14
SYSTEMS OF REGRESSION EQUATIONS

14.1 INTRODUCTION

There are many settings in which the models of the previous chapters apply to a group of related variables. In these contexts, it makes sense to consider the several models jointly. Some examples follow.

1. The capital asset pricing model of finance specifies that for a given security,

$$r_{it} - r_{ft} = \alpha_i + \beta_i(r_{mt} - r_{ft}) + \varepsilon_{it},$$

where r_{it} is the return over period t on security i, r_{ft} is the return on a risk-free security, r_{mt} is the market return, and β_i is the security's beta coefficient. The disturbances are obviously correlated across securities. The knowledge that the return on security i exceeds the risk-free rate by a given amount gives some information about the excess return of security j, at least for some j's. It may be useful to estimate the equations jointly rather than ignore this connection.

2. In the Grunfeld–Boot and de Witt investment model of Section 13.9.7, we examined a set of firms, each of which makes investment decisions based on variables that reflect anticipated profit and replacement of the capital stock. We will now specify

$$I_{it} = \beta_{1i} + \beta_{2i}F_{it} + \beta_{3i}C_{it} + \varepsilon_{it}.$$

Whether the parameter vector should be the same for all firms is a question that we shall study in this chapter. But the disturbances in the investment equations certainly include factors that are common to all the firms, such as the perceived general health of the economy, as well as factors that are specific to the particular firm or industry.

3. In a model of production, the optimization conditions of economic theory imply that if a firm faces a set of factor prices $\mathbf{p}$, then its set of cost-minimizing factor demands for producing output Y will be a set of equations of the form $x_m = f_m(Y, \mathbf{p})$. The model is

$$\begin{aligned} x_1 &= f_1(Y, \mathbf{p} : \boldsymbol{\theta}) + \varepsilon_1, \\ x_2 &= f_2(Y, \mathbf{p} : \boldsymbol{\theta}) + \varepsilon_2, \\ &\cdots \\ x_M &= f_M(Y, \mathbf{p} : \boldsymbol{\theta}) + \varepsilon_M. \end{aligned}$$

Once again, the disturbances should be correlated. In addition, the same parameters of the production technology will enter all the demand equations, so the set of equations have cross-equation restrictions. Estimating the equations separately will waste the information that the same set of parameters appears in all the equations.

All these examples have a common multiple equation structure, which we may write as

$$\begin{aligned}
\mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1,\\
\mathbf{y}_2 &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2,\\
&\;\;\vdots\\
\mathbf{y}_M &= \mathbf{X}_M\boldsymbol{\beta}_M + \boldsymbol{\varepsilon}_M.
\end{aligned} \tag{14-1}$$

There are M equations and T observations in the sample of data used to estimate them.[1] The second and third examples embody different types of constraints across equations and different structures of the disturbances. A basic set of principles will apply to them all, however.[2]

Section 14.2 below examines the general model in which each equation has its own fixed set of parameters, and examines efficient estimation techniques. Production and consumer demand models are a special case of the general model in which the equations of the model obey an adding up constraint that has important implications for specification and estimation. Some general results for demand systems are considered in Section 14.3. In Section 14.4 we examine a classic application of the model in Section 14.3 that illustrates a number of the interesting features of the current genre of demand studies in the applied literature. Section 14.4 introduces estimation of nonlinear systems, instrumental variable estimation, and GMM estimation for a system of equations.

Example 14.1 Grunfeld's Investment Data

To illustrate the techniques to be developed in this chapter, we will use the Grunfeld data first examined in Section 13.9.7 in the previous chapter. Grunfeld's model is now

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

where i indexes firms, t indexes years, and

$$\begin{aligned}
I_{it} &= \text{gross investment},\\
F_{it} &= \text{market value of the firm at the end of the previous year},\\
C_{it} &= \text{value of the stock of plant and equipment at the end of the previous year}.
\end{aligned}$$

All figures are in millions of dollars. The sample consists of 20 years of observations (1935–1954) on five firms. The model extension we consider in this chapter is to allow the coefficients to vary across firms in an unstructured fashion.

14.2 THE SEEMINGLY UNRELATED REGRESSIONS MODEL

The **seemingly unrelated regressions** (SUR) model in (14-1) is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, M, \tag{14-2}$$

[1] The use of T is not necessarily meant to imply any connection to time series. For instance, in the third example above, the data might be cross-sectional.

[2] See the surveys by Srivastava and Dwivedi (1979), Srivastava and Giles (1987), and Feibig (2001).

where

$$\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \dots, \boldsymbol{\varepsilon}_M']'$$

and

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \mathbf{0},$$
$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \boldsymbol{\Omega}.$$

We assume that a total of T observations are used in estimating the parameters of the M equations.[3] Each equation involves K_m regressors, for a total of $K = \sum_{i=1}^{n} K_i$. We will require $T > K_i$. The data are assumed to be well behaved, as described in Section 5.2.1, and we shall not treat the issue separately here. For the present, we also assume that disturbances are uncorrelated across observations. Therefore,

$$E[\varepsilon_{it}\varepsilon_{js} \mid \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \sigma_{ij}, \quad \text{if } t = s \text{ and 0 otherwise.}$$

The disturbance formulation is therefore

$$E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j' \mid \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \sigma_{ij}\mathbf{I}_T$$

or

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots & \sigma_{1M}\mathbf{I} \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots & \sigma_{2M}\mathbf{I} \\ & & \vdots & \\ \sigma_{M1}\mathbf{I} & \sigma_{M2}\mathbf{I} & \cdots & \sigma_{MM}\mathbf{I} \end{bmatrix}. \tag{14-3}$$

Note that when the data matrices are group specific observations on the same variables, as in Example 14.1, the specification of this model is precisely that of the covariance structures model of Section 13.9 save for the extension here that allows the parameter vector to vary across groups. The covariance structures model is, therefore, a testable special case.[4]

It will be convenient in the discussion below to have a term for the particular kind of model in which the data matrices are group specific data sets on the same set of variables. The Grunfeld model noted in Example 14.1 is such a case. This special case of the seemingly unrelated regressions model is a **multivariate regression model.** In contrast, the cost function model examined in Section 14.5 is not of this type—it consists of a cost function that involves output and prices and a set of cost share equations that have only a set of constant terms. We emphasize, this is merely a convenient term for a specific form of the SUR model, not a modification of the model itself.

14.2.1 GENERALIZED LEAST SQUARES

Each equation is, by itself, a classical regression. Therefore, the parameters could be estimated consistently, if not efficiently, one equation at a time by ordinary least squares.

[3]There are a few results for unequal numbers of observations, such as Schmidt (1977), Baltagi, Garvin, and Kerman (1989), Conniffe (1985), Hwang, (1990) and Im (1994). But generally, the case of fixed T is the norm in practice.

[4]This is the test of "Aggregation Bias" that is the subject of Zellner (1962, 1963). (The bias results if parameter equality is incorrectly assumed.)

The generalized regression model applies to the stacked model,

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_M\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0}\\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_M\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_M\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\\ \vdots\\ \boldsymbol{\varepsilon}_M\end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{14-4}$$

Therefore, the efficient estimator is generalized least squares.[5] The model has a particularly convenient form. For the tth observation, the $M \times M$ covariance matrix of the disturbances is

$$\boldsymbol{\Sigma} = \begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M}\\ & \vdots & & \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM}\end{bmatrix}, \tag{14-5}$$

so, in (14-3),

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}$$

and

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}. \tag{14-6}$$

Denoting the ijth element of $\boldsymbol{\Sigma}^{-1}$ by σ^{ij}, we find that the GLS estimator is

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = [\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{X}]^{-1}\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}.$$

Expanding the **Kronecker products** produces

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix}\sigma^{11}\mathbf{X}_1'\mathbf{X}_1 & \sigma^{12}\mathbf{X}_1'\mathbf{X}_2 & \cdots & \sigma^{1M}\mathbf{X}_1'\mathbf{X}_M\\ \sigma^{21}\mathbf{X}_2'\mathbf{X}_1 & \sigma^{22}\mathbf{X}_2'\mathbf{X}_2 & \cdots & \sigma^{2M}\mathbf{X}_2'\mathbf{X}_M\\ & & \vdots & \\ \sigma^{M1}\mathbf{X}_M'\mathbf{X}_1 & \sigma^{M2}\mathbf{X}_M'\mathbf{X}_2 & \cdots & \sigma^{MM}\mathbf{X}_M'\mathbf{X}_M\end{bmatrix}^{-1}\begin{bmatrix}\sum_{j=1}^{M}\sigma^{1j}\mathbf{X}_1'\mathbf{y}_j\\ \sum_{j=1}^{M}\sigma^{2j}\mathbf{X}_2'\mathbf{y}_j\\ \vdots\\ \sum_{j=1}^{M}\sigma^{Mj}\mathbf{X}_M'\mathbf{y}_j\end{bmatrix}. \tag{14-7}$$

The asymptotic covariance matrix for the GLS estimator is the inverse matrix in (14-7). All the results of Chapter 10 for the generalized regression model extend to this model (which has both heteroscedasticity and "autocorrelation").

This estimator is obviously different from ordinary least squares. At this point, however, the equations are linked only by their disturbances—hence the name *seemingly unrelated* regressions model—so it is interesting to ask just how much efficiency is gained by using generalized least squares instead of ordinary least squares. Zellner (1962) and Dwivedi and Srivastava (1978) have analyzed some special cases in detail.

[5] See Zellner (1962) and Telser (1964).

1. If the equations are actually unrelated—that is, if $\sigma_{ij} = 0$ for $i \neq j$—then there is obviously no payoff to GLS estimation of the full set of equations. Indeed, full GLS is equation by equation OLS.[6]
2. If the equations have identical explanatory variables—that is, if $\mathbf{X}_i = \mathbf{X}_j$—then OLS and GLS are identical. We will turn to this case in Section 14.2.2 and then examine an important application in Section 14.2.5.[7]
3. If the regressors in one block of equations are a subset of those in another, then GLS brings no efficiency gain over OLS in estimation of the smaller set of equations; thus, GLS and OLS are once again identical. We will look at an application of this result in Section 19.6.5.[8]

In the more general case, with unrestricted correlation of the disturbances and different regressors in the equations, the results are complicated and dependent on the data. Two propositions that apply generally are as follows:

1. The greater is the correlation of the disturbances, the greater is the efficiency gain accruing to GLS.
2. The less correlation there is between the $\mathbf{X}$ matrices, the greater is the gain in efficiency in using GLS.[9]

14.2.2 SEEMINGLY UNRELATED REGRESSIONS WITH IDENTICAL REGRESSORS

The case of **identical regressors** is quite common, notably in the capital asset pricing model in empirical finance—see Section 14.2.5. In this special case, generalized least squares is equivalent to equation by equation ordinary least squares. Impose the assumption that $\mathbf{X}_i = \mathbf{X}_j = \mathbf{X}$, so that $\mathbf{X}_i'\mathbf{X}_j = \mathbf{X}'\mathbf{X}$ for all i and j in (14-7). The inverse matrix on the right-hand side now becomes $[\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X}]^{-1}$, which, using (A-76), equals $[\boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}]$. Also on the right-hand side, each term $\mathbf{X}_i'\mathbf{y}_j$ equals $\mathbf{X}'\mathbf{y}_j$, which, in turn equals $\mathbf{X}'\mathbf{X}\mathbf{b}_j$. With these results, after moving the common $\mathbf{X}'\mathbf{X}$ out of the summations on the right-hand side, we obtain

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sigma_{11}(\mathbf{X}'\mathbf{X})^{-1} & \sigma_{12}(\mathbf{X}'\mathbf{X})^{-1} & \cdots & \sigma_{1M}(\mathbf{X}'\mathbf{X})^{-1} \\ \sigma_{21}(\mathbf{X}'\mathbf{X})^{-1} & \sigma_{22}(\mathbf{X}'\mathbf{X})^{-1} & \cdots & \sigma_{2M}(\mathbf{X}'\mathbf{X})^{-1} \\ & & \vdots & \\ \sigma_{M1}(\mathbf{X}'\mathbf{X})^{-1} & \sigma_{M2}(\mathbf{X}'\mathbf{X})^{-1} & \cdots & \sigma_{MM}(\mathbf{X}'\mathbf{X})^{-1} \end{bmatrix} \begin{bmatrix} (\mathbf{X}'\mathbf{X}) \sum_{l=1}^{M} \sigma^{1l}\mathbf{b}_l \\ (\mathbf{X}'\mathbf{X}) \sum_{l=1}^{M} \sigma^{2l}\mathbf{b}_l \\ \vdots \\ (\mathbf{X}'\mathbf{X}) \sum_{l=1}^{M} \sigma^{Ml}\mathbf{b}_l \end{bmatrix}. \tag{14-8}$$

[6]See also Baltagi (1989) and Bartels and Feibig (1991) for other cases in which OLS = GLS.

[7]An intriguing result, albeit probably of negligible practical significance, is that the result also applies if the $\mathbf{X}$'s are all nonsingular, and not necessarily identical, linear combinations of the same set of variables. The formal result which is a corollary of Kruskal's Theorem [see Davidson and MacKinnon (1993, p. 294)] is that OLS and GLS will be the same if the K columns of $\mathbf{X}$ are a linear combination of exactly K characteristic vectors of $\boldsymbol{\Omega}$. By showing the equality of OLS and GLS here, we have verified the conditions of the corollary. The general result is pursued in the exercises. The intriguing result cited is now an obvious case.

[8]The result was analyzed by Goldberger (1970) and later by Revankar (1974) and Conniffe (1982a, b).

[9]See also Binkley (1982) and Binkley and Nelson (1988).

Now, we isolate one of the subvectors, say the first, from $\hat{\boldsymbol{\beta}}$. After multiplication, the moment matrices cancel, and we are left with

$$\hat{\boldsymbol{\beta}}_1 = \sum_{j=1}^{M} \sigma_{1j} \sum_{l=1}^{M} \sigma^{j1}\mathbf{b}_l = \mathbf{b}_1\left(\sum_{j=1}^{M}\sigma_{1j}\sigma^{j1}\right) + \mathbf{b}_2\left(\sum_{j=1}^{M}\sigma_{1j}\sigma^{j2}\right) + \cdots + \mathbf{b}_M\left(\sum_{j=1}^{M}\sigma_{1j}\sigma^{jM}\right).$$

The terms in parentheses are the elements of the first row of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \mathbf{I}$, so the end result is $\hat{\boldsymbol{\beta}}_1 = \mathbf{b}_1$. For the remaining subvectors, which are obtained the same way, $\hat{\boldsymbol{\beta}}_i = \mathbf{b}_i$, which is the result we sought.[10]

To reiterate, the important result we have here is that in the SUR model, when all equations have the same regressors, the efficient estimator is single-equation ordinary least squares; OLS is the same as GLS. Also, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ for this case is given by the large inverse matrix in brackets in (14-8), which would be estimated by

$$\text{Est.Asy. Cov}[\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j] = \hat{\sigma}_{ij}(\mathbf{X}'\mathbf{X})^{-1}, \quad i, j = 1, \ldots, M, \quad \text{where } \hat{\boldsymbol{\Sigma}}_{ij} = \hat{\sigma}_{ij} = \frac{1}{T}\mathbf{e}_i'\mathbf{e}_j.$$

Except in some special cases, this general result is lost if there are any restrictions on $\boldsymbol{\beta}$, either within or across equations. We will examine one of those cases, the block of zeros restriction, in Sections 14.2.6 and 19.6.5.

14.2.3 FEASIBLE GENERALIZED LEAST SQUARES

The preceding discussion assumes that $\boldsymbol{\Sigma}$ is known, which, as usual, is unlikely to be the case. FGLS estimators have been devised, however.[11] The least squares residuals may be used (of course) to estimate consistently the elements of $\boldsymbol{\Sigma}$ with

$$\hat{\sigma}_{ij} = s_{ij} = \frac{\mathbf{e}_i'\mathbf{e}_j}{T}. \tag{14-9}$$

The consistency of s_{ij} follows from that of $\mathbf{b}_i$ and $\mathbf{b}_j$. A degrees of freedom correction in the divisor is occasionally suggested. Two possibilities are

$$s_{ij}^{*} = \frac{\mathbf{e}_i'\mathbf{e}_j}{[(T-K_i)(T-K_j)]^{1/2}} \quad \text{and} \quad s_{ij}^{**} = \frac{\mathbf{e}_i'\mathbf{e}_j}{T - \max(K_i, K_j)}.$$[12]

The second is unbiased only if i equals j or K_i equals K_j, whereas the first is unbiased only if i equals j. Whether unbiasedness of the estimate of $\boldsymbol{\Sigma}$ used for FGLS is a virtue here is uncertain. The asymptotic properties of the **feasible GLS** estimator, $\hat{\hat{\boldsymbol{\beta}}}$ do not rely on an unbiased estimator of $\boldsymbol{\Sigma}$; only consistency is required. All our results from Chapters 10–13 for FGLS estimators extend to this model, with no modification. We

[10] See Hashimoto and Ohtani (1996) for discussion of hypothesis testing in this case.

[11] See Zellner (1962) and Zellner and Huang (1962).

[12] See, as well, Judge et al. (1985), Theil (1971) and Srivistava and Giles (1987).

shall use (14-9) in what follows. With

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & s_{22} & \cdots & s_{2M} \\ & \vdots & & \\ s_{M1} & s_{M2} & \cdots & s_{MM} \end{bmatrix} \tag{14-10}$$

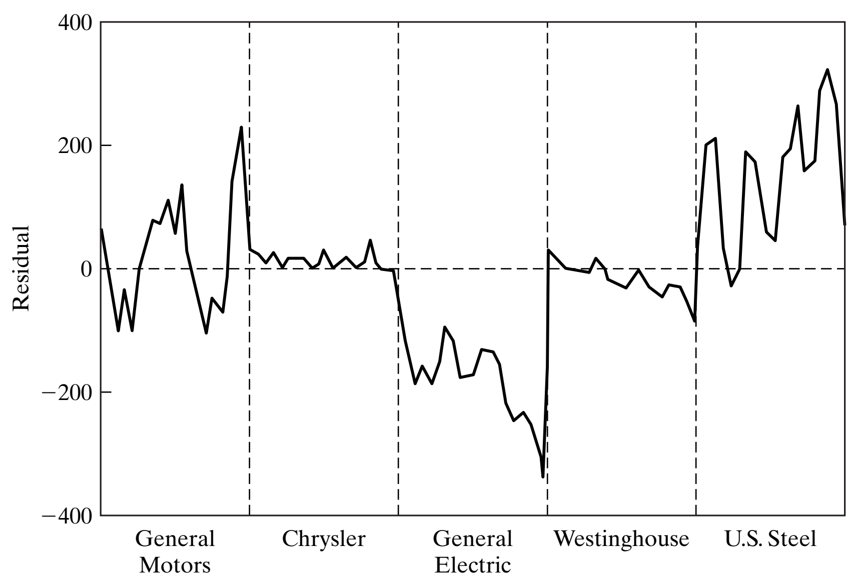
in hand, FGLS can proceed as usual. Iterated FGLS will be maximum likelihood if it is based on (14-9).

Goodness-of-fit measures for the system have been devised. For instance, McElroy (1977) suggested the systemwide measure

$$R_*^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{\varepsilon}}}{\sum_{i=1}^{M}\sum_{j=1}^{M}\hat{\sigma}^{ij}\left[\sum_{t=1}^{T}(y_{it}-\bar{y}_i)(y_{jt}-\bar{y}_j)\right]} = 1 - \frac{M}{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}_{yy})}, \tag{14-11}$$

where ˆ indicates the FGLS estimate. (The advantage of the second formulation is that it involves $M \times M$ matrices, which are typically quite small, whereas $\hat{\boldsymbol{\Omega}}$ is $MT \times MT$. In our case, M equals 5, but MT equals 100.) The measure is bounded by 0 and 1 and is related to the F statistic used to test the hypothesis that all the slopes in the model are zero. Fit measures in this generalized regression model have all the shortcomings discussed in Section 10.5.1. An additional problem for this model is that overall fit measures such as that in (14-11) will obscure the variation in fit across equations. For the investment example, using the FGLS residuals for the least restrictive model in Table 13.4 (the covariance structures model with identical coefficient vectors), McElroy's measure gives a value of 0.846. But as can be seen in Figure 14.1, this apparently good

FIGURE 14.1 FGLS Residuals with Equality Restrictions.



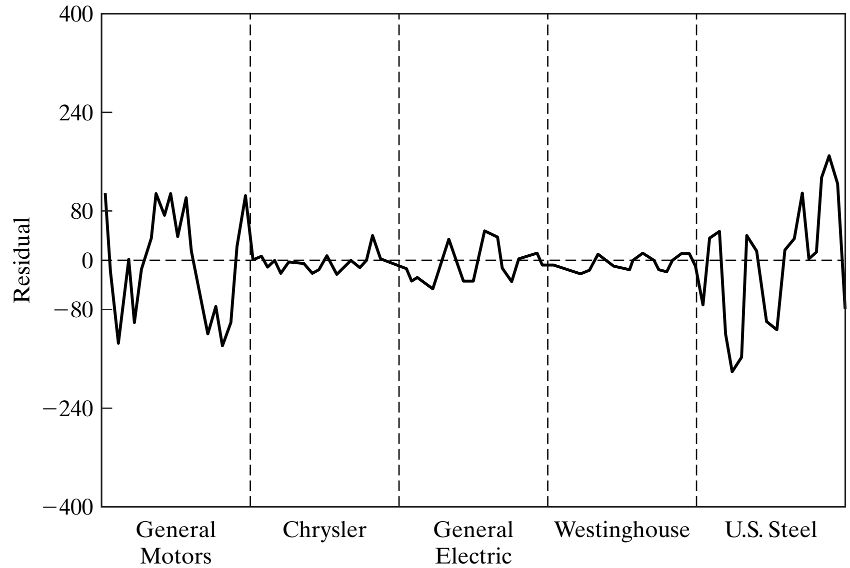


FIGURE 14.2 SUR Residuals.

overall fit is an aggregate of mediocre fits for Chrysler and Westinghouse and obviously terrible fits for GM, GE, and U.S. Steel. Indeed, the conventional measure for GE based on the same FGLS residuals, $1 - \mathbf{e}'_{GE}\mathbf{e}_{GE}/\mathbf{y}'_{GE}\mathbf{M}^0\mathbf{y}_{GE}$ is -16.7!

We might use (14-11) to compare the fit of the unrestricted model with separate coefficient vectors for each firm with the restricted one with a common coefficient vector. The result in (14-11) with the FGLS residuals based on the seemingly unrelated regression estimates in Table 14.1 (in Example 14.2) gives a value of 0.871, which compared to 0.846 appears to be an unimpressive improvement in the fit of the model. But a comparison of the residual plot in Figure 14.2 with that in Figure 14.1 shows that, on the contrary, the fit of the model has improved dramatically. The upshot is that although a fit measure for the system might have some virtue as a descriptive measure, it should be used with care.

For testing a hypothesis about $\boldsymbol{\beta}$, a statistic analogous to the F ratio in multiple regression analysis is

$$F[J, MT - K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{\varepsilon}}/(MT - K)}. \tag{14-12}$$

The computation requires the unknown $\boldsymbol{\Omega}$. If we insert the FGLS estimate $\hat{\boldsymbol{\Omega}}$ based on (14-9) and use the result that the denominator converges to one, then, in large samples, the statistic will behave the same as

$$\hat{F} = \frac{1}{J}(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q})'[\mathbf{R}\,\widehat{\mathrm{Var}}[\hat{\hat{\boldsymbol{\beta}}}]\mathbf{R}']^{-1}(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}). \tag{14-13}$$

This can be referred to the standard F table. Because it uses the estimated $\boldsymbol{\Sigma}$, even with normally distributed disturbances, the F distribution is only valid approximately. In general, the statistic $F[J, n]$ converges to $1/J$ times a chi-squared $[J]$ as $n \to \infty$.

Therefore, an alternative test statistic that has a limiting chi-squared distribution with J degrees of freedom when the hypothesis is true is

$$J\hat{F} = (\mathbf{R}\,\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q})'[\mathbf{R}\widehat{\mathrm{Var}}[\hat{\hat{\boldsymbol{\beta}}}]\mathbf{R}']^{-1}(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}). \tag{14-14}$$

This can be recognized as a **Wald statistic** that measures the distance between $\mathbf{R}\hat{\hat{\boldsymbol{\beta}}}$ and $\mathbf{q}$. Both statistics are valid asymptotically, but (14-13) may perform better in a small or moderately sized sample.[13] Once again, the divisor used in computing $\hat{\sigma}_{ij}$ may make a difference, but there is no general rule.

A hypothesis of particular interest is the **homogeneity restriction** of equal coefficient vectors in the multivariate regression model. That case is fairly common in this setting. The homogeneity restriction is that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_M, i = 1, \ldots, M-1$. Consistent with (14-13)–(14-14), we would form the hypothesis as

$$\mathbf{R}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & -\mathbf{I} \\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \cdots \\ \boldsymbol{\beta}_M \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_M \\ \boldsymbol{\beta}_2 - \boldsymbol{\beta}_M \\ \cdots \\ \boldsymbol{\beta}_{M-1} - \boldsymbol{\beta}_M \end{pmatrix} = \mathbf{0}. \tag{14-15}$$

This specifies a total of $(M-1)K$ restrictions on the $KM \times 1$ parameter vector. Denote the estimated asymptotic covariance for $(\hat{\hat{\boldsymbol{\beta}}}_i, \hat{\hat{\boldsymbol{\beta}}}_j)$ as $\hat{\mathbf{V}}_{ij}$. The bracketed matrix in (14-13) would have typical block

$$[\mathbf{R}\,\widehat{\mathrm{Var}}[\hat{\hat{\boldsymbol{\beta}}}]\mathbf{R}']_{ij} = \hat{\mathbf{V}}_{ii} - \hat{\mathbf{V}}_{ij} - \hat{\mathbf{V}}_{ji} + \hat{\mathbf{V}}_{jj}$$

This may be a considerable amount of computation. The test will be simpler if the model has been fit by maximum likelihood, as we examine in the next section.

14.2.4 MAXIMUM LIKELIHOOD ESTIMATION

The Oberhofer–Kmenta (1974) conditions (see Section 11.7.2) 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 (14-9) for estimation of σ_{ij} at each iteration. Maximum likelihood enjoys no advantages over FGLS in its asymptotic properties.[14] Whether it would be preferable in a small sample is an open question whose answer will depend on the particular data set.

By simply inserting the special form of $\boldsymbol{\Omega}$ in the log-likelihood function for the generalized regression model in (10-32), 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.

[13]See Judge et al. (1985, p. 476). The Wald statistic often performs poorly in the small sample sizes typical in this area. Feibig (2001, pp. 108–110) surveys a recent literature on methods of improving the power of testing procedures in SUR models.

[14]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.

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

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

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 $\boldsymbol{\beta}_i$ in (14-4) unless every variable happens to appear in every equation. 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 the GM and GE equations from the Boot–de Witt data in Example 14.1. The tth observation would be

$$[I_g \quad I_e]_t = [1 \quad F_g \quad C_g \quad F_e \quad C_e]_t \begin{bmatrix} \alpha_g & \alpha_e \\ \beta_{1g} & 0 \\ \beta_{2g} & 0 \\ 0 & \beta_{1e} \\ 0 & \beta_{2e} \end{bmatrix} + [\varepsilon_g \quad \varepsilon_e]_t.$$

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

$$\log L_t = -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t. \tag{14-17}$$

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

$$\log L = \sum_{t=1}^{T}\log L_t = -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t. \tag{14-18}$$

The term in the summation in (14-18) is a scalar that equals its trace. We can always permute the matrices in a trace, so

$$\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t = \sum_{t=1}^{T}\operatorname{tr}(\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t) = \sum_{t=1}^{T}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t').$$

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

$$\sum_{t=1}^{T}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = T\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\frac{1}{T}\right)\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right] = T\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{W}) \tag{14-19}$$

where

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

Since 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 εs were actually observed. Inserting this result in the log-likelihood, we have

$$\log L = -\frac{T}{2}[M \log(2\pi) + \log|\boldsymbol{\Sigma}| + \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{W})]. \tag{14-20}$$

We now consider maximizing this function.

It has been shown[15] that

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\Pi}'} &= \frac{T}{2}\mathbf{X}^{*\prime}\mathbf{E}\boldsymbol{\Sigma}^{-1}\\ \frac{\partial \log L}{\partial \boldsymbol{\Sigma}} &= -\frac{T}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{W})\boldsymbol{\Sigma}^{-1}.\end{aligned} \tag{14-21}$$

where the $\mathbf{x}_t^{*\prime}$ in (14-16) 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; $\partial \log L/\partial \boldsymbol{\Sigma} = \mathbf{0}$ implies (14-9).]

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.[16] 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,**

$$\log L_c = -\frac{T}{2}[M(1 + \log(2\pi)) + \log|\mathbf{W}|]. \tag{14-22}$$

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

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \text{Min}_{\boldsymbol{\beta}} \tfrac{1}{2}\log|\mathbf{W}|, \tag{14-23}$$

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 parameter matrix, which must be present in the MLE as well. The estimator of $\boldsymbol{\beta}$ is the set of nonzero elements in the parameter matrix in (14-16).

The **likelihood ratio statistic** is an alternative to the F statistic discussed earlier for testing hypotheses about $\boldsymbol{\beta}$. The likelihood ratio statistic is

$$\lambda = -2(\log L_r - \log L_u) = T(\log|\hat{\mathbf{W}}_r| - \log|\hat{\mathbf{W}}_u|),^{17} \tag{14-24}$$

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. The likelihood ratio statistic is asymptotically distributed as 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 covariance structures model discussed in Section 13.9 in the preceding chapter.

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

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

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

It may also be of interest to test whether $\boldsymbol{\Sigma}$ is a diagonal matrix. Two possible approaches were suggested in Section 13.9.6 [see (13-67) and (13-68)]. The unrestricted model is the one we are using here, whereas the restricted model is the groupwise heteroscedastic model of Section 11.7.2 (Example 11.5), 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

$$\lambda_{\text{LR}} = T\left[\sum_{i=1}^{M} \log \hat{\sigma}_i^2 - \log|\hat{\boldsymbol{\Sigma}}|\right], \tag{14-25}$$

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 estimator 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,**

$$\lambda_{\text{LM}} = T\sum_{i=2}^{M}\sum_{j=1}^{i-1} r_{ij}^2, \tag{14-26}$$

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 $\boldsymbol{\Sigma}$, since it is based on the OLS residuals.

Example 14.2 Estimates of a Seemingly Unrelated Regressions Model

By relaxing the constraint that all five firms have the same parameter vector, we obtain a five-equation seemingly unrelated regression model. The FGLS estimates for the system are given in Table 14.1, where we have included the equality constrained (pooled) estimator from the covariance structures model in Table 13.4 for comparison. The variables are the constant terms, *F* and *C*, respectively. The correlations of the FGLS and equality constrained FGLS residuals are given below the coefficient estimates in Table 14.1. The assumption of equal-parameter vectors appears to have seriously distorted the correlations computed earlier. We would have expected this based on the comparison of Figures 14.1 and 14.2. The diagonal elements in $\hat{\boldsymbol{\Sigma}}$ are also drastically inflated by the imposition of the homogeneity constraint. The equation by equation OLS estimates are given in Table 14.2. As expected, the estimated standard errors for the FGLS estimates are generally smaller. The *F* statistic for testing the hypothesis of equal-parameter vectors in all five equations is 129.169 with 12 and (100–15) degrees of freedom. This value is far larger than the tabled critical value of 1.868, so the hypothesis of parameter homogeneity should be rejected. We might have expected this result in view of the dramatic reduction in the diagonal elements of $\hat{\boldsymbol{\Sigma}}$ compared with those of the pooled estimator. The maximum likelihood estimates of the parameters are given in Table 14.3. The log determinant of the unrestricted maximum likelihood estimator of $\boldsymbol{\Sigma}$ is 31.71986, so the log-likelihood is

$$\log L_u = -\frac{20(5)}{2}[\log(2\pi) + 1] - \frac{20}{2}31.71986 = -459.0925.$$

The restricted model with equal-parameter vectors and correlation across equations is discussed in Section 13.9.6, and the restricted MLEs are given in Table 13.4. (The estimate of $\boldsymbol{\Sigma}$ is not shown there.) The log determinant for the constrained model is 39.1385. The log-likelihood for the constrained model is therefore -515.422. The likelihood ratio test statistic is 112.66. The 1 percent critical value from the chi-squared distribution with 12 degrees of freedom is 26.217, so the hypothesis that the parameters in all five equations are equal is (once again) rejected.

TABLE 14.1 FGLS Parameter Estimates (Standard Errors in Parentheses)

	GM	*CH*	*GE*	*WE*	*US*	*Pooled*
β_1	−162.36	0.5043	−22.439	1.0889	85.423	−28.247
	(89.46)	(11.51)	(25.52)	(6.2959)	(111.9)	(4.888)
β_2	0.12049	0.06955	0.03729	0.05701	0.1015	0.08910
	(0.0216)	(0.0169)	(0.0123)	(0.0114)	(0.0547)	(0.00507)
β_2	0.38275	0.3086	0.13078	0.0415	0.3999	0.3340
	(0.0328)	(0.0259)	(0.0221)	(0.0412)	(0.1278)	(0.0167)

FGLS Residual Covariance and Correlation Matrices [Pooled estimates]

GM	7216.04	−0.299	0.269	0.257	−0.330
	[10050.52]	[−0.349]	[−0.248]	[−.0.356]	[−0.716]
CH	−313.70	152.85	0.006,	0.238	0.384,
	[−4.8051]	[305.61]	[0.158]	[0.246]	[0.244]
GE	605.34	2.0474	700.46	0.777	0.482
	[−7160.67]	[−1966.65]	[34556.6]	[0.895]	[−0.176]
WE	129.89	16.661	200.32	94.912	0.699
	[−1400.75]	[−123.921]	[4274.0]	[833.6]	[−0.040]
US	−2686.5	455.09	1224.4	652.72	9188.2
	[4439.99]	[2158.595]	[−28722.0]	[−2893.7]	[34468.9]

TABLE 14.2 OLS Parameter Estimates (Standard Errors in Parentheses)

	GM	*CH*	*GE*	*WE*	*US*	*Pooled*
β_1	−149.78	−6.1899	−9.956	−0.5094	−30.369	−48.030
	(105.84)	(13.506)	(31.374)	(8.0152)	(157.05)	(21.480)
β_2	0.11928	0.07795	0.02655	0.05289	0.1566	0.10509
	(0.0258)	(0.0198)	(0.0157)	(0.0157)	(0.0789)	(0.01378)
β_2	0.37144	0.3157	0.15169	0.0924	0.4239	0.30537
	(0.0371)	(0.0288)	(0.0257)	(0.0561)	(0.1552)	(0.04351)
σ^2	7160.29	149.872	660.329	88.662	8896.42	15857.24

Based on the OLS results, the Lagrange multiplier statistic is 29.046, with 10 degrees of freedom. The 1 percent critical value is 23.209, so the hypothesis that $\mathbf{\Sigma}$ is diagonal can also be rejected. To compute the likelihood ratio statistic for this test, we would compute the log determinant based on the least squares results. This would be the sum of the logs of the residual variances given in Table 14.2, which is 33.957106. The statistic for the likelihood ratio test using (14–25) is therefore $20(33.95706 - 31.71986) = 44.714$. This is also larger than the critical value from the table. Based on all these results, we conclude that neither the parameter homogeneity restriction nor the assumption of uncorrelated disturbances appears to be consistent with our data.

14.2.5 AN APPLICATION FROM FINANCIAL ECONOMETRICS: THE CAPITAL ASSET PRICING MODEL

One of the growth areas in econometrics is its application to the analysis of financial markets.[18] The **capital asset pricing model** (CAPM) is one of the foundations of that field and is a frequent subject of econometric analysis.

[18]The pioneering work of Campbell, Lo, and MacKinlay (1997) is a broad survey of the field. The development in this example is based on their Chapter 5.

TABLE 14.3 Maximum Likelihood Estimates

	GM	*CH*	*GE*	*WE*	*US*	*Pooled*
β_1	−173.218	2.39111	−16.662	4.37312	136.969	−2.217
	(84.30)	(11.63)	(24.96)	(6.018)	(94.8)	(1.960)
β_2	0.122040	0.06741	0.0371	0.05397	0.08865	0.02361
	(0.02025)	(0.01709)	(0.0118)	(0.0103)	(0.0454)	(0.00429)
β_2	0.38914	0.30520	0.11723	0.026930	0.31246	0.17095
	(0.03185)	(0.02606)	(0.0217)	(0.03708)	(0.118)	(0.0152)
			Residual Covariance Matrix			
GM	7307.30					
CH	−330.55	155.08				
GE	550.27	11.429	741.22			
WE	118.83	18.376	220.33	103.13		
US	−2879.10	463.21	1408.11	734.83	9671.4	

Markowitz (1959) developed a theory of an individual investor's optimal portfolio selection in terms of the trade-off between expected return (mean) and risk (variance). Sharpe (1964) and Lintner (1965) showed how the theory could be extended to the aggregate "market" portfolio. The Sharpe and Lintner analyses produce the following model for the expected excess return from an asset i:

$$E[R_i] - R_f = \beta_i\big(E[R_m] - R_f\big),$$

where R_i is the return on asset i, R_f is the return on a "risk-free" asset, R_m is the return on the market's optimal portfolio, and β_i is the asset's market "beta,"

$$\beta_i = \frac{\text{Cov}[R_i, R_m]}{\text{Var}[R_m]}.$$

The theory states that the expected excess return on asset i will equal β_i times the expected excess return on the market's portfolio. Black (1972) considered the more general case in which there is no risk-free asset. In this instance, the observed R_f is replaced by the unobservable return on a "zero-beta" portfolio, $E[R_0] = \gamma$.

The empirical counterpart to the Sharpe and Lintner model for assets, $i = 1, \dots, N$, observed over T periods, $t = 1, \dots, T$, is a seemingly unrelated regressions (SUR) model, which we cast in the form of (14-16):

$$[y_1, y_2, \dots, y_N] = [1, z_t]\begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \beta_1 & \beta_2 & \cdots & \beta_N\end{bmatrix} + [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N]_t = \mathbf{x}_t'\mathbf{\Pi} + \boldsymbol{\varepsilon}_t',$$

where y_{it} is $R_{it} - R_{ft}$, the *observed* excess return on asset i in period t; z_t is $R_{mt} - R_{ft}$, the market excess return in period t; and disturbances ε_{it} are the deviations from the conditional means. We define the $T \times 2$ matrix $\mathbf{X} = \big([1, z_t], t = 1, \dots, T\big)$. The assumptions of the seemingly unrelated regressions model are

1. $E[\boldsymbol{\varepsilon}_t \mid \mathbf{X}] = E[\boldsymbol{\varepsilon}_t] = \mathbf{0}$,
2. $\text{Var}[\boldsymbol{\varepsilon}_t \mid \mathbf{X}] = E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' \mid \mathbf{X}] = \mathbf{\Sigma}$, a positive definite $N \times N$ matrix,
3. $\boldsymbol{\varepsilon}_t \mid \mathbf{X} \sim N[\mathbf{0}, \mathbf{\Sigma}]$.

The data are also assumed to be "well behaved" so that

4. plim $\bar{z} = E[z_t] = \mu_z$.
5. plim $s_z^2 = \text{plim}(1/T)\sum_{t=1}^{T}(z_t - \bar{z})^2 = \text{Var}[z_t] = \sigma_z^2$.

Since this model is a particular case of the one in (14-16), we can proceed to (14-20) through (14-23) for the maximum likelihood estimators of $\mathbf{\Pi}$ and $\mathbf{\Sigma}$. Indeed, since this model is an unrestricted SUR model with the same regressor(s) in every equation, we know from our results in Section 14.2.2 that the GLS and maximum likelihood estimators are simply equation by equation ordinary least squares and that the estimator of $\mathbf{\Sigma}$ is just $\mathbf{S}$, the sample covariance matrix of the least squares residuals. The asymptotic covariance matrix for the $2N \times 1$ estimator $[\mathbf{a}, \mathbf{b}]'$ will be

$$\text{Asy. Var}[\mathbf{a}, \mathbf{b}]' = \frac{1}{T}\text{plim}\left[\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1} \otimes \mathbf{\Sigma}\right] = \frac{1}{T\sigma_z^2}\begin{bmatrix}\sigma_z^2 + \mu_z^2 & \mu_z \\ \mu_z & 1\end{bmatrix} \otimes \mathbf{\Sigma},$$

which we will estimate with $(\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{S}$. [Plim $\mathbf{z}'\mathbf{z}/T = \text{plim}[(1/T)\Sigma_t(z_t - \bar{z})^2 + \bar{z}^2] = (\sigma_z^2 + \mu_z^2)$.]

The model above does not impose the Markowitz–Sharpe–Lintner hypothesis, H_0: $\boldsymbol{\alpha} = \mathbf{0}$. A Wald test of H_0 can be based on the unrestricted least squares estimates:

$$W = (\mathbf{a} - \mathbf{0})'\left\{\text{Est.Asy. Var}[\mathbf{a} - \mathbf{0}]\right\}^{-1}(\mathbf{a} - \mathbf{0}) = \mathbf{a}'[(\mathbf{X}'\mathbf{X})^{11}\mathbf{S}]^{-1}\mathbf{a} = \left(\frac{Ts_z^2}{s_z^2 + \bar{z}^2}\right)\mathbf{a}'\mathbf{S}^{-1}\mathbf{a}.$$

[To carry out this test, we now require that T be greater than or equal to N, so that $\mathbf{S} = (1/T)\Sigma_t\mathbf{e}_t\mathbf{e}_t'$ will have full rank. The assumption was not necessary until this point.] Under the null hypothesis, the statistic has a limiting chi-squared distribution with N degrees of freedom. The small-sample misbehavior of the Wald statistic has been widely observed. An alternative that is likely to be better behaved is $[(T - N - 1)/N]W$, which is exactly distributed as $F[N, T - N - 1]$ under the null hypothesis. To carry out a likelihood ratio or Lagrange multiplier test of the null hypothesis, we will require the restricted estimates. By setting $\boldsymbol{\alpha} = \mathbf{0}$ in the model, we obtain, once again, a SUR model with identical regressor, so the restricted maximum likelihood estimators are $a_{0i} = 0$ and $b_{0i} = \mathbf{y}_i'\mathbf{z}/\mathbf{z}'\mathbf{z}$. The restricted estimator of $\mathbf{\Sigma}$ is, as before, the matrix of mean squares and cross products of the residuals, now $\mathbf{S}_0$. The chi-squared statistic for the likelihood ratio test is given in (14-24); for this application, it would be

$$\lambda = N(\ln|\mathbf{S}_0| - \ln|\mathbf{S}|).$$

To compute the LM statistic, we will require the derivatives of the unrestricted log-likelihood function, evaluated at the restricted estimators, which are given in (14-21). For this model, they may be written

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{j=1}^{n}\sigma^{ij}\left(\sum_{t=1}^{T}\varepsilon_{jt}\right) = \sum_{j=1}^{N}\sigma^{ij}(T\bar{\varepsilon}_j),$$

where σ^{ij} is the ijth element of $\mathbf{\Sigma}^{-1}$, and

$$\frac{\partial \ln L}{\partial \beta_i} = \sum_{j=1}^{n}\sigma^{ij}\left(\sum_{t=1}^{T}z_t\varepsilon_{jt}\right) = \sum_{j=1}^{N}\sigma^{ij}(\mathbf{z}'\boldsymbol{\varepsilon}_j).$$

The first derivatives with respect to $\boldsymbol{\beta}$ will be zero at the restricted estimates, since the terms in parentheses are the normal equations for restricted least squares; remember, the residuals are now $e_{0it} = y_{it} - b_{0i} z_t$. The first vector of first derivatives can be written as

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}} = \boldsymbol{\Sigma}^{-1}\mathbf{E}'\mathbf{i} = \boldsymbol{\Sigma}^{-1}(T\bar{\boldsymbol{\varepsilon}}),$$

where $\mathbf{i}$ is a $T \times 1$ vector of 1s, $\mathbf{E}$ is a $T \times N$ matrix of disturbances, and $\bar{\boldsymbol{\varepsilon}}$ is the $N \times 1$ vector of means of asset specific disturbances. (The second subvector is $\partial \ln L/\partial \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\mathbf{E}'\mathbf{z}$.) Since $\partial \ln L/\partial \boldsymbol{\beta} = \mathbf{0}$ at the restricted estimates, the LM statistic involves only the upper left submatrix of $-\mathbf{H}^{-1}$. Combining terms and inserting the restricted estimates, we obtain

$$\begin{aligned}
\text{LM} &= \left[T\,\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1} : \mathbf{0}'\right]'\left[\mathbf{X}'\mathbf{X} \otimes \mathbf{S}_0^{-1}\right]^{-1}\left[T\,\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1} : \mathbf{0}'\right] \\
&= T^2(\mathbf{X}'\mathbf{X})^{11}\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1}\bar{\mathbf{e}}_0 \\
&= T\left(\frac{s_z^2 + \bar{z}^2}{s_z^2}\right)\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1}\bar{\mathbf{e}}_0.
\end{aligned}$$

Under the null hypothesis, the limiting distribution of LM is chi-squared with N degrees of freedom.

The model formulation gives $E[R_{it}] = R_{ft} + \beta_i\big(E[R_{mt}] - R_{ft}\big)$. If there is no risk-free asset but we write the model in terms of γ, the unknown return on a zero-beta portfolio, then we obtain

$$\begin{aligned}
R_{it} &= \gamma + \beta_i(R_{mt} - \gamma) + \varepsilon_{it} \\
&= (1 - \beta_i)\gamma + \beta_i R_{mt} + \varepsilon_{it}.
\end{aligned}$$

This is essentially the same as the original model, with two modifications. First, the observables in the model are real returns, not excess returns, which defines the way the data enter the model. Second, there are nonlinear restrictions on the parameters; $\alpha_i = (1 - \beta_i)\gamma$. Although the unrestricted model has $2N$ free parameters, Black's formulation implies $N - 1$ restrictions and leaves $N + 1$ free parameters. The nonlinear restrictions will complicate finding the maximum likelihood estimators. We do know from (14-21) that regardless of what the estimators of β_i and γ are, the estimator of $\boldsymbol{\Sigma}$ is still $\mathbf{S} = (1/T)\mathbf{E}'\mathbf{E}$. So, we can concentrate the log-likelihood function. The Oberhofer and Kmenta (1974) results imply that we may simply zigzag back and forth between $\mathbf{S}$ and $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$ (See Section 11.7.2.) Second, although maximization over $(\boldsymbol{\beta}, \gamma)$ remains complicated, maximization over $\boldsymbol{\beta}$ for known γ is trivial. For a given value of γ, the maximum likelihood estimator of β_i is the slope in the linear regression without a constant term of $(R_{it} - \gamma)$ on $(R_{mt} - \gamma)$. Thus, the full set of maximum likelihood estimators may be found just by scanning over the admissible range of γ to locate the value that maximizes

$$\ln L_c = -\frac{1}{2}\ln|\mathbf{S}(\gamma)|,$$

where

$$s_{ij}(\gamma) = \frac{\sum_{t=1}^{T}\left\{R_{it} - \gamma[1 - \hat{\beta}_i(\gamma)] - \hat{\beta}_i(\gamma)R_{mt}\right\}\left\{R_{jt} - \gamma[1 - \hat{\beta}_j(\gamma)] - \hat{\beta}_j(\gamma)R_{mt}\right\}}{T},$$

and

$$\hat{\beta}_i(\gamma) = \frac{\sum_{t=1}^{T}(R_{it} - \gamma)(R_{mt} - \gamma)}{\sum_{t=1}^{T}(R_{mt} - \gamma)^2}.$$

For inference purposes, an estimator of the asymptotic covariance matrix of the estimators is required. The log-likelihood for this model is

$$\ln L = -\frac{T}{2}[N \ln 2\pi + \ln|\boldsymbol{\Sigma}|] - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t$$

where the $N \times 1$ vector $\boldsymbol{\varepsilon}_t$ is $\varepsilon_{it} = [R_{it} - \gamma(1-\beta_i) - \beta_i R_{mt}]$, $i = 1, \ldots, N$. The derivatives of the log-likelihood can be written

$$\frac{\partial \ln L}{\partial[\boldsymbol{\beta}' \quad \gamma]'} = \sum_{t=1}^{T}\begin{bmatrix}(R_{mt} - \gamma)\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t \\ (\mathbf{i} - \boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\end{bmatrix} = \sum_{t=1}^{T}\mathbf{g}_t.$$

(We have omitted $\boldsymbol{\Sigma}$ from the gradient because the expected Hessian is block diagonal, and, at present, $\boldsymbol{\Sigma}$ is tangential.) With the derivatives in this form, we have

$$E[\mathbf{g}_t\,\mathbf{g}_t'] = \begin{bmatrix}(R_{mt} - \gamma)^2\boldsymbol{\Sigma}^{-1} & (R_{mt} - \gamma)\boldsymbol{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta}) \\ (R_{mt} - \gamma)(\mathbf{i} - \boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1} & (\mathbf{i} - \boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta})\end{bmatrix}. \tag{14-27}$$

Now, sum this expression over t and use the result that

$$\sum_{t=1}^{T}(R_{mt} - \gamma)^2 = \sum_{t=1}^{T}(R_{mt} - \bar{R}_m)^2 + T(\bar{R}_m - \gamma)^2 = T\left[s_{Rm}^2 + (\bar{R}_m - \gamma)^2\right]$$

to obtain the negative of the expected Hessian,

$$-E\left[\frac{\partial^2 \ln L}{\partial\begin{bmatrix}\boldsymbol{\beta}\\ \gamma\end{bmatrix}\partial\begin{bmatrix}\boldsymbol{\beta}\\ \gamma\end{bmatrix}'}\right] = T\begin{bmatrix}\left[s_{Rm}^2 + (\bar{R}_m - \gamma)^2\right]\boldsymbol{\Sigma}^{-1} & (\bar{R}_m - \gamma)\boldsymbol{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta}) \\ (\bar{R}_m - \gamma)(\mathbf{i} - \boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1} & (\mathbf{i} - \boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta})\end{bmatrix}. \tag{14-28}$$

The inverse of this matrix provides the estimator for the asymptotic covariance matrix. Using (A-74), after some manipulation we find that

$$\text{Asy. Var}[\hat{\gamma}] = \frac{1}{T}\left[1 + \frac{(\mu_{Rm} - \gamma)^2}{\sigma_{Rm}^2}\right][(\mathbf{i} - \boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta})]^{-1}.$$

where $\mu_{Rm} = \text{plim}\,\bar{R}_m$ and $\sigma_{Rm}^2 = \text{plim}\,s_{Rm}^2$.

A likelihood ratio test of the Black model requires the restricted estimates of the parameters. The unrestricted model is the SUR model for the real returns, R_{it} on the market returns, R_{mt}, with N free constants, α_i, and N free slopes, β_i. Result (14-24) provides the test statistic. Once the estimates of β_i and γ are obtained, the implied estimates of α_i are given by $\alpha_i = (1-\beta_i)\gamma$. With these estimates in hand, the LM statistic is exactly what it was before, although now all $2N$ derivatives will be required and $\mathbf{X}$ is $[\mathbf{i}, \mathbf{R}_m]$. The subscript * indicates computation at the restricted estimates;

$$\text{LM} = T\left(\frac{s_{Rm}^2 + \bar{R}_m^2}{s_{Rm}^2}\right)\bar{\mathbf{e}}_*'\mathbf{S}_*^{-1}\bar{\mathbf{e}}_* + \left(\frac{1}{Ts_{Rm}^2}\right)\mathbf{R}_m'\mathbf{E}_*\mathbf{S}_*^{-1}\mathbf{E}_*'\mathbf{R}_m - \left(\frac{2\bar{R}_m}{s_z^2}\right)\mathbf{R}_m'\mathbf{E}_*\mathbf{S}_*^{-1}\bar{\mathbf{e}}_*.$$

A Wald test of the Black model would be based on the unrestricted estimators. The hypothesis appears to involve the unknown γ, but in fact, the theory implies only the $N-1$ nonlinear restrictions: $[(\alpha_i/\alpha_N) - (1-\beta_i)/(1-\beta_N)] = 0$ or $[\alpha_i(1-\beta_N) - \alpha_N(1-\beta_i)] = 0$. Write this set of $N-1$ functions as $\mathbf{c}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{0}$. The Wald statistic based on the least squares estimates would then be

$$W = \mathbf{c}(\mathbf{a}, \mathbf{b})' \{\text{Est.Asy. Var}[\mathbf{c}(\mathbf{a}, \mathbf{b})]\}^{-1} \mathbf{c}(\mathbf{a}, \mathbf{b}).$$

Recall in the unrestricted model that Asy. Var$[\mathbf{a}, \mathbf{b}] = (1/T)\text{plim}(\mathbf{X}'\mathbf{X}/T)^{-1} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Delta}$, say. Using the delta method (see Section D.2.7), the asymptotic covariance matrix for $\mathbf{c}(\mathbf{a}, \mathbf{b})$ would be

$$\text{Asy. Var}[\mathbf{c}(\mathbf{a}, \mathbf{b})] = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}' \quad \text{where } \boldsymbol{\Gamma} = \frac{\partial \mathbf{c}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial(\boldsymbol{\alpha}, \boldsymbol{\beta})}.$$

The ith row of the $2N \times 2N$ matrix $\boldsymbol{\Gamma}$ has four only nonzero elements, one each in the ith and Nth positions of each of the two subvectors.

Before closing this lengthy example, we reconsider the assumptions of the model. There is ample evidence [e.g., Affleck–Graves and McDonald (1989)] that the normality assumption used in the preceding is not appropriate for financial returns. This fact in itself does not complicate the analysis very much. Although the estimators derived earlier are based on the normal likelihood, they are really only generalized least squares. As we have seen before (in Chapter 10), GLS is robust to distributional assumptions. The LM and LR tests we devised are not, however. Without the normality assumption, only the Wald statistics retain their asymptotic validity. As noted, the small-sample behavior of the Wald statistic can be problematic. The approach we have used elsewhere is to use an approximation, $F = W/J$, where J is the number of restrictions, and refer the statistic to the more conservative critical values of the $F[J, q]$ distribution, where q is the number of degrees of freedom in estimation. Thus, once again, the role of the normality assumption is quite minor.

The homoscedasticity and nonautocorrelation assumptions are potentially more problematic. The latter almost certainly invalidates the entire model. [See Campbell, Lo, and MacKinlay (1997) for discussion.] If the disturbances are only heteroscedastic, then we can appeal to the well-established consistency of ordinary least squares in the generalized regression model. A GMM approach might seem to be called for, but GMM estimation in this context is irrelevant. In all cases, the parameters are exactly identified. What is needed is a robust covariance estimator for our now pseudomaximum likelihood estimators. For the Sharpe–Lintner formulation, nothing more than the White estimator that we developed in Chapters 10 and 11 is required; after all, despite the complications of the models, the estimators both with and without the restrictions are ordinary least squares, equation by equation. For each equation separately, the robust asymptotic covariance matrix in (10-14) applies. For the least squares estimators $\mathbf{q}_i = (a_i, b_i)$, we seek a robust estimator of

$$\text{Asy. Cov}[\mathbf{q}_i, \mathbf{q}_j] = T\,\text{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Assuming that $E[\varepsilon_{it}\varepsilon_{jt}] = \sigma_{ij}$, this matrix can be estimated with

$$\text{Est.Asy. Cov}[\mathbf{q}_i, \mathbf{q}_j] = [(\mathbf{X}'\mathbf{X})^{-1}]\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' e_{it}e_{jt}\right)[(\mathbf{X}'\mathbf{X})^{-1}].$$

To form a counterpart for the Black model, we will once again rely on the assumption that the asymptotic covariance of the MLE of $\boldsymbol{\Sigma}$ and the MLE of $(\boldsymbol{\beta}', \gamma)$ is zero. Then the "sandwich" estimator for this M estimator (see Section 17.8) is

$$\text{Est. Asy. Var}(\hat{\beta}, \gamma) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1},$$

where $\mathbf{A}$ appears in (14-28) and $\mathbf{B}$ is in (14-27).

14.2.6 MAXIMUM LIKELIHOOD ESTIMATION OF THE SEEMINGLY UNRELATED REGRESSIONS MODEL WITH A BLOCK OF ZEROS IN THE COEFFICIENT MATRIX

In Section 14.2.2, we considered the special case of the SUR model with identical regressors in all equations. We showed there that in this case, OLS and GLS are identical. In the SUR model with normally distributed disturbances, GLS is the maximum likelihood estimator. It follows that when the regressors are identical, OLS is the maximum likelihood estimator. In this section, we consider a related case in which the coefficient matrix contains a block of zeros. The block of zeros is created by excluding the same subset of the regressors from some of but not all the equations in a model that without the exclusion restriction is a SUR with the same regressors in all equations.

This case can be examined in the context of the derivation of the GLS estimator in (14-7), but it is much simpler to obtain the result we seek for the maximum likelihood estimator. The model we have described can be formulated as in (14-16) as follows. We first transpose the equation system in (14-16) so that observation t on $y_1, \ldots, y_M$ is written

$$\mathbf{y}_t = \boldsymbol{\Pi}\mathbf{x}_t + \boldsymbol{\varepsilon}_t.$$

If we collect all T observations in this format, then the system would appear as

$$\underset{M \times T}{\mathbf{Y}'} = \underset{M \times K}{\boldsymbol{\Pi}} \quad \underset{K \times T}{\mathbf{X}'} + \underset{M \times T}{\mathbf{E}'}.$$

(Each row of $\boldsymbol{\Pi}$ contains the parameters in a particular equation.) Now, consider once again a particular observation and partition the set of dependent variables into two groups of M_1 and M_2 variables and the set of regressors into two sets of K_1 and K_2 variables. The equation system is now

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix}_t = \begin{bmatrix}\boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12}\\ \boldsymbol{\Pi}_{21} & \boldsymbol{\Pi}_{22}\end{bmatrix}\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix}_t + \begin{pmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{pmatrix}_t, \quad E\left[\begin{matrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{matrix}\,\middle|\,\mathbf{X}\right]_t = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \quad \text{Var}\left[\begin{matrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{matrix}\,\middle|\,\mathbf{X}\right]_t = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}.$$

Since this system is still a SUR model with identical regressors, the maximum likelihood estimators of the parameters are obtained using equation by equation least squares regressions. The case we are interested in here is the restricted model, with $\boldsymbol{\Pi}_{12} = \mathbf{0}$, which has the effect of excluding $\mathbf{x}_2$ from all the equations for $\mathbf{y}_1$. The results we will obtain for this case are:

1. The maximum likelihood estimator of $\boldsymbol{\Pi}_{11}$ when $\boldsymbol{\Pi}_{12} = \mathbf{0}$ is equation-by-equation least squares regression of the variables in $\mathbf{y}_1$ on $\mathbf{x}_1$ alone. That is, even with the restriction, the efficient estimator of the parameters of the first set of equations is

equation-by-equation ordinary least squares. Least squares is not the efficient estimator for the second set, however.

2. The effect of the restriction on the likelihood function can be isolated to its effect on the smaller set of equations. Thus, the hypothesis can be tested without estimating the larger set of equations.

We begin by considering maximum likelihood estimation of the unrestricted system. The log-likelihood function for this multivariate regression model is

$$\ln L = \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t})$$

where $f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t})$ is the joint normal density of the two vectors. This result is (14-17) through (14-19) in a different form. We will now write this joint normal density as the product of a marginal and a conditional:

$$f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}) = f(\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t})\, f(\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}).$$

The mean and variance of the marginal distribution for $\mathbf{y}_{1t}$ are just the upper portions of the preceding partitioned matrices:

$$E[\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}] = \mathbf{\Pi}_{11}\mathbf{x}_{1t} + \mathbf{\Pi}_{12}\mathbf{x}_{2t}, \quad \mathrm{Var}[\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}] = \mathbf{\Sigma}_{11}.$$

The results we need for the conditional distribution are given in Theorem B.6. Collecting terms, we have

$$\begin{aligned}
E[\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}] &= \left[\mathbf{\Pi}_{21} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Pi}_{11}\right]\mathbf{x}_{1t} + \left[\mathbf{\Pi}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Pi}_{12}\right]\mathbf{x}_{2t} + \left[\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\right]\mathbf{y}_{1t} \\
&= \mathbf{\Lambda}_{21}\mathbf{x}_{1t} + \mathbf{\Lambda}_{22}\mathbf{x}_{2t} + \mathbf{\Gamma}\mathbf{y}_{1t}, \\
\mathrm{Var}[\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}] &= \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} = \mathbf{\Omega}_{22}.
\end{aligned}$$

Finally, since the marginal distributions and the joint distribution are all multivariate normal, the conditional distribution is also. The objective of this partitioning is to partition the log-likelihood function likewise;

$$\begin{aligned}
\ln L &= \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}) \\
&= \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t})\, f(\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}) \\
&= \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}) + \sum_{t=1}^{T} \ln f(\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}).
\end{aligned}$$

With no restrictions on any of the parameters, we can maximize this log-likelihood by maximizing its parts separately. There are two multivariate regression systems defined by the two parts, and they have no parameters in common. Because $\mathbf{\Pi}_{21}$, $\mathbf{\Pi}_{22}$, $\mathbf{\Sigma}_{21}$, and $\mathbf{\Sigma}_{22}$ are all free, unrestricted parameters, there are no restrictions imposed on $\mathbf{\Lambda}_{21}$, $\mathbf{\Lambda}_{22}$, $\mathbf{\Gamma}$, or $\mathbf{\Omega}_{22}$. Therefore, in each case, the efficient estimators are equation-by-equation ordinary least squares. The first part produces estimates of $\mathbf{\Pi}_{11}$, $\mathbf{\Pi}_{22}$, and $\mathbf{\Sigma}_{11}$ directly. From the second, we would obtain estimates of $\mathbf{\Lambda}_{21}$, $\mathbf{\Lambda}_{22}$, $\mathbf{\Gamma}$, and $\mathbf{\Omega}_{22}$. But it is easy to see in the relationships above how the original parameters can be obtained from these mixtures:

$$\begin{aligned}
\mathbf{\Pi}_{21} &= \mathbf{\Lambda}_{21} + \mathbf{\Gamma}\mathbf{\Pi}_{11},\\
\mathbf{\Pi}_{22} &= \mathbf{\Lambda}_{22} + \mathbf{\Gamma}\mathbf{\Pi}_{12},\\
\mathbf{\Sigma}_{21} &= \mathbf{\Gamma}\mathbf{\Sigma}_{11},\\
\mathbf{\Sigma}_{22} &= \mathbf{\Omega}_{22} + \mathbf{\Gamma}\mathbf{\Sigma}_{11}\mathbf{\Gamma}'.
\end{aligned}$$

Because of the **invariance of maximum likelihood estimators** to transformation, these derived estimators of the original parameters are also maximum likelihood estimators. Thus, the result we have up to this point is that by manipulating this pair of sets of ordinary least squares estimators, we can obtain the original least squares, efficient estimators. This result is no surprise, of course, since we have just rearranged the original system and we are just rearranging our least squares estimators.

Now, consider estimation of the same system subject to the restriction $\mathbf{\Pi}_{12} = \mathbf{0}$. The second equation system is still completely unrestricted, so maximum likelihood estimates of its parameters, $\mathbf{\Lambda}_{21}$, $\mathbf{\Lambda}_{22}$ (which now equals $\mathbf{\Pi}_{22}$), $\mathbf{\Gamma}$, and $\mathbf{\Omega}_{22}$, are still obtained by equation-by-equation least squares. The equation systems have no parameters in common, so maximum likelihood estimators of the first set of parameters are obtained by maximizing the first part of the log-likelihood, once again, by equation-by-equation ordinary least squares. Thus, our first result is established. To establish the second result, we must obtain the two parts of the log-likelihood. The log-likelihood function for this model is given in (14-20). Since each of the two sets of equations is estimated by least squares, in each case (null and alternative), for each part, the term in the log-likelihood is the concentrated log-likelihood given in (14-22), where $\mathbf{W}_{jj}$ is $(1/T)$ times the matrix of sums of squares and cross products of least squares residuals. The second set of equations is estimated by regressions on $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{y}_1$ with or without the restriction $\mathbf{\Pi}_{12} = \mathbf{0}$. So, the second part of the log-likelihood is always the same,

$$\ln L_{2c} = -\frac{T}{2}[M_2(1 + \ln 2\pi) + \ln|\mathbf{W}_{22}|].$$

The concentrated log-likelihood for the first set of equations equals

$$\ln L_{1c} = -\frac{T}{2}[M_1(1 + \ln 2\pi) + \ln|\mathbf{W}_{11}|],$$

when $\mathbf{x}_2$ is included in the equations, and the same with $\mathbf{W}_{11}(\mathbf{\Pi}_{12} = \mathbf{0})$ when $\mathbf{x}_2$ is excluded. At the maximum likelihood estimators, the log-likelihood for the whole system is

$$\ln L_c = \ln L_{1c} + \ln L_{2c}.$$

The likelihood ratio statistic is

$$\lambda = -2[(\ln L_c \mid \mathbf{\Pi}_{12} = 0) - (\ln L_c)] = T[\ln|\mathbf{W}_{11}(\mathbf{\Pi}_{12} = 0)| - \ln|\mathbf{W}_{11}|].$$

This establishes our second result, since $\mathbf{W}_{11}$ is based only on the first set of equations.

The block of zeros case was analyzed by Goldberger (1970). Many regression systems in which the result might have proved useful (e.g., systems of demand equations) imposed cross-equation equality (symmetry) restrictions, so the result of the analysis was often derailed. Goldberger's result, however, is precisely what is needed in the more recent application of testing for Granger causality in the context of vector autoregressions. We will return to the issue in Section 19.6.5.

14.2.7 AUTOCORRELATION AND HETEROSCEDASTICITY

The seemingly unrelated regressions model can be extended to allow for autocorrelation in the same fashion as in Section 13.9.5. To reiterate, suppose that

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$
$$\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + u_{it},$$

where u_{it} is uncorrelated across observations. This extension will imply that the blocks in $\boldsymbol{\Omega}$ in (14-3), instead of $\sigma_{ij}\mathbf{I}$, are $\sigma_{ij}\boldsymbol{\Omega}_{ij}$, where $\boldsymbol{\Omega}_{ij}$ is given in (13-63).

The treatment developed by Parks (1967) is the one we used earlier.[19] It calls for a three-step approach:

1. Estimate each equation in the system by ordinary least squares. Compute any consistent estimators of ρ. For each equation, transform the data by the Prais–Winsten transformation to remove the autocorrelation.[20] Note that there will not be a constant term in the transformed data because there will be a column with $(1 - r_i^2)^{1/2}$ as the first observation and $(1 - r_i)$ for the remainder.
2. Using the transformed data, use ordinary least squares again to estimate $\boldsymbol{\Sigma}$.
3. Use FGLS based on the estimated $\boldsymbol{\Sigma}$ and the transformed data.

There is no benefit to iteration. The estimator is efficient at every step, and iteration does not produce a maximum likelihood estimator because of the Jacobian term in the log likelihood [see (12-30)]. After the last step, $\boldsymbol{\Sigma}$ should be reestimated with the GLS estimates. The estimated covariance matrix for $\boldsymbol{\varepsilon}$ can then be reconstructed using

$$\hat{\sigma}_{mn}(\boldsymbol{\varepsilon}) = \frac{\hat{\sigma}_{mn}}{1 - r_m r_n}.$$

As in the single equation case, opinions differ on the appropriateness of such corrections for autocorrelation. At one extreme is Mizon (1995) who argues forcefully that autocorrelation arises as a consequence of a remediable failure to include dynamic effects in the model. However, in a system of equations, the analysis that leads to this

[19] Guilkey and Schmidt (1973), Guilkey (1974) and Berndt and Savin (1977) present an alternative treatment based on $\boldsymbol{\varepsilon}_t = \mathbf{R}\boldsymbol{\varepsilon}_{t-1} + \mathbf{u}_t$, where $\boldsymbol{\varepsilon}_t$ is the $M \times 1$ vector of disturbances at time t and $\mathbf{R}$ is a correlation matrix. Extensions and additional results appear in Moschino and Moro (1994), McLaren (1996), and Holt (1998).

[20] There is a complication with the first observation that is not treated quite correctly by this procedure. For details, see Judge et al. (1985, pp. 486–489). The strictly correct (and quite cumbersome) results are for the true GLS estimator, which assumes a known $\boldsymbol{\Omega}$. It is unlikely that in a finite sample, anything is lost by using the Prais–Winsten procedure with the estimated $\boldsymbol{\Omega}$. One suggestion has been to use the Cochrane–Orcutt procedure and drop the first observation. But in a small sample, the cost of discarding the first observation is almost surely greater than that of neglecting to account properly for the correlation of the first disturbance with the other first disturbances.

TABLE 14.4 Autocorrelation Coefficients

	GM	*CH*	*GE*	*WE*	*US*
Durbin–Watson	0.9375	1.984	1.0721	1.413	0.9091
Autocorrelation	0.531	0.008	0.463	0.294	0.545
	Residual Covariance Matrix $[\hat{\sigma}_{ij}/(1-r_i r_j)]$				
GM	6679.5				
CH	−220.97	151.96			
GE	483.79	43.7891	684.59		
WE	88.373	19.964	190.37	92.788	
US	−1381.6	342.89	1484.10	676.88	8638.1
	Parameter Estimates (Standard Errors in Parentheses)				
β_1	−51.337	−0.4536	−24.913	4.7091	14.0207
	(80.62)	(11.86)	(25.67)	(6.510)	(96.49)
β_2	0.094038	0.06847	0.04271	0.05091	0.16415
	(0.01733)	(0.0174)	(0.01134)	(0.01060)	(0.0386)
β_3	0.040723	0.32041	0.10954	0.04284	0.2006
	(0.04216)	(0.0258)	(0.03012)	(0.04127)	(0.1428)

conclusion is going to be far more complex than in a single equation model.[21] Suffice to say, the issue remains to be settled conclusively.

Example 14.3 Autocorrelation in a SUR Model

Table 14.4 presents the autocorrelation-corrected estimates of the model of Example 14.2. The Durbin–Watson statistics for the five data sets given here, with the exception of Chrysler, strongly suggest that there is, indeed, autocorrelation in the disturbances. The differences between these and the uncorrected estimates given earlier are sometimes relatively large, as might be expected, given the fairly high autocorrelation and small sample size. The smaller diagonal elements in the disturbance covariance matrix compared with those of Example 14.2 reflect the improved fit brought about by introducing the lagged variables into the equation.

In principle, the SUR model can accommodate heteroscedasticity as well as autocorrelation. Bartels and Feibig (1991) suggested the generalized SUR model, $\mathbf{\Omega} = \mathbf{A}[\mathbf{\Sigma} \otimes \mathbf{I}]\mathbf{A}'$ where $\mathbf{A}$ is a block diagonal matrix. Ideally, $\mathbf{A}$ is made a function of measured characteristics of the individual and a separate parameter vector, $\boldsymbol{\theta}$, so that the model can be estimated in stages. In a first step, OLS residuals could be used to form a preliminary estimator of $\boldsymbol{\theta}$, then the data are transformed to homoscedasticity, leaving $\mathbf{\Sigma}$ and $\boldsymbol{\beta}$ to be estimated at subsequent steps using transformed data. One application along these lines is the random parameters model of Feibig, Bartels and Aigner (1991)—(13-46) shows how the random parameters model induces heteroscedasticity. Another application is Mandy and Martins–Filho, who specified $\sigma_{ij}(t) = \boldsymbol{\alpha}_{ij}'\mathbf{z}_{ij}(t)$. (The linear specification of a variance does present some problems, as a negative value is not precluded.) Kumbhakar and Heshmati (1996) proposed a cost and demand

[21] Dynamic SUR models in the spirit of Mizon's admonition were proposed by Anderson and Blundell (1982). A few recent applications are Kiviet, Phillips, and Schipp (1995) and Deschamps (1998). However, relatively little work has been done with dynamic SUR models. The VAR models in Chapter 20 are an important group of applications, but they come from a different analytical framework.

system that combined the translog model of Section 14.3.2 with the complete equation system in 14.3.1. In their application, only the cost equation was specified to include a heteroscedastic disturbance.

14.3 SYSTEMS OF DEMAND EQUATIONS: SINGULAR SYSTEMS

Most of the recent applications of the multivariate regression model[22] have been in the context of systems of demand equations, either commodity demands or factor demands in studies of production.

Example 14.4 Stone's Expenditure System

Stone's expenditure system[23] based on a set of logarithmic commodity demand equations, income Y, and commodity prices p_n is

$$\log q_i = \alpha_i + \eta_i \log\left(\frac{Y}{P}\right) + \sum_{j=1}^{M} \eta_{ij}^{*} \log\left(\frac{p_j}{P}\right),$$

where P is a generalized (share-weighted) price index, η_i is an income elasticity, and η_{ij}^{*} is a compensated price elasticity. We can interpret this system as the demand equation in real expenditure and real prices. The resulting set of equations constitutes an econometric model in the form of a set of seemingly unrelated regressions. In estimation, we must account for a number of restrictions including homogeneity of degree one in income, $\Sigma_i \eta_i = 1$, and symmetry of the matrix of compensated price elasticities, $\eta_{ij}^{*} = \eta_{ji}^{*}$.

Other examples include the system of factor demands and factor cost shares from production, which we shall consider again later. In principle, each is merely a particular application of the model of the previous section. But some special problems arise in these settings. First, the parameters of the systems are generally constrained across equations. That is, the unconstrained model is inconsistent with the underlying theory.[24] The numerous constraints in the system of demand equations presented earlier give an example. A second intrinsic feature of many of these models is that the disturbance covariance matrix $\mathbf{\Sigma}$ is singular.

[22] Note the distinction between the multi*variate* or multiple-equation model discussed here and the *multiple* regression model.

[23] A very readable survey of the estimation of systems of commodity demands is Deaton and Muellbauer (1980). The example discussed here is taken from their Chapter 3 and the references to Stone's (1954a,b) work cited therein. A counterpart for production function modeling is Chambers (1988). Recent developments in the specification of systems of demand equations include Chavez and Segerson (1987), Brown and Walker (1995), and Fry, Fry, and McLaren (1996).

[24] This inconsistency does not imply that the theoretical restrictions are not testable or that the unrestricted model cannot be estimated. Sometimes, the meaning of the model is ambiguous without the restrictions, however. Statistically rejecting the restrictions implied by the theory, which were used to derive the econometric model in the first place, can put us in a rather uncomfortable position. For example, in a study of utility functions, Christensen, Jorgenson, and Lau (1975), after rejecting the cross-equation symmetry of a set of commodity demands, stated, "With this conclusion we can terminate the test sequence, since these results invalidate the theory of demand" (p. 380). See Silver and Ali (1989) for discussion of testing symmetry restrictions.

14.3.1 COBB–DOUGLAS COST FUNCTION (EXAMPLE 7.3 CONTINUED)

Consider a Cobb–Douglas production function,

$$Y = \alpha_0 \prod_{i=1}^{M} x_i^{\alpha_i}.$$

Profit maximization with an exogenously determined output price calls for the firm to maximize output for a given cost level C (or minimize costs for a given output Y). The Lagrangean for the maximization problem is

$$\Lambda = \alpha_0 \prod_{i=1}^{M} x_i^{\alpha_i} + \lambda(C - \mathbf{p}'\mathbf{x}),$$

where $\mathbf{p}$ is the vector of M factor prices. The necessary conditions for maximizing this function are

$$\frac{\partial\Lambda}{\partial x_i} = \frac{\alpha_i Y}{x_i} - \lambda p_i = 0 \quad \text{and} \quad \frac{\partial\Lambda}{\partial\lambda} = C - \mathbf{p}'\mathbf{x} = 0.$$

The joint solution provides $x_i(Y, \mathbf{p})$ and $\lambda(Y, \mathbf{p})$. The total cost of production is

$$\sum_{i=1}^{M} p_i x_i = \sum_{i=1}^{M} \frac{\alpha_i Y}{\lambda}.$$

The cost share allocated to the ith factor is

$$\frac{p_i x_i}{\sum_{i=1}^{M} p_i x_i} = \frac{\alpha_i}{\sum_{i=1}^{M} \alpha_i} = \beta_i. \tag{14-29}$$

The full model is[25]

$$\begin{aligned} \ln C &= \beta_0 + \beta_y \ln Y + \sum_{i=1}^{M} \beta_i \ln p_i + \varepsilon_c, \\ s_i &= \beta_i + \varepsilon_i, \quad i = 1, \dots, M. \end{aligned} \tag{14-30}$$

By construction, $\sum_{i=1}^{M} \beta_i = 1$ and $\sum_{i=1}^{M} s_i = 1$. (This is the cost function analysis begun in Example 7.3. We will return to that application below.) The cost shares will also sum identically to one in the data. It therefore follows that $\sum_{i=1}^{M} \varepsilon_i = 0$ at every data point, so the system is singular. For the moment, ignore the cost function. Let the $M \times 1$ disturbance vector from the shares be $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M]'$. Since $\boldsymbol{\varepsilon}'\mathbf{i} = 0$, where $\mathbf{i}$ is a column of 1s, it follows that $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{i}] = \boldsymbol{\Sigma}\mathbf{i} = \mathbf{0}$, which implies that $\boldsymbol{\Sigma}$ is singular. Therefore, the methods of the previous sections cannot be used here. (You should verify that the *sample* covariance matrix of the OLS residuals will also be singular.)

The solution to the singularity problem appears to be to drop one of the equations, estimate the remainder, and solve for the last parameter from the other $M - 1$. The constraint $\sum_{i=1}^{M} \beta_i = 1$ states that the cost function must be homogeneous of degree one

[25] We leave as an exercise the derivation of β_0, which is a mixture of all the parameters, and β_y, which equals $1/\Sigma_m \alpha_m$.

in the prices, a theoretical necessity. If we impose the constraint

$$\beta_M = 1 - \beta_1 - \beta_2 - \cdots - \beta_{M-1}, \tag{14-31}$$

then the system is reduced to a nonsingular one:

$$\log\left(\frac{C}{p_M}\right) = \beta_0 + \beta_y \log Y + \sum_{i=1}^{M-1} \beta_i \log\left(\frac{p_i}{p_M}\right) + \varepsilon_c,$$
$$s_i = \beta_i + \varepsilon_i, \quad i = 1, \ldots, M-1$$

This system provides estimates of β_0, β_y, and $\beta_1, \ldots, \beta_{M-1}$. The last parameter is estimated using (14-31). In principle, it is immaterial which factor is chosen as the numeraire. Unfortunately, the FGLS parameter estimates in the now nonsingular system will depend on which one is chosen. Invariance is achieved by using maximum likelihood estimates instead of FGLS,[26] which can be obtained by iterating FGLS or by direct maximum likelihood estimation.[27]

Nerlove's (1963) study of the electric power industry that we examined in Example 7.3 provides an application of the Cobb–Douglas cost function model. His ordinary least squares estimates of the parameters were listed in Example 7.3. Among the results are (unfortunately) a negative capital coefficient in three of the six regressions. Nerlove also found that the simple Cobb–Douglas model did not adequately account for the relationship between output and average cost. Christensen and Greene (1976) further analyzed the Nerlove data and augmented the data set with cost share data to estimate the complete **demand system.** Appendix Table F14.2 lists Nerlove's 145 observations with Christensen and Greene's cost share data. Cost is the total cost of generation in millions of dollars, output is in millions of kilowatt-hours, the capital price is an index of construction costs, the wage rate is in dollars per hour for production and maintenance, the fuel price is an index of the cost per Btu of fuel purchased by the firms, and the data reflect the 1955 costs of production. The regression estimates are given in Table 14.5.

Least squares estimates of the Cobb–Douglas cost function are given in the first column.[28] The coefficient on capital is negative. Because $\beta_i = \beta_y \partial \ln Y / \partial \ln x_i$—that is, a positive multiple of the output elasticity of the ith factor—this finding is troubling. The third column gives the maximum likelihood estimates obtained in the constrained system. Two things to note are the dramatically smaller standard errors and the now positive (and reasonable) estimate of the capital coefficient. The estimates of economies of scale in the basic Cobb–Douglas model are $1/\beta_y = 1.39$ (column 1) and 1.25 (column 3), which suggest some increasing returns to scale. Nerlove, however, had found evidence that at extremely large firm sizes, economies of scale diminished and eventually disappeared. To account for this (essentially a classical U-shaped average cost curve), he appended a quadratic term in log output in the cost function. The single equation and maximum likelihood multivariate regression estimates are given in the second and fourth sets of results.

[26]The invariance result is proved in Barten (1969).

[27]Some additional results on the method are given by Revankar (1976).

[28]Results based on Nerlove's full data set are given in Example 7.3. We have recomputed the values given in Table 14.5. Note that Nerlove used base 10 logs while we have used natural logs in our computations.

TABLE 14.5 Regression Estimates (Standard Errors in Parentheses)

	Ordinary Least Squares				*Multivariate Regression*			
β_0	−4.686	(0.885)	−3.764	(0.702)	−7.281	(0.104)	−5.962	(0.161)
β_q	0.721	(0.0174)	0.153	(0.0618)	0.798	(0.0147)	0.303	(0.0570)
β_{qq}	—		0.0505	(0.00536)	—		0.0414	(0.00493)
β_k	−0.00847	(0.191)	0.0739	(0.150)	0.424	(0.00945)	0.424	(0.00943)
β_1	0.594	(0.205)	0.481	(0.161)	0.106	(0.00380)	0.106	(0.00380)
β_f	0.414	(0.0989)	0.445	(0.0777)	0.470	(0.0100)	0.470	(0.0100)
R^2	0.9516		0.9581		—	—		
Log $\lvert\mathbf{W}\rvert$	—		—		−12.6726		−13.02248	

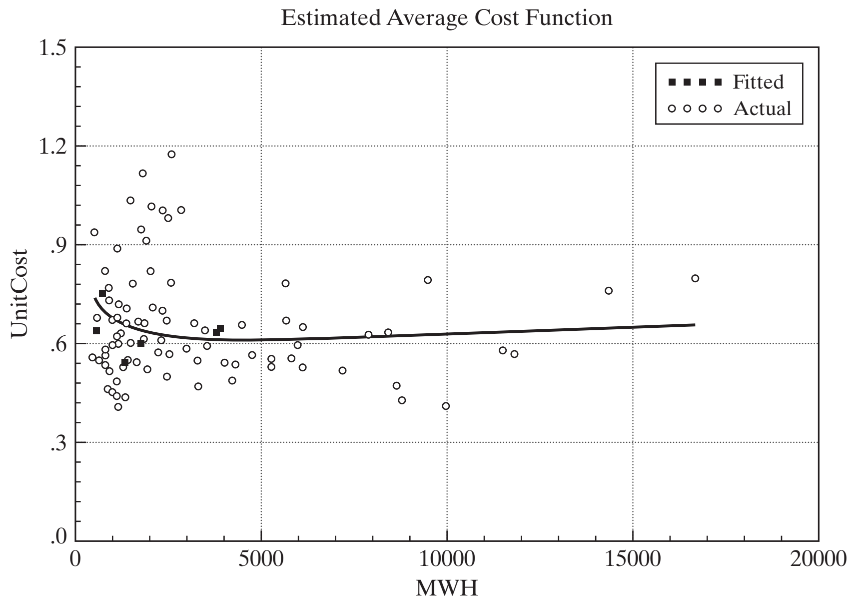


FIGURE 14.3 Predicted and Actual Average Costs.

The quadratic output term gives the cost function the expected U-shape. We can determine the point where average cost reaches its minimum by equating $\partial \ln C/\partial \ln q$ to 1. This is $q^* = \exp[(1 - \beta_q)/(2\beta_{qq})]$. For the multivariate regression, this value is $q^* = 4527$. About 85 percent of the firms in the sample had output less than this, so by these estimates, most firms in the sample had not yet exhausted the available economies of scale. Figure 14.3 shows predicted and actual average costs for the sample. (In order to obtain a reasonable scale, the smallest one third of the firms are omitted from the figure. Predicted average costs are computed at the sample averages of the input prices. The figure does reveal that that beyond a quite small scale, the economies of scale, while perhaps statistically significant, are economically quite small.

14.3.2 FLEXIBLE FUNCTIONAL FORMS: THE TRANSLOG COST FUNCTION

The literatures on production and cost and on utility and demand have evolved in several directions. In the area of models of producer behavior, the classic paper by Arrow et al. (1961) called into question the inherent restriction of the Cobb–Douglas model that all elasticities of factor substitution are equal to 1. Researchers have since developed numerous flexible functions that allow substitution to be unrestricted (i.e., not even constant).[29] Similar strands of literature have appeared in the analysis of commodity demands.[30] In this section, we examine in detail a model of production.

Suppose that production is characterized by a production function, $Y = f(\mathbf{x})$. The solution to the problem of minimizing the cost of producing a specified output rate given a set of factor prices produces the cost-minimizing set of factor demands $x_i = x_i(Y, \mathbf{p})$. The total cost of production is given by the cost function,

$$C = \sum_{i=1}^{M} p_i x_i(Y, \mathbf{p}) = C(Y, \mathbf{p}). \tag{14-32}$$

If there are constant returns to scale, then it can be shown that $C = Yc(\mathbf{p})$ or

$$C/Y = c(\mathbf{p}),$$

where $c(\mathbf{p})$ is the unit or average cost function.[31] The cost-minimizing factor demands are obtained by applying Shephard's (1970) lemma, which states that if $C(Y, \mathbf{p})$ gives the minimum total cost of production, then the cost-minimizing set of factor demands is given by

$$x_i^* = \frac{\partial C(Y, \mathbf{p})}{\partial p_i} = \frac{Y \partial c(\mathbf{p})}{\partial p_i}. \tag{14-33}$$

Alternatively, by differentiating logarithmically, we obtain the cost-minimizing factor cost shares:

$$s_i = \frac{\partial \log C(Y, \mathbf{p})}{\partial \log p_i} = \frac{p_i x_i}{C}. \tag{14-34}$$

With constant returns to scale, $\ln C(Y, \mathbf{p}) = \log Y + \log c(\mathbf{p})$, so

$$s_i = \frac{\partial \log c(\mathbf{p})}{\partial \log p_i}. \tag{14-35}$$

[29] See, in particular, Berndt and Christensen (1973). Two useful surveys of the topic are Jorgenson (1983) and Diewert (1974).

[30] See, for example, Christensen, Jorgenson, and Lau (1975) and two surveys, Deaton and Muellbauer (1980) and Deaton (1983). Berndt (1990) contains many useful results.

[31] The Cobb–Douglas function of the previous section gives an illustration. The restriction of constant returns to scale is $\beta_y = 1$, which is equivalent to $C = Yc(\mathbf{p})$. Nerlove's more general version of the cost function allows nonconstant returns to scale. See Christensen and Greene (1976) and Diewert (1974) for some of the formalities of the cost function and its relationship to the structure of production.

In many empirical studies, the objects of estimation are the elasticities of factor substitution and the own price elasticities of demand, which are given by

$$\theta_{ij} = \frac{c(\partial^2 c/\partial p_i \partial p_j)}{(\partial c/\partial p_i)(\partial c/\partial p_j)}$$

and

$$\eta_{ii} = s_i \theta_{ii}.$$

By suitably parameterizing the cost function (14-32) and the cost shares (14-33), we obtain an M or $M+1$ equation econometric model that can be used to estimate these quantities.[32]

The transcendental logarithmic, or translog, function is the most frequently used flexible function in empirical work.[33] By expanding $\log c(\mathbf{p})$ in a second-order Taylor series about the point $\log \mathbf{p} = \mathbf{0}$, we obtain

$$\log c \approx \beta_0 + \sum_{i=1}^{M} \left(\frac{\partial \log c}{\partial \log p_i}\right) \log p_i + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \left(\frac{\partial^2 \log c}{\partial \log p_i \, \partial \log p_j}\right) \log p_i \log p_j, \tag{14-36}$$

where all derivatives are evaluated at the expansion point. If we identify these derivatives as coefficients and impose the symmetry of the cross-price derivatives, then the cost function becomes

$$\begin{aligned} \log c = \beta_0 &+ \beta_1 \log p_1 + \cdots + \beta_M \log p_M + \delta_{11}\left(\tfrac{1}{2}\log^2 p_1\right) + \delta_{12} \log p_1 \log p_2 \\ &+ \delta_{22}\left(\tfrac{1}{2}\log^2 p_2\right) + \cdots + \delta_{MM}\left(\tfrac{1}{2}\log^2 p_M\right). \end{aligned} \tag{14-37}$$

This is the translog cost function. If δ_{ij} equals zero, then it reduces to the Cobb–Douglas function we looked at earlier. The cost shares are given by

$$\begin{aligned} s_1 &= \frac{\partial \log c}{\partial \log p_1} = \beta_1 + \delta_{11} \log p_1 + \delta_{12} \log p_2 + \cdots + \delta_{1M} \log p_M, \\ s_2 &= \frac{\partial \log c}{\partial \log p_2} = \beta_2 + \delta_{12} \log p_1 + \delta_{22} \log p_2 + \cdots + \delta_{2M} \log p_M, \\ &\vdots \\ s_M &= \frac{\partial \log c}{\partial \log p_M} = \beta_M + \delta_{1M} \log p_1 + \delta_{2M} \log p_2 + \cdots + \delta_{MM} \log p_M. \end{aligned} \tag{14-38}$$

[32] The cost function is only one of several approaches to this study. See Jorgenson (1983) for a discussion.

[33] See Example 2.4. The function was developed by Kmenta (1967) as a means of approximating the CES production function and was introduced formally in a series of papers by Berndt, Christensen, Jorgenson, and Lau, including Berndt and Christensen (1973) and Christensen et al. (1975). The literature has produced something of a competition in the development of exotic functional forms. The translog function has remained the most popular, however, and by one account, Guilkey, Lovell, and Sickles (1983) is the most reliable of several available alternatives. See also Example 6.2.

The cost shares must sum to 1, which requires, in addition to the symmetry restrictions already imposed,

$$\beta_1 + \beta_2 + \cdots + \beta_M = 1,$$
$$\sum_{i=1}^{M} \delta_{ij} = 0 \quad \text{(column sums equal zero)}, \tag{14-39}$$
$$\sum_{j=1}^{M} \delta_{ij} = 0 \quad \text{(row sums equal zero)}.$$

The system of share equations provides a seemingly unrelated regressions model that can be used to estimate the parameters of the model.[34] To make the model operational, we must impose the restrictions in (14-39) and solve the problem of singularity of the disturbance covariance matrix of the share equations. The first is accomplished by dividing the first $M-1$ prices by the Mth, thus eliminating the last term in each row and column of the parameter matrix. As in the Cobb–Douglas model, we obtain a nonsingular system by dropping the Mth share equation. We compute maximum likelihood estimates of the parameters to ensure invariance with respect to the choice of which share equation we drop. For the translog cost function, the elasticities of substitution are particularly simple to compute once the parameters have been estimated:

$$\theta_{ij} = \frac{\delta_{ij} + s_i s_j}{s_i s_j}, \quad \theta_{ii} = \frac{\delta_{ii} + s_i(s_i - 1)}{s_i^2}. \tag{14-40}$$

These elasticities will differ at every data point. It is common to compute them at some central point such as the means of the data.[35]

Example 14.5 A Cost Function for U.S. Manufacturing

A number of recent studies using the translog methodology have used a four-factor model, with capital K, labor L, energy E, and materials M, the factors of production. Among the first studies to employ this methodology was Berndt and Wood's (1975) estimation of a translog cost function for the U.S. manufacturing sector. The three factor shares used to estimate the model are

$$s_K = \beta_K + \delta_{KK} \log\left(\frac{p_K}{p_M}\right) + \delta_{KL} \log\left(\frac{p_L}{p_M}\right) + \delta_{KE} \log\left(\frac{p_E}{p_M}\right),$$
$$s_L = \beta_L + \delta_{KL} \log\left(\frac{p_K}{p_M}\right) + \delta_{LL} \log\left(\frac{p_L}{p_M}\right) + \delta_{LE} \log\left(\frac{p_E}{p_M}\right),$$
$$s_E = \beta_E + \delta_{KE} \log\left(\frac{p_K}{p_M}\right) + \delta_{LE} \log\left(\frac{p_L}{p_M}\right) + \delta_{EE} \log\left(\frac{p_E}{p_M}\right).$$

[34] The cost function may be included, if desired, which will provide an estimate of β_0 but is otherwise inessential. Absent the assumption of constant returns to scale, however, the cost function will contain parameters of interest that do not appear in the share equations. As such, one would want to include it in the model. See Christensen and Greene (1976) for an example.

[35] They will also be highly nonlinear functions of the parameters and the data. A method of computing asymptotic standard errors for the estimated elasticities is presented in Anderson and Thursby (1986).

TABLE 14.6 Parameter Estimates (Standard Errors in Parentheses)

β_K	0.05690	(0.00134)	δ_{KM}	−0.0189	(0.00971)
β_L	0.2534	(0.00210)	δ_{LL}	0.07542	(0.00676)
β_E	0.0444	(0.00085)	δ_{LE}	−0.00476	(0.00234)
β_M	0.6542	(0.00330)	δ_{LM}	−0.07061	(0.01059)
δ_{KK}	0.02951	(0.00580)	δ_{EE}	0.01838	(0.00499)
δ_{KL}	−0.000055	(0.00385)	δ_{EM}	−0.00299	(0.00799)
δ_{KE}	−0.01066	(0.00339)	δ_{MM}	0.09237	(0.02247)

TABLE 14.7 Estimated Elasticities

	Capital	*Labor*	*Energy*	*Materials*
		Cost Shares for 1959		
Fitted share	0.05643	0.27451	0.04391	0.62515
Actual share	0.06185	0.27303	0.04563	0.61948
		Implied Elasticities of Substitution		
Capital	−7.783			
Labor	0.9908	−1.643		
Energy	−3.230	0.6021	−12.19	
Materials	0.4581	0.5896	0.8834	−0.3623
		Implied Own Price Elasticities ($s_m\theta_{mm}$)		
	−0.4392	−0.4510	−0.5353	−0.2265

Berndt and Wood's data are reproduced in Appendix Table F14.1. Maximum likelihood estimates of the full set of parameters are given in Table 14.6.[36]

The implied estimates of the elasticities of substitution and demand for 1959 (the central year in the data) are derived in Table 14.7 using the fitted cost shares. The departure from the Cobb–Douglas model with unit elasticities is substantial. For example, the results suggest almost no substitutability between energy and labor[37] and some complementarity between capital and energy.

14.4 NONLINEAR SYSTEMS AND GMM ESTIMATION

We now consider estimation of nonlinear systems of equations. The underlying theory is essentially the same as that for linear systems. We briefly consider two cases in this section, maximum likelihood (or FGLS) estimation and GMM estimation. Since the

[36]These estimates are not the same as those reported by Berndt and Wood. To purge their data of possible correlation with the disturbances, they first regressed the prices on 10 exogenous macroeconomic variables, such as U.S. population, government purchases of labor services, real exports of durable goods, and U.S. tangible capital stock, and then based their analysis on the fitted values. The estimates given here are, in general, quite close to those given by Berndt and Wood. For example, their estimates of the first five parameters are 0.0564, 0.2539, 0.0442, 0.6455, and 0.0254.

[37]Berndt and Wood's estimate of θ_{EL} for 1959 is 0.64.

theory *is* essentially that of Section 14.2.4, most of the following will describe practical aspects of estimation.

Consider estimation of the parameters of the equation system

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{h}_1(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_1,\\ \mathbf{y}_2 &= \mathbf{h}_2(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_2,\\ &\vdots\\ \mathbf{y}_M &= \mathbf{h}_M(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_M.\end{aligned} \tag{14-41}$$

There are M equations in total, to be estimated with $t = 1, \dots, T$ observations. There are K parameters in the model. No assumption is made that each equation has "its own" parameter vector; we simply use some of or all the K elements in $\boldsymbol{\beta}$ in each equation. Likewise, there is a set of T observations on each of P independent variables $\mathbf{x}_p$, $p = 1, \dots, P$, some of or all that appear in each equation. For convenience, the equations are written generically in terms of the full $\boldsymbol{\beta}$ and $\mathbf{X}$. The disturbances are assumed to have zero means and contemporaneous covariance matrix $\boldsymbol{\Sigma}$. We will leave the extension to autocorrelation for more advanced treatments.

14.4.1 GLS ESTIMATION

In the multivariate regression model, if $\boldsymbol{\Sigma}$ is known, then the generalized least squares estimator of $\boldsymbol{\beta}$ is the vector that minimizes the generalized sum of squares

$$\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \sum_{i=1}^{M}\sum_{j=1}^{M}\sigma^{ij}[\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})]'[\mathbf{y}_j - \mathbf{h}_j(\boldsymbol{\beta}, \mathbf{X})], \tag{14-42}$$

where $\boldsymbol{\varepsilon}(\boldsymbol{\beta})$ is an $MT \times 1$ vector of disturbances obtained by stacking the equations and $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}$. [See (14-3).] As we did in Chapter 9, define the pseudoregressors as the derivatives of the $\mathbf{h}(\boldsymbol{\beta}, \mathbf{X})$ functions with respect to $\boldsymbol{\beta}$. That is, linearize each of the equations. Then the first-order condition for minimizing this sum of squares is

$$\frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{M}\sum_{j=1}^{M}\sigma^{ij}\left[2\mathbf{X}_i^{0\prime}(\boldsymbol{\beta})\boldsymbol{\varepsilon}_j(\boldsymbol{\beta})\right] = \mathbf{0}, \tag{14-43}$$

where σ^{ij} is the ijth element of $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{X}_i^0(\boldsymbol{\beta})$ is a $T \times K$ matrix of pseudoregressors from the linearization of the ith equation. (See Section 9.2.3.) If any of the parameters in $\boldsymbol{\beta}$ do not appear in the ith equation, then the corresponding column of $\mathbf{X}_i^0(\boldsymbol{\beta})$ will be a column of zeros.

This problem of estimation is doubly complex. In almost any circumstance, solution will require an iteration using one of the methods discussed in Appendix E. Second, of course, is that $\boldsymbol{\Sigma}$ is not known and must be estimated. Remember that efficient estimation in the multivariate regression model does not require an efficient estimator of $\boldsymbol{\Sigma}$, only a consistent one. Therefore, one approach would be to estimate the parameters of each equation separately using nonlinear least squares. This method will be inefficient if any of the equations share parameters, since that information will be ignored. But at this step, consistency is the objective, not efficiency. The resulting residuals can then be used to compute

$$\mathbf{S} = \frac{1}{T}\mathbf{E}'\mathbf{E}. \tag{14-44}$$

The second step of FGLS is the solution of (14-43), which will require an iterative procedure once again and can be based on **S** instead of $\boldsymbol{\Sigma}$. With well-behaved pseudoregressors, this second-step estimator is fully efficient. Once again, the same theory used for FGLS in the linear, single-equation case applies here.[38] Once the FGLS estimator is obtained, the appropriate asymptotic covariance matrix is estimated with

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}] = \left[\sum_{i=1}^{M}\sum_{j=1}^{M} s^{ij}\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{X}_j^0(\boldsymbol{\beta})\right]^{-1}.$$

There is a possible flaw in the strategy outlined above. It may not be possible to fit all the equations individually by nonlinear least squares. It is conceivable that identification of some of the parameters requires joint estimation of more than one equation. But as long as the full system identifies all parameters, there is a simple way out of this problem. Recall that all we need for our first step is a consistent set of estimators of the elements of $\boldsymbol{\beta}$. It is easy to show that the preceding defines a **GMM estimator** (see Chapter 18.) We can use this result to devise an alternative, simple strategy. The weighting of the sums of squares and cross products in (14-42) by σ^{ij} produces an efficient estimator of $\boldsymbol{\beta}$. Any other weighting based on some positive definite **A** would produce consistent, although inefficient, estimates. At this step, though, efficiency is secondary, so the choice of $\mathbf{A} = \mathbf{I}$ is a convenient candidate. Thus, for our first step, we can find $\boldsymbol{\beta}$ to minimize

$$\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \sum_{i=1}^{M}[\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})]'[\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})] = \sum_{i=1}^{M}\sum_{t=1}^{T}[y_{it} - h_i(\boldsymbol{\beta}, \mathbf{x}_{it})]^2.$$

(This estimator is just pooled nonlinear least squares, where the regression function varies across the sets of observations.) This step will produce the $\hat{\boldsymbol{\beta}}$ we need to compute **S**.

14.4.2 MAXIMUM LIKELIHOOD ESTIMATION

With normally distributed disturbances, the log-likelihood function for this model is still given by (14-18). Therefore, estimation of $\boldsymbol{\Sigma}$ is done exactly as before, using the **S** in (14-44). Likewise, the concentrated log-likelihood in (14-22) and the criterion function in (14-23) are unchanged. Therefore, one approach to maximum likelihood estimation is iterated FGLS, based on the results in Section 14.2.3. This method will require two levels of iteration, however, since for each estimated $\boldsymbol{\Sigma}(\boldsymbol{\beta}_l)$, written as a function of the estimates of $\boldsymbol{\beta}$ obtained at iteration l, a nonlinear, iterative solution is required to obtain $\boldsymbol{\beta}_{l+1}$. The iteration then returns to **S**. Convergence is based either on **S** or $\hat{\boldsymbol{\beta}}$; if one stabilizes, then the other will also.

The advantage of direct maximum likelihood estimation that was discussed in Section 14.2.4 is lost here because of the nonlinearity of the regressions; there is no

[38]Neither the nonlinearity nor the multiple equation aspect of this model brings any new statistical issues to the fore. By stacking the equations, we see that this model is simply a variant of the nonlinear regression model that we treated in Chapter 9 with the added complication of a nonscalar disturbance covariance matrix, which we analyzed in Chapter 10. The new complications are primarily practical.

convenient arrangement of parameters into a matrix $\mathbf{\Pi}$. But a few practical aspects to formulating the criterion function and its derivatives that may be useful do remain. Estimation of the model in (14-41) might be slightly more convenient if each equation did have its own coefficient vector. Suppose then that there is one underlying parameter vector $\boldsymbol{\beta}$ and that we formulate each equation as

$$h_{it} = h_i[\boldsymbol{\gamma}_i(\boldsymbol{\beta}), \mathbf{x}_{it}] + \varepsilon_{it}.$$

Then the derivatives of the log-likelihood function are built up from

$$\frac{\partial \ln|\mathbf{S}(\boldsymbol{\gamma})|}{\partial \boldsymbol{\gamma}_i} = \mathbf{d}_i = -\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{M} s^{ij}\mathbf{x}_{it}^0(\boldsymbol{\gamma}_i)e_{jt}(\boldsymbol{\gamma}_j)\right), \quad i = 1, \dots, M. \tag{14-45}$$

It remains to impose the equality constraints that have been built into the model. Since each $\boldsymbol{\gamma}_i$ is built up just by extracting elements from $\boldsymbol{\beta}$, the relevant derivative with respect to $\boldsymbol{\beta}$ is just a sum of those with respect to $\boldsymbol{\gamma}$.

$$\frac{\partial \ln L_c}{\partial \beta_k} = \sum_{i=1}^{n}\left[\sum_{g=1}^{K_i}\frac{\partial \ln L_c}{\partial \gamma_{ig}}\mathbf{1}(\gamma_{ig} = \beta_k)\right],$$

where $\mathbf{1}(\gamma_{ig} = \beta_k)$ equals 1 if γ_{ig} equals β_k and 0 if not. This derivative can be formulated fairly simply as follows. There are a total of $G = \sum_{i=1}^{n} K_i$ parameters in $\boldsymbol{\gamma}$, but only $K < G$ underlying parameters in $\boldsymbol{\beta}$. Define the matrix $\mathbf{F}$ with G rows and K columns. Then let $\mathbf{F}_{gj} = 1$ if $\gamma_g = \beta_j$ and 0 otherwise. Thus, there is exactly one 1 and $K - 1$ 0s in each row of $\mathbf{F}$. Let $\mathbf{d}$ be the $G \times 1$ vector of derivatives obtained by stacking $\mathbf{d}_i$ from (14-77). Then

$$\frac{\partial \ln L_c}{\partial \boldsymbol{\beta}} = \mathbf{F}'\mathbf{d}.$$

The Hessian is likewise computed as a simple sum of terms. We can construct it in blocks using

$$\mathbf{H}_{ij} = \frac{\partial^2 \ln L_c}{\partial \boldsymbol{\gamma}_i\,\partial \boldsymbol{\gamma}_j'} = -\sum_{t=1}^{T} s^{ij}\mathbf{x}_{it}^0(\boldsymbol{\gamma}_i)\mathbf{x}_{jt}^0(\boldsymbol{\gamma}_j)'.$$

The asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ is once again a sum of terms:

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

14.4.3 GMM ESTIMATION

All the preceding estimation techniques (including the linear models in the earlier sections of this chapter) can be obtained as GMM estimators. Suppose that in the general formulation of the model in (14-41), we allow for nonzero correlation between $\mathbf{x}_{it}^0$ and ε_{is}. (It will not always be present, but we generalize the model to allow this correlation as a possibility.) Suppose as well that there are a set of instrumental variables $\mathbf{z}_t$ such that

$$E[\mathbf{z}_t\varepsilon_{it}] = \mathbf{0}, \quad t = 1, \dots, T \quad \text{and} \quad i = 1, \dots, M. \tag{14-46}$$

(We could allow a separate set of instrumental variables for each equation, but it would needlessly complicate the presentation.)

Under these assumptions, the nonlinear FGLS and ML estimators above will be inconsistent. But a relatively minor extension of the instrumental variables technique developed for the single equation case in Section 10.4 can be used instead. The sample analog to (14-46) is

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t[y_{it} - h_i(\boldsymbol{\beta}, \mathbf{x}_t)] = \mathbf{0}, \quad i = 1, \dots, M.$$

If we use this result for each equation in the system, one at a time, then we obtain exactly the GMM estimator discussed in Section 10.4. But in addition to the efficiency loss that results from not imposing the cross-equation constraints in $\boldsymbol{\gamma}_i$, we would also neglect the correlation between the disturbances. Let

$$\frac{1}{T}\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z} = E\left[\frac{\mathbf{Z}'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'\mathbf{Z}}{T}\right]. \tag{14-47}$$

The GMM criterion for estimation in this setting is

$$\begin{aligned} q &= \sum_{i=1}^{M}\sum_{j=1}^{M}[(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X}))'\mathbf{Z}/T][\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'(\mathbf{y}_j - \mathbf{h}_j(\boldsymbol{\beta}, \mathbf{X}))/T] \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}[\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})'\mathbf{Z}/T][\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T], \end{aligned} \tag{14-48}$$

where $[\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}$ denotes the ijth block of the inverse of the matrix with the ijth block equal to $\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T$. (This matrix is laid out in full in Section 15.6.3.)

GMM estimation would proceed in several passes. To compute any of the variance parameters, we will require an initial consistent estimator of $\boldsymbol{\beta}$. This step can be done with equation-by-equation nonlinear instrumental variables—see Section 10.2.4—although if equations have parameters in common, then a choice must be made as to which to use. At the next step, the familiar White or Newey–West technique is used to compute, block by block, the matrix in (14-47). Since it is based on a consistent estimator of $\boldsymbol{\beta}$ (we assume), this matrix need not be recomputed. Now, with this result in hand, an iterative solution to the maximization problem in (14-48) can be sought, for example, using the methods of Appendix E. The first-order conditions are

$$\frac{\partial q}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{M}\sum_{j=1}^{M}\left[\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{Z}/T\right][\mathbf{Z}'\mathbf{W}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T] = \mathbf{0}. \tag{14-49}$$

Note again that the blocks of the inverse matrix in the center are extracted from the larger constructed matrix *after inversion*. [This brief discussion might understate the complexity of the optimization problem in (14-48), but that is inherent in the procedure.] At completion, the asymptotic covariance matrix for the GMM estimator is estimated with

$$\mathbf{V}_{\text{GMM}} = \frac{1}{T}\left[\sum_{i=1}^{M}\sum_{j=1}^{M}\left[\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{Z}/T\right][\mathbf{Z}'\mathbf{W}_{ij}\mathbf{Z}/T]^{ij}\left[\mathbf{Z}'\mathbf{X}_j^0(\boldsymbol{\beta})/T\right]\right]^{-1}.$$

14.5 SUMMARY AND CONCLUSIONS

This chapter has surveyed use of the seemingly unrelated regressions model. The SUR model is an application of the generalized regression model introduced in Chapter 10. The advantage of the SUR formulation is the rich variety of behavioral models that fit into this framework. We began with estimation and inference with the SUR model, treating it essentially as a generalized regression. The major difference between this set of results and the single equation model in Chapter 10 is practical. While the SUR model is, in principle a single equation GR model with an elaborate covariance structure, special problems arise when we explicitly recognize its intrinsic nature as a set of equations linked by their disturbances. The major result for estimation at this step is the feasible GLS estimator. In spite of its apparent complexity, we can estimate the SUR model by a straightforward two step GLS approach that is similar to the one we used for models with heteroscedasticity in Chapter 11. We also extended the SUR model to autocorrelation and heteroscedasticity, as in Chapters 11 and 12 for the single equation. Once again, the multiple equation nature of the model complicates these applications. Maximum likelihood is an alternative method that is useful for systems of demand equations. This chapter examined a number of applications of the SUR model. Much of the empirical literature in finance focuses on the capital asset pricing model, which we considered in Section 14.2.5. Section 14.2.6 developed an important result on estimating systems in which some equations are derived from the set by excluding some of the variables. The block of zeros case is useful in the VAR models used in causality testing in Section 19.6.5. Section 14.3 presented one of the most common recent applications of the seemingly unrelated regressions model, the estimation of demand systems. One of the signature features of this literature is the seamless transition from the theoretical models of optimization of consumers and producers to the sets of empirical demand equations derived from Roy's identity for consumers and Shephard's lemma for producers.

Key Terms and Concepts

- Autocorrelation
- Capital asset pricing model
- Concentrated log-likelihood
- Demand system
- Exclusion restriction
- Expenditure system
- Feasible GLS
- Flexible functional form
- Generalized least squares
- GMM estimator
- Heteroscedasticity
- Homogeneity restriction
- Identical regressors
- Invariance of MLE
- Kronecker product
- Lagrange multiplier statistic
- Likelihood ratio statistic
- Maximum likelihood
- Multivariate regression
- Seemingly unrelated regressions
- Wald statistic

Exercises

1. A sample of 100 observations produces the following sample data:

$$\bar{y}_1 = 1, \quad \bar{y}_2 = 2,$$
$$\mathbf{y}_1'\mathbf{y}_1 = 150,$$
$$\mathbf{y}_2'\mathbf{y}_2 = 550,$$
$$\mathbf{y}_1'\mathbf{y}_2 = 260.$$

The underlying bivariate regression model is

$$y_1 = \mu + \varepsilon_1,$$

$$y_2 = \mu + \varepsilon_2.$$

a. Compute the OLS estimate of μ, and estimate the sampling variance of this estimator.
b. Compute the FGLS estimate of μ and the sampling variance of the estimator.

2. Consider estimation of the following two equation model:

$$y_1 = \beta_1 + \varepsilon_1,$$

$$y_2 = \beta_2 x + \varepsilon_2.$$

A sample of 50 observations produces the following moment matrix:

$$\begin{array}{c} \\ 1 \\ y_1 \\ y_2 \\ x \end{array}\begin{array}{c} \begin{array}{cccc} 1 & y_1 & y_2 & x \end{array} \\ \begin{bmatrix} 50 & & & \\ 150 & 500 & & \\ 50 & 40 & 90 & \\ 100 & 60 & 50 & 100 \end{bmatrix} \end{array}.$$

a. Write the explicit formula for the GLS estimator of $[\beta_1, \beta_2]$. What is the asymptotic covariance matrix of the estimator?
b. Derive the OLS estimator and its sampling variance in this model.
c. Obtain the OLS estimates of β_1 and β_2, and estimate the sampling covariance matrix of the two estimates. Use n instead of $(n - 1)$ as the divisor to compute the estimates of the disturbance variances.
d. Compute the FGLS estimates of β_1 and β_2 and the estimated sampling covariance matrix.
e. Test the hypothesis that $\beta_2 = 1$.

3. The model

$$y_1 = \beta_1 x_1 + \varepsilon_1,$$

$$y_2 = \beta_2 x_2 + \varepsilon_2$$

satisfies all the assumptions of the classical multivariate regression model. All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$\begin{array}{c} \\ y_1 \\ y_2 \\ x_1 \\ x_2 \end{array}\begin{array}{c} \begin{array}{cccc} y_1 & y_2 & x_1 & x_2 \end{array} \\ \begin{bmatrix} 20 & 6 & 4 & 3 \\ 6 & 10 & 3 & 6 \\ 4 & 3 & 5 & 2 \\ 3 & 6 & 2 & 10 \end{bmatrix} \end{array}.$$

a. Compute the FGLS estimates of β_1 and β_2.
b. Test the hypothesis that $\beta_1 = \beta_2$.
c. Compute the maximum likelihood estimates of the model parameters.
d. Use the likelihood ratio test to test the hypothesis in part b.

4. Prove that in the model

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1,$$
$$\mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2,$$

generalized least squares is equivalent to equation-by-equation ordinary least squares if $\mathbf{X}_1 = \mathbf{X}_2$. Does your result hold if it is also known that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$?

5. Consider the two-equation system

$$\begin{aligned} y_1 &= \beta_1 x_1 \qquad\qquad\quad + \varepsilon_1, \\ y_2 &= \quad\;\; \beta_2 x_2 + \beta_3 x_3 + \varepsilon_2. \end{aligned}$$

Assume that the disturbance variances and covariance are known. Now suppose that the analyst of this model applies GLS but erroneously omits x_3 from the second equation. What effect does this specification error have on the consistency of the estimator of β_1?

6. Consider the system

$$\begin{aligned} y_1 &= \alpha_1 + \beta x + \varepsilon_1, \\ y_2 &= \alpha_2 \qquad\;\; + \varepsilon_2. \end{aligned}$$

The disturbances are freely correlated. Prove that GLS applied to the system leads to the OLS estimates of α_1 and α_2 but to a mixture of the least squares slopes in the regressions of y_1 and y_2 on x as the estimator of β. What is the mixture? To simplify the algebra, assume (with no loss of generality) that $\bar{x} = 0$.

7. For the model

$$\begin{aligned} y_1 &= \alpha_1 + \beta x + \varepsilon_1, \\ y_2 &= \alpha_2 \qquad\;\; + \varepsilon_2, \\ y_3 &= \alpha_3 \qquad\;\; + \varepsilon_3, \end{aligned}$$

assume that $y_{i2} + y_{i3} = 1$ at every observation. Prove that the sample covariance matrix of the least squares residuals from the three equations will be singular, thereby precluding computation of the FGLS estimator. How could you proceed in this case?

8. Continuing the analysis of Section 14.3.2, we find that a translog cost function for one output and three factor inputs that does not impose constant returns to scale is

$$\begin{aligned} \ln C = \alpha &+ \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln p_3 + \delta_{11}\tfrac{1}{2}\ln^2 p_1 + \delta_{12}\ln p_1 \ln p_2 \\ &+ \delta_{13}\ln p_1 \ln p_3 + \delta_{22}\tfrac{1}{2}\ln^2 p_2 + \delta_{23}\ln p_2 \ln p_3 + \delta_{33}\tfrac{1}{2}\ln^2 p_3 \\ &+ \gamma_{y1}\ln Y \ln p_1 + \gamma_{y2}\ln Y \ln p_2 + \gamma_{y3}\ln Y \ln p_3 \\ &+ \beta_y \ln Y + \beta_{yy}\tfrac{1}{2}\ln^2 Y + \varepsilon_c. \end{aligned}$$

The factor share equations are

$$\begin{aligned} S_1 &= \beta_1 + \delta_{11}\ln p_1 + \delta_{12}\ln p_2 + \delta_{13}\ln p_3 + \gamma_{y1}\ln Y + \varepsilon_1, \\ S_2 &= \beta_2 + \delta_{12}\ln p_1 + \delta_{22}\ln p_2 + \delta_{23}\ln p_3 + \gamma_{y2}\ln Y + \varepsilon_2, \\ S_3 &= \beta_3 + \delta_{13}\ln p_1 + \delta_{23}\ln p_2 + \delta_{33}\ln p_3 + \gamma_{y3}\ln Y + \varepsilon_3. \end{aligned}$$

[See Christensen and Greene (1976) for analysis of this model.]

a. The three factor shares must add identically to 1. What restrictions does this requirement place on the model parameters?
b. Show that the adding-up condition in (14-39) can be imposed directly on the model by specifying the translog model in (C/p_3), (p_1/p_3), and (p_2/p_3) and dropping the third share equation. (See Example 14.5.) Notice that this reduces the number of free parameters in the model to 10.
c. Continuing Part b, the model as specified with the symmetry and equality restrictions has 15 parameters. By imposing the constraints, you reduce this number to 10 in the estimating equations. How would you obtain estimates of the parameters not estimated directly?

The remaining parts of this exercise will require specialized software. The **E-Views, TSP, Stata** or **LIMDEP,** programs noted in the preface are four that could be used. All estimation is to be done using the data used in Section 14.3.1.

d. Estimate each of the three equations you obtained in Part b by ordinary least squares. Do the estimates appear to satisfy the cross-equation equality and symmetry restrictions implied by the theory?
e. Using the data in Section 14.3.1, estimate the full system of three equations (cost and the two independent shares), imposing the symmetry and cross-equation equality constraints.
f. Using your parameter estimates, compute the estimates of the elasticities in (14-40) at the means of the variables.
g. Use a likelihood ratio statistic to test the joint hypothesis that $\gamma_{yi} = 0$, $i = 1, 2, 3$. [Hint: Just drop the relevant variables from the model.]