14-50

There is ample evidence that the asymptotic results for these statistics are problematic in small or moderately sized samples. [See, e.g., Davidson and MacKinnon (2004, pp. 424–428).] The true distributions of all three statistics involve the data and the unknown parameters and, as suggested by the algebra, converge to the F distribution *from above*. The implication is that critical values from the chi-squared distribution are likely to be too small; that is, using the limiting chi-squared distribution in small or moderately sized samples is likely to exaggerate the significance of empirical results. Thus, in applications, the more conservative F statistic (or t for one restriction) is likely to be preferable unless one's data are plentiful.

14 ~~16~~.9.2 THE GENERALIZED REGRESSION MODEL

For the generalized regression model of Section 8.1,

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

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

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\ln|\boldsymbol{\Omega}|. \qquad (\cancel{16}\,14\text{-}48)$$

Because $\boldsymbol{\Omega}$ is a matrix of known constants, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the vector that minimizes the **generalized sum of squares,** KT

$$S_*(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

(hence the name *generalized least squares*). The necessary conditions for maximizing L are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{X}_*'(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta}) = \mathbf{0},$$
$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \qquad (\cancel{16}\,14\text{-}49)$$
$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})'(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta}) = 0.$$

The solutions are the OLS estimators using the transformed data:

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}, \qquad (\cancel{16}\,14\text{-}50)$$

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}})'(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}) \qquad (\cancel{16}\,14\text{-}51)$$
$$= \frac{1}{n}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}),$$

which implies that with normally distributed disturbances, generalized least squares is also maximum likelihood. As in the classical regression model, the maximum likelihood estimator of σ^2 is biased. An unbiased estimator is the one in (~~8~~-14). The conclusion,

9

which would be expected, is that when $\boldsymbol{\Omega}$ is known, the maximum likelihood estimator is generalized least squares. 14

When $\boldsymbol{\Omega}$ is unknown and must be estimated, then it is necessary to maximize the log-likelihood in (~~16~~-48) with respect to the full set of parameters $[\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}]$ simultaneously. Because an unrestricted $\boldsymbol{\Omega}$ alone contains $n(n+1)/2-1$ parameters, it is clear that some restriction will have to be placed on the structure of $\boldsymbol{\Omega}$ for estimation to proceed. We will examine several applications in which $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$ for some smaller vector of parameters in the next several sections. We note only a few general results at this point.

1. For a given value of $\boldsymbol{\theta}$ the estimator of $\boldsymbol{\beta}$ would be feasible GLS and the estimator of σ^2 would be the estimator in (~~16~~-51). 14
2. The likelihood equations for $\boldsymbol{\theta}$ will generally be complicated functions of $\boldsymbol{\beta}$ and σ^2, so joint estimation will be necessary. However, in many cases, for given values of $\boldsymbol{\beta}$ and σ^2, the estimator of $\boldsymbol{\theta}$ is straightforward. For example, in the model of 9 (8-15), the iterated estimator of θ when $\boldsymbol{\beta}$ and σ^2 and a prior value of θ are given is the prior value plus the slope in the regression of $(e_i^2/\hat{\sigma}_i^2 - 1)$ on z_i.

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

We will now examine the two primary, single-equation applications: heteroscedasticity and autocorrelation.

14 ~~16~~.9.2.a Multiplicative Heteroscedasticity

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

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i'\boldsymbol{\alpha}). \qquad \text{(14-52)}$$

A model with heteroscedasticity of the form

$$\sigma_i^2 = \sigma^2 \prod_{m=1}^{M} z_{im}^{\alpha_m} \qquad \text{(14-53)}$$

9

results if the logs of the variables are placed in z_i. The groupwise heteroscedasticity model described in Section 8.8.2 is produced by making $\mathbf{z}_i$ a set of group dummy variables (one must be omitted). In this case, σ^2 is the disturbance variance for the base group whereas for the other groups, $\sigma_g^2 = \sigma^2 \exp(\alpha_g)$.

14-52

We begin with a useful simplification. Let $\mathbf{z}_i$ include a constant term so that $\mathbf{z}_i' = [1, \mathbf{q}_i']$, where $\mathbf{q}_i$ is the original set of variables, and let $\boldsymbol{\gamma}' = [\ln \sigma^2, \boldsymbol{\alpha}']$. Then, the model is simply $\sigma_i^2 = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$. Once the full parameter vector is estimated, $\exp(\gamma_1)$ provides the estimator of σ^2. (This estimator uses the invariance result for maximum likelihood estimation. See Section ~~16~~ 14.4.5.d.)

The log-likelihood is

$$\begin{aligned} \ln L &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n} \ln \sigma_i^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sigma_i^2} \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n} \mathbf{z}_i'\boldsymbol{\gamma} - \frac{1}{2}\sum_{i=1}^{n} \frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}. \end{aligned} \tag{$\cancel{16}$ 14-54}$$

The likelihood equations are

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \mathbf{x}_i \frac{\varepsilon_i}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} = \mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon} = \mathbf{0}, \\ \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} &= \frac{1}{2}\sum_{i=1}^{n} \mathbf{z}_i \left(\frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1\right) = \mathbf{0}. \end{aligned} \tag{$\cancel{16}$ 14-55}$$

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

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\mathbf{x}_i\mathbf{x}_i' = -\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}, \tag{$\cancel{16}$ 14-56}$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\gamma}'} = -\sum_{i=1}^{n} \frac{\varepsilon_i}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\mathbf{x}_i\mathbf{z}_i', \tag{$\cancel{16}$ 14-57}$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma}\,\partial \boldsymbol{\gamma}'} = -\frac{1}{2}\sum_{i=1}^{n} \frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\mathbf{z}_i\mathbf{z}_i'. \tag{$\cancel{16}$ 14-58}$$

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

$$-E\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\delta}\,\partial \boldsymbol{\delta}'}\right) = \begin{bmatrix} \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2}\mathbf{Z}'\mathbf{Z} \end{bmatrix} = -\overline{\mathbf{H}}. \tag{$\cancel{16}$ 14-59}$$

KT

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

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

where $\boldsymbol{\delta}_t$ (i.e., $\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t$, and $\boldsymbol{\Omega}_t$) is the estimate at iteration t, $\mathbf{g}_t$ is the two-part vector of first derivatives $[\partial \ln L/\partial\boldsymbol{\beta}_t', \partial \ln L/\partial\boldsymbol{\gamma}_t']'$, and $\overline{\mathbf{H}}$ is partitioned likewise. [**Newton's method** KT uses the actual second derivatives in (~~16~~-56)–(~~16~~-58) rather than their expectations in (~~16~~-59). The scoring method exploits the convenience of the zero expectation of the

14 14 14

14-53

off-diagonal block (cross derivative) in (~~16~~ 14-57).] Because $\overline{\mathbf{H}}$ is block diagonal, the iteration can be written as separate equations:

$$\begin{aligned}\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\boldsymbol{\varepsilon}_t) \\ &= \boldsymbol{\beta}_t + (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_t) \\ &= (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{y} \text{ (of course).}\end{aligned} \qquad \text{14 (16-60)}$$

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

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t + [2(\mathbf{Z}'\mathbf{Z})^{-1}]\left[\frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i\left(\frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1\right)\right]. \qquad \text{14 (16-61)}$$

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

The remaining detail is to determine the starting value for the iteration. Because any consistent estimator will do, the simplest procedure is to use OLS for $\boldsymbol{\beta}$ and the slopes in a regression of the logs of the squares of the least squares residuals on $\mathbf{z}_i$ for $\boldsymbol{\gamma}$. Harvey (1976) shows that this method will produce an inconsistent estimator of $\gamma_1 = \ln \sigma^2$, but the inconsistency can be corrected just by adding 1.2704 to the value obtained.[18] Thereafter, the iteration is simply:

FN 18

FN 19

1. Estimate the disturbance variance σ_i^2 with $\exp(\mathbf{z}_i'\boldsymbol{\gamma})$.
2. Compute $\boldsymbol{\beta}_{t+1}$ by FGLS.[19]
3. Update $\boldsymbol{\gamma}_t$ using the regression described in the preceding paragraph.
4. Compute $\mathbf{d}_{t+1} = [\boldsymbol{\beta}_{t+1}, \boldsymbol{\gamma}_{t+1}] - [\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t]$. If $\mathbf{d}_{t+1}$ is large, then return to step 1.

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

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{\text{ML}}] = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1},$$

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}_{\text{ML}}] = 2(\mathbf{Z}'\mathbf{Z})^{-1}.$$

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

Testing the null hypothesis of homoscedasticity in this model,

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

in (~~16~~ 14-52), is particularly simple. The Wald test will be carried out by testing the hypothesis that the last M elements of $\boldsymbol{\gamma}$ are zero. Thus, the statistic will be

delete -1

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

replace with equation on next page msp 14-54

[18] He also presents a correction for the asymptotic covariance matrix for this first step estimator of $\boldsymbol{\gamma}$.

[19] The two-step estimator obtained by stopping here would be fully efficient if the starting value for $\boldsymbol{\gamma}$ were consistent, but it would not be the maximum likelihood estimator.

Insert on msp 14-53 where indicated

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

end of insert

14-55

1/2

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

delete -1

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

where $\mathbf{Z}_1$ is the last M columns of $\mathbf{Z}$, not including the column of ones, and $\mathbf{M}^0$ creates deviations from means. The likelihood ratio statistic is computed based on (16-54). Under both the null hypothesis (homoscedastic—using OLS) and the alternative (heteroscedastic—using MLE), the third term in $\ln L$ reduces to $-n/2$. Therefore, the statistic is simply

14c

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

14

where $s^2 = \mathbf{e}'\mathbf{e}/n$ using the OLS residuals. To compute the LM statistic, we will use the expected Hessian in (16-59). Under the null hypothesis, the part of the derivative vector in (16-55) that corresponds to $\boldsymbol{\beta}$ is $(1/s^2)\mathbf{X}'\mathbf{e} = \mathbf{0}$. Therefore, using (16-55), the LM statistic is

14 14

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

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

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

This is one-half times the explained sum of squares in the linear regression of the variable $h_i = (e_i^2/s^2 - 1)$ on $\mathbf{Z}$, which is the Breusch–Pagan/Godfrey LM statistic from Section 8.5.2.

9 14.6

Example 16.8 Multiplicative Heteroscedasticity

In Example 6.2, we fit a cost function for the U.S. airline industry of the form

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

where $C_{i,t}$ is total cost, $Q_{i,t}$ is output, and $P_{fuel,i,t}$ is the price of fuel and the 90 observations in the data set are for six firms observed for 15 years. (The model also included dummy variables for firm and year, which we will omit for simplicity.) In Example 8.4, we fit a revised model in which the load factor appears in the variance of $\varepsilon_{i,t}$ rather than in the regression function. The model is

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

Estimates were obtained by iterating the weighted least squares procedure using weights $W_{i,t} = \exp(-c_1 - c_2\, Loadfactor_{i,t})$. The estimates of γ_1 and γ_2 were obtained at each iteration by regressing the logs of the squared residuals on a constant and $Loadfactor_{it}$. It was noted at the end of the example [and is evident in (16-61)] that these would be the wrong weights to use for the iterated weighted least if we wish to compute the MLE. Table 16.2 reproduces the results from Example 8.4 and adds the MLEs produced using Harvey's method. The MLE of γ_2 is substantially different from the earlier result. The Wald statistic for testing the

14.3

TB 14.3

9.4 14

14-56

TABLE ~~16~~.2 Multiplicative Heteroscedasticity Model

	Constant	*Ln Q*	*Ln² Q*	*Ln P_f*	*R^2*	*Sum of Squares*
OLS	9.1382	0.92615	0.029145	0.41006		
ln L = 54.2747	0.24507[a]	0.032306	0.012304	0.018807	0.9861674[c]	1.577479[d]
	0.22595[b]	0.030128	0.011346	0.017524		
Two-step	9.2463	0.92136	0.024450	0.40352		
	0.21896	0.033028	0.011412	0.016974	0.986119	1.612938
Iterated[e]	9.2774	0.91609	0.021643	0.40174		
	0.20977	0.032993	0.011017	0.016332	0.986071	1.645693
MLE[f]	9.2611	0.91931	0.023281	0.40266		
ln L = 57.3122	0.2099	0.032295	0.010987	0.016304	0.986100	1.626301

[a]Conventional OLS standard errors
[b]White robust standard errors
[c]Squared correlation between actual and fitted values
[d]Sum of squared residuals
[e]Values of c_2 by iteration: 8.254344, 11.622473, 11.705029, 11.710618, 11.711012, 11.711040, 11.711042
[f]Estimate of γ_2 is 9.78076 (2.839).

homoscedasticity restriction ($\alpha = 0$) is $(9.78076/2.839)^2 = 11.869$, which is greater than 3.84, so the null hypothesis would be rejected. The likelihood ratio statistic is $-2(54.2747 - 57.3122) = 6.075$, which produces the same conclusion. However, the LM statistic is 2.96, which conflicts. This is a finite sample result that is not uncommon.

14 ~~16~~.9.2.b Autocorrelation

11.6.3

At various points in the preceding sections, we have considered models in which there is correlation across observations, including the spatial autocorrelation case in Section 9.7.2, autocorrelated disturbances in panel data models [Section 9.6.3 and in (9-28)], and in the seemingly unrelated regressions model in Section 10.2.6. The first order autoregression model examined there will be formalized in detail in Chapter ~~19~~ 20. We will briefly examine it here to highlight some useful results about the maximum likelihood estimator.

11 11 9

The linear regression model with first order autoregressive [AR(1)] disturbances is

minus

$$
\begin{aligned}
y_t &= \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t,\ t = 1, \ldots, T,\\
\varepsilon_t &= \rho\varepsilon_t + u_t,\ |\rho| < 1, \quad t-1\\
E[u_t \mid \mathbf{X}] &= 0\\
E[u_t u_s \mid \mathbf{X}] &= \sigma_u^2 \quad \text{if } t = s \quad \text{and 0 otherwise.}
\end{aligned}
$$

Feasible GLS estimation of the parameters of this model is examined in detail in Chapter ~~19~~. We now add the assumption of normality, $u_t \sim N[0, \sigma_u^2]$, and construct the maximum likelihood estimator.

20

Because every observation on y_t is correlated with every other observation, in principle, to form the likelihood function, we have the joint density of one T-variate observation. The Prais and Winsten (1954) transformation in (~~19~~-28) suggests a useful

20

way to reformulate this density. We can write

$$f(y_1, y_2, \ldots, y_T) = f(y_1) f(y_2 \mid y_1), f(y_3 \mid y_2) \ldots, f(y_T \mid y_{T-1}).$$

Because

$$\begin{aligned} \sqrt{1-\rho^2}\, y_1 &= \sqrt{1-\rho^2}\, \mathbf{x}_1'\boldsymbol{\beta} + u_1 \\ y_t \mid y_{t-1} &= \rho y_{t-1} + (\mathbf{x}_t - \rho \mathbf{x}_{t-1})'\boldsymbol{\beta} + u_t, \end{aligned}$$

14 (~~16~~-62)

and the observations on u_t are independently normally distributed, we can use these results to form the log-likelihood function,

$$\ln L = \left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_u^2 \cancel{-} + \frac{1}{2}\ln(1-\rho^2) - \frac{(1-\rho^2)(y_1 - \mathbf{x}_1'\boldsymbol{\beta})^2}{2\sigma_u^2} \right] + \sum_{t=2}^{T} \left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_u^2 - \frac{[(y_t - \rho y_{t-1}) - (\mathbf{x}_t - \rho \mathbf{x}_{t-1})'\boldsymbol{\beta}]^2}{2\sigma_{\cancel{u}\,v}^2} \right].$$

14 (~~16~~-63) v

As usual, the MLE of $\boldsymbol{\beta}$ is GLS based on the MLEs of σ_u^2 and ρ, and the MLE for σ_u^2 will be $\mathbf{u}'\mathbf{u}/T$ given $\boldsymbol{\beta}$ and ρ. The complication is how to compute ρ. As we will note in Chapter ~~19~~ 20, there is a strikingly large number of choices for consistently estimating ρ in the AR(1) model. It is tempting to choose the most convenient, and then begin the back and forth iterations between $\boldsymbol{\beta}$ and (σ_u^2, ρ) to obtain the MLE. However, this strategy will not (in general) locate the MLE unless the intermediate estimates of the variance parameters also satisfy the likelihood equation, which for ρ is

$$\frac{\partial \ln L}{\partial \rho} = \frac{\rho \varepsilon_1^2}{\sigma_u^2} - \frac{\rho}{1-\rho^2} + \sum_{t=2}^{T} \frac{u_t \varepsilon_{t-1}}{\sigma_u^2}.$$

One could sidestep the problem simply by scanning the range of ρ of $(-1, +1)$ and computing the other estimators at every point, to locate the maximum of the likelihood function by brute force. With modern computers, even with long time series, the amount of computation involved would be minor (if a bit inelegant and inefficient). Beach and MacKinnon (1978a) developed a more systematic algorithm for searching for ρ in this model. The iteration is then defined between ρ and $(\boldsymbol{\beta}, \sigma_u^2)$ as usual.

The information matrix for this log-likelihood is

$$-E\left[\frac{\partial^2 \ln L}{\partial \begin{pmatrix} \boldsymbol{\beta} \\ \sigma_u^2 \\ \rho \end{pmatrix} \partial(\boldsymbol{\beta}' \sigma_u^2\, \rho)}\right] = \begin{bmatrix} \frac{1}{\sigma_u^2}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \frac{T}{2\sigma_u^4} & \frac{\rho}{\sigma_u^2(1-\rho^2)} \\ \mathbf{0}' & \frac{\rho}{\sigma_u^2(1-\rho^2)} & \frac{T-2}{1-\rho^2} + \frac{1+\rho^2}{(1-\rho^2)^2} \end{bmatrix}.$$

14 (~~16~~-64)

Note that the diagonal elements in the matrix are $O(T)$. But the (2, 3) and (3, 2) elements are constants of $O(1)$ that will, like the second part of the (3, 3) element, become minimal as T increases. Dropping these "end effects" (and treating $T-2$ as the same as T when T increases) produces a diagonal matrix from which we extract the

14-58

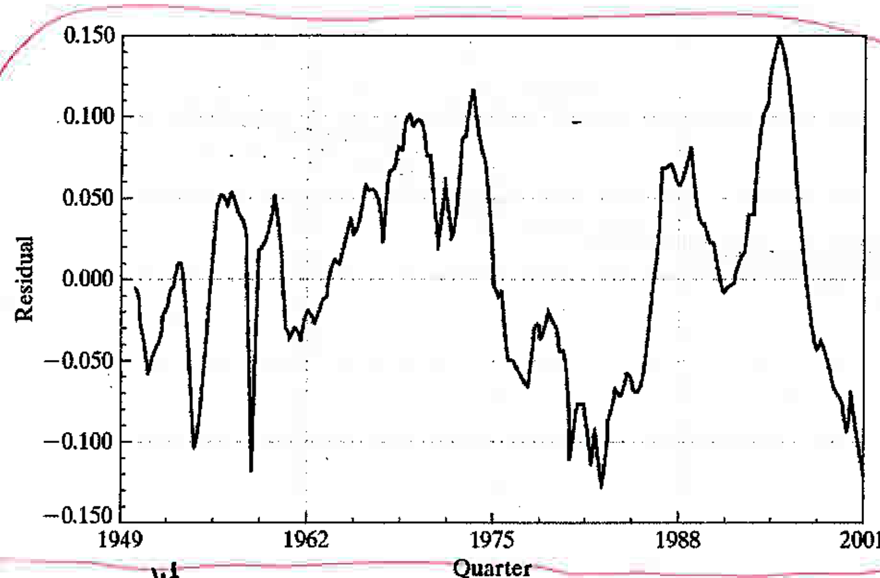

FIGURE ~~16~~14.3 Residuals from Estimated Money Demand Equation.

standard approximations for the MLEs in this model:

$$\text{Asy. Var}[\hat{\beta}] = \sigma_u^2(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1},$$
$$\text{Asy. Var}[\hat{\sigma}_u^2] = \frac{2\sigma_u^4}{T}, \qquad (\cancel{16}14\text{-}65)$$
$$\text{Asy. Var}[\hat{\rho}] = \frac{1-\rho^2}{T}.$$

Example ~~16~~14.7 Autocorrelation in a Money Demand Equation

Using the macroeconomic data in Table F5.1, we fit a money demand equation,

$$\ln(M1/CPI\_u)_t = \beta_1 + \beta_2 \ln \textit{Real GDP}_t + \beta_3 \ln \textit{T-bill rate}_t + \varepsilon_t.$$

FIG 14.3

The least squares residuals shown in Figure ~~16~~14.3 display the typical pattern for a highly autocorrelated series.

TB 14.4

The simple first-order autocorrelation of the ordinary least squares residuals is ~~r = 0.9557002.~~ r = 1 − d/2 = 0.9557, where d is the Durbin-Watson statistic in (20-23). We then refit the model using the Prais and Winsten FGLS estimator and the maximum likelihood estimator using the Beach and MacKinnon algorithm. The results are shown in Table ~~16.3~~ 14.4. Although the OLS estimator is consistent in this model, nonetheless, the FGLS and ML estimates are quite different.

~~16~~14.9.3 SEEMINGLY UNRELATED REGRESSION MODELS

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

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, M,$$
$$E[\boldsymbol{\varepsilon}_i \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_M] = 0, \qquad (\cancel{16}14\text{-}66)$$
$$E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j' \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_M] = \sigma_{ij}\mathbf{I}.$$

14-59

TABLE ~~16.3~~ 14.4 Estimates of Money Demand Equation: $T = 204$

	OLS		*Prais and Winsten*		*Maximum Likelihood*	
Variable	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Constant	−2.1316	0.09100	−1.4755	0.2550	−1.6319	0.4296
Ln real GDP	0.3519	0.01205	0.2549	0.03097	0.2731	0.0518
Ln T-bill rate	−0.1249	0.009841	−0.02666	0.007007	−0.02522	0.006941
σ_ε	0.06185		0.07767		0.07571	
σ_u	0.06185		0.01298		0.01273	
ρ	0.	0.	0.9557	0.02061	0.9858	0.01180

FGLS estimation of this model is examined in detail in Section 10.2.3. We will now add the assumption of normally distributed disturbances to the model and develop the maximum likelihood estimators. Given the covariance structure defined in (~~16~~14-66), the joint normality assumption applies to the vector of M disturbances observed at time t, which we write as

$$\boldsymbol{\varepsilon}_t \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_M \sim N[\mathbf{0}, \boldsymbol{\Sigma}], \; t = 1, \ldots, T. \qquad (\cancel{16}14\text{-}67)$$

~~16~~14.9.3.a The Pooled Model

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

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it}, \\ E[\varepsilon_{it} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_M] &= 0, \\ E[\varepsilon_{it}\varepsilon_{js} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_M] &= \sigma_{ij} \quad \text{if} \quad t = s, \quad \text{and} \quad 0 \quad \text{if} \quad t \neq s. \end{aligned} \qquad (\cancel{16}14\text{-}68)$$

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

$$\ln L = \sum_{t=1}^{T}\left[-\frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\right]. \qquad (\cancel{16}14\text{-}69)$$

As we saw earlier, the efficient estimator for this model is GLS as shown in (10-21). Because the elements of $\boldsymbol{\Sigma}$ must be estimated, the FGLS estimator based on (10-9) is used.

As we have seen in several applications now, the maximum likelihood estimator of $\boldsymbol{\beta}$, given $\boldsymbol{\Sigma}$, is GLS, based on (10-21). The maximum likelihood estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\sigma}_{ij} = \frac{(\mathbf{y}_i - \mathbf{X}_i\hat{\hat{\boldsymbol{\beta}}}_{ML})'(\mathbf{y}_j - \mathbf{X}_j\hat{\hat{\boldsymbol{\beta}}}_{ML})}{T} = \frac{\hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_j}{T} \qquad (\cancel{16}14\text{-}70)$$

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

KT

AU: Term in chap list is Oberhofer and Kmenta estimator?

14-60

(16-70) and

$$\hat{\hat{\beta}} = [\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}]. \tag{16-71}$$

The process may begin with the (consistent) ordinary least squares estimator, then (16-70), and so on. The computations are simple, using basic matrix algebra. Hypothesis tests about β may be done using the familiar Wald statistic. The appropriate estimator of the asymptotic covariance matrix is the inverse matrix in brackets in (10-21).

For testing the hypothesis that the off-diagonal elements of Σ are zero—that is, that there is no correlation across firms—there are three approaches. The likelihood ratio test is based on the statistic

$$\lambda_{\text{LR}} = T(\ln|\hat{\Sigma}_{heteroscedastic}| - \ln|\hat{\Sigma}_{general}|) = T\left(\sum_{i=1}^{M} \ln\hat{\sigma}_i^2 - \ln|\hat{\Sigma}|\right), \tag{16-72}$$

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

$$\lambda_{\text{LM}} = T\sum_{i=2}^{n}\sum_{j=1}^{i-1} r_{ij}^2, \tag{16-73}$$

where r_{ij}^2 is the ijth residual correlation coefficient. If every equation had a different parameter vector, then equation specific ordinary least squares would be efficient (and ML) and we would compute r_{ij} from the OLS residuals (assuming that there are sufficient observations for the computation). Here, however, we are assuming only a single-parameter vector. Therefore, the appropriate basis for computing the correlations is the residuals from the iterated estimator in the groupwise heteroscedastic model, that is, the same residuals used to compute $\hat{\sigma}_i^2$. (An asymptotically valid approximation to the test can be based on the FGLS residuals instead.) Note that this is not a procedure for testing all the way down to the classical, homoscedastic regression model. That case involves different LM and LR statistics based on the groupwise heteroscedasticity model. If either the LR statistic in (16-72) or the LM statistic in (16-73) are smaller than the critical value from the table, the conclusion, based on this test, is that the appropriate model is the groupwise heteroscedastic model.

16.9.3.b The SUR Model

The Oberhofer–Kmenta (1974) conditions are met for the seemingly unrelated regressions model, so maximum likelihood estimates can be obtained by iterating the FGLS procedure. We note, once again, that this procedure presumes the use of (10-9) for estimation of σ_{ij} at each iteration. Maximum likelihood enjoys no advantages over FGLS in its asymptotic properties.[20] Whether it would be preferable in a small sample is an open question whose answer will depend on the particular data set.

FN 20

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

14-61

14 ~~16~~.9.3.c Exclusion Restrictions

142

By simply inserting the special form of $\boldsymbol{\Omega}$ in the log-likelihood function for the generalized regression model in (~~16~~-48), we can consider direct maximization instead of iterated FGLS. It is useful, however, to reexamine the model in a somewhat different formulation. This alternative construction of the likelihood function appears in many other related models in a number of literatures.

Consider one observation on each of the M dependent variables and their associated regressors. We wish to arrange this observation horizontally instead of vertically. The model for this observation can be written

$$\begin{aligned}[y_1 \quad y_2 \quad \cdots \quad y_M]_t &= [\mathbf{x}_t^*]'[\boldsymbol{\pi}_1 \quad \boldsymbol{\pi}_2 \quad \cdots \quad \boldsymbol{\pi}_M] + [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_M]_t \\ &= [\mathbf{x}_t^*]'\boldsymbol{\Pi} + \mathbf{E},\end{aligned} \qquad \text{(14 ~~16~~-74)}$$

where $\mathbf{x}_t^*$ is the full set of all K^* *different* independent variables that appear in the model. The parameter matrix then has one column for each equation, but the columns are not the same as $\boldsymbol{\beta}_j$ in (~~16~~-66) unless every variable happens to appear in every equation. 14
Otherwise, in the ith equation, $\boldsymbol{\pi}_i$ will have a number of zeros in it, each one imposing an **exclusion restriction.** For example, consider a two-equation model for production costs for two airlines,

KT

$$C_{1t} = \alpha_1 + \beta_{1P}P_{1t} + \beta_{1L}LF_{1t} + \varepsilon_{1t},$$
$$C_{2t} = \alpha_2 + \beta_{2P}P_{2t} + \beta_{2L}LF_{2t} + \varepsilon_{2t},$$

where C is cost, P is fuel price, and LF is load factor. The tth observation would be

$$[C_1 \quad C_2]_t = [1 \quad P_1 \quad LF_1 \quad P_2 \quad LF_2]_t \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{1P} & 0 \\ \beta_{1L} & 0 \\ 0 & \beta_{2P} \\ 0 & \beta_{2L} \end{bmatrix} + [\varepsilon_1 \quad \varepsilon_2]_t.$$

This vector is one observation. Let $\boldsymbol{\varepsilon}_t$ be the vector of M disturbances for this observation arranged, for now, in a column. Then $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}$. The log of the joint normal density of these M disturbances is

$$\ln L_t = -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t. \qquad \text{(14 ~~16~~-75)}$$

The log-likelihood for a sample of T joint observations is the sum of these over t:

$$\ln L = \sum_{t=1}^{T}\ln L_t = -\frac{MT}{2}\ln(2\pi) - \frac{T}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t. \qquad \text{(14 ~~16~~-76)}$$

The term in the summation in (14 ~~16~~-76) is a scalar that equals its trace. We can always permute the matrices in a trace, so

$$\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t = \sum_{t=1}^{T}\text{tr}(\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t) = \sum_{t=1}^{T}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'). \qquad \text{(14 ~~16~~-77)}$$

14-62

This can be further simplified. The sum of the traces of T matrices equals the trace of the sum of the matrices [see (A-91)]. We will now also be able to move the constant matrix, $\boldsymbol{\Sigma}^{-1}$, outside the summation. Finally, it will prove useful to multiply and divide by T. Combining all three steps, we obtain

$$\sum_{t=1}^{T} \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = T \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\frac{1}{T}\right)\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right] = T\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{W}), \qquad (\cancel{16}\,14\text{-}78)$$

where

$$\mathbf{W}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{ti}\varepsilon_{tj}.$$

Because this step uses actual disturbances, $E[\mathbf{W}_{ij}] = \sigma_{ij}$; $\mathbf{W}$ is the $M \times M$ matrix we would use to estimate $\boldsymbol{\Sigma}$ if the ε's were actually observed. Inserting this result in the log-likelihood, we have

$$\ln L = -\frac{T}{2}[M\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{W})]. \qquad (\cancel{16}\,14\text{-}79)$$

We now consider maximizing this function.

FN 21

It has been shown[21] that

$$\frac{\partial \ln L}{\partial \boldsymbol{\Pi}'} = \frac{T}{2}\mathbf{X}^{*\prime}\mathbf{E}\boldsymbol{\Sigma}^{-1},$$
$$\frac{\partial \ln L}{\partial \boldsymbol{\Sigma}} = -\frac{T}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{W})\boldsymbol{\Sigma}^{-1}. \qquad (\cancel{16}\,14\text{-}80)$$

where the $\mathbf{x}_t^{*\prime}$ in ($\cancel{16}$ 14-74) is row t of $\mathbf{X}^*$. Equating the second of these derivatives to a zero matrix, we see that given the maximum likelihood estimates of the slope parameters, the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\mathbf{W}$, the matrix of mean residual sums of squares and cross products—that is, the matrix we have used for FGLS. [Notice that there is no correction for degrees of freedom; $\partial \ln L/\partial\boldsymbol{\Sigma} = \mathbf{0}$ implies (10-9).]

We also know that because this model is a generalized regression model, the maximum likelihood estimator of the parameter matrix [$\boldsymbol{\beta}$] must be equivalent to the FGLS estimator we discussed earlier.[22] It is useful to go a step further. If we insert our solution for $\boldsymbol{\Sigma}$ in the likelihood function, then we obtain the **concentrated log-likelihood**,

FN 22

AU: Term was a KT on 14-48. Here also?

$$\ln L_c = -\frac{T}{2}[M(1 + \ln(2\pi)) + \ln|\mathbf{W}|]. \qquad (\cancel{16}\,14\text{-}81)$$

We have shown, therefore, that the criterion for choosing the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \text{Min}_{\boldsymbol{\beta}}\tfrac{1}{2}\ln|\mathbf{W}|, \qquad (\cancel{16}\,14\text{-}82)$$

subject to the exclusion restrictions. This important result reappears in many other models and settings. This minimization must be done subject to the constraints in the parameter matrix. In our two-equation example, there are two blocks of zeros in the

[21] See, for example, Joreskog (1973).

[22] This equivalence establishes the Oberhofer–Kmenta conditions.

14-63

parameter matrix, which must be present in the MLE as well. The estimator of β is the set of nonzero elements in the parameter matrix in (16-74).

KT The **likelihood ratio statistic** is an alternative to the F statistic discussed earlier for testing hypotheses about β. The likelihood ratio statistic is[23]

FN 23

$$\lambda = -2(\log L_r - \log L_u) = T(\log|\hat{\mathbf{W}}_r| - \log|\hat{\mathbf{W}}_u|), \tag{16-83}$$

where $\hat{\mathbf{W}}_r$ and $\hat{\mathbf{W}}_u$ are the residual sums of squares and cross-product matrices using the constrained and unconstrained estimators, respectively. Under the null hypothesis of the restrictions, the limiting distribution of the likelihood ratio statistic is chi-squared with degrees of freedom equal to the number of restrictions. This procedure can also be used to test the homogeneity restriction in the multivariate regression model. The restricted model is the pooled model discussed in the preceding section.

It may also be of interest to test whether $\boldsymbol{\Sigma}$ is a diagonal matrix. Two possible approaches were suggested in Section 16.9.3a [see (16-72) and (16-73)]. The unrestricted model is the one we are using here, whereas the restricted model is the groupwise heteroscedastic model of Section 8.8.2 (Example 8.5), without the restriction of equal parameter vectors. As such, the restricted model reduces to separate regression models, estimable by ordinary least squares. The likelihood ratio statistic would be

$$\lambda_{\text{LR}} = T\left[\sum_{i=1}^{M} \log \hat{\sigma}_i^2 - \log|\hat{\boldsymbol{\Sigma}}|\right], \tag{16-84}$$

where $\hat{\sigma}_i^2$ is $\mathbf{e}_i'\mathbf{e}_i/T$ from the individual least squares regressions and $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimate of $\boldsymbol{\Sigma}$. This statistic has a limiting chi-squared distribution with $M(M-1)/2$ degrees of freedom under the hypothesis. The alternative suggested by Breusch and Pagan (1980) is the **Lagrange multiplier statistic,** KT

$$\lambda_{\text{LM}} = T\sum_{i=2}^{M}\sum_{j=1}^{i-1} r_{ij}^2, \tag{16-85}$$

where r_{ij} is the estimated correlation $\hat{\sigma}_{ij}/[\hat{\sigma}_{ii}\hat{\sigma}_{jj}]^{1/2}$. This statistic also has a limiting chi-squared distribution with $M(M-1)/2$ degrees of freedom. This test has the advantage that it does not require computation of the maximum likelihood estimator of $\boldsymbol{\Sigma}$, because it is based on the OLS residuals.

Example 16.8 ML Estimates of a Seemingly Unrelated Regressions Model

Although a bit dated, the Grunfeld data used in Application 9.1 have withstood the test of time and a[illegible]l the standard data set used to demonstrate the SUR model. The data in Appendix Table F9.3 are for 10 firms and 20 years (1935–1954). For the purpose of this illustration, we will use the first four firms. [The data are downloaded from the [illegible] site for Baltagi (2005), at URL http://www.wiley.com/legacy/wileychi/baltagi/supp/Grunfeld.fil.]

The model is an investment equation:

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

[23] See Attfield (1998) for refinements of this calculation to improve the small sample performance.

14-64

AU: Confirm Table 14.5 is correct number

where

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

TB 14.5

The OLS estimates for the four equations are shown in the left panel of Table ~~16.4~~ 14.5. The correlation matrix for the four OLS residual vectors is

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

14c/

Before turning to the FGLS and MLE estimates, we carry out the LM test against the null hypothesis that the regressions are actually unrelated. We leave as an exercise to show that the LM statistic in (16-85) can be computed as

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

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

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

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

where $\mathbf{E}$ is the 20×4 matrix of OLS residuals. Stacking the data in the partitioned matrices

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

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

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

The estimated asymptotic covariance matrix for the FGLS estimates is the bracketed inverse matrix. These results are shown in the center panel in Table ~~16.4~~ 14.5.

To compute the MLE, we will take advantage of the Oberhofer and Kmenta (1974) result and iterate the FGLS estimator. Using the FGLS coefficient vector, we recompute the residuals, then recompute $\mathbf{W}$, then reestimate β. The iteration is repeated until the estimated parameter vector converges. We use as our convergence measure the following criterion based on the change in the estimated parameter from iteration (s-1) to iteration (s):

minus

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

minus

The sequence of values of this criterion function are: 0.21922, 0.16318, 0.00662, 0.00037, 0.00002367825, 0.000001563348, 0.1041980×10^{-6}. We exit the iterations after iteration 7. The ML estimates are shown in the right panel of Table ~~16.4~~ 14.5

We then carry out the likelihood ratio test of the null hypothesis of a diagonal covariance matrix. The maximum likelihood estimate of Σ is

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

14-65

TABLE ~~16.4~~ 14.5 Estimated Investment Equations

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

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

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

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

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

$$H_0\text{: } \beta_1 = \beta_2 = \beta_3 = \beta_4,$$

we can formulate the hypothesis as

$$H_0\text{: } \beta_1 - \beta_4 = 0, \beta_2 - \beta_4 = 0, \beta_3 - \beta_4 = 0.$$

The Wald statistic is

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

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

14 ~~16~~.9.4 SIMULTANEOUS EQUATIONS MODELS

In Chapter ~~13~~ 10, we noted two approaches to maximum likelihood estimation in the equation system

$$\mathbf{y}_t'\boldsymbol{\Gamma} + \mathbf{x}_t'\mathbf{B} = \boldsymbol{\varepsilon}_t', \qquad (\text{14 } \cancel{16}\text{-86})$$
$$\boldsymbol{\varepsilon}_t \mid \mathbf{X} \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

The limited information maximum likelihood (LIML) estimator is a single-equation approach that estimates the parameters one equation at a time. The full information maximum likelihood (FIML) estimator analyzes the full set of equations at one step.

Derivation of the LIML estimator is quite complicated. Lengthy treatments appear in Anderson and Rubin (1948), Theil (1971), and Davidson and MacKinnon (1993, Chapter 18). The mechanics of the computation are surprisingly simple, as shown ~~earlier~~ (Section ~~13.5.4~~ 10.5.4). The LIML estimates for Klein's Model I appear in ~~Table 13.2 (Section 13.7)~~ Example 10.9 with the other single-equation and system estimators. For the practitioner, a useful result is that the asymptotic variance of the two-stage least squares (2SLS) estimator, which is yet simpler to compute, is the same as that of the LIML estimator. For practical purposes, this would generally render the LIML estimator, with its additional normality assumption, moot. The virtue of the LIML is largely theoretical—it provides a useful benchmark for the analysis of the properties of single-equation estimators. The single exception would be the invariance of the estimator to normalization of the equation (i.e., which variable appears on the left of the equals sign). This turns out to be useful ~~in the~~ context of analysis in the presence of weak instruments. (See Sections ~~12.9~~ 8.7 and ~~13.5.5~~ 10.5.6)

The FIML estimator is much simpler to derive than the LIML and considerably more difficult to implement. ~~The log-likelihood is derived and analyzed in Section 13.6.2.~~ To obtain the needed results, we first operated on the reduced form

$$\mathbf{y}_t' = \mathbf{x}_t'\mathbf{\Pi} + \mathbf{v}_t', \qquad \mathbf{v}_t \,|\, \mathbf{X} \sim N[\mathbf{0}, \mathbf{\Omega}], \tag{~~16~~14-87}$$

which is the seemingly unrelated regressions model analyzed at length in Chapter ~~12~~ 10 and in Section ~~16~~14.9.3. The complication is the restrictions imposed on the parameters,

$$\mathbf{\Pi} = -\mathbf{B}\mathbf{\Gamma}^{-1} \quad \text{and} \quad \mathbf{\Omega} = (\mathbf{\Gamma}^{-1})'\mathbf{\Sigma}(\mathbf{\Gamma}^{-1}). \tag{~~16~~14-88}$$

As is now familiar from several applications, given estimates of $\mathbf{\Gamma}$ and $\mathbf{B}$ in (~~16~~14-86), the estimator of $\mathbf{\Sigma}$ is $(1/T)\mathbf{E}'\mathbf{E}$ based on the residuals. We can even show fairly easily that given $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$, the estimator of $(-\mathbf{B})$ in (~~16~~14-86) would be provided by the results for the SUR model in Section ~~16~~14.9.3.c (where we estimate the model subject to the zero restrictions in the coefficient matrix). The complication in estimation is brought by $\mathbf{\Gamma}$; this is a Jacobian. The term $\ln|\mathbf{\Gamma}|$ appears in the log-likelihood function ~~in Section 13.6.2~~. Nonlinear optimization over the nonzero elements in a function that includes this term is exceedingly complicated. However, three-stage least squares (3SLS) has the same asymptotic efficiency as the FIML estimator, again without the normality assumption and without the practical complications.

The end result is that for the practitioner, the LIML and FIML estimators have been supplanted in the literature by much simpler GMM estimators, 2SLS, H2SLS, 3SLS, and H3SLS. Interest remains in these estimators, but largely as a component of the ongoing theoretical development.

16.9.5 MAXIMUM LIKELIHOOD ESTIMATION OF NONLINEAR REGRESSION MODELS

In Chapter ~~11~~ 7, we considered nonlinear regression models in which the nonlinearity in the parameters appeared entirely on the right-hand side of the equation. Maximum

14-67

likelihood is used when the disturbances in a regression, or the dependent variable, more generally, is not normally distributed. ~~We now consider two applications.~~ The geometric regression model provides an application.

16.9.5.a Nonnormal Disturbances—The Stochastic Frontier Model

This application will examine a regressionlike model in which the disturbances do not have a normal distribution. The model developed here also presents a convenient platform on which to illustrate the use of the invariance property of maximum likelihood estimators to simplify the estimation of the model.

A lengthy literature commencing with theoretical work by Knight (1933), Debreu (1951), and Farrell (1957) and the pioneering empirical study by Aigner, Lovell, and Schmidt (1977) has been directed at models of production that specifically account for the textbook proposition that a production function is a theoretical ideal.[24] If $y = f(\mathbf{x})$ defines a production relationship between inputs, $\mathbf{x}$, and an output, y, then for any given $\mathbf{x}$, the observed value of y must be less than or equal to $f(\mathbf{x})$. The implication for an empirical regression model is that in a formulation such as $y = h(\mathbf{x}, \boldsymbol{\beta}) + u$, u must be negative. Because the theoretical production function is an ideal—the frontier of efficient production—any nonzero disturbance must be interpreted as the result of inefficiency. A strictly orthodox interpretation embedded in a Cobb–Douglas production model might produce an empirical frontier production model such as

$$\ln y = \beta_1 + \Sigma_k \beta_k \ln x_k - u, \quad u \geq 0.$$

The gamma model described in Example 4.9 was an application. One-sided disturbances such as this one present a particularly difficult estimation problem. The primary theoretical problem is that any measurement error in $\ln y$ must be embedded in the disturbance. The practical problem is that the entire estimated function becomes a slave to any single errantly measured data point.

Aigner, Lovell, and Schmidt proposed instead a formulation within which observed deviations from the production function could arise from two sources: (1) productive inefficiency, as we have defined it earlier and that would necessarily be negative, and (2) idiosyncratic effects that are specific to the firm and that could enter the model with either sign. The end result was what they labeled the **stochastic frontier**:

$$\begin{aligned}\ln y &= \beta_1 + \Sigma_k \beta_k \ln x_k - u + v, \quad u \geq 0, \quad v \sim N[0, \sigma_v^2]. \\ &= \beta_1 + \Sigma_k \beta_k \ln x_k + \varepsilon.\end{aligned}$$

The frontier for any particular firm is $h(\mathbf{x}, \boldsymbol{\beta}) + v$, hence the name *stochastic frontier*. The inefficiency term is u, a random variable of particular interest in this setting. Because the data are in log terms, u is a measure of the percentage by which the particular observation fails to achieve the frontier, ideal production rate.

To complete the specification, they suggested two possible distributions for the inefficiency term: the absolute value of a normally distributed variable and an exponentially distributed variable. The density functions for these two compound variables are given by Aigner, Lovell, and Schmidt; let $\varepsilon = v - u$, $\lambda = \sigma_u/\sigma_v$, $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$,

[24] A survey by Greene (2007a) appears in Fried, Lovell, and Schmidt (2007). Kumbhakar and Lovell (2000) is a comprehensive reference on the subject.

14-68

Dez: check that terms are correctly marked for boldface in this section

Example 14.~~10~~9 Identification in a Loglinear Regression Model

In Example ~~7.XX~~ 7.6, we estimated an exponential regression model, of the form

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

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

$$\begin{aligned} E[Income|\mathbf{x}] &= g(\theta)\exp(\gamma_1 + \mathbf{x}'\gamma_2) \\ &= \exp(\gamma_1^* + \mathbf{x}'\gamma_2), \end{aligned}$$

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

(1) Nonlinear least squares (NLS) is robust at least to some failures of the distributional assumption. The nonlinear least squares estimator of γ_2 will be consistent and asymptotically normally distributed in all cases for which $E[Income|\mathbf{x}] = \exp(\gamma_1^* + \mathbf{x}'\gamma_2)$.

(2) The NLS estimator cannot produce a consistent estimator of γ_1; plim$c_1 = \gamma_1^*$, which varies depending on the correct distribution. In the conditional mean function, any pair of values for which $\gamma_1^* = \ln g(\theta) + \gamma_1$ is the same will lead to the same sum of squares. This is a form of multicollinearity; the pseudoregressor for θ is $\partial E[Income|\mathbf{x}]/\partial\theta = \exp(\gamma_1^* + \mathbf{x}'\gamma_2)[g'(\theta)/g(\theta)]$ while that for γ_1 is $\partial E[Income|\mathbf{x}]/\partial\gamma_1 = \exp(\gamma_1^* + \mathbf{x}'\gamma_2)$. The first is a constant multiple of the second.

NLS cannot provide separate estimates of θ and γ_1 while MLE can – see the example to follow. Second, NLS might be less efficient than MLE since it does not use the information about the distribution of the dependent variable. This second consideration is uncertain. For estimation of γ_2, the NLS estimator is less efficient for not using the distributional information. However, that shortcoming might be offset because the NLS estimator does not attempt to compute an independent estimator of the additional parameter, θ.

To illustrate, we reconsider the estimator in Example ~~7.XX~~ 7.6. The gamma regression model specifies

$$f(y\,|\,\mathbf{x}) = \frac{\mu(\mathbf{x})^{\theta}}{\Gamma(\theta)}\exp[-\mu(\mathbf{x})y]y^{\theta-1},\; y > 0,\; \theta > 0,\; \mu(\mathbf{x}) = \exp(-\gamma_1 - \mathbf{x}'\gamma_2).$$

The conditional mean function for this model is

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

TB 14.6

Table 14.6 presents estimates of θ and (γ_1, γ_2). Estimated standard errors appear in parentheses. The estimates in columns (1), (2) and (4) are all computed using nonlinear least squares. In (1), an attempt is made to estimate θ and γ_1 separately. The estimator "converged" on two values. However, the estimated standard errors are essentially infinite. The convergence to anything at all is due to rounding error in the computer. The results in column (2) are for γ_1^* and γ_2. The sums of squares for these two estimates as well as for those in (4) are all 112.19688, indicating that the three results merely show three different sets of results for which γ_1^* is the same. The full maximum likelihood estimates are presented in (3). Note that an estimate of θ is obtained here because the assumed gamma distribution provides another independent moment equation for this parameter, $\partial\ln L/\partial\theta = -n\ln\Psi(\theta) + \Sigma_i(\ln y_i - \ln\mu(\mathbf{x})) = 0$, while the normal equations for the sum of squares provides the same normal equation for θ and γ_1.

minus

Table 14.6 Estimated Gamma Regression Model

	(1) NLS	(2) Constrained NLS	(3) MLE	(4) NLS/MLE
Constant	1.22468	1.69331	3.36826	3.36380
	(47722.5)	(0.04408)	(0.05048)	(0.04408)
Age	-0.00207	-0.00207	-0.00153	-0.00207
	(0.00061)	(0.00061)	(0.00061)	(0.00061)
Education	-0.04792	-0.04792	-0.04975	-0.04792
	(0.00247)	(0.00247)	(0.00286)	(0.00247)
Female	0.00658	0.00658	0.00696	0.00658
	(0.01373)	(0.01373)	(0.01322)	(0.08677)
P	0.62699	--------	5.31474	5.31474
	(29921.3)	--------	(0.10894)	(0.00000)

14-70

TABLE 16.5 Estimated Stochastic Frontier Functions

	Least Squares			Half-Normal Model			Exponential Model		
Coefficient	Estimate	Standard Error	t Ratio	Estimate	Standard Error[a]	t Ratio	Estimate	Standard Error[a]	t Ratio
Constant	1.844	0.234	7.896	2.081	0.422	4.933	2.069	0.290	7.135
β_k	0.245	0.107	2.297	0.259	0.144	1.800	0.262	0.120	2.184
β_l	0.805	0.126	6.373	0.780	0.170	4.595	0.770	0.138	5.581
σ	0.236			0.282	0.087	3.237			
σ_u	—			0.222			0.136		
σ_v	—			0.175			0.171	0.054	3.170
λ	—			1.265	1.620	0.781			
θ	—						7.398	3.931	1.882
log L	2.2537			2.4695			2.8605		

[a] Based on BHHH estimator. Using second derivatives, standard errors would be (0.232, 0.098, 0.116, 0.0082, 0.557) for the half-normal and (0.236, 0.092, 0.111, 0.038, 3.431) for the exponential. The t ratios would be adjusted accordingly.

TABLE 16.6 Estimated Inefficiencies

State	Half-Normal	Exponential	State	Half-Normal	Exponential
Alabama	0.2011	0.1459	Maryland	0.1353	0.0925
California	0.1448	0.0972	Massachusetts	0.1564	0.1093
Connecticut	0.1903	0.1348	Michigan	0.1581	0.1076
Florida	0.5175	0.5903	Missouri	0.1029	0.0704
Georgia	0.1040	0.0714	New Jersey	0.0958	0.0659
Illinois	0.1213	0.0830	New York	0.2779	0.2225
Indiana	0.2113	0.1545	Ohio	0.2291	0.1698
Iowa	0.2493	0.2007	Pennsylvania	0.1501	0.1030
Kansas	0.1010	0.0686	Texas	0.2030	0.1455
Kentucky	0.0563	0.0415	Virginia	0.1400	0.0968
Louisiana	0.2033	0.1507	Washington	0.1105	0.0753
Maine	0.2226	0.1725	West Virginia	0.1556	0.1124
Wisconsin	0.1407	0.0971			

~~16.9.5.b ML Estimation of a Geometric Regression Model for Count Data~~

new paragraph

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

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

KT

which has **loglinear conditional mean** function $E[y_i \mid \mathbf{x}_i] = \lambda_i$. (The Poisson regression model and other specifications for data on counts are discussed at length in Chapter ~~25~~ 19. We introduce the topic here to begin development ~~of the~~ MLE in a fairly straightforward, typical nonlinear setting.) Appendix Table F11.1 presents the Riphahn et al. (2003) data, which we will use to analyze a count variable, *DocVis*, the number of visits to physicans in the survey year. The histogram in Figure 16.5 shows a distinct spike at zero followed by rapidly declining frequencies. While the Poisson distribution, which

FIG 14.5

F7.1 14

14-71

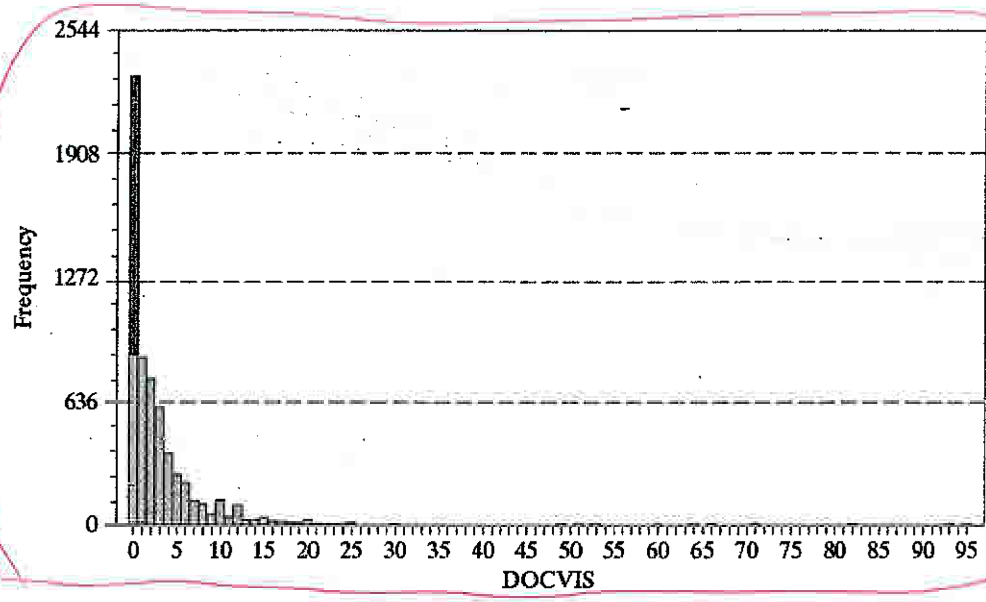

FIGURE ~~16~~.5 Histogram for Doctor Visits.
14

19

is typically hump-shaped, can accommodate this configuration if λ_i is less than one, the shape is nonetheless somewhat "non-Poisson." [So-called Zero Inflation models (discussed in Chapter ~~25~~) are often used for this situation.]

The geometric distribution,

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

is a convenient specification that produces the effect shown in Figure ~~16~~14.5. (Note that, formally, the specification is used to model the number of failures before the first success in successive independent trials each with success probability θ_i, so in fact, it is misspecified as a model for counts. The model does provide a convenient and useful illustration, however.) The conditional mean function is also $E[y_i \mid \mathbf{x}_i] = \lambda_i$. The partial effects in the model are

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i \boldsymbol{\beta},$$

so this is a distinctly nonlinear regression model. We will construct a maximum likelihood estimator, then compare the MLE to the **nonlinear least squares** and (misspecified) linear least squares estimates.

The log-likelihood function is

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

14-72

The likelihood equations are

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

Because

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

the likelihood equations simplify to

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

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

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

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

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

Finally, as noted earlier, $E[y_i \mid x_i] = \lambda_i = (1-\theta_i)/\theta_i$, is known, so we can also use the negative inverse of the expected second derivatives matrix,

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

AU: This term was already a KT on ms p 14-52. Here also?

To compute the estimates of the parameters, either **Newton's method**,

$$\hat{\beta}^{t+1} = \hat{\beta}^t - [\hat{\mathbf{H}}^t]^{-1} \hat{\mathbf{g}}^t,$$

or the method of scoring,

$$\hat{\beta}^{t+1} = \hat{\beta}^t - \{E[\hat{\mathbf{H}}^t]\}^{-1} \hat{\mathbf{g}}^t,$$

can be used, where $\mathbf{H}$ and $\mathbf{g}$ are the second and first derivatives that will be evaluated at the current estimates of the parameters. Like many models of this sort, there is a convenient set of starting values, assuming the model contains a constant term. Because $E[y_i \mid x_i] = \lambda_i$, if we start the slope parameters at zero, then a natural starting value for the constant term is the log of $\bar{y}$.

14-73

14.10

F7.1

Example ~~16.10~~ Geometric Model for Doctor Visits 14

11.14 In Example 11.10, we considered nonlinear least squares estimation of a loglinear model for the number of doctor visits variable shown in Figure ~~16.5~~. The data are drawn from the Riphahn et al. (2003) data set in Appendix Table F11.1. We will continue that analysis here by fitting a more detailed model for the count variable *DocVis*. The conditional mean analyzed here is

$$\ln E[DocVis_{it} \mid \mathbf{x}_{it}] = \beta_1 + \beta_2\, Age_{it} + \beta_3\, Educ_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it}$$ 11.14

(This differs slightly from the model in Example 11.10. For this exercise, with an eye toward the fixed effects model in Example ~~16.13~~ 14), we have specified a model that does not contain any time invariant variables, such as $Female_i$.) Sample means for the variables in the model are given in Table ~~16.7~~ 14. Note, these data are a panel. In this exercise, we are ignoring that fact, and fitting a pooled model. We will turn to panel data treatments in the next section, and revisit this application.

13 14.7

We used Newton's method for the optimization, with starting values as suggested earlier. The five iterations are as follows:

Variable	Constant	Age	Educ	Income	Kids
Start values:	.11580e+01	.00000e+00	.00000e+00	.00000e+00	.00000e+00
1st derivs.	−.25191e−08	−.61777e+05	.73202e+04	.42575e+04	.16464e+04
Parameters:	.11580e+01	.00000e+00	.00000e+00	.00000e+00	.00000e+00
Iteration 1 F =	.6287e+05	g′inv(H)g =	.4367e+02		
1st derivs.	.48616e+03	−.22449e+05	−.57162e+04	−.17112e+04	−.16521e+03
Parameters:	.11186e+01	.17563e−01	−.50263e−01	−.46274e−01	−.15609e+00
Iteration 2 F =	.6192e+05	g′inv(H)g =	.3547e+01		
1st derivs.	−.31284e+01	−.15595e+03	−.37197e+02	−.10630e+02	−.77186e+00
Parameters:	.10922e+01	.17981e−01	−.47303e−01	−.46739e−01	−.15683e+00
Iteration 3 F=	.6192e+05	g′inv(H)g =	.2598e−01		
1st derivs.	−.18417e−03	−.99368e−02	−.21992e−02	−.59354e−03	−.25994e−04
Parameters:	.10918e+01	.17988e−01	−.47274e−01	−.46751e−01	−.15686e+00
Iteration 4 F=	.6192e+05	g′inv(H)g =	.1831e−05		
1st derivs.	−.35727e−11	.86745e−10	−.26302e−10	−.61006e−11	−.15620e−11
Parameters:	.10918e+01	.17988e−01	−.47274e−01	−.46751e−01	−.15686e+00
Iteration 5 F=	.6192e+05	g′inv(H)g =	.1772e−12		

Convergence based on the LM criterion, $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ is achieved after the fourth iteration. Note that the derivatives at this point are extremely small, albeit not absolutely zero. Table ~~16.7~~ 14.7 presents the maximum likelihood estimates of the parameters. Several sets of standard errors are presented. The three sets based on different estimators of the information matrix are presented first. The fourth set are based on the cluster corrected covariance matrix discussed in Section ~~16.~~8.4. 14 Because this is actually an (unbalanced) panel data set, we anticipate correlation across observations. Not surprisingly, the standard errors rise substantially. The

TABLE ~~16~~.7 14 Estimated Geometric Regression Model Dependent Variable: DocVis: Mean = 3.18352, Standard Deviation = 5.68969

Variable	Estimate	St. Er H	St. Er. E[H]	St. Er. BHHH	St. Er. Cluster	APE	PE Mean	OLS	Mean
Constant	1.0918	0.0524	0.0524	0.0354	0.1112	—	—	2.656	
Age	0.0180	0.0007	0.0007	0.0005	0.0013	0.0572	0.0547	0.061	43.52
Education	−0.0473	0.0033	0.0033	0.0023	0.0069	−0.150	−0.144	−0.121	11.32
Income	−0.0468	0.0041	0.0042	0.0023	0.0075	−0.149	−0.142	−0.162	3.52
Kids	−0.1569	0.0156	0.0155	0.0103	0.0319	−0.499	−0.477	−0.517	0.40

14-74

partial effects listed next are computed in two ways. The "Average Partial Effect" is computed by averaging $\lambda_i \beta$ across the individuals in the sample. The "Partial Effect" is computed for the average individual by computing λ at the means of the data. The next-to-last column contains the ordinary least squares coefficients. In this model, there is no reason to expect ordinary least squares to provide a consistent estimator of β. The question might arise, What does ordinary least squares estimate? The answer is the slopes of the linear projection of DocVis on x_{it}. The resemblance of the OLS coefficients to the estimated partial effects is more than coincidental, and suggests an answer to the question.

The analysis in the table suggests three competing approaches to modeling DocVis. The results for the geometric regression model are given in Table ~~16~~14.7. At the beginning of this section, we noted that the more conventional approach to modeling a count variable such as DocVis is with the Poisson regression model. The log-likelihood function and its derivatives are even simpler than the geometric model,

$$\ln L = \sum_{i=1}^{n} y_i \ln \lambda_i - \lambda_i - \ln y_i!,$$

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

$$\partial^2 \ln L / \partial \beta \partial \beta' = \sum_{i=1}^{n} -\lambda_i \mathbf{x}_i \mathbf{x}_i'.$$

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

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

This is, in fact, the model that applies to both the geometric and Poisson cases. Under either distributional assumption, nonlinear least squares is inefficient compared to MLE. But, the distributional assumption can be dropped altogether, and the model fit as a simple exponential regression. Table ~~16~~14.8 presents the three sets of estimates.

173 14.8 14

It is not obvious how to choose among the alternatives. Of the three, the Poisson model is used most often by far. The Poisson and geometric models are not nested, so we cannot use a simple parametric test to choose between them. However, these two models will surely fit the conditions for the Vuong test described in Section ~~13.4~~ 14.6.6. To implement the test, we first computed

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

Section 14.6.6

using the respective MLEs of the parameters. The test statistic given in (~~7-14~~) is then

$$V = \frac{\left(\sqrt{\sum_{i=1}^{n} T_i}\right) \bar{V}}{s_V}.$$

TABLE ~~16~~14.8 Estimates of Three Models for DOCVIS

	Geometric Model		Poisson Model		Nonlinear Reg.	
Variable	*Estimate*	*St. Er*	*Estimate*	*St. Er.*	*Estimate*	*St. Er.*
Constant	1.0918	0.0524	1.0480	0.0272	0.9801	0.0893
Age	0.0180	0.0007	0.0184	0.0003	0.0187	0.0011
Education	−0.0473	0.0033	−0.0433	0.0017	−0.0361	0.0057
Income	−0.0468	0.0041	−0.0520	0.0022	−0.0591	0.0072
Kids	−0.1569	0.0156	−0.1609	0.0080	−0.1692	0.0264

This statistic converges to standard normal under the underlying assumptions. A large positive value favors the geometric model. The computed sample value is 37.885, which strongly favors the geometric model over the Poisson.

14 ~~16~~.9.6 PANEL DATA APPLICATIONS

Application of panel data methods to the linear panel data models we have considered so far is a fairly marginal extension. For the random effects linear model, considered in the following Section ~~16~~ 14.9.6.a, the MLE of β is, as always, FGLS given the MLEs of the variance parameters. The latter produce a fairly substantial complication, as we shall see. This extension does provide a convenient, interesting application to see the payoff to the invariance property of the MLE—we will reparameterize a fairly complicated log-likelihood function to turn it into a simple one. Where the method of maximum likelihood becomes essential is in analysis of fixed and random effects in nonlinear models. We will develop two general methods for handling these situations in generic terms in Sections ~~16~~ 14.9.6.b and ~~16~~ 14.9.6.c, then apply them in several models later in the book.

14 ~~16~~.9.6.a ML Estimation of the Linear Random Effects Model

The contribution of the ith individual to the log-likelihood for the random effects model [(~~9~~ 11-26) to (~~9~~ 11-29)] with normally distributed disturbances is

$$\begin{aligned}\ln L_i\left(\beta,\sigma_\varepsilon^2,\sigma_u^2\right) &= \frac{-1}{2}\left[T_i\ln 2\pi + \ln|\boldsymbol{\Omega}_i| + (\mathbf{y}_i-\mathbf{X}_i\beta)'\boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\beta)\right] \\ &= \frac{-1}{2}\left[T_i\ln 2\pi + \ln|\boldsymbol{\Omega}_i| + \boldsymbol{\varepsilon}_i'\boldsymbol{\Omega}_i^{-1}\boldsymbol{\varepsilon}_i\right],\end{aligned} \qquad (\cancel{16}\ 14\text{-}89)$$

where

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

and $\mathbf{i}$ denotes a $T_i \times 1$ column of ones. Note that the $\boldsymbol{\Omega}_i$ varies over i because it is $T_i \times T_i$. Baltagi (2005, pp. 19–20) presents a convenient and compact estimator for this model that involves iteration between an estimator of $\phi^2 = \left[\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + T\sigma_u^2)\right]$, based on sums of squared residuals, and $(\alpha, \beta, \sigma_\varepsilon^2)$ (α is the constant term) using FGLS. Unfortunately, the convenience and compactness come unraveled in the unbalanced case. We consider, instead, what Baltagi labels a "brute force" approach, that is, direct maximization of the log-likelihood function in (~~16~~ 14-89). (See, op. cit, pp. 169–170.)

Using (A-66), we find (in (~~9~~ 11-28) that

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

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

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

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

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