16

ESTIMATION FRAMEWORKS IN ECONOMETRICS

16.1 INTRODUCTION

This chapter begins our treatment of methods of estimation. Contemporary econometrics offers the practitioner a remarkable variety of estimation methods, ranging from tightly parameterized likelihood based techniques at one end to thinly stated nonparametric methods that assume little more than mere association between variables at the other, and a rich variety in between. Even the experienced researcher could be forgiven for wondering how they should choose from this long menu. It is certainly beyond our scope to answer this question here, but a few principles can be suggested. Recent research has leaned when possible toward methods that require few (or fewer) possibly unwarranted or improper assumptions. This explains the ascendance of the GMM estimator in situations where strong likelihood-based parameterizations can be avoided and robust estimation can be done in the presence of heteroscedasticity and serial correlation. (It is intriguing to observe that this is occurring at a time when advances in computation have helped bring about *increased* acceptance of very heavily parameterized Bayesian methods.)

As a general proposition, the progression from full to semi- to non-**parametric estimation** relaxes strong assumptions, but at the cost of weakening the conclusions that can be drawn from the data. As much as anywhere else, this is clear in the analysis of discrete choice models, which provide one of the most active literatures in the field. (A sampler appears in Chapter 21.) A formal probit or logit model allows estimation of probabilities, marginal effects, and a host of ancillary results, but at the cost of imposing the normal or logistic distribution on the data. **Semiparametric** and **nonparametric estimators** allow one to relax the restriction, but often provide, in return, only ranges of probabilities, if that, and in many cases, preclude estimation of probabilities or useful marginal effects. One does have the virtue of robustness in the conclusions, however. [See, e.g., the symposium in Angrist (2001) for a spirited discussion on these points.]

Estimation properties is another arena in which the different approaches can be compared. Within a class of estimators, one can define "the best" (most efficient) means of using the data. (See Example 16.2 below for an application.) Sometimes comparisons can be made across classes as well. For example, when they are estimating the same parameters—this remains to be established—the best parametric estimator will generally outperform the best semiparametric estimator. That is the value of the information, of course. The other side of the comparison, however, is that the semiparametric estimator will carry the day if the parametric model is misspecified in a fashion to which the semiparametric estimator is robust (and the parametric model is not).

Schools of thought have entered this conversation for a long time. Proponents of **Bayesian estimation** often took an almost theological viewpoint in their criticism of their classical colleagues. [See, for example, Poirier (1995).] Contemporary practitioners are usually more pragmatic than this. Bayesian estimation has gained currency as a set of techniques that can, in very many cases, provide both elegant and tractable solutions to problems that have heretofore been out of reach. Thus, for example, the **simulation-based estimation** advocated in the many papers of Chib and Greenberg (e.g., 1996) have provided solutions to a variety of computationally challenging problems.[1] Arguments as to the methodological virtue of one approach or the other have received much less attention than before.

Chapters 2 though 9 of this book have focused on the classical regression model and a particular estimator, least squares (linear and nonlinear). In this and the next two chapters, we will examine several general estimation strategies that are used in a wide variety of situations. This chapter will survey a few methods in the three broad areas we have listed, including Bayesian methods. Chapter 17 presents the method of **maximum likelihood,** the broad platform for parametric, classical estimation in econometrics. Chapter 18 discusses the **generalized method of moments,** which has emerged as the centerpiece of semiparametric estimation. Sections 16.2.4 and 17.8 will examine two specific estimation frameworks, one Bayesian and one classical, that are based on simulation methods. This is a recently developed body of techniques that have been made feasible by advances in estimation technology and which has made quite straightforward many estimators which were previously only scarcely used because of the sheer difficulty of the computations.

The list of techniques presented here is far from complete. We have chosen a set that constitute the mainstream of econometrics. Certainly there are others that might be considered. [See, for example, Mittelhammer, Judge, and Miller (2000) for a lengthy catalog.] Virtually all of them are the subject of excellent monographs on the subject. In this chapter we will present several applications, some from the literature, some home grown, to demonstrate the range of techniques that are current in econometric practice. We begin in Section 16.2 with parametric approaches, primarily maximum likelihood. Since this is the subject of much of the remainder of this book, this section is brief. Section 16.2 also presents Bayesian estimation, which in its traditional form, is as heavily parameterized as maximum likelihood estimation. This section focuses mostly on the **linear model.** A few applications of Bayesian techniques to other models are presented as well. We will also return to what is currently the standard toolkit in Bayesian estimation, **Markov Chain Monte Carlo** methods in Section 16.2.4. Section 16.2.3 presents an emerging technique in the classical tradition, **latent class** modeling, which makes interesting use of a fundamental result based on Bayes Theorem. Section 16.3 is on semiparametric estimation. GMM estimation is the subject of all of Chapter 18, so it is

[1]The penetration of Bayesian econometrics could be overstated. It is fairly well represented in the current journals such as the *Journal of Econometrics, Journal of Applied Econometrics, Journal of Business and Economic Statistics,* and so on. On the other hand, in the six major general treatments of econometrics published in 2000, four (Hayashi, Ruud, Patterson, Davidson) do not mention Bayesian methods at all, a buffet of 32 essays (Baltagi) devotes only one to the subject, and the one that displays any preference (Mittelhammer et al.) devotes nearly 10 percent (70) of its pages to Bayesian estimation, but all to the broad metatheory or the linear regression model and none to the more elaborate applications that form the received applications in the many journals in the field.

only introduced here. The technique of least absolute deviations is presented here as well. A range of applications from the recent literature is also surveyed. Section 16.4 describes nonparametric estimation. The fundamental tool, the kernel density estimator is developed, then applied to a problem in regression analysis. Two applications are presented here as well. Being focused on application, this chapter will say very little about the statistical theory for of these techniques—such as their asymptotic properties. (The results are developed at length in the literature, of course.) We will turn to the subject of the properties of estimators briefly at the end of the chapter, in Section 16.5, then in greater detail in Chapters 17 and 18.

16.2 PARAMETRIC ESTIMATION AND INFERENCE

Parametric estimation departs from a full statement of the **density** or probability model that provides the **data generating mechanism** for a random variable of interest. For the sorts of applications we have considered thus far, we might say that the joint density of a scalar random variable, "y" and a random vector, "$\mathbf{x}$" of interest can be specified by

$$f(y, \mathbf{x}) = g(y \mid \mathbf{x}, \boldsymbol{\beta}) \times h(\mathbf{x} \mid \boldsymbol{\theta}) \tag{16-1}$$

with unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. To continue the application that has occupied us since Chapter 2, consider the linear regression model with normally distributed disturbances. The assumption produces a full statement of the **conditional density** that is the population from which an observation is drawn;

$$y_i \mid \mathbf{x}_i \sim N[\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2].$$

All that remains for a full definition of the population is knowledge of the specific values taken by the *unknown* but *fixed* parameters. With those in hand, the conditional probability distribution for y_i is completely defined—mean, variance, probabilities of certain events, and so on. (The marginal density for the conditioning variables is usually not of particular interest.) Thus, the signature features of this modeling platform are specification of both the density and the features (parameters) of that density.

The **parameter space** for the parametric model is the set of allowable values of the parameters which satisfy some prior specification of the model. For example, in the regression model specified previously, the K regression slopes may take any real value, but the variance must be a positive number. Therefore, the parameter space for that model is $[\boldsymbol{\beta}, \sigma^2] \in \mathbb{R}^K \times \mathbb{R}_+$. "Estimation" in this context consists of specifying a criterion for ranking the points in the parameter space, then choosing that point (a point estimate) or a set of points (an interval estimate) that optimizes that criterion, that is, has the best ranking. Thus, for example, we chose linear least squares as one **estimation criterion** for the linear model. "Inference" in this setting is a process by which some regions of the (already specified) parameter space are deemed not to contain the unknown parameters, though, in more practical terms, we typically define a criterion and then, state that, by that criterion, certain regions are *unlikely* to contain the true parameters.

16.2.1 CLASSICAL LIKELIHOOD BASED ESTIMATION

The most common (by far) class of parametric estimators used in econometrics is the maximum likelihood estimators. The underlying philosophy of this class of estimators is the idea of "sample information." When the density of a sample of observations is completely specified, apart from the unknown parameters, then the joint density of those observations (assuming they are independent), is the likelihood function,

$$f(y_1, y_2, \ldots, \mathbf{x}_1, \mathbf{x}_2, \ldots) = \prod_{i=1}^{n} f(y_i, \mathbf{x}_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}), \tag{16-2}$$

This function contains all the information available in the sample about the population from which those observations were drawn. The strategy by which that information is used in estimation constitutes the estimator.

The **maximum likelihood estimator** [Fisher (1925)] is that function of the data which (as its name implies) maximizes the likelihood function (or, because it is usually more convenient, the log of the likelihood function). The motivation for this approach is most easily visualized in the setting of a discrete random variable. In this case, the likelihood function gives the joint probability for the observed sample observations, and the maximum likelihood estimator is the function of the sample information which makes the observed data most probable (at least by that criterion). Though the analogy is most intuitively appealing for a discrete variable, it carries over to continuous variables as well. Since this estimator is the subject of Chapter 17, which is quite lengthy, we will defer any formal discussion until then, and consider instead two applications to illustrate the techniques and underpinnings.

Example 16.1 The Linear Regression Model

Least squares weighs negative and positive deviations equally and gives disproportionate weight to large deviations in the calculation. This property can be an advantage or a disadvantage, depending on the data-generating process. For normally distributed disturbances, this method is precisely the one needed to use the data most efficiently. If the data are generated by a normal distribution, then the log of the likelihood function is

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

You can easily show that least squares is the estimator of choice for this model. Maximizing the function means minimizing the exponent, which is done by least squares for $\boldsymbol{\beta}$ and $\mathbf{e}'\mathbf{e}/n$ for σ^2.

If the appropriate distribution is deemed to be something other than normal—perhaps on the basis of an observation that the tails of the disturbance distribution are too thick—see Example 5.1 and Section 17.6.3—then there are three ways one might proceed. First, as we have observed, the consistency of least squares is robust to this failure of the specification, so long as the conditional mean of the disturbances is still zero. Some correction to the standard errors is necessary for proper inferences. (See Section 10.3.) Second, one might want to proceed to an estimator with better finite sample properties. The least absolute deviations estimator discussed in Section 16.3.2 is a candidate. Finally, one might consider some other distribution which accommodates the observed discrepancy. For example, Ruud (2000) examines in some detail a linear regression model with disturbances distributed according to the t distribution with v degrees of freedom. As long as v is finite, this random variable will have a larger variance than the normal. Which way should one proceed? The third approach is the least appealing. Surely if the normal distribution is inappropriate, then it would be difficult to come up with a plausible mechanism whereby the t distribution would not be. The LAD estimator might well be preferable if the sample were small. If not, then least

squares would probably remain the estimator of choice, with some allowance for the fact that standard inference tools would probably be misleading. Current practice is generally to adopt the first strategy.

Example 16.2 The Stochastic Frontier Model

The **stochastic frontier** model, discussed in detail in Section 17.6.3, is a regression-like model with a disturbance that is asymmetric and distinctly nonnormal. (See Figure 17.3.) The conditional density for the dependent variable in this model is

$$f(y \mid \mathbf{x}, \boldsymbol{\beta}, \sigma, \lambda) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left[\frac{-(y - \alpha - \mathbf{x}'\boldsymbol{\beta})^2}{2\sigma^2}\right] \Phi\left(\frac{-\lambda(y - \alpha - \mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)$$

This produces a log-likelihood function for the model,

$$\ln L = -n \ln \sigma - \frac{n}{2}\ln\frac{2}{\pi} - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma}\right)^2 + \sum_{i=1}^{n}\ln \Phi\left(\frac{-\varepsilon_i \lambda}{\sigma}\right)$$

There are at least two fully parametric estimators for this model. The maximum likelihood estimator is discussed in Section 17.6.3. Greene (1997b) presents the following **method of moments** estimator: For the regression slopes, excluding the constant term, use least squares. For the parameters α, σ, and λ, based on the second and third moments of the least squares residuals and least squares constant, solve

$$m_2 = \sigma_v^2 + [1 - 2/\pi]\sigma_u^2$$
$$m_3 = (2/\pi)^{1/2}[1 - 4/\pi]\sigma_u^3$$
$$a = \alpha + (2/\pi)^2\sigma_u$$

where $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$.

Both estimators are fully parametric. The maximum likelihood estimator is for the reasons discussed earlier. The method of moments estimators (see Section 18.2) are appropriate only for this distribution. Which is preferable? As we will see in Chapter 17, both estimators are consistent and asymptotically normally distributed. By virtue of the Cramér–Rao theorem, the maximum likelihood estimator has a smaller asymptotic variance. Neither has any small sample optimality properties. Thus, the only virtue of the method of moments estimator is that one can compute it with any standard regression/statistics computer package and a hand calculator whereas the maximum likelihood estimator requires specialized software (only somewhat—it is reasonably common).

16.2.2 BAYESIAN ESTIMATION

Parametric formulations present a bit of a methodological dilemma. They would seem to straightjacket the researcher into a fixed and immutable specification of the model. But in any analysis, there is uncertainty as to the magnitudes and even, on occasion, the signs of coefficients. 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 one of continually updating and sharpening our subjective beliefs about the state of the world.

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

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

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 the way that the joint density for our parametric model is specified in (16-1).] Thus, we write

$$\begin{aligned} P(\text{parameters} \mid \text{data}) &\propto P(\text{data} \mid \text{parameters})\,P(\text{parameters}) \\ &= \text{Likelihood function} \times \text{Prior density}. \end{aligned}$$

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 (16-1). 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 16.2.4.

16.2.2.a 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.[2]

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

[2]These sources include Judge et al. (1982, 1985), Maddala (1977a), Mittelhammer et al. (2000), and the canonical reference for econometricians, Zellner (1971). Further topics in Bayesian inference are contained in Zellner (1985). A recent treatment of both Bayesian and sampling theory approaches is Poirier (1995).

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 σ^2 given the data,** so

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

For purposes of the results below, 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

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

This density embodies all that we have to learn about the parameters from the observed data. Since the data are taken to be constants in the joint density, we may multiply this joint density by the (very carefully chosen), inessential (since it does not involve $\boldsymbol{\beta}$ or σ^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 $L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X})$ by A gives

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

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 $\boldsymbol{\beta} \mid \sigma^2$ with mean vector $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The reason will be clear shortly.

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:[3]

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

[3]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.

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. Since 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 σ^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,

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

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

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

where

$$h(\sigma^2) = \frac{[vs^2]^{v+1}}{\Gamma(v+1)} \left[\frac{1}{\sigma^2}\right]^v e^{-vs^2(1/\sigma^2)}. \quad \textbf{(16-6)}$$

For the present, we treat $h(\sigma^2)$ simply as a constant that involves σ^2, not as a probability density; (16-5) is *conditional* on σ^2. Thus, the posterior density $f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X})$ 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.

To generalize the preceding to an unknown σ^2, we specify a noninformative prior distribution for $\ln \sigma$ over the entire real line.[4] By the change of variable formula, if $g(\ln \sigma)$ is constant, then $g(\sigma^2)$ is proportional to $1/\sigma^2$.[5] Assuming that $\boldsymbol{\beta}$ and σ^2 are independent, we now have the noninformative joint prior distribution:

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

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

[5]Many treatments of this model use σ rather than σ^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.

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

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

For the same reason as before, we multiply $g_{\sigma^2}(\sigma^2)$ by a well-chosen constant, this time $vs^2\Gamma(v+1)/\Gamma(v+2) = vs^2/(v+1)$. 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 σ^2, given $\mathbf{y}$ and $\mathbf{X}$:

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

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

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

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

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

so the marginal posterior distribution for $\boldsymbol{\beta}$ is

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

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 ∞ is the negative of the integral from $(1/\sigma^2) = 0$ to ∞—we obtain

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

Reinserting the expressions for A, P, and λ produces

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

This density is proportional to a **multivariate *t* distribution**[6] 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 kth diagonal element of the covariance matrix given earlier. Once again, this is the same as our sampling theory. The difference is a matter of interpretation. In the current context, the estimated distribution is for $\boldsymbol{\beta}$ and is centered at $\mathbf{b}$.

16.2.2.b 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 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 $\mathbf{b}$, 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 σ^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 below.

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

$$E_{\boldsymbol{\beta}}[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})] = \int_{\boldsymbol{\beta}} H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) d\boldsymbol{\beta}, \tag{16-9}$$

where the weighting function is the marginal posterior density. (The joint density for $\boldsymbol{\beta}$ and σ^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,

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

To minimize this, we can use the result that

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

[6] 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 $v + K/2 = n/2$.

The minimum is found by equating this derivative to $\mathbf{0}$, whence, since $-\mathbf{W}$ is irrelevant, $\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}]$. This kind of loss function would state that errors in the positive and negative direction 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.

16.2.2.c 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 $\mathbf{b}$, with terminal values taken from the standard t tables.

16.2.2.d 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 σ^2 out of the posterior in (16-5) is complicated by itself. It is made much more so if the prior distributions of $\boldsymbol{\beta}$ and σ^2 are at all involved. Partly to offset these difficulties, researchers usually use what is called a **conjugate prior,** which is one that has the same form as the conditional density and is therefore amenable to the integration needed to obtain the marginal distributions.[7]

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 σ^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 σ^2 and the data will be normal with

$$\begin{aligned} E[\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}] &= \left\{\boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\right\}^{-1}\left\{\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\right\} \\ &= \mathbf{F}\boldsymbol{\beta}_0 + (\mathbf{I} - \mathbf{F})\mathbf{b}, \end{aligned} \tag{16-10}$$

where

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

[7] Our choice of noninformative prior for $\ln \sigma$ led to a convenient prior for σ^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.

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 covariance matrices.[8] The smaller the variance of the estimator, the larger its weight, which makes sense. Also, still taking σ^2 as known, we can write the variance of the posterior normal distribution as

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

Notice that the posterior variance combines the prior and conditional variances on the basis of their inverses.[9] 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 σ^2. If our prior over σ^2 is to be informative as well, then the resulting distribution can be extremely cumbersome. A conjugate prior for $\boldsymbol{\beta}$ and σ^2 that can be used is

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

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

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

This distribution is an **inverted gamma distribution.** It implies that $1/\sigma^2$ has a gamma distribution. The prior mean for σ^2 is σ_0^2 and the prior variance is $\sigma_0^4/(m-1)$.[10] 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 σ^2, we would obtain the prior marginal for $\boldsymbol{\beta}$ alone, which would be a multivariate t distribution.[11] Combining (16-12) with (16-13) produces the joint posterior distribution for $\boldsymbol{\beta}$ and σ^2. Finally, the marginal posterior distribution for $\boldsymbol{\beta}$ is obtained by integrating out σ^2. It has been shown that this posterior distribution is multivariate t with

$$E[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}] = \left\{[\bar{\sigma}^2\mathbf{A}]^{-1} + [\bar{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\right\}^{-1}\left\{[\bar{\sigma}^2\mathbf{A}]^{-1}\boldsymbol{\beta}^0 + [\bar{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\right\} \tag{16-14}$$

and

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

where j is a degrees of freedom parameter and $\bar{\sigma}^2$ is the Bayesian estimate of σ^2. The prior degrees of freedom m is a parameter of the prior distribution for σ^2 that would have been determined at the outset. (See the following example.) Once again, it is clear

[8] Note that it will not follow that individual elements of the posterior mean vector lie between those of $\boldsymbol{\beta}^0$ and $\mathbf{b}$. See Judge et al. (1985, pp. 109–110) and Chamberlain and Leamer (1976).

[9] 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 will see in Chapter 17.

[10] 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 λ and P are $m\sigma_0^2$ and $m+1$.

[11] 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

that as the amount of data increases, the posterior density, and the estimates thereof, converge to the sampling theory results.

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. A classical 95 percent confidence interval for β based on these estimates is (0.8780, 1.2818). (The very wide interval probably results from the obviously poor specification of the model.) Based on noninformative priors for β and σ^2, we would estimate the posterior density for β to be univariate t with 9 degrees of freedom, with mean 0.6848014 and variance $(11/9)0.061878 = 0.075628$. An HPD interval for β would coincide with the confidence interval. Using the fourth quarter (yearly) values of the 1950–2000 data used in Example 6.3, 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 β and obtain a posterior density for β based on an informative prior instead. We assume for this exercise that σ^2 may be taken as known at the sample value of 29.954. Then,

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

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 0.000071164. 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, since 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.)

16.2.2.e 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 6 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 this is what the sampling methodology mandates. (Note, for example, the uncomfortable dilemma this creates in footnote 24 in Chapter 14.) The Bayesian approach to hypothesis testing is much more appealing in this regard. Indeed, the approach might be more appropriately called "comparing hypotheses," since 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.[12] 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," since 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 (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

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

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:

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \tag{16-17}$$

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

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

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

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

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

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

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:

$$\text{Model 0: } y = \mathbf{x}_0'\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0$$

and

$$\text{Model 1: } y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1.$$

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

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 σ_0^2 and σ_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 (16-10).] Zellner (p. 310) discusses this assumption at some length. With these two assumptions, he shows that as n grows large,[13]

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

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" (i.e., the relevant part of $\boldsymbol{\beta}_0 = \mathbf{0}$, so $R_0^2 = 0$), then the ratio simplifies to

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

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 σ_j^2. Specifically, they use the 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

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

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.

16.2.3 USING BAYES THEOREM IN A CLASSICAL ESTIMATION PROBLEM: THE LATENT CLASS MODEL

Latent class modeling can be viewed as a means of modeling heterogeneity across individuals in a random parameters framework. We first encountered random parameters models in Section 13.8 in connection with panel data.[14] As we shall see, the latent class model provides an interesting hybrid of classical and Bayesian analysis. To define the *latent class model,* we begin with a random parameters formulation of the density of an observed random variable. We will assume that the data are a panel. Thus, the density of y_{it} when the parameter vector is $\boldsymbol{\beta}_i$ is $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i)$. The parameter vector $\boldsymbol{\beta}_i$ is randomly distributed over individuals according to

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i \qquad \textbf{(16-18)}$$

and where $\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i$ is the mean of the distribution, which depends on time invariant individual characteristics as well as parameters yet to be estimated, and the random

[13] 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.)

[14] In principle, the latent class model does not require panel data, but practical experience suggests that it does work best when individuals are observed more than once and is difficult to implement in a cross section.

variation comes from the individual heterogeneity, $\mathbf{v}_i$. This random vector is assumed to have mean zero and covariance matrix, $\boldsymbol{\Sigma}$. The conditional density of the parameters is

$$g(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = g(\mathbf{v}_i + \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i, \boldsymbol{\Sigma}),$$

where $g(.)$ is the underlying marginal density of the heterogeneity. The unconditional density for y_{it} is obtained by integrating over $\mathbf{v}_i$,

$$f(y_{it} \mid \mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = E_{\boldsymbol{\beta}i}[f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i)] = \int_{\mathbf{v}_i} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) g(\mathbf{v}_i + \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i, \boldsymbol{\Sigma}) d\mathbf{v}_i.$$

This result would provide the density that would enter the likelihood function for estimation of the model parameters. We will return to this model formulation in Chapter 17.

The preceding has assumed $\boldsymbol{\beta}_i$ has a continuous distribution. Suppose that $\boldsymbol{\beta}_i$ is generated from a discrete distribution with J values, or classes, so that the distribution of $\boldsymbol{\beta}$ is over these J vectors.[15] Thus, the model states that an individual belongs to one of the J latent classes, but it is unknown from the sample data exactly which one. We will use the sample data to estimate the probabilities of class membership. The corresponding model formulation is now

$$f(y_{it} \mid \mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\Delta}) = \sum_{j=1}^{J} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i) f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j)$$

where it remains to parameterize the class probabilities, p_{ij} and the structural model, $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j)$. The matrix $\boldsymbol{\Delta}$ contains the parameters of the discrete distribution. It has J rows (one for each class) and M columns for the M variables in $\mathbf{z}_i$. (The structural mean and variance parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are no longer necessary.) At a minimum, $M = 1$ and $\mathbf{z}_i$ contains a constant, if the class probabilities are fixed parameters. Finally, in order to accommodate the panel data nature of the sampling situation, we suppose that conditioned on $\boldsymbol{\beta}_j$, observations y_{it}, $t = 1, \ldots, T$ are independent. Therefore, for a group of T observations, the joint density is

$$f(y_{i1}, y_{i2}, \ldots, y_{iT} \mid \boldsymbol{\beta}_j, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}) = \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j).$$

(We will consider models that provide correlation across observations in Chapters 17 and 21.) Inserting this result in the earlier density produces the likelihood function for a panel of data,

$$\ln L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{M} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i) \prod_{t=1}^{T} g(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j) \right].$$

The class probabilities must be constrained to sum to 1. A simple approach is to reparameterize them as a set of logit probabilities,

$$p_{ij} = \frac{e^{\theta_{ij}}}{\sum_{j=1}^{J} e^{\theta_{ij}}}, \quad j = 1, \ldots, J, \quad \theta_{iJ} = 0, \quad \theta_{ij} = \boldsymbol{\delta}_j' \mathbf{z}_i, \quad (\boldsymbol{\delta}_J = \mathbf{0}). \qquad \textbf{(16-19)}$$

(See Section 21.8 for development of this model for a set of probabilities.) Note the restriction on θ_{iJ}. This is an identification restriction. Without it, the same set of

[15] One can view this as a discrete approximation to the continuous distribution. This is also an extension of Heckman and Singer's (1984b) model of latent heterogeneity, but the interpretation is a bit different here.

probabilities will arise if an arbitrary vector is added to every $\boldsymbol{\delta}_j$. The resulting log likelihood is a continuous function of the parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ and $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_J$. For all its apparent complexity, estimation of this model by direct maximization of the log likelihood is not especially difficult. [See Section E.5 and Greene (2001).] The number of classes that can be identified is likely to be relatively small (on the order of five or less), however, which is viewed as a drawback of this approach, and, in general, (as might be expected), the less rich is the panel data set in terms of cross group variation, the more difficult it is to estimate this model.

Estimation produces values for the structural parameters, $(\boldsymbol{\beta}_j, \boldsymbol{\delta}_j)$, $j = 1, \ldots, J$. With these in hand, we can compute the prior class probabilities, p_{ij} using (16-20). For prediction purposes, one might be more interested in the posterior (on the data) class probabilities, which we can compute using Bayes theorem as

$$\begin{aligned}\text{Prob(class } j \mid \text{observation } i) &= \frac{f(\text{observation } i \mid \text{class } j)\,\text{Prob(class } j)}{\sum_{j=1}^{J} f(\text{observation } i \mid \text{class } j)\,\text{Prob(class } j)} \\ &= \frac{f(y_{i1}, y_{i2}, \ldots, y_{iT} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, \boldsymbol{\beta}_j) p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)}{\sum_{j=1}^{M} f(y_{i1}, y_{i2}, \ldots, y_{iT} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, \boldsymbol{\beta}_j) p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)} \\ &= w_{ij}.\end{aligned}$$

This set of probabilities, $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iJ})$ gives the posterior density over the distribution of values of $\boldsymbol{\beta}$, that is, $[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J]$. The Bayesian estimator of the (individual specific) parameter vector would be the posterior mean

$$\hat{\boldsymbol{\beta}}_i^p = \hat{E}_j[\boldsymbol{\beta}_j \mid \text{observation } i] = \sum_{j=1}^{J} w_{ij} \hat{\boldsymbol{\beta}}_j.$$

Example 16.5 Applications of the Latent Class Model

The latent class formulation has provided an attractive platform for modeling latent heterogeneity. (See Greene (2001) for a survey.) For two examples, Nagin and Land (1993) employed the model to study age transitions through stages of criminal careers and Wang et al. (1998) and Wedel et al. (1993) and used the Poisson regression model to study counts of patents. To illustrate the estimator, we will apply the latent class model to the panel data binary choice application of firm product innovations studied by Bertschek and Lechner (1998).[16] They analyzed the dependent variable

$$y_{it} = 1 \text{ if firm } i \text{ realized a product innovation in year } t \text{ and 0 if not.}$$

Thus, this is a binary choice model. (See Section 21.2 for analysis of binary choice models.) The sample consists of 1270 German manufacturing firms observed for five years, 1984–1988. Independent variables in the model that we formulated were

x_{it1} = constant,

x_{it2} = log of sales,

x_{it3} = relative size = ratio of employment in business unit to employment in the industry,

x_{it4} = ratio of industry imports to (industry sales + imports),

x_{it5} = ratio of industry foreign direct investment to (industry sales + imports),

[16] We are grateful to the authors of this study who have generously loaned us their data for this analysis. The data are proprietary and cannot be made publicly available as are the other data sets used in our examples.

TABLE 16.2 Estimated Latent Class Model

	Probit	*Class 1*	*Class 2*	*Class 3*	*Posterior*
Constant	−1.96	−2.32	−2.71	−8.97	−3.38
	(0.23)	(0.59)	(0.69)	(2.20)	(2.14)
lnSales	0.18	0.32	0.23	0.57	0.34
	(0.022)	(0.061)	(0.072)	(0.18)	(0.09)
Rel. Size	1.07	4.38	0.72	1.42	2.58
	(0.14)	(0.89)	(0.37)	(0.76)	(1.30)
Import	1.13	0.94	2.26	3.12	1.81
	(0.15)	(0.37)	(0.53)	(1.38)	(0.74)
FDI	2.85	2.20	2.81	8.37	3.63
	(0.40)	(1.16)	(1.11)	(1.93)	(1.98)
Prod.	−2.34	−5.86	−7.70	−0.91	−5.48
	(0.72)	(2.70)	(4.69)	(6.76)	(1.78)
RawMtls	−0.28	−0.11	−0.60	0.86	−0.08
	(0.081)	(0.24)	(0.42)	(0.70)	(0.37)
Invest.	0.19	0.13	0.41	0.47	0.29
	(0.039)	(0.11)	(0.12)	(0.26)	(0.13)
ln L	−4114.05		−3503.55		
Class Prob. (Prior)		0.469	0.331	0.200	
		(0.0352)	(0.0333)	(0.0246)	
Class Prob. (Posterior)		0.469	0.331	0.200	
		(0.394)	(0.289)	(0.325)	
Pred. Count		649	366	255	

x_{it6} = productivity = ratio of industry value added to industry employment,

x_{it7} = dummy variable indicating firm is in the raw materials sector,

x_{it8} = dummy variable indicating firm is in the investment goods sector.

Discussion of the data set may be found in the article (pp. 331–332 and 370). Our central model for the binary outcome is a probit model,

$$f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j) = \text{Prob}[y_{it} \mid \mathbf{x}_{it}'\boldsymbol{\beta}_j] = \Phi[(2y_{it} - 1)\mathbf{x}_{it}'\boldsymbol{\beta}_j], \quad y_{it} = 0, 1.$$

This is the specification used by the authors. We have retained it so we can compare the results of the various models. We also fit a model with year specific dummy variables instead of a single constant and with the industry sector dummy variables moved to the latent class probability equation. See Greene (2002) for analysis of the different specifications.

Estimates of the model parameters are presented in Table 16.2. The "probit" coefficients in the first column are those presented by Bertschek and Lechner.[17] The class specific parameter estimates cannot be compared directly, as the models are quite different. The estimated posterior mean shown, which is comparable to the one class results is the sample average and standard deviation of the 1,270 firm specific posterior mean parameter vectors. They differ considerably from the probit model, but in each case, a confidence interval around the posterior mean contains the probit estimator. Finally, the (identical) prior and average of the sample posterior class probabilities are shown at the bottom of the table. The much larger empirical standard deviations reflect that the posterior estimates are based on aggregating the sample data and involve, as well, complicated functions of all the model parameters. The estimated numbers of class members are computed by assigning to each firm the predicted

[17]The authors used the robust "sandwich" estimator for the standard errors—see Section 17.9—rather than the conventional negative inverse of the Hessian.

class associated with the highest posterior class probability. Finally, to explore the difference between the probit model and the latent class model, we have computed the probability of a product innovation at the five-year mean of the independent variables for each firm using the probit estimates and the firm specific posterior mean estimated coefficient vector. The two kernel density estimates shown in Figures 16.1 and 16.2 (see Section 16.4.1) show the effect of allowing the greater between firm variation in the coefficient vectors.

FIGURE 16.1 Probit Probabilities.

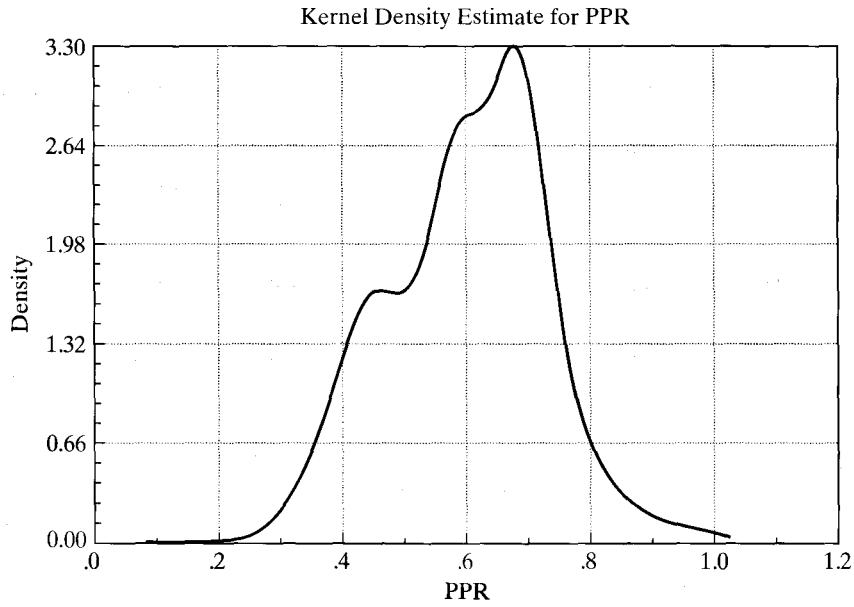
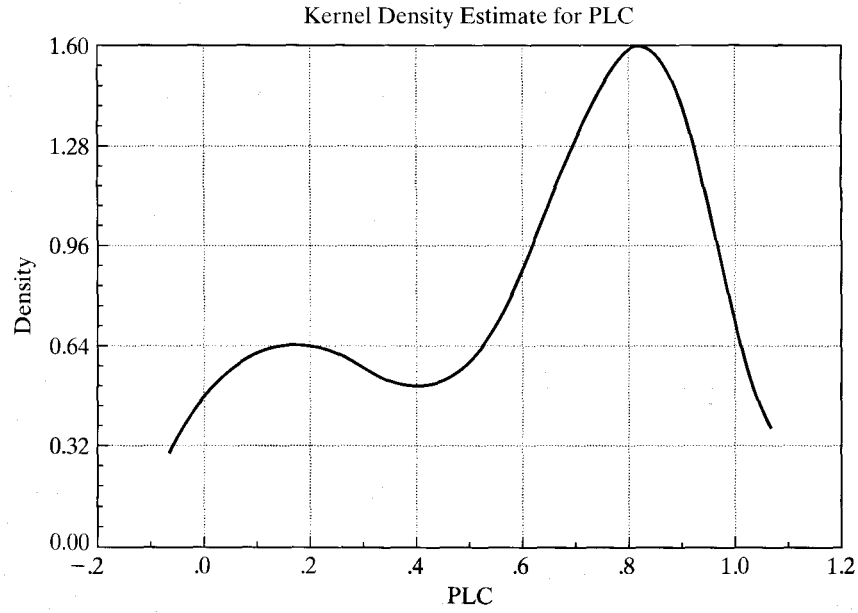


FIGURE 16.2 Latent Class Probabilities.



16.2.4 HIERARCHICAL BAYES ESTIMATION OF A RANDOM PARAMETERS MODEL BY MARKOV CHAIN MONTE CARLO SIMULATION

We now consider a Bayesian approach to estimation of the random parameters model in (16-19). For an individual i, the conditional density for the dependent variable in period t is $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i)$ 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.[18] For the sequence of T observations, assuming conditional (on $\boldsymbol{\beta}_i$) independence, person i's contribution to the likelihood for the sample is

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i). \tag{16-20}$$

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 $\boldsymbol{\beta}_i$;

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\beta}_i} \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) \phi_K[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}] \, d\boldsymbol{\beta}_i \tag{16-21}$$

where $\phi_K[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}]$ 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, $\boldsymbol{\beta}_i$ using the same method we used for the latent class model, is considered in Section 17.8. For now, we 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 $\boldsymbol{\beta}$. 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 (16-14). A full matrix $\boldsymbol{\Sigma}$ is handled by assigning to $\boldsymbol{\Sigma}$ an inverted Wishart prior density with parameters scalar K and matrix $K \times \mathbf{I}$. [The Wishart density is a multivariate counterpart to the Chi-squared distribution. Discussion may be found in Zellner (1971, pp. 389–394).] This produces the joint posterior density,

$$\Lambda(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \text{all data}) = \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) \phi_K[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}] \right\} \times p(\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{16-22}$$

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 (16-23) with respect to the components. As one might guess at this point, that integration is hopelessly complex and not remotely feasible. It is at this point that the recently

[18] In order 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 (though it is quite straightforward in the maximum simulated likelihood approach discussed in Section 17.8).

developed techniques of Markov Chain Monte Carlo (MCMC) simulation estimation and the Metropolis Hastings algorithm enter and enable us to do the estimation in a remarkably simple fashion.

The MCMC procedure makes use of a result that we have employed at many points in the preceding chapters. The joint density in (16-23) is exceedingly complex, and brute force integration is not feasible. Suppose, however, that we could draw random samples of $[\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}]$ from this population. Then, sample statistics such as means computed from these random draws would converge to the moments of the underlying population. The laws of large numbers discussed in Appendix D would apply. That partially solves the problem. The distribution remains as complex as before, however, so how to draw the sample remains to be solved. The **Gibbs sampler** and the **Metropolis–Hastings algorithm** can be used for sampling from the (hopelessly complex) joint density, $\Lambda(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \text{all data})$. The basic principle of the Gibbs sampler is described in Section E2.6. The core result is as follows: For a two-variable case, $f(x, y)$ in which $f(x \mid y)$ and $f(y \mid x)$ are known. A "Gibbs sequence" of draws, $y_0, x_0, y_1, x_1, y_2, \ldots, y_M, x_M$, is generated as follows. First, y_0 is specified "manually." Then x_0 is obtained as a random draw from the population $f(x \mid y_0)$. Then y_1 is drawn from $f(y \mid x_0)$, and so on. The iteration is, generically, as follows.

1. Draw x_j from $f(x \mid y_j)$.
2. Draw y_{j+1} from $f(y \mid x_j)$.
3. Exit or return to step 1.

If this process is repeated enough times, then at the last step, (x_j, y_j) together are a draw from the joint distribution.

Train (2001 and 2002, Chapter 12) describes how to use these results for this random parameters model.[19] 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 $\boldsymbol{\gamma} \mid \boldsymbol{\delta}$, we will use the Gibbs sampler to obtain a draw from the distribution of $(\boldsymbol{\beta} \mid \boldsymbol{\Sigma}, \boldsymbol{\delta})$ then one from the distribution of $(\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \boldsymbol{\delta})$. We will lay this out first, then turn to sampling from $\boldsymbol{\delta} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

Conditioned on $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ has a K-variate normal distribution with mean $\bar{\boldsymbol{\beta}} = (1/n)\sum_{i=1}^{n} \boldsymbol{\beta}_i$ 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.7.11.] Let $\mathbf{v}$ be a vector of K draws from the standard normal distribution. Then, $\bar{\boldsymbol{\beta}} + \mathbf{Lv}$ has mean vector $\bar{\boldsymbol{\beta}} + \mathbf{L} \times \mathbf{0} = \bar{\boldsymbol{\beta}}$ 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 of $\boldsymbol{\beta}$.

To obtain a random draw from the distribution of $\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \boldsymbol{\delta}$, we will require a random draw from the inverted Wishart distribution. The marginal posterior distribution of $\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \boldsymbol{\delta}$ is inverted Wishart with parameters scalar $K + n$ and matrix $\mathbf{W} = (K\mathbf{I} + n\mathbf{V})$

[19] Train describes use of this method for "mixed 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. In Section 17.8, we will apply this model to the probit model considered in the latent class model in Example 16.5.

where $\mathbf{V} = (1/n)\sum_{i=1}^{n}(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})'$. 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 $\mathbf{S} = \mathbf{M}(\sum_{k=1}^{K+n}\mathbf{v}_k\mathbf{v}_k')\mathbf{M}'$. 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 in $\boldsymbol{\beta}_i$, it simplifies a bit further. A draw for the nonzero kth diagonal element can be obtained using $(1 + n\mathbf{V}_{kk})/\sum_{r=1}^{K+n} v_{rk}^2$.]

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. (1995). The procedure involves the following steps:

1. Given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and "tuning constant" τ (to be described below), compute $\mathbf{d} = \tau\mathbf{Lv}$ 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 (16-21) times the joint normal prior density, $\phi_K[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}]$. Evaluate this posterior density at the trial value $\boldsymbol{\beta}_{i1}$ and the previous value $\boldsymbol{\beta}_{i0}$. Let

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

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, τ is used to control the iteration. A smaller τ increases the acceptance rate. But at the same time, a smaller τ makes new draws look more like old draws so this slows slows down the process. Gelman et al. (1995) 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.

Uses of this methodology can be found in many places in the literature. It has been particularly productive in marketing research, for example, in analyzing discrete

choice such as brand choice. The cost is in the amount of computation, which is large. Some important qualifications: As we have hinted before, in Bayesian estimation, as the amount of sample information increases, it eventually dominates the prior density, even if it is informative, so long as it is proper and has finite moments. The Bernstein–von Mises Theorem [Train (p. 5)] gives formal statements of this result, but we can summarize it with Bickel and Doksum's (2000) version, which observes that the asymptotic sampling distribution of the posterior mean is the same as the asymptotic distribution of the maximum likelihood estimator. The practical implication of this for us is that if the sample size is large, the Bayesian estimator of the parameters described here and the maximum likelihood estimator described in Section 17.9 will give the same answer.[20]

16.3 SEMIPARAMETRIC ESTIMATION

Semiparametric estimation is based on fewer assumptions than parametric estimation. In general, the distributional assumption is removed, and an estimator is devised from certain more general characteristics of the population. Intuition suggests two (correct) conclusions. First, the semiparametric estimator will be more robust than the parametric estimator—it will retain its properties, notably consistency) across a greater range of specifications. Consider our most familiar example. The least squares slope estimator is consistent whenever the data are well behaved and the disturbances and the regressors are uncorrelated. This is even true for the frontier function in Example 16.2, which has an asymmetric, nonnormal disturbance. But, second, this robustness comes at a cost. The distributional assumption usually makes the preferred estimator more efficient than a robust one. The best robust estimator in its class will usually be inferior to the parametric estimator when the assumption of the distribution is correct. Once again, in the frontier function setting, least squares may be robust for the slopes, and it is the most efficient estimator that uses only the orthogonality of the disturbances and the regressors, but it will be inferior to the maximum likelihood estimator when the two part normal distribution is the correct assumption.

16.3.1 GMM ESTIMATION IN ECONOMETRICS

Recent applications in economics include many that base estimation on the **method of moments.** The **generalized method of moments** departs from a set of model based moment equations, $E[\mathbf{m}(y_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0}$, where the set of equations specifies a relationship known to hold in the population. We used one of these in the preceding paragraph. The least squares estimator can be motivated by noting that the essential assumption is that $E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$. The estimator is obtained by seeking a parameter estimator, $\mathbf{b}$, which mimics the population result; $(1/n)\Sigma_i[\mathbf{x}_i(y_i - \mathbf{x}_i'\mathbf{b})] = \mathbf{0}$. This is, of course, the

[20]Practitioners might note, 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 program (1999)). 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 the same.

normal equations for least squares. Note that the estimator is specified without benefit of any distributional assumption. Method of moments estimation is the subject of Chapter 18, so we will defer further analysis until then.

16.3.2 LEAST ABSOLUTE DEVIATIONS ESTIMATION

Least squares can be severely distorted by outlying observations. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, all this discussion is moot.) These applications have led to the proposal of "robust" estimators that are unaffected by outlying observations.[21] In this section, we will examine one of these, the least absolute deviations, or LAD estimator.

That least squares gives such large weight to large deviations from the regression causes the results to be particularly sensitive to small numbers of atypical data points when the sample size is small or moderate. The **least absolute deviations** (LAD) estimator has been suggested as an alternative that remedies (at least to some degree) the problem. The LAD estimator is the solution to the optimization problem,

$$\text{Min}_{\mathbf{b}_0} \sum_{i=1}^{n} |y_i - \mathbf{x}_i'\mathbf{b}_0|.$$

The LAD estimator's history predates least squares (which itself was proposed over 200 years ago). It has seen little use in econometrics, primarily for the same reason that Gauss's method (LS) supplanted LAD at its origination; LS is vastly easier to compute. Moreover, in a more modern vein, its statistical properties are more firmly established than LAD's and samples are usually large enough that the small sample advantage of LAD is not needed.

The LAD estimator is a special case of the quantile regression:

$$\text{Prob}[y_i \leq \mathbf{x}_i'\boldsymbol{\beta}] = q.$$

The LAD estimator estimates the *median regression*. That is, it is the solution to the quantile regression when $q = 0.5$. Koenker and Bassett (1978, 1982), Huber (1967), and Rogers (1993) have analyzed this regression.[22] Their results suggest an estimator for the asymptotic covariance matrix of the **quantile regression** estimator,

$$\text{Est.Asy. Var}[\mathbf{b}_q] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

where $\mathbf{D}$ is a diagonal matrix containing weights

$$d_i = \left[\frac{q}{f(0)}\right]^2 \text{ if } y_i - \mathbf{x}_i'\boldsymbol{\beta} \text{ is positive and } \left[\frac{1-q}{f(0)}\right]^2 \text{ otherwise,}$$

[21] For some applications, see Taylor (1974), Amemiya (1985, pp. 70–80), Andrews (1974), Koenker and Bassett (1978), and a survey written at a very accessible level by Birkes and Dodge (1993). A somewhat more rigorous treatment is given by Hardle (1990).

[22] Powell (1984) has extended the LAD estimator to produce a robust estimator for the case in which data on the dependent variable are censored, that is, when negative values of y_i are recorded as zero. See Section 22.3.4c for discussion and Melenberg and van Soest (1996) for an application. For some related results on other semiparametric approaches to regression, see Butler, McDonald, Nelson, and White (1990) and McDonald and White (1993).

and $f(0)$ is the true density of the disturbances evaluated at 0.[23] [It remains to obtain an estimate of $f(0)$.] There is one useful symmetry in this result. Suppose that the true density were normal with variance σ^2. Then the preceding would reduce to $\sigma^2(\pi/2)(\mathbf{X}'\mathbf{X})^{-1}$, which is the result we used in Example E.1 to compare estimates of the median and the mean in a simple situation of random sampling. For more general cases, some other empirical estimate of $f(0)$ is going to be required. Nonparametric methods of density estimation are available [see Section 16.4 and, e.g., Johnston and DiNardo (1997, pp. 370–375)]. But for the small sample situations in which techniques such as this are most desirable (our application below involves 25 observations), nonparametric kernel density estimation of a single ordinate is optimistic; these are, after all, asymptotic results. But asymptotically, as suggested by Example E.1, the results begin overwhelmingly to favor least squares. For better or worse, a convenient estimator would be a kernel density estimator as described in Section 16.4.1. Looking ahead, the computation would be

$$\hat{f}(0) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{e_i}{h}\right]$$

where h is the bandwidth (to be discussed below), K[.] is a weighting, or kernel function and $e_i, i = 1, \ldots, n$ is the set of residuals. There are no hard and fast rules for choosing h; one popular choice is that used by Stata, $h = .9s/n^{1/5}$. The kernel function is likewise discretionary, though it rarely matters much which one chooses; the logit kernel (see Table 16.4) is a common choice.

The bootstrap method of inferring statistical properties is well suited for this application. Since the efficacy of the bootstrap has been established for this purpose, the search for a formula for standard errors of the LAD estimator is not really necessary. The bootstrap estimator for the asymptotic covariance matrix can be computed as follows:

$$\text{Est. Var}[\mathbf{b}_{LAD}] = \frac{1}{R}\sum_{r=1}^{R}(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})',$$

where $\mathbf{b}_{LAD}$ is the LAD estimator and $\mathbf{b}_{LAD}(r)$ is the rth LAD estimate of $\boldsymbol{\beta}$ based on a sample of n observations, drawn with replacement, from the original data set.

Example 16.6 LAD Estimation of a Cobb–Douglas Production Function

Zellner and Revankar (1970) proposed a generalization of the Cobb–Douglas production function which allows economies of scale to vary with output. Their statewide data on Y = value added (output), K = capital, L = labor, and N = the number of establishments in the transportation industry are given in Appendix Table F9.2. The generalized model is estimated in Example 17.9. For this application, estimates of the Cobb–Douglas production function,

$$\ln(Y_i/N_i) = \beta_1 + \beta_2 \ln(K_i/N_i) + \beta_3 \ln(L_i/N_i) + \varepsilon_i,$$

are obtained by least squares and LAD. The standardized least squares residuals (see Section 4.9.3) suggest that two observations (Florida and Kentucky) are outliers by the usual

[23] See Stata (2001). Koenker suggests that for independent and identically distributed observations, one should replace d_i with the constant $a = q(1-q)/[f(F^{-1}(q))]^2 = [.25/f(0)]^2$ for the median (LAD) estimator. This reduces the expression to the true asymptotic covariance matrix, $a(\mathbf{X}'\mathbf{X})^{-1}$. The one given is a sample estimator which will behave the same in large samples. (Personal communication to the author.)

TABLE 16.3 LS and LAD Estimates of a Production Function

	Least Squares			*LAD*				
					Bootstrap		*Kernel Density*	
Coefficient	*Estimate*	*Standard Error*	*t Ratio*	*Estimate*	*Std. Error*	*t Ratio*	*Std. Error*	*t Ratio*
Constant	1.844	0.234	7.896	1.806	0.344	5.244	0.320	5.639
β_k	0.245	0.107	2.297	0.205	0.128	1.597	0.147	1.398
β_l	0.805	0.126	6.373	0.849	0.163	5.201	0.173	4.903
Σe^2	1.2222			1.2407				
$\Sigma \lvert e \rvert$	4.0008			3.9927				

construction. The least squares coefficient vectors with and without these two observations are (1.844, 0.245, 0.805) and (1.764, 0.209, 0.852), respectively, which bears out the suggestion that these two points do exert considerable influence. Table 16.3 presents the LAD estimates of the same parameters, with standard errors based on 500 bootstrap replications. The LAD estimates with and without these two observations are identical, so only the former are presented. Using the simple approximation of multiplying the corresponding OLS standard error by $(\pi/2)^{1/2} = 1.2533$ produces a surprisingly close estimate of the bootstrap estimated standard errors for the two slope parameters (0.134, 0.158) compared with the bootstrap estimates of (0.128, 0.163). The second set of estimated standard errors are based on Koenker's suggested estimator, $.25/\hat{f}^2(0) = .25/1.5467^2 = 0.104502$. The bandwidth and kernel function are those suggested earlier. The results are surprisingly consistent given the small sample size.

16.3.3 PARTIALLY LINEAR REGRESSION

The proper functional form in the linear regression is an important specification issue. We examined this in detail in Chapter 7. Some approaches, including the use of dummy variables, logs, quadratics, and so on were considered as means of capturing nonlinearity. The translog model in particular (Example 2.4.) is a well-known approach to approximating an unknown nonlinear function. Even with these approaches, the researcher might still be interested in relaxing the assumption of functional form in the model. The **partially linear model** [analyzed in detail by Yatchew (1998, 2000)] is another approach. Consider a regression model in which one variable, x, is of particular interest, and the functional form with respect to x is problematic. Write the model as

$$y_i = f(x_i) + \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where the data are assumed to be well behaved and, save for the functional form, the assumptions of the classical model are met. The function $f(x_i)$ remains unspecified. As stated, estimation by least squares is not feasible until $f(x_i)$ is specified. Suppose the data were such that they consisted of pairs of observations (y_{j1}, y_{j2}), $j = 1, \ldots, n/2$ in which $x_{j1} = x_{j2}$ within every pair. If so, then estimation of $\boldsymbol{\beta}$ could be based on the simple transformed model

$$y_{j2} - y_{j1} = (\mathbf{z}_{j2} - \mathbf{z}_{j1})'\boldsymbol{\beta} + (\varepsilon_{j2} - \varepsilon_{j1}), \quad j = 1, \ldots, n/2.$$

As long as observations are independent, the constructed disturbances, v_i still have zero mean, variance now $2\sigma^2$, and remain uncorrelated across pairs, so a classical model applies and least squares is actually optimal. Indeed, with the estimate of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}_d$ in

hand, a noisy estimate of $f(x_i)$ could be estimated with $y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_d$ (the estimate contains the estimation error as well as v_i).[24]

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. The logic of the partially linear regression estimator is based on this observation nonetheless. Suppose that the observations are sorted so that $x_1 < x_2 < \cdots < x_n$. Suppose, as well, that this variable is well behaved in the sense that as the sample size increases, this sorted data vector more tightly and uniformly fills the space within which x_i is assumed to vary. Then, intuitively, the difference is "almost" right, and becomes better as the sample size grows. [Yatchew (1997, 1998) goes more deeply into the underlying theory.] A theory is also developed for a better differencing of groups of two or more observations. The transformed observation is $y_{d,i} = \sum_{m=0}^{M} d_m y_{i-m}$ where $\sum_{m=0}^{M} d_m = 0$ and $\sum_{m=0}^{M} d_m^2 = 1$. (The data are not separated into nonoverlapping groups for this transformation—we merely used that device to motivate the technique.) The pair of weights for $M = 1$ is obviously $\pm\sqrt{.5}$—this is just a scaling of the simple difference, $1, -1$. Yatchew [1998, p. 697)] tabulates "optimal" differencing weights for $M = 1, \ldots, 10$. The values for $M = 2$ are $(0.8090, -0.500, -0.3090)$ and for $M = 3$ are $(0.8582, -0.3832, -0.2809, -0.1942)$. This estimator is shown to be consistent, asymptotically normally distributed, and have asymptotic covariance matrix

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_d] = \left(1 + \frac{1}{2M}\right)\frac{\sigma_v^2}{n} E_x[\text{Var}[\mathbf{z} \mid x]].^{25}$$

The matrix can be estimated using the sums of squares and cross products of the differenced data. The residual variance is likewise computed with

$$\hat{\sigma}_v^2 = \frac{\sum_{i=M+1}^{n} (y_{d,i} - \mathbf{z}_{d,i}'\hat{\boldsymbol{\beta}}_d)^2}{n - M}.$$

Yatchew suggests that the partial residuals, $y_{d,i} - \mathbf{z}_{d,i}'\hat{\boldsymbol{\beta}}_d$ be smoothed with a kernel density estimator to provide an improved estimator of $f(x_i)$.

Example 16.7 Partially Linear Translog Cost Function

Yatchew (1998, 2000) applied this technique to an analysis of scale effects in the costs of electricity supply. The cost function, following Nerlove (1963) and Christensen and Greene (1976) was specified to be a translog model (see Example 2.4 and Section 14.3.2) involving labor and capital input prices, other characteristics of the utiity and the variable of interest, the number of customers in the system, C. We will carry out a similar analysis using Christenen and Greene's 1970 electricity supply data. The data are given in Appendix Table F5.2. (See Section 14.3.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies which are comprised of combinations of the others. In addition, there are several extremely small New England utilities whose costs are clearly unrepresentative of the best practice in the industry. We have done the analysis using firms 6-123 in the data set. Variables in the data set include Q = output, C = total cost and PK, PL, and PF = unit cost measures for capital, labor and fuel, respectively. The parametric model specified is a restricted version of the Christensen and Greene model,

$$\ln c = \beta_1 k + \beta_2 l + \beta_3 q + \beta_4 (q)^2/2 + \beta_5 + \varepsilon.$$

[24] See Estes and Honore (1995) who suggest this approach (with simple differencing of the data).

[25] Yatchew (2000, p. 191) denotes this covariance matrix $E[\text{Cov}[\mathbf{z} \mid x]]$.

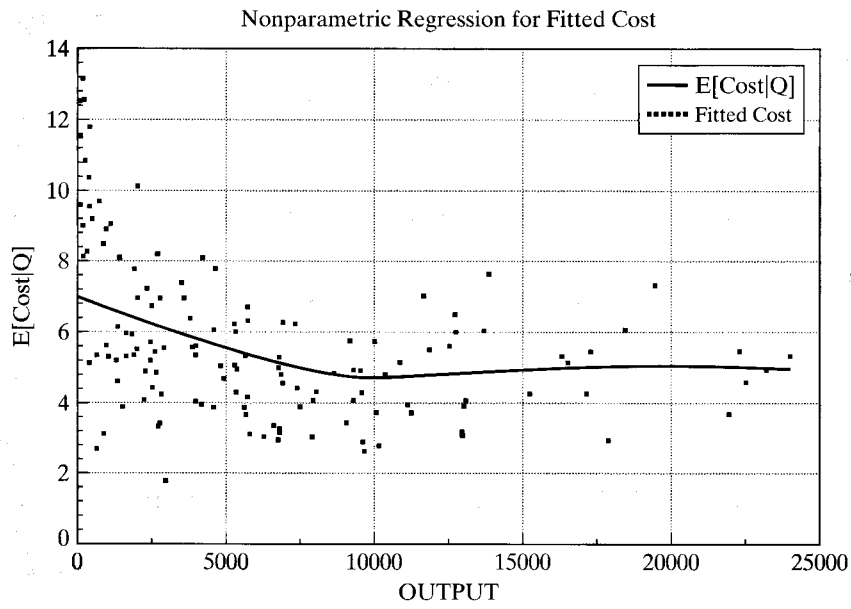


FIGURE 16.3 Smoothed Estimator for Costs.

where $c = \ln C/(Q \times PF), k = \ln(PK/PF), l = \ln(PL/PF)$ and $q = \ln Q$. The partially linear model substitutes $f(Q)$ for the last three terms. The division by PF ensures that average cost is homogeneous of degree one in the prices, a theoretical necessity. The estimated equations, with estimated standard errors are shown below.

$$\text{(parametric)} \quad \underset{}{c} = \underset{(0.353)}{-6.83} + \underset{(0.042)}{0.168k} + \underset{(0.048)}{0.146l} - \underset{(0.075)}{0.590q} + \underset{(0.010)}{0.061q^2/2} + \varepsilon, \quad s = 0.13383$$

$$\text{(partial linear)} \quad c_d = \underset{(0.049)}{0.170k_d} + \underset{(0.057)}{0.127l_d} + f(Q) + \mathrm{v}, \quad s = 0.14044$$

Yatchew's suggested smoothed kernel density estimator for the relationship between average cost and output is shown in Figure 16.3 with the unsmoothed partial residuals. We find (as did Christensen and Greene in the earlier study) that in the relatively low ranges of output, there is a fairly strong relationship between scale and average cost.

16.3.4 Kernel Density Methods

The kernel density estimator is an inherently nonparametric tool, so it fits more appropriately into the next section. But some models which use kernel methods are not completely nonparametric. The partially linear model in the preceding example is a case in point. Many models retain an index function formulation, that is, build the specification around a linear function, $\mathbf{x}'\boldsymbol{\beta}$, which makes them at least semiparametric, but nonetheless still avoid distributional assumptions by using kernel methods. Lewbel's (2000) estimator for the binary choice model is another example.

Example 16.8 Semiparametric Estimator for Binary Choice Models

The core binary choice model analyzed in Example 16.5, the probit model, is a fully parametric specification. Under the assumptions of the model, maximum likelihood is the efficient (and appropriate) estimator. However, as documented in a voluminous literature, the estimator

of β is fragile with respect to failures of the distributional assumption. We will examine a few semiparametric and nonparametric estimators in Section 21.5. To illustrate the nature of the modeling process, we consider an estimator recently suggested by Lewbel (2000). The probit model is based on the normal distribution, with $\text{Prob}[y_i = 1] = \text{Prob}[\mathbf{x}_i'\beta + \varepsilon_i > 0]$ where $\varepsilon_i \sim \text{N}[0, 1]$. The estimator of β under this specification will be inconsistent if the distribution is not normal or if ε_i is heteroscedastic. Lewbel suggests the following: If (a) it can be assumed that $\mathbf{x}_i$ contains a "special" variable, v_i, whose coefficient has a known sign—a method is developed for determining the sign and (b) the density of ε_i is independent of this variable, then a consistent estimator of β can be obtained by *linear regression* of $[y_i - s(v_i)]/f(v_i \mid \mathbf{x}_i)$ on $\mathbf{x}_i$ where $s(v_i) = 1$ if $v_i > 0$ and 0 otherwise and $f(v_i \mid \mathbf{x}_i)$ is a kernel density estimator of the density of $v_i \mid \mathbf{x}_i$. Lewbel's estimator is robust to heteroscedasticity and distribution. A method is also suggested for estimating the distribution of ε_i. Note that Lewbel's estimator is semiparametric. His underlying model is a function of the parameters β, but the distribution is unspecified.

16.4 NONPARAMETRIC ESTIMATION

Researchers have long held reservations about the strong assumptions made in parametric models fit by maximum likelihood. The linear regression model with normal disturbances is a leading example. Splines, translog models, and polynomials all represent attempts to generalize the functional form. Nonetheless, questions remain about how much generality can be obtained with such approximations. The techniques of nonparametric estimation discard essentially all fixed assumptions about functional form and distribution. Given their very limited structure, it follows that nonparametric specifications rarely provide very precise inferences. The benefit is that what information is provided is extremely robust. The centerpiece of this set of techniques is the kernel density estimator that we have used in the preceding examples. We will examine some examples, then examine an application to a bivariate regression.[26]

16.4.1 KERNEL DENSITY ESTIMATION

Sample statistics such as a mean, variance, and range give summary information about the values that a random variable may take. But, they do not suffice to show the distribution of values that the random variable takes, and these may be of interest as well. The density of the variable is used for this purpose. A fully parametric approach to density estimation begins with an assumption about the form of a distribution. Estimation of the density is accomplished by estimation of the parameters of the distribution. To take the canonical example, if we decide that a variable is generated by a normal distribution with mean μ and variance σ^2, then the density is fully characterized by these parameters. It follows that

$$\hat{f}(x) = f(x \mid \hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)^2\right].$$

One may be unwilling to make a narrow distributional assumption about the density. The usual approach in this case is to begin with a histogram as a descriptive device. Consider

[26]There is a large and rapidly growing literature in this area of econometrics. Two major references which provide an applied and theoretical foundation are Härdle (1990) and Pagan and Ullah (1999).

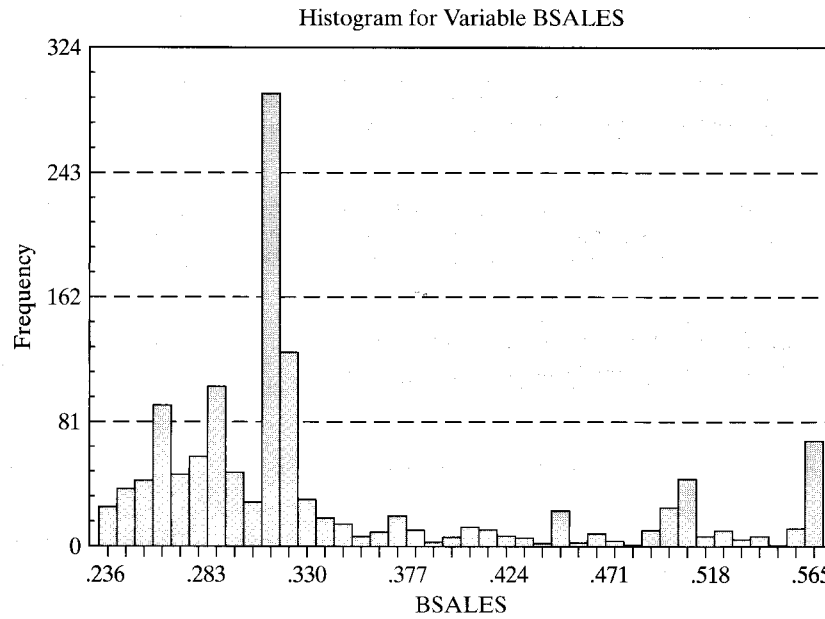


FIGURE 16.4 Histogram for Estimated Coefficients.

an example. In Example 16.5, we estimated a model that produced a posterior estimator of a slope vector for each of the 1,270 firms in our sample. We might be interested in the distribution of these estimators across firms. In particular, the posterior estimates of the estimated slope on ln*sales* for the 1,270 firms have a sample mean of 0.3428, a standard deviation of 0.08919, a minimum of 0.2361 and a maximum of 0.5664. This tells us little about the distribution of values, though the fact that the mean is well below the midrange of .4013 might suggest some skewness. The histogram in Figure 16.4 is much more revealing. Based on what we see thus far, an assumption of normality might not be appropriate. The distribution seems to be bimodal, but certainly no particular functional form seems natural.

The **histogram** is a crude density estimator. The rectangles in the figure are called bins. By construction, they are of equal width. (The parameters of the histogram are the number of bins, the bin width and the leftmost starting point. Each is important in the shape of the end result.) Since the frequency count in the bins sums to the sample size, by dividing each by n, we have a density estimator that satisfies an obvious requirement for a density; it sums (integrates) to one. We can formalize this by laying out the method by which the frequencies are obtained. Let x_k be the midpoint of the kth bin and let h be the width of the bin—we will shortly rename h to be the bandwidth for the density estimator. The distance to the left and right boundaries of the bins are $h/2$. The frequency count in each bin is the number of observations in the sample which fall in the range $x_k \pm h/2$. Collecting terms, we have our "estimator"

$$\hat{f}(x) = \frac{1}{n}\frac{\text{frequency in bin}_x}{\text{width of bin}_x} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathbf{1}\left(x - \frac{h}{2} < x_i < x + \frac{h}{2}\right)$$

where $\mathbf{1}(statement)$ denotes an indicator function which equals 1 if the statement is true and 0 if it is false and bin_x denotes the bin which has x as its midpoint. We see, then, that the histogram is an estimator, at least in some respects, like other estimators we have encountered. The event in the indicator can be rearranged to produce an equivalent form

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathbf{1}\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right).$$

This form of the estimator simply counts the number of points that are within $1/2$ bin width of x_k.

Albeit rather crude, this "naive" (its formal name in the literature) estimator is in the form of **kernel density estimators** that we have met at various points;

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{x_i - x}{h}\right], \quad \text{where } K[z] = \mathbf{1}[-1/2 < z < 1/2].$$

The naive estimator has several shortcomings. It is neither smooth nor continuous. Its shape is partly determined by where the leftmost and rightmost terminals of the histogram are set. (In constructing a histogram, one often chooses the bin width to be a specified fraction of the sample range. If so, then the terminals of the lowest and highest bins will equal the minimum and maximum values in the sample, and this will partly determine the shape of the histogram. If, instead, the bin width is set irrespective of the sample values, then this problem is resolved.) More importantly, the shape of the histogram will be crucially dependent on the bandwidth, itself. (Unfortunately, this problem remains even with more sophisticated specifications.)

The crudeness of the weighting function in the estimator is easy to remedy. Rosenblatt's (1956) suggestion was to substitute for the naive estimator some other weighting function which is continuous and which also integrates to one. A number of candidates have been suggested, including the (long) list in Table 16.4. Each of these is smooth, continuous, symmetric, and equally attractive. The Parzen, logit, and normal kernels are defined so that the weight only asymptotically falls to zero whereas the others fall to zero at specific points. It has been observed that in constructing density estimator, the choice of kernel function is rarely crucial, and is usually minor in importance compared to the more difficult problem of choosing the bandwidth. (The logit and normal kernels appear to be the default choice in many applications.)

TABLE 16.4 Kernels for Density Estimation

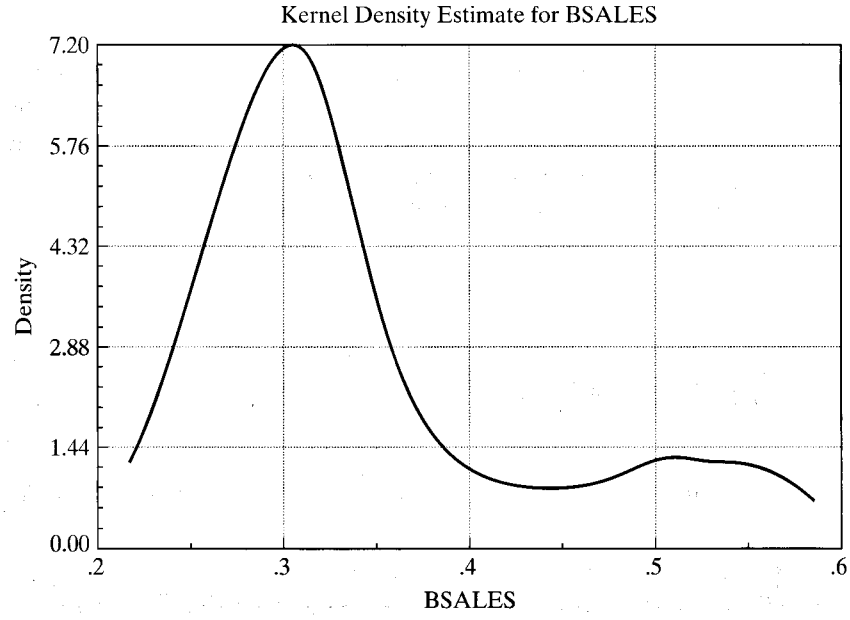
Kernel	*Formula K[z]*
Epanechnikov	$.75(1 - .2z^2)/2.236$ if $\lvert z\rvert \leq 5$, 0 else
Normal	$\phi(z)$ (normal density),
Logit	$\Lambda(z)[1 - \Lambda(z)]$ (logistic density)
Uniform	.5 if $\lvert z\rvert \leq 1$, 0 else
Beta	$(1 - z)(1 + z)/24$ if $\lvert z\rvert \leq 1$, 0 else
Cosine	$1 + \cos(2\pi z)$ if $\lvert z\rvert \leq .5$, 0 else
Triangle	$1 - \lvert z\rvert$, if $\lvert z\rvert \leq 1$, 0 else
Parzen	$4/3 - 8z^2 + 8\lvert z\rvert^3$ if $\lvert z\rvert \leq .5$, $8(1 - \lvert z\rvert)^3/3$ else

The kernel density function is an estimator. For any specific x, $\hat{f}(x)$ is a sample statistic,

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} g(x_i \mid z, h).$$

Since $g(x_i \mid z, h)$ is nonlinear, we should expect a bias in a finite sample. It is tempting to apply our usual results for sample moments, but the analysis is more complicated because the bandwidth is a function of n. Pagan and Ullah (1999) have examined the properties of kernel estimators in detail, and found that under certain assumptions the estimator is consistent and asymptotically normally distributed but biased in finite samples. The bias is a function of the bandwidth but for an appropriate choice of h, does vanish asymptotically. As intuition might suggest, the larger is the bandwidth, the greater is the bias, but at the same time, the smaller is the variance. This might suggest a search for an optimal bandwidth. After a lengthy analysis of the subject, however, the authors' conclusion provides little guidance for finding one. One consideration does seem useful. In order for the proportion of observations captured in the bin to converge to the corresponding area under the density, the width itself must shrink more slowly than $1/n$. Common applications typically use a **bandwidth** equal to some multiple of $n^{-1/5}$ for this reason. Thus, the one we used earlier is $h = 0.9 \times s/n^{1/5}$. To conclude the illustration begun earlier, Figure 16.5 is a logit based kernel density estimator for the distribution of slope estimates for the model estimated earlier. The resemblance to the histogram is to be expected.

FIGURE 16.5 Kernel Density for Coefficients.



16.4.2 NONPARAMETRIC REGRESSION

The regression function of a variable y on a single variable x is specified as

$$y = \mu(x) + \varepsilon.$$

No assumptions about distribution, homoscedasticity, serial correlation. or, most importantly, functional form are made at the outset; $\mu(x)$ may be quite nonlinear. Since this is the conditional mean, the only substantive restriction would be that deviations from the conditional mean function are not a function of (correlated with) x. We have already considered several possible strategies for allowing the conditional mean to be nonlinear, including spline functions, polynomials, logs, dummy variables, and so on. But, each of these is a "global" specification. The functional form is still the same for all values of x. Here, we are interested in methods that do not assume any particular functional form.

The simplest case to analyze would be one in which several (different) observations on y_i were made with each specific value of x_i. Then, the conditional mean function could be estimated naturally using the simple group means. The approach has two shortcomings, however. Simply connecting the points of means, $(x_i, \bar{y} \mid x_i)$ does not produce a smooth function. The method would still be assuming something specific about the function between the points, which we seek to avoid. Second, this sort of data arrangement is unlikely to arise except in an experimental situation. Given that data are not likely to be grouped, another possibility is a piecewise regression in which we define "neighborhoods" of points around each x of interest and fit a separate linear or quadratic regression in each neighborhood. This returns us to the problem of continuity that we noted earlier, but the method of splines is actually designed specifically for this purpose. Still, unless the number of neighborhoods is quite large, such a function is still likely to be crude.

Smoothing techniques are designed to allow construction of an estimator of the conditional mean function without making strong assumptions about the behavior of the function between the points. They retain the usefulness of the "**nearest neighbor**" concept, but use more elaborate schemes to produce smooth, well behaved functions. The general class may be defined by a conditional mean estimating function

$$\hat{\mu}(x^*) = \sum_{i=1}^{n} w_i(x^* \mid x_1, x_2, \ldots, x_n) y_i = \sum_{i=1}^{n} w_i(x^* \mid \mathbf{x}) y_i$$

where the weights sum to 1. The linear least squares regression line is such an estimator. The predictor is

$$\hat{\mu}(x^*) = a + bx^*$$

where a and b are the least squares constant and slope. For this function, you can show that

$$w_i(x^* \mid \mathbf{x}) = \frac{1}{n} + \frac{x^*(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The problem with this particular weighting function, which we seek to avoid here, is that it allows every x_i to be in the neighborhood of x^*, but it does not reduce the weight of any x_i when it is far from x^*. A number of **smoothing functions** have been suggested

which are designed to produce a better behaved regression function. [See Cleveland (1979) and Schimek (2000).] We will consider two.

The locally weighted smoothed regression estimator ("loess" or "lowess" depending on your source) is based on explicitly defining a neighborhood of points that is close to x^*. This requires the choice of a bandwidth, h. The neighborhood is the set of points for which $|x^* - x_i|$ is small. For example, the set of points that are within the range $x^* \pm h/2$ (as in our original histogram) might constitute the neighborhood. A suitable weight is then required. Cleveland (1979) recommends the tricube weight,

$$T_i(x^* \mid \mathbf{x}, h) = \left[1 - \left(\frac{|x_i - x^*|}{h}\right)^3\right]^3.$$

Combining terms, then the weight for the loess smoother is

$$w_i(x^* \mid \mathbf{x}, h) = \mathbf{1}(x_i \text{ in the neighborhood}) \times T_i(x^* \mid \mathbf{x}).$$

As always, the bandwidth is crucial. A wider neighborhood will produce a smoother function. But the wider neighborhood will track the data less closely than a narrower one. A second possibility, similar to the first, is to allow the neighborhood to be all points, but make the weighting function decline smoothly with the distance between x^* and any x_i. Any of the kernel functions suggested earlier will serve this purpose. This produces the kernel weighted regression estimator,

$$\hat{\mu}(x^* \mid \mathbf{x}, h) = \frac{\sum_{i=1}^{n} \frac{1}{h} K\left[\frac{x_i - x^*}{h}\right] y_i}{\sum_{i=1}^{n} \frac{1}{h} K\left[\frac{x_i - x^*}{h}\right]},$$

which has become a standard tool in nonparametric analysis.

Example 16.9 A Nonparametric Average Cost Function

In Example 16.7, we fit a partially linear regression for the relationship between average cost and output for electricity supply. Figures 16.6 and Figure 16.7 show the less ambitious nonparametric regressions of average cost on output. The overall picture is the same as in the earlier example. The kernel function is the logit density in both cases. The function in Figure 16.6 uses a bandwidth of 2,000. Since this is a fairly large proportion of the range of variation of output, the function is quite smooth. The regression in Figure 16.7 uses a bandwidth of only 200. The function tracks the data better, but at an obvious cost. The example demonstrates what we and others have noted often; the choice of bandwidth in this exercise is crucial.

Data smoothing is essentially data driven. As with most nonparametric techniques, inference is not part of the analysis—this body of results is largely descriptive. As can be seen in the example, nonparametric regression can reveal interesting characteristics of the data set. For the econometrician, however, there are a few drawbacks. Most relationships are more complicated than simple conditional mean of one variable. In the example just given, some of the variation in average cost relates to differences in factor prices (particularly fuel) and in load factors. Extensions of the fully nonparametric regression to more than one variable is feasible, but very cumbersome. [See Härdle (1990).] A promising approach is the partially linear model considered earlier.

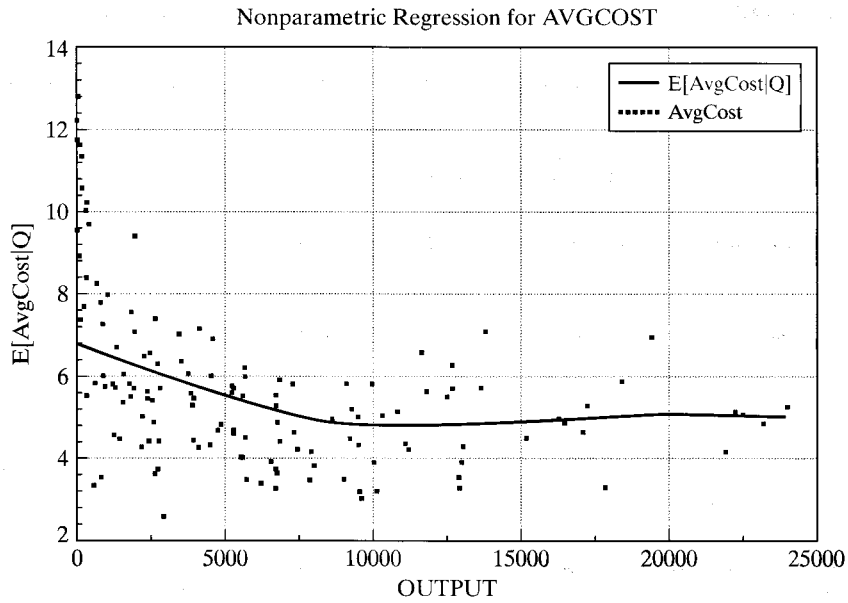
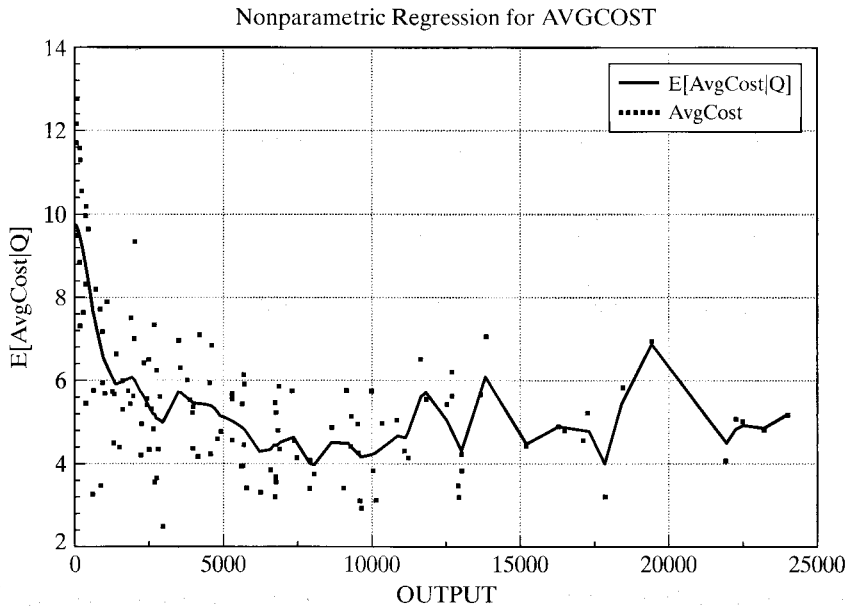


FIGURE 16.6 Nonparametric Cost Function.

FIGURE 16.7 Nonparametric Cost Function.



16.5 PROPERTIES OF ESTIMATORS

The preceding has been concerned with methods of estimation. We have surveyed a variety of techniques that have appeared in the applied literature. We have not yet examined the statistical properties of these estimators. Although, as noted earlier, we will leave extensive analysis of the asymptotic theory for more advanced treatments, it is appropriate to spend at least some time on the fundamental theoretical platform which underlies these techniques.

16.5.1 STATISTICAL PROPERTIES OF ESTIMATORS

Properties that we have considered are as follows:

- **Unbiasedness:** This is a finite sample property that can be established in only a very small number of cases. Strict unbiasedness is rarely of central importance outside the linear regression model. However, "asymptotic unbiasedness" (whereby the expectation of an estimator converges to the true parameter as the sample size grows), might be of interest. [See, e.g., Pagan and Ullah (1999, Section 2.5.1 on the subject of the kernel density estimator).] In most cases, however, discussions of asymptotic unbiasedness are actually directed toward consistency, which is a more desirable property.
- **Consistency:** This is a much more important property. Econometricians are rarely willing to place much credence in an estimator for which consistency cannot be established.
- **Asymptotic normality:** This property forms the platform for most of the statistical inference that is done with common estimators. When asymptotic normality cannot be established, for example, for the maximum score estimator discussed in Section 21.5.3, it sometimes becomes difficult to find a method of progressing beyond simple presentation of the numerical values of estimates (with caveats). However, most of the contemporary literature in macroeconomics and time series analysis is strongly focused on estimators which are decidedly not asymptotically normally distributed. The implication is that this property takes its importance only in context, not as an absolute virtue.
- **Asymptotic efficiency:** Efficiency can rarely be established in absolute terms. Efficiency within a class often can, however. Thus, for example, a great deal can be said about the relatively efficiency of maximum likelihood and GMM estimators in the class of CAN estimators. There are two important practical considerations in this setting. First, the researcher will want to know that they have not made demonstrably suboptimal use of their data. (The literature contains discussions of GMM estimation of fully specified parametric probit models—GMM estimation in this context is unambiguously inferior to maximum likelihood.) Thus, when possible, one would want to avoid obviously inefficient estimators. On the other hand, it will usually be the case that the researcher is not choosing from a list of available estimators; they have one at hand, and questions of relative efficiency are moot.

16.5.2 EXTREMUM ESTIMATORS

An **extremum estimator** is one which is obtained as the optimizer of a **criterion function** $q(\boldsymbol{\theta} \mid \textbf{data})$. Three that have occupied much of our effort thus far are

- Least squares: $\hat{\boldsymbol{\theta}}_{LS} = \text{Argmax}\left[-(1/n)\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i, \boldsymbol{\theta}_{LS}))^2\right]$,
- Maximum likelihood: $\hat{\boldsymbol{\theta}}_{ML} = \text{Argmax}\left[(1/n)\sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_{ML})\right]$,
- GMM: $\hat{\boldsymbol{\theta}}_{GMM} = \text{Argmax}[-\bar{\mathbf{m}}(\textbf{data}, \boldsymbol{\theta}_{GMM})'\mathbf{W}\bar{\mathbf{m}}(\textbf{data}, \boldsymbol{\theta}_{GMM})]$.

(We have changed the signs of the first and third only for convenience so that all three may be cast as the same type of optimization problem.) The least squares and maximum likelihood estimators are examples of **M estimators,** which are defined by optimizing over a sum of terms. Most of the familiar theoretical results developed here and in other treatises concern the behavior of extremum estimators. Several of the estimators considered in this chapter are extremum estimators, but a few, including the Bayesian estimators, some of the semiparametric estimators and all of the nonparametric estimators are not. Nonetheless. we are interested in establishing the properties of estimators in all these cases, whenever possible. The end result for the practitioner will be the set of statistical properties that will allow them to draw with confidence conclusions about the data generating process(es) that have motivated the analysis in the first place.

Derivations of the behavior of extremum estimators are pursued at various levels in the literature. (See, e.g., any of the sources mentioned in Footnote 1 of this chapter.) Amemiya (1985) and Davidson and MacKinnon (1993) are very accessible treatments. Newey and McFadden (1994) is a recent, rigorous analysis that provides a current, standard source. Our discussion at this point will only suggest the elements of the analysis. The reader is referred to one of these sources for detailed proofs and derivations.

16.5.3 ASSUMPTIONS FOR ASYMPTOTIC PROPERTIES OF EXTREMUM ESTIMATORS

Some broad results are needed in order to establish the asymptotic properties of the classical (not Bayesian) conventional extremum estimators noted above.

(a) **The parameter space** (see Section 16.2) must be convex and the parameter vector that is the object of estimation must be a point in its interior. The first requirement rules out ill defined estimation problems such as estimating a parameter which can only take one of a finite discrete set of values. Thus, searching for the date of a structural break in a time series model as if it were a conventional parameter leads to a nonconvexity. Some proofs in this context are simplified by assuming that the parameter space is compact. (A compact set is closed and bounded.) However, assuming compactness is usually restrictive, so we will opt for the weaker requirement.

(b) **The criterion function** must be concave in the parameters. (See Section A.8.2.) This assumption implies that with a given data set, the objective function has an interior optimum and that we can locate it. Criterion functions need not be "globally concave;" they may have multiple optima. But, if they are not at least "locally concave" then we cannot speak meaningfully about optimization. One would normally only encounter this problem in a badly structured model, but it is

possible to formulate a model in which the estimation criterion is monotonically increasing or decreasing in a parameter. Such a model would produce a nonconcave criterion function.[27] The distinction between compactness and concavity in the preceding condition is relevant at this point. If the criterion function is strictly continuous in a compact parameter space, then it has a maximum in that set and assuming concavity is not necessary. The problem for estimation, however, is that this does not rule out having that maximum occur on the (assumed) boundary of the parameter space. This case interferes with proofs of consistency and asymptotic normality. The overall problem is solved by assuming that the criterion function is concave in the neighborhood of the true parameter vector.

(c) Identifiability of the parameters. Any statement that begins with "the true parameters of the model, $\boldsymbol{\theta}_0$ are identified if . . ." is problematic because if the parameters are "not identified" then arguably, they are not *the* parameters of the (any) model. (For example, there is no "true" parameter vector in the unidentified model of Example 2.5.) A useful way to approach this question that avoids the ambiguity of trying to define *the* true parameter vector first and then asking if it is identified (estimable) is as follows, where we borrow from Davidson and MacKinnon (1993, p. 591): Consider the parameterized model, $\boldsymbol{M}$ and the set of allowable data generating processes for the model, μ. Under a particular parameterization μ, let there be an assumed "true" parameter vector, $\boldsymbol{\theta}(\mu)$. Consider any parameter vector $\boldsymbol{\theta}$ in the parameter space, $\boldsymbol{\Theta}$. Define

$$q_\mu(\mu, \boldsymbol{\theta}) = \text{plim}_\mu q_n(\boldsymbol{\theta} \mid \textbf{data}).$$

This function is the probability limit of the objective function under the assumed parameterization μ. If this probability limit exists (is a finite constant) and moreover, if

$$q_\mu(\mu, \boldsymbol{\theta}(\mu)) > q_\mu(\mu, \boldsymbol{\theta}) \quad \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}(\mu),$$

then if the parameter space is compact, the parameter vector is identified by the criterion function. We have not assumed compactness. For a convex parameter space, we would require the additional condition that there exist no sequences without limit points $\boldsymbol{\theta}^m$ such that $q(\mu, \boldsymbol{\theta}^m)$ converges to $q(\mu, \boldsymbol{\theta}(\mu))$.

The approach taken here is to assume first that the model has *some* set of parameters. The identifiability criterion states that assuming this is the case, the probability limit of the criterion is maximized at these parameters. This result rests on convergence of the criterion function to a finite value at any point in the interior of the parameter space. Since the criterion function is a function of the data, this convergence requires a statement of the properties of the data—e.g., well behaved in some sense. Leaving that aside for the moment, interestingly, the results to this

[27] In their Exercise 23.6, Griffiths, Hill, and Judge (1993), based (alas) on the first edition of this text, suggest a probit model for statewide voting outcomes that includes dummy variables for region, Northeast, Southeast, West, and Mountain. One would normally include three of the four dummy variables in the model, but Griffiths et al. carefully dropped two of them because in addition to the dummy variable trap, the Southeast variable is always zero when the dependent variable is zero. Inclusion of this variable produces a nonconcave likelihood function—the parameter on this variable diverges. Analysis of a closely related case appears as a caveat on page 272 of Amemiya (1985).

point already establish the consistency of the M estimator. In what might seem to be an extremely terse fashion, Amemiya (1985) defined identifiability simply as "existence of a consistent estimator." We see that identification and the conditions for consistency of the M estimator are substantively the same.

This form of identification is necessary, in theory, to establish the consistency arguments. In any but the simplest cases, however, it will be extremely difficult to verify in practice. Fortunately, there are simpler ways to secure identification that will appeal more to the intuition:

- For the least squares estimator, a sufficient condition for identification is that any two different parameter vectors, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ must be able to produce different values of the conditional mean function. This means that for any two different parameter vectors, there must be an $\mathbf{x}_i$ which produces different values of the conditional mean function. You should verify that for the linear model, this is the full rank assumption A.2. For the model in example 2.5, we have a regression in which $x_2 = x_3 + x_4$. In this case, any parameter vector of the form $(\beta_1, \beta_2 - a, \beta_3 + a, \beta_4 + a)$ produces the same conditional mean as $(\beta_1, \beta_2, \beta_3, \beta_4)$ regardless of $\mathbf{x}_i$, so this model is not identified. The full rank assumption is needed to preclude this problem. For nonlinear regressions, the problem is much more complicated, and there is no simple generality. Example 9.2 shows a nonlinear regression model that is not identified and how the lack of identification is remedied.
- For the maximum likelihood estimator, a condition similar to that for the regression model is needed. For any two parameter vectors, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ it must be possible to produce different values of the density $f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ for some data vector $(y_i, \mathbf{x}_i)$. Many econometric models that are fit by maximum likelihood are "index function" models that involve densities of the form $f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = f(y_i \mid \mathbf{x}_i'\boldsymbol{\theta})$. When this is the case, the same full rank assumption that applies to the regression model may be sufficient. (If there are no other parameters in the model, then it will be sufficient.)
- For the GMM estimator, not much simplicity can be gained. A sufficient condition for identification is that $\mathrm{E}[\bar{\mathbf{m}}(\mathbf{data}, \boldsymbol{\theta})] \neq \mathbf{0}$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

(d) Behavior of the data has been discussed at various points in the preceding text. The estimators are based on means of functions of observations. (You can see this in all three of the definitions above. Derivatives of these criterion functions will likewise be means of functions of observations.) Analysis of their large sample behaviors will turn on determining conditions under which certain sample means of functions of observations will be subject to laws of large numbers such as the Khinchine (D.5.) or Chebychev (D.6) theorems, and what must be assumed in order to assert that "root-n" times sample means of functions will obey central limit theorems such as the Lindberg–Feller (D.19) or Lyapounov (D.20) theorems for cross sections or the Martingale Difference Central Limit Theorem for dependent observations. Ultimately, this is the issue in establishing the statistical properties. The convergence property claimed above must occur in the context of the data. These conditions have been discussed in Section 5.2 and in Section 10.2.2 under the heading of "well behaved data." At this point, we will assume that the data are well behaved.

16.5.4 ASYMPTOTIC PROPERTIES OF ESTIMATORS

With all this apparatus in place, the following are the standard results on asymptotic properties of M estimators:

THEOREM 16.1 Consistency of M Estimators
If (a) the parameter space is convex and the true parameter vector is a point in its interior; (b) the criterion function is concave; (c) the parameters are identified by the criterion function; (d) the data are well behaved, then the M *estimator converges in probability to the true parameter vector.*

Proofs of consistency of M estimators rely on a fundamental convergence result that, itself, rests on assumptions (a) through (d) above. We have assumed identification. The fundamental device is the following: Because of its dependence on the data, $q(\boldsymbol{\theta} \mid \mathbf{data})$ is a random variable. We assumed in (c) that plim $q(\boldsymbol{\theta} \mid \mathbf{data}) = q_0(\boldsymbol{\theta})$ for any point in the parameter space. Assumption (c) states that the maximum of $q_0(\boldsymbol{\theta})$ occurs at $q_0(\boldsymbol{\theta}_0)$, so $\boldsymbol{\theta}_0$ is the maximizer of the probability limit. By its definition, the estimator $\hat{\boldsymbol{\theta}}$, is the maximizer of $q(\boldsymbol{\theta} \mid \mathbf{data})$. Therefore, consistency requires the limit of the maximizer, $\hat{\boldsymbol{\theta}}$ be equal to the maximizer of the limit, $\boldsymbol{\theta}_0$. Our identification condition establishes this. We will use this approach in somewhat greater detail in Section 17.4.5a where we establish consistency of the maximum likelihood estimator.

THEOREM 16.2 Asymptotic Normality of M Estimators
If

- ***(i)*** *$\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a point in the interior of the parameter space;*
- ***(ii)*** *$q(\boldsymbol{\theta} \mid \mathbf{data})$ is concave and twice continuously differentiable in $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$;*
- ***(iii)*** $\sqrt{n}[\partial q(\boldsymbol{\theta}_0 \mid \mathbf{data})/\partial \boldsymbol{\theta}_0] \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}]$;
- ***(iv)*** *for any $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$, $\lim_{n\to\infty} \Pr[|(\partial^2 q(\boldsymbol{\theta} \mid \mathbf{data})/\partial\theta_k\partial\theta_m) - h_{km}(\boldsymbol{\theta})| > \varepsilon] = 0 \; \forall \, \varepsilon > 0$ where $h_{km}(\boldsymbol{\theta})$ is a continuous finite valued function of $\boldsymbol{\theta}$;*
- ***(v)*** *the matrix of elements $\mathbf{H}(\boldsymbol{\theta})$ is nonsingular at $\boldsymbol{\theta}_0$, then* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, [\mathbf{H}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Phi}\mathbf{H}^{-1}(\boldsymbol{\theta}_0)]\}$

The proof of asymptotic normality is based on the mean value theorem from calculus and a Taylor series expansion of the derivatives of the maximized criterion function around the true parameter vector;

$$\sqrt{n}\frac{\partial q(\hat{\boldsymbol{\theta}} \mid \mathbf{data})}{\partial \hat{\boldsymbol{\theta}}} = \mathbf{0} = \sqrt{n}\frac{\partial q(\boldsymbol{\theta}_0 \mid \mathbf{data})}{\partial \boldsymbol{\theta}_0} + \frac{\partial^2 q(\bar{\boldsymbol{\theta}} \mid \mathbf{data})}{\partial \bar{\boldsymbol{\theta}} \partial \bar{\boldsymbol{\theta}}'}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

The second derivative is evaluated at a point $\bar{\theta}$ that is between $\hat{\theta}$ and θ_0, that is, $\bar{\theta} = w\hat{\theta} + (1-w)\theta_0$ for some $0 < w < 1$. Since we have assumed plim $\hat{\theta} = \theta_0$, we see that the matrix in the second term on the right must be converging to $\mathbf{H}(\theta_0)$. The assumptions in the theorem can be combined to produce the claimed normal distribution. Formal proof of this set of results appears in Newey and McFadden (1994). A somewhat more detailed analysis based on this theorem appears in Section 17.4.5b where we establish the asymptotic normality of the maximum likelihood estimator.

The preceding was restricted to M estimators, so it remains to establish counterparts for the important GMM estimator. Consistency follows along the same lines used earlier, but asymptotic normality is a bit more difficult to establish. We will return to this issue in Chapter 18, where, once again, we will sketch the formal results and refer the reader to a source such as Newey and McFadden (1994) for rigorous derivation.

The preceding results are not straightforward in all estimation problems. For example, the least absolute deviations (LAD) is not among the estimators noted earlier, but it is an M estimator and it shares the results given here. The analysis is complicated because the criterion function is not continuously differentiable. Nonetheless, consistency and asymptotic normality have been established. [See Koenker and Bassett (1982) and Amemiya (1985, pp. 152–154).] Some of the semiparametric and all of the nonparametric estimators noted require somewhat more intricate treatments. For example, Pagan and Ullah (Section 2.5 and 2.6) are able to establish the familiar desirable properties for the kernel density estimator $\hat{f}(x^*)$, but it requires a somewhat more involved analysis of the function and the data than is necessary, say, for the linear regression or binomial logit model. The interested reader can find many lengthy and detailed analyses of asymptotic properties of estimators in, for example, Amemiya (1985), Newey and McFadden (1994), Davidson and MacKinnon (1993) and Hayashi (2000). In practical terms, it is rarely possible to verify the conditions for an estimation problem at hand, and they are usually simply assumed. However, finding violations of the conditions is sometimes more straightforward, and this is worth pursuing. For example, lack of parametric identification can often be detected by analyzing the model, itself.

16.5.5 TESTING HYPOTHESES

The preceding describes a set of results that (more or less) unifies the theoretical underpinnings of three of the major classes of estimators in econometrics, least squares, maximum likelihood, and GMM. A similar body of theory has been produced for the familiar test statistics, Wald, likelihood ratio (LR), and Lagrange multiplier (LM). [See Newey and McFadden (1994).] All of these have been laid out in practical terms elsewhere in this text, so in the interest of brevity, we will refer the interested reader to the background sources listed for the technical details. Table 16.5 lists the locations in this text for various presentations of the testing procedures.

TABLE 16.5 Text References for Testing Procedures

Modeling Framework	*Wald*	*LR*	*LM*
Least Squares	6.3.1, 6.4	17.6.1	Exercise 6.7
Nonlinear LS	9.4.1	9.4.1	9.4.2
Maximum Likelihood	17.5.2	17.5.1	17.5.3
GMM	18.4.2	18.4.2	18.4.2

16.6 SUMMARY AND CONCLUSIONS

This chapter has presented a short overview of estimation in econometrics. There are various ways to approach such a survey. The current literature can be broadly grouped by three major types of estimators—parametric, semiparametric, and nonparametric. It has been suggested that the overall drift in the literature is from the first toward the third of these, but on a closer look, we see that this is probably not the case. Maximum likelihood is still the estimator of choice in many settings. New applications have been found for the GMM estimator, but at the same time, new Bayesian and simulation estimators, all fully parametric, are emerging at a rapid pace. Certainly, the range of tools that can be applied in any setting is growing steadily.

Key Terms and Concepts

- Bandwidth
- Bayesian estimation
- Bayes factor
- Bayes Theorem
- Conditional density
- Conjugate prior
- Criterion function
- Data generating mechanism
- Density
- Estimation criterion
- Extremum estimator
- Generalized method of moments
- Gibbs sampler
- Hierarchical Bayes
- Highest posterior density interval
- Histogram
- Informative prior
- Inverted gamma distribution
- Joint posterior distribution
- Kernel density estimator
- Latent class model
- Least absolute deviations
- Likelihood function
- Linear model
- Loss function
- M estimator
- Markov Chain Monte Carlo method
- Maximum likelihood estimator
- Method of moments
- Metropolis Hastings algorithm
- Multivariate t distribution
- Nearest neighbor
- Noninformative prior
- Nonparametric estimators
- Normal-gamma
- Parameter space
- Parametric estimation
- Partially linear model
- Posterior density
- Precision matrices
- Prior belief
- Prior distribution
- Prior odds ratio
- Prior probabilities
- Quantile regression
- Semiparametric estimation
- Simulation-based estimation
- Smoothing function

Exercises and Questions

1. Compare the fully parametric and semiparametric approaches to estimation of a discrete choice model such as the multinomial logit model discussed in Chapter 21. What are the benefits and costs of the semiparametric approach?
2. Asymptotics take on a different meaning in the Bayesian estimation context, since parameters do not "converge" to a population quantity. Nonetheless, in a Bayesian estimation setting, as the sample size increases, the likelihood function will dominate the posterior density. What does this imply about the Bayesian "estimator" when this occurs.
3. Referring to the situation in Question 2, one might think that an informative prior would outweigh the effect of the increasing sample size. With respect to the Bayesian analysis of the linear regression, analyze the way in which the likelihood and an informative prior will compete for dominance in the posterior mean.

The following exercises require specific software. The relevant techniques are available in several packages that might be in use, such as SAS, Stata, or LIMDEP. The exercises are suggested as departure points for explorations using a few of the many estimation techniques listed in this chapter.

4. Using the gasoline market data in Appendix Table F2.2, use the partially linear regression method in Section 16.3.3 to fit an equation of the form

$$\ln(G/Pop) = \beta_1 \ln(Income) + \beta_2 \ln P_{new\ cars} + \beta_3 \ln P_{used\ cars} + g(\ln P_{gasoline}) + \varepsilon$$

5. To continue the analysis in Question 4, consider a nonparametric regression of G/Pop on the price. Using the nonparametric estimation method in Section 16.4.2, fit the nonparametric estimator using a range of bandwidth values to explore the effect of bandwidth.
6. (You might find it useful to read the early sections of Chapter 21 for this exercise.) The extramarital affairs data analyzed in Section 22.3.7 can be reinterpreted in the context of a binary choice model. The dependent variable in the analysis is a count of events. Using these data, first recode the dependent variable 0 for none and 1 for more than zero. Now, first using the binary probit estimator, fit a binary choice model using the same independent variables as in the example discussed in Section 22.3.7. Then using a semiparametric or nonparametric estimator, estimate the same binary choice model. A model for binary choice can be fit for at least two purposes, for estimation of interesting coefficients or for prediction of the dependent variable. Use your estimated models for these two purposes and compare the two models.