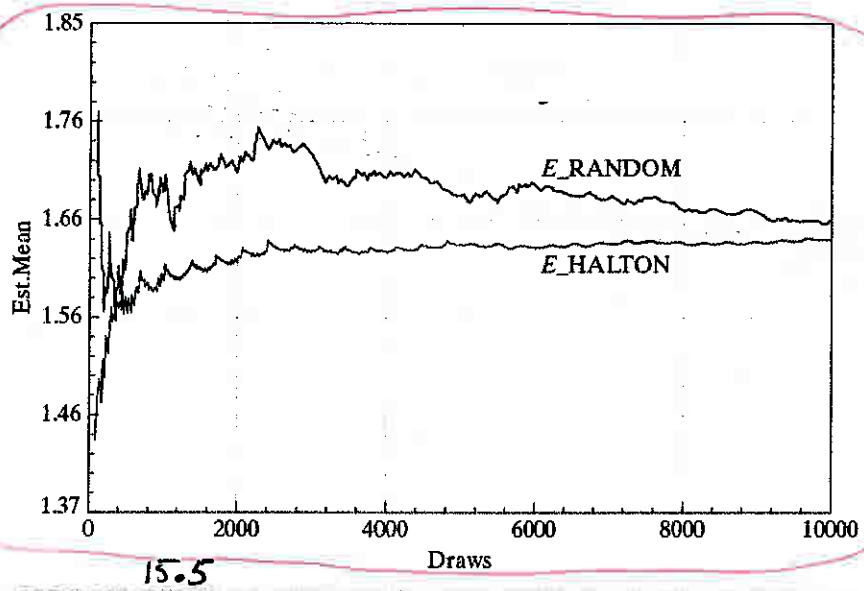

15.5

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

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 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 $n = 100, 110, \ldots, 10000$, we averaged the first n observations in the sample. Figure 17.3 plots the evolution of the sample means as a function of the sample size. The lighter trace is the sequence of Halton-based means. The greater stability of the Halton estimator is clearly evident in the figure.

FIG 15.5

lower

15.5

17.3.2 IMPORTANCE SAMPLING

Consider the general computation

$$F(x) = \int_L^U f(x)\, g(x)\, dx,$$

where $g(x)$ is a continuous function in the range $[L, U]$. (We could achieve greater generality by allowing more complicated functions, but for current purposes, we limit ourselves to straightforward cases.) Now, suppose that $g(x)$ is nonnegative in the entire range $[L, U]$. To normalize the weighting function, we suppose, as well, that

$$K = \int_L^U g(x)\, dx$$

gamma distribution with parameters $P = \lambda = \frac{1}{2}$ [see (B-39)], so

$$I(x) = \frac{\left(\frac{1}{2}\right)^{1/2}}{\Gamma\left(\frac{1}{2}\right)} x^{-1/2} e^{-(1/2)x}.$$

After a bit of manipulation, we find that

$$\frac{f(x)g(x)}{I(x)} = q(x) = e^{(1/2)[x-(\ln x)^2]} x^{1/2}.$$

Therefore, to estimate the mean of this lognormal distribution, we can draw a random sample of values x_i from the $\chi^2[1]$ distribution, which we can do by squaring the draws in a sample from the standard normal distribution, then computing the average of the sample of values, $q(x_i)$.

We carried out this experiment with 1,000 draws from a standard normal distribution. The mean of our sample was 1.6974, compared with a true mean of 1.649, so the error was less than 3 percent.

15.6.2.b

COMPUTING MULTIVARIATE NORMAL PROBABILITIES USING THE GHK SIMULATOR

is typically done using quadrature and

The computation of bivariate normal probabilities 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 suggest an alternative approach. Suppose that $\mathbf{x}$ has a K-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. (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, $\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_K < x_K < b_K]$. Our Monte Carlo integration technique is well suited for this ~~well-defined~~ 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 17.2.4. Now, define

Hw: Confirm x-ref section number is correct

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

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

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

FN 6

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

6 This method was suggested by Lerman and Manski (1981).

accurate and efficient simulation methods for this computation. A survey of the methods is given in McFadden and Ruud (1994).[7]

Among the simulation methods examined in the survey, the **GHK smooth recursive simulator** appears to be the most accurate.[8] 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** $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is a lower triangular matrix (see Section A.6.11). The elements of $\mathbf{L}$ 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, 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, again, that the range is standardized since $l_{11} = \sigma_{11}$.) ~~The draw can be obtained from a $U[0, 1]$ observation using (17-4)~~. 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 23. ~~(See Example 23.16 and Section 23.11.5.)~~ 17.

[7] 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).

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

15.6.3 SIMULATION BASED ESTIMATION OF RANDOM EFFECTS MODELS

In Section 15.4.2, (15-14) and (15-5), 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,

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

We now consider how to estimate the 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 is 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,

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

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

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

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,

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

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 [see Gourieroux and Monfort (1996)] is that $n^{1/2}/R \rightarrow 0$. 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 recreated 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

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

For the derivative term,

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

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

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

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

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

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

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

and

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

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

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

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

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

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, $\theta = (\beta', \sigma)'$ and

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

$$\mathbf{g}_{itr}(\theta) = [y_{it} - \mu_{itr}(\theta)]\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix} \tag{15-25}$$

$$\mathbf{H}_{itr}(\theta) = -\mu_{itr}(\theta)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}'$$

Estimates of the random effects model parameters would be obtained by using these expressions in the preceding general template above. 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 outlined above 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 reduntant 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 $\theta = (\beta, \sigma_u, \sigma_\varepsilon)$. We will require

$$P_{itr}(\theta) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\left[-\frac{(y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta}-\sigma_u w_{ir})^2}{2\sigma_\varepsilon^2}\right],$$

$$\mathbf{g}_{itr}(\theta) = \begin{bmatrix}\left(\frac{(y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta}-\sigma_u w_{ir})}{\sigma_\varepsilon^2}\right)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}\\ \frac{(y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta}-\sigma_u w_{ir})^2}{\sigma_\varepsilon^3}-\frac{1}{\sigma_\varepsilon}\end{bmatrix} = \begin{bmatrix}(\varepsilon_{itr}/\sigma_\varepsilon^2)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}\\ (1/\sigma_\varepsilon)[(\varepsilon_{itr}^2/\sigma_\varepsilon^2)-1]\end{bmatrix} \qquad (15\text{-}26)$$

$$\mathbf{H}_{itr}(\theta) = \begin{bmatrix}-(1/\sigma_\varepsilon^2)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}' & -(2\varepsilon_{itr}/\sigma_\varepsilon^3)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}\\ -(2\varepsilon_{itr}/\sigma_\varepsilon^3)(\mathbf{x}_{it}' \quad w_{ir}) & -(3\varepsilon_{itr}^2/\sigma_\varepsilon^4)+(1/\sigma_\varepsilon^2)\end{bmatrix}.$$

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 is changed to $-\sigma$, then the exact same value of the function and all derivatives results. The implication is that the sign of σ is not identified in this setting. With no loss of generality, it is normalized to (+) to be consistent with the underlying theory that it is a standard deviation. positive

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 11.11.1, 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, then to regress $\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{LSDV}$ 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 15.4.3 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 diagonal covariance matrix $\boldsymbol{\Gamma}$. 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. The parameter vector in the model is now

$$\boldsymbol{\theta} = (\beta_1,\ldots,\beta_K,\lambda_1,\ldots,\lambda_K,\sigma_\varepsilon).$$

(In an application, some of the γs might be fixed at zero to make the corresponding parameters nonrandom.) In order to extend the model, the disturbance in (15-16), $\varepsilon_{itr} = (y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - \sigma_u w_{ir})$, 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 (15-26) to produce

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

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, denoted $\mathbf{x}_{it} \bullet \mathbf{w}_{ir}$.

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 (15-16) 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, Γ as we did with the fixed effects case. We will now let Λ in (15-29) be the Cholesky factorization of Γ, so $\Gamma = \Lambda\Lambda'$. (This was already the case for the simpler model with diagonal Γ.) The implementation in (15-26) will be complicated a bit. The derivatives with respect to β are unchanged. For the derivatives with respect to Λ, it is useful to assume for the moment that Λ 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 Λ_{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 (β,Λ) are equated to zero. The result after some manipulation is

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

By multiplying this by σ_ε^2, we find, as usual, that σ_ε^2 is not needed for computation of the estimates of β,Λ. Thus, we can view the solution as the counterpart to least squares, which might call, instead, the minimum simulated sum of squares estimator. Once the simulated sum of squares is minimized with respect to β and Λ, then the solution for σ_ε^2 can be obtained via the likelihood equation,

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

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

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

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

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

and $Q_i = T_i/\Sigma_i T_i$ 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 11.1 and 15.6 are shown in Table 15.6. The table presents estimates based on several assumptions. The encompassing model is

$$\ln Wage_{it} = \beta_{1,i} + \beta_{2,i} Wks_{i,t} + \ldots + \beta_{12,i} Fem_i + \beta_{13,i} Blk_i + \varepsilon_{it}, \qquad (15\text{-}32)$$

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

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

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

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 heteroskedasticity 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**.

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 difference in the log likelihood functions would be attributable to this as well. Note, however, that the difference is smaller than it first appears – the comparison of 586.446 to 307.883 is misleading; the comparison should be of the difference of the two values from the log likelihood from the pooled model of -1523.254. This produces a difference of about 14%.

The full random parameters model is shown in the last two columns. Based on the likelihood ratio statistic of 2(668.630 – 568.446) = 200.368 with 12 degrees of freedom, we would reject the hypothesis that $\lambda_2 = \lambda_3 = \ldots = \lambda_{13} = 0$. 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 $\beta_{12,I}$, the coefficient on Fem_i. The estimate of the mean, β_{12}, is -0.03864 with an estimated standard error of 0.02467. This implies a confidence interval for this parameter of -0.03864 ± 1.96(0.02467) = [-0.086993,0.009713] But, this is only the location of the center of the distribution. With an estimate of λ_k of 0.2831, the random parameters model suggests that in the population, 95% of individuals have an effect of Fem_i within -0.03864 ± 1.96(0.2831) = [-0.5935,0.5163]. This is still centered near zero, but has a

different interpretation from the simple confidence interval for β, itself. 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.

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	.00422 (.00108)	.00096 (.00059)	.00084 (.00060)	.00086 (.00099)	-.00029 (.00082)	.00614 (.00042)
South	-.05564 (.01253)	-.00825 (.02246)	.00577 (.03159)	.00935 (.03106)	.04941 (.02002)	.20997 (.01702)
SMSA	.15167 (.01207)	-.02840 (.01616)	-.04748 (.01896)	-.04913 (.03710)	-.05486 (.01747)	.01165 (.02738)
MS	.04845 (.02057)	-.07090 (.01793)	-.04138 (.01899)	-.04142 (.02176)	-.06358* (.01896)	.02524 (.03190)
Exp	.04010 (.00216)	.08748 (.00225)	.10721 (.00248)	.10668 (.00290)	.09291 (.00216)	.01803 (.00092)
Exp^2	-.00067 (.0000474)	-.00076 (.0000496)	-.00051 (.0000545)	-.00050 (.0000661)	-.00019 (.0000732)	.0000812 (.00002)
Occ	-.14001 (.01466)	-.04322 (.01299)	-.02512 (.01378)	-.02437 (.02485)	-.00963 (.01331)	.02565 (.01019)
Ind	.04679 (.01179)	.00378 (.01373)	.01380 (.01529)	.01610 (.03670)	.00207 (.01357)	.02575 (.02420)
Union	.09263 (.01280)	.05835 (.01350)	.03873 (.01481)	.03724 (.02814)	.05749 (.01469)	.15260 (.02022)
Ed	.05670 (.00261)	.10707 (.00511)	.13562 (.01267)	.13952 (.03746)	.09356 (.00359)	.00409 (.00160)
Fem	-.36779 (.02510)	-.30938 (.04554)	-.17562 (.11310)	-.11694 (.10784)	-.03864 (.02467)	.28310 (.00760)
Blk	-.16694 (.02204)	-.21950 (.05252)	-.26121 (.13747)	-.15184 (.08356)	-.26864 (.03156)	.02930 (.03841)
Constant	5.25112 (.07129)	4.04144 (.08330)	3.12622 (.17761)	3.08362 (.48917)	3.81680 (.06905)	.26347 (.01628)
σ_u	.00000	.31453	.15334	.21164 (.03070)		
σ_ε	.34936	.15206	.83949	.15326 (.00217)	.14354 (.00208)	
ln*L*	-1523.254		307.873	568.446	668.630	

[a] Based on 500 Halton draws

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

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

This is the counterpart to the fixed effects model in Section 11.4. 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 $\boldsymbol{\Sigma}$ in the reparameterized model

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

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

In Example 11/19, we examined Munell's production model for gross state product,

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

The panel consists of state level data for 17 years. The model in Example 22/29 (and Munnell's) provide 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 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, in spite of the fact that both estimators are consistent and the sample is fairly large. The underlying standard deviations are computed using **G** as the covariance matrix. [For these 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 (11-86).] The increase in the standard errors is striking. This suggests that there is considerable variation in the parameters across states. We have used (11-87) to compute the estimates of the state specific coefficients.

The rightmost 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 u_i, the square roots of the diagonal elements of $\boldsymbol{\Sigma}$. These are obtained by computing the square roots of the diagonal elements of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. The estimate of $\boldsymbol{\Lambda}$ is shown here.

$$\hat{\boldsymbol{\Lambda}} = \begin{bmatrix} 0.04114 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.00715 & 0.07266 & 0 & 0 & 0 & 0 & 0 \\ -0.02446 & 0.12392 & 0.07247 & 0 & 0 & 0 & 0 \\ 0.09972 & -0.00644 & 0.31916 & 0.07614 & 0 & 0 & 0 \\ -0.08928 & 0.02143 & -0.25105 & 0.07583 & 0.04053 & 0 & 0 \\ 0.03842 & -0.06321 & -0.03992 & -0.06693 & -0.05490 & 0.00857 & 0 \\ -0.00833 & -0.00257 & -0.02478 & 0.01594 & 0.00102 & -0.00185 & 0.0018. \end{bmatrix}$$

An estimate of the correlation matrix for the parameters might also be informative. This is also derived from $\hat{\boldsymbol{\Lambda}}$ by computing $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$ 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

$$\mathbf{R} = \begin{bmatrix} 1 & & & & & & \\ 0.0979 & 1 & & & & & \\ -0.1680 & 0.83040 & 1 & & & & \\ 0.2907 & 0.00980 & 0.3983 & 1 & & & \\ -0.3180 & 0.04481 & -0.3266 & -0.8659 & 1 & & \\ 0.3176 & -0.48890 & -0.6622 & -0.3277 & -0.06073 & 1 & \\ -0.2700 & -0.10940 & -0.4253 & -0.7097 & 0.94190 & -0.08228 & 1. \end{bmatrix}$$

TABLE 15.7 Estimated Random Coefficients Production Models

	Least Squares		*Feasible GLS*			*Maximum Simulated Likelihood*	
Variable	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*	*Popn. Std. Deviation*	*Estimate*	*Std. Error*
Constant	1.9260	0.05250	1.6533	1.08331	7.0782	1.9463	0.03569
						(0.0411)	
ln pc	0.3120	0.01109	0.09409	0.05152	0.3036	0.2962	0.00882
						(0.0730)	
ln hwy	0.05888	0.01541	0.1050	0.1736	1.1112	0.09515	0.01157
						(0.146)	
ln water	0.1186	0.01236	0.07672	0.06743	0.4340	0.2434	0.01929
						(0.343)	
ln util	0.00856	0.01235	−0.01489	0.09886	0.6322	−0.1855	0.02713
						(0.281)	
ln emp	0.5497	0.01554	0.9190	0.1044	0.6595	0.6795	0.02274
						(0.121)	
unemp	−0.00727	0.001384	−0.004706	0.002067	0.01266	−0.02318	0.002712
						(0.0308)	
σ_ε	0.08542			0.2129		0.02748	
ln L	853.1372					1567.233	

5.8 HIERARCHICAL LINEAR MODELS

Example 11.20 examined an application of a "two level model," or "hierarchical model" for mortgage rates,

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

The second level equation is

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

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 (11-82) 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 following.) This model retains the earlier stochastic specification but adds the measurement equation to the generation of the random parameters. In principle, this is actually only a minor extension of the model used thus far. The 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)-(15-31) with the addition of 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 $\boldsymbol{\Delta}$, $\boldsymbol{\lambda}'$ 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,

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

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

$$\ln Price_{site,nei,com} = \pi^0_{nei,com} + \sum_{k=1}^{K} \pi^k_{nei,com} x_{k,site,nei,com} + \varepsilon_{site,nei,com},$$

$$\pi^k_{nei,com} = \beta^{0,k}_{com} + \sum_{l=1}^{L} \beta^{l,k}_{com} z_{k,nei,com} + r^k_{nei,com}, k = 0, \ldots, K,$$

$$\beta^{l,k}_{com} = \gamma^{0,l,k} + \sum_{m=1}^{M} \gamma^{m,l,k} e_{m,com} + u^{l,k}_{com}, l = 1, \ldots, L.$$

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 (9-63). 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 9.16. 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 (9-63) or 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 9.5. (The random effects model is the special case of (9-63) when only the constant term is random.) It is important to bear in mind, in all these structures, strict mean independence is maintained between 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 (9-56). The unconditional estimated

AV: Check subscripts — OK italic?

(15-35) (15-15) (15-35) (15-35) 11.5 and

AV: Provide missing text

→ the underlying integration that removes the heterogeneity, for example, in (15-13).

model is a mixture of the underlyin models, where th weights in the mixture are provided by the underlying densit of the random

15.9 NONLINEAR RANDOM PARAMETER MODELS

Most of the preceding has used the linear regression model to illustrate the procedures and demonstrate the applications. 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 17.

The simulation based random parameters estimator/model is extremely flexible. [See Train and McFadden (2000) for discussion.] 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 parameter to be consistent with a theoretical restriction. Researchers often find that the lognormal distrbution 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,

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

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

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

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. [Discussion of this sort of model construction is given in Train and Sonnier (2003) and Train (2009).] There are 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 pacakges. 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 uses the methods described here for linear, binary choice, censored data, multinomial and ordered choice, and several others. Finally, the MLWin package (http://cmm.bristol.ac.uk/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, e.g., in Example 15.13, where we estimated the parameters of the normal distribution of $\beta_{Fem,i}$. 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 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 $\hat{\mathbf{W}}_i = \left\{\sum_{i=1}^{n}\left[\hat{\boldsymbol{\Gamma}} + \hat{\sigma}_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right]^{-1}\right\}^{-1}\left[\hat{\boldsymbol{\Gamma}} + \hat{\sigma}_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right]^{-1}$

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 $E[\boldsymbol{\beta}_i|\text{information in hand about group } i]$ as

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

where (15-36)

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

The estimator of $E[\boldsymbol{\beta}_i|\mathbf{y}_i,\mathbf{X}_i]$ is equal to the estimator of the population mean 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. From the initial model assumption, we have

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

where (15-36)

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

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

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

that is, This is the contribution of group i to the likelihood function (not its log) for the sample, given $\boldsymbol{\beta}_i$. I.e., 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 $f(\beta_i) = N[\beta+\Delta z_i, \Lambda\Lambda']$. As we noted earlier in Section 15.9, some other distribution might apply. Write this generically as the marginal density of β_i, $f(\beta_i|z_i,\Omega)$, where Ω is the parameters of the underlying distribution of β_i, for example (β,Δ,Λ) in (15-37). Then, the joint distribution of y_i and β_i is

$$f(y_i,\beta_i|X_i,z_i,\theta,\Omega) = f(y_i|X_i,\beta_i,\theta)\, f(\beta_i|z_i,\Omega).$$

Bayes's

We will now use Bayes Theorem to obtain $f(\beta_i|y_i,X_i,z_i,\theta,\Omega)$:

$$\begin{aligned} f(\beta_i \mid y_i, X_i, z_i, \theta, \Omega) &= \frac{f(y_i \mid X_i, \beta_i, \theta) f(\beta_i \mid z_i, \Omega)}{f(y_i \mid X_i, z_i, \theta, \Omega)} \\ &= \frac{f(y_i \mid X_i, \beta_i, \theta) f(\beta_i \mid z_i, \Omega)}{\int_{\beta_i} f(y_i, \beta_i \mid X_i, z_i, \theta, \Omega) d\beta_i} \\ &= \frac{f(y_i \mid X_i, \beta_i, \theta) f(\beta_i \mid z_i, \Omega)}{\int_{\beta_i} f(y_i \mid X_i, \beta_i, \theta) f(\beta_i \mid z_i, \Omega) d\beta_i}. \end{aligned}$$

The denominator of this ratio is the integral of the term that appears in the log likelihood conditional on β_i. We will return momentarily to computation of the integral. We now have the conditional distribution of $\beta_i|y_i,X_i,z_i,\theta,\Omega$. The conditional expectation of $\beta_i|y_i,X_i,z_i,\theta,\Omega$ is

$$E[\beta_i \mid y_i, X_i, z_i, \theta, \Omega] = = \frac{\int_{\beta_i} \beta_i f(y_i \mid X_i, \beta_i, \theta) f(\beta_i \mid z_i, \Omega)}{\int_{\beta_i} f(y_i \mid X_i, \beta_i, \theta) f(\beta_i \mid z_i, \Omega) d\beta_i}$$

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

$$\begin{aligned} Est.E[\beta_i \mid y_i, X_i, z_i, \theta, \Omega] &= \frac{(1/R)\Sigma_{r=1}^{R} \hat{\beta}_{ir} \Pi_{t=1}^{T_i} f(y_{it} \mid x_{it}, \hat{\beta}_{ir}, \hat{\theta})}{(1/R)\Sigma_{r=1}^{R} \Pi_{t=1}^{T_i} f(y_{it} \mid x_{it}, \hat{\beta}_{ir}, \hat{\theta})} \\ &= \sum_{r=1}^{R} \hat{Q}_{ir} \hat{\beta}_{ir} \end{aligned} \qquad (15\text{-}39)$$

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

$$\hat{\beta}_{ir} = \hat{\beta} + \hat{\Delta} z_i + \hat{\Lambda} w_{ir}.$$

This can be computed after the estimation of the population parameters. (It may be more efficient to do this computation during the iterations, since 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

$$f(y_{it} \mid x_{it}, \hat{\beta}_{ir}, \hat{\theta}) = \frac{1}{\hat{\sigma}_\varepsilon \sqrt{2\pi}} \exp\left[-\frac{\left(y_{it} - x_{it}'(\hat{\beta} + \hat{\Delta} z_i + \hat{\Lambda} w_{ir})\right)^2}{2\hat{\sigma}_\varepsilon^2} \right]. \qquad (15\text{-}40)$$

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

$$Est.Var[\beta_{i,k} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega] = \begin{matrix} \{Est.E[\beta_{i,k}^2 | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega]\} - \\ \{Est.E[\beta_{i,k} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega]\}^2. \end{matrix} \quad (15\text{-}41)$$

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 β_i.

Some aspects worth noting about this computation are as follows:

- The interval suggested above is a classical (sampling theory based) counterpart to the highest posterior density interval that would be computed for β_i for a hierarchical Bayesian estimator.
- The conditional distribution from which β_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 "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 β_i, the sample distribution of the n estimated conditional means is necessarily normal. Kernel estimators based on the n estimators, for example, can have variety of shapes.
- A common misperception found in the Bayesian and classical literatures alike is that the preceding produces an estimator of β_i. In fact, it is an estimator of conditional mean of the distribution from which β_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 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 $E[\beta_{2i} \mid y_i, X_i]$ for the 48 states in the sample using (15-36) for the fixed effects estimates and (15-39) for the random effects estimates. Figures 15.6 and 15.7 examine the estimated coefficients for private capital. Figure 15.6 displays kernel density estimates for the population distributions based on the fixed and random effects 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. Figure 15.7 shows the results based on the random parameters model, using (15-39) and (15-41) to compute the estimates. As espected, the range of variation of the estimators in the conditional distributions is much smaller than the overall range of variation shown in Figure 15.6.

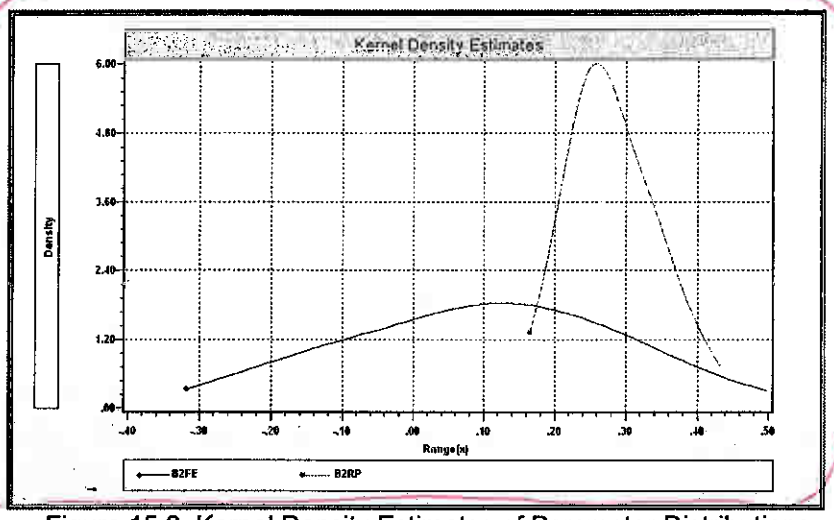

Figure 15.6 Kernel Density Estimates of Parameter Distributions

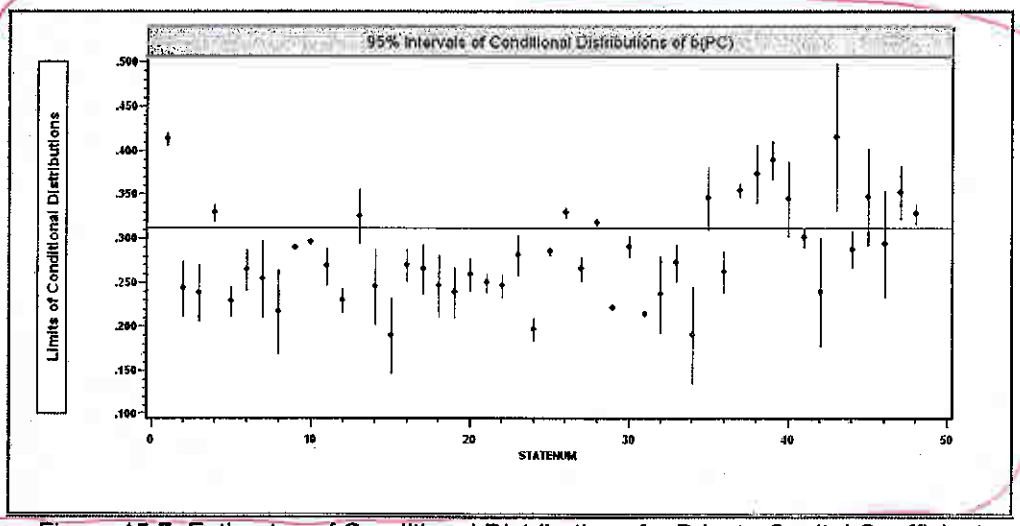

Figure 15.7 Estimates of Conditional Distributions for Private Capital Coefficient

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 and 5 and Appendix Table F3.2 for details on the location of the data.) The variables used in the analysis are

Person id,	
Education,	(time varying),
Log of hourly wage,	(time varying),
Potential experience,	(time varying),
Time trend,	(time varying),
Ability,	(time invariant),
Mother's education,	(time invariant),
Father's education,	(time invariant),
Dummy variable for residence in a broken home,	(time invariant),
Number of siblings,	(time invariant).

This is an unbalanced panel of 2,178 individuals; Figure 15.6 shows a frequency count of the numbers of observations in the sample. We will estimate the following hierarchical wage model

$$\begin{aligned}
\ln \mathit{Wage}_{it} &= \beta_{1,i} + \beta_{2,i}\,\mathit{Education}_{it} + \beta_3\,\mathit{Experience}_{it} + \beta_4\,\mathit{Experience}_{it}^2 \\
&\quad + \beta_5\,\mathit{Broken\ Home}_i + \beta_6\,\mathit{Siblings}_i + \varepsilon_{it}, \\
\beta_{1,i} &= \alpha_{1,1} + \alpha_{1,2}\,\mathit{Ability}_i + \alpha_{1,3}\,\mathit{Mother's\ education}_i + \alpha_{1,4}\,\mathit{Father's\ education}_i + u_{1,i}, \\
\beta_{2,i} &= \alpha_{2,1} + \alpha_{2,2}\,\mathit{Ability}_i + \alpha_{2,3}\,\mathit{Mother's\ education}_i + \alpha_{2,4}\,\mathit{Father's\ education}_i + u_{2,i}.
\end{aligned}$$

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 ($\Gamma_{21} = 0$), then allowing free correlation (Γ_{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 soundly rejected by the likelihood ratio test. The LM statistic [Section 11.5.4 and (11-39)] is 11,709.7, 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(10,840.18 − (−885.674)) = 23,451.71, 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 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 0.999206 and the estimated asymptotic standard error is 0.03934. Following this logic, then, the test statistic is 0.999206 / 0.03934 = 25.397. 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. 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). The simple expedient in this complex situation is to use the LM statistic, which remains consistent with the earlier conclusion.