11 Multiple Regression: Bayesian Inference

We now consider Bayesian estimation and prediction for the multiple linear regression model in which the x variables are fixed constants as in Chapters 7–9. The Bayesian statistical paradigm is conceptually simple and general because inferences involve only probability calculations as opposed to maximization of a function like the log likelihood. On the other hand, the probability calculations usually entail complicated or even intractable integrals. The Bayesian approach has become popular more recently because of the development of computer-intensive approximations to these integrals (Evans and Swartz 2000) and user-friendly programs to carry out the computations (Gilks et al. 1998). We discuss both analytical and computer-intensive approaches to the Bayesian multiple regression model.

Throughout Chapters 7 and 8 we assumed that the parameters $\boldsymbol{\beta}$ and σ^2 were unknown fixed constants. We couldn't really do otherwise because to this point (at least implicitly) we have only allowed probability distributions to represent variability due to such things as random sampling or imprecision of measurement instruments. The Bayesian approach additionally allows probability distributions to represent conjectural uncertainty. Thus $\boldsymbol{\beta}$ and σ^2 can be treated as if they are random variables because we are uncertain about their values. The technical property that allows one to treat parameters as random variables is *exchangeability* of the observational units in the study (Lindley and Smith 1972).

11.1 ELEMENTS OF BAYESIAN STATISTICAL INFERENCE

In Bayesian statistics, uncertainty about the value of a parameter is expressed using the tools of probability theory (e.g., a density function—see Section 3.2). Density functions of parameters like $\boldsymbol{\beta}$ and σ^2 reflect the current credibility of possible values for these parameters. The goal of the Bayesian approach is to use data to update the uncertainty distributions for parameters, and then draw sensible conclusions using these updated distributions.

Linear Models in Statistics, *Second Edition*, by Alvin C. Rencher and G. Bruce Schaalje

The Bayesian approach can be used in any inference situation. However, it seems especially natural in the following type of problem. Consider an industrial process in which it is desired to estimate β_0 and β_1 for the straight-line relationship in (6.1) between a response y and a predictor x for a particular batch of product. Suppose that it is known from experience that β_0 and β_1 vary randomly from batch to batch. Bayesian inference allows historical (or prior) knowledge of the distributions of β_0 and β_1 among batches to be expressed in probabilistic form, and then to be combined with (x, y) data from a specific batch in order to give improved estimates of β_0 and β_1 for that specific batch.

Bayesian inference is based on two general equations. In these equations as presented below, $\boldsymbol{\theta}$ is a vector of m continuous parameters, $\mathbf{y}$ is a vector of n continuous observations, and f, g, h, k, p, q, r and t are probability density functions.

We begin with the definition of the conditional density of $\boldsymbol{\theta}$ given $\mathbf{y}$ [see (3.18)]

$$g(\boldsymbol{\theta}\,|\,\mathbf{y}) = \frac{k(\mathbf{y}, \boldsymbol{\theta})}{h(\mathbf{y})}, \tag{11.1}$$

where $k(\mathbf{y}, \boldsymbol{\theta})$ is the joint density of $y_1, y_2, \ldots, y_n$ and $\theta_1, \theta_2, \ldots, \theta_m$. Using the definition of the conditional density $f(\mathbf{y}\,|\,\boldsymbol{\theta})$, we can write $k(\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{y}\,|\boldsymbol{\theta})p(\boldsymbol{\theta})$, and (11.1) becomes

$$g(\boldsymbol{\theta}\,|\mathbf{y}) = \frac{f(\mathbf{y}|\,\boldsymbol{\theta})p(\boldsymbol{\theta})}{h(\mathbf{y})}, \tag{11.2}$$

an expression that is commonly referred to as *Bayes' theorem*. By an extension of (3.13), the marginal density $h(\mathbf{y})$ can be obtained by integrating $\boldsymbol{\theta}$ out of $k(\mathbf{y},\boldsymbol{\theta}) = f(\mathbf{y}\,|\boldsymbol{\theta})p(\boldsymbol{\theta})$ so that (11.2) becomes

$$\begin{aligned} g(\boldsymbol{\theta}|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= cf(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \end{aligned} \tag{11.3}$$

where $d\boldsymbol{\theta} = d\theta_1 \cdots d\theta_m$. In this expression, $p(\boldsymbol{\theta})$ is known as the *prior density* of $\boldsymbol{\theta}$, and $g(\boldsymbol{\theta}\,|\mathbf{y})$ is called the *posterior density* of $\boldsymbol{\theta}$. The definite integral in the denominator of (11.3) is often replaced by a constant (c) because after integration, it no longer involves the random vector $\boldsymbol{\theta}$. This definite integral is often very complicated, but can sometimes be obtained by noting that c is a *normalizing constant*, that is, a value such that the posterior density integrates to 1. Rearranging this expression and reinterpreting the joint density function $f(\mathbf{y}\,|\boldsymbol{\theta})$ of the data as the likelihood function $L(\boldsymbol{\theta}\,|\mathbf{y})$ (see Section 7.6.2), we obtain

$$g(\boldsymbol{\theta}\,|\mathbf{y}) = cp\,(\boldsymbol{\theta})L(\boldsymbol{\theta}\,|\mathbf{y}). \tag{11.4}$$

Thus (11.2), the first general equation of Bayesian inference, merely states that the posterior density of $\boldsymbol{\theta}$ given the data (representing the updated uncertainty in $\boldsymbol{\theta}$) is proportional to the prior density of $\boldsymbol{\theta}$ times the likelihood function. Point and interval estimates of the parameters are taken as mathematical features of this joint posterior density or associated marginal posterior densities of individual parameters θ_i. For example, the mode or mean of the marginal posterior density of a parameter may be used as a point estimate of the parameter. A central or highest density interval (Gelman et al. 2004, pp. 38–39) over which the marginal posterior density of a parameter integrates to $1 - \omega$ may be taken as a $100(1 - \omega)\%$ interval estimate of the parameter.

For the second general equation of Bayesian inference, we consider a future observation y_0. In the Bayesian approach, y_0 is not independent of $\mathbf{y}$ as was assumed in Section 8.6.5 because its density depends on $\boldsymbol{\theta}$, a random vector whose current uncertainty depends on $\mathbf{y}$. Since y_0, $\mathbf{y}$ and $\boldsymbol{\theta}$ are jointly distributed, the posterior predictive density of y_0 given $\mathbf{y}$ is obtained by integrating $\boldsymbol{\theta}$ out of the joint conditional density of y_0 and $\boldsymbol{\theta}$ given $\mathbf{y}$:

$$\begin{aligned} r(y_0|\mathbf{y}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} t(y_0, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q(y_0|\boldsymbol{\theta}, \mathbf{y}) g(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \quad \text{[by (4.28)]} \end{aligned}$$

where $q(y_0|\boldsymbol{\theta}, \mathbf{y})$ is the conditional density function of the sampling distribution for a future observation y_0. Since y_0 is dependent on $\mathbf{y}$ only through $\boldsymbol{\theta}$, $q(y_0|\boldsymbol{\theta}, \mathbf{y})$ simplifies, and we have

$$r(y_0|\mathbf{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q(y_0|\boldsymbol{\theta}) g(\boldsymbol{\theta}|\mathbf{y}) \, d\boldsymbol{\theta}. \tag{11.5}$$

Equation (11.5) expresses the intuitive idea that uncertainty associated with the predicted value of a future observation has two components: sampling variability and uncertainty in the parameters. As before, point and interval predictions can be taken as mathematical features (such as the mean, mode, or specified integral) of this posterior predictive density.

11.2 A BAYESIAN MULTIPLE LINEAR REGRESSION MODEL

Bayesian multiple regression models are similar to the classical multiple regression model (see Section 7.6.1) except that they include specifications of the prior distributions for the parameters. Prior specification is an important part of the art and practice of Bayesian modeling, but since the focus of this text is the basic theory of linear models, we discuss only one set of prior specifications—one that is chosen for its mathematical convenience rather than actual prior information.

11.2.1 A Bayesian Multiple Regression Model with a Conjugate Prior

Although not necessary, it is often convenient to parameterize Bayesian models using *precision* (τ) rather than variance (σ^2), where

$$\tau = \frac{1}{\sigma^2}.$$

Using this parameterization, as an example of a Bayesian linear regression model, let

$$\mathbf{y}|\boldsymbol{\beta}, \tau \text{ be } N_n\left(X\boldsymbol{\beta}, \frac{1}{\tau}\mathbf{I}\right),$$

$$\boldsymbol{\beta}|\tau \text{ be } N_{k+1}\left(\boldsymbol{\phi}, \frac{1}{\tau}\mathbf{V}\right),$$

$$\tau \text{ be gamma}(\alpha, \delta).$$

The second and third distributions here are prior distributions, and we assume that $\boldsymbol{\phi}$, $\mathbf{V}$, α, and δ (the parameters of the prior distributions), are known. Although we will not do so here, this model could be extended by specifying *hyperprior* distributions for $\boldsymbol{\phi}$, $\mathbf{V}$, α, and δ (Lindley and Smith 1972).

As in previous chapters, the number of predictor variables is denoted by k (so that the rank of $\mathbf{X}$ is $k+1$) and the number of observations by n. The prior density function for $\boldsymbol{\beta}|\tau$ is, using (4.9)

$$p_1(\boldsymbol{\beta}|\tau) = \frac{1}{(2\pi)^{(k+1)/2}|\tau^{-1}\mathbf{V}|^{\frac{1}{2}}} e^{-\tau(\boldsymbol{\beta}-\boldsymbol{\phi})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi})/2}. \tag{11.6}$$

The prior density function for τ is the gamma density (Gelman et al. 2004, pp. 574–575)

$$p_2(\tau) = \frac{\delta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\delta\tau}, \tag{11.7}$$

where $\alpha > 0, \delta > 0$, and by definition

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \tag{11.8}$$

(see any advanced calculus text). For the gamma density in (11.7),

$$E(\tau) = \frac{\alpha}{\delta} \quad \text{and} \quad \text{var}(\tau) = \frac{\alpha}{\delta^2}.$$

These prior distributions could be formulated with small enough variances that the prior knowledge strongly influences posterior distributions of the parameters in the model. If so, they are called *informative priors*. On the other hand, both of these priors could be formulated with large variances so that they have very little effect on the posterior distributions. If so, they are called *diffuse priors*. The priors would be diffuse if, for example, $\mathbf{V}$ in (11.6) were a diagonal matrix with very large diagonal elements, and if δ in (11.7) were very close to zero.

The prior specifications in (11.6) and (11.7) are flexible and reasonable, and they also have nice mathematical properties, as will be shown in Theorem 11.2a. Other specifications for the prior distributions could be used. However, even the minor modification of proposing a prior distribution for $\boldsymbol{\beta}$ that is not conditional on τ makes the model far less mathematically tractable.

The joint prior for $\boldsymbol{\beta}$ and τ in our model is called a *conjugate prior* because its use results in a posterior distribution of the same form as the prior. We prove this in the following theorem.

Theorem 11.2a. Consider the Bayesian multiple regression model in which $\mathbf{y}|\boldsymbol{\beta}, \tau$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}|\tau$ is $N_{k+1}(\boldsymbol{\phi}, \tau^{-1}\mathbf{V})$, and τ is gamma(α, δ). The joint prior distribution is conjugate, that is, $g(\boldsymbol{\beta}, \tau|\mathbf{y})$ is of the same form as $p(\boldsymbol{\beta}, \tau)$.

Proof. Combining (11.6) and (11.7), the joint prior density is

$$\begin{aligned} p(\boldsymbol{\beta}, \tau) &= p_1(\boldsymbol{\beta}|\tau)p_2(\tau) \\ &= c_1 \tau^{(k+1/)2} e^{-\tau(\boldsymbol{\beta}-\boldsymbol{\phi})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi})/2} \tau^{\alpha-1} e^{-\delta\tau} \\ &= c_1 \tau^{(\alpha_*+k+1)/2} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi})+\delta_*]/2}, \end{aligned} \tag{11.9}$$

where $\alpha_* = 2\alpha - 2$, $\delta_* = 2\delta$ and all the factors not involving random variables are collected into the normalizing constant c_1. Using (11.4), the joint posterior density is then

$$\begin{aligned} g(\boldsymbol{\beta}, \ \tau|\mathbf{y}) &= cp(\boldsymbol{\beta}, \tau)L(\boldsymbol{\beta}, \tau\,|\,\mathbf{y}) \\ &= c_2 \tau^{(\alpha_*+k+1)/2} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi})+\delta_*]/2} \tau^{n/2} e^{-\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2} \\ &= c_2 \tau^{(\alpha_{**}+k+1)/2} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi})+(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\delta_*]/2}, \end{aligned}$$

where $\alpha_{**} = 2\alpha - 2 + n$, and all the factors not involving random variables are collected into the normalizing constant c_2. By expanding and completing the square in the exponent (Problem 11.1), we obtain

$$g(\boldsymbol{\beta}, \tau \,|\, \mathbf{y}) = c_2 \tau^{(\alpha_{**}+k+1)/2} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+\delta_{**}]/2}, \tag{11.10}$$

where $\mathbf{V}_* = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\phi}_* = \mathbf{V}_*(\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{X}'\mathbf{y})$, and $\delta_{**} = -\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_* + \boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{y}'\mathbf{y} + \delta_*$. Hence the joint posterior density has exactly the same form as the joint prior density in (11.9). □

It might seem odd to include terms like $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ in the "*constants*" of a probability distribution, while considering parameters like $\boldsymbol{\beta}$ and τ to be random, but this is completely characteristic of Bayesian inference. In this sense, inference in a Bayesian linear model is opposite to inference in the classical linear model.

11.2.2 Marginal Posterior Density of $\boldsymbol{\beta}$

In order to carry out inferences for $\boldsymbol{\beta}$, the marginal posterior density of $\boldsymbol{\beta}$ [see (3.13)] must be obtained by integrating τ out of the joint posterior density in (11.10). The following theorem gives the form of this marginal distribution.

Theorem 11.2b. Consider the Bayesian multiple regression model in which $\mathbf{y}|\boldsymbol{\beta}, \tau$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}|\tau$ is $N_{k+1}(\boldsymbol{\phi}, \tau^{-1}\mathbf{V})$, and τ is gamma(α, δ). The marginal posterior distribution $u(\boldsymbol{\beta}|\mathbf{y})$ is a multivariate t distribution with parameters $(n + 2\alpha, \boldsymbol{\phi}_*, \mathbf{W}_*)$, where

$$\boldsymbol{\phi}_* = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{X}'\mathbf{y}) \tag{11.11}$$

and

$$\mathbf{W}_* = \left[\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\phi})'(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\phi}) + 2\delta}{n + 2\alpha}\right](\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}. \tag{11.12}$$

Proof. The marginal distribution of $\boldsymbol{\beta}|\mathbf{y}$ is obtained by integration as

$$u(\boldsymbol{\beta}|\mathbf{y}) = \int_0^\infty g(\boldsymbol{\beta}, \tau|\mathbf{y})d\tau.$$

By (11.10), this becomes

$$u(\boldsymbol{\beta}|\mathbf{y}) = c_2 \int_0^\infty \tau^{(\alpha_{**}+k+1)/2} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+\delta_{**}]/2} d\tau.$$

Using (11.8) together with integration by substitution, the integral in this expression can be solved (Problem 11.2) to give the posterior distribution of $\boldsymbol{\beta}|\mathbf{y}$ as

$$\begin{aligned} u(\boldsymbol{\beta}|\mathbf{y}) &= c_2\Gamma\left(\frac{\alpha_{**}+2+k+1}{2}\right)\left[\frac{(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+\delta_{**}}{2}\right]^{-(\alpha_{**}+2+k+1)/2} \\ &= c_3[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_*+\boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi}+\mathbf{y}'\mathbf{y}+\delta_*]^{-(\alpha_{**}+2+k+1)/2}. \end{aligned}$$

To show that this is the multivariate t density, several algebraic steps are required as outlined in Problems 11.3a–c and 11.4. See also Seber and Lee (2003, pp. 100–110). After these steps, the preceding expression becomes

$$\begin{aligned} u(\boldsymbol{\beta}|\mathbf{y}) = c_3[&(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+(\mathbf{y}-\mathbf{X}\boldsymbol{\phi})'(\mathbf{I}+\mathbf{XVX}')^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\phi}) \\ &+2\delta]^{-(\alpha_{**}+2+k+1)/2}. \end{aligned}$$

Dividing the expression in square brackets by $(\mathbf{y}-\mathbf{X}\boldsymbol{\phi})'(\mathbf{I}+\mathbf{XVX}')^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\phi})+2\delta$, modifying the normalizing constant accordingly, and replacing α_{**} by $2\alpha-2+n$, we obtain

$$\begin{aligned} u(\boldsymbol{\beta}|\mathbf{y}) &= c_4\left[1+\frac{(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)/(n+2\alpha)}{[(\mathbf{y}-\mathbf{X}\boldsymbol{\phi})'(\mathbf{I}+\mathbf{XVX}')^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\phi})+2\delta]/(n+2\alpha)}\right]^{-(n+2\alpha+k+1)/2} \\ &= c_4\left(\frac{1+(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{W}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)}{n+2\alpha}\right)^{-(n+2\alpha+k+1)/2}, \end{aligned} \tag{11.13}$$

where $\mathbf{W}_*$ is as given in (11.12). The expression in (11.13) can now be recognized as the density function of the multivariate t distribution (Gelman et al. 2004, pp. 576–577; Rencher 1998, p. 56) with parameters $(n+2\alpha, \boldsymbol{\phi}_*, \mathbf{W}_*)$. Note that $\boldsymbol{\phi}_*$ is the mean vector and $[(n+2\alpha)/(n+2\alpha-2)]\mathbf{W}_*$ is the covariance matrix of $\boldsymbol{\beta}|\mathbf{y}$. □

As a historical note, the reasoning in this section is closely related to the work of W. S. Gosset or "Student" (Pearson et al. 1990, pp. 49–53, 72–73) on the small-sample distribution of

$$t = \frac{\bar{y}}{s}.$$

Gosset used Bayesian reasoning ("inverse probability") with a uniform prior distribution ("equal distribution of ignorance") to show through a combination of proof, conjecture, and simulation that the posterior density of t is related to what we now call Student's t distribution with $n - 1$ degrees of freedom.

11.2.3 Marginal Posterior Densities of τ and σ^2

Inferences regarding τ and σ^2 require knowledge of the marginal posterior distribution of $\tau|\mathbf{y}$. We derive the posterior density of $\tau|\mathbf{y}$ in the following theorem.

Theorem 11.2c. Consider the Bayesian multiple regression model in which $\mathbf{y}|\boldsymbol{\beta},\tau$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}|\tau$ is $N_{k+1}(\boldsymbol{\phi}, \tau^{-1}\mathbf{V})$, and τ is gamma(α, δ). The marginal posterior distribution $v(\tau|\mathbf{y})$ is a gamma distribution with parameters $\alpha + n/2$ and $(-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_* + \boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{y}'\mathbf{y} + 2\delta)/2$, where $\mathbf{V}_* = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$ and $\boldsymbol{\phi}_* = \mathbf{V}_*(\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{X}'\mathbf{y})$.

Proof. The marginal distribution of $\tau|\mathbf{y}$ is obtained by integration as

$$\begin{aligned}
\mathbf{v}(\tau|\mathbf{y}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{\beta},\, \tau|\mathbf{y})d\boldsymbol{\beta} \\
&= c_2 \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \tau^{(\alpha_{**}+k+1)/2} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*-1(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+\delta_{**}]/2} d\boldsymbol{\beta} \\
&= c_2 \tau^{(\alpha_{**}+k+1)/2} e^{-\tau\delta_{**}/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)]/2} d\boldsymbol{\beta}
\end{aligned}$$

where all the factors not involving random variables are collected into the normalizing constant c_2 as in (11.10). Since the integral in the preceding expression is proportional to the integral of a joint multivariate normal density, we obtain

$$\begin{aligned}
\mathbf{v}(\tau|\mathbf{y}) &= c_2 \tau^{(\alpha_{**}+k+1)/2} e^{-(\delta_{**}/2)\tau} (2\pi)^{(k+1)/2} |\mathbf{V}_*|^{1/2} \tau^{-(k+1)/2} \\
&= c_5 \tau^{(\alpha_{**}+k+1)/2-(k+1)/2} e^{-(\delta_{**}/2)\tau} \\
&= c_5 \tau^{(\alpha+n)/2-1} e^{-[(-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_*+\boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi}+\mathbf{y}'\mathbf{y}+2\delta)/2]\tau}, \qquad (11.14)
\end{aligned}$$

which is the density function of the specified gamma distribution. □

The marginal posterior density of σ^2 can now be obtained by the univariate change-of-variable technique (4.2) as

$$\mathbf{w}(\sigma^2|\mathbf{y}) = c_6(\sigma^2)^{-(\alpha+n)/2-1} e^{-[(-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_*+\boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi}+\mathbf{y}'\mathbf{y}+2\delta)/2]/\sigma^2} \qquad (11.15)$$

which is the density function of the inverse gamma distribution with parameters $\alpha + n/2$ and $(-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_* + \boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{y}'\mathbf{y} + 2\delta)/2$ (Gelman et al. 2004, pp. 574–575).

11.3 INFERENCE IN BAYESIAN MULTIPLE LINEAR REGRESSION

11.3.1 Bayesian Point and Interval Estimates of Regression Coefficients

A sensible Bayesian point estimator of $\boldsymbol{\beta}$ is the mean of the marginal posterior density in (11.13)

$$\boldsymbol{\phi}_* = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{X}'\mathbf{y}), \tag{11.16}$$

and a sensible $100(1 - \omega)\%$ Bayesian confidence region for $\boldsymbol{\beta}$ is the highest-density region $\boldsymbol{\Omega}$ such that

$$c_4 \int_{\Omega} \cdots \int \left[\frac{1 + (\boldsymbol{\beta} - \boldsymbol{\phi}_*)'\mathbf{W}_*^{-1}(\boldsymbol{\beta} - \boldsymbol{\phi}_*)}{n + 2\alpha}\right]^{-(n+2\alpha+k+1)/2} d\boldsymbol{\beta} = 1 - \omega. \tag{11.17}$$

A convenient property of the multivariate t distribution is that linear functions of the random vector follow the (univariate) t distribution. Thus, given $\mathbf{y}$,

$$\frac{\mathbf{a}'\boldsymbol{\beta} - \mathbf{a}'\boldsymbol{\phi}_*}{\mathbf{a}'\mathbf{W}_*\mathbf{a}} \quad \text{is} \quad t(\mathrm{n} + 2\alpha)$$

and, as an important special case,

$$\frac{\beta_i - \phi_{*i}}{w_{*ii}} \quad \text{is} \quad \mathrm{t}(n + 2\alpha), \tag{11.18}$$

where ϕ_{*i} is the ith element of $\boldsymbol{\phi}_*$ and w_{*ii} is the ith diagonal element of $\mathbf{W}_*$. Thus a Bayesian point estimate of β_i is ϕ_{*i} and a $100(1 - \omega)\%$ Bayesian confidence interval for β_i is

$$\phi_{*i} \pm t_{\omega/2,n+2\alpha} w_{*ii}. \tag{11.19}$$

One very appealing aspect of Bayesian inference is that intervals like (11.19) have a natural interpretation. Instead of the careful classical interpretation of a confidence interval in terms of hypothetical repeated sampling, one can simply and correctly say that the probability is $1 - \omega$ that β_i is in (11.19).

An interesting final note on Bayesian estimation of $\boldsymbol{\beta}$ is that the Bayesian estimator $\boldsymbol{\phi}_*$ in (11.16) can be obtained as the generalized least-squares estimator of $\boldsymbol{\beta}$ in (7.63). To see this, consider adding the prior information to the data as if it constituted a set of additional observations. The idea is to augment $\mathbf{y}$ with $\boldsymbol{\phi}$, and to consider the mean vector and covariance matrix of the augmented vector $\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\phi} \end{pmatrix}$ to be, respectively

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{I}_{k+1} \end{pmatrix}\boldsymbol{\beta} \quad \text{and} \quad \frac{1}{\tau}\begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{V} \end{pmatrix}.$$

Generalized least squares estimation expressed in terms of these partitioned matrices then gives $\boldsymbol{\phi}_*$ in (11.16) as an estimate of $\boldsymbol{\beta}$ (Problem 11.6). The implication of this is that prior information on the regression coefficients can be incorporated into a multiple linear regression model by the intuitive informal process of "adding" observations.

11.3.2 Hypothesis Tests for Regression Coefficients in Bayesian Inference

Classical hypothesis testing is not a natural part of Bayesian inference (Gelman et al. 2004, p. 162). Nonetheless, if the question addressed by a classical hypothesis test is whether the data support the conclusion (i.e., alternative hypothesis) that β_i is greater than β_{i0}, a sensible approach is to use the posterior distribution (in this case the t distribution with $n + 2\alpha$ degrees of freedom) to compute the probability

$$P\left(t(n + 2\alpha) > \frac{\beta_{i0} - \phi_{*i}}{w_{*ii}}\right).$$

The larger this probability is, the more credible is the hypothesis that $\beta_i > \beta_{i0}$.

If, alternatively, classical hypothesis testing is used to select a model from a set of candidate models, the corresponding Bayesian approach is to compute an information statistic for each model in question. For example, Schwarz (1978) proposed the Bayesian Information Criterion (BIC) for multiple linear regression models, and Spiegelhalter et al. (2002) proposed the Deviance Information Criterion (DIC) for more general Bayesian models. The model with the lowest value of the information criterion is selected. Model selection in Bayesian analysis is an area of current research.

11.3.3 Special Cases of Inference in Bayesian Multiple Regression Models

Two special cases of inference in this Bayesian linear model are of particular interest. First, consider the use of a diffuse prior. Let $\boldsymbol{\phi} = \mathbf{0}$, let $\mathbf{V}$ be a diagonal matrix with all diagonal elements equal to a large constant (say, 10^6), and let α and δ both be equal to a small constant (say, 10^{-6}). In this case, $\mathbf{V}^{-1}$ is close to $\mathbf{O}$, and so $\boldsymbol{\phi}_*$, the Bayesian point estimate of $\boldsymbol{\beta}$ in (11.16), is approximately equal to

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

the classical least-squares estimate. Also, since $(\mathbf{I} + \mathbf{XVX}')^{-1} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1}\mathbf{X}'$ (see Problem 11.3a), the covariance matrix $\mathbf{W}_*$ approaches

$$\begin{aligned}\mathbf{W}_* &= \frac{\mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}}{n}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{n-1}{n}s^2(\mathbf{X}'\mathbf{X})^{-1} \qquad \text{[by (7.26)].}\end{aligned}$$

Thus, in the case of diffuse priors, the Bayesian confidence region (11.17) reduces to a region similar to (8.46), and Bayesian confidence intervals for the regression coefficients in (11.19) are similar to classical confidence intervals in (8.47); the only differences are the multiplicative factor $(n-1)/n$ and the use of the t distribution with n degrees of freedom rather than $n-k-1$ degrees of freedom. If a Bayesian multiple linear regression model with independent uniformly distributed priors for $\boldsymbol{\beta}$ and $\ln(\tau^{-1})$ is considered, Bayesian confidence intervals for the regression coefficients are exactly equal to classical confidence intervals (Problem 11.5). One result of this is that simple Bayesian interpretations can be validly applied to confidence intervals for the classical linear model. In fact, most inferences for the classical linear model can be stated in terms of properties of posterior distributions.

The second special case of inference in this Bayesian linear model is the case in which $\boldsymbol{\phi} = \mathbf{0}$ and $\mathbf{V}$ is a diagonal matrix with a constant on the diagonal. Thus $\mathbf{V} = a\mathbf{I}$, where a is a positive number, and the Bayesian estimator of $\boldsymbol{\beta}$ in (11.16) becomes

$$\left(\mathbf{X}'\mathbf{X} + \frac{1}{a}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

For the centered model (Section 7.5) this estimator is also known as the "ridge estimator" (Hoerl and Kennard 1970). It was originally proposed as a method for dealing with collinearity, the situation in which the columns of the $\mathbf{X}$ matrix have near-linear dependence so that $\mathbf{X}'\mathbf{X}$ is nearly singular. However, the estimator may also be understood as a "shrinkage estimator" in which prior information causes the estimates of the coefficients to be shrunken toward zero (Seber and Lee 2003, pp. 321–322). The use of a Bayesian linear model with *hyperpriors* (prior distributions for the parameters of the prior distributions) leads to a reasonable choice of value for a in terms of variances of the prior and hyperprior distributions (Lindley and Smith 1972).

11.3.4 Bayesian Point and Interval Estimation of σ^2

A possible Bayesian point estimator of σ^2 is the mean of the marginal inverse gamma density in (11.15)

$$\frac{(-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_* + \boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{y}'\mathbf{y} + 2\delta)/2}{\alpha + n/2 - 1}$$

and a $100(1-\omega)\%$ Bayesian confidence interval for σ^2 is given by the $1-\omega/2$ and $\omega/2$ quantiles of the appropriate inverse gamma distribution.

As a special case, note that if α and δ are both close to 0, $\boldsymbol{\phi} = \mathbf{0}$, and $\mathbf{V}$ is a diagonal matrix with all diagonal elements equal to a large constant so that $\mathbf{V}^{-1}$ is close to $\mathbf{O}$, then the Bayesian point estimator of σ^2 is approximately

$$\begin{aligned}\frac{(\mathbf{y}'\mathbf{y}-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_*)/2}{n/2-1} &= \frac{\mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{n-2} \\ &= \frac{\mathbf{y}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}}{n-2} \\ &= \frac{n-k-1}{n-2}s^2,\end{aligned}$$

and the centered Bayesian confidence limits are the $1-\omega/2$ quantile and the $\omega/2$ quantile of the inverse gamma distribution with parameters $n/2$ and $\mathbf{y}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}/2$.

11.4 BAYESIAN INFERENCE THROUGH MARKOV CHAIN MONTE CARLO SIMULATION

The inability to derive a closed-form marginal posterior distribution for a parameter is extremely common in Bayesian inference (Gilks et al. 1998, p. 3). For example, if the Bayesian multiple regression model of Section 11.2.1 had involved a prior distribution for $\boldsymbol{\beta}$ that was *not* conditional on τ, closed-form marginal distributions for the parameters could not have been derived (Lindley and Smith 1972). In actual practice, the exception in Bayesian inference is to be able to derive closed-form marginal posterior distributions. However, this difficulty turns out to be only a minor hindrance when modern computing resources are available.

If it were possible, an ideal solution would be to draw a large number of samples from the joint posterior distribution. Then marginal means, marginal highest density intervals, and other properties of the posterior distribution could be approximated using sample statistics. Furthermore, functions of the sampled values could be calculated in order to approximate marginal posterior distributions of these functions. The big question, of course, is how it would be possible to draw samples from a distribution for which a familiar closed-form joint density function is not available.

A general approach for accomplishing this is referred to as *Markov Chain Monte Carlo* (MCMC) simulation (Gilks et al. 1998). A *Markov Chain* is a special sequence of random variables (Ross 2006, p. 185). Probability laws for general sequences of random variables are specified in terms of the conditional distribution of the current value in the sequence, given all past values. A Markov Chain is a simple sequence in which the conditional distribution of the current value is completely specified, given only the most recent value.

Markov Chain Monte Carlo simulation in Bayesian inference is based on sequences of alternating random draws from conditional posterior distributions of each of the parameters in the model given the most recent values of the other parameters. This process generates a Markov Chain for each parameter. Moreover, the unconditional distribution for each parameter converges to the marginal posterior distribution of the parameter, and the unconditional joint distribution of the vector of parameters for any complete iteration of MCMC converges to the joint posterior distribution. Thus after discarding a number of initial draws (the "burn-in"), draws may be considered to constitute sequences of samples from marginal posterior distributions of the parameters. The samples are not independent, but the nonindependence can be ignored if the number of draws is sufficiently large. Plots of sample values can be examined to determine whether a sufficiently large number of draws has been obtained (Gilks et al. 1998).

When the prior distributions are conjugate, closed-form density functions of the conditional posterior distributions of the parameters are available regardless of whether closed-form marginal posterior distributions can be derived. In the case of conjugate priors, a simple form of MCMC called "Gibbs sampling" (Gilks et al. 1998, Casella and George 1992) can be used by which draws are made successively from each of the conditional distributions of the parameters, given the current draws for the other parameters.

We now illustrate this procedure. Consider again the Bayesian multiple regression model in which $\mathbf{y}|\boldsymbol{\beta}, \tau$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}|\tau$ is $N_{k+1}(\boldsymbol{\phi}, \tau^{-1}\mathbf{V})$, and τ is gamma(α, δ). The joint posterior density function is given in (11.10). The conditional posterior density (or "full conditional") of $\boldsymbol{\beta}|\tau, \mathbf{y}$ can be obtained by picking the terms out of (11.10) that involve $\boldsymbol{\beta}$, and considering everything else to be part of the normalizing constant. Thus, the conditional density of $\boldsymbol{\beta}|\tau, \mathbf{y}$ is

$$\varphi(\boldsymbol{\beta}|\tau, \mathbf{y}) = c_6 e^{-\tau(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)/2}.$$

Clearly $\boldsymbol{\beta}|\tau, \mathbf{y}$ is $N_{k+1}(\boldsymbol{\phi}_*, \tau^{-1}\mathbf{V}_*)$. Similarly, the conditional posterior density for $\tau|\boldsymbol{\beta}, \mathbf{y}$ is

$$\psi(\tau|\boldsymbol{\beta}, \mathbf{y}) = c_7 \tau^{[(\alpha_{**}+k+3)/2]-1} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+\delta_{**}]/2}$$

so that $\tau|\boldsymbol{\beta}, \mathbf{y}$ can be seen to be gamma $\{(\alpha_{**} + k + 3)/2, [(\boldsymbol{\beta} - \boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta} - \boldsymbol{\phi}_*) + \delta_{**}]/2\}$.

Gibbs sampling for this model proceeds as follows:

- Specify a starting value τ_0 [possibly $1/s^2$ from (7.23)].
- For $i = 1$ to M: draw $\boldsymbol{\beta}_i$ from $N_{k+1}(\boldsymbol{\phi}_*, \tau_{i-1}^{-1}\mathbf{V}_*)$, draw τ_i from gamma $\{(\alpha_{**} + k + 3)/2, [(\boldsymbol{\beta}_i - \boldsymbol{\phi}_*)'V_*^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\phi}_*) + \delta_{**}]/2\}$.
- Discard the first Q draws (as burn-in), and consider the last $M - Q$ draws $(\boldsymbol{\beta}_i, \tau_i)$ to be draws from the joint posterior distribution. For this model, using the starting value of $1/s^2$, Q would usually be very small (say, 0), and M would be large (say, 10,000).

Bayesian inferences for all parameters of the model could now be carried out using sample statistics of this empirical joint posterior distribution. For example, a Bayesian point estimate of τ could be calculated as the sample mean or median of the draws of τ from the joint posterior distribution. If we calculate (or "monitor") $1/\tau$ on each iteration, a Bayesian point estimate of $\sigma^2 = 1/\tau$ could be calculated as the mean or

TABLE 11.1 Body Fat Data

y	x_1	x_2
11.9	19.5	29.1
22.8	24.7	28.2
18.7	30.7	37.0
20.1	29.8	31.1
12.9	19.1	30.9
21.7	25.6	23.7
27.1	31.4	27.6
25.4	27.9	30.6
21.3	22.1	23.2
19.3	25.5	24.8
25.4	31.1	30.0
27.2	30.4	28.3
11.7	18.7	23.0
17.8	19.7	28.6
12.8	14.6	21.3
23.9	29.5	30.1
22.6	27.7	25.7
25.4	30.2	24.6
14.8	22.7	27.1
21.1	25.2	27.5

median of $1/\tau$. A 95% Bayesian interval estimate of σ^2 could be computed as the central 95% interval of the sample distribution of σ^2. Other inferences could similarly be drawn on the basis of sample draws from the joint posterior distribution.

Example 11.4. Table 11.1 contains body fat data for a sample of 20 females aged 25–34 (Kutner et al. 2005, p. 256). The response variable was body fat (y), and two predictor variables were triceps skinfold thickness (x_1) and midarm circumference (x_2). The data were analyzed using the Bayesian multiple regression model of Section 11.2.1 with diffuse priors in which $\boldsymbol{\phi}' = (0, 0, 0)$, $\mathbf{V} = 10^6\mathbf{I}_3$, $\alpha = 0.0001$, and $\delta = 0.0001$. Density functions of the marginal posterior distributions of β_0, β_1, and β_2 from (11.13) as well as the marginal posterior density of σ^2 from (11.15) are graphed in Figure 11.1. Superimposed on these (and almost indistinguishable from them) are smooth estimates (Silverman 1999) of the same posterior densities based on Gibbs sampling with $Q = 0$ and $M = 10{,}000$. □

11.5 POSTERIOR PREDICTIVE INFERENCE

As a final aspect of Bayesian inference for the multiple regression model, we consider Bayesian prediction of the value of the response variable for a future individual. If we again use the Bayesian multiple regression model of Section 11.2.1 in which $\mathbf{y}|\boldsymbol{\beta}, \tau$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}|\tau$ is $N_{k+1}(\boldsymbol{\phi}, \tau^{-1}\mathbf{V})$, and τ is gamma(α, δ), the posterior predictive density for a future observation y_0 with predictor variables $\mathbf{x}_0$ can be

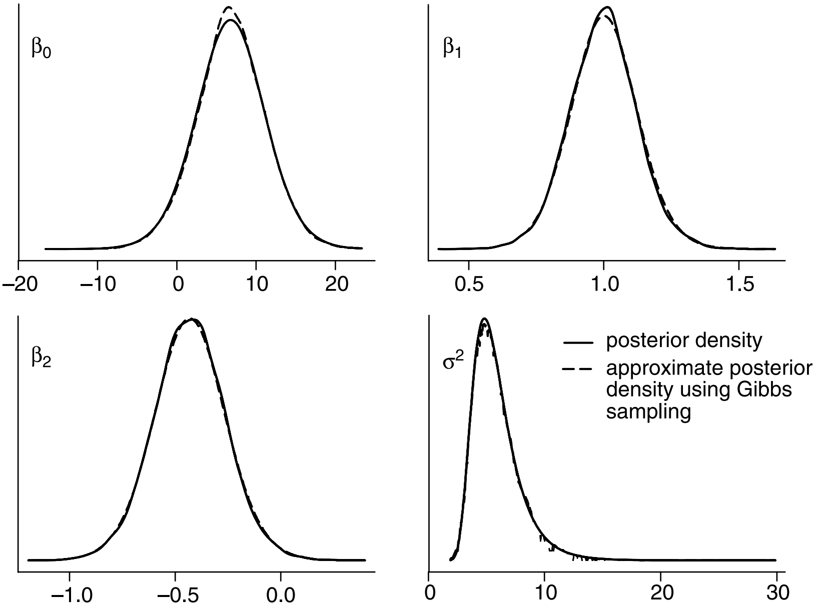

Figure 11.1 Posterior densities of parameters for the fat data in Table 11.1.

expressed using (11.5) as

$$r(y_0|\mathbf{y}) = \int_0^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} q(y_0|\boldsymbol{\beta},\tau)g(\boldsymbol{\beta},\tau|y)d\boldsymbol{\beta}\,d\tau$$

$$= c\int_0^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \tau^{1/2}e^{-\tau(y_0-\mathbf{x}_0'\boldsymbol{\beta})^2/2}\tau^{(\alpha_{**}+k+1)/2}$$

$$\times\, e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+\delta_{**}]/2}d\boldsymbol{\beta}\,d\tau$$

$$= c\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)$$

$$+\,(y_0-\mathbf{x}_0'\boldsymbol{\beta})^2+\delta_{**}]^{-(\alpha_{**}+k+4)/2}d\boldsymbol{\beta}.$$

Further analytical progress with this integral is difficult. Nonetheless, Gibbs sampling as in Section 11.4 can be easily extended to simulate the posterior predictive distribution of y_0 as follows:

- Specify a starting value τ_0 [possibly $1/s^2$ from (7.23)].
- For $i = 1$ to M: draw $\boldsymbol{\beta}_i$ from $N_{k+1}(\boldsymbol{\phi}_*, \tau_{i-1}^{-1}\mathbf{V}_*)$, draw τ_i from gamma$\{(\alpha_{**}+k+3)/2, [(\boldsymbol{\beta}_i-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\phi}_*)+\delta_{**}]/2\}$, draw y_{0i} from $N(x_0'\boldsymbol{\beta}_i, \tau_i^{-1})$.

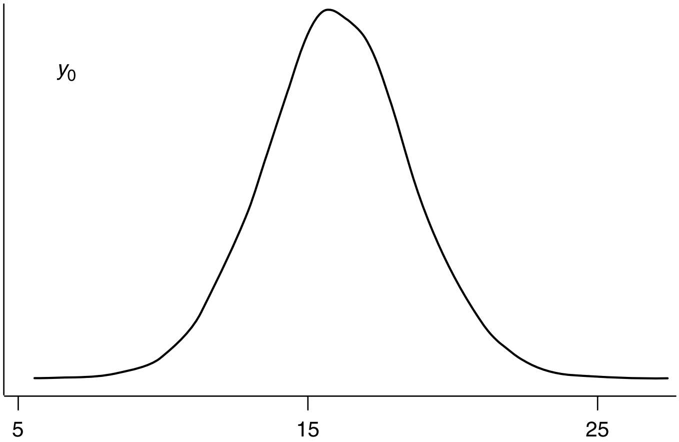

Figure 11.2 Approximate posterior predictive density using Gibbs sampling for a future observation y_0 with $\mathbf{x}_0' = (1, 20, 25)$ for the fat data in Table 11.1.

- Discard the first Q draws (as burn-in), and consider the last $M - Q$ draws of y_{0i} to be draws from the posterior predictive distribution.

Example 11.5. Example 11.4(continued). Consider a new individual with $x_1 = 20$ and $x_2 = 25$. Thus $\mathbf{x}_0' = (1, 20, 25)$. Figure 11.2 gives a smooth estimate of the posterior predictive density of y_0 based on Gibbs sampling with $Q = 0$ and $M = 10{,}000$. □

The approximate Bayesian 95% prediction interval derived from this density is (11.83, 20.15), which may be compared to the 95% prediction interval (10.46, 21.57) for the same future individual using the non-Bayesian approach (8.62).

This chapter gives a small taste of the calculations associated with the modern Bayesian multiple regression model. With very little additional work, many aspects of the model can be modified and customized, especially if the MCMC approach is used. Versatility is one of the great advantages of the Bayesian approach.

PROBLEMS

11.1 As in Theorem 11.2a, show that $(\boldsymbol{\beta} - \boldsymbol{\phi})'\mathbf{V}^{-1}(\boldsymbol{\beta} - \boldsymbol{\phi}) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \delta_* = (\boldsymbol{\beta} - \boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta} - \boldsymbol{\phi}_*) + \delta_{**}$, where $\mathbf{V}_* = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$, $\boldsymbol{\phi}_* = \mathbf{V}_*(\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{X}'\mathbf{y})$, and $\delta_{**} = -\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_* + \boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi} + \mathbf{y}'\mathbf{y} + \delta_*$.

11.2 As used in the proof to Theorem 11.2b, show that

$$\int_0^\infty t^a e^{-bt} dt = b^{-(a+1)}\Gamma(a + 1).$$

11.3 (**a**) Show that $(\mathbf{I}+\mathbf{XVX}')^{-1}=\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X}+\mathbf{V}^{-1})^{-1}\mathbf{X}'$.

(**b**) Show that $(\mathbf{I}+\mathbf{XVX}')^{-1}\mathbf{X}=\mathbf{X}(\mathbf{X}'\mathbf{X}+\mathbf{V}^{-1})^{-1}\mathbf{V}^{-1}$.

(**c**) Show that $\mathbf{V}^{-1}-\mathbf{V}^{-1}(\mathbf{X}'\mathbf{X}+\mathbf{V}^{-1})^{-1}\mathbf{V}^{-1}=\mathbf{X}'(\mathbf{I}+\mathbf{XVX}')^{-1}\mathbf{X}$.

11.4 As in the proof to Theorem 11.2b, show that $\mathbf{y}'\mathbf{y}+\boldsymbol{\phi}'\mathbf{V}^{-1}\boldsymbol{\phi}-\boldsymbol{\phi}_*'\mathbf{V}_*^{-1}\boldsymbol{\phi}_*$ $=(\mathbf{y}-\mathbf{X}\boldsymbol{\phi})'(\mathbf{I}+\mathbf{XVX}')^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\phi})$, where $\mathbf{V}_*=(\mathbf{X}'\mathbf{X}+\mathbf{V}^{-1})^{-1}$ and $\boldsymbol{\phi}_*=\mathbf{V}_*(\mathbf{X}'\mathbf{y}+\mathbf{V}^{-1}\boldsymbol{\phi})$.

11.5 Consider the Bayesian multiple linear regression model in which $\mathbf{y}|\boldsymbol{\beta},\tau$ is $N_n(\mathbf{X}\boldsymbol{\beta},\tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}$ is uniform (R^{k+1}) [i.e., uniform over $(k+1)$ -dimensional space], and $\ln(\tau^{-1})$ is uniform $(-\infty,\infty)$. Show that the marginal posterior distribution of $\boldsymbol{\beta}|\mathbf{y}$ is the multivariate t distribution with parameters $[n-k-1,\hat{\boldsymbol{\beta}},s^2(\mathbf{X}'\mathbf{X})^{-1}]$, where $\hat{\boldsymbol{\beta}}$ and s^2 are defined in the usual way [see (7.6) and (7.23)]. These prior distributions are called *improper priors* because uniform distributions must be defined for bounded sets of values. Nonetheless, the sets can be very large, and so we can proceed as if they were unbounded.

11.6 Consider the augmented data vector $\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\phi} \end{pmatrix}$ with mean vector $\begin{pmatrix} \mathbf{X} \\ \mathbf{I}_{k+1} \end{pmatrix}\boldsymbol{\beta}$ and covariance matrix

$$\begin{pmatrix} \frac{1}{\tau}\mathbf{I} & \mathbf{O} \\ \mathbf{O} & \frac{1}{\tau}\mathbf{V} \end{pmatrix}.$$

Show that the generalized least-squares estimator of $\boldsymbol{\beta}$ is the Bayesian estimator in (11.16), $(\mathbf{V}^{-1}+\mathbf{X}'\mathbf{X})^{-1}(\mathbf{V}^{-1}\boldsymbol{\phi}+\mathbf{X}'\mathbf{y})$.

11.7 Given that τ is gamma(α,δ) as in (11.7), find $E(\tau)$ and var(τ).

11.8 Use the Bayesian multiple regression model in which $\mathbf{y}|\boldsymbol{\beta},\tau$ is $N_n(\mathbf{X}\boldsymbol{\beta},\tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}|\tau$ is $N_{k+1}(\boldsymbol{\phi},\tau^{-1}\mathbf{V})$, and τ is gamma(α,δ). Derive the marginal posterior density function for $\sigma^2|\mathbf{y}$, where $\sigma^2=1/\tau$.

11.9 Consider the Bayesian simple linear regression model in which $y_i|\beta_0,\beta_1$, is $N(\beta_0+\beta_1x_i,1/\tau)$ for $i=1,\ldots,n$, $\beta_0|\tau$ is $N(a,\sigma_0^2/\tau)$, $\beta_1|\tau$ is $N(b,\sigma_1^2/\tau)$, $\mathrm{cov}(\beta_0,\beta_1|\tau)=\sigma_{12}$, and τ is gamma(α,δ).

(**a**) Find the marginal posterior density of $\beta_1|\mathbf{y}$. (Do not simplify the results.)

(**b**) Find Bayesian point and interval estimates of β_1.

11.10 Consider the Bayesian multiple regression model in which $\mathbf{y}|\boldsymbol{\beta},\tau$ is $N_n(\mathbf{X}\boldsymbol{\beta},\tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}$ is $N_{k+1}(\boldsymbol{\phi},\mathbf{V})$, and τ is gamma(α,δ). Note that this is similar to the model of Section 11.2 except that the prior distribution of $\boldsymbol{\beta}$ is *not conditional* on τ.

(**a**) Find the joint posterior density of $\boldsymbol{\beta},\tau|\mathbf{y}$ up to a normalizing constant.

(**b**) Find the conditional posterior density of $\boldsymbol{\beta}|\tau, \mathbf{y}$ up to a normalizing constant.

(**c**) Find the conditional posterior density of $\tau|\boldsymbol{\beta}, \mathbf{y}$ up to a normalizing constant.

(**d**) Develop a Gibbs sampling procedure for estimating the marginal posterior distributions of $\boldsymbol{\beta}|\mathbf{y}$ and $(1/\tau)|\mathbf{y}$.

11.11 Use the land rent data in Table 7.5.

(**a**) Find 95% Bayesian confidence intervals for β_1, β_2, and β_3 using (11.19) in connection with the model in which $\mathbf{y}|\boldsymbol{\beta}$, τ is $N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$, $\boldsymbol{\beta}|\tau$ is $N_{k+1}(\boldsymbol{\phi}, \tau^{-1}\mathbf{V})$, and τ is gamma(α, δ), where $\boldsymbol{\phi} = \mathbf{0}, \mathbf{V} = 100\mathbf{I}$, $\alpha = .0001$, and $\delta = .0001$.

(**b**) Repeat part (a), but use Gibbs sampling to approximate the confidence intervals.

(**c**) Use Gibbs sampling to obtain a 95% Bayesian posterior prediction interval for a future individual with $\mathbf{x}_0' = (1, 15, 30, .5)$.

(**d**) Repeat part (b), but use the model in which

$$\begin{aligned} \mathbf{y}|\boldsymbol{\beta}, \tau &\text{ is } N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}), \\ \boldsymbol{\beta} &\text{ is } N_{k+1}(\boldsymbol{\phi}, \mathbf{V}), \\ \tau &\text{ is gamma}(\alpha, \delta) \end{aligned} \tag{11.20}$$

where $\boldsymbol{\phi} = \mathbf{0}, \mathbf{V} = 100\mathbf{I}, \alpha = 0.0001,$ and $\delta = 0.0001$.

11.12 As in Section 11.5, show that

$$\int_0^\infty \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \tau^{1/2} e^{-\tau(y_0 - \mathbf{x}_0'\boldsymbol{\beta})^2/2} \tau^{(\alpha_{**}+k+1)/2} e^{-\tau[(\boldsymbol{\beta}-\boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\phi}_*)+\delta_{**}]/2} d\boldsymbol{\beta}\, d\tau$$

$$= c \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty [(\boldsymbol{\beta} - \boldsymbol{\phi}_*)'\mathbf{V}_*^{-1}(\boldsymbol{\beta} - \boldsymbol{\phi}_*) + (y_0 - \mathbf{x}_0'\boldsymbol{\beta})^2 + \delta_{**}]^{(-\alpha_{**}+k+4)/2} d\boldsymbol{\beta}.$$