14

MAXIMUM LIKELIHOOD ESTIMATION

14.1 INTRODUCTION

The generalized method of moments discussed in Chapter 13 and the semiparametric, nonparametric, and Bayesian estimators discussed in Chapters 12 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 14.2 through 14.6 present basic statistical results for estimation and hypothesis testing based on the maximum likelihood principle. Sections 14.7 and 14.8 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 Sections 14.9 and 14.10 apply the maximum likelihood method to most of the models in the preceding chapters and several others that illustrate different uses of the technique.

14.2 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* (i.i.d.) observations from this process is the product of the individual densities,

$$f(y_1, \ldots, y_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid \mathbf{y}). \quad \textbf{(14-1)}$$

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 will 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 highlight our interest in the parameters and the information about them that is contained in the

[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.

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

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

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}). \tag{14-2}$$

Again, to emphasize our interest in the parameters, given the observed data, we denote this function $L(\boldsymbol{\theta} \mid \textbf{data}) = L(\boldsymbol{\theta} \mid \mathbf{y})$. 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,

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n} [\ln \sigma^2 + \ln(2\pi) + (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma^2], \tag{14-3}$$

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 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 14.1 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 14.1 Identification of Parameters

For the regression model specified in (14-3), 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 (14-3) 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 17.) Consider a simple form of the regression model considered earlier, $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, where $\varepsilon_i | x_i$ has a normal distribution with zero mean and variance σ^2. To put the model in a context, consider a consumer's purchase 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^* (their reservation price) 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

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

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

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

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 unchanged, and the likelihood function does not change. This model requires a normalization. The one usually used is $\sigma = 1$, but some authors have used $\beta_1 = 1$ or $\beta_2 = 1$, instead.[2]

14.3 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

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

[2]For examples, see Horowitz (1993) and Lewbel (2014).

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

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

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

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

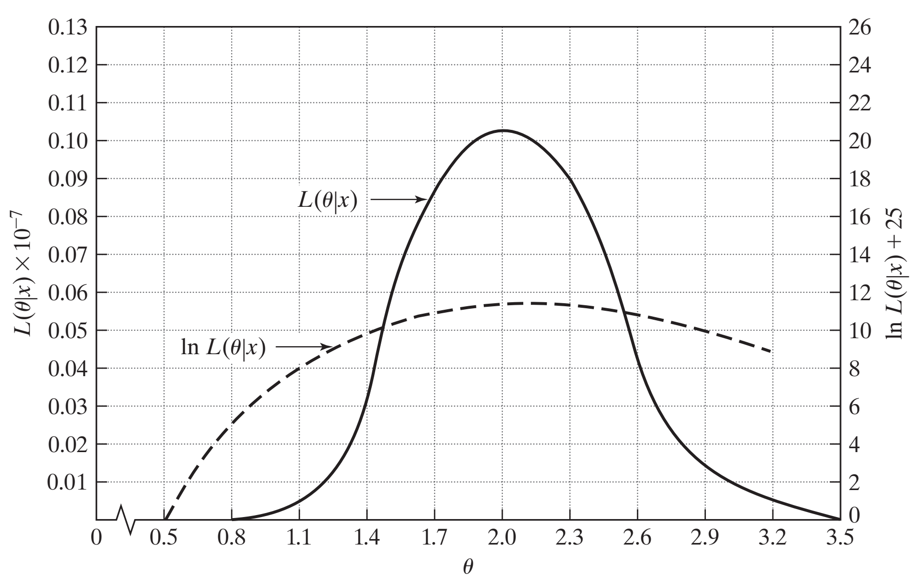
$$\frac{\partial \ln L(\theta|\mathbf{y})}{\partial\theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} y_i = 0 \Rightarrow \hat{\theta}_{ML} = \bar{y}_n.$$

For the assumed sample of observations,

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

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

FIGURE 14.1 Likelihood and Log-Likelihood Functions for a Poisson Distribution.



and

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

The solution is the same as before. Figure 14.1 also plots the log of $L(\theta|\mathbf{y})$ 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 $L(\boldsymbol{\theta}|\mathbf{data})$ or its log are the maximum likelihood estimates, denoted $\hat{\boldsymbol{\theta}}$. The logarithm is a monotonic function, so the values that maximize $L(\boldsymbol{\theta}|\mathbf{data})$ are the same as those that maximize $\ln L(\boldsymbol{\theta}|\mathbf{data})$. The necessary condition for maximizing $\ln L(\boldsymbol{\theta}|\mathbf{data})$ is

$$\frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{data})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{14-4}$$

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 data-generating process 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 14.2 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

$$\ln L(\mu, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\left[\ln(2\pi) + \ln \sigma^2 + \frac{(y_i - \mu)^2}{\sigma^2}\right], \tag{14-5}$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0, \tag{14-6}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0. \tag{14-7}$$

To solve the likelihood equations, multiply (14-6) by σ^2 and solve for $\hat{\mu}$, then insert this solution in (14-7) and solve for σ^2. The solutions are

$$\hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_n \quad \text{and} \quad \hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n).^2 \tag{14-8}$$

14.4 PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

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

DEFINITION 14.2 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.[3]

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 14.2 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.[4] Most applications in econometrics do not involve exponential families, so the appeal of the MLE remains primarily based on 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 14.1 Properties of an MLE
Under regularity, the MLE has the following asymptotic properties:

M1. Consistency: $\text{plim}\, \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

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.*

14.4.1 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 with density function $f(y_i \mid \boldsymbol{\theta}_0)$ and that the following **regularity conditions** hold.[5]

[3]*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.

[4]See Stuart and Ord (1989).

[5]Our statement of these is informal. A more rigorous treatment may be found in Stuart and Ord (1989) or Davidson and MacKinnon (2004).

DEFINITION 14.3 Regularity Conditions

R1. *The first three derivatives of* $\ln f(y_i|\boldsymbol{\theta})$ *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 to 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* $\ln f(y_i|\boldsymbol{\theta})$ *are met.*

R3. *For all values of* $\boldsymbol{\theta}$, $|\partial^3 \ln f(y_i|\boldsymbol{\theta})/\partial\theta_j\partial\theta_k\partial\theta_l|$ *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 $f(y_i|\boldsymbol{\theta})$: 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 that equates the gradient to zero. On the other hand, 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.

14.4.2 PROPERTIES OF REGULAR DENSITIES

Densities that are *regular* by Definition 14.3 have three properties that are used in establishing the properties of maximum likelihood estimators:

THEOREM 14.2 Moments of the Derivatives of the Log Likelihood

D1. $\ln f(y_i|\boldsymbol{\theta})$, $\mathbf{g}_i = \partial \ln f(y_i|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, *and* $\mathbf{H}_i = \partial^2 \ln f(y_i|\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, $i = 1, \ldots, n$, *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* $\boldsymbol{\theta}_0$. *Condition D1 is simply a consequence of the definition of the density.*

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

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

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_0 for a continuous uniform distribution with range $[0, \theta_0]$.) (In the following, the single integral $\int \ldots dy_i$ will be used to indicate the multiple integration over all the elements of a multivariate of y_i if that is necessary.) By definition,

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

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

$$\frac{\partial \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} f(y_i|\boldsymbol{\theta}_0)\,dy_i}{\partial \boldsymbol{\theta}_0} = \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} \frac{\partial f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\,dy_i + f(B(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)\frac{\partial B(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} - f(A(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)\frac{\partial A(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} = \mathbf{0}.$$

If the second and third terms go to zero, then we may interchange the operations of differentiation and integration. The necessary condition is that $\lim_{y_i \downarrow A(\boldsymbol{\theta}_0)} f(y_i|\boldsymbol{\theta}_0) = \lim_{y_i \uparrow B(\boldsymbol{\theta}_0)} f(y_i|\boldsymbol{\theta}_0) = 0$. (*Note:* 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 $\partial A(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \partial B(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \mathbf{0}$ 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,

$$\frac{\partial \int f(y_i|\boldsymbol{\theta}_0)\,dy_i}{\partial \boldsymbol{\theta}_0} = \int \frac{\partial f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\,dy_i = \int \frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} f(y_i|\boldsymbol{\theta}_0)\,dy_i$$

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

This proves D2.

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

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

But

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

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

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

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 **0** (we just showed this), the right-hand side is the variance of the first derivative vector, which proves D3,

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

14.4.3 THE LIKELIHOOD EQUATION

The log-likelihood function is

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

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

$$\mathbf{g} = \frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{g}_i. \quad \textbf{(14-9)}$$

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

$$E_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] = E_0[\mathbf{g}_0] = \mathbf{0}, \quad \textbf{(14-10)}$$

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

14.4.4 THE INFORMATION MATRIX EQUALITY

The Hessian of the log likelihood is

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

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

$$\text{Var}_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] = E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0'}\right)\right] = -E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right]. \quad \textbf{(14-11)}$$

This very useful result is known as the **information matrix equality**. It states that the variance of the first derivative of ln L equals the negative of the second derivative.

14.4.5 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.

14.4.5.a Consistency

We assume that $f(y_i|\boldsymbol{\theta}_0)$ 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 $\hat{\boldsymbol{\theta}}$ 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

$$\ln L(\hat{\boldsymbol{\theta}}) \geq \ln L(\boldsymbol{\theta}). \quad \textbf{(14-12)}$$

Consider, then, the random variable $L(\boldsymbol{\theta})/L(\boldsymbol{\theta}_0)$. Because the log function is strictly concave, from Jensen's Inequality (Theorem D.13.), we have

$$E_0\left[\ln \frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right] < \ln E_0\left[\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right]. \quad \textbf{(14-13)}$$

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

$$E_0\left[\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right] = \int\left(\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right)L(\boldsymbol{\theta}_0)d\mathbf{y} = 1 \quad \textbf{(14-14)}$$

is simply the integral of a joint density. So, the right-hand side of (14-13) equals zero. Divide the left-hand side of (14-13) by n to produce

$$E_0[1/n \ln L(\boldsymbol{\theta})] - E_0[1/n \ln L(\boldsymbol{\theta}_0)] < 0.$$

This produces a central result:

THEOREM 14.3 Likelihood Inequality

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

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

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

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

is the sample mean of n i.i.d. random variables, with expectation $E_0[(1/n) \ln L(\boldsymbol{\theta})]$. 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 $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, it follows from Theorem 14.3 that as $n \to \infty$, lim 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 converge to their 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, plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

14.4.5.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 Taylor series around the true parameters $\boldsymbol{\theta}_0$. We will use the mean value theorem to truncate the Taylor series for each element of $\mathbf{g}(\hat{\boldsymbol{\theta}})$ 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 K rows of the Hessian are each evaluated at a point $\bar{\boldsymbol{\theta}}_k$ that is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ $[\bar{\boldsymbol{\theta}}_k = w_k\hat{\boldsymbol{\theta}} + (1 - w_k)\boldsymbol{\theta}_0$ for some $0 < w_k < 1]$. (Although the vectors $\bar{\boldsymbol{\theta}}_k$ are different, they all converge to $\boldsymbol{\theta}_0$.) 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 plim$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}$, plim$(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = \mathbf{0}$ 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

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \left[-\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1}[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)]. \tag{14-15}$$

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

$$[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)] \text{ is } -E_0[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)], \text{ so}$$

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

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

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

or

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

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 14.3 Information Matrix for the Normal Distribution

For the likelihood function in Example 14.2, the second derivatives are

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

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,

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

14.4.5.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):

THEOREM 14.4 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, asymptotically normally distributed estimator, so this completes the argument.[6]

14.4.5.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

[6]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.

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 (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 19.3.3.) Suppose that the normal log likelihood in Example 14.2 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.

14.4.5.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.

14.4.6 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 $\boldsymbol{\theta}_0$ that is being estimated). If the form of the expected values of the second derivatives of the log likelihood is known, then

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left\{-E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0\,\partial\boldsymbol{\theta}_0'}\right]\right\}^{-1} \quad \textbf{(14-16)}$$

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

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = \left(-\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\,\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}. \quad \textbf{(14-17)}$$

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

$$[\hat{\hat{\mathbf{I}}}(\hat{\boldsymbol{\theta}})]^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'\right]^{-1} = [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1}, \tag{14-18}$$

where $\hat{\mathbf{g}}_i = \dfrac{\partial \ln f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}$, and $\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_n]'$ 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 (14-18) is known as the **BHHH estimator**[7] and the **outer product of gradients estimator (OPG)**.

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 (14-16) is usually unavailable, as the exact expectation of the Hessian is rarely known. Available evidence suggests that in small or moderate-sized samples, (14-17) (the Hessian) is preferable.

Example 14.4 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

$$\frac{\partial \ln L(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \frac{1}{\beta + x_i} + \sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^2} = 0, \tag{14-19}$$

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

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

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \frac{1}{(\beta + x_i)^2} - 2\sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^3}. \qquad \textbf{(14-20)}$$

Because the function $E(y_i) = \beta + x_i$ is known, the exact form of the expected value in (14-20) is known. Inserting $\hat{\beta} + x_i$ for y_i in (14-20) and taking the negative of the reciprocal yields the first variance estimate, 44.2546. Simply inserting $\hat{\beta} = 15.602727$ in (14-20) 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.

14.5 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, $f(y_i | \boldsymbol{\alpha})$. 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 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

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

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 $g(\mathbf{x}_i | \boldsymbol{\alpha})$ do not overlap with those that appear in $f(y_i | \mathbf{x}_i, \boldsymbol{\alpha})$. Thus, we partition $\boldsymbol{\alpha}$ into $[\boldsymbol{\theta}, \boldsymbol{\delta}]$ so that the log-likelihood function may be written

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

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

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

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 advanced 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. 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 14.1 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 12.5.

14.6 HYPOTHESIS AND SPECIFICATION TESTS AND FIT MEASURES

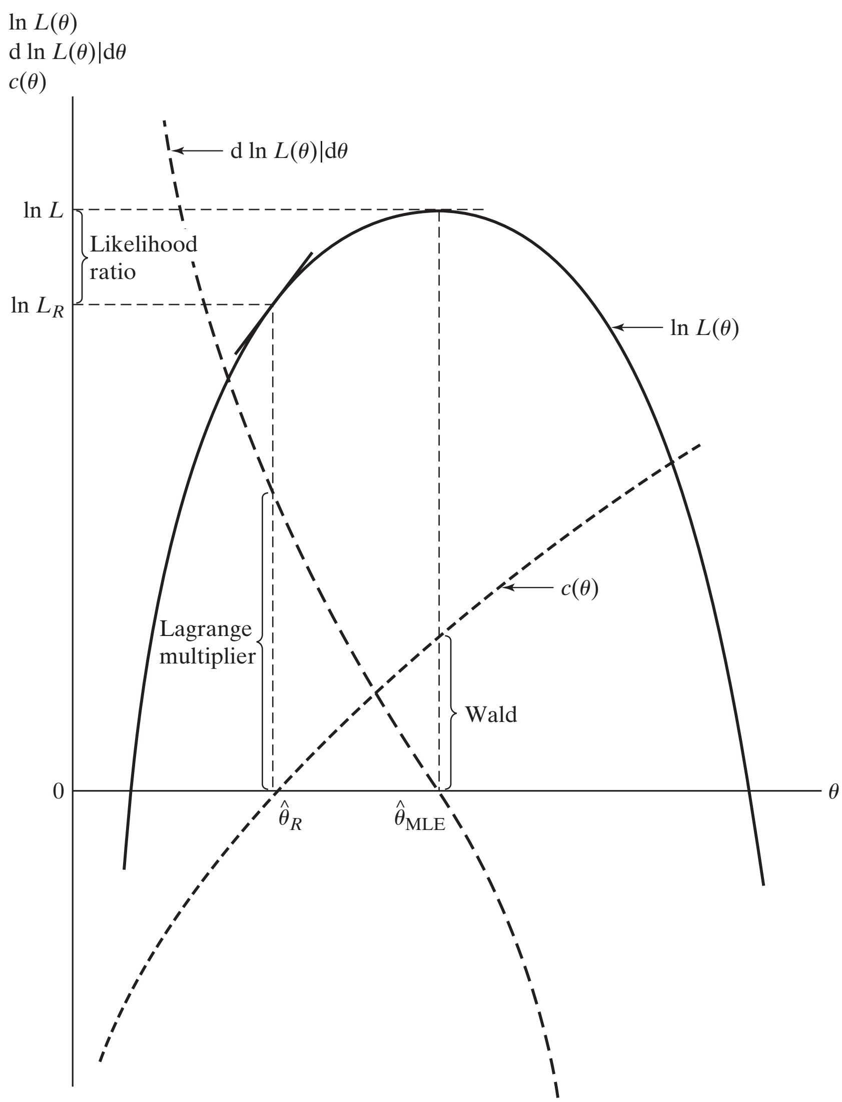
The next several sections will discuss the most commonly used test procedures: the likelihood ratio, Wald, and Lagrange multiplier tests.[8] We consider maximum likelihood estimation of a parameter θ and a test of the hypothesis H_0: $c(\theta) = 0$. The logic of the tests can be seen in Figure 14.2.[9] 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.

[8]Extensive discussion of these procedures is given in Godfrey (1988).

[9]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 14.2 Three Bases for Hypothesis Tests.



- **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.

14.6.1 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 **likelihood ratio** is

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}. \tag{14-21}$$

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 fix these ideas. In estimating from a sample of 10 from a Poisson population at the beginning of Section 14.3, 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^{-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 14.5 Limiting Distribution of the Likelihood Ratio Test Statistic
Under regularity and under H_0, the limiting 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.[10]

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.[11] 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.

14.6.2 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,

$$\text{If } \mathbf{x} \sim N_J[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \text{chi-squared}[J]. \qquad \textbf{(14-22)}$$

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.[12] 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.

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

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

[12] If the mean is not μ, then the statistic in (14-22) 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.

THEOREM 14.6 Limiting Distribution of the Wald Test Statistic
The Wald statistic is

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

Under H_0, W has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions [i.e., the number of equations in $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q} = 0$].
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 (14-22) 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.

$$\text{Est.Asy.Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] = \hat{\mathbf{C}}\,\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}]\hat{\mathbf{C}}',$$
$$\hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}\right]. \tag{14-23}$$

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.

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

$$H_0\colon \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{R}\boldsymbol{\theta} - \mathbf{q} = \mathbf{0},$$
$$\hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}\right] = \mathbf{R}, \tag{14-24}$$
$$\text{Est.Asy.Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] = \mathbf{R}\,\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}]\mathbf{R},$$

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

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

then the earlier test is based on

$$z = \frac{|\hat{\theta} - \theta_0|}{s(\hat{\theta})}, \tag{14-25}$$

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

$$W = [(\hat{\theta} - \theta_0) - 0](\text{Asy.Var}[(\hat{\theta} - \theta_0) - 0])^{-1}[(\hat{\theta} - \theta_0) - 0] = \frac{(\hat{\theta} - \theta_0)^2}{\text{Asy.Var}[\hat{\theta}]} = z^2. \quad \textbf{(14-26)}$$

Here W has a limiting chi-squared distribution with one degree of freedom, which is the distribution of the square of the standard normal test statistic in (14-25).

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.

14.6.3 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 $\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}$. 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 joint solution of

$$\frac{\partial \ln L^*}{\partial \boldsymbol{\theta}} = \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0},$$
$$\frac{\partial \ln L^*}{\partial \boldsymbol{\lambda}} = \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}, \quad \textbf{(14-27)}$$

where $\mathbf{C}'$ is the transpose of the derivatives matrix in the second line of (14-23). 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

$$\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R} = -\mathbf{C}'\hat{\boldsymbol{\lambda}} = \hat{\mathbf{g}}_R. \quad \textbf{(14-28)}$$

If the restrictions are valid, at least within 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.

THEOREM 14.7 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 (14-18) 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, because $\mathbf{i}'\mathbf{i}$ equals n, $\text{LM} = n(\mathbf{i}'\hat{G}_R[\hat{\mathbf{G}}_R'\hat{\mathbf{G}}_R]^{-1}\hat{\mathbf{G}}_R'\mathbf{i}/n) = nR_i^2$, 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.

14.6.4 AN APPLICATION OF THE LIKELIHOOD-BASED TEST PROCEDURES

Consider, again, the data in Example C.1. In Example 14.4, the parameter β in the model

$$f(y_i \mid x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)} \quad \textbf{(14-29)}$$

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

$$f(y_i \mid x_i, \beta, \rho) = \frac{\alpha_i^{\rho}}{\Gamma(\rho)} y_i^{\rho - 1} e^{-y_i \alpha_i}. \quad \textbf{(14-30)}$$

The restriction is $\rho = 1$.[13] 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

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

$$\frac{\partial \ln L}{\partial \beta} = -\rho \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} y_i \alpha_i^2, \qquad \frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^{n} \ln \alpha_i - n\Psi(\rho) + \sum_{i=1}^{n} \ln y_i, \quad \textbf{(14-31)}$$

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

[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 14.1. Three estimators are available for the asymptotic covariance matrix of the estimators of $\boldsymbol{\theta} = (\beta, \rho)'$. Using the actual Hessian as in (14-17), we compute $\mathbf{V} = [-\Sigma_i \partial^2 \ln f(y_i | x_i, \beta, \rho)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']^{-1}$ at the maximum likelihood estimates. For this model, it is easy to show that $E[y_i | x_i] = \rho(\beta + x_i)$ (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 (14-16) to compute $\mathbf{V}_E = \{-\Sigma_i E[\partial^2 \ln f(y_i | x_i, \beta, \rho)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']\}^{-1}$. Finally, by using the sums of squares and cross products of the first derivatives, we obtain the BHHH estimator in (14-18), $\mathbf{V}_B = [\Sigma_i(\partial \ln f(y_i | x_i, \beta, \rho)/\partial\boldsymbol{\theta})(\partial \ln f(y_i | x_i, \beta, \rho)/\partial\boldsymbol{\theta}')]^{-1}$. Results in Table 14.1 are based on $\mathbf{V}$.

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

$$\mathbf{V} = \begin{bmatrix} 5.499 & -1.653 \\ -1.653 & 0.6309 \end{bmatrix}, \quad \mathbf{V}_E = \begin{bmatrix} 4.900 & -1.473 \\ -1.473 & 0.5768 \end{bmatrix}, \quad \mathbf{V}_B = \begin{bmatrix} 13.370 & -4.322 \\ -4.322 & 1.537 \end{bmatrix}.$$

TABLE 14.1 Maximum Likelihood Estimates

Quantity	***Unrestricted Estimate***[a]	***Restricted Estimate***
β	−4.7185 (2.345)	15.6027 (6.794)
ρ	3.1509 (0.794)	1.0000 (0.000)
$\ln L$	−82.91605	−88.4363
$\partial \ln L/\partial\beta$	0.0000	0.0000
$\partial \ln L/\partial\rho$	0.0000	7.9145
$\partial^2 \ln L/\partial\beta^2$	−0.8557	−0.0217
$\partial^2 \ln L/\partial\rho^2$	−7.4592	−32.8987
$\partial^2 \ln L/\partial\beta\partial\rho$	−2.2420	−0.66891

[a]Estimated asymptotic standard errors based on **V** are given in parentheses.

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

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% confidence interval for ρ based on the unrestricted estimates is $3.1509 \pm 1.96\sqrt{0.6309} = [1.5941, 4.7076]$. This interval does not contain $\rho = 1$, so the hypothesis is rejected.
- **Likelihood ratio test:** The LR statistic is $\lambda = -2[-88.43626 - (-82.91604)] = 11.0404$. 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, $c(\boldsymbol{\theta}) - q = \rho - 1$, $dc(\hat{\rho})/d\hat{\rho} = 1$, $\text{Est.Asy.Var}[c(\hat{\rho}) - q] = \text{Est.Asy.Var}[\hat{\rho}] = 0.6309$, 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, $|3.1509 - 1|/\sqrt{0.6309} = 2.73335$.
- **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

$$\text{LM} = [0.0000 \quad 7.9145]\begin{bmatrix} 0.00995 & 0.26776 \\ 0.26776 & 11.199 \end{bmatrix}^{-1}\begin{bmatrix} 0.0000 \\ 7.9145 \end{bmatrix} = 15.687.$$

The conclusion is the same as before. Note that the same computation done using V rather than $\mathbf{V}_B$ produces a value of 5.1162. 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% level of significance instead of 5% 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.[14] 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.

14.6.5 COMPARING MODELS AND COMPUTING MODEL FIT

The test statistics described in Sections 14.6.1–14.6.3 are available for assessing the validity of restrictions on the parameters in a model. When the models are nested, any of the three mentioned testing procedures can be used. For nonnested models, the computation is a comparison of one model to another based on an estimation criterion to discern which is to be preferred. Two common measures that are based on the same logic as the adjusted R-squared for the linear model are

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

Akaike information criterion (AIC) $= -2 \ln L + 2K,$
Bayes (Schwarz) information criterion (BIC) $= -2 \ln L + K \ln n,$

where K is the number of parameters in the model. Choosing a model based on the lowest AIC is logically the same as using $\overline{R}^2$ in the linear model, nonstatistical, albeit widely accepted.

The AIC and BIC are information criteria, not fit measures as such. This does leave open the question of how to assess the "fit" of the model. Only the case of a linear least squares regression in a model with a constant term produces an R^2, which measures the proportion of variation explained by the regression. The ambiguity in R^2 as a fit measure arose immediately when we moved from the linear regression model to the generalized regression model in Chapter 9. The problem is yet more acute in the context of the models we consider in this chapter. For example, the estimators of the models for count data in Example 14.10 make no use of the "variation" in the dependent variable and there is no obvious measure of "explained variation."

A measure of fit that was originally proposed for discrete choice models in McFadden (1974), but surprisingly has gained wide currency throughout the empirical literature is the **likelihood ratio index**, which has come to be known as the **Pseudo R^2**. It is computed as

$$\text{Pseudo } R^2 = 1 - (\ln L)/(\ln L_0),$$

where $\ln L$ is the log likelihood for the model estimated and $\ln L_0$ is the log likelihood for the same model with only a constant term. The statistic does resemble the R^2 in a linear regression. The choice of name for this statistic is unfortunate, however, because even in the discrete choice context for which it was proposed, it has no connection to the fit of the model to the data. In discrete choice settings in which log likelihoods must be negative, the pseudo R^2 must be between zero and one and rises as variables are added to the model. It can obviously be zero, but is usually bounded below one. In the linear model with normally distributed disturbances, the maximized log likelihood is

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

With a small amount of manipulation, we find that the pseudo R^2 for the linear regression model is

$$\text{Pseudo } R^2 = \frac{-\ln(1 - R^2)}{1 + \ln 2\pi + \ln s_y^2},$$

while the *true* R^2 is $1 - \mathbf{e}'\mathbf{e}/\mathbf{e}_0'\mathbf{e}_0$. Because s_y^2 can vary independently of R^2—multiplying $\mathbf{y}$ by any scalar, A, leaves R^2 unchanged but multiplies s_y^2 by A^2—although the upper limit is one, there is no lower limit on this measure. It can even be negative. This same problem arises in any model that uses information on the scale of a dependent variable, such as the tobit model (Chapter 19). The computation makes even less sense as a fit measure in multinomial models such as the ordered probit model (Chapter 18) or the multinomial logit model. For discrete choice models, a variety of such measures are discussed in Chapter 17. For limited dependent variable and many loglinear models, some other measure that is related to a correlation between a prediction and the actual value would be more useable. Nonetheless, the measure has gained currency in the contemporary literature.[15] Notwithstanding the general contempt for the likelihood ratio index, practitioners are often interested in comparing models based on some idea of the fit of the model the data. Constructing such a measure will be specific to the context, so we will return to the issue in the discussion of specific applications such as the binary choice in Chapter 17.

14.6.6 VUONG'S TEST AND THE KULLBACK–LEIBLER INFORMATION CRITERION

Vuong's (1989) approach to testing **nonnested models** is also based on the likelihood ratio statistic. The logic of the test is similar to that which motivates the likelihood ratio test in general. Suppose that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ and $g(y_i|\mathbf{Z}_i, \boldsymbol{\gamma})$ are two competing models for the density of the random variable y_i, with f being the null model, H_0, and g being the alternative, H_1. For instance, in Example 5.7, both densities are (by assumption now) normal, y_i is consumption, C_t, $\mathbf{Z}_i$ is $[1, Y_t, Y_{t-1}, C_{t-1}]$, $\boldsymbol{\theta}$ is $(\beta_1, \beta_2, \beta_3, 0, \sigma^2)$, $\boldsymbol{\gamma}$ is $(\gamma_1, \gamma_2, 0, \gamma_3, \omega^2)$, and σ^2 and ω^2 are the respective conditional variances of the disturbances, ε_{0t} and ε_{1t}. The crucial element of Vuong's analysis is that it need not be the case that either competing model is *true*; they may both be incorrect. What we want to do is attempt to use the data to determine which competitor is closer to the truth, that is, closer to the correct (unknown) model.

We assume that observations in the sample (disturbances) are conditionally independent. Let $L_{i,0}$ denote the ith contribution to the likelihood function under the null hypothesis. Thus, the log-likelihood function under the null hypothesis is $\Sigma_i \ln L_{i,0}$. Define $L_{i,1}$ likewise for the alternative model. Now, let m_i equal $\ln L_{i,1} - \ln L_{i,0}$. If we were using the familiar likelihood ratio test, then, the likelihood ratio statistic would be simply $LR = 2\Sigma_i m_i = 2n\overline{m}$ when $L_{i,0}$ and $L_{i,1}$ are computed at the respective maximum likelihood estimators. When the competing models are nested—H_0 is a restriction on H_1—we know that $\Sigma_i m_i \geq 0$. The restrictions of the null hypothesis will never increase the likelihood function. (In the linear regression model with normally distributed disturbances that we have examined so far, the log likelihood and these results are all based on the sum of squared residuals. And, as we have seen, imposing restrictions never reduces the sum of squares.) The limiting distribution of the LR statistic under the assumption of the null hypothesis is chi squared with degrees of freedom equal to the reduction in the number of dimensions of the parameter space of the alternative hypothesis that results from imposing the restrictions.

Vuong's analysis is concerned with nonnested models for which $\Sigma_i m_i$ need not be positive. Formalizing the test requires us to look more closely at what is meant by the *right* model (and provides a convenient departure point for the discussion in the next two sections). In the context of nonnested models, Vuong allows for the possibility that neither model is *true* in the absolute sense. We maintain the classical assumption that there does exist a true model, $h(y_i|\mathbf{Z}_i, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is the true parameter vector, but possibly neither hypothesized model is that true model. The **Kullback–Leibler Information Criterion (KLIC)** measures the distance between the true model (distribution) and a

[15]The software package *Stata* reports the pseudo R^2 with every model fit by MLE, but at the same time, admonishes its users not to interpret it as anything meaningful. See, for example, www.stata.com/support/faqs/stat/pseudor2.html. Cameron and Trivedi (2005) document the pseudo R^2 at length and then give similar cautions about it and urge their readers to seek a more meaningful measure of the correlation between model predictions and the outcome variable of interest. Wooldridge (2010, p. 575) dismisses it summarily, and argues that partial effects are more important.

hypothesized model in terms of the likelihood function. Loosely, the KLIC is the log-likelihood function under the hypothesis of the true model minus the log-likelihood function for the (misspecified) hypothesized model under the assumption of the true model. Formally, for the model of the null hypothesis,

$$\text{KLIC} = E[\ln h(y_i|\mathbf{Z}_i, \boldsymbol{\alpha})|h \text{ is true}] - E[\ln f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})|h \text{ is true}].$$

The first term on the right-hand side is what we would estimate with $(1/n) \ln L$ if we maximized the log likelihood for the true model, $h(y_i|\mathbf{Z}_i, \boldsymbol{\alpha})$. The second term is what is estimated by $(1/n) \ln L$ assuming (incorrectly) that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ is the correct model. Notice that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ is written in terms of a parameter vector, $\boldsymbol{\theta}$. Because $\boldsymbol{\alpha}$ is the true parameter vector, it is perhaps ambiguous what is meant by the parameterization, $\boldsymbol{\theta}$. Vuong (p. 310) calls this the "pseudotrue" parameter vector. It is the vector of constants that the estimator converges to when one uses the estimator implied by $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$. In Example 5.7, if H_0 gives the correct model, this formulation assumes that the least squares estimator in H_1 would converge to some vector of pseudo-true parameters. But these are not the parameters of the correct model—they would be the slopes in the population linear projection of C_t on $[1, Y_t, C_{t-1}]$.

Suppose the true model is $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, with normally distributed disturbances and $\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \mathbf{w}$ is the proposed competing model. The KLIC would be the expected log-likelihood function for the true model minus the expected log-likelihood function for the second model, still assuming that the first one is the truth. By construction, the KLIC is positive. We will now say that one model is better than another if it is closer to the truth based on the KLIC. If we take the difference of the two KLICs for two models, the true log-likelihood function falls out, and we are left with

$$\text{KLIC}_1 - \text{KLIC}_0 = E[\ln f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})|h \text{ is true}] - E[\ln g(y_i|\mathbf{Z}_i, \boldsymbol{\gamma})|h \text{ is true}].$$

To compute this using a sample, we would simply compute the likelihood ratio statistic, $n\overline{m}$ (without multiplying by 2) again. Thus, this provides an interpretation of the LR statistic. But, in this context, the statistic can be negative—we don't know which competing model is closer to the truth.

Vuong's general result for nonnested models (his Theorem 5.1) describes the behavior of the statistic

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} = \sqrt{n}(\overline{m}/s_m), \quad m_i = \ln L_{i,1} - \ln L_{i,0}.$$

He finds:

1. Under the hypothesis that the models are "equivalent," $V \xrightarrow{D} N[0,1]$.
2. Under the hypothesis that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ is "better," $V \xrightarrow{A.S.} +\infty$.
3. Under the hypothesis that $g(y_i|\mathbf{Z}_i, \boldsymbol{\gamma})$ is "better," $V \xrightarrow{A.S.} -\infty$.

This test is directional. Large positive values favor the null model while large negative values favor the alternative. The intermediate values (e.g., between −1.96 and +1.96 for 95% significance) are an inconclusive region. An application appears in Example 14.8.

14.7 TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION

The applied literature contains a large and increasing number of applications in which elements of one model are embedded in another, which produces what are known as "two-step" estimation problems.[16] There are two parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The first appears in the second model, but the second does not appear in the first model. 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,

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

A two-step procedure for this kind of model could be used by estimating the parameters of model 1 first by maximizing

$$\ln L_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \ln f_1(y_{i1} \mid \mathbf{x}_{i1}, \boldsymbol{\theta}_1)$$

and then maximizing the marginal likelihood function for y_2 while embedding the consistent estimator of $\boldsymbol{\theta}_1$, treating it as given. The second step involves maximizing

$$\ln L_2(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \ln f_2(y_{i2} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2).$$

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.[17] The results given here can be found in an important reference on the subject, Murphy and Topel (2002, first published in 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 $\hat{\mathbf{V}}_1$ be n times any of the estimators of the asymptotic covariance matrix of this estimator that were discussed in Section 14.4.6.
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 $\hat{\mathbf{V}}_2$ be n times any appropriate estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_2$.

[16] Among the best known of these is Heckman's (1979) model of sample selection discussed in Example 1.1 and in Chapter 19.

[17] 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.

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 because plim $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1$, asymptotically, this line of reasoning is correct. (See point 3 of Theorem D.16.) But the same line of reasoning is not sufficient to justify using $(1/n)\mathbf{V}_2$ 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:

THEOREM 14.8 Asymptotic Distribution of the Two-Step MLE [Murphy and Topel (2002)]

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

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

where

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

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

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

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

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

A derivation of this useful result is instructive. We will rely on (14-11) and the results of Section 14.4.5.B where the asymptotic normality of the maximum likelihood estimator is developed. The first-step MLE of $\boldsymbol{\theta}_1$ is defined by

$$\frac{1}{n}\frac{\partial \ln L_1(\hat{\boldsymbol{\theta}}_1)}{\hat{\boldsymbol{\theta}}_1} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f_1(y_{i1}|\mathbf{x}_{i1}, \hat{\boldsymbol{\theta}}_1)}{\partial\hat{\boldsymbol{\theta}}_1}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_{i1}(\hat{\boldsymbol{\theta}}_1) = \bar{\mathbf{g}}_1(\hat{\boldsymbol{\theta}}_1) = \mathbf{0}.$$

Using the results in that section, we obtained the asymptotic distribution from (14-15),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} [-\mathbf{H}_{11}^{(1)}(\theta_1)]^{-1}\sqrt{n}\bar{\mathbf{g}}_1(\theta_1),$$

where the expression means that the limiting distribution of the two random vectors is the same,

and

$$\mathbf{H}_{11}^{(1)} = E\left[\frac{1}{n}\frac{\partial^2 \ln L_1(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1'}\right].$$

The second-step MLE of $\boldsymbol{\theta}_2$ is defined by

$$\frac{1}{n}\frac{\partial \ln L_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial\hat{\boldsymbol{\theta}}_2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f_2(y_{i2}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial\hat{\boldsymbol{\theta}}_2}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_{i2}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \hat{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \mathbf{0}.$$

Expand the derivative vector, $\bar{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$, in a linear Taylor series as usual, and use the results in Section 14.4.5.b once again,

$$\bar{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \left[\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right](\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) + \left[\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right](\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + o(1/n) = \mathbf{0},$$

where

$$\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_1'}\right] \text{ and } \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2'}\right].$$

To obtain the asymptotic distribution, we use the same device as before,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \xrightarrow{d} \left[-\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]^{-1}\sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \theta_2)$$
$$+ \left[-\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]^{-1}\left[\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1).$$

For convenience, denote $\mathbf{H}_{22}^{(2)} = \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\mathbf{H}_{21}^{(2)} = \mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $\mathbf{H}_{11}^{(1)} = \mathbf{H}_{11}^{(1)}(\boldsymbol{\theta}_1)$. Now substitute the first-step estimator of $\boldsymbol{\theta}_1$ in this expression to obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \xrightarrow{d} \left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\sqrt{n}\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1).$$

Consistency and asymptotic normality of the two estimators follow from our earlier results. To obtain the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_2$ we will obtain the limiting variance of the random vector in the preceding expression. The joint normal distribution of the two first derivative vectors has zero means and

$$\text{Var}\begin{bmatrix}\sqrt{n}\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1)\\ \sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1)\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}.$$

Then, the asymptotic covariance matrix we seek is

$$\begin{aligned}\text{Var}\big[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)\big] = {} & \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\boldsymbol{\Sigma}_{22}\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ & + \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\boldsymbol{\Sigma}_{11}\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ & + \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\boldsymbol{\Sigma}_{21}\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ & + \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\boldsymbol{\Sigma}_{12}\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}.\end{aligned}$$

As we found earlier, the variance of the first derivative vector of the log likelihood is the negative of the expected second derivative matrix [see (14-11)]. Therefore $\boldsymbol{\Sigma}_{22} = [-\mathbf{H}_{22}^{(2)}]$ and $\boldsymbol{\Sigma}_{11} = [-\mathbf{H}_{11}^{(1)}]$. Making the substitution we obtain

$$\begin{aligned}\text{Var}\big[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)\big] = {} & \big[-\mathbf{H}_{22}^{(2)}\big]^{-1} + \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ & + \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\boldsymbol{\Sigma}_{21}\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ & + \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\boldsymbol{\Sigma}_{12}\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}.\end{aligned}$$

From (14-15), $[-\mathbf{H}_{11}^{(1)}]^{-1}$ and $[-\mathbf{H}_{22}^{(2)}]^{-1}$ are the $\mathbf{V}_1$ and $\mathbf{V}_2$ that appear in Theorem 14.8, which further reduces the expression to

$$\begin{aligned}&\text{Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)]\\ &\quad = \mathbf{V}_2 + \mathbf{V}_2\big[\mathbf{H}_{21}^{(2)}\big]\mathbf{V}_1\big[\mathbf{H}_{21}^{(2)}\big]'\mathbf{V}_2 - \mathbf{V}_2\boldsymbol{\Sigma}_{21}\mathbf{V}_1\big[\mathbf{H}_{21}^{(2)}\big]'\mathbf{V}_2 - \mathbf{V}_2\big[\mathbf{H}_{21}^{(2)}\big]\mathbf{V}_1\boldsymbol{\Sigma}_{12}\mathbf{V}_2.\end{aligned}$$

Two remaining terms are $\mathbf{H}_{21}^{(2)}$, which is the $E[\partial^2 \ln L_2(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_1']$, which is being estimated by $-\mathbf{C}$ in the statement of the theorem [note (14-11) again for the change of sign] and $\boldsymbol{\Sigma}_{21}$, which is the covariance of the two first derivative vectors. This is being estimated by $\mathbf{R}$ in Theorem 14.8. Making these last two substitutions produces

$$\text{Var}\big[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)\big] = \mathbf{V}_2 + \mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2 - \mathbf{V}_2\mathbf{R}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2 - \mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{R}'\mathbf{V}_2,$$

which completes the derivation.

Example 14.5 Two-Step ML Estimation

A common application of the two-step method is accounting for the variation in a constructed regressor in a second-step model. In this instance, the constructed variable is often an estimate of an expected value of a variable that is likely to be endogenous in the second-step model. In this example, we will construct a rudimentary model that illustrates the computations.

In Riphahn, Wambach, and Million (RWM, 2003), the authors studied whether individuals' use of the German health care system was at least partly explained by whether or not they had purchased a particular type of supplementary health insurance. We have used their data set, German Socioeconomic Panel (GSOEP), at several points. (See, Example 7.6.) One of the variables of interest in the study is *DocVis*, the number of times an individual visits the doctor during the survey year. RWM considered the possibility that the presence of supplementary (*Addon*) insurance had an influence on the number of visits. Our simple model is as follows: The model for the number of visits is a Poisson regression (see Section 18.4.1). This is a loglinear model that we will specify as

$$E[DocVis\,|\,\mathbf{x}_2, P_{Addon}] = \mu(\mathbf{x}_2'\boldsymbol{\beta}, \gamma, \mathbf{x}_1'\boldsymbol{\alpha}) = \exp[\mathbf{x}_2'\boldsymbol{\beta} + \gamma\Lambda(\mathbf{x}_1'\boldsymbol{\alpha})].$$

The model contains the dummy variable equal to 1 if the individual has *Addon* insurance and 0 otherwise, which is likely to be endogenous in this equation. But, an estimate of $E[Addon|\mathbf{x}_1]$ from a **logistic probability model** (see Section 17.2) for whether the individual has insurance,

$$\Lambda(\mathbf{x}_1'\boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_1'\boldsymbol{\alpha})}{1 + \exp(\mathbf{x}_1'\boldsymbol{\alpha})} = \text{Prob[Individual has purchased } Addon \text{ insurance}|\mathbf{x}_1].$$

For purposes of the exercise, we will specify

$$(y_1 = Addon)\ \mathbf{x}_1 = (constant, Age, Education, Married, Kids),$$
$$(y_2 = DocVis)\ \mathbf{x}_2 = (constant, Age, Education, Income, Female).$$

As before, to sidestep issues related to the panel data nature of the data set, we will use the 4,483 observations in the 1988 wave of the data set, and drop the two observations for which *Income* is zero.

The log likelihood for the logistic probability model is

$$\ln L_1(\boldsymbol{\alpha}) = \boldsymbol{\Sigma}_i\{(1 - y_{i1}) \ln[1 - \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})] + y_{i1} \ln \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})\}.$$

The derivatives of this log likelihood are

$$\mathbf{g}_{i1}(\boldsymbol{\alpha}) = \partial \ln f_1(y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = [y_{i1} - \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]\mathbf{x}_{i1}.$$

We will maximize this log likelihood with respect to $\boldsymbol{\alpha}$ and then compute $\mathbf{V}_1$ using the BHHH estimator, as in Theorem 14.8. We will also use $\mathbf{g}_{i1}(\boldsymbol{\alpha})$ in computing $\mathbf{R}$.

The log likelihood for the Poisson regression model is

$$\ln L_2 = \Sigma_i[-\mu(\mathbf{x}_{i2}'\beta, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha}) + y_{i2} \ln \mu(\mathbf{x}_{i2}'\beta, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha}) - \ln y_{i2}'].$$

The derivatives of this log likelihood are

$$\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) = \partial \ln f_2(y_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial(\boldsymbol{\beta}', \gamma)' = [y_{i2} - \mu(\mathbf{x}_{i2}'\boldsymbol{\beta}, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha})][\mathbf{x}_{i2}', \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]'$$
$$\mathbf{g}_{i1}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) = \partial \ln f_2(y_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = [y_{i2} - \mu(\mathbf{x}_{i2}'\beta, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha})]\gamma\, \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})[1 - \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]\mathbf{x}_{i1}.$$

We will use $\mathbf{g}_{i2}^{(2)}$ for computing $\mathbf{V}_2$ and in computing $\mathbf{R}$ and $\mathbf{C}$ and $\mathbf{g}_{i1}^{(2)}$ in computing $\mathbf{C}$. In particular,

$$\mathbf{V}_1 = [(1/n)\boldsymbol{\Sigma}_i\mathbf{g}_{i1}(\boldsymbol{\alpha})\mathbf{g}_{i1}(\boldsymbol{\alpha})']^{-1},$$
$$\mathbf{V}_2 = [(1/n)\boldsymbol{\Sigma}_i\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})']^{-1},$$
$$\mathbf{C} = [(1/n)\boldsymbol{\Sigma}_i\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})\mathbf{g}_{i1}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})'],$$
$$\mathbf{R} = [(1/n)\boldsymbol{\Sigma}_i\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})\mathbf{g}_{i1}(\boldsymbol{\alpha})'].$$

Table 14.2 presents the two-step maximum likelihood estimates of the model parameters and estimated standard errors. For the first-step logistic model, the standard errors marked $\mathbf{H}_1$ vs. $\mathbf{V}_1$ compares the values computed using the negative inverse of the second derivatives matrix ($\mathbf{H}_1$) vs. the outer products of the first derivatives ($\mathbf{V}_1$). As expected with a sample this large, the difference is minor. The latter were used in computing the corrected covariance matrix at the second step. In the Poisson model, the comparison of $\mathbf{V}_2$ to $\mathbf{V}_2^*$ shows distinctly that accounting for the presence of $\hat{\boldsymbol{\alpha}}$ in the constructed regressor has a substantial impact on the standard errors, even in this relatively large sample. Note that the effect of the correction is to double the standard errors on the coefficients for the variables that the equations have in common, but it is quite minor for *Income* and *Female*, which are unique to the second-step model.

TABLE 14.2 Estimated Logistic and Poisson Models

	Logistic Model for Addon			*Poisson Model for DocVis*		
	Coefficient	***Standard Error (H_1)***	***Standard Error (V_1)***	***Coefficient***	***Standard Error (V_2)***	***Standard Error (V_2^*)***
Constant	−6.19246	0.60228	0.58287	0.77808	0.04884	0.09319
Age	0.01486	0.00912	0.00924	0.01752	0.00044	0.00111
Education	0.16091	0.03003	0.03326	−0.03858	0.00462	0.00980
Married	0.22206	0.23584	0.23523			
Kids	−0.10822	0.21591	0.21993			
Income				−0.80298	0.02339	0.02719
Female				0.16409	0.00601	0.00770
$\Lambda(\mathbf{x}_1'\boldsymbol{\alpha})$				3.91140	0.77283	1.87014

The covariance of the two gradients, $\mathbf{R}$, may converge to zero in a particular application. When the first- and second-step estimates are based on different samples, $\mathbf{R}$ is exactly zero. For example, in our earlier application, $\mathbf{R}$ is based on two residuals,

$$\mathbf{g}_{i1} = \{Addon_i - E[Addon_i|\mathbf{x}_{i1}]\} \text{ and } \mathbf{g}_{i2}^{(2)} = \{DocVis_i - E[DocVis_i|\mathbf{x}_{i2}, \boldsymbol{\Lambda}_{i1}]\}.$$

The two residuals may well be uncorrelated. This assumption would 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, $\mathbf{V}_2^{**} = (1/n)[\mathbf{V}_2 + \mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2]$. (In our application, the sample correlation between $\mathbf{g}_{i1}$ and $\mathbf{g}_{i2}^{(2)}$ is only 0.015658 and the elements of the estimate of $\mathbf{R}$ are only about 0.01 times the corresponding elements of $\mathbf{C}$—essentially about 99 percent of the correction in $\mathbf{V}_2$* is accounted for by $\mathbf{C}$.)

It has been suggested that this set of procedures might be more complicated than necessary.[18] There are two alternative approaches one might take. First, under general circumstances, the asymptotic covariance matrix of the second-step estimator could be approximated using the bootstrapping procedure that will be discussed in Section 15.4. We would note, however, if this approach is taken, then it is essential that both steps be "bootstrapped." Otherwise, taking $\hat{\boldsymbol{\theta}}_1$ as given and fixed, we will end up estimating $(1/n)\mathbf{V}_2$, not the appropriate covariance matrix. The point of the exercise is to account for the variation in $\hat{\boldsymbol{\theta}}_1$. The second possibility is to fit the full model at once. That is, use a one-step, full information maximum likelihood estimator and estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ simultaneously. Of course, this is usually the procedure we sought to avoid in the first place. And with modern software, this two-step method is often quite straightforward. Nonetheless, this is occasionally a possibility. Once again, Heckman's (1979) famous sample selection model provides an illuminating case. The two-step and full information estimators for Heckman's model are developed in Section 19.4.3.

[18]For example, Cameron and Trivedi (2005, p. 202).

14.8 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION AND ROBUST ASYMPTOTIC COVARIANCE MATRICES

Maximum likelihood estimation requires complete specification of the distribution of the observed random variable(s). 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 if the conditional mean function is $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, 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.[19]

14.8.1 A ROBUST COVARIANCE MATRIX ESTIMATOR FOR THE MLE

A heteroscedasticity robust covariance matrix for the least squares estimator was considered in Section 4.5.2. Based on the general result

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\,\Sigma_i\,\mathbf{x}_i\varepsilon_i, \tag{14-32}$$

a robust estimator of the asymptotic covariance matrix for $\mathbf{b}$ would be the White estimator,

$$\text{Est.Asy.Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}\,[\Sigma_i\,(\mathbf{x}_i e_i)(\mathbf{x}_i e_i)'](\mathbf{X}'\mathbf{X})^{-1}.$$

If $\text{Var}[\varepsilon_i|\mathbf{x}_i] = \sigma^2$ and $\text{Cov}[\varepsilon_i, \varepsilon_j|\mathbf{X}] = 0$, then we can simplify the calculation to $\text{Est.Asy.Var}[\mathbf{b}] = s^2(\mathbf{X}'\mathbf{X})^{-1}$. But the first form is appropriate in either case—it is robust, at least, to heteroscedasticity. This estimator is not robust to correlation across observations, as in a time series (considered in Chapter 20) or to clustered data (considered in the next section). The variance estimator is robust to omitted variables in the sense that $\mathbf{b}$ estimates something consistently, $\boldsymbol{\gamma}$, though generally not $\boldsymbol{\beta}$, and the variance estimator appropriately estimates the asymptotic variance of $\mathbf{b}$ around $\boldsymbol{\gamma}$. The variance estimator might be similarly robust to endogeneity of one or more variables in $\mathbf{X}$, though, again, the estimator, $\mathbf{b}$, itself does not estimate $\boldsymbol{\beta}$. This point is important for the present context. The variance estimator may still be appropriate for the asymptotic covariance matrix for $\mathbf{b}$, but $\mathbf{b}$ estimates something other than $\boldsymbol{\beta}$.

Similar considerations arise in maximum likelihood estimation. The properties of the maximum likelihood estimator are derived from (14-15). The empirical counterpart to (14-32) is

$$\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0 \approx \left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0)\right), \tag{14-33}$$

[19]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).

where $\mathbf{g}_i(\boldsymbol{\theta}_0) = \partial \ln f_i/\partial \boldsymbol{\theta}_0$, $\mathbf{H}_i(\boldsymbol{\theta}_0) = \partial^2 \ln f_i/\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'$ and $\boldsymbol{\theta}_0 = \text{plim } \hat{\boldsymbol{\theta}}_{MLE}$. Note that $\boldsymbol{\theta}_0$ is the parameter vector that is estimated by maximizing $\ln L(\boldsymbol{\theta})$, though it might not be the target parameters of the model if the log likelihood is misspecified, the MLE may be inconsistent. Assuming that $\text{plim } \frac{1}{n}\sum_{i=1}^{n} \mathbf{H}_i(\boldsymbol{\theta}_0) = \overline{\mathbf{H}}$, and the conditions needed for $\sqrt{n}\overline{\mathbf{g}} = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}_0)\right)$ to obey a central limit theorem are met, the appropriate estimator for the variance of the MLE around $\boldsymbol{\theta}_0$ would be

$$\text{Asy.Var}\left[\hat{\boldsymbol{\theta}}_{MLE}\right] = \left[-\overline{\mathbf{H}}\right]^{-1}\{\text{Asy.Var}[\overline{\mathbf{g}}]\}\left[-\overline{\mathbf{H}}\right]^{-1}. \qquad \textbf{(14-34)}$$

The missing element is what to use for the asymptotic variance of $\overline{\mathbf{g}}$. If the information matrix equality (Property D3 in Theorem 14.2) holds, then $\text{Asy.Var}[\overline{\mathbf{g}}] = (-1/n)\overline{\mathbf{H}}$, and we get the familiar result $\text{Asy.Var}\left[\hat{\boldsymbol{\theta}}_{MLE}\right] = \frac{1}{n}\left[-\overline{\mathbf{H}}\right]^{-1}$. However, (14-34) applies whether or not the information matrix equality holds. We can estimate the variance of $\overline{\mathbf{g}}$ with

$$\text{Est.Asy.Var}[\overline{\mathbf{g}}] = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)'\right]. \qquad \textbf{(14-35)}$$

The variance estimator for the MLE is then

$$\text{Est.Asy.Var}\left[\hat{\boldsymbol{\theta}}_{MLE}\right] = \left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1}\left\{\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)'\right]\right\}\left[-\frac{1}{n}\sum_{i=1}^{n} \mathbf{H}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1}. \qquad \textbf{(14-36)}$$

This is a robust covariance matrix for the maximum likelihood estimator.

If $\ln L(\boldsymbol{\theta}_0 | \mathbf{y}, \mathbf{X})$ is the appropriate conditional log likelihood, then the MLE is a consistent estimator of $\boldsymbol{\theta}_0$ and, because of the information matrix equality, the asymptotic variance of the MLE is $(1/n)$ times the bracketed term in (14-33). The issue of robustness would relate to the behavior of the estimator of $\boldsymbol{\theta}_0$ if the likelihood were misspecified. We assume that the function we are maximizing (we would now call it the *pseudo-log likelihood*) is regular enough that the maximizer that we compute converges to a parameter vector, $\boldsymbol{\beta}$. Then, by the results above, the asymptotic variance of the estimator is obtained without use of the information matrix equality. As in the case of least squares, there are two levels of robustness to be considered. To argue that the estimator, itself, is robust in this context, it must first be argued that the estimator is consistent for something that we want to estimate and that maximizing the wrong log likelihood nonetheless estimates the right parameter(s). If the model is not linear, this will generally be much more complicated to establish. For example, in the leading case, for a binary choice model, if one assumes that the probit model applies, and some other model applies, then the estimator is not robust to any of heteroscedasticity, omitted variables, autocorrelation, endogeneity, fixed or random effects, or the wrong distribution. (It is difficult to think of a model failure that the MLE is robust to.) Once the estimator, itself, is validated, then the robustness of the asymptotic covariance matrix is considered.[20]

[20]There is a trend in the current literature routinely to report "robust standard errors," based on (14-36) regardless of the likelihood function (which defines the model).

Example 14.6 A Regression with NonNormal Disturbances

If one believed that the regression disturbances were more widely dispersed than implied by the normal distribution, then the logistic or t distribution might provide an alternative specification. We consider the logistic. The model is

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,\ f(\varepsilon) = \frac{1}{\sigma}\frac{\exp(\varepsilon/\sigma)}{[1 + \exp(\varepsilon/\sigma)]^2} = \frac{1}{\sigma}\frac{\exp(w)}{[1 + \exp(w)]^2} = \frac{1}{\sigma}\Lambda(w)[1 - \Lambda(w)],$$

where $\wedge(w)$ is the logistic CDF. The logistic distribution is symmetric, as is the normal, but has a greater variance, $(\pi^2/3)\sigma^2$ compared to σ^2 for the normal, and greater kurtosis (tail thickness), 4.2 compared to 3.0 for the normal. Overall, the logistic distribution resembles a t distribution with 8 degrees of freedom, which has kurtosis 4.5 and variance $(4/3)\sigma^2$. The three densities for the standardized variable are shown in Figure 14.3.

The log-likelihood function is

$$\ln L(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \{-\ln \sigma + w_i - 2\ln[1 + \exp(w_i)]\},\ w_i = (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma. \quad \textbf{(14-37)}$$

The terms in the gradient and Hessian are

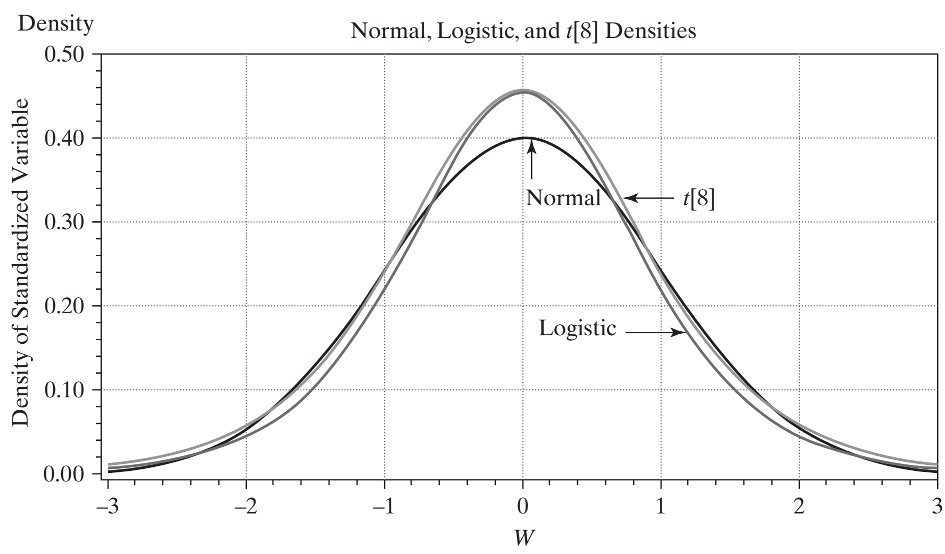
$$\mathbf{g}_i = \frac{-(1 - 2\Lambda(w_i))}{\sigma}\begin{pmatrix}\mathbf{x}_i\\ w_i\end{pmatrix} - \frac{1}{\sigma}\begin{pmatrix}\mathbf{0}\\ 1\end{pmatrix},$$

$$\mathbf{H}_i = \frac{-2\Lambda(w_i)(1 - \Lambda(w_i))}{\sigma^2}\begin{pmatrix}\mathbf{x}_i\\ w_i\end{pmatrix}\begin{pmatrix}\mathbf{x}_i\\ w_i\end{pmatrix}' + \frac{(1 - 2\Lambda(w_i))}{\sigma^2}\begin{bmatrix}0 & \mathbf{x}_i\\ \mathbf{x}_i' & 2w_i\end{bmatrix} + \frac{1}{\sigma^2}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0}' & 1\end{bmatrix}.$$

The conventional estimator of the asymptotic covariance matrix of $\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\sigma}\end{pmatrix}$ would be $\left[-\sum_{i=1}^{n}\hat{\mathbf{H}}_i\right]^{-1}$.

The robust estimator would be

FIGURE 14.3 Standardized Normal, Logistic, and t[8] Densities.



$$\text{Est.Asy.Var}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\sigma}\end{bmatrix} = \left[-\sum_{i=1}^{n}\hat{\mathbf{H}}_i\right]^{-1}\left[\sum_{i=1}^{n}\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i'\right]\left[-\sum_{i=1}^{n}\hat{\mathbf{H}}_i\right]^{-1}.$$

The data in Appendix F14.1 are a panel of 247 dairy farms in Northern Spain, observed for 6 years, 1993–1998. The model is a simple Cobb–Douglas production function,

$$\ln y_{it} = \beta_0 + \beta_1 \ln x_{1,it} + \beta_2 \ln x_{2,it} + \beta_3 \ln x_{3,it} + \beta_4 \ln x_{4,it} + \varepsilon_{it},$$

where y_{it} is the log of milk production, $x_{1,it}$ is number of cows, $x_{2,it}$ is land in hectares, $x_{3,it}$ is labor, and $x_{4,it}$ is feed. The four inputs are transformed to logs, then to deviations from the means of the logs. We then estimated $\boldsymbol{\beta}$ and σ by maximizing the log likelihood for the logistic distribution. Results are shown in Table 14.3. Standard errors are computed using $\left[-\Sigma_i\Sigma_t\hat{\mathbf{H}}_{it}\right]^{-1}$. The robust standard errors shown in column (4) are based on (14-36). They are nearly identical to the uncorrected standard errors, which suggests that the departure of the logistic distribution from the true underlying model or the influence of heteroscedasticity are minor. Column (5) reports the cluster robust standard errors based on (14-38) discussed in the next section.

The departure of the data from the logistic distribution assumed in the likelihood function seems to be minor. The log likelihood does favor the logistic distribution; however, the models cannot be compared on this basis, because the test would have zero degrees of freedom—the models are not nested. The Vuong test examined in Section 14.6.6 might be helpful. The individual terms in the log likelihood are computed using (14-37). For the normal distribution, the term in the log likelihood would be $\ln f_{it} = -(1/2)[\ln 2\pi + \ln s^2 + (y_{it} - \mathbf{x}_{it}'\mathbf{b})^2/s^2]$ where $s^2 = \mathbf{e}'\mathbf{e}/n$. Using $d_{it} = (\ln f_{it}|\text{logistic} - \ln f_{it}|\text{normal})$, the test statistic is $V = \sqrt{n}\bar{d}/s_d = 1.682$. This slightly favors the logistic distribution, but is in the inconclusive region. We conclude that for these data, the normal and logistic models are essentially indistinguishable.

14.8.2 CLUSTER ESTIMATORS

Micro-level, or individual, data are often grouped or clustered. A model of production or economic success at the firm level might be based on a group of industries, with multiple

TABLE 14.3 Maximum Likelihood Estimates of a Production Function

Estimate	*(1) Least Squares*	*(2) MLE Logistic*	*(3) Standard Error*	*(4) Robust Std. Error*	*(5) Clustered Std.Error*
β_0	11.5775	11.5826	0.00353	0.00364	0.00751
β_1	0.59518	0.58696	0.01944	0.02124	0.03697
β_2	0.02305	0.02753	0.01086	0.01104	0.01924
β_3	0.02319	0.01858	0.01248	0.01226	0.02325
β_4	0.45176	0.45671	0.01069	0.01160	0.02071
σ	0.14012[a]	0.07807	0.00169	0.00164	0.00299
R^2	0.92555	0.95253[b]			
$\ln L$	809.676	821.197			

[a]MLE of $\sigma^2 = \mathbf{e}'\mathbf{e}/n$.
[b]R^2 is computed as the squared correlation between predicted and actual values.

firms in each industry. Analyses of student educational attainment might be based on samples of entire classes, or schools, or statewide averages of schools within school districts. And, of course, such "clustering" is the defining feature of a panel data set. We considered several of these types of applications in Section 4.5.3 and in our analysis of panel data in Chapter 11. The recent literature contains many studies of clustered data in which the analyst has estimated a pooled model but sought to accommodate the expected correlation across observations with a correction to the asymptotic covariance matrix. We used this approach in computing a robust covariance matrix for the pooled least squares estimator in a panel data model [see (11-3) and Examples 11.7 and 11.11].

For the normal linear regression model, the log likelihood that we maximize with the pooled least squares estimator is

$$\ln L = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2}\frac{(y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta})^2}{\sigma^2}\right].$$

By multiplying and dividing by $(\sigma^2)^2$, the "cluster-robust" estimator in (11-3) can be written

$$\mathbf{W} = \left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left[\sum_{i=1}^{n}(\mathbf{X}_i'\mathbf{e}_i)(\mathbf{e}_i'\mathbf{X}_i)\right]\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}$$
$$= \left(-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\frac{1}{\hat{\sigma}^2}\mathbf{x}_{it}e_{it}\right)\left(\sum_{t=1}^{T_i}\frac{1}{\hat{\sigma}^2}e_{it}\mathbf{x}_{it}'\right)\right]\left(-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}.$$

The terms in the second line are the first and second derivatives of $\ln f_{it}$ for the normal distribution mean $\mathbf{x}_{it}'\boldsymbol{\beta}$ and variance σ^2 shown in (14-3). A general form of the result is

$$W = \left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\frac{\partial^2 \ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\frac{\partial \ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}}\right)\left(\sum_{t=1}^{T_i}\frac{\partial \ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}'}\right)\right]\left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\frac{\partial^2 \ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}. \quad \textbf{(14-38)}$$

This form of the correction would account for unspecified correlation across the observations (the derivatives) within the groups. [The finite population correction in (11-4) is sometimes applied.]

Example 14.7 Cluster Robust Standard Errors

The dairy farm data used in Example 14.6 are a panel of 247 farms observed in 6 consecutive years. A correction of the standard errors for possible group effects would be natural. Column (5) of Table 14.3 shows the standard errors computed using (14-38). The corrected standard errors are nearly double the values in column (5). This suggests that although the distributional specification is reasonable, there does appear to be substantial correlation across the observations.We will examine this feature of the data further in Section 19.2.4 in the discussion of the stochastic production frontier model.

Consider the specification error that the estimator is intended to accommodate for the normal linear regression. Suppose that the observations in group i were multivariate normally distributed with disturbance mean vector zero and unrestricted $T_i \times T_i$ covariance matrix, $\boldsymbol{\Sigma}_i$. Then, the appropriate log-likelihood function would be

$$\ln L = \sum_{i=1}^{n}(-T_i/2\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}\boldsymbol{\varepsilon}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\varepsilon}_i),$$

where $\boldsymbol{\varepsilon}_i$ is the $T_i \times 1$ vector of disturbances for individual i. Therefore, by using pooled least squares, we have maximized the wrong likelihood function. Indeed, the $\boldsymbol{\beta}$ that maximizes this log-likelihood function is the GLS estimator (see Chapter 9), not the OLS estimator. But OLS and the cluster corrected estimator given earlier "work" in the sense that (1) the least squares estimator is consistent in spite of the misspecification and (2) the robust estimator does, indeed, estimate the appropriate asymptotic covariance matrix.

Now, consider the more general case. Suppose the data set consists of n multivariate observations, $[y_{i,1}, \ldots, y_{i,T_i}]$, $i = 1, \ldots, n$. Each cluster is a draw from joint density $f_i(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta})$. Once again, to preserve the generality of the result, we will allow the cluster sizes to differ. The appropriate log likelihood for the sample is

$$\ln L = \sum_{i=1}^{n} \ln f_i(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta}).$$

Instead of maximizing $\ln L$, we maximize a pseudo-log likelihood

$$\ln L_P = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln g(y_{it} | \mathbf{x}_{it}, \boldsymbol{\theta}),$$

where we make the possibly unreasonable assumption that the same parameter vector, $\boldsymbol{\theta}$, enters the pseudo-log likelihood as enters the correct one. Using our familiar first-order asymptotics, the **pseudo-maximum likelihood estimator** (MLE) will satisfy

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_{P,ML} - \boldsymbol{\theta}) &\approx \left(\frac{-1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \left(\frac{-1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial \boldsymbol{\theta}} \right) + (\boldsymbol{\theta} - \boldsymbol{\beta}) \\ &= \left(\frac{-1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \mathbf{H}_{it} \right)^{-1} \left(\sum_{i=1}^{n} a_i \bar{\mathbf{g}}_i \right) + (\boldsymbol{\theta} - \boldsymbol{\beta}),\end{aligned}$$

where $a_i = T_i / \sum_{i=1}^{n} T_i$ and $\bar{\mathbf{g}}_i = (1/T_i) \sum_{t=1}^{T_i} \partial \ln f_{it} / \partial \boldsymbol{\theta}$. The trailing term in the expression is included to allow for the possibility that plim $\hat{\boldsymbol{\theta}}_{P,ML} = \boldsymbol{\beta}$, which may not equal $\boldsymbol{\theta}$.[21] Taking the expected outer product of this expression to estimate the asymptotic mean squared deviation will produce two terms—the cross term vanishes. The first will be the cluster-corrected matrix that is ubiquitous in the current literature. The second will be the squared error that may persist as n increases because the pseudo-MLE need not estimate the parameters of the model of interest.

We draw two conclusions. We can justify the cluster estimator based on this approximation. In general, it will estimate the expected squared variation of the pseudo-MLE around its probability limit. Whether it measures the variation around the appropriate parameters of the model hangs on whether the second term equals zero. In words, perhaps not surprisingly, this apparatus only works if the pseudo-MLE is consistent. Is that likely? Certainly not if the pooled model is ignoring unobservable fixed effects. Moreover, it will be inconsistent in most cases in which the misspecification is to ignore latent random effects as well. The pseudo-MLE is only consistent for random effects in a few special

[21]Note, for example, Cameron and Trivedi (2005, p. 842) specifically assume consistency in the generic model they describe.

cases, such as the linear model and Poisson and negative binomial models discussed in Chapter 18. It is not consistent in the probit and logit models in which this approach is often used. In the end, the cases in which the estimator are consistent are rarely, if ever, enumerated. The upshot is stated succinctly by Freedman (2006, p. 302): "The sandwich algorithm, under stringent regularity conditions, yields variances for the MLE that are asymptotically correct even when the specification—and hence the likelihood function—are incorrect. However, it is quite another thing to ignore bias. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter."

14.9 MAXIMUM LIKELIHOOD ESTIMATION OF LINEAR REGRESSION MODELS

We will now examine several applications of the MLE. We begin by developing the ML counterparts to most of the estimators for the classical and generalized regression models in Chapters 4 through 11. (Generally, the development for dynamic models becomes more involved than we are able to pursue here. The one exception we will consider is the standard model of autocorrelation.) We emphasize, in each of these cases, that we have already developed an efficient, generalized method of moments estimator that has the same asymptotic properties as the MLE under the assumption of normality. In more general cases, we will sometimes find that the GMM estimator is actually preferred to the MLE because of its robustness to failures of the distributional assumptions or its freedom from the necessity to make those assumptions in the first place. However, for the extensions of the classical model based on generalized least sqaures that are treated here, that is not the case. It might be argued that in these cases, the MLE is superfluous. There are occasions when the MLE will be preferred for other reasons, such as its invariance to transformation in nonlinear models and, possibly, its small sample behavior (although that is usually not the case). And, we will examine some nonlinear models in which there is no linear method of moments counterpart, so the MLE is the natural estimator. Finally, in each case, we will find some useful aspect of the estimator itself, including the development of algorithms such as Newton's method and the EM method for latent class models.

14.9.1 LINEAR REGRESSION MODEL WITH NORMALLY DISTRIBUTED DISTURBANCES

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

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

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.[22] Making the transformation, we find that the likelihood function for the n observations on the observed random variables is

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

[22]See (B-41) in Section B.5. The analysis to follow is conditioned on **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 X does not involve $\boldsymbol{\beta}$ or σ^2 and that the data are well behaved as discussed in Chapter 4.

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,

$$\begin{aligned} \ln L &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\ln 2\pi + \ln \sigma^2 + (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma^2\right]. \end{aligned} \qquad \textbf{(14-39)}$$

The necessary conditions for maximizing this log likelihood are

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

The values that satisfy these equations are

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

The slope estimator is the familiar one, whereas the variance estimator differs from the least squares value by the divisor of n instead of $n - K$.[23]

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

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

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

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

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,

$$E\left[\hat{\sigma}^2_{\mathrm{ML}}\right] = \frac{n - K}{n}\sigma^2 = \left(1 - \frac{K}{n}\right)\sigma^2 < \sigma^2. \qquad \textbf{(14-40)}$$

Despite its small-sample bias, the maximum likelihood estimator of σ^2 has the same desirable asymptotic properties. We see in (14-40) that s^2 and $\hat{\sigma}^2$ differ only by a factor $-K/n$, which

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

vanishes in large samples. It is instructive to formalize the asymptotic equivalence of the two. From (14-40), we know that

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

It follows that

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

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

14.9.2 SOME LINEAR MODELS WITH NONNORMAL DISTURBANCES

The log-likelihood function for a linear regression model with normally distributed disturbances is

$$\begin{aligned} \ln L_N(\boldsymbol{\beta}, \sigma) &= \sum_{i=1}^{n}\{-\ln \sigma - (1/2)\ln 2\pi - (1/2)w_i^2\}, \\ w_i &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma, \sigma > 0. \end{aligned} \qquad \textbf{(14-41)}$$

Example 14.6 considers maximum likelihood estimation of a linear regression model with logistically distributed disturbances. The appeal of the logistic distribution is its greater degree of kurtosis—its tails are thicker than those of the normal distribution. The log-likelihood function is

$$\begin{aligned} \ln L_L(\boldsymbol{\beta}, \sigma) &= \sum_{i=1}^{n}\{-\ln \sigma + w_i - 2\ln[1 + \exp(w_i)]\}, \\ w_i &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma, \sigma > 0. \end{aligned} \qquad \textbf{(14-42)}$$

The logistic specification fixes the shape of the distribution, as suggested earlier, similar to a $t[8]$ distribution. The t distribution with an unrestricted degrees of freedom parameter (a special case of the generalized hyperbolic distribution) allows greater flexibility in this regard. The t distribution arises as the distribution of a sum of δ squares of normally distributed variables. But the degrees of freedom parameter need not be integer valued. We allow δ to be a free parameter, though greater than 4 for the first four moments to be finite. The density of a standardized t distributed random variable with degrees of freedom parameter δ is

$$f(w \mid \delta, \sigma) = \frac{\Gamma[(\delta + 1)/2]}{\Gamma(\delta/2)\Gamma(1/2)\sqrt{\pi\delta}}\frac{1}{\sigma}\left[1 + \frac{w^2}{\delta}\right]^{-(\delta+1)/2}.$$

The log-likelihood function is

$$\begin{aligned} \ln L_t(\boldsymbol{\beta}, \sigma, \delta) &= \sum_{i=1}^{n}\begin{pmatrix} -\ln \sigma + \ln \Gamma[(\delta + 1)/2] - \ln \Gamma(\delta/2) \\ -\ln \Gamma(1/2)] - (1/2)\ln \pi - (1/2)\ln \delta \\ -[(\delta + 1)/2]\ln(1 + w_i^2/\delta) \end{pmatrix}, \\ w_i &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma, \sigma > 0, \delta > 4. \end{aligned} \qquad \textbf{(14-43)}$$

The centerpiece of the stochastic frontier model (Example 12.2 and Section 19.2.4) is a skewed distribution, the skew normal distribution,

$$f(w|\lambda, \sigma) = \frac{2}{\sigma\sqrt{2\pi}}\exp[-(1/2)w^2]\Phi(-\lambda w), \lambda \geq 0,$$

where $\Phi(z)$ is the CDF of the standard normal distribution. If the skewness parameter, λ, equals zero, this returns the standard normal distribution. The skew normal distribution arises as the distribution of $\varepsilon = \sigma_v v_i - \sigma_u |u_i|$, where v_i and u_i are standard normal variables, $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_v^2 + \sigma_u^2$. [Note that σ^2 is not the variance of ε. The variance $|u_i|$ is $(\pi - 2)/\pi$, not 1.] The log-likelihood function is

$$\ln L_{SN}(\boldsymbol{\beta}, \sigma, \lambda) = \sum_{i=1}^{n}\{-\ln \sigma - (1/2)\ln(\pi/2) - (1/2)w_i^2 + \ln \Phi(-\lambda w_i)\}, w_i = (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma. \tag{14-44}$$

Example 14.8 Logistic, t, and Skew Normal Disturbances

Table 14.4 shows the maximum likelihood estimates for the four models. There are only small differences in the slope estimators, as might be expected, at least for the first three, because the differences are in the spread of the distribution, not its shape. The skew normal density has a nonzero mean, $E[\sigma_u|u_i|) = (2/\pi)^{1/2}\sigma_u$, so the constant term has been adjusted. As noted, it is not possible directly to test the normal as a restriction on the logistic, as they have the same number of parameters. The Vuong test does not distinguish them. The *t* distribution would seem to be amenable to a direct specification test; however, the "restriction" on the *t* distribution that produces the normal is $\delta \rightarrow \infty$ which is not useable. However, we can exploit the invariance of the maximum likelihood estimator (property M4 in Table 14.1). The maximum likelihood estimator of $1/\delta$ is $1/\hat{\delta}_{MLE} = 0.101797 = \hat{\gamma}$. We can use the delta method to obtain a standard error. The estimated standard error will be $(1/\hat{\delta}_{MLE})^2(2.54296) = 0.026342$. A Wald test of H_0: $\hat{\gamma} = 0$ would test the normal versus the *t* distribution. The result is $[(0.101797 - 0)/0.026342]^2 = 14.934$, which is larger than the critical value of 3.84, so the hypothesis of normality is rejected. [There is a subtle problem with this test. The value $\gamma = 0$ is on the boundary of the parameter space, not the interior. As such, the chi-squared statistic does not have its usual properties. This issue is explored in Kodde and Palm (1988) and Coelli (1995), who suggest that an appropriate critical value for a single restriction would be 2.706, rather than 3.84.[24] The same consideration applies to the test of $\lambda = 0$ below.] We note, because the log-likelihood function could have been parameterized in terms of γ to begin with, we should be able to use a likelihood ratio test to test the same hypothesis. By the invariance result, the log likelihood in terms of γ would not change, so the test statistic will be $\lambda_{LR} = -2(809.676 - 822.192) = 25.032$. This produces the same conclusion. The normal distribution is nested within the skew normal, by $\lambda = 0$ or $\sigma_u = 0$. We can test the first of these with a likelihood ratio test; $\lambda_{LR} = -2(809.676 - 822.688) = 26.024$. The Wald statistic based on the derived estimate of σ_u would be $(0.15573/0.00279)^2 = 3115.56$.[25] The conclusion is the same for both cases. As noted, the *t* and logistic are essentially indistinguishable. The

[24]The critical value is found by solving for c in $.05 = (1/2)\text{Prob}(\chi^2[1] \geq c)$. For a chi-squared variable with one degree of freedom, the 90th percentile is 2.706.

[25]Greene and McKenzie (2015) show that for the stochastic frontier model examined here, the LM test for the hypothesis that $\sigma_u = 0$ can be based on the OLS residuals; the chi-squared statistic with one degree of freedom is $(n/6)(m_3/s^3)^2$ where m_3 is the third moment of the residuals and s^2 equals $\mathbf{e}'\mathbf{e}/n$. The value for this data set is 21.665.

remaining question, then, is whether the respecification of the model favors skewness or kurtosis. We do not have a direct statistical test available. The OLS estimator of $\boldsymbol{\beta}$ is consistent regardless, so some information might be contained in the residuals. Figure 14.4 compares the OLS residuals to the normal distribution with the same mean (zero) and standard deviation (0.14012). The figure does suggest the presence of skewness, not excess spread. Given the nature of the production function application, skewness is central to this model, so the findings so far might be expected. The development of the stochastic production frontier model is continued in Section 19.2.4.

14.9.3 HYPOTHESIS TESTS FOR REGRESSION MODELS

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

$$F[J, n - K] = \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{J}. \qquad \textbf{(14-45)}$$

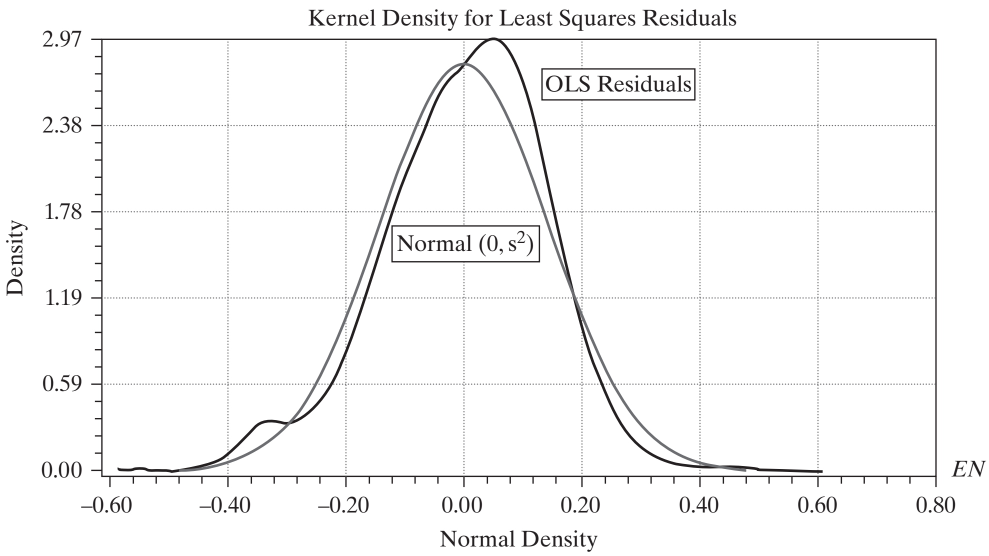
TABLE 14.4 Maximum Likelihood Estimates

(Estimated standard errors in parentheses)

Estimate	*OLS/MLE Normal*	*MLE Logistic*	*MLE t Frac. D.F.*	*MLE Skew Normal*
β_0	11.5775	11.5826	11.5813	11.6966[c]
	(0.00365)	(0.00353)	(0.00363)	(0.00447)
β_1	0.59518	0.58696	0.59042	0.58369
	(0.01958)	(0.01944)	(0.01803)	(0.01887)
β_2	0.02305	0.02753	0.02576	0.03555
	(0.01122)	(0.01086)	(0.01096)	(0.01113)
β_3	0.02319	0.01858	0.01971	0.02256
	(0.01303)	(0.01248)	(0.01299)	(0.01281)
β_4	0.45176	0.45671	0.45220	0.44948
	(0.01078)	(0.01069)	(0.00989)	(0.01035)
σ	0.14012[a]	0.07807	0.12519	0.13988[d]
	(0.00275)	(0.00169)	(0.00404)	(0.00279)
δ			9.82350	
			(2.54296)	
λ				1.50164
				(0.08748)
σ_u				0.15573[e]
				(0.00279)
R^2	0.92555	0.95253[b]	0.95254[b]	0.95250[b]
ln L	809.676	821.197	822.192	822.688

[a]MLE of $\sigma = \mathbf{e}'\mathbf{e}/n$.
[b]R^2 is computed as the squared correlation between predicted and actual values.
[c]Nonzero mean disturbance. Adjustment to β_0 is $\sigma_u(2/\pi)^{1/2} = -0.04447$.
[d]Reported $\sigma_\varepsilon = [\sigma_v^2 + \sigma_u^2(\pi - 2)/\pi]^{1/2}$. Estimated $\sigma_v = 0.10371$ (0.00418).
[e]σ_u is derived. $\sigma_u = \sigma\lambda/(1 + \lambda^2)^{1/2}$. Est.Cov$(\hat{\sigma}, \hat{\lambda}) = 2.3853\text{e}-7$. Standard error is computed using the delta method.

FIGURE 14.4 Distribution of Least Squares Residuals.



With normally distributed disturbances, the F test is valid in any sample size. The more general form of the statistic,

$$F[J, n - K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n - K)}, \quad \textbf{(14-46)}$$

is useable in large samples when the disturbances are homoscedastic even if the disturbances are not normally distributed and with nonlinear restrictions of the general form $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{0}$. In the linear regression setting with linear restrictions, the Wald statistic, $\mathbf{c}(\mathbf{b})'\{\text{Asy.Var}[\mathbf{c}(\mathbf{b})]\}^{-1}\mathbf{c}(\mathbf{b})$, equals $J \times F[J, n - K]$, so the large-sample validity extends beyond normal linear model. (See Sections 5.3.1 and 5.3.2.)

In this section, we will reconsider the Wald statistic and examine two related statistics, the likelihood ratio and Lagrange multiplier statistics. 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

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

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

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

where $\mathbf{G}(\mathbf{b}) = [\partial\mathbf{c}(\mathbf{b})/\partial\mathbf{b}']$.

The Wald statistic is based on the asymptotic distribution of the estimator. The covariance matrix can be replaced with any valid estimator of the asymptotic covariance. Also, for the same reason, the same distributional result applies to estimators based on the nonnormal distributions in Example 14.7, and indeed, for any estimator in any model setting in which $\hat{\boldsymbol{\beta}} \xrightarrow{a} N[\boldsymbol{\beta}, \mathbf{V}]$. The general result, then, is

$$W = \mathbf{c}(\hat{\boldsymbol{\beta}})'\{\mathbf{G}(\hat{\boldsymbol{\beta}})[\text{Asy.Var}(\hat{\boldsymbol{\beta}})]G(\hat{\boldsymbol{\beta}})'\}^{-1}\mathbf{c}(\hat{\boldsymbol{\beta}}) \xrightarrow{d} \chi^2[J]. \qquad \textbf{(14-47)}$$

The Wald statistic is robust in that it relies on the large sample distribution of the estimator, not on the specific distribution that underlies the likelihood function. The Wald test will be the statistic of choice in a variety of settings, not only the likelihood-based one considered here.

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 **b*** is computed (except for the linear model, which we saw earlier). The test statistic and its limiting distribution under H_0 are

$$LR = -2[\ln L_* - \ln L] \xrightarrow{d} \chi^2[J]. \qquad \textbf{(14-48)}$$

This result is general for any nested models fit by maximum likelihood. The log likelihood for the normal/linear regression model is given in (14-39). 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 $\hat{\sigma}_*^2 = (\mathbf{y} - \mathbf{Xb}_*)'(\mathbf{y} - \mathbf{Xb}_*)/n = \mathbf{e}_*'\mathbf{e}_*/n$. Evaluated at the maximum likelihood estimator, the **concentrated log likelihood**[26] 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

$$LR = n \ln[\mathbf{e}_*'\mathbf{e}_*/\mathbf{e}'\mathbf{e}] = n(\ln \hat{\sigma}_*^2 - \ln \hat{\sigma}^2) = n \ln(\hat{\sigma}_*^2/\hat{\sigma}^2). \qquad \textbf{(14-49)}$$

(Note, this is a specific result that applies to the linear or nonlinear regression model with normally distributed disturbances.)

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

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

[26]See Section E4.3.

The first-order conditions for a solution are

$$\begin{bmatrix} \dfrac{\partial \ln L_{\text{LM}}}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ln L_{\text{LM}}}{\partial \sigma^2} \\ \dfrac{\partial \ln L_{LM}}{\partial \boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} + \mathbf{G}(\boldsymbol{\beta})'\boldsymbol{\lambda} \\ \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \\ \mathbf{c}(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \end{bmatrix}. \tag{14-50}$$

The solutions to these equations give the restricted least squares estimator, **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 $\boldsymbol{\lambda}_* = [\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']^{-1}(\mathbf{Gb} - \mathbf{q})$. This equation is a linear function of the unrestricted least squares estimator. If we carry out a Wald test of the hypothesis that $\boldsymbol{\lambda}_*$ equals $\mathbf{0}$, then the statistic will be

$$\text{LM} = \boldsymbol{\lambda}_*'\{\text{Est.Var}[\boldsymbol{\lambda}_*]\}^{-1}\,\boldsymbol{\lambda}_* = (\mathbf{Gb} - \mathbf{q})'[\mathbf{G}s_*^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']^{-1}(\mathbf{Gb} - \mathbf{q}). \tag{14-51}$$

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 for the linear regression model produces an interesting result. 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

$$\text{Asy.Var}[\partial \ln L/\partial \boldsymbol{\beta}] = -E[\partial^2 \ln L/\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}']^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.^{27} \tag{14-52}$$

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

$$\text{LM} = \frac{\mathbf{e}_*'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n} = nR_*^2. \tag{14-53}$$

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. Finally, for more general models and contexts, the same principle for the LM test produces

$$\begin{aligned} LM &= \left[\bar{\mathbf{g}}\left(\hat{\boldsymbol{\theta}}_R\right)\right]'\left[\text{Est.Asy.Var}\left(\bar{\mathbf{g}}\left(\hat{\boldsymbol{\theta}}_R\right)\right)\right]^{-1}\left[\bar{\mathbf{g}}\left(\hat{\boldsymbol{\theta}}_R\right)\right] \\ &= \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)\right]'\left[\frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)'\right\}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)\right] \\ &= \mathbf{i}'\hat{\mathbf{G}}(\hat{\mathbf{G}}'\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}'\mathbf{i}, \end{aligned} \tag{14-54}$$

where $\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right) = \dfrac{\partial \ln f_i\left(\hat{\boldsymbol{\theta}}_R\right)}{\partial \hat{\boldsymbol{\theta}}_R}$, **i** is a column of ones, and $\mathbf{g}_i(\hat{\boldsymbol{\theta}}_R)'$ is the ith row of $\hat{\mathbf{G}}$.

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

There is evidence that the asymptotic results for these statistics are problematic in small or moderately sized samples.[28] 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 the critical values from the chi-squared distribution are likely to be too small; that is, using the limiting chi-squared distribution in small samples is likely to exaggerate the significance of empirical results. Thus, in applications, the more conservative F statistic (or t for one restriction) may be preferable unless one's data are plentiful.

Example 14.9 Testing for Constant Returns to Scale

The Cobb–Douglas production function estimated in Examples 14.6 and 14.7 has returns to scale parameter $\gamma = \Sigma_k \partial \ln y / \partial \ln x_k = \beta_1 + \beta_2 + \beta_3 + \beta_4$. The hypothesis of constant returns to scale, $\gamma = 1$, is routinely tested in this setting. We will carry out this test using the three procedures defined earlier. The estimation results are shown in Table 14.5. For the likelihood ratio test, the chi-squared statistic equals $-2(794.624 - 822.688) = 56.129$. The critical value for a test statistic with one degree of freedom is 3.84, so the hypothesis will be rejected on this basis. For the Wald statistic, based on the unrestricted results, $\mathbf{c}(\boldsymbol{\beta}) = [(\beta_1 + \beta_2 + \beta_3 + \beta_4) - 1]$ and $\mathbf{G} = [1, 1, 1, 1]$. The part of the asymptotic covariance matrix needed for the test is shown with Table 4.5. The statistic is

$$W = \mathbf{c}'(\hat{\boldsymbol{\beta}}_U)[\mathbf{GVG}']^{-1}\mathbf{c}(\hat{\boldsymbol{\beta}}_U) = 57.312.$$

TABLE 14.5 Testing for Constant Returns to Scale in a Production Function

(Estimated standard errors in parentheses)

Estimate	*Stochastic Frontier Unrestricted*		*Stochastic Frontier Constant Returns to Scale*	
β_0[a]	11.7014	(0.00447)	11.7022[a]	(.00457)
β_1	0.58369	(0.01887)	0.55979	(.01903)
β_2	0.03555	(0.01113)	0.00812	(.01075)
β_3	0.02256	(0.01281)	−0.04367	(.00959)
β_4	0.44948	(0.01035)	0.47575	(.00997)
σ[b]	0.13988	(0.00279)	0.18962	(.00011)
λ	1.50164	(0.08748)	1.47082	(.08576)
σ_u[c]	0.15573[d]	(0.00279)	0.15681	(0.00289)
ln L	822.688		794.624	

[a] Unadjusted for nonzero mean of ε.
[b] Reported $\sigma_\varepsilon = [\sigma_v^2 + \sigma_u^2(\pi - 2)/\pi]^{1/2}$. Estimated $\sigma_v = 0.10371$ (0.00418).
[c] σ_u is derived. $\sigma_u = \sigma\lambda/(1 + \lambda^2)^{1/2}$. Est.Cov$(\hat{\sigma}, \hat{\lambda}) = 2.3853\text{e}-7$.
Standard error is computed using the delta method.

Estimated Asy.Var[b1,b2,b3,b4] (e − n = times 10^{-n}.)

0.0003562			
−0.0001079	0.0001238		
−5.576e − 5	9.193e − 6	0.0001642	
−0.0001542	1.810e − 5	−1.235e − 5	0.0001071

[28]See, for example, Davidson and MacKinnon (2004, pp. 424–428).

For the LM test, we need the derivatives of the log-likelihood function. For the particular terms,

$$\begin{aligned} g_\beta &= \partial \ln f_i/\partial(\mathbf{x}_i'\boldsymbol{\beta}) &&= (1/\sigma)[w_i + \lambda A_i],\ A_i = \phi(-\lambda w_i)/\Phi(-\lambda w_i), \\ g_\sigma &= \partial \ln f_i/\partial\sigma &&= (1/\sigma)[-1 + w_i^2 + \lambda w_i A_i], \\ g_\lambda &= \partial \ln f_i/\partial\lambda &&= -w_i A_i. \end{aligned}$$

The calculation is in (14-48); LM = 56.398. The test results are nearly identical for the three approaches.

14.10 THE GENERALIZED REGRESSION MODEL

For the generalized regression model of Section 9.1,

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n, \\ E[\boldsymbol{\varepsilon}\,|\,\mathbf{X}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\,|\,\mathbf{X}] &= \sigma^2\boldsymbol{\Omega}, \end{aligned}$$

and as before, we first assume that $\boldsymbol{\Omega}$ is a matrix of known constants. If the disturbances are multivariate normally distributed, then the log-likelihood function for the sample is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\ln|\boldsymbol{\Omega}|. \qquad \textbf{(14-55)}$$

It might seem that simply using OLS and a heteroscedasticity robust covariance matrix (see Section 4.5) would be a preferred approach that does not rely on an assumption of normality. There are at least two situations in which GLS, and possibly MLE, might be justified. First, if there is known information about the disturbance variances, this simplicity is a minor virtue that wastes sample information. The grouped data application in Example 14.11 is such a case. Second, there are settings in which the variance itself is of interest, such as models of production risk [Asche and Tvertas (1999)] and in the heteroscedastic stochastic frontier model, which is generally based on the model in Section 14.10.3.[29]

14.10.1 GLS WITH KNOWN Ω

Because $\boldsymbol{\Omega}$ is a matrix of known constants, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the vector that minimizes the **generalized sum of squares**, $S_*(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ (hence the name *generalized least squares*). The necessary conditions for maximizing L are

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} = \frac{\mathbf{X}_*'(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})}{\sigma^2} = \mathbf{0}, \\ \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{n}{2\sigma^2}\left[\frac{(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})'(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})}{n\sigma^2} - 1\right] = 0, \end{aligned} \qquad \textbf{(14-56)}$$

[29]Just and Pope (1978, 1979).

where $\mathbf{X}_* = \boldsymbol{\Omega}^{-1/2}\mathbf{X}$ and $\mathbf{y}_* = \boldsymbol{\Omega}^{-1/2}\mathbf{y}$. The solutions are the OLS estimators using the transformed data,

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y},$$
$$\hat{\sigma}^2_{\text{ML}} = \frac{(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}})'(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}})}{n} = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n}, \quad \textbf{(14-57)}$$

which implies that with normally distributed disturbances, generalized least squares is also maximum likelihood. The maximum likelihood estimator of σ^2 is biased. An unbiased estimator is the one in (9-20). The conclusion is that when $\boldsymbol{\Omega}$ is known, the maximum likelihood estimator is generalized least squares.

14.10.2 ITERATED FEASIBLE GLS WITH ESTIMATED $\boldsymbol{\Omega}$

When $\boldsymbol{\Omega}$ is unknown and must be estimated, then it is necessary to maximize the log likelihood in (14-55) with respect to the full set of parameters $[\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}]$ simultaneously. Because an unrestricted $\boldsymbol{\Omega}$ contains $n(n + 1)/2 - 1$ free parameters, it is clear that some restriction will have to be placed on the structure of $\boldsymbol{\Omega}$ for estimation to proceed. We will examine applications in which $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$ for some smaller vector of parameters in the next several sections. We note only a few general results at this point.

1. For a given value of $\boldsymbol{\theta}$ the estimator of $\boldsymbol{\beta}$ would be GLS and the estimator of σ^2 would be the estimator in (14-57).

2. The likelihood equations for $\boldsymbol{\theta}$ will generally be complicated functions of $\boldsymbol{\beta}$ and σ^2, so joint estimation will be necessary. However, in many cases, for given values of $\boldsymbol{\beta}$ and σ^2, the estimator of $\boldsymbol{\theta}$ is straightforward. For example, in the model of (9-21), the iterated estimator of θ when $\boldsymbol{\beta}$ and σ^2 and a prior value of $\boldsymbol{\theta}$ are given is the prior value plus the slope in the regression of $(e_i^2/\hat{\sigma}_i^2 - 1)$ on $\mathbf{z}_i$.

The second step suggests a sort of back-and-forth iteration for this model that will work in many situations—starting with, say, OLS, iterating back and forth between 1 and 2 until convergence will produce the joint maximum likelihood estimator. Oberhofer and Kmenta (1974) showed that under some fairly weak requirements, most importantly that $\boldsymbol{\theta}$ not involve σ^2 or any of the parameters in $\boldsymbol{\beta}$, this procedure would produce the maximum likelihood estimator. The asymptotic covariance matrix of this estimator is the same as the GLS estimator. This is the same whether $\boldsymbol{\Omega}$ is known or estimated, which means that if $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ have no parameters in common, then exact knowledge of $\boldsymbol{\Omega}$ brings no gain in asymptotic efficiency in the estimation of $\boldsymbol{\beta}$ over estimation of $\boldsymbol{\beta}$ with a consistent estimator of $\boldsymbol{\Omega}$.

14.10.3 MULTIPLICATIVE HETEROSCEDASTICITY

Harvey's (1976) model of multiplicative heteroscedasticity is a very flexible, general model that includes many useful formulations as special cases. The general formulation is

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i'\boldsymbol{\alpha}). \quad \textbf{(14-58)}$$

A model with heteroscedasticity of the form $\sigma_i^2 = \sigma^2 \prod_{m=1}^{M} z_{im}^{\alpha_m}$ results if the logs of the variables are placed in $\mathbf{z}_i$. The groupwise heteroscedasticity model described in Section 9.7.2 is produced by making $\mathbf{z}_i$ a set of group dummy variables (one must be omitted). In this case, σ^2 is the disturbance variance for the base group whereas for the other groups $\sigma_g^2 = \sigma^2 \exp(\alpha_g)$.

Let $\mathbf{z}_i$ include a constant term so that $\mathbf{z}_i' = [1, \mathbf{q}_i']$, where $\mathbf{q}_i$ is the original set of variables, and let $\boldsymbol{\gamma}' = [\ln \sigma^2, \boldsymbol{\alpha}']$. Then, the model is simply $\sigma_i^2 = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$. Once the full parameter vector is estimated, $\exp(\gamma_1)$ provides the estimator of σ^2. (This estimator uses the invariance result for maximum likelihood estimation. See Section 14.4.5.D) The log likelihood is

$$\begin{aligned}\ln L &= -\frac{1}{2}\sum_{i=1}^{n}\left[\ln \sigma_i^2 + \ln(2\pi) - \frac{\varepsilon_i^2}{\sigma_i^2}\right]\\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\mathbf{z}_i'\boldsymbol{\gamma} + \ln(2\pi) + \frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right].\end{aligned} \quad \textbf{(14-59)}$$

The likelihood equations are

$$\begin{aligned}\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\mathbf{x}_i\frac{\varepsilon_i}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})},\\ \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} &= \frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i\left(\frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1\right) = \mathbf{0}.\end{aligned} \quad \textbf{(14-60)}$$

14.10.4 THE METHOD OF SCORING

For this model, the **method of scoring** turns out to be a particularly convenient way to maximize the log-likelihood function. The terms in the Hessian are

$$\frac{\partial^2 \ln L}{\partial\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}\partial\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}'} = -\sum_{i=1}^{n}\frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\begin{pmatrix}\mathbf{x}_i\\ \varepsilon_i\mathbf{z}_i\end{pmatrix}\begin{pmatrix}\mathbf{x}_i\\ \varepsilon_i\mathbf{z}_i\end{pmatrix}'. \quad \textbf{(14-61)}$$

The expected value of $\partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}'$ is $\mathbf{0}$ because $E[\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i] = 0$. The expected value of the fraction in $\partial^2 \ln L/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'$ is $E[\varepsilon_i^2/\sigma_i^2 | \mathbf{x}_i, \mathbf{z}_i] = 1$. Let $\boldsymbol{\delta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$. Then

$$-E\left(\frac{\partial^2 \ln L}{\partial\boldsymbol{\delta}\,\partial\boldsymbol{\delta}'}\right) = \begin{bmatrix}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} & \mathbf{0}\\ \mathbf{0}' & \frac{1}{2}\mathbf{Z}'\mathbf{Z}\end{bmatrix} = -\overline{\mathbf{H}}. \quad \textbf{(14-62)}$$

The method of scoring is an algorithm for finding an iterative solution to the likelihood equations. The iteration is

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \overline{\mathbf{H}}^{-1}\mathbf{g}_t,$$

where $\boldsymbol{\delta}_t$ (i.e., $\boldsymbol{\beta}_t$, $\boldsymbol{\gamma}_t$, and $\boldsymbol{\Omega}_t$) is the estimate at iteration t, $\mathbf{g}_t$ is the two-part vector of first derivatives $[\partial \ln L/\partial\boldsymbol{\beta}_t', \partial \ln L/\partial\boldsymbol{\gamma}_t']'$, and $\overline{\mathbf{H}}$ is partitioned likewise. [**Newton's method** uses the actual second derivatives in (14-61) rather than their expectations in (14-62). The scoring method exploits the convenience of the zero expectation of the off-diagonal block (cross derivative) in (14-62).] Because $\overline{\mathbf{H}}$ is block diagonal, the iteration can be written as separate equations,

$$\begin{aligned}\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + (\mathbf{X}\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\boldsymbol{\varepsilon}_t)\\ &= \boldsymbol{\beta}_t + (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_t)\\ &= (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{y}\text{ (of course).}\end{aligned} \quad \textbf{(14-63)}$$

Therefore, the updated coefficient vector $\boldsymbol{\beta}_{t+1}$ is computed by FGLS using the previously computed estimate of $\boldsymbol{\gamma}$ to compute $\boldsymbol{\Omega}$. We use the same approach for $\boldsymbol{\gamma}$:

$$\begin{aligned}\boldsymbol{\gamma}_{t+1} &= \boldsymbol{\gamma}_t + [2(\mathbf{Z}'\mathbf{Z})^{-1}]\left[\frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i\left(\frac{\varepsilon_{i(t)}^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma}_t)} - 1\right)\right] \\ &= \boldsymbol{\gamma}_t + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{h}_t.\end{aligned} \tag{14-64}$$

The 2 and $\frac{1}{2}$ cancel. The updated value of $\boldsymbol{\gamma}$ is computed by adding the vector of coefficients in the least squares regression of $[\varepsilon_i^2/\exp(\mathbf{z}_i'\boldsymbol{\gamma}) - 1]$ on $\mathbf{z}_i$ to the old one. Note that the correction is $2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\partial \ln L/\partial\boldsymbol{\gamma})$, so convergence occurs when the derivative is zero.

The remaining detail is to determine the starting value for the iteration. Any consistent estimator will do. The simplest procedure is to use OLS for $\boldsymbol{\beta}$ and the slopes in a regression of the logs of the squares of the least squares residuals on $\mathbf{z}_i$ for $\boldsymbol{\gamma}$. Harvey (1976) shows that this method will produce an inconsistent estimator of $\gamma_1 = \ln \sigma^2$, but the inconsistency can be corrected just by adding 1.2704 to the value obtained. Thereafter, the iteration is simply:

1. Estimate the disturbance variance σ_i^2 with $\exp(\mathbf{z}_i'\boldsymbol{\gamma})$.
2. Compute $\boldsymbol{\beta}_{t+1}$ by FGLS.[30]
3. Update $\boldsymbol{\gamma}_t$ using the regression described in the preceding paragraph.
4. Compute $\mathbf{d}_{t+1} = [\boldsymbol{\beta}_{t+1}, \boldsymbol{\gamma}_{t+1}] - [\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t]$. If $\mathbf{d}_{t+1}$ is large, then return to step 1.

If $\mathbf{d}_{t+1}$ at step 4 is sufficiently small, then exit the iteration. The asymptotic covariance matrix is simply $-\mathbf{H}^{-1}$, which is block diagonal with blocks

$$\begin{aligned}\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{\text{ML}}] &= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}, \\ \text{Asy.Var}[\boldsymbol{\gamma}_{\text{ML}}] &= 2(\mathbf{Z}'\mathbf{Z})^{-1}.\end{aligned} \tag{14-65}$$

If desired, then $\hat{\sigma}^2 = \exp(\hat{\gamma}_1)$ can be computed. The asymptotic variance would be $[\exp(\gamma_1)]^2(\text{Asy.Var}[\hat{\gamma}_{1,\text{ML}}])$.

Testing the null hypothesis of homoscedasticity in this model,

$$\text{H}_0\text{: } \boldsymbol{\alpha} = \mathbf{0}$$

in (14-58), is particularly simple. The Wald test will be carried out by testing the hypothesis that the last M elements of $\boldsymbol{\gamma}$ are zero. Thus, the statistic will be

$$\lambda_{WALD} = \hat{\boldsymbol{\alpha}}'\left\{[\mathbf{0} \quad \mathbf{I}][2(\mathbf{Z}'\mathbf{Z})]^{-1}\begin{bmatrix}\mathbf{0}' \\ \mathbf{I}\end{bmatrix}\right\}^{-1}\hat{\boldsymbol{\alpha}}.$$

Because the first column in $\mathbf{Z}$ is a constant term, this reduces to

$$\lambda_{WALD} = \frac{1}{2}\hat{\boldsymbol{\alpha}}'(\mathbf{Z}_1'\mathbf{M}^0\mathbf{Z}_1)^{-1}\hat{\boldsymbol{\alpha}},$$

where $\mathbf{Z}_1$ is the last M columns of $\mathbf{Z}$, not including the column of ones, and $\mathbf{M}^0$ creates deviations from means. The likelihood ratio statistic is computed based on (14-59).

[30]The two-step estimator obtained by stopping here would be fully efficient if the starting value for γ were consistent, but it would not be the maximum likelihood estimator.

Under both the null hypothesis (homoscedastic—using OLS) and the alternative (heteroscedastic—using MLE), the third term in ln L reduces to $-n/2$. Therefore, the statistic is simply

$$\lambda_{LR} = 2(\ln L_1 - \ln L_0) = \sum_{i=1}^{n}\left[\ln s^2 - \ln \hat{\sigma}_i^2\right] = \sum_{i=1}^{n}\ln\left(\frac{s^2}{\hat{\sigma}_i^2}\right),$$

where $s^2 = \mathbf{e}'\mathbf{e}/n$ using the OLS residuals. To compute the LM statistic, we will use the expected Hessian in (14-62). Under the null hypothesis, the part of the derivative vector in (14-60) that corresponds to $\boldsymbol{\beta}$ is $(1/s^2)\mathbf{X}'\mathbf{e} = \mathbf{0}$. Therefore, using (14-60), the LM statistic is

$$\lambda_{LM} = \left[\frac{1}{2}\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\begin{pmatrix}1\\ \mathbf{z}_{i1}\end{pmatrix}\right]'\left[\frac{1}{2}(\mathbf{Z}'\mathbf{Z})\right]^{-1}\left[\frac{1}{2}\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\begin{pmatrix}1\\ \mathbf{z}_{i1}\end{pmatrix}\right].$$

The first element in the derivative vector is zero because $\sum_i e_i^2 = ns^2$. Therefore, the expression reduces to

$$\lambda_{LM} = \frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\mathbf{z}_{i1}\right]'(\mathbf{Z}_1'\mathbf{M}^0\mathbf{Z}_1)^{-1}\left[\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\mathbf{z}_{i1}\right].$$

This is one-half times the explained sum of squares in the linear regression of the variable $h_i = (e_i^2/s^2 - 1)$ on $\mathbf{Z}$, which is the Breusch–Pagan/Godfrey LM statistic from Section 9.5.2.

Example 14.10 Multiplicative Heteroscedasticity

In Example 6.4, we fit a cost function for the U.S. airline industry of the form

$$\ln C_{it} = \beta_1 + \beta_2 \ln Q_{it} + \beta_3[\ln Q_{it}]^2 + \beta_4 \ln P_{fuel,i,t} + \beta_5\, Loadfactor_{i,t} + \varepsilon_{i,t},$$

where C_{it} is total cost, Q_{it} is output, and $P_{fuel,i,t}$ is the price of fuel, and the 90 observations in the data set are for six firms observed for 15 years. (The model also included dummy variables for firm and year, which we will omit for simplicity.) In Example 9.4, we fit a revised model in which the load factor appears in the variance of $\varepsilon_{i,t}$ rather than in the regression function. The model is

$$\sigma_{i,t}^2 = \sigma^2 \exp(\alpha\, Loadfactor_{i,t}) = \exp(\gamma_1 + \gamma_2\, Loadfactor_{i,t}).$$

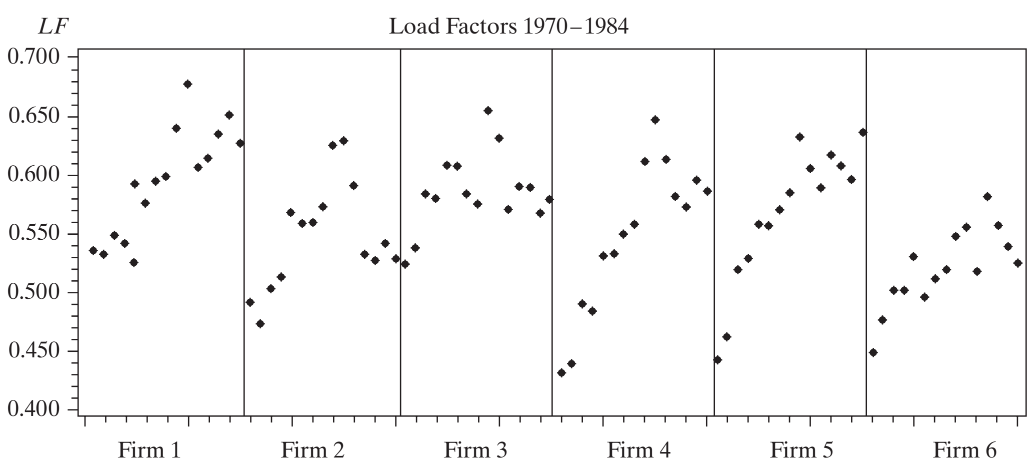
Estimates were obtained by iterating the weighted least squares procedure using weights $W_{i,t} = \exp(-c_1 - c_2\, Loadfactor_{i,t})$. The estimates of γ_1 and γ_2 were obtained at each iteration by regressing the logs of the squared residuals on a constant and $Loadfactor_{it}$. It was noted at the end of the example [and is evident in (14-61)] that these would be the wrong weights to use for iterated weighted least squares if we wish to compute the MLE. Table 14.6 reproduces the results from Example 9.4 and adds the MLEs produced using Harvey's method. The MLE of γ_2 is substantially different from the earlier result. The Wald statistic for testing the homoscedasticity restriction ($\alpha = 0$) is $(9.78076/2.839)^2 = 11.869$, which is greater than 3.84, so the null hypothesis would be rejected. The likelihood ratio statistic is $-2(54.2747 - 57.3122) = 6.075$, which produces the same conclusion. However, the LM statistic is 2.96, which conflicts. This is a finite sample result that is not uncommon. Figure 14.5 shows the pattern of load factors over the period observed. The variances of log costs would vary correspondingly. The increasing load factors in this period would have been a mixed benefit.

TABLE 14.6 Multiplicative Heteroscedasticity Model

	Constant	ln *Q*	ln² *Q*	ln *Pf*	R^{2a}	*Sum of Squares*
OLS[b]	9.13823	0.92615	0.02915	0.41006	0.986167	1.57748
Std. Error	(0.24507)	(0.03231)	(0.01230)	(0.01881)		
Het. Robust S.E.	(0.22595)	(0.03013)	(0.01135)	(0.01752)		
Cluster Robust S.E.	(0.33493)	(0.10235)	(0.04084)	(0.02477)		
Two-step	9.2463	0.92136	0.02445	0.40352	0.9861187	1.612938
Std. Error	(0.21896)	(0.03303)	(0.01141)	(0.01697)		
Iterated[c]	9.2774	0.91609	0.02164	0.40174	0.9860708	1.645693
Std. Error	(0.20977)	(0.03299)	(0.01102)	(0.01633)		
MLE[d]	9.2611	0.91931	0.02328	0.40266	0.9860099	1.626301
Std. Error	(0.2099)	(0.03229)	(0.01099)	(0.01630)		

[a]Squared correlation between actual and fitted values.
[b]$\ln L_{OLS} = 54.2747$, $\ln L_{ML} = 57.3122$.
[c]Values of c_2 by iteration: 8.25434, 11.6225, 11.7070, 11.7106, 11.7110,
[d]Estimate of γ_2 is 9.78076 (2.83945).

FIGURE 14.5 Load Factors for Six Airlines, 1970–1984.



Example 14.11 Maximum Likelihood Estimation of Gasoline Demand

In Example 9.3, we examined a two-step FGLS estimator for the OECD gasoline demand. The model is a groupwise heteroscedastic specification. In (14-58), z_{it} would be a set of country specific dummy variables. The results from Example 9.3 are shown in Table 14.7 in results (1) and (2). The maximum likelihood estimates are shown in column (3). The parameter estimates are similar, as might be expected. It appears that the standard errors of the coefficients are quite a bit smaller using MLE compared to the two-step FGLS. However, the two estimators are essentially the same. They differ numerically, as expected. However, the asymptotic properties of the two estimators are the same.

TABLE 14.7 Estimated Gasoline Consumption Equations

	(1) OLS		*(2) FGLS*		*(3) MLE*	
	Coefficient	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*
ln Income	0.66225	0.07277	0.57507	0.02927	0.45404	0.02211
ln Price	−0.32170	0.07277	−0.27967	0.03519	−0.30461	0.02578
ln Cars/Cap	−0.64048	0.03876	−0.56540	0.01613	−0.47002	0.01275

14.11 NONLINEAR REGRESSION MODELS AND QUASI-MAXIMUM LIKELIHOOD ESTIMATION

In Chapter 7, we considered nonlinear regression models in which the nonlinearity in the parameters appeared entirely on the right-hand side of the equation. Maximum likelihood is often used when the disturbance in a regression, or the dependent variable, more generally, is not normally distributed. If the distribution departs from normality, a likelihood-based approach may provide a useful, efficient way to proceed with estimation and inference. The exponential regression model provides an application.

Example 14.12 Identification in a Loglinear Regression Model

In Example 7.6, we estimated an exponential regression model, of the form

$$E[Income \mid Age, Education, Female] = \exp(\gamma_1^* + \gamma_2\, Age + \gamma_3\, Education + \gamma_4\, Female).$$

This loglinear conditional mean is consistent with several different distributions, including the lognormal, Weibull, gamma, and exponential models. In each of these cases, the conditional mean function is of the form

$$\begin{aligned} E[Income \mid \mathbf{x}] &= g(\theta)\exp(\gamma_1 + \mathbf{x}'\boldsymbol{\gamma}_2) \\ &= \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2), \end{aligned}$$

where θ is an additional parameter of the distribution and $\gamma_1^* = \ln g(\theta) + \gamma_1$. Two implications are:

1. Nonlinear least squares (NLS) is robust at least to some failures of the distributional assumption. The nonlinear least squares estimator of $\boldsymbol{\gamma}_2$ will be consistent and asymptotically normally distributed in all cases for which $E[Income \mid \mathbf{x}] = \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2)$.
2. The NLS estimator cannot produce a consistent estimator of γ_1; plim $c_1 = \gamma_1^*$, which varies depending on the correct distribution. In the conditional mean function, any pair of values (θ, γ_1) for which $\gamma_1^* = \ln g(\theta) + \gamma_1$ is the same will lead to the same sum of squares. This is a form of multicollinearity; the pseudoregressor for θ is $\partial E[Income \mid \mathbf{x}]/\partial\theta = \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2)[g'(\theta)/g(\theta)]$ while that for γ_1 is $\partial E[Income \mid \mathbf{x}]/\partial\gamma_1 = \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2)$. The first is a constant multiple of the second. NLS cannot provide separate estimates of θ and γ_1 while MLE can—see the example to follow. Second, NLS might be less efficient than MLE because it does not use the information about the distribution of the dependent variable. This second consideration is uncertain. For estimation of $\boldsymbol{\gamma}_2$, the NLS estimator is less efficient for not using the distributional information. However, that shortcoming might be offset because the NLS estimator does not attempt to compute an independent estimator of the additional parameter, θ.

To illustrate, we reconsider the estimator in Example 7.6. The gamma regression model specifies

$$f(y\,|\,\mathbf{x}) = \frac{1}{\Gamma(\theta)\mu(\mathbf{x})^{\theta}}\exp[-y/\mu(\mathbf{x})]y^{\theta-1},\, y > 0,\, \theta > 0,\, \mu(\mathbf{x}) = \exp(\gamma_1 + \mathbf{x}'\boldsymbol{\gamma}_2).$$

The conditional mean function for this model is

$$E[y\,|\,\mathbf{x}] = \theta/\mu(\mathbf{x}) = \theta\exp(\gamma_1 + \mathbf{x}'\gamma_2) = \exp(\gamma_1^* + \mathbf{x}'\gamma_2).$$

Table 14.8 presents estimates of θ and $(\gamma_1, \boldsymbol{\gamma}_2)$. Estimated standard errors appear in parentheses. The estimates in columns (1), (2), and (4) are all computed using nonlinear least squares. In (1), an attempt was made to estimate θ and γ_1 separately. The estimator converged on two values. However, the estimated standard errors are essentially infinite. The convergence to anything at all is due to rounding error in the computer. The results in column (2) are for γ_1^* and γ_2. The sums of squares for these two estimates as well as for those in (4) are all 112.19688, indicating that the three results merely show three different sets of results for which γ_1^* is the same. The full maximum likelihood estimates are presented in column (3). Note that an estimate of θ is obtained here because the assumed gamma distribution provides another independent moment equation for this parameter; $\partial \ln L/\partial\theta = -n \ln \Psi(\theta) + \Sigma_i(\ln y_i - \ln \mu(\mathbf{x})) = 0$, while the normal equations for the sum of squares provide the same equations for θ and γ_1.

14.11.1 MAXIMUM LIKELIHOOD ESTIMATION

The standard approach to modeling counts of events begins with the Poisson regression model,

$$\text{Prob}[Y = y_i\,|\,\mathbf{x}_i] = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!},\, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}),\, y_i = 0, 1, \ldots,$$

which has **loglinear conditional mean** function $E[y_i\,|\,\mathbf{x}_i] = \lambda_i$. (The Poisson regression model and other specifications for data on counts are discussed at length in Chapter 18. We

TABLE 14.8 Estimated Gamma Regression Model

	(1) NLS	(2) Constrained NLS	(3) MLE	(4) NLS/MLE
Constant	1.22468	−1.69331	−3.36826	−3.36380
	(47722.5)[a]	(0.04408)	(0.05048)	(0.04408)
Age	0.00207	0.00207	0.00153	0.00207
	(0.00061)[b]	(0.00061)	(0.00061)	(0.00061)
Education	0.04792	0.04792	0.04975	0.04792
	(0.00247)[b]	(0.00247)	(0.00286)	(0.00247)
Female	−0.00658	−0.00658	0.00696	−0.00658
	(0.01373)[b]	(0.01373)	(0.01322)	(0.08677)
θ	0.62699	—	5.31474	5.31474[c]
	(29921.3)[a]	—	(0.10894)	(0.00000)

[a]Reported value is not meaningful; this is rounding error. See text for description.
[b]Standard errors are the same as in column (2).
[c]Fixed at this value.

introduce the topic here to begin development of the MLE in a fairly straightforward, typical nonlinear setting.) Appendix Table F7.1 presents the Riphahn et al. (2003) data, which we will use to analyze a count variable, *DocVis*, the number of visits to physicans in the survey year. We are using the 1988 wave of the panel, with 4,483 observations. The histogram in Figure 14.6 shows a distinct spike at zero followed by rapidly declining frequencies. While the Poisson distribution, which is typically hump shaped, can accommodate this configuration if λ_i is less than one, the shape is nonetheless somewhat "non-Poisson."[31]

The geometric distribution,

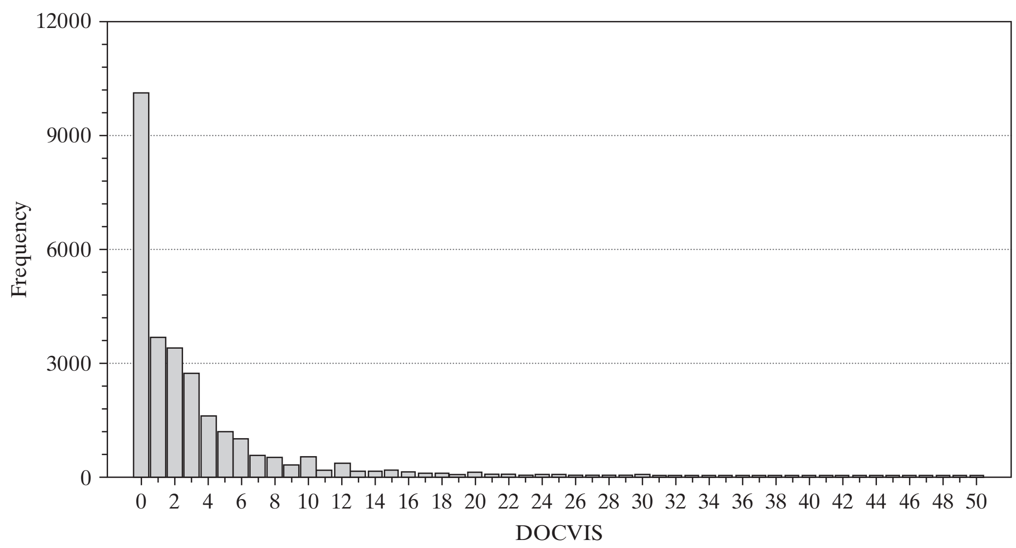
$$f(y_i|\mathbf{x}_i) = \theta_i(1 - \theta_i)^{y_i}, \theta_i = 1/(1 + \lambda_i), \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), y_i = 0, 1, \ldots,$$

is a convenient specification that produces the effect shown in Figure 14.4. (Note that, formally, the specification is used to model the number of failures before the first success in successive independent trials each with success probability θ_i, so in fact, it is misspecified as a model for counts. The model does provide a convenient and useful illustration, however. Moreover, it will turn out that the specification can deliver a consistent estimator of the parameters of interest even if the Poisson is the right model.) The conditional mean function is also $E[y_i|\mathbf{x}_i] = \lambda_i$. The partial effects in the model are $\partial E[y_i|\mathbf{x}_i]/\partial\mathbf{x}_i = \lambda_i\boldsymbol{\beta}$, so this is a distinctly nonlinear regression model. We will construct a maximum likelihood estimator, then compare the MLE to the **nonlinear least squares** and (mis-specified) linear least squares estimates.

The log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \theta_i + y_i \ln(1 - \theta_i).$$

FIGURE 14.6 Histogram for Doctor Visits.



[31] So-called Hurdle and Zero Inflation models (discussed in Chapter 18) are often used for this situation.

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left(\frac{1}{\theta_i} - \frac{y_i}{1 - \theta_i}\right)\frac{d\theta_i}{d\lambda_i}\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

Because

$$\frac{d\theta_i}{d\lambda_i}\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} = \left(\frac{-1}{(1 + \lambda_i)^2}\right)\lambda_i \mathbf{x}_i = -\theta_i(1 - \theta_i)\mathbf{x}_i,$$

the likelihood equations simplify to

$$\begin{aligned}\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}(\theta_i y_i - (1 - \theta_i))\mathbf{x}_i \\ &= \sum_{i=1}^{n}(\theta_i(1 + y_i) - 1)\mathbf{x}_i.\end{aligned}$$

To estimate the asymptotic covariance matrix, we can use any of the estimators of Asy.Var$[\hat{\boldsymbol{\beta}}_{\text{MLE}}]$ discussed earlier. The BHHH estimator would be

$$\begin{aligned}\text{Est.Asy.Var}_{\text{BHHH}}[\hat{\boldsymbol{\beta}}_{\text{MLE}}] &= \left[\sum_{i=1}^{n}\left(\frac{\partial \ln f(y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)\left(\frac{\partial \ln f(y_i, \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)'\right]^{-1} \\ &= \left[\sum_{i=1}^{n}(\hat{\theta}_i(1 + y_i) - 1)^2 \mathbf{x}_i \mathbf{x}_i'\right] \\ &= [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1}.\end{aligned}$$

The negative inverse of the second derivatives matrix evaluated at the MLE is

$$\left[-\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}\, \partial \hat{\boldsymbol{\beta}}'}\right]^{-1} = \left[\sum_{i=1}^{n}(1 + y_i)\hat{\theta}_i(1 - \hat{\theta}_i)\mathbf{x}_i \mathbf{x}_i'\right]^{-1} = [-\hat{\mathbf{H}}]^{-1}.$$

As noted earlier, $E[y_i \mid \mathbf{x}_i] = \lambda_i = (1 - \theta_i)/\theta_i$ is known, so we can also use the negative inverse of the expected second derivatives matrix,

$$\left[-E\left(\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}\, \partial \hat{\boldsymbol{\beta}}}\right)\right]^{-1} = \left[\sum_{i=1}^{n}(1 - \hat{\theta}_i)\mathbf{x}_i \mathbf{x}_i'\right]^{-1} = \{-E[\hat{\mathbf{H}}]\}^{-1}.$$

Finally, although we are confident in the form of the conditional mean function, but uncertain about the distribution, it might make sense to use the robust estimator in (14-36),

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = [-\hat{\mathbf{H}}]^{-1}[\hat{\mathbf{G}}'\hat{\mathbf{G}}][-\hat{\mathbf{H}}]^{-1}.$$

To compute the estimates of the parameters, either Newton's method, $\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^{t} - [\hat{\mathbf{H}}^{t}]^{-1}\hat{\mathbf{g}}^{t}$, or the method of scoring, $\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^{t} - \{E[\hat{\mathbf{H}}^{t}]\}^{-1}\hat{\mathbf{g}}^{t}$, can be used, where $\mathbf{H}$ and $\mathbf{g}$ are the second and first derivatives that will be evaluated at the current estimates of the parameters. Like many models of this sort, there is a convenient set of starting values, assuming the model contains a constant term. Because $E[y_i \mid \mathbf{x}_i] = \lambda_i$, if we start the slope parameters at zero, then a natural starting value for the constant term is the log of $\bar{y}$.

14.11.2 QUASI-MAXIMUM LIKELIHOOD ESTIMATION

If one is confident in the form of the conditional mean function (and that is the function of interest), but less sure about the appropriate distribution, one might seek a robust approach. That is precisely the situation that arose in the preceding example. Given that *DocVis* is a nonnegative count, the exponential mean function makes sense. But we gave equal plausibility to a Poisson model, a geometric model, and a semiparametric approach based on nonlinear least squares. The conditional mean function is correctly specified, but each of these three approaches has a significant shortcoming. The Poisson model imposes an "equidispersion" (variance equal to the mean) that is likely to be transparently inconsistent with the data; the geometric model is manifestly an inappropriate specification, and the nonlinear least squares estimator ignores all information in the sample save for the form of the conditional mean function. A **quasi-MLE**(QMLE) approach based on linear exponential forms provides a somewhat robust approach in this sort of circumstance.

The exponential family of distributions is defined in Definition 13.1. For a random variable, y with density $f(y|\boldsymbol{\theta})$, the exponential family of distributions is

$$\ln f(y|\boldsymbol{\theta}) = a(y) + b(\boldsymbol{\theta}) + \Sigma_k c_k(y) s_k(\boldsymbol{\theta}).$$

Many familiar distributions are in this class, including the normal, logistic, Bernoulli, Poisson, gamma, exponential, Weibull, and others. Based on this framework, Gourieroux, Monfort, and Trognon (1984) proposed the class of conditional linear exponential families,

$$\ln f(y|\mu(\mathbf{x}, \boldsymbol{\beta})) = a(y) + b(\mu(\mathbf{x}, \boldsymbol{\beta})) + ys(\mu(\mathbf{x}, \boldsymbol{\beta})),$$

where the conditional mean function is $E[y|\mathbf{x}, \boldsymbol{\beta}] = \mu(\mathbf{x}, \boldsymbol{\beta})$. The usefulness of this class of specifications is that maximizing the implied log likelihood produces a consistent estimator of $\boldsymbol{\beta}$ even if the true distribution of $y|\mathbf{x}$ is not $f(y|\mu(\mathbf{x}, \boldsymbol{\beta}))$, so long as the mean is correctly specified.

Example 14.13 examines a count variable, $DocVis =$ the number of doctor visits. The assumed conditional mean function is $E[y_i|\mathbf{x}_i] = \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, but we are uncertain of the distribution. Two candidates are considered, geometric with $f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \theta_i(1 - \theta_i)^{y_i}$ with $\theta_i = 1/(1 + \lambda_i)$, and Poisson with $f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = \exp(-\lambda_i)\lambda_i^{y_i}/\Gamma(y_i + 1)$. Both of these distributions are in the LEF family; for the geometric, $\ln f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = \ln[\theta_i/(1-\theta_i)] + y_i \ln \theta_i$, and for the Poisson, $\ln f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = -\lambda_i + y_i \ln \lambda_i - \ln \Gamma(y_i + 1)$. Because both are LEFs involving the same mean function, either log likelihood will produce a consistent estimator of the same $\boldsymbol{\beta}$.

The conditional variance is unspecified so far. In the two cases considered, the variance is a simple function of the mean. For the geometric distribution, $\text{Var}[y|\mathbf{x}] = \lambda(1 + \lambda)$; for the Poisson, $\text{Var}[y|\mathbf{x}] = E[y|\mathbf{x}] = \lambda$. This relationship will hold in general for linear exponential families. For another example, the Bernoulli distribution for a binary or fractional variable, $f(y|\mathbf{x}) = P_i^{y_i}(1 - P_i)^{(1-y_i)}$, where $P_i = \lambda_i/(1 + \lambda_i)$ has conditional variance, $\lambda_i/[1 + \lambda_i]^2 = P_i(1 - P_i)$. The other models examined below, gamma, Weibull, and negative binomial, all behave likewise. The conventional estimator of the asymptotic variance based on the information matrix, (14-16) or (14-17), would apply if the distribution of the LEF were the actual distribution of y_i. However, because the variance has not actually been specified, this may not be the case. Thus, the heteroscedasticity makes the robust variance matrix estimator in (14-36) a logical choice.

An apparently minor extension is needed to accommodate distributions that have an additional parameter, typically a shape parameter, such as the Weibull distribution,

$$f(y_i|\mathbf{x}_i) = \frac{\theta}{\lambda_i}\left(\frac{y_i}{\lambda_i}\right)^{\theta-1}\exp\left[-\left(\frac{y_i}{\lambda_i}\right)^{\theta}\right],$$

for which $E[y_i|\mathbf{x}_i] = \lambda_i\,\Gamma(1 + 1/\theta)$, or the gamma distribution,

$$f(y_i|\mathbf{x}_i) = \frac{y_i^{\theta-1}}{\lambda_i^{\theta}\Gamma(\theta)}\exp\left(-\frac{y_i}{\lambda_i}\right),$$

for which $E[y_i|\mathbf{x}_i] = \lambda_i\theta$. These random variables satisfy the assumptions of the LEF models, but the more detailed specifications create complications both for estimation and inference. First, for these models, the mean, λ_i, is no longer correctly specified. In the cases shown, there is a scaling parameter. If $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ as is typical, and $\boldsymbol{\beta}$ contains a constant term, then the constant term is offset by the log of that scaling term. For the Weibull model, the constant term is offset by $\ln\Gamma(1 + 1/\theta)$ while for the gamma model, the offset is $\ln\theta$. These would seem to be innocuous; however, if the conditional mean itself or partial effects of the mean are the objects of estimation, this is a potentially serious shortcoming. These two models noted are, like the candidates noted earlier, also heteroscedastic; for the gamma, $\text{Var}[y|\mathbf{x}] = \theta\lambda^2$, while for the Weibull, $\text{Var}[y|x] = \lambda\{\Gamma(1 + 2/\theta) - \Gamma^2(1 + 1/\theta)\}$. The robust estimator of the asymptotic covariance matrix in (14-36) for the QMLEs would still be preferred.

These four distributions noted and the others listed below are all members of the LEF, which would suggest that any of them could form the basis of a quasi-MLE for $(y|\mathbf{x})$. The distributions listed are, in principle, for binary (Bernoulli), count (Poisson, geometric), and continuous (gamma, Weibull, normal) random variables. The LEF approach should work best if the random variable studied is of the type that is natural for the form of the distribution used, or at least closest to it. Thus, in the example below, we have modeled the count variable using the geometric and Poisson. One could use the Bernoulli framework for a binary or fractional variable as the basis for the quasi-MLE. Given the results thus far, the Bernoulli LEF could also be used for a continuous variable, but the gamma or Weibull distribution would be a better choice. In general, the support of the observed variable should match that of the variable that underlies the candidate distribution, for example, the nonnegative integers in Example 14.13. [Continuity is not essential; the Poisson (exponential) LEF would work for a continuous (discrete) nonnegative variable.]

If interest centers on estimation of $\boldsymbol{\beta}$, our results would seem to imply that several of these distributions would suffice as the vehicle for estimation in a given situation. But intuition should suggest (no doubt correctly) that some choices should be better than others. On the other hand, why not just use nonlinear least squares (GMM) in all cases if only the conditional mean has been specified? The argument so far does not distinguish any of these estimators; they are all consistent. The criterion function chosen implies a weighting of the observations, and it would seem that some weighting schemes would be better (more efficient) than others, based on the same logic that makes generalized least squares better than ordinary least squares.

The preceding efficiency argument is somewhat ambiguous. It remains a question why one would use this approach instead of nonlinear least squares. The leading application of these methods [and the focus of Gourieroux et al. (1984) who developed them] is about modeling counts such as our doctor visits variable, in the presence of unmeasured heterogeneity. Consider that in the canonical model for counts, the Poisson regression, there is no explicit place in the specification for unmeasured heterogeneity. The entire specification builds off the conditional mean, $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, and the marginal Poisson distribution. A natural way to extend the Poisson regression specification is $\lambda_i | \varepsilon_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i)$. The conditional mean function is $\Lambda_i = \mathrm{E}[\exp(\varepsilon_i)]\lambda_i$. If the model contains a constant term, then nothing is lost by assuming that $\mathrm{E}[\exp(\varepsilon_i)] = 1$, so $\Lambda_i = \lambda_i$. Left unspecified are the variance of $y_i | \mathbf{x}_i$ and the distribution of ε_i. We assume that ε_i is a conventional disturbance, exogenous to the rest of the model. Thus, the conditional (on $\mathbf{x}$) mean is correctly specified by λ_i which implies that the Poisson QMLE is a robust estimator for this model with only vaguely specified heterogeneity—it is exogenous and has mean 1.

The marginal distribution is $f(y_i | \mathbf{x}_i) = \int_{\varepsilon_i} f(y_i | \mathbf{x}_i, \varepsilon_i) g(\varepsilon_i) d\varepsilon_i$. If $\exp(\varepsilon_i)$ has a gamma distribution with mean 1, $G(\theta, \theta)$, this produces the negative binomial type 2 regression model,

$$f(y_i | \mathbf{x}_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1)\Gamma(\theta)} \left(\frac{\lambda_i}{\lambda_i + \theta} \right)^{y_i} \left(\frac{\theta}{\lambda_i + \theta} \right)^{\theta}, \; y_i = 0, 1, \ldots.$$

This random variable has mean λ_i and variance $= \lambda_i[1 + \lambda_i/\theta]$. The negative binomial density is a member of the LEF. The advantage of this formulation for count data is that the Poisson quasi-log likelihood will produce a consistent estimator of $\boldsymbol{\beta}$ regardless of the distribution of ε_i as long as ε_i is exogenous, homoscedastic (with respect to $\mathbf{x}_i$), and is parameterized free of $\boldsymbol{\beta}$.

To conclude, the QMLE would seem to be a competitor to the GMM estimator for certain kinds of models. In the leading application, it is a robust estimator that follows the form of the random variable while nonlinear least squares does not.

Example 14.13 Geometric Regression Model for Doctor Visits

In Example 7.6, we considered nonlinear least squares estimation of a loglinear model for the number of doctor visits variable shown in Figure 14.6. (41 observations for which DocVis > 50 out of 27,326 in total are omitted from the figure). The data are drawn from the Riphahn et al. (2003) data set in Appendix Table F7.1. We will continue that analysis here by fitting a more detailed model for the count variable *DocVis*. The conditional mean analyzed here is

$$\ln E[DocVis_{it} | \mathbf{x}_{it}] = \beta_1 + \beta_2\, Age_{it} + \beta_3\, Educ_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it}.$$

(This differs slightly from the model in Example 11.16.) For this exercise, with an eye toward the fixed effects model in Example 14.13, we have specified a model that does not contain any time-invariant variables, such as *Female.* (Also, for this application, we will use the entire sample.) Sample means for the variables in the model are given in Table 14.9. Note, these data are a panel. In this exercise, we are ignoring that fact, and fitting a pooled model. We will turn to panel data treatments in the next section, and revisit this application.

We used Newton's method for the optimization, with starting values as suggested earlier. The five iterations are shown in Table 14.9.

Convergence based on the LM criterion, $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$, is achieved after the fourth iteration. Note that the derivatives at this point are extremely small, albeit not absolutely zero. Table 14.10 presents the quasi-maximum likelihood estimates of the parameters. Several sets

TABLE 14.9 Newton Iterations

Start values:	0.11580e+1	0.00000	0.00000	0.00000	0.00000
1st derivatives	0.00000	−0.61777e+5	0.73202e+4	0.42575e+4	0.16464e+4
Parameters:	0.11580e+1	0.00000	0.00000	0.00000	0.00000
Iteration 1 F =	0.6287e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.1907e+4			
1st derivatives	0.48616e+3	−0.22449e+5	0.57162e+4	−0.17112e+3	−0.16521e+3
Parameters:	0.11186e+1	0.1762e−1	−0.50263e−1	−0.46274e−1	−0.15609
Iteration 2 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.1258e+2			
1st derivatives	−0.31284e+1	−0.15595e+3	−0.37197e+2	−0.10630e+1	−0.77186
Parameters:	0.10922e+1	0.17981e−1	−0.47303e−1	−0.46739e−1	−0.15683
Iteration 3 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.6759e−3			
1st derivatives	−0.18417e−3	−0.99368e−2	−0.21992e−2	−0.59354e−4	−0.25994e−4
Parameters:	0.10918e+1	0.17988e−1	−0.47274e−1	−0.46751e−1	−0.15686
Iteration 4 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.1831e−8			
1st derivatives	−0.35727e−11	0.86745e−10	−0.26302e−10	−0.61006e−11	−0.15620e−11
Parameters:	0.10918e+1	0.17988e−1	−0.47274e−1	−0.46751e−1	−0.15686
Iteration 5 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.177e−12			

of standard errors are presented. The three sets based on different estimators of the information matrix are presented first. The fourth set is based on the cluster corrected covariance matrix discussed in Section 14.8.4. Because this is actually an (unbalanced) panel data set, we anticipate correlation across observations. Not surprisingly, the standard errors rise substantially. The partial effects listed next are computed in two ways. The *average partial effect* is computed by averaging $\lambda_i\boldsymbol{\beta}$ across the individuals in the sample. The *partial effect* is computed for the average individual by computing λ at the means of the data. The next-to-last column contains the ordinary least squares coefficients. In this model, there is no reason to expect ordinary least squares to provide a consistent estimator of $\boldsymbol{\beta}$. The question might arise, What does ordinary least squares estimate? The answer is the slopes of the linear projection of DocVis on x_{it}. The resemblance of the OLS coefficients to the estimated partial effects is more than coincidental, and suggests an answer to the question.

The analysis in Table 14.11 suggests three competing approaches to modeling DocVis. The results for the geometric regression model are given first in Table 14.10. At the beginning of this section, we noted that the more conventional approach to modeling a count variable such as DocVis is with the Poisson regression model. The quasi-log-likelihood function and its derivatives are even simpler than the geometric model

$$\ln L = \sum_{i=1}^{n} y_i \ln\lambda_i - \lambda_i - \ln y_i!,$$
$$\partial \ln L/\partial\boldsymbol{\beta} = \sum_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i,$$
$$\partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = \sum_{i=1}^{n} - \lambda_i \mathrm{x}_i \mathrm{x}_i'.$$

A third approach might be a semiparametric, nonlinear regression model,

$$\mathrm{y}_{it} = \exp(x_{it}'\boldsymbol{\beta}) + \varepsilon_{it}.$$

TABLE 14.10 Estimated Geometric Regression Model Dependent Variable: DocVis: Mean = 3.18352, Standard Deviation = 5.68969, n = 27,326

Variable	Estimate	Std. Err. H	Std. Err. E[H]	Std. Err. BHHH	Std. Err. Cluster	APE	PE Mean	OLS	Var. Mean
Constant	1.0918	0.0524	0.0524	0.0354	0.1083	—	—	2.656	
Age	0.0180	0.0007	0.0007	0.0005	0.0013	0.0572	0.057	0.061	43.52
Education	−0.0473	0.0033	0.0033	0.0023	0.0067	−0.150	−0.144	−0.121	11.32
Income	−0.4684	0.0411	0.0423	0.0278	0.0727	−1.490	−1.424	−1.621	0.352
Kids	−0.1569	0.0156	0.0155	0.0103	0.0306	−0.487	−0.477	−0.517	0.403

TABLE 14.11 Estimates of Three Models for DocVis

	Geometric Model			Poisson Model			Nonlinear Reg.		
Variable	Estimate	Std. Err.	APE	Estimate	Std. Err.	APE	Estimate	Std. Err.	APE
Constant	1.0918	0.1083		0.10480	0.1137		0.9802	0.1814	
Age	0.0180	0.0013	0.057	0.0184	0.0013	0.060	0.0187	0.0020	0.060
Education	−0.0473	0.0067	−0.150	−0.0433	0.0070	−0.138	−0.0361	0.0123	−0.115
Income	−0.4684	0.0727	−1.490	−0.5207	0.0822	−1.658	−0.5919	0.1283	−1.884
Kids	−0.1569	0.0306	−0.487	−0.1609	0.0312	−0.500	−0.1693	0.0488	−0.539

Without the distributional assumption, nonlinear least squares is robust, but inefficient compared to the QMLE. But the distributional assumption can be dropped altogether, and the model fit as a simple exponential regression. Note the similarity of the Poisson QMLE and the NLS estimator. For the QMLE, the likelihood equations, $\Sigma_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i = \mathbf{0}$, imply that at the solution, the residuals, $(y_i - \lambda_i)$, are orthogonal to the actual regressors, $\mathbf{x}_i$. The NLS normal equations, $\Sigma_{i=1}^{n}(y_i - \lambda_i)\lambda_i\mathbf{x}_i = \Sigma_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i^0 = \mathbf{0}$ will imply that at the solutions, the residuals are orthogonal to the pseudo-regressors, $\lambda_i\mathbf{x}_i$.

Table 14.11 presents the three sets of estimates. It is not obvious how to choose among the alternatives. Of the three, the Poisson model is used most often by far. The Poisson and geometric models are not nested, so we cannot use a simple parametric test to choose between them. However, these two models will surely fit the conditions for the Vuong test described in Section 14.6.6. To implement the test, we first computed

$$V_{it} = \ln f_{it}|\text{geometric} - \ln f_{it}|\text{Poisson}$$

using the respective QMLEs of the parameters. The test statistic given in Section 14.6.6 is then

$$V = \frac{(\sqrt{n})\overline{V}}{s_V}.$$

This statistic converges to standard normal under the underlying assumptions. A large positive value favors the geometric model. The computed sample value is 37.885, which strongly favors the geometric model over the Poisson. Figure 14.6 suggests an explanation for this finding. The very large mass at $DocVis = 0$ is distinctly non-Poisson. This would motivate an extended model such as the negative binomial model, or more likely a two-part model such as the hurdle model examined in Section 18.4.8. The geometric model would likely provide a better fit to a data set such as this one. The three approaches do display a substantive difference. The average partial effects in Table 14.11 differ noticeably for the three specifications.

14.12 SYSTEMS OF REGRESSION EQUATIONS

The general form of the seemingly unrelated regression (SUR) model is given in (10-1) through (10-3),

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, i = 1, \dots, M, \\ E[\boldsymbol{\varepsilon}_i \mid \mathbf{X}_1, \dots, \mathbf{X}_M] &= 0, \\ E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j' \mid \mathbf{X}_1, \dots, \mathbf{X}_M] &= \sigma_{ij}\mathbf{I}. \end{aligned} \quad \textbf{(14-66)}$$

FGLS estimation of this model is examined in detail in Section 10.2.3. We will now add the assumption of normally distributed disturbances to the model and develop the maximum likelihood estimators. This suggests a general approach for multiple equation systems. Given the covariance structure defined in (14-66), the joint normality assumption applies to the vector of M disturbances observed at time t, which we write as

$$\boldsymbol{\varepsilon}_t \mid \mathbf{X}_1, \dots, \mathbf{X}_M \sim N[\mathbf{0}, \boldsymbol{\Sigma}], t = 1, \dots, T. \quad \textbf{(14-67)}$$

14.12.1 THE POOLED MODEL

The pooled model, in which all coefficient vectors are equal, provides a convenient starting point. With the assumption of equal coefficient vectors, the regression model becomes

$$\begin{aligned} &y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, i = 1, \dots, M, t = 1, \dots, T, \\ &E[\boldsymbol{\varepsilon}_{it} \mid \mathbf{X}_1, \dots, \mathbf{X}_M] = 0, \\ &E[\varepsilon_{it}\varepsilon_{js} \mid \mathbf{X}_1, \dots, \mathbf{X}_M] = \sigma_{ij} \quad \text{if} \quad t = s, \quad \text{and} \quad 0 \quad \text{if} \quad t \neq s. \end{aligned} \quad \textbf{(14-68)}$$

This is a model of heteroscedasticity and cross-sectional correlation. With multivariate normality, the log likelihood is

$$\ln L = \sum_{t=1}^{T}\left[-\frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t \right]. \quad \textbf{(14-69)}$$

As we saw earlier, the efficient estimator for this model is GLS, as shown in (10-22). Because the elements of $\boldsymbol{\Sigma}$ must be estimated, the FGLS estimator based on (10-23) and (10-13) is used.

The maximum likelihood estimator of $\boldsymbol{\beta}$, given $\boldsymbol{\Sigma}$, is GLS, based on (10-22). The maximum likelihood estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\sigma}_{ij} = \frac{(\mathbf{y}_i' - \mathbf{X}_i\hat{\hat{\boldsymbol{\beta}}}_{ML})'(\mathbf{y}_j - \mathbf{X}_j\hat{\hat{\boldsymbol{\beta}}}_{ML})}{T} = \frac{\hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_j}{T}, \quad \textbf{(14-70)}$$

based on the MLE of $\boldsymbol{\beta}$. If each MLE requires the other, how can we proceed to obtain both? The answer is provided by **Oberhofer and Kmenta** (1974), who show that for certain models, including this one, one can iterate back and forth between the two estimators. Thus, the MLEs are obtained by iterating to convergence between (14-70) and

$$\hat{\hat{\boldsymbol{\beta}}} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}]. \quad \textbf{(14-71)}$$

The process may begin with the (consistent) ordinary least squares estimator, then (14-70), and so on. The computations are simple, using basic matrix algebra. Hypothesis tests about $\boldsymbol{\beta}$ may be done using the familiar Wald statistic. The appropriate estimator of the asymptotic covariance matrix is the inverse matrix in brackets in (10-22).

For testing the hypothesis that the off-diagonal elements of $\boldsymbol{\Sigma}$ are zero—that is, that there is no correlation across groups—there are three approaches. The likelihood ratio test is based on the statistic

$$\lambda_{\text{LR}} = T(\ln|\hat{\boldsymbol{\Sigma}}_{heteroscedastic}| - \ln|\hat{\boldsymbol{\Sigma}}_{general}|) = T\left(\sum_{i=1}^{M}\ln\hat{\sigma}_i^2 - \ln|\hat{\boldsymbol{\Sigma}}|\right), \quad \textbf{(14-72)}$$

where $\hat{\sigma}_i^2$ are the estimates of σ_i^2 obtained from the maximum likelihood estimates of the groupwise heteroscedastic model and $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimator in the unrestricted model.[32] The large-sample distribution of the statistic is chi squared with $M(M-1)/2$ degrees of freedom. The Lagrange multiplier test developed by Breusch and Pagan (1980) provides an alternative. The general form of the statistic is

$$\lambda_{\text{LM}} = T\sum_{i=2}^{m}\sum_{j=1}^{i-1} r_{ij}^2, \quad \textbf{(14-73)}$$

where r_{ij}^2 is the ijth residual correlation coefficient. If every equation had a different parameter vector, then equation-specific ordinary least squares would be efficient (and ML) and we would compute r_{ij} from the OLS residuals (assuming that there are sufficient observations for the computation). Here, however, we are assuming only a single-parameter vector. Therefore, the appropriate basis for computing the correlations is the residuals from the iterated estimator in the groupwise heteroscedastic model, that is, the same residuals used to compute $\hat{\sigma}_i^2$. (An asymptotically valid approximation to the test can be based on the FGLS residuals instead.) Note that this is not a procedure for testing all the way down to the homoscedastic regression model. That case involves different LM and LR statistics based on the groupwise heteroscedasticity model. If either the LR statistic in (14-72) or the LM statistic in (14-73) is smaller than the critical value from the table, the conclusion, based on this test, is that the appropriate model is the groupwise heteroscedastic model.

14.12.2 THE SUR MODEL

The Oberhofer–Kmenta (1974) conditions are met for the seemingly unrelated regressions model, so maximum likelihood estimates can be obtained by iterating the FGLS procedure. We note, once again, that this procedure presumes the use of (10-11) for estimation of σ_{ij} at each iteration. Maximum likelihood enjoys no advantages over FGLS in its asymptotic properties.[33] Whether it would be preferable in a small sample is an open question whose answer will depend on the particular data set.

[32] *Note:* The excess variation produced by the restrictive model is used to construct the test.

[33] Jensen (1995) considers some variation on the computation of the asymptotic covariance matrix for the estimator that allows for the possibility that the normality assumption might be violated.

Example 14.14 ML Estimates of a Seemingly Unrelated Regressions Model

Although a bit dated, the Grunfeld data used in Application 11.2 have withstood the test of time and are still a standard data set used to demonstrate the SUR model. The data in Appendix Table F10.4 are for 10 firms and 20 years (1935–1954). For the purpose of this illustration, we will use the first four firms.[34]

The model is an investment equation,

$$I_{it} = \beta_{1i} + \beta_{2i}F_{it} + \beta_{3i}C_{it} + \varepsilon_{it}, t = 1, \ldots, 20, i = 1, \ldots, 10,$$

where

$$I_{it} = \text{real gross investment for firm } i \text{ in year } t,$$
$$F_{it} = \text{real value of the firm-shares outstanding,}$$
$$C_{it} = \text{real value of the capital stock.}$$

The OLS estimates for the four equations are shown in the left panel of Table 14.12. The correlation matrix for the four OLS residual vectors is

$$\mathbf{R_e} = \begin{bmatrix} 1 & -0.261 & 0.279 & -0.273 \\ -0.261 & 1 & 0.428 & 0.338 \\ 0.279 & 0.428 & 1 & -0.0679 \\ -0.273 & 0.338 & -0.0679 & 1 \end{bmatrix}.$$

Before turning to the FGLS and MLE estimates, we carry out the LM test against the null hypothesis that the regressions are actually unrelated. We leave as an exercise to show that the LM statistic in (14-73) can be computed as

$$\lambda_{LM} = (T/2)[\text{trace}(\mathbf{R_e'R_e}) - M] = 10.451.$$

The 95% critical value from the chi-squared distribution with 6 degrees of freedom is 12.59, so at this point, it appears that the null hypothesis is not rejected. We will proceed in spite of this finding.

TABLE 14.12 Estimated Investment Equations

		OLS		*FGLS*		*MLE*	
Firm	***Variable***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
	Constant	−149.78	97.58	−160.68	90.41	−179.41	86.66
1	*F*	0.1192	0.02382	0.1205	0.02187	0.1248	0.02086
	C	0.3714	0.03418	0.3800	0.03311	0.3802	0.03266
	Constant	−49.19	136.52	21.16	116.18	36.46	106.18
2	*F*	0.1749	0.06841	0.1304	0.05737	0.1244	0.05191
	C	0.3896	0.1312	0.4485	0.1225	0.4367	0.1171
	Constant	−9.956	28.92	−19.72	26.58	−24.10	25.80
3	*F*	0.02655	0.01435	0.03464	0.01279	0.03808	0.01217
	C	0.1517	0.02370	0.1368	0.02249	0.1311	0.02223
	Constant	−6.190	12.45	0.9366	11.59	2.581	11.54
4	*F*	0.07795	0.01841	0.06785	0.01705	0.06564	0.01698
	C	0.3157	0.02656	0.3146	0.02606	0.3137	0.02617

[34]The data are downloaded from the Web site for Baltagi (2005) at www.wiley.com/legacy/wileychi/baltagi/supp/Grunfeld.fil. See also Kleiber and Zeileis (2010).

The next step is to compute the covariance matrix for the OLS residuals using

$$\mathbf{W} = (1/T)\,\mathbf{E}'\mathbf{E} = \begin{bmatrix} \mathbf{7160.29} & -1967.05 & 607.533 & -282.756 \\ -1967.05 & \mathbf{7904.66} & 978.45 & 367.84 \\ 607.533 & 978.45 & \mathbf{660.829} & -21.3757 \\ -282.756 & 367.84 & -21.3757 & \mathbf{149.872} \end{bmatrix},$$

where **E** is the 20 × 4 matrix of OLS residuals. Stacking the data in the partitioned matrices,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_4 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{bmatrix},$$

we now compute $\hat{\boldsymbol{\Omega}} = \mathbf{W} \otimes \mathbf{I}_{20}$ and the FGLS estimates,

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}.$$

The estimated asymptotic covariance matrix for the FGLS estimates is the bracketed inverse matrix. These results are shown in the center panel in Table 14.12. To compute the MLE, we will take advantage of the Oberhofer and Kmenta (1974) result and iterate the FGLS estimator. Using the FGLS coefficient vector, we recompute the residuals, then recompute **W**, then reestimate $\boldsymbol{\beta}$. The iteration is repeated until the estimated parameter vector converges. We use as our convergence measure the following criterion based on the change in the estimated parameter from iteration $(s - 1)$ to iteration $(\mathbf{s})$:

$$\delta = [\hat{\boldsymbol{\beta}}(s) - \hat{\boldsymbol{\beta}}(s - 1)]'[\mathbf{X}'[\hat{\boldsymbol{\Omega}}(s)]^{-1}\mathbf{X}][\hat{\boldsymbol{\beta}}(s) - \hat{\boldsymbol{\beta}}(s - 1)].$$

The sequence of values of this criterion function are: 0.21922, 0.16318, 0.00662, 0.00037, 0.00002367825, 0.000001563348, 0.1041980×10^{-6}. We exit the iterations after iteration 7. The ML estimates are shown in the right panel of Table 14.12. We then carry out the likelihood ratio test of the null hypothesis of a diagonal covariance matrix. The maximum likelihood estimate of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \mathbf{7235.46} & -2455.13 & 615.167 & -325.413 \\ -2455.13 & \mathbf{8146.41} & 1288.66 & 427.011 \\ 615.167 & 1288.66 & \mathbf{702.268} & 2.51786 \\ -325.413 & 427.011 & 2.51786 & \mathbf{153.889} \end{bmatrix}.$$

The estimate for the constrained model is the diagonal matrix formed from the diagonals of **W** shown earlier for the OLS results. (The estimates are shown in boldface in the preceding matrix, **W**.) The test statistic is then

$$\text{LR} = T(\ln|\text{diag}(\mathbf{W})| - \ln|\hat{\boldsymbol{\Sigma}}|) = 18.55.$$

Recall that the critical value is 12.59. The results contradict the LM statistic. The hypothesis of diagonal covariance matrix is now rejected.

Note that aside from the constants, the four sets of coefficient estimates are fairly similar. Because of the constants, there seems little doubt that the pooling restriction will be rejected. To find out, we compute the Wald statistic based on the MLE results. For testing

$$H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{\beta}_4,$$

we can formulate the hypothesis as

$$H_0: \boldsymbol{\beta}_1 - \boldsymbol{\beta}_4 = \mathbf{0}, \boldsymbol{\beta}_2 - \boldsymbol{\beta}_4 = \mathbf{0}, \boldsymbol{\beta}_3 - \boldsymbol{\beta}_4 = \mathbf{0}.$$

The Wald statistic is

$$\lambda_W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{RVR}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = 2190.96,$$

where $\mathbf{R} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0} & \mathbf{0} & -\mathbf{I}_3 \\ \mathbf{0} & \mathbf{I}_3 & \mathbf{0} & -\mathbf{I}_3 \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_3 & -\mathbf{I}_3 \end{bmatrix}$, $\mathbf{q} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$, and $\mathbf{V} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}$. Under the null hypothesis, the Wald statistic has a limiting chi-squared distribution with 9 degrees of freedom. The critical value is 16.92, so, as expected, the hypothesis is rejected. It may be that the difference is due to the different constant terms. To test the hypothesis that the four pairs of slope coefficients are equal, we replaced the $\mathbf{I}_3$ in $\mathbf{R}$ with $[[\mathbf{0}, \mathbf{I}_2]]$, the $\mathbf{0}$'s with 2×3 zero matrices, and $\mathbf{q}$ with a 6×1 zero vector. The resulting chi-squared statistic equals 229.005. The critical value is 12.59, so this hypothesis is rejected as well.

14.13 SIMULTANEOUS EQUATIONS MODELS

In Chapter 10, we noted two approaches to maximum likelihood estimation of the equation
system,

$$\mathbf{y}_t'\boldsymbol{\Gamma} + \mathbf{x}_t'\mathbf{B} = \boldsymbol{\varepsilon}_t',$$
$$\boldsymbol{\varepsilon}_t|\mathbf{X} \sim N[\mathbf{0}, \boldsymbol{\Sigma}]: \tag{14-73}$$

full information maximum likelihood (FIML) and limited information maximum likelihood (LIML). The FIML approach simultaneously estimates all model parameters. The FIML estimator for a linear equation system is extremely complicated both theoretically and practically. However, its asymptotic properties are identical to three-stage least squares (3SLS), which is straightforward and a standard feature of modern econometric software. (See Section 10.4.5.) Thus, the additional assumption of normality in the system brings no theoretical or practical advantage.

The LIML estimator is a single-equation approach that estimates the parameters of the model one equation at a time. We examined two approaches to computing the LIML estimator, both straightforward, when the equations are linear. The least variance ratio approach shown in Section 10.4.4 is based on some basic matrix algebra calculations—the only unconventional calculation involves the characteristic roots of an asymmetric matrix (or obtaining the matrix square root of a symmetric matrix). The more direct approach in Section (8.4.3) provides some useful results for interpreting the model.

The leading application of LIML estimation is for an equation that contains one endogenous variable. (This is the application in most of Chapter 8.) Let that be the first equation in (14-73),

$$y_1\gamma_{11} + y_2\gamma_{21} + \mathbf{x}_1'\mathbf{b}_1 = \varepsilon_1.$$

Normalize the equation, so the coefficient on y_1 is 1 and the other variables apprear on the right-hand side. Then,

$$y_1 = y_2\delta_1 + \mathbf{x}_1{}'\boldsymbol{\beta}_1 + w_1. \tag{14-74}$$

This is the structural form for the first equation that contains a single included endogenous variable. The reduced form for the entire system is $\mathbf{y}' = \mathbf{x}'(-\mathbf{B}\boldsymbol{\Gamma}^{-1}) + \mathbf{v}'$. [See Section 10.4.2 and (10-36).] The second equation in the reduced form is

$$y_2 = \mathbf{x}'\boldsymbol{\pi}_2 + u_2. \tag{14-75}$$

Note that the structural equation for y_1 involves only some of the exogenous variables in the system while the reduced form involves all of them including at least one that is not contained in $\mathbf{x}_1$. As we developed in Section 10.4.3, there must be exogenous variables in the system that are excluded from the y_1 equation—this is the order condition for identification. The disturbances in the two equations are linear functions of the disturbances in (14-73), so with normality, the disturbances in (14-74) and (14-75) are joint normal.

The two-equation system (14-74,14-75) is precisely the same as the one we examined in Section 8.4.3,

$$y = \mathbf{x}_1{}'\boldsymbol{\beta} + x_2\lambda + \varepsilon \tag{14-76}$$

$$x_2 = \mathbf{z}'\boldsymbol{\gamma} + u, \tag{14-77}$$

where y_2 in (14-74) is the x_2 in (14-76) and $\mathbf{z} = (\mathbf{x}_1, \ldots)$. Equation (14-77) is the reduced form equation for y_2. This formalizes the results for an equation in a simultaneous equations model that contains one endogenous variable. The estimator is actually based on two equations, the structural equation of interest and the reduced form for the endogenous variable that appears in that equation. The log-likelihood function for the LIML estimator for this (actually) two-equation system is shown in (8-17). In the typical equation, (14-76) and (14-77) might well be the recursive structure. This construction of the model underscores the point that in a model that contains an endogenous variable, there is a second equation that "explains" the endogeneity.

For the practitioner, a useful result is that the asymptotic variance of the two-stage least squares (2SLS) estimator is the same as that of the LIML estimator. This would generally render the LIML estimator, with its additional normality assumption, moot. The exception would be the invariance of the LIML estimator to normalization of the equation (i.e., which variable appears on the left of the equals sign). This turns out to be useful in the context of analysis in the presence of weak instruments. (See Section 8.7.) More generally, the LIML and FIML estimators have been supplanted in the literature by much simpler GMM estimators, 2SLS, 3SLS, and extensions that accommodate heteroscedasticity. Interest remains in these estimators, but largely as a component of the ongoing theoretical research.

14.14 PANEL DATA APPLICATIONS

Application of panel data methods to the linear panel data models we have considered so far is a fairly marginal extension. For the random effects linear model, considered in the following Section 14.14.1, the MLE of $\boldsymbol{\beta}$ is, as always, FGLS given the MLEs of the variance parameters. The latter produce a fairly substantial complication, as we shall see. This extension does provide a convenient, interesting application to see the payoff to the invariance property of the MLE—we will reparameterize a fairly complicated log-likelihood function to turn it into a simple one. Where the method of maximum likelihood becomes essential is in analysis of fixed and random effects in nonlinear models. We will develop two general methods for handling these situations in generic terms in Sections 14.14.3 and 14.14.4, then apply them in several models later in the book.

14.14.1 ML ESTIMATION OF THE LINEAR RANDOM EFFECTS MODEL

The contribution of the ith individual to the log likelihood for the random effects model [(11-28) to (11-32)] with normally distributed disturbances is

$$\begin{aligned}\ln L_i(\boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_u^2) &= -\frac{1}{2}[T_i \ln 2\pi + \ln|\boldsymbol{\Omega}_i| + (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})] \\ &= -\frac{1}{2}[T_i \ln 2\pi + \ln|\boldsymbol{\Omega}_i| + \boldsymbol{\varepsilon}_i'\boldsymbol{\Omega}_i^{-1}\boldsymbol{\varepsilon}_i],\end{aligned} \quad \textbf{(14-78)}$$

where

$$\boldsymbol{\Omega}_i = \sigma_\varepsilon^2\mathbf{I}_{Ti} + \sigma_u^2\mathbf{ii}',$$

and $\mathbf{i}$ denotes a $T_i \times 1$ column of ones. Note that the $\boldsymbol{\Omega}_i$ varies over i because it is $T_i \times T_i$. Baltagi (2013) presents a convenient and compact estimator for this model that involves iteration between an estimator of $\phi^2 = [\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + T\sigma_u^2)]$, based on sums of squared residuals, and $(\alpha, \boldsymbol{\beta}, \sigma_\varepsilon^2)$ (α is the constant term) using FGLS. Unfortunately, the convenience and compactness come unraveled in the unbalanced case. We consider, instead, what Baltagi labels a "brute force" approach, that is, direct maximization of the log-likelihood function in (14-78). (See, Baltagi, pp. 169–170.)

Using (A-66), we find that

$$\boldsymbol{\Omega}_i^{-1} = \frac{1}{\sigma_\varepsilon^2}\left[\mathbf{I}_{T_i} - \frac{\sigma_u^2}{\sigma_\varepsilon^2 + T_i\sigma_u^2}\mathbf{ii}'\right].$$

We will also need the determinant of $\boldsymbol{\Omega}_i$. To obtain this, we will use the product of its characteristic roots. First, write

$$|\boldsymbol{\Omega}_i| = (\sigma_\varepsilon^2)^{T_i}|\mathbf{I} + \gamma\mathbf{ii}'|,$$

where $\gamma = \sigma_u^2/\sigma_\varepsilon^2$. To find the characteristic roots of the matrix, use the definition

$$[\mathbf{I} + \gamma\mathbf{ii}']\mathbf{c} = \lambda\mathbf{c},$$

where $\mathbf{c}$ is a characteristic vector and λ is the associated characteristic root. The equation implies that $\gamma\mathbf{ii}'\mathbf{c} = (\lambda - 1)\mathbf{c}$. Premultiply by $\mathbf{i}'$ to obtain $\gamma(\mathbf{i}'\mathbf{i})(\mathbf{i}'\mathbf{c}) = (\lambda - 1)(\mathbf{i}'\mathbf{c})$. Any vector $\mathbf{c}$ with elements that sum to zero will satisfy this equality. There will be $T_i - 1$ such vectors and the associated characteristic roots will be $(\lambda - 1) = 0$ or $\lambda = 1$. For the remaining root, divide by the nonzero $(\mathbf{i}'\mathbf{c})$ and note that $\mathbf{i}'\mathbf{i} = T_i$, so the last root is $T_i\gamma = \lambda - 1$ or $\lambda = (1 + T_i\gamma)$.[35] It follows that the log of the determinant is

$$\ln|\boldsymbol{\Omega}_i| = T_i \ln \sigma_\varepsilon^2 + \ln(1 + T_i\gamma).$$

[35]By this derivation, we have established a useful general result. The characteristic roots of a $T \times T$ matrix of the form $\mathbf{A} = (\mathbf{I} + a\mathbf{bb}')$ are 1 with multiplicity $(T - 1)$ and $a\mathbf{b}'\mathbf{b}$ with multiplicity 1. The proof follows precisely along the lines of our earlier derivation.

Expanding the parts and multiplying out the third term gives the log-likelihood function

$$\ln L = \sum_{i=1}^{n} \ln L_i$$
$$= -\frac{1}{2}\left[(\ln 2\pi + \ln \sigma_\varepsilon^2)\sum_{i=1}^{n} T_i + \sum_{i=1}^{n}\ln(1 + T_i\gamma)\right] - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i - \frac{\sigma_u^2(T_i\bar{\varepsilon}_i)^2}{\sigma_\varepsilon^2 + T_i\sigma_u^2}\right].$$

Note that in the third term, we can write $\sigma_\varepsilon^2 + T_i\sigma_u^2 = \sigma_\varepsilon^2(1 + T_i\gamma)$ and $\sigma_u^2 = \sigma_\varepsilon^2\gamma$. After inserting these, two appearances of σ_ε^2 in the square brackets will cancel, leaving

$$\ln L = -\frac{1}{2}\sum_{i=1}^{n}\left(T_i(\ln 2\pi + \ln \sigma_\varepsilon^2) + \ln(1 + T_i\gamma) + \frac{1}{\sigma_\varepsilon^2}\left[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i - \frac{\gamma(T_i\bar{\varepsilon}_i)^2}{1 + T_i\gamma}\right]\right).$$

Now, let $\theta = 1/\sigma_\varepsilon^2$, $R_i = 1 + T_i\gamma$, and $Q_i = \gamma/R_i$. The individual contribution to the log likelihood becomes

$$\ln L_i = -\frac{1}{2}[\theta(\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i - Q_i(T_i\bar{\varepsilon}_i)^2) + \ln R_i - T_i \ln \theta + T_i \ln 2\pi].$$

The likelihood equations are

$$\frac{\partial \ln L_i}{\partial \boldsymbol{\beta}} = \theta\left[\sum_{t=1}^{T_i}\mathbf{x}_{it}\varepsilon_{it}\right] - \theta\left[Q_i\left(\sum_{t=1}^{T_i}\mathbf{x}_{it}\right)\left(\sum_{t=1}^{Ti}\varepsilon_{it}\right)\right],$$
$$\frac{\partial \ln L_i}{\partial \theta} = -\frac{1}{2}\left[\left(\sum_{t=1}^{T_i}\varepsilon_{it}^2\right) - Q_i\left(\sum_{t=1}^{T_i}\varepsilon_{it}\right)^2 - \frac{T_i}{\theta}\right],$$
$$\frac{\partial \ln L_i}{\partial \gamma} = \frac{1}{2}\left[\theta\left(\frac{1}{R_i^2}\left(\sum_{t=1}^{T_i}\varepsilon_{it}\right)^2\right) - \frac{T_i}{R_i}\right].$$

These will be sufficient for programming an optimization algorithm such as DFP or BFGS. (See Section E3.3.) We could continue to derive the second derivatives for computing the asymptotic covariance matrix, but this is unnecessary. For $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, we know that because this is a generalized regression model, the appropriate asymptotic covariance matrix is

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{\text{MLE}}] = \left[\sum_{i=1}^{n}\mathbf{X}_i'\hat{\boldsymbol{\Omega}}_i^{-1}\mathbf{X}_i\right]^{-1}.$$

(See Section 11.5.2.) We also know that the MLEs of the variance components estimators will be asymptotically uncorrelated with the MLE of $\boldsymbol{\beta}$. In principle, we could continue to estimate the asymptotic variances of the MLEs of σ_ε^2 and σ_u^2. It would be necessary to derive these from the estimators of θ and γ, which one would typically do in any event. However, statistical inference about the disturbance variance, σ_ε^2, in a regression model, is typically of no interest. On the other hand, one might want to test the hypothesis that σ_u^2 equals zero, or $\gamma = 0$. Breusch and Pagan's (1979) LM statistic in (11-42) extended to the unbalanced panel case considered here would be

$$LM = \frac{\left(\sum_{i=1}^{N} T_i\right)^2}{\left[2\sum_{i=1}^{N} T_i(T_i - 1)\right]}\left[\frac{\sum_{i=1}^{N}(T_i\bar{e}_i)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T_i} e_{it}^2} - 1\right]^2$$

$$= \frac{\left(\sum_{i=1}^{N} T_i\right)^2}{\left[2\sum_{i=1}^{N} T_i(T_i - 1)\right]}\left[\frac{\sum_{i=1}^{N}[(T_i\bar{e}_i)^2 - \mathbf{e}_i'\mathbf{e}_i]}{\sum_{i=1}^{N}\mathbf{e}_i'\mathbf{e}_i}\right]^2.$$

Example 14.15 Maximum Likelihood and FGLS Estimates of A Wage Equation

Example 11.11 presented FGLS estimates of a wage equation using Cornwell and Rupert's panel data. We have reestimated the wage equation using maximum likelihood instead of FGLS. The parameter estimates appear in Table 14.13, with the FGLS and pooled OLS estimates. The estimates of the variance components are shown in the table as well. The similarity of the MLEs and FGLS slope estimates is to be expected given the large sample size. The difference in the estimates of σ_u is perhaps surprising. The estimator is not based on a simple sum of squares, however, so this kind of variation is common. The LM statistic for testing for the presence of the common effects is 3,497.02, which is far larger than the critical value of 3.84. With the MLE, we can also use an LR test to test for random effects against the null hypothesis of no effects. The chi-squared statistic based on the two log likelihoods is 3,662.25, which leads to the same conclusion.

TABLE 14.13 Wage Equation Estimated by FGLS and MLE

Variable	*Least Squares Estimate*	*Clustered Std. Error*	*Random Effects FGLS*	*Standard Error*	*Random Effects MLE*	*Standard Error*
Constant	5.25112	0.12355	4.04144	0.08330	3.12622	0.17761
Exp	0. 04010	0.00408	0.08748	0.00225	0.10721	0.00248
ExpSq	−0.00067	0.00009	−0.00076	0.00005	−0.00051	0.00005
Wks	0.00422	0.00154	0.00096	0.00059	0.00084	0.00060
Occ	−0.14001	0.02724	−0.04322	0.01299	−0.02512	0.01378
Ind	0.04679	0.02366	0.00378	0.01373	0.01380	0.01529
South	−0.05564	0.02616	−0.00825	0.02246	0.00577	0.03159
SMSA	0.15167	0.02410	−0.02840	0.01616	−0.04748	0.01896
MS	0.04845	0.04094	−0.07090	0.01793	−0.04138	0.01899
Union	0.09263	0.02367	0.05835	0.01350	0.03873	0.01481
Ed	0.05670	0.00556	0.10707	0.00511	0.13562	0.01267
Fem	−0.36779	0.04557	−0.30938	0.04554	−0.17562	0.11310
Blk	−0.16694	0.04433	−0.21950	0.05252	−0.26121	0.13747
θ					42.5265	
γ					29.9705	
σ_ε	0.34936		0.15206		0.15335	
σ_u	0.00000		0.31453		0.83949	

14.14.2 NESTED RANDOM EFFECTS

Consider a data set on test scores for multiple school districts in a state. To establish a notation for this complex model, we define a four-level unbalanced structure,

$$\begin{aligned} Z_{ijkt} &= \text{test score for student } t, \text{ teacher } k, \text{ school } j, \text{ district } i, \\ L &= \text{school districts}, i = 1, \ldots, L, \\ M_i &= \text{schools in each district}, j = 1, \ldots, M_i, \\ N_{ij} &= \text{teachers in each school}, k = 1, \ldots, N_{ij}, \\ T_{ijk} &= \text{students in each class}, t = 1, \ldots, T_{ijk}. \end{aligned}$$

Thus, from the outset, we allow the model to be unbalanced at all levels. In general terms, then, the random effects regression model would be

$$y_{ijkt} = \mathbf{x}'_{ijkt}\boldsymbol{\beta} + u_{ijk} + v_{ij} + w_i + \varepsilon_{ijkt}.$$

Strict exogeneity of the regressors is assumed at all levels. All parts of the disturbance are also assumed to be uncorrelated. (A normality assumption will be added later as well.) From the structure of the disturbances, we can see that the overall covariance matrix, $\boldsymbol{\Omega}$, is block diagonal over i, with each diagonal block itself block diagonal in turn over j, each of these is block diagonal over k, and, at the lowest level, the blocks, for example, for the class in our example, have the form for the random effects model that we saw earlier.

Generalized least squares has been well worked out for the balanced case.[36] Define the following to be constructed from the variance components, σ_ε^2, σ_u^2, σ_v^2, and σ_w^2:

$$\begin{aligned} \sigma_1^2 &= T\sigma_u^2 + \sigma_\varepsilon^2, \\ \sigma_2^2 &= NT\sigma_v^2 + T\sigma_u^2 + \sigma_\varepsilon^2 = \sigma_1^2 + NT\sigma_v^2, \\ \sigma_3^2 &= MNT\sigma_w^2 + NT\sigma_v^2 + T\sigma_u^2 + \sigma_\varepsilon^2 = \sigma_2^2 + MNT\sigma_w^2. \end{aligned}$$

Then, full generalized least squares is equivalent to OLS regression of

$$\widetilde{y}_{ijkt} = y_{ijkt} - \left(1 - \frac{\sigma_\varepsilon}{\sigma_1}\right)\overline{y}_{ijk\cdot} - \left(\frac{\sigma_\varepsilon}{\sigma_1} - \frac{\sigma_\varepsilon}{\sigma_2}\right)\overline{y}_{ij\cdot\cdot} - \left(\frac{\sigma_\varepsilon}{\sigma_2} - \frac{\sigma_\varepsilon}{\sigma_3}\right)\overline{y}_{i\cdots}$$

on the same transformation of $\mathbf{x}_{ijkt}$. FGLS estimates are obtained by three groupwise between estimators and the within estimator for the innermost grouping.

The counterparts for the unbalanced case can be derived, but the degree of complexity rises dramatically.[37] As Antwiler (2001) shows, however, if one is willing to assume normality of the distributions, then the log likelihood is very tractable. (We note an intersection of practicality with nonrobustness.) Define the variance ratios

$$\rho_u = \frac{\sigma_u^2}{\sigma_\varepsilon^2}, \rho_v = \frac{\sigma_v^2}{\sigma_\varepsilon^2}, \rho_w = \frac{\sigma_w^2}{\sigma_\varepsilon^2}.$$

Construct the following intermediate results

$$\theta_{ijk} = 1 + T_{ijk}\rho_u, \phi_{ij} = \sum_{k=1}^{N_{ij}} \frac{T_{ijk}}{\theta_{ijk}}, \theta_{ij} = 1 + \phi_{ij}\rho_v, \phi_i = \sum_{j=1}^{M_i} \frac{\phi_{ij}}{\theta_{ij}}, \theta_i = 1 + \rho_w\phi_i$$

[36] See, for example, Baltagi, Song, and Jung (2001), who also provide results for the three-level unbalanced case.

[37] See Baltagi et al. (2001).

and sums of squares of the disturbances $e_{ijkt} = y_{ijkt} - \mathbf{x}'_{ijkt}\boldsymbol{\beta}$,

$$A_{ijk} = \sum_{t=1}^{T_{ijk}} e_{ijkt}^2,$$

$$B_{ijk} = \sum_{t=1}^{T_{ijk}} e_{ijkt},\ B_{ij} = \sum_{k=1}^{N_{ij}} \frac{B_{ijk}}{\theta_{ijk}},\ B_i = \sum_{j=1}^{M_i} \frac{B_{ij}}{\theta_{ij}}.$$

The log likelihood is

$$\ln L = -\frac{1}{2}H\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2}\Bigg[\sum_{i=1}^{L}\Bigg\{\ln\theta_i + \sum_{j=1}^{M_i}\Bigg\{\ln\theta_{ij} + \sum_{k=1}^{N_{ij}}{}'$$
$$\Bigg\{\ln\theta_{ijk} + \frac{A_{ijk}}{\sigma_\varepsilon^2} - \frac{\rho_u}{\theta_{ijk}}\frac{B_{ijk}^2}{\sigma_\varepsilon^2}\Bigg\} - \frac{\rho_v}{\theta_{ij}}\frac{B_{ij}^2}{\sigma_\varepsilon^2}\Bigg\} - \frac{\rho_w}{\theta_i}\frac{B_i^2}{\sigma_\varepsilon^2}\Bigg\}\Bigg],$$

where H is the total number of observations. (For three levels, $L = 1$ and $\rho_w = 0$.) Antwiler (2001) provides the first derivatives of the log-likelihood function needed to maximize $\ln L$. However, he does suggest that the complexity of the results might make numerical differentiation attractive. On the other hand, he finds the second derivatives of the function intractable and resorts to numerical second derivatives in his application. The complex part of the Hessian is the cross derivatives between $\boldsymbol{\beta}$ and the variance parameters, and the lower-right part for the variance parameters themselves. However, these are not needed. As in any generalized regression model, the variance estimators and the slope estimators are asymptotically uncorrelated. As such, one need only invert the part of the matrix with respect to $\boldsymbol{\beta}$ to get the appropriate asymptotic covariance matrix. The relevant block is

$$-\frac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}}\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\mathbf{x}'_{ijkt} - \frac{\rho_w}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\right)\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\right)$$
$$-\frac{\rho_v}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\right)\right)\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\right)\right) \qquad \textbf{(14-79)}$$
$$-\frac{\rho_u}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\left(\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\right)\right)\right)\left(\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\right)\right)\right).$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is FGLS based on the maximum likelihood estimators of the variance parameters. Thus, expression (14-79) provides the appropriate covariance matrix for the GLS or maximum likelihood estimator. The difference will be in how the variance components are computed. Baltagi et al. (2001) suggest a variety of methods for the three-level model. For more than three levels, the MLE becomes more attractive.

Eample 14.16 Statewide Productivity

Munnell (1990) analyzed the productivity of public capital at the state level using a Cobb–Douglas production function. We will use the data from that study to estimate a three-level log linear regression model,

$$\ln gsp_{jkt} = \alpha + \beta_1 \ln pc_{jkt} + \beta_2 \ln hwy_{jkt} + \beta_3 \ln water_{jkt}$$
$$+ \beta_4 \ln util_{jkt} + \beta_5 \ln emp_{jkt} + \beta_6\, unemp_{jkt} + \varepsilon_{jkt} + u_{jk} + v_j,$$
$$j = 1, \ldots, 9;\ t = 1, \ldots, 17,\ k = 1, \ldots, N_j,$$

where the variables in the model are

$$\begin{aligned} gsp &= \text{gross state product}, \\ p\_cap &= \text{public capital} = \text{hwy} + \text{water} + \text{util}, \\ hwy &= \text{highway capital}, \\ water &= \text{water utility capital}, \\ util &= \text{utility capital}, \\ pc &= \text{private capital}, \\ emp &= \text{employment (labor)}, \\ unemp &= \text{unemployment rate}, \end{aligned}$$

and we have defined $M = 9$ regions each consisting of a group of the 48 contiguous states:

$$\begin{aligned} Gulf &= \text{AL, FL, LA, MS}, \\ Midwest &= \text{IL, IN, KY, MI, MN, OH, WI}, \\ Mid\ Atlantic &= \text{DE, MD, NJ, NY, PA, VA}, \\ Mountain &= \text{CO, ID, MT, ND, SD, WY}, \\ New\ England &= \text{CT, ME, MA, NH, RI, VT}, \\ South &= \text{GA, NC, SC, TN, WV}, \\ Southwest &= \text{AZ, NV, NM, TX, UT}, \\ Tornado\ Alley &= \text{AR, IA, KS, MO, NE, OK} \\ West\ Coast &= \text{CA, OR, WA}. \end{aligned}$$

We have 17 years of data, 1970 to 1986, for each state."[38] The two- and three-level random effects models were estimated by maximum likelihood. The two-level model was also fit by FGLS, using the methods developed in Section 11.5.3.

Table 14.14 presents the estimates of the production function using pooled OLS, OLS for the fixed effects model, and both FGLS and maximum likelihood for the random effects models. Overall, the estimates are similar, though the OLS estimates do stand somewhat apart. This suggests, as one might suspect, that there are omitted effects in the pooled model. The F statistic for testing the significance of the fixed effects is 76.712 with 47 and 762 degrees of freedom. The critical value from the table is 1.379, so on this basis, one would reject the hypothesis of no common effects. Note, as well, the extremely large differences between the conventional OLS standard errors and the robust (cluster) corrected values. The three- or four-fold differences strongly suggest that there are latent effects at least at the regional level. It remains to consider which approach, fixed or random effects, is preferred. The Hausman test for fixed vs. random effects produces a chi-squared value of 18.987. The critical value is 12.592. This would imply that the fixed effects model would be the preferred specification. When we repeat the calculation of the Hausman statistic using the three-level estimates in the last column of Table 14.14, the statistic falls slightly to 15.327. Finally, note the similarity of all three sets of random effects estimates. In fact, under the hypothesis of mean independence, all three are consistent estimators. It is tempting at this point to carry out a likelihood ratio test of the hypothesis of the two-level model against the broader alternative three-level model. The test statistic would be twice the difference of the log-likelihoods, which is 2.46. For one degree of freedom, the critical chi squared with one degree of freedom is 3.84, so on this basis, we would not reject the hypothesis of the two-level model. We note, however, that there is a problem with this testing procedure. The hypothesis that a variance is zero is not well defined for the likelihood ratio test—the parameter under the null hypothesis is on the boundary of the parameter space ($\sigma_v^2 \geq 0$). In this instance, the familiar distribution theory does not apply. The results of Kodde and Palm (1988) in Example 14.8 can be used instead of the standard test.

[38]The data were downloaded from the Web site for Baltagi (2005) at www.wiley.com/legacy/wileychi/baltagi3e/. See Appendix Table F10.1.

TABLE 14.14 Estimated Statewide Production Function

	OLS		Fixed Effects	Random Effects FGLS	Random Effects ML	Nested Random Effects
	Estimate	*Std. Err.*[a]	*Estimate (Std. Err.)*	*Estimate (Std. Err.)*	*Estimate (Std. Err.)*	*Estimate (Std. Err.)*
α	1.9260	0.05250		2.1608	2.1759	2.1348
		(0.2143)		(0.1380)	(0.1477)	(0.1514)
β_1	0.3120	0.01109	0.2350	0.2755	0.2703	0.2724
		(0.04678)	(0.02621)	(0.01972)	(0.02110)	(0.02141)
β_2	0.05888	0.01541	0.07675	0.06167	0.06268	0.06645
		(0.05078)	(0.03124)	(0.02168)	(0.02269)	(0.02287)
β_3	0.1186	0.01236	0.0786	0.07572	0.07545	0.07392
		(0.03450)	(0.0150)	(0.01381)	(0.01397)	(0.01399)
β_4	0.00856	0.01235	−0.11478	−0.09672	−0.1004	−0.1004
		(0.04062)	(0.01814)	(0.01683)	(0.01730)	(0.01698)
β_5	0.5497	0.01554	0.8011	0.7450	0.7542	0.7539
		(0.06770)	(0.02976)	(0.02482)	(0.02664)	(0.02613)
β_6	−0.00727	0.001384	−0.005179	−0.005963	−0.005809	−0.005878
		(0.002946)	(0.000980)	(0.0008814)	(0.0009014)	(0.0009002)
σ_ε	0.085422		0.03676493	0.0367649	0.0366974	0.0366964
σ_u				0.0771064	0.0875682	0.0791243
σ_v						0.0386299
ln L	853.1372		1565.501		1429.075	1430.30576

[a]Robust (cluster) standard errors in parentheses. The covariance matrix is multiplied by a degrees of freedom correction, $nT/(nT - k) = 816/810$.

14.14.3 CLUSTERING OVER MORE THAN ONE LEVEL

Given the complexity of (14-79), one might prefer simply to use OLS in spite of its inefficiency. As might be expected, the standard errors will be biased owing to the correlation across observations; there is evidence that the bias is downward.[39] In that event, the robust estimator in (11-4) would be the natural alternative. In the example given earlier, the nesting structure was obvious. In other cases, such as our application in Example 11.16, that might not be true. In Example 14.16 and in the application in Baltagi (2013), statewide observations are grouped into regions based on intuition. The impact of an incorrect grouping is unclear. Both OLS and FGLS would remain consistent—both are equivalent to GLS with the wrong weights, which we considered earlier. However, the impact on the asymptotic covariance matrix for the estimator remains to be analyzed.

The nested structure of the data would call the clustering computation in (11-4) into question. If the grouping is done only on the innermost level (on teachers in our example), then the assumption that the clusters are independent is incorrect (teachers in the same school in our example). A two- or more level grouping might be called for in this case. For two levels, as in clusters within stratified data (such as panels on firms within industries) or panel data on individuals within neighborhoods), a reasonably

[39]See Moulton (1986).

compact procedure can be constructed. [See, e.g., Cameron and Miller (2015).] The pseudo-log-likelihood function is

$$\ln L = \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{i=1}^{N_{cs}} \ln f(y_{ics} \mid \mathbf{x}_{ics}, \boldsymbol{\theta}), \tag{14-80}$$

where there are S strata, $s = 1, \ldots, S$, C_s clusters in stratum s, $c = 1, \ldots, C_s$ and N_{cs} individual observations in cluster c in stratum s, $i = 1, \ldots, N_{cs}$. We emphasize, this is not the true log likelihood for the sample; the assumed clustering and stratification of the data imply that observations are correlated. Let

$$\mathbf{g}_{ics} = \frac{\partial \ln f(y_{ics} \mid \mathbf{x}_{ics}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \mathbf{g}_{cs} = \sum_{i=1}^{N_{cs}} \mathbf{g}_{ics}, \mathbf{g}_s = \sum_{c=1}^{C_s} \mathbf{g}_{cs},$$

$$\mathbf{G}_s = \left(\sum_{c=1}^{C_s} \mathbf{g}_{cs}\mathbf{g}'_{cs} \right) - \frac{1}{C_s}\mathbf{g}_s\mathbf{g}'_s, \qquad \mathbf{G} = \sum_{s=1}^{S} \mathbf{G}_s,$$

$$\mathbf{H} = \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{i=1}^{N_{cs}} \frac{\partial^2 \ln f(y_{ics} \mid \mathbf{x}_{ics}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{i=1}^{N_{cs}} \mathbf{H}_{ics}. \tag{14-81}$$

Then, the corrected covariance matrix for the pseudo-MLE would be

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}] = [-\hat{\mathbf{H}}]^{-1}[\hat{\mathbf{G}}][-\hat{\mathbf{H}}]^{-1} \tag{14-82}$$

For a linear model estimated using least squares, we would use $\mathbf{g}_{ics} = (e_{ics}/s^2)\mathbf{x}_{ics}$ and $\mathbf{H}_{ics} = (1/s^2)\mathbf{x}_{ics}\mathbf{x}'_{ics}$. The appearances of s^2 would cancel out in the final result. One last consideration concerns some finite population corrections. The terms in $\mathbf{G}$ might be weighted by a factor $w_s = (1 - C_s/C*)$ if stratum s consists of a finite set of C^* clusters of which C_s is a significant proportion, times the within cluster correction, $C_s/(C_s - 1)$, that appears in (11-4), and finally, times $(n - 1)/(n - K)$, where n is the full sample size and K is the number of parameters estimated.

14.14.4 RANDOM EFFECTS IN NONLINEAR MODELS: MLE USING QUADRATURE

Example 14.13 describes a nonlinear model for panel data, the geometric regression model,

$$\text{Prob}[Y_{it} = y_{it} \mid \mathbf{x}_{it}] = \theta_{it}(1 - \theta_{it})^{y_{it}}, y_{it} = 0, 1, \ldots, i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$\theta_{it} = 1/(1 + \lambda_{it}), \lambda_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta}).$$

As noted, this is a panel data model, although as stated, it has none of the features we have used for the panel data in the linear case. It is a regression model,

$$E[y_{it} \mid \mathbf{x}_{it}] = \lambda_{it},$$

which implies that

$$y_{it} = \lambda_{it} + \varepsilon_{it}.$$

This is simply a tautology that defines the deviation of y_{it} from its conditional mean. It might seem natural at this point to introduce a common fixed or random effect, as we did earlier in the linear case, as in

$$y_{it} = \lambda_{it} + \varepsilon_{it} + c_i.$$

However, the difficulty in this specification is that whereas ε_{it} is defined residually just as the difference between y_{it} and its mean, c_i is a freely varying random variable. Without extremely complex constraints on how c_i varies, the model as stated cannot prevent y_{it} from being negative. When building the specification for a nonlinear model, greater care must be taken to preserve the internal consistency of the specification. A frequent approach in **index function models** such as this one is to introduce the common effect in the conditional mean function. The random effects geometric regression model, for example, might appear

$$\text{Prob}[Y_{it} = y_{it}|\mathbf{x}_{it}] = \theta_{it}(1 - \theta_{it})^{y_{it}}, y_{it} = 0, 1, \ldots; i = 1, \ldots, n, t = 1, \ldots, T_i,$$

$$\theta_{it} = 1/(1 + \lambda_{it}), \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i),$$

$$f(u_i) = \text{the specification of the distribution of random effects over individuals.}$$

By this specification, it is now appropriate to state the model specification as

$$\text{Prob}[Y_{it} = y_{it}, |\mathbf{x}_{it}, u_i] = \theta_{it}(1 - \theta_{it})^{y_{it}}.$$

That is, our statement of the probability is now conditioned on both the observed data and the unobserved random effect. The random common effect can then vary freely and the inherent characteristics of the model are preserved.

Two questions now arise:

- How does one obtain maximum likelihood estimates of the parameters of the model? We will pursue that question now.
- If we ignore the individual heterogeneity and simply estimate the pooled model, will we obtain consistent estimators of the model parameters? The answer is sometimes, but usually not. The favorable cases are the simple loglinear models such as the geometric and Poisson models that we consider in this chapter. The unfavorable cases are most of the other common applications in the literature, including, notably, models for binary choice, censored regressions, two-part models, sample selection, and, generally, nonlinear models that do not have simple exponential means.[40]

We will now develop a maximum likelihood estimator for a nonlinear random effects model. To set up the methodology for applications later in the book, we will do this in a generic specification, then return to the specific application of the geometric regression model in Example 14.13. Assume, then, that the panel data model defines the probability distribution of a random variable, y_{it}, conditioned on a data vector, $\mathbf{x}_{it}$, and an unobserved common random effect, u_i. As always, there are T_i observations in the group, and the data on $\mathbf{x}_{it}$ and now u_i are assumed to be strictly exogenously determined. Our model for one individual is, then,

$$p(y_{it}|\mathbf{x}_{it}, u_i) = f(y_{it}|\mathbf{x}_{it}, u_i, \boldsymbol{\theta}),$$

where $p(y_{it}|\mathbf{x}_{it}, u_i)$ indicates that we are defining a conditional density while $f(y_{it}|\mathbf{x}_{it}, u_i, \theta)$ defines the functional form and emphasizes the vector of parameters to be estimated. We are also going to assume that, but for the common u_i, observations within a group would be independent—the dependence of observations in the group arises through the

[40]*Note:* This is the crucial issue in the consideration of robust covariance matrix estimation in Section 14.8. See, as well, Freedman (2006).

presence of the common u_i. The joint density of the T_i observations on y_{it} given u_i under these assumptions would be

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \mathbf{X}_i, u_i) = \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}),$$

because conditioned on u_i, the observations are independent. But because u_i is part of the observation on the group, to construct the log likelihood, we will require the joint density,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}, u_i, | \mathbf{X}_i) = \left[\prod_{t=1}^{T_i} f(y_{it}, | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i).$$

The likelihood function is the joint density for the observed random variables. Because u_i is an unobserved random effect, to construct the likelihood function, we will then have to integrate it out of the joint density. Thus,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \mathbf{X}_i) = \int_{u_i} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i) du_i.$$

The contribution to the log-likelihood function of group i is, then,

$$\ln L_i = \ln \int_{u_i} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i) du_i.$$

There are two practical problems to be solved to implement this estimator. First, it will be rare that the integral will exist in closed form. (It does when the density of y_{it} is normal with linear conditional mean and the random effect is normal, because, as we have seen, this is the random effects linear model.) As such, the practical complication that arises is how the integrals are to be computed. Second, it remains to specify the distribution of u_i over which the integration is taken. The distribution of the common effect is part of the model specification. Several approaches for this model have now appeared in the literature. The one we will develop here extends the random effects model with normally distributed effects that we have analyzed in the previous section. The technique is **Butler and Moffitt's method** (1982). It was originally proposed for extending the random effects model to a binary choice setting (see Chapter 17), but, as we shall see presently, it is straightforward to extend it to a wide range of other models. The computations center on a technique for approximating integrals known as **Gauss–Hermite quadrature**.

We assume that u_i is normally distributed with mean zero and variance σ_u^2. Thus,

$$f(u_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{u_i^2}{2\sigma_u^2} \right).$$

With this assumption, the ith term in the log likelihood is

$$\ln L_i = \ln \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{u_i^2}{2\sigma_u^2} \right) du_i.$$

To put this function in a form that will be convenient for us later, we now let $w_i = u_i/(\sigma_u\sqrt{2})$ so that $u_i = \sigma_u\sqrt{2}w_i = \phi w_i$ and the Jacobian of the transformation from u_i to w_i is $du_i = \phi dw_i$. Now, we make the change of variable in the integral to produce the function

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi w_i, \boldsymbol{\theta}) \right] \exp(-w_i^2) dw_i.$$

For the moment, let

$$g(w_i) = \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi w_i, \boldsymbol{\theta}).$$

Then, the function we are manipulating is

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} g(w_i) \exp(-w_i^2) dw_i.$$

The payoff to all this manipulation is that integrals of this form can be computed very accurately by Gauss–Hermite quadrature. Gauss–Hermite quadrature replaces the integration with a weighted sum of the functions evaluated at a specific set of points. For the general case, this is

$$\int_{-\infty}^{\infty} g(w_i) \exp(-w_i^2) dw_i \approx \sum_{h=1}^{H} z_h g(v_h),$$

where z_h is the weight and v_h is the node. Tables of the nodes and weights are found in popular sources such as Abramovitz and Stegun (1971). For example, the nodes and weights for a four-point quadrature are

$$v_h = \pm 0.52464762327529002 \quad \text{and} \quad \pm 1.6506801238857849,$$
$$z_h = 0.80491409000549996 \quad \text{and} \quad 0.081312835447250001.$$

In practice, it is common to use eight or more points, up to a practical limit of about 96. Assembling all of the parts, we obtain the approximation to the contribution to the log likelihood,

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\theta}) \right].$$

The Hermite approximation to the log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\theta}) \right]. \qquad \textbf{(14-83)}$$

This function is now to be maximized with respect to $\boldsymbol{\theta}$ and ϕ. Maximization is a complex problem. However, it has been automated in contemporary software for some models, notably the binary choice models mentioned earlier, and is in fact quite straightforward to implement in many other models as well. The first and second derivatives of the log-likelihood function are correspondingly complex but still computable using quadrature. The estimate of σ_u and an appropriate standard error are obtained from $\hat{\phi}$ using the result $\phi = \sigma_u \sqrt{2}$. The hypothesis of no cross-period correlation can be tested with a likelihood ratio test.

Example 14.17 Random Effects Geometric Regression Model

We will use the preceding to construct a random effects model for the *DocVis* count variable analyzed in Example 14.10. Using (14-90), the approximate log-likelihood function will be

$$\ln L_H = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} \theta_{it}(1 - \theta_{it})^{y_{it}} \right],$$

$$\theta_{it} = 1/(1 + \lambda_{it}),\ \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \phi v_h).$$

The derivatives of the log likelihood are approximated as well. The following is the general result—development is left as an exercise:

$$\frac{\partial \log L}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}} = \sum_{i=1}^{n} \frac{1}{L_i} \frac{\partial L_i}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}}$$

$$\approx \sum_{i=1}^{n} \frac{\left\{ \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta}) \right] \left[\sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta})}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}} \right] \right\}}{\left\{ \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta}) \right] \right\}}.$$

It remains only to specialize this to our geometric regression model. For this case, the density is given earlier. The missing components of the preceding derivatives are the partial derivatives with respect to $\boldsymbol{\beta}$ and ϕ that were obtained in Section 14.14.4. The necessary result is

$$\frac{\partial \ln f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta})}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}} = [\theta_{it}(1 + y_{it}) - 1] \begin{pmatrix} \mathbf{x}_{it} \\ v_h \end{pmatrix}.$$

Maximum likelihood estimates of the parameters of the random effects geometric regression model are given in Example 14.13 with the fixed effects estimates for this model.

14.14.5 FIXED EFFECTS IN NONLINEAR MODELS: THE INCIDENTAL PARAMETERS PROBLEM

Using the same modeling framework that we used in the previous section, we now define a fixed effects model as an index function model with a group-specific constant term. As before, the model is the assumed density for a random variable,

$$p(y_{it} | d_{it}, \mathbf{x}_{it}) = f(y_{it} | \alpha_i d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where d_{it} is a dummy variable that takes the value one in every period for individual i and zero otherwise. (In more involved models, such as the censored regression model we examine in Chapter 19, there might be other parameters, such as a variance. For now, it is convenient to omit them—the development can be extended to add them later.) For convenience, we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model.[41] The

[41]In estimating a fixed effects linear regression model in Section 11.4, we found that it was not possible to analyze models with time-invariant variables. The same limitation applies in the nonlinear case, for essentially the same reasons. The time-invariant effects are absorbed in the constant term. In estimation, the columns of the derivatives matrix corresponding to time-invariant variables will be transformed to columns of zeros when we compute derivatives of the log-likelihood function.

parameters to be estimated are the K elements of $\boldsymbol{\beta}$ and the n individual constant terms, α_i. The log-likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln f(y_{it}|\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where $f(.)$ is the probability density function of the observed outcome, for example, the geometric regression model that we used in our previous example. It will be convenient to let

$$z_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} \text{ so that } p(y_{it}|d_{it}, \mathbf{x}_{it}) = f(y_{it}|z_{it}).$$

In the fixed effects linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means that eliminated the person-specific constants from the equation. (See Section 11.4.1.) In a few cases of nonlinear models, it is also possible to eliminate the fixed effects from the likelihood function, although in general not by taking deviations from means. One example is the **exponential regression model** that is used in duration modeling, for example for lifetimes of electronic components and electrical equipment such as light bulbs,

$$f(y_{it}|\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}) = \theta_{it}\exp(-\theta_{it}y_{it}), \theta_{it} = \exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}), y_{it} \geq 0.$$

It will be convenient to write $\theta_{it} = \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) = \gamma_i\Delta_{it}$. We are exploiting the invariance property of the MLE—estimating $\gamma_i = \exp(\alpha_i)$ is the same as estimating α_i. The log likelihood is

$$\begin{aligned}\ln L &= \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln \theta_{it} - \theta_{it}y_{it} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln(\gamma_i\Delta_{it}) - (\gamma_i\Delta_{it})y_{it}.\end{aligned} \tag{14-84}$$

The MLE will be found by equating the $n + K$ partial derivatives with respect to γ_i and $\boldsymbol{\beta}$ to zero. For each constant term,

$$\frac{\partial \ln L}{\partial \gamma_i} = \sum_{t=1}^{T_i}\left(\frac{1}{\gamma_i} - \Delta_{it}y_{it}\right).$$

Equating this to zero provides a solution for γ_i in terms of the data and $\boldsymbol{\beta}$,

$$\gamma_i = \frac{T_i}{\sum_{t=1}^{T_i}\Delta_{it}y_{it}}. \tag{14-85}$$

[Note the analogous result for the linear model in (11-16b).] Inserting this solution back in the log-likelihood function in (14-84), we obtain the concentrated log likelihood,

$$\ln L_C = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[\ln\left(\frac{T_i\Delta_{it}}{\sum_{s=1}^{T_i}\Delta_{is}y_{is}}\right) - \left(\frac{T_i\Delta_{it}}{\sum_{s=1}^{T_i}\Delta_{is}y_{is}}\right)y_{it}\right], \tag{14-86}$$

which is now only a function of $\boldsymbol{\beta}$. This function can now be maximized with respect to $\boldsymbol{\beta}$ alone. The MLEs for α_i are then found as the logs of the results of (14-92). Note, once again, we have eliminated the constants from the estimation problem, but not by computing deviations from group means. That is specific to the linear model.

The concentrated log likelihood is only obtainable in only a small handful of cases, including the linear model, the exponential model (as just shown), the Poisson regression model, and a few others. Lancaster (2000) lists some of these and discusses the underlying methodological issues. In most cases, if one desires to estimate the parameters of a fixed effects model, it will be necessary to actually compute the possibly huge number of constant terms, α_i, at the same time as the main parameters, $\boldsymbol{\beta}$. This has widely been viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception.[42] The likelihood equations for the general fixed effects, index function model are

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it} | z_{it})}{\partial z_{it}} \frac{\partial z_{it}}{\partial \alpha_i} = \sum_{t=1}^{T_i} g_{it} = g_{i.} = 0,$$

and

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it} | z_{it})}{\partial z_{it}} \frac{\partial z_{it}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} g_{it} \mathbf{x}_{it} = \mathbf{0}.$$

The second derivatives matrix is

$$\frac{\partial^2 \ln L}{\partial \alpha_i^2} = \sum_{t=1}^{T_i} \frac{\partial^2 \ln f(y_{it} | z_{it})}{\partial z_{it}^2} = \sum_{t=1}^{T_i} h_{it} = h_{i.} < 0,$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \alpha_i} = \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it},$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}_{it}' = \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}'},$$

where $\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}'}$ is a negative definite matrix. The likelihood equations are a large system, but the solution turns out to be surprisingly straightforward.[43]

By using the formula for the partitioned inverse, we find that the $K \times K$ submatrix of the inverse of the Hessian that corresponds to $\boldsymbol{\beta}$, which would provide the asymptotic covariance matrix for the MLE, is

$$\mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} = \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}_{it}' - \frac{1}{h_{i.}} \left(\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \right) \left(\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}' \right) \right] \right\}^{-1},$$

$$= \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it} (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)' \right] \right\}^{-1}, \quad \text{where} \quad \overline{\mathbf{x}}_i = \frac{\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}}{h_{i.}}.$$

Note the striking similarity to the result we had in (11-4) for the fixed effects model in the linear case.[44] By assembling the Hessian as a partitioned matrix for $\boldsymbol{\beta}$ and the full vector of constant terms, then using (A-66b) and the preceding definitions to isolate one diagonal element, we find

$$\mathbf{H}^{\alpha_i \alpha_i} = \frac{1}{h_{i.}} + \overline{\mathbf{x}}_i' \mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} \overline{\mathbf{x}}_i.$$

[42]See, for example, Maddala (1987), p. 317.

[43]See Greene (2004a).

[44]A similar result is noted briefly in Chamberlain (1984).

Once again, the result has the same format as its counterpart in the linear model. In principle, the negatives of these would be the estimators of the asymptotic variances of the maximum likelihood estimators. (Asymptotic properties in this model are problematic, as we consider shortly.)

All of these can be computed quite easily once the parameter estimates are in hand, so that in fact, practical estimation of the model is not really the obstacle. [This must be qualified, however. Consider the likelihood equation for one of the constants in the geometric regression model. This would be

$$\sum_{t=1}^{T_i}[\theta_{it}(1 + y_{it}) - 1] = 0.$$

Suppose y_{it} equals zero in every period for individual i. Then, the solution occurs where $\mathbf{\Sigma}_i(\theta_{it} - 1) = 0$. But θ_{it} is between zero and one, so the sum must be negative and cannot equal zero. The likelihood equation has no solution with finite coefficients. Such groups would have to be removed from the sample to fit this model.]

It is shown in Greene (2004a) that, in spite of the potentially large number of parameters in the model, Newton's method can be used with the following iteration, which uses only the $K \times K$ matrix computed earlier and a few $K \times 1$ vectors:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(s+1)} &= \hat{\boldsymbol{\beta}}^{(s)} - \left\{\sum_{i=1}^{n}\left[\sum_{t=1}^{T_i} h_{it}(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)'\right]\right\}^{-1}\left\{\sum_{i=1}^{n}\left[\sum_{t=1}^{T_i} g_{it}(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\right]\right\}\\ &= \hat{\boldsymbol{\beta}}^{(s)} + \mathbf{\Delta}_{\beta}^{(s)},\end{aligned}$$

and

$$\hat{\alpha}_l^{(s+1)} = \hat{\alpha}_l^{(s)} - [(g_{i.}/h_{i.}) + \overline{\mathbf{x}}_i'\mathbf{\Delta}_{\boldsymbol{\beta}}^{(s)}].$$[45]

This is a large amount of computation involving many summations, but it is linear in the number of parameters and does not involve any $n \times n$ matrices.

In addition to the theoretical virtues and shortcomings (yet to be addressed) of this model, we note the practical aspect of estimation of what are possibly a huge number of parameters, $n + K$. In the fixed effects case, n is not limited, and could be in the thousands in a typical application. In Examples 14.15 and 14.16, n is 7,293. Two large applications of the method described here are Kingdon and Cassen's (2007) study, in which they fit a fixed effects probit model with well over 140,000 dummy variable coefficients, and Fernandez-Val's (2009) study, which analyzes a model with 500,000 groups.

The problems with the fixed effects estimator are statistical, not practical.[46] The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. In this setting, not only is T_i fixed, it is also likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). There is, as well, a small sample (small T_i) bias in the slope estimators. This is the **incidental parameters problem**.[47] The source of the

[45]Similar results appear in Prentice and Gloeckler (1978) who attribute it to Rao (1973) and Chamberlain (1980, 1984).

[46]See Vytlacil, Aakvik, and Heckman (2005), Chamberlain (1980, 1984), Newey (1994), Bover and Arellano (1997), Chen (1998), and Fernandez-Val (2009) for some extensions of parametric and semiparametric forms of the binary choice models with fixed effects.

[47]See Neyman and Scott (1948) and Lancaster (2000).

problem appears to arise from estimating $n + K$ parameters with n multivariate observations—the number of parameters estimated grows with the sample size. The precise implication of the incidental parameters problem differs from one model to the next. In general, the slope estimators in the fixed effects model do converge to a parameter vector, but not to $\boldsymbol{\beta}$. In the most familiar cases, binary choice models such as probit and logit, the small T bias in the coefficient estimators appears to be proportional (e.g., 100% when $T = 2$), and away from zero, and to diminish monotonically with T, becoming essentially negligible as T reaches 15 or 20. In other cases involving continuous variables, the slope coefficients appear not to be biased at all, but the impact is on variance and scale parameters. The linear fixed effects model noted in Footnote 12 in Chapter 11 is an example; the stochastic frontier model (Section 19.2) is another. Yet, in models for truncated variables (Section 19.2), the incidental parameters bias appears to affect both the slopes (biased toward zero) and the variance parameters (also attenuated). We will examine the incidental parameters problem in more detail in Section 15.5.2.

Example 14.18 Fixed and Random Effects Geometric Regression

Example 14.13 presents pooled estimates for a geometric regression model,

$$f(y_{it} \mid \mathbf{x}_{it}) = \theta_{it}(1 - \theta_{it})^{y_{it}},\ \theta_{it} = 1/(1 + \lambda_{it}),\ \lambda_{it} = \exp(c_i + \mathbf{x}_{it}'\boldsymbol{\beta}),\ y_{it} = 0, 1, \ldots.$$

We will now reestimate the model under the assumptions of the random and fixed effects specifications. The methods of the preceding two sections are applied directly—no modification of the procedures was required. Table 14.15 presents the three sets of maximum likelihood estimates. The estimates vary considerably. The average group size is about five. This implies that the fixed effects estimator may well be subject to a small sample bias. Save for the coefficient on *Kids*, the fixed effects and random effects estimates are quite similar. On the other hand, the two panel models give similar results to the pooled model except for the *Income* coefficient. On this basis, it is difficult to see, based solely on the results, which should be the preferred model. The model is nonlinear to begin with, so the pooled model, which might otherwise be preferred on the basis of computational ease, now has no redeeming virtues. None of the three models is robust to misspecification. Unlike the linear model, in this and other nonlinear models, the fixed effects *estimator* is inconsistent when T is small in both random and fixed effects cases. The random effects estimator is consistent in the random effects model, but, as usual, not in the fixed effects model. The pooled estimator is inconsistent in both random and fixed effects cases (which calls into question the virtue of the robust covariance matrix). It might be tempting to use a Hausman specification test (see Section 11.5.5); however, the conditions that underlie the test are not met—unlike the linear model where the fixed effects estimator is consistent in both cases, here it is inconsistent in both cases. For better or worse, that leaves the analyst with the need to choose the model based on the underlying theory.

TABLE 14.15 Panel Data Estimates of a Geometric Regression for *DOCVIS*

	Pooled		*Random Effects*[a]		*Fixed Effects*	
Variable	*Estimate*	*Std. Err.*[b]	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
Constant	1.09189	0.10828	0.39936	0.09530		
Age	0.01799	0.00130	0.02209	0.00122	0.04845	0.00351
Education	−0.04725	0.00671	−0.04506	0.00626	−0.05434	0.03721
Income	−0.46836	0.07265	−0.19569	0.06106	−0.18760	0.09134
Kids	−0.15684	0.03055	−0.12434	0.02336	−0.00253	0.03687

[a]Estimated $\sigma_u = 0.95441$.
[b]Standard errors corrected for clusters in the panel.

14.15 LATENT CLASS AND FINITE MIXTURE MODELS

In this final application of maximum likelihood estimation, rather than explore a particular model, we will develop a technique that has been used in many different settings. The latent class modeling framework specifies that the distribution of the observed data is a mixture of a finite number of underlying populations. The model can be motivated in several ways:

- In the classic application of the technique, the observed data are drawn from a mixture of distinct underlying populations. Consider, for example, a historical or fossilized record of the intersection (or collision) of two populations.[48] The anthropological record consists of measurements on some variable that would differ imperfectly, but substantively, between the populations. However, the analyst has no definitive marker for which subpopulation an observation is drawn from. Given a sample of observations, they are interested in two statistical problems: (1) estimate the parameters of the underlying populations (models) and (2) classify the observations in hand as having originated in which population. The technique has seen a number of recent applications in health econometrics. For example, in a study of obesity, Greene, Harris, Hollingsworth, and Maitra (2008) speculated that their ordered choice model (see Chapter 19) might systematically vary in a sample that contained (it was believed) some individuals who have a genetic predisposition toward obesity and most that did not. In another application, Lambert (1992) studied the number of defective outcomes in a production process. When a "zero defectives" condition is observed, it could indicate either regime 1, "the process is under control," or regime 2, "the process is not under control but just happens to produce a zero observation."
- In a narrower sense, one might view parameter heterogeneity in a population as a form of discrete mixing. We have modeled parameter heterogeneity using continuous distributions in Section 11.10. The "finite mixture" approach takes the distribution of parameters across individuals to be discrete. (Of course, this is another way to interpret the first point.)
- The finite mixing approach is a means by which a distribution (model) can be constructed from a mixture of underlying distributions. Quandt and Ramsey's mixture of normals model in Example 13.4 is a case in which a nonnormal distribution is created by mixing two normal distributions with different parameters.

14.15.1 A FINITE MIXTURE MODEL

To lay the foundation for the more fully developed model that follows, we revisit the mixture of normals model from Example 13.4. Consider a population that consists of a latent mixture of two underlying normal distributions. Neglecting for the moment that it is unknown which applies to a given individual, we have, for individual i, one of the following:

[48]The first application of these methods was Pearson's (1894) analysis of 1,000 measures of the "forehead breadth to body length" of two intermingled species of crabs in the Bay of Naples.

$$f(y_i | class_i = 1) = N[\mu_1, \sigma_1^2] = \frac{\exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}},$$

or **(14-87)**

$$f(y_i | class_i = 2) = N[\mu_2, \sigma_2^2] = \frac{\exp[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}}.$$

The contribution to the likelihood function is $f(y_i | class_i = 1)$ for an individual in class 1 and $f(y_i | class = 2)$ for an individual in class 2. Assume that there is a true proportion $\lambda = \text{Prob}(class_i = 1)$ of individuals in the population that are in class 1, and $(1 - \lambda)$ in class 2. Then, the unconditional (marginal) density for individual i is

$$\begin{aligned} f(y_i) &= \lambda f(y_i | class_i = 1) + (1 - \lambda) f(y_i | class_i = 2) \\ &= E_{classes}\, f(y_i | class_i). \end{aligned} \quad \textbf{(14-88)}$$

The parameters to be estimated are $\lambda, \mu_1, \mu_2, \sigma_1$, and σ_2. Combining terms, the log likelihood for a sample of n individual observations would be

$$\ln L = \sum_{i=1}^{n} \ln\left(\frac{\lambda \exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}} + \frac{(1 - \lambda)\exp[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}} \right). \quad \textbf{(14-89)}$$

This is the mixture density that we saw in Example 13.4. We suggested the method of moments as an estimator of the five parameters in that example. However, this appears to be a straightforward problem in maximum likelihood estimation.

Example 14.19 A Normal Mixture Model for Grade Point Averages

Appendix Table F14.1 contains a data set of 32 observations used by Spector and Mazzeo (1980) to study whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses. Variables in the data set include

GPA = the student's grade point average,
GRADE = dummy variable for whether the student's grade in Intermediate Macroeconomics was higher than in the principles course,
PSI = dummy variable for whether the individual participated in the PSI,
TUCE = the student's score on a pretest in economics.

We will use these data to develop a finite mixture normal model for the distribution of grade point averages.

We begin by computing maximum likelihood estimates of the parameters in (14-89). To estimate the parameters using an iterative method, it is necessary to devise a set of starting values. It might seem natural to use the simple values from a one-class model, $\bar{y}$ and s_y, and a value such as 1/2 for λ. However, the optimizer will immediately stop on these values, as the derivatives will be zero at this point. Rather, it is common to use some value near these—perturbing them slightly (a few percent), just to get the iterations started. Table 14.16 contains the estimates for this two-class finite mixture model. The estimates for the one-class model are the sample mean and standard deviation of *GPA*. [Because these are the MLEs, $\hat{\sigma}^2 = (1/n)\Sigma_{i=1}^{32}(GPA_i - \overline{GPA})^2$.] The means and standard deviations of the two classes are noticeably different—the model appears to be revealing a distinct splitting of the data into two classes. (Whether two is the appropriate number of classes is considered in Section 14.15.5.) It is tempting at this point to identify the two classes with some other covariate, either in the data set or not, such as *PSI*. However, at this point, there is no basis for doing so—the classes are "latent." As the analysis continues, however, we will want to investigate whether any observed data help predict the class membership.

14.15.2 MODELING THE CLASS PROBABILITIES

The development thus far has assumed that the analyst has no information about class membership. Estimation of the prior probabilities (λ in the preceding example) is part of the estimation problem. There may be some, albeit imperfect, information about class membership in the sample as well. For our earlier example of grade point averages, we also know the individual's score on a test of economic literacy (*TUCE*). Use of this information might sharpen the estimates of the class probabilities. The mixture of normals distribution, for example, might be formulated

$$f(y_i|\mathbf{z}_i) = \left(\begin{array}{c} \dfrac{\text{Prob}(class = 1|\mathbf{z}_i)\exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}} \\ + \dfrac{[1 - \text{Prob}(class = 1|\mathbf{z}_i)]\exp[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}} \end{array} \right),$$

where $\mathbf{z}_i$ is the vector of variables that help explain the class probabilities. To make the mixture model amenable to estimation, it is necessary to parameterize the probabilities. The logit probability model is a common device. [See Section 17.2. For applications, see Greene (2005, Section 2.3.3) and references cited.] For the two-class case, this might appear as follows:

$$\text{Prob}(class = 1|\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})},$$
$$\text{Prob}(class = 2|\mathbf{z}_i) = 1 - \text{Prob}(class = 1|\mathbf{z}_i). \quad \textbf{(14-90)}$$

(The more general J class case is shown in Section 14.15.6.) The log likelihood for the mixture of two normal densities becomes

$$\ln L = \sum_{i=1}^{n} \ln L_i = \sum_{i=1}^{n} \ln \left(\begin{array}{c} \left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})}\right)\dfrac{\exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}} \\ + \left(\dfrac{1}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})}\right)\dfrac{\exp[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}} \end{array} \right). \quad \textbf{(14-91)}$$

The log likelihood is now maximized with respect to $\mu_1, \sigma_1, \mu_2, \sigma_2$, and $\boldsymbol{\theta}$. If $\mathbf{z}_i$ contains a constant term and some other observed variables, then the earlier model returns if the coefficients on those other variables all equal zero. In this case, it follows that

TABLE 14.16 Estimated Normal Mixture Model

	One Class		*Latent Class 1*		*Latent Class 2*	
Parameter	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
$\boldsymbol{\mu}$	3.1172	0.08251	3.64187	0.3452	2.8894	0.2514
$\boldsymbol{\sigma}$	0.4594	0.04070	0.2524	0.2625	0.3218	0.1095
Probability	1.0000	0.0000	0.3028	0.3497	0.6972	0.3497
ln L	−20.51274			−19.63654		

$\lambda = \ln[\theta/(1 - \theta)]$. (This device is usually used to ensure that $0 < \lambda < 1$ in the earlier model.)

14.15.3 LATENT CLASS REGRESSION MODELS

To complete the construction of the latent class model, we note that the means (and, in principle, the variances) in the original model could be conditioned on observed data as well. For our normal mixture models, we might make the marginal mean, μ_j, a conditional mean,

$$\mu_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j.$$

In the data of Example 14.17, we also observe an indicator of whether the individual has participated in a special program designed to enhance the economics program (PSI). We might modify the model,

$$f(y_i | class_i = 1, PSI_i) = N[\mu_{i1}, \sigma_1^2] = \frac{\exp[-\frac{1}{2}(y_i - \beta_{1,1} - \beta_{2,1}PSI_i)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}},$$

and similarly for $f(y_i | class_i = 2, PSI_i)$. The model is now a **latent class linear regression model**.

More generally, as we will see shortly, the latent class, or **finite mixture model** for a variable y_i can be formulated as

$$f(y_i | class_i = j, \mathbf{x}_i) = h_j(y_i, \mathbf{x}_i, \boldsymbol{\gamma}_j),$$

where h_j denotes the density conditioned on class j—indexed by j to indicate, for example, the jth parameter vector $\boldsymbol{\gamma}_j = (\boldsymbol{\beta}_j, \sigma_j)$ and so on. The marginal class probabilities are

$$\text{Prob}(class_i = j | \mathbf{z}_i) = p_j(j, \mathbf{z}_i, \boldsymbol{\theta}).$$

The methodology can be applied to any model for y_i. In the example in Section 14.15.6, we will model a binary dependent variable with a probit model. The methodology has been applied in many other settings, such as stochastic frontier models [Orea and Kumbhakar (2004), Greene (2004)], Poisson regression models [Wedel et al. (1993)], and a wide variety of count, discrete choice, and limited dependent variable models [McLachlan and Peel (2000), Greene (2007b)].

Example 14.20 Latent Class Regression Model for Grade Point Averages

Combining 14.15.2 and 14.15.3, we have a latent class model for grade point averages,

$$f(GPA_i | class_i = j, PSI_i) = \frac{\exp[-\frac{1}{2}(y_i - \beta_{1j} - \beta_{2j}PSI_i)^2/\sigma_j^2]}{\sigma_j\sqrt{2\pi}}, j = 1, 2,$$

$$\text{Prob}(class_i = 1 | TUCE_i) = \frac{\exp(\theta_1 + \theta_2 TUCE_i)}{1 + \exp(\theta_1 + \theta_2 TUCE_i)},$$

$$\text{Prob}(class_i = 2 | TUCE_i) = 1 - \text{Prob}(class = 1 | TUCE_i).$$

The log likelihood is now

$$\ln L = \sum_{i=1}^{n} \ln \left(\begin{array}{l} \left(\dfrac{\exp(\theta_1 + \theta_2 TUCE_i)}{1 + \exp(\theta_1 + \theta_2 TUCE_i)}\right)\dfrac{\exp[-\frac{1}{2}(y_i - \beta_{1,1} - \beta_{2,1}PSI_i)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}} \\ + \left(\dfrac{1}{1 + \exp(\theta_1 + \theta_2 TUCE_i)}\right)\dfrac{\exp[-\frac{1}{2}(y_i - \beta_{1,2} - \beta_{2,2}PSI_i)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}} \end{array} \right).$$

Maximum likelihood estimates of the parameters are given in Table 14.17.

TABLE 14.17 Estimated Latent Class Linear Regression Model for GPA

	One Class		*Latent Class 1*		*Latent Class 2*	
Parameter	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
β_1	3.1011	0.1117	3.3928	0.1733	2.7926	0.04988
β_2	0.03675	0.1689	−0.1074	0.2006	−0.5703	0.07553
σ	0.4443	0.0003086	0.3812	0.09337	0.1119	0.04487
θ_1	0.0000	0.0000	−6.8392	3.07867	0.0000	0.0000
θ_2	0.0000	0.0000	0.3518	0.1601	0.0000	0.0000
$P(\text{class}\,\vert\, TUCE)$	1.0000		0.7063		0.2937	
ln L	−20.48752		−13.39966			

14.15.4 PREDICTING CLASS MEMBERSHIP AND β_i

The model in (14-91) now characterizes two random variables, y_i, the outcome variable of interest, and $class_i$, the indicator of which class the individual resides in. We have a joint distribution, $f(y_i, class_i)$, which we are modeling in terms of the conditional density, $f(y_i \mid class_i)$ in (14-87), and the marginal density of $class_i$ in (14-90). We have initially assumed the latter to be a simple Bernoulli distribution with $\text{Prob}(class_i = 1) = \lambda$, but then modified in the previous section to equal $\text{Prob}(class_i = 1 \mid \mathbf{z}_i) = \Lambda(\mathbf{z}_i'\boldsymbol{\theta})$. These can be viewed as the prior probabilities in a Bayesian sense. If we wish to make a prediction as to which class the individual came from, using all the information that we have on that individual, then the prior probability is going to waste some information; it wastes the information on the observed outcome. The posterior, or conditional (on the remaining data) probability,

$$\text{Prob}(class_i = 1 \mid \mathbf{z}_i\, y_i) = \frac{f(y_i, class = 1 \mid \mathbf{z}_i)}{f(y_i)},$$

will be based on more information than the marginal probabilities. We have the elements that we need to compute this conditional probability. Use Bayes's theorem to write this as

$$\text{Prob}(class_i = 1 \mid \mathbf{z}_i, y_i) = \frac{f(y_i \mid \text{class}_\text{i} = 1, \mathbf{z}_i)\,\text{Prob}(class_i = 1 \mid \mathbf{z}_i)}{f(y_i \mid class_i = 1, \mathbf{z}_i)\,\text{Prob}(class_i = 1 \mid \mathbf{z}_i) + f(y_i \mid class_i = 2, \mathbf{z}_i)\,\text{Prob}(class_i = 2 \mid \mathbf{z}_i)}.$$

The denominator is L_i (not ln L_i) from (14-91). The numerator is the first term in L_i. To continue our mixture of two normals example, the conditional (posterior) probability is

$$\text{Prob}(class_i = 1 \mid \mathbf{z}_i, y_i) = \frac{\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})}\right)\dfrac{\exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}}}{L_i},$$

while the unconditional probability is in (14-90). The conditional probability for the second class is computed using the other two marginal densities in the numerator (or by subtraction from one). Note that the conditional probabilities are functions of the data even if the unconditional ones are not. To come to the problem suggested at the outset, then, the natural predictor of $class_i$ is the class associated with the largest estimated posterior probability.

In random parameter settings, we have also been interested in predicting $E[\boldsymbol{\beta}_i | y_i, \mathbf{X}_i]$. There are two candidates for the latent class model. Having made the best guess as to which specific class an individual resides in, a natural estimator of β_i would be the β_j associated with that class. A preferable estimator that uses more information would be the posterior expected value,

$$\hat{E}[\boldsymbol{\beta}_i | y_i, \mathbf{X}_i, z_i] = \sum_{j=1}^{J} \hat{\pi}_{ij}(\hat{\boldsymbol{\Theta}}, \mathbf{z}_i)\hat{\boldsymbol{\beta}}_j.$$

Example 14.21 Predicting Class Probabilities

Table 14.18 lists the observations sorted by GPA. The predictions of class membership reflect what one might guess from the coefficients in the table of coefficients. Class 2 members on average have lower GPAs than in class 1. The listing in Table 14.18 shows this clustering. It

TABLE 14.18 Estimated Latent Class Probabilities

GPA	*TUCE*	*PSI*	*CLASS*	P_1	P_1^*	P_2	P_2^*
2.06	22	1	2	0.7109	0.0116	0.2891	0.9884
2.39	19	1	2	0.4612	0.0467	0.5388	0.9533
2.63	20	0	2	0.5489	0.1217	0.4511	0.8783
2.66	20	0	2	0.5489	0.1020	0.4511	0.8980
2.67	24	1	1	0.8325	0.9992	0.1675	0.0008
2.74	19	0	2	0.4612	0.0608	0.5388	0.9392
2.75	25	0	2	0.8760	0.3499	0.1240	0.6501
2.76	17	0	2	0.2975	0.0317	0.7025	0.9683
2.83	19	0	2	0.4612	0.0821	0.5388	0.9179
2.83	27	1	1	0.9345	1.0000	0.0655	0.0000
2.86	17	0	2	0.2975	0.0532	0.7025	0.9468
2.87	21	0	2	0.6336	0.2013	0.3664	0.7987
2.89	14	1	1	0.1285	1.0000	0.8715	0.0000
2.89	22	0	2	0.7109	0.3065	0.2891	0.6935
2.92	12	0	2	0.0680	0.0186	0.9320	0.9814
3.03	25	0	1	0.8760	0.9260	0.1240	0.0740
3.10	21	1	1	0.6336	1.0000	0.3664	0.0000
3.12	23	1	1	0.7775	1.0000	0.2225	0.0000
3.16	25	1	1	0.8760	1.0000	0.1240	0.0000
3.26	25	0	1	0.8760	0.9999	0.1240	0.0001
3.28	24	0	1	0.8325	0.9999	0.1675	0.0001
3.32	23	0	1	0.7775	1.0000	0.2225	0.0000
3.39	17	1	1	0.2975	1.0000	0.7025	0.0000
3.51	26	1	1	0.9094	1.0000	0.0906	0.0000
3.53	26	0	1	0.9094	1.0000	0.0906	0.0000
3.54	24	1	1	0.8325	1.0000	0.1675	0.0000
3.57	23	0	1	0.7775	1.0000	0.2225	0.0000
3.62	28	1	1	0.9530	1.0000	0.0470	0.0000
3.65	21	1	1	0.6336	1.0000	0.3664	0.0000
3.92	29	0	1	0.9665	1.0000	0.0335	0.0000
4.00	21	0	1	0.6336	1.0000	0.3664	0.0000
4.00	23	1	1	0.7775	1.0000	0.2225	0.0000

also suggests how the latent class model is using the sample information. If the results in Table 14.16—just estimating the means, constant class probabilities—are used to produce the same table, when sorted, the highest 10 GPAs are in class 1 and the remainder are in class 2. The more elaborate model is adding information on *TUCE* to the computation. A low *TUCE* score can push a high GPA individual into class 2. (Of course, this is largely what multiple linear regression does as well.)

14.15.5 DETERMINING THE NUMBER OF CLASSES

There is an unsolved inference issue remaining in the specification of the model. The number of classes has been taken as a known parameter—two in our main example thus far, three in the following application. Ideally, one would like to determine the appropriate number of classes statistically. However, J is not a parameter in the model. A likelihood ratio test, for example, will not provide a valid result. Consider the original model in Example 14.17. The model has two classes and five parameters in total. It would seem natural to test down to a one-class model that contains only the mean and variance using the LR test. However, the number of restrictions here is actually ambiguous. If $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, then the mixing probability is irrelevant—the two class densities are the same, and it is a one-class model. Thus, the number of restrictions needed to get from the two-class model to the one-class model is ambiguous. It is neither two nor three. One strategy that has been suggested is to test upward, adding classes until the marginal class insignificantly changes the log likelihood or one of the information criteria such as the AIC or BIC (see Section 14.6.5). Unfortunately, this approach is likewise problematic because the estimates from any specification that is too short are inconsistent. The alternative would be to test down from a specification known to be too large. Heckman and Singer (1984b) discuss this possibility and note that when the number of classes becomes larger than appropriate, the estimator should break down. In our Example 14.15, if we expand to four classes, the optimizer breaks down, and it is no longer possible to compute the estimates. A five-class model does produce estimates, but some are nonsensical. This does provide at least the directions to seek a viable strategy. The authoritative treatise on finite mixture models by McLachlan and Peel (2000, Chapter 6) contains extensive discussion of this issue.

14.15.6 A PANEL DATA APPLICATION

The latent class model is a useful framework for applications in panel data. The class probabilities partly play the role of common random effects, as we will now explore. The latent class model can be interpreted as a random parameters model with a discrete distribution of the parameters.

Suppose that $\boldsymbol{\beta}_j$ is generated from a discrete distribution with J outcomes, or classes, so that the distribution of $\boldsymbol{\beta}_j$ is over these classes. Thus, the model states that an individual belongs to one of the J latent classes, indexed by the parameter vector, but it is unknown from the sample data exactly which one. We will use the sample data to estimate the parameter vectors, the parameters of the underlying probability distribution and the probabilities of class membership. The corresponding model formulation is now

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\Delta}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J) = \sum_{j=1}^{J} p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta}) f(y_{it}|class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j), \quad \textbf{(14-92)}$$

where it remains to parameterize the class probabilities, p_{ij}, and the structural model, $f(y_{it}|class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)$. The parameter matrix, $\boldsymbol{\Delta}$, contains the parameters of the discrete probability distribution. It has J rows, one for each class, and M columns, for the M variables in $\mathbf{z}_i$. At a minimum, $M = 1$ and $\mathbf{z}_i$ contains a constant term if the class probabilities are fixed parameters as in Example 14.17. Finally, to accommodate the panel data nature of the sampling situation, we suppose that conditioned on $\boldsymbol{\beta}_j$, that is, on membership in class j, which is fixed over time, the observations on y_{it} are independent. Therefore, for a group of T_i observations, the joint density is

$$f(y_{i1}, y_{i2}, \ldots, y_{t,T_i} | \, class = j, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) = \prod_{t=1}^{T_i} f(y_{it} | \, class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j).$$

The log-likelihood function for a panel of data is

$$\ln L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{J} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i) \prod_{t=1}^{T_i} f(y_{it} | \, class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j) \right]. \quad \textbf{(14-93)}$$

The class probabilities must be constrained be in (0,1) and to sum to 1. The approach that is usually used is to reparameterize them as a set of logit probabilities, as we did in the preceding examples. Then,

$$p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta}) = \frac{\exp(\theta_{ij})}{\sum_{j=1}^{J} \exp(\theta_{ij})}, j = 1, \ldots, J, \theta_{ij} = \mathbf{z}_i' \boldsymbol{\delta}_j, \theta_{iJ} = 0 \; (\boldsymbol{\delta}_J = \mathbf{0}). \quad \textbf{(14-94)}$$

(See Section 18.2.2 for development of this model for the set of probabilities.) Note the restriction on θ_{ij}. This is an identification restriction. Without it, the same set of probabilities will arise if an arbitrary vector is added to every $\boldsymbol{\delta}_j$. The resulting log likelihood is a continuous function of the parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ and $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_J$. For all its apparent complexity, estimation of this model by direct maximization of the log likelihood is not especially difficult.[49] The number of classes that can be identified is likely to be relatively small (on the order of 5 or 10 at most), however, which has been viewed as a drawback of the approach. In general, the more complex the model for y_{it}, the more difficult it becomes to expand the number of classes. Also, as might be expected, the less rich the data set in terms of cross-group variation, the more difficult it is to estimate latent class models.

Estimation produces values for the structural parameters, $(\boldsymbol{\beta}_j, \boldsymbol{\delta}_j), j = 1, \ldots, J$. With these in hand, we can compute the prior class probabilities, p_{ij}, using (14-94). For prediction purposes, we are also interested in the posterior (to the data) class probabilities, which we can compute using Bayes' theorem [see (14-93)]. The conditional probability is

$$\begin{aligned} \text{Prob}(class = j | \, \text{observation } i) &\\ &= \frac{f(\text{observation } i | \, class = j)\text{Prob}(class\ j)}{\sum_{j=1}^{J} f(\text{observation } i | \, class = j)\ \text{Prob}(class\ j)} \\ &= \frac{f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_j, \boldsymbol{\Delta})}{\sum_{j=1}^{J} f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_j, \boldsymbol{\Delta})} \\ &= w_{ij}. \end{aligned} \quad \textbf{(14-95)}$$

[49] See Section E.3 and Greene (2001, 2007b). The EM algorithm discussed in Section E.3.7 is especially well suited for estimating the parameters of latent class models. See McLachlan and Peel (2000).

The set of probabilities, $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iJ})$, gives the posterior density over the distribution of values of $\boldsymbol{\beta}$, that is, $[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J]$. For a particular model and allowing for grouping within a panel data set, the posterior probability for class j is found as

$$\text{Prob}(class = j \mid \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i) = \frac{p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)\prod_{t=1}^{T_i} f(y_{it} \mid class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)}{\sum_{j=1}^{J} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)\prod_{t=1}^{T_i} f(y_{it} \mid class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)}$$
$$= \frac{\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\Delta}_j)}{\Sigma_{m=1}^{J}\exp(\mathbf{z}_i'\boldsymbol{\Delta}_m)}\right)\prod_{t=1}^{T_i} f(y_{it} \mid class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)}{\sum_{j=1}^{J}\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\Delta}_j)}{\Sigma_{m=1}^{J}\exp(\mathbf{z}_i'\boldsymbol{\Delta}_m)}\right)\prod_{t=1}^{T_i} f(y_{it} \mid class = m, \mathbf{x}_{it}, \boldsymbol{\beta}_m)}. \quad \textbf{(14-96)}$$

Example 14.22 A Latent Class Two-Part Model for Health Care Utilization

Jones and Bago D'Uva (2009) examined health care utilization in Europe using 8 waves of the ECHP panel data set. The variable of interest was numbers of visits to the physician. They examined two outcomes, visits to general practitioners and visits to specialists. The modeling framework was the latent class model in (14-92). The class-specific model was a two-part, negative binomial "hurdle" model for counts,

$$\text{Prob}(y_{it} = 0 \mid \mathbf{x}_{it}, \boldsymbol{\beta}_{1j}) = \frac{1}{1 + \lambda_{1it,j}}, \; \lambda_{1it,j} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}_{1j})$$

$$\text{Prob}(y_{it} \mid y_{it} > 0, \mathbf{x}_{it}, \boldsymbol{\beta}_{2j}, \alpha_j) = \frac{(\alpha_j\lambda_{2it,j} + 1)^{-1/\alpha_j}\,\Gamma(y_{it} + 1/\alpha_j)[1 + (\lambda_{2it,j}^{-1}/\alpha_j)]^{-y_{it}}}{\Gamma(1/\alpha_j)\Gamma(y_{it} + 1)[1 - (\alpha_j\lambda_{2it,j} + 1)^{-1/\alpha_j}]},$$

$$\lambda_{2it,j} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}_{2j}),\; \alpha_j > 0.$$

[This is their equation (2) with $k = 0$.] The first equation is a participation equation, for whether the number of doctor visits equals 0 or some positive value. The second equation is the intensity equation that predicts the number of visits, given that the number of visits is positive. The count model is a *negative binomial model*. This is an extension of the Poisson regression model. The Poisson model is a limiting case when $\alpha_j \rightarrow 0$. The hurdle and count equations involve different coefficient vectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, so that the determinants of care have different effects on the two stages. Interpretation of this model is complicated by the results that variables appear in both equations, and that the conditional mean function is complex. The simple conditional mean, if there were no hurdle effects, would be $E[y_{it} \mid \mathbf{x}_{it}] = \lambda_{2it}$. However, with the hurdle effects,

$$E[y_{it} \mid \mathbf{x}_{it}] = \text{Prob}(y_{it} > 0 \mid \mathbf{x}_{it}) \times E[y_{it} \mid y_{it} > 0, \mathbf{x}_{it}].$$

The authors examined the two components of this result separately. (The elasticity of the mean would be the sum of these two elasticities.) The mixture model involves two classes (as typical in this literature) A sampling of their results appears in Table 14.19 below. (The results are extracted from their Table 8.) Note that separate tables are given for "Low Users" and "High Users." The results in Section 14.15.4 are used to classify individuals into class 1 and class 2. It is then discovered that the average usage of those individuals classified as in class 1 is far lower than the average use of those in class 2.

TABLE 14.19 Country-Specific Estimated Income Coefficients and Elasticities for GP Visits

		Low Users		*High Users*	
Country		*Coefficient*	*Elasticity*	*Coefficient*	*Elasticity*
Austria	$P(y > 0)$	−0.051	−0.012	−0.109	−0.005
	$E[y \mid y > 0]$	0.012	0.009	0.039	0.035
Denmark	$P(y > 0)$	0.083	0.033	0.261	0.023
	$E[y \mid y > 0]$	0.042	0.021	−0.030	−0.024
The Netherlands	$P(y > 0)$	0.082	0.035	0.094	0.009
	$E[y \mid y > 0]$	−0.037	−0.019	−0.085	−0.068

Example 14.23 Latent Class Models for Health Care Utilization

In Examples 7.6 and 11.21, we proposed an exponential regression model,

$$y_{it} = DocVis_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) + \varepsilon_{it},$$

for the variable *DocVis*, the number of visits to the doctor, in the German health care data. (See Example 11.20 for details.) The regression results for the specification,

$$\mathbf{x}_{it} = (1, Age_{it}, Education_{it}, Income_{it}, Kids_{it}),$$

are repeated (in parentheses) in Table 14.20 for convenience. The nonlinear least squares estimator is only semiparametric; it makes no assumption about the distribution of $DocVis_{it}$ or about ε_{it}. We do see striking increases in the standard errors when the "cluster robust" asymptotic covariance matrix is used. (The estimates are given in parentheses.) The analysis at this point assumes that the nonlinear least squares estimator remains consistent in the presence of the cross-observation correlation. Given the way the model is specified, that is, only in terms of the conditional mean function, this is probably reasonable. The extension would imply a nonlinear generalized regression as opposed to a nonlinear ordinary regression.

TABLE 14.20 Panel Data Estimates of a Geometric Regression for DOCVIS

	Pooled		*Random Effects*[a]		*Fixed Effects*	
Variable	*Estimate*	*Std. Err.*[b]	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
Constant	1.09189 (0.98017)[c]	0.10828 (0.18137)	0.39936	0.09530		
Age	0.01799 (0.01873)	0.00130 (0.00198)	0.02209	0.00122	0.04845	0.00351
Education	−0.04725 (−0.03609)	0.00671 (0.01287)	−0.04506	0.00626	−0.05434	0.03721
Income	−0.46836 (−0.59189)	0.07265 (0.12827)	−0.19569	0.06106	−0.18760	0.09134
Kids	−0.15684 (−0.16930)	0.03055 (0.04882)	−0.12434	0.02336	−0.00253	0.03687

[a]Estimated $\sigma_u = 0.95441$.
[b]Standard errors corrected for clusters in the panel.
[c]Nonlinear least squares results in parentheses.

In Example 14.13, we narrowed this model by assuming that the observations on doctor visits were generated by a geometric distribution,

$$f(y_i|\mathbf{x}_i) = \theta_i(1 - \theta_i)^{y_i}, \theta_i = 1/(1 + \lambda_i), \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), y_i = 0, 1, \ldots .$$

The conditional mean is still $\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$, but this specification adds the structure of a particular distribution for outcomes. The pooled model was estimated in Example 14.13. Examples 14.17 and 14.18 added the panel data assumptions of random, then fixed effects, to the model. The model is now

$$f(y_{it}|\mathbf{x}_{it}) = \theta_{it}(1 - \theta_{it})^{y_{it}}, \theta_{it} = 1/(1 + \lambda_{it}), \lambda_{it} = \exp(c_i + \mathbf{x}_{it}'\boldsymbol{\beta}), y_{it} = 0, 1, \ldots .$$

The pooled, random effects and fixed effects estimates appear in Table 14.17. The pooled estimates, where the standard errors are corrected for the panel data grouping, are comparable to the nonlinear least squares estimates with the robust standard errors. The parameter estimates are similar—both are consistent and this is a very large sample. The smaller standard errors seen for the MLE are the product of the more detailed specification. We will now relax the specification by assuming a two-class finite mixture model. We also specify that the class probabilities are functions of gender and marital status. For the latent class specification,

$$\text{Prob}(class_i = 1|\mathbf{z}_i) = \Lambda(\theta_1 + \theta_2 Female_i + \theta_3 Married_i).$$

The model structure is the geometric regression as before. Estimates of the parameters of the latent class model are shown in Table 14.21. See Section E3.7 for discussion of estimation methods.

Deb and Trivedi (2002) and Bago D'Uva and Jones (2009) suggested that a meaningful distinction between groups of health care system users would be between *infrequent* and *frequent* users. To investigate whether our latent class model is picking up this distinction in the data, we used (14-96) to predict the class memberships (class 1 or 2). We then linearly regressed $DocVis_{it}$ on a constant and a dummy variable for class 2. The results are

$$DocVis_{it} = 5.8034\ (0.0465) - 4.7801\ (0.06282) Class2_i + e_{it},$$

TABLE 14.21 Estimated Latent Class Geometric Regression Model for DocVis

	One Class		*Latent Class 1*		*Latent Class 2*	
Parameter	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
β_1	1.0918	0.1082	1.6423	0.05351	−0.3344	0.09288
β_2	0.0180	0.0013	0.01691	0.0007324	0.02649	0.001248
β_3	−0.0473	0.0067	−0.04473	0.003451	−0.06502	0.005739
β_4	−0.4687	0.0726	−0.4567	0.04688	0.01395	0.06964
β_5	−0.1569	0.0306	−0.1177	0.01611	−0.1388	0.02738
θ_1	0.000	0.000	−0.4280	0.06938	0.0000	0.0000
θ_2	0.000	0.000	0.8255	0.06322	0.0000	0.0000
θ_3	0.000	0.000	−0.07829	0.07143	0.0000	0.0000
Prob$\|z$	1.0000		0.47697		0.52303	
ln L	−61917.97		−58708.63			

where estimated standard errors are in parentheses. The linear regression suggests that the class membership dummy variable is strongly segregating the observations into frequent and infrequent users. The information in the regression is summarized in the descriptive statistics in Table 14.22.

Finally, we did a specification search for the number of classes. Table 14.23 reports the log likelihoods and AICs for models with 1 to 8 classes. The lowest value of the AIC occurs with 7 classes, although the marginal improvement ends near to $J = 4$. The rightmost 8 columns show the averages of the conditional probabilities, which equal the unconditional probabilities. Note that when $J = 8$, three of the classes (2, 5, and 6) have extremely small probabilities. This suggests that the model might be overspecified. We will see another indicator in the next section.

14.15.7 A SEMIPARAMETRIC RANDOM EFFECTS MODEL

Heckman and Singer (1984a,b) suggested a semiparametric maximum likelihood approach to modeling latent heterogeneity in a duration model (Section 19.5) for unemployment spells. The methodology applies equally well to other settings, such as the one we are examining here. Their method can be applied as a finite mixture model in which only the constant term varies across classes. The log likelihood in this case would be

$$\ln L = \sum_{i=1}^{n} \ln \sum_{j=1}^{J} \pi_j \left(\prod_{t=1}^{T_i} f(y_{it} \mid \alpha_j + \mathbf{x}_{it}'\boldsymbol{\beta}) \right). \tag{14-97}$$

TABLE 14.22 Descriptive Statistics for Doctor Visits

Class	*Mean*	*Standard Deviation*
All, $n = 27{,}326$	3.18352	5.68979
Class 1, $n = 12{,}349$	5.80347	7.47579
Class 2, $n = 14{,}977$	1.02330	1.63076

TABLE 14.23 Specification Search for Number of Latent Classes

J	*ln L*	*AIC*	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
1	−61917.77	1.23845	1.0000							
2	−58708.48	1.17443	0.4770	0.5230						
3	−58036.15	1.16114	0.2045	0.6052	0.1903					
4	−57953.02	1.15944	0.1443	0.5594	0.2407	0.0601				
5	−57866.34	1.15806	0.0708	0.0475	0.4107	0.3731	0.0979			
6	−57829.96	1.15749	0.0475	0.0112	0.2790	0.1680	0.4380	0.0734		
7	−57808.50	1.15723	0.0841	0.0809	0.0512	0.3738	0.0668	0.0666	0.2757	
8	−57808.07	1.15738	0.0641	0.0038	0.4434	0.3102	0.0029	0.0002	0.1115	0.0640

This is a restricted form of (14-93). The specification is a random effects model in which the heterogeneity has a discrete, multinomial distribution with unconditional mixing probabilities.

Example 14.24 Semiparametric Random Effects Model

Estimates of a random effects geometric regression model are given in Table 14.17. The random effect (random constant term) is assumed to be normally distributed; the estimated standard deviation is 0.95441. Tables 14.24 and 14.25 present estimates of the semiparametric random effects model. The estimated constant terms and class probabilities are shown in Table 14.24. We fit mixture models for 2 through 7 classes. The AIC stopped falling at $J = 7$. The results for 6 and 7 are shown in the table. Note in the 7 class model, the estimated standard errors for the constants for classes 2 and 4 are essentially infinite—the values shown are the result of rounding error. As Heckman and Singer noted, this should be taken as evidence of overfitting the data. The remaining coefficients for the parametric parts of the model are shown in Table 14.25. The two approaches to fitting the random effects model produce similar results. The coefficients on the regressors and their estimated standard errors are very similar. The random effects in the normal model are estimated to have a mean of 0.39936 and standard deviation of 0.95441. The multinomial distribution in the mixture model has estimated mean 0.27770 and standard deviation 1.2333. Figure 14.7 shows a comparison of the two estimated distributions.[50]

TABLE 14.24 Heckman and Singer Semiparametric Random Effects Model

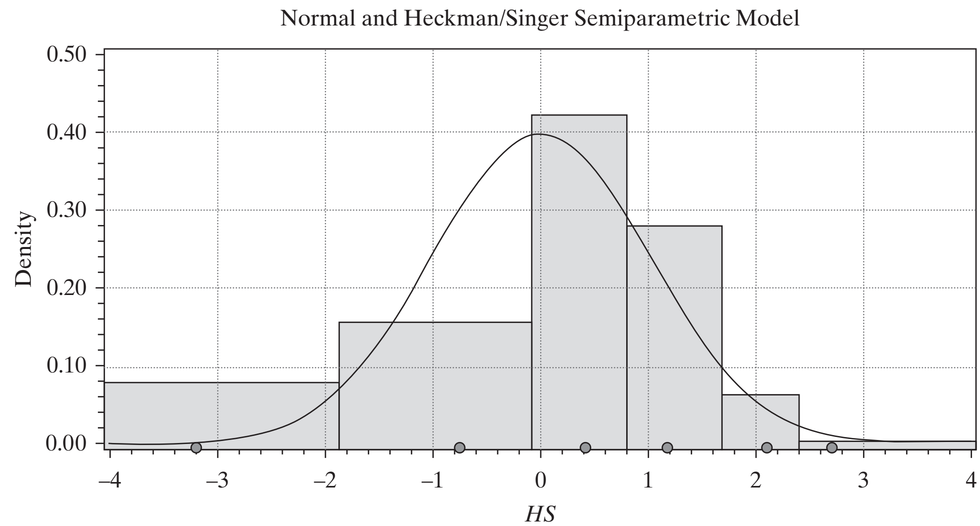
Class	α	*Std. Err.*	*P(class)*	α	*Std. Err.*	*P(class)*
1	−3.17815	0.28542	0.07394	−0.72948	0.16886	0.16825
2	−0.72948	0.15847	0.16825	1.23774	358561.2	0.04030
3	0.38886	0.11867	0.41734	0.38886	0.15112	0.41734
4	1.23774	0.12295	0.28452	1.23774	59175.41	0.24421
5	2.11958	0.28568	0.05183	2.11958	0.41549	0.05183
6	2.69846	0.98622	0.00412	2.69846	1.17124	0.00412
7				−3.17815	0.28863	0.07394

TABLE 14.25 Estimated Random Effects Exponential Count Data Model

	Finite Mixture Model		*Normal Random Effects Model*	
	Estimate	*Std. Err.*	*Estimate*	*Std. Err .*
Constant	$\bar{\hat{\alpha}} = 0.277697$		0.39936	0.09530
Age	0.02136	0.00115	0.02209	0.00122
Educ.	−0.03877	0.00607	−0.04506	0.00626
Income	−0.23729	0.05972	−0.19569	0.06106
Kids	−0.12611	0.02280	−0.12434	0.02336
	$s_\alpha = 1.23333$		$\sigma_u = 0.95441$	

[50]The multinomial distribution has interior boundaries at the midpoints between the estimated constants. The mass points have heights equal to the probabilities. The rectangles sum to slightly more than one—about 1.15. The figure is only a sketch of an implied approximation to the normal distribution in the parametric model.

FIGURE 14.7 Estimated Distributions of Random Effects.



14.16 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. Chapter 13 considered GMM estimation techniques that are less parametric, but more robust to variation in the underlying data-generating process. Together, ML and GMM estimation account for the large majority of empirical estimation in econometrics.

Key Terms and Concepts

- Asymptotic variance
- BHHH estimator
- Butler and Moffitt's method
- Concentrated log likelihood
- Efficient score
- Finite mixture model
- Gauss–Hermite quadrature
- Generalized sum of squares
- Incidental parameters problem
- Index function model
- Information matrix equality
- Kullback–Leibler information criterion (KLIC)
- Lagrange multiplier statistic
- Lagrange multiplier (LM) test
- Latent class linear regression model
- Likelihood equation
- Likelihood ratio
- Likelihood ratio index
- Likelihood ratio statistic
- Likelihood ratio (LR) test
- Limited Information Maximum Likelihood
- Logistic probability model
- Loglinear conditional mean
- Maximum likelihood
- Method of scoring
- Newton's method
- Noncentral chi-squared distribution

- Nonlinear least squares
- Nonnested models
- Oberhofer–Kmenta estimator
- Outer product of gradients estimator (OPG)
- Precision parameter
- Pseudo-log-likelihood function
- Pseudo-MLE
- Quasi-MLE
- Random effects
- Regularity conditions
- Score test
- Score vector
- Vuong 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 $f(y|x)$ is of the form

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

   Prove that $f(y|x)$ integrates to 1. Find the maximum likelihood estimator of λ and its asymptotic distribution. (*Hint:* In the conditional distribution, just carry the x's 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

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

   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. Because 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 14.3 holds for the Poisson distribution used in Section 14.3 by showing that $E[(1/n) \ln L(\theta|y)]$ is uniquely maximized at $\theta = \theta_0$. (*Hint:* First show that the expectation is $-\theta + \theta_0 \ln\theta - E_0[\ln y_i]$.) Show that the likelihood inequality in Theorem 14.3 holds for the normal distribution.
8. For random sampling from the classical regression model in (14-3), 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.
9. 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 estimators of the parameters are $\hat{\mu}_m = \bar{y}_m$, and

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

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

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

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

Applications

1. **Binary Choice.** This application will be based on the health care data analyzed in Example 14.13 and several others. Details on obtaining the data are given in Appendix F Table 7.1. We consider analysis of a dependent variable, y_{it}, that takes values 1 and 0 with probabilities $F(\mathbf{x}_i'\boldsymbol{\beta})$ and $1 - F(\mathbf{x}_i'\boldsymbol{\beta})$, where F is a function that defines a probability. The dependent variable, *yit*, is constructed from the count variable *DocVis*, which is the number of visits to the doctor in the given year. Construct the binary variable

$$y_{it} = 1 \text{ if } DocVis > 0, 0 \text{ otherwise.}$$

We will build a model for the probability that y_{it} equals one. The independent variables of interest will be

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, female_t, married_{it}, hsat_{it}).$$

a. According to the model, the theoretical density for y_{it} is

$$f(y_{it}|\mathbf{x}_{it}) = F(\mathbf{x}_{it}'\boldsymbol{\beta}) \text{ for } y_{it} = 1 \text{ and } 1 - F(\mathbf{x}_{it}'\boldsymbol{\beta}) \text{ for } y_{it} = 0.$$

We will assume that a "logit model" (see Section 17.2) is appropriate, so that

$$F(\mathbf{x}_{it}'\boldsymbol{\beta}) = \Lambda(\mathbf{x}_{it}'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta})}{1 - \exp(\mathbf{x}_{it}'\boldsymbol{\beta})}.$$

Show that for the two outcomes, the probabilities may be combined into the density function

$$f(y_{it}|\mathbf{x}_{it}) = g(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}) = \Lambda[(2y_{it} - 1)\mathbf{x}_{it}'\boldsymbol{\beta}].$$

Now, use this result to construct the log-likelihood function for a sample of data on $(y_{it}, \mathbf{x}_{it})$. (*Note:* We will be ignoring the panel aspect of the data set. Build the model as if this were a cross section.)

b. Derive the likelihood equations for estimation of $\boldsymbol{\beta}$.
c. Derive the second derivatives matrix of the log-likelihood function. (*Hint:* The following will prove useful in the derivation: $d\Lambda(t)/dt = \Lambda(t)[1 - \Lambda(t)]$.)
d. Show how to use Newton's method to estimate the parameters of the model.
e. Does the method of scoring differ from Newton's method? Derive the negative of the expectation of the second derivatives matrix.
f. Obtain maximum likelihood estimates of the parameters for the data and variables noted. Report your results, estimates, standard errors, and so on, as well as the value of the log likelihood.
g. Test the hypothesis that the coefficients on female and marital status are zero. Show how to do the test using Wald, LM, and LR tests, and then carry out the tests.
h. Test the hypothesis that all the coefficients in the model save for the constant term are equal to zero.

2. The geometric distribution used in Examples 14.13, 14.17, 14.18, and 14.22 would not be the typical choice for modeling a count such as DocVis. The Poisson model suggested at the beginning of Section 14.11.1 would be the more natural choice (at least at the first step in an analysis). Redo the calculations in Exercises 14.13 and 14.17 using a Poisson model rather than a geometric model. Do the results change very much? It is difficult to tell from the coefficient estimates. Compute the partial effects for the Poisson model and compare them to the partial effects shown in Table 14.11.
3. (This application will require an optimizer. Maximization of a user-supplied function is provided by commands in *Stata*, *R*, *SAS*, *EViews* or *NLOGIT*.) Use the following pseudo-code to generate a random sample of 1,000 observations on y from a mixed normals population:

```
Set the seed of the random number generator at any specific value.
Generate two sets of 1,000 random draws from normal populations
  with standard deviations 1. For the means, use 1 for y1 and
  5 for y2.
Generate a set of 1,000 random draws, c, from uniform(0,1)
  population.
```

```
For each observation, if c < .3, y = y1; if c ≥ .3, use y = y2.
```

The log-likelihood function for the mixture of two normals is given in (14-89). (The first step sets the seed at a particular value so that you can replicate your calculation of the data sets.)

a. Find the values that maximize the log-likelihood function. As starting values, use the sample mean of y (the same value) and sample standard deviation of y (again, same value) and 0.5 for π.
b. You should have observed the iterations in part a never get started. Try again using $0.9\overline{y}$, $.9s_y$, $1.1\overline{y}$, $1.1s_y$, and 0.5. This should be much more satisfactory.
c. Experiment with the estimator by generating y_1 and y_2 with more similar means, such as 1 and 3, or 1 and 2.