14-76

AU: Check renumbering of footnotes

where $\mathbf{c}$ is a characteristic vector and λ is the associated characteristic root. The equation implies that $\gamma\mathbf{ii}'\mathbf{c} = (\lambda - 1)\mathbf{c}$. Premultiply by $\mathbf{i}'$ to obtain $\gamma(\mathbf{i}'\mathbf{i})(\mathbf{i}'\mathbf{c}) = (\lambda - 1)(\mathbf{i}'\mathbf{c})$. Any vector $\mathbf{c}$ with elements that sum to zero will satisfy this equality. There will be $T_i - 1$ such vectors and the associated characteristic roots will be $(\lambda - 1) = 0$ or $\lambda = 1$. For the remaining root, divide by the nonzero $(\mathbf{i}'\mathbf{c})$ and note that $\mathbf{i}'\mathbf{i} = T_i$, so the last root is $T_i\gamma = \lambda - 1$ or $\lambda = (1 + T_i\gamma)$.[~~26~~] 24 It follows that the determinant is

FN 24

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

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

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

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

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

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

minus

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

The likelihood equations are

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

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

24

[~~26~~] By this derivation, we have established a useful general result. The characteristic roots of a $T \times T$ matrix of the form $\mathbf{A} = (\mathbf{I} + a\mathbf{bb}')$ are 1 with multiplicity $(T - 1)$ and $a\mathbf{b}'\mathbf{b}$ with multiplicity 1. The proof follows precisely along the lines of our earlier derivation.

14-77

covariance matrix is

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

(See Section ~~9~~11.5.1.) We also know that the MLEs of the variance components estimators will be asymptotically uncorrelated with that of β. In principle, we could continue to estimate the asymptotic variances of the MLEs of σ_ε^2 and σ_u^2. It would be necessary to derive these from the estimators of θ and γ, which one would typically do in any event. However, statistical inference about the disturbance variance, σ_ε^2 in a regression model, is typically of no interest. On the other hand, one might want to test the hypothesis that σ_u^2 equals zero, or $\gamma = 0$. Breusch and Pagan's (1979) LM statistic in (9-39) extended to the unbalanced panel case considered here would be

11

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

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

Example ~~16~~14.11 Maximum Likelihood and FGLS Estimates of a Wage Equation

TB 14.9

Example ~~9~~11.6 presented FGLS estimates of a wage equation using Cornwell and Rupert's panel data. We have reestimated the wage equation using maximum likelihood instead of FGLS. The parameter estimates appear in Table ~~16~~14.9, with the FGLS and pooled OLS estimates. The estimates of the variance components are shown in the table as well. The similarity of the MLEs and FGLS estimates is to be expected given the large sample size. The LM statistic for testing for the presence of the common effects is 3,881.34, which is far larger than the critical value of 3.84. With the MLE, we can also use an LR test to test for random effects against the null hypothesis of no effects. The chi-squared statistic based on the two log-likelihoods is 4297.57, which leads to the same conclusion.

TABLE ~~16~~14.9 Estimates of the Wage Equation

	Pooled Least Squares		*Random Effects MLE*		*Random Effects FGLS*	
Variable	*Estimate*	*Std. Error*[a]	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Exp	0.0361	0.004533	0.1078	0.002480	0.08906	0.002280
Exp^2	−0.0006550	0.0001016	−0.0005054	0.00005452	−0.0007577	0.00005036
Wks	0.004461	0.001728	0.0008663	0.0006031	0.001066	0.0005939
Occ	−0.3176	0.02726	−0.03954	0.01374	−0.1067	0.01269
Ind	0.03213	0.02526	0.008807	0.01531	−0.01637	0.01391
South	−0.1137	0.02868	−0.01615	0.03201	−0.06899	0.02354
SMSA	0.1586	0.02602	−0.04019	0.01901	−0.01530	0.01649
MS	0.3203	0.03494	−0.03540	0.01880	−0.02398	0.01711
Union	0.06975	0.02667	0.03306	0.01482	0.03597	0.01367
Constant	5.8802	0.09673	4.8197	0.06035	5.3455	0.04361
σ_ε^2	0.146119		0.023436 ($\theta = 42.66926$)		0.023102	
σ_u^2	0		0.876517 ($\gamma = 37.40035$)		0.838361	
ln L	−1899.537		2162.938		—	

[a] Robust standard errors

249.25

14-78

14.9.6.b ~~9.7.1~~ NESTED RANDOM EFFECTS~~[24]~~

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

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

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

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

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

AU: OK to spell out "e.g." in text?

Generalized least squares has been well worked out for the balanced case. [See, e.g., Baltagi, Song, and Jung (2001), who also provide results for the three-level unbalanced case.] Define the following to be constructed from the variance components, $\sigma_\varepsilon^2, \sigma_u^2, \sigma_v^2$, and σ_w^2:

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

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

$$\tilde{y}_{ijkt} = y_{ijkt} - \left(1 - \frac{\sigma_\varepsilon}{\sigma_1}\right)\bar{y}_{ijk\cdot} - \left(\frac{\sigma_\varepsilon}{\sigma_1} - \frac{\sigma_\varepsilon}{\sigma_2}\right)\bar{y}_{ij\cdot\cdot} - \left(\frac{\sigma_\varepsilon}{\sigma_2} - \frac{\sigma_\varepsilon}{\sigma_3}\right)\bar{y}_{i\cdots} \qquad \text{~~(9-45)~~}$$

AU: Are ellipses OK as set?

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

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

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

~~[24] This development is based on maximum likelihood estimation, which is presented in Chapter 16.~~

Construct the following intermediate results:

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

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

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

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

The log likelihood is

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

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

$$-\frac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}}\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\mathbf{x}'_{ijkt} - \frac{\rho_w}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\right)\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\right)$$
$$-\frac{\rho_v}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\right)\right)\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\right)\right) \qquad \text{~~(9-46)~~ (14-90)}$$
$$-\frac{\rho_u}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\left(\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\right)\right)\right)\left(\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\left(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\left(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\right)\right)\right).$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is FGLS based on the maximum likelihood estimators of the variance parameters. Thus, expression ~~(9-46)~~ provides the appropriate covariance matrix for the GLS or maximum likelihood estimator. The difference will be in how the variance components are computed. Baltagi et al. (2001) suggest a variety

(14-90)

14-80

of methods for the three-level model. For more than three levels, the MLE becomes more attractive.

Given the complexity of the results, one might prefer simply to use OLS in spite of its inefficiency. As might be expected, the standard errors will be biased owing to the correlation across observations; there is evidence that the bias is downward. [See Moulton (1986).] In that event, the robust estimator in ~~(9-3)~~ 11-~~3~~ 4 would be the natural alternative. In the example given earlier, the nesting structure was obvious. In other cases, such as our application in Example ~~9.10~~ 11.12, that might not be true. In Example ~~9.9~~ ~~11.12~~ 14.12 [and in the application in Baltagi (2005)], statewide observations are grouped into regions based on intuition. The impact of an incorrect grouping is unclear. Both OLS and FGLS would remain consistent—both are equivalent to GLS with the wrong weights, which we considered earlier. However, the impact on the asymptotic covariance matrix for the estimator remains to be analyzed.

Example ~~9.9~~ ~~11.12~~ 14.12 Statewide Productivity

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

← pc

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

where the variables in the model are

gsp	= gross state product,
p_cap	= public capital, = hwy + water + util
hwy	= highway capital,
water	= water utility capital,
util	= utility capital,
pc	= private capital,
emp	= employment (labor),
unemp	= unemployment rate,

and ~~there are~~ we have defined M = 9 regions each consisting of a group of the 48 continental states:

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

For each state, we have 17 years of data, from 1970 to 1986.[~~26~~ 27 25] The two- and three-level random effects models were estimated by maximum likelihood. The two-level model was also fit by FGLS using the methods developed in Section ~~9.5.2.~~ 11.5.3

FN 25

[27 ~~26~~ 25] The data were downloaded from the website for Baltagi (2005) at http://www.wiley.com/legacy/wileychi/baltagi3e/. See Appendix Table F~~9.2~~. 11.5.3

ORL

14.10

TABLE ~~9.6~~ Estimated Statewide Production Function

	OLS		*Fixed Effects*	*Random Effects FGLS*	*Random Effects ML*	*Nested Random Effects*
	Estimate	*Std.Err.*[a]	*Estimate (Std.Err.)*	*Estimate (Std.Err.)*	*Estimate (Std.Err.)*	*Estimate (Std.Err.)*
α	1.9260	0.05250		2.1608	2.1759	2.1348
		(0.2143)		(0.1380)	(0.1477)	(0.1514)
β_1	0.3120	0.01109	0.2350	0.2755	0.2703	0.2724
		(0.04678)	(0.02621)	(0.01972)	(0.02110)	(0.02141)
β_2	0.05888	0.01541	0.07675	0.06167	0.06268	0.06645
		(0.05078)	(0.03124)	(0.02168)	(0.02269)	(0.02287)
β_3	0.1186	0.01236	0.0786	0.07572	0.07545	0.07392
		(0.03450)	(0.0150)	(0.01381)	(0.01397)	(0.01399)
β_4	0.00856	0.01235	−0.11478	−0.09672	−0.1004	−0.1004
		(0.04062)	(0.01814)	(0.01683)	(0.01730)	(0.01698)
β_5	0.5497	0.01554	0.8011	0.7450	0.7542	0.7539
		(0.06770)	(0.02976)	(0.02482)	(0.02664)	(0.02613)
β_6	−0.00727	0.001384	−0.005179	−0.005963	−0.005809	−0.005878
		(0.002946)	(0.000980)	(0.0008814)	(0.0009014)	(0.0009002)
σ_ε	0.985422		0.03676493	0.0367649	0.0366974	0.0366964
σ_u	↑0			0.0771064	0.0875682	0.0791243
σ_v						0.0386299
ln *L*	853.1372		1565.501		1429.075	1430.30576

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

TB 14.10

Table 14.10 presents the estimates of the production function using pooled OLS, OLS for the fixed effects model and both FGLS and maximum likelihood for the random effects models. Overall, the estimates are similar, though the OLS estimates do stand somewhat apart. This suggests, as one might suspect, that there are omitted effects in the pooled model. The F statistic for testing the significance of the fixed effects is 76.712 with 47 and 762 degrees of freedom. The critical value from the table is 1.379, so on this basis, one would reject the hypothesis of no common effects. Note, as well, the extremely large differences between the conventional OLS standard errors and the robust (cluster) corrected values. The three or four fold differences strongly suggest that there are latent effects at least at the state level. It remains to consider which approach, fixed or random effects is preferred. The Hausman test for fixed vs. random effects produces a chi-squared value of 18.987. The critical value is 12.592. This would imply that the fixed effects model would be the preferred specification. When we repeat the calculation of the Hausman statistic using the three-level estimates in the last column of Table ~~9.6~~, the statistic falls slightly to 15.327. Finally, note the similarity of all three sets of random effects estimates. In fact, under the hypothesis of mean independence, all three are consistent estimators. It is tempting at this point to carry out a likelihood ratio test of the hypothesis of the two-level model against the broader alternative three-level model. The test statistic would be twice the difference of the log likelihoods, which is 2.46. For one degree of freedom, the critical chi-squared with one degree of freedom is 3.84, so on this basis, we would not reject the hypothesis of the two-level model. We note, however, that there is a problem with this testing procedure. The hypothesis that a variance is zero is not well defined for the likelihood ratio test—the parameter under the null hypothesis is on the boundary of the parameter space ($\sigma_v^2 \geq 0$). In this instance, the familiar distribution theory does not apply. ~~We will revisit this issue in Chapter 16 in our study of the method of maximum likelihood~~. 11.9

14-82

4297.54

~~random effects against the null hypothesis of no effects. The chi-squared statistic based on the two log-likelihoods is 8,124.949, which leads to the same conclusion.~~

14 ~~16~~.9.6.~~b~~ c Random Effects in Nonlinear Models: MLE using Quadrature

Section 1~~6~~ 14 .9.5.b describes a nonlinear model for panel data, the geometric regression model,

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

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

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

which implies that

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

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

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

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

KT

$$\text{Prob}[Y_{it} = y_{it}\,|\,\mathbf{x}_{it}] = \theta_{it}(1-\theta_{it})^{y_{it}},\ y_{it} = 0, 1, \ldots; i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$\theta_{it} = 1/(1+\lambda_{it}),\ \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i),$$
$$f(u_i) = \text{the specification of the distribution of random effects over individuals.} \qquad \text{~~(16-90)~~}$$

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

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

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

Two questions now arise:

- How does one obtain maximum likelihood estimates of the parameters of the model? We will pursue that question now.

14-83

- If we ignore the individual heterogeneity and simply estimate the pooled model, will we obtain consistent estimators of the model parameters? The answer is sometimes, but usually not. The favorable cases are the simple loglinear models such as the geometric and Poisson models that we consider in this chapter. The unfavorable cases are most of the other common applications in the literature, including, notably, models for binary choice, censored regressions, sample selection, and, generally, nonlinear models that do not have simple exponential means. [Note that this is the crucial issue in the consideration of robust covariance matrix estimation in Sections ~~16~~14.8.3 and ~~16~~14.8.4. See, as well, Freedman (2006).] ✓

We will now develop a maximum likelihood estimator for a nonlinear random effects model. To set up the methodology for applications later in the book, we will do this in a generic specification, then return to the specific application of the geometric regression model in Example ~~16~~.12. Assume, then, that the panel data model defines 14 the probability distribution of a random variable, y_{it}, conditioned on a data vector, $\mathbf{x}_{it}$, and an unobserved common random effect, u_i. As always, there are T_i observations in the group, and the data on $\mathbf{x}_{it}$ and now u_i are assumed to be strictly exogenously determined. Our model for one individual is, then,

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

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

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

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

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

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

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

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

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

14-84

There are two practical problems to be solved to implement this estimator. First, it will be rare that the integral will exist in closed form. (It does when the density of y_{it} is normal with linear conditional mean and the random effect is normal, because, as we have seen, this is the random effects linear model.) As such, the practical complication that arises is how the integrals are to be computed. Second, it remains to specify the distribution of u_i over which the integration is taken. The distribution of the common effect is part of the model specification. Several approaches for this model have now appeared in the literature. The one we will develop here extends the random effects model with normally distributed effects that we have analyzed in the previous section. The technique is **Butler and Moffitt's** (1982) **method.** It was originally proposed for extending the random effects model to a binary choice setting (see Chapter ~~23~~ 17), but, as we shall see presently, it is straightforward to extend it to a wide range of other models. The computations center on a technique for approximating integrals known as **Gauss–Hermite quadrature.**

AU: Term in chap. list is "Butler and Moffitt's model"?

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

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

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

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

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

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

For the moment, let

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

Then, the function we are manipulating is

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

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

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

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

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

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

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

The Hermite approximation to the log-likelihood function is

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

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

Example ~~16.12~~ 14.13 Random Effects Geometric Regression Model

We will use the preceding to construct a random effects model for the *DocVis* count variable analyzed in Example ~~16~~ 14.10. Using (~~16~~ 14-90), the approximate log-likelihood function will be

$$\ln L_H = \frac{1}{\sqrt{\pi}} \sum_{i=1}^{n} \ln \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} \theta_{it}(1-\theta_{it})^{y_{it}} \right],$$
$$\theta_{it} = 1/(1+\lambda_{it}), \ \lambda_{it} = \exp(\mathbf{x}_{it}'\beta + \phi v_h).$$

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

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

It remains only to specialize this to our geometric regression model. For this case, the density is given earlier. The missing components of the preceding derivatives are the partial derivatives with respect to β and ϕ that were obtained in Section ~~16~~.9.5.b. The necessary result is 14

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

Maximum likelihood estimates of the parameters of the random effects geometric regression model are given in Example ~~16~~ 14.13 with the fixed effects estimates for this model.

14 ~~16~~.9.6.~~c~~ d Fixed Effects in Nonlinear Models: Full MLE

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

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

where d_{it} is a dummy variable that takes the value one in every period for individual i and zero otherwise. (In more involved models, such as the censored regression model we examine in Chapter ~~24~~ 18, there might be other parameters, such as a variance. For now, it is convenient to omit them—the development can be extended to add them later.) For convenience, we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model.~~[27]~~ [28] 26 FN 26 The parameters to be estimated are the K elements of β and the n individual constant terms. The log-likelihood function for the fixed effects model is

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

where $f(.)$ is the probability density function of the observed outcome, for example, the geometric regression model that we used in our previous example. It will be convenient to let $z_{it} = \alpha_i + \mathbf{x}_{it}'\beta$ so that $p(y_{it}\,|\,d_{it}, \mathbf{x}_{it}) = f(y_{it}\,|\,z_{it})$.

In the fixed effects linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means that eliminated the person-specific constants from the equation. (See Section ~~9.4.1~~ 11.4.1.) In a few cases of nonlinear models, it is also possible to eliminate the fixed effects from the likelihood function, although in general not by taking deviations from means. One example is the **exponential regression model** that is used for lifetimes of electronic components and electrical equipment such as light bulbs:

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

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

~~[27]~~ [28] 26 In estimating a fixed effects linear regression model in Section ~~9.4~~ 11.4 we found that it was not possible to analyze models with time-invariant variables. The same limitation applies in the nonlinear case, for essentially the same reasons. The time-invariant effects are absorbed in the constant term. In estimation, the columns of the data matrix with time-invariant variables will be transformed to columns of zeros when we compute derivatives of the log-likelihood function.

log-likelihood is

$$\ln L = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln \theta_{it} - \theta_{it} y_{it}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln(\gamma_i \Delta_{it}) - (\gamma_i \Delta_{it}) y_{it}. \qquad 14\ (~~16~~\text{-}91)$$

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

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

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

$$\gamma_i = \frac{T_i}{\sum_{t=1}^{T_i} \Delta_{it} y_{it}}. \quad \backslash\backslash \qquad 14\ (~~16~~\text{-}92)$$

✓ [Note the analogous result for the linear model in (~~9~~-15).] Inserting this solution back in the log-likelihood function in (~~16~~-91), we obtain the concentrated log-likelihood,

14

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

which is now only a function of $\boldsymbol{\beta}$. This function can now be maximized with respect to $\boldsymbol{\beta}$ alone. The MLEs for α_i are then found as the logs of the results of 14 (~~16~~-91). Note, once again, we have eliminated the constants from the estimation problem, but not by computing deviations from group means. That is specific to the linear model.

✓ The concentrated log-likelihood is only obtainable in only a small handful of cases, including the linear model, the exponential model (as just shown), the Poisson regression model, and a few others. Lancaster (2000) lists some of these and discusses the underlying methodological issues. In most cases, if one desires to estimate the parameters of a fixed effects model, it will be necessary to actually compute the possibly huge number of constant terms, α_i, at the same time as the main parameters, $\boldsymbol{\beta}$. This has widely been viewed as a practical obstacle to estimation of this model because of the need to ✓ invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The likelihood equations for the fixed effects model are

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

and

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

14-88

The second derivatives matrix is

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

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

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

where $\mathbf{H}_{\beta\beta'}$ is a negative definite matrix. The likelihood equations are a large system, but the solution turns out to be surprisingly straightforward. [See Greene (2001).]

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

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

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

Note the striking similarity to the result we had in (9-18) for the fixed effects model in the linear case. [A similar result is noted briefly in Chamberlain (1984).] By assembling the Hessian as a partitioned matrix for $\boldsymbol{\beta}$ and the full vector of constant terms, then using (A-66b) and the preceding definitions to isolate one diagonal element, we find

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

Once again, the result has the same format as its counterpart in the linear model. [See (9-18).] In principle, the negatives of these would be the estimators of the asymptotic variances of the maximum likelihood estimators. (Asymptotic properties in this model are problematic, as we consider shortly.) 11

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

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

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

14-89

It is shown in Greene (2001) in spite of the potentially large number of parameters in the model, Newton's method can be used with the following iteration, which uses only the $K \times K$ matrix computed earlier and a few $K \times 1$ vectors:

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

and

FN 27

$$\hat{\alpha}_l^{(s+1)} = \hat{\alpha}_l^{(s)} - [(g_{ii}/h_{ii}) + \overline{\mathbf{x}}_i' \Delta_{\beta}^{(s)}].$$ [~~26~~] 27 29

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

In addition to the theoretical virtues and shortcomings of this model, we note the practical aspect of estimation of what are possibly a huge number of parameters, $n + K$. In the fixed effects case, n is not limited, and could be in the thousands in a typical application. [In Example ~~16~~.13, n is 7,293. As of this writing, the largest application of the method described here that we are aware of is Kingdon and Cassen's (2007) study in which they fit a fixed effects probit model with well over 140,000 dummy variable coefficients.] The problems with the fixed effects estimator are statistical, not practical.[~~28~~] 28 30 14

FN 28

The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. In this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). There is, as well, a small sample (small T_i) bias in the slope estimators. This is the **incidental parameters problem.** [See Neyman and Scott (1948) and Lancaster (2000).] We will examine the incidental parameters problem in a bit more detail with a Monte Carlo study in Section ~~17.4~~ 15.3. also

Example ~~16.18~~ 14.14 Fixed and Random Effects Geometric Regression

Example ~~16~~.10 presents pooled estimates for the geometric regression model 14

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

TB 14.11

We will now reestimate the model under the assumptions of the random and fixed effects specifications. The methods of the preceding two sections are applied directly—no modification of the procedures was required. Table ~~16.10~~ 14.11 presents the three sets of maximum likelihood estimates. The estimates vary considerably. The average group size is about five. This implies that the fixed effects estimator may well be subject to a small sample bias. Save for the coefficient on *Kids*, the fixed effects and random effects estimates are quite similar. On the other hand, the two panel models give similar results to the pooled model except for the *Income* coefficient. On this basis, it is difficult to see, based solely on the results, which should be the preferred model. The model is nonlinear to begin with, so the pooled model, which might otherwise be preferred on the basis of computational ease, now has no

27 29 [~~26~~] Similar results appear in Prentice and Gloeckler (1978) who attribute it to Rao (1973) and Chamberlain (1980, 1984).

28 30 [~~28~~] See Vytlacil, Aakvik, and Heckman (2005), Chamberlain (1980, 1984), Newey (1994), Bover and Arellano (1997), and Chen (1998) for some extensions of parametric and semiparametric forms of the binary choice models with fixed effects.

14-90

TABLE ~~16.10~~ 14.11 Panel Data Estimates of a Geometric Regression for DOCVIS

	Pooled		Random Effects[a]		Fixed Effects	
Variable	*Estimate*	*St. Er.*	*Estimate*	*St. Er.*	*Estimate*	*St. Er.*
Constant	1.0918	0.1112	0.3998	0.09531		
Age	0.0180	0.0013	0.02208	0.001220	0.04845	0.003511
Education	−0.0473	0.0069	−0.04507	0.006262	−0.05437	0.03721
Income	−0.0468	0.0075	−0.1959	0.06103	−0.1892	0.09127
Kids	−0.1569	0.0319	−0.1242	0.02336	−0.002543	0.03687

[a] Estimated $\sigma_u = 0.9542921$.

redeeming virtues. None of the three models is robust to misspecification. Unlike the linear model, in this and other nonlinear models, the fixed effects estimator is inconsistent when T is small in both random and fixed effects models. The random effects estimator is consistent in the random effects model, but, as usual, not in the fixed effects model. The pooled estimator is inconsistent in both random and fixed effects cases (which calls into question the virtue of the robust covariance matrix). It might be tempting to use a Hausman specification 11.5.5 test (see Section ~~9.5.4~~); however, the conditions that underlie the test are not met—unlike the linear model where the fixed effects is consistent in both cases, here it is inconsistent in both cases. For better or worse, that leaves the analyst with the need to choose the model based on the underlying theory.

14.10 ~~16.7~~ 14 LATENT CLASS AND FINITE MIXTURE MODELS

The latent class model specifies that the distribution of the observed data is a mixture of a finite number of underlying distributions. The model can be motivated in several ways:

- In the classic application of the technique, the observed data are drawn from a mix of distinct underlying populations. Consider, for example, a historical or fossilized record of the intersection (or collision) of two populations. The anthropological record consists of measurements on some variable that would differ distinctly between the populations. However, the analyst has no definitive marker for which subpopulation an observation is drawn from. Given a sample of observations, they are interested in two statistical problems: (1) estimate the parameters of the underlying populations and (2) classify observations in hand as having originated in which population. In another contemporary application, Lambert (1992) studied the number of defective outcomes in a production process. When a "zero defectives" condition is observed, it could indicate either regime 1, "the process is under control," or regime 2, "the process is not under control but just happens to produce a zero observation."
- In a narrower sense, one might view parameter heterogeneity in a population as a form of discrete mixing. We have modeled parameter heterogeneity using continuous distributions in Chapter ~~9~~ 11. The "finite mixture" approach takes the distribution of parameters across individuals to be discrete. (Of course, this is another way to interpret the first point.)
- The finite mixing approach is a means by which a distribution (model) can be constructed from a mixture of underlying distributions. Goldfeld and Quandt's mixture of normals model in Example ~~15~~.4 is a case in which a nonnormal distribution is created by mixing two normal distributions with different parameters.

13

14.10 LATENT CLASS AND FINITE MIXTURE MODELS

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

- In the classic application of the technique, the observed data are drawn from a mix of distinct underlying populations. Consider, for example, a historical or fossilized record of the intersection (or collision) of two populations. The anthropological record consists of measurements on some variable that would differ imperfectly, but substantively, between the populations. However, the analyst has no definitive marker for which subpopulation an observation is drawn from. Given a sample of observations, they are interested in two statistical problems: (1) estimate the parameters of the underlying populations and (2) classify the observations in hand as having originated in which population. The technique has seen a number of recent applications in health econometrics. For example, in a study of obesity, Greene, Harris, Hollingsworth and Maitra (2008) speculated that their ordered choice model (see Chapter 17) might systematically vary in a sample that contained (it was believed) some individuals who have a genetic predisposition toward obesity and most that did not. In another contemporary application, Lambert (1992) studied the number of defective outcomes in a production process. When a "zero defectives" condition is observed, it could indicate either regime 1, "the process is under control," or regime 2, "the process is not under control but just happens to produce a zero observation."

- In a narrower sense, one might view parameter heterogeneity in a population as a form of discrete mixing. We have modeled parameter heterogeneity using continuous distributions in Chapter 11 and 15. The "finite mixture" approach takes the distribution of parameters across individuals to be discrete. (Of course, this is another way to interpret the first point.)

- The finite mixing approach is a means by which a distribution (model) can be constructed from a mixture of underlying distributions. Goldfeld and Quandt's mixture of normals model in Example 13.4 is a case in which a nonnormal distribution is created by mixing two normal distributions with different parameters.

14.10.1 ~~16.9.7.a~~ 14 A Finite Mixture Model

To lay the foundation for the more fully developed model that follows, we revisit the mixture of normals model from Example ~~15~~13.4. Consider a population that consists of a latent mixture of two underlying normal distributions. Neglecting for the moment that it is unknown which applies to a given individual, we have, for individual i,

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

and (~~16~~14-93)

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

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

$$\begin{aligned} f(y_i) &= \lambda f(y_i \mid class_i = 1) + (1 - \lambda) f(y_i \mid class_i = 2) \\ &= E_{classes}\, f(y_i \mid class_i). \end{aligned} \qquad (\text{~~16~~}14\text{-}94)$$

The parameters to be estimated are λ, μ_1, μ_2, σ_1, and σ_2. Combining terms, the log-likelihood for a sample of n individual observations would be

$$\ln L = \sum_{i=1}^{n} \ln\left(\frac{\lambda \exp\left[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} + \frac{(1 - \lambda)\exp\left[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2\right]}{\sigma_2\sqrt{2\pi}}\right). \qquad (\text{~~16~~}14\text{-}95)$$

This is the mixture density that we saw in Example ~~15~~13.4. We suggested the method of moments as an estimator of the five parameters in that example. However, this appears to be a straightforward problem in maximum likelihood estimation.

Example ~~16.14~~ 14.15 Latent Class Model for Grade Point Averages

Appendix Table F~~16~~14.1 contains a data set of 32 observations used by Spector and Mazzeo (1980) to study whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses. Variables in the data set include

GPA_i = the student's grade point average,
$GRADE_i$ = dummy variable for whether the student's grade in intermediate macroeconomics was higher than in the principles course,
PSI_i = dummy variable for whether the individual participated in the PSI,
$TUCE_i$ = the student's score on a pretest in economics.

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

We begin by computing maximum likelihood estimates of the parameters in (~~16~~14-95). To estimate the parameters using an iterative method, it is necessary to devise a set of starting values. It ~~is~~ might seem natural to use the simple values from a one-class model, $\bar{y}$ and s_y, and a value such as 1/2 for λ. However, the optimizer will immediately stop on these values, as the derivatives will be zero at this point. Rather, it is common to use some value near these—perturbing them slightly (a few percent), just to get the iterations started. Table ~~16.11~~ 14.12 contains the estimates for this two-class finite mixture model. The estimates for the one-class model are the sample mean and standard deviations of GPA. [Because these are the MLEs,

TB 14.12

14-93

14.12

TABLE 1~~6~~.11 Estimated Normal Mixture Model

	One Class		Latent Class 1		Latent Class 2	
Parameter	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
μ	3.1172	0.08251	3.64187	0.3452	2.8894	0.2514
σ	0.4594	0.04070	0.2524	0.2625	0.3218	0.1095
Probability	1.0000	0.0000	0.3028	0.3497	0.6972	0.3497
Rom *ln L*	−20.51274		−19.63654			

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(GPA_i - \overline{GPA})^2$.] The means and standard deviations of the two classes are noticeably different—the model appears to be revealing a distinct splitting of the data into two classes. (Whether two is the appropriate number of classes is considered in Section ~~16~~.9.7.e). 14 It is tempting at this point to identify the two classes with some other covariate, either in the data set or not, such as *PSI*. However, at this point, there is no basis for doing so—the classes are "latent." As the analysis continues, however, we will want to investigate whether any observed data help to predict the class membership.

14.10.2 14 ~~16.9.7.b~~ Measured and Unmeasured Heterogeneity

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

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

where $\mathbf{z}_i$ is the vector of variables that help to explain the class probabilities. To make the mixture model amenable to estimation, it is necessary to parameterize the probabilities. The logit probability model is a common device. (See Section ~~23.4~~. 17.4 For applications, see Greene (2007d, Section 2.3.3) and references cited.) For the two-class case, this might appear as follows:

$$\text{Prob}(class = 1 \mid \mathbf{z}_i) = \frac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})},\ \text{Prob}(class = 2 \mid \mathbf{z}_i) = 1 - \text{Prob}(class = 1 \mid \mathbf{z}_i). \quad \text{14 } (~~16~~\text{-96})$$

(The more general J class case is shown in Section ~~16.9.7.f~~.) 14.10.6 The log-likelihood for our mixture of two normals example becomes

$$\begin{aligned} \ln L &= \sum_{i=1}^{n}\ln L_i \\ &= \sum_{i=1}^{n}\ln\left(\begin{array}{l}\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1+\exp(\mathbf{z}_i'\boldsymbol{\theta})}\right)\dfrac{\exp\left[-\frac{1}{2}(y_i-\mu_1)^2/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} \\ + \left(\dfrac{1}{1+\exp(\mathbf{z}_i'\boldsymbol{\theta})}\right)\dfrac{\exp\left[-\frac{1}{2}(y_i-\mu_2)^2/\sigma_2^2\right]}{\sigma_2\sqrt{2\pi}}\end{array}\right). \end{aligned} \quad \text{14 } (~~16~~\text{-97})$$

14-94

The log-likelihood is now maximized with respect to $\mu_1, \sigma_1, \mu_2, \sigma_2$, and $\boldsymbol{\theta}$. If $\mathbf{z}_i$ contains a constant term and some other observed variables, then the earlier model returns if the coefficients on those other variables all equal zero. In this case, it follows that $\lambda = \ln[\theta/(1-\theta)]$. (This device is usually used to ensure that $0 < \lambda < 1$ in the earlier model.)

14.10.3 ~~16.9.7.c~~ Predicting Class Membership

14c

The model in (16-97) now characterizes two random variables, y_i, the outcome variable of interest, and $class_i$, the indicator of which class the individual resides in. We have a joint distribution, $f(y_i, class_i)$, which we are modeling in terms of the conditional density, $f(y_i \mid class_i)$ in (~~16~~-93), and the marginal density of $class_i$ in (~~16~~-96). We have initially assumed the latter to be a simple Bernoulli distribution with $\text{Prob}(class_i = 1) = \lambda$, but then modified in the previous section to equal $\text{Prob}(class_i = 1 \mid \mathbf{z}_i) = \Lambda(\mathbf{z}_i'\boldsymbol{\theta})$. These can be viewed as the "prior" probabilities in a Bayesian sense. If we wish to make a prediction as to which class the individual came from, using all the information that we have on that individual, then the prior probability is going to waste some information. The "posterior," or conditional (on the remaining data) probability,

14 14

$$\text{Prob}(class_i = 1 \mid \mathbf{z}_i\, y_i) = \frac{f(y_i, class = 1 \mid \mathbf{z}_i)}{f(y_i)}, \qquad (\cancel{16}\text{-98})$$

14

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

KT

Bayes's

$$\text{Prob}(class_i = 1 \mid \mathbf{z}_i, y_i) = \frac{f(y_i \mid class_i = 1, \mathbf{z}_i)\text{Prob}(class_i = 1 \mid \mathbf{z}_i)}{f(y_i \mid class_i = 1, \mathbf{z}_i)\text{Prob}(class_i = 1 \mid \mathbf{z}_i) + f(y_i \mid class_i = 2, \mathbf{z}_i)\text{Prob}(class_i = 2 \mid \mathbf{z}_i)}. \qquad (\cancel{16}\text{-99})$$

14

The denominator is L_i (not $\ln L_i$) from (~~16~~ 14-97). The numerator is the first term in L_i. To continue our mixture of two normals example, the conditional (posterior) probability is

14

$$\text{Prob}(class_i = 1 \mid \mathbf{z}_i, y_i) = \frac{\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1+\exp(\mathbf{z}_i'\boldsymbol{\theta})}\right)\dfrac{\exp\left[-\frac{1}{2}(y_i-\mu_1)^2/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}}}{L_i}, \qquad (\cancel{16}\text{-100})$$

14

while the unconditional probability is in (~~16~~ 14-96). The conditional probability for the second class is computed using the other two marginal densities in the numerator (or by subtraction from one). Note that the conditional probabilities are functions of the data even if the unconditional ones are not. To come to the problem suggested at the outset, then, the natural predictor of $class_i$ is the class associated with the largest estimated posterior probability.

14.10.4 ~~16.9.7.d~~ A Conditional Latent Class Model

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

14-95

conditional mean:

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

In the data of Example 14 ~~16~~.14, we also observe an indicator of whether the individual has participated in a special program designed to enhance the economics program (PSI). We might modify the model,

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

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

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

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

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

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

The methodology can be applied to any model for y_i. In the example in Section ~~16.9.7.f~~, 16.10.6 we will model a binary dependent variable with a probit model. The methodology has been applied in many other settings, such as stochastic frontier models [Orea and ✓✓✓ Kumbhakar (2004), Greene (2004)], Poisson regression models [Wedel et al. (1993)], and a wide variety of count, discrete choice, and limited dependent variable models ✓✓ [McLachlan and Peel (2000), Greene (2007b)].

Example ~~16.15~~ 14.16 Latent Class Regression Model for Grade Point Averages

Combining ~~16.9.7.b~~ 14.10.2 and ~~16.9.7.d,~~ 14.10.4 we have a latent class model for grade point averages,

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

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

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

The log-likelihood is now

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

TB 14.13 Maximum likelihood estimates of the parameters are given in Table ~~16.12~~ 14.13.

TB 14.14 14.14 Table ~~16.13~~ lists the observations sorted by GPA. The predictions of class membership reflect what one might guess from the coefficients in the table of coefficients. Class 2 members on average have lower GPAs than in class 1. The listing in Table ~~16~~.13 14.14 shows this clustering. It also suggests how the latent class model is using the sample information. If the results in

14-96

TABLE ~~16.12~~ 14.13 Estimated Latent Class Linear Regression Model for GPA

	One Class		*Latent Class 1*		*Latent Class 2*	
Parameter	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
β_1	3.1011	0.1117	3.3928	0.1733	2.7926	0.04988
β_2	0.03675	0.1689	−0.1074	0.2006	−0.5703	0.07553
$\sigma = e'e/n$	0.4443	0.0003086	0.3812	0.09337	0.1119	0.04487
θ_1	0.0000	0.0000	−6.8392	3.07867	0.0000	0.0000
θ_2	0.0000	0.0000	0.3518	0.1601	0.0000	0.0000
Prob | $\overline{TUCE}$	1.0000		0.7063		0.2937	
ln *L*	−20.48752		−13.39966			

TABLE ~~16.13~~ 14.14 Estimated Latent Class Probabilities

GPA	*TUCE*	*PSI*	*CLASS*	*P1*	*P1**	*P2*	*P2**
2.06	22	1	2	0.7109	0.0116	0.2891	0.9884
2.39	19	1	2	0.4612	0.0467	0.5388	0.9533
2.63	20	0	2	0.5489	0.1217	0.4511	0.8783
2.66	20	0	2	0.5489	0.1020	0.4511	0.8980
2.67	24	1	1	0.8325	0.9992	0.1675	0.0008
2.74	19	0	2	0.4612	0.0608	0.5388	0.9392
2.75	25	0	2	0.8760	0.3499	0.1240	0.6501
2.76	17	0	2	0.2975	0.0317	0.7025	0.9683
2.83	19	0	2	0.4612	0.0821	0.5388	0.9179
2.83	27	1	1	0.9345	1.0000	0.0655	0.0000
2.86	17	0	2	0.2975	0.0532	0.7025	0.9468
2.87	21	0	2	0.6336	0.2013	0.3664	0.7987
2.89	14	1	1	0.1285	1.0000	0.8715	0.0000
2.89	22	0	2	0.7109	0.3065	0.2891	0.6935
2.92	12	0	2	0.0680	0.0186	0.9320	0.9814
3.03	25	0	1	0.8760	0.9260	0.1240	0.0740
3.10	21	1	1	0.6336	1.0000	0.3664	0.0000
3.12	23	1	1	0.7775	1.0000	0.2225	0.0000
3.16	25	1	1	0.8760	1.0000	0.1240	0.0000
3.26	25	0	1	0.8760	0.9999	0.1240	0.0001
3.28	24	0	1	0.8325	0.9999	0.1675	0.0001
3.32	23	0	1	0.7775	1.0000	0.2225	0.0000
3.39	17	1	1	0.2975	1.0000	0.7025	0.0000
3.51	26	1	1	0.9094	1.0000	0.0906	0.0000
3.53	26	0	1	0.9094	1.0000	0.0906	0.0000
3.54	24	1	1	0.8325	1.0000	0.1675	0.0000
3.57	23	0	1	0.7775	1.0000	0.2225	0.0000
3.62	28	1	1	0.9530	1.0000	0.0470	0.0000
3.65	21	1	1	0.6336	1.0000	0.3664	0.0000
3.92	29	0	1	0.9665	1.0000	0.0335	0.0000
4.00	21	0	1	0.6336	1.0000	0.3664	0.0000
4.00	23	1	1	0.7775	1.0000	0.2225	0.0000

Table ~~16.11~~ 14.12—just estimating the means, constant class probabilities—are used to produce the same table, when sorted, the highest 10 GPAs are in class 1 and the remainder are in class 2. The more elaborate model is adding information on *TUCE* to the computation. A low *TUCE* score can push a high GPA individual into class 2. (Of course, this is largely what multiple linear regression does as well).

14-97

14.10.5 14 ~~16.9.7.6~~ Determining the Number of Classes

There is an unsolved inference issue remaining in the specification of the model. The number of classes has been taken as a known parameter—two in our main example thus far, three in the following application. Ideally, one would like to determine the appropriate number of classes statistically. However, J is not a parameter in the model. A likelihood ratio test, for example, will not provide a valid result. Consider the original model in Example ~~16~~.14. The model has two classes and five parameters in total. It would
14 seem natural to test down to a one-class model that contains only the mean and variance using the LR test. However, the number of restrictions here is actually ambiguous. If $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, then the mixing probability is irrelevant—the two class densities are the same, and it is a one-class model. Thus, the number of restrictions needed to get from the two-class model to the one-class model is ambiguous. It is neither two nor three. One strategy that has been suggested is to test upward, adding classes until the marginal class insignificantly changes the log-likelihood or one of the information 14
criteria such as the AIC or BIC (see Section ~~16~~.6.5). Unfortunately, this approach is likewise problematic because the estimates from any specification that is too short are inconsistent. The alternative would be to test down from a specification known to be too large. Heckman and Singer (1984b) discuss this possibility and note that when the number of classes becomes larger than appropriate, the estimator should break down. In our Example ~~16~~.14, if we expand to four classes, the optimizer breaks down, and it is no
14 longer possible to compute the estimates. A five-class model does produce estimates, but some are nonsensical. This does provide at least the directions to seek a viable strategy. The authoritative treatise on finite mixture models by McLachlan and Peel (2000, Chapter 6) contains extensive discussion of this issue.

14.10.6 14 ~~16.9.7.7~~ A Panel Data Application

The latent class model is a useful framework for applications in panel data. The class probabilities partly play the role of common random effects, as we will now explore. The latent class model can be interpreted as a random parameters model, as suggested in Section ~~9~~.8.2, with a discrete distribution of the parameters.

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

$$f(y_{it}\,|\,\mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\Delta}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J) = \sum_{j=1}^{J} p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta})\, f(y_{it}\,|\,class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j),$$

where it remains to parameterize the class probabilities, p_{ij}, and the structural model, $f(y_{it}\,|\,class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)$. The parameter matrix, $\boldsymbol{\Delta}$, contains the parameters of the discrete probability distribution. It has J rows, one for each class, and M columns, for the M variables in $\mathbf{z}_i$. At a minimum, $M = 1$ and $\mathbf{z}_i$ contains a constant term if the class probabilities are fixed parameters as in Example ~~16.14~~. Finally, to accommodate the panel data nature of the sampling situation, we suppose that conditioned on $\boldsymbol{\beta}_j$, that is, on membership in class j, which is fixed over time, the observations on y_{it} are

14.15

14-98

independent. Therefore, for a group of T_i observations, the joint density is

$$f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} \mid class = j, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) = \prod_{t=1}^{T_i} f(y_{it} \mid class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j).$$

The log-likelihood function for a panel of data is

$$\ln L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{J} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i) \prod_{t=1}^{T_i} f(y_{it} \mid class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j) \right].$$

The class probabilities must be constrained to sum to 1. The approach that is usually used is to reparameterize them as a set of logit probabilities, as we did in the preceding examples. Then,

$$p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta}) = \frac{\exp(\theta_{ij})}{\sum_{j=1}^{J} \exp(\theta_{ij})}, J = 1, \ldots, J, \theta_{ij} = \mathbf{z}_i'\boldsymbol{\delta}_j, \theta_{iJ} = 0 \, (\boldsymbol{\delta}_J = \mathbf{0}). \tag{$\overset{14}{\cancel{16}}$-101}$$

(See Section ~~23~~17.11 for development of this model for the set of probabilities.) Note the restriction on θ_{ij}. This is an identification restriction. Without it, the same set of probabilities will arise if an arbitrary vector is added to every $\boldsymbol{\delta}_j$. The resulting log likelihood is a continuous function of the parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ and $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_J$. For all its apparent complexity, estimation of this model by direct maximization of the log-likelihood is not especially difficult. [See Section E.3 and Greene (2001, 2007b). The EM algorithm discussed in Section E.3.7 is especially well suited for estimating the parameters of latent class models. See McLachlan and Peel (2000).] The number of classes that can be identified is likely to be relatively small (on the order of 5 or 10 at most), however, which has been viewed as a drawback of the approach. In general, the more complex the model for y_{it}, the more difficult it becomes to expand the number of classes. Also, as might be expected, the less rich the data set in terms of cross-group variation, the more difficult it is to estimate latent class models.

Estimation produces values for the structural parameters, $(\boldsymbol{\beta}_j, \boldsymbol{\delta}_j)$, $j = 1, \ldots, J$. With these in hand, we can compute the prior class probabilities, p_{ij} using (~~16~~14-101). For prediction purposes, we are also interested in the posterior (on the data) class probabilities, which we can compute using Bayes theorem [see (~~16~~14-99)]. The conditional probability is

$$\begin{aligned} &\text{Prob}(class = j \mid \text{observation } i) \\ &\quad = \frac{f(\text{observation } i \mid \text{class} = j)\text{Prob}(\text{class } j)}{\sum_{j=1}^{J} f(\text{observation } i \mid \text{class} = j)\text{Prob}(\text{class } j)} \\ &\quad = \frac{f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta})}{\sum_{j=1}^{J} f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta})} \\ &\quad = w_{ij}. \end{aligned} \tag{$\overset{14}{\cancel{16}}$-102}$$

The set of probabilities, $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iJ})$ gives the posterior density over the distribution of values of $\boldsymbol{\beta}$, that is, $[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J]$.

14-99

TABLE ~~16.14~~ 14.15 Panel Data Estimates of a Geometric Regression for DocVis

	Pooled MLE (Nonlinear Least Squares)		*Random Effects*[a]		*Fixed Effects*	
Variable	*Estimate*	*St. Er*	*Estimate*	*St. Er.*	*Estimate*	*St. Er.*
Constant	1.0918 (0.9801)	0.1082 (0.1813)	0.3998	0.09531		
Age	0.0180 (0.01873)	0.0013 (0.00198)	0.02208	0.001220	0.04845	0.003511
Education	−0.0473 (−0.03613)	0.0067 (0.01228)	−0.04507	0.006262	−0.05437	0.03721
Income	−0.4687 (−0.5911)	0.0726 (0.1282)	−0.1959	0.06103	−0.1982	0.09127
Kids	−0.1569 (−0.1692)	0.0306 (0.04882)	−0.1242	0.02336	−0.002543	0.03687

[a] Estimated $\sigma_u = 0.9542921$.

Example ~~16.16~~ 14.17 Latent Class Model for Health Care Utilization

In Example ~~11.10~~ 11.13, we proposed an exponential regression model,

$$y_{it} = DocVis_{it} = \exp(\mathbf{x}_{it}'\beta) + \varepsilon_{it},$$

for the variable DocVis, the number of visits to the doctor, in the German health care data. (See Example ~~11.10~~ 11.13 for details.) The regression results for the specification,

$$\mathbf{x}_{it} = (1, Age_{it}, Education_{it}, Income_{it}, Kids_{it})$$

TB 14.15

are repeated (in parentheses) in Table ~~16.14~~ 14.15 for convenience. The nonlinear least squares estimator is only semiparametric; it makes no assumption about the distribution of $DocVis_{it}$ or about ε_{it}. We do see striking increases in the standard errors when the "~~cluster~~ robust" asymptotic covariance matrix is used. (The estimates are given in Example ~~11.10~~ 11.13.) The analysis at this point assumes that the nonlinear least squares estimator remains consistent in the presence of the cross-observation correlation. Given the way the model is specified, that is, only in terms of the conditional mean function, this is probably reasonable. The extension would imply a nonlinear generalized regression as opposed to a nonlinear ordinary regression.

In Example ~~16~~14.10, we narrowed this model by assuming that the observations on doctor visits were generated by a geometric distribution,

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

The conditional mean is still $\exp(\mathbf{x}_{it}'\beta)$, but this specification adds the structure of a particular distribution for outcomes. The pooled model was estimated in Example ~~16~~14.10. Example ~~16.13~~ 14.14 added the panel data assumptions of random then fixed effects to the model. The model is now

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

The pooled, random effects and fixed effects estimates appear in Table ~~16.14~~ 14.15. The pooled estimates, where the standard errors are corrected for the panel data grouping, are comparable to the nonlinear least squares estimates with the robust standard errors. The parameter estimates are similar—both are consistent and this is a very large sample. The smaller standard errors seen for the MLE are the product of the more detailed specification.

We will now relax the specification by assuming a two-class finite mixture model. We also specify that the class probabilities are functions of gender and marital status. For the latent

TABLE ~~16.15~~ 14.16 Estimated Latent Class Linear Regression Model for GPA

	One Class		Latent Class 1		Latent Class 2	
Parameter	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
β_1	1.0918	0.1082	1.6423	0.05351	−0.3344	0.09288
β_2	0.0180	0.0013	0.01691	0.0007324	0.02649	0.001248
β_3	−0.0473	0.0067	−0.04473	0.003451	−0.06502	0.005739
β_4	−0.4687	0.0726	−0.4567	0.04688	0.01395	0.06964
β_5	−0.1569	0.0306	−0.1177	0.01611	−0.1388	0.02738
θ_1	0.0000	0.0000	−0.4280	0.06938	0.0000	0.0000
θ_2	0.0000	0.0000	0.8255	0.06322	0.0000	0.0000
θ_3	0.0000	0.0000	−0.07829	0.07143	0.0000	0.0000
Prob $\mid z$	1.0000		0.47697		0.52303	
ln *L*	−61917.97		−58708.63			

TABLE ~~16.16~~ 14.17 Descriptive Statistics for Doctor Visits

Class	*Mean*	*Standard Deviation*
All, $n = 27{,}326$	3.18352	7.47579
Class 1, $n = 12{,}349$	5.80347	1.63076
Class 2, $n = 14{,}977$	1.02330	3.18352

class specification,

$$\text{Prob}(class_i = 1 \mid \mathbf{z}_i) = \Lambda(\theta_1 + \theta_2 \, Female_i + \theta_3 \, Married_i).$$

TB 14.16

The model structure is the geometric regression as before. Estimates of the parameters of the latent class model are shown in Table ~~16.15~~ 14.16. See Section E3.7 for discussion of estimation methods.

Deb and Trivedi (2002) suggested that a meaningful distinction between groups of health care system users would be between "infrequent" and "frequent" users. To investigate whether our latent class model is picking up this distinction in the data, we used (~~16~~ 14-102) to predict the class memberships (class 1 or 2). We then linearly regressed $DocVis_{it}$ on a constant and a dummy variable for class 2. The results are

$$DocVis_{it} = 5.8034\,(0.0465) - 4.7801\,(0.06282)\,Class2_i + e_{it},$$

TB 14.17

where estimated standard errors are in parentheses. The linear regression suggests that the class membership dummy variable is strongly segregating the observations into frequent and infrequent users. The information in the regression is summarized in the descriptive statistics in Table ~~16.16~~ 14.17.

14.11 ~~16.10~~ SUMMARY AND CONCLUSIONS

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

14-101

previous chapter considered GMM estimation techniques which are less parametric, but more robust to variation in the underlying data generating process. Together, ML and GMM estimation account for the large majority of empirical estimation in econometrics.

AU: Terms not marked as bold KTs in text are indicated by blue check

Key Terms and Concepts

- AIC
- ✓ Asymptotic efficiency
- ✓ Asymptotic normality
- Asymptotic variance
- ✓ Autocorrelation
- Bayes theorem
- BHHH estimator
- BIC
- Butler and Moffitt's model
- ✓ Cluster estimator
- Concentrated log-likelihood
- ✓ Conditional likelihood
- ✓ Consistency
- Cramér–Rao lower bound
- Efficient score
- ✓ Estimable parameters
- Exclusion restriction
- Exponential regression model
- Finite mixture model
- ✓ Fixed effects
- Full information maximum likelihood {FIML}
- Gauss–Hermite quadrature
- Generalized sum of squares
- ✓ Geometric regression
- ✓ GMM estimator
- Identification
- Incidental parameters problem
- Index function model
- ✓ Information matrix
- Information matrix equality
- ✓ Invariance
- Jacobian
- ② Lagrange multiplier statistic {LM}
- Lagrange multiplier test
- ✓ Latent class model
- Latent class linear regression model
- Likelihood equation
- Likelihood function
- ✓ Likelihood inequality
- Likelihood ratio index
- Likelihood ratio statistic
- Likelihood ratio test {LR}
- ✓ Limited information maximum likelihood
- ✓ Logit model
- Loglinear conditional mean
- Maximum likelihood
- Maximum likelihood estimator
- *M* estimator
- Method of scoring
- ✓ Murphy and Topel estimator
- Newton's method
- Nonlinear least squares
- Noncentral chi-squared distribution
- ① Normalization
- Oberhofer–Kmenta estimator
- Outer product of gradients estimator (OPG)
- ✓ Parameter space
- Pseudo-log likelihood function
- Pseudo MLE
- Pseudo R squared
- ✓ Quadrature
- ✓ Random effects
- Regularity conditions
- Sandwich estimator
- Score test
- Score vector
- ✓ Stochastic frontier
- ✓ Two-step maximum likelihood
- Wald statistic
- ✓ Wald test
- ✓ Vuong test

✓ • Latent regression

① • Nonnested models

② • Kullback–Leibler information criterion

Bayes's

acute

Insert terms where indicated. DNS numbers

Exercises

1. Assume that the distribution of x is $f(x) = 1/\theta, 0 \le x \le \theta$. In random sampling from this distribution, prove that the sample maximum is a consistent estimator of θ. Note: You can prove that the maximum is the maximum likelihood estimator of θ. But the usual properties do not apply here. Why not? (Hint: Attempt to verify that the expected first derivative of the log-likelihood with respect to θ is zero.)

that/

2. In random sampling from the exponential distribution $f(x) = (1/\theta)e^{-x/\theta}, x \ge 0, \theta > 0$, find the maximum likelihood estimator of θ and obtain the asymptotic distribution of this estimator.
3. *Mixture distribution.* Suppose that the joint distribution of the two random variables x and y is

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

AU: Check exercise titles-bold or italic?

14-102

a. Find the maximum likelihood estimators of β and θ and their asymptotic joint distribution.

b. Find the maximum likelihood estimator of $\theta/(\beta+\theta)$ and its asymptotic distribution.

c. Prove that $f(x)$ is of the form

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

and find the maximum likelihood estimator of γ and its asymptotic distribution.

d. Prove that $f(y\,|\,x)$ is of the form

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

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

e. Prove that

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

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

f. Prove that

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

Based on this distribution, what is the maximum likelihood estimator of β?

4. Suppose that x has the Weibull distribution

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

a. Obtain the log-likelihood function for a random sample of n observations.

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

c. Obtain the second derivatives matrix of the log-likelihood with respect to α and β. The exact expectations of the elements involving β involve the derivatives of the gamma function and are quite messy analytically. Of course, your exact result provides an empirical estimator. How would you estimate the asymptotic covariance matrix for your estimators in part b?

d. Prove that $\alpha\beta\text{Cov}[\ln x, x^\beta]=1$. (Hint: The expected first derivatives of the log-likelihood function are zero.)

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

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

a. Obtain the maximum likelihood estimates of α and β, and estimate the asymptotic covariance matrix for the estimates.

b. Carry out a Wald test of the hypothesis that $\beta=1$.

c. Obtain the maximum likelihood estimate of α under the hypothesis that $\beta=1$.

14-103

d. Using the results of Parts a and c, carry out a likelihood ratio test of the hypothesis that $\beta = 1$.

e. Carry out a Lagrange multiplier test of the hypothesis that $\beta = 1$.

6. **Limited Information Maximum Likelihood Estimation.** Consider a bivariate distribution for x and y that is a function of two parameters, α and β. The joint density is $f(x, y \mid \alpha, \beta)$. We consider maximum likelihood estimation of the two parameters. The full information maximum likelihood estimator is the now familiar maximum likelihood estimator of the two parameters. Now, suppose that we can factor the joint distribution as done in Exercise 3, but in this case, we have $f(x, y \mid \alpha, \beta) = f(y \mid x, \alpha, \beta) f(x \mid \alpha)$. That is, the conditional density for y is a function of both parameters, but the marginal distribution for x involves only α.
   a. Write down the general form for the log-likelihood function using the joint density.
   b. Because the joint density equals the product of the conditional times the marginal, the log-likelihood function can be written equivalently in terms of the factored density. Write this down, in general terms.
   c. The parameter α can be estimated by itself using only the data on x and the log likelihood formed using the marginal density for x. It can also be estimated with β by using the full log-likelihood function and data on both y and x. Show this.
   d. Show that the first estimator in part c has a larger asymptotic variance than the second one. This is the difference between a limited information maximum likelihood estimator and a full information maximum likelihood estimator.
   e. Show that if $\partial^2 \ln f(y \mid x, \alpha, \beta)/\partial\alpha\partial\beta = 0$, then the result in part d is no longer true.
7. Show that the likelihood inequality in Theorem ~~16~~14.3 holds for the Poisson distribution used in Section ~~16~~14.3 by showing that $E[(1/n)\ln L(\theta \mid y)]$ is uniquely maximized at $\theta = \theta_0$. (Hint: First show that the expectation is $-\theta + \theta_0 \ln\theta - E_0[\ln y_i!]$.)
8. Show that the likelihood inequality in Theorem ~~16~~14.3 holds for the normal distribution.
9. For random sampling from the classical regression model in (~~16~~14-3), reparameterize the likelihood function in terms of $\eta = 1/\sigma$ and $\delta = (1/\sigma)\boldsymbol{\beta}$. Find the maximum likelihood estimators of η and $\boldsymbol{\delta}$ and obtain the asymptotic covariance matrix of the estimators of these parameters.
10. Consider sampling from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_M)$ and covariance matrix $\sigma^2\mathbf{I}$. The log-likelihood function is

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

Show that the maximum likelihood estimates of the parameters are $\hat{\mu} = \bar{y}_m$, and

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

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

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

14-104

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

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

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

11. Prove the result claimed in Example 4.~~9~~7.

Applications

1. **Binary Choice.** This application will be based on the health care data analyzed in Example ~~16~~14.15 and several others. Details on obtaining the data are given in Example 11.10. We consider analysis of a dependent variable, y_{it}, that takes values 11.14 and 1 and 0 with probabilities $F(\mathbf{x}_i'\boldsymbol{\beta})$ and $1 - F(\mathbf{x}_i'\boldsymbol{\beta})$, where F is a function that defines a probability. The dependent variable, y_{it}, is constructed from the count variable *DocVis*, which is the number of visits to the doctor in the given year. Construct the binary variable

$$y_{it} = 1 \text{ if } DocVis_{it} > 0, 0 \text{ otherwise.}$$

We will build a model for the probability that y_{it} equals one. The independent variables of interest will be,

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, female_{it}, married_{it}, hsat_{it}).$$

a. According to the model, the theoretical density for y_{it} is

$$f(y_{it}\,|\,\mathbf{x}_{it}) = F(\mathbf{x}_{it}'\boldsymbol{\beta}) \text{ for } y_{it} = 1 \text{ and } 1 - F(\mathbf{x}_{it}'\boldsymbol{\beta}) \text{ for } y_{it} = 0.$$

We will assume that a "logit model" (see Section ~~23~~17.4) is appropriate, so that

$$F(\mathbf{x}_{it}'\boldsymbol{\beta}) = \Lambda(\mathbf{x}_{it}'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta})}{1 - \exp(\mathbf{x}_{it}'\boldsymbol{\beta})}.$$

Show that for the two outcomes, the probabilities may be may be combined into the density function

$$f(y_{it}\,|\,\mathbf{x}_{it}) = g(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}) = \Lambda[(2y_{it} - 1)\mathbf{x}_{it}'\boldsymbol{\beta}].$$

Now, use this result to construct the log-likelihood function for a sample of data on $(y_{it}, \mathbf{x}_{it})$. (Note that we will be ignoring the panel aspect of the data set. Build the model as if this were a cross section.)

b. Derive the likelihood equations for estimation of $\boldsymbol{\beta}$.

c. Derive the second derivatives matrix of the log likelihood function. (Hint: The following will prove useful in the derivation: $d\Lambda(t)/dt = \Lambda(t)[1 - \Lambda(t)]$.)

d. Show how to use Newton's method to estimate the parameters of the model.

e. Does the method of scoring differ from Newton's method? Derive the negative of the expectation of the second derivatives matrix.

f. Obtain maximum likelihood estimates of the parameters for the data and variables noted. Report your results: estimates, standard errors, etc., as well as the value of the log-likelihood.

AU: Is "d" in part c OK Roman?

14-105

END 14

g. Test the hypothesis that the coefficients on female and marital status are zero. Show how to do the test using Wald, LM, and LR tests, and then carry out the tests.

h. Test the hypothesis that all the coefficients in the model save for the constant term are equal to zero.

2. ~~**Stochastic Frontier Model.** Section 10.4.1 presents estimates of a Cobb-Douglas cost function using Nerlove's 1955 data on the U.S. electric power industry. Christensen and Greene's 1976 update of this study used 1970 data for this industry. The Christensen and Greene data are given in Appendix Table F4.3. These data have provided a standard test data set for estimating different forms of production and cost functions, including the stochastic frontier model examined in Example 16.9. It has been suggested that one explanation for the apparent finding of economies of scale in these data is that the smaller firms were inefficient for other reasons. The stochastic frontier might allow one to disentangle these effects. Use these data to fit a frontier cost function which includes a quadratic term in log output in addition to the linear term and the factor prices. Then examine the estimated Jondrow et al. residuals to see if they do indeed vary negatively with output, as suggested. (This will require either some programming on your part or specialized software. The stochastic frontier model is provided as an option in Stata, TSP and LIMDEP. Or, the likelihood function can be programmed fairly easily for RATS, MatLab or GAUSS. Note, for a cost frontier as opposed to a production frontier, it is necessary to reverse the sign on the argument in the Φ function that appears in the log-likelihood.)~~