17
MAXIMUM LIKELIHOOD ESTIMATION

17.1 INTRODUCTION

The generalized method of moments discussed in Chapter 18 and the semiparametric, nonparametric, and Bayesian estimators discussed in Chapter 16 are becoming widely used by model builders. Nonetheless, the maximum likelihood estimator discussed in this chapter remains the preferred estimator in many more settings than the others listed. As such, we focus our discussion of generally applied estimation methods on this technique. Sections 17.2 through 17.5 present statistical results for estimation and hypothesis testing based on the maximum likelihood principle. After establishing some general results for this method of estimation, we will then extend them to the more familiar setting of econometric models. Some applications are presented in Section 17.6. Finally, three variations on the technique, maximum simulated likelihood, two-step estimation and pseudomaximum likelihood estimation are described in Sections 17.7 through 17.9.

17.2 THE LIKELIHOOD FUNCTION AND IDENTIFICATION OF THE PARAMETERS

The probability density function, or pdf for a random variable y, conditioned on a set of parameters, $\boldsymbol{\theta}$, is denoted $f(y \mid \boldsymbol{\theta})$.[1] This function identifies the data generating process that underlies an observed sample of data and, at the same time, provides a mathematical description of the data that the process will produce. The joint density of *n independent* and *identically distributed* (iid) observations from this process is the product of the individual densities;

$$f(y_1, \ldots, y_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid \mathbf{y}). \qquad \textbf{(17-1)}$$

This joint density is the **likelihood function,** defined as a function of the unknown parameter vector, $\boldsymbol{\theta}$, where $\mathbf{y}$ is used to indicate the collection of sample data. Note that we write the joint density as a function of the data conditioned on the parameters whereas when we form the likelihood function, we write this function in reverse, as a function of the parameters, conditioned on the data. Though the two functions are the same, it is to be emphasized that the likelihood function is written in this fashion to

[1] Later we will extend this to the case of a random vector, $\mathbf{y}$, with a multivariate density, but at this point, that would complicate the notation without adding anything of substance to the discussion.

highlight our interest in the parameters and the information about them that is contained in the observed data. However, it is understood that the likelihood function is not meant to represent a probability density for the parameters as it is in Section 16.2.2. In this classical estimation framework, the parameters are assumed to be fixed constants which we hope to learn about from the data.

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}). \tag{17-2}$$

Again, to emphasize our interest in the parameters, given the observed data, we denote this function $L(\boldsymbol{\theta} \mid \mathbf{data}) = L(\boldsymbol{\theta} \mid \mathbf{y})$. The likelihood function and its logarithm, evaluated at $\boldsymbol{\theta}$, are sometimes denoted simply $L(\boldsymbol{\theta})$ and $\ln L(\boldsymbol{\theta})$, respectively or, where no ambiguity can arise, just L or $\ln L$.

It will usually be necessary to generalize the concept of the likelihood function to allow the density to depend on other conditioning variables. To jump immediately to one of our central applications, suppose the disturbance in the classical linear regression model is normally distributed. Then, conditioned on it's specific $\mathbf{x}_i$, y_i is normally distributed with mean $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ and variance σ^2. That means that the observed random variables are not iid; they have different means. Nonetheless, the observations are independent, and as we will examine in closer detail,

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{n} [\ln \sigma^2 + \ln(2\pi) + (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma^2], \tag{17-3}$$

where $\mathbf{X}$ is the $n \times K$ matrix of data with ith row equal to $\mathbf{x}_i'$.

The rest of this chapter will be concerned with obtaining estimates of the parameters, $\boldsymbol{\theta}$ and in testing hypotheses about them and about the data generating process. Before we begin that study, we consider the question of whether estimation of the parameters is possible at all—the question of **identification.** Identification is an issue related to the formulation of the model. The issue of identification must be resolved before estimation can even be considered. The question posed is essentially this: Suppose we had an infinitely large sample—that is, for current purposes, all the information there is to be had about the parameters. Could we uniquely determine the values of $\boldsymbol{\theta}$ from such a sample? As will be clear shortly, the answer is sometimes no.

DEFINITION 17.1 Identification
*The parameter vector $\boldsymbol{\theta}$ is identified (**estimable**) if for any other parameter vector, $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, for some data $\mathbf{y}$, $L(\boldsymbol{\theta}^* \mid \mathbf{y}) \neq L(\boldsymbol{\theta} \mid \mathbf{y})$.*

This result will be crucial at several points in what follows. We consider two examples, the first of which will be very familiar to you by now.

Example 17.1 Identification of Parameters
For the regression model specified in (17-3), suppose that there is a nonzero vector $\mathbf{a}$ such that $\mathbf{x}_i'\mathbf{a} = 0$ for every $\mathbf{x}_i$. Then there is another "parameter" vector, $\boldsymbol{\gamma} = \boldsymbol{\beta} + \mathbf{a} \neq \boldsymbol{\beta}$ such that

$\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{x}_i'\boldsymbol{\gamma}$ for every $\mathbf{x}_i$. You can see in (17-3) that if this is the case, then the log-likelihood is the same whether it is evaluated at $\boldsymbol{\beta}$ or at $\boldsymbol{\gamma}$. As such, it is not possible to consider estimation of $\boldsymbol{\beta}$ in this model since $\boldsymbol{\beta}$ cannot be distinguished from $\boldsymbol{\gamma}$. This is the case of perfect collinearity in the regression model which we ruled out when we first proposed the linear regression model with "Assumption 2. Identifiability of the Model Parameters."

The preceding dealt with a necessary characteristic of the sample data. We now consider a model in which identification is secured by the specification of the parameters in the model. (We will study this model in detail in Chapter 21.) Consider a simple form of the regression model considered above, $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, where $\varepsilon_i \mid x_i$ has a normal distribution with zero mean and variance σ^2. To put the model in a context, consider a consumer's purchases of a large commodity such as a car where x_i is the consumer's income and y_i is the difference between what the consumer is willing to pay for the car, p_i^*, and the price tag on the car, p_i. Suppose rather than observing p_i^* or p_i, we observe only whether the consumer actually purchases the car, which, we assume, occurs when $y_i = p_i^* - p_i$ is positive. Collecting this information, our model states that they will purchase the car if $y_i > 0$ and not purchase it if $y_i \leq 0$. Let us form the likelihood function for the observed data, which are (purchase or not) and income. The random variable in this model is "purchase" or "not purchase"—there are only two outcomes. The probability of a purchase is

$$\begin{aligned}
\text{Prob(purchase} \mid \beta_1, \beta_2, \sigma, x_i) &= \text{Prob}(y_i > 0 \mid \beta_1, \beta_2, \sigma, x_i) \\
&= \text{Prob}(\beta_1 + \beta_2 x_i + \varepsilon_i > 0 \mid \beta_1, \beta_2, \sigma, x_i) \\
&= \text{Prob}[\varepsilon_i > -(\beta_1 + \beta_2 x_i) \mid \beta_1, \beta_2, \sigma, x_i] \\
&= \text{Prob}[\varepsilon_i/\sigma > -(\beta_1 + \beta_2 x_i)/\sigma \mid \beta_1, \beta_2, \sigma, x_i] \\
&= \text{Prob}[z_i > -(\beta_1 + \beta_2 x_i)/\sigma \mid \beta_1, \beta_2, \sigma, x_i]
\end{aligned}$$

where z_i has a standard normal distribution. The probability of not purchase is just one minus this probability. The likelihood function is

$$\prod_{i=\text{purchased}} [\text{Prob(purchase} \mid \beta_1, \beta_2, \sigma, x_i)] \prod_{i=\text{not purchased}} [1 - \text{Prob(purchase} \mid \beta_1, \beta_2, \sigma, x_i)].$$

We need go no further to see that the parameters of this model are not identified. If β_1, β_2 and σ are all multiplied by the same nonzero constant, regardless of what it is, then Prob(purchase) is unchanged, 1 − Prob(purchase) is also, and the likelihood function does not change. This model requires a **normalization.** The one usually used is $\sigma = 1$, but some authors [e.g., Horowitz (1993)] have used $\beta_1 = 1$ instead.

17.3 EFFICIENT ESTIMATION: THE PRINCIPLE OF MAXIMUM LIKELIHOOD

The principle of **maximum likelihood** provides a means of choosing an asymptotically efficient estimator for a parameter or a set of parameters. The logic of the technique is easily illustrated in the setting of a discrete distribution. Consider a random sample of the following 10 observations from a Poisson distribution: 5, 0, 1, 1, 0, 3, 2, 3, 4, and 1. The density for each observation is

$$f(y_i \mid \theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}.$$

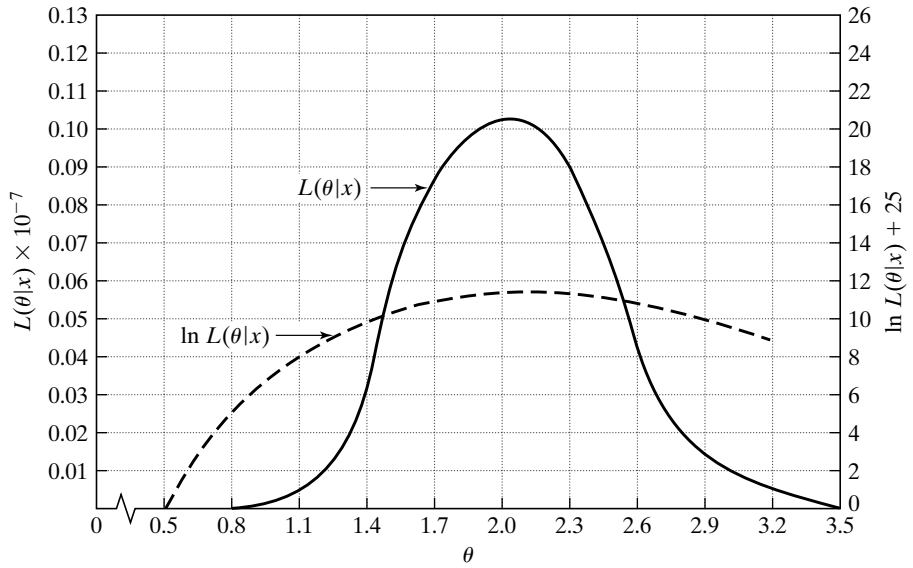


FIGURE 17.1 Likelihood and Log-likelihood Functions for a Poisson Distribution.

Since the observations are independent, their joint density, which is the likelihood for this sample, is

$$f(y_1, y_2, \ldots, y_{10} \mid \theta) = \prod_{i=1}^{10} f(y_i \mid \theta) = \frac{e^{-10\theta}\theta^{\Sigma_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!} = \frac{e^{-10\theta}\theta^{20}}{207,360}.$$

The last result gives the probability of observing *this particular sample,* assuming that a Poisson distribution with as yet unknown parameter θ generated the data. What value of θ would make this sample most probable? Figure 17.1 plots this function for various values of θ. It has a single mode at $\theta = 2$, which would be the **maximum likelihood estimate,** or MLE, of θ.

Consider maximizing $L(\theta \mid \mathbf{y})$ with respect to θ. Since the log function is monotonically increasing and easier to work with, we usually maximize $\ln L(\theta \mid \mathbf{y})$ instead; in sampling from a Poisson population,

$$\ln L(\theta \mid \mathbf{y}) = -n\theta + \ln\theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!),$$

$$\frac{\partial \ln L(\theta \mid \mathbf{y})}{\partial \theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} y_i = 0 \Rightarrow \hat{\theta}_{\mathrm{ML}} = \bar{y}_n.$$

For the assumed sample of observations,

$$\ln L(\theta \mid \mathbf{y}) = -10\theta + 20\ln\theta - 12.242,$$

$$\frac{d \ln L(\theta \mid \mathbf{y})}{d\theta} = -10 + \frac{20}{\theta} = 0 \Rightarrow \hat{\theta} = 2,$$

and

$$\frac{d^2 \ln L(\theta \mid \mathbf{y})}{d\theta^2} = \frac{-20}{\theta^2} < 0 \Rightarrow \text{this is a maximum.}$$

The solution is the same as before. Figure 17.1 also plots the log of $L(\theta \mid \mathbf{y})$ to illustrate the result.

The reference to the probability of observing the given sample is not exact in a continuous distribution, since a particular sample has probability zero. Nonetheless, the principle is the same. The values of the parameters that maximize $L(\boldsymbol{\theta} \mid \mathbf{data})$ or its log are the maximum likelihood estimates, denoted $\hat{\boldsymbol{\theta}}$. Since the logarithm is a monotonic function, the values that maximize $L(\boldsymbol{\theta} \mid \mathbf{data})$ are the same as those that maximize $\ln L(\boldsymbol{\theta} \mid \mathbf{data})$. The necessary condition for maximizing $\ln L(\boldsymbol{\theta} \mid \mathbf{data})$ is

$$\frac{\partial \ln L(\boldsymbol{\theta} \mid \mathbf{data})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{17-4}$$

This is called the **likelihood equation.** The general result then is that the MLE is a root of the likelihood equation. The application to the parameters of the dgp for a discrete random variable are suggestive that maximum likelihood is a "good" use of the data. It remains to establish this as a general principle. We turn to that issue in the next section.

Example 17.2 Log Likelihood Function and Likelihood Equations for the Normal Distribution

In sampling from a normal distribution with mean μ and variance σ^2, the log-likelihood function and the likelihood equations for μ and σ^2 are

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - \mu)^2}{\sigma^2}\right], \tag{17-5}$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0, \tag{17-6}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0. \tag{17-7}$$

To solve the likelihood equations, multiply (17-6) by σ^2 and solve for $\hat{\mu}$, then insert this solution in (17-7) and solve for σ^2. The solutions are

$$\hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_n \quad \text{and} \quad \hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2. \tag{17-8}$$

17.4 PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

Maximum likelihood estimators (MLEs) are most attractive because of their large-sample or asymptotic properties.

DEFINITION 17.2 Asymptotic Efficiency
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed (CAN), and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.[2]

If certain regularity conditions are met, the MLE will have these properties. The finite sample properties are sometimes less than optimal. For example, the MLE may be biased; the MLE of σ^2 in Example 17.2 is biased downward. The occasional statement that the properties of the MLE are *only* optimal in large samples is not true, however. It can be shown that when sampling is from an exponential family of distributions (see Definition 18.1), there will exist sufficient statistics. If so, MLEs will be functions of them, which means that when minimum variance unbiased estimators exist, they will be MLEs. [See Stuart and Ord (1989).] Most applications in econometrics do not involve exponential families, so the appeal of the MLE remains primarily its asymptotic properties.

We use the following notation: $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator; $\boldsymbol{\theta}_0$ denotes the true value of the parameter vector; $\boldsymbol{\theta}$ denotes another possible value of the parameter vector, not the MLE and not necessarily the true values. Expectation based on the true values of the parameters is denoted $E_0[.]$. If we assume that the regularity conditions discussed below are met by $f(\mathbf{x}, \boldsymbol{\theta}_0)$, then we have the following theorem.

THEOREM 17.1 Properties of an MLE
Under regularity, the maximum likelihood estimator (MLE) has the following asymptotic properties:

M1. ***Consistency:*** $\operatorname{plim} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

M2. ***Asymptotic normality:*** $\hat{\boldsymbol{\theta}} \overset{a}{\sim} N[\boldsymbol{\theta}_0, \{\mathbf{I}(\boldsymbol{\theta}_0)\}^{-1}]$, *where*

$$\mathbf{I}(\boldsymbol{\theta}_0) = -E_0[\partial^2 \ln L/\partial\boldsymbol{\theta}_0 \partial\boldsymbol{\theta}_0'].$$

M3. ***Asymptotic efficiency:*** $\hat{\boldsymbol{\theta}}$ *is asymptotically efficient and achieves the* ***Cramér–Rao lower bound*** *for consistent estimators, given in M2 and Theorem C.2.*

M4. ***Invariance:*** *The maximum likelihood estimator of* $\boldsymbol{\gamma}_0 = \mathbf{c}(\boldsymbol{\theta}_0)$ *is* $\mathbf{c}(\hat{\boldsymbol{\theta}})$ *if* $\mathbf{c}(\boldsymbol{\theta}_0)$ *is a continuous and continuously differentiable function.*

17.4.1 REGULARITY CONDITIONS

To sketch proofs of these results, we first obtain some useful properties of probability density functions. We assume that $(y_1, \ldots, y_n)$ is a random sample from the population

[2] *Not larger* is defined in the sense of (A-118): The covariance matrix of the less efficient estimator equals that of the efficient estimator plus a nonnegative definite matrix.

with density function $f(y_i \mid \boldsymbol{\theta}_0)$ and that the following **regularity conditions** hold. [Our statement of these is informal. A more rigorous treatment may be found in Stuart and Ord (1989) or Davidson and MacKinnon (1993).]

DEFINITION 17.3 Regularity Conditions

R1. *The first three derivatives of* $\ln f(y_i \mid \boldsymbol{\theta})$ *with respect to* $\boldsymbol{\theta}$ *are continuous and finite for almost all* y_i *and for all* $\boldsymbol{\theta}$*. This condition ensures the existence of a certain Taylor series approximation and the finite variance of the derivatives of* $\ln L$*.*

R2. *The conditions necessary to obtain the expectations of the first and second derivatives of* $\ln f(y_i \mid \boldsymbol{\theta})$ *are met.*

R3. *For all values of* $\boldsymbol{\theta}$, $|\partial^3 \ln f(y_i \mid \boldsymbol{\theta})/\partial\theta_j\partial\theta_k\partial\theta_l|$ *is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.*

With these regularity conditions, we will obtain the following fundamental characteristics of $f(y_i \mid \boldsymbol{\theta})$: D1 is simply a consequence of the definition of the likelihood function. D2 leads to the moment condition which defines the maximum likelihood estimator. On the one hand, the MLE is found as the maximizer of a function, which mandates finding the vector which equates the gradient to zero. On the other, D2 is a more fundamental relationship which places the MLE in the class of generalized method of moments estimators. D3 produces what is known as the **Information matrix equality.** This relationship shows how to obtain the asymptotic covariance matrix of the MLE.

17.4.2 PROPERTIES OF REGULAR DENSITIES

Densities that are "regular" by Definition 17.3 have three properties which are used in establishing the properties of maximum likelihood estimators:

THEOREM 17.2 Moments of the Derivatives of the Log-Likelihood

D1. $\ln f(y_i \mid \boldsymbol{\theta})$, $\mathbf{g}_i = \partial \ln f(y_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}$, *and* $\mathbf{H}_i = \partial^2 \ln f(y_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, $i = 1, \ldots, n$, *are all random samples of random variables. This statement follows from our assumption of random sampling. The notation* $\mathbf{g}_i(\boldsymbol{\theta}_0)$ *and* $\mathbf{H}_i(\boldsymbol{\theta}_0)$ *indicates the derivative evaluated at* $\boldsymbol{\theta}_0$*.*

D2. $E_0[\mathbf{g}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$.

D3. $\text{Var}[\mathbf{g}_i(\boldsymbol{\theta}_0)] = -E[\mathbf{H}_i(\boldsymbol{\theta}_0)]$.

Condition D1 is simply a consequence of the definition of the density.

For the moment, we allow the range of y_i to depend on the parameters; $A(\boldsymbol{\theta}_0) \leq y_i \leq B(\boldsymbol{\theta}_0)$. (Consider, for example, finding the maximum likelihood estimator of θ/break

for a continuous uniform distribution with range $[0, \theta_0]$.) (In the following, the single integral $\int \ldots dy_i$, would be used to indicate the multiple integration over all the elements of a multivariate of y_i if that were necessary). By definition,

$$\int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} f(y \mid \boldsymbol{\theta}_0)\, dy_i = 1.$$

Now, differentiate this expression with respect to $\boldsymbol{\theta}_0$. Leibnitz's theorem gives

$$\begin{aligned}\frac{\partial \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} f(y_i \mid \boldsymbol{\theta}_0)\, dy_i}{\partial \boldsymbol{\theta}_0} &= \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} \frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\, dy_i + f(B(\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0)\frac{\partial B(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \\ &\quad - f(A(\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0)\frac{\partial A(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \\ &= \mathbf{0}.\end{aligned}$$

If the second and third terms go to zero, then we may interchange the operations of differentiation and integration. The necessary condition is that $\lim_{y_i \to A(\boldsymbol{\theta}_0)} f(y_i \mid \boldsymbol{\theta}_0) = \lim_{y_i \to B(\boldsymbol{\theta}_0)} f(y_i \mid \boldsymbol{\theta}_0) = 0$. (Note that the uniform distribution suggested above violates this condition.) Sufficient conditions are that the range of the observed random variable, y_i, does not depend on the parameters, which means that $\partial A(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \partial B(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \mathbf{0}$ or that the density is zero at the terminal points. This condition, then, is regularity condition R2. The latter is usually assumed, and we will assume it in what follows. So,

$$\frac{\partial \int f(y_i \mid \boldsymbol{\theta}_0)\, dy_i}{\partial \boldsymbol{\theta}_0} = \int \frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\, dy_i = \int \frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} f(y_i \mid \boldsymbol{\theta}_0)\, dy_i = E_0\left[\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right] = \mathbf{0}.$$

This proves D2.

Since we may interchange the operations of integration and differentiation, we differentiate under the integral once again to obtain

$$\int \left[\frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} f(y_i \mid \boldsymbol{\theta}_0) + \frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\right] dy_i = \mathbf{0}.$$

But

$$\frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} = f(y_i \mid \boldsymbol{\theta}_0)\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'},$$

and the integral of a sum is the sum of integrals. Therefore,

$$-\int \left[\frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right] f(y_i \mid \boldsymbol{\theta}_0)\, dy_i = \int \left[\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\right] f(y_i \mid \boldsymbol{\theta}_0)\, dy_i = [\mathbf{0}].$$

The left-hand side of the equation is the negative of the expected second derivatives matrix. The right-hand side is the expected square (outer product) of the first derivative vector. But, since this vector has expected value $\mathbf{0}$ (we just showed this), the right-hand side is the variance of the first derivative vector, which proves D3:

$$\mathrm{Var}_0\left[\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right] = E_0\left[\left(\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\right)\right] = -E\left[\frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right].$$

17.4.3 THE LIKELIHOOD EQUATION

The log-likelihood function is

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}).$$

The first derivative vector, or **score vector,** is

$$\mathbf{g} = \frac{\partial \ln L(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{g}_i. \tag{17-9}$$

Since we are just adding terms, it follows from D1 and D2 that at $\boldsymbol{\theta}_0$,

$$E_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] = E_0[\mathbf{g}_0] = \mathbf{0}. \tag{17-10}$$

which is the **likelihood equation** mentioned earlier.

17.4.4 THE INFORMATION MATRIX EQUALITY

The Hessian of the log-likelihood is

$$\mathbf{H} = \frac{\partial^2 \ln L(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^{N} \mathbf{H}_i.$$

Evaluating once again at $\boldsymbol{\theta}_0$, by taking

$$E_0[\mathbf{g}_0 \mathbf{g}_0'] = E_0\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{g}_{0i}\mathbf{g}_{0j}'\right]$$

and, because of D1, dropping terms with unequal subscripts we obtain

$$E_0[\mathbf{g}_0 \mathbf{g}_0'] = E_0\left[\sum_{i=1}^{n} \mathbf{g}_{0i}\mathbf{g}_{0i}'\right] = E_0\left[\sum_{i=1}^{n}(-\mathbf{H}_{0i})\right] = -E_0[\mathbf{H}_0],$$

so that

$$\begin{aligned}\mathrm{Var}_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] &= E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0'}\right)\right] \\ &= -E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right].\end{aligned} \tag{17-11}$$

This very useful result is known as the **information matrix equality.**

17.4.5 ASYMPTOTIC PROPERTIES OF THE MAXIMUM LIKELIHOOD ESTIMATOR

We can now sketch a derivation of the asymptotic properties of the MLE. Formal proofs of these results require some fairly intricate mathematics. Two widely cited derivations are those of Cramér (1948) and Amemiya (1985). To suggest the flavor of the exercise,

we will sketch an analysis provided by Stuart and Ord (1989) for a simple case, and indicate where it will be necessary to extend the derivation if it were to be fully general.

17.4.5.a CONSISTENCY

We assume that $f(\mathbf{y}_i \mid \boldsymbol{\theta}_0)$ is a possibly multivariate density which at this point does not depend on covariates, $\mathbf{x}_i$. Thus, this is the iid, random sampling case. Since $\hat{\boldsymbol{\theta}}$ is the MLE, in any finite sample, for any $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$ (including the true $\boldsymbol{\theta}_0$) it must be true that

$$\ln L(\hat{\boldsymbol{\theta}}) \geq \ln L(\boldsymbol{\theta}). \tag{17-12}$$

Consider, then, the random variable $L(\boldsymbol{\theta})/L(\boldsymbol{\theta}_0)$. Since the log function is strictly concave, from Jensen's Inequality (Theorem D.8.), we have

$$E_0\left[\log \frac{L(\theta)}{L(\theta_0)}\right] < \log E_0\left[\frac{L(\theta)}{L(\theta_0)}\right]. \tag{17-13}$$

The expectation on the right hand side is exactly equal to one, as

$$E_0\left[\frac{L(\theta)}{L(\theta_0)}\right] = \int \left(\frac{L(\theta)}{L(\theta_0)}\right) L(\theta_0)\, d\mathbf{y} = 1 \tag{17-14}$$

is simply the integral of a joint density. Now, take logs on both sides of (17-13), insert the result of (17-14), then divide by n to produce

$$E_0[1/n \ln L(\theta)] - E_0[1/n \ln L(\theta_0)] < 0. \tag{17-15}$$

This produces a central result:

THEOREM 17.3 Likelihood Inequality

$E_0[(1/n) \ln L(\theta_0)] > E_0[(1/n) \ln L(\theta)]$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ (including $\hat{\boldsymbol{\theta}}$).

This result is (17-15).

In words, *the expected value of the log-likelihood is maximized at the true value of the parameters*.

For any $\boldsymbol{\theta}$, including $\hat{\boldsymbol{\theta}}$,

$$[(1/n) \ln L(\boldsymbol{\theta})] = (1/n) \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta})$$

is the sample mean of n iid random variables, with expectation $E_0[(1/n) \ln L(\boldsymbol{\theta})]$. Since the sampling is iid by the regularity conditions, we can invoke the Khinchine Theorem, D.5; the sample mean converges in probability to the population mean. Using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, it follows from Theorem 17.3 that as $n \to \infty$, $\lim \text{Prob}\{[(1/n) \ln L(\hat{\boldsymbol{\theta}})] < [(1/n) \ln L(\boldsymbol{\theta}_0)]\} = 1$ if $\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$. But, $\hat{\boldsymbol{\theta}}$ is the MLE, so for every n, $(1/n) \ln L(\hat{\boldsymbol{\theta}}) \geq (1/n) \ln L(\boldsymbol{\theta}_0)$. The only way these can both be true is if $(1/n)$ times the sample log-likelihood evaluated at the MLE converges to the population expectation of $(1/n)$ times the log-likelihood evaluated at the true parameters. There remains one final step.

Does $(1/n)\ln L(\hat{\boldsymbol{\theta}}) \to (1/n)\ln L(\boldsymbol{\theta}_0)$ imply that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$? If there is a single parameter and the likelihood function is one to one, then clearly so. For more general cases, this requires a further characterization of the likelihood function. If the likelihood is strictly continuous and twice differentiable, which we assumed in the regularity conditions, and if the parameters of the model are identified which we assumed at the beginning of this discussion, then yes, it does, so we have the result.

This is a heuristic proof. As noted, formal presentations appear in more advanced treatises than this one. We should also note, we have assumed at several points that sample means converged to the population expectations. This is likely to be true for the sorts of applications usually encountered in econometrics, but a fully general set of results would look more closely at this condition. Second, we have assumed iid sampling in the preceding—that is, the density for $\mathbf{y}_i$ does not depend on any other variables, $\mathbf{x}_i$. This will almost never be true in practice. Assumptions about the behavior of these variables will enter the proofs as well. For example, in assessing the large sample behavior of the least squares estimator, we have invoked an assumption that the data are "well behaved." The same sort of consideration will apply here as well. We will return to this issue shortly. With all this in place, we have property M1, $\operatorname{plim}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

17.4.5.b ASYMPTOTIC NORMALITY

At the maximum likelihood estimator, the gradient of the log-likelihood equals zero (by definition), so

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

(This is the sample statistic, not the expectation.) Expand this set of equations in a second-order Taylor series around the true parameters $\boldsymbol{\theta}_0$. We will use the mean value theorem to truncate the Taylor series at the second term.

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{g}(\boldsymbol{\theta}_0) + \mathbf{H}(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}.$$

The Hessian is evaluated at a point $\bar{\boldsymbol{\theta}}$ that is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ ($\bar{\boldsymbol{\theta}} = w\hat{\boldsymbol{\theta}} + (1-w)\boldsymbol{\theta}_0$ for some $0 < w < 1$). We then rearrange this function and multiply the result by $\sqrt{n}$ to obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = [-\mathbf{H}(\bar{\boldsymbol{\theta}})]^{-1}[\sqrt{n}\mathbf{g}(\boldsymbol{\theta}_0)].$$

Because $\operatorname{plim}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}$, $\operatorname{plim}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = 0$ as well. The second derivatives are continuous functions. Therefore, if the limiting distribution exists, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} [-\mathbf{H}(\boldsymbol{\theta}_0)]^{-1}[\sqrt{n}\mathbf{g}(\boldsymbol{\theta}_0)].$$

By dividing $\mathbf{H}(\boldsymbol{\theta}_0)$ and $\mathbf{g}(\boldsymbol{\theta}_0)$ by n, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \left[-\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1}[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)].$$

We may apply the Lindberg–Levy central limit theorem (D.18) to $[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)]$, since it is $\sqrt{n}$ times the mean of a random sample; we have invoked D1 again. The limiting variance of $[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)]$ is $-E_0[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)]$, so

$$\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0) \xrightarrow{d} N\left\{\mathbf{0}, -E_0\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}.$$

By virtue of Theorem D.2, $\text{plim}[-(1/n)\mathbf{H}(\boldsymbol{\theta}_0)] = -E_0[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)]$. Since this result is a constant matrix, we can combine results to obtain

$$\left[-\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1}\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\left\{-E_0\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}\left\{-E_0\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\right],$$

or

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\right],$$

which gives the asymptotic distribution of the MLE:

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N[\boldsymbol{\theta}_0, \{\mathbf{I}(\boldsymbol{\theta}_0)\}^{-1}].$$

This last step completes M2.

Example 17.3 Information Matrix for the Normal Distribution

For the likelihood function in Example 17.2, the second derivatives are

$$\frac{\partial^2 \ln L}{\partial \mu^2} = \frac{-n}{\sigma^2},$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2,$$

$$\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = \frac{-1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu).$$

For the **asymptotic variance** of the maximum likelihood estimator, we need the expectations of these derivatives. The first is nonstochastic, and the third has expectation 0, as $E[x_i] = \mu$. That leaves the second, which you can verify has expectation $-n/(2\sigma^4)$ because each of the n terms $(x_i - \mu)^2$ has expected value σ^2. Collecting these in the information matrix, reversing the sign, and inverting the matrix gives the asymptotic covariance matrix for the maximum likelihood estimators:

$$\left\{-E_0\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}_0 \, \partial \boldsymbol{\theta}_0'}\right]\right\}^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}.$$

17.4.5.c ASYMPTOTIC EFFICIENCY

Theorem C.2 provides the lower bound for the variance of an unbiased estimator. Since the asymptotic variance of the MLE achieves this bound, it seems natural to extend the result directly. There is, however, a loose end in that the MLE is almost never unbiased. As such, we need an asymptotic version of the bound, which was provided by Cramér (1948) and Rao (1945) (hence the name):

THEOREM 17.4 Cramér–Rao Lower Bound

Assuming that the density of y_i satisfies the regularity conditions R1–R3, the asymptotic variance of a consistent and asymptotically normally distributed estimator of the parameter vector $\boldsymbol{\theta}_0$ will always be at least as large as

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left(-E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \, \partial \boldsymbol{\theta}_0'}\right]\right)^{-1} = \left(E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)'\right]\right)^{-1}.$$

The asymptotic variance of the MLE is, in fact, equal to the Cramér–Rao Lower Bound for the variance of a consistent estimator, so this completes the argument.[3]

17.4.5.d INVARIANCE

Lastly, the invariance property, M4, is a mathematical result of the method of computing MLEs; it is not a statistical result as such. More formally, the MLE is invariant to *one-to-one* transformations of $\boldsymbol{\theta}$. Any transformation that is not one to one either renders the model inestimable if it is one to many or imposes restrictions if it is many to one. Some theoretical aspects of this feature are discussed in Davidson and MacKinnon (1993, pp. 253–255). For the practitioner, the result can be extremely useful. For example, when a parameter appears in a likelihood function in the form $1/\theta_j$, it is usually worthwhile to reparameterize the model in terms of $\gamma_j = 1/\theta_j$. In an important application, Olsen (1978) used this result to great advantage. (See Section 22.2.3.) Suppose that the normal log-likelihood in Example 17.2 is parameterized in terms of the **precision parameter,** $\theta^2 = 1/\sigma^2$. The log-likelihood becomes

$$\ln L(\mu, \theta^2) = -(n/2)\ln(2\pi) + (n/2)\ln\theta^2 - \frac{\theta^2}{2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

The MLE for μ is clearly still $\bar{x}$. But the likelihood equation for θ^2 is now

$$\partial \ln L(\mu, \theta^2)/\partial\theta^2 = \frac{1}{2}\left[n/\theta^2 - \sum_{i=1}^{n}(y_i - \mu)^2\right] = 0,$$

which has solution $\hat{\theta}^2 = n/\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 1/\hat{\sigma}^2$, as expected. There is a second implication. If it is desired to analyze a function of an MLE, then the function of $\hat{\boldsymbol{\theta}}$ will, itself, be the MLE.

17.4.5.e CONCLUSION

These four properties explain the prevalence of the maximum likelihood technique in econometrics. The second greatly facilitates hypothesis testing and the construction of interval estimates. The third is a particularly powerful result. The MLE has the minimum variance achievable by a consistent and asymptotically normally distributed estimator.

17.4.6 ESTIMATING THE ASYMPTOTIC VARIANCE OF THE MAXIMUM LIKELIHOOD ESTIMATOR

The asymptotic covariance matrix of the maximum likelihood estimator is a matrix of parameters that must be estimated (that is, it is a function of the $\boldsymbol{\theta}_0$ that is being estimated). If the form of the expected values of the second derivatives of the log-likelihood is known, then

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left\{-E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0\,\partial\boldsymbol{\theta}_0'}\right]\right\}^{-1} \qquad \textbf{(17-16)}$$

[3] A result reported by LeCam (1953) and recounted in Amemiya (1985, p. 124) suggests that in principle, there do exist CAN functions of the data with smaller variances than the MLE. But the finding is a narrow result with no practical implications. For practical purposes, the statement may be taken as given.

can be evaluated at $\hat{\boldsymbol{\theta}}$ to estimate the covariance matrix for the MLE. This estimator will rarely be available. The second derivatives of the log-likelihood will almost always be complicated nonlinear functions of the data whose exact expected values will be unknown. There are, however, two alternatives. A second estimator is

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = \left(-\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}\, \partial \hat{\boldsymbol{\theta}}'}\right)^{-1}. \tag{17-17}$$

This estimator is computed simply by evaluating the actual (not expected) second derivatives matrix of the log-likelihood function at the maximum likelihood estimates. It is straightforward to show that this amounts to estimating the expected second derivatives of the density with the sample mean of this quantity. Theorem D.4 and Result (D-5) can be used to justify the computation. The only shortcoming of this estimator is that the second derivatives can be complicated to derive and program for a computer. A third estimator based on result D3 in Theorem 17.2, that the expected second derivatives matrix is the covariance matrix of the first derivatives vector is

$$[\hat{\hat{\mathbf{I}}}(\hat{\boldsymbol{\theta}})]^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'\right]^{-1} = [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1}, \tag{17-18}$$

where

$$\hat{\mathbf{g}}_i = \frac{\partial \ln f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}$$

and

$$\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_n]'.$$

$\hat{\mathbf{G}}$ is an $n \times K$ matrix with ith row equal to the transpose of the ith vector of derivatives in the terms of the log-likelihood function. For a single parameter, this estimator is just the reciprocal of the sum of squares of the first derivatives. This estimator is extremely convenient, in most cases, because it does not require any computations beyond those required to solve the likelihood equation. It has the added virtue that it is always nonnegative definite. For some extremely complicated log-likelihood functions, sometimes because of rounding error, the *observed* Hessian can be indefinite, even at the maximum of the function. The estimator in (17-18) is known as the **BHHH** estimator[4] and the **outer product of gradients,** or **OPG,** estimator.

None of the three estimators given here is preferable to the others on statistical grounds; all are asymptotically equivalent. In most cases, the BHHH estimator will be the easiest to compute. One caution is in order. As the example below illustrates, these estimators can give different results in a finite sample. This is an unavoidable finite sample problem that can, in some cases, lead to different statistical conclusions. The example is a case in point. Using the usual procedures, we would reject the hypothesis that $\beta = 0$ if either of the first two variance estimators were used, but not if the third were used. The estimator in (17-16) is usually unavailable, as the exact expectation of the Hessian is rarely known. Available evidence suggests that in small or moderate sized samples, (17-17) (the Hessian) is preferable.

[4]It appears to have been advocated first in the econometrics literature in Berndt et al. (1974).

Example 17.4 Variance Estimators for an MLE

The sample data in Example C.1 are generated by a model of the form

$$f(y_i, x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)},$$

where $y =$ income and $x =$ education. To find the maximum likelihood estimate of β, we maximize

$$\ln L(\beta) = -\sum_{i=1}^{n} \ln(\beta + x_i) - \sum_{i=1}^{n} \frac{y_i}{\beta + x_i}.$$

The likelihood equation is

$$\frac{\partial \ln L(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \frac{1}{\beta + x_i} + \sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^2} = 0, \tag{17-19}$$

which has the solution $\hat{\beta} = 15.602727$. To compute the asymptotic variance of the MLE, we require

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \frac{1}{(\beta + x_i)^2} - 2\sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^3}. \tag{17-20}$$

Since the function $E(y_i) = \beta + x_i$ is known, the exact form of the expected value in (17-20) is known. Inserting $\beta + x_i$ for y_i in (17-20) and taking the reciprocal yields the first variance estimate, 44.2546. Simply inserting $\hat{\beta} = 15.602727$ in (17-20) and taking the negative of the reciprocal gives the second estimate, 46.16337. Finally, by computing the reciprocal of the sum of squares of first derivatives of the densities evaluated at $\hat{\beta}$,

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})]^{-1} = \frac{1}{\sum_{i=1}^{n} [-1/(\hat{\beta} + x_i) + y_i/(\hat{\beta} + x_i)^2]^2},$$

we obtain the BHHH estimate, 100.5116.

17.4.7 CONDITIONAL LIKELIHOODS AND ECONOMETRIC MODELS

All of the preceding results form the statistical underpinnings of the technique of maximum likelihood estimation. But, for our purposes, a crucial element is missing. We have done the analysis in terms of the density of an observed random variable and a vector of parameters, $f(y_i \mid \boldsymbol{\alpha})$. But, econometric models will involve exogenous or predetermined variables, $\mathbf{x}_i$, so the results must be extended. A workable approach is to treat this modeling framework the same as the one in Chapter 5, where we considered the large sample properties of the linear regression model. Thus, we will allow $\mathbf{x}_i$ to denote a mix of random variables and constants that enter the conditional density of y_i. By partitioning the joint density of y_i and $\mathbf{x}_i$ into the product of the conditional and the marginal, the log-likelihood function may be written

$$\ln L(\boldsymbol{\alpha} \mid \mathbf{data}) = \sum_{i=1}^{n} \ln f(y_i, \mathbf{x}_i \mid \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\alpha}) + \sum_{i=1}^{n} \ln g(\mathbf{x}_i \mid \boldsymbol{\alpha}),$$

where any nonstochastic elements in $\mathbf{x}_i$ such as a time trend or dummy variable, are being carried as constants. In order to proceed, we will assume as we did before that the

process generating $\mathbf{x}_i$ takes place outside the model of interest. For present purposes, that means that the parameters that appear in $g(\mathbf{x}_i \mid \boldsymbol{\alpha})$ do not overlap with those that appear in $f(y_i \mid \mathbf{x}_i, \boldsymbol{\alpha})$. Thus, we partition $\boldsymbol{\alpha}$ into $[\boldsymbol{\theta}, \boldsymbol{\delta}]$ so that the log-likelihood function may be written

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \textbf{data}) = \sum_{i=1}^{n} \ln f(y_i, \mathbf{x}_i \mid \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \ln g(\mathbf{x}_i \mid \boldsymbol{\delta}).$$

As long as $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ have no elements in common and no restrictions connect them (such as $\theta + \delta = 1$), then the two parts of the log likelihood may be analyzed separately. In most cases, the marginal distribution of $\mathbf{x}_i$ will be of secondary (or no) interest.

Asymptotic results for the maximum conditional likelihood estimator must now account for the presence of $\mathbf{x}_i$ in the functions and derivatives of $\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$. We will proceed under the assumption of well behaved data so that sample averages such as

$$(1/n) \ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$$

and its gradient with respect to $\boldsymbol{\theta}$ will converge in probability to their population expectations. We will also need to invoke central limit theorems to establish the asymptotic normality of the gradient of the log likelihood, so as to be able to characterize the MLE itself. We will leave it to more advance treatises such as Amemiya (1985) and Newey and McFadden (1994) to establish specific conditions and fine points that must be assumed to claim the "usual" properties for maximum likelihood estimators. For present purposes (and the vast bulk of empirical applications), the following minimal assumptions should suffice:

- **Parameter space.** Parameter spaces that have gaps and nonconvexities in them will generally disable these procedures. An estimation problem that produces this failure is that of "estimating" a parameter that can take only one among a discrete set of values. For example, this set of procedures does not include "estimating" the timing of a structural change in a model. (See Section 7.4.) The likelihood function must be a continuous function of a convex parameter space. We allow unbounded parameter spaces, such as $\sigma > 0$ in the regression model, for example.
- **Identifiability.** Estimation must be feasible. This is the subject of definition 17.1 concerning identification and the surrounding discussion.
- **Well behaved data.** Laws of large numbers apply to sample means involving the data and some form of central limit theorem (generally Lyapounov) can be applied to the gradient. Ergodic stationarity is broad enough to encompass any situation that is likely to arise in practice, though it is probably more general than we need for most applications, since we will not encounter dependent observations specifically until later in the book. The definitions in Chapter 5 are assumed to hold generally.

With these in place, analysis is essentially the same in character as that we used in the linear regression model in Chapter 5 and follows precisely along the lines of Section 16.5.

17.5 THREE ASYMPTOTICALLY EQUIVALENT TEST PROCEDURES

The next several sections will discuss the most commonly used test procedures: the likelihood ratio, Wald, and Lagrange multiplier tests. [Extensive discussion of these procedures is given in Godfrey (1988).] We consider maximum likelihood estimation of a parameter θ and a test of the hypothesis H_0: $c(\theta) = 0$. The logic of the tests can be seen in Figure 17.2.[5] The figure plots the log-likelihood function $\ln L(\theta)$, its derivative with respect to θ, $d \ln L(\theta)/d\theta$, and the constraint $c(\theta)$. There are three approaches to testing the hypothesis suggested in the figure:

- **Likelihood ratio test.** If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference, $\ln L_U - \ln L_R$, where L_U is the value of the likelihood function at the unconstrained value of θ and L_R is the value of the likelihood function at the restricted estimate.
- **Wald test.** If the restriction is valid, then $c(\hat{\theta}_{\text{MLE}})$ should be close to zero since the MLE is consistent. Therefore, the test is based on $c(\hat{\theta}_{\text{MLE}})$. We reject the hypothesis if this value is significantly different from zero.
- **Lagrange multiplier test.** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave rather differently in a small sample. Unfortunately, their small-sample properties are unknown, except in a few special cases. As a consequence, the choice among them is typically made on the basis of ease of computation. The likelihood ratio test requires calculation of both restricted and unrestricted estimators. If both are simple to compute, then this way to proceed is convenient. The Wald test requires only the unrestricted estimator, and the Lagrange multiplier test requires only the restricted estimator. In some problems, one of these estimators may be much easier to compute than the other. For example, a linear model is simple to estimate but becomes nonlinear and cumbersome if a nonlinear constraint is imposed. In this case, the Wald statistic might be preferable. Alternatively, restrictions sometimes amount to the removal of nonlinearities, which would make the Lagrange multiplier test the simpler procedure.

17.5.1 THE LIKELIHOOD RATIO TEST

Let $\boldsymbol{\theta}$ be a vector of parameters to be estimated, and let H_0 specify some sort of restriction on these parameters. Let $\hat{\boldsymbol{\theta}}_U$ be the maximum likelihood estimator of $\boldsymbol{\theta}$ obtained without regard to the constraints, and let $\hat{\boldsymbol{\theta}}_R$ be the constrained maximum likelihood estimator. If $\hat{L}_U$ and $\hat{L}_R$ are the likelihood functions evaluated at these two estimates, then the

[5] See Buse (1982). Note that the scale of the vertical axis would be different for each curve. As such, the points of intersection have no significance.

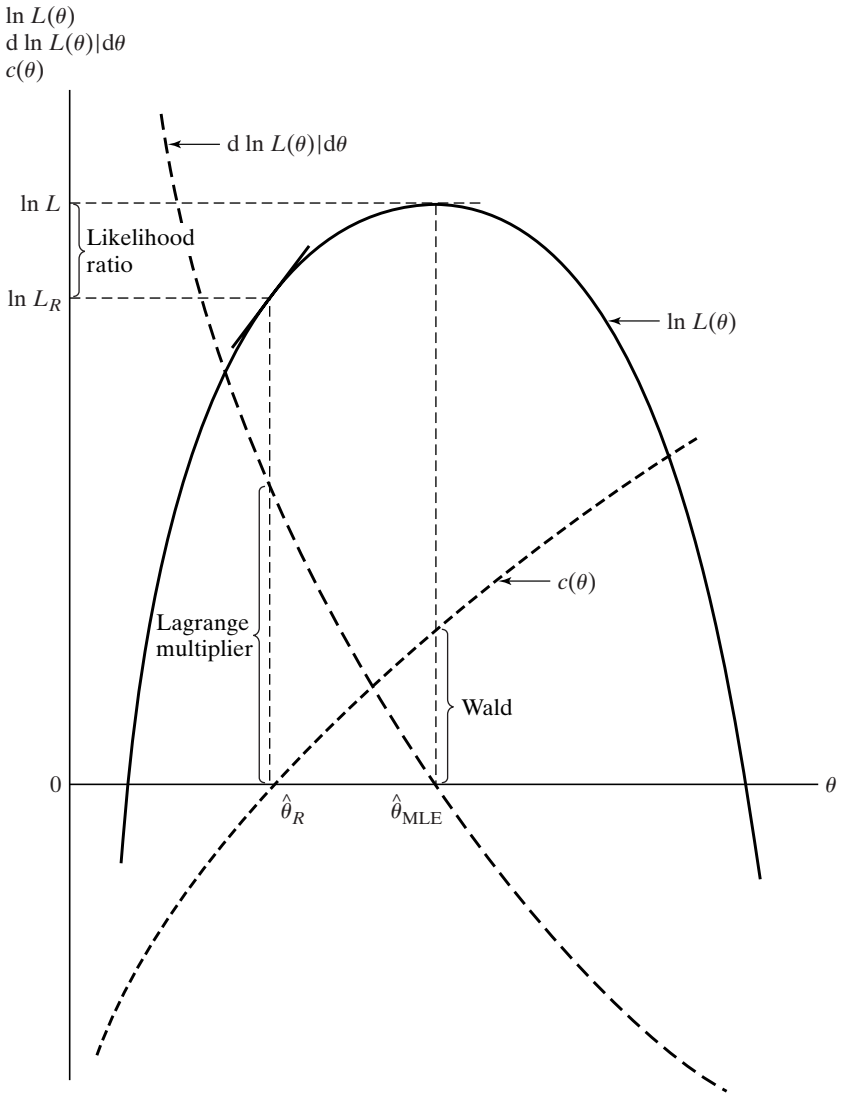


FIGURE 17.2 Three Bases for Hypothesis Tests.

likelihood ratio is

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}. \tag{17-21}$$

This function must be between zero and one. Both likelihoods are positive, and $\hat{L}_R$ cannot be larger than $\hat{L}_U$. (A restricted optimum is never superior to an unrestricted one.) If λ is too small, then doubt is cast on the restrictions.

An example from a discrete distribution helps to fix these ideas. In estimating from a sample of 10 from a Poisson distribution at the beginning of Section 17.3, we found the

MLE of the parameter θ to be 2. At this value, the likelihood, which is the probability of observing the sample we did, is 0.104×10^{-8}. Are these data consistent with $H_0: \theta = 1.8$? $L_R = 0.936 \times 10^{-9}$, which is, as expected, smaller. This particular sample is somewhat less probable under the hypothesis.

The formal test procedure is based on the following result.

THEOREM 17.5 Limiting Distribution of the Likelihood Ratio Test Statistic

Under regularity and under H_0, the large sample distribution of $-2 \ln \lambda$ is chi-squared, with degrees of freedom equal to the number of restrictions imposed.

The null hypothesis is rejected if this value exceeds the appropriate critical value from the chi-squared tables. Thus, for the Poisson example,

$$-2 \ln \lambda = -2 \ln \left(\frac{0.0936}{0.104} \right) = 0.21072.$$

This chi-squared statistic with one degree of freedom is not significant at any conventional level, so we would not reject the hypothesis that $\theta = 1.8$ on the basis of this test.[6]

It is tempting to use the likelihood ratio test to test a simple null hypothesis against a simple alternative. For example, we might be interested in the Poisson setting in testing $H_0: \theta = 1.8$ against $H_1: \theta = 2.2$. But the test cannot be used in this fashion. The degrees of freedom of the chi-squared statistic for the likelihood ratio test equals the reduction in the number of dimensions in the parameter space that results from imposing the restrictions. In testing a simple null hypothesis against a simple alternative, this value is zero.[7] Second, one sometimes encounters an attempt to test one distributional assumption against another with a likelihood ratio test; for example, a certain model will be estimated assuming a normal distribution and then assuming a t distribution. The ratio of the two likelihoods is then compared to determine which distribution is preferred. This comparison is also inappropriate. The parameter spaces, and hence the likelihood functions of the two cases, are unrelated.

17.5.2 THE WALD TEST

A practical shortcoming of the likelihood ratio test is that it usually requires estimation of both the restricted and unrestricted parameter vectors. In complex models, one or the other of these estimates may be very difficult to compute. Fortunately, there are two alternative testing procedures, the Wald test and the Lagrange multiplier test, that circumvent this problem. Both tests are based on an estimator that is asymptotically normally distributed.

[6] Of course, our use of the large-sample result in a sample of 10 might be questionable.

[7] Note that because both likelihoods are restricted in this instance, there is nothing to prevent $-2 \ln \lambda$ from being negative.

These two tests are based on the distribution of the full rank quadratic form considered in Section B.11.6. Specifically,

$$\text{If } \mathbf{x} \sim N_J[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \text{chi-squared}[J]. \tag{17-22}$$

In the setting of a hypothesis test, under the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$, the quadratic form has the chi-squared distribution. If the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$ is false, however, then the quadratic form just given will, on average, be larger than it would be if the hypothesis were true.[8] This condition forms the basis for the test statistics discussed in this and the next section.

Let $\hat{\boldsymbol{\theta}}$ be the vector of parameter estimates obtained without restrictions. We hypothesize a set of restrictions

$$H_0: \mathbf{c}(\boldsymbol{\theta}) = \mathbf{q}.$$

If the restrictions are valid, then at least approximately $\hat{\boldsymbol{\theta}}$ should satisfy them. If the hypothesis is erroneous, however, then $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}$ should be farther from $\mathbf{0}$ than would be explained by sampling variability alone. The device we use to formalize this idea is the Wald test.

THEOREM 17.6 Limiting Distribution of the Wald Test Statistic
The Wald statistic is

$$W = [\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}]'\big(\text{Asy.Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}]\big)^{-1}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}].$$

Under H_0, in large samples, W has a chi-squared distribution with degrees of freedom equal to the number of restrictions [i.e., the number of equations in $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q} = \mathbf{0}$]. A derivation of the limiting distribution of the Wald statistic appears in Theorem 6.15.

This test is analogous to the chi-squared statistic in (17-22) if $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}$ is normally distributed with the hypothesized mean of $\mathbf{0}$. A large value of W leads to rejection of the hypothesis. Note, finally, that W only requires computation of the unrestricted model. One must still compute the covariance matrix appearing in the preceding quadratic form. This result is the variance of a possibly nonlinear function, which we treated earlier.

$$\begin{aligned}\text{Est. Asy. Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] &= \hat{\mathbf{C}}\ \text{Est. Asy. Var}[\hat{\boldsymbol{\theta}}]\hat{\mathbf{C}}', \\ \hat{\mathbf{C}} &= \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}\right].\end{aligned} \tag{17-23}$$

That is, $\mathbf{C}$ is the $J \times K$ matrix whose jth row is the derivatives of the jth constraint with respect to the K elements of $\boldsymbol{\theta}$. A common application occurs in testing a set of linear restrictions.

[8] If the mean is not $\boldsymbol{\mu}$, then the statistic in (17-22) will have a **noncentral chi-squared distribution.** This distribution has the same basic shape as the central chi-squared distribution, with the same degrees of freedom, but lies to the right of it. Thus, a random draw from the noncentral distribution will tend, on average, to be larger than a random observation from the central distribution.

For testing a set of linear restrictions $\mathbf{R}\boldsymbol{\theta} = \mathbf{q}$, the Wald test would be based on

$$H_0: \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{R}\boldsymbol{\theta} - \mathbf{q} = \mathbf{0},$$
$$\hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}\right] = \mathbf{R}', \tag{17-24}$$
$$\text{Est. Asy. Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] = \mathbf{R}\ \text{Est. Asy. Var}[\hat{\boldsymbol{\theta}}]\mathbf{R},$$

and

$$W = [\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}]'[\mathbf{R}\ \text{Est. Asy. Var}(\hat{\boldsymbol{\theta}})\mathbf{R}']^{-1}[\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}].$$

The degrees of freedom is the number of rows in $\mathbf{R}$.

If $\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q}$ is a single restriction, then the Wald test will be the same as the test based on the confidence interval developed previously. If the test is

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_1: \theta \neq \theta_0,$$

then the earlier test is based on

$$z = \frac{|\hat{\theta} - \theta_0|}{s(\hat{\theta})}, \tag{17-25}$$

where $s(\hat{\theta})$ is the estimated asymptotic standard error. The test statistic is compared to the appropriate value from the standard normal table. The Wald test will be based on

$$\text{W} = [(\hat{\theta} - \theta_0) - 0]\big(\text{Asy. Var}[(\hat{\theta} - \theta_0) - 0]\big)^{-1}[(\hat{\theta} - \theta_0) - 0] = \frac{(\hat{\theta} - \theta_0)^2}{\text{Asy. Var}[\hat{\theta}]} = z^2. \tag{17-26}$$

Here W has a chi-squared distribution with one degree of freedom, which is the distribution of the square of the standard normal test statistic in (17-25).

To summarize, the Wald test is based on measuring the extent to which the unrestricted estimates fail to satisfy the hypothesized restrictions. There are two shortcomings of the Wald test. First, it is a pure significance test against the null hypothesis, not necessarily for a specific alternative hypothesis. As such, its power may be limited in some settings. In fact, the test statistic tends to be rather large in applications. The second shortcoming is not shared by either of the other test statistics discussed here. The Wald statistic is not invariant to the formulation of the restrictions. For example, for a test of the hypothesis that a function $\theta = \beta/(1 - \gamma)$ equals a specific value q there are two approaches one might choose. A Wald test based directly on $\theta - q = 0$ would use a statistic based on the variance of this nonlinear function. An alternative approach would be to analyze the linear restriction $\beta - q(1 - \gamma) = 0$, which is an equivalent, but linear, restriction. The Wald statistics for these two tests could be different and might lead to different inferences. These two shortcomings have been widely viewed as compelling arguments against use of the Wald test. But, in its favor, the Wald test does not rely on a strong distributional assumption, as do the likelihood ratio and Lagrange multiplier tests. The recent econometrics literature is replete with applications that are based on distribution free estimation procedures, such as the GMM method. As such, in recent years, the Wald test has enjoyed a redemption of sorts.

17.5.3 THE LAGRANGE MULTIPLIER TEST

The third test procedure is the **Lagrange multiplier (LM)** or **efficient score** (or just **score**) test. It is based on the restricted model instead of the unrestricted model. Suppose that we maximize the log-likelihood subject to the set of constraints $\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}$. Let $\boldsymbol{\lambda}$ be a vector of Lagrange multipliers and define the Lagrangean function

$$\ln L^*(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) + \boldsymbol{\lambda}'(\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q}).$$

The solution to the constrained maximization problem is the root of

$$\begin{aligned} \frac{\partial \ln L^*}{\partial \boldsymbol{\theta}} &= \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0}, \\ \frac{\partial \ln L^*}{\partial \boldsymbol{\lambda}} &= \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}, \end{aligned} \tag{17-27}$$

where $\mathbf{C}'$ is the transpose of the derivatives matrix in the second line of (17-23). If the restrictions are valid, then imposing them will not lead to a significant difference in the maximized value of the likelihood function. In the first-order conditions, the meaning is that the second term in the derivative vector will be small. In particular, $\boldsymbol{\lambda}$ will be small. We could test this directly, that is, test H_0: $\boldsymbol{\lambda} = \mathbf{0}$, which leads to the Lagrange multiplier test. There is an equivalent simpler formulation, however. At the restricted maximum, the derivatives of the log-likelihood function are

$$\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R} = -\hat{\mathbf{C}}'\hat{\boldsymbol{\lambda}} = \hat{\mathbf{g}}_R. \tag{17-28}$$

If the restrictions are valid, at least within the range of sampling variability, then $\hat{\mathbf{g}}_R = \mathbf{0}$. That is, the derivatives of the log-likelihood evaluated at the restricted parameter vector will be approximately zero. The vector of first derivatives of the log-likelihood is the vector of **efficient scores.** Since the test is based on this vector, it is called the **score test** as well as the Lagrange multiplier test. The variance of the first derivative vector is the information matrix, which we have used to compute the asymptotic covariance matrix of the MLE. The test statistic is based on reasoning analogous to that underlying the Wald test statistic.

THEOREM 17.7 Limiting Distribution of the Lagrange Multiplier Statistic

The Lagrange multiplier test statistic is

$$\text{LM} = \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R}\right)' [\mathbf{I}(\hat{\boldsymbol{\theta}}_R)]^{-1} \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R}\right).$$

Under the null hypothesis, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. All terms are computed at the restricted estimator.

The LM statistic has a useful form. Let $\hat{\mathbf{g}}_{iR}$ denote the ith term in the gradient of the log-likelihood function. Then,

$$\hat{\mathbf{g}}_R = \sum_{i=1}^{n} \hat{\mathbf{g}}_{iR} = \hat{\mathbf{G}}_R'\mathbf{i},$$

where $\hat{\mathbf{G}}_R$ is the $n \times K$ matrix with ith row equal to $\mathbf{g}_{iR}'$ and $\mathbf{i}$ is a column of 1s. If we use the BHHH (outer product of gradients) estimator in (17-18) to estimate the Hessian, then

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = [\hat{\mathbf{G}}_R'\hat{\mathbf{G}}_R]^{-1}$$

and

$$\text{LM} = \mathbf{i}'\hat{\mathbf{G}}_R[\hat{\mathbf{G}}_R'\hat{\mathbf{G}}_R]^{-1}\hat{\mathbf{G}}_R'\mathbf{i}.$$

Now, since $\mathbf{i}'\mathbf{i}$ equals n, $\text{LM} = n(\mathbf{i}'\hat{\mathbf{G}}_R[\hat{\mathbf{G}}_R'\hat{\mathbf{G}}_R]^{-1}\hat{\mathbf{G}}_R'\mathbf{i}/n) = nR_{\mathbf{i}}^2$, which is n times the uncentered squared multiple correlation coefficient in a linear regression of a column of 1s on the derivatives of the log-likelihood function computed at the restricted estimator. We will encounter this result in various forms at several points in the book.

17.5.4 AN APPLICATION OF THE LIKELIHOOD BASED TEST PROCEDURES

Consider, again, the data in Example C.1. In Example 17.4, the parameter β in the model

$$f(y_i|x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)} \qquad \textbf{(17-29)}$$

was estimated by maximum likelihood. For convenience, let $\beta_i = 1/(\beta + x_i)$. This exponential density is a restricted form of a more general gamma distribution,

$$f(y_i|x_i, \beta, \rho) = \frac{\beta_i^{\rho}}{\Gamma(\rho)} y_i^{\rho-1} e^{-y_i\beta_i}. \qquad \textbf{(17-30)}$$

The restriction is $\rho = 1$.[9] We consider testing the hypothesis

$$H_0: \rho = 1 \quad \text{versus} \quad H_1: \rho \neq 1$$

using the various procedures described previously. The log-likelihood and its derivatives are

$$\ln L(\beta, \rho) = \rho \sum_{i=1}^{n} \ln \beta_i - n \ln \Gamma(\rho) + (\rho - 1) \sum_{i=1}^{n} \ln y_i - \sum_{i=1}^{n} y_i\beta_i,$$

$$\frac{\partial \ln L}{\partial \beta} = -\rho \sum_{i=1}^{n} \beta_i + \sum_{i=1}^{n} y_i\beta_i^2, \qquad \frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^{n} \ln \beta_i - n\Psi(\rho) + \sum_{i=1}^{n} \ln y_i, \qquad \textbf{(17-31)}$$

$$\frac{\partial^2 \ln L}{\partial \beta^2} = \rho \sum_{i=1}^{n} \beta_i^2 - 2 \sum_{i=1}^{n} y_i\beta_i^3, \qquad \frac{\partial^2 \ln L}{\partial \rho^2} = -n\Psi'(\rho), \qquad \frac{\partial^2 \ln L}{\partial \beta \partial \rho} = -\sum_{i=1}^{n} \beta_i.$$

[9] The gamma function $\Gamma(\rho)$ and the gamma distribution are described in Sections B.4.5 and E.5.3.

TABLE 17.1 Maximum Likelihood Estimates

Quantity	*Unrestricted Estimate*[a]	*Restricted Estimate*
β	−4.7198 (2.344)	15.6052 (6.794)
ρ	3.1517 (0.7943)	1.0000 (0.000)
$\ln L$	−82.91444	−88.43771
$\partial \ln L/\partial\beta$	0.0000	0.0000
$\partial \ln L/\partial\rho$	0.0000	7.9162
$\partial^2 \ln L/\partial\beta^2$	−0.85628	−0.021659
$\partial^2 \ln L/\partial\rho^2$	−7.4569	−32.8987
$\partial^2 \ln L/\partial\beta\partial\rho$	−2.2423	−0.66885

[a]Estimated asymptotic standard errors based on $\mathbf{V}$ are given in parentheses.

[Recall that $\Psi(\rho) = d \ln \Gamma(\rho)/d\rho$ and $\Psi'(\rho) = d^2 \ln \Gamma(\rho)/d\rho^2$.] Unrestricted maximum likelihood estimates of β and ρ are obtained by equating the two first derivatives to zero. The restricted maximum likelihood estimate of β is obtained by equating $\partial \ln L/\partial\beta$ to zero while fixing ρ at one. The results are shown in Table 17.1. Three estimators are available for the asymptotic covariance matrix of the estimators of $\boldsymbol{\theta} = (\beta, \rho)'$. Using the actual Hessian as in (17-17), we compute $\mathbf{V} = [-\Sigma_i \partial^2 \ln L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']^{-1}$ at the maximum likelihood estimates. For this model, it is easy to show that $E[y_i \mid x_i] = \rho(\beta + x_i)$ (either by direct integration or, more simply, by using the result that $E[\partial \ln L/\partial\beta] = 0$ to deduce it). Therefore, we can also use the expected Hessian as in (17-16) to compute $\mathbf{V}_E = \{-\Sigma_i E[\partial^2 \ln L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']\}^{-1}$. Finally, by using the sums of squares and cross products of the first derivatives, we obtain the BHHH estimator in (17-18), $\mathbf{V}_B = [\Sigma_i(\partial \ln L/\partial\boldsymbol{\theta})(\partial \ln L/\partial\boldsymbol{\theta}')]^{-1}$. Results in Table 17.1 are based on $\mathbf{V}$.

The three estimators of the asymptotic covariance matrix produce notably different results:

$$\mathbf{V} = \begin{bmatrix} 5.495 & -1.652 \\ -1.652 & 0.6309 \end{bmatrix}, \quad \mathbf{V}_E = \begin{bmatrix} 4.897 & -1.473 \\ -1.473 & 0.5770 \end{bmatrix}, \quad \mathbf{V}_B = \begin{bmatrix} 13.35 & -4.314 \\ -4.314 & 1.535 \end{bmatrix}.$$

Given the small sample size, the differences are to be expected. Nonetheless, the striking difference of the BHHH estimator is typical of its erratic performance in small samples.

- **Confidence Interval Test:** A 95 percent confidence interval for ρ based on the unrestricted estimates is $3.1517 \pm 1.96\sqrt{0.6309} = [1.5942, 4.7085]$. This interval does not contain $\rho = 1$, so the hypothesis is rejected.
- **Likelihood Ratio Test:** The LR statistic is $\lambda = -2[-88.43771 - (-82.91444)] = 11.0465$. The table value for the test, with one degree of freedom, is 3.842. Since the computed value is larger than this critical value, the hypothesis is again rejected.
- **Wald Test:** The Wald test is based on the unrestricted estimates. For this restriction, $c(\boldsymbol{\theta}) - q = \rho - 1$, $dc(\hat{\rho})/d\hat{\rho} = 1$, Est.Asy. Var$[c(\hat{\rho}) - q]$ = Est.Asy. Var$[\hat{\rho}] = 0.6309$, so $W = (3.1517 - 1)^2/[0.6309] = 7.3384$.

The critical value is the same as the previous one. Hence, H_0 is once again rejected. Note that the Wald statistic is the square of the corresponding test statistic that would be used in the confidence interval test, $|3.1517 - 1|/\sqrt{0.6309} = 2.70895$.

- **Lagrange Multiplier Test:** The Lagrange multiplier test is based on the restricted estimators. The estimated asymptotic covariance matrix of the derivatives used to compute the statistic can be any of the three estimators discussed earlier. The BHHH estimator, $\mathbf{V}_B$, is the empirical estimator of the variance of the gradient and is the one usually used in practice. This computation produces

$$\text{LM} = [0.0000 \quad 7.9162] \begin{bmatrix} 0.0099438 & 0.26762 \\ 0.26762 & 11.197 \end{bmatrix}^{-1} \begin{bmatrix} 0.0000 \\ 7.9162 \end{bmatrix} = 15.687.$$

  The conclusion is the same as before. Note that the same computation done using $\mathbf{V}$ rather than $\mathbf{V}_B$ produces a value of 5.1182. As before, we observe substantial small sample variation produced by the different estimators.

The latter three test statistics have substantially different values. It is possible to reach different conclusions, depending on which one is used. For example, if the test had been carried out at the 1 percent level of significance instead of 5 percent and LM had been computed using $\mathbf{V}$, then the critical value from the chi-squared statistic would have been 6.635 and the hypothesis would not have been rejected by the LM test. Asymptotically, all three tests are equivalent. But, in a finite sample such as this one, differences are to be expected.[10] Unfortunately, there is no clear rule for how to proceed in such a case, which highlights the problem of relying on a particular significance level and drawing a firm reject or accept conclusion based on sample evidence.

17.6 APPLICATIONS OF MAXIMUM LIKELIHOOD ESTIMATION

We now examine three applications of the maximum likelihood estimator. The first extends the results of Chapters 2 through 5 to the linear regression model with normally distributed disturbances. In the second application, we fit a nonlinear regression model by maximum likelihood. This application illustrates the effect of transformation of the dependent variable. The third application is a relatively straightforward use of the maximum likelihood technique in a nonlinear model that does not involve the normal distribution. This application illustrates the sorts of extensions of the MLE into settings that depart from the linear model of the preceding chapters and that are typical in econometric analysis.

17.6.1 THE NORMAL LINEAR REGRESSION MODEL

The linear regression model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i.$$

The likelihood function for a sample of n independent, identically and normally distributed disturbances is

$$L = (2\pi\sigma^2)^{-n/2} e^{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/(2\sigma^2)}. \qquad \textbf{(17-32)}$$

[10] For further discussion of this problem, see Berndt and Savin (1977).

The transformation from ε_i to y_i is $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, so the **Jacobian** for each observation, $|\partial\varepsilon_i/\partial y_i|$, is one.[11] Making the transformation, we find that the likelihood function for the n observations on the observed random variable is

$$L = (2\pi\sigma^2)^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \tag{17-33}$$

To maximize this function with respect to $\boldsymbol{\beta}$, it will be necessary to maximize the exponent or minimize the familiar sum of squares. Taking logs, we obtain the log-likelihood function for the classical regression model:

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}. \tag{17-34}$$

The necessary conditions for maximizing this log-likelihood are

$$\begin{bmatrix} \dfrac{\partial \ln L}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}. \tag{17-35}$$

The values that satisfy these equations are

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b} \quad \text{and} \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{\mathbf{e}'\mathbf{e}}{n}. \tag{17-36}$$

The slope estimator is the familiar one, whereas the variance estimator differs from the least squares value by the divisor of n instead of $n - K$.[12]

The Cramér–Rao bound for the variance of an unbiased estimator is the negative inverse of the expectation of

$$\begin{bmatrix} \dfrac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \dfrac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\sigma^2} \\ \dfrac{\partial^2 \ln L}{\partial\sigma^2\partial\boldsymbol{\beta}'} & \dfrac{\partial^2 \ln L}{\partial(\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & -\dfrac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sigma^4} \\ -\dfrac{\boldsymbol{\varepsilon}'\mathbf{X}}{\sigma^4} & \dfrac{n}{2\sigma^4} - \dfrac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^6} \end{bmatrix}. \tag{17-37}$$

In taking expected values, the off-diagonal term vanishes leaving

$$[\mathbf{I}(\boldsymbol{\beta}, \sigma^2)]^{-1} = \begin{bmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & 2\sigma^4/n \end{bmatrix}. \tag{17-38}$$

The least squares slope estimator is the maximum likelihood estimator for this model. Therefore, it inherits all the desirable *asymptotic* properties of maximum likelihood estimators.

We showed earlier that $s^2 = \mathbf{e}'\mathbf{e}/(n-K)$ is an unbiased estimator of σ^2. Therefore, the maximum likelihood estimator is biased toward zero:

$$E\left[\hat{\sigma}^2_{\mathrm{ML}}\right] = \frac{n-K}{n}\sigma^2 = \left(1 - \frac{K}{n}\right)\sigma^2 < \sigma^2. \tag{17-39}$$

[11] See (B-41) in Section B.5. The analysis to follow is conditioned on $\mathbf{X}$. To avoid cluttering the notation, we will leave this aspect of the model implicit in the results. As noted earlier, we assume that the data generating process for $\mathbf{X}$ does not involve $\boldsymbol{\beta}$ or σ^2 and that the data are well behaved as discussed in Chapter 5.

[12] As a general rule, maximum likelihood estimators do not make corrections for degrees of freedom.

Despite its small-sample bias, the maximum likelihood estimator of σ^2 has the same desirable asymptotic properties. We see in (17-39) that s^2 and $\hat{\sigma}^2$ differ only by a factor $-K/n$, which vanishes in large samples. It is instructive to formalize the asymptotic equivalence of the two. From (17-38), we know that

$$\sqrt{n}\left(\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2\right) \xrightarrow{d} N[0, 2\sigma^4].$$

It follows

$$z_n = \left(1 - \frac{K}{n}\right)\sqrt{n}\left(\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2\right) + \frac{K}{\sqrt{n}}\sigma^2 \xrightarrow{d} \left(1 - \frac{K}{n}\right)N[0, 2\sigma^4] + \frac{K}{\sqrt{n}}\sigma^2.$$

But $K/\sqrt{n}$ and K/n vanish as $n \to \infty$, so the limiting distribution of z_n is also $N[0, 2\sigma^4]$. Since $z_n = \sqrt{n}(s^2 - \sigma^2)$, we have shown that the asymptotic distribution of s^2 is the same as that of the maximum likelihood estimator.

The standard test statistic for assessing the validity of a set of linear restrictions in the linear model, $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$, is the F ratio,

$$F[J, n-K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)} = \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{J}.$$

With normally distributed disturbances, the F test is valid in any sample size. There remains a problem with nonlinear restrictions of the form $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{0}$, since the counterpart to F, which we will examine here, has validity only asymptotically even with normally distributed disturbances. In this section, we will reconsider the Wald statistic and examine two related statistics, the likelihood ratio statistic and the Lagrange multiplier statistic. These statistics are both based on the likelihood function and, like the Wald statistic, are generally valid only asymptotically.

No simplicity is gained by restricting ourselves to linear restrictions at this point, so we will consider general hypotheses of the form

$$H_0: \mathbf{c}(\boldsymbol{\beta}) = \mathbf{0},$$
$$H_1: \mathbf{c}(\boldsymbol{\beta}) \neq \mathbf{0}.$$

The **Wald statistic** for testing this hypothesis and its limiting distribution under H_0 would be

$$W = \mathbf{c}(\mathbf{b})'\{\mathbf{C}(\mathbf{b})[\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{C}(\mathbf{b})'\}^{-1}\mathbf{c}(\mathbf{b}) \xrightarrow{d} \chi^2[J], \tag{17-40}$$

where

$$\mathbf{C}(\mathbf{b}) = [\partial \mathbf{c}(\mathbf{b})/\partial \mathbf{b}']. \tag{17-41}$$

The **likelihood ratio (LR) test** is carried out by comparing the values of the log-likelihood function with and without the restrictions imposed. We leave aside for the present how the restricted estimator $\mathbf{b}_*$ is computed (except for the linear model, which we saw earlier). The test statistic and it's limiting distribution under H_0 are

$$\mathrm{LR} = -2[\ln L_* - \ln L] \xrightarrow{d} \chi^2[J]. \tag{17-42}$$

The log-likelihood for the regression model is given in (17-34). The first-order conditions imply that regardless of how the slopes are computed, the estimator of σ^2 without

restrictions on $\boldsymbol{\beta}$ will be $\hat{\sigma}^2 = (\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})/n$ and likewise for a restricted estimator $\hat{\sigma}_*^2 = (\mathbf{y}-\mathbf{Xb}_*)'(\mathbf{y}-\mathbf{Xb}_*)/n = \mathbf{e}_*'\mathbf{e}_*/n$. The **concentrated log-likelihood**[13] will be

$$\ln L_c = -\frac{n}{2}[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)]$$

and likewise for the restricted case. If we insert these in the definition of LR, then we obtain

$$\text{LR} = n\ln[\mathbf{e}_*'\mathbf{e}_*/\mathbf{e}'\mathbf{e}] = n\,(\ln\hat{\sigma}_*^2 - \ln\hat{\sigma}^2) = n\ln(\hat{\sigma}_*^2/\hat{\sigma}^2). \tag{17-43}$$

The **Lagrange multiplier (LM)** test is based on the gradient of the log-likelihood function. The principle of the test is that if the hypothesis is valid, then at the restricted estimator, the derivatives of the log-likelihood function should be close to zero. There are two ways to carry out the LM test. The log-likelihood function can be maximized subject to a set of restrictions by using

$$\ln L_{\text{LM}} = -\frac{n}{2}\left[\ln 2\pi + \ln\sigma^2 + \frac{[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]/n}{\sigma^2}\right] + \boldsymbol{\lambda}'\mathbf{c}(\boldsymbol{\beta}).$$

The first-order conditions for a solution are

$$\begin{bmatrix} \dfrac{\partial \ln L_{\text{LM}}}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ln L_{\text{LM}}}{\partial \sigma^2} \\ \dfrac{\partial \ln L_{\text{LM}}}{\partial \boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} + \mathbf{C}(\boldsymbol{\beta})'\boldsymbol{\lambda} \\ \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \\ \mathbf{c}(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \end{bmatrix}. \tag{17-44}$$

The solutions to these equations give the restricted least squares estimator, $\mathbf{b}_*$; the usual variance estimator, now $\mathbf{e}_*'\mathbf{e}_*/n$; and the Lagrange multipliers. There are now two ways to compute the test statistic. In the setting of the classical linear regression model, when we actually compute the Lagrange multipliers, a convenient way to proceed is to test the hypothesis that the multipliers equal zero. For this model, the solution for $\boldsymbol{\lambda}_*$ is $\boldsymbol{\lambda}_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})$. This equation is a linear function of the least squares estimator. If we carry out a *Wald* test of the hypothesis that $\boldsymbol{\lambda}_*$ equals $\mathbf{0}$, then the statistic will be

$$\text{LM} = \boldsymbol{\lambda}_*'\{\text{Est. Var}[\boldsymbol{\lambda}_*]\}^{-1}\boldsymbol{\lambda}_* = (\mathbf{Rb}-\mathbf{q})'[\mathbf{R}\,s_*^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q}). \tag{17-45}$$

The disturbance variance estimator, s_*^2, based on the restricted slopes is $\mathbf{e}_*'\mathbf{e}_*/n$.

An alternative way to compute the LM statistic often produces interesting results. In most situations, we maximize the log-likelihood function without actually computing the vector of Lagrange multipliers. (The restrictions are usually imposed some other way.) An alternative way to compute the statistic is based on the (general) result that under the hypothesis being tested,

$$E\,[\partial \ln L/\partial\boldsymbol{\beta}] = E\,[(1/\sigma^2)\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0}$$

and

$$\text{Asy. Var}[\partial \ln L/\partial\boldsymbol{\beta}] = -E\,[\partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}']^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.^{14} \tag{17-46}$$

[13] See Section E.6.3.

[14] This makes use of the fact that the Hessian is block diagonal.

We can test the hypothesis that at the restricted estimator, the derivatives are equal to zero. The statistic would be

$$\text{LM} = \frac{\mathbf{e}_*'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n} = nR_*^2. \tag{17-47}$$

In this form, the LM statistic is n times the coefficient of determination in a regression of the residuals $e_{i*} = (y_i - \mathbf{x}_i'\mathbf{b}_*)$ on the full set of regressors.

With some manipulation we can show that $W = [n/(n-K)]JF$ and LR and LM are approximately equal to this function of F.[15] All three statistics converge to JF as n increases. The linear model is a special case in that the LR statistic is based only on the unrestricted estimator and does not actually require computation of the restricted least squares estimator, although computation of F does involve most of the computation of $\mathbf{b}_*$. Since the log function is concave, and $W/n \geq \ln(1 + W/n)$, Godfrey (1988) also shows that $W \geq \text{LR} \geq \text{LM}$, so for the linear model, we have a firm ranking of the three statistics.

There is ample evidence that the asymptotic results for these statistics are problematic in small or moderately sized samples. [See, e.g., Davidson and MacKinnon (1993, pp. 456–457).] The true distributions of all three statistics involve the data and the unknown parameters and, as suggested by the algebra, converge to the F distribution *from above*. The implication is that critical values from the chi-squared distribution are likely to be too small; that is, using the limiting chi-squared distribution in small or moderately sized samples is likely to exaggerate the significance of empirical results. Thus, in applications, the more conservative F statistic (or t for one restriction) is likely to be preferable unless one's data are plentiful.

17.6.2 MAXIMUM LIKELIHOOD ESTIMATION OF NONLINEAR REGRESSION MODELS

In Chapter 9, we considered nonlinear regression models in which the nonlinearity in the parameters appeared entirely on the right-hand side of the equation. There are models in which parameters appear nonlinearly in functions of the dependent variable as well.

Suppose that, in general, the model is

$$g(y_i, \boldsymbol{\theta}) = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i.$$

One approach to estimation would be least squares, minimizing

$$S(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{n} [g(y_i, \boldsymbol{\theta}) - h(\mathbf{x}_i, \boldsymbol{\beta})]^2.$$

There is no reason to expect this **nonlinear least squares** estimator to be consistent, however, though it is difficult to show this analytically. The problem is that nonlinear least squares ignores the Jacobian of the transformation. Davidson and MacKinnon (1993, p. 244) suggest a qualitative argument, which we can illustrate with an example. Suppose y is positive, $g(y, \theta) = \exp(\theta y)$ and $h(\mathbf{x}, \boldsymbol{\beta}) = \beta x$. In this case, an obvious "solution" is

[15] See Godfrey (1988, pp. 49–51).

$\beta = 0$ and $\theta \to -\infty$, which produces a sum of squares of zero. "Estimation" becomes a nonissue. For this type of regression model, however, maximum likelihood estimation is consistent, efficient, and generally not appreciably more difficult than least squares.

For normally distributed disturbances, the density of y_i is

$$f(y_i) = \left|\frac{\partial \varepsilon_i}{\partial y_i}\right| (2\pi\sigma^2)^{-1/2} e^{-[g(y_i,\boldsymbol{\theta}) - h(\mathbf{x}_i,\boldsymbol{\beta})]^2/(2\sigma^2)}.$$

The Jacobian of the transformation [see (3-41)] is

$$J(y_i, \boldsymbol{\theta}) = \left|\frac{\partial \varepsilon_i}{\partial y_i}\right| = \left|\frac{\partial g(y_i, \boldsymbol{\theta})}{\partial y_i}\right| = J_i.$$

After collecting terms, the log-likelihood function will be

$$\ln L = \sum_{i=1}^{n} -\frac{1}{2}[\ln 2\pi + \ln \sigma^2] + \sum_{i=1}^{n} \ln J(y_i, \boldsymbol{\theta}) - \frac{\sum_{i=1}^{n}[g(y_i, \boldsymbol{\theta}) - h(\mathbf{x}_i, \boldsymbol{\beta})]^2}{2\sigma^2}. \quad \textbf{(17-48)}$$

In many cases, including the applications considered here, there is an inconsistency in the model in that the transformation of the dependent variable may rule out some values. Hence, the assumed normality of the disturbances cannot be strictly correct. In the generalized production function, there is a singularity at $y_i = 0$ where the Jacobian becomes infinite. Some research has been done on specific modifications of the model to accommodate the restriction [e.g., Poirier (1978) and Poirier and Melino (1978)], but in practice, the typical application involves data for which the constraint is inconsequential.

But for the Jacobians, nonlinear least squares would be maximum likelihood. If the Jacobian terms involve $\boldsymbol{\theta}$, however, then *least squares is not maximum likelihood*. As regards σ^2, this likelihood function is essentially the same as that for the simpler nonlinear regression model. The maximum likelihood estimator of σ^2 will be

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[g(y_i, \hat{\boldsymbol{\theta}}) - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})]^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \quad \textbf{(17-49)}$$

The likelihood equations for the unknown parameters are

$$\begin{bmatrix} \dfrac{\partial \ln L}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ln L}{\partial \boldsymbol{\theta}} \\ \dfrac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n} \dfrac{\varepsilon_i \partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ \displaystyle\sum_{i=1}^{n} \frac{1}{J_i}\left(\frac{\partial J_i}{\partial \boldsymbol{\theta}}\right) - \left(\frac{1}{\sigma^2}\right)\sum_{i=1}^{n} \varepsilon_i \frac{\partial g(y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \dfrac{-n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n} \varepsilon_i^2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ 0 \end{bmatrix}. \quad \textbf{(17-50)}$$

These equations will usually be nonlinear, so a solution must be obtained iteratively. One special case that is common is a model in which $\boldsymbol{\theta}$ is a single parameter. Given a particular value of θ, we would maximize $\ln L$ with respect to $\boldsymbol{\beta}$ by using nonlinear least squares. [It would be simpler yet if, in addition, $h(\mathbf{x}_i, \boldsymbol{\beta})$ were linear so that we could use linear least squares. See the following application.] Therefore, a way to maximize L for all the parameters is to scan over values of θ for the one that, with the associated least squares estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, gives the highest value of $\ln L$. (Of course, this requires that we know roughly what values of θ to examine.)

If $\boldsymbol{\theta}$ is a vector of parameters, then direct maximization of L with respect to the full set of parameters may be preferable. (Methods of maximization are discussed in Appendix E.) There is an additional simplification that may be useful. Whatever values are ultimately obtained for the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the estimate of σ^2 will be given by (17-49). If we insert this solution in (17-48), then we obtain the **concentrated log-likelihood,**

$$\ln L_c = \sum_{i=1}^{n} \ln J(y_i, \boldsymbol{\theta}) - \frac{n}{2}[1 + \ln(2\pi)] - \frac{n}{2} \ln \left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \right]. \tag{17-51}$$

This equation is a function only of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. We can maximize it with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and obtain the estimate of σ^2 as a by-product. (See Section E.6.3 for details.)

An estimate of the asymptotic covariance matrix of the maximum likelihood estimators can be obtained by inverting the estimated information matrix. It is quite likely, however, that the Berndt et al. (1974) estimator will be much easier to compute. The log of the density for the ith observation is the ith term in (17-50). The derivatives of $\ln L_i$ with respect to the unknown parameters are

$$\mathbf{g}_i = \begin{bmatrix} \partial \ln L_i / \partial \boldsymbol{\beta} \\ \partial \ln L_i / \partial \boldsymbol{\theta} \\ \partial \ln L_i / \partial \sigma^2 \end{bmatrix} = \begin{bmatrix} (\varepsilon_i / \sigma^2)[\partial h(\mathbf{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}] \\ (1/J_i)[\partial J_i / \partial \boldsymbol{\theta}] - (\varepsilon_i / \sigma^2)[\partial g(y_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}] \\ (1/(2\sigma^2))\left[\varepsilon_i^2 / \sigma^2 - 1\right] \end{bmatrix}. \tag{17-52}$$

The asymptotic covariance matrix for the maximum likelihood estimators is estimated using

$$\text{Est.Asy. Var[MLE]} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right]^{-1} = (\hat{\mathbf{G}}' \hat{\mathbf{G}})^{-1}. \tag{17-53}$$

Note that the preceding includes of a row and a column for σ^2 in the covariance matrix. In a model that transforms y as well as $\mathbf{x}$, the Hessian of the log-likelihood is generally not block diagonal with respect to $\boldsymbol{\theta}$ and σ^2. When y is transformed, the maximum likelihood estimators of $\boldsymbol{\theta}$ and σ^2 are positively correlated, because both parameters reflect the scaling of the dependent variable in the model. This result may seem counterintuitive. Consider the difference in the variance estimators that arises when a linear and a loglinear model are estimated. The variance of $\ln y$ around its mean is obviously different from that of y around its mean. By contrast, consider what happens when only the independent variables are transformed, for example, by the Box–Cox transformation. The slope estimators vary accordingly, but in such a way that the variance of y around its conditional mean will stay constant.[16]

Example 17.5 A Generalized Production Function

The Cobb–Douglas function has often been used to study production and cost. Among the assumptions of this model is that the average cost of production increases or decreases monotonically with increases in output. This assumption is in direct contrast to the standard textbook treatment of a U-shaped average cost curve as well as to a large amount of empirical evidence. (See Example 7.3 for a well-known application.) To relax this assumption, Zellner

[16] See Seaks and Layson (1983).

TABLE 17.2 Generalized Production Function Estimates

	Maximum Likelihood			
	Estimate	***SE(1)***	***SE(2)***	***Nonlinear Least Squares***
β_1	2.914822	0.44912	0.12534	2.108925
β_2	0.350068	0.10019	0.094354	0.257900
β_3	1.092275	0.16070	0.11498	0.878388
θ	0.106666	0.078702		−0.031634
σ^2	0.0427427			0.0151167
$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$	1.068567			0.7655490
ln L	−8.939044			−13.621256

and Revankar (1970) proposed a generalization of the Cobb–Douglas production function.[17] Their model allows economies of scale to vary with output and to increase and then decrease as output rises:

$$\ln y + \theta y = \ln \gamma + \alpha(1 - \delta) \ln K + \alpha\delta \ln L + \varepsilon.$$

Note that the right-hand side of their model is intrinsically linear according to the results of Section 7.3.3. The model as a whole, however, is intrinsically nonlinear due to the parametric transformation of y appearing on the left.

For Zellner and Revankar's production function, the Jacobian of the transformation from ε_i to y_i is $\partial\varepsilon_i/\partial y_i = (\theta + 1/y_i)$. Some simplification is achieved by writing this as $(1 + \theta y_i)/y_i$. The log-likelihood is then

$$\ln L = \sum_{i=1}^{n} \ln(1 + \theta y_i) - \sum_{i=1}^{n} \ln y_i - \frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2,$$

where $\varepsilon_i = (\ln y_i + \theta y_i - \beta_1 - \beta_2 \ln \text{capital}_i - \beta_3 \ln \text{labor}_i)$. Estimation of this model is straightforward. For a given value of θ, $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are estimated by linear least squares. Therefore, to estimate the full set of parameters, we could scan over the range of zero to one for θ. The value of θ that, with its associated least squares estimates of $\boldsymbol{\beta}$ and σ^2, maximizes the log-likelihood function provides the maximum likelihood estimate. This procedure was used by Zellner and Revankar. The results given in Table 17.2 were obtained by maximizing the log-likelihood function directly, instead. The statewide data on output, capital, labor, and number of establishments in the transportation industry used in Zellner and Revankar's study are given in Appendix Table F9.2 and Example 16.6. For this application, y = value added per firm, K = capital per firm, and L = labor per firm.

Maximum likelihood and nonlinear least squares estimates are shown in Table 17.2. The asymptotic standard errors for the maximum likelihood estimates are labeled SE(1). These are computed using the BHHH form of the asymptotic covariance matrix. The second set, SE(2), are computed treating the estimate of θ as fixed; they are the usual linear least squares results using $(\ln y + \theta y)$ as the dependent variable in a linear regression. Clearly, these results would be very misleading. The final column of Table 10.2 lists the simple nonlinear least squares estimates. No standard errors are given, because there is no appropriate formula for computing the asymptotic covariance matrix. The sum of squares does not provide an appropriate method for computing the pseudoregressors for the parameters in the transformation. The last two rows of the table display the sum of squares and the log-likelihood function evaluated at the parameter estimates. As expected, the log-likelihood is much larger at the maximum likelihood estimates. In contrast, the nonlinear least squares estimates lead to a much lower sum of squares; least squares is still *least* squares.

[17] An alternative approach is to model costs directly with a flexible functional form such as the translog model. This approach is examined in detail in Chapter 14.

Example 17.6 An LM Test for (Log-) Linearity

A natural generalization of the **Box–Cox regression model** (Section 9.3.2) is

$$y^{(\lambda)} = \boldsymbol{\beta}'\mathbf{x}^{(\lambda)} + \varepsilon. \tag{17-54}$$

where $z^{(\lambda)} = (z^\lambda - 1)/\lambda$. This form includes the linear ($\lambda = 1$) and loglinear ($\lambda = 0$) models as special cases. The Jacobian of the transformation is $|d\varepsilon/dy| = y^{\lambda-1}$. The log-likelihood function for the model with normally distributed disturbances is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + (\lambda - 1)\sum_{i=1}^{n}\ln y_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \boldsymbol{\beta}'\mathbf{x}_i^{(\lambda)}\right)^2. \tag{17-55}$$

The MLEs of λ and $\boldsymbol{\beta}$ are computed by maximizing this function. The estimator of σ^2 is the mean squared residual as usual. We can use a one-dimensional grid search over λ—for a given value of λ, the MLE of $\boldsymbol{\beta}$ is least squares using the transformed data. It must be remembered, however, that the criterion function includes the Jacobian term.

We will use the BHHH estimator of the asymptotic covariance matrix for the maximum likelihood. The derivatives of the log likelihood are

$$\begin{bmatrix} \dfrac{\partial \ln L}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ln L}{\partial \lambda} \\ \dfrac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} \dfrac{\varepsilon_i \mathbf{x}_i^{(\lambda)}}{\sigma^2} \\ \ln y_i - \dfrac{\varepsilon_i}{\sigma^2}\left[\dfrac{\partial y_i^{(\lambda)}}{\partial \lambda} - \displaystyle\sum_{k=1}^{K}\beta_k \frac{\partial x_{ik}^{(\lambda)}}{\partial \lambda}\right] \\ \dfrac{1}{2\sigma^2}\left[\dfrac{\varepsilon_i^2}{\sigma^2} - 1\right] \end{bmatrix} = \sum_{i=1}^{n}\mathbf{g}_i \tag{17-56}$$

where

$$\frac{\partial[z^\lambda - 1]/\lambda}{\partial\lambda} = \frac{\lambda z^\lambda \ln z - (z^\lambda - 1)}{\lambda^2} = \frac{1}{\lambda}\left(z^\lambda \ln z - z^{(\lambda)}\right). \tag{17-57}$$

(See Exercise 6 in Chapter 9.) The estimator of the asymptotic covariance matrix for the maximum likelihood estimator is given in (17-53).

The Box–Cox model provides a framework for a specification test of linearity versus log-linearity. To assemble this result, consider first the basic model

$$y = f(x, \beta_1, \beta_2, \lambda) + \varepsilon = \beta_1 + \beta_2 x^{(\lambda)} + \varepsilon.$$

The pseudoregressors are $x_1^* = 1, x_2^* = x^{(\lambda)}, x_3^* = \beta_2(\partial x^{(\lambda)}/\partial\lambda)$ as given above. We now consider a Lagrange multiplier test of the hypothesis that λ equals zero. The test is carried out by first regressing y on a constant and $\ln x$ (i.e., the regressor evaluated at $\lambda = 0$) and then computing nR_*^2 in the regression of the residuals from this first regression on x_1^*, x_2^*, and x_3^*, also evaluated at $\lambda = 0$. The first and second of these are 1 and $\ln x$. To obtain the third, we require $x_{3\,|\,\lambda=0}^* = \beta_2 \lim_{\lambda\to 0}(\partial x^{(\lambda)}/\partial\lambda)$. Applying L'Hôpital's rule to the right-hand side of (12-57), differentiate numerator and denominator with respect to λ. This produces

$$\lim_{\lambda\to 0}\frac{\partial x^{(\lambda)}}{\partial\lambda} = \lim_{\lambda\to 0}\left[x^\lambda(\ln x)^2 - \frac{\partial x^{(\lambda)}}{\partial\lambda}\right] = \frac{1}{2}\lim_{\lambda\to 0}x^\lambda(\ln x)^2 = \frac{1}{2}(\ln x)^2.$$

Therefore, $\lim_{\lambda\to 0} x_3^* = \beta_2[\frac{1}{2}(\ln x)^2]$. The Lagrange multiplier test is carried out in two steps. First, we regress y on a constant and $\ln x$ and compute the residuals. Second, we regress these residuals on a constant, $\ln x$, and $b_2(\frac{1}{2}\ln^2 x)$, where b_2 is the coefficient on $\ln x$ in the first regression. The Lagrange multiplier statistic is nR^2 from the second regression. To generalize this procedure to several regressors, we would use the logs of all the regressors at the first step. Then, the additional regressor for the second regression would be

$$x_\lambda^* = \sum_{k=1}^{K} b_k(\tfrac{1}{2}\ln^2 x_k),$$

where the sum is taken over all the variables that are transformed in the original model and the b_k's are the least squares coefficients in the first regression.

By extending this process to the model of (17-54), we can devise a bona fide test of log-linearity (against the more general model, not linearity). [See Davidson and MacKinnon (1985). A test of linearity can be conducted using $\lambda = 1$, instead.) Computing the various terms at $\lambda = 0$ again, we have

$$\hat{\varepsilon}_i = \ln y_i - \hat{\beta}_1 - \hat{\beta}_2 \ln x_i,$$

where as before, $\hat{\beta}_1$ and $\hat{\beta}_2$ are computed by the least squares regression of $\ln y$ on a constant and $\ln x$. Let $\hat{\varepsilon}_i^* = \frac{1}{2}\ln^2 y_i - \hat{\beta}_2(\frac{1}{2}\ln^2 x_i)$. Then

$$\hat{\mathbf{g}}_i = \begin{bmatrix} \hat{\varepsilon}_i/\hat{\sigma}^2 \\ (\ln x_i)\hat{\varepsilon}_i/\hat{\sigma}^2 \\ \ln y_i - \hat{\varepsilon}_i\hat{\varepsilon}_i^*/\hat{\sigma}^2 \\ \left(\hat{\varepsilon}_i^2/\hat{\sigma}^2 - 1\right)/(2\hat{\sigma}^2) \end{bmatrix}.$$

If there are K regressors in the model, then the second component in $\hat{\mathbf{g}}_i$ will be a vector containing the logs of the variables, whereas $\hat{\varepsilon}_i^*$ in the third becomes

$$\hat{\varepsilon}_i^* = \frac{1}{2}\ln^2 y_i - \sum_{k=1}^{K} \hat{\beta}_k \left(\frac{1}{2}\ln^2 x_{ik}\right).$$

Using the Berndt et al. estimator given in (10-54), we can now construct the Lagrange multiplier statistic as

$$\text{LM} = \chi^2[1] = \left(\sum_{i=1}^{n} \hat{\mathbf{g}}_i\right)' \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'\right]^{-1} \left(\sum_{i=1}^{n} \hat{\mathbf{g}}_i\right) = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i},$$

where $\mathbf{G}$ is the $n \times (K + 2)$ matrix whose columns are $\mathbf{g}_1$ through $\mathbf{g}_{K+2}$ and $\mathbf{i}$ is a column of 1s. The usefulness of this approach for either of the models we have examined is that in testing the hypothesis, it is not necessary to compute the nonlinear, unrestricted, Box–Cox regression.

17.6.3 NONNORMAL DISTURBANCES—THE STOCHASTIC FRONTIER MODEL

This final application will examine a regressionlike model in which the disturbances do not have a normal distribution. The model developed here also presents a convenient platform on which to illustrate the use of the invariance property of maximum likelihood estimators to simplify the estimation of the model.

A lengthy literature commencing with theoretical work by Knight (1933), Debreu (1951), and Farrell (1957) and the pioneering empirical study by Aigner, Lovell, and Schmidt (1977) has been directed at models of production that specifically account for the textbook proposition that a production function is a theoretical ideal.[18] If $y = f(\mathbf{x})$ defines a production relationship between inputs, $\mathbf{x}$, and an output, y, then for any given $\mathbf{x}$, the observed value of y must be less than or equal to $f(\mathbf{x})$. The implication for an empirical regression model is that in a formulation such as $y = h(\mathbf{x}, \boldsymbol{\beta}) + u$, u must be negative. Since the theoretical production function is an ideal—the frontier of efficient

[18] A survey by Greene (1997b) appears in Pesaran and Schmidt (1997). Kumbhakar and Lovell (2000) is a comprehensive reference on the subject.

production—any nonzero disturbance must be interpreted as the result of inefficiency. A strictly orthodox interpretation embedded in a Cobb–Douglas production model might produce an empirical frontier production model such as

$$\ln y = \beta_1 + \Sigma_k \beta_k \ln x_k - u, \quad u \geq 0.$$

The gamma model described in Example 5.1 was an application. One-sided disturbances such as this one present a particularly difficult estimation problem. The primary theoretical problem is that any measurement error in $\ln y$ must be embedded in the disturbance. The practical problem is that the entire estimated function becomes a slave to any single errantly measured data point.

Aigner, Lovell, and Schmidt proposed instead a formulation within which observed deviations from the production function could arise from two sources: (1) productive inefficiency as we have defined it above and that would necessarily be negative; and (2) idiosyncratic effects that are specific to the firm and that could enter the model with either sign. The end result was what they labeled the "stochastic frontier":

$$\begin{aligned}\ln y &= \beta_1 + \Sigma_k \beta_k \ln x_k - u + v, \quad u \geq 0, \quad v \sim N[0, \sigma_v^2].\\ &= \beta_1 + \Sigma_k \beta_k \ln x_k + \varepsilon.\end{aligned}$$

The frontier for any particular firm is $h(\mathbf{x}, \boldsymbol{\beta}) + v$, hence the name stochastic frontier. The inefficiency term is u, a random variable of particular interest in this setting. Since the data are in log terms, u is a measure of the percentage by which the particular observation fails to achieve the frontier, ideal production rate.

To complete the specification, they suggested two possible distributions for the inefficiency term, the absolute value of a normally distributed variable and an exponentially distributed variable. The density functions for these two compound distributions are given by Aigner, Lovell, and Schmidt; let $\varepsilon = v - u$, $\lambda = \sigma_u/\sigma_v$, $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$, and $\Phi(z) =$ the probability to the left of z in the standard normal distribution [see Sections B.4.1 and E.5.6]. For the "half-normal" model,

$$\ln h(\varepsilon_i \mid \boldsymbol{\beta}, \lambda, \sigma) = \left[-\ln \sigma - \left(\frac{1}{2}\right) \log \frac{2}{\pi} - \frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2 + \ln \Phi\left(\frac{-\varepsilon_i \lambda}{\sigma}\right)\right],$$

whereas for the exponential model

$$\ln h(\varepsilon_i \mid \boldsymbol{\beta}, \theta, \sigma_v) = \left[\ln \theta + \frac{1}{2}\theta^2 \sigma_v^2 + \theta \varepsilon_i + \ln \Phi\left(-\frac{\varepsilon_i}{\sigma_v} - \theta \sigma_v\right)\right].$$

Both these distributions are asymmetric. We thus have a regression model with a nonnormal distribution specified for the disturbance. The disturbance, ε, has a nonzero mean as well; $E[\varepsilon] = -\sigma_u(2/\pi)^{1/2}$ for the half-normal model and $-1/\theta$ for the exponential model. Figure 17.3 illustrates the density for the half-normal model with $\sigma = 1$ and $\lambda = 2$. By writing $\beta_0 = \beta_1 + E[\varepsilon]$ and $\varepsilon^* = \varepsilon - E[\varepsilon]$, we obtain a more conventional formulation

$$\ln y = \beta_0 + \Sigma_k \beta_k \ln x_k + \varepsilon^*$$

which does have a disturbance with a zero mean but an asymmetric, nonnormal distribution. The asymmetry of the distribution of ε^* does not negate our basic results for least squares in this classical regression model. This model satisfies the assumptions of the

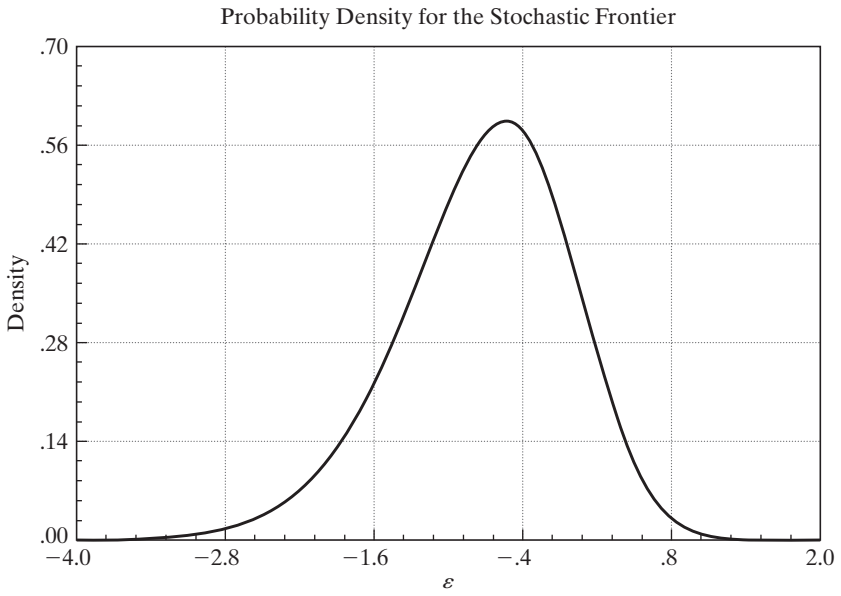


FIGURE 17.3 Density for the Disturbance in the Stochastic Frontier Model.

Gauss–Markov theorem, so least squares is unbiased and consistent (save for the constant term), and efficient among linear unbiased estimators. In this model, however, the maximum likelihood estimator is not linear, and it is more efficient than least squares.

We will work through maximum likelihood estimation of the half-normal model in detail to illustrate the technique. The log likelihood is

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

This is not a particularly difficult log-likelihood to maximize numerically. Nonetheless, it is instructive to make use of a convenience that we noted earlier. Recall that maximum likelihood estimators are invariant to one-to-one transformation. If we let $\theta = 1/\sigma$ and $\boldsymbol{\gamma} = (1/\sigma)\boldsymbol{\beta}$, the log-likelihood function becomes

$$\ln L = n \ln \theta - \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} (\theta y_i - \boldsymbol{\gamma}' \mathbf{x}_i)^2 + \sum_{i=1}^{n} \ln \Phi[-\lambda(\theta y_i - \boldsymbol{\gamma}' \mathbf{x}_i)].$$

As you could verify by trying the derivations, this transformation brings a dramatic simplification in the manipulation of the log-likelihood and its derivatives. We will make repeated use of the functions

$$\alpha_i = \varepsilon_i / \sigma = \theta y_i - \boldsymbol{\gamma}' \mathbf{x}_i,$$

$$\delta(y_i, \mathbf{x}_i, \lambda, \theta, \boldsymbol{\gamma}) = \frac{\phi[-\lambda \alpha_i]}{\Phi[-\lambda \alpha_i]} = \delta_i.$$

$$\Delta_i = -\delta_i(-\lambda \alpha_i + \delta_i)$$

(The second of these is the derivative of the function in the final term in log L. The third is the derivative of δ_i with respect to its argument; $\Delta_i < 0$ for all values of $\lambda\alpha_i$.) It will also be convenient to define the $(K+1) \times 1$ columns vectors $\mathbf{z}_i = (\mathbf{x}_i', -y_i)'$ and $\mathbf{t}_i = (\mathbf{0}', 1/\theta)'$. The likelihood equations are

$$\frac{\partial \ln L}{\partial(\boldsymbol{\gamma}', \theta)'} = \sum_{i=1}^{n} \mathbf{t}_i + \sum_{i=1}^{n} \alpha_i \mathbf{z}_i + \lambda \sum_{i=1}^{n} \boldsymbol{\delta}_i \mathbf{z}_i = \mathbf{0},$$

$$\frac{\partial \ln L}{\partial \lambda} = -\sum_{i=1}^{n} \delta_i \alpha_i = 0$$

and the second derivatives are

$$\mathbf{H}(\boldsymbol{\gamma}, \theta, \lambda) = \sum_{i=1}^{n} \left\{ \begin{bmatrix} (\lambda^2 \Delta_i - 1)\mathbf{z}_i \mathbf{z}_i' & (\delta_i - \lambda\alpha_i \Delta_i)\mathbf{z}_i \\ (\delta_i - \lambda\alpha_i \Delta_i)\mathbf{z}_i' & \alpha_i^2 \Delta_i \end{bmatrix} - \begin{bmatrix} \mathbf{t}_i \mathbf{t}_i' & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix} \right\}.$$

The estimator of the asymptotic covariance matrix for the directly estimated parameters is

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\gamma}}', \hat{\theta}, \hat{\lambda}]' = \left\{ -\mathbf{H}[\hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\theta}}, \hat{\lambda}] \right\}^{-1}.$$

There are two sets of transformations of the parameters in our formulation. In order to recover estimates of the original structural parameters $\sigma = 1/\theta$ and $\boldsymbol{\beta} = \boldsymbol{\gamma}/\theta$ we need only transform the MLEs. Since these transformations are one to one, the MLEs of σ and $\boldsymbol{\beta}$ are $1/\hat{\theta}$ and $\hat{\boldsymbol{\gamma}}/\hat{\theta}$. To compute an asymptotic covariance matrix for these estimators we will use the delta method, which will use the derivative matrix

$$\mathbf{G} = \begin{bmatrix} \partial\hat{\boldsymbol{\beta}}/\partial\hat{\boldsymbol{\gamma}}' & \partial\hat{\boldsymbol{\beta}}/\partial\hat{\theta} & \partial\hat{\boldsymbol{\beta}}/\partial\hat{\lambda} \\ \partial\hat{\sigma}/\partial\hat{\boldsymbol{\gamma}}' & \partial\hat{\sigma}/\partial\hat{\theta} & \partial\hat{\sigma}/\partial\hat{\lambda} \\ \partial\hat{\lambda}/\partial\hat{\boldsymbol{\gamma}}' & \partial\hat{\lambda}/\partial\hat{\theta} & \partial\hat{\lambda}/\partial\hat{\lambda} \end{bmatrix} = \begin{bmatrix} (1/\hat{\theta})\mathbf{I} & -(1/\hat{\theta}^2)\hat{\boldsymbol{\gamma}} & \mathbf{0} \\ \mathbf{0}' & -(1/\hat{\theta}^2) & 0 \\ \mathbf{0}' & 0 & 1 \end{bmatrix}.$$

Then, for the recovered parameters, we

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}', \hat{\sigma}, \hat{\lambda}]' = \mathbf{G} \times \left\{ -\mathbf{H}[\hat{\boldsymbol{\gamma}}', \hat{\theta}, \hat{\lambda}] \right\}^{-1} \times \mathbf{G}'.$$

For the half-normal model, we would also rely on the invariance of maximum likelihood estimators to recover estimates of the deeper variance parameters, $\sigma_v^2 = \sigma^2/(1+\lambda^2)$ and $\sigma_u^2 = \sigma^2\lambda^2/(1+\lambda^2)$.

The stochastic frontier model is a bit different from those we have analyzed previously in that the disturbance is the central focus of the analysis rather than the catchall for the unknown and unknowable factors omitted from the equation. Ideally, we would like to estimate u_i for each firm in the sample to compare them on the basis of their productive efficiency. (The parameters of the production function are usually of secondary interest in these studies.) Unfortunately, the data do not permit a direct estimate, since with estimates of $\boldsymbol{\beta}$ in hand, we are only able to compute a direct estimate of $\varepsilon = y - \mathbf{x}'\boldsymbol{\beta}$. Jondrow et al. (1982), however, have derived a useful approximation that is now the standard measure in these settings,

$$E[u \mid \varepsilon] = \frac{\sigma\lambda}{1+\lambda^2}\left[\frac{\phi(z)}{1-\Phi(z)} - z\right], \quad z = \frac{\varepsilon\lambda}{\sigma},$$

TABLE 17.3 Estimated Stochastic Frontier Functions

	Least Squares			*Half-Normal Model*			*Exponential Model*		
		Standard			*Standard*			*Standard*	
Coefficient	*Estimate*	*Error*	*t Ratio*	*Estimate*	*Error*	*t Ratio*	*Estimate*	*Error*	*t Ratio*
Constant	1.844	0.234	7.896	2.081	0.422	4.933	2.069	0.290	7.135
β_k	0.245	0.107	2.297	0.259	0.144	1.800	0.262	0.120	2.184
β_l	0.805	0.126	6.373	0.780	0.170	4.595	0.770	0.138	5.581
σ	0.236			0.282	0.087	3.237			
σ_u	—			0.222			0.136		
σ_v	—			0.190			0.171	0.054	3.170
λ	—			1.265	1.620	0.781			
θ	—						7.398	3.931	1.882
log L	2.2537			2.4695			2.8605		

for the half normal-model, and

$$E[u \mid \varepsilon] = z + \sigma_v \frac{\phi(z/\sigma_v)}{\Phi(z/\sigma_v)}, \qquad z = \varepsilon - \theta\sigma_v^2$$

for the exponential model. These values can be computed using the maximum likelihood estimates of the structural parameters in the model. In addition, a structural parameter of interest is the proportion of the total variance of ε that is due to the inefficiency term. For the half-normal model, $\text{Var}[\varepsilon] = \text{Var}[u] + \text{Var}[v] = (1 - 2/\pi)\sigma_u^2 + \sigma_v^2$, whereas for the exponential model, the counterpart is $1/\theta^2 + \sigma_v^2$.

Example 17.7 Stochastic Frontier Model

Appendix Table F9.2 lists 25 statewide observations used by Zellner and Revankar (1970) to study production in the transportation equipment manufacturing industry. We have used these data to estimate the stochastic frontier models. Results are shown in Table 17.3.[19] The Jondrow, et al. (1982) estimates of the inefficiency terms are listed in Table 17.4. The estimates of the parameters of the production function, β_1, β_2, and β_3 are fairly similar, but the variance parameters, σ_u and σ_v, appear to be quite different. Some of the parameter difference is illusory, however. The variance components for the half-normal model are $(1 - 2/\pi)\sigma_u^2 = 0.0179$ and $\sigma_v^2 = 0.0361$, whereas those for the exponential model are $1/\theta^2 = 0.0183$ and $\sigma_v^2 = 0.0293$. In each case, about one-third of the total variance of ε is accounted for by the variance of u.

17.6.4 CONDITIONAL MOMENT TESTS OF SPECIFICATION

A spate of studies has shown how to use **conditional moment restrictions** for specification testing as well as estimation.[20] The logic of the conditional moment (CM) based specification test is as follows. The model specification implies that certain moment restrictions will hold in the population from which the data were drawn. If the specification

[19] N is the number of establishments in the state. Zellner and Revankar used per establishment data in their study. The stochastic frontier model has the intriguing property that if the least squares residuals are skewed in the positive direction, then least squares with $\lambda = 0$ maximizes the log-likelihood. This property, in fact, characterizes the data above when scaled by N. Since that leaves a not particularly interesting example and it does not occur when the data are not normalized, for purposes of this illustration we have used the unscaled data to produce Table 17.3. We do note that this result is a common, vexing occurrence in practice.

[20] See, for example, Pagan and Vella (1989).

TABLE 17.4 Estimated Inefficiencies

State	*Half-Normal*	*Exponential*	*State*	*Half-Normal*	*Exponential*
Alabama	0.2011	0.1459	Maryland	0.1353	0.0925
California	0.1448	0.0972	Massachusetts	0.1564	0.1093
Connecticut	0.1903	0.1348	Michigan	0.1581	0.1076
Florida	0.5175	0.5903	Missouri	0.1029	0.0704
Georgia	0.1040	0.0714	New Jersey	0.0958	0.0659
Illinois	0.1213	0.0830	New York	0.2779	0.2225
Indiana	0.2113	0.1545	Ohio	0.2291	0.1698
Iowa	0.2493	0.2007	Pennsylvania	0.1501	0.1030
Kansas	0.1010	0.0686	Texas	0.2030	0.1455
Kentucky	0.0563	0.0415	Virginia	0.1400	0.0968
Louisiana	0.2033	0.1507	Washington	0.1105	0.0753
Maine	0.2226	0.1725	West Virginia	0.1556	0.1124
Wisconsin	0.1407	0.0971			

is correct, then the sample data should mimic the implied relationships. For example, in the classical regression model, the assumption of homoscedasticity implies that the disturbance variance is independent of the regressors. As such,

$$E\{\mathbf{x}_i[(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 - \sigma^2]\} = E[\mathbf{x}_i(\varepsilon_i^2 - \sigma^2)] = \mathbf{0}.$$

If, on the other hand, the regression is heteroscedastic *in a way that depends on* $\mathbf{x}_i$, then this covariance will not be zero. If the hypothesis of homoscedasticity is correct, then we would expect the sample counterpart to the moment condition,

$$\bar{\mathbf{r}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(e_i^2 - s^2),$$

where e_i is the OLS residual, to be close to zero. (This computation appears in Breusch and Pagan's LM test for homoscedasticity. See Section 11.4.3.) The practical problems to be solved are (1) to formulate suitable moment conditions that do correspond to the hypothesis test, which is usually straightforward; (2) to devise the appropriate sample counterpart; and (3) to devise a suitable measure of closeness to zero of the sample moment estimator. The last of these will be in the framework of the Wald statistics that we have examined at various points in this book. So the problem will be to devise the appropriate covariance matrix for the sample moments.

Consider a general case in which the moment condition is written in terms of variables in the model $[y_i, \mathbf{x}_i, \mathbf{z}_i]$ and parameters (as in the linear regression model) $\hat{\boldsymbol{\theta}}$. The sample moment can be written

$$\bar{\mathbf{r}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i(y_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{r}}_i. \tag{17-58}$$

The hypothesis is that based on the true $\boldsymbol{\theta}$, $E[\mathbf{r}_i] = \mathbf{0}$. Under the null hypothesis that $E[\mathbf{r}_i] = \mathbf{0}$ and assuming that plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and that a central limit theorem (Theorem D.18 or D.19) applies to $\sqrt{n}\,\bar{\mathbf{r}}(\boldsymbol{\theta})$ so that

$$\sqrt{n}\bar{\mathbf{r}}(\boldsymbol{\theta}) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}]$$

for some covariance matrix $\boldsymbol{\Sigma}$ that we have yet to estimate, it follows that the Wald statistic,

$$n\bar{\mathbf{r}}'\hat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{r}} \xrightarrow{d} \chi^2(J), \tag{17-59}$$

where the degrees of freedom J is the number of moment restrictions being tested and $\hat{\boldsymbol{\Sigma}}$ is an estimate of $\boldsymbol{\Sigma}$. Thus, the statistic can be referred to the chi-squared table.

It remains to determine the estimator of $\boldsymbol{\Sigma}$. The full derivation of $\boldsymbol{\Sigma}$ is fairly complicated. [See Pagan and Vella (1989, pp. S32–S33).] But when the vector of parameter estimators is a maximum likelihood estimator, as it would be for the least squares estimator with normally distributed disturbances and for most of the other estimators we consider, a surprisingly simple estimator can be used. Suppose that the parameter vector used to compute the moments above is obtained by solving the equations

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(y_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_i = \mathbf{0}, \tag{17-60}$$

where $\hat{\boldsymbol{\theta}}$ is the estimated parameter vector [e.g., $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ in the linear model]. For the linear regression model, that would be the normal equations

$$\frac{1}{n}\mathbf{X}'\mathbf{e} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\mathbf{b}) = \mathbf{0}.$$

Let the matrix $\mathbf{G}$ be the $n \times K$ matrix with ith row equal to $\hat{\mathbf{g}}_i'$. In a maximum likelihood problem, $\mathbf{G}$ is the matrix of derivatives of the individual terms in the log-likelihood function with respect to the parameters. This is the $\mathbf{G}$ used to compute the BHHH estimator of the information matrix. [See (17-18).] Let $\mathbf{R}$ be the $n \times J$ matrix whose ith row is $\hat{\mathbf{r}}_i'$. Pagan and Vella show that for maximum likelihood estimators, $\boldsymbol{\Sigma}$ can be estimated using

$$\mathbf{S} = \frac{1}{n}[\mathbf{R}'\mathbf{R} - \mathbf{R}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{R}].^{21} \tag{17-61}$$

This equation looks like an involved matrix computation, but it is simple with any regression program. Each element of $\mathbf{S}$ is the mean square or cross-product of the least squares residuals in a linear regression of a column of $\mathbf{R}$ on the variables in $\mathbf{G}$.[22] Therefore, the operational version of the statistic is

$$C = n\bar{\mathbf{r}}'\mathbf{S}^{-1}\bar{\mathbf{r}} = \frac{1}{n}\mathbf{i}'\mathbf{R}[\mathbf{R}'\mathbf{R} - \mathbf{R}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{R}]^{-1}\mathbf{R}'\mathbf{i}, \tag{17-62}$$

where $\mathbf{i}$ is an $n \times 1$ column of ones, which, once again, is referred to the appropriate critical value in the chi-squared table. This result provides a joint test that all the moment conditions are satisfied simultaneously. An individual test of just one of the moment

[21] It might be tempting just to use $(1/n)\mathbf{R}'\mathbf{R}$. This idea would be incorrect, because $\mathbf{S}$ accounts for $\mathbf{R}$ being a function of the estimated parameter vector that is converging to its probability limit at the same rate as the sample moments are converging to theirs.

[22] If the estimator is not an MLE, then estimation of $\boldsymbol{\Sigma}$ is more involved but also straightforward using basic matrix algebra. The advantage of (17-62) is that it involves simple sums of variables that have already been computed to obtain $\hat{\boldsymbol{\theta}}$ and $\bar{\mathbf{r}}$. Note, as well, that if $\boldsymbol{\theta}$ has been estimated by maximum likelihood, then the term $(\mathbf{G}'\mathbf{G})^{-1}$ is the BHHH estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. If it were more convenient, then this estimator could be replaced with any other appropriate estimator of Asy. Var$[\hat{\boldsymbol{\theta}}]$.

restrictions in isolation can be computed even more easily than a joint test. For testing one of the L conditions, say the ℓth one, the test can be carried out by a simple t test of whether the constant term is zero in a linear regression of the ℓth column of $\mathbf{R}$ on a constant term and all the columns of $\mathbf{G}$. In fact, the test statistic in (17-62) could also be obtained by stacking the J columns of $\mathbf{R}$ and treating the L equations as a seemingly unrelated regressions model with $(\mathbf{i}, \mathbf{G})$ as the (identical) regressors in each equation and then testing the joint hypothesis that all the constant terms are zero. (See Section 14.2.3.)

Example 17.8 Testing for Heteroscedasticity in the Linear Regression Model

Suppose that the linear model is specified as

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i .$$

To test whether

$$E\left[z_i^2\left(\varepsilon_i^2 - \sigma^2\right)\right] = 0,$$

we linearly regress $z_i^2(e_i^2 - s^2)$ on a constant, e_i, $x_i e_i$, and $z_i e_i$. A standard t test of whether the constant term in this regression is zero carries out the test. To test the joint hypothesis that there is no heteroscedasticity with respect to both x and z, we would regress both $x_i^2(e_i^2 - s^2)$ and $z_i^2(e_i^2 - s^2)_i$ on $[1, e_i, x_i e_i, z_i e_i]$ and collect the two columns of residuals in $\mathbf{V}$. Then $\mathbf{S} = (1/n)\mathbf{V}'\mathbf{V}$. The moment vector would be

$$\bar{\mathbf{r}} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_i \\ z_i \end{bmatrix}\left(e_i^2 - s^2\right).$$

The test statistic would now be

$$C = n\bar{\mathbf{r}}'\mathbf{S}^{-1}\bar{\mathbf{r}} = n\bar{\mathbf{r}}'\left[\frac{1}{n}\mathbf{V}'\mathbf{V}\right]^{-1}\bar{\mathbf{r}}.$$

We will examine other conditional moment tests using this method in Section 22.3.4 where we study the specification of the censored regression model.

17.7 TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION

The applied literature contains a large and increasing number of models in which one model is embedded in another, which produces what are broadly known as "two-step" estimation problems. Consider an (admittedly contrived) example in which we have the following.

Model 1. Expected number of children $= E[y_1 \mid \mathbf{x}_1, \boldsymbol{\theta}_1]$.
Model 2. Decision to enroll in job training $= y_2$, a function of $\left(\mathbf{x}_2, \boldsymbol{\theta}_2, E[y_1 \mid \mathbf{x}_1, \boldsymbol{\theta}_1]\right)$.

There are two parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The first appears in the second model, although not the reverse. In such a situation, there are two ways to proceed. **Full information maximum likelihood (FIML)** estimation would involve forming the joint distribution $f(y_1, y_2 \mid \mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ of the two random variables and then maximizing

the full log-likelihood function,

$$\ln L = \sum_{i=1}^{n} f(y_{i1}, y_{i2} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

A second, or two-step, **limited information maximum likelihood (LIML)** procedure for this kind of model could be done by estimating the parameters of model 1, since it does not involve $\boldsymbol{\theta}_2$, and then maximizing a conditional log-likelihood function using the estimates from Step 1:

$$\ln \hat{L} = \sum_{i=1}^{n} f[y_{i2} \mid \mathbf{x}_{i2}, \boldsymbol{\theta}_2, (\mathbf{x}_{i1}, \hat{\boldsymbol{\theta}}_1)].$$

There are at least two reasons one might proceed in this fashion. First, it may be straightforward to formulate the two separate log-likelihoods, but very complicated to derive the joint distribution. This situation frequently arises when the two variables being modeled are from different kinds of populations, such as one discrete and one continuous (which is a very common case in this framework). The second reason is that maximizing the separate log-likelihoods may be fairly straightforward, but maximizing the joint log-likelihood may be numerically complicated or difficult.[23] We will consider a few examples. Although we will encounter FIML problems at various points later in the book, for now we will present some basic results for two-step estimation. Proofs of the results given here can be found in an important reference on the subject, Murphy and Topel (1985).

Suppose, then, that our model consists of the two marginal distributions, $f_1(y_1 \mid \mathbf{x}_1, \boldsymbol{\theta}_1)$ and $f_2(y_2 \mid \mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Estimation proceeds in two steps.

1. Estimate $\boldsymbol{\theta}_1$ by maximum likelihood in Model 1. Let $(1/n)\hat{\mathbf{V}}_1$ be n times any of the estimators of the asymptotic covariance matrix of this estimator that were discussed in Section 17.4.6.
2. Estimate $\boldsymbol{\theta}_2$ by maximum likelihood in model 2, with $\hat{\boldsymbol{\theta}}_1$ inserted in place of $\boldsymbol{\theta}_1$ as if it were known. Let $(1/n)\hat{\mathbf{V}}_2$ be n times any appropriate estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_2$.

The argument for consistency of $\hat{\boldsymbol{\theta}}_2$ is essentially that if $\boldsymbol{\theta}_1$ *were* known, then all our results for MLEs would apply for estimation of $\boldsymbol{\theta}_2$, and since plim $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1$, asymptotically, this line of reasoning is correct. But the same line of reasoning is not sufficient to justify using $(1/n)\hat{\mathbf{V}}_2$ as the estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_2$. Some correction is necessary to account for an estimate of $\boldsymbol{\theta}_1$ being used in estimation of $\boldsymbol{\theta}_2$. The essential result is the following.

[23] There is a third possible motivation. If either model is misspecified, then the FIML estimates of both models will be inconsistent. But if only the second is misspecified, at least the first will be estimated consistently. Of course, this result is only "half a loaf," but it may be better than none.

THEOREM 17.8 Asymptotic Distribution of the Two-Step MLE [Murphy and Topel (1985)]

If the standard regularity conditions are met for both log-likelihood functions, then the second-step maximum likelihood estimator of $\boldsymbol{\theta}_2$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\mathbf{V}_2^* = \frac{1}{n}\left[\mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}' - \mathbf{R}\mathbf{V}_1\mathbf{C}' - \mathbf{C}\mathbf{V}_1\mathbf{R}']\mathbf{V}_2\right],$$

where

$$\mathbf{V}_1 = \text{Asy.Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)] \text{ based on } \ln L_1,$$

$$\mathbf{V}_2 = \text{Asy.Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)] \text{ based on } \ln L_2 \mid \boldsymbol{\theta}_1,$$

$$\mathbf{C} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_1'}\right)\right], \quad \mathbf{R} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_1}{\partial \boldsymbol{\theta}_1'}\right)\right].$$

The correction of the asymptotic covariance matrix at the second step requires some additional computation. Matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ are estimated by the respective uncorrected covariance matrices. Typically, the BHHH estimators,

$$\hat{\mathbf{V}}_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)\right]^{-1}$$

and

$$\hat{\mathbf{V}}_2 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2'}\right)\right]^{-1}$$

are used. The matrices $\mathbf{R}$ and $\mathbf{C}$ are obtained by summing the individual observations on the cross products of the derivatives. These are estimated with

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)$$

and

$$\hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)$$

Example 17.9 Two-Step ML Estimation

Continuing the example discussed at the beginning of this section, we suppose that y_{i2} is a binary indicator of the choice whether to enroll in the program ($y_{i2} = 1$) or not ($y_{i2} = 0$) and that the probabilities of the two outcomes are

$$\text{Prob}[y_{i2} = 1 \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}] = \frac{e^{\mathbf{x}_{i2}'\boldsymbol{\beta} + \gamma E[y_{i1} \mid \mathbf{x}_{i1}']}}{1 + e^{\mathbf{x}_{i2}'\boldsymbol{\beta} + \gamma E[y_{i1} \mid \mathbf{x}_{i1}']}}$$

and Prob$[y_{i2} = 0 \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}] = 1 -$ Prob$[y_{i2} = 1 \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}]$, where $\mathbf{x}_{i2}$ is some covariates that might influence the decision, such as marital status or age and $\mathbf{x}_{i1}$ are determinants of family size. This setup is a **logit** model. We will develop this model more fully in Chapter 21. The *expected value* of y_{i1} appears in the probability. (Remark: The expected, rather than the actual value was chosen deliberately. Otherwise, the models would differ substantially. In our case, we might view the difference as that between an ex ante decision and an ex post one.) Suppose that the number of children can be described by a Poisson distribution (see Section B.4.8) dependent on some variables $\mathbf{x}_{i1}$ such as education, age, and so on. Then

$$\text{Prob}[y_{i1} = j \mid \mathbf{x}_{i1}] = \frac{e^{-\lambda_i}\lambda_i^j}{j!}, \quad j = 0, 1, \ldots,$$

and suppose, as is customary, that

$$E[y_{i1}] = \lambda_i = \exp(\mathbf{x}_{i1}'\boldsymbol{\delta}).$$

The models involve $\boldsymbol{\theta} = [\boldsymbol{\delta}, \boldsymbol{\beta}, \gamma]$, where $\boldsymbol{\theta}_1 = \boldsymbol{\delta}$. In fact, it is unclear what the joint distribution of y_1 and y_2 might be, but two-step estimation is straightforward. For model 1, the log-likelihood and its first derivatives are

$$\begin{aligned}
\ln L_1 &= \sum_{i=1}^{n} \ln f_1(y_{i1} \mid \mathbf{x}_{i1}, \boldsymbol{\delta}) \\
&= \sum_{i=1}^{n} [-\lambda_i + y_{i1} \ln \lambda_i - \ln y_{i1}!] = \sum_{i=1}^{n} [-\exp(\mathbf{x}_{i1}'\boldsymbol{\delta}) + y_{i1}(\mathbf{x}_{i1}'\boldsymbol{\delta}) - \ln y_{i1}!], \\
\frac{\partial \ln L_1}{\partial \boldsymbol{\delta}} &= \sum_{i=1}^{n} (y_{i1} - \lambda_i)\mathbf{x}_{i1} = \sum_{i=1}^{n} u_i \mathbf{x}_{i1}.
\end{aligned}$$

Computation of the estimates is developed in Chapter 21. Any of the three estimators of $\mathbf{V}_1$ is also easy to compute, but the BHHH estimator is most convenient, so we use

$$\hat{\mathbf{V}}_1 = \left[\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \mathbf{x}_{i1}\mathbf{x}_{i1}' \right]^{-1}.$$

[In this and the succeeding summations, we are actually estimating expectations of the various matrices.]

We can write the density function for the second model as

$$f_2(y_{i2} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}) = P_i^{y_{i2}} \times (1 - P_i)^{1 - y_{i2}},$$

where $P_i = \text{Prob}[y_{i2} = 1 \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}]$ as given earlier. Then

$$\ln L_2 = \sum_{i=1}^{n} y_{i2} \ln P_i + (1 - y_{i2}) \ln(1 - P_i).$$

For convenience, let $\hat{\mathbf{x}}_{i2}^* = [\mathbf{x}_{i2}', \exp(\mathbf{x}_{i1}'\hat{\boldsymbol{\delta}})]'$, and recall that $\boldsymbol{\theta}_2 = [\boldsymbol{\beta}, \gamma]'$. Then

$$\ln \hat{L}_2 = \sum_{i=1}^{n} y_{i2}[\hat{\mathbf{x}}_{i2}^{*\prime}\boldsymbol{\theta}_2 - \ln(1 + \exp(\hat{\mathbf{x}}_{i2}^{*\prime}\boldsymbol{\theta}_2))] + (1 - y_{i2})[-\ln(1 + \exp(\hat{\mathbf{x}}_{i2}^{*\prime}\boldsymbol{\theta}_2))].$$

So, at the second step, we create the additional variable, append it to $\mathbf{x}_{i2}$, and estimate the logit model as if $\boldsymbol{\delta}$ (and this additional variable) were actually observed instead of estimated. The maximum likelihood estimates of $[\boldsymbol{\beta}, \boldsymbol{\gamma}]$ are obtained by maximizing this function. (See

Chapter 21.) After a bit of manipulation, we find the convenient result that

$$\frac{\partial \ln \hat{L}_2}{\partial \boldsymbol{\theta}_2} = \sum_{i=1}^{n} (y_{i2} - P_i)\hat{\mathbf{x}}_{i2}^* = \sum_{i=1}^{n} v_i \hat{\mathbf{x}}_{i2}^*.$$

Once again, any of the three estimators could be used for estimating the asymptotic covariance matrix, but the BHHH estimator is convenient, so we use

$$\hat{\mathbf{V}}_2 = \left[\frac{1}{n} \sum_{i=1}^{n} \hat{v}_i^2 \hat{\mathbf{x}}_{i2}^* \hat{\mathbf{x}}_{i2}^{*\prime} \right]^{-1}.$$

For the final step, we must correct the asymptotic covariance matrix using $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$. What remains to derive—the few lines are left for the reader—is

$$\frac{\partial \ln L_2}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} v_i [\gamma \exp(\mathbf{x}_{i1}'\boldsymbol{\delta})]\mathbf{x}_{i1}.$$

So, using our estimates,

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \hat{v}_i^2 [\exp(\mathbf{x}_{i1}'\hat{\boldsymbol{\delta}})]\hat{\mathbf{x}}_{i2}^* \mathbf{x}_{i1}', \quad \text{and} \quad \hat{\mathbf{R}} = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i \hat{v}_i \hat{\mathbf{x}}_{i2}^* \mathbf{x}_{i1}'.$$

We can now compute the correction.

In many applications, the covariance of the two gradients $\mathbf{R}$ converges to zero. When the first and second step estimates are based on different samples, $\mathbf{R}$ is exactly zero. For example, in our application above, $\mathbf{R} = \sum_{i=1}^{n} u_i v_i \mathbf{x}_{i2}^* \mathbf{x}_{i1}'$. The two "residuals," u and v, may well be uncorrelated. This assumption must be checked on a model-by-model basis, but in such an instance, the third and fourth terms in $\mathbf{V}_2^*$ vanish asymptotically and what remains is the simpler alternative,

$$\mathbf{V}_2^{**} = (1/n)[\mathbf{V}_2 + \mathbf{V}_2 \mathbf{C} \mathbf{V}_1 \mathbf{C}' \mathbf{V}_2].$$

We will examine some additional applications of this technique (including an empirical implementation of the preceding example) later in the book. Perhaps the most common application of two-step maximum likelihood estimation in the current literature, especially in regression analysis, involves inserting a prediction of one variable into a function that describes the behavior of another.

17.8 MAXIMUM SIMULATED LIKELIHOOD ESTIMATION

The technique of maximum simulated likelihood (MSL) is essentially a classical sampling theory counterpart to the hierarchical Bayesian estimator we considered in Section 16.2.4. Since the celebrated paper of Berry, Levinsohn, and Pakes (1995), and a related literature advocated by McFadden and Train (2000), maximum simulated likelihood estimation has been used in a large and growing number of studies based on log-likelihoods that involve integrals that are expectations.[24] In this section, we will lay out some general results for MSL estimation by developing a particular application,

[24] A major reference for this set of techniques is Gourieroux and Monfort (1996).

the random parameters model. This general modeling framework has been used in the majority of the received applications. We will then continue the application to the discrete choice model for panel data that we began in Section 16.2.4.

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

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i$$

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 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 denoted

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

where $g(.)$ is the underlying marginal density of the heterogeneity. For the T observations in group i, the joint conditional density is

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

The unconditional density for $\mathbf{y}_i$ is obtained by integrating over $\boldsymbol{\beta}_i$,

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = E_{\boldsymbol{\beta}_i}[f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i)] = \int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i) g(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) d\boldsymbol{\beta}_i.$$

Collecting terms, and making the transformation from $\mathbf{v}_i$ to $\boldsymbol{\beta}_i$, the true log-likelihood would be

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} \ln \left\{ \int_{\mathbf{v}_i} \left[\prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i) \right] g(\mathbf{v}_i \mid \boldsymbol{\Sigma}) d\mathbf{v}_i \right\} \\ &= \sum_{i=1}^{n} \ln \left\{ \int_{\mathbf{v}_i} f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i) g(\mathbf{v}_i \mid \boldsymbol{\Sigma}) d\mathbf{v}_i \right\}.\end{aligned}$$

Each of the n terms involves an expectation over $\mathbf{v}_i$. The end result of the integration is a function of $(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma})$ which is then maximized.

As in the previous applications, it will not be possible to maximize the log-likelihood in this form because there is no closed form for the integral. We have considered two approaches to maximizing such a log-likelihood. In the latent class formulation, it is assumed that the parameter vector takes one of a discrete set of values, and the log-likelihood is maximized over this discrete distribution as well as the structural parameters. (See Section 16.2.3.) The hierarchical Bayes procedure used Markov Chain–Monte Carlo methods to sample from the joint posterior distribution of the underlying parameters and used the empirical mean of the sample of draws as the estimator. We now consider a third approach to estimating the parameters of a model of this form, maximum simulated likelihood estimation.

The terms in the log-likelihood are each of the form

$$\ln L_i = E_{\mathbf{v}_i}[f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i)].$$

As noted, we do not have a closed form for this function, so we cannot compute it directly. Suppose we could sample randomly from the distribution of $\mathbf{v}_i$. If an appropriate law

of large numbers can be applied, then

$$\lim_{R\to\infty} \frac{1}{R}\sum_{r=1}^{R} f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_{ir}) = E_{\mathbf{v}_i}[f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i)]$$

where $\mathbf{v}_{ir}$ is the rth random draw from the distribution. This suggests a strategy for computing the log-likelihood. We can substitute this approximation to the expectation into the log-likelihood function. With sufficient random draws, the approximation can be made as close to the true function as desired. [The theory for this approach is discussed in Gourieroux and Monfort (1996), Bhat (1999), and Train (1999, 2002). Practical details on applications of the method are given in Greene (2001).] A detail to add concerns how to sample from the distribution of $\mathbf{v}_i$. There are many possibilities, but for now, we consider the simplest case, the multivariate normal distribution. Write $\boldsymbol{\Sigma}$ in the Cholesky form $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$ where $\mathbf{L}$ is a lower triangular matrix. Now, let $\mathbf{u}_{ir}$ be a vector of K independent draws from the standard normal distribution. Then a draw from the multivariate distribution with covariance matrix $\boldsymbol{\Sigma}$ is simply $\mathbf{v}_{ir} = \mathbf{L}\mathbf{u}_{ir}$. The simulated log-likelihood is

$$\ln L_S = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R}\sum_{r=1}^{R} \left[\prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{L}\mathbf{u}_{ir}) \right] \right\}.$$

The resulting function is maximized with respect to $\boldsymbol{\beta}$, $\boldsymbol{\Delta}$ and $\mathbf{L}$. This is obviously not a simple calculation, but it is feasible, and much easier than trying to manipulate the integrals directly. In fact, for most problems to which this method has been applied, the computations are surprisingly simple. The intricate part is obtaining the function and its derivatives. But, the functions are usually index function models that involve $\mathbf{x}_{it}'\boldsymbol{\beta}_i$ which greatly simplifies the derivations.

Inference in this setting does not involve any new results. The estimated asymptotic covariance matrix for the estimated parameters is computed by manipulating the derivatives of the simulated log-likelihood. The Wald and likelihood ratio statistics are also computed the way they would usually be. As before, we are interested in estimating person specific parameters. A prior estimate might simply use $\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i$, but this would not use all the information in the sample. A posterior estimate would compute

$$\hat{E}_{\mathbf{v}_i}[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Delta}, \mathbf{z}_i, \boldsymbol{\Sigma}] = \frac{\sum_{r=1}^{R} \hat{\boldsymbol{\beta}}_{ir} f(\mathbf{y}_i \mid \mathbf{X}_i, \hat{\boldsymbol{\beta}}_{ir})}{\sum_{r=1}^{R} f(\mathbf{y}_i \mid \mathbf{X}_i, \hat{\boldsymbol{\beta}}_{ir})}, \qquad \hat{\boldsymbol{\beta}}_{ir} = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Delta}}\mathbf{z}_i + \hat{\mathbf{L}}\mathbf{u}_{ir}.$$

Mechanical details on computing the MSLE are omitted. The interested reader is referred to Gourieroux and Monfort (1996), Train (2000, 2002), and Greene (2001, 2002) for details.

Example 17.10 Maximum Simulated Likelihood Estimation of a Binary Choice Model

We continue Example 16.5 where estimates of a binary choice model for product innovation are obtained. The model is for Prob[$y_{it} = 1 \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i$] where

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

The independent variables in the model are

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),

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

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

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

The sample consists of 1,270 German manufacturing firms observed for five years, 1984–1988. The density that enters the log-likelihood is

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

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim \text{N}[\mathbf{0}, \boldsymbol{\Sigma}].$$

To be consistent with Bertschek and Lechner (1998) we did not fit any firm-specific, time-invariant components in the main equation for $\boldsymbol{\beta}_i$.

Table 17.5 presents the estimated coefficients for the basic probit model in the first column. The estimates of the means, $\boldsymbol{\beta}$ are shown in the second column. There appear to be large differences in the parameter estimates, though this can be misleading since there is large variation across the firms in the posterior estimates. The third column presents the square roots of the implied diagonal elements of $\boldsymbol{\Sigma}$ computed as the diagonal elements of $\mathbf{LL}'$. These estimated standard deviations are for the underlying distribution of the parameter in the model—they are not estimates of the standard deviation of the sampling distribution of the estimator. For the mean parameter, that is shown in parentheses in the second column. The fourth column presents the sample means and standard deviations of the 1,270 estimated posterior

TABLE 17.5 Estimated Random Parameters Model

	Probit	*RP Means*	*RP Std. Devs.*	*Empirical Distn.*	*Posterior*
Constant	−1.96	−3.91	2.70	−3.27	−3.38
	(0.23)	(0.20)		(0.57)	(2.14)
lnSales	0.18	0.36	0.28	0.32	0.34
	(0.022)	(0.019)		(0.15)	(0.09)
Rel.Size	1.07	6.01	5.99	3.33	2.58
	(0.14)	(0.22)		(2.25)	(1.30)
Import	1.13	1.51	0.84	2.01	1.81
	(0.15)	(0.13)		(0.58)	(0.74)
FDI	2.85	3.81	6.51	3.76	3.63
	(0.40)	(0.33)		(1.69)	(1.98)
Prod.	−2.34	−5.10	13.03	−8.15	−5.48
	(0.72)	(0.73)		(8.29)	(1.78)
RawMtls	−0.28	−0.31	1.65	−0.18	−0.08
	(0.081)	(0.075)		(0.57)	(0.37)
Invest.	0.19	0.27	1.42	0.27	0.29
	(0.039)	(0.032)		(0.38)	(0.13)
ln L	−4114.05		−3498.654		

estimates of the coefficients. The last column repeats the estimates for the latent class model. The agreement in the two sets of estimates is striking in view of the crude approximation given by the latent class model.

Figures 17.4a and b present kernel density estimators of the firm-specific probabilities computed at the 5-year means for the random parameters model and with the original probit estimates. The estimated probabilities are strikingly similar to the latent class model, and also fairly similar to, though smoother than the probit estimates.

FIGURE 17.4a Probit Probabilities.

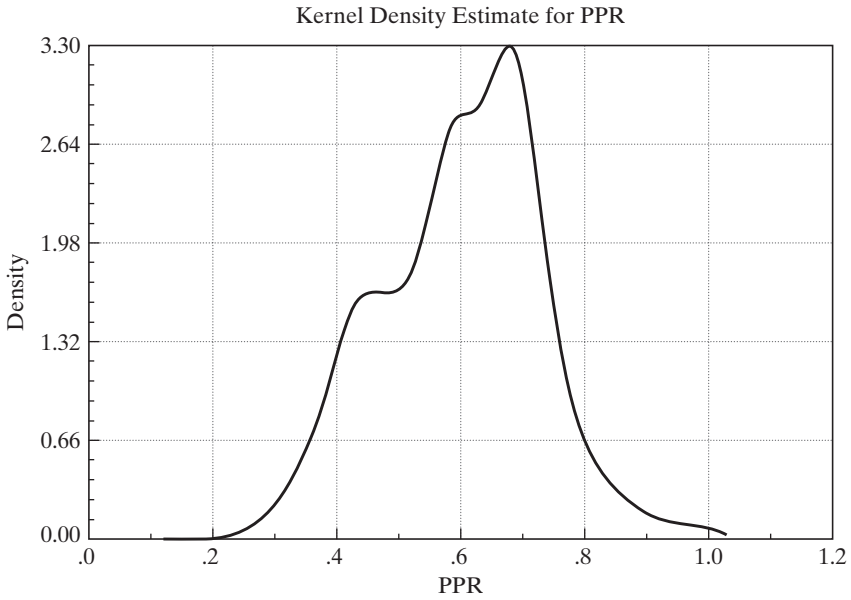


FIGURE 17.4b Random Parameters Probabilities.

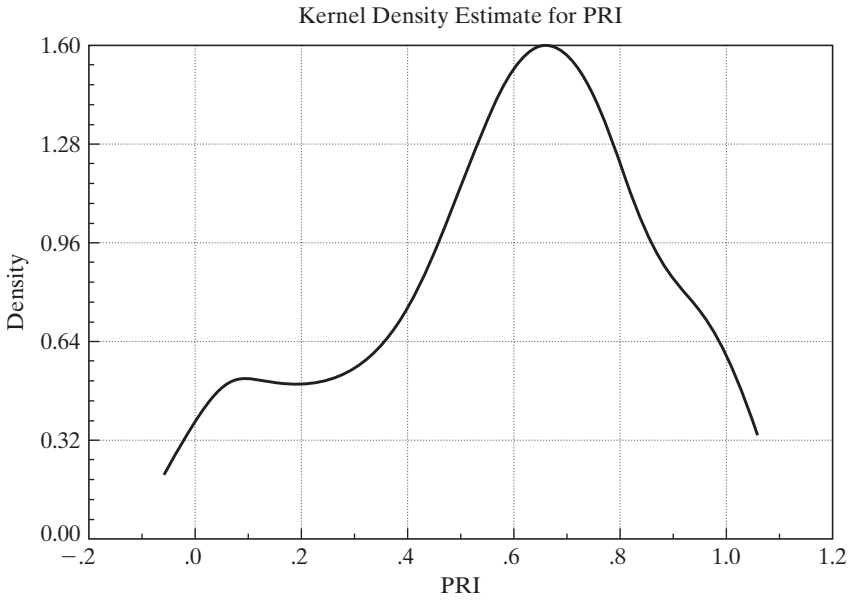


Figure 17.5 shows the kernel density estimate for the firm-specific estimates of the log sales coefficient. The comparison to Figure 16.5 shows some striking difference. The random parameters model produces estimates that are similar in magnitude, but the distributions are actually quite different. Which should be preferred? Only on the basis that the three point discrete latent class model is an approximation to the continuous variation model, we would prefer the latter.

FIGURE 17.5a Random Parameters, β_{sales}.

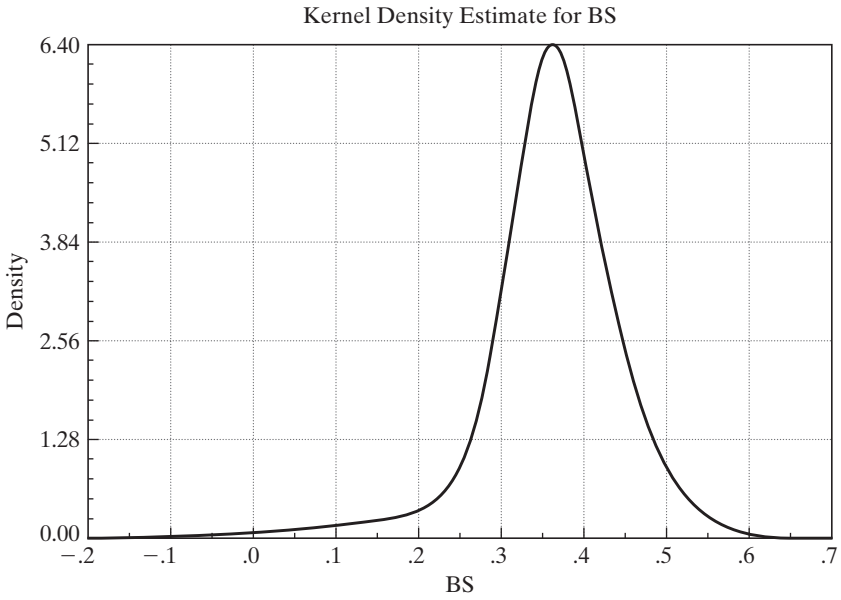
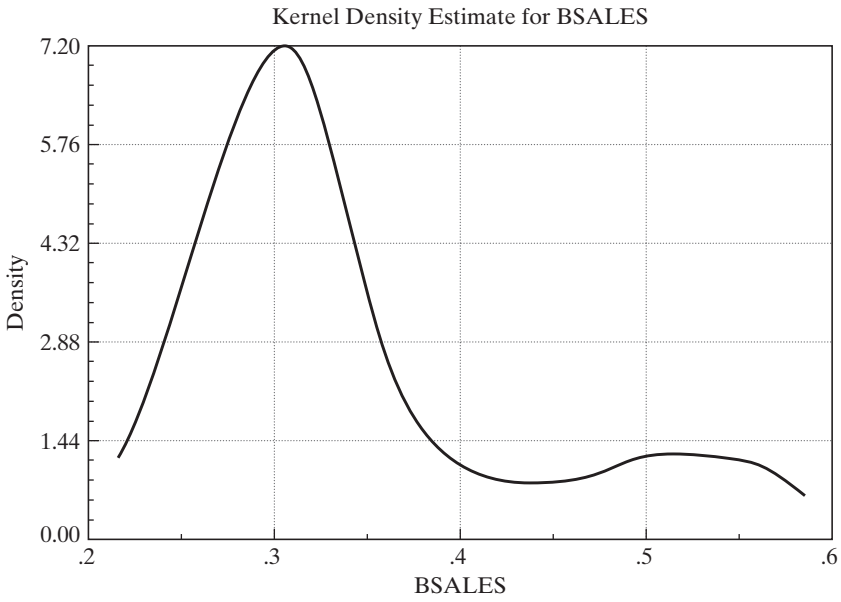


FIGURE 17.5b Latent Class Model, β_{sales}.



17.9 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION AND ROBUST ASYMPTOTIC COVARIANCE MATRICES

Maximum likelihood estimation requires complete specification of the distribution of the observed random variable. If the correct distribution is something other than what we assume, then the likelihood function is misspecified and the desirable properties of the MLE might not hold. This section considers a set of results on an estimation approach that is robust to some kinds of model misspecification. For example, we have found that in a model, if the conditional mean function is $E[y \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, then certain estimators, such as least squares, are "robust" to specifying the wrong distribution of the disturbances. That is, LS is MLE if the disturbances are normally distributed, but we can still claim some desirable properties for LS, including consistency, even if the disturbances are not normally distributed. This section will discuss some results that relate to what happens if we maximize the "wrong" log-likelihood function, and for those cases in which the estimator is consistent despite this, how to compute an appropriate asymptotic covariance matrix for it.[25]

Let $f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})$ be the true probability density for a random variable y_i given a set of covariates $\mathbf{x}_i$ and parameter vector $\boldsymbol{\beta}$. The log-likelihood function is $(1/n) \log L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = (1/n) \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$. The MLE, $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, is the sample statistic that maximizes this function. (The division of $\log L$ by n does not affect the solution.) We maximize the log-likelihood function by equating its derivatives to zero, so the MLE is obtained by solving the set of empirical moment equations

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{ML}})}{\partial \hat{\boldsymbol{\beta}}_{\mathrm{ML}}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) = \bar{\mathbf{d}}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) = \mathbf{0}.$$

The population counterpart to the sample moment equation is

$$E\left[\frac{1}{n}\frac{\partial \log L}{\partial \boldsymbol{\beta}}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(\boldsymbol{\beta})\right] = E[\bar{\mathbf{d}}(\boldsymbol{\beta})] = \mathbf{0}.$$

Using what we know about GMM estimators, if $E[\bar{\mathbf{d}}(\boldsymbol{\beta})] = \mathbf{0}$, then $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is consistent and asymptotically normally distributed, with asymptotic covariance matrix equal to

$$\mathbf{V}_{\mathrm{ML}} = [\mathbf{G}(\boldsymbol{\beta})'\mathbf{G}(\boldsymbol{\beta})]^{-1}\mathbf{G}(\boldsymbol{\beta})'\{\mathrm{Var}[\bar{\mathbf{d}}(\boldsymbol{\beta})]\}\mathbf{G}(\boldsymbol{\beta})[\mathbf{G}(\boldsymbol{\beta})'\mathbf{G}(\boldsymbol{\beta})]^{-1},$$

where $\mathbf{G}(\boldsymbol{\beta}) = \operatorname{plim} \partial\bar{\mathbf{d}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$. Since $\bar{\mathbf{d}}(\boldsymbol{\beta})$ is the derivative vector, $\mathbf{G}(\boldsymbol{\beta})$ is $1/n$ times the expected Hessian of $\log L$; that is, $(1/n)E[\mathbf{H}(\boldsymbol{\beta})] = \bar{\mathbf{H}}(\boldsymbol{\beta})$. As we saw earlier, $\mathrm{Var}[\partial \log L/\partial\boldsymbol{\beta}] = -E[\mathbf{H}(\boldsymbol{\beta})]$. Collecting all seven appearances of $(1/n)E[\mathbf{H}(\boldsymbol{\beta})]$, we obtain the familiar result $\mathbf{V}_{\mathrm{ML}} = \{-E[\mathbf{H}(\boldsymbol{\beta})]\}^{-1}$. [All the ns cancel and $\mathrm{Var}[\bar{\mathbf{d}}] = (1/n)\bar{\mathbf{H}}(\boldsymbol{\beta})$.] Note that this result depends crucially on the result $\mathrm{Var}[\partial \log L/\partial\boldsymbol{\beta}] = -E[\mathbf{H}(\boldsymbol{\beta})]$.

[25] The following will sketch a set of results related to this estimation problem. The important references on this subject are White (1982a); Gourieroux, Monfort, and Trognon (1984); Huber (1967); and Amemiya (1985). A recent work with a large amount of discussion on the subject is Mittelhammer et al. (2000). The derivations in these works are complex, and we will only attempt to provide an intuitive introduction to the topic.

The maximum likelihood estimator is obtained by maximizing the function $\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n} \log f(y_i, \mathbf{x}_i, \boldsymbol{\beta})$. This function converges to its expectation as $n \to \infty$. Since this function is the log-likelihood for the sample, it is also the case (not proven here) that as $n \to \infty$, it attains its unique maximum at the true parameter vector, $\boldsymbol{\beta}$. (We used this result in proving the consistency of the maximum likelihood estimator.) Since $\operatorname{plim} \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = E[\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$, it follows (by interchanging differentiation and the expectation operation) that $\operatorname{plim} \partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial\boldsymbol{\beta} = E[\partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}]$. But, if this function achieves its *maximum* at $\boldsymbol{\beta}$, then it must be the case that $\operatorname{plim} \partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$.

An estimator that is obtained by maximizing a criterion function is called an M estimator [Huber (1967)] or an extremum estimator [Amemiya (1985)]. Suppose that we obtain an estimator by maximizing some other function, $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$ that, although not the log-likelihood function, also attains its unique maximum at the true $\boldsymbol{\beta}$ as $n \to \infty$. Then the preceding argument might produce a consistent estimator with a known asymptotic distribution. For example, the log-likelihood for a linear regression model with normally distributed disturbances with *different* variances, $\sigma^2\omega_i$, is

$$\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-1}{2}\left[\log(2\pi\sigma^2\omega_i) + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2\omega_i}\right]\right\}.$$

By maximizing this function, we obtain the maximum likelihood estimator. But we also examined another estimator, simple least squares, which maximizes $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = -(1/n)\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$. As we showed earlier, least squares is consistent and asymptotically normally distributed even with this extension, so it qualifies as an M estimator of the sort we are considering here.

Now consider the general case. Suppose that we estimate $\boldsymbol{\beta}$ by maximizing a criterion function

$$M_n(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \log g(y_i|\mathbf{x}_i, \boldsymbol{\beta}).$$

Suppose as well that $\operatorname{plim} M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = E[M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$ and that as $n \to \infty$, $E[M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$ attains its unique maximum at $\boldsymbol{\beta}$. Then, by the argument we used above for the MLE, $\operatorname{plim} \partial M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial\boldsymbol{\beta} = E[\partial M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}] = \mathbf{0}$. Once again, we have a set of moment equations for estimation. Let $\hat{\boldsymbol{\beta}}_E$ be the estimator that maximizes $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$. Then the estimator is defined by

$$\frac{\partial M_n(\mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E} = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_E) = \mathbf{0}.$$

Thus, $\hat{\boldsymbol{\beta}}_E$ is a GMM estimator. Using the notation of our earlier discussion, $\mathbf{G}(\hat{\boldsymbol{\beta}}_E)$ is the symmetric Hessian of $E[M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$, which we will denote $(1/n)E[\mathbf{H}_M(\hat{\boldsymbol{\beta}}_E)] = \bar{\mathbf{H}}_M(\hat{\boldsymbol{\beta}}_E)$. Proceeding as we did above to obtain $\mathbf{V}_{\mathrm{ML}}$, we find that the appropriate asymptotic covariance matrix for the extremum estimator would be

$$\mathbf{V}_E = [\bar{\mathbf{H}}_M(\boldsymbol{\beta})]^{-1}\left(\frac{1}{n}\boldsymbol{\Phi}\right)[\mathbf{H}_M(\boldsymbol{\beta})]^{-1}$$

where $\boldsymbol{\Phi} = \operatorname{Var}[\partial \log g(y_i|\mathbf{x}_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}]$, and, as before, the asymptotic distribution is normal.

The Hessian in $\mathbf{V}_E$ can easily be estimated by using its empirical counterpart,

$$\text{Est.}[\bar{\mathbf{H}}_M(\hat{\boldsymbol{\beta}}_E)] = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}}_E)}{\partial\hat{\boldsymbol{\beta}}_E\partial\hat{\boldsymbol{\beta}}'_E}.$$

But, $\boldsymbol{\Phi}$ remains to be specified, and it is unlikely that we would know what function to use. The important difference is that in this case, the variance of the first derivatives vector need not equal the Hessian, so $\mathbf{V}_E$ does not simplify. We can, however, consistently estimate $\boldsymbol{\Phi}$ by using the sample variance of the first derivatives,

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}}\right]\left[\frac{\partial \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}'}\right].$$

If this were the maximum likelihood estimator, then $\hat{\boldsymbol{\Phi}}$ would be the BHHH estimator that we have used at several points. For example, for the least squares estimator in the heteroscedastic linear regression model, the criterion is $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = -(1/n)\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$, the solution is $\mathbf{b}$, $\mathbf{G}(\mathbf{b}) = (-2/n)\mathbf{X}'\mathbf{X}$, and

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{i=1}^{n}[2\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})][2\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})]' = \frac{4}{n}\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i'.$$

Collecting terms, the 4s cancel and we are left precisely with the White estimator of (11-13)!

At this point, we consider the motivation for all this weighty theory. One disadvantage of maximum likelihood estimation is its requirement that the density of the observed random variable(s) be fully specified. The preceding discussion suggests that in some situations, we can make somewhat fewer assumptions about the distribution than a full specification would require. The extremum estimator is robust to some kinds of specification errors. One useful result to emerge from this derivation is an estimator for the asymptotic covariance matrix of the extremum estimator that is robust at least to some misspecification. In particular, if we obtain $\hat{\boldsymbol{\beta}}_E$ by maximizing a criterion function that satisfies the other assumptions, then the appropriate estimator of the asymptotic covariance matrix is

$$\text{Est. }\mathbf{V}_E = \frac{1}{n}[\bar{\mathbf{H}}(\hat{\boldsymbol{\beta}}_E)]^{-1}\hat{\boldsymbol{\Phi}}(\hat{\boldsymbol{\beta}}_E)[\bar{\mathbf{H}}(\hat{\boldsymbol{\beta}}_E)]^{-1}.$$

If $\hat{\boldsymbol{\beta}}_E$ is the true MLE, then $\mathbf{V}_E$ simplifies to $\{-[\mathbf{H}(\hat{\boldsymbol{\beta}}_E)]\}^{-1}$. In the current literature, this estimator has been called the "sandwich" estimator. There is a trend in the current literature to compute this estimator routinely, regardless of the likelihood function. It is worth noting that if the log-likelihood is not specified correctly, then the parameter estimators are likely to be inconsistent, save for the cases such as those noted below, so robust estimation of the asymptotic covariance matrix may be misdirected effort. But if the likelihood function is correct, then the sandwich estimator is unnecessary. This method is not a general patch for misspecified models. Not every likelihood function qualifies as a consistent extremum estimator *for the parameters of interest in the model.*

One might wonder at this point how likely it is that the conditions needed for all this to work will be met. There are applications in the literature in which this machinery has been used that probably do not meet these conditions, such as the tobit model of Chapter 22. We have seen one important case. Least squares in the generalized

regression model passes the test. Another important application is models of "individual heterogeneity" in cross-section data. Evidence suggests that simple models often overlook unobserved sources of variation across individuals in cross sections, such as unmeasurable "family effects" in studies of earnings or employment. Suppose that the correct model for a variable is $h(y_i|\mathbf{x}_i, \mathbf{v}_i, \boldsymbol{\beta}, \theta)$, where $\mathbf{v}_i$ is a random term that is not observed and θ is a parameter of the distribution of $\mathbf{v}$. The correct log-likelihood function is $\Sigma_i \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \theta) = \Sigma_i \log \int_v h(y_i|\mathbf{x}_i, \mathbf{v}_i, \boldsymbol{\beta}, \theta) f(\mathbf{v}_i)\, dv_i$. Suppose that we maximize some other pseudo-log-likelihood function, $\Sigma_i \log g(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ and then use the sandwich estimator to estimate the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Does this produce a consistent estimator of the true parameter vector? Surprisingly, sometimes it does, even though it has ignored the nuisance parameter, θ. We saw one case, using OLS in the GR model with heteroscedastic disturbances. Inappropriately fitting a Poisson model when the negative binomial model is correct—see Section 21.9.3—is another case. For some specifications, using the wrong likelihood function in the probit model with proportions data (Section 21.4.6) is a third. [These two examples are suggested, with several others, by Gourieroux, Monfort, and Trognon (1984).] We do emphasize once again that the sandwich estimator, in and of itself, is not necessarily of any virtue if the likelihood function is misspecified and the other conditions for the M estimator are not met.

17.10 SUMMARY AND CONCLUSIONS

This chapter has presented the theory and several applications of maximum likelihood estimation, which is the most frequently used estimation technique in econometrics after least squares. The maximum likelihood estimators are consistent, asymptotically normally distributed, and efficient among estimators that have these properties. The drawback to the technique is that it requires a fully parametric, detailed specification of the data generating process. As such, it is vulnerable to misspecification problems. The next chapter considers GMM estimation techniques which are less parametric, but more robust to variation in the underlying data generating process.

Key Terms and Concepts

- Asymptotic efficiency
- Asymptotic normality
- Asymptotic variance
- BHHH estimator
- Box–Cox model
- Conditional moment restrictions
- Concentrated log-likelihood
- Consistency
- Cramér–Rao lower bound
- Efficient score
- Estimable parameters
- Full information maximum likelihood
- Identification
- Information matrix
- Information matrix equality
- Invariance
- Jacobian
- Lagrange multiplier test
- Likelihood equation
- Likelihood function
- Likelihood inequality
- Likelihood ratio test
- Limited information maximum likelihood
- Maximum likelihood estimator
- Nonlinear least squares
- Outer product of gradients estimator
- Regularity conditions
- Score test
- Stochastic frontier
- Two-step maximum likelihood
- Wald statistic
- Wald test

Exercises

1. Assume that the distribution of x is $f(x) = 1/\theta, 0 \leq x \leq \theta$. In random sampling from this distribution, prove that the sample maximum is a consistent estimator of θ. Note: You can prove that the maximum is the maximum likelihood estimator of θ. But the usual properties do not apply here. Why not? [Hint: Attempt to verify that the expected first derivative of the log-likelihood with respect to θ is zero.]
2. In random sampling from the exponential distribution $f(x) = (1/\theta)e^{-x/\theta}, x \geq 0$, $\theta > 0$, find the maximum likelihood estimator of θ and obtain the asymptotic distribution of this estimator.
3. *Mixture distribution.* Suppose that the joint distribution of the two random variables x and y is

$$f(x, y) = \frac{\theta e^{-(\beta+\theta)y}(\beta y)^x}{x!}, \quad \beta, \theta > 0, y \geq 0, x = 0, 1, 2, \ldots.$$

   a. Find the maximum likelihood estimators of β and θ and their asymptotic joint distribution.
   b. Find the maximum likelihood estimator of $\theta/(\beta + \theta)$ and its asymptotic distribution.
   c. Prove that $f(x)$ is of the form

$$f(x) = \gamma(1 - \gamma)^x, x = 0, 1, 2, \ldots,$$

   and find the maximum likelihood estimator of γ and its asymptotic distribution.
   d. Prove that $f(y \mid x)$ is of the form

$$f(y \mid x) = \frac{\lambda e^{-\lambda y}(\lambda y)^x}{x!}, \quad y \geq 0, \lambda > 0.$$

   Prove that $f(y \mid x)$ integrates to 1. Find the maximum likelihood estimator of λ and its asymptotic distribution. [Hint: In the conditional distribution, just carry the xs along as constants.]
   e. Prove that

$$f(y) = \theta e^{-\theta y}, \quad y \geq 0, \quad \theta > 0.$$

   Find the maximum likelihood estimator of θ and its asymptotic variance.
   f. Prove that

$$f(x \mid y) = \frac{e^{-\beta y}(\beta y)^x}{x!}, \quad x = 0, 1, 2, \ldots, \beta > 0.$$

   Based on this distribution, what is the maximum likelihood estimator of β?
4. Suppose that x has the Weibull distribution

$$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}, \quad x \geq 0, \alpha, \beta > 0.$$

   a. Obtain the log-likelihood function for a random sample of n observations.
   b. Obtain the likelihood equations for maximum likelihood estimation of α and β. Note that the first provides an explicit solution for α in terms of the data and β. But, after inserting this in the second, we obtain only an implicit solution for β. How would you obtain the maximum likelihood estimators?

c. Obtain the second derivatives matrix of the log-likelihood with respect to α and β. The exact expectations of the elements involving β involve the derivatives of the gamma function and are quite messy analytically. Of course, your exact result provides an empirical estimator. How would you estimate the asymptotic covariance matrix for your estimators in Part b?
d. Prove that $\alpha\beta\text{Cov}[\ln x, x^\beta] = 1$. [Hint: The expected first derivatives of the log-likelihood function are zero.]

5. The following data were generated by the Weibull distribution of Exercise 4:

1.3043	0.49254	1.2742	1.4019	0.32556	0.29965	0.26423
1.0878	1.9461	0.47615	3.6454	0.15344	1.2357	0.96381
0.33453	1.1227	2.0296	1.2797	0.96080	2.0070	

a. Obtain the maximum likelihood estimates of α and β, and estimate the asymptotic covariance matrix for the estimates.
b. Carry out a Wald test of the hypothesis that $\beta = 1$.
c. Obtain the maximum likelihood estimate of α under the hypothesis that $\beta = 1$.
d. Using the results of Parts a and c, carry out a likelihood ratio test of the hypothesis that $\beta = 1$.
e. Carry out a Lagrange multiplier test of the hypothesis that $\beta = 1$.

6. (**Limited Information Maximum Likelihood Estimation**). Consider a bivariate distribution for x and y that is a function of two parameters, α and β. The joint density is $f(x, y \mid \alpha, \beta)$. We consider maximum likelihood estimation of the two parameters. The full information maximum likelihood estimator is the now familiar maximum likelihood estimator of the two parameters. Now, suppose that we can factor the joint distribution as done in Exercise 3, but in this case, we have $f(x, y \mid \alpha, \beta) = f(y \mid x, \alpha, \beta) f(x \mid \alpha)$. That is, the conditional density for y is a function of both parameters, but the marginal distribution for x involves only α.
a. Write down the general form for the log likelihood function using the joint density.
b. Since the joint density equals the product of the conditional times the marginal, the log-likelihood function can be written equivalently in terms of the factored density. Write this down, in general terms.
c. The parameter α can be estimated by itself using only the data on x and the log likelihood formed using the marginal density for x. It can also be estimated with β by using the full log-likelihood function and data on both y and x. Show this.
d. Show that the first estimator in Part c has a larger asymptotic variance than the second one. This is the difference between a limited information maximum likelihood estimator and a full information maximum likelihood estimator.
e. Show that if $\partial^2 \ln f(y \mid x, \alpha, \beta)/\partial\alpha\partial\beta = 0$, then the result in Part d is no longer true.

7. Show that the likelihood inequality in Theorem 17.3 holds for the Poisson distribution used in Section 17.3 by showing that $E[(1/n) \ln L(\theta \mid y)]$ is uniquely maximized at $\theta = \theta_0$. Hint: First show that the expectation is $-\theta + \theta_0 \ln \theta - E_0[\ln y_i!]$.
8. Show that the likelihood inequality in Theorem 17.3 holds for the normal distribution.
9. For random sampling from the classical regression model in (17-3), reparameterize the likelihood function in terms of $\eta = 1/\sigma$ and $\boldsymbol{\delta} = (1/\sigma)\boldsymbol{\beta}$. Find the maximum

likelihood estimators of η and $\boldsymbol{\delta}$ and obtain the asymptotic covariance matrix of the estimators of these parameters.

10. Section 14.3.1 presents estimates of a Cobb–Douglas cost function using Nerlove's 1955 data on the U.S. electric power industry. Christensen and Greene's 1976 update of this study used 1970 data for this industry. The Christensen and Greene data are given in Table F5.2. These data have provided a standard test data set for estimating different forms of production and cost functions, including the stochastic frontier model examined in Example 17.5. It has been suggested that one explanation for the apparent finding of economies of scale in these data is that the smaller firms were inefficient for other reasons. The stochastic frontier might allow one to disentangle these effects. Use these data to fit a frontier cost function which includes a quadratic term in log output in addition to the linear term and the factor prices. Then examine the estimated Jondrow et al. residuals to see if they do indeed vary negatively with output, as suggested. (This will require either some programming on your part or specialized software. The stochastic frontier model is provided as an option in TSP and LIMDEP. Or, the likelihood function can be programmed fairly easily for RATS or GAUSS. Note, for a cost frontier as opposed to a production frontier, it is necessary to reverse the sign on the argument in the Φ function.)
11. Consider, sampling from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_M)$ and covariance matrix $\sigma^2\mathbf{I}$. The log-likelihood function is

$$\ln L = \frac{-nM}{2}\ln(2\pi) - \frac{nM}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})'(\mathbf{y}_i - \boldsymbol{\mu}).$$

Show that the maximum likelihood estimates of the parameters are

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{\sum_{i=1}^{n}\sum_{m=1}^{M}(y_{im} - \bar{y}_m)^2}{nM} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}(y_{im} - \bar{y}_m)^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}^2_m.$$

Derive the second derivatives matrix and show that the asymptotic covariance matrix for the maximum likelihood estimators is

$$\left\{-E\left[\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right]\right\}^{-1} = \begin{bmatrix}\sigma^2\mathbf{I}/n & \mathbf{0}\\ \mathbf{0} & 2\sigma^4/(nM)\end{bmatrix}.$$

Suppose that we wished to test the hypothesis that the means of the M distributions were all equal to a particular value μ^0. Show that the Wald statistic would be

$$\mathbf{W} = (\bar{\mathbf{y}} - \mu^0\mathbf{i})'\left(\frac{\hat{\sigma}^2}{n}\mathbf{I}\right)^{-1}(\bar{\mathbf{y}} - \mu^0\mathbf{i}), = \left(\frac{n}{s^2}\right)(\bar{\mathbf{y}} - \mu^0\mathbf{i})'(\bar{\mathbf{y}} - \mu^0\mathbf{i}),$$

where $\bar{\mathbf{y}}$ is the vector of sample means.