|\lambda) = \prod_{i=1}^{n} g_i(\theta_i|\lambda).$$

3. Prior:
   
   Given $f(y|\theta)$ and $g(\theta|\lambda)$, the parameters $\lambda \in \Lambda$ are assigned a prior distribution $\pi(\lambda)$. 

4. Posterior:

Using the above formulation, the posterior density or mass function of $\lambda$ may be derived as,

$$p(\lambda|y) = \frac{f(y|\theta) g(\theta|\lambda) \pi(\lambda)}{\int_{\lambda} \int_{\Theta} f(y|\theta) g(\theta|\lambda) \pi(\lambda) d\theta d\lambda}. \quad (10.4)$$

If the $Y_i$ are conditionally independent and the $\theta_i$ are independent, then,

$$p(\lambda|y) = \prod_{i=1}^{n} f_i(y_i|\theta_i) g_i(\theta_i|\lambda) \pi(\lambda) \int_{\Lambda} \left\{ \prod_{i=1}^{n} \int_{\Theta_i} f_i(y_i|\theta_i) g_i(\theta_i|\lambda) d\theta_i \right\} \pi(\lambda) d\lambda.$$

Now notice that, in the above progression, we could have equivalently combined items 1 and 2 into the overall mixture model

$$h(y|\lambda) = \int_{\Theta} f(y|\theta) g(\theta|\lambda) d\theta,$$

or, in the case of conditional independence of $\{Y_i : i = 1, \ldots, n\}$ and independence of $\{\theta_i : i = 1, \ldots, n\}$,

$$h(y|\lambda) = \prod_{i=1}^{n} \int_{\Theta_i} f_i(y_i|\theta_i) g_i(\theta_i|\lambda) d\theta_i.$$

The posterior of $\lambda$ could then be expressed, in a manner entirely analogous with expression (10.1), as,

$$p(\lambda|y) = \frac{h(y|\lambda) \pi(\lambda)}{\int_{\Lambda} h(y|\lambda) \pi(\lambda) d\lambda}, \quad (10.5)$$

or, in the case of independence,

$$p(\lambda|y) = \prod_{i=1}^{n} \frac{h_i(y_i|\lambda) \pi(\lambda)}{\int_{\Omega_i} h_i(y_i|\lambda) \pi(\lambda) dy_i}.$$
The preceding expressions are a direct expression of a Bayesian analysis of a mixture model. The mixture model is given by $h(y|\lambda)$, and we simply assign the parameters of this model, namely $\lambda$, a prior distribution $\pi(\lambda)$ and derive the corresponding posterior $p(\lambda|y)$.

We are not yet entirely prepared to address the possible issue of the introduction, but notice that, in principle, other distributions are available to us, such as

$$p(\theta|y) = \frac{\int_{\Lambda} f(y|\theta) g(\theta|\lambda) \pi(\lambda) d\lambda}{\int_{\Lambda} \int_\Theta f(y|\lambda) g(\theta|\lambda) \pi(\lambda) d\theta d\lambda},$$

$$p(\theta|\lambda, y) = \frac{f(y|\theta) g(\theta|\lambda) \pi(\lambda)}{\int_\Theta f(y|\lambda) g(\theta|\lambda) \pi(\lambda) d\theta}. \quad (10.6)$$

The first of these expressions is sometimes called the *marginal* posterior of $\theta$ and the second the *conditional* posterior of $\theta$. This last expression for the conditional posterior $p(\theta|\lambda, y)$ is of particular interest, since it can also be written as,

$$p(\theta|\lambda, y) = \frac{m(\theta, y|\lambda)}{\int_\Theta m(\theta, y|\lambda) d\theta} = \frac{m(\theta, y|\lambda)}{h(y|\lambda)},$$

$$= \frac{f(y|\theta) g(\theta|\lambda)}{h(y|\lambda)}. \quad (10.7)$$

where $h(y|\lambda)$ is the same as the “marginal” density of $Y$ discussed in marginal maximum likelihood analysis of mixture models. Notice that, to use $p(\theta|\lambda, y)$
directly for inference, we would need to have an estimate of \( \lambda \) to use in this density function as \( p(\theta | \hat{\lambda}, y) \).

Finally, we also have the modeled distribution of \( \theta \) as \( g(\theta | \lambda) \) and, given an estimate of \( \lambda \) we could focus attention on this distribution with estimated \( \lambda \) as \( g(\theta | \hat{\lambda}) \); this, in fact, is what we have done in Part II of the course in presenting plots of, e.g., beta mixing densities, using \( \hat{\lambda} \) as given by marginal maximum likelihood estimates. It would certainly be possible, however, to take \( \hat{\lambda} \) to be the mode or expectation of the posterior distribution \( p(\lambda | y) \) in a Bayesian approach. A portion of the issue introduced in the introduction to this discussion of Bayesian analysis concerns whether inference about \( \theta \) in a Bayesian analysis of a mixture model (or hierarchical model) should be made based on \( p(\theta | y) \), \( p(\theta | \hat{\lambda}, y) \) or \( g(\theta | \hat{\lambda}) \).

### 10.2 Bayesian Analysis of Unknowns

What I call here the Bayesian analysis of unknowns is not so much a different approach to that of the strict Bayesian analysis of Chapter 10.1 as it is just a slightly different angle on what is being accomplished. The philosophy of this viewpoint is that statistical models contain unknown quantities. Whether some might be considered random variables and others parameters under the strict Bayesian interpretation is not material; everything we do not know is simply an unknown quantity. Probability and, in particular, epistemic probability, is the way that we describe uncertainty. Thus, anything we do not know the value of can be assigned a “prior” distribution to represent our current knowledge and, given a subsequent set of observed data \( \{y_1, \ldots, y_n\} \), these can be updated to be “posterior” distributions. This seems to be the view taken in several recent texts on Bayesian data analysis (Carlin and Louis, 2000; Gelman, Carlin, Stern
and Rubin, 1995). In fact, Carlin and Louis (2000, p.17) state that

In the Bayesian approach, in addition to specifying the model for the observed data $y = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\theta$, usually in the form of a probability distribution $f(y|\theta)$, we suppose that $\theta$ is a random quantity as well, . . .

It is unclear whether Carlin and Louis are asserting that $\theta$ is a parameter or a random variable (in the sense we have used these words in this class). My own interpretation of their words is that they just don’t really care. If one wishes to think of $\theta$ as a parameter, that’s fine, or if one wishes to think of $\theta$ as a random variable that’s fine too. Either way, $\theta$ is unknown, and so we use probability to describe our uncertainty about its value.

Gelman, Carlin (a different Carlin from Carlin and Louis), Stern and Rubin (1995) are quite circumspect in talking about random variables per se, more often using phrases such as “random observables”, “sampling models”, or simply the unadorned “variable” (reference to random variables does pop up occasionally, e.g., p. 20, but that phrase is nearly absent from the entire text). These authors present (Chapter 1.5 of that text) a clear exposition of the basic notion of probability as a representation of uncertainty for any quantity about which we are uncertain. The implication is again that we need not be overly concerned about assigning some mathematical notion of “type” (e.g., random variable or parameter) to quantities in order to legitimately use probability as a measure of uncertainty.

This notion of using probability as a measure of uncertainty without getting all tangled up in whether the object of our uncertainty should be considered a parameter, random variable, or some other type of quantity is seductively simple, perhaps too much so, as I will argue below. It does have the attractive
flavor that all quantities involved in an analysis are put on an equal footing. The distinction is only that some of those quantities will be observable (i.e., the data) while others will not. Simplicity in analytical approach (but not necessarily practice) is then achieved through the assertion that the entire process of statistical inference is just a matter of conditioning our uncertainty (in the form of probability) about unobservable quantities on the values obtained for observable quantities.

This view of Bayesian analysis does not result in any differences with that of the strict Bayesian view in terms of manipulations of distributions, derivation of posteriors and so forth. In simple problems that involve only a data or observational model and a prior, such as those described in Example 10.1, there is only one possible way to condition the distribution of (our knowledge or uncertainty about) the unobservable $\theta$ on that of the data. In these situations, I see no possible conflicts between the strict Bayesian view and that considered in this section.

In a hierarchical situation involving data model $f(y|\theta)$ and additional distributions $g(\theta|\lambda)$ and $\pi(\lambda)$, there may be some slightly different implications for inference of the strict Bayesian view and that of this section. In particular, as noted above, Bayesian analysis of unknowns would indicate that the proper distribution on which to base inference is the joint posterior $p(\theta, \lambda|y)$, that is, the conditional distribution of unobservables on observables. If one is uninterested in $\lambda$, for example, the corresponding marginal $p(\theta|y)$ is the posterior of concern. The possibilities of making use of the conditional posterior $p(\theta|\lambda, y)$ or (especially) the model distribution $g(\theta|\lambda)$ are not in total concert with Bayesian analysis of unknowns.

In fact, the motivation of specifying a distribution $g(\theta|\lambda)$ can be considered somewhat differently under this view of Bayesian analysis. In a strict
Bayes view, \( g(\theta|\lambda) \) is thought of as a mixing distribution that represents, in the model, a scientific mechanism or phenomenon of interest. The fixed parameter \( \lambda \) then embodies the immutable mechanism, while \( g(\theta|\lambda) \) describes the ways that the mechanism is manifested in different situations. Here, we are hesitant to assign such responsibilities to \( \lambda \) and \( g(\theta|\lambda) \). The value of \( \theta \) is an unobservable quantity that controls the description of our uncertainty about the observable quantity \( y \) (or quantities \( y_1, \ldots, y_n \)). Our uncertainty about \( \theta \) is therefore assigned a distribution \( g(\theta|\lambda) \), which may depend on an additional unknown quantity \( \lambda \). Our uncertainty about \( \lambda \) is then also assigned a distribution \( \pi(\lambda) \). It is quite natural, then, to refer to prior distributions for both \( \theta \) and \( \lambda \). The distribution \( g(\theta|\lambda) \) would be the conditional prior for \( \theta \), while the marginal prior for \( \theta \) would be,

\[
\pi_\theta(\theta) = \int_{\Lambda} g(\theta|\lambda) \, \pi(\lambda) \, d\lambda.
\]

One could then derive a posterior as in expression (10.1) using data model \( f(y|\theta) \) and prior \( \pi_\theta(\theta) \), resulting in the posterior

\[
p(\theta|y) = \frac{f(y|\theta) \, \pi_\theta(\theta)}{\int_{\Theta} f(y|\theta) \, \pi_\theta(\theta) \, d\theta} = \frac{\int_{\Lambda} f(y|\theta) \, g(\theta|\lambda) \, \pi(\lambda) \, d\lambda}{\int_{\Lambda} \int_{\Theta} f(y|\lambda) \, g(\theta|\lambda) \, \pi(\lambda) \, d\theta \, d\lambda}
\]

(10.8)

Notice that (10.8) is exactly the same marginal posterior for \( \theta \) given in expression (10.6), illustrating the point that all of the mathematics of Chapter 10.1 also apply here.

What I want to contrast (10.8) with, however, is the posterior for \( \lambda \) in expression (10.5). Expression (10.5) was developed by combining \( f(y|\theta) \) and
10.3. SUMMAR Y OF THE VIEWPOINTS

$g(\theta|\lambda)$ into the mixture model $h(y|\lambda)$ which was then used with the prior $\pi(\lambda)$ to derive the posterior $p(\lambda|y)$. Expression (10.8) has been developed by combining $g(\theta|\lambda)$ and $\pi(\lambda)$ into the prior $\pi_\theta(\theta)$ which was then used with the data model $f(y|\theta)$ to derive the posterior $p(\theta|y)$. That is, to get to (10.5) most directly, $g(\theta|\lambda)$ was considered part of the model $h(y|\lambda) = \int f(y|\theta)g(\theta|\lambda)\,d\theta$, to which was assigned the prior $\pi(\lambda)$. To get to (10.8) most directly, $g(\theta|\lambda)$ was considered part of the prior $\pi_\theta(\theta) = \int g(\theta|\lambda)\pi(\lambda)\,d\lambda$, which was applied to the model $f(y|\theta)$.

While these progressions are not inherent parts of the two viewpoints that I have called here strict Bayesian analysis and Bayesian analysis of unknowns they do indicate a somewhat different slant to the way in which hierarchical models and, in particular, the roles of $p(\theta|y)$, $p(\theta|y,\lambda)$ and $g(\theta|\lambda)$, are interpreted.

10.3 Summary of the Viewpoints

To encapsulate what has been presented in the two previous sections, there appear to be several angles from which Bayesian analysis of hierarchical models can be approached. Both depend on the derivation of posterior distributions for unknown quantities in a statistical model. Both collapse to the same viewpoint in the case of simple models with a given data or observational model and a fixed parameter that controls that model. The potential differences arise in consideration of multi-level or hierarchical models. The following points seem relevant.

1. Under either viewpoint, all of the same distributions are available. Distributions of $f(y|\theta)$, $g(\theta|\lambda)$ and $\pi(\lambda)$ constitute the statistical model.
Distributions $p(\theta, \lambda|y)$ and the associated marginals $p(\theta|y)$ and $p(\lambda|y)$, the conditional form $p(\theta|\lambda, y)$, and the model distribution $g(\theta|\hat{\lambda})$ are available and identical under both viewpoints. For the purposes of inference, these last two would require a plug-in estimator of $\lambda$ as $p(\theta|\hat{\lambda}, y)$ and $g(\theta|\hat{\lambda})$.

2. Under what I have called strict Bayesian analysis there is really only one prior distribution, that being for whatever quantities are considered fixed parameters in a model. If we have a data model $f(y|\theta)$ for which $\theta$ is fixed, and to which we wish to assign the prior $g(\theta|\lambda)$ but have uncertainty about an appropriate value for $\lambda$, then we might combine $g(\theta|\lambda)$ and $\pi(\lambda)$ into the prior $\pi_{\theta}(\theta)$ just as in the progression given for the viewpoint titled Bayesian analysis of unknowns. This seems unlikely, however. If I know enough about the possible values of $\theta$ to assign them (my knowledge of them) a prior $g(\theta|\lambda)$ that takes a particular form, surely I must also have some idea what the value of $\lambda$ might be. Thus, it would be somewhat unnatural (not necessarily wrong or inappropriate, but perhaps a bit odd), given a strict Bayesian view of the world, to make use of the argument of Carlin and Louis (2000, p.19-20) that hierarchical models result from uncertainty about the appropriate values of the parameter $\lambda$ to use in the “first-stage” prior $g(\theta|\lambda)$. A counter-argument to my assertion is possibly offered by a theorem due to de Finetti (1974) which we will discuss under the heading of exchangeability in Chapter 12. This theorem may be used to justify the formulation of a prior for $\theta$ in the data model $f(y|\theta)$ through the use of a mixture such as $\pi_{\theta}(\theta)$ developed just before expression (10.8).

On the other hand, if we have a data model that consists of the obser-
vation model \( f(y|\theta) \) and a mixing distribution \( g(\theta|\lambda) \), then we are in one sense assigning a prior only to \( \lambda \), and considering \( \theta \) to be a random variable in the usual sense of the term, although typically a “latent” random variable that is not connected to observable quantities. In this latter setting, \( p(\lambda|y) \) is certainly meaningful. A question arises, however, concerning \( p(\theta|y) \) versus \( p(\theta|\lambda, y) \) versus \( g(\theta|\lambda) \) where, for the purpose of inference, a plug-in estimate of \( \lambda \) would be needed in the latter two distributions.

3. Under what I have called Bayesian analysis of unknowns there is no distinction made between the status of quantities as random variables or parameters, or the mixing model idea of “random variables that play the role of parameters”. Probability is epistemic in nature wherever it is applied; even to observable quantities. If my knowledge about the values of observable quantities, represented through the observation model \( f(y|\theta) \) happens to agree with what would result from repeated observation of the same situation (i.e., relative frequency) then so much the better; I then just have some empirical justification for my belief. Given this viewpoint, the appropriate quantities for inference would seem to be \( p(\theta, \lambda|y) \) and the associated marginals \( p(\theta|y) \) and \( p(\lambda|y) \). There is probably little role for \( p(\theta|\lambda, y) \) and almost certainly little use for \( g(\theta|\lambda) \), where these latter two quantities would again be used with a plug-in estimator of \( \lambda \).

4. What can be said about relations between \( p(\theta|y) \), \( p(\theta|\lambda, y) \) and \( g(\theta|\lambda) \) in the mathematical sense? Such relations must apply regardless of the
viewpoint toward analysis that is being taken. First, we have that,

\[
p(\theta | y) = \frac{\int_\Lambda m(y, \theta, \lambda) d\lambda}{\int_\Lambda \int_\Theta m(y, \theta, \lambda) d\theta d\lambda} = \int_\Lambda p(\theta | \lambda, y) p(\lambda | y) h(y) d\lambda \]

Thus, as it should be, the marginal posterior \( p(\theta | y) \) is the expected value of the conditional posterior \( p(\theta | \lambda, y) \) taken over the distribution of \( \lambda \) given \( y \). The variance of \( p(\theta | y) \) will then be greater than that of \( p(\theta | \lambda, y) \).

Secondly, directly from (10.7) we have that,

\[
g(\theta | \lambda) = \frac{h(y | \lambda)}{f(y | \theta)} p(\theta | \lambda, y).\]

Since \( h(y | \lambda) \) is more diffuse (variable) than \( f(y | \theta) \), so \( g(\theta | \lambda) \) will be more diffuse than \( p(\theta | \lambda, y) \). Thus, both \( p(\theta | y) \) and \( g(\theta | \lambda) \) represent greater uncertainty about \( \theta \) than does \( p(\theta | \lambda, y) \), which is intuitive. The unknown relation, at this time, is between \( p(\theta | y) \) and \( g(\theta | \lambda) \) when the latter is evaluated at a value \( \hat{\lambda} \), which will be either the expectation or mode of the posterior \( p(\lambda | y) \).

5. Under what I have called a strict Bayesian viewpoint, inference about \( \lambda \) is most naturally based on \( p(\lambda | y) \). Inference about \( \theta \) is most naturally based on either \( g(\theta | \hat{\lambda}) \) or possibly \( p(\theta | y) \). The role, if any, for \( p(\theta | \hat{\lambda}, y) \) is not clear. Under what I have called the Bayesian analysis of unknowns, inference for \( \lambda \) is most naturally based on \( p(\lambda | y) \), although this will
typically not be of much interest. Inference for $\theta$ should probably be based on $p(\theta|y)$. There is little, if any, role for $p(\theta|\lambda, y)$ and almost certainly no use for $g(\theta|\lambda)$. Note here, however, that using $p(\theta|\lambda, y)$ is one formulation of methods that are called *empirical Bayes* (e.g., Carlin and Louis, 2000, Chapter 3).
Chapter 11

Sequential Bayes

Perhaps one of the strongest arguments in favor of a Bayesian approach to analysis is provided by the similarity between the view of science as a progressive “building up” of knowledge, and a sequential use of Bayesian analysis. To see this clearly, adopt for the moment the strict Bayesian viewpoint that our concern is learning about the value of a fixed “state of nature” represented by a parameter $\theta$. That is, suppose we have a problem in which data are “generated” by a observational model $f(y|\theta)$ and we have specified a prior $\pi(\theta)$ for $\theta$. Now, $\pi(\theta)$ represents our knowledge about $\theta$ before any observations are available. Given observations $y$, we update that knowledge in the form of the posterior $p(\theta|y)$. Now, suppose that we (or someone else) are able to repeat the study that led to the observations $y$, or at least conduct a similar study in which $\theta$ is a controlling parameter in an observational model and has the same scientific meaning it did in the first study. Then it would be natural to take $p(\theta|y)$ from the first study as representing our current knowledge about $\theta$ (i.e., the posterior from the first study becomes the prior for the second study). In a sequence of $k$ such studies, then, we could conduct an overall
analysis using data models \( f_1(\mathbf{y}_1), \ldots, f_k(\mathbf{y}_k) \) and a “cascade” of prior and posterior distributions, in the following way:

<table>
<thead>
<tr>
<th>Study</th>
<th>Prior</th>
<th>Data</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \pi_1(\theta) )</td>
<td>( f_1(\mathbf{y}_1</td>
<td>\theta) )</td>
</tr>
<tr>
<td>2</td>
<td>( \pi_2(\theta) = p_1(\theta</td>
<td>\mathbf{y}_1) )</td>
<td>( f_2(\mathbf{y}_2</td>
</tr>
<tr>
<td>3</td>
<td>( \pi_3(\theta) = p_2(\theta</td>
<td>\mathbf{y}_2) )</td>
<td>( f_3(\mathbf{y}_3</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( k )</td>
<td>( \pi_k(\theta) = p_{k-1}(\theta</td>
<td>\mathbf{y}_{k-1}) )</td>
<td>( f_k(\mathbf{y}_k</td>
</tr>
</tbody>
</table>

At each step in this progression we would have the basic Bayesian update of expression (10.1), namely,

\[
p_j(\theta|\mathbf{y}_j) = \frac{f_j(\mathbf{y}_j|\theta) \pi_j(\theta)}{\int_{\Theta} f_j(\mathbf{y}_j|\theta) \pi_j(\theta) \, d\theta}; \quad j = 1, \ldots, k.
\]

Implementation of this progression is made much easier if all of the data models have the same form, \( f(\cdot) = f_1(\cdot) = f_2(\cdot) = \ldots = f_k(\cdot) \), and if the common data model \( f(\cdot) \) and prior \( \pi_1(\cdot) \) have a property called “conjugacy” which means that the posterior \( p_1(\theta|\mathbf{y}) \) has the same form as the prior \( \pi_1(\theta) \) (more on this in Chapter 12.2). This then implies that all of the priors \( \pi_1(\cdot), \pi_2(\cdot), \ldots, \pi_k(\cdot) \) all have the same form as well.

Example 11.1

The sex ratio at birth of various species is of interest to ecologists in understanding evolutionary pressures and the way in which organisms have adapted in response to those pressures. A study was conducted over several years on the South American Guanaco (one of the South American “camels”), the others
being the Llama, Alpaca, and Vicuna) to determine the proportion of males at birth. It is fairly well established that mortality is higher during the first year of life for males than for females, because males tend to venture farther from the protection of the adult “herd” than do females, and are thus more often attacked by predators (your own interpretation of this being because males are “adventurous”, “developing skills to protect the herd when adults”, or “just stupid” may well depend on your own sex). At any rate, this has led ecologists and geneticists to believe that the ratio of males to females at birth should be greater than 0.50. The basic idea has nothing to do with the “good of the species” but rather the genetic pay-off to adult females who produce male or female offspring. Under random assortment, each female contributes one \(X\) chromosome and each male either an \(X\) or \(Y\) chromosome with equal probability. Skewed sex ratio at birth could be due to viability of implantation, development during gestation, and so forth if there is a “physiological edge” to being a male with \(XY\) chromosomes. That is, if the sex ratio at conception is 50/50 but more males die before reaching breeding age, then a female who “decides” (in an evolutionary sense) to produce a male is playing the “genetic lottery” (less chance of having any grandchildren, but more of them if it happens). Females should “decide” to take this gamble up to a point, but no more. Given the pace of evolution in mammals, the ratio of male to female births should represent an “Evolutionary Stable Strategy”, that is, a “true state of nature” at least for the duration of multiple human generations. The study under discussion was conducted out of what was then the Department of Animal Ecology at Iowa State University (under the direction of Dr. William Franklin) to determine if there is evidence that such a strategy has developed in Guanacos so that more males than females are born. I have been told that geneticists have, for reasons unclear to me, predicted that the proportion of
males at birth should be 0.524.

The study design was quite simple. Field workers located and observed adult female Guanacos over a period of 4 years, recording the number of male and female offspring in each year. Each female Guanaco that breeds in a given year produces one offspring. Since we are assuming random assortment in the genetic process, which males were involved is irrelevant (at least to us, maybe not to them). The number of male and female births recorded in this study are given in the following table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>1988</td>
<td>47</td>
<td>42</td>
</tr>
<tr>
<td>1989</td>
<td>51</td>
<td>42</td>
</tr>
<tr>
<td>1990</td>
<td>53</td>
<td>46</td>
</tr>
</tbody>
</table>

Since each female in a given year produces one offspring, and we are assuming that the male parent is not a factor in sex of that offspring, in a given year $i$ it is not unreasonable to formulate an observation model for the number of male offspring as $Y_i \sim Bin(\theta, n_i)$ where $n_i$ is taken as fixed by the number of births observed. Although some females may have been observed several times over the 4 year period, individual identification was either not possible or not recorded. Thus, from a statistical perspective, we will assume that $Y_1, Y_2, Y_3$ and $Y_4$ are exchangeable (more on exchangability in the next chapter). Our absolute prior knowledge about the value of $\theta$ before any data are observed is that $0 < \theta < 1$, and we will choose a beta distribution to represent our knowledge of this parameter, which represents the “true state of nature”, based on arguments from the ecological and evolutionary sciences (as sketched above). We might, in an effort to appear “objective”, choose to take the prior
for 1987 to have parameters $\alpha_0 = 1$ and $\beta_0 = 1$, resulting in a uniform distribution on $(0, 1)$ for $\theta$. From the discussion of Example 10.1, the posterior after observation in 1987 is a beta distribution with parameters $\alpha_1 = \alpha_0 + y_1$ and $\beta_1 = \beta_0 + n_1 - y_1$. Proceeding in a sequential manner as illustrated in the table of page 801, we obtain the following posterior parameters:

<table>
<thead>
<tr>
<th>Year</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>34</td>
<td>31</td>
<td>0.5231</td>
<td>0.0038</td>
</tr>
<tr>
<td>1988</td>
<td>81</td>
<td>73</td>
<td>0.5260</td>
<td>0.0016</td>
</tr>
<tr>
<td>1989</td>
<td>132</td>
<td>115</td>
<td>0.5344</td>
<td>0.0010</td>
</tr>
<tr>
<td>1990</td>
<td>185</td>
<td>161</td>
<td>0.5347</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

Graphs of the posterior densities (which by conjugacy will all be beta densities) are shown in Figure 11.1; the prior for the 1987 sample is a uniform and not shown in this figure.

Credible intervals for $\theta$ (we will get to these in Chapter 13.2) were,

<table>
<thead>
<tr>
<th>Year</th>
<th>95% Interval</th>
<th>90% Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>(0.402, 0.642)</td>
<td>(0.421, 0.624)</td>
</tr>
<tr>
<td>1988</td>
<td>(0.447, 0.604)</td>
<td>(0.460, 0.592)</td>
</tr>
<tr>
<td>1989</td>
<td>(0.472, 0.596)</td>
<td>(0.482, 0.586)</td>
</tr>
<tr>
<td>1990</td>
<td>(0.482, 0.587)</td>
<td>(0.490, 0.579)</td>
</tr>
</tbody>
</table>

As can be seen from the above tables and Figure 11.1, the posterior distributions appear to be “narrowing in” on a value of $\theta$ that is above the value of 0.50, although after 4 years of data the value 0.50 is still included in both 95% and 90% credible intervals.

While the results of Example 11.1 are a pleasing illustration of a sequential process of gaining more and more knowledge about a parameter $\theta$, as evidenced by the decreasing variances of the posterior distributions in moving from 1987
Figure 11.1: Sequential posterior densities for the analysis of sex ratio at birth in Guanacos. The initial prior was a uniform distribution on the interval (0, 1).
to 1990, one has to wonder if this truly represents some kind of scientific support for the concept that $\theta$ represents an evolutionarily stable strategy on which we are “zeroing in” as more data are accumulated. An examination of the model (and evidenced by the values of $\hat{\alpha}$ and $\hat{\beta}$ in the previous table), shows that the beta distribution parameters $\alpha$ and $\beta$ will both increase as more data are collected. Does this essentially imply that the posterior variance will decrease as a function of amount of data, regardless of what those data indicate about the value of $\theta$? This question is difficult to answer analytically, but an example will suffice to illustrate the point.

**Example 11.2**

First consider a simulated version of the situation of Example 11.1, in which 5 successive values $y_1, \ldots, y_5$ were independently simulated from a binomial distribution with parameter $\theta = 0.55$ and binomial sample size fixed at $n = 30$ for each value. Beginning with a uniform (0, 1) prior, analysis of these data in the same manner as that of Example 11.1 produced the following posterior values:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$y_i$</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>0.469</td>
<td>0.0075</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>0.548</td>
<td>0.0039</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>0.565</td>
<td>0.0026</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>0.557</td>
<td>0.0020</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>0.539</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

We see in this table essentially the same behavior as that of the table of posterior values for Example 11.1, which is good, although in that example we do not know the actual value of $\theta$, while here we know the true value is $\theta = 0.55$. Nevertheless, what we are doing in this example is essentially trying
Figure 11.2: Posterior densities for the first set of simulated data in Example 11.2. The true value of $\theta = 0.55$ is shown by the solid vertical line.

to mimic the behavior of data and estimators from that situation with real data. Graphs of the sequence of prior/posterior densities for the simulated data are shown in Figure 11.2, which again looks quite similar to Figure 11.1.

Now consider values $y_1, \ldots, y_5$ simulated from a beta-binomial model in which the observation or data model is taken to be $\text{Bin}(\theta_i, n)$, again with $n = 30$, and $\theta_1, \ldots, \theta_5$ are values from a beta distribution with parameters $\alpha = 4$ and $\beta = 3.2727$. These values are the $\lambda$ in the mixing distribution.
$g(\theta|\lambda)$ and produce $E(\theta) = 0.55$ and $\text{var}(\theta) = 0.0299$. Suppose, however, that we apply the same model used previously in which each $y_i; \ i = 1, \ldots, 5$ is taken to be from the same binomial with parameter $\theta$ and $n = 30$. In this case, a sequential analysis, proceeding exactly as above yields the following results:

<table>
<thead>
<tr>
<th></th>
<th>$y_i$</th>
<th>$\theta_i$</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>0.407</td>
<td>0.375</td>
<td>0.0071</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>0.412</td>
<td>0.452</td>
<td>0.0039</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>0.532</td>
<td>0.500</td>
<td>0.0027</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>0.855</td>
<td>0.582</td>
<td>0.0020</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.704</td>
<td>0.599</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

While the posterior expectations in the above table are perhaps somewhat more variable than those from the previous table (although this would be exceedingly difficult to detect without having the previous values available), what I want to draw your attention to are the posterior variances, which are so similar to the first portion of this example as to be discomforting. Graphs of the posterior densities for these data from fitting the (incorrect) sequential model are shown in Figure 11.3.

Figure 11.3 is, again, quite similar to Figure 11.2, particularly if one did not have the solid vertical line showing, in this case, the true expected value of $\theta; \ i = 1, \ldots, 5$. In fact, without knowledge of the true data generating mechanism (in this case the simulation distributions) we would not consider the results of this sequential analysis applied to the two sets of data different at all, and that is exactly the point. Although the data were simulated from quite distinct models, an analysis supposing a single binomial parameter gave quite similar results.
Figure 11.3: Posterior densities for the second set of simulated data in Example 11.2. The true value of $E(\theta) = 0.55$ is shown by the solid vertical line.
If we superimpose on the graphs of Figure 11.3 the true distribution \( g(\theta|\lambda) \) = \( g(\theta|\alpha, \beta) \) we obtain what is shown in Figure 11.4. If inferences were made about a value of \( \theta \) that was (incorrectly) assumed to be constant for these data, we would base that inference on the densities shown by the solid curves of Figure 11.4, which would clearly be misleading. Relative to the analysis of Example 11.1 about sex ratio at birth in Guanacos, the implications of this exercise is that we should take no comfort from the progression of posterior densities that appear to be closing in on a given value of \( \theta \). Rather, support for that conclusion must come from the scientific argument that \( \theta \) does in fact represent an evolutionarily stable strategy, and hence should be considered the same for each year of observation.

The message of this example is not that Bayesian analysis can result in misleading inferences. The same difficulty illustrated above could easily be encountered in a non-Bayesian analysis of this problem. The message is that uncritical Bayesian analysis can lead to such difficulties, just as uncritical non-Bayesian analysis. That is, there is no special protection against the deleterious effects of model misspecification offered by taking a Bayesian approach. No matter what approach is taken toward estimation and inference of parametric statistical models, the modeling process itself is critical. Gelman, Carlin, Stern and Rubin (1995, p. 161) make this same point in emphasizing that the use of sensitivity analysis in a Bayesian approach should consider not only the effects of prior specification, but also the likelihood that results from a given data model.

Now consider a Bayesian analysis of the actual model used to simulate the second data set of Example 11.2, which was a beta-binomial model. We would take the model to consist of binomial specifications for independent random variables \( Y_1, \ldots, Y_5 \), each with its own data model parameter \( \theta_1, \ldots, \theta_5 \), which
Figure 11.4: Posterior densities for the second set of simulated data in Example 1.2 with the true mixing density for values of $\theta$ overlaid. The true value of $E(\theta) = 0.55$ is shown by the solid vertical line.
follow the same beta distribution with parameters $\alpha$ and $\beta$. A prior would be specified for the distribution of $\lambda \equiv (\alpha, \beta)^T$ as $\pi(\lambda)$. The posterior distribution $p(\lambda|y)$ would represent our knowledge about $\lambda$ in the light of the observed data $y_1, \ldots, y_5$. But what about the “posterior” $p(\theta|y)$? Here, given independence of the $Y_i$s and exchangeability or independence of the $\theta_i$s, this distribution would have the form,

$$p(\theta|y) = \prod_{i=1}^5 p_i(\theta_i|y) = \prod_{i=1}^5 p_i(\theta_i|y_i).$$

What interpretation should be given to the $p_i(\theta_i|y_i)$ in this case? This is the conditional distribution of $\theta_i$ given the observed value $y_i$, but, if each $\theta_i$ is representation of a unique random variable, what does this conditional distribution (i.e., this “posterior”) tell us? In the model, the common distribution of the $\theta_i$s, namely $g(\theta|\lambda)$ represented the (distribution of the) ways in which the scientific mechanism or phenomenon of interest is expressed in observable situations. But here, each $p_i(\theta_i|y_i)$ will be a different distribution. Does $p_i(\theta_i|y_i)$ then represent what we know about how the mechanism was manifested in the particular situation that led to the observed value $y_i$? Perhaps. Does $p_i(\theta_i|y_i)$ represent what we know about the “true state of nature” in the situation that led to the observation $y_i$? Perhaps. But, extending the pure reductionist view that everything would be deterministic if we understood all of the forces at work in particular situations, the “true state of nature” should represent forces or factors (i.e., scientific mechanisms or phenomena) that are larger in scope than can be seen in those particular instances, that is, commonalities or fundamental laws about nature. And, from this perspective, the conditional distributions $p_i(\theta_i|y_i)$ seem to be less of any type of “posterior” distributions
than simply mathematical expressions of questionable value.

A number of points are relevant at this juncture:

1. It is difficult to fit the above situation into the framework offered by Carlin and Louis (2000) that in hierarchical models $g(\theta|\lambda)$ is a true prior on $\theta$ that is controlled by a parameter $\lambda$ about which we have additional uncertainty and thus use a “second-stage” prior $\pi(\lambda)$.

2. There does seem to be a role for $g(\theta|\lambda)$ in inference, since this distribution is thought (from the modeling exercise) to apply to a broader set of situations than those under observation, while $p(\theta|y)$ concerns only the particular values of $\theta$ that may have been associated with what was observed.

3. Clearly, whatever view we have of Bayesian (or even non-Bayesian) analysis, a fundamental concern is model formulation. This is one of the reasons I indicated previously that the viewpoint of “simply putting distributions on unknown quantities” and then making inference by conditioning unobservable quantities on observable quantities was perhaps a bit overly simplistic. Without careful consideration of how a model is connected to a problem under study, which seems most natural by considering the concepts of random variables, random parameters, and fixed parameters, no approach to analysis will be guaranteed to provide reasonable scientific inferences.

4. Coming full circle, I now repeat the claim made in the first sentence of this part of the course (Chapter 10) that what we are concerned with is the Bayesian analysis of models, not the analysis of Bayesian models.
Chapter 12

Prior Distributions

Regardless of the overall view one might take toward Bayesian analysis, there is a need to specify prior distributions for one or more quantities in a model. In this chapter we consider several topics that are involved in this process of assigning priors.

12.1 Exchangeability

A common assumption in Bayesian analyses is that the “observable” random variables $Y_1, \ldots, Y_n$ are exchangeable. The meaning of exchangeable is given in the following definition.

Definition:

1. $Y_1, \ldots, Y_n$ are marginally exchangeable if, for a probability density or mass function $m(\cdot)$ and permutation operator $\mathcal{P}$,

\[
m(y_1, \ldots, y_n) = m(\mathcal{P}(y_1, \ldots, y_n)).
\]
2. $Y_1,\ldots,Y_n$ are conditionally exchangeable given $z$ if, for a probability density or mass function $m(\cdot)$ and permutation operator $\mathcal{P}$,

$$m(y_1,\ldots,y_n|z) = m(\mathcal{P}(y_1,\ldots,y_n)|z).$$

The interpretation of these definitions needs clarification. For any valid joint distribution it is always true that the indices of variables may be permuted. That is, for random variables $X_1$ and $X_2$ it is always true that

$$Pr(X_1 = x_1, X_2 = x_2) = Pr(X_2 = x_2, X_1 = x_1).$$

This is trivial, not exchangeability. What exchangeability implies is that

$$Pr(X_1 = x_1, X_2 = x_2) = Pr(X_1 = x_2, X_2 = x_1),$$

which is a quite different condition. The implication of exchangeability is that the probability with which random variables assume various values does not depend on the “identity” of the random variables involved; this is essentially a “symmetry” condition.

It is true that iid random variables are exchangeable, and we often assume the condition of independent and identical distribution, but we should realize that exchangeability is not the same as either independence or identical distribution, as shown by the following example.

Example 12.1

1. Exchangeable but not Independent Random Variables. Let the pair of random variables $(X,Y)$ be bivariate with possible values

$$(X, Y) \in \{(0, 1), (0, -1), (1, 0), (-1, 0)\},$$

which is a quite different condition. The implication of exchangeability is that the probability with which random variables assume various values does not depend on the “identity” of the random variables involved; this is essentially a “symmetry” condition.
such that each possible value has probability 0.25.

**Exchangeability:**
Clearly, \( Pr(X = x, Y = y) = Pr(X = y, Y = x) \), since each possible value has the same probability.

**Lack of Independence:**
\( Pr(X = 1) = 0.25 \) and \( Pr(Y = 0) = 0.5 \), but \( Pr(X = 1, Y = 0) = 0.25 \neq 0.25(0.5) \)

2. Independent but not Exchangeable Random Variables.

Let \( X \) and \( Y \) be any two independent random variables with \( X \) being discrete with probability mass function \( f_X(x); \ x \in \Omega_X \) and \( Y \) being continuous with probability density function \( f_Y(y); \ y \in \Omega_Y \).

**Independence:** Independence is by assumption, so that the joint (mixed) density (and mass function) is
\[
m(x, y) = f_X(x)f_Y(y); \ (x, y) \in \Omega_X \times \Omega_Y.
\]

**Lack of Exchangeability:**
For any \( y \notin \Omega_X \) we would have
\[
f_X(y)f_Y(x) = 0 \neq f_X(x)f_Y(y),
\]
so that \( X \) and \( Y \) cannot be exchangeable.

In the first portion of this example, what “messes up” independence is the lack of what has previously been called the positivity condition. That is, while \( \Omega_X \equiv \{-1, 0, 1\} \) and \( \Omega_Y \equiv \{-1, 0, 1\} \), and the probability distributions of \( X \) and \( Y \) on these sets is the same (i.e., \( X \) and \( Y \) are identically distributed), it is not true that \( \Omega_{X,Y} = \Omega_X \times \Omega_Y \). In the second portion of the example, what
“messes up” exchangeability is that the sets of possible values $\Omega_X$ and $\Omega_Y$ are not the same, although the positivity condition does hold in this case.

In general, random variables that are not identically distributed cannot be exchangeable, random variables that are independent and identically distributed are exchangeable, but exchangeability is not the same property as independence.

The role of exchangeability in formulation of prior distributions follows from a famous theorem of de Finetti (1974) which essentially states that, if quantities $\theta_1, \ldots, \theta_n$ are exchangeable following the same distribution $m(\theta_i|\lambda)$ where $\lambda$ is unknown, then any suitably “well-behaved” joint distribution for these quantities can be written in the mixture form,

$$m(\theta) = \int_{\Lambda} \left\{ \prod_{i=1}^{n} m(\theta_i|\lambda) \right\} \pi(\lambda) d\lambda,$$

as $n \to \infty$.

What this theorem basically justifies is the use of a prior as a mixing distribution in distributions that have the form of an iid mixture. We relied on exchangeability of $Y_1, \ldots, Y_4$ in Example 12.1 for the sequential analysis of sex ratio at birth in Guanacos. In hierarchical models, exchangeability of the data model parameters $\{\theta_i : i = 1, \ldots, n\}$ and de Finetti’s theorem does lend some credence to thinking of a distribution,

$$\pi_{\theta}(\theta) = \int_{\Lambda} \left\{ \prod_{i=1}^{n} g(\theta_i|\lambda) \right\} \pi(\lambda) d\lambda,$$

as a prior to be applied to the data model $f(y|\theta)$. 
12.2 Conjugate Priors

We have already seen the use of conjugate priors in Example 10.1 and Examples 11.1 and 11.2 on sequential Bayes. To expand on conjugacy, consider a simple setting with observation model \( f(\mathbf{y}|\mathbf{\theta}) \) and prior \( \pi(\theta|\lambda_0) \), where we have written the prior as a parameterized distribution, but are considering \( \lambda_0 \) 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,

\[
p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta) \pi(\theta|\lambda_0)}{\int_\Theta f(\mathbf{y}|\theta) \pi(\theta|\lambda_0) d\theta} = \pi(\theta|h(\mathbf{y}, \lambda_0)),
\]

where \( h(\mathbf{y}, \lambda_0) \) is some function of \( \mathbf{y} \) and \( \lambda_0 \). That is, if in the transition from prior to posterior, the effect of the data \( \mathbf{y} \) is only to modify the parameter values of the prior, not its functional form, then the prior \( \pi(\cdot) \) is said to be conjugate for the given data model \( f(\cdot|\cdot) \).

Example 12.2

Consider a data model consisting of \( Y_1, \ldots, Y_n \sim iid N(\mu, \sigma^2) \) where \( \sigma^2 \) is considered known, and our interest is in the fixed parameter \( \mu \). Let the prior for \( \mu \) be specified as \( \mu \sim N(\lambda, \tau^2) \), where both \( \lambda \) and \( \tau^2 \) are specified values. The posterior of \( \mu \) is easily shown to be a normal distribution with parameters

\[
\frac{\sigma^2 \lambda + \tau^2 n \bar{y}}{n \tau^2 + \sigma^2}; \quad \frac{\sigma^2 \tau^2}{n \tau^2 + \sigma^2}.
\]

Note also in this example that the posterior mean may be written as a weighted average of the prior mean \( \lambda \) and the usual data estimator \( \bar{y} \) as

\[
\frac{n}{\bar{y}} \bar{y} + \frac{1}{\tau^2} \lambda
\]

\[
\frac{n}{\bar{y} \tau^2} + \frac{1}{\tau^2}.
\]
This form also indicates why many statisticians write model formulations to which a Bayesian analysis is to be applied in terms of “precision” parameters (here $1/\sigma^2$ and $1/\tau^2$) rather than variances.

12.3 Noninformative Priors

In some ways the phrase “noninformative prior” is an unfortunate title for several of the methods for prior formulation discussed in this section. In the sense that noninformative implies providing no information, not all “noninformative” priors should be considered as such. Nevertheless, this has become the standard heading under which to consider the methods for formulating priors that we consider here.

12.3.1 Proper Uniform Priors

The parameter space of some data or observation models is bounded both above and below. For example, a binomial model with parameter $\theta$ implies that $0 < \theta < 1$. A multinomial model with $k + 1$ categories and parameters $\theta_1, \ldots, \theta_k$ implies that both $0 < \theta_j < 1$ for $j = 1, \ldots, k$ and that $\sum_j \theta_j \leq 1$. In these situations, placing a uniform prior on the allowable interval allows us to express a prior belief that gives no preference to any of the possible values of the parameter $\theta$. We have already seen examples of a uniform priors used in this manner in previous examples that combine binomial data models with priors that take $\theta \sim Unif(0, 1)$.

In other models, even when the data model parameter space is unbounded in one or both directions, it may be possible to determine an interval $(a, b)$ such that it is either physically impossible, or scientifically implausible that
\[ \theta < a \text{ or } \theta > b. \] In these cases, it may be reasonable to assign \( \theta \) a prior distribution that is uniform on the given interval.

Uniform priors sometimes (but not always) simplify calculation of the integral that appears in the denominator of expression (10.1) since then we have,

\[
\int_{\Theta} f(y|\theta)\pi(\theta) d\theta = \frac{1}{b-a} \int_{a}^{b} f(y|\theta) d\theta.
\]

An objection to uniform priors as “noninformative” is that uniform distributions are not invariant to transformation. For example, if \( \theta \sim U(0, 1) \) then \( \eta = 1/\theta \) has density \( h(\eta) = 1/\eta^2; \quad 1 < \eta < \infty. \) Thus, the indifference that would seem to be expressed by the uniform prior on \( \theta \) does not translate into indifference about values of \( \eta \), although the data model may be equivalently expressed as either \( f(y|\theta) \) or \( f(y|\eta) \). Note that this is an objection to uniform priors being thought of as noninformative in nature, not as an objection to uniform priors per se.

### 12.3.2 Improper Priors

For data models that have parameters with unbounded parameter space from one or both directions, an extension of the idea of giving all possible values equal prior weight results in improper priors of the form

\[ \pi(\theta) = 1; \quad \theta \in \Theta, \]

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

\[
\int_{\Theta} f(y|\theta) d\theta = K(\theta) < \infty,
\]
then the posterior distribution

$$p(\theta|y) = \frac{f(y|\theta)}{\int_\Theta f(y|\theta) \, d\theta} = \frac{f(y|\theta)}{K(\theta)},$$

will exist and will integrate to 1. It has actually become quite popular to use improper priors, particularly for some elements of the parameter vector $\theta$ in complex models that involve a parameter $\theta$ of high dimension.

Improper priors perhaps deserve the label “noninformative” in that posterior distributions in simple cases with improper priors often result in essentially the same conclusions that would be reached under the sampling distribution of non-Bayesian estimators.

**Example 12.3**

Consider again the situation of Example 12.2 in which $Y_1, \ldots, Y_n \sim iid N(\mu, \sigma^2)$ with $\sigma^2$ 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 $\mu$ as $\pi(\mu) = 1; \ -\infty < \mu < \infty$. The resulting posterior is,

$$p(\mu|y) = \frac{\exp\left\{ -\frac{n}{2\sigma^2}(\bar{y} - \mu)^2 \right\}}{\int_{-\infty}^{\infty} \exp\left\{ -\frac{n}{2\sigma^2}(\bar{y} - \mu)^2 \right\}}$$

$$\propto \exp\left\{ -\frac{n}{2\sigma^2}(\mu - \bar{y})^2 \right\},$$

which is the density of a normal distribution with mean $\bar{y}$ and variance $\sigma^2/n$, that is, $p(\mu|y)$ is $N(\bar{y}, \sigma^2/n)$. While we have not yet discussed Bayesian estimation and inference, it should be intuitive that a reasonable point estimate of $\mu$ is the expectation of the posterior distribution which would be, in this case, $\bar{y}$ which agrees with any sensible non-Bayesian estimation. Similarly, a reasonable 90% interval estimate would be the central 90% of the posterior
density or distribution, namely $\bar{\bar{y}} \pm 1.645 (\sigma^2/n)$ which again agrees with what would be obtained from a non-Bayesian approach.

As illustrated by Example 12.3, improper priors often lead to situations in which the prior actually plays no role at all or is, in truth, “noninformative”. It may be the case that we wish to take this approach for certain elements of $\theta$, while placing informative proper prior distributions on our knowledge of other elements of $\theta$. The standard caution in use of improper priors is that one must be certain that the resulting posterior is in fact a distribution. This is not always a trivial matter.

12.3.3 Jeffreys’ Priors

Consider again the objection to using uniform prior distribution that they are not invariant to transformation in expressing lack of knowledge or indifference about values of a parameter. Jeffreys (1961) introduced a procedure for prior formulation based on the idea that any noninformative prior should be equivalent, in terms of expression of prior knowledge, on different scales.

To understand this, consider a data model with scalar parameter, and a procedure for assigning a prior, such as priors that are uniform on the range of the parameter space. If applied to the data model $f(y|\theta)$ with parameter space $\theta \in \Theta \equiv (\theta_1, \theta_2)$, this procedure results in the prior $\pi_\theta(\theta)$. Now consider an alternative parameterization using $\eta \equiv h(\theta)$ for some one-to-one transformation $h(\cdot)$. An equivalent model is now $f(y|\eta)$ with parameter space $\eta \in (h(\theta_1), h(\theta_2))$. Applying the same procedure for prior formulation as used under the model $f(y|\theta)$ results in a prior $\pi_\eta(\eta)$. But, the original prior $\pi_\theta$ also implies a distribution for $\eta \equiv h(\theta)$ as,

$$\pi'_\eta(\eta) = \pi_\theta(h^{-1}(\eta)) \left| \frac{dh^{-1}(\eta)}{d\eta} \right|$$
Jeffreys’ idea was that the procedure for assigning priors $\pi_\theta(\theta)$ and $\pi_\eta(\eta)$ is invariant under transformation if

$$\pi_\eta(\eta) = \pi_\eta'(\eta),$$

that is, if the prior assigned under the model $f(y|\eta)$ is the same as the distribution that results from the prior assigned under model $f(y|\theta)$ by transforming $\theta$ into $\eta$. As we have seen in Section 12.3.1, assigning uniform priors on the ranges of parameter spaces does not result in this property.

The suggestion Jeffreys gave for a procedure to assign priors that would result in this property was to take, under a model $f(y|\theta)$,

$$[\pi_\theta(\theta)]^2 \propto E \left[ \left( \frac{d \log f(y|\theta)}{d \theta} \right)^2 \right]$$

$$= -E \left[ \frac{d^2 \log f(y|\theta)}{d \theta^2} \right]$$

$$= I(\theta),$$

or,

$$\pi_\theta(\theta) = \{I(\theta)\}^{1/2}. \quad (12.2)$$

The form (12.2) is thus known as Jeffreys prior.

To verify that the procedure (12.2) does result in the desired property for priors, apply this procedure to the model $f(y|\eta)$, which gives,

$$[\pi_\eta(\eta)]^2 \propto E \left[ \left( \frac{d \log f(y|\eta)}{d \eta} \right)^2 \right]$$
12.3. NONINFORMATIVE PRIORS

\[
= E \left[ \left( \frac{d \log f(y|\theta)}{d \theta} \right)^2 \left| \frac{d \theta}{d \eta} \right|^2 \right]
\]

\[
= E \left[ \left( \frac{d \log f(y|\theta)}{d \theta} \right)^2 \right] \left| \frac{d \theta}{d \eta} \right|^2
\]

\[
= I(\theta) \left| \frac{d \theta}{d \eta} \right|^2,
\]

or,

\[
\pi_\eta(\eta) = \{I(\theta)\}^{1/2} \left| \frac{d \theta}{d \eta} \right|.
\]

Now, using (12.1), the distribution for \(\eta\) implied by (12.2) is

\[
\pi'(\eta) = \{I(\theta)\}^{1/2} \left| \frac{d \theta}{d \eta} \right|.
\]

Thus, \(\pi'_\eta(\eta) = \pi_\eta(\eta)\) and the approach suggested by Jeffreys does result in the property desired.

Example 12.4

Suppose that we have a single observation corresponding to the data model

\[ Y \sim Bin(\theta, n) \]

where \(n\) is fixed. We now have two choices for assigning \(\theta\) a so-called noninformative prior distribution. The first would be to take \(\pi_1(\theta) = 1; \ 0 < \theta < 1\), while the second would be to use the procedure of Jeffreys. In this case,

\[
I(\theta) = -E \left[ \frac{d^2 \log f(y|\theta)}{d\theta^2} \right]
\]

\[
= \frac{n}{\theta(1-\theta)},
\]

so that Jeffreys prior would be \(\pi_2(\theta) \propto \{\theta(1-\theta)\}^{-1/2}\). As we have already seen in Example 10.1, the uniform prior results in a beta posterior with para-
meters $1 + y$ and $1 + n - y$. A parallel derivation gives the posterior associated with the Jeffreys prior as beta with parameters $(1/2) + y$ and $(1/2) + n - y$.

In principle, Jeffreys’ method for forming noninformative priors can be extended to the case of vector valued $\theta$ by taking

$$
\pi_\theta(\theta) \propto |I(\theta)|^{1/2},
$$

where $|I(\theta)|$ here denotes the determinant of the expected information matrix $I(\theta)$. This type of multidimensional prior can be difficult to achieve in practice, however, and Jeffreys priors are usually seen in simpler cases with scalar $\theta$.

### 12.4 Priors for Vector Parameters

Although much of what we have presented applies in principle to parameter vectors $\theta$ (e.g., Jeffreys priors), the practical assignment of a joint prior to a $p$-dimensional parameter $\theta$ can become less than an easy matter. Two techniques that are often used in such situations are to take joint priors as products of individual (marginal) priors in the same way we would form a joint distribution for independent random variables, and to specify priors for some of the components of $\theta$ 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 $\theta \equiv (\theta_1, \theta_2)^T$.

**Example 12.5**

Consider again the beta-binomial mixture model of Example 10.2. There we had

$$
Y_1, \ldots, Y_m \sim \text{indep Bin}(\theta_i) \\
\theta_1, \ldots, \theta_m \sim \text{iid Beta}(\alpha, \beta)
$$
What is necessary to conduct a Bayesian analysis for this model is a prior for the parameter $\lambda \equiv (\alpha, \beta)^T$. Now, the parameter space is $\alpha > 0$, $\beta > 0$, but if we first reparameterize the beta mixing distribution in terms of parameters,

$$
\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \eta = \frac{1}{\alpha + \beta + 1},
$$

then $0 < \mu < 1$ and $0 < \eta < 1$. We might then assign the joint prior as

$$
\pi(\mu, \eta) = \pi_\mu(\mu) \pi_\eta(\eta),
$$

where both $\pi_\mu(\cdot)$ and $\pi_\eta(\cdot)$ are uniform distributions on the interval $(0, 1)$.

Derivation of the posterior $p(\alpha, \beta|y)$ would, in this example, require the use of simulation methods.

### Example 12.6

Consider again the normal one-sample problem of Example 12.3, but now not assuming that the variance $\sigma^2$ is known. Here, it would not be possible to consider only the distribution of $\bar{Y}$ in the likelihood, since $\bar{Y}$ is sufficient for $\mu$ but not $\sigma^2$. Thus, we must work with the full joint distribution of $Y_1, \ldots, Y_n$, which can be written as,

$$
f(y|\mu, \sigma^2) = \{2\pi\sigma^2\}^{n/2} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right]
$$

$$
= \{2\pi\sigma^2\}^{n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \right]
$$

$$
- \frac{1}{2\sigma^2} n (\bar{y} - \mu)^2.
$$

One way to assign the joint prior $\pi(\mu, \sigma^2)$ to this model is to use the conditional prior $\pi_1(\mu|\sigma^2)$ and the marginal prior $\pi_2(\sigma^2)$ as,

$$
\pi_1(\mu|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{\kappa_0}{2\sigma^2} \{\mu - \mu_0\}^2 \right]
$$
\[ \pi_2(\sigma^2) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \{\sigma^2\}^{-(\alpha_0+1)} \exp\{-\beta_0/\sigma^2\}. \]

Here, \( \pi_1(\cdot) \) is normal with parameters \( \mu_0 \) and \( \sigma^2/\kappa_0 \), while \( \pi_2(\cdot) \) is inverse gamma with parameters \( \alpha_0 \) and \( \beta_0 \), which are conjugate for \( \mu \) in a model with \( \sigma^2 \) assumed known and \( \sigma^2 \) with \( \mu \) 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 \( p(\mu|y) \) is a \( t- \) distribution, the marginal posterior \( p(\sigma^2|y) \) is an inverse gamma distribution, and the conditional posterior \( p(\mu|\sigma^2, y) \) is a normal distribution (e.g., Gelman, Carlin, Stern and Rubin, 1995, pp. 72-73). What is important for us at this point is the use of the conditional prior \( \pi_1(\mu|\sigma^2) \). First, the fact that \( \sigma^2 \) appears in this prior indicates that \( \mu \) and \( \sigma^2 \) are not independent in the joint prior \( \pi(\mu, \sigma^2) \).

Secondly, the fixed constant \( \kappa_0 \) plays the role of the “number of observations” we believe our prior information about \( \mu \) is “worth” in the overall problem; note that \( \pi_1(\cdot) \) has the form of the distribution of the sample mean of \( \kappa_0 \) observations taken from a \( N(\mu, \sigma^2) \) distribution.

Gelman, Carlin, Stern and Rubin (1995) use this example also as a simple case in which simulation-based derivation of the posterior is valuable. While the marginal posteriors are of relative “nice” forms, and the joint posterior can be derived in closed form, this joint posterior is not necessarily easily manipulated to find, for example, expectations, quantiles, and so forth. But the fact that the conditional posterior \( p(\mu|\sigma^2, y) \) is normal and the marginal \( p(\sigma^2|y) \) is inverse gamma indicates that these two univariate distributions may be easily simulated from. An algorithm to simulate values from the joint posterior is then also easily constructed as follows:

1. Generate (or draw or simulate) a value \( \sigma^{2*} \) from the marginal posterior
2. Using this value, generate a value \( \mu^* \) from the conditional posterior \( p(\mu|\sigma^2, y) \).

3. The resulting pair \( (\mu^*, \sigma^2*) \) is one value from the joint posterior \( P(\mu, \sigma^2|y) \).

4. Repeat this process a large number \( M \) times, and make inference based on the empirical distribution of the set of values \( \{ (\mu_j^*, \sigma_j^2) : j = 1, \ldots, M \} \).

It is also instructive to contrast Example 12.6 to what would happen if, instead of specifying \( \pi_1(\mu|\sigma^2) = N(\mu_0, \sigma^2/\kappa_0) \) we would use independent priors and simply take \( p(\mu) \) as \( N(\mu_0, \tau^2_0) \) for example. In this case (e.g., Gelman, Carlin, Stern and Rubin, 1995, Chapter 3.4) \( \mu \) and \( \sigma^2 \) are still dependent in the joint posterior, the conditional posterior of \( \mu \) given \( \sigma^2 \) is again normal, but the marginal posterior of \( \sigma^2 \) cannot be derived in closed form. The simulation algorithm given above can also be used in this situation, but the initial step of generating \( \sigma^2* \) becomes more difficult, and must make use of an indirect method of simulation (e.g., inversion, rejection sampling, etc.).
Chapter 13

Basic Estimation and Inference

In some ways, discussing Bayesian methods of inference is a very short topic. Under the concept of epistemic probability, a posterior distribution represents our knowledge about a parameter of interest. That is, given a posterior $p(\theta|y)$, we are free to make probability statements about $\theta$ with the understanding that our statements actually refer to our knowledge about the value of $\theta$ which, under the strict Bayesian viewpoint is a fixed parameter and under the viewpoint of Bayesian analysis of uncertainty is just some unknown quantity. That is, in this context it is perfectly acceptable to write a statement such as

$$Pr(a \leq \theta \leq b) = \alpha,$$

where $a$, $b$, and $\alpha$ are all particular real numbers. Nevertheless, there are a few particular issues relative to Bayesian inference that merit brief consideration. Before moving on to these, we will mention that, although inference is to be based on a posterior, there are several broad methods for obtaining those posteriors, several of which have been eluded to in the previous sections. We may find a posterior distribution through one of three avenues:

1. Analytical derivation.
2. Approximation.

3. Simulation.

The first of these, analytical derivation, we have seen in a number of the examples presented. Here, a posterior is derived in closed mathematical form, usually as a probability density function. Simple models with conjugate priors are a prime example of situations in which this approach is useful. We will not discuss approximation in this course, but will only mention that the most popular such approximation is known as the *Laplace Approximation*. For a very accessible introduction to this approximation see Carlin and Louis (2000) Chapter 5.2.2. Finally, simulation of posteriors has become the most widely used method for finding posteriors in use today. This is because it may be applied to a huge variety of models, many of which could not be approached by any other technique. Our department offers a course (the old Stat 601, the number of which was usurped for this current course, and has yet to receive a new designation) in which simulation of posteriors using methods known collectively as Markov Chain Monte Carlo (MCMC) are the bulk of what is discussed.

### 13.1 Point Estimation

While it is well and fine to indicate that, given a posterior distribution, one is free to make any probability statement desired, there is typically still a desire for concise summarization of a posterior. For example, we may desire as a summarization of the location of a posterior distribution a single value or “point estimate”. Here, I am about to start sounding like I’m teaching Statistics 101 because the values that are used are the posterior mean, median,
or mode. The median is not used that frequently, the mode is still popular,
but less than it was before the advent of MCMC simulation, and the mean is
often the quantity of choice in Bayesian analyses.

The posterior mode is often the easiest to find in an analytical approach,
because it does not depend on finding the posterior “normalizing constant”,
the denominator of

\[ p(\theta|y) = \frac{f(y|\theta) \pi(\theta)}{\int_{\Theta} f(y|\theta) \pi(\theta) d\theta}. \]

As shown for one case in Example 12.3, if \( \pi(\theta) \propto 1 \) for \( \theta \in \Theta \), then the
posterior mode is equal to the maximum likelihood estimate, since then,

\[ \max_{\theta} p(\theta|y) = \max_{\theta} f(y|\theta). \]

Use of the posterior mean or expected value can be justified based on
decision-theoretic grounds, if one considers squared error loss (e.g., Berger,
1985, p. 161). For scalar \( \theta \), the posterior mean is given in the obvious way as,

\[ E\{\theta|y\} = \int_{\Theta} \theta p(\theta|y) d\theta, \quad (13.1) \]

and is sometimes reported along with the posterior variance,

\[ \text{var}(\theta|y) = \int_{\Theta} (\theta - E\{\theta|y\})^2 d\theta. \quad (13.2) \]

For vector-valued parameters \( \theta \),

\[ E\{\theta|y\} = (E\{\theta_1|y\}, \ldots, E\{\theta_p|y\}), \]

and expression (13.1) continues to hold for the component quantities with \( \theta \)
replaced by \( \theta_j \) and \( p(\theta|y) \) replaced by \( p_j(\theta_j|y) \), the marginal posterior of \( \theta_j \);
\( j = 1, \ldots, p \). The same is true for the variances and expression (13.2), with
the additional covariances given as,
\[
\text{cov}(\theta_j, \theta_k) = \int_{\Theta} \int_{\Theta} (\theta_j - E\{\theta_j|y\})(\theta_k - E\{\theta_k|y\})p_{j,k}(\theta_j, \theta_k|y) \, d\theta_j \, d\theta_k,
\]
where \(p_{j,k}(\theta_j, \theta_k|y)\) is the joint marginal posterior of \(\theta_j\) and \(\theta_k\).

### 13.2 Interval Estimation

Although posterior variances and covariances can be, and often are, computed as given in expressions (13.2) and (13.3) they are typically not used to form interval estimates of \(\theta\) 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 \(\theta\). The Bayesian analog of confidence sets or intervals are typically called “credible sets” or “credible intervals”. All that is needed is a sensible way to make probability statements such as \(Pr(a \leq \theta \leq b) = 1 - \alpha\) and find the appropriate values of \(a\) and \(b\). The basic definition of a credible set for \(\theta\) is a set \(C\) such that
\[
1 - \alpha \leq Pr(\theta \in C|y) = \int_C p(\theta|y) \, d\theta.
\]
(13.4)

If \(\theta\) should happen to be discrete the integral in (13.4) is replaced with a summation.

For a given posterior \(p(\theta|y)\) there may be many sets \(C\) that satisfy (13.4). One technique that has been used to help get around this difficulty is to define a Highest Posterior Density credible set as a set
\[
C^* = \{\theta : p(\theta|y) \geq k(\alpha)\},
\]
where \(k(\alpha)\) is the largest constant such that \(C^*\) is a credible set. What this
means is that, for any \( \theta^* \in C^* \) and any other \( \theta \in C^{*'} \), where \( C^{*'} \) is the complement of \( C \),

\[
p(\theta^* | y) \geq p(\theta | y).
\]

In other words, the posterior density for any value of \( \theta \) included in the credible set is at least as great as that for any value not in the credible set.

While highest posterior density (HPD) credible sets are not hard to find for scalar \( \theta \), they can be quite difficult to determine in higher dimensions. In addition, HPD credible sets are not invariant to transformation of \( \theta \). For a more complete discussion of issues involved with credible sets, HPD credible sets and their extension to “optimal” credible sets see Berger (1985).

In many applications and, in particular, those in which the posterior is found through the use of simulation, a common practice is to use the “central” \( 1 - \alpha \) interval for any component of \( \theta \), regardless of whether it would qualify as an HPD interval or not. That is, if we wish a \( (1 - \alpha)100\% \) credible interval for \( \theta_j \), that interval is given by \( (L, U) \) where

\[
1 - \alpha/2 = \int_{-\infty}^{L} p(\theta_j | y) \, d\theta_j
\]

\[
1 - \alpha/2 = \int_{U}^{\infty} p(\theta_j | y) \, d\theta_j
\]

and where \( p(\theta_j | y) \) is the marginal posterior of \( \theta_j \).

### 13.3 Model Comparison

Suppose that we have two competing models denoted as \( M_1 \) and \( M_2 \) that we would like to compare in light of a set of observations \( y \). These models may differ in the number of parameters associated with covariates (e.g., a typical “variable selection” problem in regression), by one model having fixed values
for a portion of a vector-valued parameter (e.g., \( \mu = 0 \) in a normal data model), by having parameter values that are restricted to different regions of a partitioned parameter space (e.g., \( p < 0.5 \) versus \( p > 0.5 \) in a binomial model), by having different priors, or by having different data models \( f_1(y|\theta_1) \) and \( f_2(y|\theta_2) \) (e.g., gamma versus lognormal). Notice that, in particular, we do not require nested parameter spaces for our competing models.

Now, suppose we represent our beliefs about the possible models \( M_1 \) and \( M_2 \) in terms of a prior distribution, which will necessarily place distinct probabilities on the possible models \( M_1 \) and \( M_2 \) (e.g., \( \pi(M_1) = \gamma \) and \( \pi(M_2) = 1-\gamma \)). Let the values of this prior be represented as \( Pr(M_1) \) and \( Pr(M_2) \). The two models may then be compared by taking the ratio of posterior probabilities of the models as the “posterior odds ratio”.

\[
\frac{Pr(M_1|y)}{Pr(M_2|y)} = \frac{Pr(M_1) Pr(y|M_1)}{Pr(M_2) Pr(y|M_2)} = \frac{Pr(M_1)}{Pr(M_2)} BF(M_1, M_2)
\]

(13.6)

where \( BF(M_1, M_2) \) denotes the “Bayes Factor” of model \( M_1 \) relative to model \( M_2 \), namely,

\[
BF(M_1, M_2) = \frac{Pr(y|M_1)}{Pr(y|M_2)}.
\]

(13.7)

Ways to interpret the concept quantified in a Bayes factor include:

1. BF is ratio of posterior odds in favor of model \( M_1 \) to the prior odds in favor of model \( M_1 \), as,

\[
BF(M_1, M_2) = \frac{Pr(M_1|y)/Pr(M_2|y)}{Pr(M_1)/Pr(M_2)}.
\]

2. BF is a likelihood ratio, which is a direct interpretation of expression (13.7), that is, the ratio of the likelihood of the data \( y \) under model \( M_1 \)
13.3. MODEL COMPARISON

to the likelihood of the data $y$ under model $M_2$. Note, however, that
the Bayes Factor of (13.7) is written in terms of probabilities rather than
densities.

If the Bayes factor $BF(M_1, M_2)$ is greater than 1, then the posterior odds
in favor of model $M_1$ are increased from the prior odds. If the prior odds ratio
is taken to be 1 (i.e., $Pr(M_1) = Pr(M_2) = 0.5$) the the Bayes factor is equal
to the posterior odds ratio in favor of $M_1$. How big should $BF(M_1, M_2)$ be
in order for us to have substantially greater belief in $M_1$ than $M_2$? Kass and
Raftery (1995) give a slightly modified version of a scale suggested by Jeffreys
(1961) which suggests that values from 3.2 to 10 provide some evidence in favor
of $M_1$, values from 10 to 100 provide strong evidence, and values greater than
100 provide “decisive” evidence. These authors also suggest their own scale
which results in the same categories of evidence for ranges of Bayes factors 3 to
20 (some evidence), 20 to 150 (strong evidence) and greater than 150 (decisive
evidence).

Now, in a situation in which a model is formulated through density func-
tions, a given model $M_i$ is embodied through its data model $f_i(y|\theta_i)$ and its
prior $\pi_i(\theta_i)$ so that $Pr(y|M_i)$ is associated with the density,

$$h_i(y|M_i) = \int f_i(y|\theta_i) \pi_i(\theta_i) \, d\theta_i,$$

or, in the case of a hierarchical model,

$$h_i(y|M_i) = \int f_i(y|\theta_i) g(\theta_i|\lambda_i) \pi_i(\lambda_i) \, d\theta_i \, d\lambda_i.$$

The Bayes factor for models $M_1$ and $M_2$ is then often written in terms of these
densities as,

$$BF(M_1, M_2) = \frac{h_1(y|M_1)}{h_2(y|M_2)},$$

(13.8)
CHAPTER 13. BASIC ESTIMATION AND INFERENCE

where the notation $h_i(\cdot); i = 1, 2$ has been used to emphasize that these densities are not necessarily of the same form.

Now, the Bayes factor of (13.8) is based on densities, not probabilities (cf. 13.7), but to have interpretation as developed from the posterior odds ratio of expression (13.6) we still need it to be true that,

$$Pr(M_1|y) = \frac{Pr(M_1)h_1(y|M_1)}{Pr(M_1)h_1(y|M_1) + Pr(M_2)h_2(y|M_2)}.$$  

(13.9)

and similarly for $Pr(M_2|y)$. What is needed in order for (13.9) to hold? Essentially the same type of conditions as were needed in order for likelihoods to be proportional to the probability of data $y$ for given parameters $\theta$ in Section 8.3.1 of Part II of these notes. Here, we extend the basic idea of using a linear approximation to the integral mean value theorem (what we actually depended on for likelihood is usually called the intermediate value theorem) from the case of independent variables to that of a joint density $h_i(y|M_i)$.

Consider a joint density $h_i(y|M_i)$ evaluated at the observed values $y$. In order to apply the mean value or intermediate value theorems for multiple integrals, we can use the following:

(i) The value $y$ implies that there exists an $n$—dimensional ball $\delta_y$, centered at $y$, such that for the random variable $Y$ associated with the situation that led to $y$, 

$$Pr(Y \in \delta_y) = \int_{\delta_y} h_i(t|M_i) dt.$$  

(ii) The set $\delta_y$ is connected in the sense that any two points of $\delta_y$ can be joined by an arc that lies entirely in $\delta_y$. 


(iii) The density function $h_i(\cdot)$ is continuous on $\delta_y$.

Under these conditions, and the assumption that $\delta_y$ is of sufficiently small volume, a similar argument as used to connect likelihood with probability gives that

$$\int_{\delta_y} h_i(t|M_i) \, dt \approx |\delta_y|h_i(y|M_i),$$

or $Pr(y|M_i) \propto h_i(y|M_i)$. Now, the last thing needed in order to make (13.9) hold is then that the volume $|\delta_y|$ be the same for models $M_1$ and $M_2$. If this is the case, then (13.9) may be used on the right hand side of (13.6) to give,

$$\frac{Pr(M_1|y)}{Pr(M_2|y)} = \frac{Pr(M_1)}{Pr(M_2)} \frac{h_1(y|M_1)}{h_2(y|M_2)} = \frac{Pr(M_1)}{Pr(M_2)} BF(M_1, M_2),$$

with the Bayes factor $BF$ defined as in expression (13.8). For the majority of models we deal with, the conditions needed to take the Bayes factor as a ratio of densities is not likely to raise any concern. But, if competing data models would happen to differ in functional form, then greater caution is needed.

Example 13.1

In the continuation of Example 5.2 contained in Section 7.3.1 of these notes we considered a number of possible models for analysis of a hypothetical study of the effect of violent cartoons and stories on aggression in children. The measured quantities were responses of “like”, “do not like”, or “don’t care about” made to pictures shown to children who had watched or been read “violent” or “happy” cartoons or stories. In considering possible random variables that could be constructed for this setting, and possible distributions that might be assigned to those random variables, we discussed possible models based on binomial, multinomial, and beta distributions. Any of these models could be
analyzed through a Bayesian procedure. If we were to attempt a comparison of two of these models, say the beta model and the multinomial model, through a Bayes factor, the issue described immediately above would become important. Suppose, for example, that we took $y$ to be the proportion of the total “score” possible over 20 pictures by making “don’t like” correspond to 0, “don’t care” correspond to 1 and “like” correspond to 2, then summing these scores across pictures for each child and dividing by 40 (the maximum “aggression score” possible).

We might then take $f_1(y|\theta_1)$ to be beta with a (joint) prior for $\theta_1$ formulated as a product of uniform $(0, 1)$ densities as described in Example 12.5 in Chapter 12.4. Also suppose that we took $f_2(y|\theta_2)$ to be multinomial with a (joint) prior for $\theta_2$ given as a Dirichlet. It would be possible, of course, to derive $h_1(y|M_1)$ from the beta model and $h_2(y|M_2)$ from the multinomial model, and then just slap these distributions into expression (13.8) calling the result a Bayes factor. This would clearly be a mistake, and the formal reason for this is the above discussion.

Another potential difficulty with the use of Bayes factors occurs with the use of improper prior distributions. In this case it may be true that numerical values may be computed for the integrals preceding expression (13.8) (this will be the case whenever improper priors lead to proper posteriors) but the Bayes factor is nevertheless undefined, since these integrals cannot be considered proportional to $Pr(y|M_i)$ as in expression (13.7). The distinction is that, in derivation of a posterior, we consider

$$p(\theta|y) \propto f(y|\theta),$$

where the proportionality constant is the integral

$$\int_{\Theta} f(y|\theta)\pi(\theta) \, d\theta.$$
Thus, so long as this integral is finite, the posterior exists as a proper distribution. For use in a Bayes factor, however, we need this same integral to result in a density for the argument \( y \) as,

\[
h(y) = \int_{\Theta} f(y|\theta)\pi(\theta) \, d\theta,
\]

such that

\[
\int_{-\infty}^{\infty} h(y) \, dy = 1.
\]

If \( f(y|\theta) \) is a proper density with support \( \Omega \) and \( \pi(\theta) \) is an improper prior over \( \Theta \), the function \( h(\cdot) \) cannot be a density with argument \( y \) because then,

\[
\int_{-\infty}^{\infty} h(y) \, dy = \int_{\Theta} \int_{\Omega} f(y|\theta)\pi(\theta) \, d\theta \, dy
\]

\[
= \int_{\Theta} \int_{\Omega} f(y|\theta)\pi(\theta) \, dy \, d\theta
\]

\[
= \int_{\Theta} \pi(\theta) \, d\theta,
\]

which is not finite.

Despite the potential difficulties with Bayes Factors they can often be a useful method for comparing among models. In addition, for cases in which Bayes factors are defined, they may also be used to conduct what might be thought of as Bayesian tests of hypotheses.

**Example 13.2**

Consider again the analysis of sex ratio at birth in Guanacos of Example 11.1 discussed at some length in Chapter 11. With \( \theta \) being defined as the probability of a male birth, we may formulate a hypothesis in this example of \( H_0 : \theta \leq 0.5 \) and an alternative hypothesis of \( H_1 : \theta > 0.5 \). Let these hypotheses correspond to the models \( M_2 \) and \( M_1 \), respectively. That is, model \( M_2 \) corresponds
to $\theta \leq 0.5$ while model $M_1$ corresponds to $\theta > 0.5$. Suppose that, \textit{a priori} we give these two models equal weight, so that $Pr(M_1) = Pr(M_2) = 0.5$. Since, in this case, models correspond to values of $\theta$, these hypotheses are naturally reflected in a uniform prior for $\theta$ on the interval $(0, 1)$; there are other priors that could also reflect these hypotheses (anything with a density symmetric about 0.5 would suffice), but the uniform serves nicely in this case. The prior odds of $M_1$ to $M_2$ is then $0.5/0.5 = 1$. In this case, then, expression (13.6) indicates that the posterior odds of $M_1$ to $M_2$ become

$$\frac{Pr(M_1|y)}{Pr(M_2|y)} = BF(M_1, M_2).$$

Now,

$$Pr(M_1|y) = Pr(\theta > 0.5|y),$$

and,

$$Pr(M_2|y) = Pr(\theta \leq 0.5|y).$$

At the end of four years of data collection (1990) we had that, beginning with a uniform prior, the posterior distribution of $\theta$ was beta with parameters $\alpha = 185$ and $\beta = 161$ (see Chapter 11). Thus, the posterior odds, or Bayes factor, in favor of model $M_1$ are

$$\frac{Pr(\theta > 0.5|y)}{Pr(\theta \leq 0.5|y)} = \frac{Pr(\theta > 0.5|\alpha = 185, \beta = 161)}{Pr(\theta \leq 0.5|\alpha = 185, \beta = 161)}.$$

These values are easily computed for a beta distribution to be,

$$\frac{Pr(\theta > 0.5|y)}{Pr(\theta \leq 0.5|y)} = \frac{0.90188}{0.09812} = 9.1913.$$

Using typical scales for “strength of evidence” as discussed earlier in this Section we would conclude that there is some, but not strong, evidence against $M_2$ in favor of $M_1$, which agrees with our assessment from the credible intervals of Chapter 11.
13.4 Predictive Inference

It is often the case that we desire a predictive distribution for a “new” observation $y^*$, presumed to follow the same model as the components of $y$. Such predictive distributions may be useful in their own right for the purposes of forecasting if $y^*$ lies outside of the extent (or spatial and temporal window) of the available observations $y$, or prediction if $y^*$ lies within the extent of the available data but corresponds to a location or time or general “position” that is not observed in the data. In addition, predictive distributions may be useful in model assessment from the viewpoint that a good model predicts well. Quantifying the agreement of “replicated” data with actual data, where the replicated data are simulated from a predictive distribution is one way to accomplish this type of model assessment (e.g., Gelman, Carlin, Stern and Rubin 1995, Chapter 6.3).

The fundamental distribution used in predictive inference is the *posterior predictive* distribution $p(y^*|y)$, which is in concert with the viewpoint of Bayesian analysis of unknowns in that inference about unobserved quantities are made on the basis of the conditional distributions of those quantities given the observed data. Consider either a simple model consisting of $f(y|\theta)$ and the prior $\pi(\theta)$, or a hierarchical model in which we view the mixture $\int g(\theta|\lambda)\pi(\lambda)\,d\lambda$ to constitute a prior for $\theta$ as $\pi_\theta(\theta)$. Using the notation $p(\cdot)$ as a generic probability density or mass function, and assuming that $y^*$ is conditionally independent of $y$ given $\theta$,

$$p(y^*|y) = \frac{p(y^*, y)}{p(y)}$$

$$= \frac{1}{p(y)} \int_{\Theta} p(y^*, y, \theta) \,d\theta$$

$$= \frac{1}{p(y)} \int_{\Theta} p(y^*|\theta)p(y|\theta)p(\theta) \,d\theta$$
\[ = \frac{1}{p(y)} \int_{\Theta} p(y^*|\theta)p(\theta,y) \, d\theta \]
\[ = \int_{\Theta} p(y^*|\theta)p(\theta|y) \, d\theta. \]

(13.11)

Conditional independence of \( y^* \) and \( y \) is used in the transition from line 2 to line 3 of expression (13.11). Notice that (13.11) indicates the posterior predictive density of \( y^* \) is the expected value of \( p(y^*|\theta) \), which will be of the same form as the data model \( f(y|\theta) \), taken with respect to our knowledge about \( \theta \) as represented by the posterior distribution \( p(\theta|y) \). Now, reverting to the use of \( f(\cdot) \) for the data model density and \( p(\cdot|y) \) for a posterior density, the posterior predictive density of \( y^* \) may be written as,

\[ p(y^*|y) = \int_{\Theta} f(y^*|\theta)p(\theta|y) \, d\theta. \]

(13.12)

In a hierarchical model it is also possible to derive a conditional posterior predictive density for \( y^* \) as,

\[ p(y^*|y,\lambda) = \int_{\Theta} f(y^*|\theta)p(\theta|\lambda,y) \, d\theta. \]

(13.13)

Expression (13.12) is the same as expression (13.11) with notation to make the roles of the data model \( f(y^*|\theta) \) and posterior \( p(\theta|y) \) explicit. Expression (13.13) can be derived along the lines of (13.11), leading to the conditional posterior predictive distribution in notation parallel to that of the ordinary or marginal posterior predictive density (13.12). While the predictive density (13.13) is certainly less common in typical Bayesian analysis, it does appear on occasion, for example in the dynamic models of West and Harrison (1989). To make use of the conditional distribution (13.13) one would need a plug-in value for \( \lambda \) in the same way that this is required for use of the conditional posterior \( p(\theta|\lambda,y) \) of expression (10.7).
Chapter 14

Simulation of Posterior Distributions

The basic idea of simulation of a posterior distribution is quite simple. Suppose that we have a model that consists of a data model \( f(y|\theta) \) and a prior \( \pi(\theta) \). The posterior is

\[
p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int_{\Theta} f(y|\theta)\pi(\theta) d\theta} \propto f(y|\theta)\pi(\theta).
\]  

(14.1)

Note that both the left and right hand sides of this expression must be considered as functions of \( \theta \). Suppose that the integral in the denominator of the first line of (14.1) is intractable so that the posterior \( p(\theta|y) \) cannot be derived in closed form. A typical goal in simulation of posteriors is then to use the last line of expression (14.1) to allow simulation of values from \( p(\theta|y) \) even in such cases. In a hierarchical model we have a data model \( f(y|\theta) \), an intermediate distribution \( g(\theta|\lambda) \), which can be considered either a part of
the model as a mixing distribution or a part of the prior, depending on the viewpoint taken, (see Chapter 10) and a prior or final stage prior \( \pi(\lambda) \). For this type of model with \( \theta \in \Theta \) and \( \lambda \in \Lambda \), we may have interest in the joint posterior,

\[
    p(\theta, \lambda | y) = \frac{f(y|\theta)g(\theta|\lambda)\pi(\lambda)}{\int_{\Lambda} \int_{\Theta} f(y|\theta)g(\theta|\lambda)\pi(\lambda) \, d\theta \, d\lambda}
\]

\( \propto f(y|\theta)g(\theta|\lambda)\pi(\lambda). \) (14.2)

Again, the integrals in the first line of (14.2) may be intractable, and a common goal is to use the last line of (14.2) to allow simulation from \( p(\theta, \lambda | y) \) without explicitly evaluating these integrals. There are several ways that simulation from the posteriors (14.1) and (14.2) can be accomplished, but before going into these methods it is useful to identify several basic principles of simulation.

### 14.1 Fundamental Principles of Simulation

It is useful to set forth several basic truths of simulation procedures, which we do in this section, using generic notation for random variables \( X, Y, \) and \( Z \) and their density functions as \( f(x), f(y) \) and \( f(z) \).

1. Averaging over Simulations Estimates Expectations

   This first principle embodies the fundamental idea of Monte Carlo estimation, which we have already encountered in Chapter 8.6 in discussing Parametric Bootstrap methods. Consider first the case of a univariate random variable \( X \) with distribution function \( F(x) \). If we obtain simulated values \( \{x_j^* : j = 1, \ldots, M\} \) as independent and identical realizations
from $F(x)$, a Monte Carlo estimate of the expected value $E(X)$ is,

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

Indeed, for any suitable function $q(X)$,

$$\hat{E}_M\{q(X)\} = \frac{1}{M} \sum_{j=1}^{M} q(x^*_j). \quad (14.3)$$

Thus, simulation and averaging accomplishes estimation of the integral

$$E\{q(X)\} = \int q(x) \, dF(x),$$

where it is assumed that the dominating measure is Lebesgue or counting measure. That $\hat{E}_M\{q(X)\}$ is consistent for $E\{q(X)\}$ follows immediately from the law of large numbers. Similarly, the empirical distribution function of the values $\{x^*_j : j = 1, \ldots, M\}$,

$$F_M(x) = \frac{1}{M} \sum_{j=1}^{M} I(x^*_j \leq x),$$

converges to $F(x)$ for each fixed $x$ as $M \to \infty$. The Gilvenko-Cantelli Theorem gives that this convergence is uniform in $x$ (e.g., Billingsley, 1986, p. 275).

These results, which form the basis for Monte Carlo estimation, may be extended to sequences of random variables that are not independent, if such sequences have a property called ergodicity. A complete coverage of ergodicity is beyond the scope of these notes, but an intuitive understanding of the fundamental idea can be gained as follows. Consider a distribution $F(x); x \in \Omega$ from which we would like to simulate a sample $\{x^*_j : j = 1, \ldots, M\}$ so that the above Monte Carlo estimators $\hat{E}_M\{q(X)\}$ and $F_M(x)$ converge to $E\{q(X)\}$ and $F(x)$ as in the case of independent
realizations. Now, suppose we are unable to simulate values from \( F(x) \) directly, but we are able to construct a sequence of random variables \( X(t) \equiv \{ X(t) : t = 0, 1, \ldots \} \) called a \textit{chain} in such a way that the above results continue to hold using values simulated from \( X(t) \) rather than \( F(x) \). This can only occur, for dependent \( X(t) \), if the sequence “mixes” over the set \( \Omega \) in the proper manner. Suppose we partition \( \Omega \) 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 \( \Omega \), then the results desired will continue to hold using \( \{ x^*(t) : t = B, B + 1, \ldots, B + M \} \) in place of \( \{ x^*_j : j = 1, \ldots, M \} \). What is needed, then, is for the sequence \( X(t) \) to “visit” or “mix over” each of the subsets \( \Omega_1, \ldots, \Omega_k \) with the correct frequencies, and with sufficient rapidity that we don’t have to wait until \( M \) becomes too large for the approximations to be adequate. Sequences \( X(t) \) have have these behaviors are called ergodic.

The construction of what are called \textit{Markov chain samplers} is concerned with developing chains that are ergodic and, importantly, mimic the probabilistic behavior of the distribution \( F(x) \), simulation from which was goal in the first place. In this context, \( F(x) \) is often called the “target” distribution. We will encounter several Markov chain samplers in Chapters 14.4 and 14.5.

2. Expectations wrt to \( F \) Can be Estimated by Simulating from \( G \).

Consider basic Monte Carlo estimation of \( E\{q(X)\} \) as described in the first part of item 1 above. There, it was assumed that values \( \{ x^*_j : j =
14.1. FUNDAMENTAL PRINCIPLES OF SIMULATION

..., \ldots, M\} had been simulated from \( F(x) \), the distribution of \( X \). Assume
that \( F(x) \) has a corresponding density function \( f(x) \). The expected value
of \( q(X) \) may than also be written as,

\[
E\{q(X)\} = \int q(x) \frac{f(x)}{g(x)} g(x) \, d\mu(x),
\]

for some function \( g(\cdot) \) with domain matching the support \( \Omega \) of \( f(\cdot) \), and
\( \mu \) either Lebesgue or counting measure. If \( g(\cdot) \) corresponds to a density
over \( \Omega \), then a Monte Carlo estimate of the expectation can be computed
as

\[
\hat{E}_M\{q(X)\} = \frac{1}{M} \sum_{j=1}^{M} q(x^*_j) \frac{f(x^*_j)}{g(x^*_j)},
\]  

(14.4)

where \( \{x^*_j : j = 1, \ldots, M\} \) have been simulated from the distribution
with density \( g(x); x \in \Omega \). Expression (14.4) is called an importance
sampling estimator of \( E\{q(X)\} \), and \( g(x) \) is the importance (or impor-
tance sampling) distribution. While importance sampling had its origins
in techniques useful to reduce the variance of Monte Carlo estimators, in
most statistical applications it appears in one of two situations. First, notice
that importance sampling can be used to formulate a Monte Carlo
estimator of the integral of nearly any function, regardless of whether
that integral is originally expressed as an expectation or not. That is,
the integral of a function \( h(x) \) can be expressed as the expectation of
\( h(x)/g(x) \) with respect to a distribution having density \( g(\cdot) \) as long as
the support of \( g \) matches the limits of the original integration. Secondly,
and perhaps most commonly, importance sampling is useful in situations
where the distribution with density \( f(\cdot) \) in (14.4) is difficult to simulate
from, but for which we can find an importance distribution with density
\( g(\cdot) \), with the same support, but from which it is easy to simulate. A key
to the successful application of importance sampling, however, is not only
finding an importance distribution from which it is easy to sample, but also one that is “good” in the sense that it results in rapid convergence of \( \hat{E}_M\{q(X)\} \) to \( E\{q(X)\} \).


For two random variables \( X \) and \( Y \) with marginal and conditional densities \( f(x) \) and \( f(y|x) \), simulation of a value \( x^* \) from \( f(x) \) followed by simulation of a value \( y^* \) from \( f(y|x^*) \), gives one value \( y^* \) from the marginal density

\[
f(y) = \int f(y|x) f(x) \, dx.
\]  
(14.5)

Thus, if we want a set of simulated values \( \{y_j^* : j = 1, \ldots, M\} \) from the distribution with density \( f(y) \), we simulate \( M \) values \( \{x_j^* : j = 1, \ldots, M\} \) independently from \( f(x) \) and, for each of these values, simulate a value \( y_j^* \) from \( f(y|x_j^*) \).

Similarly, simulation of one value \( z^* \) from \( f(z^*) \), followed by simulation of one value \( x^* \) from \( f(x|z^*) \), followed in turn by one value \( y^* \) simulated from \( f(y|x^*) \) produces one value \( y^* \) from the distribution,

\[
f(y) = \int \int f(y|x) f(x|z) f(z) \, dx \, dz,
\]  
(14.6)

where a crucial assumption is that \( f(y|x,z) = f(y|x) \), that is, \( y \) depends on \( z \) only through its effect on \( x \). To simulate a set of \( M \) values we simulate \( \{z_j^* : j = 1, \ldots, M\} \) from \( f(z) \), simulate \( \{x_j^* : j = 1, \ldots, M\} \) from \( f(x|z_j^*) \), and then simulate \( \{y_j^* : j = 1, \ldots, M\} \) from \( f(y|x_j^*) \).

This principle of simulation perhaps comes into play directly most often in simulating values from a given model, but it can also be useful in situations for which it is possible to derive in closed form marginal and
conditional posteriors \( p(\lambda|y) \) and \( p(\theta|\lambda, y) \). In particular, for scalar \( \theta \) and \( \lambda \), the above prescription can be used to simulate values from \( p(\theta|y) \) by simulating \( \lambda^* \) from \( p(\lambda|y) \) and then \( \theta^* \) from \( p(\theta|\lambda^*, y) \). Here, \( p(\lambda|y) \) plays the role of \( f(x) \) in (14.5) while \( p(\theta|\lambda, y) \) plays the role of \( f(y|x) \).

4. Simulation of Joint Distributions Accomplishes Simulation of Marginal Distributions.

Consider a joint distribution \( F(x, y, z) \) for random variables \( X, Y, \) and \( Z \). Suppose we are able to simulate the values \( \{(x_j^*, y_j^*, z_j^*): j = 1, \ldots, M\} \) from this joint distribution. Then, an estimator of \( E\{q(X)\} \), for example, would be the same as given in expression (14.3) using only the values \( \{x_j^*: j = 1, \ldots, M\} \) and ignoring the \( y_j^* \) and \( z_j^* \) values. Estimators for the expected values of functions of \( Y \) and \( Z \) would be formed in a similar manner. In fact, the empirical distribution of the values \( \{x_j^*: j = 1, \ldots, M\} \) would approximate the true marginal distribution \( F_X(x) \) say, and the same would be true for the empirical distributions of the values \( \{y_j^*: j = 1, \ldots, M\} \) and \( \{z_j^*: j = 1, \ldots, M\} \) as estimators of the marginals \( F_Y(y) \) and \( F_Z(z) \). Thus, if we are able to simulate from a joint distribution we have also accomplished simulation from any of the marginals.

This principle of simulation comes into play in that if, for a model with multiple parameter elements, say \( \theta \equiv (\theta_1, \ldots, \theta_p)^T \), simulation from \( p(\theta|y) \) also provides simulated values from \( p(\theta_j|y) \) for \( j = 1, \ldots, p \). Similarly, if we are able to simulate from the joint posterior \( p(\theta, \lambda|y) \) from a hierarchical model, then we have also simulated from the marginal posteriors \( p(\theta|y) \) and \( p(\lambda|y) \) and, by what is immediately above, also the marginals of any of the elements of \( \theta \) and \( \lambda \).
14.2 Basic Methods of Simulation

14.2.1 Inversion

14.2.2 Composition

14.2.3 Basic Rejection Sampling

14.2.4 Ratio of Uniforms

14.2.5 Adaptive Rejection Sampling

14.3 The Method of Successive Substitution

14.4 The Gibbs Sampler

14.5 Metropolis Hastings