15

SIMULATION-BASED ESTIMATION AND INFERENCE AND RANDOM PARAMETER MODELS

15.1 INTRODUCTION

Simulation-based methods have become increasingly popular in econometrics. They are extremely computer intensive, but steady improvements in recent years in computation hardware and software have reduced that cost enormously. The payoff has been in the form of methods for solving estimation and inference problems that have previously been unsolvable in analytic form. The methods are used for two main functions. First, **simulation**-based methods are used to infer the characteristics of random variables, including estimators, functions of estimators, test statistics, and so on, by sampling from their distributions. Second, simulation is used in constructing estimators that involve complicated integrals that do not exist in a closed form that can be evaluated. In such cases, when the integral can be written in the form of an expectation, simulation methods can be used to evaluate it to within acceptable degrees of approximation by estimating the expectation as the mean of a random sample. The technique of maximum simulated likelihood (MSL) is essentially a classical sampling theory counterpart to the hierarchical Bayesian estimator considered in Chapter 16. Since the celebrated paper of Berry, Levinsohn, and Pakes (1995), and the review by McFadden and Train (2000), maximum simulated likelihood estimation has been used in a large and growing number of studies.

The following are three examples from earlier chapters that have relied on simulation methods.

Example 15.1 Inferring the Sampling Distribution of the Least Squares Estimator

In Example 4.1, we demonstrated the idea of a sampling distribution by drawing several thousand samples from a population and computing a least squares coefficient with each sample. We then examined the distribution of the sample of linear regression coefficients. A histogram suggested that the distribution appeared to be normal and centered over the true population value of the coefficient.

Example 15.2 Bootstrapping the Variance of the LAD Estimator

In Example 4.3, we compared the asymptotic variance of the least absolute deviations (LAD) estimator to that of the ordinary least squares (OLS) estimator. The form of the asymptotic variance of the LAD estimator is not known except in the special case of normally distributed disturbances. We relied, instead, on a random sampling method to approximate features of the sampling distribution of the LAD estimator. We used a device (bootstrapping) that allowed us to draw a sample of observations from the population that produces the estimator. With that random sample, by computing the corresponding sample statistics, we can infer characteristics of the distribution such as its variance and its 2.5th and 97.5th percentiles, which can be used to construct a confidence interval.

Example 15.3 Least Simulated Sum of Squares

Familiar estimation and inference methods, such as least squares and maximum likelihood, rely on closed form expressions that can be evaluated exactly [at least in principle—likelihood equations such as (14-4) may require an iterative solution]. Model building and analysis often require evaluation of expressions that cannot be computed directly. Familiar examples include expectations that involve integrals with no closed form such as the random effects nonlinear regression model presented in Section 14.14.4. The estimation problem posed there involved nonlinear least squares estimation of the parameters of

$$E[y_{it}|\mathbf{x}_{it}, u_i] = h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i).$$

Minimizing the sum of squares,

$$S(\boldsymbol{\beta}) = \sum_i \sum_t [y_{it} - h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)]^2,$$

is not feasible because u_i is not observed. In this formulation,

$$E[y|\mathbf{x}_{it}] = E_u E[y_{it}|\mathbf{x}_{it}, u_i] = \int_u E[y_{it}|\mathbf{x}_{it}, u_i] f(u_i) du_i,$$

so the feasible estimation problem would involve the sum of squares,

$$S^*(\boldsymbol{\beta}) = \sum_i \sum_t \left[y_{it} - \int_u h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) f(u_i) du_i \right]^2.$$

When the function is linear and u_i is normally distributed, this is a simple problem—it reduces to ordinary least squares. If either condition is not met, then the integral generally remains in the estimation problem. Although the integral,

$$E_u[h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)] = \int_u h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) f(u_i) du_i,$$

cannot be computed, if a large sample of R observations from the population of u_i, that is, u_{ir}, $r = 1, \ldots R$, were observed, then by virtue of the law of large numbers, we could rely on

$$\begin{aligned} \text{plim}(1/R)\sum_r h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_{ir}) &= E_u E[y_{it}|\mathbf{x}_{it}, u_i] \\ &= \int_u h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) f(u_i) du_i. \end{aligned} \qquad \textbf{(15-1)}$$

We are suppressing the extra parameter, σ_u, which would become part of the estimation problem. A convenient way to formulate the problem is to write $u_i = \sigma_u v_i$ where v_i has zero mean and variance one. By using this device, integrals can be replaced with sums that are feasible to compute. Our "simulated sum of squares" becomes

$$S_{\text{simulated}}(\boldsymbol{\beta}) = \sum_i \sum_t \left[y_{it} - (1/R)\sum_r h(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u v_{ir}) \right]^2, \qquad \textbf{(15-2)}$$

which can be minimized by conventional methods. As long as (15-1) holds, then

$$\frac{1}{nT}\sum_i \sum_t \left[y_{it} - (1/R)\sum_r h(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u v_{ir}) \right]^2 \xrightarrow{p} \frac{1}{nT}\sum_i \sum_t \left[y_{it} - \int_v h(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u v_i) f(v_i) dv_i \right]^2 \qquad \textbf{(15-3)}$$

and it follows that with sufficiently increasing R, the $\boldsymbol{\beta}$ that minimizes the left-hand side converges (in nT) to the same parameter vector that minimizes the probability limit of the right-hand side. We are thus able to substitute a computer simulation for the intractable computation on the right-hand side of the expression.

This chapter will describe some of the common applications of simulation methods in econometrics. We begin in Section 15.2 with the essential tool at the heart of all the computations, random number generation. Section 15.3 describes simulation-based inference using the method of Krinsky and Robb as an alternative to the delta method (see Section 4.4.4). The method of bootstrapping for inferring the features of the distribution of an estimator is described in Section 15.4. In Section 15.5, we will use a Monte Carlo study to learn about the behavior of a test statistic and the behavior of the fixed effects estimator in some nonlinear models. Sections 15.6 through 15.9 present simulation-based estimation methods. The essential ingredient of this entire set of results is the computation of integrals. Section 15.6.1 describes an application of a simulation-based estimator, a nonlinear random effects model. Section 15.6.2 discusses methods of integration. Then, the methods are applied to the estimation of the random effects model. Sections 15.7 through 15.9 describe several techniques and applications, including maximum simulated likelihood estimation for random parameter and hierarchical models. A third major (perhaps *the* major) application of simulation-based estimation in the current literature is Bayesian analysis using Markov Chain Monte Carlo (MCMC or MC^2) methods. Bayesian methods are discussed separately in Chapter 16. Sections 15.10 and 15.11 consider two remaining aspects of modeling parameter heterogeneity, estimation of individual specific parameters, and a comparison of modeling with continuous distributions to less parametric modeling with discrete distributions using latent class models.

15.2 RANDOM NUMBER GENERATION

All the techniques we will consider here rely on samples of observations from an underlying population. We will sometimes call these *random samples*, though it will emerge shortly that they are never actually random. One of the important aspects of this entire body of research is the need to be able to replicate one's computations. If the samples of draws used in any kind of simulation-based analysis were truly random, then that would be impossible. Although the samples we consider here will appear to be random, they are, in fact, deterministic—the samples can be replicated. For this reason, the sampling methods described in this section are more often labeled *pseudo–random number generators*. (This does raise an intriguing question: Is it possible to generate truly random draws from a population with a computer? The answer for practical purposes is no.) This section will begin with a description of some of the mechanical aspects of random number generation. We will then detail the methods of generating particular kinds of random samples.[1]

15.2.1 GENERATING PSEUDO-RANDOM NUMBERS

Data are generated internally in a computer using **pseudo–random number generators**. These computer programs generate sequences of values that appear to be strings of draws from a specified probability distribution. There are many types of random

[1] See Train (2009, Chapter 3) for extensive further discussion.

number generators, but most take advantage of the inherent inaccuracy of the digital representation of real numbers. The method of generation is usually by the following steps:

1. Set a **seed**.
2. Update the seed by $\text{seed}_j = \text{seed}_{j-1} \times s$ value.
3. $x_j = \text{seed}_j \times x$ value.
4. Transform x_j if necessary, and then move x_j to desired place in memory.
5. Return to step 2, or exit if no additional values are needed.

Random number generators produce sequences of values that resemble strings of random draws from the specified distribution. In fact, the sequence of values produced by the preceding method is not truly random at all; it is a deterministic **Markov chain** of values. The set of 32 bits in the random value only appear random when subjected to certain tests.[2] Because the series is, in fact, deterministic, at any point that this type of generator produces a value it has produced before, it must thereafter replicate the entire sequence. Because modern digital computers typically use 32-bit double words to represent numbers, it follows that the longest string of values that this kind of generator can produce is $2^{32} - 1$ (about 4.3 billion). This length is the **period** of a random number generator. (A generator with a shorter period than this would be inefficient, because it is possible to achieve this period with some fairly simple algorithms.) Some improvements in the periodicity of a generator can be achieved by the method of **shuffling**. By this method, a set of, say, 128 values is maintained in an array. The random draw is used to select one of these 128 positions from which the draw is taken and then the value in the array is replaced with a draw from the generator. The period of the generator can also be increased by combining several generators.[3] The most popular random number generator in current use is the **Mersenne Twister**,[4] which has a period of about $2^{20,000}$.

The deterministic nature of pseudo–random number generators is both a flaw and a virtue. Many Monte Carlo studies require billions of draws, so the finite period of any generator represents a nontrivial consideration. On the other hand, being able to reproduce a sequence of values just by resetting the seed to its initial value allows the researcher to replicate a study.[5] The seed itself can be a problem. It is known that certain seeds in particular generators will produce shorter series or series that do not pass randomness tests. For example, *congruential* generators of the sort just discussed should be started from odd seeds.

15.2.2 SAMPLING FROM A STANDARD UNIFORM POPULATION

The output of the generator described in Section 15.2.1 will be a pseudo-draw from the $U\,[0, 1]$ population. (In principle, the draw should be from the closed interval [0, 1]. However, the actual draw produced by the generator will be strictly between zero and one with probability just slightly below one. In the application described, the draw will be constructed from the sequence of 32 bits in a double word. All

[2] See Press et al. (1986).

[3] See L'Ecuyer (1998), Gentle (2002, 2003), and Greene (2007b).

[4] See Matsumoto, and Nishimura (1998).

[5] Readers of empirical studies are often interested in replicating the computations. In Monte Carlo studies, at least in principle, data can be replicated efficiently merely by providing the random number generator and the seed.

but two of the $2^{31} - 1$ strings of bits will produce a value in (0, 1). The practical result is consistent with the theoretical one, that the probabilities attached to the terminal points are zero also.) When sampling from a standard uniform, U [0, 1] population, the sequence is a kind of difference equation, because given the initial seed, x_j is ultimately a function of x_{j-1}. In most cases, the result at step 3 is a pseudo-draw from the continuous uniform distribution in the range zero to one, which can then be transformed to a draw from another distribution by using the fundamental probability transformation.

15.2.3 SAMPLING FROM CONTINUOUS DISTRIBUTIONS

One is usually interested in obtaining a sequence of draws, $x_1, \ldots, x_R$, from some particular population such as the normal with mean μ and variance σ^2. A sequence of draws from $U[0, 1]$, $u_1, \ldots, u_R$, produced by the random number generator, is an intermediate step. These will be transformed into draws from the desired population. A common approach is to use the **fundamental probability transformation**. For continuous distributions, this is done by treating the draw, $u_r = F_r$, as if F_r was $F(x_r)$, where F (.) is the cdf of x. For example, if we desire draws from the exponential distribution with known θ, then $F(x) = 1 - \exp(-\theta x)$. The inverse transform is $x = (-1/\theta)\ln(1 - F)$. For example, for a draw of $u = 0.4$ with $\theta = 5$, the associated x would be $(-1/5)\ln(1 - 0.4) = 0.1022$. For the logistic population with cdf $F(x) = \Lambda(x) = \exp(x)/[1 + \exp(x)]$, the inverse transformation is $x = \ln[F/(1 - F)]$. There are many references, for example, Evans, Hastings, and Peacock (2010) and Gentle (2003), that contain tables of inverse transformations that can be used to construct random number generators.

One of the most common applications is the draws from the standard normal distribution. This is complicated because there is no closed form for $\Phi^{-1}(F)$. There are several ways to proceed. A well-known approximation to the inverse function is given in Abramovitz and Stegun (1971),

$$\Phi^{-1}(F) = x \approx T - \frac{c_0 + c_1 T + c_2 T^2}{1 + d_1 T + d_2 T^2 + d_3 T^3},$$

where $T = [\ln(1/H^2)]^{1/2}$ and $H = F$ if $F > 0.5$ and $1 - F$ otherwise. The sign is then reversed if $F < 0.5$. A second method is to transform the $U[0, 1]$ values directly to a standard normal value. The Box–Muller (1958) method is $z = (-2 \ln u_1)^{1/2} \cos(2\pi\, u_2)$, where u_1 and u_2 are two independent $U[0, 1]$ draws. A second $N[0, 1]$ draw can be obtained from the same two values by replacing cos with sin in the transformation. The Marsaglia–Bray (1964) generator is $z_i = x_i[-(2/v)\ln v]^{1/2}$, where $x_i = 2u_i - 1$, u_i is a random draw from $U[0, 1]$ and $v = u_1^2 + u_2^2, i = 1, 2$. The pair of draws is rejected and redrawn if $v \geq 1$.

Sequences of draws from the standard normal distribution can easily be transformed into draws from other distributions by making use of the results in Section B.4. For example, the square of a standard normal draw will be a draw from chi-squared [1], and the sum of K chi-squared [1] is chi-squared [K]. From this relationship, it is possible to produce samples from the chi-squared [K], *t[n]*, and *F[K,n]* distributions.

A related problem is obtaining draws from the truncated normal distribution. The random variable with truncated normal distribution is obtained from one with a normal distribution by discarding the part of the range above a value U and below a value L. The density of the resulting random variable is that of a normal distribution restricted to the range $[L, U]$. The truncated normal density is

$$f(x \,|\, L \leq x \leq U) = \frac{f(x)}{\text{Prob}[L \leq x \leq U]} = \frac{(1/\sigma)\phi[(x - \mu)/\sigma]}{\Phi[(U - \mu)/\sigma] - \Phi[(L - \mu)/\sigma]},$$

where $\phi(t) = (2\pi)^{-1/2}\exp(-t^2/2)$ and $\Phi(t)$ is the cdf. An obviously inefficient (albeit effective) method of drawing values from the truncated normal $[\mu, \sigma^2]$ distribution in the range $[L,U]$ is simply to draw F from the $U[0, 1]$ distribution and transform it first to a standard normal variate as discussed previously and then to the $N[\mu, \sigma^2]$ variate by using $x = \mu + \sigma\Phi^{-1}(F)$. Finally, the value x is retained if it falls in the range $[L,U]$ and discarded otherwise. This rejection method will require, on average, $1/\{\Phi[(U - \mu)/\sigma] - \Phi[(L - \mu)/\sigma]\}$ draws per observation, which could be substantial. A direct transformation that requires only one draw is as follows: Let $P_j = \Phi[(j - \mu)/\sigma], j = L, U$. Then

$$x = \mu + \sigma\Phi^{-1}[P_L + F \times (P_U - P_L)]. \tag{15-4}$$

15.2.4 SAMPLING FROM A MULTIVARIATE NORMAL POPULATION

Many applications, including the method of Krinsky and Robb in Section 15.3, involve draws from a multivariate normal distribution with specified mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. To sample from this K-variate distribution, we begin with a draw, $\mathbf{z}$, from the K-variate standard normal distribution. This is done by first computing K independent standard normal draws, $z_1, \ldots, z_K$, using the method of the previous section and stacking them in the vector $\mathbf{z}$. Let $\mathbf{C}$ be a square root of $\boldsymbol{\Sigma}$ such that $\mathbf{CC}' = \boldsymbol{\Sigma}$. The desired draw is then $\mathbf{x} = \boldsymbol{\mu} + \mathbf{Cz}$, which will have covariance matrix $E[(\mathbf{x} - \boldsymbol{\mu}), (\mathbf{x} - \boldsymbol{\mu})'] = \mathbf{C}E[\mathbf{zz}']\mathbf{C}' = \mathbf{CIC}' = \boldsymbol{\Sigma}$. For the square root matrix, the usual choice is the **Cholesky decomposition**, in which $\mathbf{C}$ is a lower triangular matrix. (See Section A.6.11.) For example, suppose we wish to sample from the bivariate normal distribution with mean vector $\boldsymbol{\mu}$, unit variances, and correlation coefficient ρ. Then,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{bmatrix}.$$

The transformation of two draws z_1 and z_2 is $x_1 = \mu_1 + z_1$ and $x_2 = \mu_2 + [\rho z_1 + (1 - \rho^2)^{1/2} z_2]$. Section 15.3 and Example 15.4 following show a more involved application.

15.2.5 SAMPLING FROM DISCRETE POPULATIONS

There is generally no inverse transformation available for discrete distributions, such as the Poisson. An inefficient, though usually unavoidable, method for some distributions is to draw the F and then search sequentially for the smallest value that has cdf equal to or greater than F. For example, a generator for the Poisson distribution is constructed as follows. The pdf is $\text{Prob}[x = j] = p_j = \exp(-\mu)\mu^j/j!$ where μ is the mean of the random variable. The generator will use the recursion $p_j = p_{j-1} \times \mu/j, j = 1, \ldots$ beginning with $p_0 = \exp(-\mu)$. An algorithm that requires only a single random draw is as follows:

```
Initialize $c = \exp(-\mu), p = c, x = 0$;
Draw $F$ from $U[0, 1]$;
Deliver $x$; $*$ exit with draw $x$ if $c > F$;
Iterate: set $x = x + 1, p = p \times \mu/x, c = c + p$;
Return to $*$.
```

This method is based explicitly on the pdf and cdf of the distribution. Other methods are suggested by Knuth (1997) and Press et al. (2007).

The most common application of random sampling from a discrete distribution is, fortunately, also the simplest. The method of bootstrapping, and countless other applications involve random samples of draws from the **discrete uniform distribution**, $\text{Prob}(x = j) = 1/n, j = 1, \ldots, n$. In the bootstrapping application, we are going to draw random samples of observations from the sequence of integers $1, \ldots, n$, where each value must be equally likely. In principle, the random draw could be obtained by partitioning the unit interval into n equal parts, $[0, a_1), [a_1, a_2), \ldots, [a_{n-2}, a_{n-1}), [a_{n-1}, 1]; a_j = j/n, j = 1, \ldots, n - 1$. Then, random draw F delivers $x = j$ if F falls into interval j. This would entail a search, which could be time consuming. However, a simple method that will be much faster is simply to deliver $x =$ the integer part of $(n \times F + 1.0)$. (Once again, we are making use of the practical result that F will equal exactly 1.0—and x will equal $n + 1$—with ignorable probability.)

15.3 SIMULATION-BASED STATISTICAL INFERENCE: THE METHOD OF KRINSKY AND ROBB

Most of the theoretical development in this text has concerned the statistical properties of estimators—that is, the characteristics of sampling distributions such as the mean (probability limits), variance (asymptotic variance), and quantiles (such as the boundaries for confidence intervals). In cases in which these properties cannot be derived explicitly, it is often possible to infer them by using random sampling methods to draw samples from the population that produced an estimator and deduce the characteristics from the features of such a random sample. In Example 4.4, we computed a set of least squares regression coefficients, $b_1, \ldots, b_K$, and then examined the behavior of a nonlinear function $c_k = b_k/(1 - b_m)$ using the delta method. In some cases, the asymptotic properties of nonlinear functions such as these are difficult to derive directly from the theoretical distribution of the parameters. The sampling methods described here can be used for that purpose. A second common application is learning about the behavior of test statistics. For example, in Sections 5.3.3 and 14.6.3 [see (14-53)], we defined a Lagrange multiplier statistic for testing the hypothesis that certain coefficients are zero in a linear regression model. Under the assumption that the disturbances are normally distributed, the statistic has a limiting chi-squared distribution, which implies that the analyst knows what critical value to employ if he uses this statistic. Whether the statistic has this distribution if the disturbances are not normally distributed is unknown. Monte Carlo methods can be helpful in determining if the guidance of the chi-squared result is useful in more general cases. Finally, in Section 14.7, we defined a two-step maximum likelihood estimator. Computation of the asymptotic variance of such an estimator can be challenging. Monte Carlo methods, in particular, bootstrapping methods, can be used as an effective substitute for the intractable derivation of the appropriate asymptotic distribution of an estimator. This and the next two sections will detail these three procedures and develop applications to illustrate their use.

The method of Krinsky and Robb is suggested as a way to estimate the asymptotic covariance matrix of $\mathbf{c} = \mathbf{f}(\mathbf{b})$, where $\mathbf{b}$ is an estimated parameter vector with asymptotic covariance matrix $\boldsymbol{\Sigma}$ and $\mathbf{f}(\mathbf{b})$ defines a set of possibly nonlinear functions of $\mathbf{b}$. We assume that $\mathbf{f}(\mathbf{b})$ is a set of continuous and continuously differentiable functions that do not involve the sample size and whose derivatives do not equal zero at $\boldsymbol{\beta} = \text{plim } \mathbf{b}$. (These are the conditions underlying the Slutsky theorem in Section D.2.3.) In Section 4.6, we used the delta method to estimate the asymptotic covariance matrix of $\mathbf{c}$; $\text{Est.Asy.Var}[\mathbf{c}] = \mathbf{GSG}'$, where $\mathbf{S}$ is the estimate of $\boldsymbol{\Sigma}$ and $\mathbf{G}$ is the matrix of partial derivatives, $\mathbf{G} = \partial\mathbf{f}(\mathbf{b})/\partial\mathbf{b}'$. The recent literature contains some occasional skepticism about the accuracy of the delta method. The method of Krinsky and Robb (1986, 1990, 1991) is often suggested as an alternative. In a study of the behavior of estimated elasticities based on a translog model, the authors (1986) advocated an alternative approach based on Monte Carlo methods and the law of large numbers. We have consistently estimated $\boldsymbol{\beta}$ and $(\sigma^2/n)\mathbf{Q}^{-1}$, the mean and variance of the asymptotic normal distribution of the estimator $\mathbf{b}$, with $\mathbf{b}$ and $s^2(\mathbf{X}'\mathbf{X})^{-1}$. It follows that we could estimate the mean and variance of the distribution of a function of $\mathbf{b}$ by drawing a random sample of observations from the asymptotic normal population generating $\mathbf{b}$, and using the empirical mean and variance of the sample of functions to estimate the parameters of the distribution of the function. The quantiles of the sample of draws, for example, the 0.025th and 0.975th quantiles, can be used to estimate the boundaries of a confidence interval of the functions. The multivariate normal sample would be drawn using the method described in Section 15.2.4.

Krinsky and Robb (1986) reported huge differences in the standard errors produced by the delta method compared to the simulation-based estimator. In a subsequent paper (1990), they reported that the entire difference could be attributed to a bug in the software they used—upon redoing the computations, their estimates were essentially the same with the two methods. It is difficult to draw a conclusion about the effectiveness of the delta method based on the received results—it does seem at this juncture that the delta method remains an effective device that can often be employed with a hand calculator as opposed to the much more computation-intensive Krinsky and Robb (1986) technique. Unfortunately, the results of any comparison will depend on the data, the model, and the functions being computed. The amount of nonlinearity in the sense of the complexity of the functions seems not to be the answer. Krinsky and Robb's case was motivated by the extreme complexity of the elasticities in a translog model. In another study, Hole (2006) examines a similarly complex problem and finds that the delta method still appears to be the more accurate procedure.

Example 15.4 Long-Run Elasticities

A dynamic version of the demand for gasoline model is estimated in Example 4.7. The model is

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(Income/Pop)_t + \beta_4 \ln P_{nc,t} \\ + \beta_5 \ln P_{uc,t} + \gamma \ln (G/Pop)_{t-1} + \varepsilon_t.$$

In this model, the short-run price and income elasticities are β_2 and β_3. The long-run elasticities are $\phi_2 = \beta_2/(1 - \gamma)$ and $\phi_3 = \beta_3/(1 - \gamma)$, respectively. To estimate the long-run elasticities, we estimated the parameters by least squares and then computed these two nonlinear functions of the estimates. Estimates of the full set of model parameters and the estimated asymptotic covariance matrix are given in Example 4.7. The delta method was used to estimate the asymptotic standard errors for the estimates of ϕ_2 and ϕ_3. The three estimates of the specific parameters and the 3 × 3 submatrix of the estimated asymptotic covariance matrix are

$$Est.\begin{pmatrix}\beta_2\\ \beta_3\\ \gamma\end{pmatrix} = \begin{pmatrix}b_2\\ b_3\\ c\end{pmatrix} = \begin{pmatrix}-0.069532\\ 0.164047\\ 0.830971\end{pmatrix},$$

$$Est.\,Asy.\,Var\begin{pmatrix}b_2\\ b_3\\ c\end{pmatrix} = \begin{pmatrix}0.00021705 & 1.61265e{-5} & -0.0001109\\ 1.61265e{-5} & 0.0030279 & -0.0021881\\ -0.0001109 & -0.0021881 & 0.0020943\end{pmatrix}.$$

The method suggested by Krinsky and Robb would use a random number generator to draw a large trivariate sample, $(b_2, b_3, c)_r$, $r = 1, \ldots, R$, from the normal distribution with this mean vector and covariance matrix, and then compute the sample of observations on f_2 and f_3 and obtain the empirical mean and variance and the 0.025 and 0.975 quantiles from the sample. The method of drawing such a sample is shown in Section 15.2.4. We will require the square root of the covariance matrix. The Cholesky matrix is

$$\mathbf{C} = \begin{pmatrix}0.0147326 & 0 & 0\\ 0.00109461 & 0.0550155 & 0\\ -0.0075275 & -0.0396227 & 0.0216259\end{pmatrix}$$

The sample is drawn by obtaining vectors of three random draws from the standard normal population, $\mathbf{v}_r = (v_1, v_2, v_3)'_r$, $r = 1, \ldots, R$. The draws needed for the estimation are then obtained by computing $\mathbf{b}_r = \mathbf{b} + \mathbf{Cv_r}$, where $\mathbf{b}$ is the set of least squares estimates. We then compute the sample of estimated long-run elasticities, $f_{2r} = b_{2r}/(1 - c_r)$ and $f_{3r} = b_{3r}/(1 - c_r)$. The mean and standard deviation of the sample observations constitute the estimates of the functions and asymptotic standard errors.

Table 15.1 shows the results of these computations based on 1,000 draws from the underlying distribution. The estimates from Example 4.4 using the delta method are shown as well. The two sets of estimates are in quite reasonable agreement. For a 95% confidence interval for ϕ_2 based on the estimates, the t distribution with $51 - 6 = 45$ degrees of freedom and the delta method would be $-0.411358 \pm 2.014(0.152296)$. The result for ϕ_3 would be $0.970522 \pm 2.014(0.162386)$. These are shown in Table 15.2 with the same computation

TABLE 15.1 Simulation Results

	Regression Estimate		*Simulated Values*	
	Estimate	***Std.Err.***	***Mean***	***Std.Dev.***
β_2	−0.069532	0.0147327	−0.068791	0.0138485
β_3	0.164047	0.0550265	0.162634	0.0558856
γ	0.830971	0.0457635	0.831083	0.0460514
ϕ_2	−0.411358	0.152296	−0.453815	0.219110
ϕ_3	0.970522	0.162386	0.950042	0.199458

TABLE 15.2 Estimated Confidence Intervals

	ϕ_2		ϕ_3	
	Lower	***Upper***	***Lower***	***Upper***
Delta Method	−0.718098	−0.104618	0.643460	1.297585
Krinsky and Robb	−0.895125	−0.012505	0.548313	1.351772
Sample Quantiles	−0.983866	−0.209776	0.539668	1.321617

using the Krinsky and Robb estimated standard errors. The table also shows the empirical estimates of these quantiles computed using the 26th and 975th values in the samples. There is reasonable agreement in the estimates, though a considerable amount of sample variability is also evident, even in a sample as large as 1,000.

We note, finally, that it is generally not possible to replicate results such as these across software platforms because they use different random number generators. Within a given platform, replicability can be obtained by setting the seed for the random number generator.

15.4 BOOTSTRAPPING STANDARD ERRORS AND CONFIDENCE INTERVALS

The technique of bootstrapping is used to obtain a description of the sampling properties of empirical estimators using the sample data themselves, rather than broad theoretical results.[6] Suppose that $\hat{\boldsymbol{\theta}}_n$ is an estimator of a parameter vector $\boldsymbol{\theta}$ based on a sample, $\mathbf{Z} = [(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)]$. An approximation to the statistical properties of $\hat{\boldsymbol{\theta}}_n$ can be obtained by studying a sample of bootstrap estimators $\hat{\boldsymbol{\theta}}(b)_m$, $b = 1, \ldots, B$, obtained by sampling m observations, with replacement, from $\mathbf{Z}$ and recomputing $\hat{\boldsymbol{\theta}}$ with each sample. After a total of B times, the desired sampling characteristic is computed from

$$\hat{\boldsymbol{\Theta}} = [\hat{\boldsymbol{\theta}}(1)_m, \hat{\boldsymbol{\theta}}(2)_m, \ldots, \hat{\boldsymbol{\theta}}(B)_m].$$

The most common application of bootstrapping for consistent estimators when n is reasonably large is approximating the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\theta}}_n$ with

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}_n] = \frac{1}{B-1}\sum_{b=1}^{B}[\hat{\boldsymbol{\theta}}(b)_m - \bar{\hat{\boldsymbol{\theta}}}_B][\hat{\boldsymbol{\theta}}(b)_m - \bar{\hat{\boldsymbol{\theta}}}_B]', \qquad \textbf{(15-5)}$$

where $\bar{\hat{\boldsymbol{\theta}}}_B$ is the average of the B bootstrapped estimates of $\boldsymbol{\theta}$. There are few theoretical prescriptions for the number of replications, B. Andrews and Buchinsky (2000) and Cameron and Trivedi (2005, pp. 361–362) make some suggestions for particular applications; Davidson and MacKinnon (2006) recommend at least 399. Several hundred is the norm; we have used 1,000 in our application to follow.[7] An application to the least absolute deviations estimator in the linear model is shown in the following example and in Chapter 4.

[6] See Efron (1979), Efron and Tibshirani (1994), and Davidson and Hinkley (1997), Brownstone and Kazimi (1998), Horowitz (2001), MacKinnon (2002), and Davidson and MacKinnon (2006).

[7] For applications, see, for example, Veall (1987, 1992), Vinod (1993), and Vinod and Raj (1994). Extensive surveys of uses and methods in econometrics appear in Cameron and Trivedi (2005), Horowitz (2001), and Davidson and MacKinnon (2006).

15.4.1 TYPES OF BOOTSTRAPS

The preceding is known as a **paired bootstrap**. The pairing is the joint sampling of y_i and $\mathbf{x}_i$. An alternative approach in a regression context would be to sample the observations on $\mathbf{x}_i$ once and then with each $\mathbf{x}_i$ sampled, generate the accompanying y_i by randomly generating the disturbance, then $\hat{y}_i(b) = \mathbf{x}_i(b)'\hat{\boldsymbol{\theta}}_n + \hat{\varepsilon}_i(b)$. This would be a **parametric bootstrap** in that in order to simulate the disturbances, we need either to know (or assume) the data-generating process that produces ε_i. In other contexts, such as in discrete choice modeling in Chapter 17, one would bootstrap sample the exogenous data in the model and then generate the dependent variable by this method using the appropriate underlying DGP. This is the approach used in 15.5.2 and in Greene (2004b) in a study of the incidental parameters problem in several limited dependent variable models. The obvious disadvantage of the parametric bootstrap is that one cannot learn of the influence of an unknown DGP for ε by assuming it is known. For example, if the bootstrap is being used to accommodate unknown heteroscedasticity in the model, then a parametric bootstrap that assumes homoscedasticity would defeat the purpose. The more natural application would be a **nonparametric bootstrap**, in which both $\mathbf{x}_i$ and y_i, and, implicitly, ε_i, are sampled simultaneously.

Example 15.5 Bootstrapping the Variance of the Median

There are few cases in which an exact expression for the sampling variance of the median is known. Example 15.7 examines the case of the median of a sample of 500 observations from the *t* distribution with 10 degrees of freedom. This is one of those cases in which there is no exact formula for the asymptotic variance of the median. However, we can use the bootstrap technique to estimate one empirically. In one run of the experiment, we obtained a sample of 500 observations for which we computed the median, −0.00786. We drew 100 samples of 500 with replacement from this sample of 500 and recomputed the median with each of these samples. The empirical square root of the mean squared deviation around this estimate of −0.00786 was 0.056. In contrast, consider the same calculation for the mean. The sample mean is −0.07247. The sample standard deviation is 1.08469, so the standard error of the mean is 0.04657. (The bootstrap estimate of the standard error of the mean was 0.052.) This agrees with our expectation in that the sample mean should generally be a more efficient estimator of the mean of the distribution in a large sample. There is another approach we might take in this situation. Consider the regression model $y_i = \alpha + \varepsilon_i$, where ε_i has a symmetric distribution with finite variance. The least absolute deviations estimator of the coefficient in this model is an estimator of the median (which equals the mean) of the distribution. So, this presents another estimator. Once again, the bootstrap estimator must be used to estimate the asymptotic variance of the estimator. Using the same data, we fit this regression model using the LAD estimator. The coefficient estimate is −0.05397 with a bootstrap estimated standard error of 0.05872. The estimated standard error agrees with the earlier one. The difference in the estimated coefficient stems from the different computations—the regression estimate is the solution to a linear programming problem while the earlier estimate is the actual sample median.

15.4.2 BIAS REDUCTION WITH BOOTSTRAP ESTIMATORS

The bootstrap estimation procedure has also been suggested as a method of reducing bias. In principle, we would compute $\hat{\boldsymbol{\theta}}_n - \text{bias}(\hat{\boldsymbol{\theta}}_n) = \hat{\boldsymbol{\theta}}_n - \{E[\hat{\boldsymbol{\theta}}_n] - \boldsymbol{\theta}\}$. Because neither $\boldsymbol{\theta}$ nor the exact expectation of $\hat{\boldsymbol{\theta}}_n$ is known, we estimate the first with the mean of the bootstrap replications and the second with the estimator itself. The revised estimator is

$$\hat{\boldsymbol{\theta}}_{n,B} = \hat{\boldsymbol{\theta}}_n - \left[\frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\theta}}(b)_m - \hat{\boldsymbol{\theta}}_n\right] = 2\hat{\boldsymbol{\theta}}_n - \bar{\hat{\boldsymbol{\theta}}}_B. \tag{15-6}$$

[Efron and Tibshirani (1994, p. 138) provide justification for what appears to be the wrong sign on the correction.] Davidson and MacKinnon (2006) argue that the smaller bias of the corrected estimator is offset by an increased variance compared to the uncorrected estimator.[8] The authors offer some other cautions for practitioners contemplating use of this technique. First, perhaps obviously, the extension of the method to samples with dependent observations presents some obstacles. For time-series data, the technique makes little sense—none of the bootstrapped samples will be a time series, so the properties of the resulting estimators will not satisfy the underlying assumptions needed to make the technique appropriate.

15.4.3 BOOTSTRAPPING CONFIDENCE INTERVALS

A second common application of bootstrapping methods is the computation of confidence intervals for parameters. This calculation will be useful when the underlying data-generating process is unknown, and the bootstrap method is being used to obtain appropriate standard errors for estimated parameters. A natural approach to bootstrapping confidence intervals for parameters would be to compute the estimated asymptotic covariance matrix using (15-5) and then form confidence intervals in the usual fashion. An improvement in terms of the bias of the estimator is provided by the **percentile method**.[9] By this technique, during each bootstrap replication, we compute

$$t_k^*(b) = \frac{\hat{\theta}_k(b) - \hat{\theta}_{n,k}}{se.(\hat{\theta}_{n,k})}, \tag{15-7}$$

where "k" indicates the kth parameter in the model, and $\hat{\theta}_{n,k}$, $s.e.(\hat{\theta}_{n,k})$ and $\hat{\theta}_k(b)$ are the original estimator and estimated standard error from the full sample and the bootstrap replicate. Then, with all B replicates in hand, the bootstrap confidence interval is

$$\hat{\theta}_{n,k} + t^*_{k[\alpha/2]}se.(\hat{\theta}_{n,k}) \text{ to } \hat{\theta}_{n,k} + t^*_{k[1-\alpha/2]}s.e.(\hat{\theta}_{n,k}). \tag{15-8}$$

(Note that $t^*_{k[\alpha/2]}$ is negative, which explains the plus sign in the left term.) For example, in our next application, next, we compute the estimator and the asymptotic covariance matrix using the full sample. We compute 1,000 bootstrap replications, and compute the t ratio in (15-7) for the education coefficient in each of the 1,000 replicates. After the bootstrap samples are accumulated, we sorted the results from (15-7), and the 25th and 975th largest values provide the values of t^*.

15.4.4 BOOTSTRAPPING WITH PANEL DATA: THE BLOCK BOOTSTRAP

Example 15.6 demonstrates the computation of a confidence interval for a coefficient using the bootstrap. The application uses the Cornwell and Rupert panel data set used in Example 11.4 and several later applications. There are 595 groups of seven observations in the data set. Bootstrapping with panel data requires an additional element in the computations. The bootstrap replications are based on sampling over i, not t. Thus, the bootstrap sample consists of n blocks of T (or T_i) observations—the ith group as a whole is sampled. This produces, then, a **block bootstrap** sample.

[8] See, as well, Cameron and Trivedi (2005).

[9] See Cameron and Trivedi (2005, p. 364).

Example 15.6 Block Bootstrapping Standard Errors and Confidence Intervals in a Panel

Example 11.4 presents least squares estimates and robust standard errors for the labor supply equation using Cornwell and Rupert's panel data set. There are 595 individuals and seven periods in the data set. As seen in the results in Table 11.1 (reproduced below), using a clustering correction in a robust covariance matrix for the least squares estimator produces substantial changes in the estimated standard errors. Table 15.3 reproduces the least squares coefficients and the standard errors associated with the conventional $s^2(\mathbf{X}'\mathbf{X})^{-1}$ and the robust standard errors using the clustering correction in column (3). The block bootstrapped standard errors using 1,000 bootstrap replications are shown in column (4). The ability of the bootstrapping procedure to detect and mimic the effect of the clustering that is evident in columns (3) and (4). Note, as well, the resemblance to the naïve bootstrap estimates in column (5) and the conventional, uncorrected standard errors in column (2).

We also computed a confidence interval for the coefficient on *Ed* using the conventional, symmetric approach, $b_{Ed} \pm 1.96s(b_{Ed})$, and the percentile method in (15-7) and (15-8). For the conventional estimator, we use $0.05670 \pm 1.96(0.00556) = [0.04580, 0.06760]$. For the bootstrap confidence interval method, we first computed and sorted the 1,000 *t* statistics based on (15-7). The 25th and 975th values were −2.148 and +1.966. The confidence interval is [0.04476, 0.06802].

Figure 15.1 shows a kernel density estimator of the distribution of the *t* statistics computed using (15-7) with the (approximate) standard normal density.

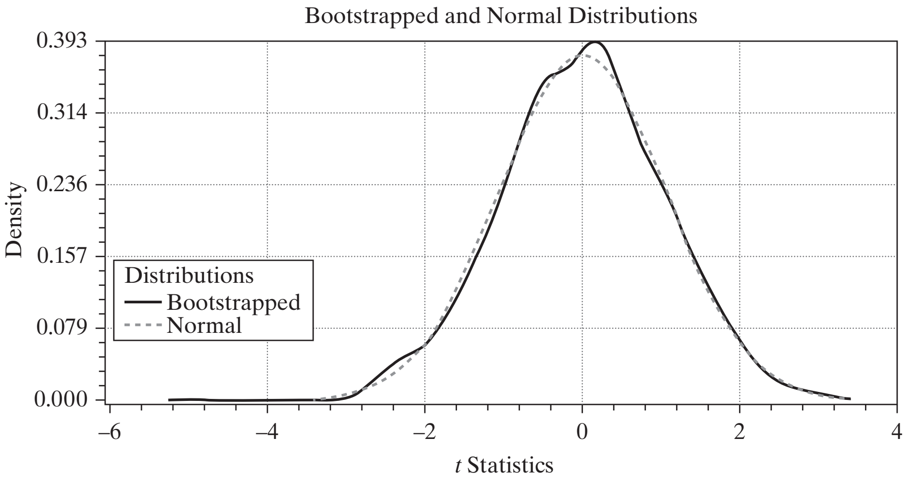
15.5 MONTE CARLO STUDIES

Simulated data generated by the methods of the preceding sections have various uses in econometrics. One of the more common applications is the analysis of the properties of estimators or in obtaining comparisons of the properties of estimators. For example,

TABLE 15.3 Bootstrap Estimates of Standard Errors for a Wage Equation

Variable	*(1) Least Squares Estimate*	*(2) Least Squares Standard Error*	*(3) Cluster Robust Standard Error*	*(4) Block Bootstrap Standard Error*	*(5) Simple Bootstrap Standard Error*
Constant	5.25112	0.07129	0.12355	0.12421	0.07761
Wks	0.00422	0.00108	0.00154	0.00159	0.00115
South	−0.05564	0.01253	0.02616	0.02557	0.01284
SMSA	0.15167	0.01207	0.02410	0.02383	0.01200
MS	0.04845	0.02057	0.04094	0.04208	0.02010
Exp	0.04010	0.00216	0.00408	0.00418	0.00213
Exp^2	−0.00067	0.00005	0.00009	0.00009	0.00005
Occ	−0.14001	0.01466	0.02724	0.02733	0.01539
Ind	0.04679	0.01179	0.02366	0.02350	0.01183
Union	0.09263	0.01280	0.02367	0.02390	0.01203
Ed	0.05670	0.00261	0.00556	0.00576	0.00273
Fem	−0.36779	0.02510	0.04557	0.04562	0.02390
Blk	−0.16694	0.02204	0.04433	0.04663	0.02103

FIGURE 15.1 Distributions of Test Statistics.



in time-series settings, most of the known results for characterizing the sampling distributions of estimators are asymptotic, large-sample results. But the typical time series is not very long, and descriptions that rely on T, the number of observations, going to infinity may not be very accurate. Exact finite-sample properties are usually intractable, however, which leaves the analyst with only the choice of learning about the behavior of the estimators experimentally.

In the typical application, one would either compare the properties of two or more estimators while holding the sampling conditions fixed or study how the properties of an estimator are affected by changing conditions such as the sample size or the value of an underlying parameter.

Example 15.7 Monte Carlo Study of the Mean Versus the Median

In Example D.8, we compared the asymptotic distributions of the sample mean and the sample median in random sampling from the normal distribution. The basic result is that both estimators are consistent, but the mean is asymptotically more efficient by a factor of

$$\frac{\text{Asy.Var[Median]}}{\text{Asy.Var[Mean]}} = \frac{\pi}{2} = 1.5708.$$

This result is useful, but it does not tell which is the better estimator in small samples, nor does it suggest how the estimators would behave in some other distribution. It is known that the mean is affected by outlying observations whereas the median is not. The effect is averaged out in large samples, but the small-sample behavior might be very different. To investigate the issue, we constructed the following experiment: We sampled 500 observations from the t distribution with d degrees of freedom by sampling $d + 1$ values from the standard normal distribution and then computing

$$t_{ir} = \frac{z_{ir,d+1}}{\sqrt{\frac{1}{d}\sum_{l=1}^{d} z_{ir,l}^2}}, \quad i = 1, \ldots, 500, \quad r = 1, \ldots, 100.$$

The t distribution with a low value of d was chosen because it has very thick tails and because large outlying values have high probability. For each value of d, we generated $R = 100$ replications. For each of the 100 replications, we obtained the mean and median. Because both are unbiased, we compared the mean squared errors around the true expectations using

$$M_d = \frac{(1/R)\sum_{r=1}^{R}(\text{median}_r - 0)^2}{(1/R)\sum_{r=1}^{R}(\bar{x}_r - 0)^2}.$$

We obtained ratios of 0.6761, 1.2779, and 1.3765 for $d = 3$, 6, and 10, respectively. (You might want to repeat this experiment with different degrees of freedom.) These results agree with what intuition would suggest. As the degrees of freedom parameter increases, which brings the distribution closer to the normal distribution, the sample mean becomes more efficient—the ratio should approach its limiting value of 1.5708 as d increases. What might be surprising is the apparent overwhelming advantage of the median when the distribution is very nonnormal even in a sample as large as 500.

The preceding is a very small application of the technique. In a typical study, there are many more parameters to be varied and more dimensions upon which the results are to be studied. One of the practical problems in this setting is how to organize the results. There is a tendency in Monte Carlo work to proliferate tables indiscriminately. It is incumbent on the analyst to collect the results in a fashion that is useful to the reader. For example, this requires some judgment on how finely one should vary the parameters of interest. One useful possibility that will often mimic the thought process of the reader is to collect the results of bivariate tables in carefully designed contour plots.

There are any number of situations in which Monte Carlo simulation offers the only method of learning about finite-sample properties of estimators. Still, there are a number of problems with Monte Carlo studies. To achieve any level of generality, the number of parameters that must be varied and hence the amount of information that must be distilled can become enormous. Second, they are limited by the design of the experiments, so the results they produce are rarely generalizable. For our example, we may have learned something about the t distribution, but the results that would apply in other distributions remain to be described. And, unfortunately, real data will rarely conform to any specific distribution, so no matter how many other distributions we analyze, our results would still only be suggestive. In more general terms, this problem of **specificity** [Hendry (1984)] limits most Monte Carlo studies to quite narrow ranges of applicability. There are very few that have proved general enough to have provided a widely cited result.

15.5.1 A MONTE CARLO STUDY: BEHAVIOR OF A TEST STATISTIC

Monte Carlo methods are often used to study the behavior of test statistics when their true properties are uncertain. This is often the case with Lagrange multiplier statistics. For example, Baltagi (2005) reports on the development of several new test statistics for panel data models such as a test for serial correlation. Examining the behavior of a test statistic is fairly straightforward. We are interested in two characteristics: the true **size of the test**—that is, the probability that it rejects the null hypothesis when that hypothesis is actually true (the probability of a type 1 error) and the **power of the test**—that is the probability that it will correctly reject a false null hypothesis (one minus the probability of a type 2 error). As we will see, the power of a test is a function of the alternative against which the null is tested.

To illustrate a Monte Carlo study of a test statistic, we consider how a familiar procedure behaves when the model assumptions are incorrect. Consider the linear regression model

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i, \quad \varepsilon_i \,|\, (x_i, z_i) \sim \text{N}[0, \sigma^2].$$

The Lagrange multiplier statistic for testing the null hypothesis that γ equals zero for this model is

$$\text{LM} = \mathbf{e}_0'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_0/(\mathbf{e}_0'\mathbf{e}_0/n),$$

where $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \mathbf{z})$ and $\mathbf{e}_0$ is the vector of least squares residuals obtained from the regression of $\mathbf{y}$ on the constant and $\mathbf{x}$ (and not $\mathbf{z}$). [See (14-53).] Under the assumptions of the preceding model, the large sample distribution of the LM statistic is chi squared with one degree of freedom. Thus, our testing procedure is to compute LM and then reject the null hypothesis $\gamma = 0$ if LM is greater than the critical value. We will use a nominal size of 0.05, so the critical value is 3.84. The theory for the statistic is well developed when the specification of the model is correct.[10] We are interested in two specification errors. First, how does the statistic behave if the normality assumption is not met? Because the LM statistic is based on the likelihood function, if some distribution other than the normal governs ε_i, then the LM statistic would not be based on the OLS estimator. We will examine the behavior of the statistic under the true specification that ε_i comes from a t distribution with five degrees of freedom. Second, how does the statistic behave if the homoscedasticity assumption is not met? The statistic is entirely wrong if the disturbances are heteroscedastic. We will examine the case in which the conditional variance is $\text{Var}[\varepsilon_i \,|\, x_i, z_i] = \sigma^2[\exp(0.2x_i)]^2$.

The design of the experiment is as follows: We will base the analysis on a sample of 50 observations. We draw 50 observations on x_i and z_i from independent N[0, 1] populations at the outset of each cycle. For each of 1,000 replications, we draw a sample of 50 ε_i's according to the assumed specification. The LM statistic is computed and the proportion of the computed statistics that exceed 3.84 is recorded. The experiment is repeated for $\gamma = 0$ to ascertain the true size of the test and for values of γ including $-1, \ldots, -0.2, -0.1, 0, 0.1, 0.2, \ldots, 1.0$ to assess the power of the test. The cycle of tests is repeated for the two scenarios, the $t[5]$ distribution and the model with heteroscedasticity.

Table 15.4 lists the results of the experiment. The "Normal" column in each panel shows the expected results for the LM statistic under the model assumptions for which it is appropriate. The size of the test appears to be in line with the theoretical results. Comparing the first and third columns in each panel, it appears that the presence of heteroscedasticity seems not to degrade the power of the statistic. But the different distributional assumption does. Figure 15.2 plots the values in the table, and displays the characteristic form of the power function for a test statistic.

15.5.2 A MONTE CARLO STUDY: THE INCIDENTAL PARAMETERS PROBLEM

Section 14.14.5 examines the maximum likelihood estimator of a panel data model with fixed effects,

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

[10] See, for example, Godfrey (1988).

TABLE 15.4 Power Functions for LM Test

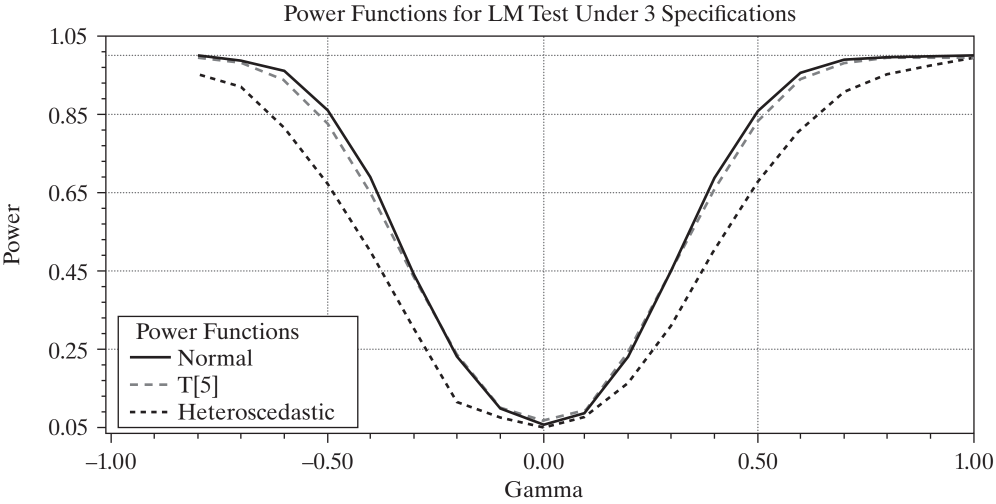
	Model				*Model*		
γ	***Normal***	***t[5]***	***Het.***	γ	***Normal***	***t[5]***	***Het.***
−1.0	1.000	0.993	1.000	0.1	0.090	0.083	0.098
−0.9	1.000	0.984	1.000	0.2	0.235	0.169	0.249
−0.8	0.999	0.953	0.996	0.3	0.464	0.320	0.457
−0.7	0.989	0.921	0.985	0.4	0.691	0.508	0.666
−0.6	0.961	0.822	0.940	0.5	0.859	0.680	0.835
−0.5	0.863	0.677	0.832	0.6	0.957	0.816	0.944
−0.4	0.686	0.500	0.651	0.7	0.989	0.911	0.984
−0.3	0.451	0.312	0.442	0.8	0.998	0.956	0.995
−0.2	0.236	0.177	0.239	0.9	1.000	0.976	0.998
−0.1	0.103	0.080	0.107	1.0	1.000	0.994	1.000
0.0	0.059	0.052	0.071				

where the individual effects may be correlated with x_{it}. The extra parameter vector $\boldsymbol{\theta}$ represents M other parameters that might appear in the model, such as the disturbance variance, σ_ε^2, in a linear regression model with normally distributed disturbance. The development there considers the mechanical problem of maximizing the log likelihood

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

with respect to the $n + K + M$ parameters $(\alpha_1, \ldots, \alpha_n, \boldsymbol{\beta}, \boldsymbol{\theta})$. A statistical problem with this estimator that was suggested was that there is a phenomenon labeled the **incidental**

FIGURE 15.2 Power Functions.



parameters problem.[11] With the exception of a very small number of specific models (such as the Poisson regression model in Section 18.4.1), the *brute force*, unconditional maximum likelihood estimator of the parameters in this model is inconsistent. The result is straightforward to visualize with respect to the individual effects. Suppose that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ were actually known. Then, each α_i would be estimated with T_i observations. Because T_i is assumed to be fixed (and small), there is no asymptotic result to provide consistency for the MLE of α_i. But $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are estimated with $\Sigma_i T_i = N$ observations, so their large sample behavior is less transparent. One known result concerns the logit model for binary choice (see Sections 17.2–17.4). Kalbfleisch and Sprott (1970), Andersen (1973), Hsiao (1996), and Abrevaya (1997) have established that in the binary logit model, if $T_i = 2$, then plim $\hat{\boldsymbol{\beta}}_{\text{MLE}} = 2\boldsymbol{\beta}$. Two other cases are known with certainty. In the linear regression model with fixed effects and normally distributed disturbances, the slope estimator, $\mathbf{b}_{\text{LSDV}}$, is unbiased and consistent, however, the MLE of the variance, σ^2, converges to $(T - 1)\sigma^2/T$. (The degrees of freedom correction will adjust for this, but the MLE does not correct for degrees of freedom.) Finally, in the Poisson regression model (Section 18.4.7.b), the unconditional MLE is consistent.[12] Almost nothing else is known with certainty—that is, as a firm theoretical result—about the behavior of the maximum likelihood estimator in the presence of fixed effects. The literature appears to take as given the qualitative wisdom of Hsiao and Abrevaya, that the FE/MLE is inconsistent when T is small and fixed. (The implication that the severity of the inconsistency declines as T increases makes sense, but, again, remains to be shown analytically.)

The result for the two-period binary logit model is a standard result for discrete choice estimation. Several authors, all using Monte Carlo methods, have pursued the result for the logit model for larger values of T.[13] Greene (2004) analyzed the incidental parameters problem for other discrete choice models using Monte Carlo methods. We will examine part of that study.

The current studies are preceded by a small study in Heckman (1981) which examined the behavior of the fixed effects MLE in the following experiment:

$$\begin{aligned}
z_{it} &= 0.1t + 0.5z_{i,t-1} + u_{it},\ z_{i0} = 5 + 10.0u_{i0},\\
u_{it} &\sim U[-0.5, 0.5],\ i = 1, \ldots, 100,\ t = 0, \ldots, 8,\\
Y_{it} &= \sigma_t\tau_i + \beta z_{it} + \varepsilon_{it},\ \tau_i \sim \text{N}[0,1],\ \varepsilon_{it} \sim \text{N}[0,1],\\
y_{it} &= 1 \text{ if } Y_{it} > 0,\ 0 \text{ otherwise.}
\end{aligned}$$

Heckman attempted to learn something about the behavior of the MLE for the probit model with $T = 8$. He used values of $\beta = -1.0, -0.1$, and 1.0 and $\sigma_\tau = 0.5, 1.0$, and 3.0. The mean values of the maximum likelihood estimates of β for the nine cases are as follows:

	$\beta = -1.0$	$\beta = -0.1$	$\beta = 1.0$
$\sigma_\tau = 0.5$	-0.96	-0.10	0.93
$\sigma_\tau = 1.0$	-0.95	-0.09	0.91
$\sigma_\tau = 3.0$	-0.96	-0.10	0.90.

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

[12] See Cameron and Trivedi (1988).

[13] See, for example, Katz (2001).

The findings here disagree with the received wisdom. Where there appears to be a bias (i.e., excluding the center column), it seems to be quite small, and toward, not away from, zero.

The Heckman study used a very small sample and, moreover, analyzed the fixed effects estimator in a random effects model. (*Note:* τ_i is independent of z_{it}.) Greene (2004a), using the same parameter values, number of replications, and sample design, found persistent biases away from zero on the order of 15 to 20%. Numerous authors have extended the logit result for $T = 2$ with larger values of T, and likewise persistently found biases away from zero that diminish with increases in T. Greene (2004a) redid the experiment for the logit model and then replicated it for the probit and ordered probit models. The experiment is designed as follows: All models are based on the same index function

$$\begin{aligned} w_{it} &= \alpha_i + \beta x_{it} + \delta d_{it}, \quad \text{where } \beta = \delta = 1,\\ x_{it} &\sim \text{N}[0,1],\, d_{it} = \mathbf{1}[x_{it} + h_{it} > 0], \qquad \text{where } h_{it} \sim \text{N}[0,1],\\ \alpha_i &= \sqrt{T}\,\bar{x}_i + v_i,\, v_i \sim \text{N}[0,1]. \end{aligned}$$

The regressors d_{it} and x_{it} are constructed to be correlated. The random term h_{it} is used to produce independent variation in d_{it}. There is, however, no within group correlation in x_{it} or d_{it} built into the data generator. (Other experiments suggested that the marginal distribution of x_{it} mattered little to the outcome of the experiment.) The correlations between the variables are approximately 0.7 between x_{it} and d_{it}, 0.4 between α_i and x_{it}, and 0.2 between α_i and d_{it}. The individual effect is produced from independent variation, v_i as well as the group mean of x_{it}. The latter is scaled by $\sqrt{T}$ to maintain the unit variances of the two parts—without the scaling, the covariance between α_i and x_{it} falls to zero as T increases and $\bar{x}_i$ converges to its mean of zero. Thus, the data generator for the index function satisfies the assumptions of the fixed effects model. The sample used for the results below contains $n = 1{,}000$ individuals. The data-generating processes for the discrete dependent variables are as follows:

$$\begin{aligned} &\textit{probit}: && y_{it} = \mathbf{1}[w_{it} + \varepsilon_{it} > 0],\, \varepsilon_{it} \sim \text{N}[0,1],\\ &\textit{ordered probit}: && y_{it} = \mathbf{1}[w_{it} + \varepsilon_{it} > 0] + \mathbf{1}[w_{it} + \varepsilon_{it} > 3],\, \varepsilon_{it} \sim \mathbf{N}[0,1],\\ &\textit{logit}: && y_{it} = \mathbf{1}[w_{it} + v_{it} > 0],\, v_{it} = \log[u_{it}/(1 - u_{it})],\, u_{it} \sim \text{U}[0,1]. \end{aligned}$$

(The three discrete dependent variables are described in Chapters 17 and 18.)

Table 15.5 reports the results of computing the MLE with 200 replications. Models were fit with $T = 2, 3, 5, 8, 10$, and 20. (*Note:* This includes Heckman's experiment.) Each model specification and group size (T) is fit 200 times with random draws for ε_{it} or u_{it}. The data on the regressors were drawn at the beginning of each experiment (that is, for each T) and held constant for the replications. The table contains the average estimate of the coefficient and, for the binary choice models, the partial effects. The coefficients for the probit and logit models with $T = 2$ correspond to the received result, a 100% bias. The remaining values show, as intuition would suggest, that the bias decreases with increasing T. The benchmark case of $T = 8$ appears to be less benign than Heckman's results suggested. One encouraging finding for the model builder is that the biases in the estimated marginal effects appears to be somewhat less than for the coefficients. Greene (2004b) extends this analysis to some other models, including the tobit and truncated regression

TABLE 15.5 Means of Empirical Sampling Distributions, $N = 1{,}000$ Individuals Based on 200 Replications

	Logit		*Probit*		*Ord. Probit*
Periods	*Coefficient*	*Partial Effect*[a]	*Coefficient*	*Partial Effect*[a]	*Coefficient*
$T = 2$					
β	2.020	1.676	2.083	1.474	2.328
δ	2.027	1.660	1.938	1.388	2.605
$T = 3$					
β	1.698	1.523	1.821	1.392	1.592
δ	1.668	1.477	1.777	1.354	1.806
$T = 5$					
β	1.379	1.319	1.589	1.406	1.305
δ	1.323	1.254	1.407	1.231	1.415
$T = 8$					
β	1.217	1.191	1.328	1.241	1.166
δ	1.156	1.128	1.243	1.152	1.220
$T = 10$					
β	1.161	1.140	1.247	1.190	1.131
δ	1.135	1.111	1.169	1.110	1.158
$T = 20$					
β	1.069	1.034	1.108	1.088	1.058
δ	1.062	1.052	1.068	1.047	1.068

[a] Average ratio of estimated partial effect to true partial effect.

models discussed in Chapter 19. The results there suggest that the conventional wisdom for the tobit model may not be correct—the incidental parameters (IP) problem seems to appear in the estimator of σ^2 in the tobit model, not in the estimators of the slopes.[14] This is consistent with the linear regression model, but not with the binary choice models.

15.6 SIMULATION-BASED ESTIMATION

Sections 15.3 through 15.5 developed a set of tools for inference about model parameters using simulation methods. This section will describe methods for using simulation as part of the estimation process. The modeling framework arises when integrals that cannot be computed directly appear in the estimation criterion function (sum of squares, log likelihood, and so on). To begin the development, in Section 15.6.1, we will construct a nonlinear model with random effects. Section 15.6.2 will describe how simulation is used to evaluate integrals for maximum likelihood estimation. Section 15.6.3 will develop an application, the random effects regression model.

[14] Research on the incidental parameters problem in discrete choice models, such as Fernandez-Val (2009), focuses on the slopes in the models. However, in all cases examined, the incidental parameters problem shows up as a proportional bias, which would seem to relate to an implicit scaling. The IP problem in the linear regression affects only the estimator of the disturbance variance.

15.6.1 RANDOM EFFECTS IN A NONLINEAR MODEL

In Example 11.20, we considered a nonlinear regression model for the number of doctor visits in the German Socioeconomic Panel. The basic form of the nonlinear regression model is

$$E[y_{it}|\mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}), t = 1, \ldots, T_i, i = 1, \ldots, n.$$

In order to accommodate unobserved heterogeneity in the panel data, we extended the model to include a random effect,

$$E[y_{it}|\mathbf{x}_{it}, u_i] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i), \tag{15-9}$$

where u_i is an unobserved random effect with zero mean and constant variance, possibly normally distributed—we will turn to that shortly. We will now go a step further and specify a particular probability distribution for y_{it}. Because doctor visits is a count, the Poisson regression model would be a natural choice,

$$p(y_{it}|\mathbf{x}_{it}, u_i) = \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i). \tag{15-10}$$

Conditioned on $\mathbf{x}_{it}$ and u_i, the T_i observations for individual i are independent. That is, by conditioning on u_i, we treat them as data, the same as $\mathbf{x}_{it}$. Thus, the T_i observations are independent when they are conditioned on $\mathbf{x}_{it}$ and u_i. The joint density for the T_i observations for individual i is the product,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i, u_i) = \prod_{t=1}^{T_i}\frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i), t = 1, \ldots, T_i. \tag{15-11}$$

In principle at this point, the log-likelihood function to be maximized would be

$$\ln L = \sum_{i=1}^{n}\ln\left[\prod_{t=1}^{T_i}\frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}\right], \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i). \tag{15-12}$$

But it is not possible to maximize this log likelihood because the unobserved $u_i, i = 1, \ldots, n$, appears in it. The joint distribution of $(y_{i1}, y_{i2}, \ldots, y_{i,Ti}, u_i)$ is equal to the marginal distribution of u_i times the conditional distribution of $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,Ti})$ given u_i,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}, u_i|\mathbf{X}_i) = p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i, u_i)f(u_i),$$

where $f(u_i)$ is the marginal density for u_i. Now, we can obtain the marginal distribution of $(y_{i1}, y_{i2}, \ldots, y_{i,Ti})$ without u_i by

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i) = \int_{u_i} p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i, u_i)f(u_i)du_i.$$

For the specific application, with the Poisson conditional distributions for $y_{it}|u_i$ and a normal distribution for the random effect,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i) = \int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i}\frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}\right]\frac{1}{\sigma}\phi\left(\frac{u_i}{\sigma}\right)du_i, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i).$$

The log-likelihood function will now be

$$\ln L = \sum_{i=1}^{n} \ln \left\{ \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!} \right] \frac{1}{\sigma} \phi\left(\frac{u_i}{\sigma}\right) du_i \right\}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i). \qquad \textbf{(15-13)}$$

The optimization problem is now free of the unobserved u_i, but that complication has been traded for another one, the integral that remains in the function.

To complete this part of the derivation, we will simplify the log-likelihood function slightly in a way that will make it fit more naturally into the derivations to follow. Make the change of variable $u_i = \sigma w_i$, where w_i has mean zero and standard deviation one. Then, the Jacobian is $du_i = \sigma dw_i$, and the limits of integration for w_i are the same as for u_i. Making the substitution and multiplying by the Jacobian, the log-likelihood function becomes

$$\ln L = \sum_{i=1}^{n} \ln \left\{ \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!} \right] \phi(w_i) dw_i \right\}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i). \qquad \textbf{(15-14)}$$

The log likelihood is then maximized over $(\boldsymbol{\beta}, \sigma)$. The purpose of the simplification is to parameterize the model so that the distribution of the variable that is being integrated out has no parameters of its own. Thus, in (15-14), w_i is normally distributed with mean zero and variance one.

In the next section, we will turn to how to compute the integrals. Section 14.14.4 analyzes this model and suggests the **Gauss–Hermite quadrature** method for computing the integrals. In this section, we will derive a method based on simulation, **Monte Carlo integration**.[15]

15.6.2 MONTE CARLO INTEGRATION

Integrals often appear in econometric estimators in *open form*, that is, in a form for which there is no specific closed form function that is equivalent to them. For example, the integral, $\int_0^t \theta \exp(-\theta w) dw = 1 - \exp(-\theta t)$, is in closed form. The integral in (15-14) is in open form. There are various devices available for approximating open form integrals—Gauss–Hermite and Gauss–Laguerre quadrature noted in Section 14.14.4 and in Appendix E2.4 are two. The technique of Monte Carlo integration can often be used when the integral is in the form

$$h(y) = \int_w g(y \mid w) f(w) dw = E_w[g(y \mid w)],$$

where $f(w)$ is the density of w and and w is a random variable that can be simulated.[16]

If $w_1, w_2, \ldots, w_n$ are a random sample of observations on the random variable w and $g(w)$ is a function of w with finite mean and variance, then by the law of large numbers [Theorem D.4 and the corollary in (D-5)],

[15] The term *Monte Carlo* is in reference to the casino at Monte Carlo, where random number generation is a crucial element of the business.

[16] There are some necessary conditions on w and $g(y \mid w)$ that will be met in the applications that interest us here. Some details appear in Cameron and Trivedi (2005) and Train (2009).

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n} g(w_i) = E[g(w)].$$

The function in (15-14) is in this form,

$$\int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}]}{y_{it}!}\right]\phi(w_i)dw_i$$
$$= E_{w_i}[g(y_{i1}, y_{i2}, \ldots, y_{iT_i}|w_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma)],$$

where

$$g(y_{i1}, y_{i2}, \ldots, y_{iT_i}|w_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma) = \prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}]}{y_{it}!}$$

and w_i is a random variable with standard normal distribution. It follows, then, that

$$\text{plim}\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}]}{y_{it}!}$$
$$= \int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}]}{y_{it}!}\right]\phi(w_i)dw_i. \qquad \textbf{(5-15)}$$

This suggests the strategy for computing the integral. We can use the methods developed in Section 15.2 to produce the necessary set of random draws on w_i from the standard normal distribution and then compute the approximation to the integral according to (15-15).

Example 15.8 Fractional Moments of the Truncated Normal Distribution

The following function appeared in Greene's (1990) study of the stochastic frontier model:

$$h(M, \varepsilon) = \frac{\int_0^{\infty} z^M \frac{1}{\sigma}\phi\left[\frac{z - (-\varepsilon - \theta\sigma^2)}{\sigma}\right]dz}{\int_0^{\infty} \frac{1}{\sigma}\phi\left[\frac{z - (-\varepsilon - \theta\sigma^2)}{\sigma}\right]dz}.$$

The integral only exists in closed form for integer values of M. However, the weighting function that appears in the integral is of the form

$$f(z|z > 0) = \frac{f(z)}{\text{Prob}[z > 0]} = \frac{\frac{1}{\sigma}\phi\left(\frac{z - \mu}{\sigma}\right)}{\int_0^{\infty}\frac{1}{\sigma}\phi\left(\frac{z - \mu}{\sigma}\right)dz}.$$

This is a truncated normal distribution. It is the distribution of a normally distributed variable z with mean μ and standard deviation σ, conditioned on z being greater than zero. The integral is equal to the expected value of z^M given that z is greater than zero when z is normally distributed with mean $\mu = -\varepsilon - \theta\sigma^2$ and variance σ^2.

The truncated normal distribution is examined in Section 19.2. The function $h(M, \varepsilon)$ is the expected value of z^M when z is the truncation of a normal random variable with mean μ and standard deviation σ. To evaluate the integral by Monte Carlo integration, we would require a sample $z_1, \ldots, z_R$ from this distribution. We have the results we need in (15-4) with $L = 0$, so $P_L = \Phi[0 - (-\varepsilon - \theta\sigma^2)/\sigma] = \Phi(\varepsilon/\sigma + \theta\sigma)$ and $U = +\infty$ so $P_U = 1$. Then, a draw on z is obtained by

$$z = \mu + \sigma\Phi^{-1}[P_L + F(1 - P_L)],$$

where F is the primitive draw from U[0, 1]. Finally, the integral is approximated by the simple average of the draws,

$$h(M, \varepsilon) \approx \frac{1}{R}\sum_{r=1}^{R} z[\varepsilon, \theta, \sigma, F_r]^M.$$

This is an application of Monte Carlo integration. In certain cases, an integral can be approximated by computing the sample average of a set of function values. The approach taken here was to interpret the integral as an expected value. Our basic statistical result for the behavior of sample means implies that, with a large enough sample, we can approximate the integral as closely as we like. The general approach is widely applicable in Bayesian econometrics and classical statistics and econometrics as well.[17]

15.6.2a HALTON SEQUENCES AND RANDOM DRAWS FOR SIMULATION-BASED INTEGRATION

Monte Carlo integration is used to evaluate the expectation

$$E[g(x)] = \int_x g(x)f(x)dx,$$

where $f(x)$ is the density of the random variable x and $g(x)$ is a smooth function. The Monte Carlo approximation is

$$E[g(x)] \approx \frac{1}{R}\sum_{r=1}^{R} g(x_r).$$

Convergence of the approximation to the expectation is based on the law of large numbers—a random sample of draws on $g(x)$ will converge in probability to its expectation. The standard approach to simulation-based integration is to use random draws from the specified distribution. Conventional simulation-based estimation uses a random number generator to produce the draws from a specified distribution. The central component of this approach is drawn from the standard continuous uniform distribution, $U[0,1]$. Draws from other distributions are obtained from these draws by using transformations. In particular, for a draw from the normal distribution, where u_i is one draw from $U[0, 1]$, $v_i = \Phi^{-1}(u_i)$. Given that the initial draws satisfy the necessary assumptions, the central issue for purposes of specifying the simulation is the number of draws. Good performance in this connection requires large numbers of draws. Results differ on the number needed in a given application, but the general finding is that when simulation is done in this fashion, the number is large (hundreds or thousands). A consequence of this is that for large-scale problems, the amount of computation time in simulation-based estimation can be extremely large. Numerous methods have been devised for reducing the numbers of draws needed to obtain a satisfactory approximation. One such method is to introduce some autocorrelation into the draws—a small amount of negative correlation across the draws will reduce the variance of the simulation. **Antithetic draws**, whereby each draw in a sequence is included with its mirror image (w_i and $-w_i$ for normally distributed draws, w_i and $1 - w_i$ for uniform, for example), is one such method.[18]

[17] See Geweke (1986, 1988, 1989, 2005) for discussion and applications. A number of other references are given in Poirier (1995, p. 654) and Koop (2003). See, as well, Train (2009).

[18] See Geweke (1988) and Train (2009, Chapter 9).

Procedures have been devised in the numerical analysis literature for taking intelligent draws from the uniform distribution, rather than random ones.[19] An emerging literature has documented dramatic speed gains with no degradation in simulation performance through the use of a smaller number of **Halton draws** or other constructed, nonrandom sequences instead of a large number of random draws. These procedures appear to vastly reduce the number of draws needed for estimation (sometimes by a factor of 90% or more) and reduce the simulation error associated with a given number of draws. In one application of the method to be discussed here, Bhat (1999) found that 100 Halton draws produced lower simulation error than 1,000 random numbers.

A Halton sequence is generated as follows: Let r be a prime number. Expand the sequence of integers $g = 1, 2, \ldots$ in terms of the base r as

$$g = \sum_{i=0}^{I} b_i r^i \text{ where, by construction, } 0 \leq b_i \leq r - 1 \text{ and } r^I \leq g < r^{I+1}.$$

The Halton sequence of values that corresponds to this series is

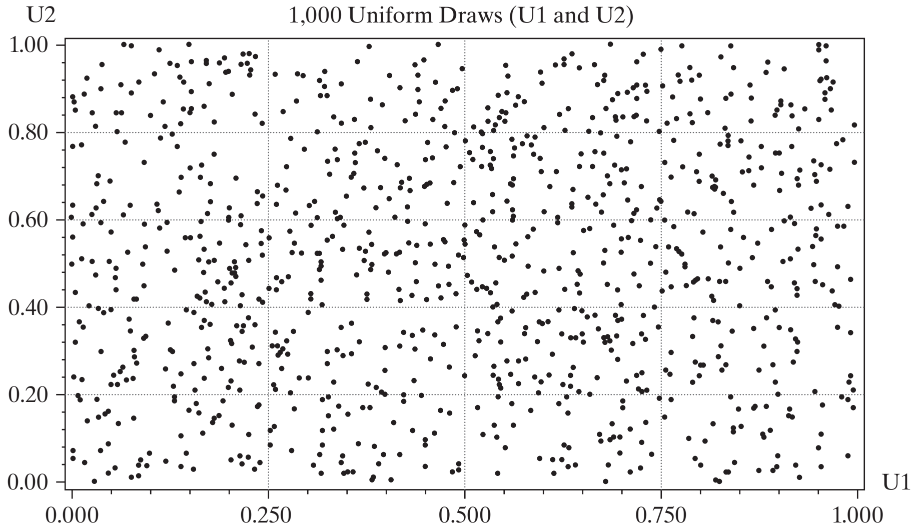
$$H(g) = \sum_{i=0}^{I} b_i r^{-i-1}.$$

For example, using base 5, the integer 37 has $b_0 = 2$, $b_1 = 2$, and $b_2 = 1$. Then

$$H_5(37) = 2 \times 5^{-1} + 2 \times 5^{-2} + 1 \times 5^{-3} = 0.488.$$

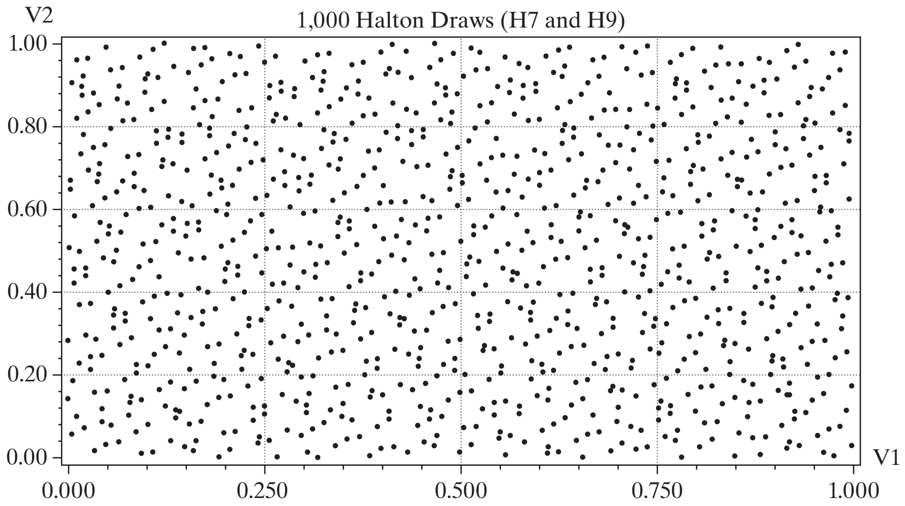
The sequence of Halton values is efficiently spread over the unit interval. The sequence is not random as the sequence of pseudo-random numbers is; it is a well-defined deterministic sequence. But randomness is not the key to obtaining accurate approximations to integrals. Uniform coverage of the support of the random variable is the central requirement. The large numbers of random draws are required to obtain smooth and dense coverage of the unit interval. Figures 15.3 and 15.4 show two sequences

FIGURE 15.3 Bivariate Distribution of Random Uniform Draws.



[19] See Train (1999, 2009) and Bhat (1999) for extensive discussion and further references.

FIGURE 15.4 Bivariate Distribution of Halton (7) and Halton (9).



of 1,000 Halton draws and two sequences of 1,000 pseudo-random draws. The Halton draws are based on $r = 7$ and $r = 9$. The clumping evident in the first figure is the feature (among others) that mandates large samples for simulations.

Example 15.9 Estimating the Lognormal Mean

We are interested in estimating the mean of a standard lognormally distributed variable. Formally, this result is

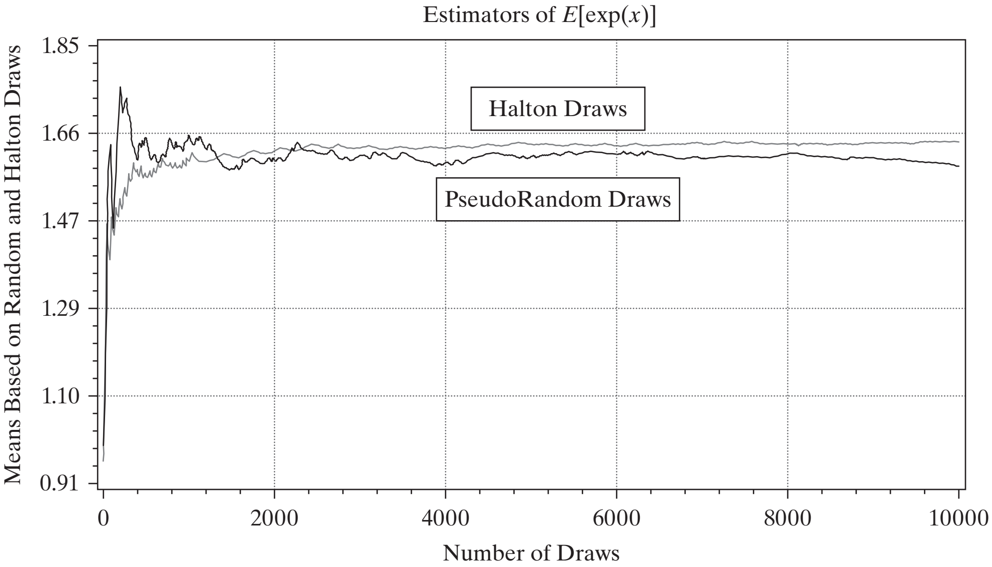
$$E[y] = \int_{-\infty}^{\infty} \exp(x) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] dx = 1.649.$$

To use simulation for the estimation, we will average n draws on $y = \exp(x)$ where x is drawn from the standard normal distribution. To examine the behavior of the Halton sequence as compared to that of a set of pseudo-random draws, we did the following experiment. Let $x_{i,t}$ = the sequence of values for a standard normally distributed variable. We draw $t = 1, \ldots, 10{,}000$ draws. For $i = 1$, we used a random number generator. For $i = 2$, we used the sequence of the first 10,000 Halton draws using $r = 7$. The Halton draws were converted to standard normal using the inverse normal transformation. To finish preparation of the data, we transformed $x_{i,t}$ to $y_{i,t} = \exp(x_{i,t})$. Then, for $n = 100, 110, \ldots, 10{,}000$, we averaged the first n observations in the sample. Figure 15.5 plots the evolution of the sample means as a function of the sample size. The lower trace is the sequence of Halton-based means. The greater stability of the Halton estimator is clearly evident in the figure.

15.6.2.b COMPUTING MULTIVARIATE NORMAL PROBABILITIES USING THE GHK SIMULATOR

The computation of bivariate normal probabilities is typically done using quadrature and requires a large amount of computing effort. Quadrature methods have been developed for trivariate probabilities as well, but the amount of computing effort needed at this level is enormous. For integrals of level greater than three, satisfactory (in terms of speed and accuracy) direct approximations remain to be developed. Our work thus far does

FIGURE 15.5 Estimates of $E[\exp(x)]$ Based on Random Draws and Halton Sequences, by Sample Size.



suggest an alternative approach. Suppose that $\mathbf{x}$ has a K-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. (No generality is sacrificed by the assumption of a zero mean, because we could just subtract a nonzero mean from the random vector wherever it appears in any result.) We wish to compute the K-variate probability, $\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_K < x_K < b_K]$. The Monte Carlo integration technique is well suited for this problem. As a first approach, consider sampling R observations, $\mathbf{x}_r$, $r = 1, \ldots, R$, from this multivariate normal distribution, using the method described in Section 15.2.4. Now, define

$$d_r = \mathbf{1}[a_1 < x_{r1} < b_1, a_2 < x_{r2} < b_2, \ldots, a_K < x_{rK} < b_K].$$

(That is, $d_r = 1$ if the condition is true and 0 otherwise.) Based on our earlier results, it follows that

$$\text{plim}\ \bar{d} = \text{plim}\ \frac{1}{R}\sum_{r=1}^{R} d_r = \text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_K < x_K < b_K].^{20}$$

This method is valid in principle, but in practice it has proved to be unsatisfactory for several reasons. For large-order problems, it requires an enormous number of draws from the distribution to give reasonable accuracy. Also, even with large numbers of draws, it appears to be problematic when the desired tail area is very small. Nonetheless, the idea is sound, and recent research has built on this idea to produce some quite accurate and efficient simulation methods for this computation. A survey of the methods is given in McFadden and Ruud (1994).[21]

[21] A symposium on the topic of simulation methods appears in *Review of Economic Statistics*, Vol. 76, November 1994. See, especially, McFadden and Ruud (1994), Stern (1994), Geweke, Keane, and Runkle (1994), and Breslaw (1994). See, as well, Gourieroux and Monfort (1996).

Among the simulation methods examined in the survey, the **GHK smooth recursive simulator** appears to be the most accurate.[22] The method is surprisingly simple. The general approach uses

$$\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_K < x_K < b_K] \approx \frac{1}{R}\sum_{r=1}^{R}\prod_{k=1}^{K} Q_{rk},$$

where Q_{rk} are easily computed univariate probabilities. The probabilities Q_{rk} are computed according to the following recursion: We first factor $\boldsymbol{\Sigma}$ using the **Cholesky factorization** $\boldsymbol{\Sigma} = \mathbf{CC}'$, where $\mathbf{C}$ is a lower triangular matrix (see Section A.6.11). The elements of $\mathbf{C}$ are l_{km}, where $l_{km} = 0$ if $m > k$. Then we begin the recursion with

$$Q_{r1} = \Phi(b_1/l_{11}) - \Phi(a_1/l_{11}).$$

Note that $l_{11} = \sigma_{11}$, so this is just the marginal probability, $\text{Prob}[a_1 < x_1 < b_1]$. Now, using (15-4), we generate a random observation ε_{r1} from the truncated standard normal distribution in the range

$$A_{r1} \text{ to } B_{r1} = a_1/l_{11} \text{ to } b_1/l_{11}.$$

(*Note:* The range is standardized because $l_{11} = \sigma_{11}$.) For steps $k = 2, \ldots, K$, compute

$$A_{rk} = \left[a_k - \sum_{m=1}^{k-1} l_{km}\varepsilon_{rm}\right]\Big/ l_{kk},$$

$$B_{rk} = \left[b_k - \sum_{m=1}^{k-1} l_{km}\varepsilon_{rm}\right]\Big/ l_{kk}.$$

Then,

$$Q_{rk} = \Phi(B_{rk}) - \Phi(A_{rk}).$$

Finally, in preparation for the next step in the recursion, we generate a random draw from the truncated standard normal distribution in the range A_{rk} to B_{rk}. This process is replicated R times, and the estimated probability is the sample average of the simulated probabilities.

The GHK simulator has been found to be impressively fast and accurate for fairly moderate numbers of replications. Its main usage has been in computing functions and derivatives for maximum likelihood estimation of models that involve multivariate normal integrals. We will revisit this in the context of the method of simulated moments when we examine the probit model in Chapter 17.

15.6.3 SIMULATION-BASED ESTIMATION OF RANDOM EFFECTS MODELS

In Section 15.6.2, (15-10), and (15-14), we developed a random effects specification for the Poisson regression model. For feasible estimation and inference, we replace the log-likelihood function,

$$\ln L = \sum_{i=1}^{n}\ln\left\{\int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}}{y_{it}!}\right]\phi(w_i)dw_i\right\},$$

with the simulated log-likelihood function,

[22] See Geweke (1989), Hajivassiliou (1990), and Keane (1994). Details on the properties of the simulator are given in Börsch-Supan and Hajivassiliou (1993).

$$\ln L_S = \sum_{i=1}^{n} \ln\left\{\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}}{y_{it}!}\right\}. \tag{15-16}$$

We now consider how to estimate the parameters via maximum simulated likelihood. In spite of its complexity, the simulated log likelihood will be treated in the same way that other log likelihoods were handled in Chapter 14. That is, we treat ln L_S as a function of the unknown parameters conditioned on the data, ln $L_S(\boldsymbol{\beta}, \sigma)$, and maximize the function using the methods described in Appendix E, such as the DFP or BFGS gradient methods. What is needed here to complete the derivation are expressions for the derivatives of the function. We note that the function is a sum of n terms; asymptotic results will be obtained in n; each observation can be viewed as one T_i-variate observation.

In order to develop a general set of results, it will be convenient to write each single density in the simulated function as

$$P_{itr}(\boldsymbol{\beta}, \sigma) = f(y_{it} | \mathbf{x}_{it}, w_{ir}, \boldsymbol{\beta}, \sigma) = P_{itr}(\boldsymbol{\theta}) = P_{itr}.$$

For our specific application in (15-16),

$$P_{itr} = \frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}]}{y_{it}!}.$$

The simulated log likelihod is, then,

$$\ln L_S = \sum_{i=1}^{n} \ln\left\{\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta})\right\}. \tag{15-17}$$

Continuing this shorthand, then, we will also define

$$P_{ir} = P_{ir}(\boldsymbol{\theta}) = \prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta}),$$

so that

$$\ln L_S = \sum_{i=1}^{n} \ln\left\{\frac{1}{R}\sum_{r=1}^{R} P_{ir}(\boldsymbol{\theta})\right\}.$$

And, finally,

$$P_i = P_i(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R} P_{ir},$$

so that

$$\ln L_S = \sum_{i=1}^{n} \ln P_i(\boldsymbol{\theta}). \tag{15-18}$$

With this general template, we will be able to accommodate richer specifications of the index function, now $\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i$, and other models such as the linear regression, binary choice models, and so on, simply by changing the specification of P_{itr}.

The algorithm will use the usual procedure,

$$\hat{\boldsymbol{\theta}}^{(k)} = \hat{\boldsymbol{\theta}}^{(k-1)} + \textit{update vector},$$

starting from an initial value, $\hat{\boldsymbol{\theta}}^{(0)}$, and will exit when the update vector is sufficiently small. A natural initial value would be from a model with no random effects; that is, the pooled estimator for the linear or Poisson or other model with $\sigma = 0$. Thus, at entry to the iteration (update), we will compute

$$\ln \hat{L}_S^{(k-1)} = \sum_{i=1}^{n} \ln\left\{ \frac{1}{R}\sum_{r=1}^{R} \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\sigma}^{(k-1)}w_{ir})][\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\sigma}^{(k-1)}w_{ir})]^{y_{it}}}{y_{it}!}\right\}.$$

To use a gradient method for the update, we will need the first derivatives of the function. Computation of an asymptotic covariance matrix may require the Hessian, so we will obtain this as well.

Before proceeding, we note two important aspects of the computation. First, a question remains about the number of draws, R, required for the maximum simulated likelihood estimator to be consistent. The approximated function,

$$\hat{E}_w[f(y\,|\,\mathbf{x}, w)] = \frac{1}{R}\sum_{r=1}^{R} f(y\,|\,\mathbf{x}, w_r),$$

is an unbiased estimator of $E_w[f(y\,|\,\mathbf{x}, w)]$. However, what appears in the simulated log likelihood is $\ln E_w[f(y\,|\,\mathbf{x}, w)]$, and the log of the estimator is a biased estimator of the log of its expectation. To maintain the asymptotic equivalence of the MSL estimator of $\boldsymbol{\theta}$ and the true MLE (if w were observed), it is necessary for the estimators of these terms in the log likelihood to converge to their expectations faster than the expectation of ln L converges to its expectation. The requirement is that $n^{1/2}/R \rightarrow 0$.[23] The estimator remains consistent if $n^{1/2}$ and R increase at the same rate; however, the asymptotic covariance matrix of the MSL estimator will then be larger than that of the true MLE. In practical terms, this suggests that the number of draws be on the order of $n^{.5+\delta}$ for some positive δ. [This does not state, however, what R should be for a given n; it only establishes the properties of the MSL estimator as n increases. For better or worse, researchers who have one sample of n observations often rely on the numerical stability of the estimator with respect to changes in R as their guide. Hajivassiliou (2000) gives some suggestions.] Note, as well, that the use of Halton sequences or any other autocorrelated sequences for the simulation, which is becoming more prevalent, interrupts this result. The appropriate counterpart to the Gourieroux and Monfort result for random sampling remains to be derived. One might suspect that the convergence result would persist, however. The usual standard is several hundred.

Second, it is essential that the same (pseudo- or Halton) draws be used every time the function or derivatives or any function involving these is computed for observation i. This can be achieved by creating the pool of draws for the entire sample before the optimization begins, and simply dipping into the same point in the pool each time a computation is required for observation i. Alternatively, if computer memory is an issue and the draws are re-created for each individual each time, the same practical result can be achieved by setting a preassigned seed for individual i, $seed(i) = s(i)$ for some simple monotonic function of i, and resetting the seed when draws for individual i are needed.

To obtain the derivatives, we begin with

$$\frac{\partial \ln L_S}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{(1/R)\sum_{r=1}^{R} \partial\left(\prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta})\right)/\partial \boldsymbol{\theta}}{(1/R)\sum_{r=1}^{R}\prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta})}. \qquad \textbf{(15-19)}$$

[23] See Gourieroux and Monfort (1996).

For the derivative term,

$$\begin{aligned}\partial\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} &= \left(\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})\right)\partial\left(\ln\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})\right)/\partial\boldsymbol{\theta}\\ &= \left(\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})\right)\sum_{t=1}^{T_i}\partial\ln P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\\ &= P_{ir}(\boldsymbol{\theta})\left(\sum_{t=1}^{T_i}\partial\ln P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right) = P_{ir}(\boldsymbol{\theta})\sum_{t=1}^{T_i}\mathbf{g}_{itr}(\boldsymbol{\theta})\\ &= P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta}).\end{aligned} \tag{15-20}$$

Now, insert the result of (15-20) in (15-19) to obtain

$$\frac{\partial\ln L_S(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n}\frac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})}. \tag{15-21}$$

Define the weight $Q_{ir}(\boldsymbol{\theta}) = P_{ir}(\boldsymbol{\theta})/\Sigma_{r=1}^{R}P_{ir}(\boldsymbol{\theta})$ so that $0 < Q_{ir}(\boldsymbol{\theta}) < 1$ and $\Sigma_{r=1}^{R}Q_{ir}(\boldsymbol{\theta}) = 1$. Then,

$$\frac{\partial\ln L_S(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n}\sum_{r=1}^{R}Q_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\overline{\mathbf{g}}_i(\boldsymbol{\theta}). \tag{15-22}$$

To obtain the second derivatives, define $\mathbf{H}_{itr}(\boldsymbol{\theta}) = \partial^2\ln P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ and let

$$\mathbf{H}_{ir}(\boldsymbol{\theta}) = \sum_{t=1}^{T_i}\mathbf{H}_{itr}(\boldsymbol{\theta})$$

and

$$\overline{\mathbf{H}}_i(\boldsymbol{\theta}) = \sum_{r=1}^{R}Q_{ir}(\boldsymbol{\theta})\mathbf{H}_{ir}(\boldsymbol{\theta}). \tag{15-23}$$

Then, working from (15-21), the second derivatives matrix breaks into three parts as follows:

$$\frac{\partial^2\ln L_s(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{i=1}^{n}\left[\begin{array}{l}\dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{H}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})} + \\ \dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})'}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})} - \left[\dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})}\right]\left[\dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})}\right]'\end{array}\right].$$

We can now use (15-20) through (15-23) to combine these terms;

$$\frac{\partial^2\ln L_S}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{i=1}^{n}\left\{\overline{\mathbf{H}}_i(\boldsymbol{\theta}) + \sum_{r=1}^{R}Q_{ir}(\boldsymbol{\theta})[\mathbf{g}_{ir}(\boldsymbol{\theta}) - \overline{\mathbf{g}}_i(\boldsymbol{\theta})][\mathbf{g}_{ir}(\boldsymbol{\theta}) - \overline{\mathbf{g}}_i(\boldsymbol{\theta})]'\right\}. \tag{15-24}$$

An estimator of the asymptotic covariance matrix for the MSLE can be obtained by computing the negative inverse of this matrix.

Example 15.10 Poisson Regression Model with Random Effects

For the Poisson regression model, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma)'$ and

$$\begin{aligned}
P_{itr}(\boldsymbol{\theta}) &= \frac{\exp[-\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}]}{y_{it}!} = \frac{\exp[-\mu_{itr}(\theta)]\mu_{itr}(\theta)^{y_{it}}}{y_{it}!} \\
\mathbf{g}_{itr}(\boldsymbol{\theta}) &= [y_{it} - \mu_{itr}(\boldsymbol{\theta})]\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix} \qquad \textbf{(15-25)} \\
\mathbf{H}_{itr}(\boldsymbol{\theta}) &= -\mu_{itr}(\boldsymbol{\theta})\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix}'.
\end{aligned}$$

Estimates of the random effects model parameters would be obtained by using these expressions in the preceding general template. We will apply these results in an application in Chapter 19 where the Poisson regression model is developed in greater detail.

Example 15.11 Maximum Simulated Likelihood Estimation of the Random Effects Linear Regression Model

The preceding method can also be used to estimate a linear regression model with random effects. We have already seen two ways to estimate this model, using two-step FGLS in Section 11.5.3 and by (closed form) maximum likelihood in Section 14.9.6.a. It might seem redundant to construct yet a third estimator for the model. However, this third approach will be the only feasible method when we generalize the model to have other random parameters in the next section. To use the simulation estimator, we define $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_\varepsilon)$. We will require

$$\begin{aligned}
P_{itr}(\boldsymbol{\theta}) &= \frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\left[-\frac{(y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta} - \sigma_u w_{ir})^2}{2\sigma_\varepsilon^2}\right], \\
\mathbf{g}_{itr}(\boldsymbol{\theta}) &= \begin{bmatrix} \left(\dfrac{(y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta} - \sigma_u w_{ir}}{\sigma_\varepsilon^2}\right)\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix} \\ \dfrac{(y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta} - \sigma_u w_{ir})^2}{\sigma_\varepsilon^3} - \dfrac{1}{\sigma_\varepsilon} \end{bmatrix} = \begin{bmatrix} (\varepsilon_{itr}/\sigma_\varepsilon^2)\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix} \\ (1/\sigma_\varepsilon)[(\varepsilon_{itr}^2/\sigma_\varepsilon^2) - 1] \end{bmatrix}, \qquad \textbf{(15-26)} \\
\mathbf{H}_{itr}(\boldsymbol{\theta}) &= \begin{bmatrix} -(1/\sigma_\varepsilon^2)\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix}' & -(2\varepsilon_{itr}/\sigma_\varepsilon^3)\begin{pmatrix}\mathbf{x}_{it} \\ w_{ir}\end{pmatrix} \\ -(2\varepsilon_{itr}/\sigma_\varepsilon^3)(\mathbf{x}'_{it}w_{ir}) & -(3\varepsilon_{itr}^2/\sigma_\varepsilon^4) + (1/\sigma_\varepsilon^2) \end{bmatrix}.
\end{aligned}$$

Note in the computation of the disturbance variance, σ_ε^2, we are using the sum of squared simulated residuals. However, the estimator of the variance of the heterogeneity, σ_u, is not being computed as a mean square. It is essentially the regression coefficient on w_{ir}. One surprising implication is that the actual estimate of σ_u can be negative. This is the same result that we have encountered in other situations. In no case is there a natural estimator of σ_u^2 that is based on a sum of squares. However, in this context, there is yet another surprising aspect of this calculation. In the simulated log-likelihood function, if every w_{ir} for every individual were changed to $-w_{ir}$ and σ_u is changed to $-\sigma_u$, then the exact same value of the function and all derivatives results. The implication is that the sign of σ_u is not identified in this setting. With no loss of generality, it is normalized to positive (+) to be consistent with the underlying theory that it is a standard deviation.

15.7 A RANDOM PARAMETERS LINEAR REGRESSION MODEL

We will slightly reinterpret the random effects model as

$$\begin{aligned} y_{it} &= \beta_{0i} + \mathbf{x}'_{it1}\boldsymbol{\beta}_1 + \varepsilon_{it}, \\ \beta_{0i} &= \beta_0 + u_i. \end{aligned} \tag{15-27}$$

This is equivalent to the random effects model, though in (15-27), we reinterpret it as a regression model with a randomly distributed constant term. In Section 11.10.1, we built a linear regression model that provided for parameter heterogeneity across individuals,

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta}_i + \varepsilon_{it}, \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \mathbf{u}_i, \end{aligned} \tag{15-28}$$

where $\mathbf{u}_i$ has mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma}$. In that development, we took a fixed effects approach in that no restriction was placed on the covariance between $\mathbf{u}_i$ and $\mathbf{x}_{it}$. Consistent with these assumptions, we constructed an estimator that involved n regressions of $\mathbf{y}_i$ on $\mathbf{X}_i$ to estimate $\boldsymbol{\beta}$ one unit at a time. Each estimator is consistent in T_i. (This is precisely the approach taken in the fixed effects model, where there are n unit specific constants and a common $\boldsymbol{\beta}$. The approach there is to estimate $\boldsymbol{\beta}$ first and then to regress $\mathbf{y}_i - \mathbf{X}_i, \mathbf{b}_{\text{LSDV}}$ on $\mathbf{d}_i$ to estimate α_i.) In the same way that assuming that u_i is uncorrelated with $\mathbf{x}_{it}$ in the fixed effects model provided a way to use FGLS to estimate the parameters of the random effects model, if we assume in (15-28) that $\mathbf{u}_i$ is uncorrelated with $\mathbf{X}_i$, we can extend the random effects model in Section 15.6.3 to a model in which some or all of the other coefficients in the regression model, not just the constant term, are randomly distributed. The theoretical proposition is that the model is now extended to allow individual heterogeneity in all coefficients.

To implement the extended model, we will begin with a simple formulation in which $\mathbf{u}_i$ has a diagonal covariance matrix—this specification is quite common in the literature. The implication is that the random parameters are uncorrelated; $\beta_{i,k}$ has mean β_k and variance γ_k^2. The model in (15-26) can modified to allow this case with a few minor changes in notation. Write

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_i, \tag{15-29}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the standard deviations $(\gamma_1, \gamma_2, \ldots, \gamma_K)$ of $(u_{i1}, \ldots, u_{iK})$ on the diagonal and $\mathbf{w}_i$ is now a random vector with zero means and unit standard deviations. Then, $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. The parameter vector in the model is now

$$\boldsymbol{\theta} = (\beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_K, \sigma_\varepsilon)'.$$

(In an application, some of the γ's might be fixed at zero to make the corresponding parameters nonrandom.) In order to extend the model, the disturbance in (15-26), $\varepsilon_{itr} = (y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - \sigma_u w_{ir})$, becomes

$$\varepsilon_{itr} = y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{ir}). \tag{15-30}$$

Now, combine (15-17) and (15-29) with (15-30) to produce

$$\ln L_S = \sum_{i=1}^{n} \ln\left\{\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\left[\frac{(y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{ir}))^2}{2\sigma_\varepsilon^2}\right]\right\}. \tag{15-31}$$

In the derivatives in (15-26), the only change needed to accommodate this extended model is that the scalar w_{ir} becomes the vector $(w_{ir,1}x_{it1}, w_{ir,2}x_{it,2}, \ldots, w_{ir,K}x_{it,K})$. This is the element-by-element product of the regressors, $\mathbf{x}_{it}$, and the vector of random draws, $\mathbf{w}_{ir}$, which is the **Hadamard product**, **direct product**, or **Schur product** of the two vectors, usually denoted $\mathbf{x}_{it} \circ \mathbf{w}_{ir}$.

Although only a minor change in notation in the random effects template in (15-26), this formulation brings a substantial change in the formulation of the model. The integral in ln L is now a K dimensional integral. Maximum simulated likelihood estimation proceeds as before, with potentially much more computation as each draw now requires a K-variate vector of pseudo-random draws.

The random parameters model can now be extended to one with a full covariance matrix, $\mathbf{\Gamma}$ as we did with the fixed effects case. We will now let $\mathbf{\Gamma}$ in (15-29) be the Cholesky factorization of $\mathbf{\Gamma}$, so $\mathbf{\Gamma} = \mathbf{\Lambda}\mathbf{\Lambda}'$. (This was already the case for the simpler model with diagonal $\mathbf{\Gamma}$.) The implementation in (15-26) will be a bit complicated. The derivatives with respect to $\boldsymbol{\beta}$ are unchanged. For the derivatives with respect to $\mathbf{\Lambda}$, it is useful to assume for the moment that $\mathbf{\Lambda}$ is a full matrix, not a lower triangular one. Then, the scalar w_{ir} in the derivative expression becomes a $K^2 \times 1$ vector in which the $(k - 1) \times K + l^{\text{th}}$ element is $x_{it,k} \times w_{ir,l}$. The full set of these is the **Kronecker product** of $\mathbf{x}_{it}$ and $\mathbf{w}_{ir}$, $\mathbf{x}_{it} \otimes \mathbf{w}_{ir}$. The necessary elements for maximization of the log-likelihood function are then obtained by discarding the elements for which $\mathbf{\Lambda}_{kl}$ are known to be zero—these correspond to $l > k$.

In (15-26), for the full model, for computing the MSL estimators, the derivatives with respect to $(\boldsymbol{\beta}, \mathbf{\Lambda})$. are equated to zero. The result after some manipulation is

$$\frac{\partial \ln L_S}{\partial(\boldsymbol{\beta}, \mathbf{\Lambda})} = \sum_{i=1}^{n}\frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T_i}\frac{(y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{w}_{ir}))}{\sigma_\varepsilon^2}\begin{bmatrix}\mathbf{x}_{it}\\ \mathbf{x}_{it} \otimes \mathbf{w}_{ir}\end{bmatrix} = \mathbf{0}.$$

By multiplying this by σ_ε^2, we find, as usual, that σ_ε^2 is not needed for computation of the estimates of $(\boldsymbol{\beta}, \mathbf{\Lambda})$. Thus, we can view the solution as the counterpart to least squares, which might call, instead, the least simulated sum of squares estimator. Once the simulated sum of squares is minimized with respect to $\boldsymbol{\beta}$ and $\mathbf{\Lambda}$, then the solution for σ_ε^2 can be obtained via the likelihood equation,

$$\frac{\partial \ln L_S}{\partial \sigma_\varepsilon^2} = \sum_{i=1}^{n}\left\{\frac{1}{R}\sum_{r=1}^{R}\left[\frac{-T_i}{2\sigma_\varepsilon^2} + \frac{\sum_{t=1}^{T_i}(y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{w}_{i,r}))^2}{2\sigma_\varepsilon^4}\right]\right\} = 0.$$

Multiply both sides of this equation by $-2\sigma_\varepsilon^4$ to obtain the equivalent condition

$$\frac{\partial \ln L_S}{\partial \sigma_\varepsilon^2} = \sum_{i=1}^{n}\left\{\frac{1}{R}\sum_{r=1}^{R}T_i\left[-\sigma_\varepsilon^2 + \frac{\sum_{t=1}^{T_i}(y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{w}_{i,r}))^2}{T_i}\right]\right\} = 0.$$

By expanding this expression and manipulating it a bit, we find the solution for σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{n}F_i\frac{1}{R}\sum_{r=1}^{R}\hat{\sigma}_{\varepsilon,\,ir}^2, \text{ where } \hat{\sigma}_{\varepsilon,\,ir}^2 = \frac{\sum_{t=1}^{T_i}(y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{w}_{i,r}))^2}{T_i}$$

and $F_i = T_i/\Sigma_i T_i$ is a weight for each group that equals $1/n$ if T_i is the same for all i.

Example 15.12 Random Parameters Wage Equation

Estimates of the random effects log wage equation from the Cornwell and Rupert study in Examples 11.7 and 15.6 are shown in Table 15.6. The table presents estimates based on several assumptions. The encompassing model is

$$\ln Wage_{it} = \beta_{1,i} + \beta_{2,i}Wks_{i,t} + \cdots + \beta_{12,i}Fem_i + \beta_{13,i}Blk_i + \varepsilon_{it}, \tag{15-32}$$

$$\beta_{k,i} = \beta_k + \lambda_k w_{ik},\ w_{ik} \sim N[0, 1],\ k = 1, \ldots, 13. \tag{15-33}$$

TABLE 15.6 Estimated Wage Equations (Standard errors in parentheses)

Variable	*Pooled OLS*	*Feasible Two-Step GLS*	*Maximum Likelihood*	*Maximum Simulated Likelihood*[a]	*Random Parameters Max. Simulated Likelihood*[a]	
					β	λ
Wks	0.00422	0.00096	0.00084	0.00086	−0.00029	0.00614
	(0.00108)	(0.00059)	(0.00060)	(0.00047)	(0.00082)	(0.00042)
South	−0.05564	−0.00825	0.00577	0.00935	0.04941	0.20997
	(0.01253)	(0.02246)	(0.03159)	(0.00508)	(0.02002)	(0.01702)
SMSA	0.15167	−0.02840	−0.04748	−0.04913	−0.05486	0.01165
	(0.01207)	(0.01616)	(0.01896)	(0.00507)	(0.01747)	(0.02738)
MS	0.04845	−0.07090	−0.04138	−0.04142	−0.06358	0.02524
	(0.02057)	(0.01793)	(0.01899)	(0.00824)	(0.01896)	(0.03190)
Exp	0.04010	0.08748	0.10721	0.10668	0.09291	0.01803
	(0.00216)	(0.00225)	(0.00248)	(0.00096)	(0.00216)	(0.00092)
Exp^2	−0.00067	−0.00076	−0.00051	−0.00050	−0.00019	0.00008
	(0.00005)	(0.00005)	(0.00005)	(0.00002)	(0.00007	(0.00002)
Occ	−0.14001	−0.04322	−0.02512	−0.02437	−0.00963	0.02565
	(0.01466)	(0.01299)	(0.01378)	(0.00593)	(0.01331)	(0.01019)
Ind	0.04679	0.00378	0.01380	0.01610	0.00207	0.02575
	(0.01179)	(0.01373)	(0.01529)	(0.00490)	(0.01357)	(0.02420)
Union	0.09263	0.05835	0.03873	0.03724	0.05749	0.15260
	(0.01280)	(0.01350)	(0.01481)	(0.00509)	(0.01469)	(0.02022)
Ed	0.05670	0.10707	0.13562	0.13952	0.09356	0.00409
	(0.00261)	(0.00511)	(0.01267)	(0.01170)	(0.00359)	(0.00160)
Fem	−0.36779	−0.30938	−0.17562	−0.11694	−0.03864	0.28310
	(0.02510)	(0.04554)	(0.11310)	(0.01060)	(0.02467)	(0.00760)
Blk	−0.16694	−0.21950	−0.26121	−0.15184	−0.26864	0.02930
	(0.02204)	(0.05252)	(0.13747)	(0.00979)	(0.03156)	(0.03841)
Constant	5.25112	4.04144	3.12622	3.08362	3.81680	0.26347
	(0.07129)	(0.08330)	(0.17761)	(0.03276)	(0.06905)	(0.01628)
σ_u	0.00000	0.31453	0.83932	0.80926		
σ_ε	0.34936	0.15206	0.15334	0.15326	0.14354	
	LM = 3497.02				(0.00208)	
Ln L	−1523.254		307.873	309.173	365.313	

[a] Based on 500 Halton draws.

Under the assumption of homogeneity, that is, $\lambda_k = 0$, the pooled OLS estimator is consistent and efficient. As we saw in Chapter 11, under the random effects assumption, that is $\lambda_k = 0$ for $k = 2, \ldots, 13$ but $\lambda_1 \neq 0$, the OLS estimator is consistent, as are the next three estimators that explicitly account for the heterogeneity. To consider the full specification, write the model in the equivalent form

$$\begin{aligned} \ln Wage_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \left(\lambda_1 w_{i,1} + \sum_{k=2}^{13} \lambda_k w_{i,k} x_{it,k}\right) + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + W_{it} + \varepsilon_{it}. \end{aligned}$$

This is still a regression: $E[W_{it} + \varepsilon_{it} | \mathbf{X}] = 0$. (For the product terms, $E[\lambda_k w_{i,k} x_{it,k} | \mathbf{X}] = \lambda_k x_{it,k} E[w_{i,k} | x_{itk}] = 0$.) Therefore, even OLS remains consistent. The heterogeneity induces heteroscedasticity in W_{it} so the OLS estimator is inefficient and the conventional covariance matrix will be inappropriate. The random effects estimators of $\boldsymbol{\beta}$ in the center three columns of Table 15.6 are also consistent, by a similar logic. However, they likewise are inefficient. The result at work, which is specific to the linear regression model, is that we are estimating the mean parameters, β_k, and the variance parameters, λ_k and σ_ε, separately. Certainly, if λ_k is nonzero for $k = 2, \ldots, 13$, then the pooled and RE estimators that assume they are zero are all inconsistent. With $\boldsymbol{\beta}$ estimated consistently in an otherwise misspecified model, we would call the MLE and MSLE **pseudo-maximum likelihood estimators**.See Section 14.8.

Comparing the ML and MSL estimators of the random effects model, we find the estimates are similar, though in a few cases, noticeably different nonetheless. The estimates tend to differ most when the estimates themselves have large standard errors (small *t* ratios). This is partly due to the different methods of estimation in a finite sample of 595 observations. We could attribute at least some of the difference to the approximation error in the simulation compared to the exact evaluation of the (closed form) integral in the MLE. The full random parameters model is shown in the last two columns. Based on the likelihood ratio statistic of $2(365.312 - 309.173) = 112.28$ with 12 degrees of freedom, we would reject the hypothesis that $\lambda_2 = \lambda_3 = \cdots = \lambda_{13} = 0$. The 95% critical value with 12 degrees of freedom is 21.03. This random parameters formulation of the model suggests a need to reconsider the notion of *statistical significance* of the estimated parameters. In view of (15-33), it may be the case that the mean parameter might well be significantly different from zero while the corresponding standard deviation, λ, might be large as well, suggesting that a large proportion of the population remains statistically close to zero. Consider the estimate of $\beta_{3,\mathrm{i}}$, the coefficient on $South_{it}$. The estimate of the mean, β_3, is 0.04941, with an estimated standard error of 0.02002. This implies a confidence interval for this parameter of $0.04941 \pm 1.96(0.02002) = [0.01017, 0.08865]$. But this is only the location of the center of the distribution. With an estimate of λ_k of 0.20997, the random parameters model suggests that in the population, 95% of individuals have an effect of *South* within $0.04941 \pm 1.96(0.20997) = [-0.36213, 0.46095]$. This is still centered near zero but has a different interpretation from the simple confidence interval for β itself. Most of the population is less than two standard deviations from zero. This analysis suggests that it might be an interesting exercise to estimate β_i rather than just the parameters of the distribution. We will consider that estimation problem in Section 15.10.

The next example examines a random parameters model in which the covariance matrix of the random parameters is allowed to be a free, positive definite matrix. That is,

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it} \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \mathbf{u}_i, E[\mathbf{u}_i | \mathbf{X}] = \mathbf{0}, \mathrm{Var}[\mathbf{u}_i | \mathbf{X}] = \boldsymbol{\Gamma}. \end{aligned} \qquad \textbf{(15-34)}$$

This is the random effects counterpart to the fixed effects model in Section 11.10.1. Note that the difference in the specifications is the random effects assumption, $E[\mathbf{u}_i|\mathbf{X}] = \mathbf{0}$. We continue to use the Cholesky decomposition of $\mathbf{\Gamma}$ in the reparameterized model

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{w}_i, E[\mathbf{w}_i|\mathbf{X}] = \mathbf{0}, \text{Var}[\mathbf{w}_i|\mathbf{X}] = \mathbf{I}.$$

Example 15.13 Least Simulated Sum of Squares Estimates of a Production Function Model

In Example 11.22, we examined Munnell's production model for gross state product,

$$\ln gsp_{it} = \beta_1 + \beta_2 \ln pc_{it} + \beta_3 \ln hwy_{it} + \beta_4 \ln water_{it}$$
$$+ \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7\, unemp_{it} + \varepsilon_{it}, i = 1, \dots, 48; t = 1, \dots, 17.$$

The panel consists of state-level data for 17 years. The model in Example 11.19 (and Munnell's) provides no means for parameter heterogeneity save for the constant term. We have reestimated the model using the Hildreth and Houck approach. The OLS, feasible GLS, and maximum likelihood estimates are given in Table 15.7. (The OLS and FGLS results are reproduced from Table 11.21.) The chi-squared statistic for testing the null hypothesis of parameter homogeneity is 25,556.26, with 7(47) = 329 degrees of freedom. The critical value from the table is 372.299, so the hypothesis would be rejected. Unlike the other cases we have examined in this chapter, the FGLS estimates are very different from OLS in these estimates. The FGLS estimates correspond to a fixed effects view, as they do not assume that the variation in the coefficients is unrelated to the exogenous variables. The underlying standard deviations are computed using **G** as the covariance matrix. [For these

TABLE 15.7 Estimated Random Coefficients Models

	Least Squares		*Feasible GLS*		*Maximum Simulated Likelihood*	
Variable	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*
Constant	1.9260	0.05250	1.6533	1.08331	2.02319 (0.53228)	0.03801
ln pc	0.3120	0.01109	0.09409	0.05152	0.32049 (0.15871)	0.00621
ln hwy	0.05888	0.01541	0.1050	0.1736	0.01215 (0.19212)	0.00909
ln water	0.1186	0.01236	0.07672	0.06743	0.07612 (0.17484)	0.00600
ln util	0.00856	0.01235	−0.01489	0.09886	−0.04665 (0.78196)	0.00850
ln emp	0.5497	0.01554	0.9190	0.1044	0.67568 (0.82133)	0.00984
unemp	−0.00727	0.001384	−0.004706	0.002067	−0.00791 (0.02171)	0.00093
σ_ε	0.08542			0.2129	0.02360	
ln L	853.1372				1527.196	

data, subtracting the second matrix rendered **G** not positive definite so, in the table, the standard deviations are based on the estimates using only the first term in (11-88).] The increase in the standard errors is striking. This suggests that there is considerable variation in the parameters across states. We have used (11-89) to compute the estimates of the state-specific coefficients.

The rightmost two columns of Table 15.7 present the maximum simulated likelihood estimates of the random parameters production function model. They somewhat resemble the OLS estimates, more so than the FGLS estimates, which are computed by an entirely different method. The values in parentheses under the parameter estimates are the estimates of the standard deviations of the distribution of $\mathbf{u}_i$, the square roots of the diagonal elements of Γ. These are obtained by computing the square roots of the diagonal elements of $\mathbf{\Lambda\Lambda}'$. The estimate of $\mathbf{\Lambda}$ is shown here.

$$\hat{\mathbf{\Lambda}} = \begin{bmatrix} 0.53228 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.12511 & 0.09766 & 0 & 0 & 0 & 0 & 0 \\ 0.17529 & -0.07196 & 0.03169 & 0 & 0 & 0 & 0 \\ 0.03467 & 0.03306 & 0.15498 & 0.06522 & 0 & 0 & 0 \\ 0.16413 & -0.03030 & -0.08889 & 0.59745 & 0.46772 & 0 & 0 \\ 0.14750 & -0.02049 & 0.05248 & 0.67429 & 0.44158 & 0.00167 & 0 \\ 0.00427 & -0.00337 & 0.00181 & 0.01640 & 0.01277 & 0.00239 & 0.00083 \end{bmatrix}.$$

An estimate of the correlation matrix for the parameters might also be informative. This is also derived from $\hat{\mathbf{\Lambda}}$ by computing $\hat{\mathbf{\Gamma}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$ and then transforming the covariances to correlations by dividing by the products of the respective standard deviations (the values in parentheses in Table 15.7). The result is

$$\mathbf{R} = \begin{bmatrix} 1 & & & & & & \\ -0.7883 & 1 & & & & & \\ 0.9124 & -0.9497 & 1 & & & & \\ 0.1983 & -0.0400 & 0.2563 & 1 & & & \\ 0.2099 & -0.1893 & 0.1873 & 0.2186 & 1 & & \\ 0.1796 & -0.1569 & 0.1837 & 0.3938 & 0.9802 & 1 & \\ 0.1966 & -0.2504 & 0.2512 & 0.3654 & 0.9669 & 0.9812 & 1 \end{bmatrix}.$$

15.8 HIERARCHICAL LINEAR MODELS

Example 11.23 examined an application of a two-level model, or hierarchical model, for mortgage rates,

$$RM_{it} = \beta_{1i} + \beta_{2,i}J_{it} + \text{ various terms relating to the mortgage } + \varepsilon_{it}.$$

The second-level equation is

$$\begin{aligned} \beta_{2,i} = \alpha_1 + \alpha_2\text{GFA}_i + \alpha_3 \text{ one-year treasury rate } + \alpha_4 \text{ ten-year treasury rate} \\ + \alpha_5 \text{ credit risk } + \alpha_6 \text{ prepayment risk } + \cdots + u_i. \end{aligned}$$

Recent research in many fields has extended the idea of hierarchical modeling to the full set of parameters in the model. (Depending on the field studied, the reader may find these labeled *hierarchical models*, **mixed models**, *random parameters models,* or *random effects models*. The last of these generalizes our notion of random effects.) A two-level formulation of the model in (15-34) might appear as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it},$$
$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{u}_i.$$

(A three-level model is shown in Example 15.14.) This model retains the earlier stochastic specification but adds the measurement equation to the generation of the random parameters. model of the previous section now becomes

$$y_{it} = \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i) + \varepsilon_{it},$$

which is essentially the same as our earlier model in (15-28) to (15-31) with the addition of the product (interaction) terms of the form $\delta_{kl}x_{itk}z_{il}$, which suggests how it might be estimated (simply by adding the interaction terms to the previous formulation). In the template in (15-26), the term $\sigma_u w_{ir}$ becomes $\mathbf{x}_{it}'(\boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i)$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\delta}', \boldsymbol{\lambda}', \sigma_\varepsilon)'$, where $\boldsymbol{\delta}'$ is a row vector composed of the rows of $\boldsymbol{\Delta}$, and $\boldsymbol{\lambda}'$ is a row vector composed of the rows of $\boldsymbol{\Lambda}$. The scalar term w_{ir} in the derivatives is replaced by a column vector of terms contained in $(\mathbf{x}_{it} \otimes \mathbf{z}_i, \mathbf{x}_{it} \otimes \mathbf{w}_{ir})$.

The hierarchical model can be extended in several useful directions. Recent analyses have expanded the model to accommodate multilevel stratification in data sets such as those we considered in the treatment of nested random effects in Section 14.9.6.b. A three-level model would appear as in the next example that relates to home sales,

$$\begin{aligned} y_{ijt} &= \mathbf{x}_{ijt}'\boldsymbol{\beta}_{ij} + \varepsilon_{it}, t = \textit{site}, j = \textit{neighborhood}, i = \textit{community}, \\ \boldsymbol{\beta}_{ij} &= \boldsymbol{\phi}_i + \boldsymbol{\Delta}\mathbf{z}_{ij} + \mathbf{u}_{ij} \\ \boldsymbol{\phi}_i &= \boldsymbol{\pi} + \boldsymbol{\Phi}\mathbf{r}_i + \mathbf{v}_i. \end{aligned} \tag{15-35}$$

Example 15.14 Hierarchical Linear Model of Home Prices

Beron, Murdoch, and Thayer (1999) used a hedonic pricing model to analyze the sale prices of 76,343 homes in four California counties: Los Angeles, San Bernardino, Riverside, and Orange. The data set is stratified into 2,185 census tracts and 131 school districts. Home prices are modeled using a three-level random parameters pricing model. (We will change their notation somewhat to make roles of the components of the model more obvious.) Let *site* denote the specific location (sale), *nei* denote the neighborhood, and *com* denote the community, the highest level of aggregation. The pricing equation is

$$\ln \textit{Price}_{site,\, nei,\, com} = \pi^0_{nei,\, com} + \sum_{k=1}^{K} \pi^k_{nei,\, com} x_{k,site,nei,com} + \varepsilon_{site,nei,com},$$
$$\pi^k_{nei,\, com} = \beta^{0,k}_{com} + \sum_{l=1}^{L} \beta^{l,k}_{com} z_{k,nei,\, com} + r^k_{nei,com},\ k = 0, \ldots, K,$$
$$\beta^{l,k}_{com} = \gamma^{0,l,k} + \sum_{m=1}^{M} \gamma^{m,l,k} e_{m,com} + u^{l,k}_{com},\ l = 1, \ldots, L.$$

There are K level-one variables, x_k, and a constant in the main equation, L level-two variables, z_l, and a constant in the second-level equations, and M level-three variables, e_m, and a constant in the third-level equations. The variables in the model are as follows. The level-one variables define the hedonic pricing model,

x = house size, number of bathrooms, lot size, presence of central heating, presence of air conditioning, presence of a pool, quality of the view, age of the house, distance to the nearest beach.

Levels two and three are measured at the neighborhood and community levels,

z = percentage of the neighborhood below the poverty line, racial makeup of the neighborhood, percentage of residents over 65, average time to travel to work

and

e = FBI crime index, average achievement test score in school district, air quality measure, visibility index.

The model is estimated by maximum simulated likelihood.

The **hierarchical linear model** analyzed in this section is also called a *mixed model* and *random parameters model*. Although the three terms are usually used interchangeably, each highlights a different aspect of the structural model in (15-35). The hierarchical aspect of the model refers to the layering of coefficients that is built into stratified and panel data structures, such as in Example 15.4. The random parameters feature is a signature feature of the model that relates to the modeling of heterogeneity across units in the sample. Note that the model in (15-35) and Beron et al.'s application could be formulated without the random terms in the lower-level equations. This would then provide a convenient way to introduce interactions of variables in the linear regression model. The addition of the random component is motivated on precisely the same basis that u_i appears in the familiar random effects model in Section 11.5 and (15-39). It is important to bear in mind, in all these structures, strict mean independence is maintained between $\mathbf{u}_i$ and all other variables in the model. In most treatments, we go yet a step further and assume a particular distribution for u_i, typically joint normal. Finally, the mixed model aspect of the specification relates to the underlying integration that removes the heterogeneity, for example, in (15-13). The unconditional estimated model is a mixture of the underlying models, where the weights in the mixture are provided by the underlying density of the random component.

15.9 NONLINEAR RANDOM PARAMETER MODELS

Most of the preceding applications have used the linear regression model to illustrate and demonstrate the procedures. However, the template used to build the model has no intrinsic features that limit it to the linear regression. The initial description of the model and the first example were applied to a nonlinear model, the Poisson regression. We will examine a random parameters binary choice model in the next section as well. This random parameters model has been used in a wide variety of settings. One of the most common is the multinomial choice models that we will discuss in Chapter 18.

The simulation-based random parameters estimator/model is extremely flexible.[24] The simulation method, in addition to extending the reach of a wide variety of model classes, also allows great flexibility in terms of the model itself. For example, constraining a parameter to have only one sign is a perennial issue. Use of a lognormal specification of the parameter, $\beta_i = \exp(\beta + \sigma w_i)$, provides one method of restricting a random

[24] See Train and McFadden (2000) for discussion.

parameter to be consistent with a theoretical restriction. Researchers often find that the lognormal distribution produces unrealistically large values of the parameter. A model with parameters that vary in a restricted range that has found use is the random variable with symmetric about zero triangular distribution,

$$f(w) = \mathbf{1}[-a \le w \le 0](a + w)/a^2 + \mathbf{1}[0 < w \le a](a - w)/a^2.$$

A draw from this distribution with $a = 1$ can be computed as

$$w = \mathbf{1}[u \le .5][(2u)^{1/2} - 1] + \mathbf{1}[u > .5][1 - (2(1 - u))^{1/2}],$$

where u is the $U[0,1]$ draw. Then, the parameter restricted to the range $\beta \pm \lambda$ is obtained as $\beta + \lambda w$. A further refinement to restrict the sign of the random coefficient is to force $\lambda = \beta$, so that β_i ranges from 0 to 2λ.[25] There is a large variety of methods for simulation that allow the model to be extended beyond the linear model and beyond the simple normal distribution for the random parameters.

Random parameters models have been implemented in several contemporary computer packages. The PROC MIXED package of routines in SAS uses a kind of generalized least squares for linear, Poisson, and binary choice models. The GLAMM program—Rabe-Hesketh, Skrondal, and Pickles (2005)—written for *Stata* uses quadrature methods for several models including linear, Poisson, and binary choice. The RPM and RPL procedures in *LIMDEP/NLOGIT* use the methods described here for linear, binary choice, censored data, multinomial, ordered choice, and several others. Finally, the *MLWin* package (www.bristol.ac.uk/cmm/software/mlwin/) is a large implementation of some of the models discussed here. *MLWin* uses MCMC methods with noninformative priors to carry out maximum simulated likelihood estimation.

15.10 INDIVIDUAL PARAMETER ESTIMATES

In our analysis of the various random parameters specifications, we have focused on estimation of the population parameters $\boldsymbol{\beta}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Lambda}$ in the model,

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i,$$

for example, in Example 15.13, where we estimated a model of production. At a few points, it is noted that it might be useful to estimate the individual specific $\boldsymbol{\beta}_i$. We did a similar exercise in analyzing the Hildreth/Houck/Swamy model in Example 11.19 in Section 11.11.1. The model is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$
$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i,$$

where no restriction is placed on the correlation between $\mathbf{u}_i$ and $\mathbf{X}_i$. In this fixed effects case, we obtained a feasible GLS estimator for the population mean, $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\hat{\mathbf{W}}_i\mathbf{b}_i,$$

where

$$\hat{\mathbf{W}}_i = \left\{\sum_{i=1}^{n}[\hat{\boldsymbol{\Gamma}} + \hat{\sigma}_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}]^{-1}\right\}^{-1}[\hat{\boldsymbol{\Gamma}} + \hat{\sigma}_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}]^{-1}$$

[25] Discussion of this sort of model construction is given in Train and Sonnier (2003) and Train (2009).

and

$$\mathbf{b}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i.$$

For each group, we then proposed an estimator of $E[\boldsymbol{\beta}_i \mid \text{information in hand about group } i]$ as

$$\text{Est. } E[\boldsymbol{\beta}_i \mid \mathbf{y}_i, \mathbf{X}_i] = \mathbf{b}_i + \hat{\mathbf{Q}}_i(\hat{\boldsymbol{\beta}} - \mathbf{b}_i),$$

where

$$\hat{\mathbf{Q}}_i = \{[s_i^2(\mathbf{X}_i'\mathbf{X}_i)]^{-1} + \hat{\boldsymbol{\Gamma}}^{-1}\}^{-1}\hat{\boldsymbol{\Gamma}}^{-1}. \qquad \textbf{(15-36)}$$

The estimator of $E[\boldsymbol{\beta}_i \mid \mathbf{y}_i, \mathbf{X}_i]$ is equal to the least squares estimator plus a proportion of the difference between $\hat{\boldsymbol{\beta}}$ and $\mathbf{b}_i$. (The matrix $\hat{\mathbf{Q}}_i$ is between $\mathbf{0}$ and $\mathbf{I}$. If there were a single column in $\mathbf{X}_i$, then $\hat{q}_i$ would equal $(1/\hat{\gamma})/\{(1/\hat{\gamma}) + [1/(s_i^2/\mathbf{x}_i'\mathbf{x}_i)]\}$.)

We can obtain an analogous result for the mixed models we have examined in this chapter.[26] From the initial model assumption, we have

$$f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\theta}),$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i \qquad \textbf{(15-37)}$$

and $\boldsymbol{\theta}$ is any other parameters in the model, such as σ_ε in the linear regression model. For a panel, because we are conditioning on $\boldsymbol{\beta}_i$, that is, on $\mathbf{w}_i$, the T_i observations are independent, and it follows that

$$f(y_{i1}, y_{i2}, \ldots, y_{iT_i} \mid \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) = f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) = \Pi_t f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\theta}). \qquad \textbf{(15-38)}$$

This is the contribution of group i to the likelihood function (not its log) for the sample, given $\boldsymbol{\beta}_i$; that is, note that the log of this term is what appears in the simulated log-likelihood function in (15-31) for the normal linear model and in (15-16) for the Poisson model. The marginal density for $\boldsymbol{\beta}_i$ is induced by the density of $\mathbf{w}_i$ in (15-37). For example, if $\mathbf{w}_i$ is joint normally distributed, then $f(\boldsymbol{\beta}_i) = N[\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i, \boldsymbol{\Lambda}\boldsymbol{\Lambda}']$. As we noted earlier in Section 15.9, some other distribution might apply. Write this generically as the marginal density of $\boldsymbol{\beta}_i$, $f(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is the parameters of the underlying distribution of $\boldsymbol{\beta}_i$, for example $(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Lambda})$ in (15-37). Then, the joint distribution of $\mathbf{y}_i$ and $\boldsymbol{\beta}_i$ is

$$f(\mathbf{y}_i, \boldsymbol{\beta}_i \mid \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}) = f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\Omega}).$$

We will now use Bayes' theorem to obtain $f(\boldsymbol{\beta}_i \mid \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega})$:

$$\begin{aligned} f(\boldsymbol{\beta}_i \mid \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}) &= \frac{f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\Omega})}{f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega})} \\ &= \frac{f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\Omega})}{\int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i, \boldsymbol{\beta}_i \mid \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}) d\boldsymbol{\beta}_i} \\ &= \frac{f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\Omega})}{\int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\Omega}) d\boldsymbol{\beta}_i}. \end{aligned}$$

[26] See Revelt and Train (2000) and Train (2009).

The denominator of this ratio is the integral of the term that appears in the log-likelihood conditional on $\boldsymbol{\beta}_i$. We will return momentarily to computation of the integral. We now have the conditional distribution of $\boldsymbol{\beta}_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}$. The conditional expectation of $\boldsymbol{\beta}_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}$ is

$$E[\boldsymbol{\beta}_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] = \frac{\int_{\boldsymbol{\beta}_i} \boldsymbol{\beta}_i f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i | \mathbf{z}_i, \boldsymbol{\Omega})}{\int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i | \mathbf{z}_i, \boldsymbol{\Omega}) d\boldsymbol{\beta}_i}.$$

Neither of these integrals will exist in closed form. However, using the methods already developed in this chapter, we can compute them by simulation. The simulation estimator will be

$$\begin{aligned}\text{Est.}E[\boldsymbol{\beta}_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] &= \frac{(1/R)\sum_{r=1}^{R} \hat{\boldsymbol{\beta}}_{ir} \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \hat{\boldsymbol{\beta}}_{ir}, \hat{\boldsymbol{\theta}})}{(1/R)\sum_{r=1}^{R} \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \hat{\boldsymbol{\beta}}_{ir}, \hat{\boldsymbol{\theta}})} \\ &= \sum_{r=1}^{R} \hat{Q}_{ir} \hat{\boldsymbol{\beta}}_{ir},\end{aligned} \tag{15-39}$$

where $\hat{Q}_{ir}$ is defined in (15-20), (15-21), and

$$\hat{\boldsymbol{\beta}}_{ir} = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Delta}}\mathbf{z}_i + \hat{\boldsymbol{\Lambda}}\mathbf{w}_{ir}.$$

This can be computed after the estimation of the population parameters. (It may be more efficient to do this computation during the iterations because everything needed to do the calculation will be in place and available while the iterations are proceeding.) For example, for the random parameters linear model, we will use

$$f(y_{it} | \mathbf{x}_{it}, \hat{\boldsymbol{\beta}}_{ir}, \hat{\boldsymbol{\theta}}) = \frac{1}{\hat{\sigma}_\varepsilon \sqrt{2\pi}} \exp\left[-\frac{(y_{it} - \mathbf{x}_{it}'(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Delta}}\mathbf{z}_i + \hat{\boldsymbol{\Lambda}}\mathbf{w}_{ir}))^2}{2\hat{\sigma}_\varepsilon^2} \right]. \tag{15-40}$$

We can also estimate the conditional variance of $\boldsymbol{\beta}_i$ by estimating first, one element at a time, $E[\beta_{i,k}^2 | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}]$, then, again, one element at a time,

$$\begin{aligned}\text{Est.Var}[\beta_{\text{i, k}} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] = &\{\text{Est.}E[\beta_{i,k}^2 | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}]\} - \\ &\{\text{Est.}E[\beta_{i,k} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}]\}^2.\end{aligned} \tag{15-41}$$

With the estimates of the conditional mean and conditional variance in hand, we can then compute the limits of an interval that resembles a confidence interval as the mean plus and minus two estimated standard deviations. This will construct an interval that contains at least 95% of the conditional distribution of $\boldsymbol{\beta}_i$.

Some aspects worth noting about this computation are as follows:

- The preceding suggested interval is a classical (sampling-theory-based) counterpart to the highest posterior density interval that would be computed for $\boldsymbol{\beta}_i$ for a hierarchical Bayesian estimator.
- The conditional distribution from which $\boldsymbol{\beta}_i$ is drawn might not be symmetric or normal, so a symmetric interval of the mean plus and minus two standard deviations may pick up more or less than 95% of the actual distribution. This is likely to be a

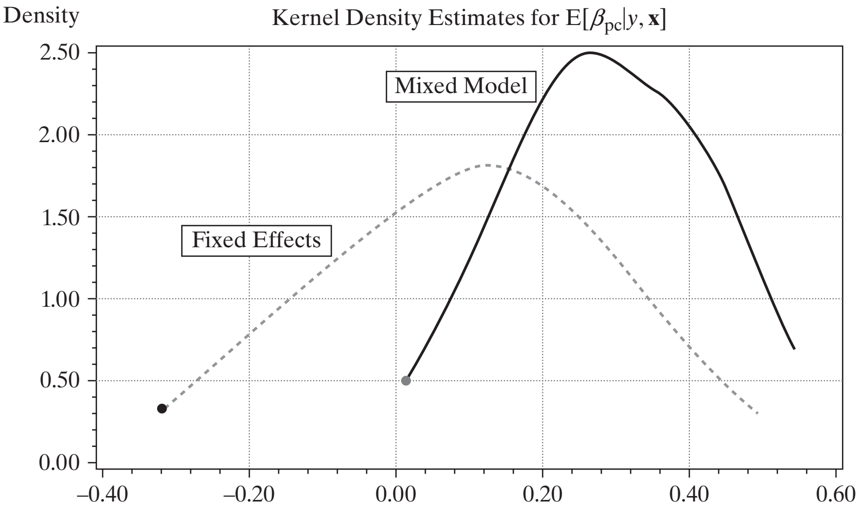
small effect. In any event, in any population, whether symmetric or not, the mean plus and minus two standard deviations will typically encompass at least 95% of the mass of the distribution.

- It has been suggested that this classical interval is too narrow because it does not account for the sampling variability of the parameter estimators used to construct it. But the suggested computation should be viewed as a *point estimate* of the interval, not an interval estimate as such. Accounting for the sampling variability of the estimators might well suggest that the endpoints of the interval should be somewhat farther apart. The Bayesian interval that produces the same estimation would be narrower because the estimator is posterior to, that is, applies only to the sample data.
- Perhaps surprisingly so, even if the analysis departs from normal marginal distributions $\boldsymbol{\beta}_i$, the sample distribution of the n estimated conditional means is not necessarily normal. Kernel estimators based on the n estimators, for example, can have a variety of shapes.
- A common misperception found in the Bayesian and classical literatures alike is that the preceding produces an estimator of $\boldsymbol{\beta}_i$. In fact, it is an estimator of conditional mean of the distribution from which $\boldsymbol{\beta}_i$ is an observation. By construction, for example, every individual with the same ($\mathbf{y}_i$. $\mathbf{X}_i$, $\mathbf{z}_i$) has the same prediction even though the $\mathbf{w}_i$ and any other stochastic elements of the model, such as ε_i, will differ across individuals.

Example 15.15 Individual State Estimates of a Private Capital Coefficient

Example 15.13 presents feasible GLS and maximum simulated likelihood estimates of Munnell's state production model. We have computed the estimates of $E[\beta_{2i}|\mathbf{y}_i, \mathbf{X}_i]$ for the 48 states in the sample using (15-36) for the fixed effects estimates and (15-39) for the random effects estimates. Figure 15.6 examines the estimated coefficients for private capital. Figure 15.6 displays kernel density estimates for the population distributions based on the fixed and random effects

FIGURE 15.6 Kernel Density Estimates of Parameter Distributions.



estimates computed using (15-36) and (15-39). The much narrower distribution corresponds to the random effects estimates. The substantial overall difference of the distributions is presumably due in large part to the difference between the fixed effects and random effects assumptions. One might suspect on this basis that the random effects assumption is restrictive.

Example 15.16 Mixed Linear Model for Wages

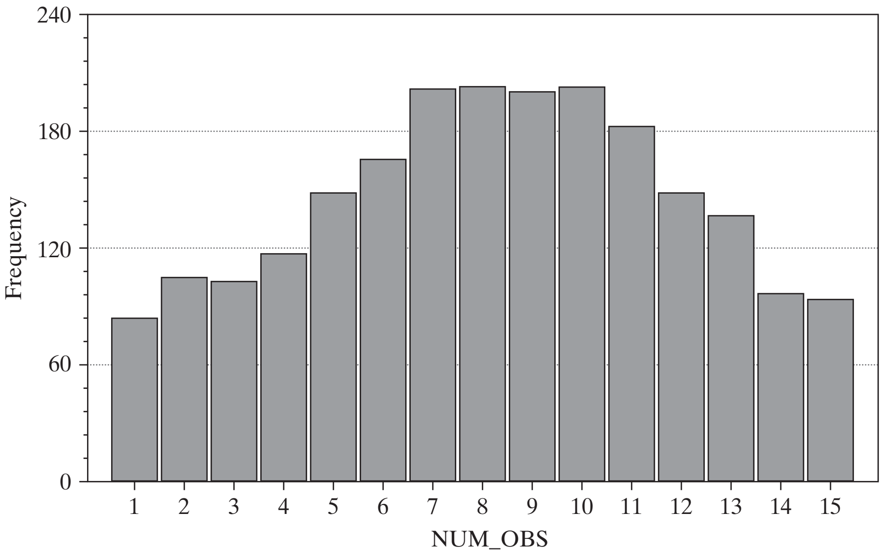
Koop and Tobias (2004) analyzed a panel of 17,919 observations in their study of the relationship between wages and education, ability, and family characteristics. (See the end of chapter applications in Chapters 3, 5, and 11 and Appendix Table F3.2 for details on the location of the data.) The variables used in the analysis are:

		Mean	*Reported mean*
Person id	(time invariant)		
Education	(time varying)	12.68	12.68
Log of hourly wage	(time varying)	2.297	2.30
Potential experience	(time varying)	8.363	8.36
Time trend	(time varying)		
Ability	(time invariant)	0.0524	0.239
Mother's education	(time invariant)	11.47	12.56
Father's education	(time invariant)	11.71	13.17
Broken home dummy	(time invariant)	0.153	0.157
Number of siblings	(time invariant)	3.156	2.83

This is an unbalanced panel of 2,178 individuals. The means in the list are computed from the sample data. The authors report the second set of means based on a subsample of 14,170 observations whose parents have at least 9 years of education. Figure 15.7 shows a frequency count of the numbers of observations in the sample.

We will estimate the following hierarchical wage model:

FIGURE 15.7 Group Sizes for Wage Data Panel.



$$\ln Wage_{it} = \beta_{1,i} + \beta_{2,i}\, Education_{it} + \beta_3\, Experience_{it} + \beta_4\, Experience_{it}^2$$
$$+\beta_5\, Broken\ Home_i + \beta_6\, Siblings_i + \varepsilon_{it},$$
$$\beta_{1,i} = \alpha_{1,1} + \alpha_{1,2}\, Ability_i + \alpha_{1,3}\, Mother's\ education_i + \alpha_{1,4}\, Father's\ education_i + u_{1,i},$$
$$\beta_{2,i} = \exp(\alpha_{2,1} + \alpha_{2,2}\, Ability_i + \alpha_{2,3}\, Mother's\ education_i + \alpha_{2,4}\, Father's\ education_i + u_{2,i}).$$

We anticipate that the education effect will be nonnegative for everyone in the population, so we have built that effect into the model by using a lognormal specification for this coefficient. Estimates are computed using the maximum simulated likelihood method described in Sections 15.6.3 and 15.7. Estimates of the model parameters appear in Table 15.8. The four models in Table 15.8 are the pooled OLS estimates, the random effects model, and the random parameters models, first assuming that the random parameters are uncorrelated ($\Gamma_{21} = 0$) and then allowing free correlation ($\Gamma_{21} =$ nonzero). The differences between the conventional and the robust standard errors in the pooled model are fairly large, which suggests the presence of latent common effects. The formal estimates of the random effects model confirm this. There are only minor differences between the FGLS and the ML estimates of the random effects model. But the hypothesis of the pooled model is decisively rejected by the likelihood ratio test. The LM statistic [Section 11.5.5 and (11-42)] is 19,353.51, which is far larger than the critical value of 3.84. So, the hypothesis of the pooled model is firmly rejected. The likelihood ratio statistic based on the MLEs is $2(12300.51 - 8013.43) = 8{,}574.16$, which produces the same conclusion. An alternative approach would be to test the hypothesis that $\sigma_u^2 = 0$ using a Wald statistic—the standard t test. The software used for this exercise reparameterizes the log likelihood in terms of $\theta_1 = \sigma_u^2/\sigma_\varepsilon^2$ and $\theta_2 = 1/\sigma_\varepsilon^2$. One approach, based on the delta method (see Section 4.4.4), would be to estimate σ_u^2 with the MLE of θ_1/θ_2. The

TABLE 15.8 Estimated Random Parameter Models

Variable	*Pooled OLS*	*RE/FGLS*	*RE/MLE*	*RE/MSL*	*Random Parameters*
Exp	0.09089	0.10272	0.10289	0.10277	0.10531
	(0.00431)	(0.00260)	(0.00261)	(0.00165)	(0.00165)
Exp^2	−0.00305	−0.00363	−0.00364	−0.00364	−0.00375
	(0.00025)	(0.00014)	(0.00014)	(0.000093)	(0.000093)
Broken Home	−0.05603	−0.06328	−0.06360	−0.05675	−0.04816
	(0.02178)	(0.02171)	(0.02252)	(0.00667)	(0.00665)
Siblings	−0.00202	−0.00664	−0.00675	−0.00841	−0.00125
	(0.00407)	(0.00384)	(0.00398)	(0.00116)	(0.00121)
Constant	0.69271	0.60995	0.61223	0.60346	*
	(0.05876)	(0.04665)	(0.04781)	(0.01744)	*
Education	0.08869	0.08954	0.08929	0.08982	*
	(0.00433)	(0.00337)	(0.00346)	(0.00123)	*
σ_ε	0.48079	0.328699	0.32913	0.32979	0.32949
σ_u	0.00000	0.350882	0.036580	0.37922	*
LM	19353.51				
ln *L*	−12300.51446		−8013.43044	−8042.97734	−7983.57355

*** Random Parameters**

$\hat{\beta}_{1,i} = 0.83417 + 0.02870\, Ability_i - 0.01355\, Mother's\ Ed_i + 0.00878\, Father's\ Ed_i + 0.30857 u_{1,i}$

(.04952) (0.01304) (0.00463) (0.00372)

$\hat{\beta}_{2,i} = \exp[-2.78412 + 0.05680\, Ability_i + 0.01960\, Mother's\ Ed_i - 0.00370\, Father's\ Ed_i + 0.10178\, u_{2,i})$

(.05582) (0.01505) (0.00503) (0.00388)

asymptotic variance of this estimator would be estimated using Theorem 4.5. Alternatively, we might note that σ_ε^2 must be positive in this model, so it is sufficient simply to test the hypothesis that $\theta_1 = 0$. Our MLE of θ_1 is 9.23137 and the estimated asymptotic standard error is 0.10427. Following this logic, then, the test statistic is 88.57. This is far larger than the critical value of 1.96, so, once again, the hypothesis is rejected. We do note a problem with the LR and Wald tests: The hypothesis that $\sigma_u^2 = 0$ produces a nonstandard test under the null hypothesis because $\sigma_u^2 = 0$ is on the boundary of the parameter space. Our standard theory for likelihood ratio testing (see Chapter 14) requires the restricted parameters to be in the interior of the parameter space, not on the edge. The distribution of the test statistic under the null hypothesis is not the familiar chi squared.[27] The simple expedient in this complex situation is to use the LM statistic, which remains consistent with the earlier conclusion.

The fifth model in Table 15.8 presents the mixed model estimates. The mixed model allows Λ_{21} to be a free parameter. The implied estimators for σ_{u1}, σ_{u2}, and $\sigma_{u,21}$ are the elements of $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$, where $\hat{\mathbf{\Lambda}} = \begin{bmatrix} 0.30857 & 0.00000 \\ -0.06221 & 0.08056 \end{bmatrix}$. Then, $\hat{\sigma}_{u1} = \sqrt{\hat{\Lambda}_{11}^2} = 0.30857$ and $\hat{\sigma}_{u2} = \sqrt{\hat{\Lambda}_{21}^2 + \hat{\Lambda}_{22}^2} = 0.10178$.

Note that for both random parameters models, the estimate of σ_ε is relatively unchanged. The models decompose the variation across groups in the parameters differently, but the overall variation of the dependent variable is largely the same.

The interesting coefficient in the model is $\beta_{2,i}$. The coefficient on education in the model is $\beta_{2,i} = \exp(\alpha_{2,1} + \alpha_{2,2}\, \textit{Ability} + \alpha_{2,3}\, \textit{Mother's education} + \alpha_{2,4}\, \textit{Father's education} + u_{2,i})$. The raw coefficients are difficult to interpret. The expected value of $\beta_{2\mathrm{i}}$ equals $\exp(\boldsymbol{\alpha}_2'\mathbf{z}_\mathrm{i} + \sigma_{\mathrm{u}2}^2/2)$. The sample means for the three variables are 0.052374, 11.4719, and 11.7092, respectively. With these values, and $\sigma_{\mathrm{u}2} = 0.10178$, the population mean value for the education coefficient is approximately 0.0727, which is in line with expectations. This is comparable to, though somewhat smaller than, the estimates for the pooled and random effects model. Of course, variation in this parameter across the sample individuals was the objective of this specification. Figure 15.8 plots a kernel density estimate for the estimated conditional means for the 2,178 sample individuals. The figure shows the range of variation in the sample estimates.

The authors of this study used Bayesian methods, but a very similar specification to ours to study heterogeneity in the returns to education. They proposed several specifications, including a latent class approach that we will consider momentarily. Their *massively* preferred specification[28] is similar to the one we used in our random parameters specification,

$$\begin{aligned} \ln \textit{Wage}_{it} &= \theta_{1,i} + \theta_{2,i}\, \textit{Education}_{it} + \boldsymbol{\gamma}'\mathbf{z}_{it} + \varepsilon_{it}, \\ \theta_{1,i} &= \theta_{1,0} + u_{1,i}, \\ \theta_{2,i} &= \theta_{2,0} + u_{2,i}. \end{aligned}$$

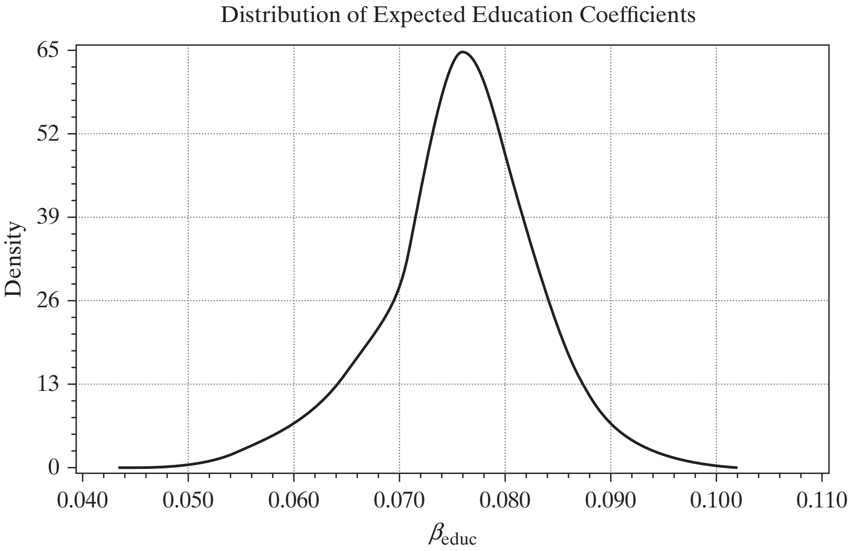
Among the preferred alternatives in their specification is a Heckman and Singer (1984) style (Section 14.15.7) latent class model with 10 classes. The specification would be

$$\begin{aligned} &\ln(\textit{Wage}_{it} \mid \textit{class} = j) = \theta_{1,j} + \theta_{2,j}\, \textit{Education}_{it} + \gamma' z_{it} + \varepsilon_{it}, \\ &\mathrm{Prob}(\textit{class} = j) = \pi_j, j = 1, \ldots, 10. \end{aligned}$$

[27] This issue is confronted in Breusch and Pagan (1980) and Godfrey (1988) and analyzed at (great) length by Andrews (1998, 1999, 2000, 2001, 2002) and Andrews and Ploberger (1994, 1995).

[28] The model selection criterion used is the Bayesian information criterion, $2\ln f(\mathbf{data} \mid \mathbf{parameters}) - K \ln n$, where the first term would be the posterior density for the data, K is the number of parameters in the model, and n is the sample size. For frequentist methods such as those we use here, the first term would be twice the log likelihood. The authors report a BIC of $-16{,}528$ for their preferred model. The log likelihood for the 5 class latent class model reported below is -8053.676. With 22 free parameters (8 common parameters in the regression $+\ 5(\theta_1$ and $\theta_2) + 4$ free class probabilities), the BIC for our model is $-16{,}275.45$.

FIGURE 15.8 Kernel Density Estimate for Education Coefficient.



We fit this alternative model to explore the sensitivity of the returns coefficient to the specification. With 10 classes, the frequentist approach converged, but several of the classes were estimated to be extremely small—on the order of 0.1% of the population, and these segments produced nonsense values of θ_2 such as -5.0. Results for a finite mixture model with 5 classes are as follows (the other model coefficients are omitted):

Class	θ_{Ed}	π
1	0.09447	0.32211
2	0.05354	0.03644
3	0.09988	0.09619
4	0.07155	0.33285
5	0.05677	0.21241

The weighted average of these results is 0.07789. The numerous estimates of the returns to education computed in this example are in line with other studies, in this paper, elsewhere in the book, and in other studies. What we have found here is that the estimated returns, for example, by OLS in Table 15.8, are a bit lower when the model accounts for heterogeneity in the population.

15.11 MIXED MODELS AND LATENT CLASS MODELS

Sections 15.7 through 15.10 examined different approaches to modeling parameter heterogeneity. The fixed effects approach begun in Section 11.4 is extended to include the full set of regression coefficients in Section 11.10.1 where

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$
$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i,$$

and no restriction is placed on $E[\mathbf{u}_i|\mathbf{X}_i]$. Estimation produces a feasible GLS estimate of $\boldsymbol{\beta}$. Estimation of $\boldsymbol{\beta}$ begins with separate least squares estimation with each group, i—because of the correlation between $\mathbf{u}_i$ and $\mathbf{x}_{it}$, the pooled estimator is not consistent. The efficient estimator of $\boldsymbol{\beta}$ is then a mixture of the $\mathbf{b}_i$'s. We also examined an estimator of $\boldsymbol{\beta}_i$, using the optimal predictor from the conditional distributions, (15-39). The crucial assumption underlying the analysis is the possible correlation between $\mathbf{X}_i$ and $\mathbf{u}_i$. We also considered two modifications of this random coefficients model. First, a restriction of the model in which some coefficients are nonrandom provides a useful simplification. The familiar fixed effects model of Section 11.4 is such a case, in which only the constant term varies across individuals. Second, we considered a hierarchical form of the model

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{u}_i. \tag{15-42}$$

This approach is applied to an analysis of mortgage rates in Example 11.23.

A second approach to random parameters modeling builds from the crucial assumption added to (15-42) that $\mathbf{u}_i$ and $\mathbf{X}_i$ are uncorrelated. The general model is defined in terms of the conditional density of the random variable, $f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\theta})$, and the marginal density of the random coefficients, $f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})$, in which $\boldsymbol{\Omega}$ is the separate parameters of this distribution. This leads to the mixed models examined in this chapter. The random effects model that we examined in Section 11.5 and several other points is a special case in which only the constant term is random (like the fixed effects model). We also considered the specific case in which u_i is distributed normally with variance σ_u^2.

A third approach to modeling heterogeneity in parametric models is to use a discrete distribution, either as an approximation to an underlying continuous distribution, or as the model of the data-generating process in its own right. (See Section 14.15.) This model adds to the preceding a nonparametric specification of the variation in $\boldsymbol{\beta}_i$,

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_j|\mathbf{z}_i) = \pi_{ij}, j = 1, \ldots, J.$$

A somewhat richer, semiparametric form that mimics (15-42) is

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_j|\mathbf{z}_i) = \pi_j(\mathbf{z}_i, \boldsymbol{\Omega}), j = 1, \ldots, J.$$

We continue to assume that the process generating variation in $\boldsymbol{\beta}_i$ across individuals is independent of the process that produces $\mathbf{X}_i$—that is, in a broad sense, we retain the random effects approach. In the last example of this chapter, we will examine a comparison of mixed and finite mixture models for a nonlinear model.

Example 15.17 Maximum Simulated Likelihood Estimation of a Binary Choice Model

Bertschek and Lechner (1998) analyzed the innovations of a sample of German manufacturing firms. They used a probit model (Sections 17.2–17.4) to study firm innovations. The model is for Prob($y_{it} = 1|\mathbf{x}_{it}, \boldsymbol{\beta}_i$) where

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

The independent variables in the model are

$x_{it,1}$ = constant,
$x_{it,2}$ = log of sales,

$x_{it,3}$ = relative size = ratio of employment in business unit to employment in the industry,
$x_{it,4}$ = ratio of industry imports to (industry sales + imports),
$x_{it,5}$ = ratio of industry foreign direct investment to (industry sales + imports),
$x_{it,6}$ = productivity = ratio of industry value added to industry employment,
$x_{it,7}$ = dummy variable indicating firm is in the raw materials sector,
$x_{it,8}$ = dummy variable indicating the firm is in the investment goods sector.

The sample consists of 1,270 German firms observed for five years, 1984–1988. (See Appendix Table F15.1.) The density that enters the log likelihood is

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

where

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

To be consistent with Bertschek and Lechner (1998) we did not fit any firm-specific time-invariant components in the main equation for $\boldsymbol{\beta}_i$. Table 15.9 presents the estimated coefficients for the basic probit model in the first column. These are the values reported in the 1998 study. The estimates of the means, $\boldsymbol{\beta}$**,** are shown in the second column. There appear to be large differences in the parameter estimates, although this can be misleading as there is large variation across the firms in the posterior estimates. The third column presents the square roots of the implied diagonal elements of $\boldsymbol{\Sigma}$ computed as the diagonal elements of **CC′.** These estimated standard deviations are for the underlying distribution of the parameter in the model—they are not estimates of the standard deviation of the sampling distribution of the estimator. That is shown for the mean parameter in the second column. The fourth column presents the sample means and standard deviations of the 1,270 estimated conditional estimates of the coefficients.

TABLE 15.9 Estimated Random Parameters Model

	Probit	*RP Mean*	*RP Std. Dev.*	*Empirical Distn.*
Constant	−1.96031	−3.43237	0.44947	−3.42768
	(0.37298)	(0.28187)	(0.02121)	(0.15151)
ln Sales	0.17711	0.31054	0.09014	0.31113
	(0.03580)	(0.02757)	(0.00242)	(0.06206)
Relative Size	1.07274	4.36456	3.91986	4.37532
	(0.26871)	(0.27058)	(0.23881)	(1.03431)
Import	1.13384	1.69975	0.93927	1.70413
	(0.24331)	(0.18440)	(0.07287)	(0.20289)
FDI	2.85318	2.91042	0.93468	2.91600
	(0.64233)	(0.47161)	(0.32610)	(0.15182)
Productivity	−2.34116	−4.05320	2.52542	−4.02747
	(1.11575)	(1.04683)	(0.21665)	(0.54492)
Raw materials	−0.27858	−0.42055	0.34962	−0.41966
	(0.12656)	(0.10694)	(0.06926)	(0.05948)
Investment	0.18796	0.30491	0.04672	0.30477
	(0.06287)	(0.04756)	(0.02812)	(0.00812)
ln L	−4114.05		−3524.66	

TABLE 15.10 Estimated Latent Class Model

	Class 1	*Class 2*	*Class 3*	*Posterior*
Constant	−2.32073	−2.70546	−8.96773	−3.77582
	(0.65898)	(0.73335)	(2.46099)	(2.14253)
ln Sales	0.32265	0.23337	0.57148	0.34283
	(0.06516)	(0.06790)	(0.19448)	(0.08919)
Relative Size	4.37802	0.71974	1.41997	2.57719
	(0.87099)	(0.29163)	(0.71765)	(1.29454)
Import	0.93572	2.25770	3.12177	1.80964
	(0.41140)	(0.50726)	(1.33320)	(0.74348)
FDI	2.19747	2.80487	8.37073	3.63157
	(1.58729)	(1.02824)	(2.09091)	(1.98176)
Productivity	−5.86238	−7.70385	−0.91043	−5.48219
	(1.53051)	(4.10134)	(1.46314)	(1.78348)
Raw Materials	−0.10978	−0.59866	0.85608	−0.07825
	(0.17459)	(0.37942)	(0.40407)	(0.36666)
Investment	0.13072	0.41353	0.46904	0.29184
	(0.11851)	(0.12388)	(0.23876)	(0.12462)
ln L = −3503.55				
Class Prob (Prior)	0.46950	0.33073	0.19977	
	(0.03762)	(0.03407)	(0.02629)	
Class Prob (Posterior)	0.46950	0.33073	0.19976	
	(0.39407)	(0.28906)	(0.32492)	
Pred. Count	649	366	255	

The latent class formulation developed in Section 14.15 provides an alternative approach for modeling latent parameter heterogeneity.[29] To illustrate the specification, we will reestimate the random parameters innovation model using a three-class latent class model. Estimates of the model parameters are presented in Table 15.10. The estimated conditional mean shown, which is comparable to the empirical means in the rightmost column in Table 15.9 for the random parameters model, are the sample average and standard deviation of the 1,270 firm-specific posterior mean parameter vectors. They are computed using $\hat{\boldsymbol{\beta}}_i = \Sigma_{j=1}^{3}\hat{\boldsymbol{\pi}}_{ij}\,\hat{\boldsymbol{\beta}}_j$, where $\hat{\pi}_{ij}$ is the conditional estimator of the class probabilities in (14-97). These estimates differ considerably from the probit model, but they are quite similar to the empirical means in Table 15.9. In each case, a confidence interval around the posterior mean contains the one-class pooled probit estimator. Finally, the (identical) prior and average of the sample posterior class probabilities are shown at the bottom of the table. The much larger empirical standard deviations reflect that the posterior estimates are based on aggregating the sample data and involve, as well, complicated functions of all the model parameters. The estimated numbers of class members are computed by assigning to each firm the predicted class associated with the highest posterior class probability.

[29] See Greene (2001) for a survey. For two examples, Nagin and Land (1993) employed the model to study age transitions through stages of criminal careers and Wang et al. (1998) and Wedel et al. (1993) used the Poisson regression model to study counts of patents.

15.12 SUMMARY AND CONCLUSIONS

This chapter has outlined several applications of simulation-assisted estimation and inference. The essential ingredient in any of these applications is a random number generator. We examined the most common method of generating what appear to be samples of random draws from a population—in fact, they are deterministic Markov chains that only appear to be random. Random number generators are used directly to obtain draws from the standard uniform distribution. The inverse probability transformation is then used to transform these to draws from other distributions. We examined several major applications involving random sampling:

- Random sampling, in the form of bootstrapping, allows us to infer the characteristics of the sampling distribution of an estimator, in particular its asymptotic variance. We used this result to examine the sampling variance of the median in random sampling from a nonnormal population. Bootstrapping is also a useful, robust method of constructing confidence intervals for parameters.
- Monte Carlo studies are used to examine the behavior of statistics when the precise sampling distribution of the statistic cannot be derived. We examined the behavior of a certain test statistic and of the maximum likelihood estimator in a fixed effects model.
- Many integrals that do not have closed forms can be transformed into expectations of random variables that can be sampled with a random number generator. This produces the technique of Monte Carlo integration. The technique of maximum simulated likelihood estimation allows the researcher to formulate likelihood functions (and other criteria such as moment equations) that involve expectations that can be integrated out of the function using Monte Carlo techniques. We used the method to fit random parameters models.

  The techniques suggested here open up a vast range of applications of Bayesian statistics and econometrics in which the characteristics of a posterior distribution are deduced from random samples from the distribution, rather than brute force derivation of the analytic form. Bayesian methods based on this principle are discussed in Chapter 16.

Key Terms and Concepts

- Antithetic draws
- Block bootstrap
- Cholesky decomposition
- Cholesky factorization
- Direct product
- Discrete uniform distribution
- Fundamental probability transformation
- Gauss–Hermite quadrature
- GHK smooth recursive simulator
- Hadamard product
- Halton draws
- Hierarchical linear model
- Incidental parameters problem
- Kronecker product
- Markov chain
- Mersenne Twister
- Mixed model
- Monte Carlo integration
- Nonparametric bootstrap
- Paired bootstrap
- Parametric bootstrap
- Percentile method
- Period
- Power of a test
- Pseudo maximum likelihood estimator
- Pseudo–random number generator
- Schur product
- Seed
- Simulation
- Size of a test
- Shuffling
- Specificity

Exercises

1. The exponential distribution has density $f(x) = \theta \exp(-\theta x)$. How would you obtain a random sample of observations from an exponential population?
2. The Weibull population has survival function $S(x) = \lambda p \exp(-(\lambda x)p)$. How would you obtain a random sample of observations from a Weibull population? (The survival function equals one minus the cdf.)
3. Derive the first-order conditions for nonlinear least squares estimation of the parameters in (15-2). How would you estimate the asymptotic covariance matrix for your estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$?

Applications

1. Does the Wald statistic reject the null hypothesis too often? Construct a Monte Carlo study of the behavior of the Wald statistic for testing the hypothesis that γ equals zero in the model of Section 15.5.1. Recall that the Wald statistic is the square of the *t* ratio on the parameter in question. The procedure of the test is to reject the null hypothesis if the Wald statistic is greater than 3.84, the critical value from the chi-squared distribution with one degree of freedom. Replicate the study in Section 15.5.1 that is for all three assumptions about the underlying data.
2. A regression model that describes income as a function of experience is

$$\ln Income_i = \beta_1 + \beta_2\, Experience_i + \beta_3\, Experience_i^2 + \varepsilon_i.$$

3. The model implies that ln *Income* is largest when ∂ ln *Income*/∂ *Experience* equals zero. The value of *Experience* at which this occurs is where $\beta_4 + 2\beta_5\, Experience = 0$, or $Experience* = -\beta_2/\beta_3$. Describe how to use the delta method to obtain a confidence interval for *Experience**. Now, describe how to use bootstrapping for this computation. A model of this sort using the Cornwell and Rupert data appears in Example 15.6. Using your proposals here, carry out the computations for that model using the Cornwell and Rupert data.