6

The Linear Regression Model with General Error Covariance Matrix

6.1 INTRODUCTION

In this chapter we return to the linear regression model

$$y = X\beta + \varepsilon \tag{6.1}$$

In previous chapters, we assumed ε to be $N(0_N, h^{-1}I_N)$. This statement is really a combination of several assumptions, some of which we might want to relax. The assumption that the errors have mean zero is an innocuous one. If a model has errors with a non-zero mean, this non-zero mean can be incorporated into the intercept. To be precise, a new model, which is identical to the old except for the intercept, can be created which does have mean zero errors. However, the assumption that the covariance matrix of the errors is $h^{-1}I_N$ might not be innocuous in many applications. Similarly, the assumption that the errors have a Normal distribution is one which might be worth relaxing in many cases. In this chapter, we consider several empirically-relevant ways of relaxing these assumptions and describe Bayesian inference in the resulting models.

All the models in this chapter are based on (6.1) and the following assumptions:

1. ε has a multivariate Normal distribution with mean 0_N and covariance matrix $h^{-1}\Omega$, where Ω is an $N \times N$ positive definite matrix.
2. All elements of X are either fixed (i.e. not random variables) or, if they are random variables, they are independent of all elements of ε with a probability density function, $p(X|\lambda)$, where λ is a vector of parameters that does not include β and h.

Note that these assumptions are identical to those made in Chapters 2, 3 and 4, except for the assumption about the error covariance matrix. However, as we shall show in this chapter, assumptions about this error covariance matrix are closely related to distributional assumptions. Hence, we can use this framework to free up the assumption that the errors are Normally distributed.

The various models we discuss differ in the precise form that Ω takes. After discussing some general theory which is relevant for any choice of Ω, we examine several specific choices which arise in many applications. We begin by considering *heteroskedasticity*, which is the name given for cases where the error variances differ across observations. We consider two types of heteroskedasticity: one where its form is known and one where it is unknown. The latter case allows us to free up the Normality assumption, and we discuss, in particular, how a certain model with heteroskedasticity of unknown form is equivalent to a linear regression models with Student-t errors. This model allows us to introduce the concept of a *hierarchical prior*, which will be used extensively in the remainder of this book. Subsequently, we consider a case where the errors are correlated with one another. In particular, we discuss the Normal linear regression model with *autoregressive* or *AR* errors. In addition to being of interest in and of themselves, AR models are important time series models and provide us with a convenient starting point for an introduction to time series methods. The final model considered in this chapter is the *seemingly unrelated regressions* or *SUR* model. This is a model which has several equations corresponding to multiple dependent variables and is a component of models considered in future chapters.

6.2 THE MODEL WITH GENERAL Ω

6.2.1 Introduction

Before discussing the likelihood function, prior, posterior and computational methods, we present a general result which has implications for both interpretation and computation for this model. Since Ω is a positive definite matrix, it follows from Appendix A, Theorem A.10 that an $N \times N$ matrix P exists with the property that $P\Omega P' = I_N$. If we multiply both sides of (6.1) by P, we obtain a transformed model

$$y^* = X^*\beta + \varepsilon^* \tag{6.2}$$

where $y^* = Py$, $X^* = PX$ and $\varepsilon^* = P\varepsilon$. It can be verified that ε^* is $N(0_N, h^{-1}I_N)$. Hence, the transformed model given in (6.2) is identical to the Normal linear regression model discussed in Chapters 2, 3 and 4. This has two important implications. First, if Ω is known, Bayesian analysis of the Normal linear regression model with nonscalar error covariance matrix is straightforward. The researcher can transform her data and carry out Bayesian inference using the methods of earlier chapters. Secondly, if Ω is unknown, (6.2) suggests methods for Bayesian computation. That is, conditional on Ω, (6.2) implies that the posteriors of β and h will be of the same form as in previous chapters and, hence, these earlier results can be used for derivations relating to β and h. If the prior for β and h is $NG(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu})$, then all the results of Chapters 2 and 3 are applicable *conditional upon* Ω and we can draw upon these results to derive a posterior simulator. For instance, (3.14) can be used to verify that $p(\beta|y, \Omega)$ is a multivariate t distribution and this, combined with a posterior simulator for $p(\Omega|y)$ can be used to carry out posterior inference on β and Ω. This is done in Griffiths (2001) for the noninformative limiting case of the natural conjugate prior. In this chapter we use a prior of the independent Normal-Gamma form of Chapter 4, Section 4.2, and a Gibbs sampler which sequentially draws from $p(\beta|y, h, \Omega)$, $p(h|y, \beta, \Omega)$ and $p(\Omega|y, \beta, h)$ can be set up. The first two of these posterior conditionals will be Normal and Gamma, as in Section 4.2.2 of Chapter 4, while $p(\Omega|y, \beta, h)$ depends upon the precise form of Ω. Hence, the only new derivations which are required relate to this latter distribution. Similar considerations hold for priors which impose inequality constraints (see Chapter 4, Section 4.3).

6.2.2 The Likelihood Function

Using the properties of the multivariate Normal distribution, the likelihood function can be seen to be:

$$p(y|\beta, h, \Omega) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}}|\Omega|^{-\frac{1}{2}}\left\{\exp\left[-\frac{h}{2}(y - X\beta)'\Omega^{-1}(y - X\beta)\right]\right\} \tag{6.3}$$

or, in terms of the transformed data,

$$p(y^*|\beta, h, \Omega) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}}\left\{\exp\left[-\frac{h}{2}(y^* - X^*\beta)'(y^* - X^*\beta)\right]\right\} \tag{6.4}$$

In Chapter 3, we showed how the likelihood function could be written in terms of OLS quantities (see (3.4)–(3.7)). Here an identical derivation using the transformed model yields a likelihood function written in terms of Generalized Least Squares[1] (GLS) quantities:

$$\nu = N - k \tag{6.5}$$

$$\widehat{\beta}(\Omega) = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \tag{6.6}$$

and

$$\begin{aligned} s^2(\Omega) &= \frac{(y^* - X^*\widehat{\beta}(\Omega))'(y^* - X^*\widehat{\beta}(\Omega))}{\nu} \\ &= \frac{(y - X\widehat{\beta}(\Omega))'\Omega^{-1}(y - X\widehat{\beta}(\Omega))}{\nu} \end{aligned} \tag{6.7}$$

---

[1] For the reader unfamiliar with the concept of a Generalized Least Squares estimator, any frequentist econometrics textbook such as Green (2000) will provide a detailed discussion. Knowledge of this material is not necessary to understand the material in this chapter.

as:

$$
\begin{aligned}
p(y|\beta, h, \Omega) = {} & \frac{1}{(2\pi)^{\frac{N}{2}}} \\
& \times \left\{ h^{\frac{1}{2}} \exp\left[-\frac{h}{2}(\beta - \widehat{\beta}(\Omega))' X' \Omega^{-1} X (\beta - \widehat{\beta}(\Omega)) \right] \right\} \\
& \times \left\{ h^{\frac{\nu}{2}} \exp\left[-\frac{h\nu}{2s(\Omega)^{-2}} \right] \right\}
\end{aligned} \tag{6.8}
$$

6.2.3 The Prior

Here we use an independent Normal-Gamma prior for β and h (see Chapter 4, Section 4.2.1), and use the general notation, $p(\Omega)$, to indicate the prior for Ω. In other words, the prior used in this section is

$$p(\beta, h, \Omega) = p(\beta)p(h)p(\Omega)$$

where

$$p(\beta) = f_N(\beta|\underline{\beta}, \underline{V}) \tag{6.9}$$

and

$$p(h) = f_G(h|\underline{\nu}, \underline{s}^{-2}) \tag{6.10}$$

6.2.4 The Posterior

The posterior is proportional to the prior times the likelihood and is of the form

$$
\begin{aligned}
p(\beta, h, \Omega|y) \propto {} & p(\Omega) \\
& \times \left\{ \exp\left[-\frac{1}{2}\{h(y^* - X^*\beta)'(y^* - X^*\beta) \right.\right. \\
& \left.\left. + (\beta - \underline{\beta})' \underline{V}^{-1} (\beta - \underline{\beta})\} \right] \right\} \\
& \times h^{\frac{N+\underline{\nu}-2}{2}} \exp\left[-\frac{h\underline{\nu}}{2\underline{s}^{-2}} \right]
\end{aligned} \tag{6.11}
$$

This posterior is written based on the likelihood function expressed as in (6.4). Alternative expressions based on (6.3) or (6.8) can be written out. However, we do not do this, since this joint posterior density for β, h and Ω does not take the form of any well-known and understood density and, hence, cannot be directly used in a simple way for posterior inference. At least some of the conditionals of the posterior are, however, simple. Proceeding in the same manner as in Chapter 4 (see (4.4)–(4.10) and surrounding discussion), it can be verified that the posterior of β, conditional on the other parameters of the model is multivariate Normal:

$$\beta|y, h, \Omega \sim N(\overline{\beta}, \overline{V}) \tag{6.12}$$

where

$$\overline{V} = (\underline{V}^{-1} + hX'\Omega^{-1}X)^{-1} \tag{6.13}$$

and

$$\overline{\beta} = \overline{V}(\underline{V}^{-1}\underline{\beta} + hX'\Omega^{-1}X\widehat{\beta}(\Omega)) \tag{6.14}$$

The posterior for h conditional on the other parameters in the model is Gamma:

$$h|y, \beta, \Omega \sim G(\overline{s}^{-2}, \overline{\nu}) \tag{6.15}$$

where

$$\overline{\nu} = N + \underline{\nu} \tag{6.16}$$

and

$$\overline{s}^2 = \frac{(y - X\beta)'\Omega^{-1}(y - X\beta) + \underline{\nu}\underline{s}^2}{\overline{\nu}} \tag{6.17}$$

The posterior for Ω conditional on β and h has a kernel of the form

$$p(\Omega|y, \beta, h) \propto p(\Omega)|\Omega|^{-\frac{1}{2}} \left\{\exp\left[-\frac{h}{2}(y - X\beta)'\Omega^{-1}(y - X\beta)\right]\right\} \tag{6.18}$$

In general, this conditional posterior does not take any easily recognized form. In future sections of this chapter we consider particular forms for Ω and derive appropriate posterior simulators. At this stage, we only note that, if we could take posterior draws from $p(\Omega|y, \beta, h)$, then a Gibbs sampler for this model could be set up in a straightforward manner, since $p(\beta|y, h, \Omega)$ is Normal and $p(h|y, \beta, \Omega)$ is Gamma.

6.3 HETEROSKEDASTICITY OF KNOWN FORM

6.3.1 Introduction

Heteroskedasticity is said to occur if the error variances differ across observations. The models in previous chapters all had error variances which were identical across observations and were, thus, *homoskedastic*. A couple of examples will serve to motivate how heteroskedasticity might arise in practice. Consider first a microeconomic example where the dependent variable is company sales. If errors are proportionate to firm size, then errors for small firms will tend to smaller than those for large firms. Secondly, heteroskedasticity might arise in a study involving data from many countries. Since developed countries have better agencies for collecting statistics than developing countries, it might be the case that errors are smaller in the former countries.

In terms of our regression model, heteroskedasticity occurs if

$$\Omega = \begin{bmatrix} \omega_1 & 0 & \cdot & \cdot & 0 \\ 0 & \omega_2 & 0 & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & \omega_N \end{bmatrix} \tag{6.19}$$

In other words, the Normal linear regression model with heteroskedastic errors is identical to that studied in Chapters 2–4, except that we now assume $var(\varepsilon_i) = h^{-1}\omega_i$ for $i = 1, \ldots, N$.

The examples above indicate that we often know (or at least suspect), what form this heteroskedasticity will take. For instance, ω_i might depend upon whether firm i is small or large or whether country i is developing or developed. Here we will assume that

$$\omega_i = h(z_i, \alpha) \tag{6.20}$$

where $h()$ is a positive function which depends on parameters α and a p-vector of data, z_i. z_i may include some or all of the explanatory variables, x_i. A common choice for $h()$, which ensures that the error variances are positive is:

$$h(z_i, \alpha) = (1 + \alpha_1 z_{i1} + \alpha_2 z_{i2} + \cdots + \alpha_p z_{ip})^2 \tag{6.21}$$

but the discussion of this section works for other choices.

The prior, likelihood and posterior for this model are simply those in Section 6.2 with the expression for Ω given in (6.19) plugged in. Accordingly, we do not write them out here. Note, however, that in the present section Ω depends upon α and, hence, the formulae below are written as depending on α.

To carry out Bayesian inference in the present heteroskedastic model, a posterior simulator is required. The previous discussion suggests that a Metropolis-within-Gibbs algorithm (see Section 5.5.3) might be appropriate. In particular, as noted in (6.12) and (6.15), $p(\beta|y, h, \alpha)$ is Normal and $p(h|y, \beta, \alpha)$ is Gamma, and we require only a method for taking draws from $p(\alpha|y, \beta, h)$ to have a complete posterior simulator. Unfortunately, if we plug (6.19) and (6.20) into (6.18) to obtain an expression for $p(\alpha|y, \beta, h)$ the result does not take the form of any convenient density. Nevertheless, a Metropolis–Hastings algorithm can be developed. In the empirical illustration which follows, a Random Walk Chain Metropolis–Hastings algorithm (see Section 5.5.2 of Chapter 5) is used although other algorithms are possible. Bayes factors for any hypothesis of interest (e.g. $\alpha_1 = \cdots = \alpha_p = 0$ which is the hypothesis that heteroskedasticity does not exist) can be calculated using the Gelfand–Dey approach. Alternatively, posterior predictive p-values or HPDIs can be calculated to shed light on the fit and appropriateness of the model. Predictive inference in this model can be carried out using the strategy outlined in Chapter 4, Section 4.2.6.

6.3.2 Empirical Illustration: Heteroskedasticity of a Known Form

We use the house price data set introduced in Chapter 3, to illustrate the use of Gibbs sampling in the Normal linear regression model with heteroskedasticity of known form. The reader is referred to Section 3.9 of Chapter 3 for a precise description of the dependent and explanatory variables for this data set. We assume the heteroskedasticity takes the form given in (6.21) and that $z_i = (x_{i2}, \dots, x_{ik})'$. The priors for β and h are given in (6.9) and (6.10) and we use the same values for hyperparameters as in Chapter 4, Section 4.2.7. We use a noninformative prior for α of the form

$$p(\alpha) \propto 1$$

Note that this prior is improper and, hence, we cannot calculate meaningful Bayes factors for hypotheses involving the elements of α. Accordingly, we present 95% HPDIs along with posterior means and standard deviations in Table 6.1.

The posterior simulator is a Metropolis-within-Gibbs algorithm, with draws of β and h taken from (6.12) and (6.15), respectively. Draws from $p(\alpha|y, \beta, h)$ are taken using a Random Walk Chain Metropolis–Hastings algorithm with a Normal increment random variable (see Chapter 5, (5.10)). $p(\alpha|y, \beta, h)$ is given in (6.18) with (6.21) providing the precise form for Ω. Equation (6.18), evaluated at old and candidate draws, is used to calculate the acceptance probability (see Chapter 5, (5.11)). The variance of the proposal density, labelled Σ in (5.12), is chosen by first setting $\Sigma = cI$ and experimenting with different values of the scalar c until a value is found which yields reasonable acceptance probabilities. The posterior simulator is then run using this value to yield an estimate of the posterior variance of α, $\widehat{var}(\alpha|y)$. We then set $\Sigma = c\widehat{var}(\alpha|y)$ and experiment with different values of c until we find one which yields an average acceptance probability of roughly 0.50. Then a final long run of 30 000 replications, with 5000 burn-in replications

Table 6.1 Posterior Results for β, h and α

	Mean	Standard Deviation	95% HPDI
β_1	−5453.92	2976.04	[−10 310, 557]
β_2	6.12	0.40	[5.42, 6.82]
β_3	3159.52	1025.63	[1477, 4850]
β_4	14 459.34	1672.43	[11 742, 17 224]
β_5	7851.11	939.34	[6826, 9381]
h	1.30×10^{-7}	4.05×10^{-8}	$[7 \times 10^{-8}, 2 \times 10^{-7}]$
α_1	5.49×10^{-4}	1.36×10^{-4}	$[3 \times 10^{-4}, 8 \times 10^{-4}]$
α_2	0.68	0.32	[0.21, 1.26]
α_3	0.70	0.42	[0.08, 1.40]
α_4	−0.35	0.33	[−0.89, 0.18]

discarded, is taken. MCMC diagnostics indicate convergence of the Metropolis-within-Gibbs algorithm and numerical standard errors indicate an approximation error which is small relative to posterior standard deviations of all parameters.

Table 6.1 indicates that heteroskedasticity does seem to exist for this data set. That is, the 95% HPDIs do not include zero for α_1, α_2 and α_3 indicating that lot size, number of bedrooms and number of bathrooms have significant explanatory power in the equation for heteroskedasticity. The fact that all of these coefficients are positive indicates that the error variance for large houses tends to be bigger than for small houses. In previous chapters, we ignored heteroskedasticity when working with this data set. To see what effect this omission had, you may wish to compare the results in Table 6.1 with those in Table 4.1. The latter table contains results for the homoskedastic version of the model, but uses the same data and the same prior for β and h. It can be seen that including heteroskedasticity has some effect on the posterior of β. For instance, the posterior mean of β_4 was 16 133 in the homoskedastic model and is 14 459 in the heteroskedastic one. However, for many purposes, such differences might be fairly small and the researcher might conclude that the incorporation of heteroskedasticity has not had an enormous effect on results relating to β.

6.4 HETEROSKEDASTICITY OF AN UNKNOWN FORM: STUDENT-t ERRORS

6.4.1 General Discussion

In the previous section, we assumed that the heteroskedasticity was of a form given by (6.20). The question arises as to how to proceed if you suspect heteroskedasticity is present, but of unknown form. In other words, you are willing to assume (6.19), but unwilling to assume a functional form as in (6.20). With N observations and $N + k + 1$ parameters to estimate (i.e. β, h and $\omega = (\omega_1, \ldots, \omega_N)'$), treatment of heteroskedasticity of unknown form may sound like a difficult task. However, as we shall see, it is not too difficult to extend the techniques of the previous sections of this chapter to be applicable to this model. Furthermore, the method developed to handle this case is quite important for two reasons. Firstly, the method involves the use of a *hierarchical prior*. This is a concept we will use again and again throughout the remainder of this book. Hierarchical priors have played a big role in many recent developments in Bayesian statistical theory and are gradually becoming more popular in econometrics as well. They are commonly used as a way of making flexible, parameter-rich models more amenable to statistical analysis.[2] Secondly, this model also allows us to

---

[2] Frequentist econometricians also work with models that are hierarchical in structure and very similar to ones discussed in this book. However, the frequentist statistical theory surrounding these models is often quite difficult. Accordingly, Bayesian methods are particularly popular in this area of the statistical literature.

introduce concepts relating to flexible econometric modelling (see Chapter 10) and, in particular, allows us to free up the assumption of Normal errors that we have used so far.

We begin by eliciting $p(\omega)$, the prior for the N-dimensional vector ω. As in previous chapters, it proves convenient to work with error precisions rather than variances and, hence, we define $\lambda \equiv (\lambda_1, \lambda_2, \ldots, \lambda_N)' \equiv (\omega_1^{-1}, \omega_2^{-1}, \ldots, \omega_N^{-1})'$. Consider the following prior for λ:

$$p(\lambda) = \prod_{i=1}^{N} f_G(\lambda_i|1, \nu_\lambda) \tag{6.22}$$

Note that the prior for λ depends upon a hyperparameter, ν_λ, which is chosen by the researcher and assumes each λ_i comes from the same distribution. In other words, (6.22) implies that the λ_is are i.i.d. draws from the Gamma distribution. This assumption (or something similar) is necessary to deal with the problems caused by the high-dimensionality of λ. Intuitively, if we were to simply treat $\lambda_1, \ldots, \lambda_N$ as N completely independent and unrestricted parameters, we would not have enough observations to estimate each one of them. Equation (6.22) puts some structure which allows for estimation. It allows for all the error variances to be different from one another, but says they are all drawn from the same distribution. Thus, we can have a very flexible model, but enough structure is still imposed to allow for statistical inference.

You may be wondering why we chose the particular form given in (6.22). For instance, why should the λ_is be i.i.d. draws from the Gamma distribution with mean 1.0? Rather remarkably, it turns out that this model, with likelihood given by (6.3) and prior given by (6.9), (6.10) and (6.22) is *exactly the same* as the linear regression model with i.i.d. Student-t errors with ν_λ degrees of freedom. In other words, if we had begun by assuming

$$p(\varepsilon_i) = f_t(\varepsilon_i|0, h^{-1}, \nu_\lambda) \tag{6.23}$$

for $i = 1, \ldots, N$, derived the likelihood and used (6.9) and (6.10) as priors for β and h, respectively, we would have ended up with exactly the same posterior. We will not formally prove this statement and the interested reader is referred to Geweke (1993) for proofs and further explanation. Note, however, the power and convenience of this result. The Student-t distribution is similar to the Normal, but has fatter tails and is more flexible. In fact, the Normal distribution is a special case of the Student-t which occurs as $\nu_\lambda \to \infty$. Thus, we have a model that allows for a more flexible error distribution, but we have achieved this result without leaving our familiar Normal linear regression model framework. Furthermore, we can draw on the computational methods derived above to develop a posterior simulator for the linear regression model with independent Student-t errors. For this reason, an explicit statement of the likelihood function for this model is not given here.

In Chapter 10 we discuss several ways of making models more flexible. However, it is worthwhile briefly noting that the model discussed here involves a *mixture of Normals* distribution of a particular sort. Intuitively, if a Normal distribution is too restrictive, you can create a more flexible distribution by taking a weighted average of more than one Normal distribution. As more and more Normals are mixed, the distribution becomes more and more flexible and, as discussed in Chapter 10, can approximate any distribution to a high degree of accuracy. Thus, mixtures of Normals models are a powerful tool for use when economic theory does not suggest a particular form for the likelihood function and you wish to be very flexible. Our treatment of heteroskedasticity of an unknown form is equivalent to a *scale mixture of Normals*. This means that the assumption that ε_i are independent $N(0, h^{-1}\lambda_i^{-1})$ with prior for λ_i given in (6.22) is equivalent to the assumption that the error distribution is a weighted average (or mixture) of different Normal distributions which have different variances (i.e. different scales) but the same means (i.e. all errors have mean zero). When this mixing is done using $f_G(\lambda_i|1, \nu_\lambda)$ densities, the mixture of Normals ends up being equivalent to the t distribution. However, using densities other than the $f_G(\lambda_i|1, \nu_\lambda)$ will yield other distributions more flexible than the Normal. See Chapter 10 for further details.

The previous discussion assumed that ν_λ was known. In practice, this would usually not be a reasonable assumption, and it is, thus, desirable to treat it as an unknown parameter. In the Bayesian framework, every parameter requires a prior distribution and, at this stage, we will use the general notation $p(\nu_\lambda)$. Note that, if we do this, the prior for λ is specified in two steps, the first being (6.22), the other being $p(\nu_\lambda)$. Alternatively, the prior for λ can be written as $p(\lambda|\nu_\lambda)p(\nu_\lambda)$. Priors written in two (or more) steps in this way are referred to as hierarchical priors. Writing a prior as a hierarchical prior is often a convenient way of expressing prior information and many of the models discussed in future chapters will be written in this way. However, we do stress the convenience aspect of hierarchical priors. It is never necessary to use a hierarchical prior, since the laws of probability imply that every hierarchical prior can be written in a non-hierarchical fashion. In the present case, the result $p(\lambda) = \int p(\lambda|\nu_\lambda)p(\nu_\lambda)d\nu_\lambda$ could be used to derive the non-hierarchical version of our prior for λ.

In all of the previous empirical illustrations, we have presented posterior means as point estimates of parameters and posterior standard deviations as measures of the uncertainty associated with the point estimates. However, as mentioned in Chapter 1, means and standard deviations do not exist for all valid probability density functions. The present model is the first one we have considered where means and standard deviations do not necessarily exist. In particular, Geweke (1993) shows that if you use a common noninformative prior for β (i.e. $p(\beta) \propto 1$ on the interval $(-\infty, \infty)$), then the posterior mean does not exist, unless $p(\nu_\lambda)$ is zero on the interval $(0, 2]$. The posterior standard deviation does not exist unless $p(\nu_\lambda)$ is zero on the interval $(0, 4]$. Hence, the researcher who wants to use a noninformative prior for β should either use a prior which excludes small values for ν_λ or present posterior medians and interquartile ranges (which will exist for any valid p.d.f.). With an informative Normal prior for β like (6.9), the posterior mean and standard deviation of β will exist.

It is also risky to use a noninformative prior for ν_λ. A naive researcher who wishes to be noninformative might use an improper Uniform prior:

$$p(\nu_\lambda) \propto 1 \text{ for } \nu_\lambda \in (0, \infty)$$

thinking that it would allocate equal prior weight to every interval of equal length. But the Student-t distribution with ν_λ degrees of freedom approaches the Normal distribution as $\nu_\lambda \to \infty$. In practice, the Student-t is virtually identical to the Normal for $\nu_\lambda > 100$. Our naive 'noninformative' prior allocates virtually all its weight to this region (i.e. $\frac{p(\nu_\lambda \leq 100)}{p(\nu_\lambda > 100)} = 0$). So this prior, far from being noninformative, is extremely informative: it is saying the errors are Normally distributed! This illustrates one of the problems with trying to come up with noninformative priors. There is a large Bayesian literature on how to construct noninformative priors (Zellner, 1971, provides an introduction to this). A detailed discussion of this issue is beyond the scope of the present book (although see Chapter 12, Section 12.3). However, it is worth noting that extreme care must be taken when trying to elicit a noninformative prior.

6.4.2 Bayesian Computation

In this subsection, we develop a Gibbs sampler for posterior analysis of β, h, λ and ν_λ. The Gibbs sampler requires the derivation of the full conditional posterior distributions of these parameters. We have already derived some of these as $p(\beta|y, h, \lambda)$ and $p(h|y, \beta, \lambda)$ are given in (6.12) and (6.15), respectively.[3] Hence, we focus on $p(\lambda|y, \beta, h, \nu_\lambda)$ and $p(\nu_\lambda|y, \beta, h, \lambda)$. The former of these can be derived by plugging the prior given in (6.22) into the general form for the conditional posterior given in (6.18). An examination of the resulting density shows that the λ_is are independent of one another (conditional on the other parameters of the model) and each of the conditional posteriors for λ_i has the form of a Gamma density. Formally, we have

$$p(\lambda|y, \beta, h, \nu_\lambda) = \prod_{i=1}^{N} p(\lambda_i|y, \beta, h, \nu_\lambda) \tag{6.24}$$

and

$$p(\lambda_i|y, \beta, h, \nu_\lambda) = f_G\left(\lambda_i|\frac{\nu_\lambda + 1}{h\varepsilon_i^2 + \nu_\lambda}, \nu_\lambda + 1\right) \tag{6.25}$$

Note that, conditional on knowing β, ε_i can be calculated and, hence, the parameters of the Gamma density in (6.25) can be calculated in the Gibbs sampler.

---

[3] Formally, the full conditionals to be used in the Gibbs sampler should be $p(\beta|y, h, \lambda, \nu_\lambda)$ and $p(h|y, \beta, \lambda, \nu_\lambda)$. However, conditional on λ, ν_λ adds no new information and, thus, $p(\beta|y, h, \lambda, \nu_\lambda) = p(\beta|y, h, \lambda)$ and $p(h|y, \beta, \lambda, \nu_\lambda) = p(h|y, \beta, \lambda)$.

Up until now, we have said nothing about the prior for ν_λ, and its precise form has no relevance for the posterior conditional for the other parameters. However, the form of $p(\nu_\lambda)$ does, of course, affect $p(\nu_\lambda|y, \beta, h, \lambda)$ and, hence, we must specify it here. Since we must have $\nu_\lambda > 0$, we use an exponential distribution for the prior. As noted in Appendix B, Theorem B.7, the exponential density is simply the Gamma with two degrees of freedom. Hence, we write

$$p(\nu_\lambda) = f_G(\nu_\lambda|\underline{\nu}_\lambda, 2) \tag{6.26}$$

Other priors can be handled with small changes in the following posterior simulation algorithm.

$p(\nu_\lambda|y, \beta, h, \lambda)$ is relatively easy to derive, since ν_λ does not enter the likelihood and it can be confirmed that $p(\nu_\lambda|y, \beta, h, \lambda) = p(\nu_\lambda|\lambda)$. It follows from Bayes theorem that

$$p(\nu_\lambda|\lambda) \propto p(\lambda|\nu_\lambda)p(\nu_\lambda)$$

and, thus, the kernel of the posterior conditional of ν_λ is simply times (6.22) times (6.26). Thus, we obtain

$$p(\nu_\lambda|y, \beta, h, \lambda) \propto \left(\frac{\nu_\lambda}{2}\right)^{\frac{N\nu_\lambda}{2}} \Gamma\left(\frac{\nu_\lambda}{2}\right)^{-N} \exp(-\eta\nu_\lambda) \tag{6.27}$$

where

$$\eta = \frac{1}{\underline{\nu}_\lambda} + \frac{1}{2}\sum_{i=1}^{N}[\ln(\lambda_i^{-1}) + \lambda_i]$$

This density is a non-standard one. Hence, we will use a Metropolis–Hastings algorithm to take draws from (6.27). However, it should be mentioned in passing that Geweke (1993) recommends use of another useful computational technique called *acceptance sampling*. This technique is very useful when the non-standard distribution that the researcher wishes to draw from is univariate and can be bounded. We will not discuss it here, but Geweke (1993) provides more detail on acceptance sampling as it relates to the present model (see also Chapter 12, Section 12.1). Devroye (1986) offers a thorough discussion of acceptance sampling in general.

For many hypotheses (e.g. $\beta_j = 0$) the Savage–Dickey density ratio can be used for model comparison. It can be calculated as described in Chapter 4, Section 4.2.5. However, not all hypotheses are easily calculated using the Savage–Dickey ratio. For instance, in many cases you might be interested in seeing whether there is any evidence of departures from Normality. In this case, you would wish to compare $M_1\colon \nu_\lambda \to \infty$ to $M_2\colon \nu_\lambda$ is finite. These models do not easily fit in the nested model comparison framework for which the Savage–Dickey density ratio is suitable. However, the Bayes factor comparing these two models can be calculated using the Gelfand–Dey approach. Note that this would require a posterior simulator for each model (i.e. the posterior simulator in Chapter 4, Section 4.2 for M_1 and the one described in this section for M_2). Alternatively, posterior predictive p-values or HPDIs can be calculated to shed light on the fit and appropriateness of the model. Predictive inference in this model can be carried out using the strategy outlined in Chapter 4, Section 4.2.6.

6.4.3 Empirical Illustration: The Regression Model with Student-t Errors

We return to our familiar house price data set introduced in Chapter 3 to illustrate the use of Gibbs sampling in the linear regression model with independent Student-t errors (or, equivalently, the Normal linear regression model with heteroskedasticity of unknown form). The reader is referred to Section 3.9 for a precise description of the dependent and explanatory variables for this data set. The priors for β and h are given in (6.9) and (6.10) and we use the same values for hyperparameters as in Chapter 4, Section 4.2.7. The prior for ν_λ depends upon the hyperparameter $\underline{\nu}_\lambda$, its prior mean. We set $\underline{\nu}_\lambda = 25$, a value which allocates substantial prior weight both to very fat-tailed error distributions (e.g. $\nu_\lambda < 10$), as well as error distributions which are roughly Normal (e.g. $\nu_\lambda > 40$).

The posterior simulator is a Metropolis-within-Gibbs algorithm, with draws of β and h taken from (6.12) and (6.15), respectively. Draws from $p(\lambda|y, \beta, h, \nu_\lambda)$ are taken using (6.25). For $p(\nu_\lambda|y, \beta, h, \lambda)$, we use a Random Walk Chain Metropolis–Hastings algorithm with a Normal increment random variable (see Chapter 5, (5.10)) . Equation (6.27), evaluated at old and candidate draws, is used to calculate the acceptance probability (see Chapter 5, (5.11)). Candidate draws of ν_λ which are less than or equal to zero have the acceptance probability set to zero. The variance of the proposal density, labelled Σ in (5.12), is chosen by first setting $\Sigma = c$ and experimenting with different values of the scalar c until a value is found which yields reasonable acceptance probabilities. The posterior simulator is then run using this value to yield an estimate of the posterior variance of ν_λ, $\widehat{var(\nu_\lambda|y)}$. We then set $\Sigma = c\widehat{var(\nu_\lambda|y)}$ and experiment with different values of c until we find one which yields an average acceptance probability of roughly 0.50. Then a final long run of 30 000 replications, with 5000 burn-in replications discarded, is taken. MCMC diagnostics indicate convergence of the Metropolis-within-Gibbs algorithm and numerical standard errors indicate an approximation error which is small relative to posterior standard deviations of all parameters.

Table 6.2 contains posterior results for the key parameters and it can be seen that, although posteriors for the elements of β are qualitatively similar to those presented in Tables 4.1 and 6.1, the posterior for ν_λ indicates the errors exhibit substantial deviations from Normality. Since this crucial parameter is univariate, we also plot a histogram approximation to its posterior. Figure 6.1 indicates that $p(\nu_\lambda|y)$ has a shape which is quite skewed and confirms the finding that virtually all of the posterior probability is allocated to small values for the degrees of freedom parameter. Note, however, that there is virtually no support for extremely small values which would imply extremely fat tails. Remember that the Cauchy distribution is the Student-t with $\nu_\lambda = 1$. It has such fat tails that its mean does not exist. There is no evidence for this sort of extreme behavior in the errors for the present data set.

Table 6.2 Posterior Results for β and ν_λ

	Mean	Standard Deviation	95% HPDI
β_1	−413.15	2898.24	[−5153, 4329]
β_2	5.24	0.36	[4.65, 5.83]
β_3	2118.02	972.84	[501, 3709]
β_4	14 910.41	1665.83	[12 188, 17 631]
β_5	8108.53	955.74	[6706, 9516]
ν_λ	4.31	0.85	[3.18, 5.97]

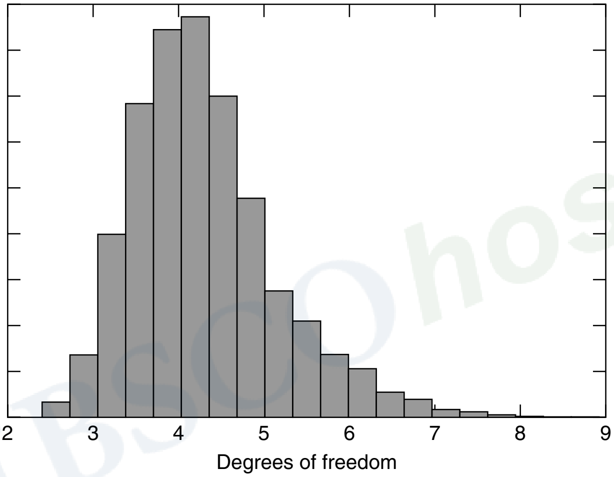

Figure 6.1 Posterior Density for Degrees of Freedom

6.5 AUTOCORRELATED ERRORS

6.5.1 Introduction

Many time series variables are correlated over time due to factors such as habit persistence or the time taken for adjustments to take place. This correlation between values of a variable at different times can spill over to the error. It is thus desirable to consider forms for the error covariance matrix which allow for this. In earlier chapters we assumed ε to be $N(0_N, h^{-1}I_N)$. In the previous sections of the present chapter, we relaxed this to allow for the error covariance matrix to be diagonal. However, so far, we have always assumed the errors to be uncorrelated with one another (i.e. $E(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$). In this section, we consider a model which relaxes this assumption.

Following common practice, we will use a subscript t to indicate time. That is, y_t for $t = 1, \ldots, T$ indicates observations on the dependent variable from period 1 through T (e.g. annual observations on GDP from 1946–2001). A simple manner of allowing for the errors to be correlated is to assume they follow an *autoregressive process of order 1* or *AR(1)* process:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \tag{6.28}$$

where u_t is i.i.d. $N(0, h^{-1})$. This specification allows for the error in one period to depend on that in the previous period.

The time series literature has developed a myriad of tools to aid in a more formal understanding the properties of various time series models and we digress briefly to introduce a few of them here using a general notation, z_t to indicate a time series.[4] In this section, we set $z_t = \varepsilon_t$, but the concepts are generally relevant and will be used in later chapters. It is standard to assume that the process generating the time series has been running from time period $-\infty$ and will run until period ∞. The econometrician observes this process for periods $t = 1, \ldots, T$. z_t is said to be *covariance stationary* if, for every t and s:

$$E(z_t) = E(z_{t-s}) = \mu$$

$$var(z_t) = var(z_{t-s}) = \gamma_0$$

and

$$cov(z_t, z_{t-s}) = \gamma_s$$

where μ, γ_0 and γ_s are all finite. In words, a time series is covariance stationary if it has a constant mean, variance and the covariance between any two observations depends only upon the number of periods apart they are. Many time series variables in economics do seem to be stationary or, if not, can be *differenced* to stationarity. The *first difference* of z_t is denoted by Δz_t and is defined by

$$\Delta z_t = z_t - z_{t-1}$$

In a similar fashion, we can define mth order differences for $m > 1$ as

$$\Delta^m z_t = \Delta^{m-1} z_t - \Delta^{m-1} z_{t-1}$$

To understand the economic relevance of differencing, suppose that z_t is the log of the price level, then Δz_t is (approximately) the percentage change in prices which is inflation. $\Delta^2 z_t$ would then be the percentage change in the rate of inflation. Any or all of these might be important in a macroeconomic model.

A common tool for examining the properties of stationary time series variables is γ_s which is referred to as the *autocovariance function*. Closely related is the *autocorrelation function,* which calculates correlations between observations s periods apart (i.e. it is defined as $\frac{\gamma_s}{\gamma_0}$ for $s = 0, \ldots, \infty$). These are both functions of s and it is common to plot either of them to see how they

---

[4] Space precludes a detailed discussion of time series methods. Bauwens, Lubrano and Richard (1999) provide an excellent Bayesian discussion of time series methods and the reader is referred to this book for more detail. Enders (1995) is a fine non-Bayesian book.

change as s increases. For instance, with macroeconomic variables we typically find autocorrelation functions decrease with s since recent happenings have more impact on current macroeconomic conditions than things that happened long ago.

Let us now return to the AR(1) process for the errors given in (6.28). To figure out its properties it is convenient to write ε_t in terms of u_{t-s} for $s = 0, \ldots, \infty$. This can be done by noting $\varepsilon_{t-1} = \rho\varepsilon_{t-2} + u_{t-1}$ and substituting this expression into (6.28), yielding

$$\varepsilon_t = \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t$$

If we then substitute in the expression for ε_{t-2} we obtain an equation involving ε_{t-3} which we can substitute in for. Successively substituting in expressions for ε_{t-s} in this manner, (6.28) can be written as

$$\varepsilon_t = \sum_{s=0}^{\infty} \rho^s u_{t-s} \tag{6.29}$$

Written in this form, you can see that problems will occur if you try and calculate the mean, variance and covariance of ε_t since ρ^s will become infinite if $|\rho| > 1$. Even if $\rho = 1$, such calculations will involve infinite sums of finite terms. In fact, $|\rho| < 1$ is required for the time series to be stationary.

If we impose $|\rho| < 1$ it can be confirmed that $E(\varepsilon_t) = 0$

$$\gamma_0 = var(\varepsilon_t) = h^{-1}\sum_{s=0}^{\infty} \rho^{2s} = \frac{1}{h(1-\rho^2)}$$

and

$$\gamma_s = cov(\varepsilon_t, \varepsilon_{t-s}) = \frac{\rho^s}{h(1-\rho^2)}$$

Note that, since $|\rho| < 1$, the autocovariance function γ_s declines as s increases. Intuitively, with an AR(1) process, the influence of the past gradually dies away.

These results can be used to write the covariance matrix of ε as $h^{-1}\Omega$, where

$$\Omega = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdot & \rho^{T-1} \\ \rho & 1 & \rho & \cdot & \cdot \\ \rho^2 & \rho & \cdot & \cdot & \rho^2 \\ \cdot & \cdot & \cdot & \cdot & \rho \\ \rho^{T-1} & \cdot & \rho^2 & \rho & 1 \end{bmatrix} \tag{6.30}$$

The AR(1) model can be extended to include more past time periods or *lags*. We can define the autoregressive process of order p or *AR(p) process* as

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \cdots + \rho_p\varepsilon_{t-p} + u_t \tag{6.31}$$

and methods similar to those above can be used to calculate the mean, variance and autocovariance function. As will be shown in the next section, we do not need to know the exact form of the autocovariance function in order to do

Bayesian inference with the AR(p) process. Hence, we do not write it out here. The interested reader is referred to a time series book such as Enders (1995) for detail. Suffice it to note here that the AR(p) process has similar properties to the AR(1), but is more flexible.

This is a convenient place to introduce some more time series notation. The *lag operator* is denoted by L and has the property that $L\varepsilon_t = \varepsilon_{t-1}$ or, more generally, $L^m\varepsilon_t = \varepsilon_{t-m}$. The AR(p) model can thus be written as

$$(1 - \rho_1 L - \cdots - \rho_p L^p)\ \varepsilon_t = u_t$$

or

$$\rho(L)\varepsilon_t = u_t$$

where $\rho(L) = (1 - \rho_1 L - \cdots - \rho_p L^p)$ is a polynomial of order p in the lag operator. It can be verified that an AR(p) process is stationary if the roots of the equation $\rho(z) = 0$ are all greater than one in absolute value. For future reference, define $\rho = (\rho_1, \ldots, \rho_p)'$ and let Φ denote the stationary region for this model.

6.5.2 Bayesian Computation

A posterior simulator which allows for Bayesian inference in the Normal linear regression model with AR(p) errors can be developed by adapting the formulae for the general case with Ω unspecified given in (6.12), (6.15) and (6.18). If we make one approximation, these posterior conditionals assume a simple form. This approximation involves the treatment of the initial conditions. To understand what is meant by this statement, consider how we would transform the model as in (6.2). We can do this by working out the form of Ω when AR(p) errors are present and then deriving the matrix P such that $P\Omega P' = I$. Alternatively, let us write the regression model as

$$y_t = x_t'\beta + \varepsilon_t \tag{6.32}$$

where $x_t = (1, x_{t2}, \ldots, x_{tk})'$. Multiplying both sides of (6.32) by $\rho(L)$ and defining $y_t^* = \rho(L)y_t$ and $x_t^* = \rho(L)x_t$ we obtain

$$y_t^* = x_t^{*\prime}\beta + u_t \tag{6.33}$$

We have assumed that u_t is i.i.d. $N(0, h^{-1})$ and, thus, the transformed model given in (6.33) is simply a Normal linear regression model with i.i.d. errors. Note, however, what happens to this transformation for values of $t \le p$. y_1^*, for instance, depends upon $y_0, \ldots, y_{1-p}$. Since our data runs from $t = 1, \ldots, T$, these so-called initial conditions, $y_0, \ldots, y_{1-p}$, are not observed. The treatment of initial conditions is a subtle issue, especially if the AR process is non-stationary or nearly so. The interested reader is referred to Bauwens, Lubrano and Richard (1999) or Schotman (1994) for more detail. Here, we will assume stationarity of errors, so the treatment of initial conditions is of less importance. Accordingly, we will follow a common practice and work with the likelihood function based on data from $t = p + 1, \ldots, T$ rather than $t = 1, \ldots, T$. Provided p is small relative to T, this will result in an approximate likelihood which is quite close to the true likelihood. Since y_t^* and x_t^* for $t = p + 1, \ldots, T$ do not depend upon unobserved lagged values, the transformation given in (6.33) can be done in a straightforward fashion.

To keep the notation as simple as possible, we will not introduce a new notation for the likelihood, posterior, etc. for data from $t = p + 1, \ldots, T$. Instead, for the remainder of this section, we will simply interpret y, y^*, ε and ε^* as $(T - p)$-vectors (i.e. the first p elements have been removed). X and X^* will be $(T - p) \times k$ matrices. With these changes, a Gibbs sampling algorithm can be derived in a straightforward fashion using previous results. Intuitively, $p(\beta|y, h, \rho)$ and $p(h|y, \beta, \rho)$ are given in (6.12) and (6.15). $p(\rho|y, \beta, h)$ can be derived by noting that, conditional on β and h, ε_t for $t = p + 1, \ldots, T$ is known and (6.31) is simply a Normal linear regression model (with known error variance) with coefficients given by ρ. Thus, standard Bayesian results from previous chapters can be used to derive $p(\rho|y, \beta, h)$.

Formally, using the independent Normal-Gamma prior for β and h given in (6.9) and (6.10), the results of Section 6.2 can be modified to the present case as

$$\beta|y, h, \rho \sim N(\overline{\beta}, \overline{V}) \tag{6.34}$$

where

$$\overline{V} = (\underline{V}^{-1} + hX^{*\prime}X^*)^{-1} \tag{6.35}$$

and

$$\overline{\beta} = \overline{V}(\underline{V}^{-1}\underline{\beta} + hX^{*\prime}y^*) \tag{6.36}$$

The posterior for h conditional on the other parameters in the model is Gamma:

$$h|y, \beta, \rho \sim G(\overline{s}^{-2}, \overline{\nu}) \tag{6.37}$$

where

$$\overline{\nu} = T - p + \underline{\nu} \tag{6.38}$$

and

$$\overline{s}^2 = \frac{(y^* - X^*\beta)'(y^* - X^*\beta) + \underline{\nu}\,\underline{s}^2}{\overline{\nu}} \tag{6.39}$$

The posterior for ρ depends upon its prior which, of course, can be anything which reflects the researcher's non-data information. Here we assume it is multivariate Normal, truncated to the stationary region. That is,

$$p(\rho) \propto f_N(\rho|\underline{\rho}, \underline{V}_\rho)1(\rho \in \Phi) \tag{6.40}$$

where $1(\rho \in \Phi)$ is the indicator function which equals 1 for the stationary region and zero otherwise. With this prior, it is straightforward to derive

$$p(\rho|y, \beta, h) \propto f_N(\rho|\overline{\rho}, \overline{V}_\rho)1(\rho \in \Phi) \tag{6.41}$$

where

$$\overline{V}_\rho = (\underline{V}_\rho^{-1} + hE'E)^{-1} \tag{6.42}$$

$$\overline{\rho} = \overline{V}\rho(\underline{V}_\rho^{-1}\underline{\rho} + hE'\varepsilon) \tag{6.43}$$

and E is a $(T - p) \times k$ matrix with tth row given by $(\varepsilon_{t-1}, \ldots, \varepsilon_{t-p})$.

The Gibbs sampler involves sequentially drawing from (6.34), (6.37) and (6.41). The fact that (6.41) is truncated multivariate Normal, rather than simply multivariate Normal adds a slight complication. However, drawing from the truncated multivariate Normal distribution can be done by drawing from the untruncated variant and simply discarding the draws which fall outside the stationary region. Provided $\overline{\rho}$ lies within (or not too far outside) the stationary region, this strategy should work well. Alternatively, a Metropolis–Hastings algorithm can be derived or the methods of Geweke (1991) for drawing from the truncated multivariate Normal can be used. Predictive inference in this model can be carried out using the strategy outlined in Chapter 4, Section 4.2.6. Posterior predictive p-values or HPDIs can be calculated to shed light on the fit and appropriateness of the model. Bayes factors for any hypothesis of interest can be calculated using either the Savage–Dickey density ratio or the Gelfand–Dey approach. The fact that (6.41) provides only the kernel of $p(\rho|y, \beta, h)$ makes the use of the Savage–Dickey density ratio a little more complicated. Remember (see Chapter 4, Section 4.2.5) that the Savage–Dickey density ratio requires you to know the complete densities (not just the kernel), $p(\rho|y, \beta, h)$ or $p(\rho|y)$. For $p = 1$, the integrating constant can be easily calculated since $p(\rho|y, \beta, h)$ is a univariate truncated Normal and the properties of this univariate density are well known (see Poirier, 1995, p. 115). However, for $p > 1$ the stationary region is nonlinear and $p(\rho|y, \beta, h)$ is harder to work with analytically. Nevertheless, it is straightforward to calculate the necessary integrating constant through posterior simulation. That is, the density corresponding to (6.41) is

$$p(\rho|y, \beta, h) = \frac{f_N(\rho|\overline{\rho}, \overline{V}_\rho)1(\rho \in \Phi)}{\int_\Phi f_N(\rho|\overline{\rho}, \overline{V}_\rho)d\rho}$$

A common posterior simulator involves drawing from $f_N(\rho|\overline{\rho}, \overline{V}_\rho)$ and discarding draws outside the stationary region. But, $\int_\Phi f_N(\rho|\overline{\rho}, \overline{V}_\rho)d\rho$ is simply the proportion of draws retained. This can be estimated by, at every pass through the Gibbs sampler, calculating the number of rejected draws before an acceptable one is found. $1 - \int_\Phi f_N(\rho|\overline{\rho}, \overline{V}_\rho)d\rho$ is approximated by the number of rejected draws divided by the number of rejected draws plus one. As the number of Gibbs replications goes to infinity, the approximation error will go to zero. In general, the integrating constant of any truncated density can always be found by drawing from its untruncated counterpart and calculating the proportion of draws within the truncated region. Depending on the prior used, such a strategy may be necessary for calculating its integrating constant.

6.5.3 Empirical Illustration: The Normal Regression Model with Autocorrelated Errors

To illustrate Bayesian inference in the Normal regression model with autocorrelated errors, we use a data set pertaining to baseball. The dependent variable is the winning percentage of the New York Yankees baseball team every year between 1903 and 1999. Interest centers on explaining the Yankees' performance using various measures of team offensive and defensive performance. Thus

- y_t = winning percentage (PCT) in year t = wins/(wins + losses),
- x_{t2} = team on-base percentage (OBP) in year t,
- x_{t3} = team slugging average (SLG) in year t,
- x_{t4} = team earned run average (ERA) in year t.

A knowledge of baseball is not necessary to understand this empirical example. You need only note that the explanatory variables are all measures of team performance. We would expect x_{t2} and x_{t3} to be positively associated with winning percentage while x_{t4} should exhibit a negative association. Despite the prior information revealed in the previous sentence, we use a noninformative prior for β and set $\underline{V}^{-1} = 0_{k\times k}$. We also use a noninformative prior for the error precision and set $\underline{\nu} = 0$. With these choices, the values of $\underline{\beta}$ and $\underline{s}^{-2}$ are irrelevant. We use the technique described in the previous subsection to calculate the Savage–Dickey density ratios comparing models with $\rho_j = 0$ for $j = 1, \ldots, p$ to unrestricted models. This requires an informative prior for ρ and, thus, we set $\underline{\rho} = 0$ and $\underline{V}_\rho = cI_p$. Various values of c are chosen below in a prior sensitivity analysis. Throughout, we set $p = 1$. In preliminary runs with larger values of p, Bayes factors and HPDIs provided no evidence for autocorrelation of an order higher than one. To help provide intuition, note that the stationarity condition with $p = 1$ implies $|\rho_1| < 1$ and values of ρ_1 near one can be considered as implying a large degree of autocorrelation.

All results are based on 30 000 replications, with 5000 burn-in replications discarded and 25 000 replications retained. MCMC diagnostics indicate convergence of the Gibbs sampler, and numerical standard errors indicate an approximation error which is small relative to posterior standard deviations of all parameters.

Table 6.3 presents posterior results for β with $c = 0.09$, a reasonably small value reflecting a prior belief that autocorrelation in the errors is fairly small (i.e. the prior standard deviation of ρ_1 is 0.3). It can be seen that the results are as expected in that OBP and SLG are positive and ERA is negatively associated with winning.

At the beginning of the book, we emphasized the importance of doing prior sensitivity analysis. For the sake of space, our previous empirical illustrations did not include any investigation of prior sensitivity. However, we will do one here with regards to the AR(1) coefficient. Table 6.4 contains results from a prior sensitivity analysis where various values of c are used. This table reveals that prior information has little affect on the posterior, unless prior information is

Table 6.3 Posterior Results for β

	Mean	Standard Deviation	95% HPDI
β_1	0.01	0.07	$[-0.11, 0.12]$
β_2	1.09	0.35	$[0.52, 1.66]$
β_3	1.54	0.18	$[1.24, 1.83]$
β_4	-0.12	0.01	$[-0.13, -0.10]$

Table 6.4 Posterior Results for ρ_1

	Mean	Standard Deviation	95% HPDI	Bayes Factor for $\rho_1 = 0$
$c = 0.01$	0.10	0.07	$[-0.02, 0.23]$	0.49
$c = 0.09$	0.20	0.10	$[0.03, 0.36]$	0.43
$c = 0.25$	0.21	0.11	$[0.04, 0.39]$	0.56
$c = 1.0$	0.22	0.11	$[0.05, 0.40]$	0.74
$c = 100$	0.22	0.11	$[0.05, 0.40]$	0.84

extremely strong as in the $c = 0.01$ case. This can be seen by noting that posterior means, standard deviations and HPDIs are almost the same for all values of c between 0.09 and 100. The latter is a very large value which, to all intents and purposes, implies that the prior is flat and noninformative over the stationary region. The Bayes factors are also fairly robust to changes in the prior. As an aside, it is worth noting that this robustness of the Bayes factor is partly to do with the fact that the prior is truncated to a bounded interval – the stationary region. Don't forget the problems that can occur with Bayes factors when you use noninformative improper priors on parameters whose support is unbounded (e.g. see Chapter 3, Section 3.6.2).

6.6 THE SEEMINGLY UNRELATED REGRESSIONS MODEL

6.6.1 Introduction

The final model considered in this chapter is the Seemingly Unrelated Regressions (SUR) model. It is a multiple equation model which is both interesting in and of itself and is a component of other common models. In economics, multiple equation models arise in many contexts. For instance, in a study of consumption, the researcher may wish to estimate an equation for each category of consumption (i.e. food, consumer durables, non-durables, etc.). In a microeconomic application, the researcher may wish to estimate a factor demand equation for each factor of production.[5] In many cases, simply working with one equation at a time using the techniques of previous chapters will not lead the researcher too far wrong. However, working with all the equations together can improve estimation. This section discusses how to do so.

The SUR model can be written as

$$y_{mi} = x'_{mi}\beta_m + \varepsilon_{mi} \tag{6.44}$$

with $i = 1, \ldots, N$ observations for $m = 1, \ldots, M$ equations. y_{mi} is the ith observation on the dependent variable in equation m, x_{mi} is a k_m-vector containing the ith observation of the vector of explanatory variables in the mth equation and β_m is a k_m-vector of regression coefficients for the mth equation.[6] Note that this framework allows for the number of explanatory variables to differ across equations, but some or all of them may be the same in different equations.

We can put the SUR model in a familiar form. To do this we stack all equations into vectors/matrices as $y_i = (y_{1i}, \ldots, y_{Mi})'$, $\varepsilon_i = (\varepsilon_{1i}, \ldots, \varepsilon_{Mi})'$

$$\beta = \begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \beta_M \end{pmatrix}$$

$$X_i = \begin{pmatrix} x'_{1i} & 0 & \cdot & \cdot & 0 \\ 0 & x'_{2i} & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & x'_{Mi} \end{pmatrix}$$

and define $k = \sum_{m=1}^{M} k_m$. Using this notation, it can be verified that (6.44) can be written as

$$y_i = X_i\beta + \varepsilon_i \tag{6.45}$$

We now stack all the observations together as

$$y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{pmatrix}$$

---

[5] For the reader with additional knowledge of econometrics, the reduced form of a simultaneous equations model is in the form of a SUR model. Similarly, a Vector Autoregression or VAR is also a SUR model (see Chapter 12, Section 12.4).

[6] Note that we have slightly changed notation from that used previously. In this section, x_{mi} is a vector and the first subscript indicates the equation number. Previously, x_{ij} was a scalar indicating the ith observation on the jth explanatory variable.

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \varepsilon_N \end{pmatrix}$$

$$X = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ X_N \end{pmatrix}$$

and write

$$y = X\beta + \varepsilon$$

Thus, the SUR model can be written as our familiar linear regression model.

So far we have said nothing about error properties of this model. If we were to assume ε_{mi} to be i.i.d. $N(0, h^{-1})$ for all i and m, then we would simply have the Normal linear regression model of Chapters 2, 3 and 4. However, in many applications, it is common for the errors to be correlated across observations and, thus, we assume ε_i to be i.i.d. $N(0, H^{-1})$ for $i = 1, \ldots, N$ where H is an $M \times M$ error precision matrix. With this assumption it can be seen that ε is $N(0, \Omega)$ where Ω is an $NM \times NM$ block-diagonal matrix given by

$$\Omega = \begin{pmatrix} H^{-1} & 0 & \cdot & \cdot & 0 \\ 0 & H^{-1} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & H^{-1} \end{pmatrix} \tag{6.46}$$

Hence, the SUR model lies in the class of models being studied in this chapter and the prior, likelihood and posterior have been discussed in Section 6.2. One minor issue you may have noticed is that there is no h in this model. This is not a substantive difference in that h was merely a scalar that was factored out for convenience in the previous sections. In this model, it is not convenient to factor out a scalar in this way (although we could have done this if we had wanted to).

6.6.2 The Prior

It is worthwhile to briefly discuss prior elicitation in the SUR model as this is a topic which has received a great deal of attention in the literature. In this section, we will use an extended version of our familiar independent Normal-Gamma prior, the independent Normal-Wishart prior:

$$p(\beta, H) = p(\beta)p(H)$$

where

$$p(\beta) = f_N(\beta|\underline{\beta}, \underline{V}) \tag{6.47}$$

and

$$p(H) = f_W(H|\underline{\nu}, \underline{H}) \tag{6.48}$$

The Wishart distribution, which is a matrix generalization of the Gamma distribution, is defined in Appendix B, Definition B.27. For prior elicitation, the most important things to note are that $E(H) = \underline{\nu H}$ and that noninformativeness is achieved by setting $\underline{\nu} = 0$ and $\underline{H}^{-1} = 0_{M\times M}$ (see Appendix B, Theorem B.16).

However, many other priors have been suggested for this model. In particular, a Normal-Wishart natural conjugate prior exists for this model analogous to that used in Chapter 3. This prior has the advantage that analytical results are available so that a posterior simulator is not required. However, the natural conjugate prior for the SUR model has been found by many to be too restrictive. For instance, it implies that the prior covariances between coefficients in each pair of equations (i.e. β_m and β_j for $j \neq m$) are all proportional to the same matrix. For this reason, only the noninformative variant of the natural conjugate prior has received much attention in empirical work. Furthermore, there have been various attempts to derive extended versions of the natural conjugate prior which are less restrictive. Readers interested in learning more about this area of the literature are referred to Dreze and Richard (1983) or Richard and Steel (1988).

6.6.3 Bayesian Computation

Bayesian computation in this model can be implemented with a Gibbs sampler using (6.12) and (6.18) based on the prior given in (6.47) and (6.48). However, both of these posterior conditionals involving inverting the $NM \times NM$ matrix Ω, which is computationally difficult. However, the block-diagonal structure of Ω allows the matrix inversion to be partly done analytically. If we do this, $p(\beta|y, H)$ and $p(H|y, \beta)$ take convenient forms. In particular,

$$\beta|y, H \sim N(\overline{\beta}, \overline{V}) \tag{6.49}$$

where

$$\overline{V} = \left(\underline{V}^{-1} + \sum_{i=1}^{N} X_i{}'HX_i\right)^{-1} \tag{6.50}$$

and

$$\overline{\beta} = \overline{V}\left(\underline{V}^{-1}\underline{\beta} + \sum_{i=1}^{N} X_i'Hy_i\right) \tag{6.51}$$

The posterior for H conditional on β is Wishart:

$$H|y, \beta \sim W(\overline{\nu}, \overline{H}) \tag{6.52}$$

where

$$\overline{\nu} = N + \underline{\nu} \tag{6.53}$$

and

$$\overline{H} = \left[\underline{H}^{-1} + \sum_{i=1}^{N} (y_i - X_i\beta)(y_i - X_i\beta)' \right]^{-1} \tag{6.54}$$

Since random number generators for the Wishart distribution are available (e.g. a MATLAB variant is available in James LeSage's Econometrics Toolbox), a Gibbs sampler which successively draws from $p(\beta|y, H)$ and $p(H|y, \beta)$ can easily be developed.

Predictive inference in this model can be carried out using the strategy outlined in Chapter 4, Section 4.2.6. Posterior predictive p-values or HPDIs can be calculated to shed light on the fit and appropriateness of the model. The Savage–Dickey density ratio is particularly easy to calculate should you wish to calculate posterior odds ratios.

6.6.4 Empirical Illustration: The Seemingly Unrelated Regressions Model

To illustrate Bayesian inference in the SUR model we use an extended version of the baseball data set used in the autocorrelated errors example. In that example, we chose one baseball team, the Yankees, and investigated how team winning percentage (PCT) depended upon team on-base percentage (OBP), slugging average (SLG) and earned run average (ERA). The former two of these explanatory variables are measures of offensive performance, the last defensive performance. In the current example, we add a second equation for a second team, the Boston Red Sox (the arch-rivals of the Yankees). Hence, we have two equations, one for each team, with explanatory variables in each equation being the relevant team's OBP, SLG and ERA. Section 6.5.3 provides further detail about the data.

We use a noninformative prior for H and set $\underline{\nu} = 0$ and $\underline{H}^{-1} = 0_{2\times2}$. For the regression coefficients, we set $\underline{\beta} = 0_k$ and $\underline{V} = 4I_k$. This prior reflects relatively noninformative prior beliefs. That is the regression coefficients are all centered over points which imply the explanatory variable has no effect on the dependent variable. But each coefficient has prior standard deviation of 2, a value which allows for the explanatory variables to have quite large impacts on the dependent variable.

Table 6.5 presents posterior results obtained from 30 000 replications from the Gibbs sampler outlined above, with 5000 burn-in replications discarded and 25 000 replications retained. MCMC diagnostics indicate convergence of the Gibbs sampler and numerical standard errors indicate an approximation error which is small relative to the posterior standard deviations of all parameters. Instead of presenting posterior results for H, which may be hard to interpret, we focus on the correlation between the errors in the two equations (i.e. corr($\varepsilon_{1i}, \varepsilon_{2i}$) which is assumed to be the same for all $i = 1, \ldots, N$). If this correlation is equal to zero, then there is no benefit to using the SUR model over simply doing posterior inference on each equation separately. As we have emphasized throughout this book (see, e.g., Chapter 1, Section 1.2 or Chapter 3, Section 3.8), posterior

Table 6.5 Posterior Results for β and Error Correlation

	Mean	Standard Deviation	95% HPDI
		Yankees Equation	
β_1	0.03	0.06	$[-0.06, 0.13]$
β_2	0.92	0.30	$[0.43, 1.41]$
β_3	1.61	0.15	$[1.36, 1.86]$
β_4	−0.12	0.01	$[-0.13, -0.10]$
		Red Sox Equation	
β_5	−0.15	0.06	$[-0.26, -0.05]$
β_6	1.86	0.28	$[1.41, 2.32]$
β_7	1.24	0.15	$[0.99, 1.50]$
β_8	−0.11	0.01	$[-0.12, -0.10]$
		Cross-equation Error Correlation	
$\mathrm{corr}(\varepsilon_1, \varepsilon_2)$	−0.01	0.11	$[-0.18, 0.17]$

simulator output can be used to do posterior inference on any function of the parameters of the model. Hence, the Gibbs draws of H can be used to derive posterior properties of $\mathrm{corr}(\varepsilon_{1i}, \varepsilon_{2i})$. It can be seen that, with this data set, the correlation between the errors in the two equations is very near to zero. Thus, there is minimal benefit to working with the SUR model. If we had used an informative prior for H, we could have calculated a Bayes factor using the Savage–Dickey density ratio. This Bayes factor would have provided more formal evidence in favor of the hypothesis that the errors in the two equations are uncorrelated.

The regression coefficients measure the impacts of OBP, SLG and ERA on team performance. For both teams, results are sensible, indicating that higher OBP and SLG and lower ERA are associated with a higher team winning percentage. A baseball enthusiast might be interested in whether these coefficients are different in the two equations. After all, baseball wisdom has it that in some stadiums it is important to have power hitters, in others pitching is a relatively important key to success, etc. An examination of Table 6.5 indicates that, with one exception, the posterior means of comparable regression coefficients are roughly the same across equations, relative to their standard deviations. Furthermore, 95% HPDIs for comparable coefficients in different equations exhibit a large degree of overlap. The one exception is OBP where β_2 and β_6 are quite different from one another. The question of whether comparable coefficients are different in the two equations can be formally addressed by calculating Bayes factors comparing: $M_1 : \beta_j - \beta_{k_1+j} = 0$ for $j = 1, \ldots, k_1$ against M_2 where the coefficients are left unrestricted. This can be calculated using the Savage–Dickey density ratio implemented as outlined in Chapter 4, Section 4.2.5. Since the prior and conditional posterior of β are both Normal (see (6.47) and (6.49)), the prior