# 16
# BAYESIAN ESTIMATION AND INFERENCE

## 16.1 INTRODUCTION

The preceding chapters (and those that follow this one) are focused primarily on parametric specifications and classical estimation methods. These elements of the econometric method present a bit of a methodological dilemma for the researcher. They appear to straightjacket the analyst into a fixed and immutable specification of the model. But in any analysis, there is uncertainty as to the magnitudes, sometimes the signs and, at the extreme, even the meaning of parameters. It is rare that the presentation of a set of empirical results has not been preceded by at least some exploratory analysis. Proponents of the Bayesian methodology argue that the process of *estimation* is not one of deducing the values of fixed parameters, but rather, in accordance with the scientific method, one of continually updating and sharpening our subjective beliefs about the state of the world. Of course, this adherence to a subjective approach to model building is not necessarily a virtue. If one holds that *models* and *parameters* represent objective truths that the analyst seeks to discover, then the subjectivity of Bayesian methods may be less than perfectly comfortable.

Contemporary applications of Bayesian methods typically advance little of this theological debate. The modern practice of Bayesian econometrics is much more pragmatic. As we will see in several of the following examples, Bayesian methods have produced some remarkably efficient solutions to difficult estimation problems. Researchers often choose the techniques on practical grounds, rather than in adherence to their philosophical basis; indeed, for some, the Bayesian estimator is merely an algorithm.[1]

Bayesian methods have have been employed by econometricians since well before Zellner's classic (1971) presentation of the methodology to economists, but until fairly recently, were more or less at the margin of the field. With recent advances in technique (notably the Gibbs sampler) and the advance of computer software and hardware that has made simulation-based estimation routine, Bayesian methods that rely heavily on both have become widespread throughout the social sciences. There are libraries of work on Bayesian econometrics, a rapidly expanding applied literature.[2] This chapter will introduce the vocabulary and techniques of Bayesian econometrics. Section 16.2

[1]For example, the Website of MLWin, a widely used program for random parameters modeling, www.bristol.ac.uk/cmm/software/mlwin/features/mcmc.html, states that their use of diffuse priors for Bayesian models produces approximations to maximum likelihood estimators. Train (2001) is an interesting application that compares Bayesian and classical estimators of a random parameters model. Another comparison appears in Example 16.7 below.

[2]Recent additions to the dozens of books on the subject include Gelman et al. (2004), Geweke (2005), Gill (2002), Koop (2003), Lancaster (2004), Congdon (2005), and Rossi et al. (2005). Readers with a historical bent will find Zellner (1971) and Leamer (1978) worthwhile reading. There are also many methodological surveys. Poirier and Tobias (2006) as well as Poirier (1988, 1995) sharply focus the nature of the methodological distinctions between the classical (frequentist) and Bayesian approaches.

lays out the essential foundation for the method. The canonical application, the linear regression model, is developed in Section 16.3. Section 16.4 continues the methodological development. The fundamental tool of contemporary Bayesian econometrics, the Gibbs sampler, is presented in Section 16.5. Three applications and several more limited examples are presented in Sections 16.6 through 16.8. Section 16.6 shows how to use the Gibbs sampler to estimate the parameters of a probit model without maximizing the likelihood function. This application also introduces the technique of data augmentation. Bayesian counterparts to the panel data random and fixed effects models are presented in Section 16.7. A hierarchical Bayesian treatment of the random parameters model is presented in Section 16.8 with a comparison to the classical treatment of the same model. Some conclusions are drawn in Section 16.9. The presentation here is nontechnical. A much more extensive entry-level presentation is given by Lancaster (2004). Intermediate-level presentations appear in Cameron and Trivedi (2005, Chapter 13), and Koop (2003). A more challenging treatment is offered in Geweke (2005). The other sources listed in footnote 2 are oriented to applications.

## 16.2 BAYES' THEOREM AND THE POSTERIOR DENSITY

The centerpiece of the Bayesian methodology is **Bayes' theorem**: for events $A$ and $B$, the conditional probability of event $A$ given that $B$ has occurred is

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{16-1}$$

Paraphrased for our applications here, we would write

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})}.$$

In this setting, the data are viewed as constants whose distributions do not involve the parameters of interest. For the purpose of the study, we treat the data as only a fixed set of additional information to be used in updating our beliefs about the parameters. Note the similarity to (12-1). Thus, we write

$$\begin{aligned} P(\text{parameters} \mid \text{data}) &\propto P(\text{data} \mid \text{parameters})P(\text{parameters}) \\ &= \textbf{Likelihood function} \times \textbf{Prior density}. \end{aligned} \tag{16-2}$$

The symbol $\propto$ means "is proportional to." In the preceding equation, we have dropped the marginal density of the data, so what remains is not a proper density until it is scaled by what will be an inessential proportionality constant. The first term on the right is the joint distribution of the observed random variables $\mathbf{y}$, given the parameters. As we shall analyze it here, this distribution is the normal distribution we have used in our previous analysis—see (12-1). The second term is the **prior beliefs** of the analyst. The left-hand side is the **posterior density** of the parameters, given the current body of data, or our revised beliefs about the distribution of the parameters after seeing the data. The posterior is a mixture of the prior information and the current information, that is, the data. Once obtained, this posterior density is available to be the **prior density** function

when the next body of data or other usable information becomes available. The principle involved, which appears nowhere in the classical analysis, is one of continual accretion of knowledge about the parameters.

Traditional Bayesian estimation is heavily parameterized. The prior density and the likelihood function are crucial elements of the analysis, and both must be fully specified for estimation to proceed. The Bayesian estimator is the mean of the posterior density of the parameters, a quantity that is usually obtained either by integration (when closed forms exist), approximation of integrals by numerical techniques, or by Monte Carlo methods, which are discussed in Section 15.6.2.

### *Example 16.1 Bayesian Estimation of a Probability*

Consider estimation of the probability that a production process will produce a defective product. In case 1, suppose the sampling design is to choose $N = 25$ items from the production line and count the number of defectives. If the probability that any item is defective is a constant $\theta$ between zero and one, then the likelihood for the sample of data is

$$L(\theta \,|\, \textbf{data}) = \theta^D (1 - \theta)^{25 - D},$$

where $D$ is the number of defectives, say, 8. The maximum likelihood estimator of $\theta$ will be $p = D/25 = 0.32$, and the asymptotic variance of the maximum likelihood estimator is estimated by $p(1 - p)/25 = 0.008704$.

Now, consider a Bayesian approach to the same analysis. The posterior density is obtained by the following reasoning:

$$p(\theta \,|\, \textbf{data}) = \frac{p(\theta, \textbf{data})}{p(\textbf{data})} = \frac{p(\theta, \textbf{data})}{\displaystyle\int_\theta p(\theta, \textbf{data}) d\theta} = \frac{p(\textbf{data} \,|\, \theta) p(\theta)}{p(\textbf{data})}$$
$$= \frac{\mathit{Likelihood}(\textbf{data} \,|\, \theta) \times p(\theta)}{p(\textbf{data})},$$

where $p(\theta)$ is the prior density assumed for $\theta$. [We have taken some license with the terminology, because the **likelihood function** is conventionally defined as $L(\theta \,|\, \textbf{data})$.] Inserting the results of the sample first drawn, we have the posterior density,

$$p(\theta \,|\, \textbf{data}) = \frac{\theta^D (1 - \theta)^{N - D} p(\theta)}{\displaystyle\int_\theta \theta^D (1 - \theta)^{N - D} p(\theta) d\theta}.$$

What follows depends on the assumed prior for $\theta$. Suppose we begin with a noninformative prior that treats all *allowable* values of $\theta$ as equally likely. This would imply a uniform distribution over (0,1). Thus, $p(\theta) = 1$, $0 \leq \theta \leq 1$. The denominator with this assumption is a beta integral (see Section E2.3) with parameters $a = D + 1$ and $b = N - D + 1$, so the posterior density is

$$p(\theta \,|\, \textbf{data}) = \frac{\theta^D (1 - \theta)^{N - D}}{\left( \dfrac{\Gamma(D + 1)\Gamma(N - D + 1)}{\Gamma(D + 1 + N - D + 1)} \right)} = \frac{\Gamma(N + 2) \theta^D (1 - \theta)^{N - D}}{\Gamma(D + 1)\Gamma(N - D + 1)}.$$

This is the density of a random variable with a beta distribution with parameters $(\alpha, \beta) = (D + 1, N - D + 1)$. (See Section B.4.6.) The mean of this random variable is $(D + 1)/(N + 2) = 9/27 = 0.3333$ (as opposed to 0.32, the MLE). The posterior variance is $[(D + 1)/(N - D + 1)]/[(N + 3)(N + 2)^2] = 0.007936$ compared to 0.00874 for the MLE.

There is a loose end in this example. If the uniform prior were truly noninformative, that would mean that the only information we had was in the likelihood function. Why didn't the Bayesian estimator and the MLE coincide? The reason is that the uniform prior over [0,1] is not really noninformative. It did introduce the information that $\theta$ must fall in the unit interval. The prior mean is 0.5 and the prior variance is 1/12. The posterior mean is an average of the MLE and the prior mean. Another less than obvious aspect of this result is the smaller variance of the Bayesian estimator. The principle that lies behind this (aside from the fact that the prior did in fact introduce some certainty in the estimator) is that the Bayesian estimator is conditioned on the specific sample data. The theory behind the classical MLE implies that it averages over the entire population that generates the data. This will always introduce a greater degree of uncertainty in the classical estimator compared to its Bayesian counterpart.

## 16.3 BAYESIAN ANALYSIS OF THE CLASSICAL REGRESSION MODEL

The complexity of the algebra involved in Bayesian analysis is often extremely burdensome. For the linear regression model, however, many fairly straightforward results have been obtained. To provide some of the flavor of the techniques, we present the full derivation only for some simple cases. In the interest of brevity, and to avoid the burden of excessive algebra, we refer the reader to one of the several sources that present the full derivation of the more complex cases.[3]

The classical normal regression model we have analyzed thus far is constructed around the conditional multivariate normal distribution $N[\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}]$. The interpretation is different here. In the sampling theory setting, this distribution embodies the information about the observed sample data given the assumed distribution and the fixed, albeit unknown, parameters of the model. In the Bayesian setting, this function summarizes the information that a particular realization of the data provides about the assumed distribution of the model parameters. To underscore that idea, we rename this joint density the *likelihood for* $\boldsymbol{\beta}$ *and* $\sigma^2$ *given the data*, so

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{-[(1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}. \tag{16-3}$$

For purposes of the following results, some reformulation is useful. Let $d = n - K$ (the degrees of freedom parameter), and substitute

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b})$$

in the exponent. Expanding this produces

$$\left(-\frac{1}{2\sigma^2}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \left(-\frac{1}{2}ds^2\right)\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)(\boldsymbol{\beta} - \mathbf{b}).$$

After a bit of manipulation (note that $n/2 = d/2 + K/2$), the likelihood may be written

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = [2\pi]^{-d/2}[\sigma^2]^{-d/2} e^{-(d/2)(s^2/\sigma^2)} [2\pi]^{-K/2}[\sigma^2]^{-K/2} e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}.$$

This density embodies all that we have to learn about the parameters from the observed data. Because the data are taken to be constants in the joint density, we may multiply

[3]These sources include Judge et al. (1982, 1985), Maddala (1977a), Mittelhammer et al. (2000), and the canonical reference for econometricians, Zellner (1971). A remarkable feature of the current literature is the degree to which the analytical components have become ever simpler while the applications have become progressively more complex. This will become evident in Sections 16.5–16.7.

this joint density by the (very carefully chosen), inessential (because it does not involve $\boldsymbol{\beta}$ or $\sigma^2$) constant function of the observations,

$$A = \frac{\left(\frac{d}{2}s^2\right)^{(d/2)+1}}{\Gamma\left(\frac{d}{2}+1\right)}[2\pi]^{(d/2)}|\mathbf{X}'\mathbf{X}|^{-1/2}.$$

For convenience, let $v = d/2$. Then, multiplying $L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$ by $A$ gives

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \frac{[vs^2]^{v+1}}{\Gamma(v+1)}\left(\frac{1}{\sigma^2}\right)^{v} e^{-vs^2(1/\sigma^2)}[2\pi]^{-K/2}|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \times e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}. \quad \textbf{(16-4)}$$

The likelihood function is proportional to the product of a gamma density for $z = 1/\sigma^2$ with parameters $\lambda = vs^2$ and $P = v + 1$ [see (B-39); this is an **inverted gamma distribution**] and a $K$-variate normal density for $\boldsymbol{\beta}|\sigma^2$ with mean vector $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The reason will be clear shortly.

### 16.3.1 ANALYSIS WITH A NONINFORMATIVE PRIOR

The departure point for the Bayesian analysis of the model is the specification of a **prior distribution**. This distribution gives the analyst's prior beliefs about the parameters of the model. One of two approaches is generally taken. If no prior information is known about the parameters, then we can specify a **noninformative prior** that reflects that. We do this by specifying a flat prior for the parameter in question:[4]

$$g(\text{parameter}) \propto \text{constant}.$$

There are different ways that one might characterize the lack of prior information. The implication of a flat prior is that within the range of valid values for the parameter, all intervals of equal length—hence, in principle, all values—are equally likely. The second possibility, an **informative prior**, is treated in the next section. The posterior density is the result of combining the likelihood function with the prior density. Because it pools the full set of information available to the analyst, once the data have been drawn, the posterior density would be interpreted the same way the prior density was before the data were obtained.

To begin, we analyze the case in which $\sigma^2$ is assumed to be known. This assumption is obviously unrealistic, and we do so only to establish a point of departure. Using Bayes' theorem, we construct the posterior density,

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2) = \frac{L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})g(\boldsymbol{\beta}|\sigma^2)}{f(\mathbf{y})} \propto L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})g(\boldsymbol{\beta}|\sigma^2),$$

assuming that the distribution of $\mathbf{X}$ does not depend on $\boldsymbol{\beta}$ or $\sigma^2$. Because $g(\boldsymbol{\beta}|\sigma^2) \propto$ a constant, this density is the one in (16-4). For now, write

$$f(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X}) \propto h(\sigma^2)[2\pi]^{-K/2}|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}, \quad \textbf{(16-5)}$$

[4]That this *improper* density might not integrate to one is only a minor difficulty. Any constant of integration would ultimately drop out of the final result. See Zellner (1971, pp. 41–53) for a discussion of noninformative priors.

where

$$h(\sigma^2) = \frac{[\nu s^2]^{\nu+1}}{\Gamma(\nu+1)}\left[\frac{1}{\sigma^2}\right]^{\nu} e^{-\nu s^2(1/\sigma^2)}. \tag{16-6}$$

For the present, we treat $h(\sigma^2)$ simply as a constant that involves $\sigma^2$, not as a probability density; (16-5) is conditional on $\sigma^2$. Thus, the posterior density $f(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})$ is proportional to a multivariate normal distribution with mean $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

This result is familiar, but it is interpreted differently in this setting. First, we have combined our prior information about $\boldsymbol{\beta}$ (in this case, no information) and the sample information to obtain a posterior distribution. Thus, on the basis of the sample data in hand, we obtain a distribution for $\boldsymbol{\beta}$ with mean $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The result is dominated by the sample information, as it should be if there is no prior information. In the absence of any prior information, the mean of the posterior distribution, which is a type of Bayesian point estimate, is the sampling theory estimator, $\mathbf{b}$.

To generalize the preceding to an unknown $\sigma^2$, we specify a noninformative prior distribution for $\ln \sigma$ over the entire real line.[5] By the change of variable formula, if $g(\ln \sigma)$ is constant, then $g(\sigma^2)$ is proportional to $1/\sigma^2$.[6] Assuming that $\boldsymbol{\beta}$ and $\sigma^2$ are independent, we now have the noninformative joint prior distribution,

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}}(\boldsymbol{\beta}) g_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}.$$

We can obtain the **joint posterior distribution** for $\boldsymbol{\beta}$ and $\sigma^2$ by using

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X}) g_{\sigma^2}(\sigma^2) \propto L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X}) \times \frac{1}{\sigma^2}. \tag{16-7}$$

For the same reason as before, we multiply $g_{\sigma^2}(\sigma^2)$ by a well-chosen constant, this time $\nu s^2\Gamma(\nu+1)/\Gamma(\nu+2) = \nu s^2/(\nu+1)$. Multiplying (16-5) by this constant times $g_{\sigma^2}(\sigma^2)$ and inserting $h(\sigma^2)$ gives the joint posterior for $\boldsymbol{\beta}$ and $\sigma^2$, given $\mathbf{y}$ and $\mathbf{X}$,

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \frac{[\nu s^2]^{\nu+2}}{\Gamma(\nu+2)}\left[\frac{1}{\sigma^2}\right]^{\nu+1} e^{-\nu s^2(1/\sigma^2)}[2\pi]^{-K/2}|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \times e^{-(1/2)(\boldsymbol{\beta}-b)'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}.$$

To obtain the marginal posterior distribution for $\boldsymbol{\beta}$, it is now necessary to integrate $\sigma^2$ out of the joint distribution (and vice versa to obtain the marginal distribution for $\sigma^2$). By collecting the terms, $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$ can be written as

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto A \times \left(\frac{1}{\sigma^2}\right)^{P-1} e^{-\lambda(1/\sigma^2)},$$

[5]See Zellner (1971) for justification of this prior distribution.

[6]Many treatments of this model use $\sigma$ rather than $\sigma^2$ as the parameter of interest. The end results are identical. We have chosen this parameterization because it makes manipulation of the likelihood function with a gamma prior distribution especially convenient. See Zellner (1971, pp. 44–45) for discussion.

where

$$A = \frac{[vs^2]^{v+2}}{\Gamma(v+2)}[2\pi]^{-K/2}|(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2},$$
$$P = v + 2 + K/2 = (n-K)/2 + 2 + K/2 = (n+4)/2,$$

and

$$\lambda = vs^2 + \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}).$$

The marginal posterior distribution for $\boldsymbol{\beta}$ is

$$\int_0^\infty f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})d\sigma^2 \propto A\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{P-1} e^{-\lambda(1/\sigma^2)}d\sigma^2.$$

To do the integration, we have to make a change of variable; $d(1/\sigma^2) = -(1/\sigma^2)^2 d\sigma^2$, so $d\sigma^2 = -(1/\sigma^2)^{-2}\, d(1/\sigma^2)$. Making the substitution—the sign of the integral changes twice, once for the Jacobian and back again because the integral from $\sigma^2 = 0$ to $\infty$ is the negative of the integral from $(1/\sigma^2) = 0$ to $\infty$—we obtain

$$\int_0^\infty f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})d\sigma^2 \propto A\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{P-3} e^{-\lambda(1/\sigma^2)}d\left(\frac{1}{\sigma^2}\right)$$
$$= A \times \frac{\Gamma(P-2)}{\lambda^{P-2}}.$$

Reinserting the expressions for $A$, $P$, and $\lambda$ produces

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \propto \frac{\dfrac{[vs^2]^{v+2}\Gamma(v+K/2)}{\Gamma(v+2)}[2\pi]^{-K/2}|X'X|^{-1/2}}{\left[vs^2 + \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})\right]^{v+K/2}}. \tag{16-8}$$

This density is proportional to a **multivariate *t* distribution**[7] and is a generalization of the familiar univariate distribution we have used at various points. This distribution has a degrees of freedom parameter, $d = n - K$, mean $\mathbf{b}$, and covariance matrix $(d/(d-2)) \times [s^2(\mathbf{X}'\mathbf{X})^{-1}]$. Each element of the $K$-element vector $\boldsymbol{\beta}$ has a marginal distribution that is the univariate $t$ distribution with degrees of freedom $n - K$, mean $b_k$, and variance equal to the $k$th diagonal element of the covariance matrix given earlier. Once again, this is the same as our sampling theory result. The difference is a matter of interpretation. In the current context, the estimated distribution is for $\boldsymbol{\beta}$ and is centered at $\mathbf{b}$.

### 16.3.2 ESTIMATION WITH AN INFORMATIVE PRIOR DENSITY

Once we leave the simple case of noninformative priors, matters become quite complicated, both at a practical level and, methodologically, in terms of just where the prior comes from. The integration of $\sigma^2$ out of the posterior in (16-7) is complicated by itself. It is made much more so if the prior distributions of $\boldsymbol{\beta}$ and $\sigma^2$ are at all involved. Partly to offset these difficulties, researchers have used **conjugate priors**, which are ones

[7] See, for example, Judge et al. (1985) for details. The expression appears in Zellner (1971, p. 67). Note that the exponent in the denominator is $v + K/2 = n/2$.

that have the same form as the conditional density and are therefore amenable to the integration needed to obtain the marginal distributions.[8]

### *Example 16.2 Estimation with a Conjugate Prior*

We continue Example 16.1, but we now assume a conjugate prior. For likelihood functions involving proportions, the beta prior is a common device, for reasons that will emerge shortly. The beta prior is

$$p(\theta) = \frac{\Gamma(\alpha + \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}.$$

Then the posterior density becomes

$$\frac{\theta^{D}(1-\theta)^{N-D}\dfrac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}{\displaystyle\int_0^1 \theta^{D}(1-\theta)^{N-D}\frac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}d\theta} = \frac{\theta^{D+\alpha-1}(1-\theta)^{N-D+\beta-1}}{\displaystyle\int_0^1 \theta^{D+\alpha-1}(1-\theta)^{N-D+\beta-1}d\theta}.$$

The posterior density is, once again, a beta distribution, with parameters $(D + \alpha, N - D + \beta)$. The posterior mean is

$$E[\theta \,|\, \mathbf{data}] = \frac{D+\alpha}{N+\alpha+\beta}.$$

(Our previous choice of the uniform density was equivalent to $\alpha = \beta = 1$.) Suppose we choose a prior that conforms to a prior mean of 0.5, but with less mass near zero and one than in the center, such as $\alpha = \beta = 2$. Then the posterior mean would be $(8 + 2)/(25 + 3) = 0.33571$. (This is yet larger than the previous estimator. The reason is that the prior variance is now smaller than 1/12, so the prior mean, still 0.5, receives yet greater weight than it did in the previous example.)

Suppose that we assume that the prior beliefs about $\boldsymbol{\beta}$ may be summarized in a $K$-variate normal distribution with mean $\boldsymbol{\beta}_0$ and variance matrix $\boldsymbol{\Sigma}_0$. Once again, it is illuminating to begin with the case in which $\sigma^2$ is assumed to be known. Proceeding in exactly the same fashion as before, we would obtain the following result: The posterior density of $\boldsymbol{\beta}$ conditioned on $\sigma^2$ and the data will be normal with

$$\begin{aligned} E[\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}, \mathbf{X}] &= \{\boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}\{\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\} \\ &= \mathbf{F}\boldsymbol{\beta}_0 + (\mathbf{I} - \mathbf{F})\mathbf{b}, \end{aligned} \tag{16-9}$$

where

$$\begin{aligned} \mathbf{F} &= \{\boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}\boldsymbol{\Sigma}_0^{-1} \\ &= \{[\text{prior variance}]^{-1} + [\text{conditional variance}]^{-1}\}^{-1}[\text{prior variance}]^{-1}. \end{aligned} \tag{16-10}$$

This vector is a matrix weighted average of the prior and the least squares (sample) coefficient estimates, where the weights are the inverses of the prior and the conditional

[8]Our choice of noninformative prior for $\ln \sigma$ led to a convenient prior for $\sigma^2$ in our derivation of the posterior for $\boldsymbol{\beta}$. The idea that the prior can be specified arbitrarily in whatever form is mathematically convenient is very troubling; it is supposed to represent the accumulated prior belief about the parameter. On the other hand, it could be argued that the conjugate prior is the posterior of a previous analysis, which could justify its form. The issue of how priors should be specified is one of the focal points of the methodological debate. Non-Bayesians argue that it is disingenuous to claim the methodological high ground and then base the crucial prior density in a model purely on the basis of mathematical convenience. In a small sample, this assumed prior is going to dominate the results, whereas in a large one, the sampling theory estimates will dominate anyway.

covariance matrices.[9] The smaller the variance of the estimator, the larger its weight, which makes sense. Also, still taking $\sigma^2$ as known, we can write the variance of the posterior normal distribution as

$$\text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] = \{\boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}. \quad \textbf{(16-11)}$$

Notice that the posterior variance combines the prior and conditional variances on the basis of their inverses.[10] We may interpret the noninformative prior as having infinite elements in $\boldsymbol{\Sigma}_0$. This assumption would reduce this case to the earlier one.

Once again, it is necessary to account for the unknown $\sigma^2$. If our prior over $\sigma^2$ is to be informative as well, then the resulting distribution can be extremely cumbersome. A conjugate prior for $\boldsymbol{\beta}$ and $\sigma^2$ that can be used is

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2)g_{\sigma^2}(\sigma^2), \quad \textbf{(16-12)}$$

where $g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2)$ is normal, with mean $\boldsymbol{\beta}^0$ and variance $\sigma^2\mathbf{A}$ and

$$g_{\sigma^2}(\sigma^2) = \frac{[m\sigma_0^2]^{m+1}}{\Gamma(m+1)}\left(\frac{1}{\sigma^2}\right)^m e^{-m\sigma_0^2(1/\sigma^2)}. \quad \textbf{(16-13)}$$

This distribution is an inverted gamma distribution. It implies that $1/\sigma^2$ has a gamma distribution. The prior mean for $\sigma^2$ is $\sigma_0^2$ and the prior variance is $\sigma_0^4/(m-1)$.[11] The product in (16-12) produces what is called a **normal-gamma prior**, which is the natural conjugate prior for this form of the model. By integrating out $\sigma^2$, we would obtain the prior marginal for $\boldsymbol{\beta}$ alone, which would be a multivariate $t$ distribution.[12] Combining (16-12) with (16-13) produces the joint posterior distribution for $\boldsymbol{\beta}$ and $\sigma^2$. Finally, the marginal posterior distribution for $\boldsymbol{\beta}$ is obtained by integrating out $\sigma^2$. It has been shown that this posterior distribution is multivariate $t$ with

$$E[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}] = \{[\overline{\sigma}^2\mathbf{A}]^{-1} + [\overline{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}\{[\overline{\sigma}^2\mathbf{A}]^{-1}\boldsymbol{\beta}_0 + [\overline{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\} \quad \textbf{(16-14)}$$

and

$$\text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}] = \left(\frac{j}{j-2}\right)\{[\overline{\sigma}^2\mathbf{A}]^{-1} + [\overline{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}, \quad \textbf{(16-15)}$$

where $j$ is a degrees of freedom parameter and $\overline{\sigma}^2$ is the Bayesian estimate of $\sigma^2$. The prior degrees of freedom $m$ is a parameter of the prior distribution for $\sigma^2$ that would have been determined at the outset. (See the following example.) Once again, it is clear that as the amount of data increases, the posterior density, and the estimates thereof, converge to the sampling theory results.

---

[9]Note that it will not follow that individual elements of the posterior mean vector lie between those of $\boldsymbol{\beta}_0$ and b. See Judge et al. (1985, pp. 109–110) and Chamberlain and Leamer (1976).

[10]Precisely this estimator was proposed by Theil and Goldberger (1961) as a way of combining a previously obtained estimate of a parameter and a current body of new data. They called their result a "mixed estimator." The term "mixed estimation" takes an entirely different meaning in the current literature, as we saw in Chapter 15.

[11]You can show this result by using gamma integrals. Note that the density is a function of $1/\sigma^2 = 1/x$ in the formula of (B-39), so to obtain $E[\sigma^2]$, we use the analog of $E[1/x] = \lambda/(P-1)$ and $E[(1/x)^2] = \lambda^2/[(P-1)(P-2)]$. In the density for $(1/\sigma^2)$, the counterparts to $\lambda$ and $P$ are $m\sigma_0^2$ and $m+1$.

[12]Full details of this (lengthy) derivation appear in Judge et al. (1985, pp. 106–110) and Zellner (1971).

**TABLE 16.1** Estimates of the MPC

| *Years* | *Estimated MPC* | *Variance of b* | *Degrees of Freedom* | *Estimated σ* |
|---|---|---|---|---|
| 1940–1950 | 0.6848014 | 0.061878 | 9 | 24.954 |
| 1950–2000 | 0.92481 | 0.000065865 | 49 | 92.244 |

### *Example 16.3 Bayesian Estimate of the Marginal Propensity to Consume*

In Example 3.2, an estimate of the marginal propensity to consume is obtained using 11 observations from 1940 to 1950, with the results shown in the top row of Table 16.1. [Referring to Example 3.2, the variance is (6,848.975/9)/12,300.182.] A classical 95% confidence interval for $\beta$ based on these estimates is (0.1221, 1.2475). (The very wide interval probably results from the obviously poor specification of the model.) Based on noninformative priors for $\beta$ and $\sigma^2$, we would estimate the posterior density for $\beta$ to be univariate $t$ with nine degrees of freedom, with mean 0.6848014 and variance (11/9)0.061878 $=$ 0.075628. An HPD interval for $\beta$ would coincide with the confidence interval. Using the fourth quarter (yearly) values of the 1950–2000 data used in Example 5.3, we obtain the new estimates that appear in the second row of the table.

We take the first estimate and its estimated distribution as our prior for $\beta$ and obtain a posterior density for $\beta$ based on an informative prior instead. We assume for this exercise that $\sigma$ may be taken as known at the sample value of 24.954. Then,

$$\bar{b} = \left[\frac{1}{0.061878} + \frac{1}{0.000065865}\right]^{-1}\left[\frac{0.6848014}{0.061878} + \frac{0.92481}{0.000065865}\right] = 0.92455,$$

The weighted average is overwhelmingly dominated by the far more precise sample estimate from the larger sample. The posterior variance is the inverse in brackets, which is 0.000065795. This is close to the variance of the latter estimate. An HPD interval can be formed in the familiar fashion. It will be slightly narrower than the confidence interval, because the variance of the posterior distribution is slightly smaller than the variance of the sampling estimator. This reduction is the value of the prior information. (As we see here, the prior is not particularly informative.)

## 16.4 BAYESIAN INFERENCE

The posterior density is the Bayesian counterpart to the likelihood function. It embodies the information that is available to make inference about the econometric model. As we have seen, the mean and variance of the posterior distribution correspond to the classical (sampling theory) point estimator and asymptotic variance, although they are interpreted differently. Before we examine more intricate applications of Bayesian inference, it is useful to formalize some other components of the method, point and interval estimation and the Bayesian equivalent of testing a hypothesis.[13]

### 16.4.1 POINT ESTIMATION

The posterior density function embodies the prior and the likelihood and therefore contains all the researcher's information about the parameters. But for purposes of presenting

[13]We do not include prediction in this list. The Bayesian approach would treat the prediction problem as one of estimation in the same fashion as parameter estimation. The value to be forecasted is among the unknown elements of the model that would be characterized by a prior and would enter the posterior density in a symmetric fashion along with the other parameters.

results, the density is somewhat imprecise, and one normally prefers a point or interval estimate. The natural approach would be to use the mean of the posterior distribution as the estimator. For the noninformative prior, we use **b**, the **sampling theory** estimator.

One might ask at this point, why bother? These Bayesian point estimates are identical to the sampling theory estimates. All that has changed is our interpretation of the results. This situation is, however, exactly the way it should be. Remember that we entered the analysis with noninformative priors for $\boldsymbol{\beta}$ and $\sigma^2$. Therefore, the only information brought to bear on estimation is the sample data, and it would be peculiar if anything other than the sampling theory estimates emerged at the end. The results do change when our prior brings out of sample information into the estimates, as we shall see later.

The results will also change if we change our motivation for estimating $\boldsymbol{\beta}$. The parameter estimates have been treated thus far as if they were an end in themselves. But in some settings, parameter estimates are obtained so as to enable the analyst to make a decision. Consider then, a **loss function**, $H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, which quantifies the cost of basing a decision on an estimate $\hat{\boldsymbol{\beta}}$ when the parameter is $\boldsymbol{\beta}$. The expected, or average, loss is

$$E_{\boldsymbol{\beta}}[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})] = \int_{\boldsymbol{\beta}} H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) f(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) d\boldsymbol{\beta}, \tag{16-16}$$

where the weighting function, $f$, is the marginal posterior density. (The joint density for $\boldsymbol{\beta}$ and $\sigma^2$ would be used if the loss were defined over both.) The Bayesian point estimate is the parameter vector that minimizes the expected loss. If the loss function is a quadratic form in $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, then the mean of the posterior distribution is the *minimum expected loss* (MELO) estimator. The proof is simple. For this case,

$$E[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) | \mathbf{y}, \mathbf{X}] = E\left[\tfrac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) | \mathbf{y}, \mathbf{X}\right].$$

To minimize this, we can use the result that

$$\begin{aligned} \partial E[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) | \mathbf{y}, \mathbf{X}] / \partial \hat{\boldsymbol{\beta}} &= E[\partial H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) / \partial \hat{\boldsymbol{\beta}} | \mathbf{y}, \mathbf{X}] \\ &= E[-\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) | \mathbf{y}, \mathbf{X}]. \end{aligned}$$

The minimum is found by equating this derivative to $\mathbf{0}$, whence, because $-\mathbf{W}$ is irrelevant, $\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}]$. This kind of loss function would state that errors in the positive and negative directions are equally bad, and large errors are much worse than small errors. If the loss function were a linear function instead, then the MELO estimator would be the median of the posterior distribution. These results are the same in the case of the noninformative prior that we have just examined.

### 16.4.2 INTERVAL ESTIMATION

The counterpart to a confidence interval in this setting is an interval of the posterior distribution that contains a specified probability. Clearly, it is desirable to have this interval be as narrow as possible. For a unimodal density, this corresponds to an interval within which the density function is higher than any points outside it, which justifies the term **highest posterior density (HPD) interval**. For the case we have analyzed, which involves a symmetric distribution, we would form the HPD interval for $\boldsymbol{\beta}$ around the least squares estimate **b**, with terminal values taken from the standard $t$ tables. Section 4.8.3 shows the (classical) derivation of an HPD interval for an asymmetric distribution, in that case for a prediction of $y$ when the regression models ln $y$.

### 16.4.3 HYPOTHESIS TESTING

The Bayesian methodology treats the classical approach to hypothesis testing with a large amount of skepticism. Two issues are especially problematic. First, a close examination of only the work we have done in Chapter 5 will show that because we are using consistent estimators, with a large enough sample, we will ultimately reject any (nested) hypothesis unless we adjust the significance level of the test downward as the sample size increases. Second, the all-or-nothing approach of either rejecting or not rejecting a hypothesis provides no method of simply sharpening our beliefs. Even the most committed of analysts might be reluctant to discard a strongly held prior based on a single sample of data, yet that is what the sampling methodology mandates. The Bayesian approach to hypothesis testing is much more appealing in this regard. Indeed, the approach might be more appropriately called *comparing hypotheses*, because it essentially involves only making an assessment of which of two hypotheses has a higher probability of being correct.

The Bayesian approach to hypothesis testing bears large similarity to Bayesian estimation.[14] We have formulated two hypotheses, a null, denoted $H_0$, and an alternative, denoted $H_1$. These need not be complementary, as in $H_0$: "statement $A$ is true" versus $H_1$: "statement $A$ is not true," because the intent of the procedure is not to reject one hypothesis in favor of the other. For simplicity, however, we will confine our attention to hypotheses about the parameters in the regression model, which often are complementary. Assume that before we begin our experimentation (i.e., data gathering, statistical analysis) we are able to assign **prior probabilities** $P(H_0)$ and $P(H_1)$ to the two hypotheses. The **prior odds ratio** is simply the ratio

$$\text{Odds}_{prior} = \frac{P(H_0)}{P(H_1)}. \tag{16-17}$$

For example, one's uncertainty about the sign of a parameter might be summarized in a prior odds over $H_0: \beta \geq 0$ versus $H_1: \beta < 0$ of $0.5/0.5 = 1$. After the sample evidence is gathered, the prior will be modified, so the posterior is, in general,

$$\text{Odds}_{posterior} = B_{01} \times \text{Odds}_{prior}.$$

The value $B_{01}$ is called the **Bayes factor** for comparing the two hypotheses. It summarizes the effect of the sample data on the prior odds. The end result, $\text{Odds}_{posterior}$, is a new odds ratio that can be carried forward as the prior in a subsequent analysis.

The Bayes factor is computed by assessing the likelihoods of the data observed under the two hypotheses. We return to our first departure point, the likelihood of the data, given the parameters,

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \tag{16-18}$$

Based on our priors for the parameters, the expected, or average likelihood, assuming that hypothesis $j$ is true ($j = 0, 1$), is

$$f(\mathbf{y} \mid \mathbf{X}, H_j) = E_{\boldsymbol{\beta},\sigma^2}[f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j)] = \int_{\sigma^2}\int_{\boldsymbol{\beta}} f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j) g(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2.$$

[14] For extensive discussion, see Zellner and Siow (1980) and Zellner (1985, pp. 275–305).

(This conditional density is also the **predictive density** for $\mathbf{y}$.) Therefore, based on the observed data, we use Bayes's theorem to reassess the probability of $H_j$; the posterior probability is

$$P(H_j|\mathbf{y}, \mathbf{X}) = \frac{f(\mathbf{y}|\mathbf{X}, H_j)P(H_j)}{f(\mathbf{y})}.$$

The posterior odds ratio is $P(H_0|\mathbf{y}, \mathbf{X})/P(H_1|\mathbf{y}, \mathbf{X})$, so the Bayes factor is

$$B_{01} = \frac{f(\mathbf{y}|\mathbf{X}, H_0)}{f(y|\mathbf{X}, H_1)}.$$

### *Example 16.4 Posterior Odds for the Classical Regression Model*

Zellner (1971) analyzes the setting in which there are two possible explanations for the variation in a dependent variable $y$:

$$\text{Model0: } y = \mathbf{x}_0'\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0$$

and

$$\text{Model1: } y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1.$$

We will briefly sketch his results. We form *informative priors* for $[\boldsymbol{\beta}, \sigma^2]_j$, $j = 0, 1$, as specified in (16-12) and (16-13), that is, multivariate normal and inverted gamma, respectively. Zellner then derives the Bayes factor for the posterior odds ratio. The derivation is lengthy and complicated, but for large $n$, with some simplifying assumptions, a useful formulation emerges. First, assume that the priors for $\sigma_0^2$ and $\sigma_1^2$ are the same. Second, assume that $[|\mathbf{A}_0^{-1}|/|\mathbf{A}_0^{-1} + \mathbf{X}_0'\mathbf{X}_0|]/[|\mathbf{A}_1^{-1}|/|\mathbf{A}_1^{-1} + \mathbf{X}_1'\mathbf{X}_1|] \rightarrow 1$. The first of these would be the usual situation, in which the uncertainty concerns the covariation between $y_i$ and $\mathbf{x}_i$, not the amount of residual variation (lack of fit). The second concerns the relative amounts of information in the prior ($\mathbf{A}$) versus the likelihood ($\mathbf{X}'\mathbf{X}$). These matrices are the inverses of the covariance matrices, or the **precision matrices**. [Note how these two matrices form the matrix weights in the computation of the posterior mean in (16-9).] Zellner (p. 310) discusses this assumption at some length. With these two assumptions, he shows that as $n$ grows large,[15]

$$B_{01} \approx \left(\frac{s_0^2}{s_1^2}\right)^{-(n+m)/2} = \left(\frac{1 - R_0^2}{1 - R_1^2}\right)^{-(n+m)/2}.$$

Therefore, the result favors the model that provides the better fit using $R^2$ as the fit measure. If we stretch Zellner's analysis a bit by interpreting model 1 as *the model* and model 0 as "no model" (that is, the relevant part of $\boldsymbol{\beta}_0 = \mathbf{0}$, so $R_0^2 = 0$), then the ratio simplifies to

$$B_{01} = (1 - R_1^2)^{(n+m)/2}.$$

Thus, the better the fit of the regression, the lower the Bayes factor in favor of model 0 (no model), which makes intuitive sense.

Zellner and Siow (1980) have continued this analysis with noninformative priors for $\boldsymbol{\beta}$ and $\sigma_j^2$. Specifically, they use a flat prior for $\ln \sigma$ [see (16-7)] and a multivariate Cauchy prior (which has infinite variances) for $\boldsymbol{\beta}$. Their main result (3.10) is

[15] A ratio of exponentials that appears in Zellner's result (his equation 10.50) is omitted. To the order of approximation in the result, this ratio vanishes from the final result. (Personal correspondence from A. Zellner to the author.)

$$B_{01} = \frac{\frac{1}{2}\sqrt{\pi}}{\Gamma[(k+1)/2]}\left(\frac{n-K}{2}\right)^{k/2}(1-R^2)^{(n-K-1)/2}.$$

This result is very much like the previous one, with some slight differences due to degrees of freedom corrections and the several approximations used to reach the first one.

### 16.4.4 LARGE-SAMPLE RESULTS

Although all statistical results for Bayesian estimators are necessarily "finite sample" (they are conditioned on the sample data), it remains of interest to consider how the estimators behave in large samples.[16] Do Bayesian estimators "converge" to something? To do this exercise, it is useful to envision having a sample that is the entire population. Then, the posterior distribution would characterize this entire population, not a sample from it. It stands to reason in this case, at least intuitively, that the posterior distribution should coincide with the likelihood function. It will (as usual) save for the influence of the prior. But as the sample size grows, one should expect the likelihood function to overwhelm the prior. It will, unless the strength of the prior grows with the sample size (that is, for example, if the prior variance is of order $1/n$). An informative prior will still fade in its influence on the posterior unless it becomes *more* informative as the sample size grows.

The preceding suggests that the posterior mean will converge to the maximum likelihood estimator. The MLE is the parameter vector that is at the mode of the likelihood function. The Bayesian estimator is the **posterior mean**, not the mode, so a remaining question concerns the relationship between these two features. The **Bernstein–von Mises "theorem"** [See Cameron and Trivedi (2005, p. 433) and Train (2003, Chapter 12)] states that the posterior mean and the maximum likelihood estimator will coverge to the same probability limit and have the same limiting normal distribution. A form of central limit theorem is at work.

But for remaining philosophical questions, the results suggest that for large samples, the choice between Bayesian and frequentist methods can be one of computational efficiency. (This is the thrust of the application in Section 16.8. Note, as well, footnote 1 at the beginning of this chapter. In an infinite sample, the maintained uncertainty of the Bayesian estimation framework would have to arise from deeper questions about the model. For example, the mean of the entire population is its mean; there is no uncertainty about the parameter.)

## 16.5 POSTERIOR DISTRIBUTIONS AND THE GIBBS SAMPLER

The foregoing analysis has proceeded along a set of steps that includes formulating the likelihood function (the model), the prior density over the objects of estimation, and the posterior density. To complete the inference step, we then analytically derived the characteristics of the posterior density of interest, such as the mean or mode, and the

[16]The standard preamble in econometric studies, that the analysis to follow is "exact" as opposed to approximate or "large sample," refers to this aspect—the analysis is conditioned on and, by implication, applies only to the sample data in hand. Any inference outside the sample, for example, to hypothesized random samples is, like the sampling theory counterpart, approximate.

variance. The complicated element of any of this analysis is determining the moments of the posterior density, for example, the mean,

$$\hat{\theta} = E[\theta|\text{data}] = \int_{\theta} \theta\, p(\theta|\textbf{data}) d\theta. \quad \textbf{(16-19)}$$

There are relatively few applications for which integrals such as this can be derived in closed form. (This is one motivation for conjugate priors.) The modern approach to Bayesian inference takes a different strategy. The result in (16-19) is an expectation. Suppose it were possible to obtain a random sample, as large as desired, from the population defined by $p(\theta|\textbf{data})$. Then, using the same strategy we used throughout Chapter 15 for simulation-based estimation, we could use that sample's characteristics, such as mean, variance, quantiles, and so on, to infer the characteristics of the posterior distribution. Indeed, with an (essentially) infinite sample, we would be freed from having to limit our attention to a few simple features such as the mean and variance and we could view any features of the posterior distribution that we like. The (much less) complicated part of the analysis is the formulation of the posterior density.

It remains to determine how the sample is to be drawn from the posterior density. This element of the strategy is provided by a remarkable (and remarkably useful) result known as the **Gibbs sampler**.[17] The central result of the Gibbs sampler is as follows: We wish to draw a random sample from the joint population $(x, y)$. The joint distribution of $x$ and $y$ is either unknown or intractable and it is not possible to sample from the joint distribution. However, assume that the conditional distributions $f(x|y)$ and $f(y|x)$ are known and simple enough that it is possible to draw univariate random samples from both of them. The following iteration will produce a bivariate random sample from the joint distribution:

Gibbs Sampler:

1. Begin the cycle with a value of $x_0$ that is in the right range of $x|y$,
2. Draw an observation $y_0|x_0$, from the known population $y|x$,
3. Draw an observation $x_t|y_{t-1}$, from the known population $x|y$,
4. Draw an observation $y_t|x_t$ from the known population of $y|x$.

Iteration of steps 3 and 4 for several thousand cycles will eventually produce a random sample from the joint distribution. (The first several thousand draws are discarded to avoid the influence of the initial conditions—this is called the **burn in**.) [Some technical details on the procedure appear in Cameron and Trivedi (Section 13.5).]

### *Example 16.5 Gibbs Sampling from the Normal Distribution*

To illustrate the mechanical aspects of the Gibbs sampler, consider random sampling from the joint normal distribution. We consider the bivariate normal distribution first. Suppose we wished to draw a random sample from the population

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

As we have seen in Chapter 15, a direct approach is to use the fact that linear functions of normally distributed variables are normally distributed. [See (B-80).] Thus, we might

[17] See Casella and George (1992).

transform a series of independent normal draws $(u_1, u_2)'$ by the Cholesky decomposition of the covariance matrix,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_i = \begin{bmatrix} 1 & 0 \\ \theta_1 & \theta_2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_i = \mathbf{L}\mathbf{u}_i,$$

where $\theta_1 = \rho$ and $\theta_2 = \sqrt{1 - \rho^2}$. The Gibbs sampler would take advantage of the result

$$x_1 | x_2 \sim N[\rho x_2, (1 - \rho^2)],$$

and

$$x_2 | x_1 \sim N[\rho x_1, (1 - \rho^2)].$$

To sample from a trivariate, or multivariate population, we can expand the Gibbs sequence in the natural fashion. For example, to sample from a trivariate population, we would use the Gibbs sequence

$$\begin{aligned} x_1 | x_2, x_3 &\sim N[\beta_{1,2}x_2 + \beta_{1,3}x_3, \boldsymbol{\Sigma}_{1|2,3}], \\ x_2 | x_1, x_3 &\sim N[\beta_{2,1}x_1 + \beta_{2,3}x_3, \boldsymbol{\Sigma}_{2|1,3}], \\ x_3 | x_1, x_2 &\sim N[\beta_{3,1}x_1 + \beta_{3,2}x_2, \boldsymbol{\Sigma}_{3|1,2}], \end{aligned}$$

where the conditional means and variances are given in Theorem B.7. This defines a three-step cycle.

The availability of the Gibbs sampler frees the researcher from the necessity of deriving the analytical properties of the full, joint posterior distribution. Because the formulation of conditional priors is straightforward, and the derivation of the conditional posteriors is only slightly less so, this tool has facilitated a vast range of applications that previously were intractable. For an example, consider, once again, the classical normal regression model. From (16-7), the joint posterior for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto\ & \frac{[\nu s^2]^{\nu+2}}{\Gamma(\nu + 2)} \left[\frac{1}{\sigma^2}\right]^{\nu+1} \exp(-\nu s^2/\sigma^2)[2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \\ & \times \exp(-(1/2)(\boldsymbol{\beta} - \mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta} - \mathbf{b}). \end{aligned}$$

If we wished to use a simulation approach to characterizing the posterior distribution, we would need to draw a $K + 1$ variate sample of observations from this intractable distribution. However, with the assumed priors, we found the conditional posterior for $\boldsymbol{\beta}$ in (16-5):

$$p(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) = N[\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

From (16-6), we can deduce that the conditional posterior for $\sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}$ is an inverted gamma distribution with parameters $m\sigma_0^2 = \nu\hat{\sigma}^2$ and $m = \nu$ in (16-13):

$$p(\sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = \frac{[\nu\hat{\sigma}^2]^{\nu+1}}{\Gamma(\nu + 1)} \left[\frac{1}{\sigma^2}\right]^{\nu} \exp(-\nu\hat{\sigma}^2/\sigma^2), \quad \hat{\sigma}^2 = \frac{\Sigma_{i=1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{n - K}.$$

This sets up a Gibbs sampler for sampling from the joint posterior of $\boldsymbol{\beta}$ and $\sigma^2$. We would cycle between random draws from the multivariate normal for $\boldsymbol{\beta}$ and the inverted gamma distribution for $\sigma^2$ to obtain a $K + 1$ variate sample on $(\boldsymbol{\beta}, \sigma^2)$. [Of course, for this application, we do know the marginal posterior distribution for $\boldsymbol{\beta}$—see (16-8).]

The Gibbs sampler is not truly a random sampler; it is a Markov chain—each "draw" from the distribution is a function of the draw that precedes it. The random input at each cycle provides the randomness, which leads to the popular name for this strategy, **Markov chain Monte Carlo** or **MCMC** or **MC**$^2$ (pick one) estimation. In its simplest form, it provides a remarkably efficient tool for studying the posterior distributions in very complicated models. The example in the next section shows a striking example of how to locate the MLE for a probit model without computing the likelihood function or its derivatives. In Section 16.8, we will examine an extension and refinement of the strategy, the Metropolis–Hasting algorithm.

In the next several sections, we will present some applications of Bayesian inference. In Section 16.9, we will return to some general issues in classical and Bayesian estimation and inference. At the end of the chapter, we will examine Koop and Tobias's (2004) Bayesian approach to the analysis of heterogeneity in a wage equation based on panel data. We used classical methods to analyze these data in Example 15.16.

## 16.6 APPLICATION: BINOMIAL PROBIT MODEL

Consider inference about the binomial probit model for a dependent variable that is generated as follows (see Sections 17.2–17.4):

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim \mathrm{N}[0, 1], \tag{16-20}$$

$$y_i = 1 \quad \text{if} \quad y_i^* > 0, \text{otherwise } y_i = 0. \tag{16-21}$$

(Theoretical motivation for the model appears in Section 17.3.) The data consist of $(\mathbf{y}, \mathbf{X}) = (y_i, \mathbf{x}_i), i = 1, \ldots, n$. The random variable $y_i$ has a Bernoulli distribution with probabilities

$$\begin{aligned}\text{Prob}[y_i = 1|\mathbf{x}_i] &= \Phi(\mathbf{x}_i'\boldsymbol{\beta}),\\ \text{Prob}[y_i = 0|\mathbf{x}_i] &= 1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}).\end{aligned}$$

The likelihood function for the observed data is

$$L(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n}[\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}.$$

(Once again, we cheat a bit on the notation—the likelihood function is actually the joint density for the data, given $\mathbf{X}$ and $\boldsymbol{\beta}$.) Classical maximum likelihood estimation of $\boldsymbol{\beta}$ is developed in Section 17.3. To obtain the posterior mean (Bayesian estimator), we assume a noninformative, flat (improper) prior for $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}) \propto 1.$$

The posterior density would be

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \frac{\prod_{i=1}^{n}[\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}(1)}{\int_{\boldsymbol{\beta}}\prod_{i=1}^{n}[\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}(1)d\boldsymbol{\beta}},$$

and the estimator would be the posterior mean,

$$\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}] = \frac{\int_{\boldsymbol{\beta}} \boldsymbol{\beta} \prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i} d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}} \prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i} d\boldsymbol{\beta}}. \tag{16-22}$$

Evaluation of the integrals in (16-22) is hopelessly complicated, but a solution using the Gibbs sampler and a technique known as **data augmentation**, pioneered by Albert and Chib (1993a), is surprisingly simple. We begin by treating the unobserved $y_i^*$'s as unknowns to be estimated, along with $\boldsymbol{\beta}$. Thus, the $(K + n) \times 1$ parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{y}^*)$. We now construct a Gibbs sampler. Consider, first, $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X})$. If $y_i^*$ is known, then $y_i$ is known [see (16-21)]. It follows that

$$p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X}) = p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{X}).$$

This posterior defines a linear regression model with normally distributed disturbances and known $\sigma^2 = 1$. It is precisely the model we saw in Section 16.3.1, and the posterior we need is in (16-5), with $\sigma^2 = 1$. So, based on our earlier results, it follows that

$$p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X}) = \mathrm{N}[\mathbf{b}^*, (\mathbf{X}'\mathbf{X})^{-1}], \tag{16-23}$$

where

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*.$$

For $y_i^*$, ignoring $y_i$ for the moment, it would follow immediately from (16-20) that

$$p(y_i^*|\boldsymbol{\beta}, \mathbf{X}) = \mathrm{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1].$$

However, $y_i$ is informative about $y_i^*$. If $y_i$ equals one, we know that $y_i^* > 0$ and if $y_i$ equals zero, then $y_i^* \leq 0$. The implication is that conditioned on $\boldsymbol{\beta}$, $\mathbf{X}$, and $\mathbf{y}$, $y_i^*$ has the truncated (above or below zero) normal distribution that is developed in Sections 19.2.1 and 19.2.2. The standard notation for this is

$$\begin{aligned} p(y_i^*|y_i = 1, \boldsymbol{\beta}, \mathbf{x}_i) &= \mathrm{N}^+[\mathbf{x}_i'\boldsymbol{\beta}, 1], \\ p(y_i^*|y_i = 0, \boldsymbol{\beta}, \mathbf{x}_i) &= \mathrm{N}^-[\mathbf{x}_i'\boldsymbol{\beta}, 1]. \end{aligned} \tag{16-24}$$

Results (16-23) and (16-24) set up the components for a Gibbs sampler that we can use to estimate the posterior means $E[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}]$ and $E[\mathbf{y}^*|\mathbf{y}, \mathbf{X}]$. The following is our algorithm:

*Gibbs Sampler for the Binomial Probit Model*

1. Compute $\mathbf{X}'\mathbf{X}$ once at the outset and obtain $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}' = (\mathbf{X}'\mathbf{X})^{-1}$ (Cholesky decomposition).
2. Start $\boldsymbol{\beta}$ at any value such as $\mathbf{0}$.
3. Result (15-4) shows how to transform a draw from U[0, 1] to a draw from the truncated normal with underlying mean $\mu$ and standard deviation $\sigma$. For this application, the draw is

$$\begin{aligned} y_{i,r}^*(r) &= \mathbf{x}_i'\boldsymbol{\beta}_{r-1} + \Phi^{-1}[1 - (1 - U)\Phi(\mathbf{x}_i'\boldsymbol{\beta}_{r-1})] \quad \text{if } y_i = 1, \\ y_{i,r}^*(r) &= \mathbf{x}_i'\boldsymbol{\beta}_{r-1} + \Phi^{-1}[U\Phi(-\mathbf{x}_i'\boldsymbol{\beta}_{r-1})] \quad \text{if } y_i = 0. \end{aligned}$$

This step is used to draw the $n$ observations on $y_{i,r}^*(r)$.

4. Section 15.2.4 shows how to draw an observation from the multivariate normal population. For this application, we use the results at step 3 to compute $\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*(r)$. We obtain a vector, $\mathbf{v}$, of $K$ draws from the N[0, 1] population, then $\boldsymbol{\beta}(r) = \mathbf{b}^* + \mathbf{L}\mathbf{v}$.

The iteration cycles between steps 3 and 4. This should be repeated several thousand times, discarding the burn-in draws, then the estimator of $\boldsymbol{\beta}$ is the sample mean of the retained draws. The posterior variance is computed with the variance of the retained draws. Posterior estimates of $y_i^*$ would typically not be useful.

### *Example 16.6 Gibbs Sampler for a Probit Model*

In Examples 14.19 through 14.21, we examined Spector and Mazzeo's (1980) widely traveled data on a binary choice outcome. (The example used the data for a different model.) The binary probit model studied in the paper was

$$\text{Prob}(GRADE_i = 1 \,|\, \boldsymbol{\beta}, \mathbf{x}_i) = \Phi(\beta_1 + \beta_2 GPA_i + \beta_3 TUCE_i + \beta_4 PSI_i).$$

The variables are defined in Example 14.19. Their probit model is studied in Example 17.3. The sample contains 32 observations. Table 16.2 presents the maximum likelihood estimates and the posterior means and standard deviations for the probit model. For the Gibbs sampler, we used 5,000 draws, and discarded the first 1,000.

The results in Table 16.2 suggest the similarity of the posterior mean estimated with the Gibbs sampler to the maximum likelihood estimate. However, the sample is quite small, and the differences between the coefficients are still fairly substantial. For a striking example of the behavior of this procedure, we now revisit the German health care data examined in Example 14.23 and several other examples throughout the book. The probit model to be estimated is

$$\begin{aligned}\text{Prob}(Doctor\ visits_{it} > 0) = \Phi(&\beta_1 + \beta_2\, Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income_{it} \\ &+ \beta_5\, Kids_{it} + \beta_6\, Married_{it} + \beta_7\, Female_{it}).\end{aligned}$$

The sample contains data on 7,293 families and a total of 27,326 observations. We are pooling the data for this application. Table 16.3 presents the probit results for this model using the same procedure as before. (We used only 500 draws and discarded the first 100.)

The similarity is what one would expect given the large sample size. We note before proceeding to other applications, notwithstanding the striking similarity of the Gibbs sampler to the MLE, that this is not an efficient method of estimating the parameters of a probit model. The estimator requires generation of thousands of samples of potentially thousands of observations. We used only 500 replications to produce Table 16.3. The computations took about five minutes. Using Newton's method to maximize the log likelihood directly took less than five seconds. Unless one is wedded to the Bayesian paradigm, on strictly practical grounds, the MLE would be the preferred estimator.

**TABLE 16.2** Probit Estimates for Grade Equation

| | *Maximum Likelihood* | | *Posterior Means and Std. Devs.* | |
|---|---|---|---|---|
| ***Variable*** | ***Estimate*** | ***Std. Error*** | ***Posterior Mean*** | ***Posterior S.D.*** |
| *Constant* | −7.4523 | 2.5425 | −8.6286 | 2.7995 |
| *GPA* | 1.6258 | 0.6939 | 1.8754 | 0.7668 |
| *TUCE* | 0.0517 | 0.0839 | 0.0628 | 0.0869 |
| *PSI* | 1.4263 | 0.5950 | 1.6072 | 0.6257 |

**TABLE 16.3** Probit Estimates for Doctor Visits Equation

| | *Maximum Likelihood* | | *Posterior Means and Std. Devs.* | |
|---|---|---|---|---|
| ***Variable*** | ***Estimate*** | ***Std. Error*** | ***Posterior Mean*** | ***Posterior S.D.*** |
| *Constant* | −0.124332 | 0.058146 | −0.126287 | 0.054759 |
| *Age* | 0.011892 | 0.000796 | 0.011979 | 0.000801 |
| *Education* | −0.014959 | 0.003575 | −0.015142 | 0.003625 |
| *Income* | −0.132595 | 0.046552 | −0.126693 | 0.047979 |
| *Kids* | −0.152114 | 0.018327 | −0.151492 | 0.018400 |
| *Married* | 0.073518 | 0.020644 | 0.071977 | 0.020852 |
| *Female* | 0.355906 | 0.016017 | 0.355828 | 0.015913 |

This application of the Gibbs sampler demonstrates in an uncomplicated case how the algorithm can provide an alternative to actually maximizing the log likelihood. We do note that the similarity of the method to the EM algorithm in Section E.3.7 is not coincidental. Both procedures use an estimate of the unobserved, censored data, and both estimate $\boldsymbol{\beta}$ by using OLS using the predicted data.

## 16.7 PANEL DATA APPLICATION: INDIVIDUAL EFFECTS MODELS

We consider a panel data model with common individual effects,

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathrm{N}[0, \sigma_\varepsilon^2].$$

In the Bayesian framework, there is no need to distinguish between fixed and random effects. The classical distinction results from an asymmetric treatment of the data and the parameters. So, we will leave that unspecified for the moment. The implications will emerge later when we specify the prior densities over the model parameters.

The likelihood function for the sample under normality of $\varepsilon_{it}$ is

$$p(\mathbf{y}|\alpha_1, \ldots, \alpha_n, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mathbf{X}) = \prod_{i=1}^{n}\prod_{t=1}^{T_i} \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} \exp\left(-\frac{(y_{it} - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta})^2}{2\sigma_\varepsilon^2}\right).$$

The remaining analysis hinges on the specification of the prior distributions. We will consider three cases. Each illustrates an aspect of the methodology.

First, group the full set of location (regression) parameters in one $(n + K) \times 1$ slope vector, $\boldsymbol{\gamma}$. Then, with the disturbance variance, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2) = (\boldsymbol{\gamma}, \sigma_\varepsilon^2)$. Define a conformable data matrix, $\mathbf{Z} = (\mathbf{D}, \mathbf{X})$, where $\mathbf{D}$ contains the $n$ dummy variables so that we may write the model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

in the familiar fashion for our common effects linear regression. (See Chapter 11.) We now assume the **uniform-inverse gamma prior** that we used in our earlier treatment of the linear model,

$$p(\boldsymbol{\gamma}, \sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2.$$

The resulting (marginal) posterior density for $\boldsymbol{\gamma}$ is precisely that in (16-8) (where now the slope vector includes the elements of $\boldsymbol{\alpha}$). The density is an $(n + K)$ variate $t$ with mean equal to the OLS estimator and covariance matrix $[(\Sigma_i T_i - n - K)/(\Sigma_i T_i - n - K - 2)]s^2(\mathbf{Z}'\mathbf{Z})^{-1}$.

Because OLS in this model as stated means the within estimator, the implication is that with this noninformative prior over $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the model is equivalent to the fixed effects model. Note, again, this is not a consequence of any assumption about correlation between effects and included variables. That has remained unstated; though, by implication, we would allow correlation between **D** and **X**.

Some observers are uncomfortable with the idea of a **uniform prior** over the entire real line.[18] Formally, our assumption of a uniform prior over the entire real line is an **improper prior** because it cannot have a positive density and integrate to one over the entire real line. As such, the posterior appears to be ill defined. However, note that the "improper" uniform prior will, in fact, fall out of the posterior, because it appears in both numerator and denominator. The practical solution for location parameters, such as a vector of regression slopes, is to assume a nearly flat, "almost uninformative" prior. The usual choice is a conjugate normal prior with an arbitrarily large variance. (It should be noted, of course, that as long as that variance is finite, even if it is large, the prior is informative. We return to this point in Section 16.9.)

Consider, then, the conventional normal-gamma prior over $(\boldsymbol{\gamma}, \sigma_\varepsilon^2)$ where the conditional (on $\sigma_\varepsilon^2$) prior normal density for the slope parameters has mean $\boldsymbol{\gamma}_0$ and covariance matrix $\sigma_\varepsilon^2\mathbf{A}$, where the $(n + K) \times (n + K)$ matrix, **A**, is yet to be specified. [See the discussion after (16-13).] The marginal posterior mean and variance for $\boldsymbol{\gamma}$ for this set of assumptions are given in (16-14) and (16-15). We reach a point that presents two rather serious dilemmas for the researcher. The posterior was simple with our uniform, noninformative prior. Now, it is necessary actually to specify **A**, which is potentially large. (In one of our main applications in this text, we are analyzing models with $n = 7{,}293$ constant terms and about $K = 7$ regressors.) It is hopelessly optimistic to expect to be able to specify all the variances and covariances in a matrix this large, unless we actually have the results of an earlier study (in which case we would also have a prior estimate of $\boldsymbol{\gamma}$). A practical solution that is frequently chosen is to specify **A** to be a diagonal matrix with extremely large diagonal elements, thus emulating a uniform prior without having to commit to one. The second practical issue then becomes dealing with the actual computation of the order $(n + K)$ inverse matrix in (16-14) and (16-15). Under the strategy chosen, to make **A** a multiple of the identity matrix, however, there are forms of partitioned inverse matrices that will allow solution to the actual computation.

Thus far, we have assumed that each $\alpha_i$ is generated by a different normal distribution, $-\boldsymbol{\gamma}_0$ and **A**, however specified, have (potentially) different means and variances for the elements of $\boldsymbol{\alpha}$. The third specification we consider is one in which all $\alpha_i$'s in the model are assumed to be draws from the same population. To produce this specification, we use a **hierarchical prior** for the individual effects. The full model will be

$$
\begin{aligned}
y_{it} &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it} \sim \mathrm{N}[0, \sigma_\varepsilon^2], \\
p(\boldsymbol{\beta}|\sigma_\varepsilon^2) &= \mathrm{N}[\boldsymbol{\beta}_0, \sigma_\varepsilon^2\mathbf{A}], \\
p(\sigma_\varepsilon^2) &= \mathrm{Gamma}(\sigma_0^2, m), \\
p(\alpha_i) &= \mathrm{N}[\mu_\alpha, \tau_\alpha^2], \\
p(\mu_\alpha) &= \mathrm{N}[a, Q], \\
p(\tau_\alpha^2) &= \mathrm{Gamma}(\tau_0^2, v).
\end{aligned}
$$

[18] See, for example, Koop (2003, pp. 22–23), Zellner (1971, p. 20), and Cameron and Trivedi (2005, pp. 425–427).

We will not be able to derive the posterior density (joint or marginal) for the parameters of this model. However, it is possible to set up a Gibbs sampler that can be used to infer the characteristics of the posterior densities statistically. The sampler will be driven by conditional normal posteriors for the location parameters, $[\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma_\varepsilon^2, \mu_\alpha, \tau_\alpha^2]$, $[\alpha_i|\boldsymbol{\beta}, \sigma_\varepsilon^2, \mu_\alpha, \tau_\alpha^2]$, and $[\mu_\alpha|\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\varepsilon^2, \tau_\alpha^2]$ and conditional gamma densities for the scale (variance) parameters, $[\sigma_\varepsilon^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_\alpha, \tau_\alpha^2]$ and $[\tau_\alpha^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mu_\alpha]$.[19] The assumption of a common distribution for the individual effects and an independent prior for $\boldsymbol{\beta}$ produces a Bayesian counterpart to the random effects model.

## 16.8 HIERARCHICAL BAYES ESTIMATION OF A RANDOM PARAMETERS MODEL

We now consider a Bayesian approach to estimation of the random parameters model.[20] For an individual $i$, the conditional density for the dependent variable in period $t$ is $f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i)$, where $\boldsymbol{\beta}_i$ is the individual specific $K \times 1$ parameter vector and $\mathbf{x}_{it}$ is individual specific data that enter the probability density.[21] For the sequence of $T$ observations, assuming conditional (on $\boldsymbol{\beta}_i$) independence, person $i$'s contribution to the likelihood for the sample is

$$f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=1}^{T} f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i), \tag{16-25}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$ and $\mathbf{X}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}]$. We will suppose that $\boldsymbol{\beta}_i$ is distributed normally with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$. (This is the "hierarchical" aspect of the model.) The unconditional density would be the expected value over the possible values of $\boldsymbol{\beta}_i$,

$$f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\beta}_i} \prod_{t=1}^{T} f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i) \phi_K[\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}] d\boldsymbol{\beta}_i, \tag{16-26}$$

where $\phi_K[\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}]$ denotes the $K$ variate normal prior density for $\boldsymbol{\beta}_i$ given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Maximum likelihood estimation of this model, which entails estimation of the deep parameters, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, then estimation of the individual specific parameters, $\boldsymbol{\beta}_i$ is considered in Sections 15.7 through 15.11. We now consider the Bayesian approach to estimation of the parameters of this model.

To approach this from a Bayesian viewpoint, we will assign noninformative prior densities to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. As is conventional, we assign a flat (noninformative) prior to $\boldsymbol{\beta}$.

[19]The procedure is developed at length by Koop (2003, pp. 152–153).

[20]Note that there is occasional confusion as to what is meant by *random parameters* in a random parameters (RP) model. In the Bayesian framework we discuss in this chapter, the "randomness" of the random parameters in the model arises from the uncertainty of the analyst. As developed at several points in this book (and in the literature), the randomness of the parameters in the RP model is a characterization of the heterogeneity of parameters across individuals. Consider, for example, in the Bayesian framework of this section, in the RP model, each vector $\boldsymbol{\beta}_i$ is a random vector with a distribution (defined hierarchically). In the classical framework, each $\boldsymbol{\beta}_i$ represents a single draw from a parent population.

[21]To avoid a layer of complication, we will embed the time-invariant effect $\Delta\mathbf{z}_i$ in $\mathbf{x}_{it}'\boldsymbol{\beta}$. A full treatment in the same fashion as the latent class model would be substantially more complicated in this setting (although it is quite straightforward in the maximum simulated likelihood approach discussed in Section 15.11).

The variance parameters are more involved. If it is assumed that the elements of $\boldsymbol{\beta}_i$ are conditionally independent, then each element of the (now) diagonal matrix $\boldsymbol{\Sigma}$ may be assigned the inverted gamma prior that we used in (16-13). A full matrix $\boldsymbol{\Sigma}$ is handled by assigning to $\boldsymbol{\Sigma}$ an **inverted Wishart** prior density with parameters scalar $K$ and matrix $K \times \mathbf{I}$.[22] This produces the joint posterior density,

$$\Lambda(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \,|\, \text{all data}) = \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T} f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}_i) \phi_K[\boldsymbol{\beta}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma}] \right\} \times p(\boldsymbol{\beta}, \boldsymbol{\Sigma}). \quad \textbf{(16-27)}$$

This gives the joint density of all the unknown parameters conditioned on the observed data. Our Bayesian estimators of the parameters will be the posterior means for these $(n + 1)K + K(K + 1)/2$ parameters. In principle, this requires integration of (16-27) with respect to the components. As one might guess at this point, that integration is hopelessly complex and not remotely feasible.

However, the techniques of Markov chain Monte Carlo (MCMC) simulation estimation (the Gibbs sampler) and the **Metropolis–Hastings algorithm** enable us to sample from the (only seemingly hopelessly complex) joint density $\Lambda(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \,|\, \textbf{all data})$ in a remarkably simple fashion. Train (2001 and 2002, Chapter 12) describes how to use these results for this random parameters model.[23] The usefulness of this result for our current problem is that it is, indeed, possible to partition the joint distribution, and we can easily sample from the conditional distributions. We begin by partitioning the parameters into $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\delta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$. Train proposes the following strategy: To obtain a draw from $\boldsymbol{\gamma} | \boldsymbol{\delta}$, we will use the Gibbs sampler to obtain a draw from the distribution of $(\boldsymbol{\beta} | \boldsymbol{\Sigma}, \boldsymbol{\delta})$ and then one from the distribution of $(\boldsymbol{\Sigma} | \boldsymbol{\beta}, \boldsymbol{\delta})$. We will lay out this first, then turn to sampling from $\boldsymbol{\delta} | \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

Conditioned on $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ has a $K$-variate normal distribution with mean $\overline{\boldsymbol{\beta}} = (1/n) | \Sigma_{i=1}^{n} \boldsymbol{\beta}_i$ and covariance matrix $(1/n)\boldsymbol{\Sigma}$. To sample from this distribution we will first obtain the Cholesky factorization of $\boldsymbol{\Sigma} = \mathbf{LL}'$ where $\mathbf{L}$ is a lower triangular matrix. (See Section A.6.11.) Let $\mathbf{v}$ be a vector of $K$ draws from the standard normal distribution. Then, $\overline{\boldsymbol{\beta}} + \mathbf{Lv}$ has mean vector $\overline{\boldsymbol{\beta}} + \mathbf{L} \times \mathbf{0} = \overline{\boldsymbol{\beta}}$ and covariance matrix $\mathbf{LIL}' = \boldsymbol{\Sigma}$, which is exactly what we need. So, this shows how to sample a draw from the conditional distribution $\boldsymbol{\beta}$.

To obtain a random draw from the distribution of $\boldsymbol{\Sigma} | \boldsymbol{\beta}, \boldsymbol{\delta}$, we will require a random draw from the inverted Wishart distribution. The marginal posterior distribution of $\boldsymbol{\Sigma} | \boldsymbol{\beta}, \boldsymbol{\delta}$ is inverted Wishart with parameters scalar $K + n$ and matrix $\mathbf{W} = (K\mathbf{I} + n\mathbf{V})$, where $\mathbf{V} = (1/n) \sum_{i=1}^{n} (\boldsymbol{\beta}_i - \overline{\boldsymbol{\beta}})(\boldsymbol{\beta}_i - \overline{\boldsymbol{\beta}})'$. Train (2001) suggests the following strategy for sampling a matrix from this distribution: Let $\mathbf{M}$ be the lower triangular Cholesky factor of $\mathbf{W}^{-1}$, so $\mathbf{MM}' = \mathbf{W}^{-1}$. Obtain $K + n$ draws of $\mathbf{v}_k = K$ standard normal variates. Then, obtain $\mathbf{S} = \mathbf{M}\left( \sum_{k=1}^{K+n} \mathbf{v}_k \mathbf{v}_k' \right) \mathbf{M}'$. Then $\boldsymbol{\Sigma}^j = \mathbf{S}^{-1}$ is a draw from the inverted Wishart distribution. [This is fairly straightforward, as it involves only random sampling from the standard normal distribution. For a diagonal $\boldsymbol{\Sigma}$ matrix, that is, uncorrelated parameters

[22] The Wishart density is a multivariate counterpart to the chi-squared distribution. Discussion may be found in Zellner (1971, pp. 389–394) and Gelman (2003).

[23] Train describes the use of this method for mixed (random parameters) multinomial logit models. By writing the densities in generic form, we have extended his result to any general setting that involves a parameter vector in the fashion described above. The classical version of this appears in Section 15.11 for the binomial probit model and in Section 18.2.7 for the mixed logit model.

in $\boldsymbol{\beta}_i$, it simplifies a bit further. A draw for the nonzero $k$th diagonal element can be obtained using $(1 + n\mathbf{V}_{kk})/\sum_{k=1}^{K+n} v_{rk}^2$.]

The difficult step is sampling $\boldsymbol{\beta}_i$. For this step, we use the Metropolis–Hastings (M–H) algorithm suggested by Chib and Greenberg (1995, 1996) and Gelman et al. (2004). The procedure involves the following steps:

1. Given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and "tuning constant" $\tau$ (to be described next), compute $\mathbf{d} = \tau\mathbf{L}\mathbf{v}$ where $\mathbf{L}$ is the Cholesky factorization of $\boldsymbol{\Sigma}$ and $\mathbf{v}$ is a vector of $K$ independent standard normal draws.
2. Create a trial value $\boldsymbol{\beta}_{i1} = \boldsymbol{\beta}_{i0} + \mathbf{d}$ where $\boldsymbol{\beta}_{i0}$ is the previous value.
3. The posterior distribution for $\boldsymbol{\beta}_i$ is the likelihood that appears in (16-26) times the joint normal prior density, $\phi_K[\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}]$. Evaluate this posterior density at the trial value $\boldsymbol{\beta}_{i1}$ and the previous value $\boldsymbol{\beta}_{i0}$. Let

$$R_{10} = \frac{f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_{i1})\phi_K(\boldsymbol{\beta}_{i1}|\boldsymbol{\beta}, \boldsymbol{\Sigma})}{f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_{i0})\phi_K(\boldsymbol{\beta}_{i0}|\boldsymbol{\beta}, \boldsymbol{\Sigma})}.$$

4. Draw one observation, $u$, from the standard uniform distribution, $U[0, 1]$.
5. If $u < R_{10}$, then accept the trial (new) draw. Otherwise, reuse the old one.

This M–H iteration converges to a sequence of draws from the desired density. Overall, then, the algorithm uses the Gibbs sampler and the Metropolis–Hastings algorithm to produce the sequence of draws for all the parameters in the model. The sequence is repeated a large number of times to produce each draw from the joint posterior distribution. The entire sequence must then be repeated $N$ times to produce the sample of $N$ draws, which can then be analyzed, for example, by computing the posterior mean.

Some practical details remain. The tuning constant, $\tau$, is used to control the iteration. A smaller $\tau$ increases the acceptance rate. But at the same time, a smaller $\tau$ makes new draws look more like old draws so this slows down the process. Gelman et al. (2004) suggest $\tau = 0.4$ for $K = 1$ and smaller values down to about 0.23 for higher dimensions, as will be typical. Each multivariate draw takes many runs of the MCMC sampler. The process must be started somewhere, though it does not matter much where. Nonetheless, a "burn-in" period is required to eliminate the influence of the starting value. Typical applications use several draws for this burn-in period for each run of the sampler. How many sample observations are needed for accurate estimation is not certain, though several hundred would be a minimum. This means that there is a huge amount of computation done by this estimator. However, the computations are fairly simple. The only complicated step is computation of the acceptance criterion at step 3 of the M–H iteration. Depending on the model, this may, like the rest of the calculations, be quite simple.

### *Example 16.7 Bayesian and Classical Estimation of Heterogeneity in the Returns to Education*

Koop and Tobias (2004) study individual heterogeneity in the returns to education using a panel data set from the National Longitudinal Survey of Youth (NLSY). In a wage equation such as

$$\ln Wage_{it} = \theta_{1,i} + \theta_{2,i}\, Education_{it} + \gamma_1 Experience_{it} + \gamma_2 Experience_{it}^2 + \gamma_3 Time_{it} + \gamma_4 Unemp_{it} + \varepsilon_{it}, \quad \textbf{(16-28)}$$

individual heterogeneity appears in the intercept and in the returns to education. Received estimates of the returns to education, $\theta_2$ here, computed using OLS, are biased due to the

endogeneity of *Education* in the equation. The missing variables would include ability and motivation. Instrumental variable approaches will mitigate the problem (and IV estimators are typically larger than OLS), but the authors are concerned that the results might be specific to the instrument used. They cite the example of using as an instrumental variable a dummy variable for presence of a college in the county of residence, by which the IV estimator will deliver the returns to education for those who attend college given that there is a college in their county, but not for others (the *local average treatment effect* rather than the *average treatment effect*). They propose a structural approach based on directly modeling the heterogeneity. They examine several models including random parameters (continuous variation) and latent class (discrete variation) specifications. They propose extensions of the familiar models by introducing covariates into the heterogeneity model (see Example 15.16) and by exploiting time variation in schooling as part of the identification strategy. Bayesian methods are used for the estimation and inference.[24]

Several models are considered. The one most preferred is the hierarchical linear model examined in Example 15.16:

$$\begin{aligned}\theta_{1,i} &= \theta_{1,0} + \lambda_{1,1}\textit{Ability}_i + \lambda_{1,2}\textit{Mother's Education}_i + \lambda_{1,3}\,\textit{Father's Education}_i + \\ &\quad + \lambda_{1,4}\,\textit{Broken Home}_i + \lambda_{1,5}\,\textit{Siblings}_i + u_{1,i},\\ \theta_{2,i} &= \theta_{2,0} + \lambda_{2,1}\,\textit{Ability}_i + \lambda_{2,2}\,\textit{Mother's Education}_i + \lambda_{2,3}\,\textit{Father's Education}_i \\ &\quad + \lambda_{2,4}\,\textit{Broken Home}_i + \lambda_{2,5}\,\textit{Siblings}_i + u_{2,i}. \end{aligned} \quad \textbf{(16-29)}$$

The candidate models are framed as follows:

| | | |
|---|---|---|
| $y_{it}\vert\mathbf{x}_{it},\mathbf{z}_{it},\boldsymbol{\theta}_i,\boldsymbol{\gamma},\boldsymbol{\sigma}_\varepsilon^2$ | $\sim N[\mathbf{x}_{it}'\boldsymbol{\theta}_i + \mathbf{z}_{it}'\boldsymbol{\gamma}, \boldsymbol{\sigma}_\varepsilon^2]$ | (main regression model), |
| $\boldsymbol{\gamma}\vert\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma$ | $\sim N[\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma]$ | (normal distribution for location parameters), |
| $\sigma_\varepsilon^{-2}\vert\ s_\varepsilon^{-2}, \eta_\varepsilon$ | $\sim G(s_\varepsilon^{-2},\eta_\varepsilon)$ | (gamma distribution for $1/\sigma_\varepsilon^2$), |
| $\boldsymbol{\theta}_i\vert\boldsymbol{\lambda}, \mathbf{w}_i$ | $\sim f(\boldsymbol{\theta}_i\vert\boldsymbol{\lambda},\mathbf{w}_i)$ | (varies by model, discrete or continuous), |
| $\boldsymbol{\lambda}\vert\underline{\boldsymbol{\lambda}}$ | $\sim g(\underline{\boldsymbol{\lambda}})$ | (varies by model). |

The models for $\boldsymbol{\theta}_i\vert\boldsymbol{\lambda},\mathbf{w}_i$ are either discrete or continuous distributions, parameterized in terms of a vector of parameters, $\boldsymbol{\lambda}$ and a vector of time-invariant variables, $w_i$. [*Note,* for example, (16-29).] The model for the regression slopes, $\underline{\boldsymbol{\gamma}}$, and the regression variance, $\sigma_\varepsilon^2$, will be common to all the specifications. The models for the heterogeneity, $\boldsymbol{\theta}_i\vert\boldsymbol{\lambda},\mathbf{w}_i$ and for $\boldsymbol{\lambda}\vert\underline{\boldsymbol{\lambda}}$ will vary with the specification. The models considered are:

**1.** $\theta_{1,i} = \theta_{1,0}$ and $\theta_{2,i} = \theta_{2,0}$, no heterogeneity (−24,212),
**2.** $\boldsymbol{\theta}_i \sim N[\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_u]$, a simple random parameters model (−15,886),
**3.** $\theta_{1,i} \sim N[\theta_1,\sigma_{u1^2}]$, $\theta_{2,i} = \theta_{2,0}$, a random effects model (−16,501),
**4.** $\boldsymbol{\theta}_i = \boldsymbol{\theta}_g^0$ with probability $\pi_g$, a latent class model(−16,528),
**5.** $f(\boldsymbol{\theta}_i) = \sum_g \pi_g\,\phi(\boldsymbol{\theta}_i\vert\boldsymbol{\theta}_g^0, \Sigma_g)$, a finite mixture of normal (−15,898).

(The BIC values for model selection reported in the study are shown in parentheses. These are discussed further below.) The preferred model is model 2 with mean function $\boldsymbol{\theta}_0 + \boldsymbol{\Lambda}\mathbf{w}_i$. This is

[24]The authors note, "Although the length of our panel is rather short, this does not create a significant problem for us as we employ a Bayesian approach which provides exact finite sample results." It is not clear at this point what problem is caused by the short panel—actually, for most of the sample the panel is reasonably long (see Figure 15.7)—or how exact inference mitigates that problem. Likewise, "estimates of the individual-level parameters obtained from our hierarchical model incorporate not only information from the outcomes of that individual, but also incorporate information obtained from the other individuals in the sample." As the authors carefully note later, they do not actually compute individual specific estimates, but rather conditional means for individuals with specific characteristics. (Both from p. 828.)

(16-28) and (16-29). Model 4 could be also augmented with $\mathbf{w}_i$. This would be a latent class model with $prob(\theta_i = \theta_g^0) = \exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)/\Sigma_{g=1}^{G}\exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)$. This model is developed in Section 14.15.2. Estimates based on this latent class formulation are shown below.

The data set is an unbalanced panel of 2,178 individuals, altogether 17,919 person-year observations with $T_i$ ranging from 1 to 15. (See Figure 15.7.) Means of the data are given in Example 15.16.[25] Most of the analysis is based on the full data set. However, models involving the time-invariant variables were estimated using 1,694 individuals (14,170 person-year observations) whose parents have at least 9 years of education. A Gibbs sampler is used with 11,000 repetitions; the first 1,000 are discarded as the burn-in. (The Gibbs sampler, priors, and other computational details are provided in an appendix in the paper.) Two devices are proposed to choose among the models. First, the posterior odds ratio in Section 16.4.3 is computed. With equal priors for the models, the posterior odds equals the likelihood ratio, which is computed for two models, $A$ and $B$, as $\exp(\ln L^A - \ln L^B)$. The log likelihoods for models 1, 2, and 3 are −12,413, −8,046, and −8,153. Small differences in the log likelihoods always translate to huge differences in the posterior odds. For these cases, the posterior odds in favor of model 2 against model 3 are exp(107), which is overwhelming ("massive"). (The log likelihood for the version of model 2 in Example 15.16 is −7,983, which is also vastly better than the model 2 here by this criterion.) A second criterion is the Bayesian information criterion, which is $2\ln L - K\ln n$, where $K$ is the number of parameters estimated and $n$ is the number of individuals (2,170 or 1,694). The BICs for the five models are listed above with the model specifications. The model with no heterogeneity is clearly rejected. Among the others, Model 2, the random parameters specification, is preferred by a wide margin. Model 5, the mixture of two normal distributions with heterogeneous means, is second, followed by Model 3, the random effects model. Model 4, the latent class model, is clearly the least preferred.

**Continuous Distribution of Heterogeneity**

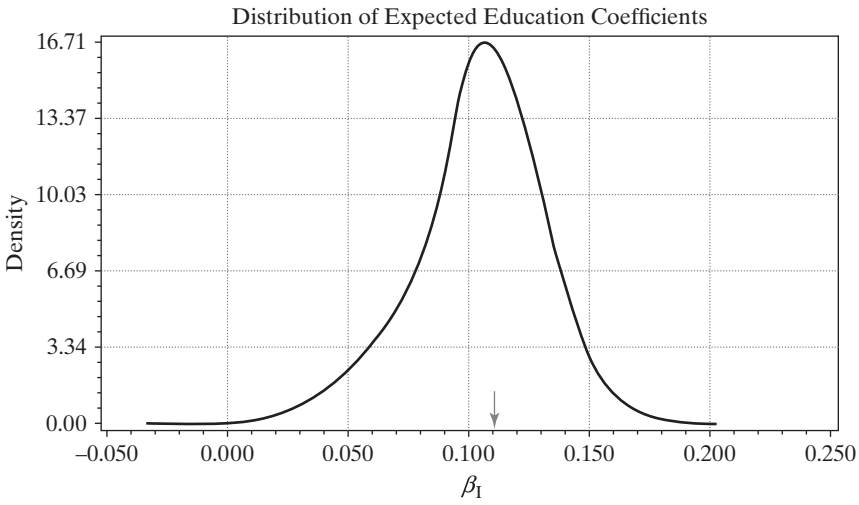
The main results for the study are based on the subsample and Model 2. The reported posterior means of the coefficient distributions of (16-29) are shown in the right panel in Table 16.4. (Results are extracted from Tables IV and V in the paper.) We re-estimated (16-28) and (16-29) using the methods of Sections 15.7 and 15.8. The estimates of the parameters

**TABLE 16.4** Estimated Wage Equations

| | *Random Parameters Model* | | *Koop–Tobias Posterior Means* | |
|---|---|---|---|---|
| ***Variable*** | ***Constant*** | ***Education*** | ***Constant*** | ***Education*** |
| *Exp* | 0.12621 | | 0.126 | |
| $Exp^2$ | −0.00388 | | −0.004 | |
| *Time* | −0.01787 | | −0.024 | |
| *Unemployment* | | | −0.004 | |
| *Constant* | 0.39277 | 0.09578 | 0.797 | 0.070 |
| *Ability* | −0.13177 | 0.01568 | −0.073 | 0.0125 |
| *Mother's Educ* | 0.02864 | −0.00167 | 0.021 | −0.001 |
| *Father's Educ* | 0.00242 | −0.00022 | −0.022 | 0.002 |
| *Broken Home* | 0.12963 | −0.01640 | 0.115 | −0.015 |
| *Number Siblings* | −0.08323 | 0.00659 | −0.079 | 0.007 |

[25]The unemployment rate variable in (16-28) is not included in the JAE archive data set that we have used to partially replicate this study in Example 15.16 and here.

**FIGURE 16.1** Random Parameters Estimate of Expected Returns.




of the model are shown in the left panel of Table 16.4. Overall, the mean return is about 11% (0.11). We did the same analysis with the classical results based on Section 15.10. The individual specific estimates are summarized in Figure 16.1 (which is nearly identical to the authors' Figure 3). The results are essentially the same as Koop and Tobias's. The differences are attributable to the different methodologies – the prior distributions will have at least some influence on the results – and to our omission of the unemployment rate from the main equation. The authors' reported results suggest that the impact of the unemployment rate on the results is minor, which would suggest that the differences in the estimated results primarily reflect the different approaches to the analysis. The similarity of the end results would be anticipated by the Bernstein–von Mises theorem. (See Section 16.4.4.)

**Discrete Distribution of Heterogeneity**

Model 4 in the study is a latent class model. The authors fit a model with $G = 10$ classes. The model is a Heckman and Singer style (Section 14.15.7) specification in that the coefficients on the time-varying variables are the same in all 10 classes. The class probabilities are specified as fixed constants. This provides a discrete distribution for the heterogeneity in $\boldsymbol{\theta}_i$. Model 4 was the least preferred model among the candidates.

We fit a 5 segment latent class model based on (16-28) and (16-29). The parameters on the time-varying variables in (16-28) are the same in all classes—only the constant terms and the education coefficients differ across the classes. The class probabilities are built on the time-invariant effects, ability, parent's education, etc. (The authors do not report a model with this form of heterogeneity.) The log likelihood for this extension of the model is

$$\ln L = \sum_{i=1}^{n} \ln \sum_{g=1}^{G} \pi_{ig}(\mathbf{w}_i)\left(\prod_{t=1}^{T_i} f(y_{it}|\theta_{0,g} + \theta_{1,g}\, Education_{it} + \mathbf{z}_{it}'\boldsymbol{\gamma})\right) \qquad \textbf{(16-30)}$$

$$\pi_{ig}(\mathbf{w}_i) = \frac{\exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)}{\sum_{g=1}^{G}\exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)}.$$

Using the suggested subsample, the log likelihood for the model in (16-30) is 6235.02. When the time-invariant variables are not included in the class probabilities, the log likelihood falls to 6192.66. By a standard likelihood ratio test, the chi squared is 84.72, with 20 degrees of freedom (the 5 additional coefficients in G-1 of the class probabilities). The critical chi squared is 31.02. We computed $E[\theta_{1,i} | \mathbf{data}]$ for each individual based on the estimated posterior class probabilities as

$$\hat{E}[\theta_{1,i}] = \Sigma_{g=1}^{G} \hat{\pi}_{ig}(\theta_{1,g} | \mathbf{w}_i, \mathbf{data}_i)\hat{\theta}_{1,g}. \quad \textbf{(16-31)}$$

(See Section 14.15.4.) The overall estimate of returns to education is the sample average of these, 0.107. Figure 16.2 shows the results of this computation for the 1,694 individuals. We then used the method in Section 14.15.4. to estimate the class assignments and computed the means of the expected returns for the individuals assigned to each of the 5 classes. The results are shown in Table 16.5. Finally, because we now have a complete (estimated) assignment of the individuals, we constructed in Figure 16.3 a comparison of distributions of the expected coefficients in each of the 5 classes.

This analysis has examined the heterogeneity in the returns to education by a variety of model specifications. In the end, the results are quite consistent across the different models and based on the two methodologies.

## 16.9 SUMMARY AND CONCLUSIONS

This chapter has introduced the major elements of the Bayesian approach to estimation and inference. The contrast between Bayesian and classical, or frequentist, approaches to the analysis has been the subject of a decades-long dialogue among

**FIGURE 16.2** Estimated Distribution of Expected Returns Based on Latent Class Model.

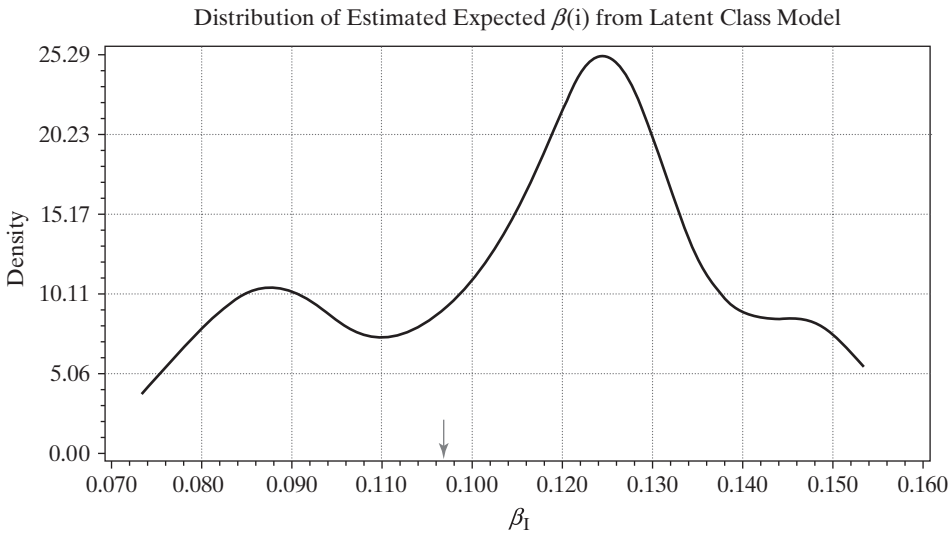
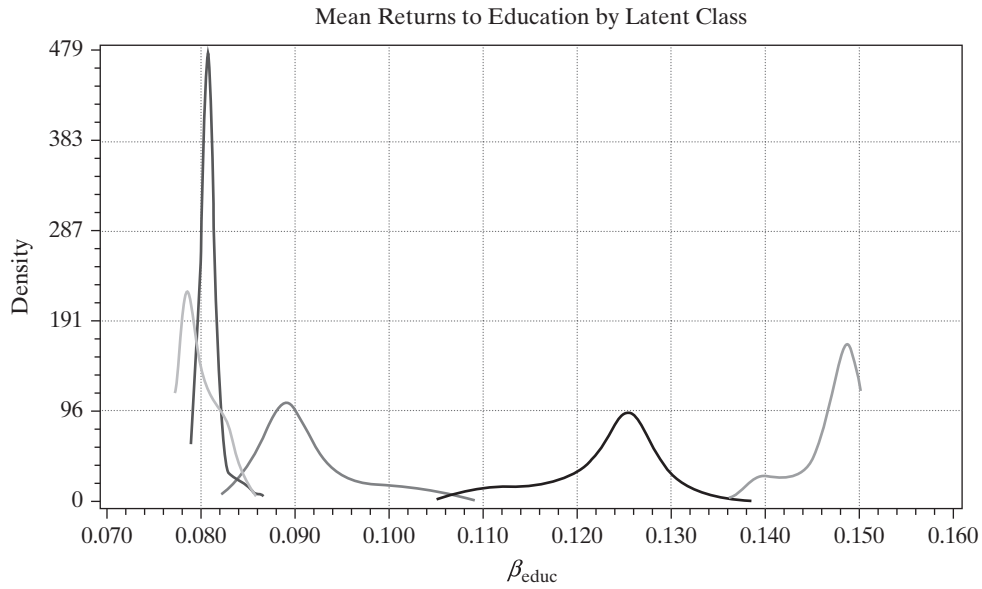


**TABLE 16.5** Estimated Expected Returns to Schooling by Class

| *Class* | *Mean* | $n_g$ |
|---|---|---|
| 1 | 0.147 | 167 |
| 2 | 0.092 | 608 |
| 3 | 0.080 | 189 |
| 4 | 0.123 | 640 |
| 5 | 0.083 | 90 |
| Full Sample | 0.107 | 1,694 |

**FIGURE 16.3** Kernel Density Estimates of Expected Returns by Class.




practitioners and philosophers. As the frequency of applications of Bayesian methods has grown dramatically in the modern literature, however, the approach to the body of techniques has typically become more pragmatic. The Gibbs sampler and related techniques including the Metropolis–Hastings algorithm have enabled some remarkable simplifications of previously intractable problems. For example, recent developments in commercial software have produced a wide choice of mixed estimators which are various implementations of the maximum likelihood procedures and hierarchical Bayes procedures (such as the *Sawtooth and MLWin* programs). Unless one is dealing with a small sample, the choice between these can be based on convenience. There is little methodological difference. This returns us to the practical point noted earlier. The choice between the Bayesian approach and the sampling theory method in this application would not be based on a fundamental methodological criterion, but on purely practical considerations—the end result is largely the same.

This chapter concludes our survey of estimation and inference methods in econometrics. We will now turn to two major areas of applications, microeconometrics in Chapters 17–19. which is primarily oriented to cross-section and panel data applications, and time series and (broadly) macroeconometrics in Chapters 20 and 21.

## Key Terms and Concepts

- Bayes factor
- Bayes' theorem
- Bernstein–von Mises theorem
- Burn in
- Conjugate prior
- Data augmentation
- Gibbs sampler
- Hierarchical prior
- Highest posterior density (HPD) interval
- Improper prior
- Informative prior
- Inverted gamma distribution
- Inverted Wishart
- Joint posterior distribution
- Likelihood function
- Loss function
- Markov chain Monte Carlo (MCMC)
- Metropolis–Hastings algorithm
- Multivariate *t* distribution
- Noninformative prior
- Normal-gamma prior
- Posterior density
- Posterior mean
- Precision matrix
- Predictive density
- Prior beliefs
- Prior density
- Prior distribution
- Prior odds ratio
- Prior probabilities
- Sampling theory
- Uniform-inverse gamma prior
- Uniform prior

## Exercise

1. Suppose the distribution of $y_i|\lambda$ is Poisson,

$$f(y_i|\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} = \frac{\exp(-\lambda)\lambda^{y_i}}{\Gamma(y_i + 1)}, \quad y_i = 0, 1, \ldots, \lambda > 0.$$

We will obtain a sample of observations, $y_i, \ldots, y_n$. Suppose our prior for $\lambda$ is the inverted gamma, which will imply

$$p(\lambda) \propto \frac{1}{\lambda}.$$

a. Construct the likelihood function, $p(y_1, \ldots, y_n|\lambda)$.
b. Construct the posterior density,

$$p(\lambda|y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n|\lambda)p(\lambda)}{\int_0^\infty p(y_1, \ldots, y_n|\lambda)p(\lambda)d\lambda}.$$

c. Prove that the Bayesian estimator of $\lambda$ is the posterior mean, $E[\lambda|y_1, \ldots, y_n] = \bar{y}$.
d. Prove that the posterior variance is $\text{Var}[\lambda|y_l, \ldots, y_n] = \bar{y}/n$.
(*Hint:* You will make heavy use of gamma integrals in solving this problem. Also, you will find it convenient to use $\Sigma_i y_i = n\bar{y}$.)

## Applications

1. Consider a model for the mix of male and female children in families. Let $K_i$ denote the family size (number of children), $K_i = 1, \ldots$. Let $F_i$ denote the number of female children, $F_i = 0, \ldots, K_i$. Suppose the density for the number of female children in a family with $K_i$ children is binomial with constant success probability $\theta$:

$$p(F_i | K_i, \theta) = \binom{K_i}{F_i} \theta^{F_i} (1 - \theta)^{K_i - F_i}.$$

We are interested in analyzing the "probability," $\theta$. Suppose the (conjugate) prior over $\theta$ is a beta distribution with parameters $a$ and $b$:

$$p(\theta) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1}.$$

Your sample of 25 observations is given here:

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K_i$ | 2 | 1 | 1 | 5 | 5 | 4 | 4 | 5 | 1 | 2 | 4 | 4 | 2 | 4 | 3 | 2 | 3 | 2 | 3 | 5 | 3 | 2 | 5 | 4 | 1 |
| $F_i$ | 1 | 1 | 1 | 3 | 2 | 3 | 2 | 4 | 0 | 2 | 3 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 2 | 4 | 2 | 1 | 1 | 4 | 1 |

   a. Compute the classical maximum likelihood estimate of $\theta$.
   b. Form the posterior density for $\theta$ given $(K_i, F_i)$, $i = 1, \ldots, 25$ conditioned on $a$ and $b$.
   c. Using your sample of data, compute the posterior mean assuming $a = b = 1$.
   d. Using your sample of data, compute the posterior mean assuming $a = b = 2$.
   e. Using your sample of data, compute the posterior mean assuming $a = 1$ and $b = 2$.