$$P(\gamma|s) = [(1, x*)P^{-1}(\theta|s)(1, x*)']^{-1} \tag{3.39}$$

and one would usually estimate γ by (l, x*)μ* and an HPD interval for γ is easily found from the Student t tables. For example, a $1 - \Delta$ ($0 \leqslant \Delta \leqslant 1$) HPD interval for γ is

$$(1, x*)\mu* \pm t_{\Delta/2, n+2\alpha} P(\gamma/s)^{-1/2}, \tag{3.40}$$

where $P(\gamma|s)$ is given by (3.39).

The analysis of a simple linear regression model is completed with the development of a forecasting procedure to predict a future value w of the dependent variable Y when the regressor $x = Y_{n+1}$ and x_{n+1} is known. Recall from the section on predictive analysis in Chapter 1 (see pp. 14–18) the derivation of the Bayesian predictive density, with which a future value of Y is to be predicted. One obtains the future predictive density of Y if one substitutes the appropriate quantities into formula (1.39), which is based on the conjugate prior density with hyperparameters α, β, μ, and p.

Use the following substitutions: k = 1, $z = (1, x_{n+1})$, x is the $n \times 2$ matrix following formula (3.30), and Y is the $n \times 1$ vector of observations. The predictive density of $Y_{n+1} = w$ is a t-distribution with $n + 2\alpha$ degrees of freedom, location $A^{-}B$ and precision $(n + 2\alpha)A.(C - B'A^{-}B)^{-}$, where A, B, and C are explained by formula (1.39).

Notice, letting $\beta \to 0$, $p \to 0(2 \times 2)$ and $\alpha \to -1$,

$$A \to 1 - (1, x_{n+1})(x'x + z'z)^{-1} \begin{pmatrix} 1 \\ x_{n+1} \end{pmatrix} \tag{1.41}$$

$$B \to (1, x_{n+1})(z'z + x'x)^{-1} x'y \tag{1.42}$$

$$C \to \sum_{1}^{n} y_i^2 - y'x(z'z + x'x)^{-1} x'y, \tag{1.43}$$

where

$$z'z = \begin{pmatrix} 1 & x_{n+1} \\ x_{n+1} & x_{n+1}^2 \end{pmatrix}$$

and

$$x'y = \begin{matrix} \sum_{1}^{n} y_i \\ \sum_{1}^{n} x_i y_i \end{matrix}.$$

The predictive density of w which corresponds to the improper prior density is a univariate t density with $n - 2$ ($n > 2$) degrees of freedom, location $A^{-1}B = E(w|s)$ and precision $P(w|s) = (n - 2)A.(C - B'A^{-1}B)^{-1}$ where A, B, and C are now given by (1.41), (1.42), and (1.43), respectively, and a $1 - \Delta (0 \leqslant \Delta \leqslant 1)$ HPD interval for a future observation w (when $x = x_{n+1}$) is given by

$$E(w|s) \pm t_{\Delta/2, n-2} \sqrt{A^{-1}\left(C - B'A^{-1}B\right)(n-2)^{-1}}, \quad n > 3. \tag{1.44}$$

How does this interval compare to the "usual" prediction interval of a future observation? See Draper and Smith (1966) for details about the conventional simple linear regression analysis.

An Example of Simple Linear Regression

Consider the model

$$y_i = \theta_1 + \theta_2 x_i + e_i, \quad i = 1, 2, \ldots, 30, \tag{3.41}$$

$x_i = i$, the e_i are n.i.d. (0, 1), $\theta = 1$, and $\theta_2 = 2$. Thirty y_i values were generated with this model and the parameters of the normal-gamma prior density were determined empirically from the first six (x_i, y_i) pairs. The values of the hyperparameters were

$$\mu = (.03958, \quad 27426), \tag{3.42}$$
$$p = \begin{pmatrix} 5.58765 & 19.5568 \\ 19.5568 & 84.7461 \end{pmatrix},$$

$$\alpha = 1,$$

and

$$\beta = .931275,$$

where these values were calculated as

$$\mu = (z'z)^{-1}z'y^*,$$
$$P = z'z/s^2,$$
$$\beta = (s^2)^{-1},$$
$$s^2 = \frac{y^{*\prime}y^* - y^{*\prime}z(z'z)^{-1}z'y^*}{4},$$

and α was set equal to one. Also,

$$y^* = (y_1, y_1, \ldots, y_6)'$$

and

$$z = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_6 \end{pmatrix}.$$

The posterior analysis is based on the normal-gamma prior density, which was constructed from the first six observations and the likelihood function, which is based on the 7th through 29th pairs (i, y_i), $i = 7, 8, \ldots, 29$, selected from (3.41). Combining the prior and likelihood function gave the following values for the parameters of the joint posterior distribution:

$$E(\theta|s) = \begin{pmatrix} .970903 \\ 2.005566 \end{pmatrix} = \begin{pmatrix} E(\theta_1|s) \\ E(\theta_2|s) \end{pmatrix}, \tag{3.43}$$
$$P(\theta|s) = \begin{pmatrix} 23.1789 & 351.528 \\ 351.528 & 6931.33 \end{pmatrix},$$
$$P(\theta_1|s) = 5.3085,$$
$$P(\theta_2|s) = 1600.1.$$

The average value of a future observation y_{30} given $x_{30} = 30$ is

$$E(w|s) = 61.1376$$

compared to $y_{30} = 61.1289$, and the precision of the predictive distribution of y_{30} is $P(w|s) = .707206$, where $s = \{(i, y_i): i = y, \ldots, 29\}$.

If one uses a Jeffreys' prior density (3.33), the following values of the parameters of the posterior distribution were calculated:

$$\begin{pmatrix} E(\theta_1|s) \\ E(\theta_2|s) \end{pmatrix} = \begin{pmatrix} 1.0334 \\ 2.00196 \end{pmatrix}. \tag{3.44}$$
$$P(\theta|s) = \begin{pmatrix} 17.3861 & 312.95 \\ 312.95 & 6398.09 \end{pmatrix},$$
$$P(\theta_1|s) = 2.07877,$$
$$P(\theta_2|s) = 764.989,$$

and the mean of the predictive distribution of y_{30} is

$$E(w|s) = 61.0922, \text{ when } x_{30} = 30$$

when actually $y_{30} = 61.1289$, and the precision of w is $P(w|s) = .637491$.

By using a normal-gamma prior density with parameters fitted to the first six observations (i, y_i), $i = 1, 2, 3, 4, 5, 6$, one obtains prior estimates (3.42) of θ_1 and θ_2 which are very close to the true values of the intercept $\theta_1 = 1$ and slope $\theta_2 = 2$ of the regression model (3.41), therefore it is not surprising that the posterior estimates (3.43) are even closer to the true values.

On the other hand, when the prior information is vague, the posterior estimates (3.44) of the regression coefficients are not as close as they were when the prior was fitted to the first six observations. Also, we see the posterior precisions of and are less when Jeffreys' density expresses our prior information.

Suppose θ_1 and θ_2 are estimated by 95% HPD intervals. The marginal posterior distribution of is a t with $n + 2\alpha$ degrees of freedom, location $E(\theta_1|s)$

and precision $P(\theta_1|s)$, thus

$$E(\theta_1|s) \pm t_{.05/2,n+2\alpha}\sqrt{P^{-1}(\theta_1/s)}$$

is a 95% HPD region for and θ_1 reduces to (n = 23, α = 1)

$$1.0334 \pm (2.060)\sqrt{.693580} \quad \text{or} \quad (-0.39537,\ 2.462175)$$

As for θ_2 the 95% HPD interval is 2.00556 ± (2.060) (.02499) or (1.95407, 2.05709), which is a much shorter interval than the HPD interval for because the latter has a much smaller precision.

Now suppose y_{30} is to be predicted with a 95% HPD interval, then

$$E(w|s) \pm t_{.025,n+2\alpha}\sqrt{P^{-1}(w|s)}$$

is the appropriate formula. Substituting n = 23 and α = 1, $E(w|s) = 61.1376$ and $P(w|s) = .707206$ gives 61.1376 ± 2.4495 as the interval for forecasting the future observation y_{30} when $x_{30} = 30$.

What are the 95% HPD intervals for θ_1, θ_2 and $w|x = 30$ when vague prior information is used? The following y values were generated from the model (3.41) and were used in the preceding calculations. Beginning with y_1 and continuing through y_{30}, they are:

2.70178, 4.07126, 7.64881, 7.57121, 12.31, 13.6939, 14.7727, 16.6751, 19.1657, 20.4161, 24.6735, 26.1047, 26.4773, 28.2187, 29.5846, 34.3475, 35.337, 36.2295, 40.5608, 40.6826, 42.5413, 46.2414, 47.9887, 47.5131, 50.6753, 55.0602, 54.9478, 54.8281, 60.5314, 60.0476.

We see they are increasing, as they should because so are the x values since $x_i = i$, i = 1, 2, …, 30.

MULTIPLE LINEAR REGRESSION

Consider a linear regression model with two independent variables, then

$$y_i = \theta_1 + \theta_2 x_{i1} + \theta_3 x_{i2} + e_i, \tag{3.45}$$

where the e_i are n.i.d. $(0, \tau^{-1})$. The n triples (y_i, x_{i1}, x_{i2}), i = 1, 2, … n, are such that y_i is the i-th observation on a dependent variable y and x_{i1} and x_{i2} are the i-th observations on two nonstochastic regressor variables x_1 and x_2, respectively. The unknown parameters are $\theta = (\theta_1, \theta_2, \theta_3)$ and τ where $\theta \in R^3$ and τ > 0, and given the sample $S = \{(y_i, x_{i1}, x_{i2}): i = 1, 2, \ldots, n\}$, what inferences can be made about the parameters and how does one forcast future values of the dependent variable? Of course, these questions are answered in much the same way as one would answer the same questions about the simple linear regression model and the theory concerning the prior, predictive, and posterior analysis is covered in Chapter 1.

With the simple linear regression model, it was shown the prior analysis is based on the normal-gamma conjugate prior density or the Jeffreys' improper prior, the posterior analysis consists of studying the normal-gamma posterior distribution of θ and τ, and the predictive analysis is then done with the aid of the predictive t distribution. All these techniques apply to the analysis of multiple linear regression models; however, because the model contains more than one independent variable (as compared to a simple linear regression model), the posterior analysis becomes more involved since there are now more parameters in the model and the posterior distribution of θ has one more dimension.

The Posterior and Predictive Analysis

Let us briefly consider the posterior analysis if an improper prior density

$$\xi(\theta,\tau) \propto 1/\tau, \qquad \tau > 0, \quad \theta \in R^3 \tag{3.46}$$

is used in assessing one's prior information.

The joint posterior density of the parameter is

$$\xi(\theta,\tau|s) \propto \tau^{(n/2)-1}\exp\ -\ \tfrac{\tau}{2}\ \sum_{i=1}^{n}[y_i - \theta_1 - \theta_2 x_{i1} - \theta_3 x_{i2}]^2, \tag{3.47}$$

where $\theta_i \in R$ (i = 1, 2, 3) and τ > 0, and is a normal-gamma density, thus the marginal posterior distribution of θ is a trivariate t with n − 3 degrees of freedom, location vector

$$E(\theta|s) = A^{-1}B \tag{3.48}$$

and precision matrix

$$P(\theta|s) = (n-3)A\left(C - B'A^{-1}B\right)^{-1}, \quad n > 3 \tag{3.49}$$

where

$$A = \begin{pmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i1}x_{i2} \\ \Sigma x_{i2} & \Sigma x_{i1}x_{i2} & \Sigma x_{i2}^2 \end{pmatrix},$$

$$B = \begin{pmatrix} \Sigma y_i \\ \Sigma y_i x_{i1} \\ \Sigma y_i x_{i2} \end{pmatrix},$$

and

$$C = \sum_1^n y_i^2.$$

The three marginal posterior univariate distributions are t distributions. For example, the marginal posterior distribution of θ_2 is a t with n − 3 degrees of freedom, location

$$E(\theta_2|s) = (0,1,0)E(\theta|s) \tag{3.50}$$

and precision

$$P(\theta_2|s) = [(0,1,0)P^{-1}(\theta|s)(0,1,0)']^{-1} \tag{3.51}$$

and al−Δ($0 \leqslant \Delta \leqslant 1$) HPD region for θ_2 is given by

$$E(\theta_2|s) \pm t_{\Delta/2,n-3}\sqrt{P^{-1}(\theta_2|s)} \tag{3.52}$$

Consider a test of H_0: $\theta_2 = 0$ versus the alternative H_0: $\theta_2 \neq 0$, then if $0 \notin HPD_\Delta(\theta_2)$, H_0 is rejected at the "significance level" Δ. An HPD region for $(\theta_1, \theta_2, \theta_3)$ is found as follows.

Since $\theta|s \sim t_3[n\text{-}3E(\theta|s), P(\theta|s)]$,

$$F(\theta)\tfrac{1}{3}[\theta - E(\theta|s)]'P(\theta|s)(n-3)^{-1}[\theta - E(\theta|s)] \tag{3.53}$$

has an F-distribution with 3 and n − 3 degrees of freedom, and a $1 - \Delta$ ($0 < \Delta < 1$) HPD region for θ is

$$HPD_\Delta(\theta) = \{\theta : F(\theta) \leqslant F_{3,n-3,\Delta}\} \tag{3.54}$$

and one would reject H_0: $\theta = 0$ versus H_a: $\theta \neq 0$ whenever $F(0) > F_{\Delta;\,3,\,n-3}$. One may show this is equivalent to the size Δ likelihood-ratio test of H_0 versus H_a. This is, of course, the so-called test for significance of regression and the reader is referred to Draper and Smith (1966) for an explanation of the "usual" approach to tests of hypotheses in regression analysis.

Since the marginal posterior distribution of τ is gamma with parameters α' = (n − 3)/2 and β', where 2β' = C − B'A^{-1}B, point and interval estimators of this parameter are easily found.

In forecasting a future value of w of Y where

$$w = \theta_1 + \theta_2 x_1^* + \theta_3 x_2^* + e$$

and $e \sim n(0, \tau^{-1})$, one would use the Bayesian predictive distribution (1.39) of Chapter 1, where $z = (1, x^*_1, x^*_2)$, k = 1, p = 0(3 × 3), β = 0, α = −3/2, x is n × 3, namely

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$$

and $Y = (y_1, y_2, \ldots, y_n)'$ and a $1 - \gamma$($0 < \gamma < 1$) HPD region for w is given by (1.40).

An important problem in multiple linear regression is choosing the "best" subset of independent variables to put into the model. That is with two variables x_1 and x_2 one could have the models (always including the intercept)

$$y = \theta_1 + e_1 \quad (\text{neither } x_1 \text{ nor } x_2)$$

or

$$y = \theta_{11} + \theta_{12}x_1 + e_2 \quad (\text{only } x_1 \text{ included})$$

$$y = \theta_{11} + \theta_{12}x_1 + e_2 \quad (\text{only } x_1 \text{ included})$$

or

$$y = \theta_{21} + \theta_{22}x_2 + e_3 \quad (\text{only } x_2 \text{ included})$$

or

$$y = \theta_{31} + \theta_{32}x_1 + \theta_{33}x_2 + e_4 \quad (\text{both } x_1 \text{ and } x_2 \text{ included})$$

and the problem is to choose the appropriate model.

Of course, one's choice of a model depends on one's criterion of choosing the model, and some argue, see Lindley (1968), that one's criterion of choosing a model depends, in turn, on one's purpose of the regression analysis. If the main objective is to estimate the parameters, one's criterion of choosing a model will be based on the posterior distribution of the parameters, but on the other hand, if the main goal is to forecast future observations, one's criterion will involve the Bayesian predictive distribution of those future observations. This is Lindley's (1968) way to solve the problem, however, he uses a decision-theory approach, involving loss functions, and his method will not be taken in this book. Instead a more informal way is introduced,