MAXIMUM LIKELIHOOD ESTIMATION

16.1 INTRODUCTION

The generalized method of moments discussed in Chapter 15 and the semiparametric, nonparametric, and Bayesian estimators discussed in Chapters 14 and 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 16.2 through 16.6 present basic statistical results for estimation and hypothesis testing based on the maximum likelihood principle. Sections 16.7 and 16.8 present two extensions of the method, two-step estimation and pseudo maximum likelihood estimation. After establishing the general results for this method of estimation, we will then apply them to the more familiar setting of econometric models. The applications presented in Section 16.9 apply the maximum likelihood method to most of the models in the preceding chapters and several others that illustrate different uses of the technique.

16.2 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, θ, is denoted $f(y|\theta)$.[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* (i.i.d.) observations from this process is the product of the individual densities;

$$f(y_1, \ldots, y_n | \theta) = \prod_{i=1}^{n} f(y_i | \theta) = L(\theta | \mathbf{y}). \tag{16-1}$$

This joint density is the **likelihood function**, defined as a function of the unknown parameter vector, θ, 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 will write this function in reverse,

[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.

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 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 Chapter 18. In this classical estimation framework, the parameters are assumed to be fixed constants that we hope to learn about from the data.

AW: Confirm x-ref to Chap. 18

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

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}). \tag{16-2}$$

Again, to emphasize our interest in the parameters, given the observed data, we denote this function $L(\boldsymbol{\theta} \mid \text{data}) = L(\boldsymbol{\theta} \mid \mathbf{y})$. 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 its 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 i.i.d.; they have different means. Nonetheless, the observations are independent, and as we will examine in closer detail,

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}[\ln \sigma^2 + \ln(2\pi) + (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma^2], \tag{16-3}$$

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 16.1 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 16.1 Identification of Parameters

For the regression model specified in (16-3), 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 (16-3) 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 because $\boldsymbol{\beta}$ 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 23.) Consider a simple form of the regression model considered earlier, $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, 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.

16.3 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

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

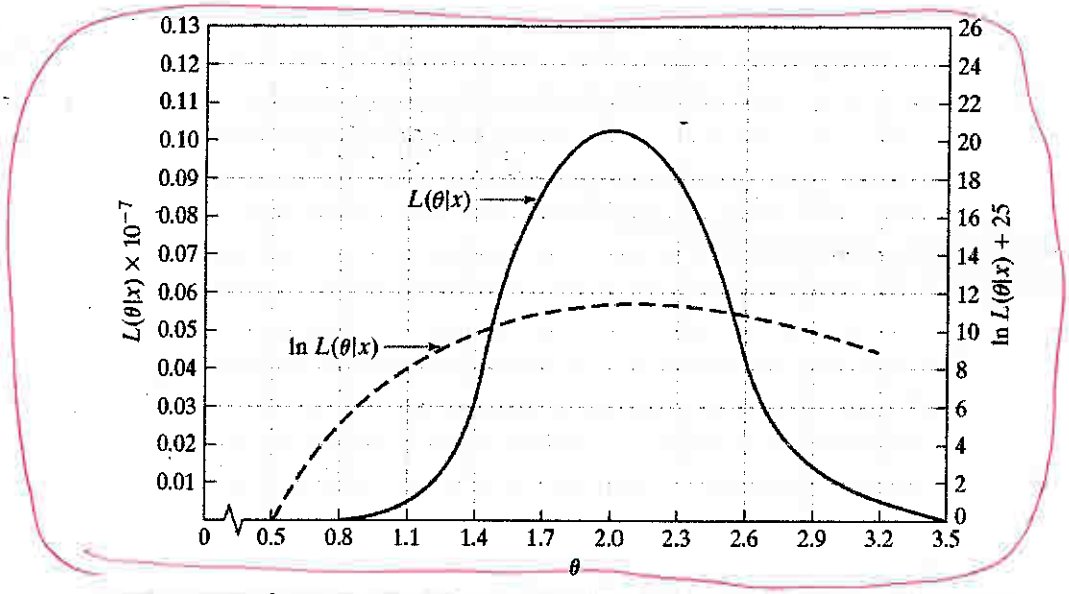


FIGURE 16.1 Likelihood and Log-Likelihood Functions for a Poisson Distribution.

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

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

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 16.1 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. Because 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,

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

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

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 16.1 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, because a particular sample has probability zero. Nonetheless, the principle is the same. The values of the parameters that maximize $L(\theta \mid \mathbf{data})$ or its log are the maximum likelihood estimates, denoted $\hat{\theta}$. The logarithm is a monotonic function, so the values that maximize $L(\theta \mid \mathbf{data})$ are the same as those that maximize $\ln L(\theta \mid \mathbf{data})$. The necessary condition for maximizing $\ln L(\theta \mid \mathbf{data})$ is

$$\frac{\partial \ln L(\theta \mid \mathbf{data})}{\partial \theta} = 0. \tag{16-4}$$

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 16.2 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

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - \mu)^2}{\sigma^2}\right], \tag{16-5}$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0, \tag{16-6}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0. \tag{16-7}$$

To solve the likelihood equations, multiply (16-6) by σ^2 and solve for $\hat{\mu}$, then insert this solution in (16-7) and solve for σ^2. The solutions are

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_n \quad \text{and} \quad \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2. \tag{16-8}$$

16.4 PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

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

DEFINITION 16.2 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 16.2 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 13.1), 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 momentarily are met by $f(\mathbf{x}, \boldsymbol{\theta}_0)$, then we have the following theorem.

THEOREM 16.1 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.*

16.4.1 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 $f(y_i \mid \theta_0)$ 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 (2004).]

DEFINITION 16.3 Regularity Conditions

R1. *The first three derivatives of* $\ln f(y_i \mid \theta)$ *with respect to* θ *are continuous and finite for almost all* y_i *and for all* θ. *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* $\ln f(y_i \mid \theta)$ *are met.*

R3. *For all values of* θ, $|\partial^3 \ln f(y_i \mid \theta)/\partial\theta_j\partial\theta_k\partial\theta_l|$ *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 $f(y_i \mid \theta)$: 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 that equates the gradient to zero. On the other, D2 is a more fundamental relationship that 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.

16.4.2 PROPERTIES OF REGULAR DENSITIES

Densities that are "regular" by Definition 16.3 have three properties that are used in establishing the properties of maximum likelihood estimators:

THEOREM 16.2 Moments of the Derivatives of the Log-Likelihood

D1. $\ln f(y_i \mid \theta)$, $\mathbf{g}_i = \partial \ln f(y_i \mid \theta)/\partial\theta$, *and* $\mathbf{H}_i = \partial^2 \ln f(y_i \mid \theta)/\partial\theta\partial\theta'$, $i = 1, \ldots, n$, *are all random samples of random variables. This statement follows from our assumption of random sampling. The notation* $\mathbf{g}_i(\theta_0)$ *and* $\mathbf{H}_i(\theta_0)$ *indicates the derivative evaluated at* θ_0.

D2. $E_0[\mathbf{g}_i(\theta_0)] = \mathbf{0}$.

D3. $\text{Var}[\mathbf{g}_i(\theta_0)] = -E[\mathbf{H}_i(\theta_0)]$.

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; $A(\theta_0) \leq y_i \leq B(\theta_0)$. (Consider, for example, finding the maximum likelihood estimator of θ_0

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,

$$\int_{A(\theta_0)}^{B(\theta_0)} f(y_i \mid \theta_0)\, dy_i = 1.$$

Now, differentiate this expression with respect to θ_0. Leibnitz's theorem gives

$$\begin{aligned}\frac{\partial \int_{A(\theta_0)}^{B(\theta_0)} f(y_i \mid \theta_0)\, dy_i}{\partial \theta_0} &= \int_{A(\theta_0)}^{B(\theta_0)} \frac{\partial f(y_i \mid \theta_0)}{\partial \theta_0}\, dy_i + f(B(\theta_0) \mid \theta_0)\frac{\partial B(\theta_0)}{\partial \theta_0} \\ &\quad - f(A(\theta_0) \mid \theta_0)\frac{\partial A(\theta_0)}{\partial \theta_0} \\ &= \mathbf{0}.\end{aligned}$$

If the second and third terms go to zero, then we may interchange the operations of differentiation and integration. The necessary condition is that $\lim_{y_i \to A(\theta_0)} f(y_i \mid \theta_0) = \lim_{y_i \to B(\theta_0)} f(y_i \mid \theta_0) = 0$. (Note that the uniform distribution suggested earlier 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 $\partial A(\theta_0)/\partial \theta_0 = \partial B(\theta_0)/\partial \theta_0 = \mathbf{0}$ 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,

$$\begin{aligned}\frac{\partial \int f(y_i \mid \theta_0)\, dy_i}{\partial \theta_0} &= \int \frac{\partial f(y_i \mid \theta_0)}{\partial \theta_0}\, dy_i = \int \frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0} f(y_i \mid \theta_0)\, dy_i \\ &= E_0\left[\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0}\right] = \mathbf{0}.\end{aligned}$$

This proves D2.

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

$$\int \left[\frac{\partial^2 \ln f(y_i \mid \theta_0)}{\partial \theta_0 \partial \theta_0'} f(y_i \mid \theta_0) + \frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0}\frac{\partial f(y_i \mid \theta_0)}{\partial \theta_0'}\right] dy_i = \mathbf{0}.$$

But

$$\frac{\partial f(y_i \mid \theta_0)}{\partial \theta_0'} = f(y_i \mid \theta_0)\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0'},$$

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

$$-\int \left[\frac{\partial^2 \ln f(y_i \mid \theta_0)}{\partial \theta_0 \partial \theta_0'}\right] f(y_i \mid \theta_0)\, dy_i = \int \left[\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0}\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0'}\right] f(y_i \mid \theta_0)\, dy_i.$$

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, because 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:

$$\mathrm{Var}_0\left[\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0}\right] = E_0\left[\left(\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0}\right)\left(\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta_0'}\right)\right] = -E\left[\frac{\partial^2 \ln f(y_i \mid \theta_0)}{\partial \theta_0 \partial \theta_0'}\right].$$

14-9

16.4.3 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

$$\mathbf{g} = \frac{\partial \ln L(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{g}_i. \tag{16-9}$$

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

$$E_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] = E_0[\mathbf{g}_0] = \mathbf{0}. \tag{16-10}$$

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

Au: Term "likelihood equation" was a KT on m>p 14-5. Here also?

16.4.4 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

$$\begin{aligned}\mathrm{Var}_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] &= E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0'}\right)\right] \\ &= -E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right].\end{aligned} \tag{16-11}$$

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

Au: Term "information matrix equality" was KT on m>p 14-5. Here also?

16.4.5 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.

16.4.5.a Consistency

We assume that $f(\mathbf{y}_i \mid \theta_0)$ is a possibly multivariate density that at this point does not depend on covariates, $\mathbf{x}_i$. Thus, this is the i.i.d., random sampling case. Because $\hat{\theta}$ is the MLE, in any finite sample, for any $\theta \neq \hat{\theta}$ (including the true θ_0) it must be true that

$$\ln L(\hat{\theta}) \geq \ln L(\theta). \tag{16-12}$$

Consider, then, the random variable $L(\theta)/L(\theta_0)$. Because the log function is strictly concave, from Jensen's Inequality (Theorem D.13.), we have

$$E_0\left[\ln \frac{L(\theta)}{L(\theta_0)}\right] < \ln E_0\left[\frac{L(\theta)}{L(\theta_0)}\right]. \tag{16-13}$$

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

$$E_0\left[\frac{L(\theta)}{L(\theta_0)}\right] = \int \left(\frac{L(\theta)}{L(\theta_0)}\right) L(\theta_0)\, d\mathbf{y} = 1 \tag{16-14}$$

is simply the integral of a joint density. Now, take logs on both sides of (16-13), insert the result of (16-14), then divide by n to produce

$$E_0[1/n \ln L(\theta)] - E_0[1/n \ln L(\theta_0)] < 0.$$

This produces a central result:

THEOREM 16.3 Likelihood Inequality

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

This result is (16-15).

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

For any θ, including $\hat{\theta}$,

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

is the sample mean of n i.i.d. random variables with expectation $E_0[(1/n) \ln L(\theta)]$. Because the sampling is i.i.d. by the regularity conditions, we can invoke the Khinchine theorem, D.5; the sample mean converges in probability to the population mean. Using $\theta = \hat{\theta}$, it follows from Theorem 16.3 that as $n \to \infty$, $\lim \text{Prob}[[(1/n) \ln L(\hat{\theta})] < [(1/n) \ln L(\theta_0)]] = 1$ if $\hat{\theta} \neq \theta_0$. But, $\hat{\theta}$ is the MLE, so for every n, $(1/n) \ln L(\hat{\theta}) \geq (1/n) \ln L(\theta_0)$. 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 $(1/n) \ln L(\hat{\theta}) \to (1/n) \ln L(\theta_0)$ imply that $\hat{\theta} \to \theta_0$? 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

14-11

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 i.i.d. 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, $\text{plim}\,\hat{\theta} = \theta_0$.

14

16.4.5.b Asymptotic Normality

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

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

(This is the sample statistic, not the expectation.) Expand this set of equations in a Taylor series around the true parameters θ_0. We will use the mean value theorem to truncate the Taylor series at the second term,

$$\mathbf{g}(\hat{\theta}) = \mathbf{g}(\theta_0) + \mathbf{H}(\bar{\theta})(\hat{\theta} - \theta_0) = \mathbf{0}.$$

The Hessian is evaluated at a point $\bar{\theta}$ that is between $\hat{\theta}$ and θ_0 [$\bar{\theta} = w\hat{\theta} + (1 - w)\theta_0$ for some $0 < w < 1$]. We then rearrange this function and multiply the result by $\sqrt{n}$ to obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) = [-\mathbf{H}(\bar{\theta})]^{-1}[\sqrt{n}\mathbf{g}(\theta_0)].$$

Because $\text{plim}(\hat{\theta} - \theta_0) = \mathbf{0}$, $\text{plim}(\hat{\theta} - \bar{\theta}) = \mathbf{0}$ as well. The second derivatives are continuous functions. Therefore, if the limiting distribution exists, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} [-\mathbf{H}(\theta_0)]^{-1}[\sqrt{n}\mathbf{g}(\theta_0)].$$

By dividing $\mathbf{H}(\theta_0)$ and $\mathbf{g}(\theta_0)$ by n, we obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \left[-\tfrac{1}{n}\mathbf{H}(\theta_0)\right]^{-1}[\sqrt{n}\,\bar{\mathbf{g}}(\theta_0)]. \qquad (14\text{-}15)$$

We may apply the Lindeberg–Levy central limit theorem (D.18) to $[\sqrt{n}\,\bar{\mathbf{g}}(\theta_0)]$, because it is $\sqrt{n}$ times the mean of a random sample; we have invoked D1 again. The limiting variance of $[\sqrt{n}\,\bar{\mathbf{g}}(\theta_0)]$ is $-E_0[(1/n)\mathbf{H}(\theta_0)]$, so

$$\sqrt{n}\,\bar{\mathbf{g}}(\theta_0) \xrightarrow{d} N\{\mathbf{0}, -E_0[\tfrac{1}{n}\mathbf{H}(\theta_0)]\}.$$

By virtue of Theorem D.2, $\text{plim}[-(1/n)\mathbf{H}(\theta_0)] = -E_0[(1/n)\mathbf{H}(\theta_0)]$. This result is a constant matrix, so we can combine results to obtain

$$\left[-\tfrac{1}{n}\mathbf{H}(\theta_0)\right]^{-1}\sqrt{n}\,\bar{\mathbf{g}}(\theta_0) \xrightarrow{d} N[\mathbf{0}, \{-E_0[\tfrac{1}{n}\mathbf{H}(\theta_0)]\}^{-1}\{-E_0[\tfrac{1}{n}\mathbf{H}(\theta_0)]\}\{-E_0[\tfrac{1}{n}\mathbf{H}(\theta_0)]\}^{-1}],$$

or

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\tfrac{1}{n}\mathbf{H}(\theta_0)\right]\right\}^{-1}\right],$$

which gives the asymptotic distribution of the MLE:

$$\hat{\theta} \overset{a}{\sim} N[\theta_0, \{\mathbf{I}(\theta_0)\}^{-1}].$$

This last step completes M2.

14

Example 16.3 Information Matrix for the Normal Distribution

For the likelihood function in Example 16.2, the second derivatives are

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

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

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

KT

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 $E[y_i] = \mu$. That leaves the second, which you can verify has expectation $-n/(2\sigma^4)$ because each of the n terms $(y_i - \mu)^2$ 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:

$$\left\{-E_0\left[\frac{\partial^2 \ln L}{\partial \theta_0 \, \partial \theta_0'}\right]\right\}^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}.$$

14

16.4.5.c Asymptotic Efficiency

Theorem C.2 provides the lower bound for the variance of an unbiased estimator. Because 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):

14

THEOREM 16.4 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 θ_0 will always be at least as large as

$$[\mathbf{I}(\theta_0)]^{-1} = \left(-E_0\left[\frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \, \partial \theta_0'}\right]\right)^{-1} = \left(E_0\left[\left(\frac{\partial \ln L(\theta_0)}{\partial \theta_0}\right)\left(\frac{\partial \ln L(\theta_0)}{\partial \theta_0}\right)'\right]\right)^{-1}.$$

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

16.4.5.d Invariance

Last, 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 θ. 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 (2004, pp. 446, 539–540). 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 24.3.3.) Suppose that the normal log-likelihood in Example 16.2 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 $\hat{\theta}$ will, itself, be the MLE.

16.4.5.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.

16.4.6 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 (i.e., it is a function of the θ_0 that is being estimated). If the form of the expected values of the second derivatives of the

[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.

log-likelihood is known, then

$$[\mathbf{I}(\theta_0)]^{-1} = \left\{ -E_0 \left[\frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \, \partial \theta_0'} \right] \right\}^{-1} \tag{16-16}$$

can be evaluated at $\hat{\theta}$ 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

$$[\hat{\mathbf{I}}(\hat{\theta})]^{-1} = \left(-\frac{\partial^2 \ln L(\hat{\theta})}{\partial \hat{\theta} \, \partial \hat{\theta}'} \right)^{-1}. \tag{16-17}$$

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 16.2, that the expected second derivatives matrix is the covariance matrix of the first derivatives vector, is

$$[\hat{\hat{\mathbf{I}}}(\hat{\theta})]^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right]^{-1} = [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1}, \tag{16-18}$$

where

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

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 (16-18) 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 following example 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 (16-16) is usually unavailable, as the exact expectation of

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

the Hessian is rarely known. Available evidence suggests that in small or moderate-sized samples, (16-17) (the Hessian) is preferable.

Example 16.4 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

$$\frac{\partial \ln L(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \frac{1}{\beta + x_i} + \sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^2} = 0, \qquad (16\text{-}19)$$

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

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \frac{1}{(\beta + x_i)^2} - 2\sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^3}. \qquad (16\text{-}20)$$

Because the function $E(y_i) = \beta + x_i$ is known, the exact form of the expected value in (16-20) is known. Inserting $\hat{\beta} + x_i$ for y_i in (16-20) and taking the negative of the reciprocal yields the first variance estimate, 44.2546. Simply inserting $\hat{\beta} = 15.602727$ in (16-20) 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 $\hat{\beta}$,

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

we obtain the BHHH estimate, 100.5116.

16.5 CONDITIONAL LIKELIHOODS, ECONOMETRIC MODELS, AND THE GMM ESTIMATOR

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, $f(y_i \mid \alpha)$. 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 4, 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

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

14-16

where any nonstochastic elements in $\mathbf{x}_i$ such as a time trend or dummy variable are being carried as constants. 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 $g(\mathbf{x}_i \mid \alpha)$ do not overlap with those that appear in $f(y_i \mid \mathbf{x}_i, \alpha)$. Thus, we partition α into $[\theta, \delta]$ so that the log-likelihood function may be written

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

As long as θ and δ 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 $\ln f(y_i \mid \mathbf{x}_i, \theta)$. We will proceed under the assumption of well-behaved data so that sample averages such as

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

and its gradient with respect to θ 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:

advanced

- **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. 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 16.1 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, because we will not encounter dependent observations specifically until later in the book. The definitions in Chapter 4 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 4 and follows precisely along the lines of Section 14.5.

12

16.6 HYPOTHESIS AND SPECIFICATION TESTS AND FIT MEASURES

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 16.2.[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:

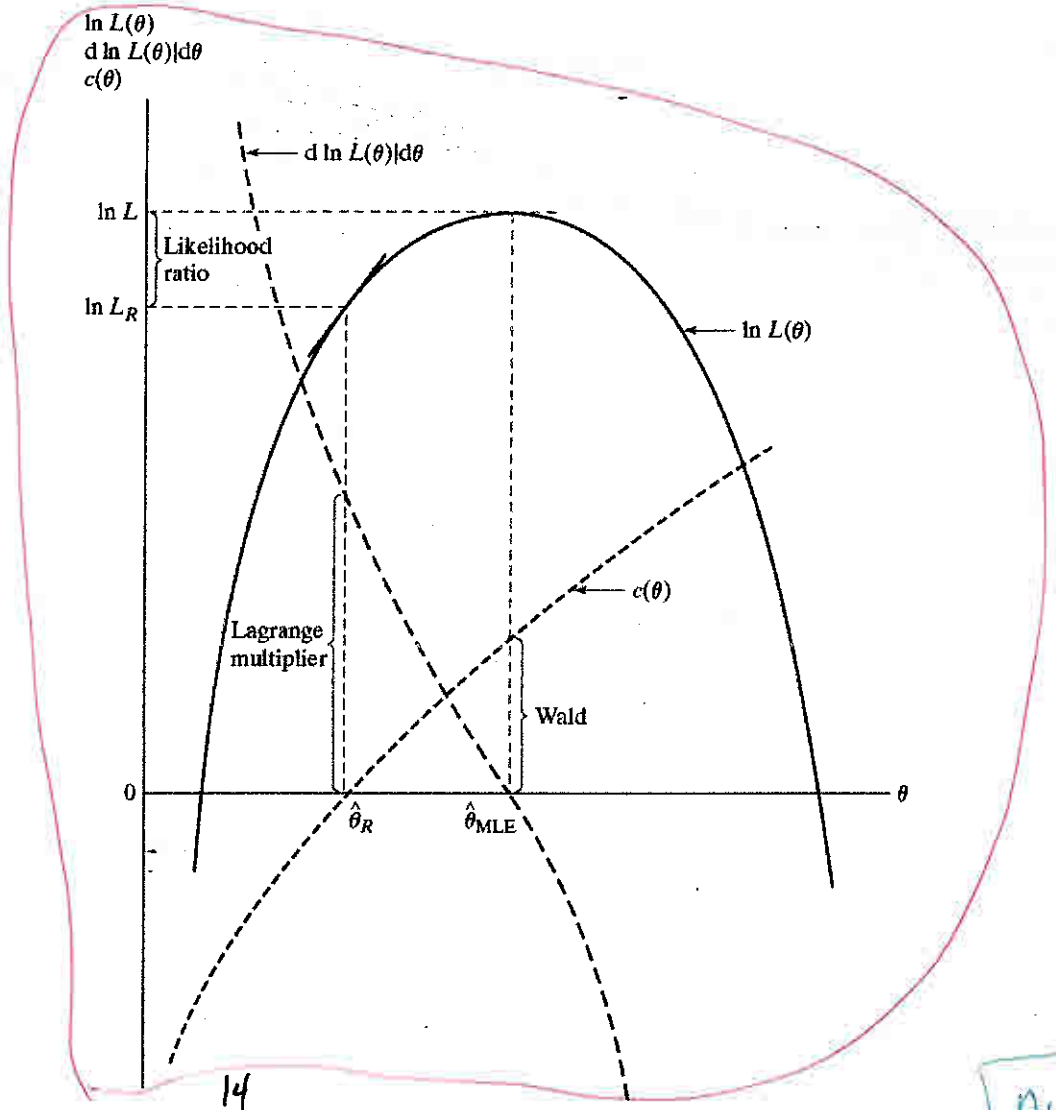
- **Likelihood ratio test.** If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference, $\ln L_U - \ln L_R$, where L_U is the value of the likelihood function at the unconstrained value 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 because 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.

16.6.1 THE LIKELIHOOD RATIO TEST

Let θ be a vector of parameters to be estimated, and let H_0 specify some sort of restriction on these parameters. Let $\hat{\theta}_U$ be the maximum likelihood estimator of θ obtained without regard to the constraints, and let $\hat{\theta}_R$ 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.



14

FIGURE 16.2 Three Bases for Hypothesis Tests.

RT

likelihood ratio is

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}. \qquad (16\text{-}21)$$

14

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 16.3, we found the

14

Aus Term "likelihood ratio" is not in the chap. list

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^{-7}. Are these data consistent with $H_0: \theta = 1.8$? $L_R = 0.936 \times 10^{-8}$, 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 16.5 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.

16.6.2 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.

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

$$\text{If } \mathbf{x} \sim N_J[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \text{chi-squared}[J]. \tag{16-22}$$

[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.

In the setting of a hypothesis test, under the hypothesis that $E(\mathbf{x}) = \mu$, the quadratic form has the chi-squared distribution. If the hypothesis that $E(\mathbf{x}) = \mu$ 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 $\hat{\theta}$ be the vector of parameter estimates obtained without restrictions. We hypothesize a set of restrictions

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

If the restrictions are valid, then at least approximately $\hat{\theta}$ should satisfy them. If the hypothesis is erroneous, however, then $\mathbf{c}(\hat{\theta}) - \mathbf{q}$ 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 16.6 Limiting Distribution of the Wald Test Statistic

The Wald statistic is

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

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 $\mathbf{c}(\hat{\theta}) - \mathbf{q} = \mathbf{0}$]. A derivation of the limiting distribution of the Wald statistic appears in Theorem 5.1.

This test is analogous to the chi-squared statistic in (16-22) if $\mathbf{c}(\hat{\theta}) - \mathbf{q}$ 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.

$$\text{Est. Asy. Var}[\mathbf{c}(\hat{\theta}) - \mathbf{q}] = \hat{\mathbf{C}}\,\text{Est. Asy. Var}[\hat{\theta}]\hat{\mathbf{C}}', \qquad \hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\theta})}{\partial \hat{\theta}'}\right]. \tag{16-23}$$

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 θ. A common application occurs in testing a set of linear restrictions.

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

$$H_0: \mathbf{c}(\theta) - \mathbf{q} = \mathbf{R}\theta - \mathbf{q} = \mathbf{0}, \qquad \hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\theta})}{\partial \hat{\theta}'}\right] = \mathbf{R}', \tag{16-24}$$

$$\text{Est. Asy. Var}[\mathbf{c}(\hat{\theta}) - \mathbf{q}] = \mathbf{R}\,\text{Est. Asy. Var}[\hat{\theta}]\mathbf{R}',$$

[8] If the mean is not μ, then the statistic in (16-22) 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.

and

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

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

If $\mathbf{c}(\theta) = \mathbf{q}$ 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

$$z = \frac{|\hat{\theta} - \theta_0|}{s(\hat{\theta})}, \tag{16-25}$$

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

$$W = [(\hat{\theta} - \theta_0) - 0]\big(\text{Asy. Var}[(\hat{\theta} - \theta_0) - 0]\big)^{-1}[(\hat{\theta} - \theta_0) - 0] = \frac{(\hat{\theta} - \theta_0)^2}{\text{Asy. Var}[\hat{\theta}]} = z^2. \tag{16-26}$$

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 (16-25).

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.

16.6.3 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 $\mathbf{c}(\theta) - \mathbf{q} = \mathbf{0}$. Let $\boldsymbol{\lambda}$ be a vector of Lagrange multipliers and define the Lagrangean function

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

The solution to the constrained maximization problem is the root of

$$\frac{\partial \ln L^*}{\partial \boldsymbol{\theta}} = \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0},$$
$$\frac{\partial \ln L^*}{\partial \boldsymbol{\lambda}} = \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}, \quad (16\text{-}27)$$

where $\mathbf{C}'$ is the transpose of the derivatives matrix in the second line of (16-23). 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

$$\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R} = -\hat{\mathbf{C}}'\hat{\boldsymbol{\lambda}} = \hat{\mathbf{g}}_R. \quad (16\text{-}28)$$

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.** Because 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.

AU: Terms "efficient scores" & "score test" were KTs on p. 14-21. Here also?

THEOREM 16.7 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 $\hat{\mathbf{g}}_{iR}'$ and $\mathbf{i}$ is a column of 1s. If we use the BHHH (outer product of gradients) estimator in (16-18) to estimate the Hessian,

then

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

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, because $\mathbf{i}'\mathbf{i}$ equals n, $\text{LM} = n(\mathbf{i}'\hat{\mathbf{G}}_R[\hat{\mathbf{G}}_R'\hat{\mathbf{G}}_R]^{-1}\hat{\mathbf{G}}_R'\mathbf{i}/n) = nR_i^2$, 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.

16.6.4 AN APPLICATION OF THE LIKELIHOOD-BASED TEST PROCEDURES

Consider, again, the data in Example C.1. In Example 16.4, the parameter β in the model

$$f(y_i \mid x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)} \qquad (16\text{-}29)$$

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,

$$f(y_i \mid x_i, \beta, \rho) = \frac{\beta_i^\rho}{\Gamma(\rho)} y_i^{\rho-1} e^{-y_i\beta_i}. \qquad (16\text{-}30)$$

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

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

$$\frac{\partial \ln L}{\partial \beta} = -\rho \sum_{i=1}^n \beta_i + \sum_{i=1}^n y_i \beta_i^2, \qquad \frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^n \ln \beta_i - n\Psi(\rho) + \sum_{i=1}^n \ln y_i, \qquad (16\text{-}31)$$

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

[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 16.1. Three estimators are available for the asymptotic covariance matrix of the estimators of $\theta = (\beta, \rho)'$. Using the actual Hessian as in (16-17), we compute $\mathbf{V} = [-\Sigma_i \partial^2 \ln f(y_i \mid x_i, \beta, \rho)/\partial\theta\partial\theta']^{-1}$ at the maximum likelihood estimates. For this model, it is easy to show that

[9] The gamma function $\Gamma(\rho)$ and the gamma distribution are described in Sections B.4.5 and E2.3.

TABLE 16.1 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 V are given in parentheses.

replace table entries. values on next page MS p 14-25

$E[y_i \mid x_i] = \rho(\beta + x_i)$ (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 (16-16) to compute $\mathbf{V}_E = \{-\Sigma_i E[\partial^2 \ln f(y_i \mid x_i, \beta, \rho)/\partial\theta\partial\theta']\}^{-1}$. Finally, by using the sums of squares and cross products of the first derivatives, we obtain the BHHH estimator in (16-18), $\mathbf{V}_B = [\Sigma_i(\partial \ln f(y_i \mid x_i, \beta, \rho)/\partial\theta)(\partial \ln f(y_i \mid x_i, \beta, \rho)/\partial\theta')]^{-1}$. Results in Table 16.1 are based on $\mathbf{V}$.

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

$$\mathbf{V} = \begin{bmatrix} 5.497 & -1.652 \\ -1.652 & 0.6309 \end{bmatrix}, \quad \mathbf{V}_E = \begin{bmatrix} 4.897 & -1.473 \\ -1.473 & 0.5770 \end{bmatrix}, \quad \mathbf{V}_B = \begin{bmatrix} 13.33 & -4.314 \\ -4.314 & 1.535 \end{bmatrix}.$$

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 $3.1517 \pm 1.96\sqrt{0.6309} = [1.5942, 4.7092]$. 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. The computed value is larger than this critical value, so the hypothesis is again rejected.
- **Wald test:** The Wald test is based on the unrestricted estimates. For this restriction, $c(\theta) - q = \rho - 1$, $dc(\hat{\rho})/d\hat{\rho} = 1$, 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, $|3.1517 - 1|/\sqrt{0.6309} = 2.70895$. 2.73335
- **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

$$\text{LM} = [0.0000 \quad 7.9162] \begin{bmatrix} 0.009938 & 0.26762 \\ 0.26762 & 11.197 \end{bmatrix}^{-1} \begin{bmatrix} 0.0000 \\ 7.9162 \end{bmatrix} = 15.681.$$

Insert in TB 14.1 on msp 14-24 where indicated

14-25

-4.7185	(2.345)	15.6027	(6.794)
3.1509	(0.794)	1.0000	(0.000)
-82.91605		-88.43626	
0.0000		0.0000	
0.0000		7.9145	
-0.85570		-0.02166	
-7.4592		-32.8987	
-2.2420		0.66891	

end of insert