12

ESTIMATION FRAMEWORKS IN ECONOMETRICS

12.1 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 to choose from this long menu. It is certainly beyond our scope to answer this question here, but a few principles will 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 semiparametric to nonparametric 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 17.) A formal probit or logit model allows estimation of probabilities, partial effects, and a host of ancillary results, but at the cost of imposing the normal or logistic distribution on the data. **Semiparametric estimators** 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 partial effects. The conclusions drawn based on the nonparametric and semiparametric estimators, such as they are, are robust.[1]

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 12.2 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 additional information used by the parametric estimator, 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).

[1]See, for example, the symposium in Angrist and Pischke (2010) for a spirited discussion on these points.

Schools of thought have punctuated this conversation. Proponents of **Bayesian estimation** often took an almost theological viewpoint in their criticism of their classical colleagues.[2] 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.[3] Thus, for example, the **simulation-based estimation** advocated in the many papers of Chib and Greenberg (for example, 1996) have provided solutions to a variety of computationally challenging problems. Arguments as to the methodological virtue of one approach or the other have received much less attention than before.

Chapters 2 through 7 of this book have focused on the classical regression model and a particular estimator, least squares (linear and nonlinear). In this and the next four 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. Chapter 13 discusses the **generalized method of moments**, which has emerged as the centerpiece of semiparametric estimation. Chapter 14 presents the method of **maximum likelihood**, the broad platform for parametric, classical estimation in econometrics. Chapter 15 discusses simulation-based estimation and bootstrapping. This is a body of techniques that have been made feasible by advances in estimation technology and which have made quite straightforward many estimators that were previously only scarcely used because of the sheer difficulty of the computations. Finally, Chapter 16 introduces the methods of Bayesian econometrics.

The list of techniques presented here is far from complete. We have chosen a set that constitutes the mainstream of econometrics. Certainly there are others that might be considered.[4] Virtually all of them are the subjects 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 12.2 with parametric approaches, primarily maximum likelihood. Because this is the subject of much of the remainder of this book, this section is brief. Section 12.2 also introduces Bayesian estimation, which in its traditional form is as heavily parameterized as maximum likelihood estimation. Section 12.3 is on semiparametric estimation. GMM estimation is the subject of all of Chapter 13, so it is 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 12.4 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 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 12.5, then in greater detail in Chapters 13 through 16.

[2]See, for example, Poirier (1995).

[3]The penetration of Bayesian methods in econometrics could be overstated. It is quite well represented in 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, of 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. Likewise, Wooldridge's (2010) widely cited treatise contains no mention of Bayesian econometrics. The one that displays any preference [for example, Mittelhammer et al. (2000)] devotes nearly 10% (70) of its pages to Bayesian estimation, but all to the broad metatheory of the linear regression model and none to the more elaborate applications that form the received applications in the many journals in the field.

[4]See, for example, Mittelhammer, Judge, and Miller (2000) for a lengthy catalog.

12.2 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

$$f(y, \mathbf{x}) = g(y|\mathbf{x}, \boldsymbol{\beta}) \times h(\mathbf{x}|\boldsymbol{\theta}), \tag{12-1}$$

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,

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

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 specifications 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 that 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 $[\beta, \sigma^2] \in \mathbb{R}^K \times \mathbb{R}+$. *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.

12.2.1 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**

$$f(y_1, y_2, \ldots, \mathbf{x}_1, \mathbf{x}_2, \ldots) = \prod_{i=1}^{n} f(y_i, \mathbf{x}_i|\boldsymbol{\beta}, \boldsymbol{\theta}). \tag{12-2}$$

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 the function of the data that (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 sample data, and the maximum likelihood estimator is the function of the sample information that 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. Because this estimator is the subject of Chapter 14, which is quite lengthy, we will defer any formal discussion until then and consider instead two applications to illustrate the techniques and underpinnings.

Example 12.1 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 $\boldsymbol{\beta}$, then $\mathbf{e}'\mathbf{e}/n$ follows as the estimator 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 14.8 and Section 14.9.2) 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. Second, one might want to proceed to an estimator with better finite sample properties. The least absolute deviations estimator discussed in Section 12.3.3 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 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 12.2 The Stochastic Frontier Model

The **stochastic frontier model**, discussed in detail in Chapter 19, is a regression-like model with a disturbance distribution that is asymmetric and distinctly nonnormal. The conditional density for the dependent variable in this skew normal model is

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

This produces a log-likelihood function for the model,

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

There are at least two fully parametric estimators for this model. The maximum likelihood estimator is discussed in Section 19.2.4. Greene (2007a) 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 the 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 13.2) are appropriate only for this distribution. Which is preferable? As we will see in Chapter 19, 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).

12.2.2 MODELING JOINT DISTRIBUTIONS WITH COPULA FUNCTIONS

Specifying the likelihood function commits the analyst to a possibly strong assumption about the distribution of the random variable of interest. The payoff, of course, is the stronger inferences that this permits. However, when there is more than one random variable of interest, such as in a joint household decision on health care usage in the example to follow, formulating the full likelihood involves specifying the marginal distributions, which might be comfortable, and a full specification of the joint distribution, which is likely to be less so. In the typical situation, the model might involve two similar random variables and an ill-formed specification of correlation between them. Implicitly, this case involves specification of the marginal distributions. The joint distribution is an empirical necessity to allow the correlation to be nonzero. The **copula function** approach provides a mechanism that the researcher can use to steer around this situation.

Trivedi and Zimmer (2007) suggest a variety of applications that fit this description:

- Financial institutions are often concerned with the prices of different, related (dependent) assets. The typical multivariate normality assumption is problematic because of GARCH effects (see Section 20.13) and thick tails in the distributions. While specifying appropriate marginal distributions may be reasonably straightforward, specifying the joint distribution is anything but that. Klugman and Parsa (2000) is an application.
- There are many microeconometric applications in which straightforward marginal distributions cannot be readily combined into a natural joint distribution. The bivariate event count model analyzed in Munkin and Trivedi (1999) and in the next example is an application.
- In the linear self-selection model of Chapter 19, the necessary joint distribution is part of a larger model. The likelihood function for the observed outcome involves the joint distribution of a variable of interest, hours, wages, income, and so on, and

the probability of observation. The typical application is based on a joint normal distribution. Smith (2003, 2005) suggests some applications in which a flexible copula representation is more appropriate. [In an intriguing early application of copula modeling that was not labeled as such, since it greatly predates the econometric literature, Lee (1983) modeled the outcome variable in a selectivity model as normal, the observation probability as logistic, and the connection between them using what amounted to the "Gaussian" copula function shown next.]

- Cheng and Trivedi (2015) used a copula function as an alternative to the bivariate normal distribution in analyzing attrition in a panel data set. (This application is examined in Example 11.3.)

Although the antecedents in the statistics literature date to Sklar's (1973) derivations, the applications in econometrics and finance are quite recent, with most applications appearing since 2000.[5]

Consider a modeling problem in which the marginal cdfs of two random variables can be fully specified as $F_1(y_1|\bullet)$ and $F_2(y_2|\bullet)$, where we condition on sample information (data) and parameters denoted "$\bullet$." For the moment, assume these are continuous random variables that obey all the axioms of probability. The bivariate cdf is $F_{12}(y_1, y_2|\bullet)$. A (bivariate) copula function (the results also extend to multivariate functions) is a function $C(u_1, u_2)$ defined over the unit square $[(0 \leq u_1 \leq 1) \times (0 \leq u_2 \leq 1)]$ that satisfies

$$\begin{aligned}
&(1)\quad C(1, u_2) = u_2 \text{ and } C(u_1, 1) = u_1,\\
&(2)\quad C(0, u_2) = C(u_1, 0) = 0,\\
&(3)\quad \partial C(u_1, u_2)/\partial u_1 \geq 0 \text{ and } \partial C(u_1, u_2)/\partial u_2 \geq 0.
\end{aligned}$$

These are properties of bivariate cdfs for random variables u_1 and u_2 that are bounded in the unit square. It follows that the copula function is a two-dimensional cdf defined over the unit square that has one-dimensional marginal distributions that are standard uniform in the unit interval [that is, property (1)]. To make profitable use of this relationship, we note that the cdf of a random variable, $F_1(y_1|\bullet)$, is, itself, a uniformly distributed random variable. This is the **fundamental probability transform** that we use for generating random numbers. (See Section 15.2.) In **Sklar's theorem**(1973), the marginal cdfs play the roles of u_1 and u_2. The theorem states that there exists a copula function, $C(...,...)$ such that

$$F_{12}(y_1, y_2|\bullet) = C[F_1(y_1|\bullet), F_2(y_2|\bullet)].$$

If $F_{12}(y_1, y_2|\bullet) = C[F_1(y_1|\bullet), F_2(y_2|\bullet)]$ is continuous and if the marginal cdfs have quantile (inverse) functions $F_j^{-1}(u_j)$ where $0 \leq u_j \leq 1$, then the copula function can be expressed as

$$\begin{aligned}
F_{12}(y_1, y_2|\bullet) &= F_{12}[F_1^{-1}(u_1|\bullet), F_2^{-1}(u_2|\bullet)]\\
&= \text{Prob}[U_1 \leq u_1, U_2 \leq u_2]\\
&= C(u_1, u_2).
\end{aligned}$$

[5]See the excellent survey by Trivedi and Zimmer (2007) for an extensive description.

In words, the theorem implies that the joint density can be written as the copula function evaluated at the two cumulative probability functions.

Copula functions allow the analyst to assemble joint distributions when only the marginal distributions can be specified. To fill in the desired element of correlation between the random variables, the copula function is written

$$F_{12}(y_1, y_2|\bullet) = C[F_1(y_1|\bullet), F_2(y_2|\bullet), \theta],$$

where θ is a dependence parameter. For continuous random variables, the joint pdf is then the mixed partial derivative,

$$\begin{aligned} f_{12}(y_1, y_2|\bullet) &= c_{12}[F_1(y_1|\bullet), F_2(y_2|\bullet), \theta] \\ &= \partial^2 C[F_1(y_1|\bullet), F_2(y_2|\bullet), \theta]/\partial y_1 \partial y_2 \\ &= [\partial^2 C(.,.,\theta)/\partial F_1 \partial F_2] f_1(y_1|\bullet) f_2(y_2|\bullet). \end{aligned} \tag{12-3}$$

A log-likelihood function can now be constructed using the logs of the right-hand sides of (12-3). Taking logs of (12-3) reveals the utility of the copula approach. The contribution of the joint observation to the log likelihood is

$$\ln f_{12}(y_1, y_2|\bullet) = \ln[\partial^2 C(.,.,\theta)/\partial F_1 \partial F_2] + \ln f_1(y_1|\bullet) + \ln f_2(y_2|\bullet).$$

Some of the common copula functions that have been used in applications are as follows:

Product: $C[u_1, u_2, \theta] = u_1 u_2,$

FGM: $C[u_1, u_2, \theta] = u_1 u_2[1 + \theta(1 - u_1)(1 - u_2)],$

Gaussian: $C[u_1, u_2, \theta] = \Phi_2[\Phi^{-1}(u_1), \Phi^{-1}(u_2), \theta],$

Clayton: $C[u_1, u_2, \theta] = [u_1^{-\theta} + u_2^{-\theta} - 1]^{-1/\theta},$

Frank: $C[u_1, u_2, \theta] = \dfrac{1}{\theta} \ln\left[1 + \dfrac{\exp(\theta u_1 - 1)\exp(\theta u_2 - 1)}{\exp(\theta) - 1}\right],$

Plackett: $C[u_1, u_2, \theta] = \dfrac{1 + (\theta - 1)(u_1 + u_2) - \sqrt{[1 + (\theta - 1)(u_1 + u_2)]^2 - 4\theta(\theta - 1)(u_1 u_2)}}{2(\theta - 1)}.$

The product copula implies that the random variables are independent because it implies that the joint cdf is the product of the marginals. In the FGM (Fairlie, Gumbel, Morgenstern) copula, it can be seen that $\theta = 0$ implies the product copula, or independence. The same result can be shown for the Clayton copula. Independence in the Plackett copula follows if $\theta = 1$. In the Gaussian function, the copula is the bivariate normal cdf if the marginals happen to be normal to begin with. The essential point is that the marginals need not be normal to construct the copula function, so long as the marginal cdfs can be specified. (The dependence parameter is not the correlation between the variables. Trivedi and Zimmer provide transformations of θ that are closely related to correlations for each copula function listed.)

The essence of the copula technique is that the researcher can specify and analyze the marginals and the copula functions separately. The likelihood function is obtained by formulating the cdfs [or the densities because the differentiation in (12-3) will reduce the joint density to a convenient function of the marginal densities] and the copula.

Example 12.3 Joint Modeling of a Pair of Event Counts

The standard regression modeling approach for a random variable, y, that is a count of events is the Poisson regression model,

$$\text{Prob}[Y = y|\mathbf{x}] = \exp(-\lambda)\lambda^y/y!, \text{ where } \lambda = \exp(\mathbf{x}'\boldsymbol{\beta}), y = 0, 1, \ldots.$$

More intricate specifications use the negative binomial model (version 2, NB2),

$$\text{Prob}[Y = y|\mathbf{x}] = \frac{\Gamma(y + \alpha)}{\Gamma(\alpha)\Gamma(y + 1)}\left(\frac{\alpha}{\lambda + \alpha}\right)^{\alpha}\left(\frac{\lambda}{\lambda + \alpha}\right)^{y}, y = 0, 1, \ldots,$$

where α is an overdispersion parameter. (See Section 18.4.) A satisfactory, appropriate specification for bivariate outcomes has been an ongoing topic of research. Early suggestions were based on a latent mixture model,

$$y_1 = z + w_1,$$
$$y_2 = z + w_2,$$

where w_1 and w_2 have the Poisson or NB2 distributions specified earlier with conditional means λ_1 and λ_2 and z is taken to be an unobserved Poisson or NB variable. This formulation induces correlation between the variables but is unsatisfactory because that correlation must be positive. In a natural application, y_1 is doctor visits and y_2 is hospital visits. These could be negatively correlated. Munkin and Trivedi (1999) specified the jointness in the conditional mean functions, in the form of latent, common heterogeneity,

$$\lambda_j = \exp(\mathbf{x}_j'\boldsymbol{\beta}_j + \varepsilon),$$

where ε is common to the two functions. Cameron et al. (2004) used a bivariate copula approach to analyze Australian data on self-reported and actual physician visits (the latter maintained by the Health Insurance Commission). They made two adjustments to the preceding model we developed above. First, they adapted the basic copula formulation to these discrete random variables. Second, the variable of interest to them was not the actual or self-reported count but the difference. Both of these are straightforward modifications of the basic copula model.

Example 12.4 The Formula That Killed Wall Street[6]

David Li (2000) designed a bivariate normal (Gaussian) copula model for the pricing of collateralized debt obligations (CDOs) such as mortgage-backed securities. The methodology he proposed became a widely used tool in the mortgage-backed securities market. The model appeared to work well when markets were stable, but failed spectacularly in the turbulent period around the financial crisis of 2008–2009. Li has been (surely unfairly) deemed partly to blame for the financial crash of 2008.[7]

12.3 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.

[6]Salmon (2000) and Li (1999, 2000).

[7]For example, Lee (2009), Hombrook (2009), Jones (2009), many others. From the CBC article: "… David Li is a Canadian math whiz who, some now say, developed the risk formula that destroyed Wall Street."

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 12.2, 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.

12.3.1 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, $(1/n)\Sigma_i[\mathbf{x}_i(y_i - \mathbf{x}_i'\mathbf{b})] = \mathbf{0}$. These are, of course, the 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 13, so we will defer further analysis until then.

12.3.2 MAXIMUM EMPIRICAL LIKELIHOOD ESTIMATION

Empirical likelihood methods are suggested as a semiparametric alternative to maximum likelihood. As we shall see shortly, the estimator is closely related to the GMM estimator. Let π_i denote generically the probability that $y_i|\mathbf{x}_i$ takes the realized value in the sample. Intuition suggests (correctly) that with no further information, π_i will equal $1/n$. The **empirical likelihood function** is

$$EL = \prod_{i=1}^{n} \pi_i^{1/n}.$$

The **maximum empirical likelihood estimator** maximizes EL. Equivalently, we maximize the log of the empirical likelihood,

$$ELL = \frac{1}{n}\sum_{i=1}^{n} \ln \pi_i.$$

As a maximization problem, this program lacks sufficient structure to admit a solution—the solutions for π_i are unbounded. If we impose the restrictions that π_i are probabilities that sum to one, we can use a Lagrangean formulation to solve the optimization problem,

$$ELL = \left[\frac{1}{n}\sum_{i=1}^{n} \ln \pi_i\right] + \lambda\left[1 - \sum_{i=1}^{n} \pi_i\right].$$

This slightly restricts the problem since with $0 < \pi_i < 1$ and $\Sigma_i \pi_i = 1$, the solution suggested earlier becomes obvious. (There is nothing in the problem that differentiates the π_i's so they must all be equal to each other.) Inserting this result in the derivative with respect to any specific π_i produces the remaining result, $\lambda = 1$.

The maximization problem becomes meaningful when we impose a structure on the data. To develop an example, we'll recall Example 7.6, a nonlinear regression equation for *Income* for the German Socioeconomic Panel data, where we specified

$$E[Income \,|\, Age, Sex, Education] = \exp(\mathbf{x}'\boldsymbol{\beta}) = h(\mathbf{x}, \boldsymbol{\beta}).$$

For an example, assume that *Education* may be endogenous in this equation, but we have available a set of instruments, $\mathbf{z}$, say (*Age*, *Health*, *Sex*, *Market Condition*). We have assumed that there are more instruments (4) than included variables (3), so that the parameters will be overidentified (and the example will be complicated enough to be interesting). (See Sections 8.3.4 and 8.9.) The orthogonality conditions for nonlinear instrumental variable estimation are that the disturbances be uncorrelated with the instrumental variables, so

$$E\{\mathbf{z}_i[Income_i - h(\mathbf{x}_i, \boldsymbol{\beta})]\} = E[\mathbf{m}_i(\boldsymbol{\beta})] = \mathbf{0}.$$

The nonlinear least squares solution to this problem was developed in Section 8.9. A GMM estimator will minimize with respect to $\boldsymbol{\beta}$ the criterion function

$$q = \overline{\mathbf{m}}'(\boldsymbol{\beta})\mathbf{A}\overline{m}(\boldsymbol{\beta}),$$

where $\mathbf{A}$ is the chosen weighting matrix. Note that for our example, including the constant term, there are four elements in $\boldsymbol{\beta}$ and five moment equations, so the parameters are overidentified.

If, instead, we impose the restrictions implied by our moment equations on the empirical likelihood function, we obtain the population moment condition,

$$\left[\sum_{i=1}^{n} \pi_i \mathbf{z}_i \times (Income_i - h(\mathbf{x}_i, \boldsymbol{\beta}))\right] = \mathbf{0}.$$

(The probabilities are population quantities, so this is the expected value.) This produces the constrained empirical log likelihood,

$$ELL = \left[\frac{1}{n}\sum_{i=1}^{n} \ln \pi_i\right] + \lambda\left[1 - \sum_{i=1}^{n} \pi_i\right] + \boldsymbol{\gamma}'\left[\sum_{i=1}^{n} \pi_i \mathbf{z}_i \,(Income_i - h(\mathbf{x}_i, \boldsymbol{\beta}))\right].$$

The function is now maximized with respect to π_i, λ, $\boldsymbol{\beta}$ (K elements), and $\boldsymbol{\gamma}$ (L elements, the number of instrumental variables). At the solution, the values of π_i provide, essentially, a set of weights. Cameron and Trivedi (2005, p. 205) provide a solution for $\hat{\pi}_i$ in terms of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and show, once again, that $\lambda = 1$. The concentrated ELL function with these inserted provides a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ that remains to be maximized.

The empirical likelihood estimator has the same asymptotic properties as the GMM estimator. (This makes sense, given the resemblance of the estimation criteria—ultimately, both are focused on the moment equations.) There is evidence that, at least in some cases, the finite sample properties of the empirical likelihood estimator might be better than GMM. A survey appears in Imbens (2002). One suggested modification of the procedure is to replace the core function in $(1/n)\Sigma_i \ln \pi_i$ with the **entropy** measure,

$$Entropy = -(1/n)\Sigma_i \pi_i \ln \pi_i.$$

The **maximum entropy** estimator is developed in Golan, Judge, and Miller (1996) and Golan (2009).

12.3.3 LEAST ABSOLUTE DEVIATIONS ESTIMATION AND QUANTILE REGRESSION

Least squares can be severely distorted by outlying observations in a small sample. 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, this discussion is moot.) These applications have led to the proposal of robust estimators that are unaffected by outlying observations. One of these, the least absolute deviations, or LAD estimator discussed in Section 7.3.1, is also useful in its own right as an estimator of the conditional median function in the modified model

$$\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}_{.50}.$$

That is, rather than providing a robust alternative to least squares as an estimator of the slopes of $E[y|\mathbf{x}]$, LAD is an estimator of a different feature of the population. This is essentially a semiparametric specification in that it specifies only a particular feature of the distribution, its median, but not the distribution itself. It also specifies that the conditional median be a linear function of $\mathbf{x}$.

The median, in turn, is only one possible quantile of interest. If the model is extended to other quantiles of the conditional distribution, we obtain

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\boldsymbol{\beta}_q|\mathbf{x}] = q, 0 < q < 1.$$

This is essentially a semiparametric specification. No assumption is made about the distribution of $y|\mathbf{x}$ or about its conditional variance. The fact that q can vary continuously (strictly) between zero and one means that there is an infinite number of possible parameter vectors. It seems reasonable to view the coefficients, which we might write $\boldsymbol{\beta}(q)$ less as fixed parameters, as we do in the linear regression model, than loosely as features of the distribution of $y|\mathbf{x}$. For example, it is not likely to be meaningful to view $\boldsymbol{\beta}(.49)$ to be discretely different from $\boldsymbol{\beta}(.50)$ or to compute precisely a particular difference such as $\boldsymbol{\beta}(.5) - \boldsymbol{\beta}(.3)$. On the other hand, the qualitative difference, or possibly the lack of a difference, between $\boldsymbol{\beta}(.3)$ and $\boldsymbol{\beta}(.5)$ may well be an interesting characteristic of the population. The quantile regression model is examined in Section 7.3.2.

12.3.4 KERNEL DENSITY METHODS

The kernel density estimator is an inherently nonparametric tool, so it fits more appropriately into the next section. But some models that use kernel methods are not completely nonparametric. The partially linear model in Section 7.4 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 12.5 Semiparametric Estimator for Binary Choice Models

The core binary choice model analyzed in Section 17.3, 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 $\boldsymbol{\beta}$ is fragile with respect to failures of the distributional assumption. We will examine

a few semiparametric and nonparametric estimators in Section 17.4.7. To illustrate the nature of the modeling process, we consider an estimator suggested by Lewbel (2000). The probit model is based on the normal distribution, with $\text{Prob}[y_i = 1 | \mathbf{x}_i] = \text{Prob}[\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0]$ where $\varepsilon_i \sim \mathbf{N}[0, 1]$. The estimator of $\boldsymbol{\beta}$ under this specification may 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 $\boldsymbol{\beta}$ can be obtained by *regression* of $[y_i - s(v_i)]/f(v_i | \mathbf{x}_i)$ on $\mathbf{x}_i$ where $s(v_i) = 1$ if $v_i > 0$ and 0 otherwise and $f(v_i | \mathbf{x}_i)$ is a kernel density estimator of the density of $v_i | \mathbf{x}_i$. 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 $\boldsymbol{\beta}$ but the distribution is unspecified.

12.3.5 COMPARING PARAMETRIC AND SEMIPARAMETRIC ANALYSES

It is often of interest to compare the outcomes of parametric and semiparametric models. As we have noted earlier, the strong assumptions of the fully parametric model come at a cost; the inferences from the model are only as robust as the underlying assumptions. Of course, the other side of that argument is that when the assumptions are met, parametric models represent efficient strategies for analyzing the data. The alternative, semiparametric approaches, relax assumptions such as normality and homoscedasticity. It is important to note that the model extensions to which semiparametric estimators are typically robust render the more heavily parameterized estimators inconsistent. The comparison is not just one of efficiency. As a consequence, comparison of parameter estimates can be misleading—the parametric and semiparametric estimators are often estimating very different quantities.

Example 12.6 A Model of Vacation Expenditures

Melenberg and van Soest (1996) analyzed the 1981 vacation expenditures of a sample of 1,143 Dutch families. The important feature of the data that complicated the analysis was that 37% (423) of the families reported zero expenditures. A linear regression that ignores this feature of the data would be heavily skewed toward underestimating the response of expenditures to the covariates such as total family expenditures (budget), family size, age, or education. (See Section 19.3.) The standard parametric approach to analyzing data of this sort is the Tobit, or censored, regression model,

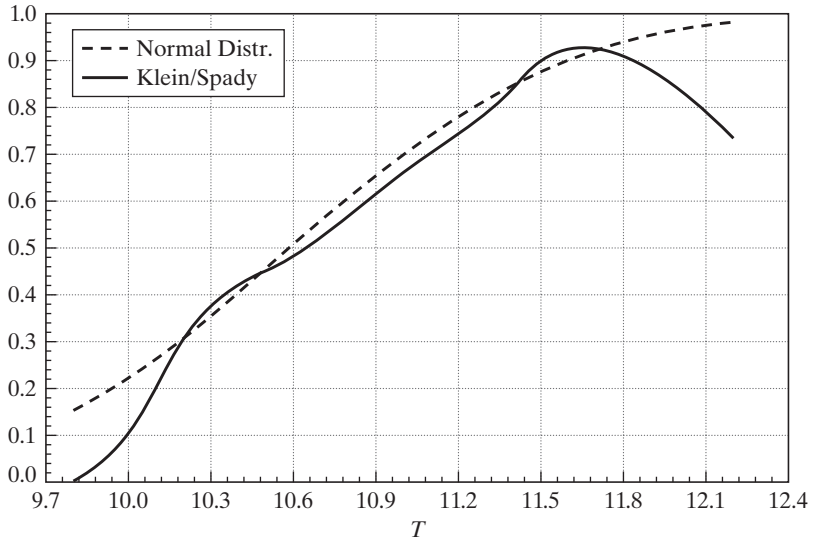
$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim \text{N}[0, \sigma^2]$$
$$y_i = \max(0, y_i^*),$$

or a two-part model that models the participation (zero or positive expenditure) and intensity (expenditure given positive expenditure) as separate decisions. (Maximum likelihood estimation of this model is examined in detail in Section 19.3.) The model rests on two strong assumptions, normality and homoscedasticity. Both assumptions can be relaxed in a more elaborate parametric framework, but the authors found that test statistics persistently rejected one or both of the assumptions even with the extended specifications. An alternative approach that is robust to both is Powell's (1984, 1986a,b) censored least absolute deviations estimator, which is a more technically demanding computation based on the LAD estimator in Section 7.3.1. Not surprisingly, the parameter estimates produced by the two approaches vary widely. The authors computed a variety of estimators of $\boldsymbol{\beta}$. A useful exercise that they

did not undertake would be to compare the partial effects from the different models. This is a benchmark on which the differences between the different estimators can sometimes be reconciled. In the Tobit model, $\partial E[y_i|\mathbf{x}_i]/\partial\mathbf{x}_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)\boldsymbol{\beta}$ (see Section 19.3). It is unclear how to compute the counterpart in the semiparametric model, since the underlying specification holds only that $\text{Med}[\varepsilon_i|\mathbf{x}_i] = 0$. (The authors report on the *Journal of Applied Econometrics* data archive site that these data are proprietary. As such, we were unable to extend the analysis to obtain estimates of partial effects.) This highlights a significant difficulty with the semiparametric approach to estimation. In a nonlinear model such as this one, it is often the partial effects that are of interest, not the coefficients. But one of the byproducts of the more robust specification is that the partial effects are not defined.

In a second stage of the analysis, the authors decomposed their expenditure equation into a participation equation that modeled probabilities for the binary outcome "expenditure $= 0$ or > 0" and a conditional expenditure equation for those with positive expenditure.[8] For this step, the authors once again used a parametric model based on the normal distribution (the probit model—see Section 17.3) and a semiparametric model that is robust to distribution and heteroscedasticity developed by Klein and Spady (1993). As before, the coefficient estimates differ substantially. However, in this instance, the specification tests are considerably more sympathetic to the parametric model. Figure 12.1, which reproduces their Figure 2, compares the predicted probabilities from the two models. The dashed curve is the probit model. Within the range of most of the data, the models give quite similar predictions. Once again, however, it is not possible to compare partial effects. The interesting outcome from this part of the analysis seems to be that the failure of the parametric specification resides more in the modeling of the continuous expenditure variable than with the model that separates the two subsamples based on zero or positive expenditures.

FIGURE 12.1 Predicted Probabilities of Positive Expenditure.



[8]In Section 18.4.8, we will label this a "hurdle" model. See Mullahy (1986).

12.4 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.[9]

12.4.1 KERNEL DENSITY ESTIMATION

Sample statistics such as 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

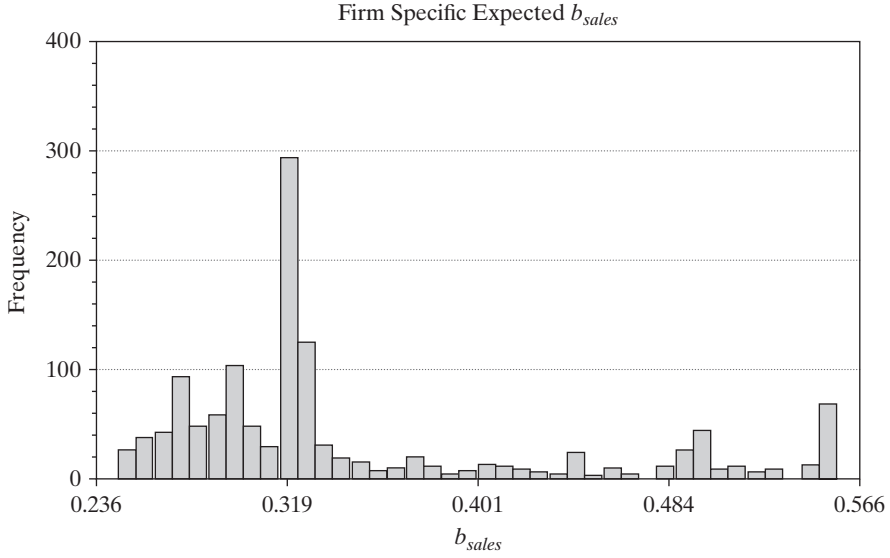
$$\hat{f}(x) = f(x|\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\mu}}{\hat{\sigma}} \right)^2 \right].$$

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 an example. In Example 15.17 and in Greene (2004c), we estimate a model that produces a conditional estimator of a slope vector for each of the 1,270 firms in the sample. We might be interested in the distribution of these estimators across firms. In particular, the conditional 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 0.4013 might suggest some skewness. The histogram in Figure 12.2 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.) Because 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

[9]The set of literature in this area of econometrics is large and rapidly growing. Major references which provide an applied and theoretical foundation are Härdle (1990), Pagan and Ullah (1999), and Li and Racine (2007).

FIGURE 12.2 Histogram for Estimated b_{sales} Coefficients.



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 distances 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 that 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 one half-bin width of x_k.

Albeit rather crude, this "naïve" (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 naïve 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 naïve 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 12.1. Each of these is smooth, continuous, symmetric, and equally attractive. The 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 a 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, normal and Epanechnikov kernels appear to be the default choices in many applications.)

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

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

Because $g(x_i|\, z, h)$ 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.[10] The bias is a function of the bandwidth, but for an appropriate choice of h, the bias does vanish asymptotically. As intuition might suggest, the larger the bandwidth, the greater the bias, but at the same time, the smaller 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. For the proportion of observations captured in the bin to converge to the

TABLE 12.1 Kernel Functions for Density Estimation

Kernel	***Formula K[z]***
Epanechnikov	$0.75(1 - 0.2z^2)/\sqrt{5}$ if $\lvert z\rvert \leq \sqrt{5}$, 0 else
Normal	$\phi(z)$ (normal density)
Logit	$\Lambda(z)[1 - \Lambda(z)]$ (logistic density)
Uniform	0.5 if $\lvert z\rvert \leq 1$, 0 else
Beta	$0.75(1 - z)(1 + z)$ if $\lvert z\rvert \leq 1$, 0 else
Cosine	$1 + \cos(2\pi z)$ if $\lvert z\rvert \leq 0.5$, 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 $\lvert z\rvert \leq 0.5$, $8(1 - \lvert z\rvert)^3/3$ if $0.5 < \lvert z\rvert \leq 1$, 0 else

[10]See also Li and Racine (2007) and Henderson and Parmeter (2015).

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 Silverman's (1986) bandwidth, $h = 0.9 \times s/n^{1/5}$. To conclude the illustration begun earlier, Figure 12.3 is a logit-based kernel density estimator for the distribution of slope estimates for the model estimated earlier. The resemblance to the histogram in Figure 12.2 is to be expected.

12.5 PROPERTIES OF ESTIMATORS

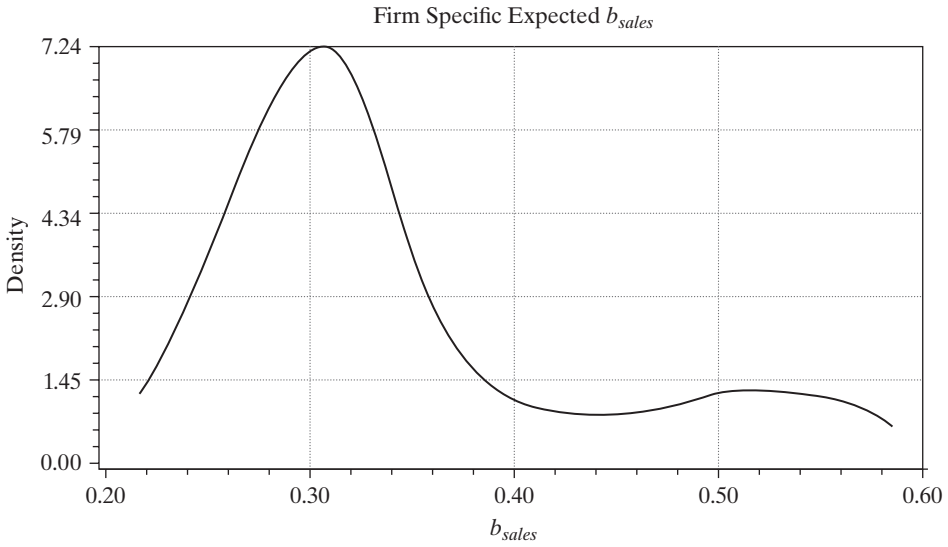
The preceding has been concerned with methods of estimation. We have surveyed a variety of techniques that have appeared in the applied literature. We have not yet examined the statistical properties of these estimators. Although, as noted earlier, we will leave extensive analysis of the asymptotic theory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

12.5.1 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.[11] In most cases, however, discussions of asymptotic

FIGURE 12.3 Kernel Density for b_{sales}.



[11]See, for example, Pagan and Ullah (1999, Section 2.5.1) and Henderson and Parmeter (2015, Section 2.2) on the subject of the kernel density estimator.

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. In some instances, the inconsistency can be more precisely quantified. For example, the "incidental parameters problem" (see Section 17.7.3) relates to estimation of fixed effects models in panel data settings in which an estimator is inconsistent for fixed T but is consistent in T (and tolerably biased for moderate sized T).
- **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, 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 that 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 relative efficiency of maximum likelihood and GMM estimators in the class of consistent and asymptotically normally distributed (CAN) estimators. There are two important practical considerations in this setting. First, the researcher will want to know that he or she has not made demonstrably suboptimal use of the 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; he or she has one at hand, and questions of relative efficiency are moot.

12.5.2 EXTREMUM ESTIMATORS

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

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

(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 him or her 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, for example, any of the sources mentioned in Footnote 1 of this chapter.) Amemiya (1985) and Davidson and MacKinnon (2004) are very accessible treatments. Newey and McFadden (1994) is a 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.

12.5.3 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.

1. **The parameter space** (see Section 12.2) 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.
2. **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.[12] 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.
3. **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

[12]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 in Amemiya (1985, p. 272).

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, 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, Θ. Define

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

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. Because the criterion function is a function of the data, this convergence requires a statement of the properties of the data, for example, well behaved in some sense. Leaving that aside for the moment, interestingly, the results to this 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 $\mathbf{x}_i$ that 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 7.2 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 $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ 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 $f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = f(y_i|\mathbf{x}_i'\boldsymbol{\theta})$. 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 $\mathrm{E}[\overline{\mathbf{m}}(\mathbf{data}, \boldsymbol{\theta})] \neq \mathbf{0}$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

4. **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 preceding definitions. 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 Lindeberg–Feller (D.19) or Lyapounov (D.20) theorems for cross sections or the Martingale Difference Central Limit theorem for dependent observations (Theorem 20.3). 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 Sections 4.4.1 and 4.4.2 under the heading of "well-behaved data." At this point, we will assume that the data are well behaved.

12.5.4 ASYMPTOTIC PROPERTIES OF ESTIMATORS

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

THEOREM 12.1 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, and (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) in Theorem 12.1. We have assumed identification. The fundamental device is the following: Because of its dependence on the data, $q(\boldsymbol{\theta}|\mathbf{data})$ is a random variable. We assumed in (c) that plim $q(\boldsymbol{\theta}|\mathbf{data}) = q_0(\boldsymbol{\theta})$ 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 $q(\boldsymbol{\theta}|\mathbf{data})$. 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 14.4.5.a where we establish consistency of the maximum likelihood estimator.

THEOREM 12.2 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) $q(\boldsymbol{\theta}|\mathbf{data})$ *is concave and twice continuously differentiable in* $\boldsymbol{\theta}$ *in a neighborhood of* $\boldsymbol{\theta}_0$;

(iii) $\sqrt{n}[\partial q(\boldsymbol{\theta}_0|\mathbf{data})/\partial\boldsymbol{\theta}_0] \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}]$;

(iv) *for any* $\boldsymbol{\theta}$ *in* $\boldsymbol{\Theta}$, $\lim_{n\to\infty} \Pr[|(\partial^2 q(\boldsymbol{\theta}|\mathbf{data})/\partial\theta_k\partial\theta_m) - h_{km}(\theta)| > \varepsilon] = 0 \,\forall\, \varepsilon > 0$ *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,

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

Each derivative is evaluated at a point $\overline{\boldsymbol{\theta}}$ that is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, that is, $\overline{\boldsymbol{\theta}} = w\hat{\boldsymbol{\theta}} + (1 - w)\boldsymbol{\theta}_0$ for some $0 < w < 1$. Because we have assumed plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, we see that the matrix in the second term on the right must be converging to $\mathbf{H}(\boldsymbol{\theta}_0)$. 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 14.4.5.b, 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 13, 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.[13] Some of the semiparametric and all of the nonparametric estimators noted require somewhat more intricate treatments. For example, Pagan and Ullah (1999, Sections 2.5–2.6) and Li and Racine (2007, Sections 1.9–1.12) 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

[13]See Koenker and Bassett (1982) and Amemiya (1985, pp. 152–154).

and MacKinnon (2004), 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.

12.5.5 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).[14] 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.

12.6 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

- Bayesian estimation
- Conditional density
- Copula function
- Criterion function
- Data-generating mechanism
- Density
- Empirical likelihood function
- Entropy
- Estimation criterion
- Extremum estimator
- Fundamental probability transform
- Generalized method of moments
- Histogram
- Kernel density estimator
- M estimator
- Maximum empirical likelihood estimator
- Maximum entropy
- Maximum likelihood estimator
- Method of moments
- Nonparametric estimators
- Semiparametric estimation
- Simulation-based estimation
- Sklar's theorem
- Stochastic frontier model

Exercise and Question

1. Compare the fully parametric and semiparametric approaches to estimation of a discrete choice model such as the multinomial logit model discussed in Chapter 17. What are the benefits and costs of the semiparametric approach?

[14] See Newey and McFadden (1994).