at $\boldsymbol{X}\hat{\beta}$ instead of at zero. The right angle formed by $\boldsymbol{y} - \boldsymbol{X}\hat{\beta}$ and $\mathcal{S}(\boldsymbol{X})$ is the key feature of least squares. At any other point in $\mathcal{S}(\boldsymbol{X})$, such as $\boldsymbol{X}\beta'$ in the figure, $\boldsymbol{y} - \boldsymbol{X}\beta'$ does not form a right angle with $\mathcal{S}(\boldsymbol{X})$ and, as a consequence, $\|\boldsymbol{y} - \boldsymbol{X}\beta'\|$ must necessarily be larger than $\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|$.

The vector of derivatives of the SSR (1.02) with respect to the elements of $\boldsymbol{\beta}$ is

$$-2\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

which must equal $\boldsymbol{0}$ at a minimum. Since we have assumed that the columns of $\boldsymbol{X}$ are linearly independent, the matrix $\boldsymbol{X}^{\top}\boldsymbol{X}$ must have full rank. This, combined with that fact that any matrix of the form $\boldsymbol{X}^{\top}\boldsymbol{X}$ is necessarily nonnegative definite, implies that the sum of squared residuals is a strictly convex function of $\boldsymbol{\beta}$ and must therefore have a unique minimum. Thus $\hat{\boldsymbol{\beta}}$ is uniquely determined by the **normal equations**

$$\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}. \tag{1.03}$$

These normal equations say that the vector $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ must be orthogonal to all of the columns of $\boldsymbol{X}$ and hence to any vector that lies in the space spanned by those columns. The normal equations (1.03) are thus simply a way of stating algebraically what Figure 1.2 showed geometrically, namely, that $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ must form a right angle with $\mathcal{S}(\boldsymbol{X})$.

Since the matrix $\boldsymbol{X}^{\top}\boldsymbol{X}$ has full rank, we can always invert it to solve the normal equations for $\hat{\boldsymbol{\beta}}$. We obtain the standard formula:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}. \tag{1.04}$$

Even if $\boldsymbol{X}$ is not of full rank, the fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ are uniquely defined, because $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is simply the point in $\mathcal{S}(\boldsymbol{X})$ that is closest to $\boldsymbol{y}$. Look again at Figure 1.2 and suppose that $\boldsymbol{X}$ is an $n \times 2$ matrix, but of rank only one. The geometrical point $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is still uniquely defined. However, since $\boldsymbol{\beta}$ is now a 2–vector and $\mathcal{S}(\boldsymbol{X})$ is just one-dimensional, the vector $\hat{\boldsymbol{\beta}}$ is not uniquely defined. Thus the requirement that $\boldsymbol{X}$ have full rank is a purely algebraic requirement that is needed to obtain unique estimates $\hat{\boldsymbol{\beta}}$.

If we substitute the right-hand side of (1.04) for $\hat{\boldsymbol{\beta}}$ into $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, we obtain

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y} \equiv \boldsymbol{P}_X\boldsymbol{y}. \tag{1.05}$$

This equation defines the $n \times n$ matrix $\boldsymbol{P}_X \equiv \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}$, which **projects** the vector $\boldsymbol{y}$ orthogonally onto $\mathcal{S}(\boldsymbol{X})$. The matrix $\boldsymbol{P}_X$ is an example of an **orthogonal projection matrix**. Associated with every linear subspace of E^n are two such matrices, one of which projects any point in E^n onto that subspace, and one of which projects any point in E^n onto its orthogonal complement. The matrix that projects onto $\mathcal{S}^{\perp}(\boldsymbol{X})$ is

$$\boldsymbol{M}_X \equiv \mathbf{I} - \boldsymbol{P}_X \equiv \mathbf{I} - \boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top},$$

The modified version is known as the **centered** R^2, and we will denote it by R_c^2. It is defined as

$$R_c^2 \equiv 1 - \frac{\|\boldsymbol{M}_X \boldsymbol{y}\|^2}{\|\boldsymbol{M}_\iota \boldsymbol{y}\|^2}, \tag{1.09}$$

where

$$\boldsymbol{M}_\iota \equiv \mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top = \mathbf{I} - n^{-1} \boldsymbol{\iota} \boldsymbol{\iota}^\top$$

is the matrix that projects off the space spanned by the constant vector $\boldsymbol{\iota}$, which is simply a vector of n ones. When any vector is multiplied by $\boldsymbol{M}_\iota$, the result is a vector of deviations from the mean. Thus what the centered R^2 measures is the proportion of the total sum of squares of the regressand *around its mean* that is explained by the regressors.

An alternative expression for R_c^2 is

$$\frac{\|\boldsymbol{P}_X \boldsymbol{M}_\iota \boldsymbol{y}\|^2}{\|\boldsymbol{M}_\iota \boldsymbol{y}\|^2}, \tag{1.10}$$

but this is equal to (1.09) only if $\boldsymbol{P}_X \boldsymbol{\iota} = \boldsymbol{\iota}$, which means that $\mathcal{S}(\boldsymbol{X})$ must include the vector $\boldsymbol{\iota}$ (so that either one column of $\boldsymbol{X}$ must be a constant, or some linear combination of the columns of $\boldsymbol{X}$ must equal a constant). In this case, the equality must hold, because

$$\boldsymbol{M}_X \boldsymbol{M}_\iota \boldsymbol{y} = \boldsymbol{M}_X (\mathbf{I} - \boldsymbol{P}_\iota) \boldsymbol{y} = \boldsymbol{M}_X \boldsymbol{y},$$

the second equality here being a consequence of the fact that $\boldsymbol{M}_X$ annihilates $\boldsymbol{P}_\iota$ when $\boldsymbol{\iota}$ belongs to $\mathcal{S}(\boldsymbol{X})$. When this is not the case and (1.10) is not valid, there is no guarantee that R_c^2 will be positive. After all, there will be many cases in which a regressand $\boldsymbol{y}$ is better explained by a constant term than by some set of regressors that does not include a constant term. Clearly, if (1.10) is valid, R_c^2 must lie between 0 and 1, since (1.10) is then simply the uncentered R^2 for a regression of $\boldsymbol{M}_\iota \boldsymbol{y}$ on $\boldsymbol{X}$.

The use of the centered R^2 when $\boldsymbol{X}$ does not include a constant term or the equivalent is thus fraught with difficulties. Some programs for statistics and econometrics refuse to print an R^2 at all in this circumstance; others print R_u^2 (without always warning the user that they are doing so); some print R_c^2, defined as (1.09), which may be either positive or negative; and some print still other quantities, which would be equal to R_c^2 if $\boldsymbol{X}$ included a constant term but are not when it does not. Users of statistical software, be warned!

Notice that R^2 is an interesting number only because we used the least squares estimator $\hat{\boldsymbol{\beta}}$ to estimate $\boldsymbol{\beta}$. If we chose an estimate of $\boldsymbol{\beta}$, say $\tilde{\boldsymbol{\beta}}$, in any other way, so that the triangle in Figure 1.3 were no longer a right-angled triangle, we would find that the equivalents of the two definitions of R^2, (1.09) and (1.10), were not the same:

$$1 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{\|\boldsymbol{y}\|^2} \neq \frac{\|\boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{\|\boldsymbol{y}\|^2}.$$

seem to have introduced, and then reintroduced, it to econometricians. The theorem is much more general, and much more generally useful, than a casual reading of those papers might suggest, however. Among other things, it almost totally eliminates the need to invert partitioned matrices when one is deriving many standard results about ordinary (and nonlinear) least squares.

The FWL Theorem applies to any regression where there are two or more regressors, and these can logically be broken up into two groups. The regression can thus be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals}, \tag{1.18}$$

where $\boldsymbol{X}_1$ is $n \times k_1$ and $\boldsymbol{X}_2$ is $n \times k_2$, with $\boldsymbol{X} \equiv [\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ and $k = k_1 + k_2$. For example, $\boldsymbol{X}_1$ might be seasonal dummy variables or trend variables and $\boldsymbol{X}_2$ genuine economic variables. This was in fact the type of situation dealt with by Frisch and Waugh (1933) and Lovell (1963). Another possibility is that $\boldsymbol{X}_1$ might be regressors, the joint significance of which we desire to test, and $\boldsymbol{X}_2$ might be other regressors that are not being tested. Or $\boldsymbol{X}_1$ might be regressors that are known to be orthogonal to the regressand, and $\boldsymbol{X}_2$ might be regressors that are not orthogonal to it, a situation which arises very frequently when we wish to test nonlinear regression models; see Chapter 6.

Now consider another regression,

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals}, \tag{1.19}$$

where $\boldsymbol{M}_1$ is the matrix that projects off $\mathcal{S}(\boldsymbol{X}_1)$. In (1.19) we have first regressed $\boldsymbol{y}$ and each of the k_2 columns of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$ and then regressed the vector of residuals $\boldsymbol{M}_1\boldsymbol{y}$ on the $n \times k_2$ matrix of residuals $\boldsymbol{M}_1\boldsymbol{X}_2$. The FWL Theorem tells us that the residuals from regressions (1.18) and (1.19), and the OLS estimates of $\boldsymbol{\beta}_2$ from those two regressions, will be *numerically identical*. Geometrically, in regression (1.18) we project $\boldsymbol{y}$ directly onto $\mathcal{S}(\boldsymbol{X}) \equiv \mathcal{S}(\boldsymbol{X}_1, \boldsymbol{X}_2)$, while in regression (1.19) we first project $\boldsymbol{y}$ and all of the columns of $\boldsymbol{X}_2$ off $\mathcal{S}(\boldsymbol{X}_1)$ and then project the residuals $\boldsymbol{M}_1\boldsymbol{y}$ onto the span of the matrix of residuals, $\mathcal{S}(\boldsymbol{M}_1\boldsymbol{X}_2)$. The FWL Theorem tells us that these two apparently rather different procedures actually amount to the same thing.

The FWL Theorem can be proved in several different ways. One standard proof is based on the algebra of partitioned matrices. First, observe that the estimate of $\boldsymbol{\beta}_2$ from (1.19) is

$$\left(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y}. \tag{1.20}$$

This simple expression, which we will make use of many times, follows immediately from substituting $\boldsymbol{M}_1\boldsymbol{X}_2$ for $\boldsymbol{X}$ and $\boldsymbol{M}_1\boldsymbol{y}$ for $\boldsymbol{y}$ in expression (1.04) for the vector of OLS estimates. The algebraic proof would now use results on the inverse of a partitioned matrix (see Appendix A) to demonstrate that the OLS estimate from (1.18), $\hat{\boldsymbol{\beta}}_2$, is identical to (1.20) and would then go

OLS estimate of $\boldsymbol{\gamma}$ is

$$\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{Q}^{\top}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}^{\top}\boldsymbol{y} = \boldsymbol{Q}^{\top}\boldsymbol{y},$$

which is trivial to compute. It is equally easy to compute the fitted values $\boldsymbol{Q}\hat{\boldsymbol{\gamma}}$ and the residuals

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{Q}\hat{\boldsymbol{\gamma}} = \boldsymbol{y} - \boldsymbol{Q}\boldsymbol{Q}^{\top}\boldsymbol{y}. \tag{1.36}$$

Thus, if we are simply interested in residuals and/or fitted values, we do not need to compute $\hat{\boldsymbol{\beta}}$ at all.

Notice from (1.36) that the projection matrices $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ are equal to $\boldsymbol{Q}\boldsymbol{Q}^{\top}$ and $\mathbf{I} - \boldsymbol{Q}\boldsymbol{Q}^{\top}$, respectively. The simplicity of these expressions follows from the fact that $\boldsymbol{Q}$ forms an orthonormal basis for $\mathcal{S}(\boldsymbol{X})$. Geometrically, nothing would change in any of the figures we have drawn if we used $\boldsymbol{Q}$ instead of $\boldsymbol{X}$ as the matrix of regressors, since $\mathcal{S}(\boldsymbol{Q}) = \mathcal{S}(\boldsymbol{X})$. If we were to show the columns of $\boldsymbol{Q}$ in the figures, each column would be a point in $\mathcal{S}(\boldsymbol{X})$ located on the unit sphere (i.e., the sphere with radius one centered at the origin) and at right angles to the points representing the other columns of $\boldsymbol{Q}$.

In order to calculate $\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}$, which, along with the residuals and the fitted values, allow us to calculate all the main quantities of interest, we make use of the facts that $\hat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\hat{\boldsymbol{\gamma}}$ and

$$\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1} = \left(\boldsymbol{R}^{\top}\boldsymbol{Q}^{\top}\boldsymbol{Q}\boldsymbol{R}\right)^{-1} = \left(\boldsymbol{R}^{\top}\boldsymbol{R}\right)^{-1} = \boldsymbol{R}^{-1}(\boldsymbol{R}^{-1})^{\top}.$$

Thus, once we have computed $\boldsymbol{R}^{-1}$, we can very easily calculate the least squares estimates $\hat{\boldsymbol{\beta}}$ and their estimated covariance matrix (see Chapter 2). Since $\boldsymbol{R}$ is a triangular matrix, its inverse is very easily and cheaply computed; we do not even have to check for possible singularity, since $\boldsymbol{R}$ will fail to have full rank only if $\boldsymbol{X}$ does not have full rank, and that will already have shown up and been dealt with when we formed $\boldsymbol{Q}$ and $\boldsymbol{R}$.

The most costly part of these procedures is forming the matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ from $\boldsymbol{X}$. This requires a number of arithmetic operations that is roughly proportional to nk^2. Forming the matrix of sums and cross-products, which is the first step for methods based on solving the normal equations, also requires a number of operations proportional to nk^2, although the factor of proportionality is smaller. Thus linear regression by any method can become expensive when the number of regressors is large and/or the sample size is very large. If one is going to calculate many regressions using the same large data set, it makes sense to economize by doing the expensive calculations only once. Many regression packages allow users first to form the matrix of sums of squares and cross-products for all the variables in a data set and then to calculate estimates for a variety of regressions by retrieving the relevant rows and columns and using normal equation methods. If this approach is used, it is particularly important that the data be scaled so that the various regressors are not too dissimilar in mean and variance.

if $\boldsymbol{X}$ simply consisted of a constant vector, $\boldsymbol{e}_t^\top \boldsymbol{P}_X \boldsymbol{e}_t$ would equal $1/n$. Even when there is no constant term, h_t can never be 0 unless every element of $\boldsymbol{X}_t$ is 0. However, it is evidently quite possible for h_t to equal 1. Suppose, for example, that one column of $\boldsymbol{X}$ is the dummy variable $\boldsymbol{e}_t$. In that case, $h_t = \boldsymbol{e}_t^\top \boldsymbol{P}_X \boldsymbol{e}_t = \boldsymbol{e}_t^\top \boldsymbol{e}_t = 1$.

It is interesting to see what happens when we add a dummy variable $\boldsymbol{e}_t$ to a regression. It turns out that $\hat{u}_t$ will equal zero and that the tth observation will have no effect at all on any coefficient except the one corresponding to the dummy variable. The latter simply takes on whatever value is needed to make $\hat{u}_t = 0$, and the remaining coefficients are those that minimize the SSR for the remaining $n-1$ observations. These results are easily established by using the FWL Theorem.

Consider the following two regressions, where for ease of notation the data have been ordered so that observation t is the last observation, and $\boldsymbol{y}_{(t)}$ and $\boldsymbol{X}_{(t)}$ denote the first $n-1$ rows of $\boldsymbol{y}$ and $\boldsymbol{X}$, respectively:

$$\begin{bmatrix} \boldsymbol{y}_{(t)} \\ y_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{(t)} \\ \boldsymbol{X}_t \end{bmatrix} \boldsymbol{\beta} + \text{residuals,} \tag{1.43}$$

and

$$\begin{bmatrix} \boldsymbol{y}_{(t)} \\ y_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{(t)} & \boldsymbol{0} \\ \boldsymbol{X}_t & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \alpha \end{bmatrix} + \text{residuals.} \tag{1.44}$$

Regression (1.43) is simply the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, which yields parameter estimates $\hat{\boldsymbol{\beta}}$ and least squares residuals $\hat{\boldsymbol{u}}$. Regression (1.44) is regression (1.43) with $\boldsymbol{e}_t$ as an additional regressor. By the FWL Theorem, the estimate of $\boldsymbol{\beta}$ from (1.44) must be identical to the estimate of $\boldsymbol{\beta}$ from the regression

$$\boldsymbol{M}_t \begin{bmatrix} \boldsymbol{y}_{(t)} \\ y_t \end{bmatrix} = \boldsymbol{M}_t \begin{bmatrix} \boldsymbol{X}_{(t)} \\ \boldsymbol{X}_t \end{bmatrix} \boldsymbol{\beta} + \text{residuals,} \tag{1.45}$$

where $\boldsymbol{M}_t$ is the matrix that projects orthogonally onto $\mathbb{S}^\perp(\boldsymbol{e}_t)$. Multiplying any vector by $\boldsymbol{M}_t$ merely annihilates the last element of that vector. Thus regression (1.45) is simply

$$\begin{bmatrix} \boldsymbol{y}_{(t)} \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{(t)} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{\beta} + \text{residuals.} \tag{1.46}$$

The last observation, in which the regressand and all regressors are zero, obviously has no effect at all on parameter estimates. Regression (1.46) is therefore equivalent to regressing $\boldsymbol{y}_{(t)}$ on $\boldsymbol{X}_{(t)}$ and so must yield OLS estimates $\hat{\boldsymbol{\beta}}^{(t)}$. For regression (1.46), the residual for observation t is clearly zero; the FWL Theorem then implies that the residual for observation t from regression (1.44) must likewise be zero, which implies that $\hat{\alpha}$ must equal $y_t - \boldsymbol{X}_t \hat{\boldsymbol{\beta}}^{(t)}$.

These results make it easy to derive the results (1.40) and (1.41), which were earlier stated without proof. Readers who are not interested in the

with the values of certain variables. They may be the only variables about which we have information or the only ones that we are interested in for a particular purpose. If we had more information about potential explanatory variables, we might very well specify $x_t(\boldsymbol{\beta})$ differently so as to make use of that additional information.

It is sometimes desirable to make explicit the fact that $x_t(\boldsymbol{\beta})$ represents the **conditional mean** of y_t, that is, the mean of y_t conditional on the values of a number of other variables. The set of variables on which y_t is conditioned is often referred to as an **information set**. If Ω_t denotes the information set on which the expectation of y_t is to be conditioned, one could define $x_t(\boldsymbol{\beta})$ formally as $E(y_t \,|\, \Omega_t)$. There may be more than one such information set. Thus we might well have both

$$x_{1t}(\boldsymbol{\beta}_1) \equiv E(y_t \,|\, \Omega_{1t}) \quad \text{and} \quad x_{2t}(\boldsymbol{\beta}_2) \equiv E(y_t \,|\, \Omega_{2t}),$$

where Ω_{1t} and Ω_{2t} denote two different information sets. The functions $x_{1t}(\boldsymbol{\beta}_1)$ and $x_{2t}(\boldsymbol{\beta}_2)$ might well be quite different, and we might want to estimate both of them for different purposes. There are many circumstances in which we might not want to condition on all available information. For example, if the ultimate purpose of specifying a regression function is to use it for forecasting, there may be no point in conditioning on information that will not be available at the time the forecast is to be made. Even when we do want to take account of all available information, the fact that a certain variable belongs to Ω_t does not imply that it will appear in $x_t(\boldsymbol{\beta})$, since its value may tell us nothing useful about the conditional mean of y_t, and including it may impair our ability to estimate how other variables affect that conditional mean.

For any given dependent variable y_t and information set Ω_t, one is always at liberty to consider the difference $y_t - E(y_t \,|\, \Omega_t)$ as the error term associated with the tth observation. But for a *regression model* to be applicable, these differences must generally have the i.i.d. property. Actually, it is possible, when the sample size is large, to deal with cases in which the error terms are independent, but identically distributed only as regards their means, and not necessarily as regards their variances. We will discuss techniques for dealing with such cases in Chapters 16 and 17, in the latter of which we will also relax the independence assumption. As we will see in Chapter 3, however, conventional techniques for making inferences from regression models are unreliable when models lack the i.i.d. property, even when the regression function $x_t(\boldsymbol{\beta})$ is "correctly" specified. Thus we are in general not at liberty to choose an arbitrary information set and estimate a properly specified regression function based on it if we want to make inferences using conventional procedures.

There are, however, exceptional cases in which we can choose any information set we like, because models based on different information sets will always be mutually consistent. For example, suppose that the vector consisting of y_t and each of x_{1t} through x_{mt} is independently and identically

Chapter 3

Inference in
Nonlinear Regression Models

3.1 INTRODUCTION

Suppose that one is given a vector $\boldsymbol{y}$ of observations on some dependent variable, a vector $\boldsymbol{x}(\boldsymbol{\beta})$ of, in general nonlinear, regression functions, which may and normally will depend on independent variables, and the data needed to evaluate $\boldsymbol{x}(\boldsymbol{\beta})$. Then, assuming that these data allow one to identify all elements of the parameter vector $\boldsymbol{\beta}$ and that one has access to a suitable computer program for nonlinear least squares and enough computer time, one can always obtain NLS estimates $\hat{\boldsymbol{\beta}}$. In order to interpret these estimates, one generally makes the heroic assumption that the model is "correct," which means that $\boldsymbol{y}$ is in fact generated by a DGP from the family

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}). \tag{3.01}$$

Without this assumption, or some less restrictive variant, it would be very difficult to say anything about the properties of $\hat{\boldsymbol{\beta}}$, although in certain special cases one can do so.

It is clear that $\hat{\boldsymbol{\beta}}$ must be a vector of random variables, since it will depend on $\boldsymbol{y}$ and hence on the vector of error terms $\boldsymbol{u}$. Thus, if we are to make inferences about $\boldsymbol{\beta}$, we must recognize that $\hat{\boldsymbol{\beta}}$ is random and quantify its randomness. In Chapter 5, we will demonstrate that it is reasonable, when the sample size is large enough, to treat $\hat{\boldsymbol{\beta}}$ as being normally distributed around the true value of $\boldsymbol{\beta}$, which we may call $\boldsymbol{\beta}_0$. Thus the only thing we need to know if we are to make asymptotically valid inferences about $\boldsymbol{\beta}$ is the **covariance matrix** of $\hat{\boldsymbol{\beta}}$, say $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$. In the next section, we discuss how this covariance matrix may be estimated for linear and nonlinear regression models. In Section 3.3, we show how the resulting estimates may be used to make inferences about $\boldsymbol{\beta}$. In Section 3.4, we discuss the basic ideas that underlie all types of hypothesis testing. In Section 3.5, we then discuss procedures for testing hypotheses in linear regression models. In Section 3.6, we discuss similar procedures for testing hypotheses in nonlinear regression models. The latter section provides an opportunity to introduce the three

fundamental principles on which most hypothesis tests are based: the Wald, Lagrange multiplier, and likelihood ratio principles. Finally, in Section 3.7, we discuss the effects of imposing incorrect restrictions and introduce the notion of preliminary test estimators.

3.2 Covariance Matrix Estimation

In the case of the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \tag{3.02}$$

it is well known that when the DGP satisfies (3.02) for specific parameter values $\boldsymbol{\beta}_0$ and σ_0, the covariance matrix of the vector of OLS estimates $\hat{\boldsymbol{\beta}}$ is

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \sigma_0^2 \big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1}. \tag{3.03}$$

The proof of this familiar result is quite straightforward. The covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$ is defined as the expectation of the outer product of $\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})$ with itself, conditional on the independent variables $\boldsymbol{X}$. Starting with this definition and using the fact that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$, we first replace $\hat{\boldsymbol{\beta}}$ by what it is equal to under the DGP, then take expectations conditional on $\boldsymbol{X}$, and finally simplify the algebra to obtain (3.03):

$$
\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\beta}}) &\equiv E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \\
&= E\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{\beta}_0 \big)\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{\beta}_0 \big)^\top \\
&= E\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}) - \boldsymbol{\beta}_0 \big)\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}) - \boldsymbol{\beta}_0 \big)^\top \\
&= E\big(\boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} - \boldsymbol{\beta}_0 \big)\big(\boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} - \boldsymbol{\beta}_0 \big)^\top \\
&= E\big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1} \boldsymbol{X}^\top \boldsymbol{u} \boldsymbol{u}^\top \boldsymbol{X} \big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1} \\
&= \big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1} \boldsymbol{X}^\top (\sigma_0^2 \boldsymbol{I}) \boldsymbol{X} \big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1} \\
&= \sigma_0^2 \big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1} \\
&= \sigma_0^2 \big(\boldsymbol{X}^\top \boldsymbol{X} \big)^{-1}.
\end{aligned}
$$

Deriving an analogous result for the nonlinear regression model (3.01) requires a few concepts of asymptotic analysis that we have not yet developed, plus a certain amount of mathematical manipulation. We will therefore postpone this derivation until Chapter 5 and merely state an approximate result here.

For a nonlinear model, we cannot in general obtain an exact expression for $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$ in the finite-sample case. In Chapter 5, on the assumption that the data are generated by a DGP which is a special case of (3.01), we will, however, obtain an asymptotic result which allows us to state that

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \cong \sigma_0^2 \big(\boldsymbol{X}^\top(\boldsymbol{\beta}_0) \boldsymbol{X}(\boldsymbol{\beta}_0) \big)^{-1}, \tag{3.04}$$

region for the entire parameter vector $\boldsymbol{\beta}$, implying that $l = k$. For concreteness, we will also assume that the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})$, although it could just as well be $\boldsymbol{V}_s(\hat{\boldsymbol{\beta}})$.

Let us denote the true (but unknown) value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0$. Consider the quadratic form

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \hat{\boldsymbol{V}}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{3.13}$$

This is just a random scalar that depends on the random vector $\hat{\boldsymbol{\beta}}$. For neither a linear nor a nonlinear regression will it actually have the χ^2 distribution with l degrees of freedom in finite samples. But it is reasonable to hope that it will be approximately distributed as $\chi^2(l)$, and in fact such an approximation is valid when the sample is large enough; see Section 5.7. Consequently, with just as much justification (or lack of it) as for the case of a single parameter, the confidence region for $\boldsymbol{\beta}$ is constructed as if (3.13) did indeed have the $\chi^2(l)$ distribution.[4]

For a given set of estimates $\hat{\boldsymbol{\beta}}$, the (approximate) confidence region at level α can be defined as the set of vectors $\boldsymbol{\beta}$ for which the value of (3.13) with $\boldsymbol{\beta}_0$ replaced by $\boldsymbol{\beta}$ is less than some critical value, say $c_\alpha(l)$. This critical value will be such that, if z is a random variable with the $\chi^2(l)$ distribution,

$$\Pr\big(z > c_\alpha(l)\big) = \alpha.$$

The confidence region is therefore the set of all $\boldsymbol{\beta}$ for which

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \hat{\boldsymbol{V}}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq c_\alpha(l). \tag{3.14}$$

Since the left-hand side of this inequality is quadratic in $\boldsymbol{\beta}$, the region is, for $l = 2$, the interior of an ellipse and, for $l > 2$, the interior of an l–dimensional ellipsoid.

Figure 3.2 illustrates what a confidence ellipse can look like in the two-parameter case. In this case, the two parameter estimates are negatively correlated, and the ellipse is centered at the parameter estimates $(\hat{\beta}_1, \hat{\beta}_2)$. Confidence intervals for β_1 and β_2 are also shown, and it should now be clear why it can be misleading to consider only these rather than the confidence ellipse. On the one hand, there are clearly many points, such as (β_1^*, β_2^*), that lie outside the confidence ellipse but inside the two confidence intervals, and on the other hand there are points, like (β_1', β_2'), that are contained in the ellipse but lie outside one or both of the confidence intervals.

---

[4] It is also possible, of course, to construct an approximate confidence region by using the F distribution with l and $n - k$ degrees of freedom, and this might well provide a better approximation in finite samples. Our discussion utilizes the χ^2 distribution primarily because it simplifies the exposition.

obtained by differentiating (3.42) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and setting the derivatives to zero are

$$-\boldsymbol{X}^\top(\tilde{\boldsymbol{\beta}})\big(\boldsymbol{y} - \boldsymbol{x}(\tilde{\boldsymbol{\beta}})\big) + \boldsymbol{R}^\top\tilde{\boldsymbol{\lambda}} = \boldsymbol{0} \tag{3.43}$$

$$\boldsymbol{R}\tilde{\boldsymbol{\beta}} - \boldsymbol{r} = \boldsymbol{0}, \tag{3.44}$$

where $\tilde{\boldsymbol{\beta}}$ denotes the restricted estimates and $\tilde{\boldsymbol{\lambda}}$ denotes the estimated Lagrange multipliers. From (3.43), we see that

$$\boldsymbol{R}^\top\tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{X}}^\top(\boldsymbol{y} - \tilde{\boldsymbol{x}}), \tag{3.45}$$

where, as usual, $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{X}}$ denote $\boldsymbol{x}(\tilde{\boldsymbol{\beta}})$ and $\boldsymbol{X}(\tilde{\boldsymbol{\beta}})$. The expression on the right-hand side of (3.45) is minus the k–vector of the derivatives of $\frac{1}{2}SSR(\boldsymbol{\beta})$ with respect to all the elements of $\boldsymbol{\beta}$, evaluated at $\tilde{\boldsymbol{\beta}}$. This vector is often called the **score vector**. Since $\boldsymbol{y} - \tilde{\boldsymbol{x}}$ is simply a vector of residuals, which should converge asymptotically under H_0 to the vector of error terms $\boldsymbol{u}$, it seems plausible that the asymptotic covariance matrix of the vector of scores is

$$\sigma_0^2\boldsymbol{X}^\top(\boldsymbol{\beta}_0)\boldsymbol{X}(\boldsymbol{\beta}_0). \tag{3.46}$$

Subject to certain asymptotic niceties, that is indeed the case, and a more rigorous version of this result will be proved in Chapter 5.

The obvious way to estimate (3.46) is to use $\tilde{s}^2\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$, where $\tilde{s}^2$ is $SSR(\tilde{\boldsymbol{\beta}})/(n - k + r)$. Putting this estimate together with the expressions on each side of (3.45), we can construct two apparently different, but numerically identical, test statistics. The first of these is

$$\tilde{\boldsymbol{\lambda}}^\top\boldsymbol{R}\big(\tilde{s}^2\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\big)^{-1}\boldsymbol{R}^\top\tilde{\boldsymbol{\lambda}} = \frac{1}{\tilde{s}^2}\tilde{\boldsymbol{\lambda}}^\top\boldsymbol{R}\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\big)^{-1}\boldsymbol{R}^\top\tilde{\boldsymbol{\lambda}}. \tag{3.47}$$

In this form, the test statistic is clearly a Lagrange multiplier statistic. Since $\tilde{\boldsymbol{\lambda}}$ is an r–vector, it should not be surprising that this statistic would be asymptotically distributed as $\chi^2(r)$. A proof that this is the case follows from essentially the same arguments used in the case of the Wald test, since (3.47) is a quadratic form similar to (3.37). Of course, the result depends critically on the vector $\tilde{\boldsymbol{\lambda}}$ being asymptotically normally distributed, something that we will prove in Chapter 5.

The second test statistic, which we stress is numerically identical to the first, is obtained by substituting $\tilde{\boldsymbol{X}}^\top(\boldsymbol{y} - \tilde{\boldsymbol{x}})$ for $\boldsymbol{R}^\top\tilde{\boldsymbol{\lambda}}$ in (3.47). The result, which is the **score form** of the LM statistic, is

$$\frac{1}{\tilde{s}^2}(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top\tilde{\boldsymbol{X}}\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\big)^{-1}\tilde{\boldsymbol{X}}^\top(\boldsymbol{y} - \tilde{\boldsymbol{x}}) = \frac{1}{\tilde{s}^2}(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top\tilde{\boldsymbol{P}}_X(\boldsymbol{y} - \tilde{\boldsymbol{x}}), \tag{3.48}$$

where $\tilde{\boldsymbol{P}}_X \equiv \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\top$. It is evident that this expression is simply the explained sum of squares from the **artificial linear regression**

$$\frac{1}{\tilde{s}}(\boldsymbol{y} - \tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{X}}\boldsymbol{b} + \text{residuals}, \tag{3.49}$$

with $\boldsymbol{\beta}_{20} \neq 0$. Then it is easy to see that the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ will, in general, be biased. Under this DGP,

$$
\begin{aligned}
E(\tilde{\boldsymbol{\beta}}_1) &= E\Big(\big(\boldsymbol{X}_1^\top \boldsymbol{X}_1\big)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y}\Big) \\
&= E\Big(\big(\boldsymbol{X}_1^\top \boldsymbol{X}_1\big)^{-1} \boldsymbol{X}_1^\top (\boldsymbol{X}_1 \boldsymbol{\beta}_{10} + \boldsymbol{X}_2 \boldsymbol{\beta}_{20} + \boldsymbol{u})\Big) \qquad (3.57) \\
&= \boldsymbol{\beta}_{10} + \big(\boldsymbol{X}_1^\top \boldsymbol{X}_1\big)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_{20}.
\end{aligned}
$$

Unless $\boldsymbol{X}_1^\top \boldsymbol{X}_2$ is a zero matrix or $\boldsymbol{\beta}_{20}$ is a zero vector, $\tilde{\boldsymbol{\beta}}_1$ will be a biased estimator. The magnitude of the bias will depend on the matrices $\boldsymbol{X}_1^\top \boldsymbol{X}_1$ and $\boldsymbol{X}_1^\top \boldsymbol{X}_2$ and the vector $\boldsymbol{\beta}_{20}$.

Results very similar to (3.57) are available for all types of restrictions, not just for linear restrictions, and for all sorts of models in addition to linear regression models. We will not attempt to deal with nonlinear models here because that requires a good deal of technical apparatus, which will be developed in Chapter 12. Results analogous to (3.57) for nonlinear regression models and other types of nonlinear models may be found in Kiefer and Skoog (1984). The important point is that imposition of false restrictions on some of the parameters of a model generally causes all of the parameter estimates to be biased. This bias does not go away as the sample size gets larger.

Even though $\tilde{\boldsymbol{\beta}}_1$ is biased when the DGP is (3.56), it is still of interest to ask how well it performs. The analog of the covariance matrix for a biased estimator is the **mean squared error matrix**, which in this case is

$$
\begin{aligned}
& E\big(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\big)\big(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\big)^\top \\
&= E\big((\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top (\boldsymbol{X}_2 \boldsymbol{\beta}_{20} + \boldsymbol{u})\big)\big((\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top (\boldsymbol{X}_2 \boldsymbol{\beta}_{20} + \boldsymbol{u})\big)^\top \\
&= \sigma_0^2 \big(\boldsymbol{X}_1^\top \boldsymbol{X}_1\big)^{-1} + \big(\boldsymbol{X}_1^\top \boldsymbol{X}_1\big)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_{20} \boldsymbol{\beta}_{20}^\top \boldsymbol{X}_2^\top \boldsymbol{X}_1 \big(\boldsymbol{X}_1^\top \boldsymbol{X}_1\big)^{-1}. \quad (3.58)
\end{aligned}
$$

The third line here is the sum of two matrices: the covariance matrix of $\tilde{\boldsymbol{\beta}}_1$ when the DGP satisfies the restrictions, and the outer product of the second term in the last line of (3.57) with itself. It is possible to compare (3.58) with $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_1)$, the covariance matrix of the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$, only if σ_0 and $\boldsymbol{\beta}_{20}$ are known. Since the first term of (3.58) is smaller in the matrix sense than $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_1)$, it is clear that if $\boldsymbol{\beta}_{20}$ is small enough (3.58) will be smaller than $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_1)$. Thus it may be desirable to use the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ when the restrictions are false, provided they are not too false.

Applied workers frequently find themselves in a situation like the one we have been discussing. They want to estimate $\boldsymbol{\beta}_1$ and do not know whether or not $\boldsymbol{\beta}_2 = 0$. It then seems natural to define a new estimator,

$$
\check{\boldsymbol{\beta}}_1 = \begin{cases} \tilde{\boldsymbol{\beta}}_1 & \text{if } F_{\boldsymbol{\beta}_2 = \boldsymbol{0}} < c_\alpha; \\ \hat{\boldsymbol{\beta}}_1 & \text{if } F_{\boldsymbol{\beta}_2 = \boldsymbol{0}} \geq c_\alpha. \end{cases}
$$

Here $F_{\boldsymbol{\beta}_2 = \mathbf{0}}$ is the usual F test statistic for the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$, and c_α is the critical value for a test of size α given by the $F(r, n-k)$ distribution. Thus $\check{\boldsymbol{\beta}}_1$ will be the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ when the F test does not reject the hypothesis that the restrictions are satisfied and will be the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$ when the F test does reject that hypothesis. It is an example of what is called a **preliminary test estimator** or **pretest estimator**.

Pretest estimators are used all the time. Whenever we test some aspect of a model's specification and then decide, on the basis of the test results, what version of the model to estimate or what estimation method to use, we are employing a pretest estimator. Unfortunately, the properties of pretest estimators are, in practice, very difficult to know. The problems can be seen from the example we have been studying. Suppose the restrictions hold. Then the estimator we would like to use is the restricted estimator, $\tilde{\boldsymbol{\beta}}_1$. But, $\alpha\%$ of the time, the F test will incorrectly reject the null hypothesis and $\check{\boldsymbol{\beta}}_1$ will be equal to the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$ instead. Thus $\check{\boldsymbol{\beta}}_1$ must be less efficient than $\tilde{\boldsymbol{\beta}}_1$ when the restrictions do in fact hold. Moreover, since the estimated covariance matrix reported by the regression package will not take the pretesting into account, inferences about $\check{\boldsymbol{\beta}}_1$ may be misleading.

On the other hand, when the restrictions do not hold, we may or may not want to use the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$. Depending on how much power the F test has, $\check{\boldsymbol{\beta}}_1$ will sometimes be equal to $\tilde{\boldsymbol{\beta}}_1$ and sometimes be equal to $\hat{\boldsymbol{\beta}}_1$. It will certainly not be unbiased, because $\tilde{\boldsymbol{\beta}}_1$ is not unbiased, and it may be more or less efficient (in the sense of mean squared error) than the unrestricted estimator. Inferences about $\check{\boldsymbol{\beta}}_1$ based on the usual estimated OLS covariance matrix for whichever of $\tilde{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_1$ it turns out to be equal to may be misleading, because they fail to take into account the pretesting that occurred previously.

In practice, there is often not very much that we can do about the problems caused by pretesting, except to recognize that pretesting adds an additional element of uncertainty to most problems of statistical inference. Since α, the level of the preliminary test, will affect the properties of $\check{\boldsymbol{\beta}}_1$, it may be worthwhile to try using different values of α. Conventional significance levels such as .05 are certainly not optimal in general, and there is a literature on how to choose better ones in specific cases; see, for example, Toyoda and Wallace (1976). However, real pretesting problems are much more complicated than the one we have discussed as an example or the ones that have been studied in the literature. Every time one subjects a model to any sort of test, the result of that test may affect the form of the final model, and the implied pretest estimator therefore becomes even more complicated. It is hard to see how this can be analyzed formally.

Our discussion of pretesting has been very brief. More detailed treatments may be found in Fomby, Hill, and Johnson (1984, Chapter 7), Judge, Hill, Griffiths, Lütkepohl, and Lee (1985, Chapter 21), and Judge and Bock (1978). In the remainder of this book, we entirely ignore the problems caused

condition. Unlike asymptotic equality, the big-O relation does not require that the ratio $f(n)/g(n)$ should have any limit. It may have, but it may also oscillate boundedly for ever.

The relations we have defined so far are for nonstochastic real-valued sequences. Of greater interest to econometricians are the so-called **stochastic order relations**. These are perfectly analogous to the relations we have defined but instead use one or other of the forms of stochastic convergence. Formally:

Definition 4.8.

If $\{a_n\}$ is a sequence of random variables, and $g(n)$ is a real-valued function of the positive integer argument n, then the notation $a_n = o_p\big(g(n)\big)$ means that

$$\plim_{n\to\infty}\left(\frac{a_n}{g(n)}\right) = 0.$$

Similarly, the notation $a_n = O_p\big(g(n)\big)$ means that, for all $\varepsilon > 0$, there exist a constant K and a positive integer N such that

$$\Pr\left(\left|\frac{a_n}{g(n)}\right| > K\right) < \varepsilon \quad \text{for all } n > N.$$

If $\{b_n\}$ is another sequence of random variables, the notation $a_n \overset{a}{=} b_n$ means that

$$\plim_{n\to\infty}\left(\frac{a_n}{b_n}\right) = 1.$$

Comparable definitions may be written down for almost sure convergence and convergence in distribution, but we will not use these. In fact, after this section we will not bother to use the subscript p in the stochastic order symbols, because it will always be plain when random variables are involved. When they are, $O(\cdot)$ and $o(\cdot)$ should be read as $O_p(\cdot)$ and $o_p(\cdot)$.

The order symbols are very easy to manipulate, and we now present a few useful rules for doing so. For simplicity, we restrict ourselves to functions $g(n)$ that are just powers of n, for that is all we use in this book. The rules for addition and subtraction are

$$O(n^p) \pm O(n^q) = O\big(n^{\max(p,q)}\big);$$
$$o(n^p) \pm o(n^q) = o\big(n^{\max(p,q)}\big);$$
$$O(n^p) \pm o(n^q) = O(n^p) \quad \text{if } p \geq q;$$
$$O(n^p) \pm o(n^q) = o(n^q) \quad \text{if } p < q.$$

The rules for multiplication, and by implication for division, are

$$O(n^p)O(n^q) = O(n^{p+q});$$
$$o(n^p)o(n^q) = o(n^{p+q});$$
$$O(n^p)o(n^q) = o(n^{p+q}).$$

A comparison of (4.17) and (4.18) reveals that the behavior of the estimator $\hat{\alpha}$ is quite different under the two different rules for sample-size extension.

There is not always a simple resolution to the sort of problem posed in the above example. It is *usually* unrealistic to assume that linear time trends of the form of $\boldsymbol{\tau}$ will continue to increase forever, but it suffices to look at price series in the twentieth century (and many other centuries) to realize that some economic variables do not seem to have natural upper bounds. Even quantity series such as real GNP or personal consumption are sometimes fruitfully considered as being unbounded. Nevertheless, although the asymptotic theories resulting from different kinds of rules for extending DGPs to arbitrarily large samples can be very different, it is important to be clear that deciding among competing asymptotic theories of this sort is *not* an empirical issue. For any given empirical investigation, the sample size is what it is, even if the *possibility* of collecting further relevant data exists. The issue is always one of selecting a suitable *model*, not only for the data that exist, but for a set of economic *phenomena*, of which the data are supposed to be a manifestation. There is always an infinity of models (not all plausible of course) that are compatible with any finite data set. As a consequence, the issue of model selection among a set of such models can be decided only on the basis of such criteria as the explanatory power of the concepts used in the model, simplicity of expression, or ease of interpretation, but not on the basis of the information contained in the data themselves.

Although, in the model (4.14), the assumption that the time trend variable goes to infinity with the sample size may seem more plausible than the fixed-in-repeated-samples assumption, we will throughout most of this book assume that the DGP is of the latter rather than the former type. The problem with allowing τ_t to go to infinity with the sample size is that each additional observation gives us more information about the value of α than any of the preceding observations. That is why $\mathrm{Var}(\hat{\alpha})$ turned out to be $O(n^{-3})$ when we made that assumption about the DGP. It seems much more plausible in most cases that each additional observation should, on average, give us the same amount of information as the preceding observations. This implies that the variance of parameter estimates will be $O(n^{-1})$, as was $\mathrm{Var}(\hat{\alpha})$ when we assumed that the DGP was of the fixed-in-repeated-samples type. Our general assumptions about DGPs will likewise lead to the conclusion that the variance of parameter estimates is $O(n^{-1})$, although we will consider DGPs that do not lead to this conclusion in Chapter 20, which deals with dynamic models.

4.5 CONSISTENCY AND LAWS OF LARGE NUMBERS

We begin this section by introducing the notion of **consistency**, one of the most basic ideas of asymptotic theory. When one is interested in estimating parameters from data, it is desirable that the parameter estimates should have certain properties. In Chapters 2 and 3, we saw that, under certain regularity

interested in the nondegenerate asymptotic distribution of the sample mean as an estimator. We saw in Section 4.3 that for this purpose we should look at the distribution of $n^{1/2}(m_1 - \mu)$, where m_1 is the sample mean. Specifically, we wish to study

$$n^{1/2}(m_1 - \mu) = n^{-1/2} \sum_{t=1}^{n} (y_t - \mu),$$

where $y_t - \mu$ has variance σ_t^2.

We begin by stating the following simple **central limit theorem**.

Theorem 4.2. Simple Central Limit Theorem. (Lyapunov)

Let $\{y_t\}$ be a sequence of independent, centered random variables with variances σ_t^2 such that $\underline{\sigma}^2 \leq \sigma_t^2 \leq \overline{\sigma}^2$ for two finite positive constants, $\underline{\sigma}^2$ and $\overline{\sigma}^2$, and absolute third moments μ_3 such that $\mu_3 \leq \bar{\mu}_3$ for a finite constant $\bar{\mu}_3$. Further, let

$$\sigma_0^2 \equiv \lim_{n \to \infty} \left(\frac{1}{n} \sum_{t=1}^{n} \sigma_t^2 \right)$$

exist. Then the sequence

$$\left\{ n^{-1/2} \sum_{t=1}^{n} y_t \right\}$$

tends in distribution to a limit characterized by the normal distribution with mean zero and variance σ_0^2.

Theorem 4.2 applies directly to the example (4.26). Thus our hypothetical investigator may, within the limits of asymptotic theory, use the $N(0, \sigma_0^2)$ distribution for statistical inference on the estimate m_1 via the random variable $n^{1/2}(m_1 - \mu)$. Knowledge of σ_0^2 is not necessary, provided that it can be estimated consistently.

Although we do not intend to offer a formal proof of even this simple central limit theorem, in view of the technicalities that such a proof would entail, it is not difficult to give a general idea of why the result is true. For simplicity, let us consider the case in which all the variables y_t of the sequence $\{y_t\}$ have the same distribution with variance σ^2. Then clearly the variable

$$S_n \equiv n^{-1/2} \sum_{t=1}^{n} y_t$$

has mean zero and variance σ^2 for each n. But what of the higher moments of S_n? By way of an example, consider the fourth moment. It is

$$E(S_n^4) = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} \sum_{t=1}^{n} \sum_{u=1}^{n} E(y_r y_s y_t y_u). \tag{4.27}$$

Another important consequence of the definition of a conditional expectation is the so-called **law of iterated expectations**, which can be stated as follows:

$$E\big(E(y \mid \boldsymbol{z})\big) = E(y).$$

The proof of this is an immediate consequence of using the whole of $\mathbb{R}^k$ as the set G in (4.29).

The definitions which follow are rather technical, as are the statements of the laws of large numbers that make use of them. Some readers may therefore wish to skip over them and the discussion of central limit theorems to the definitions of the two sets of regularity conditions, which we call WULLN and CLT, presented at the end of this section. Such readers may return to this point when some reference to it is made later in the book.

Definition 4.10.

The sequence $\{y_t\}$ is said to be **stationary** if for all finite k the joint distribution of the linked set $\{y_t, y_{t+1}, \ldots, y_{t+k}\}$ is independent of the index t.

Definition 4.11.

The stationary sequence $\{y_t\}$ is said to be **ergodic** if, for any two bounded mappings $Y \colon \mathbb{R}^{k+1} \to \mathbb{R}$ and $Z \colon \mathbb{R}^{l+1} \to \mathbb{R}$,

$$\lim_{n\to\infty} \big| E\big(Y(y_i, \ldots, y_{i+k}) Z(y_{i+n}, \ldots, y_{i+n+l})\big)\big|$$
$$= \big| E\big(Y(y_i, \ldots, y_{i+k})\big)\big| \, \big| E\big(Z(y_i, \ldots, y_{i+l})\big)\big|.$$

Definition 4.12.

The sequence $\{y_t\}$ is said to be **uniformly mixing**, or **ϕ–mixing**, if there is a sequence of positive numbers $\{\phi_n\}$, convergent to zero, such that, for any two bounded mappings $Y \colon \mathbb{R}^{k+1} \to \mathbb{R}$ and $Z \colon \mathbb{R}^{l+1} \to \mathbb{R}$,

$$\big| E\big(Y(y_t, \ldots, y_{t+k}) \mid Z(y_{t+n}, \ldots, y_{t+n+l})\big) - E\big(Y(y_t, \ldots, y_{t+k})\big)\big| < \phi_n.$$

The symbol $E(\cdot \mid \cdot)$ denotes a conditional expectation, as defined above.

Definition 4.13.

The sequence $\{y_t\}$ is said to be **α–mixing** if there is a sequence of positive numbers $\{\alpha_n\}$, convergent to zero, such that, if Y and Z are as in the preceding definition, then

$$\big| E\big(Y(y_t, \ldots, y_{t+k}) Z(y_{t+n}, \ldots, y_{t+n+l})\big) - E\big(Y(\cdot)\big) E\big(Z(\cdot)\big)\big| < \alpha_n.$$

The last three definitions can be thought of as defining various forms of **asymptotic independence**. According to them, random variables y_t and y_s are more nearly independent (in some sense) the farther apart are the indices t

Theorem 4.7. (Lindeberg-Lévy)

If the variables of the random sequence $\{y_t\}$ are independent and have the same distribution with mean μ and variance v, then S_n converges in distribution to the standard normal distribution $N(0,1)$.

This theorem has minimal requirements for the moments of the variables but maximal requirements for their homogeneity. Note that, in this case,

$$S_n = (nv)^{-1/2} \sum_{t=1}^{n} (y_t - \mu).$$

The next theorem allows for much heterogeneity but still requires independence.

Theorem 4.8. (Lyapunov)

For each positive integer n let the finite sequence $\{y_t^n\}_{t=1}^n$ consist of independent centered random variables possessing variances v_t^n. Let $s_n^2 \equiv \sum_{t=1}^n v_t^n$ and let the **Lindeberg condition** be satisfied, namely, that for all $\varepsilon > 0$

$$\lim_{n \to \infty} \left(\sum_{t=1}^{n} s_n^{-2} E\big((y_t^n)^2 I_G(y_t^n)\big) \right) = 0,$$

where the set G used in the indicator function is $\{y : |y| \geq \varepsilon s_n\}$. Then $s_n^{-1} \sum_{t=1}^n y_t^n$ converges in distribution to $N(0,1)$.

Our last central limit theorem allows for dependent sequences.

Theorem 4.9. (McLeish)

For each positive integer n let the finite sequences $\{y_t^n\}_{t=1}^n$ be martingale difference sequences with $v_t^n \equiv \mathrm{Var}\,(y_t^n) < \infty$, and $s_n^2 \equiv \sum_{t=1}^n v_t^n$. If for all $\varepsilon > 0$

$$\lim_{n \to \infty} \left(s_n^{-2} \sum_{t=1}^{n} E\big((y_t^n)^2 I_G(y_t^n)\big) \right) = 0,$$

where again the set $G \equiv \{y : |y| \geq \varepsilon s_n\}$, and if the sequence

$$\left\{ \sum_{t=1}^{n} \frac{(y_t^n)^2}{s_n^2} \right\}$$

obeys a law of large numbers and thus converges to 1, then $s_n^{-1} \sum_{t=1}^n y_t^n$ converges in distribution to $N(0,1)$.

See McLeish (1974). Observe the extra condition needed in this theorem, which ensures that the variance of the limiting distribution is the same as the limit of the variances of the variables in $s_n^{-1} \sum_{t=1}^n y_t^n$.

since the distribution of the u_t's has not been specified. Thus, for a sample of size n, the model $\mathbb{M}$ described by (5.08) is the set of all DGPs generating samples $\boldsymbol{y}$ of size n such that the expectation of y_t conditional on some information set Ω_t that includes $\boldsymbol{Z}_t$ is $x_t(\boldsymbol{\beta})$ for some parameter vector $\boldsymbol{\beta} \in \mathbb{R}^k$, and such that the differences $y_t - x_t(\boldsymbol{\beta})$ are independently distributed error terms with common variance σ^2, usually unknown.

It will be convenient to generalize this specification of the DGPs in $\mathbb{M}$ a little, in order to be able to treat **dynamic models**, that is, models in which there are **lagged dependent variables**. Therefore, we explicitly recognize the possibility that the regression function $x_t(\boldsymbol{\beta})$ may include among its (until now implicit) dependences an arbitrary but bounded number of lags of the dependent variable itself. Thus x_t may depend on $y_{t-1}, y_{t-2}, \ldots, y_{t-l}$, where l is a fixed positive integer that does not depend on the sample size. When the model uses time-series data, we will therefore take $x_t(\boldsymbol{\beta})$ to mean the expectation of y_t conditional on an information set that includes the entire past of the dependent variable, which we can denote by $\{y_s\}_{s=1}^{t-1}$, and also the entire history of the exogenous variables up to and including the period t, that is, $\{\boldsymbol{Z}_s\}_{s=1}^{t}$. The requirements on the disturbance vector $\boldsymbol{u}$ are unchanged.

For asymptotic theory to be applicable, we must next provide a rule for extending (5.08) to samples of arbitrarily large size. For models which are not dynamic (including models estimated with cross-section data, of course), so that there are no time trends or lagged dependent variables in the regression functions x_t, there is nothing to prevent the simple use of the fixed-in-repeated-samples notion that we discussed in Section 4.4. Specifically, we consider only sample sizes that are integer multiples of the actual sample size m and then assume that $x_{Nm+t}(\boldsymbol{\beta}) = x_t(\boldsymbol{\beta})$ for $N > 1$. This assumption makes the asymptotics of nondynamic models very simple compared with those for dynamic models.[3]

Some econometricians would argue that the above solution is too simple-minded when one is working with time-series data and would prefer a rule like the following. The variables $\boldsymbol{Z}_t$ appearing in the regression functions will usually themselves display regularities as time series and may be susceptible to modeling as one of the standard stochastic processes used in time-series analysis; we will discuss these standard processes at somewhat greater length in Chapter 10. In order to extend the DGP (5.08), the out-of-sample values for the $\boldsymbol{Z}_t$'s should themselves be regarded as random, being generated by appropriate processes. The introduction of this additional randomness complicates the asymptotic analysis a little, but not really a lot, since one would always assume that the stochastic processes generating the $\boldsymbol{Z}_t$'s were independent of the stochastic process generating the disturbance vector $\boldsymbol{u}$.

---

[3] Indeed, even for *linear* dynamic models it is by no means trivial to show that least squares yields consistent, asymptotically normal estimates. The classic reference on this subject is Mann and Wald (1943).

The result (5.44) essentially proves the Gauss-Markov Theorem, since it implies that

$$
\begin{aligned}
& E(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \\
& = E\Big(\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u} + \boldsymbol{C}\boldsymbol{u}\big)\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u} + \boldsymbol{C}\boldsymbol{u}\big)^\top\Big) \qquad (5.45) \\
& = \sigma_0^2\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1} + \sigma_0^2\boldsymbol{C}\boldsymbol{C}^\top.
\end{aligned}
$$

Thus the difference between the covariance matrices of $\check{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ is $\sigma_0^2\boldsymbol{C}\boldsymbol{C}^\top$, which is a positive semidefinite matrix. Notice that the assumption that $E(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma_0^2\mathbf{I}$ is crucial here. If instead we had $E(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}$, with $\boldsymbol{\Omega}$ an arbitrary $n \times n$ positive definite matrix, the last line of (5.45) would be

$$
\begin{aligned}
& \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1} \\
& \quad + \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}^\top + \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{C}^\top + \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1},
\end{aligned}
$$

and we could draw no conclusion about the relative efficiency of $\hat{\boldsymbol{\beta}}$ and $\check{\boldsymbol{\beta}}$.

As a simple example of the Gauss-Markov Theorem in action, suppose that $\check{\boldsymbol{\beta}}$ is the OLS estimator obtained by regressing $\boldsymbol{y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$ jointly, where $\boldsymbol{Z}$ is a matrix of regressors such that $E(\boldsymbol{y}\,|\,\boldsymbol{X}, \boldsymbol{Z}) = E(\boldsymbol{y}\,|\,\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$. Since the information that $\boldsymbol{Z}$ does not belong in the regression is being ignored when we construct $\check{\boldsymbol{\beta}}$, the latter must in general be inefficient. Using the FWL Theorem, we find that

$$
\check{\boldsymbol{\beta}} = \big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{y}, \qquad (5.46)
$$

where, as usual, $\boldsymbol{M}_Z$ is the matrix that projects orthogonally onto $\mathcal{S}^\perp(\boldsymbol{Z})$. If we write $\check{\boldsymbol{\beta}}$ as in (5.42), we obtain

$$
\begin{aligned}
\check{\boldsymbol{\beta}} & = \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{y} + \big((\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\boldsymbol{y} \\
& = \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{y} + \big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}\big(\boldsymbol{X}^\top\boldsymbol{M}_Z - \boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\boldsymbol{y} \\
& = \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{y} + \big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}\Big(\boldsymbol{X}^\top\boldsymbol{M}_Z\big(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\Big)\boldsymbol{y} \\
& = \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{y} + \big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{M}_X\boldsymbol{y} \\
& = \hat{\boldsymbol{\beta}} + \boldsymbol{C}\boldsymbol{y}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.47)
\end{aligned}
$$

Thus, in this case, the matrix $\boldsymbol{C}$ is the matrix $(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{M}_X$. We see that the inefficient estimator $\check{\boldsymbol{\beta}}$ is equal to the efficient estimator $\hat{\boldsymbol{\beta}}$ plus a random component which is uncorrelated with it. That $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{C}\boldsymbol{y}$ are uncorrelated follows from the fact (required for $\boldsymbol{C}\boldsymbol{y}$ to have mean zero) that $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}$, which is true because $\boldsymbol{M}_X$ annihilates $\boldsymbol{X}$. Further, we see that

$$
\begin{aligned}
E(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top & = \sigma_0^2\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1} \\
& + \sigma_0^2\big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{M}_X\boldsymbol{M}_Z\boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}.
\end{aligned} \qquad (5.48)
$$

the residual $\hat{u}_t$. But this expansion is still unnecessarily complicated, because we have

$$\boldsymbol{X}_t^* = \boldsymbol{X}_{0t} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{A}_t^* = \boldsymbol{X}_{0t} + O(n^{-1/2})$$

by Taylor's Theorem and the fact that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O(n^{-1/2})$; recall that $\boldsymbol{A}_t$ is the Hessian of the regression function $x_t(\boldsymbol{\beta})$. Thus (5.56) can be written more simply as

$$\hat{u}_t = u_t - n^{-1/2} \boldsymbol{X}_{0t} \big(n^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0\big)^{-1} n^{-1/2} \boldsymbol{X}_0^\top \boldsymbol{u} + o(n^{-1/2}).$$

Since this is true for all t, we have the vector equation

$$\hat{\boldsymbol{u}} = \boldsymbol{u} - \boldsymbol{X}_0 \big(\boldsymbol{X}_0^\top \boldsymbol{X}_0\big)^{-1} \boldsymbol{X}_0^\top \boldsymbol{u} + o(n^{-1/2}),$$

where the small-order symbol is now to be interpreted as an n–vector, each component of which is $o(n^{-1/2})$. This equation can be rewritten in terms of the projection $\boldsymbol{P}_0 \equiv \boldsymbol{X}_0(\boldsymbol{X}_0^\top \boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^\top$ and its complementary projection $\boldsymbol{M}_0 \equiv \boldsymbol{I} - \boldsymbol{P}_0$:

$$\hat{\boldsymbol{u}} = \boldsymbol{u} - \boldsymbol{P}_0 \boldsymbol{u} + o(n^{-1/2}) = \boldsymbol{M}_0 \boldsymbol{u} + o(n^{-1/2}). \tag{5.57}$$

This is the asymptotic equivalent of the exact result that, for linear models, the OLS residuals are the orthogonal projection of the disturbances off the regressors. Recall that if one runs the regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, and the DGP is indeed a special case of this model, then we have exactly that

$$\hat{\boldsymbol{u}} = \boldsymbol{M}_X \boldsymbol{u}. \tag{5.58}$$

The result (5.57) reduces to this when the model is linear. The projection matrix $\boldsymbol{M}_0$ is now equal to $\boldsymbol{M}_X$, and the $o(n^{-1/2})$ term, which was due only to the nonlinearity of $\boldsymbol{x}(\boldsymbol{\beta})$, no longer appears.

Now let us substitute the right-most expression of (5.57) into (5.53). The latter becomes

$$n^{-1/2} \boldsymbol{a}^\top \hat{\boldsymbol{u}} = n^{-1/2} \boldsymbol{a}^\top \boldsymbol{M}_0 \boldsymbol{u} + n^{-1/2} \sum_{t=1}^{n} o(n^{-1/2}). \tag{5.59}$$

The first term on the right-hand side here is clearly $O(1)$, while the second is $o(1)$. Thus, in contrast to what happened when we simply replaced $\hat{u}_t$ by u_t, we can ignore the second term on the right-hand side of (5.59). So the result (5.57) provides what we need if we are to undertake asymptotic analysis of expressions like (5.53).

We should pause for a moment here in order to make clear the relation between the asymptotic result (5.57), the exact linear result (5.58), and two other results. These other results are (1.03), which states that the OLS residuals are orthogonal to the regressors, and (2.05), which we may express

term. The sort of result displayed in (5.68) occurs very frequently. The *twice* continuous differentiability of $r(\beta)$ means that Taylor's Theorem can be applied to order two, and then it is possible to discover from the last term in that expansion exactly the order of the error, in this case $O(n^{-1})$, committed by neglecting it. In future we will not be explicit about this reasoning and will simply mention that twice continuous differentiability gives a result similar to (5.68).

The quantities in (5.66) other than $\hat{r}$ are **asymptotically nonstochastic**. By this we mean that

$$\hat{R} = R_0 + O(n^{-1/2}) \quad \text{and} \quad \hat{X} = X_0 + O(n^{-1/2}). \tag{5.69}$$

Again, a short Taylor-series argument, this time only to first order, produces these results. They are to be interpreted component by component for the matrices R and X. This is not a matter of consequence for the $r \times k$ matrix R, but it is for the $n \times k$ matrix X. We have to be careful because in matrix products like $\hat{X}^\top \hat{X}$ we run across sums of n terms, which will of course have different orders in general from the terms of the sums. However, if we explicitly use the fact that $\hat{r} = O(n^{-1/2})$ to rewrite (5.66) as

$$\left(n^{1/2}\hat{r}\right)^\top \left(\hat{\sigma}^2 \hat{R}(n^{-1}\hat{X}^\top \hat{X})^{-1}\hat{R}^\top\right)^{-1}\left(n^{1/2}\hat{r}\right), \tag{5.70}$$

we see that we are concerned, not with $\hat{X}^\top \hat{X}$ itself, but rather with $n^{-1}\hat{X}^\top \hat{X}$, and the latter *is* asymptotically nonstochastic:

$$
\begin{aligned}
n^{-1}(\hat{X}^\top \hat{X})_{ij} &= n^{-1}\sum_{t=1}^{n} \hat{X}_{ti}\hat{X}_{tj} \\
&= n^{-1}\sum_{t=1}^{n}\left(X_{ti}^0 + O(n^{-1/2})\right)\left(X_{tj}^0 + O(n^{-1/2})\right) \\
&= n^{-1}\sum_{t=1}^{n} X_{ti}^0 X_{tj}^0 + O(n^{-1/2}) \\
&= n^{-1}(X_0^\top X_0)_{ij} + O(n^{-1/2}),
\end{aligned}
$$

where X_{ti}^0 denotes the ti^{th} element of X_0. The second line uses (5.69). The third line follows because the sum of n terms of order $n^{-1/2}$ can be at most of order $n^{1/2}$; when divided by n, it becomes of order $n^{-1/2}$. Note that $n^{-1}X_0^\top X_0$ itself is $O(1)$.

Next, we use the asymptotic normality result (5.39) to obtain a more convenient expression for $n^{1/2}\hat{r}$. We have

$$n^{1/2}\hat{r} = R_0\left(n^{-1}X_0^\top X_0\right)^{-1}n^{-1/2}X_0^\top u + o(1). \tag{5.71}$$

since $\boldsymbol{P}_1$ plays the same role for the manifold $\mathcal{R}$ as does $\boldsymbol{P}_0$ for $\mathcal{X}$. The LM statistic (3.48) is

$$\frac{1}{\tilde{\sigma}^2}(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{P}}_X(\boldsymbol{y} - \tilde{\boldsymbol{x}}). \tag{5.76}$$

If we express the statistic in terms of quantities that are $O(1)$, we obtain

$$\frac{1}{\tilde{\sigma}^2} n^{-1/2}(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{X}} \left(n^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\right)^{-1} n^{-1/2} \tilde{\boldsymbol{X}}^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}}). \tag{5.77}$$

Like $\hat{\boldsymbol{X}}_t$, $\tilde{\boldsymbol{X}}_t$ is asymptotically nonstochastic. Therefore, from (5.75),

$$\begin{aligned}
n^{-1/2} \tilde{\boldsymbol{X}}^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}}) &= n^{-1/2} \sum_{t=1}^{n} \tilde{\boldsymbol{X}}_t^\top \tilde{u}_t \\
&= n^{-1/2} \sum_{t=1}^{n} \boldsymbol{X}_{0t}^\top (\boldsymbol{M}_1 \boldsymbol{u})_t + o(1) \\
&= n^{-1/2} \sum_{t=1}^{n} (\boldsymbol{M}_1 \boldsymbol{X}_0)_t\, u_t + o(1) \\
&= n^{-1/2} \boldsymbol{X}_0^\top \boldsymbol{M}_1 \boldsymbol{u} + o(1).
\end{aligned}$$

The matrix $n^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ is asymptotically nonstochastic, just as $n^{-1} \hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}}$ is, and so the LM statistic (5.77) is asymptotically equivalent to

$$\boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{X}_0 \left(\sigma_0^2 \boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{X}_0^\top \boldsymbol{M}_1 \boldsymbol{u} = \sigma_0^{-2} \boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{P}_0 \boldsymbol{M}_1 \boldsymbol{u}. \tag{5.78}$$

Since $\mathcal{S}(\boldsymbol{X}_1)$ is a subspace of $\mathcal{S}(\boldsymbol{X}_0)$, we have $\boldsymbol{P}_1 \boldsymbol{P}_0 = \boldsymbol{P}_0 \boldsymbol{P}_1 = \boldsymbol{P}_1$, from which it follows that $\boldsymbol{M}_1 \boldsymbol{P}_0 \boldsymbol{M}_1 = \boldsymbol{P}_0 - \boldsymbol{P}_1$. Expression (5.78) thus becomes

$$\sigma_0^{-2} \boldsymbol{u}^\top (\boldsymbol{P}_0 - \boldsymbol{P}_1) \boldsymbol{u} = \sigma_0^{-2} \boldsymbol{u}^\top \boldsymbol{P}_2 \boldsymbol{u}. \tag{5.79}$$

Comparison of (5.79) with (5.72) shows that the LM statistic is asymptotically equal to the Wald statistic. Thus it too is asymptotically $\chi^2(r)$ under the null hypothesis.

The third of the three test statistics discussed in Section 3.6 was the one based on the likelihood ratio principle, the pseudo-F statistic (3.50). Since we are interested in asymptotic results only, we rewrite it here in a form in which it should be asymptotically distributed as $\chi^2(r)$:

$$\frac{1}{s^2}\left(SSR(\tilde{\boldsymbol{\beta}}) - SSR(\hat{\boldsymbol{\beta}})\right) \tag{5.80}$$

and will (somewhat loosely) refer to it as the LR statistic. We have already seen that $s^2 \to \sigma_0^2$ as $n \to \infty$. It remains to show that $SSR(\tilde{\boldsymbol{\beta}}) - SSR(\hat{\boldsymbol{\beta}})$, when divided by σ_0^2, is asymptotically $\chi^2(r)$. From (5.64), we have

$$\hat{\sigma}^2 = \frac{1}{n} \boldsymbol{u}^\top \boldsymbol{M}_0 \boldsymbol{u} + o(n^{-1}),$$

difference is that the regressand has not been divided by an estimate of σ. As we will see below, the test statistic is no more difficult to calculate by running (6.17) than by running (3.49).

Limiting our attention to zero restrictions makes it possible for us to gain a little more insight into the connection between the GNR and LM tests. Using the FWL Theorem, we see that regression (6.17) will yield exactly the same estimates of $\boldsymbol{b}_2$, namely $\tilde{\boldsymbol{b}}_2$, and exactly the same sum of squared residuals as the regression

$$\boldsymbol{y} - \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{M}}_1 \tilde{\boldsymbol{X}}_2 \boldsymbol{b}_2 + \text{residuals}, \tag{6.18}$$

where $\tilde{\boldsymbol{M}}_1$ is the matrix that projects onto $\mathcal{S}^\perp(\tilde{\boldsymbol{X}}_1)$. The regressand here is not multiplied by $\tilde{\boldsymbol{M}}_1$ because the first-order conditions imply that $\boldsymbol{y} - \tilde{\boldsymbol{x}}$ already lies in $\mathcal{S}^\perp(\tilde{\boldsymbol{X}}_1)$, which in turn implies that $\tilde{\boldsymbol{M}}_1(\boldsymbol{y} - \tilde{\boldsymbol{x}}) = \boldsymbol{y} - \tilde{\boldsymbol{x}}$. The sum of squared residuals from regression (6.18) is

$$(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}}) - (\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{X}}_2 \big(\tilde{\boldsymbol{X}}_2^\top \tilde{\boldsymbol{M}}_1 \tilde{\boldsymbol{X}}_2 \big)^{-1} \tilde{\boldsymbol{X}}_2^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}}).$$

Since $\boldsymbol{y} - \tilde{\boldsymbol{x}}$ lies in $\mathcal{S}^\perp(\tilde{\boldsymbol{X}}_1)$, it is orthogonal to $\tilde{\boldsymbol{X}}_1$. Thus, if we had not included $\tilde{\boldsymbol{X}}_2$ in the regression, the SSR would have been $(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}})$. Hence the reduction in the SSR of regression (6.17) brought about by the inclusion of $\tilde{\boldsymbol{X}}_2$ is

$$(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{X}}_2 \big(\tilde{\boldsymbol{X}}_2^\top \tilde{\boldsymbol{M}}_1 \tilde{\boldsymbol{X}}_2 \big)^{-1} \tilde{\boldsymbol{X}}_2^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}}). \tag{6.19}$$

This quantity is also the explained sum of squares (around zero) from regression (6.17), again because $\tilde{\boldsymbol{X}}_1$ has no explanatory power. We can now show directly that this quantity, divided by any consistent estimate of σ^2, is asymptotically distributed as $\chi^2(r)$ under the null hypothesis. We already showed this in Section 5.7, but the argument that the number of degrees of freedom is r was an indirect one.

First, observe that

$$n^{-1/2}(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{X}}_2 \stackrel{a}{=} n^{-1/2} \boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 \equiv \boldsymbol{\nu}^\top,$$

where $\boldsymbol{M}_1 \equiv \boldsymbol{M}_1(\boldsymbol{\beta}_0)$ and $\boldsymbol{X}_2 \equiv \boldsymbol{X}_2(\boldsymbol{\beta}_0)$. The asymptotic equality here follows from the fact that $\tilde{\boldsymbol{u}} \stackrel{a}{=} \boldsymbol{M}_1 \boldsymbol{u}$, which is the result (6.09) for the case in which the model is estimated subject to the restrictions that $\boldsymbol{\beta}_2 = \boldsymbol{0}$. The covariance matrix of the $r \times 1$ random vector $\boldsymbol{\nu}$ is

$$E(\boldsymbol{\nu}\boldsymbol{\nu}^\top) = E\big(n^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{u} \boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 \big) = n^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 (\sigma_0^2 \mathbf{I}) \boldsymbol{M}_1 \boldsymbol{X}_2$$

$$= n^{-1} \sigma_0^2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2) \equiv \sigma_0^2 \boldsymbol{V}.$$

The consistency of $\tilde{\boldsymbol{\beta}}$ and the regularity conditions for Theorem 5.1 imply that

$$n^{-1} \tilde{\boldsymbol{X}}_2^\top \tilde{\boldsymbol{M}}_1 \tilde{\boldsymbol{X}}_2 \stackrel{a}{=} n^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2 = \boldsymbol{V}.$$

Thus, when b_2 is a scalar, the t statistic on $\tilde{b}_2$ from the GNR (6.17) is just as valid as any of the test statistics we have been discussing.

Why does regressing residuals from the restricted model on the derivatives of $\boldsymbol{x}(\boldsymbol{\beta})$ allow us to compute valid test statistics? Why do we need to include all the derivatives and not merely those that correspond to the parameters which were restricted? The above discussion has provided formal answers to these questions, but perhaps not ones that are intuitively appealing. Let us therefore consider the matter from a slightly different point of view. In Section 5.7, we showed that Wald, LR, and LM statistics for testing the same set of restrictions are all asymptotically equal to the same random variable under the null hypothesis and that this random variable is asymptotically distributed as $\chi^2(r)$. For the nonlinear regression models we have been discussing, the LR statistic is simply the difference between $SSR(\hat{\boldsymbol{\beta}})$ and $SSR(\tilde{\boldsymbol{\beta}})$, divided by any consistent estimate of σ^2. To see why the LM statistic is valid and why the GNR must include the derivatives with respect to all parameters, we will view the LM statistic based on the GNR as a quadratic approximation to this LR statistic. That this should be the case makes sense, since the GNR itself is a linear approximation to the nonlinear regression model.

One way to view the Gauss-Newton regression is to think of it as a way of approximating the function $SSR(\boldsymbol{\beta})$ by a quadratic function that has the same first derivatives and, asymptotically, the same second derivatives at the point $\tilde{\boldsymbol{\beta}}$. This quadratic approximating function, which we will call $SSR^*(\tilde{\boldsymbol{\beta}}, \boldsymbol{b})$, is simply the sum-of-squares function for the artificial regression. It is defined as

$$SSR^*(\tilde{\boldsymbol{\beta}}, \boldsymbol{b}) = (\boldsymbol{y} - \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{X}}\boldsymbol{b})^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{X}}\boldsymbol{b}).$$

The explained sum of squares from the GNR is precisely the difference between $SSR(\tilde{\boldsymbol{\beta}})$ and $SSR^*(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}})$. If $\tilde{\boldsymbol{\beta}}$ is reasonably close to $\hat{\boldsymbol{\beta}}$, $SSR^*(\cdot)$ should provide a good approximation to $SSR(\cdot)$ in the neighborhood of $\hat{\boldsymbol{\beta}}$. Indeed, provided that the restrictions are true and that the sample size is sufficiently large, $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ must be close to each other because they are both consistent for $\boldsymbol{\beta}_0$. Therefore, $SSR^*(\cdot)$ *must* provide a good approximation to $SSR(\cdot)$. This implies that $SSR^*(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}})$ will be close to $SSR(\hat{\boldsymbol{\beta}})$ and that the explained sum of squares from the GNR will provide a good approximation to $SSR(\tilde{\boldsymbol{\beta}}) - SSR(\hat{\boldsymbol{\beta}})$. When we divide the explained sum of squares by a consistent estimate of σ^2, the resulting LM test statistic should therefore be similar to the LR test statistic.

It should now be clear why the GNR has to include $\tilde{\boldsymbol{X}}_1$ as well as $\tilde{\boldsymbol{X}}_2$. If it did not, the GNR would not be minimizing $SSR^*(\tilde{\boldsymbol{\beta}}, \boldsymbol{b})$, but rather *another* approximation to $SSR(\boldsymbol{\beta})$,

$$SSR^{**}(\tilde{\boldsymbol{\beta}}, \boldsymbol{b}_2) = (\boldsymbol{y} - \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{X}}_2\boldsymbol{b}_2)^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{X}}_2\boldsymbol{b}_2).$$

Although $SSR^*(\cdot)$ should normally provide a reasonably good approximation to $SSR(\cdot)$, $SSR^{**}(\cdot)$ normally will not, because it does not have enough free

since the equilibrium price depends, in part, on the error term in the demand equation. Hence the standard assumption that error terms and regressors are independent is violated in this (and every) system of simultaneous equations. Thus, if we attempt to take the plim of the right-hand side of (7.14), we will find that the second term is not zero. It follows that $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ will be inconsistent.

The results of this simple example are true in general. Since they are determined simultaneously, all the endogenous variables in a simultaneous equation system generally depend on the error terms in all the equations. Thus, except perhaps in a few very special cases, the right-hand side endogenous variables in a structural equation from such a system will always be correlated with the error terms. As a consequence, application of OLS to such an equation will always yield biased and inconsistent estimates.

We have now seen two important situations in which explanatory variables will be correlated with the error terms of regression equations, and are ready to take up the main topic of this chapter, namely, the method of instrumental variables. This method can be used whenever the error terms are correlated with one or more explanatory variables, regardless of how that correlation may have arisen. It is remarkably simple, general, and powerful.

7.4 Instrumental Variables: The Linear Case

The fundamental ingredient of any IV procedure is a matrix of **instrumental variables** (or simply **instruments**, for short). We will call this matrix $\boldsymbol{W}$ and specify that it is $n \times l$. The columns of $\boldsymbol{W}$ are simply exogenous and/or predetermined variables that are known (or at least assumed) to be independent of the error terms $\boldsymbol{u}$. In the context of the simultaneous equations model, a natural choice for $\boldsymbol{W}$ is the matrix of all the exogenous and predetermined variables in the model. There must be at least as many instruments as there are explanatory variables in the equation to be estimated. Thus, if the equation to be estimated is the linear regression model (7.01), with $\boldsymbol{X}$ having k columns, we require that $l \geq k$. This is an identification condition; see Section 7.8 for further discussion of conditions for identification in models estimated by IV. Some of the explanatory variables may appear among the instruments. Indeed, as we will see below, any column of $\boldsymbol{X}$ that is known to be exogenous or predetermined should be included in $\boldsymbol{W}$ if we want to obtain asymptotically efficient estimates.

The intuition behind IV procedures is the following. Least squares minimizes the distance between $\boldsymbol{y}$ and $\mathcal{S}(\boldsymbol{X})$, which leads to inconsistent estimates because $\boldsymbol{u}$ is correlated with $\boldsymbol{X}$. The n–dimensional space in which $\boldsymbol{y}$ is a point can be divided into two orthogonal subspaces, $\mathcal{S}(\boldsymbol{W})$ and $\mathcal{S}^{\perp}(\boldsymbol{W})$. Instrumental variables minimizes only the portion of the distance between $\boldsymbol{y}$ and $\mathcal{S}(\boldsymbol{X})$ that lies in $\mathcal{S}(\boldsymbol{W})$. Provided that $\boldsymbol{u}$ is independent of $\boldsymbol{W}$, as assumed, any

variables in the entire system. Then the second-stage regression for $\boldsymbol{y}$ can simply be written as

$$\boldsymbol{y} = \boldsymbol{P}_W \boldsymbol{X} \boldsymbol{\beta} + \text{residuals}. \tag{7.28}$$

The OLS estimator of $\boldsymbol{\beta}$ from this regression is just the IV estimator (7.17):

$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{y}.$$

Notice, however, that the OLS covariance matrix estimate from (7.28) is not the estimate we want. This estimate will be

$$\frac{\| \boldsymbol{y} - \boldsymbol{P}_W \boldsymbol{X} \tilde{\boldsymbol{\beta}} \|^2}{n - k} \left(\boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{X} \right)^{-1}, \tag{7.29}$$

while the estimate (7.24) that was derived earlier can be written as

$$\frac{\| \boldsymbol{y} - \boldsymbol{X} \tilde{\boldsymbol{\beta}} \|^2}{n} \left(\boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{X} \right)^{-1}. \tag{7.30}$$

These two estimates are not the same. They would be the same only if IV and OLS were identical, that is, if $\boldsymbol{X} = \boldsymbol{P}_W \boldsymbol{X}$. In addition, n would have to be replaced by $n - k$ in (7.30). The problem is that the second-stage OLS regression provides an incorrect estimate of σ^2; it uses $\boldsymbol{y} - \boldsymbol{P}_W \boldsymbol{X} \tilde{\boldsymbol{\beta}}$ rather than $\boldsymbol{y} - \boldsymbol{X} \tilde{\boldsymbol{\beta}}$ as the vector of residuals. The second-stage residuals $\boldsymbol{y} - \boldsymbol{P}_W \boldsymbol{X} \tilde{\boldsymbol{\beta}}$ may be either too large or too small, asymptotically. Whether they are too large or too small will depend on σ^2, on the variance of the elements of $\boldsymbol{M}_W \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X} \boldsymbol{\beta} - \boldsymbol{P}_W \boldsymbol{X} \boldsymbol{\beta}$, and on the correlation between $\boldsymbol{M}_W \boldsymbol{X} \boldsymbol{\beta}$ and $\boldsymbol{u}$. If one actually performs 2SLS in two stages, rather than relying on a preprogrammed 2SLS or IV procedure, one must be careful to use (7.30) rather than (7.29) for the estimated covariance matrix.[2] Programs for 2SLS estimation normally replace $\boldsymbol{P}_W \boldsymbol{X} \tilde{\boldsymbol{\beta}}$ by $\boldsymbol{X} \tilde{\boldsymbol{\beta}}$ before calculating the explained sum of squares, the sum of squared residuals, the R^2, and other statistics that depend on these quantities.

There has been an enormous amount of work on the finite-sample properties of 2SLS, that is, the IV estimator $\tilde{\boldsymbol{\beta}}$. A few of the many papers in this area are Anderson (1982), Anderson and Sawa (1979), Mariano (1982), Phillips (1983), and Taylor (1983). Unfortunately, many of the results of this literature are very model-specific. One important result (Kinal, 1980) is that the mth moment of the 2SLS estimator exists if and only if

$$m < l - k + 1.$$

---

[2] 2SLS is a special case of a regression with what Pagan (1984b, 1986) calls "generated regressors." Even when such regressions provide consistent parameter estimates, they usually provide inconsistent estimates of the covariance matrix of the parameter estimates. The inconsistency of (7.29) provides a simple example of this phenomenon.

The ultimate result is

$$n^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{a}{\sim} N\Big(\mathbf{0}, \sigma_0^2 \plim_{n \to \infty} \big(n^{-1}\boldsymbol{X}_0^\top \boldsymbol{P}_W \boldsymbol{X}_0\big)^{-1}\Big), \tag{7.34}$$

which closely resembles (7.23) for the linear case.

The **nonlinear IV estimator** based on minimizing the criterion function (7.32) was proposed by Amemiya (1974), who very misleadingly called it the **nonlinear two-stage least squares estimator**, or **NL2SLS**. In fact, it is *not* computed in two stages at all. Attempting to compute an estimator analogous to linear 2SLS would in general result in an inconsistent estimator very different from nonlinear IV.

It is illuminating to see why this is so. We must make explicit the dependence of $\boldsymbol{x}(\boldsymbol{\beta})$ on explanatory variables. Thus the model (7.31) may be rewritten as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\beta})$ is a vector with typical element $x_t(\boldsymbol{Z}_t, \boldsymbol{\beta})$, $\boldsymbol{Z}$ being a matrix of observations on explanatory variables, with t^{th} row $\boldsymbol{Z}_t$, some columns of which may be correlated with $\boldsymbol{u}$. The $\boldsymbol{Z}$ matrix is not necessarily $n \times k$, because there may be more or fewer parameters than explanatory variables. A 2SLS procedure would regress those columns of $\boldsymbol{Z}$ that are potentially correlated with $\boldsymbol{u}$ on the matrix of instruments $\boldsymbol{W}$ so as to obtain $\boldsymbol{P}_W \boldsymbol{Z}$. It would then minimize the objective function

$$\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{P}_W \boldsymbol{Z}, \boldsymbol{\beta})\big)^\top \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{P}_W \boldsymbol{Z}, \boldsymbol{\beta})\big). \tag{7.35}$$

This procedure would yield consistent estimates if the regression functions $x_t(\boldsymbol{Z}_t, \boldsymbol{\beta})$ were linear in all the endogenous elements of $\boldsymbol{Z}_t$. But if the regression functions were nonlinear in any of the endogenous elements of $\boldsymbol{Z}_t$, minimizing (7.35) would not yield consistent estimates, because even though $\boldsymbol{P}_W \boldsymbol{Z}$ would be asymptotically orthogonal to $\boldsymbol{u}$, $\boldsymbol{X}(\boldsymbol{Z}, \boldsymbol{\beta}) \boldsymbol{P}_W$ would not be.

As a very simple example, suppose that the regression function $x_t(\boldsymbol{Z}_t, \boldsymbol{\beta})$ were βz_t^2. Thus there would be just one independent variable, which is correlated with u_t, and one parameter. The theory for linear regressions is applicable to this example, since the regression function is linear with respect to the parameter β. What is needed to obtain a consistent estimate of β is to minimize $\|\boldsymbol{P}_W(\boldsymbol{y} - \beta \boldsymbol{z}^2)\|^2$ with respect to β, where $\boldsymbol{z}^2$ means the vector with typical element z_t^2. In contrast, if one first projected $\boldsymbol{z}$ onto $\boldsymbol{W}$ in a 2SLS procedure, one would be minimizing $\|\boldsymbol{y} - \beta(\boldsymbol{P}_W \boldsymbol{z})^2\|^2$, where $(\boldsymbol{P}_W \boldsymbol{z})^2$ means the vector with typical element $(\boldsymbol{P}_W \boldsymbol{z})_t^2$. The latter minimization is evidently not restricted to the subspace $\mathcal{S}(\boldsymbol{W})$, and so it will not in general yield consistent estimates of β.

In many cases, the biggest problem with nonlinear IV procedures is how to choose $\boldsymbol{W}$. With a linear model, it is relatively easy to do so. If the equation to be estimated comes from a system of linear simultaneous equations,

and that these are estimated by IV using the instrument matrix $\boldsymbol{W}$. Now suppose that the estimates are actually obtained by two-stage least squares. It is easy to see that the sum of squared residuals from the second-stage regression for (7.43), in which $\boldsymbol{X}_1$ is replaced by $\boldsymbol{P}_W\boldsymbol{X}_1$, will be

$$\text{RSSR}^* \equiv \boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{y}, \tag{7.45}$$

where $\boldsymbol{M}_1$ denotes the matrix that projects orthogonally onto $\mathcal{S}^\perp(\boldsymbol{P}_W\boldsymbol{X}_1)$. Similarly, it can be shown (doing so is a good exercise) that the sum of squared residuals from the second-stage regression for (7.44) will be

$$\text{USSR}^* \equiv \boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{P}_W \boldsymbol{X}_2 \big(\boldsymbol{X}_2^\top \boldsymbol{P}_W \boldsymbol{M}_1 \boldsymbol{P}_W \boldsymbol{X}_2 \big)^{-1} \boldsymbol{X}_2^\top \boldsymbol{P}_W \boldsymbol{M}_1 \boldsymbol{y}. \tag{7.46}$$

The difference between (7.45) and (7.46) is

$$\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{P}_W \boldsymbol{X}_2 \big(\boldsymbol{X}_2^\top \boldsymbol{P}_W \boldsymbol{M}_1 \boldsymbol{P}_W \boldsymbol{X}_2 \big)^{-1} \boldsymbol{X}_2^\top \boldsymbol{P}_W \boldsymbol{M}_1 \boldsymbol{y}, \tag{7.47}$$

which bears a striking and by no means coincidental resemblance to expression (7.41). Under the null hypothesis (7.43), $\boldsymbol{y}$ is equal to $\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{u}$. Since $\boldsymbol{P}_W\boldsymbol{M}_1$ annihilates $\boldsymbol{X}_1$, (7.47) reduces to

$$\boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{P}_W \boldsymbol{X}_2 \big(\boldsymbol{X}_2^\top \boldsymbol{P}_W \boldsymbol{M}_1 \boldsymbol{P}_W \boldsymbol{X}_2 \big)^{-1} \boldsymbol{X}_2^\top \boldsymbol{P}_W \boldsymbol{M}_1 \boldsymbol{u}$$

under the null. It should be easy to see that, under reasonable assumptions, this quantity, divided by anything which estimates σ^2 consistently, will be asymptotically distributed as $\chi^2(r)$. The needed assumptions are essentially (7.18a)–(7.18c), plus assumptions sufficient for a central limit theorem to apply to $n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}$.

The problem, then, is to estimate σ^2. Notice that $\text{USSR}^*/(n-k)$ does *not* estimate σ^2 consistently, for the reasons discussed in Section 7.5. As we saw there, the residuals from the second-stage regression may be either too large or too small. Thus estimates of σ^2 must be based on the set of residuals $\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ rather than the set $\boldsymbol{y} - \boldsymbol{P}_W\boldsymbol{X}\tilde{\boldsymbol{\beta}}$. One valid estimate is $\text{USSR}/(n-k)$, where

$$\text{USSR} \equiv \big\| \boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\tilde{\boldsymbol{\beta}}_2 \big\|^2.$$

The analog of (7.42) would then be

$$\frac{(\text{RSSR}^* - \text{USSR}^*)/r}{\text{USSR}/(n-k)} \overset{a}{\sim} F(r, n-k). \tag{7.48}$$

Notice that the numerator and denominator of this test statistic are based on different sets of residuals. The numerator is $1/r$ times the difference between the sums of squared residuals from the second-stage regressions, while the denominator is $1/(n-k)$ times the sum of squared residuals that would be printed by a program for IV estimation.

We must now show that the SSR from regression (7.50) is asymptotically equal to minus the second term in expression (7.49). This SSR is

$$\left\| \boldsymbol{P}_W\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\check{\beta}}) - \boldsymbol{\check{X}}\boldsymbol{\check{b}}\big) \right\|^2,$$

where $\boldsymbol{\check{b}}$ is the vector of parameter estimates from OLS estimation of (7.50). Recall from the results of Section 6.6 on one-step estimation that $(\boldsymbol{\tilde{\beta}} - \boldsymbol{\check{\beta}})$ is asymptotically equal to the estimate $\boldsymbol{\check{b}}$ from the GNR (7.38). Thus

$$\boldsymbol{P}_W\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\check{\beta}}) - \boldsymbol{\check{X}}\boldsymbol{\check{b}}\big) \overset{a}{=} \boldsymbol{P}_W\boldsymbol{y} - \boldsymbol{P}_W\boldsymbol{x}(\boldsymbol{\check{\beta}}) - \boldsymbol{P}_W\boldsymbol{\check{X}}(\boldsymbol{\tilde{\beta}} - \boldsymbol{\check{\beta}}). \qquad (7.52)$$

But a first-order Taylor expansion of $\boldsymbol{x}(\boldsymbol{\tilde{\beta}})$ about $\boldsymbol{\beta} = \boldsymbol{\check{\beta}}$ gives

$$\boldsymbol{x}(\boldsymbol{\tilde{\beta}}) \cong \boldsymbol{x}(\boldsymbol{\check{\beta}}) + \boldsymbol{X}(\boldsymbol{\check{\beta}})(\boldsymbol{\tilde{\beta}} - \boldsymbol{\check{\beta}}).$$

Subtracting the right-hand side of this expression from $\boldsymbol{y}$ and multiplying by $\boldsymbol{P}_W$ yields the right-hand side of (7.52). Thus we see that the SSR from regression (7.50) is asymptotically equal to

$$\left\| \boldsymbol{P}_W\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\tilde{\beta}})\big) \right\|^2,$$

which is the second term of (7.49). We have therefore proved that the difference between the restricted and unrestricted values of the criterion function, expression (7.49), is asymptotically equivalent to the explained sum of squares from the GNR (7.38). Since the latter can be used to construct a valid test statistic, so can the former.

This result is important. It tells us that we can always construct a test of a hypothesis about $\boldsymbol{\beta}$ by taking the difference between the restricted and unrestricted values of the criterion function for IV estimation and dividing it by anything that estimates σ^2 consistently. Moreover, such a test will be asymptotically equivalent to taking the explained sum of squares from the GNR evaluated at $\boldsymbol{\check{\beta}}$ and treating it in the same way. Either of these tests can be turned into an asymptotic F test by dividing numerator and denominator by their respective degrees of freedom, r and $n - k$. Whether this is actually a good thing to do in finite samples is unclear, however.

7.8 Identification and Overidentifying Restrictions

Identification is a somewhat more complicated matter in models estimated by IV than in models estimated by least squares, because the choice of instruments affects whether the model is identified or not. A model that would not be identified if it were estimated by least squares will also not be identified if it is estimated by IV. However, a model that would be identified if it were estimated by least squares may not be identified if it is estimated by IV using a

useful when it is not clear whether it is safe to use least squares rather than instrumental variables.

Regression (7.63) deserves further comment. It has the remarkable feature that the OLS estimates of $\boldsymbol{\beta}$ are numerically identical to the IV estimates of $\boldsymbol{\beta}$ in the original model (7.01). Moreover, the estimated covariance matrices are also the same, except that the OLS estimate from (7.63) uses an inconsistent estimator for σ^2. These results are easy to obtain. Denote by $\boldsymbol{M}^*$ the orthogonal projection onto the space $\mathcal{S}^{\perp}(\boldsymbol{M}_W\boldsymbol{X}^*)$. Then, by the FWL Theorem, the OLS estimates from (7.63) must be identical to those from the regression

$$\boldsymbol{M}^*\boldsymbol{y} = \boldsymbol{M}^*\boldsymbol{X}\boldsymbol{\beta} + \text{residuals}. \tag{7.65}$$

Now

$$\boldsymbol{M}^*\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{M}_W\boldsymbol{X}^*\big(\boldsymbol{X}^{*\top}\boldsymbol{M}_W\boldsymbol{X}^*\big)^{-1}\boldsymbol{X}^{*\top}\boldsymbol{M}_W\boldsymbol{X}.$$

From the fact that $\boldsymbol{M}_W\boldsymbol{X} = [\boldsymbol{M}_W\boldsymbol{X}^* \quad \boldsymbol{0}]$, it follows that

$$\boldsymbol{X}^{*\top}\boldsymbol{M}_W\boldsymbol{X} = \boldsymbol{X}^{*\top}\boldsymbol{M}_W\big[\boldsymbol{X}^* \quad \boldsymbol{0}\big].$$

Consequently, we obtain

$$\boldsymbol{M}^*\boldsymbol{X} = \boldsymbol{X} - \big[\boldsymbol{M}_W\boldsymbol{X}^* \quad \boldsymbol{0}\big] = \boldsymbol{X} - \boldsymbol{M}_W\boldsymbol{X} = \boldsymbol{P}_W\boldsymbol{X}.$$

Then the OLS estimate of $\boldsymbol{\beta}$ from (7.65) is seen to be

$$\big(\boldsymbol{X}^{\top}\boldsymbol{M}^*\boldsymbol{X}\big)^{-1}\boldsymbol{X}^{\top}\boldsymbol{M}^*\boldsymbol{y} = \big(\boldsymbol{X}^{\top}\boldsymbol{P}_W\boldsymbol{X}\big)^{-1}\boldsymbol{X}^{\top}\boldsymbol{P}_W\boldsymbol{y}. \tag{7.66}$$

The right-hand side of (7.66) is of course the expression for the IV or 2SLS estimate of $\boldsymbol{\beta}$, expression (7.17).

By an extension of this argument, it is easy to see that the estimated OLS covariance matrix of $\hat{\boldsymbol{\beta}}$ from (7.63) will be

$$\tilde{s}^2\big(\boldsymbol{X}^{\top}\boldsymbol{P}_W\boldsymbol{X}\big)^{-1}, \tag{7.67}$$

where $\tilde{s}^2$ denotes the OLS estimate of the error variance in (7.63). Expression (7.67) looks just like the IV covariance matrix (7.24), except that $\tilde{s}^2$ appears instead of $\tilde{\sigma}^2$. When $\boldsymbol{\eta}$ is nonzero (so that IV estimation is necessary), the variance of the errors in (7.63) will be less than σ^2. As a consequence, $\tilde{s}^2$ will be biased downward as an estimator of σ^2. Of course, it would be easy to obtain a valid estimated covariance matrix by multiplying (7.67) by $\tilde{\sigma}^2/\tilde{s}^2$.

We now return to the DWH test. A variant of this test is applicable to nonlinear models like (7.31) as well as to linear ones. The test would then be based on a variant of the Gauss-Newton regression. If the null were that the NLS estimates $\hat{\boldsymbol{\beta}}$ were consistent, an appropriate test statistic would be an asymptotic F test for $\boldsymbol{c} = \boldsymbol{0}$ in the GNR

$$\boldsymbol{y} - \hat{\boldsymbol{x}} = \hat{\boldsymbol{X}}\boldsymbol{b} + \boldsymbol{M}_W\hat{\boldsymbol{X}}^*\boldsymbol{c} + \text{residuals}, \tag{7.68}$$

We leave it as an exercise to prove that a test of whether the vector (7.69) has mean zero asymptotically may be accomplished by testing whether the q–vector $\boldsymbol{\delta}$ is equal to zero in the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_2\boldsymbol{X}^*\boldsymbol{\delta} + \text{residuals.} \tag{7.70}$$

Here $\boldsymbol{P}_2\boldsymbol{X}^*$ consists of the q columns of $\boldsymbol{P}_2\boldsymbol{X}$ that are not annihilated by $\boldsymbol{M}_{P_1X}$. Regression (7.70) must be estimated by IV using $\boldsymbol{W}_1$ as the matrix of instruments, and any of the tests discussed in Section 7.7 may then be used to test whether $\boldsymbol{\delta} = \boldsymbol{0}$.

7.10 Conclusion

This chapter has introduced all of the important concepts associated with the technique of instrumental variables estimation. For a more detailed treatment, see Bowden and Turkington (1984). Another useful reference is Godfrey (1988, Chapter 5), which discusses a large number of specification tests for both linear and nonlinear models that have been estimated by IV.

In this chapter, we applied the method of instrumental variables only to univariate linear and nonlinear regression models with i.i.d. errors. We will encounter numerous other applications later in the book, notably in Chapters 17 and 18, in which we discuss GMM estimation and simultaneous equations models, respectively. In many other cases, we will state a result in the context of OLS or NLS estimation and point out that it goes through with minor modification in the context of IV estimation as well.

Terms and Concepts

criterion function
Durbin-Wu-Hausman (DWH) tests
errors in variables
exactly identified (just identified)
    model
Gauss-Newton regression (GNR)
generalized IV estimator
identification: local, global, and
    asymptotic
instrumental variables (IV) estimator
instruments (instrumental variables)
nonlinear IV estimator
nonlinear two-stage least squares
    (NL2SLS) estimator
normalization (of a simultaneous
    equations model)

overidentified model
overidentifying restrictions
predetermined variable
reduced form (of a simultaneous
    equations model)
restricted reduced form (RRF)
simple IV estimator
simultaneous equations bias
simultaneous equations model
structural form (of a simultaneous
    equations model)
two-stage least squares (2SLS)
    estimator
unrestricted reduced form (URF)
vector of contrasts

In words, the limiting Hessian matrix is the negative of the limiting information matrix. An analogous result is true for individual observations:

$$E_0\big(D_{\theta\theta}^2 \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\big) = -E_0\big(D_\theta^\top \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0) D_\theta \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\big). \tag{8.44}$$

The latter result clearly implies the former, given the assumptions that permit the application of a law of large numbers to the sequences $\{D_{\theta\theta}^2 \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\}_{t=1}^\infty$ and $\{D_\theta^\top \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0) D_\theta \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\}_{t=1}^\infty$.

The result (8.44) is proved by an argument very similar to that used at the beginning of the last section in order to show that the expectation of the CG matrix is zero. From the fact that

$$\frac{\partial \ell_t}{\partial \theta_i} = \frac{1}{L_t} \frac{\partial L_t}{\partial \theta_i},$$

we obtain after a further differentiation that

$$\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} = \frac{1}{L_t} \frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j} - \frac{1}{L_t^2} \frac{\partial L_t}{\partial \theta_i} \frac{\partial L_t}{\partial \theta_j}.$$

Consequently,

$$\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} + \frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} = \frac{1}{L_t} \frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j}. \tag{8.45}$$

If now we take the expectation of (8.45) for the DGP characterized by the same value of the parameter vector $\boldsymbol{\theta}$ as that at which the functions ℓ_t and L_t are evaluated (which as usual we denote by E_θ), we find that

$$\begin{aligned} E_\theta \left(\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} + \frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} \right) &= \int L_t \frac{1}{L_t} \frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j} \, dy_t \\ &= \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int L_t \, dy_t = 0, \end{aligned} \tag{8.46}$$

provided that, as for (8.34), the interchange of the order of differentiation and integration can be justified. The result (8.46) now establishes (8.44), since it implies that

$$E_\theta \left(\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} \right) = 0 - E_\theta \left(\frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} \right) = -E_\theta \left(\frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} \right).$$

In order to establish (8.43), recall that, from (8.19) and the law of large numbers,

$$\begin{aligned} \mathcal{H}_{ij}(\boldsymbol{\theta}) &= \lim_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^n E_\theta \left(\frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \right) \\ &= -\lim_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^n E_\theta \left(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_j} \right) \right) \\ &= -\mathcal{I}_{ij}(\boldsymbol{\theta}), \end{aligned}$$

where the last line follows immediately from the definition of the limiting information matrix, (8.22). This then establishes (8.43).

By substituting either $-\mathcal{H}(\boldsymbol{\theta}_0)$ for $\mathcal{I}(\boldsymbol{\theta}_0)$ or $\mathcal{I}(\boldsymbol{\theta}_0)$ for $-\mathcal{H}(\boldsymbol{\theta}_0)$ in (8.42), it is now easy to conclude that the asymptotic covariance matrix of the ML estimator is given by either of the two equivalent expressions $-\mathcal{H}(\boldsymbol{\theta}_0)^{-1}$ and $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$. Formally, we may write

$$\boldsymbol{V}^{\infty}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = \mathcal{I}^{-1}(\boldsymbol{\theta}_0) = -\mathcal{H}^{-1}(\boldsymbol{\theta}_0).$$

In order to perform any statistical inference, it is necessary to be able to *estimate* $\mathcal{I}^{-1}(\boldsymbol{\theta}_0)$ or $-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)$. One estimator which suggests itself at once is $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$, that is, the inverse of the limiting information matrix evaluated at the MLE, $\hat{\boldsymbol{\theta}}$. Notice that the matrix function $\mathcal{I}(\boldsymbol{\theta})$ is *not* a sample-dependent object. It can, in principle, be computed theoretically as a matrix function of the model parameters from the (sequence of) loglikelihood functions ℓ^n. For some models, this is an entirely feasible computation, and then it yields what is often the preferred estimator of the asymptotic covariance matrix. But for many models the computation, even if feasible, would be excessively laborious, and in these cases it is convenient to have available other consistent estimators of $\mathcal{I}(\boldsymbol{\theta}_0)$ and consequently of the asymptotic covariance matrix.

One common estimator is the negative of the so-called **empirical Hessian**. This matrix is defined as

$$\hat{\mathcal{H}} \equiv \frac{1}{n}\sum_{t=1}^{n} D_{\theta\theta}^2 \ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}). \tag{8.47}$$

The consistency of $\hat{\boldsymbol{\theta}}$ and the application of a law of large numbers to the right-hand side guarantees the consistency of (8.47) for $\mathcal{H}(\boldsymbol{\theta}_0)$. When the empirical Hessian is readily available, as it will be if maximization routines that use second derivatives are employed, minus its inverse can provide a very convenient way to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$. However, the Hessian is often difficult to compute, and if it is not already being calculated for other purposes, it probably does not make sense to compute it just to estimate a covariance matrix.

Another commonly used estimator of the information matrix is known as the **outer-product-of-the-gradient estimator**, or **OPG estimator**. It is based on the definition

$$\mathcal{I}(\boldsymbol{\theta}) \equiv \lim_{n \to \infty}\left(\frac{1}{n}\sum_{t=1}^{n} E_{\theta}\big(D_{\theta}^{\top}\ell_t(\boldsymbol{\theta}) D_{\theta}\ell_t(\boldsymbol{\theta})\big)\right).$$

The OPG estimator is

$$\hat{\mathcal{I}}_{\text{OPG}} \equiv \frac{1}{n}\sum_{t=1}^{n} D_{\theta}^{\top}\ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) D_{\theta}\ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{n}\boldsymbol{G}^{\top}(\hat{\boldsymbol{\theta}})\boldsymbol{G}(\hat{\boldsymbol{\theta}}), \tag{8.48}$$

to be numerically identical if the same estimate of the information matrix is used to calculate them. One form, originally proposed by Rao (1948), is called the **score form of the LM test**, or simply the **score test**, and is calculated using the gradient or score vector of the unrestricted model evaluated at the restricted estimates. The other form, which gives the test its name, was proposed by Aitchison and Silvey (1958, 1960) and Silvey (1959). This latter form is calculated using the vector of Lagrange multipliers which emerge if one maximizes the likelihood function subject to constraints by means of a Lagrangian. Econometricians generally use the LM test in its score form but nevertheless insist on calling it an LM test, perhaps because Lagrange multipliers are so widely used in economics. References on LM tests in econometrics include Breusch and Pagan (1980) and Engle (1982a, 1984). Buse (1982) provides an intuitive discussion of the relationships among the LR, LM, and Wald tests.

One way to maximize $\ell(\boldsymbol{\theta})$ subject to the exact restrictions

$$r(\boldsymbol{\theta}) = \mathbf{0}, \tag{8.71}$$

where $r(\boldsymbol{\theta})$ is an r–vector with $r \leq k$, is simultaneously to maximize the Lagrangian

$$\ell(\boldsymbol{\theta}) - r^{\top}(\boldsymbol{\theta})\boldsymbol{\lambda}$$

with respect to $\boldsymbol{\theta}$ and minimize it with respect to the r–vector of Lagrange multipliers $\boldsymbol{\lambda}$. The first-order conditions that characterize the solution to this problem are

$$\begin{aligned} g(\tilde{\boldsymbol{\theta}}) - \boldsymbol{R}^{\top}(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} &= \mathbf{0} \\ r(\tilde{\boldsymbol{\theta}}) &= \mathbf{0}, \end{aligned} \tag{8.72}$$

where $\boldsymbol{R}(\boldsymbol{\theta})$ is a $r \times k$ matrix with typical element $\partial r_i(\boldsymbol{\theta})/\partial \theta_j$.

We are interested in the distribution of $\tilde{\boldsymbol{\lambda}}$ under the null hypothesis, so we will suppose that the DGP satisfies (8.71) with parameter vector $\boldsymbol{\theta}_0$. The value of the vector of Lagrange multipliers $\boldsymbol{\lambda}$ if $\tilde{\boldsymbol{\theta}}$ were equal to $\boldsymbol{\theta}_0$ would be zero. Thus it seems natural to take a first-order Taylor expansion of the first-order conditions (8.72) around the point $(\boldsymbol{\theta}_0, \mathbf{0})$. This yields

$$g(\boldsymbol{\theta}_0) + \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \boldsymbol{R}^{\top}(\bar{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} = \mathbf{0}$$

$$\boldsymbol{R}(\ddot{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0},$$

where $\bar{\boldsymbol{\theta}}$ and $\ddot{\boldsymbol{\theta}}$ denote values of $\boldsymbol{\theta}$ that lie between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. These equations may be rewritten as

$$\begin{bmatrix} -\boldsymbol{H}(\bar{\boldsymbol{\theta}}) & \boldsymbol{R}^{\top}(\bar{\boldsymbol{\theta}}) \\ \boldsymbol{R}(\ddot{\boldsymbol{\theta}}) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \tilde{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} g(\boldsymbol{\theta}_0) \\ \mathbf{0} \end{bmatrix}. \tag{8.73}$$

If we multiply $\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ by n^{-1}, $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ by $n^{1/2}$, $g(\boldsymbol{\theta}_0)$ by $n^{-1/2}$, and $\tilde{\boldsymbol{\lambda}}$ by $n^{-1/2}$, we do not change the equality in (8.73), and we render all quantities that

The LM statistic (8.76) is numerically equal to a test based on the score vector $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$. By the first set of first-order conditions (8.72), $\boldsymbol{g}(\tilde{\boldsymbol{\theta}}) = \tilde{\boldsymbol{R}}^{\top}\tilde{\boldsymbol{\lambda}}$. Substituting $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$ for $\tilde{\boldsymbol{R}}^{\top}\tilde{\boldsymbol{\lambda}}$ in (8.76) yields the score form of the LM test,

$$\frac{1}{n}\tilde{\boldsymbol{g}}^{\top}\tilde{\mathcal{I}}^{-1}\tilde{\boldsymbol{g}}. \tag{8.77}$$

In practice, this score form is often more useful than the LM form because, since restricted estimates are rarely obtained via a Lagrangian, $\tilde{\boldsymbol{g}}$ is generally readily available while $\tilde{\boldsymbol{\lambda}}$ typically is not. However, deriving the test via the Lagrange multipliers is illuminating, because this derivation makes it quite clear why the test has r degrees of freedom.

The third of the three classical tests is the **Wald test**. This test is very easy to derive. It asks whether the vector of restrictions, evaluated at the unrestricted estimates, is close enough to a zero vector for the restrictions to be plausible. In the case of the restrictions (8.71), the Wald test is based on the vector $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$, which should tend to a zero vector asymptotically if the restrictions hold. As we have seen in Sections 8.5 and 8.6,

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N\big(\boldsymbol{0},\, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)\big).$$

A Taylor-series approximation of $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ yields $\boldsymbol{r}(\hat{\boldsymbol{\theta}}) \cong \boldsymbol{R}_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Therefore,

$$\boldsymbol{V}\big(n^{1/2}\boldsymbol{r}(\hat{\boldsymbol{\theta}})\big) \overset{a}{=} \boldsymbol{R}_0\,\mathcal{I}_0^{-1}\boldsymbol{R}_0^{\top}.$$

It follows that an appropriate test statistic is

$$n\boldsymbol{r}^{\top}(\hat{\boldsymbol{\theta}})\big(\hat{\boldsymbol{R}}\hat{\mathcal{I}}^{-1}\hat{\boldsymbol{R}}^{\top}\big)^{-1}\boldsymbol{r}(\hat{\boldsymbol{\theta}}), \tag{8.78}$$

where $\hat{\mathcal{I}}$ denotes any consistent estimate of $\mathcal{I}(\boldsymbol{\theta}_0)$ based on the unrestricted estimates $\hat{\boldsymbol{\theta}}$. Different variants of the Wald test will use different estimates of $\mathcal{I}(\boldsymbol{\theta}_0)$. It is easy to see that given suitable regularity the test statistic (8.78) will be asymptotically distributed as $\chi^2(r)$ under the null.

The fundamental property of the three classical test statistics is that under the null hypothesis, as $n \to \infty$, they all tend to the same random variable, which is distributed as $\chi^2(r)$. We will prove this result in Chapter 13. The implication is that, in large samples, it does not really matter which of the three tests we use. If both $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are easy to compute, it is attractive to use the LR test. If $\tilde{\boldsymbol{\theta}}$ is easy to compute but $\hat{\boldsymbol{\theta}}$ is not, as is often the case for tests of model specification, then the LM test becomes attractive. If on the other hand $\hat{\boldsymbol{\theta}}$ is easy to compute but $\tilde{\boldsymbol{\theta}}$ is not, as may be the case when we are interested in nonlinear restrictions on a linear model, then the Wald test becomes attractive. When the sample size is not large, choice among the three tests is complicated by the fact that they may have very different finite-sample properties, which may further differ greatly among the alternative variants of the LM and Wald tests. This makes the choice of tests rather more complicated in practice than asymptotic theory would suggest.

same answer, if it is feasible to calculate $\mathcal{I}(\boldsymbol{\theta})$ at all, although one approach may be easier than the other in any given situation.

For the nonlinear regression model (8.79), the parameter vector $\boldsymbol{\theta}$ is the vector $[\boldsymbol{\beta} \vdots \sigma]$. We now calculate the limiting information matrix $\mathcal{I}(\boldsymbol{\beta}, \sigma)$ for this model using the second method, based on the CG matrix, which requires only first derivatives. It is a good exercise to repeat the derivation using the Hessian, which requires second derivatives, and verify that it yields the same results. The first derivative of $\ell_t(y_t, \boldsymbol{\beta}, \sigma)$ with respect to β_i is

$$\frac{\partial \ell_t}{\partial \beta_i} = \frac{1}{\sigma^2}\big(y_t - x_t(\boldsymbol{\beta})\big)X_{ti}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\,e_t(\boldsymbol{\beta})X_{ti}(\boldsymbol{\beta}), \qquad (8.83)$$

where $e_t(\boldsymbol{\beta}) \equiv y_t - x_t(\boldsymbol{\beta})$ and, as usual, $X_{ti}(\boldsymbol{\beta}) \equiv \partial x_t(\boldsymbol{\beta})/\partial \beta_i$. The first derivative of $\ell_t(y_t, \boldsymbol{\beta}, \sigma)$ with respect to σ is

$$\frac{\partial \ell_t}{\partial \sigma} = -\frac{1}{\sigma} + \frac{\big(y_t - x_t(\boldsymbol{\beta})\big)^2}{\sigma^3} = -\frac{1}{\sigma} + \frac{e_t^2(\boldsymbol{\beta})}{\sigma^3}. \qquad (8.84)$$

Expressions (8.83) and (8.84) are all that we need to calculate the information matrix using the CG matrix. The column of that matrix which corresponds to σ will have typical element (8.84), while the remaining k columns, which correspond to the β_i's, will have typical element (8.83).

The element of $\mathcal{I}(\boldsymbol{\beta}, \sigma)$ corresponding to β_i and β_j is

$$\mathcal{I}(\beta_i, \beta_j) = \operatorname*{plim}_{n \to \infty}\left(\frac{1}{n}\sum_{t=1}^{n} \frac{e_t^2(\boldsymbol{\beta})}{\sigma^4} X_{ti}(\boldsymbol{\beta})X_{tj}(\boldsymbol{\beta})\right).$$

Since $e_t^2(\boldsymbol{\beta})$ has expectation σ^2 under the DGP characterized by $(\boldsymbol{\beta}, \sigma)$ and is independent of $\boldsymbol{X}(\boldsymbol{\beta})$, we can replace it by σ^2 here to yield

$$\mathcal{I}(\beta_i, \beta_j) = \operatorname*{plim}_{n \to \infty}\left(\frac{1}{n}\sum_{t=1}^{n} \frac{1}{\sigma^2} X_{ti}(\boldsymbol{\beta})X_{tj}(\boldsymbol{\beta})\right).$$

Thus we see that the whole $(\boldsymbol{\beta}, \boldsymbol{\beta})$ block of the limiting information matrix is

$$\frac{1}{\sigma^2} \operatorname*{plim}_{n \to \infty}\left(\frac{1}{n}\boldsymbol{X}^{\top}(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})\right). \qquad (8.85)$$

The element of $\mathcal{I}(\boldsymbol{\beta}, \sigma)$ corresponding to σ is

$$\begin{aligned}
\mathcal{I}(\sigma, \sigma) &= \operatorname*{plim}_{n \to \infty}\left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{\sigma^2} + \frac{e_t^4(\boldsymbol{\beta})}{\sigma^6} - \frac{2e_t^2(\boldsymbol{\beta})}{\sigma^4}\right)\right) \\
&= \frac{1}{n}\left(\frac{n}{\sigma^2} + \frac{3n\sigma^4}{\sigma^6} - \frac{2n\sigma^2}{\sigma^4}\right) \qquad (8.86) \\
&= \frac{2}{\sigma^2}.
\end{aligned}$$

over all t and then taking the logarithm yields the Jacobian term that appears in (8.92).

Concentrating the loglikelihood function with respect to σ yields

$$
\begin{aligned}
\ell^c(\boldsymbol{\beta}, \gamma) = C - \tfrac{n}{2} \log\left(\sum_{t=1}^{n} \left(y_t^\gamma - \beta_0 - \beta_1 x_t\right)^2\right) \\
+ n \log |\gamma| + (\gamma - 1) \sum_{t=1}^{n} \log(y_t).
\end{aligned}
\tag{8.93}
$$

Maximizing this with respect to γ and $\boldsymbol{\beta}$ is straightforward. If a suitable nonlinear optimization program is not available, one can simply do a one-dimensional search over γ, calculating β_0 and β_1 conditional on γ by means of least squares, so as to find the value $\hat{\gamma}$ that maximizes (8.93). Of course, one cannot use the OLS covariance matrix obtained in this way, since it treats $\hat{\gamma}$ as fixed. The information matrix is *not* block-diagonal between $\boldsymbol{\beta}$ and the other parameters of (8.91), so one must calculate and invert the full information matrix to obtain an estimated covariance matrix.

ML estimation works in this case because of the Jacobian term that appears in (8.92) and (8.93). It vanishes when $\gamma = 1$ but plays an extremely important role for all other values of γ. We saw in Section 8.1 that if one applied NLS to (8.01) and all the y_t's were greater than unity, one would end up with an infinitely large and negative estimate of γ. That will not happen if one uses maximum likelihood, because the term $(\gamma - 1) \sum_{t=1}^{n} \log(y_t)$ will tend to minus infinity as $\gamma \to -\infty$ much faster than $-n/2$ times the logarithm of the sum-of-squares term tends to plus infinity. This example illustrates how useful ML estimation can be for dealing with modified regression models in which the dependent variable is subject to a transformation. We will encounter other problems of this type in Chapter 14.

ML estimation can also be very useful when it is believed that the error terms are nonnormal. As an extreme example, consider the following model:

$$
y_t = \boldsymbol{X}_t \boldsymbol{\beta} + \alpha \varepsilon_t, \quad f(\varepsilon_t) = \frac{1}{\pi(1 + \varepsilon_t^2)},
\tag{8.94}
$$

where $\boldsymbol{\beta}$ is a k–vector and $\boldsymbol{X}_t$ is the tth row of an $n \times k$ matrix. The density of ε_t here is the Cauchy density (see Section 4.6) and ε_t therefore has no finite moments. The parameter α is simply a scale parameter, *not* the standard error of the error terms; since the Cauchy distribution has no moments, the error terms do not have a standard error.

If we write ε_t as a function of y_t, we find that

$$
\varepsilon_t = \frac{y_t - \boldsymbol{X}_t \boldsymbol{\beta}}{\alpha}.
$$

Consider the class of models

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim N\big(\boldsymbol{0}, \boldsymbol{\Omega}(\boldsymbol{\alpha})\big). \tag{9.31}$$

By modifying the loglikelihood function (9.03) slightly, we find that the loglikelihood function corresponding to (9.31) is

$$
\begin{aligned}
\ell^n(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = {}& -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Omega}(\boldsymbol{\alpha})| \\
& -\tfrac{1}{2}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big).
\end{aligned}
\tag{9.32}
$$

There will be two sets of first-order conditions, one for $\boldsymbol{\alpha}$ and one for $\boldsymbol{\beta}$. The latter will be similar to the first-order conditions (9.05) for GNLS:

$$\boldsymbol{X}^{\top}(\hat{\boldsymbol{\beta}})\,\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})\big(\boldsymbol{y} - \boldsymbol{x}(\hat{\boldsymbol{\beta}})\big) = \boldsymbol{0}.$$

The former will be rather complicated and will depend on precisely how $\boldsymbol{\Omega}$ is related to $\boldsymbol{\alpha}$. For a more detailed treatment, see Magnus (1978).

In Section 8.10, we saw that the information matrix for $\boldsymbol{\beta}$ and σ in a nonlinear regression model with covariance matrix $\sigma^2 \mathbf{I}$ is block-diagonal between $\boldsymbol{\beta}$ and σ. An analogous result turns out to be true for the model (9.31) as well: The information matrix is block-diagonal between $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. This means that, asymptotically, the vectors $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ are independent. Thus the fact that $\hat{\boldsymbol{\alpha}}$ is estimated jointly with $\hat{\boldsymbol{\beta}}$ can be ignored, and $\hat{\boldsymbol{\beta}}$ will have the same properties asymptotically as the GNLS estimator $\hat{\boldsymbol{\beta}}$ and the feasible GNLS estimator $\tilde{\boldsymbol{\beta}}$.

The above argument does not require that the error terms u_t actually be normally distributed. All that we require is that the vectors $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ be asymptotically independent and $O(1)$ under whatever DGP actually generated the data. It can be shown that this is in fact the case under fairly general conditions, similar to the conditions detailed in Chapter 5 for least squares to be consistent and asymptotically normal; see White (1982) and Gouriéroux, Monfort, and Trognon (1984) for fundamental results in this area. As we saw in Section 8.1, when the method of maximum likelihood is applied to a data set for which the DGP was not in fact a special case of the model being estimated, the resulting estimator is called a quasi-ML, or QML, estimator. In practice, of course, almost all the ML estimators we use are actually QML estimators, since some of the assumptions of our models are almost always wrong. It is therefore comforting that in certain common situations, including this one, the properties of QML estimators are very similar to those of genuine ML estimators, although asymptotic efficiency is of course lost.

As a concrete example of GLS, feasible GLS, and ML estimation, consider the model

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \ \ \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}), \ \ \Omega_{tt} = \sigma^2 w_t^{\alpha}, \ \ \Omega_{ts} = 0 \ \text{for all} \ t \neq s. \tag{9.33}$$

introduced in (9.52), as follows:

$$\boldsymbol{X}_i(\boldsymbol{\beta}) = \sum_{j=1}^{m} \boldsymbol{Z}_j(\boldsymbol{\beta})\psi_{ji}.$$

Then the stacked GNR is

$$\begin{bmatrix} (\boldsymbol{Y} - \boldsymbol{\xi}(\boldsymbol{\beta}))\boldsymbol{\psi}_1 \\ \vdots \\ (\boldsymbol{Y} - \boldsymbol{\xi}(\boldsymbol{\beta}))\boldsymbol{\psi}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1(\boldsymbol{\beta}) \\ \vdots \\ \boldsymbol{X}_m(\boldsymbol{\beta}) \end{bmatrix} \boldsymbol{b} + \text{residuals}. \tag{9.58}$$

The OLS estimates from the GNR (9.58) will be defined by the first-order conditions

$$\left(\sum_{i=1}^{m} \boldsymbol{X}_i^{\top}(\boldsymbol{\beta}) \boldsymbol{X}_i(\boldsymbol{\beta}) \right) \ddot{\boldsymbol{b}} = \sum_{i=1}^{m} \boldsymbol{X}_i^{\top}(\boldsymbol{\beta}) (\boldsymbol{Y} - \boldsymbol{\xi}(\boldsymbol{\beta})) \boldsymbol{\psi}_i. \tag{9.59}$$

Some manipulation of (9.59) based on the definition of the $\boldsymbol{X}_i$'s and of $\boldsymbol{\psi}$ shows that this is equivalent to

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \sigma^{ij} \boldsymbol{Z}_i^{\top}(\boldsymbol{\beta}) (\boldsymbol{y}_j - \boldsymbol{x}_j(\boldsymbol{\beta}) - \boldsymbol{Z}_j(\boldsymbol{\beta})\boldsymbol{b}) = \boldsymbol{0}. \tag{9.60}$$

Thus we see that regression (9.58) has all the properties we have come to expect from the Gauss-Newton regression. If we evaluate it at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$, the regression will have no explanatory power at all, because (9.60) is satisfied with $\boldsymbol{b} = \boldsymbol{0}$ by the first-order conditions (9.53). The estimated covariance matrix from regression (9.58) with $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ will be

$$\tilde{s}^2 \left(\sum_{i=1}^{m} \sum_{j=1}^{m} \sigma^{ij} \tilde{\boldsymbol{Z}}_i^{\top} \tilde{\boldsymbol{Z}}_j \right)^{-1}, \tag{9.61}$$

where $\tilde{s}^2$ is the estimate of the variance that the regression package will generate, which will evidently tend to 1 asymptotically if $\boldsymbol{\Sigma}$ is in fact the contemporaneous covariance matrix of $\boldsymbol{U}_t$. If (9.61) is rewritten as a sum of contributions from the successive observations, the result is

$$\tilde{s}^2 \left(\sum_{t=1}^{n} \tilde{\boldsymbol{\Xi}}_t \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\Xi}}_t^{\top} \right)^{-1},$$

from which it is clear that (9.61) is indeed the proper GNLS covariance matrix estimator.

the ML estimates $\hat{\boldsymbol{\beta}}$, the estimated error variance for it, $\hat{s}^2$, will be equal to

$$
\frac{1}{mn-k} \sum_{t=1}^{n} \big(\boldsymbol{Y}_t - \hat{\boldsymbol{\xi}}_t\big)\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^{\top}\big(\boldsymbol{Y}_t - \hat{\boldsymbol{\xi}}_t\big)^{\top}
$$
$$
= \frac{1}{mn-k} \sum_{t=1}^{n} \big(\boldsymbol{Y}_t - \hat{\boldsymbol{\xi}}_t\big)\hat{\boldsymbol{\Sigma}}^{-1}\big(\boldsymbol{Y}_t - \hat{\boldsymbol{\xi}}_t\big)^{\top} = \frac{mn}{mn-k}.
$$

(9.70)

The last equality here follows from an argument almost identical to the one used to establish (9.65). Since it is evident that (9.70) tends asymptotically to 1, expression (9.61), which is in this case

$$
\frac{mn}{mn-k}\left(\sum_{t=1}^{n} \hat{\boldsymbol{\Xi}}_t\,\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Xi}}_t^{\top}\right)^{-1},
$$

provides a natural and very convenient way to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$.

We have now established all the principal results of interest concerning the estimation of multivariate nonlinear regression models. Since those results have been in terms of a rather general and abstract model, it may help to make them more concrete if we indicate precisely how our general notation relates to the case of the linear expenditure system that we discussed earlier. For concreteness, we will assume that $m = 2$, which means that there is a total of three commodities. Then we see that

$$
\boldsymbol{Y}_t = [s_{t1} \quad s_{t2}];
$$

$$
\boldsymbol{\beta} = [\alpha_1 \;\vdots\; \alpha_2 \;\vdots\; \gamma_1 \;\vdots\; \gamma_2 \;\vdots\; \gamma_3];
$$

$$
\boldsymbol{\xi}_t(\boldsymbol{\beta}) = \left[\frac{\gamma_1 p_{1t}}{E_t} + \frac{\alpha_1}{E_t}\left(E_t - \sum_{j=1}^{3} p_{jt}\gamma_j\right) \quad \frac{\gamma_2 p_{2t}}{E_t} + \frac{\alpha_2}{E_t}\left(E_t - \sum_{j=1}^{3} p_{jt}\gamma_j\right)\right];
$$

$$
\boldsymbol{\Xi}_t(\boldsymbol{\beta}) = \begin{bmatrix} \big(E_t - \sum_{j=1}^{3} p_{jt}\gamma_j\big)/E_t & 0 \\ 0 & \big(E_t - \sum_{j=1}^{3} p_{jt}\gamma_j\big)/E_t \\ (1-\alpha_1)p_{1t}/E_t & -\alpha_2 p_{1t}/E_t \\ -\alpha_1 p_{2t}/E_t & (1-\alpha_2)p_{2t}/E_t \\ -\alpha_1 p_{3t}/E_t & -\alpha_2 p_{3t}/E_t \end{bmatrix}.
$$

It may be a useful exercise to set up the GNR for testing the hypothesis that $\gamma_1 = \gamma_2 = \gamma_3 = 0$, where estimates subject to that restriction have been obtained.

Our treatment of multivariate models has been relatively brief. A much fuller treatment, but only for linear SUR models, may be found in Srivastava and Giles (1987), which is also an excellent source for references to the econometric and statistical literature on the subject.

where the matrix $\boldsymbol{M}_D$ is simply the matrix that takes deviations from the group means $\bar{\boldsymbol{X}}_{.i}$ for $i = 1, \ldots, n$. Thus a typical element of $\boldsymbol{M}_D \boldsymbol{X}$ is

$$(\boldsymbol{M}_D \boldsymbol{X})_{ti} = \boldsymbol{X}_{ti} - \bar{\boldsymbol{X}}_{.i}.$$

This makes it easy to compute $\acute{\boldsymbol{\beta}}$ even when n is so large that it would be infeasible to run regression (9.74). One simply has to compute the group means $y_{.i}$ and $\boldsymbol{X}_{.i}$ for all i and then regress $y_{ti} - \bar{y}_{.i}$ on $\boldsymbol{X}_{ti} - \bar{\boldsymbol{X}}_{.i}$ for all t and i. The estimated covariance matrix should then be adjusted to reflect the fact that the number of degrees of freedom used in estimation is actually $n + k$ rather than k.

Because the fixed-effects estimator (9.75) depends only on the deviations of the regressand and regressors from their respective group means, it is sometimes called the **within-groups estimator**. As this name implies, it makes no use of the fact that the group means are in general different for different groups. This property of the estimator can be an advantage or a disadvantage, depending on circumstances. As we mentioned above, it may well be the case that the cross-sectional effects v_i are correlated with the regressors $\boldsymbol{X}_{ti}$ and consequently also with the respective group means of the regressors. In that event the OLS estimator (without fixed effects) based on the full sample would be inconsistent, but the within-groups estimator would remain consistent. However, if on the contrary the fixed effects *are* independent of the regressors, the within-groups estimator is not fully efficient. In the extreme case in which any one of the independent variables does not vary at all within groups, but only between groups, then the coefficient corresponding to that variable will not even be identifiable by the within-groups estimator.

An alternative inefficient estimator that uses only the variation among the group means is called the **between-groups estimator**. It may be written as

$$\dot{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{P}_D \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{P}_D \boldsymbol{y}. \tag{9.76}$$

Since $\boldsymbol{P}_D \boldsymbol{X}_{ti} = \bar{\boldsymbol{X}}_{.i}$, this estimator really involves only n distinct observations rather than nT. It will clearly be inconsistent if the cross-sectional effects, the v_i's, are correlated with the group means of the regressors, the $\bar{\boldsymbol{X}}_{.i}$'s. The OLS estimator can be written as a matrix-weighted average of the within-groups and between-groups estimators:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$
$$= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{M}_D \boldsymbol{y} + \boldsymbol{X}^\top \boldsymbol{P}_D \boldsymbol{y}\right)$$
$$= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{M}_D \boldsymbol{X} \acute{\boldsymbol{\beta}} + \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{P}_D \boldsymbol{X} \dot{\boldsymbol{\beta}}.$$

Thus we see immediately that OLS will be inconsistent whenever the between-groups estimator (9.76) is inconsistent.

where $\boldsymbol{\Omega}_0$ is the matrix $\boldsymbol{\Omega}(\rho)$ defined in (10.05), evaluated at ρ_0 and ω_0. Evidently, (10.07) will in general not be consistently estimated by the OLS covariance matrix estimator $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. Except in special cases, it is not possible to say whether the incorrect standard error estimates obtained using OLS will be larger or smaller than the correct ones obtained by taking the square roots of the diagonal elements of (10.07). However, analysis of special cases suggests that for values of ρ greater than 0 (the most commonly encountered case) the incorrect OLS standard errors are usually too small; see, among others, Nicholls and Pagan (1977), Sathe and Vinod (1974), and Vinod (1976).

Expression (10.07) applies to any situation in which OLS is incorrectly used in place of GLS and not merely to situations in which the errors follow an AR(1) process. So does the previous result that $\hat{\boldsymbol{\beta}}$ is unbiased if $\boldsymbol{X}$ is fixed and $E(\boldsymbol{X}^\top\boldsymbol{u}) = \boldsymbol{0}$. But recall from Section 9.5 that, even when these conditions are satisfied, $\hat{\boldsymbol{\beta}}$ may fail to be consistent if the errors are correlated enough among themselves. We may conclude that, when the regressors are fixed and the covariance matrix of the error terms is such that there is not too much correlation of the error terms, the OLS estimates will be consistent, but the OLS covariance matrix estimate will not be. A consistent estimate of the covariance matrix of the OLS estimator can usually be found. However, since the proof of the Gauss-Markov Theorem depended on the assumption that $E(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2\boldsymbol{I}$, OLS is not the best linear unbiased estimator when this assumption does not hold.

The preceding discussion assumed that there were no lagged dependent variables among the columns of $\boldsymbol{X}$. When this assumption is dropped, the results change drastically, and OLS is seen to be both biased and inconsistent. The simplest way to see this is to think about an element of $\boldsymbol{X}^\top\boldsymbol{u}$ corresponding to the lagged dependent variable (or to one of the lagged dependent variables if there is more than one). If the dependent variable is lagged j periods, this element is

$$\sum_{t=1}^{n} y_{t-j}u_t. \tag{10.08}$$

Now recall expression (10.03), in which we expressed u_t as a function of u_{t-j} and of all the innovations between periods $t - j + 1$ and t. Since y_{t-j} is equal to $\boldsymbol{X}_{t-j}\boldsymbol{\beta} + u_{t-j}$, it is clear from (10.03) that (10.08) cannot possibly have expectation zero. Thus we conclude that when $\boldsymbol{X}$ includes lagged dependent variables and u_t is serially correlated,

$$\plim_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{u}\right) \neq \boldsymbol{0}, \tag{10.09}$$

which implies that

$$\plim_{n\to\infty}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \plim_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\plim_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{u}\right) \neq \boldsymbol{0}. \tag{10.10}$$

the reported covariance matrix from the Cochrane-Orcutt and Hildreth-Lu procedures will thus be invalid in many cases. One either has to calculate the GNR (10.29) oneself or use nonlinear least squares from the beginning so that the regression package will do so.

When the conditional covariance matrix estimate is invalid, reported standard errors are always too small (asymptotically). In fact, the covariance matrix estimate produced by the GNR (10.29) for the estimates of $\boldsymbol{\beta}$ differs from that produced by (10.27) by a positive definite matrix, if we ignore the fact that the degrees of freedom are different. To see this, notice that the Gauss-Newton regression (10.29) has the same regressors as (10.27), plus one additional regressor, $\hat{\boldsymbol{u}}_{-1}$. If we apply the FWL Theorem to (10.29), we see that the covariance matrix estimate from it is the same as that from a regression in which all the variables are projected onto the orthogonal complement of $\hat{\boldsymbol{u}}_{-1}$. The residuals are unchanged by the projection and so are identical to those of (10.27), as we saw above. The difference between the covariance matrix estimates for $\hat{\boldsymbol{\beta}}$ from (10.29) and (10.27) is therefore proportional to

$$\left(\boldsymbol{X}^{*\top}(\hat{\rho})\boldsymbol{M}_{\hat{u}_{-1}}\boldsymbol{X}^{*}(\hat{\rho})\right)^{-1} - \left(\boldsymbol{X}^{*\top}(\hat{\rho})\boldsymbol{X}^{*}(\hat{\rho})\right)^{-1}, \tag{10.32}$$

except for an asymptotically negligible effect due to the different degrees-of-freedom factors. If we subtract the inverses of the two matrices in (10.32) in the opposite order, we obtain

$$\boldsymbol{X}^{*\top}(\hat{\rho})\boldsymbol{P}_{\hat{u}_{-1}}\boldsymbol{X}^{*}(\hat{\rho}),$$

which is evidently positive semidefinite. It then follows from a result proved in Appendix A that (10.32) is itself positive semidefinite. If $\hat{\boldsymbol{u}}_{-1}$ is substantially correlated with the columns of $\boldsymbol{X}^{*}(\hat{\rho})$, the incorrect variance estimate from regression (10.27) may be much smaller than the correct variance estimate from the GNR (10.29).

The Gauss-Newton regressions (10.26) and (10.29) yield estimated standard errors for $\hat{\rho}$ as well as for $\hat{\boldsymbol{\beta}}$. If the covariance matrix is asymptotically block-diagonal between ρ and $\boldsymbol{\beta}$, we see from (10.31) that the asymptotic variance of $n^{1/2}(\hat{\rho} - \rho_0)$ will be equal to

$$\omega^2 \plim_{n \to \infty} \left(\frac{\hat{\boldsymbol{u}}_{-1}^{\top}\hat{\boldsymbol{u}}_{-1}}{n-1}\right)^{-1} = \omega^2 \left(\frac{1-\rho_0^2}{\omega^2}\right) = 1 - \rho_0^2. \tag{10.33}$$

Thus, in this special case, the variance of $\hat{\rho}$ can be estimated by

$$\frac{1-\hat{\rho}^2}{n-1}. \tag{10.34}$$

It may seem puzzling that neither the asymptotic variance $1 - \rho_0^2$ nor the estimate (10.34) depends on ω^2. After all, we normally expect the variance

where L denotes the **lag operator**. The lag operator L has the property that when L multiplies anything with a time subscript, this subscript is lagged one period. Thus

$$Lu_t = u_{t-1}, \quad L^2 u_t = u_{t-2}, \quad L^p u_t = u_{t-p},$$

and so on. The expression in parentheses in (10.36) is a polynomial in the lag operator L, with coefficients 1 and $-\rho_1, \ldots, -\rho_p$. If we define $A(L, \boldsymbol{\rho})$ as being equal to this polynomial, $\boldsymbol{\rho}$ representing the vector $[\rho_1 \vdots \rho_2 \vdots \cdots \vdots \rho_p]$, we can write (10.36) even more compactly as

$$A(L, \boldsymbol{\rho}) u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \tag{10.37}$$

For the same reasons that we wish to impose the condition $|\rho_1| < 1$ on AR(1) processes so as to ensure that they are stationary, we would like to impose stationarity conditions on general AR(p) processes. The stationarity condition for such processes may be expressed in several ways; one of them is that all the roots of the polynomial equation in z,

$$A(z, \boldsymbol{\rho}) \equiv 1 - \rho_1 z - \rho_2 z^2 - \cdots - \rho_p z^p = 0 \tag{10.38}$$

must lie **outside the unit circle**, which simply means that all of the roots of (10.38) must be greater than 1 in absolute value. This condition can lead to quite complicated restrictions on $\boldsymbol{\rho}$ for general AR(p) processes.

It rarely makes sense to specify a high-order AR(p) process (i.e., one with p a large number) when trying to model the error terms associated with a regression model. The AR(2) process is much more flexible, but also much more complicated, than the AR(1) process; it is often all that is needed when the latter is too restrictive. The additional complexity of the AR(2) process is easily seen. For example, the variance of u_t, assuming stationarity, is

$$\sigma^2 = \frac{1 - \rho_2}{1 + \rho_2} \times \frac{\omega^2}{(1 - \rho_2)^2 - \rho_1^2},$$

which is substantially more complicated than the corresponding expression (10.02) for the AR(1) case, and stationarity now requires that three conditions hold:

$$\rho_1 + \rho_2 < 1; \quad \rho_2 - \rho_1 < 1; \quad \rho_2 > -1. \tag{10.39}$$

Conditions (10.39) define a **stationarity triangle**. This triangle has vertices at $(-2, -1)$, $(2, -1)$, and $(0, 1)$. Provided that the point (ρ_1, ρ_2) lies within the triangle, the AR(2) process will be stationary.

Autoregressive processes of order higher than 2 arise quite frequently with time-series data that exhibit seasonal variation. It is not uncommon, for example, for error terms in models estimated using quarterly data apparently to follow the **simple AR(4) process**

$$u_t = \rho_4 u_{t-4} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \tag{10.40}$$

where $\hat{\boldsymbol{X}}^*$ denotes the $n \times k$ matrix of the derivatives of the vector of nonlinear functions $\boldsymbol{x}^*(\boldsymbol{\beta}, \rho)$, defined in (10.46), with respect to the elements of $\boldsymbol{\beta}$, evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\rho})$, and

$$\hat{\boldsymbol{V}}(\hat{\rho}, \hat{\omega}) = \begin{bmatrix} \dfrac{n}{1 - \hat{\rho}^2} + \dfrac{3\hat{\rho}^2 - 1}{(1 - \hat{\rho}^2)^2} & \dfrac{2\hat{\rho}}{\hat{\omega}(1 - \hat{\rho}^2)} \\[2ex] \dfrac{2\hat{\rho}}{\hat{\omega}(1 - \hat{\rho}^2)} & \dfrac{2n}{\hat{\omega}^2} \end{bmatrix}^{-1} .$$

The estimated covariance matrix (10.54) is block-diagonal between $\boldsymbol{\beta}$ and ρ and between $\boldsymbol{\beta}$ and ω (recall that we have ruled out lagged dependent variables). However, unlike the situation with regression models, it is not block-diagonal between ρ and ω. The off-diagonal terms in the (ρ, ω) block of the information matrix are $O(1)$, while the diagonal terms are $O(n)$. Thus $\boldsymbol{V}(\hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\omega})$ will be asymptotically block-diagonal between $\boldsymbol{\beta}$, ρ, and ω. This is what we would expect, since it is only the first observation, which is asymptotically negligible, that prevents (10.54) from being block-diagonal in the first place.

It is an excellent exercise to derive the estimated covariance matrix (10.54). One starts by taking the second derivatives of (10.51) with respect to all of the parameters of the model to find the Hessian, then takes expectations of minus it to obtain the information matrix. One then replaces parameters by their ML estimates and inverts the information matrix to obtain (10.54). Although this exercise is straightforward, there are plenty of opportunities to make mistakes. For example, Beach and MacKinnon (1978a) fail to take all possible expectations and, as a result, end up with an excessively complicated estimated covariance matrix.

The preceding discussion makes it clear that taking the first observation into account is significantly harder than ignoring it. Even if an appropriate computer program is available, so that estimation is straightforward, one runs into trouble when one wants to test the model. Since the transformed model is no longer a regression model, the Gauss-Newton regression no longer applies and cannot be used to do model specification tests; see Sections 10.8 and 10.9. One could of course estimate the model twice, once taking account of the first observation, in order to obtain the most efficient possible estimates, and once dropping it, in order to be able to test the specification, but this clearly involves some extra work. The obvious question that arises, then, is whether the additional trouble of taking the first observation into account is worth it.

There is a large literature on this subject, including Kadiyala (1968), Rao and Griliches (1969), Maeshiro (1976, 1979), Beach and MacKinnon (1978a), Chipman (1979), Spitzer (1979), Park and Mitchell (1980), Ansley and Newbold (1980), Poirier (1978a), Magee (1987), and Thornton (1987). In many cases, retaining the first observation yields more efficient estimates but not by very much. However, when the sample size is modest and there is one or

in month t would affect the value of instruments maturing in months t, $t+1$, and $t+2$ but would not directly affect the value of instruments maturing later, because the latter would not yet have been issued. This suggests that the error term should be modeled by an MA(2) process; see Frankel (1980) and Hansen and Hodrick (1980). Moving average errors also arise when data are gathered using a survey that includes some of the same respondents in consecutive periods, such as the labor force surveys in both the United States and Canada, which are used to estimate unemployment rates; see Hausman and Watson (1985).

It is generally somewhat harder to estimate regression models with moving average errors than to estimate models with autoregressive errors. To see why, suppose that we want to estimate the model

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t = \varepsilon_t - \alpha\varepsilon_{t-1}, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \tag{10.61}$$

Compared with (10.57), we have dropped the subscript from α and changed its sign for convenience; the sign change is of course purely a normalization. Let us make the asymptotically innocuous assumption that the unobserved innovation ε_0 is equal to zero (techniques that do not make this assumption will be discussed below). Then we see that

$$
\begin{aligned}
y_1 &= x_1(\boldsymbol{\beta}) + \varepsilon_1 \\
y_2 &= x_2(\boldsymbol{\beta}) - \alpha\big(y_1 - x_1(\boldsymbol{\beta})\big) + \varepsilon_2 \\
y_3 &= x_3(\boldsymbol{\beta}) - \alpha\big(y_2 - x_2(\boldsymbol{\beta})\big) - \alpha^2\big(y_1 - x_1(\boldsymbol{\beta})\big) + \varepsilon_3,
\end{aligned}
\tag{10.62}
$$

and so on. By making the definitions

$$
\begin{aligned}
y_0^* &= 0; \quad y_t^* = y_t + \alpha y_{t-1}^*, \ \ t = 1, \ldots, n; \\
x_0^* &= 0; \quad x_t^*(\boldsymbol{\beta}, \alpha) = x_t(\boldsymbol{\beta}) + \alpha x_{t-1}^*(\boldsymbol{\beta}, \alpha), \ \ t = 1, \ldots, n,
\end{aligned}
\tag{10.63}
$$

we can write equations (10.62) in the form

$$y_t = -\alpha y_{t-1}^* + x_t^*(\boldsymbol{\beta}, \alpha) + \varepsilon_t, \tag{10.64}$$

which makes it clear that we have a nonlinear regression model. But the regression function depends on the entire sample up to period t, since y_{t-1}^* depends on all previous values of y_t and x_t^* depends on $x_{t-i}(\boldsymbol{\beta})$ for all $i \geq 0$. In the by no means unlikely case in which $|\alpha| = 1$, the dependence of y_t on past values does not even tend to diminish as those values recede into the distant past. If we have a specialized program for estimation with MA(1) errors, or a smart nonlinear least squares program that allows us to define the regression function recursively, as in (10.63), estimating (10.64) need not be any more difficult than estimating other nonlinear regression models. But if appropriate software is lacking, this estimation can be quite difficult.

If we assume that the error terms are normally distributed, the model (10.61) becomes

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t = \varepsilon_t - \alpha\varepsilon_{t-1}, \quad \varepsilon_t \sim \text{NID}(0, \omega^2). \qquad (10.65)$$

We previously made the asymptotically innocuous assumption that the unobserved innovation ε_0 is equal to zero. Although asymptotically innocuous, that assumption is clearly false, since according to (10.65) ε_0 must be distributed as $N(0, \omega^2)$. The simplest way to take proper account of this fact was suggested by MacDonald and MacKinnon (1985); our treatment follows theirs.

The concentrated loglikelihood function for the model (10.65) is

$$C - \tfrac{n}{2}\log\Big(\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^{\top}\boldsymbol{\Delta}^{-1}(\alpha)\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)\Big) - \tfrac{1}{2}\log|\boldsymbol{\Delta}(\alpha)|, \qquad (10.66)$$

where $\omega^2\boldsymbol{\Delta}(\alpha)$ is the covariance matrix of the vector of error terms $\boldsymbol{u}$, expression (10.60).[4] As discussed by Box and Jenkins (1976) and others, the Jacobian term $-\tfrac{1}{2}\log|\boldsymbol{\Delta}(\alpha)|$ is

$$\tfrac{1}{2}\log\big(1 - \alpha^2\big) - \tfrac{1}{2}\log\big(1 - \alpha^{2n+2}\big). \qquad (10.67)$$

When $|\alpha| = 1$, both terms in (10.67) are undefined. In that case, by using l'Hôpital's Rule, one can show that

$$\lim_{|\alpha|\to 1}\Big(\tfrac{1}{2}\log\big(1 - \alpha^2\big) - \tfrac{1}{2}\log\big(1 - \alpha^{2n+2}\big)\Big) = -\tfrac{1}{2}\log(n + 1).$$

This result allows the loglikelihood function (10.66) to be evaluated for any value of α in the invertibility region $-1 \leq \alpha \leq 1$.

It is important to be able to deal with the case in which $|\alpha| = 1$, since in practice one not infrequently obtains ML estimates with $|\hat{\alpha}| = 1$, especially when the sample size is small; see, for example, Osborn (1976) and Davidson (1981). The reason for this is that if we concentrate the loglikelihood function with respect to $\boldsymbol{\beta}$ and ω to obtain $\ell^c(\alpha)$, we will find that $\ell^c(\alpha)$ has the same value for α and $1/\alpha$. That, of course, is the reason for imposing the invertibility condition that $|\alpha| \leq 1$. Thus, if $\ell^c(\alpha)$ is rising as $\alpha \to 1$ or as $\alpha \to -1$, it must have a maximum precisely at $\alpha = 1$ or $\alpha = -1$. This is a distinctly undesirable feature of the model (10.65). When $|\hat{\alpha}| = 1$, one cannot make inferences about α in the usual way, since $\hat{\alpha}$ is then on the boundary of the parameter space. Since $\hat{\alpha}$ can equal ± 1 with finite probability,

---

[4] In fact, expression (10.66) could be the concentrated loglikelihood function for a nonlinear regression model with error terms that follow any sort of **autoregressive moving average**, or **ARMA**, process, provided that $\boldsymbol{\Delta}(\alpha)$ were replaced by the covariance matrix for $\boldsymbol{u}$ implied by that ARMA process.

using the normal distribution to approximate its finite-sample distribution is a somewhat dubious procedure. Thus, if $\hat{\alpha}$ is equal to or even close to 1 in absolute value, the investigator should exercise care in making inferences about α. Of course, as $n \to \infty$ the fact that $\hat{\alpha}$ is consistent means that the number of times that $|\hat{\alpha}| = 1$ tends to zero, unless $|\alpha_0| = 1$.

It is not easy to evaluate (10.66) directly; see Pesaran (1973), Osborn (1976), and Balestra (1980), among others.[5] We therefore use a trick that provides an alternative way to do so. Recall equations (10.62), in which we explicitly wrote $y_1, \ldots, y_n$ as functions of current and lagged values of $x_t(\boldsymbol{\beta})$ and lagged values of y_t. We may rewrite these equations, taking account of observation zero, as

$$
\begin{aligned}
0 &= -\upsilon + \varepsilon_0 \\
y_1 &= x_1(\boldsymbol{\beta}) - \alpha\upsilon + \varepsilon_1 \\
y_2 &= x_2(\boldsymbol{\beta}) - \alpha\big(y_1 - x_1(\boldsymbol{\beta})\big) - \alpha^2\upsilon + \varepsilon_2 \\
y_3 &= x_3(\boldsymbol{\beta}) - \alpha\big(y_2 - x_2(\boldsymbol{\beta})\big) - \alpha^2\big(y_1 - x_1(\boldsymbol{\beta})\big) - \alpha^3\upsilon + \varepsilon_3,
\end{aligned}
\tag{10.68}
$$

and so on. Here we have added both one observation and one parameter to equations (10.62). The extra observation is observation zero, which as written here simply says that the unknown parameter υ is *defined* to equal the error term ε_0. This unknown parameter also appears in all subsequent observations, multiplied by larger and larger powers of α, to reflect the dependence of y_t for all observations on ε_0. Notice that because we have added both an extra parameter and an extra observation, we have not changed the number of degrees of freedom (i.e., the number of observations minus the number of parameters estimated) at all.

If we make the definitions

$$
\begin{aligned}
y_0^* &= 0; \quad y_t^* = y_t + \alpha y_{t-1}^*, \ \ t = 1, \ldots, n; \\
x_0^* &= 0; \quad x_t^*(\boldsymbol{\beta}, \alpha) = x_t(\boldsymbol{\beta}) + \alpha x_{t-1}^*(\boldsymbol{\beta}, \alpha), \ \ t = 1, \ldots, n; \\
z_0^* &= -1; \quad z_t^* = \alpha z_{t-1}^*,
\end{aligned}
$$

we can write equations (10.68) in the form

$$
y_t^*(\alpha) = x_t^*(\boldsymbol{\beta}, \alpha) + \upsilon z_t^* + \varepsilon_t,
\tag{10.69}
$$

making them look like very much like a nonlinear regression model. The sum of squared residuals would then be

$$
\sum_{t=0}^{n} \big(y_t^*(\alpha) - x_t^*(\boldsymbol{\beta}, \alpha) - \upsilon z_t^*\big)^2.
\tag{10.70}
$$

---

[5] Another approach to the estimation of models with moving average errors has been proposed by Harvey and Phillips (1979) and by Gardner, Harvey, and Phillips (1980). It requires specialized software.

to the one for testing against AR(q) errors. Perhaps more surprisingly, the same artificial regression also turns out to be appropriate for testing against ARMA(p, q) errors, with $\max(p, q)$ lags of $\tilde{u}$ now being included in the regression. For more details, see Godfrey (1978b, 1988).

Using something very like the Gauss-Newton regression to test for serial correlation was first suggested by Durbin (1970) in a paper that also introduced what has become known as **Durbin's h test**. The latter procedure, which we will not discuss in detail, is an asymptotic test for AR(1) errors that can be used when the null hypothesis is a linear regression model which includes the dependent variable lagged once, and possibly more than once as well, among the regressors. The h test can be calculated with a hand calculator from the output for the original regression printed by most regression packages, although in some cases it cannot be calculated at all because it would be necessary to compute the square root of a negative number. For reasons that today seem hard to understand (but are presumably related to the primitive state of computer hardware and econometric software in the early 1970s), Durbin's h test became widely used, while his so-called **alternative procedure**, a t test based on the modified GNR (10.77), was all but ignored for quite some time.[8] It was finally rediscovered and extended by Breusch (1978) and Godfrey (1978a, 1978b). All of these papers assumed that the error terms ε_t were normally distributed, and they developed tests based on the GNR as Lagrange multiplier tests based on maximum likelihood estimation. The normality assumption is of course completely unnecessary.

Equally unnecessary is any assumption about the presence or absence of lagged dependent variables in the regression function $x_t(\boldsymbol{\beta})$. All we require is that this function satisfy the regularity conditions of Chapter 5, in order that nonlinear least squares estimates will be consistent and asymptotically normal under both the null and alternative hypotheses. As the above history implies, and as we will discuss below, many tests for serial correlation require that $x_t(\boldsymbol{\beta})$ not depend on lagged dependent variables, and all of the literature cited in the previous paragraph was written with the specific aim of handling the case in which $x_t(\boldsymbol{\beta})$ is linear and depends on one or more lagged values of the dependent variable.

The problem with tests based on the GNR is that they are valid only asymptotically. This is true whether or not $x_t(\boldsymbol{\beta})$ is linear, because $\tilde{\boldsymbol{u}}_{-1}$ is only an estimate of $\boldsymbol{u}_{-1}$. Indeed, as we saw in Section 5.6, $\tilde{\boldsymbol{u}} \overset{a}{=} \boldsymbol{M}_0 \boldsymbol{u}$, where $\boldsymbol{M}_0 \equiv \mathbf{I} - \boldsymbol{X}_0(\boldsymbol{X}_0^{\top}\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^{\top}$ and $\boldsymbol{X}_0 \equiv \boldsymbol{X}(\boldsymbol{\beta}_0)$. This is just the asymptotic equality (5.57). The asymptotic equality is replaced by an exact equality if $\boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$.

---

[8] Maddala and Rao (1973), Spencer (1975), and Inder (1984), among others, have provided Monte Carlo evidence on Durbin's h test as compared with the test based on the GNR. This evidence does not suggest any strong reason to prefer one test over the other. Thus the greater convenience and more general applicability of the test based on the GNR are probably the main factors in its favor.

underlying regression model is linear, and $\boldsymbol{X}$ contains only fixed regressors. This distribution necessarily depends on $\boldsymbol{X}$. The calculation uses the fact that the d statistic can be written as

$$\frac{\boldsymbol{u}^{\top} \boldsymbol{M}_X \boldsymbol{A} \boldsymbol{M}_X \boldsymbol{u}}{\boldsymbol{u}^{\top} \boldsymbol{M}_X \boldsymbol{u}}, \tag{10.82}$$

where $\boldsymbol{A}$ is the $n \times n$ matrix

$$\begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

From (10.82), the d statistic is seen to be a ratio of quadratic forms in normally distributed random variables, and the distributions of such ratios can be evaluated using several numerical techniques; see Durbin and Watson (1971) and Savin and White (1977) for references.

Most applied workers never attempt to calculate the exact distribution of the d statistic corresponding to their particular $\boldsymbol{X}$ matrix. Instead, they use the fact that the critical values for its distribution are known to fall between two bounding values, d_L and d_U, which depend on the sample size, n, the number of regressors, k, and whether or not there is a constant term. Tables of d_L and d_U may be found in some econometrics textbooks and in papers such as Durbin and Watson (1951) and Savin and White (1977). As an example, when $n = 50$ and $k = 6$ (counting the constant term as one of the regressors), for a test against $\rho > 0$ at the .05 level, $d_L = 1.335$ and $d_U = 1.771$. Thus, if one calculated a d statistic for this sample size and number of regressors and it was less than 1.335, one could confidently decide to reject the null hypothesis of no serial correlation at the .05 level. If the statistic was greater than 1.771, one could confidently decide not to reject. However, if the statistic was in the "inconclusive region" between 1.335 and 1.771, one would be unsure of whether to reject or not. When the sample size is small, and especially when it is small relative to the number of regressors, the inconclusive region can be very large. This means that the d statistic may not be very informative when used in conjunction with the tables of d_L and d_U.[9] In such cases, one may have no choice but to calculate the exact distribution of the statistic, if one wants to make inferences from the d statistic in a small sample. A few software packages, such as SHAZAM, allow one to do this. Of course,

---

[9] There is reason to believe that when the regressors are slowly changing, a situation which may often be the case with time-series data, d_U provides a better approximation than d_L. See Hannan and Terrell (1966).

be less than $2k+1$. The easiest way to see why this will almost always be the case is to consider an example.

Suppose that the regression function $x_t(\boldsymbol{\beta})$ for the original H_0 model is

$$\beta_0 + \beta_1 z_t + \beta_2 t + \beta_3 z_{t-1} + \beta_4 y_{t-1}, \tag{10.93}$$

where z_t is the tth observation on an economic time series, and t is the tth observation on a linear time trend. The regression function for the unrestricted H_2 model which corresponds to (10.93) is

$$\begin{aligned}
\beta_0 + \beta_1 z_t + \beta_2 t + \beta_3 z_{t-1} + \beta_4 y_{t-1} + \rho y_{t-1} \\
+ \gamma_0 + \gamma_1 z_{t-1} + \gamma_2(t-1) + \gamma_3 z_{t-2} + \gamma_4 y_{t-2}.
\end{aligned} \tag{10.94}$$

This regression function appears to have 11 parameters, but 4 of them are in fact unidentifiable. It is obvious that we cannot estimate both β_0 and γ_0, since there cannot be two constant terms. Similarly, we cannot estimate both β_3 and γ_1, since there cannot be two coefficients on z_{t-1}, and we cannot estimate both β_4 and ρ, since there cannot be two coefficients on y_{t-1}. We also cannot estimate γ_2 along with β_2 and the constant, because t, $t-1$ and the constant term are perfectly collinear, since $t - (t-1) = 1$. Thus the version of H_2 that can actually be estimated has the regression function

$$\delta_0 + \beta_1 z_t + \delta_1 t + \delta_2 z_{t-1} + \delta_3 y_{t-1} + \gamma_3 z_{t-2} + \gamma_4 y_{t-2}, \tag{10.95}$$

where

$$\delta_0 = \beta_0 + \gamma_0 - \gamma_2; \quad \delta_1 = \beta_2 + \gamma_2; \quad \delta_2 = \beta_3 + \gamma_1; \quad \text{and } \delta_3 = \rho + \beta_4.$$

We see that (10.95) has seven identifiable parameters: β_1, γ_3, γ_4, and δ_0 through δ_3, instead of the eleven parameters, many of them not identifiable, of (10.94). The regression function for the restricted model, H_1, is

$$\begin{aligned}
\beta_0 + \beta_1 z_t + \beta_2 t + \beta_3 z_{t-1} + \beta_4 y_{t-1} + \rho y_{t-1} \\
- \rho \beta_0 - \rho \beta_1 z_{t-1} - \rho \beta_2 (t-1) - \rho \beta_3 z_{t-2} - \rho \beta_4 y_{t-2},
\end{aligned}$$

and it has six parameters, ρ and β_0 through β_4. Thus, in this case, l, the number of restrictions that H_1 imposes on H_2, is just 1.

While this is a slightly extreme example, similar problems arise in almost every attempt to test common factor restrictions. Constant terms, many types of dummy variables (notably seasonal dummies and time trends), lagged dependent variables, and independent variables that appear with more than one time subscript almost always result in an unrestricted model H_2 of which not all parameters will be identifiable. Luckily, it is very easy to deal with these problems when one does an F test; one simply has to omit the redundant regressors when estimating H_2. One can then calculate l as the number of

parameters in H_2 minus the number in H_1, which is $k + 1$. Since many regression packages automatically drop redundant regressors, one naive but often effective approach is simply to attempt to estimate H_2 in something close to its original form and then to count the number of parameters that the regression package is actually able to estimate.

The F test (10.92) is not the only way to test common factor restrictions. Since the regression function for H_2 is linear in all parameters, while the one for H_1 is nonlinear, it is natural to try to base tests on the OLS estimates of H_2 alone. One approach to this problem is discussed by Sargan (1980a), but it is quite complicated and requires specialized computer software. A simpler approach is to use a one-step estimator of H_1. Consistent estimates of the parameters of H_1 may be obtained from the estimates of H_2, as discussed in Section 10.3, and the GNR (10.19) is then used to obtain one-step estimates. These estimates themselves are not necessarily of interest. All that is needed is the sum of squared residuals from the GNR, which may be used in place of SSR$_1$ in the formula (10.92) for the F test. However, since it is generally neither difficult nor expensive to estimate H_1 with modern computers and software packages, situations in which there is a significant advantage from the use of this one-step procedure are likely to be rare.

Something very like a test of common factor restrictions can be employed even when the original (H_0) model is nonlinear. In this case, the H_1 model can be written as

$$(1 - \rho L)y_t = (1 - \rho L)x_t(\boldsymbol{\beta}) + \varepsilon_t. \tag{10.96}$$

A version of (10.96) in which the common factor restriction does not hold is

$$(1 - \rho L)y_t = (1 - \delta L)x_t(\boldsymbol{\beta}) + \varepsilon_t. \tag{10.97}$$

Evidently, (10.96) is just (10.97) subject to the restriction that $\delta = \rho$. This restriction can be tested by a Gauss-Newton regression in the usual way. This GNR is

$$\boldsymbol{y} - \hat{\boldsymbol{x}} - \hat{\rho}(\boldsymbol{y}_{-1} - \hat{\boldsymbol{x}}_{-1}) = (\hat{\boldsymbol{X}} - \hat{\rho}\hat{\boldsymbol{X}}_{-1})\boldsymbol{b}$$
$$+ r(\boldsymbol{y}_{-1} - \hat{\boldsymbol{x}}_{-1}) + d\hat{\boldsymbol{x}}_{-1} + \text{residuals}, \tag{10.98}$$

where $\hat{\rho}$ and $\hat{\boldsymbol{\beta}}$ are the NLS estimates of H_1, and $\hat{\boldsymbol{x}} \equiv \boldsymbol{x}(\hat{\boldsymbol{\beta}})$. Regression (10.98) looks exactly like the GNR (10.26), which we used to calculate the covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\rho}$, with the addition of the extra regressor $\hat{\boldsymbol{x}}_{-1}$, the coefficient of which is d. The t statistic for $d = 0$ will be an asymptotically valid test statistic.

Notice that this GNR could be used even if $x_t(\boldsymbol{\beta})$ were a linear function. Since this variant of the common factor restrictions test necessarily has only one degree of freedom, it would not be the same as the usual form of the test, discussed above, for any model with $l > 1$. The difference arises because

the test based on (10.98) is testing against a less general alternative than the usual form of the test. When $x_t(\boldsymbol{\beta})$ is linear, (10.97) can be written as

$$(1 - \rho L)y_t = \boldsymbol{X}_t\boldsymbol{\beta} - \delta \boldsymbol{X}_{t-1}\boldsymbol{\beta} + \varepsilon_t, \tag{10.99}$$

which is in general (but not when $l = 1$) more restrictive than equation (10.89). Thus consideration of the nonlinear regression case reveals that there are really two different tests of common factor restrictions when the original model is linear. The first, which tests (10.88) against (10.89), is the F test (10.92). It will have l degrees of freedom, where $1 \leq l \leq k$. The second, which tests (10.88) against (10.99), is the t test of $d = 0$ in the Gauss-Newton regression (10.98). It will always have one degree of freedom. Either test might perform better than the other, depending on how the data were actually generated; see Chapter 12. When $l = 1$, the two tests will coincide, a fact that it may be a good exercise to demonstrate.

10.10 Instrumental Variables and Serial Correlation

So far in this chapter, we have assumed that the regression function $\boldsymbol{x}(\boldsymbol{\beta})$ depends only on exogenous and predetermined variables. However, there is no reason for serially correlated errors not to occur in models for which current endogenous variables appear in the regression function. As we discussed in Chapter 7, the technique of instrumental variables (IV) estimation is commonly used to obtain consistent estimates for such models. In this section, we briefly discuss how IV methods can be used to estimate univariate regression models with errors that are serially correlated and to test for serial correlation in such models.

Suppose that we wish to estimate the model (10.12) by instrumental variables. Then, as we saw in Section 7.6, the IV estimates may be obtained by minimizing, with respect to $\boldsymbol{\beta}$ and ρ, the criterion function

$$\big(\boldsymbol{y} - \boldsymbol{x}'(\boldsymbol{\beta}, \rho)\big)^{\top}\boldsymbol{P}_W\big(\boldsymbol{y} - \boldsymbol{x}'(\boldsymbol{\beta}, \rho)\big), \tag{10.100}$$

where the regression function $\boldsymbol{x}'(\boldsymbol{\beta}, \rho)$ is defined by (10.13), and $\boldsymbol{P}_W$ is the matrix that projects orthogonally onto the space spanned by $\boldsymbol{W}$, a suitable matrix of instruments. The IV form of the Gauss-Newton regression can be used as the basis for an algorithm to minimize (10.100). Given suitable regularity conditions on $x_t(\boldsymbol{\beta})$, and assuming that $|\rho| < 1$, these estimates will be consistent and asymptotically normal. See Sargan (1959) for a full treatment of the case in which $\boldsymbol{x}(\boldsymbol{\beta})$ is linear.

The only potential difficulty with this IV procedure is that one has to find a "suitable" matrix of instruments $\boldsymbol{W}$. For asymptotic efficiency, one always wants the instruments to include all the exogenous and predetermined variables that appear in the regression function. From (10.13), we see that more

such variables appear in the regression function $x_t'(\boldsymbol{\beta}, \rho)$ for the transformed model than in the original regression function $x_t(\boldsymbol{\beta})$. Thus the optimal choice of instruments may differ according to whether one takes account of serial correlation or assumes that it is absent.

To make this point more clearly, let us assume that the original model is linear, with regression function

$$x_t(\boldsymbol{\beta}) = \boldsymbol{Z}_t \boldsymbol{\beta}_1 + \boldsymbol{Y}_t \boldsymbol{\beta}_2, \tag{10.101}$$

where $\boldsymbol{Z}_t$ is a row vector of explanatory variables that are exogenous or predetermined, and $\boldsymbol{Y}_t$ is a row vector of current endogenous variables; the dimension of $\boldsymbol{\beta} \equiv [\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2]$ is k. The regression function for the transformed model is then

$$x_t'(\boldsymbol{\beta}, \rho) = \rho y_{t-1} + \boldsymbol{Z}_t \boldsymbol{\beta}_1 + \boldsymbol{Y}_t \boldsymbol{\beta}_2 - \rho \boldsymbol{Z}_{t-1} \boldsymbol{\beta}_1 - \rho \boldsymbol{Y}_{t-1} \boldsymbol{\beta}_2. \tag{10.102}$$

In (10.101), the only exogenous or predetermined variables were the variables in $\boldsymbol{Z}_t$. In (10.102), however, they are y_{t-1} and the variables in $\boldsymbol{Z}_t$, $\boldsymbol{Z}_{t-1}$, and $\boldsymbol{Y}_{-1}$ (the same variables may occur in more than one of these, of course; see the discussion of common factor restrictions in the previous section). All these variables would normally be included in the matrix of instruments $\boldsymbol{W}$. Since the number of these variables is almost certain to be greater than $k+1$, it would not normally be necessary to include any additional instruments to ensure that all parameters are identified.

For more discussion of the estimation of single linear equations with serially correlated errors and current endogenous regressors, see Sargan (1959, 1961), Amemiya (1966), Fair (1970), Dhrymes, Berner, and Cummins (1974), Hatanaka (1976), and Bowden and Turkington (1984).

Testing for serial correlation in models estimated by IV is straightforward if one uses a variant of the Gauss-Newton regression. In Section 7.7, we discussed the GNR (7.38), in which the regressand and regressors are evaluated at the restricted estimates, and showed how it can be used to calculate test statistics. Testing for serial correlation is simply an application of this procedure. Suppose we want to test a nonlinear regression model for AR(1) errors. The alternative model is given by (10.12), for observations 2 through n, with the null hypothesis being that $\rho = 0$. In this case, the GNR (7.38) is

$$\tilde{\boldsymbol{u}} = \boldsymbol{P}_W \tilde{\boldsymbol{X}} \boldsymbol{b} + r \boldsymbol{P}_W \tilde{\boldsymbol{u}}_{-1} + \text{residuals}, \tag{10.103}$$

where $\tilde{\boldsymbol{\beta}}$ denotes the IV estimates under the null hypothesis of no serial correlation, $\tilde{\boldsymbol{u}}$ denotes $\boldsymbol{y} - \boldsymbol{x}(\tilde{\boldsymbol{\beta}})$, and $\tilde{\boldsymbol{X}}$ denotes $\boldsymbol{X}(\tilde{\boldsymbol{\beta}})$. This is clearly the IV analog of regression (10.76); if the two occurrences of $\boldsymbol{P}_W$ were removed, (10.76) and (10.103) would be identical. The t statistic on the estimate of r from this regression will be a valid test statistic. This will be true both when (10.103) is estimated explicitly by OLS and when $\tilde{\boldsymbol{u}}$ is regressed on $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{u}}_{-1}$ using

where $\hat{\boldsymbol{\beta}}$ denotes the NLS estimates of $\boldsymbol{\beta}$ for the whole sample. The GNR (11.04) may be written more compactly as

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{X}}\boldsymbol{b} + \boldsymbol{\delta} * \hat{\boldsymbol{X}}\boldsymbol{c} + \text{residuals,} \tag{11.05}$$

where $\hat{\boldsymbol{u}}$ has typical element $y_t - x_t(\hat{\boldsymbol{\beta}})$, and $\hat{\boldsymbol{X}}$ has typical element $\boldsymbol{X}_t(\hat{\boldsymbol{\beta}})$. Here $*$ denotes the **direct product** of two matrices. Since $\delta_t X_{ti}(\hat{\boldsymbol{\beta}})$ is a typical element of $\boldsymbol{\delta} * \hat{\boldsymbol{X}}$, $\delta_t * \hat{\boldsymbol{X}}_t = \hat{\boldsymbol{X}}_t$ when $\delta_t = 1$ and $\delta_t * \hat{\boldsymbol{X}}_t = \boldsymbol{0}$ when $\delta_t = 0$. To perform the test, we simply have to estimate the model using the entire sample and regress the residuals from that estimation on the matrix of derivatives $\hat{\boldsymbol{X}}$ and on that matrix with the rows which correspond to group 1 observations set to zero. We do not have to reorder the data. As usual, there are several asymptotically valid test statistics, the best probably being the ordinary F statistic for the null hypothesis that $\boldsymbol{c} = \boldsymbol{0}$. In the usual case with k less than $\min(n_1, n_2)$, that test statistic will have k degrees of freedom in the numerator and $n - 2k$ degrees of freedom in the denominator.

Notice that the sum of squared residuals from regression (11.05) is equal to the SSR from the GNR

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{X}}\boldsymbol{b} + \text{residuals} \tag{11.06}$$

run over observations 1 to n_1 plus the SSR from the same GNR run over observations $n_1 + 1$ to n. This is the unrestricted sum of squared residuals for the F test of $\boldsymbol{c} = \boldsymbol{0}$ in (11.05). The restricted sum of squared residuals for that test is simply the SSR from (11.06) run over all n observations, which is the same as the SSR from nonlinear estimation of the null hypothesis H_0. Thus the ordinary Chow test for the GNR (11.06) will be numerically identical to the F test of $\boldsymbol{c} = \boldsymbol{0}$ in (11.05). This provides the easiest way to calculate the test statistic.

As we mentioned above, the ordinary Chow test (11.03) is not applicable if $\min(n_1, n_2) < k$. Using the GNR framework, it is easy to see why this is so. Suppose that $n_2 < k$ and $n_1 > k$, without loss of generality, since the numbering of the two groups of observations is arbitrary. Then the matrix $\boldsymbol{\delta} * \hat{\boldsymbol{X}}$, which has k columns, will have $n_2 < k$ rows that are not just rows of zeros and hence will have rank at most n_2. Thus, when equation (11.05) is estimated, at most n_2 elements of $\boldsymbol{c}$ will be identifiable, and the residuals corresponding to all observations that belong to group 2 will be zero. The number of degrees of freedom for the numerator of the F statistic must therefore be at most n_2. In fact, it will be equal to the rank of $[\hat{\boldsymbol{X}} \quad \boldsymbol{\delta} * \hat{\boldsymbol{X}}]$ minus the rank of $\hat{\boldsymbol{X}}$, which might be less than n_2 in some cases. The number of degrees of freedom for the denominator will be the number of observations for which (11.05) has nonzero residuals, which will normally be n_1, minus the number of regressors that affect those observations, which will be k, for a total of $n_1 - k$. Thus we can use the GNR whether or not $\min(n_1, n_2) < k$, provided that we use the appropriate numbers of degrees of freedom for the numerator and denominator of the F test.

than the other may be seen as a deficiency of these tests. That is so only if one misinterprets their nature. Nonnested hypothesis tests are specification tests, and since there is almost never any reason a priori to believe that either of the models actually generated the data, it is appropriate that nonnested tests, like other model specification tests, may well tell us that neither model seems to be compatible with the data.

It is important to stress that the purpose of nonnested tests is *not* to choose one out of a fixed set of models as the "best" one. That is the subject of an entirely different strand of the econometric literature, which deals with criteria for **model selection**. We will not discuss the rather large literature on model selection in this book. Two useful surveys are Amemiya (1980) and Leamer (1983), and an interesting recent paper is Pollak and Wales (1991).

It is of interest to examine more closely the case in which both models are linear, that is, $x(\beta) = X\beta$ and $z(\gamma) = Z\gamma$. This will allow us to see why the J and P tests (which in this case are identical) are asymptotically valid and also to see why these tests may not always perform well in finite samples. The J-test regression for testing H_1 against H_2 is

$$y = Xb + \alpha P_Z y + \text{residuals}, \tag{11.16}$$

where $P_Z = Z(Z^\top Z)^{-1}Z^\top$ and $b = (1 - \alpha)\beta$. Using the FWL Theorem, we see that the estimate of α from (11.16) will be the same as the estimate from the regression

$$M_X y = \alpha M_X P_Z y + \text{residuals}. \tag{11.17}$$

Thus, if $\acute{s}$ denotes the OLS estimate of σ from (11.16), the t statistic for $\alpha = 0$ will be

$$\frac{y^\top P_Z M_X y}{\acute{s}(y^\top P_Z M_X P_Z y)^{1/2}}. \tag{11.18}$$

First of all, notice that when only one column of Z, say Z_1, does not belong to $\mathcal{S}(X)$, it must be the case that

$$\mathcal{S}(X, P_Z y) = \mathcal{S}(X, Z) = \mathcal{S}(X, Z_1).$$

Therefore, the J-test regression (11.16) must yield exactly the same SSR as the regression

$$y = Xb + \delta Z_1 + \text{residuals}. \tag{11.19}$$

Thus, in this special case, the J test is equal in absolute value to the t statistic on the estimate of δ from (11.19).

When two or more columns of Z do not belong to $\mathcal{S}(X)$, this special result is no longer available. If the data were actually generated by H_1, we can replace y in the numerator of (11.18) by $X\beta + u$. Since $M_X X\beta = 0$, that numerator becomes

$$\beta^\top X^\top P_Z M_X u + u^\top P_Z M_X u. \tag{11.20}$$

The two terms of (11.20) are of different orders. The first term is a weighted sum of the elements of the vector $\boldsymbol{u}$, each of which has mean zero. Thus, under suitable regularity conditions, it is easy to see that

$$n^{-1/2}\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{u} \overset{a}{\sim} N\Big(\boldsymbol{0},\ \plim_{n\to\infty}\big(n^{-1}\sigma_1^2\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{P}_Z\boldsymbol{X}\boldsymbol{\beta}\big)\Big).$$

This first term is thus $O(n^{1/2})$. The second term, in contrast, is $O(1)$, since

$$\plim_{n\to\infty}\big(\boldsymbol{u}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{u}\big) = \plim_{n\to\infty}\big(\boldsymbol{u}^{\top}\boldsymbol{P}_Z\boldsymbol{u} - \boldsymbol{u}^{\top}\boldsymbol{P}_Z\boldsymbol{P}_X\boldsymbol{u}\big)$$
$$= \sigma_1^2 k_2 - \sigma_1^2 \lim_{n\to\infty}\big(\text{Tr}\,(\boldsymbol{P}_Z\boldsymbol{P}_X)\big),$$

and the trace of $\boldsymbol{P}_Z\boldsymbol{P}_X$ is $O(1)$. Thus, asymptotically, it is only the first term in (11.20) that matters.

Similarly, under H_1 the factor in parentheses in the denominator of (11.18) is equal to

$$\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{P}_Z\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{P}_Z\boldsymbol{u} + \boldsymbol{u}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{P}_Z\boldsymbol{u}. \qquad (11.21)$$

By arguments similar to those used in connection with the numerator, the first of the three terms in (11.21) may be shown to be $O(n)$, the second $O(n^{1/2})$, and the third $O(1)$. Moreover, it is clear that $\acute{s} \to \sigma_1$ under H_1. Thus, asymptotically under H_1, the test statistic (11.18) tends to the random variable

$$\frac{\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{u}}{\sigma_1\big(\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{P}_Z\boldsymbol{X}\boldsymbol{\beta}\big)^{1/2}},$$

which can be shown to be distributed asymptotically as $N(0,1)$.

This analysis not only makes it clear why the J and P tests are valid asymptotically but also indicates why they may not be well behaved in finite samples. When the sample size is small or $\boldsymbol{Z}$ contains many regressors that are not in $\mathcal{S}(\boldsymbol{X})$, the quantity $\boldsymbol{u}^{\top}\boldsymbol{P}_Z\boldsymbol{M}_X\boldsymbol{u}$, which is asymptotically negligible, may actually be large and positive. Hence, in such circumstances, the J-test statistic (11.18) may have a mean that is substantially greater than zero.

Several ways of reducing or eliminating this bias have been suggested. The simplest, which was first proposed by Fisher and McAleer (1981) and further studied by Godfrey (1983), is to replace $\hat{\boldsymbol{\gamma}}$ in the J-test and P-test regressions by $\tilde{\boldsymbol{\gamma}}$, which is the estimate of $\boldsymbol{\gamma}$ obtained by minimizing

$$\big(\hat{\boldsymbol{x}} - \boldsymbol{z}(\boldsymbol{\gamma})\big)^{\top}\big(\hat{\boldsymbol{x}} - \boldsymbol{z}(\boldsymbol{\gamma})\big).$$

Thus $\tilde{\boldsymbol{\gamma}}$ is the NLS estimate of $\boldsymbol{\gamma}$ obtained when one uses the fitted values $\hat{\boldsymbol{x}}$ instead of the dependent variable $\boldsymbol{y}$. In the linear case, this means that the J-test regression (11.16) is replaced by the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \alpha\boldsymbol{P}_Z\boldsymbol{P}_X\boldsymbol{y} + \text{residuals}. \qquad (11.22)$$

This regression yields what is called the **J_A test** because Fisher and McAleer attributed the basic idea to Atkinson (1970). Godfrey (1983) showed, using a result of Milliken and Graybill (1970), that the t statistic on the estimate of α from regression (11.22) actually has the t distribution in finite samples under the usual conditions for t statistics to have this distribution ($\boldsymbol{u}$ normally distributed, $\boldsymbol{X}$ and $\boldsymbol{Z}$ independent of $\boldsymbol{u}$). The intuition for this result is quite simple. The vector of fitted values $\boldsymbol{P}_X \boldsymbol{y}$ contains only the part of $\boldsymbol{y}$ that lies in $\mathcal{S}(\boldsymbol{X})$. It must therefore be independent of $\boldsymbol{M}_X \boldsymbol{y}$, which is what the residuals from (11.22) would be if $\alpha = 0$. Therefore, we can treat $\boldsymbol{P}_Z \boldsymbol{P}_X \boldsymbol{y}$ (or any other regressor that depends on $\boldsymbol{y}$ only through $\boldsymbol{P}_X \boldsymbol{y}$) as if it were a fixed regressor.[4] The **P_A test** is to the P test as the J_A test is to the J test.

Unfortunately, the J_A and P_A tests are in many circumstances much less powerful than the ordinary J and P tests; see Davidson and MacKinnon (1982) and Godfrey and Pesaran (1983). Thus if, for example, the J test rejects the null hypothesis and the J_A test does not, it is hard to know whether this is because the former is excessively prone to commit a Type I error or because the latter is excessively prone to commit a Type II error.

A second approach is to estimate the expectation of $\boldsymbol{u}^\top \boldsymbol{M}_X \boldsymbol{P}_Z \boldsymbol{u}$, subtract it from $\boldsymbol{y}^\top \boldsymbol{M}_X \boldsymbol{P}_Z \boldsymbol{y}$, and then divide it by an estimate of the square root of the variance of the resulting quantity so as to obtain a test statistic that would be asymptotically $N(0,1)$. This approach was originally proposed in a somewhat more complicated form by Godfrey and Pesaran (1983); a simpler version may be found in the "Reply" of MacKinnon (1983). This second approach is a good deal harder to use than the J_A test, since it involves matrix calculations that cannot be performed by a sequence of regressions, and it does not yield an exact test. It also requires the assumption of normality. However, it does seem to yield a test with much better finite-sample properties under the null than the J test and, at least in some circumstances, much better power than the J_A test.

The vector $\tilde{\boldsymbol{\gamma}}$ is of interest in its own right. The original Cox test used the fact that, under H_1,

$$\operatorname*{plim}_{n \to \infty} \left(\tilde{\boldsymbol{\gamma}} \right) = \operatorname*{plim}_{n \to \infty} \left(\hat{\boldsymbol{\gamma}} \right).$$

It is possible to construct a test based directly on the difference between $\hat{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\gamma}}$. Such a test, originally proposed by Dastoor (1983) and developed further by Mizon and Richard (1986), looks at whether the value of $\boldsymbol{\gamma}$ predicted by the H_1 model (i.e., $\tilde{\boldsymbol{\gamma}}$) is the same as the value obtained by direct estimation of H_2 (i.e., $\hat{\boldsymbol{\gamma}}$). These tests are called **encompassing tests**, because if H_1 does explain the performance of H_2, it may be said to "encompass" it; see Mizon (1984). The principle on which they are based is sometimes called the **encompassing principle**.

---

[4] By the same argument, the RESET test discussed in Section 6.5 is exact in finite samples whenever an ordinary t test would be exact.

For a regression model like (11.27), it is easy to compute a DWH test by means of an artificial regression. We saw some examples of this in Section 7.9 and will discuss further examples below. However, there is another way to compute DWH tests, and it can be more convenient in some cases. For some model that need not necessarily be a regression model, let $\hat{\boldsymbol{\theta}}$ denote an efficient estimator of the model parameters and $\check{\boldsymbol{\theta}}$ an estimator that is less efficient but consistent under weaker conditions than those of the model. Let us denote the vector of contrasts between $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ by $\boldsymbol{e}$. Then we have seen that

$$n^{1/2}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + n^{1/2}\boldsymbol{e}, \tag{11.30}$$

where $n^{1/2}\boldsymbol{e}$ is asymptotically uncorrelated with $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. This result was proved for models estimated by maximum likelihood in Section 8.8; its finite-sample equivalent for linear regression models was proved as part of the proof of the Gauss-Markov Theorem in Section 5.5. Because the two terms on the right-hand side of (11.30) are asymptotically uncorrelated, the asymptotic covariance matrix of the left-hand side is just the sum of the asymptotic covariance matrices of those two terms. Therefore, we obtain

$$\lim_{n\to\infty} \boldsymbol{V}\big(n^{1/2}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = \lim_{n\to\infty} \boldsymbol{V}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) + \lim_{n\to\infty} \boldsymbol{V}(n^{1/2}\boldsymbol{e}),$$

from which, in simplified notation, we may deduce the asymptotic covariance matrix of the vector of contrasts:

$$\boldsymbol{V}^\infty(\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = \boldsymbol{V}^\infty(\check{\boldsymbol{\theta}}) - \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}}). \tag{11.31}$$

In words, the asymptotic covariance matrix of the difference between $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ is equal to the difference of their respective asymptotic covariance matrices. This important result is due to Hausman (1978).

The result (11.31) can be used to construct DWH tests of the form

$$(\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \big(\check{\boldsymbol{V}}(\check{\boldsymbol{\theta}}) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}})\big)^{-1}(\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}), \tag{11.32}$$

where $\check{\boldsymbol{V}}(\check{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}})$ denote estimates of the covariance matrices of $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$, respectively. The test statistic (11.32) will be asymptotically distributed as chi-squared with as many degrees of freedom as the rank of $\boldsymbol{V}^\infty(\check{\boldsymbol{\theta}}) - \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}})$. Note that the inverse in (11.32) will have to be replaced by a generalized inverse if, as is often the case, the rank of $\boldsymbol{V}^\infty(\check{\boldsymbol{\theta}}) - \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}})$ is less than the number of parameters in $\boldsymbol{\theta}$; see Hausman and Taylor (1982). There can be practical difficulties with (11.32) if $\check{\boldsymbol{V}}(\check{\boldsymbol{\theta}}) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}})$ is not positive semidefinite or if the rank of $\check{\boldsymbol{V}}(\check{\boldsymbol{\theta}}) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}})$ differs from the rank of $\boldsymbol{V}^\infty(\check{\boldsymbol{\theta}}) - \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}})$. That is why we emphasize the approach based on artificial regressions.

In the case of the linear regression (11.27), where the two estimators are (11.28) and (11.29), the DWH test is based on the vector of contrasts

$$\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \big(\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}\big)^{-1} \boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{M}_X \boldsymbol{y}. \tag{11.33}$$

This looks just like expression (7.59), with $\boldsymbol{A}$ replacing $\boldsymbol{P}_W$, and may be derived in exactly the same way. The first factor in (11.33), $(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X})^{-1}$, is simply a $k \times k$ matrix with full rank, which will have no effect on any test statistic that we might compute. Therefore, what we really want to do is test whether the vector

$$n^{-1/2}\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{M}_X\boldsymbol{y} \tag{11.34}$$

has mean zero asymptotically. This vector has k elements, but even if $\boldsymbol{A}\boldsymbol{X}$ has full rank, not all those elements may be random variables, because $\boldsymbol{M}_X$ may annihilate some columns of $\boldsymbol{A}\boldsymbol{X}$. Suppose that k^* is the number of linearly independent columns of $\boldsymbol{A}\boldsymbol{X}$ that are not annihilated by $\boldsymbol{M}_X$. Then testing (11.34) is equivalent to testing whether the vector

$$n^{-1/2}\boldsymbol{X}^{*\top}\boldsymbol{A}\boldsymbol{M}_X\boldsymbol{y} \tag{11.35}$$

has mean zero asymptotically, where $\boldsymbol{X}^*$ denotes k^* columns of $\boldsymbol{X}$ with the property that none of the columns of $\boldsymbol{A}\boldsymbol{X}^*$ is annihilated by $\boldsymbol{M}_X$.

Now consider the artificial regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{X}^*\boldsymbol{\delta} + \text{residuals}. \tag{11.36}$$

It is easily shown by using the FWL Theorem that the OLS estimate of $\boldsymbol{\delta}$ is

$$\acute{\boldsymbol{\delta}} = \big(\boldsymbol{X}^{*\top}\boldsymbol{A}\boldsymbol{M}_X\boldsymbol{A}\boldsymbol{X}^*\big)^{-1}\boldsymbol{X}^{*\top}\boldsymbol{A}\boldsymbol{M}_X\boldsymbol{y},$$

and it is evident that, in general, $\text{plim}(\acute{\boldsymbol{\delta}}) = \boldsymbol{0}$ if and only if (11.35) has mean zero asymptotically. The ordinary F statistic for $\boldsymbol{\delta} = \boldsymbol{0}$ in (11.36) is

$$\frac{\boldsymbol{y}^\top\boldsymbol{P}_{M_X A X^*}\boldsymbol{y}/k^*}{\boldsymbol{y}^\top\boldsymbol{M}_{X,\,M_X A X^*}\boldsymbol{y}/(n-k-k^*)}, \tag{11.37}$$

where $\boldsymbol{P}_{M_X A X^*}$ is the matrix that projects onto $\mathcal{S}(\boldsymbol{M}_X\boldsymbol{A}\boldsymbol{X}^*)$, and $\boldsymbol{M}_{X,\,M_X A X^*}$ is the matrix that projects onto $\mathcal{S}^\perp(\boldsymbol{X}, \boldsymbol{M}_X\boldsymbol{A}\boldsymbol{X}^*)$. If (11.27) actually generated the data, the statistic (11.37) will certainly be valid asymptotically, since the denominator will then consistently estimate σ^2. It will be exactly distributed as $F(k^*, n-k-k^*)$ in finite samples if the u_t's in (11.27) are normally distributed and $\boldsymbol{X}$ and $\boldsymbol{A}$ can be treated as fixed. Regression (11.36) and expression (11.37) are essentially the same as regression (7.62) and expression (7.64), respectively; the latter are special cases of the former.

The most common type of DWH test is the one we dealt with in Section 7.9, which asks whether least squares estimates are consistent when some of the regressors may be correlated with the error terms. However, there are numerous other possibilities. For example, $\breve{\boldsymbol{\beta}}$ might be the OLS estimator for $\boldsymbol{\beta}$ in the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \tag{11.38}$$

would be to have only one column in $\boldsymbol{Z}$, that column being proportional to $\boldsymbol{a}$. In practice, we are rarely in that happy position. There are normally a number of things that we suspect might be wrong with our model and hence a large number of regression directions in which to test. Faced with this situation, there are at least two ways to proceed.

One approach is to test against each type of potential misspecification separately, with each test having only one or a few degrees of freedom. If the model is in fact wrong in one or a few of the regression directions in which these tests are carried out, such a procedure is as likely as any to inform us of that fact. However, the investigator must be careful to control the overall size of the test, since when one does, say, 10 different tests each at the .05 level, the overall size could be as high as .40; see Savin (1980). Moreover, one should avoid jumping to the conclusion that the model is wrong in a particular way just because a certain test statistic is significant. One must remember that $\cos^2\phi$ will often be well above zero for *many* tests, even if only one thing is wrong with the model.

Alternatively, it is possible to test for a great many types of misspecification at once by putting all the regression directions we want to test against into one big $\boldsymbol{Z}$ matrix. This maximizes $\cos^2\phi$ and hence maximizes the chance that the test is consistent, and it also makes it easy to control the size of the test. But because such a test will have many degrees of freedom, power may be poor except when the sample size is large. Moreover, if such a test rejects the null, that rejection gives us very little information as to what may be wrong with the model. Of course, the coefficients on the individual columns of $\boldsymbol{Z}$ in the test regression may well be informative.

This raises the question of what to do when one or more tests reject the null hypothesis. That is a very difficult question, and we will discuss it in Section 12.7.

12.6 Asymptotic Relative Efficiency

Since all consistent tests reject with probability one as the sample size tends to infinity, it is not obvious how to compare the power of tests of which the distributions are known only asymptotically. Various approaches have been proposed in the statistical literature, of which the best known is probably the concept of **asymptotic relative efficiency**, or **ARE**. This concept, which is closely related to the idea of local alternatives, is due to Pitman (1949), and has since been developed by many other authors; see Kendall and Stuart (1979, Chapter 25). Suppose that we have two test statistics, say τ_1 and τ_2, both of which have the same asymptotic distribution under the null and both of which, like all the test statistics we have discussed in this chapter, are root-n consistent. This means that, for the test to have a nondegenerate asymptotic distribution, the drifting DGP must approach the simple null hypothesis at a

rate proportional to $n^{-1/2}$. In this case, the asymptotic efficiency of τ_2 relative to τ_1 is defined as

$$\text{ARE}_{21} = \lim_{n \to \infty} \left(\frac{n_1}{n_2} \right),$$

where n_1 and n_2 are sample sizes such that τ_1 and τ_2 have the same power, and the limit is taken as both n_1 and n_2 tend to infinity. If, for example, ARE_{21} were 0.25, τ_2 would asymptotically require 4 times as large a sample as τ_1 to achieve the same power.

For tests with the same number of degrees of freedom, it is easy to see that

$$\text{ARE}_{21} = \frac{\cos^2 \phi_2}{\cos^2 \phi_1}.$$

Recall from expression (12.23) that the NCP is proportional to $\cos^2 \phi$. If the DGP did not drift, it would also be proportional to the sample size. If τ_1 and τ_2 are to be equally powerful in this case, they must have the same NCP. This means that n_1/n_2 must be equal to $\cos^2 \phi_2 / \cos^2 \phi_1$. Suppose, for example, that $\cos^2 \phi_1 = 1$ and $\cos^2 \phi_2 = 0.5$. Then the implicit alternative hypothesis for τ_1 must include the DGP, while the implicit alternative for τ_2 does not. Thus the directions in which τ_1 is testing explain all of the divergence between the null hypothesis and the DGP, while the directions in which τ_2 is testing explain only half of it. But we can compensate for this reduced explanatory power by making n_2 twice as large as n_1, so as to make both tests equally powerful asymptotically. Hence ARE_{21} must be 0.5. See Davidson and MacKinnon (1987) for more on this special case.

In the more general case in which τ_1 and τ_2 have different numbers of degrees of freedom, calculating the ARE becomes more complicated. The optimal test will be one for which the implicit alternative hypothesis includes the drifting DGP (so that $\cos^2 \phi = 1$) and that involves only one degree of freedom. There may of course be many asymptotically equivalent tests that satisfy these criteria, or there may in practice be none at all. Tests that involve more than one degree of freedom, or have $\cos^2 \phi < 1$, will be asymptotically less efficient than the optimal test and hence will have AREs less than 1.

The consequences of using tests with $r > 1$ and/or $\cos^2 \phi < 1$ are illustrated in Table 12.1. The effect of changing $\cos^2 \phi$ does not depend on either the size or power of the test, but the effect of changing r depends on both of these; see Rothe (1981) and Saikkonen (1989). The table was calculated for a size of .05 and powers of .90 (the first entry in each cell) and .50 (the second entry). Each entry in the table is the ARE for the specified test relative to that for the optimal one. Thus each entry may be interpreted as the factor by which the sample size for the optimal test may be smaller than the sample size for the nonoptimal test if both are to have equal asymptotic power.

From Table 12.1, we see that the cost of using a test with a needlessly large number of degrees of freedom, or with a value of $\cos^2 \phi$ less than 1,

Our objective is to calculate the NCPs and corresponding values of $\cos^2\phi$ for tests of H_0 against both H_1 and H_2 when the data are generated by (12.28). Thus we will suppose that the data are generated by a drifting DGP that is a special case of H_2. This drifting DGP can be written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta}_0 + \alpha_0 n^{-1/2}(\boldsymbol{X}_{t-1}\boldsymbol{\beta}_0 + u_{t-1}) + u_t, \;\; u_t \sim \text{IID}(0, \sigma_0^2). \qquad (12.29)$$

Note that this DGP does not involve the recursive calculation of y_t, as (12.28) seems to require, because (12.29) is locally equivalent to (12.28) in the neighborhood of $\delta = 0$ and $\alpha_0 = 0$.

When we test H_0 against H_2, we will be testing in the direction of the DGP and $\cos^2\phi$ will evidently be unity. Using expression (12.25), we see that the NCP for this test is

$$\Lambda_{22} \equiv \frac{\alpha_0^2}{\sigma_0^2} \plim_{n\to\infty} \Big(\frac{1}{n}\big(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}\big)^{\top}\boldsymbol{M}_X\big(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}\big)\Big), \qquad (12.30)$$

where $\boldsymbol{u}_{-1}$ and $\boldsymbol{X}_{-1}$ denote, respectively, the vector with typical element u_{t-1} and the matrix with typical row $\boldsymbol{X}_{t-1}$. Here $\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}$ is playing the role of the vector $\boldsymbol{a}$ in expression (12.25). The notation Λ_{22} means that H_2 is the alternative against which we are testing and that the DGP belongs to H_2. Taking the probability limit, (12.30) becomes

$$\begin{aligned}
\Lambda_{22} &= \frac{\alpha_0^2}{\sigma_0^2}\Big(\sigma_0^2 + \plim_{n\to\infty}\frac{1}{n}\big\|\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0\big\|^2\Big) \\
&= \alpha_0^2\Big(1 + \sigma_0^{-2}\plim_{n\to\infty}\frac{1}{n}\big\|\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0\big\|^2\Big).
\end{aligned}$$

Now let us see what happens when we test H_0 against H_1. In the neighborhood of H_0, the latter is locally equivalent to

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \rho\boldsymbol{u}_{-1} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \qquad (12.31)$$

which avoids the recursive calculation that (12.27) seems to require. Because AR(1) and MA(1) processes are locally equivalent near the point where their respective parameters are zero, this looks like a model with an MA(1) error process. We see from (12.31) that $\boldsymbol{u}_{-1}$ plays the role of $\boldsymbol{Z}$. Once again, $\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}$ plays the role of $\boldsymbol{a}$. Thus, from (12.18), the NCP is given by

$$\begin{aligned}
\Lambda_{12} &= \frac{\alpha_0^2}{\sigma_0^2}\plim_{n\to\infty}\Big(\frac{1}{n}\big(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}\big)^{\top}\boldsymbol{M}_X\boldsymbol{u}_{-1}\Big)\plim_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{u}_{-1}{}^{\top}\boldsymbol{M}_X\boldsymbol{u}_{-1}\Big)^{-1} \\
&\quad \times \plim_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{u}_{-1}{}^{\top}\boldsymbol{M}_X\big(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}\big)\Big). \qquad (12.32)
\end{aligned}$$

Because

$$\operatorname*{plim}_{n\to\infty}\Big(\tfrac{1}{n}(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1})^{\top}\boldsymbol{M}_X\boldsymbol{u}_{-1}\Big)$$

$$= \operatorname*{plim}_{n\to\infty}\Big(\tfrac{1}{n}\big(\boldsymbol{\beta}_0^{\top}\boldsymbol{X}_{-1}^{\top}\boldsymbol{M}_X\boldsymbol{u}_{-1} + \boldsymbol{u}_{-1}^{\top}\boldsymbol{M}_X\boldsymbol{u}_{-1}\big)\Big) = \sigma_0^2,$$

expression (12.32) simplifies to

$$\frac{\alpha_0^2}{\sigma_0^2}\sigma_0^2(\sigma_0^{-2})\sigma_0^2 = \alpha_0^2.$$

Since the data were generated by a special case of H_2, $\cos^2\phi$ for the test against H_1 is simply the ratio of the NCP Λ_{12} to the NCP Λ_{22}. Thus

$$
\begin{aligned}
\cos^2\phi &= \alpha_0^2\Big(\alpha_0^2\Big(1 + \sigma_0^{-2}\operatorname{plim}\tfrac{1}{n}\big\|\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0\big\|^2\Big)\Big)^{-1}\\
&= \Big(1 + \frac{\operatorname{plim} n^{-1}\|\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0\|^2}{\sigma_0^2}\Big)^{-1}.
\end{aligned}
\tag{12.33}
$$

The second line of (12.33) provides a remarkably simple expression for $\cos^2\phi$ for this special case. It depends only on the ratio of the probability limit of n^{-1} times the squared length of the vector $\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0$ to the variance of the error terms in the DGP (12.29). As this ratio tends to zero, $\cos^2\phi$ tends to unity. Conversely, as this ratio tends to infinity, $\cos^2\phi$ tends to zero. The intuition is very simple. As the ratio of $\operatorname{plim} n^{-1}\|\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0\|^2$ to σ_0^2 tends to zero, because for instance $\boldsymbol{\beta}_0$ tends to zero, $\boldsymbol{M}_X\boldsymbol{y}_{-1}$ (where $\boldsymbol{y}_{-1}$ has typical element y_{t-1}) becomes indistinguishable from $\boldsymbol{M}_X\boldsymbol{u}_{-1}$. When that happens, a test against H_1 becomes indistinguishable from a test against H_2. On the other hand, as the ratio tends in the other direction toward infinity, the correlation between y_{t-1} and u_{t-1} tends to zero, and the directions in which H_1 and H_2 differ from H_0 tend to become mutually orthogonal.

The foregoing analysis could just as easily have been performed under the assumption that the data were generated by a special case of H_1. The drifting DGP would then be

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta}_0 + \rho_0 n^{-1/2}u_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma_0^2).$$

When we test H_0 against H_1, $\cos^2\phi$ is now unity, and by an even simpler argument than the one that led to (12.32) we see that the NCP is

$$\Lambda_{11} = \frac{\rho_0^2}{\sigma_0^2}\operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{u}_{-1}^{\top}\boldsymbol{M}_X\boldsymbol{u}_{-1}\Big) = \rho_0^2.$$

Similarly, when we test H_0 against H_2, the NCP is

$$\Lambda_{21} = \frac{\rho_0^2}{\sigma_0^2} \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} \boldsymbol{u}_{-1}^{\top} \boldsymbol{M}_X (\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}) \right)$$

$$\times \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} (\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1})^{\top} \boldsymbol{M}_X (\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}) \right)^{-1}$$

$$\times \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} (\boldsymbol{X}_{-1}\boldsymbol{\beta}_0 + \boldsymbol{u}_{-1})^{\top} \boldsymbol{M}_X \boldsymbol{u}_{-1} \right).$$

This simplifies to

$$\frac{\rho_0^2}{\sigma_0^2} \sigma_0^2 \left(\sigma_0^2 + \operatorname{plim} \frac{1}{n} \|\boldsymbol{M}_X \boldsymbol{X}_{-1}\boldsymbol{\beta}_0\|^2 \right)^{-1} \sigma_0^2$$

$$= \rho_0^2 \left(1 + \sigma_0^{-2} \operatorname{plim} \frac{1}{n} \|\boldsymbol{M}_X \boldsymbol{X}_{-1}\boldsymbol{\beta}_0\|^2 \right)^{-1}.$$

Evidently, $\cos^2\phi$ for the test of H_0 against H_2 is the right-hand expression here divided by ρ_0^2, which is

$$\left(1 + \frac{\operatorname{plim} n^{-1} \|\boldsymbol{M}_X \boldsymbol{X}_{-1}\boldsymbol{\beta}_0\|^2}{\sigma_0^2} \right)^{-1}. \tag{12.34}$$

This last result is worth comment. We have found that $\cos^2\phi$ for the test against H_2 when the data were generated by H_1, expression (12.34), is identical to $\cos^2\phi$ for the test against H_1 when the data were generated by H_2, expression (12.33). This result is true not just for this example, but for every case in which both alternatives involve one-degree-of-freedom tests. Geometrically, this equivalence simply reflects the fact that when $\boldsymbol{z}$ is a vector, the angle between $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{a}$ and the projection of $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{a}$ onto $\mathcal{S}(\boldsymbol{X}, \boldsymbol{z})$, which is

$$\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{z} (\boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{z})^{-1} \boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{a},$$

is the same as the angle between $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{a}$ and $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{z}$. The reason for this is that $(\boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{z})^{-1} \boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{a}$ is a scalar when $\boldsymbol{z}$ is a vector. Hence, if we reverse the roles of $\boldsymbol{a}$ and $\boldsymbol{z}$, the angle is unchanged. This geometrical fact also results in two numerical facts. First, in the regressions

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \gamma \boldsymbol{z} + \text{residuals} \quad \text{and}$$

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \delta \boldsymbol{y} + \text{residuals},$$

the t statistic on $\boldsymbol{z}$ in the first is equal to that on $\boldsymbol{y}$ in the second. Second, in the regressions

$$\boldsymbol{M}_X \boldsymbol{y} = \gamma \boldsymbol{M}_X \boldsymbol{z} + \text{residuals} \quad \text{and}$$

$$\boldsymbol{M}_X \boldsymbol{z} = \delta \boldsymbol{M}_X \boldsymbol{y} + \text{residuals},$$

the t statistics on γ and δ are numerically identical and so are the uncentered R^2's.

The analysis of power for this example illustrates the simplicity and generality of the idea of drifting DGPs. Although the case considered is rather simple, it is very commonly encountered in applied work. Regression models with time-series data frequently display evidence of serial correlation in the form of low Durbin-Watson statistics or other significant test statistics for AR(1) errors. We have seen that (except when $\operatorname{plim} n^{-1}\|\boldsymbol{M}_X \boldsymbol{X}_{-1}\boldsymbol{\beta}_0\|^2$ is large relative to σ_0^2) this evidence is almost as consistent with the hypothesis that the model should have included a lagged dependent variable as with the hypothesis that the error terms actually follow an AR(1) process. Thus one should be very cautious indeed when one has to interpret the results of a test against AR(1) errors that rejects the null. One would certainly want to consider several possible alternative models in addition to the alternative that the errors actually follow an AR(1) process. At the very least, before even tentatively accepting that alternative, one would want to subject it to the tests for common factor restrictions that we discussed in Section 10.9.

In the foregoing example, it was easy to evaluate analytically the values of Λ and $\cos^2\phi$ in which we were interested. This will of course not always be the case. However, it is always possible to calculate approximations to these quantities numerically. To do this one simply has to run regression (12.20), evaluating $\boldsymbol{X}(\boldsymbol{\beta})$, $\boldsymbol{a}$, and $\boldsymbol{Z}$ at assumed (or estimated) parameter values. If $\boldsymbol{a}$ and/or $\boldsymbol{Z}$ were stochastic, one would have to generate them randomly and use a very large number of generated observations (which can be obtained by repeating the actual observations as many times as necessary) so as to approximate the desired probability limits. The uncentered R^2 from the regression approximates $\cos^2\phi$ and the explained sum of squares approximates Λ.

12.8 Test Statistics that Do Not Reject the Null

For most of this chapter, we have been concerned with how to interpret test statistics that reject the null hypothesis. In many instances, of course, test statistics fail to reject. Thus it is just as important to know how to interpret a failure to reject as it is to know how to interpret a rejection. Even though we may sometimes speak about "accepting" a null hypothesis when one or more tests fail to reject it, any such acceptance should obviously be provisional and tempered with caution. Just how cautious we should be depends on the power of the test or tests that did not reject the null. We can be most confident about the validity of the null hypothesis if tests that are known to have high power against the alternatives of interest fail to reject it.

As we have seen, the power of a test depends on the way the data are actually generated. In a recent paper, Andrews (1989) has suggested that, as an aid to interpreting nonrejection of a null hypothesis by a particular test, one might consider the power the test would have under the DGPs associated with alternative hypotheses of interest. It seems reasonable that such alternatives

Similarly, when we test H_0 against H_2, the NCP is

$$\Lambda_{21} = \frac{\rho_0^2}{\sigma_0^2} \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} \boldsymbol{u}_{-1}^{\top} \boldsymbol{M}_X (\boldsymbol{X}_{-1} \boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}) \right)$$

$$\times \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} (\boldsymbol{X}_{-1} \boldsymbol{\beta}_0 + \boldsymbol{u}_{-1})^{\top} \boldsymbol{M}_X (\boldsymbol{X}_{-1} \boldsymbol{\beta}_0 + \boldsymbol{u}_{-1}) \right)^{-1}$$

$$\times \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} (\boldsymbol{X}_{-1} \boldsymbol{\beta}_0 + \boldsymbol{u}_{-1})^{\top} \boldsymbol{M}_X \boldsymbol{u}_{-1} \right).$$

This simplifies to

$$\frac{\rho_0^2}{\sigma_0^2} \sigma_0^2 \left(\sigma_0^2 + \operatorname{plim} \frac{1}{n} \left\| \boldsymbol{M}_X \boldsymbol{X}_{-1} \boldsymbol{\beta}_0 \right\|^2 \right)^{-1} \sigma_0^2$$

$$= \rho_0^2 \left(1 + \sigma_0^{-2} \operatorname{plim} \frac{1}{n} \left\| \boldsymbol{M}_X \boldsymbol{X}_{-1} \boldsymbol{\beta}_0 \right\|^2 \right)^{-1}.$$

Evidently, $\cos^2\phi$ for the test of H_0 against H_2 is the right-hand expression here divided by ρ_0^2, which is

$$\left(1 + \frac{\operatorname{plim} n^{-1} \| \boldsymbol{M}_X \boldsymbol{X}_{-1} \boldsymbol{\beta}_0 \|^2}{\sigma_0^2} \right)^{-1}. \tag{12.34}$$

This last result is worth comment. We have found that $\cos^2\phi$ for the test against H_2 when the data were generated by H_1, expression (12.34), is identical to $\cos^2\phi$ for the test against H_1 when the data were generated by H_2, expression (12.33). This result is true not just for this example, but for every case in which both alternatives involve one-degree-of-freedom tests. Geometrically, this equivalence simply reflects the fact that when $\boldsymbol{z}$ is a vector, the angle between $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{a}$ and the projection of $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{a}$ onto $\mathcal{S}(\boldsymbol{X}, \boldsymbol{z})$, which is

$$\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{z} \left(\boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{z} \right)^{-1} \boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{a},$$

is the same as the angle between $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{a}$ and $\alpha n^{-1/2} \boldsymbol{M}_X \boldsymbol{z}$. The reason for this is that $(\boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{z})^{-1} \boldsymbol{z}^{\top} \boldsymbol{M}_X \boldsymbol{a}$ is a scalar when $\boldsymbol{z}$ is a vector. Hence, if we reverse the roles of $\boldsymbol{a}$ and $\boldsymbol{z}$, the angle is unchanged. This geometrical fact also results in two numerical facts. First, in the regressions

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \gamma\boldsymbol{z} + \text{residuals} \quad \text{and}$$

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \delta\boldsymbol{y} + \text{residuals},$$

the t statistic on $\boldsymbol{z}$ in the first is equal to that on $\boldsymbol{y}$ in the second. Second, in the regressions

$$\boldsymbol{M}_X \boldsymbol{y} = \gamma\boldsymbol{M}_X \boldsymbol{z} + \text{residuals} \quad \text{and}$$

$$\boldsymbol{M}_X \boldsymbol{z} = \delta\boldsymbol{M}_X \boldsymbol{y} + \text{residuals},$$

the t statistics on γ and δ are numerically identical and so are the uncentered R^2's.

the standard normal distribution, this probability is

$$P(\alpha, \lambda) \equiv 1 - \Phi(c_\alpha - \lambda) + \Phi(-c_\alpha - \lambda). \tag{12.36}$$

In order to find the inverse power function corresponding to (12.36), we let $P(\alpha, \lambda) = \pi$ for some desired level of power π. This equation implicitly defines the inverse power function. It is easy to check from (12.36) that $P(\alpha, -\lambda) = P(\alpha, \lambda)$. Thus, if $P(\alpha, \lambda) = \pi$, then $P(\alpha, -\lambda) = \pi$ also. However, the nonuniqueness of λ would not arise if we were to square the test statistic to obtain a χ^2 form. No closed-form expression exists giving the (absolute) value of λ as a function of α and π in the present example, but for any given arguments λ is not hard to calculate numerically.

What interpretation should we give to the resulting function $\lambda(\alpha, \pi)$? If we square the asymptotically normal statistic (12.35) in order to obtain a χ^2 form, the result will have a limiting distribution of $\chi^2(1, \Lambda)$ with $\Lambda = \lambda^2$. Then it appears that $\Lambda = (\lambda(\alpha, \pi))^2$ is asymptotically the smallest NCP needed in order that a test of size α based on the square of (12.35) should have probability at least π of rejecting the null.

Let the nonlinear regression model be written, as usual, as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \tag{12.37}$$

where the parameter of interest θ is a component of the parameter vector $\boldsymbol{\beta}$. If we denote by $\boldsymbol{X}_\theta$ the derivative of the vector $\boldsymbol{x}(\boldsymbol{\beta})$ with respect to θ, evaluated at the parameters $\boldsymbol{\beta}_0$, and by $\boldsymbol{M}_X$ the projection off all the columns of $\boldsymbol{X}(\boldsymbol{\beta})$ other than $\boldsymbol{X}_\theta$, then the asymptotic variance of the least squares estimator $\hat{\theta}$ is $\sigma_0^2(\boldsymbol{X}_\theta^\top \boldsymbol{M}_X \boldsymbol{X}_\theta)^{-1}$, where σ_0^2 is the variance of the components of $\boldsymbol{u}$. If we consider a DGP with a parameter $\theta \neq \theta_0$, then for a given sample size n, the parameter δ of the drifting DGP becomes $n^{1/2}(\theta - \theta_0)$, and $\Lambda = \lambda^2$ becomes

$$\Lambda = \frac{1}{\sigma_0^2}(\theta - \theta_0)^2 \boldsymbol{X}_\theta^\top \boldsymbol{M}_X \boldsymbol{X}_\theta. \tag{12.38}$$

This may be compared with the general expression (12.26). Now let $\theta(\alpha, \pi)$ be the value of θ that makes Λ in (12.38) equal to $(\lambda(\alpha, \pi))^2$ as given above by the inverse power function. We see that, within an asymptotic approximation, DGPs with values of θ closer to the θ_0 of the null hypothesis than $\theta(\alpha, \pi)$ will have probability less than π of rejecting the null on a test of size α.

We should be unwilling to regard a failure to reject the null as evidence against some other DGP or set of DGPs if, under the latter, there is not a fair probability of rejecting the null. What do we mean by a "fair probability" here? Some intuition on this matter can be obtained by considering what we would learn in the present context by using a standard tool of conventional statistical inference, namely, a confidence interval. Armed with the estimate $\hat{\theta}$ and an estimate of its standard error, $\hat{\sigma}_\theta$, we can form a confidence interval

If we write the LR statistic (13.08) in partitioned form, we obtain

$$
\begin{aligned}
LR &= (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^{\top} \boldsymbol{I} (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \\
&= \begin{bmatrix} \tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1 \\ \tilde{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_2 \end{bmatrix}^{\top} \begin{bmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1 \\ \tilde{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_2 \end{bmatrix} \\
&= (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^{\top} \boldsymbol{I}_{11} (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) - 2(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^{\top} \boldsymbol{I}_{12} \hat{\boldsymbol{\theta}}_2 + \hat{\boldsymbol{\theta}}_2^{\top} \boldsymbol{I}_{22} \hat{\boldsymbol{\theta}}_2 .
\end{aligned}
$$

where the last line uses the fact that $\tilde{\boldsymbol{\theta}}_2 = \boldsymbol{0}$. Making use of the result (13.11), the LR statistic can then be rewritten as

$$
\begin{aligned}
LR &= (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^{\top} \boldsymbol{I}_{11} (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) - 2(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^{\top} \boldsymbol{I}_{11} (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) + \hat{\boldsymbol{\theta}}_2^{\top} \boldsymbol{I}_{22} \hat{\boldsymbol{\theta}}_2 \\
&= \hat{\boldsymbol{\theta}}_2^{\top} \boldsymbol{I}_{22} \hat{\boldsymbol{\theta}}_2 - (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^{\top} \boldsymbol{I}_{11} (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) .
\end{aligned} \tag{13.12}
$$

We now show that the Wald statistic is equal to (13.12). Since the restrictions take the form (13.09), we see that $\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{\theta}_2$ and $\hat{\boldsymbol{r}} = \hat{\boldsymbol{\theta}}_2$. This implies that the matrix $\boldsymbol{R}$ can be written as

$$
\boldsymbol{R}(\boldsymbol{\theta}) = [\boldsymbol{0} \quad \boldsymbol{I}],
$$

where the $\boldsymbol{0}$ matrix is $r \times (k - r)$, and the identity matrix $\boldsymbol{I}$ is $r \times r$. Then the expression $\hat{\boldsymbol{R}} \boldsymbol{I}^{-1} \hat{\boldsymbol{R}}^{\top}$ that appears in the Wald statistic (13.05) is just the $(2, 2)$ block of the inverse matrix $\boldsymbol{I}^{-1}$. By the results in Appendix A on partitioned matrices, we obtain

$$
\left(\hat{\boldsymbol{R}} \boldsymbol{I}^{-1} \hat{\boldsymbol{R}}^{\top} \right)^{-1} = \left((\boldsymbol{I}^{-1})_{22} \right)^{-1} = \boldsymbol{I}_{22} - \boldsymbol{I}_{21} \boldsymbol{I}_{11}^{-1} \boldsymbol{I}_{12} . \tag{13.13}
$$

This result allows us to put (13.05) in the form

$$
W = \hat{\boldsymbol{\theta}}_2^{\top} \left(\boldsymbol{I}_{22} - \boldsymbol{I}_{21} \boldsymbol{I}_{11}^{-1} \boldsymbol{I}_{12} \right) \hat{\boldsymbol{\theta}}_2 .
$$

By (13.11), this last expression is equal to

$$
\hat{\boldsymbol{\theta}}_2^{\top} \boldsymbol{I}_{22} \hat{\boldsymbol{\theta}}_2 - \left(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1 \right)^{\top} \boldsymbol{I}_{11} \left(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1 \right),
$$

which is the same as (13.12). The proof of the equality of the three classical statistics for the quadratic loglikelihood function (13.07) is therefore complete.

It is of interest to see how the three classical test statistics are related geometrically. Figure 13.1 depicts the graph of a loglikelihood function $\ell(\boldsymbol{y}, \theta_1, \theta_2)$. It is drawn for a *given* sample vector $\boldsymbol{y}$ and consequently a given sample size n. For simplicity, the parameter space has been supposed to be two-dimensional. There is only one restriction, which is that the second component of the parameter vector, θ_2, is equal to zero. Therefore, the function ℓ can be treated as a function of the two variables θ_1 and θ_2 only, and its

Thus we write for observation t

$$\ell_t = \ell_t(\boldsymbol{y}^t, \boldsymbol{\theta}_0) + n^{-1/2} a_t(\boldsymbol{y}^t). \qquad (13.27)$$

We can see from this that the log of the density of the tth observation is taken to be as given by a parametrized model for a parameter vector $\boldsymbol{\theta}_0$ satisfying the restrictions of the null hypothesis, plus a term that vanishes with $n^{-1/2}$ as $n \to \infty$. The fact that any density function is normalized so as to integrate to unity means that the functions a_t in (13.27) must be chosen so as to obey the normalization condition

$$\int \exp\big(\ell_t + n^{-1/2} a_t\big) dy_t = 1.$$

It can readily be shown that this implies that

$$E_0\big(a_t(\boldsymbol{y}^t)\big) = O(n^{-1/2}), \qquad (13.28)$$

where E_0 denotes an expectation calculated using $\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta}_0)$ as log density. To leading order asymptotically, then, the random variables a_t have mean zero.

The fact that ℓ_t is written in (13.27) as the sum of two terms does not restrict the applicability of the analysis at all, because one can think of (13.27) as arising from a first-order Taylor-series approximation to any drifting DGP. An example would be the sequence of local alternatives

$$\ell_t\big(\boldsymbol{y}^t, \boldsymbol{\theta}_0 + n^{-1/2} \boldsymbol{\delta}\big).$$

By arguments similar to those of Section 12.3, one can show that a Taylor-series approximation to this can be written in the form of (13.27).

We will now state without proof the results that correspond to equations (12.11), (12.12), and (12.13) in the NLS context. They are discussed and proved in Davidson and MacKinnon (1987), a paper that many readers may, however, find somewhat difficult because of the nature of the mathematics employed. These results provide asymptotically valid expressions for the various ingredients of the classical test statistics under the drifting DGP specified by (13.27). The first result is that the estimators $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are still root-n consistent for $\boldsymbol{\theta}_0$:

$$\ddot{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + O(n^{-1/2}),$$

from which we may conclude that $\ddot{\mathfrak{I}}$ and $\ddot{\boldsymbol{R}}$ are consistent for $\mathfrak{I}_0$ and $\boldsymbol{R}_0$, just as they are under the null hypothesis:

$$\ddot{\mathfrak{I}} = \mathfrak{I}_0 + O(n^{-1/2}); \quad \text{and} \quad \ddot{\boldsymbol{R}} = \boldsymbol{R}_0 + O(n^{-1/2}).$$

We may also conclude from the consistency of $\ddot{\boldsymbol{\theta}}$ that all the Taylor expansions used in developing equations (13.23), (13.25), and (13.26) are still valid, as are these equations themselves.

By the FWL Theorem,

$$\hat{\boldsymbol{\beta}}_2 = \big(\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2\big)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{y} \quad \text{and}$$

$$\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\big)_{22} = \big(\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2\big)^{-1}.$$

Thus (13.41) becomes

$$W = n\left(\frac{\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{M}_X \boldsymbol{y}}\right) = n\left(\frac{\boldsymbol{y}^\top \boldsymbol{P}_{M_1 X_2} \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{M}_X \boldsymbol{y}}\right).$$

From (13.37) and (13.39), we obtain

$$W = \left(\frac{rn}{n-k}\right) F; \quad LR = n \log\left(1 + \frac{W}{n}\right). \tag{13.42}$$

Since W is equal to $n/(n-k)$ times rF, it is evident that

$$W = rF + O(n^{-1}).$$

Finally, we turn to the LM statistic. We first observe from (8.83) that the gradient with respect to the regression parameters $\boldsymbol{\beta}$ of the loglikelihood function for a linear regression model with normal errors is

$$\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^2} \sum_{t=1}^{n} \boldsymbol{X}_t^\top (y_t - \boldsymbol{X}_t \boldsymbol{\beta}) = \sigma^{-2} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Thus, from (13.03), the LM statistic is

$$\begin{aligned}
LM &= \tilde{\boldsymbol{g}}_2^\top \big(\tilde{\boldsymbol{I}}^{-1}\big)_{22} \tilde{\boldsymbol{g}}_2 \\
&= \tilde{\sigma}^{-4} (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top \boldsymbol{X}_2 \big(\tilde{\sigma}^2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1}\big) \boldsymbol{X}_2^\top (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).
\end{aligned} \tag{13.43}$$

Since the LM test is based on the restricted model (13.35), we use the ML estimate of σ from that model:

$$\tilde{\sigma}^2 = \tfrac{1}{n} \boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{y}.$$

Substituting this into (13.43), we see that

$$\begin{aligned}
LM &= n\left(\frac{\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{y}}\right) \\
&= n\left(\frac{\boldsymbol{y}^\top \boldsymbol{P}_{M_1 X_2} \boldsymbol{y} / \boldsymbol{y}^\top \boldsymbol{M}_X \boldsymbol{y}}{1 + \boldsymbol{y}^\top \boldsymbol{P}_{M_1 X_2} \boldsymbol{y} / \boldsymbol{y}^\top \boldsymbol{M}_X \boldsymbol{y}}\right) \\
&= n\left(\frac{rF}{n - k + rF}\right).
\end{aligned} \tag{13.44}$$

from which we may derive that

$$e^{2\hat{\tau}} = \hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} y_t^2. \tag{13.50}$$

For this parametrization, the information matrix, which has only one element, is constant and equal to 2:

$$\mathcal{I} = -\frac{1}{n}E(D_\tau^2 \ell) = \frac{2}{n}\sum_{t=1}^{n} e^{-2\tau}E(y_t^2) = 2.$$

Notice that, although $\mathcal{I}$ is constant, the loglikelihood function is *not* a quadratic function of τ. We now consider various classical tests for the null hypothesis that $\tau = 0$, or, equivalently, that $\sigma^2 = 1$. Despite the simplicity of this example, we will uncover a bewildering variety of test statistics.

Initially, we will work with the τ parametrization. It is not necessary to do any estimation at all in order to find restricted estimates, since $\tilde{\tau} = 0$. For the Wald and LR tests we need to find $\hat{\tau}$. From (13.50), it is

$$\hat{\tau} = \frac{1}{2}\log\left(\frac{1}{n}\sum_{t=1}^{n} y_t^2\right).$$

The restricted "maximum" of the loglikelihood function is just the value of the function at $\tau = 0$:

$$\tilde{\ell} = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n} y_t^2 = -\frac{n}{2}\log 2\pi - \frac{n}{2}e^{2\hat{\tau}}. \tag{13.51}$$

Although this is the restricted maximum, it is convenient to express it, as we have done here, in terms of the unrestricted estimate, $\hat{\tau}$. The unrestricted maximum, $\hat{\ell}$, is given by

$$-\frac{n}{2}\log 2\pi - n\hat{\tau} - \frac{1}{2}e^{-2\hat{\tau}}\sum_{t=1}^{n} y_t^2 = -\frac{n}{2}\log 2\pi - n\hat{\tau} - \frac{n}{2}, \tag{13.52}$$

where the equality uses (13.50).

We may proceed at once to obtain the LR statistic, which is twice the difference between (13.52) and (13.51):

$$LR = 2(\hat{\ell} - \tilde{\ell}) = n(e^{2\hat{\tau}} - 1 - 2\hat{\tau})$$
$$= 2n\hat{\tau}^2 + o(1). \tag{13.53}$$

The second line of (13.53) is a Taylor expansion of the statistic in powers of $\hat{\tau}$. This is of interest because, *under the null hypothesis*, we expect $\hat{\tau}$, which is

both the estimate itself and the *difference* between the estimate and the true value of the parameter, to be of order $n^{-1/2}$. It follows that $2n\hat{\tau}^2$ will be of order unity and that higher terms in the expansion of the exponential function in (13.53) will be of lower order. Thus, if the various forms of the classical test do indeed yield asymptotically equal expressions, we may expect that the leading term of all of them will be $2n\hat{\tau}^2$.

Let us next consider the LM statistic. The essential piece of it is the derivative of the loglikelihood function (13.49) with respect to τ, evaluated at $\tau = 0$. We find that

$$\frac{\partial \ell}{\partial \tau} = -n + e^{-2\tau} \sum_{t=1}^{n} y_t^2 \quad \text{and} \quad \frac{\partial \ell}{\partial \tau}\bigg|_{\tau=0} = n\big(e^{2\hat{\tau}} - 1\big). \qquad (13.54)$$

If for the variance of $\partial \ell/\partial \tau$ we use n times the true, constant, value of the single element of the information matrix, 2, the LM statistic is the square of $(\partial \ell/\partial \tau)|_{\tau=0}$, given by (13.54), divided by $2n$:

$$LM_1 = \frac{n}{2}\big(e^{2\hat{\tau}} - 1\big)^2 = 2n\hat{\tau}^2 + o(1).$$

This variant of the LM statistic has the same leading term as the LR statistic (13.53) but will of course differ from it in finite samples.

Instead of the true information matrix, an investigator might prefer to use the negative of the empirical Hessian to estimate the information matrix; see equations (8.47) and (8.49). Because the loglikelihood function is not exactly quadratic, this estimator does *not* coincide numerically with the true value. Since

$$\frac{\partial^2 \ell}{\partial \tau^2} = -2e^{-2\tau} \sum_{t=1}^{n} y_t^2, \qquad (13.55)$$

which at $\tau = 0$ is $-2ne^{2\hat{\tau}}$, the LM test calculated in this fashion is

$$LM_2 = \frac{n}{2}e^{-2\hat{\tau}}\big(e^{2\hat{\tau}} - 1\big)^2 = 2n\hat{\tau}^2 + o(1). \qquad (13.56)$$

The leading term is as in LR and LM_1, but LM_2 will differ from both those statistics in finite samples.

Another possibility is to use the OPG estimator of the information matrix; see equations (8.48) and (8.50). This estimator is

$$\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial \ell}{\partial \tau}\right)^2 = \frac{1}{n}\sum_{t=1}^{n}\big(y_t^2 e^{-2\tau} - 1\big)^2,$$

which, when evaluated at $\tau = 0$, is equal to

$$\frac{1}{n}\sum_{t=1}^{n}\big(y_t^2 - 1\big)^2.$$

both the estimate itself and the *difference* between the estimate and the true value of the parameter, to be of order $n^{-1/2}$. It follows that $2n\hat{\tau}^2$ will be of order unity and that higher terms in the expansion of the exponential function in (13.53) will be of lower order. Thus, if the various forms of the classical test do indeed yield asymptotically equal expressions, we may expect that the leading term of all of them will be $2n\hat{\tau}^2$.

Let us next consider the LM statistic. The essential piece of it is the derivative of the loglikelihood function (13.49) with respect to τ, evaluated at $\tau = 0$. We find that

$$\frac{\partial \ell}{\partial \tau} = -n + e^{-2\tau} \sum_{t=1}^{n} y_t^2 \quad \text{and} \quad \frac{\partial \ell}{\partial \tau}\Big|_{\tau=0} = n\big(e^{2\hat{\tau}} - 1\big). \qquad (13.54)$$

If for the variance of $\partial \ell/\partial \tau$ we use n times the true, constant, value of the single element of the information matrix, 2, the LM statistic is the square of $(\partial \ell/\partial \tau)|_{\tau=0}$, given by (13.54), divided by $2n$:

$$LM_1 = \frac{n}{2}\big(e^{2\hat{\tau}} - 1\big)^2 = 2n\hat{\tau}^2 + o(1).$$

This variant of the LM statistic has the same leading term as the LR statistic (13.53) but will of course differ from it in finite samples.

Instead of the true information matrix, an investigator might prefer to use the negative of the empirical Hessian to estimate the information matrix; see equations (8.47) and (8.49). Because the loglikelihood function is not exactly quadratic, this estimator does *not* coincide numerically with the true value. Since

$$\frac{\partial^2 \ell}{\partial \tau^2} = -2e^{-2\tau} \sum_{t=1}^{n} y_t^2, \qquad (13.55)$$

which at $\tau = 0$ is $-2ne^{2\hat{\tau}}$, the LM test calculated in this fashion is

$$LM_2 = \frac{n}{2}e^{-2\hat{\tau}}\big(e^{2\hat{\tau}} - 1\big)^2 = 2n\hat{\tau}^2 + o(1). \qquad (13.56)$$

The leading term is as in LR and LM_1, but LM_2 will differ from both those statistics in finite samples.

Another possibility is to use the OPG estimator of the information matrix; see equations (8.48) and (8.50). This estimator is

$$\frac{1}{n} \sum_{t=1}^{n} \left(\frac{\partial \ell_t}{\partial \tau}\right)^2 = \frac{1}{n} \sum_{t=1}^{n} \big(y_t^2 e^{-2\tau} - 1\big)^2,$$

which, when evaluated at $\tau = 0$, is equal to

$$\frac{1}{n} \sum_{t=1}^{n} \big(y_t^2 - 1\big)^2.$$

This expression cannot even be expressed as a function of $\hat{\tau}$ alone. To obtain an expansion of the test statistic that makes use of it, we must make use of the property of the normal distribution which tells us that $E(y_t^4) = 3\sigma^4$, or, in terms of τ, $3e^{4\tau}$.[4] Using this property, we can invoke a law of large numbers and conclude that the OPG information matrix estimator is indeed equal to $2 + o(1)$ at $\tau = 0$. Thus the third variant of the LM test statistic is

$$LM_3 = \frac{n^2\left(e^{2\hat{\tau}} - 1\right)^2}{\sum_{t=1}^{n}\left(y_t^2 - 1\right)^2} = 2n\hat{\tau}^2 + o(1).$$

Once again, the leading term is $2n\hat{\tau}^2$, but the form of LM_3 is otherwise quite different from that of LM_1 or LM_2.

Just as there are various forms of the LM test, so are there various forms of the Wald test. Any one of these may be formed by combining the unrestricted estimate $\hat{\tau}$ with some estimate of the information matrix, which in this case is actually a scalar. The simplest choice is just the true information matrix, that is, 2. With this we obtain

$$W_1 = 2n\hat{\tau}^2. \tag{13.57}$$

It is easy to see that W_2, which uses the empirical Hessian, is identical to W_1, because (13.55) evaluated at $\tau = \hat{\tau}$ is just $-2n$. On the other hand, use of the OPG estimator yields

$$W_3 = \hat{\tau}^2 \sum_{t=1}^{n}\left(y_t^2 e^{-2\hat{\tau}} - 1\right)^2,$$

which is quite different from W_1 and W_2.

All of the above test statistics were based on τ as the single parameter of the model, but we could just as well use σ or σ^2 as the model parameter. Ideally, we would like test statistics to be invariant to such reparametrizations. The LR statistic is always invariant, since $\hat{\ell}$ and $\tilde{\ell}$ do not change when the model is reparametrized. But all forms of the Wald statistic, and some forms of the LM statistic, are in general not invariant, as we now illustrate.

Suppose we take σ^2 to be the parameter of the model. The information matrix is not constant in this new parametrization, and so we must evaluate it at the *estimate* $\hat{\sigma}^2$. It is easy to see that the information matrix, as a

---

[4] Note that it was *not* necessary to use special properties of the normal distribution in order to expand the previous statistics, which were in fact all functions of one and only one random variable, namely $\hat{\tau}$. In general, in less simple situations, this agreeable feature of the present example is absent and special properties must be invoked in order to discover the behavior of all the various test statistics.

function of σ^2, is $1/(2\sigma^4)$. If we use this expression for the information matrix, evaluated at $\hat{\sigma}^2$, the Wald test becomes

$$W_1 = \frac{n}{2}\hat{\sigma}^{-4}(\hat{\sigma}^2 - 1)^2 = \frac{n}{2}e^{-4\hat{\tau}}(e^{2\hat{\tau}} - 1)^2 = 2n\hat{\tau}^2 + o(1).$$

Since this differs from (13.57), we have shown that different parametrizations lead to numerically different Wald statistics even if the true information matrix, evaluated at the MLE of the model parameter, is used in both cases.

As we will see in the next section, the LM test is invariant if it is based on the true information matrix evaluated at the MLE. But if some other information matrix estimator is used, the LM test can also be parametrization dependent. Suppose that we use the empirical Hessian. From (13.48), the first two derivatives of ℓ with respect to σ^2, evaluated at $\sigma^2 = 1$, are

$$\frac{\partial \ell}{\partial \sigma^2}\bigg|_{\sigma^2=1} = -\frac{1}{2}\left(n - \sum_{t=1}^{n} y_t^2\right) = \frac{n}{2}(e^{2\hat{\tau}} - 1) \quad \text{and}$$

$$\frac{\partial^2 \ell}{(\partial \sigma^2)^2}\bigg|_{\sigma^2=1} = \frac{n}{2}(1 - 2e^{2\hat{\tau}}).$$

From this, we find that the statistic LM_2 calculated as was (13.56) but for the σ^2 parametrization, is

$$LM_2 = \frac{n(e^{2\hat{\tau}} - 1)^2}{2(2e^{2\hat{\tau}} - 1)} = 2n\hat{\tau}^2 + o(1). \tag{13.58}$$

The leading term is correct, as it must be, but (13.58) is numerically different from (13.56).

Plainly, there are still more forms of both the LM and Wald tests, some but not all of which will coincide with one of the versions we have already computed. The interested reader is invited to try out, for example, the effects of using σ itself, rather than σ^2, as the model parameter.

This example illustrates the fact that there may be many different classical tests, which are numerically different but asymptotically equivalent. The fact that there are so many different tests creates the problem of how to choose among them. One would prefer to use tests that are easy to compute and for which the finite-sample distribution is well approximated by the asymptotic distribution. Unfortunately, it frequently requires considerable effort to determine the finite-sample properties of asymptotic tests. Any method of analysis tends to be restricted to very special cases, such as the case of linear regression models with normal errors discussed in Section 13.4. One generally applicable approach is to use computer simulation (Monte Carlo experiments); see Chapter 21.

statistic will be the same. This result assumes that we are using the efficient score form of the LM test. If we based the test on estimates of the information matrix, the two LM statistics might not be numerically the same, although they would still be the same asymptotically.

Geometrically, two different alternative hypotheses are locally equivalent if they **touch** at the null hypothesis. By this we mean not merely that the two alternative hypotheses yield the same values of their respective loglikelihood functions when restricted by the null hypothesis, as will always be the case, but also that the gradients of the two loglikelihood functions are the same, since the gradients are *tangents* to the two models that touch at the null model. In these circumstances, the two LM tests must be numerically identical.

What does it mean for two models to touch, or, to use the nongeometrical term for the property, to be locally equivalent? A circular definition would simply be that their gradients are the same at all DGPs at which the two models intersect. Statistically, it means that if one departs only slightly from the null hypothesis while respecting one of the two alternative hypotheses, then one departs from the other alternative hypothesis by an amount that is of the second order of small quantities. For instance, an AR(1) process characterized by a small autoregressive parameter ρ differs from some MA(1) process to an extent proportional only to ρ^2. To prove this formally would entail a formal definition of the distance between two DGPs, but our earlier circular definition is an operational one: If the gradient $\tilde{\boldsymbol{g}}^1$ calculated for the first alternative is the same as the gradient $\tilde{\boldsymbol{g}}^2$ for the second, then the two alternatives touch at the null. It should now be clear that this requirement is too strong: It is enough if the components of $\tilde{\boldsymbol{g}}^2$ are all linear combinations of those of $\tilde{\boldsymbol{g}}^1$ and vice versa. An example of this last possibility is provided by the local equivalence, around the null of white noise errors, of regression models with ARMA(p, q) errors on the one hand and with AR$(\max(p, q))$ errors on the other; see Section 10.8. For more examples, see Godfrey (1981) and Godfrey and Wickens (1982).

Both the geometrical and algebraic aspects of the invariance of LM tests under local equivalence are expressed by means of one simple remark: The LM test can be constructed solely on the basis of the restricted ML estimates and the *first* derivatives of the loglikelihood function evaluated at those estimates. This implies that the LM test takes no account of the curvature of the alternative hypothesis near the null.

We may summarize the results of this section as follows:

1. The LR test depends only on two maximized loglikelihood functions. It therefore cannot depend either on the parametrization of the model or on the way in which the restrictions are formulated in terms of those parameters.

2. The efficient score form of the LM test is constructed out of two ingredients, the gradient and the information matrix, which do alter under

can be used with any model estimated by maximum likelihood. The OPG regression was first used as a means of computing test statistics by Godfrey and Wickens (1981). This artificial regression, which is very easy indeed to set up for most models estimated by maximum likelihood, can be used for the same purposes as the GNR: verification of first-order conditions for the maximization of the loglikelihood function, covariance matrix estimation, one-step efficient estimation, and, of greatest immediate interest, the computation of test statistics.

Suppose that we are interested in the parametrized model (13.01). Let $G(\boldsymbol{\theta})$ be the CG matrix associated with the loglikelihood function $\ell^n(\boldsymbol{\theta})$, with typical element

$$G_{ti}(\boldsymbol{\theta}) \equiv \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i}; \quad t = 1, \ldots, n, \; i = 1, \ldots, k,$$

where k is the number of elements in the parameter vector $\boldsymbol{\theta}$. Then the OPG regression associated with the model (13.01) can be written as

$$\boldsymbol{\iota} = G(\boldsymbol{\theta})\boldsymbol{c} + \text{residuals}. \tag{13.81}$$

Here $\boldsymbol{\iota}$ is an n–vector of which each element is unity and $\boldsymbol{c}$ is a k–vector of artificial parameters. The product of the matrix of regressors with the regressand is the gradient $\boldsymbol{g}(\boldsymbol{\theta}) \equiv G^{\top}(\boldsymbol{\theta})\boldsymbol{\iota}$. The matrix of sums of squares and cross-products of the regressors, $G^{\top}(\boldsymbol{\theta})G(\boldsymbol{\theta})$, when divided by n, consistently estimates the information matrix $\mathfrak{I}(\boldsymbol{\theta})$. These two features are essentially all that is required for (13.81) to be a valid artificial regression.[6] As with the GNR, the regressors of the OPG regression depend on the vector $\boldsymbol{\theta}$. Therefore, before the artificial regression is run, these regressors must be evaluated at some chosen parameter vector.

One possible choice for this parameter vector is $\hat{\boldsymbol{\theta}}$, the ML estimator for the model (13.01). In this case, the regressor matrix is $\hat{G} \equiv G(\hat{\boldsymbol{\theta}})$ and the artificial parameter estimates, which we will denote by $\hat{\boldsymbol{c}}$, are identically zero:

$$\hat{\boldsymbol{c}} = \left(\hat{G}^{\top}\hat{G}\right)^{-1}\hat{G}^{\top}\boldsymbol{\iota} = \left(\hat{G}^{\top}\hat{G}\right)^{-1}\hat{\boldsymbol{g}} = \boldsymbol{0}.$$

Since $\hat{\boldsymbol{g}}$ here is the gradient of the loglikelihood function evaluated at $\hat{\boldsymbol{\theta}}$, the last equality above is a consequence of the first-order conditions for the maximum of the likelihood. As with the GNR, then, running the OPG regression with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ provides a simple way to test how well the first-order conditions are in fact satisfied by a set of estimates calculated by means of some computer program. The t statistics again provide the most suitable check. They should not exceed a number around 10^{-2} or 10^{-3} in absolute value if a good approximation to the maximum has been found.

[6] Precise conditions for a regression to be called "artificial" are provided by Davidson and MacKinnon (1990); see Section 14.4.

Since the estimates $\hat{\boldsymbol{c}}$ for regression (13.81) are zero when the regressors are $\hat{\boldsymbol{G}}$, those regressors have no explanatory power for $\boldsymbol{\iota}$, and the sum of squared residuals is therefore equal to the total sum of squares. Because the latter is

$$\boldsymbol{\iota}^{\top}\boldsymbol{\iota} = \sum_{t=1}^{n} 1 = n,$$

the ML estimate of the residual variance in (13.81) is just unity:

$$\tfrac{1}{n}\text{SSR} = \tfrac{1}{n}\boldsymbol{\iota}^{\top}\boldsymbol{\iota} = \tfrac{1}{n}n = 1.$$

The OLS variance estimate, which is $\text{SSR}/(n-k) = n/(n-k)$, is asymptotically equivalent to this, but it will simplify the exposition if we suppose that the ML estimate is used. The covariance matrix estimate for the vector $\hat{\boldsymbol{c}}$ from (13.81) is then

$$\left(\hat{\boldsymbol{G}}^{\top}\hat{\boldsymbol{G}}\right)^{-1}.$$

It is this expression that gives the OPG regression its name, for its inverse is precisely the OPG estimator of the information matrix; see (8.48) and (8.50).[7] It follows that, as with the GNR, n^{-1} times the covariance matrix estimator from the OPG regression is asymptotically equal to the covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

The property just established is not the only one shared by the Gauss-Newton and OPG regressions. We will now establish two further properties of the OPG regression that are in fact shared by all regressions to which we give the name "artificial." The first of these properties is what allows one to use artificial regressions to perform one-step efficient estimation. According to this property, if the OPG regression (13.81) is evaluated at some parameter vector $\acute{\boldsymbol{\theta}}$ that is root-n consistent for $\boldsymbol{\theta}_0$, so that $\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O(n^{-1/2})$, then the artificial parameter estimates $\acute{\boldsymbol{c}}$ are such that

$$n^{1/2}\acute{\boldsymbol{c}} \overset{a}{=} n^{1/2}(\hat{\boldsymbol{\theta}} - \acute{\boldsymbol{\theta}}), \tag{13.82}$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$. This result is essentially the same as the one proved for the Gauss-Newton regression in Section 6.6.

The result (13.82) is important. Because of it, we can proceed in one step from any root-n consistent estimator $\acute{\boldsymbol{\theta}}$ to an estimator asymptotically equivalent to the asymptotically efficient estimator $\hat{\boldsymbol{\theta}}$. The one-step estimator $\grave{\boldsymbol{\theta}}$ defined by $\grave{\boldsymbol{\theta}} \equiv \acute{\boldsymbol{\theta}} + \acute{\boldsymbol{c}}$ has the property that

$$n^{1/2}(\grave{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o(1), \tag{13.83}$$

---

[7] As we noted in Section 8.6, some authors refer to the OPG estimator of the information matrix as the BHHH estimator, after Berndt, Hall, Hall, and Hausman (1974), who advocated its use, although they did not explicitly make use of the OPG regression itself.
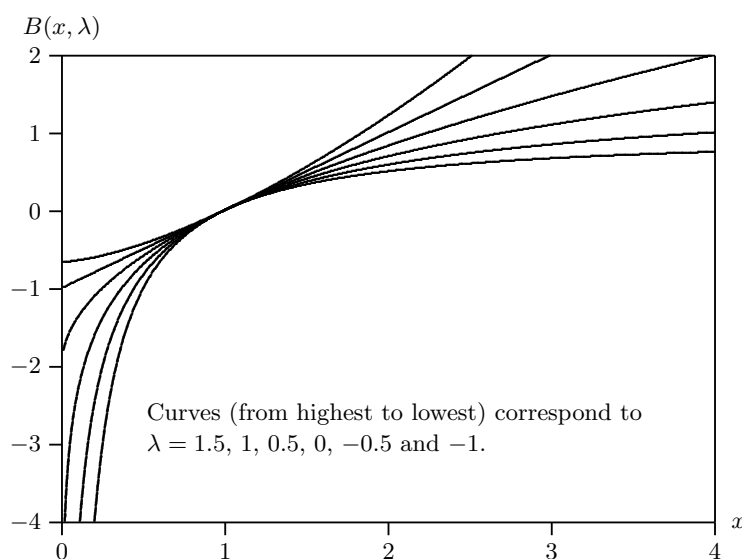
Figure 14.1 Box-Cox transformations for various values of λ

the regressors include a constant term, subjecting the dependent variable to a Box-Cox transformation with $\lambda = 1$ is equivalent to not transforming it at all. Subjecting it to a Box-Cox transformation with $\lambda = 0$ is equivalent to using $\log y_t$ as the regressand. Since these are both very plausible special cases, it is attractive to use a transformation that allows for both of them. Even when it is not considered plausible in its own right, the conventional Box-Cox model provides a convenient alternative against which to test the specification of linear and loglinear regression models; see Section 14.6.

The Box-Cox transformation is not without some serious disadvantages, however. Consider the simple Box-Cox model

$$B(y_t, \lambda) = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \tag{14.07}$$

For most values of λ (but not for $\lambda = 0$ or $\lambda = 1$) the value of $B(y_t, \lambda)$ is bounded either from below or above; specifically, when $\lambda > 0$, $B(y_t, \lambda)$ cannot be less than $-1/\lambda$ and, when $\lambda < 0$, $B(y_t, \lambda)$ cannot be greater than $-1/\lambda$. However, if u_t is normally distributed, the right-hand side of (14.07) is not bounded and could, at least in principle, take on arbitrarily large positive or negative values. Thus, strictly speaking, (14.07) is logically impossible as a model for y_t. This remains true if we replace $x_t(\boldsymbol{\beta})$ by a regression function that depends on λ.

One way to deal with this problem is to assume that data on y_t are observed only when the bounds are not violated, as in Poirier (1978b) and Poirier and Ruud (1979). This leads to loglikelihood functions similar to

The fundamental result that makes the DLR possible is that, for this class of models, the information matrix $\mathfrak{I}(\boldsymbol{\theta})$ satisfies the equality

$$\mathfrak{I}(\boldsymbol{\theta}) = \plim_{n\to\infty} \left(\tfrac{1}{n} \left(\boldsymbol{F}^\top(\boldsymbol{y},\boldsymbol{\theta}) \boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) + \boldsymbol{K}^\top(\boldsymbol{y},\boldsymbol{\theta}) \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \right) \right) \qquad (14.20)$$

and so can be consistently estimated by

$$\tfrac{1}{n} \left(\boldsymbol{F}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}}) \boldsymbol{F}(\boldsymbol{y},\ddot{\boldsymbol{\theta}}) + \boldsymbol{K}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}}) \boldsymbol{K}(\boldsymbol{y},\ddot{\boldsymbol{\theta}}) \right), \qquad (14.21)$$

where $\ddot{\boldsymbol{\theta}}$ is any consistent estimator of $\boldsymbol{\theta}$. We are interested in the implications of (14.20) rather than how it is derived. The derivation makes use of some rather special properties of the normal distribution and may be found in Davidson and MacKinnon (1984a).

The principal implication of (14.20) is that a certain artificial regression, which we call the DLR, has all the properties that we expect an artificial regression to have. The DLR may be written as

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{\iota} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix} \boldsymbol{b} + \text{residuals.} \qquad (14.22)$$

This artificial regression has $2n$ **artificial observations**. The regressand is $f_t(y_t,\boldsymbol{\theta})$ for observation t and unity for observation $t+n$, and the regressors corresponding to $\boldsymbol{\theta}$ are $-\boldsymbol{F}_t(\boldsymbol{y},\boldsymbol{\theta})$ for observation t and $\boldsymbol{K}_t(\boldsymbol{y},\boldsymbol{\theta})$ for observation $t+n$, where $\boldsymbol{F}_t$ and $\boldsymbol{K}_t$ denote, respectively, the tth rows of $\boldsymbol{F}$ and $\boldsymbol{K}$. Intuitively, the reason we need a double-length regression here is that each genuine observation makes two contributions to the loglikelihood function: a sum-of-squares term $-\tfrac{1}{2} f_t^2$ and a Jacobian term k_t. As a result, the gradient and the information matrix each involve two parts as well, and the way to take both of these into account is to incorporate two artificial observations into the artificial regression for each genuine one.

Why is (14.22) a valid artificial regression? As we noted when we discussed the OPG regression in Section 13.7, there are two principal conditions that an artificial regression must satisfy. It is worth stating these conditions somewhat more formally here.[4] Let $\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta})$ denote the regressand for some artificial regression and let $\boldsymbol{R}(\boldsymbol{y},\boldsymbol{\theta})$ denote the matrix of regressors. Let the number of rows of both $\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta})$ and $\boldsymbol{R}(\boldsymbol{y},\boldsymbol{\theta})$ be n^*, which will generally be either n or an integer multiple of n. The regression of $\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta})$ on $\boldsymbol{R}(\boldsymbol{y},\boldsymbol{\theta})$ will have the properties of an artificial regression if

$$\boldsymbol{R}^\top(\boldsymbol{y},\boldsymbol{\theta}) \boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta}) = \rho(\boldsymbol{\theta}) \boldsymbol{g}(\boldsymbol{y},\boldsymbol{\theta}) \quad \text{and} \qquad (14.23)$$

$$\plim_{n\to\infty} \boldsymbol{\theta} \left(\tfrac{1}{n} \boldsymbol{R}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}}) \boldsymbol{R}(\boldsymbol{y},\ddot{\boldsymbol{\theta}}) \right) = \rho(\boldsymbol{\theta}) \mathfrak{I}(\boldsymbol{\theta}), \qquad (14.24)$$

---

[4] For a fuller treatment of this topic, see Davidson and MacKinnon (1990).

where $\ddot{\boldsymbol{\theta}}$ denotes any consistent estimator of $\boldsymbol{\theta}$. The notation $\text{plim}_{\boldsymbol{\theta}}$ indicates, as usual, that the probability limit is being taken under the DGP characterized by the parameter vector $\boldsymbol{\theta}$, and $\rho(\boldsymbol{\theta})$ is a scalar defined as

$$\rho(\boldsymbol{\theta}) \equiv \underset{n\to\infty}{\text{plim}_{\boldsymbol{\theta}}}\left(\frac{1}{n^*}\,\boldsymbol{r}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta})\right).$$

Because $\rho(\boldsymbol{\theta})$ is equal to unity for both the OPG regression and the DLR, those two artificial regressions satisfy the simpler conditions

$$\boldsymbol{R}^\top(\boldsymbol{y},\boldsymbol{\theta})\,\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{y},\boldsymbol{\theta}) \quad\text{and} \tag{14.25}$$

$$\underset{n\to\infty}{\text{plim}_{\boldsymbol{\theta}}}\left(\frac{1}{n}\,\boldsymbol{R}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\boldsymbol{R}(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\right) = \mathfrak{I}(\boldsymbol{\theta}), \tag{14.26}$$

as well as the original conditions (14.23) and (14.24). However, these simpler conditions are not satisfied by the GNR and are thus evidently too simple in general.

It is now easy to see that the DLR (14.21) satisfies conditions (14.25) and (14.26). For the first of these, simple calculation shows that

$$\begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{f}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{\iota} \end{bmatrix} = -\boldsymbol{F}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{y},\boldsymbol{\theta}) + \boldsymbol{K}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{\iota},$$

which by (14.19) is equal to the gradient $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{\theta})$. For the second, we see that

$$\begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix} = \boldsymbol{F}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) + \boldsymbol{K}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}).$$

The right-hand side here is just the expression that appears in the fundamental result (14.20). Hence it is clear that the DLR must satisfy (14.26). All this discussion assumes, of course, that the matrices $\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta})$ and $\boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta})$ satisfy appropriate regularity conditions, which may not always be easy to verify in practice; see Davidson and MacKinnon (1984a).

The DLR can be used in all the same ways that the GNR and the OPG regression can be used. In particular, it can be used

(i) to verify that the first-order conditions for a maximum of the log-likelihood function are satisfied sufficiently accurately,

(ii) to calculate estimated covariance matrices,

(iii) to calculate test statistics,

(iv) to calculate one-step efficient estimates, and

(v) as a key part of procedures for finding ML estimates.

Hence the DLR for the simple Box-Cox model, (14.04) with $\tau(y_t, \lambda)$ given by the Box-Cox transformation, is

$$
\begin{bmatrix} \frac{1}{\sigma} u_t(y_t, \boldsymbol{\beta}, \lambda) \\ 1 \end{bmatrix}
\tag{14.33}
$$

$$
= \begin{bmatrix} \frac{1}{\sigma} \boldsymbol{X}_t(\boldsymbol{\beta}) & \dfrac{-(\lambda y_t^\lambda \log y_t - y_t^\lambda + 1)}{\sigma \lambda^2} & \dfrac{u_t(y_t, \boldsymbol{\beta}, \lambda)}{\sigma^2} \\ \boldsymbol{0} & \log y_t & -\dfrac{1}{\sigma} \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ a \\ s \end{bmatrix} + \text{residuals},
$$

where $\boldsymbol{b}$ is a k–vector of coefficients corresponding to $\boldsymbol{\beta}$, a and s are scalar coefficients corresponding to λ and σ, and

$$
u_t(y_t, \boldsymbol{\beta}, \lambda) \equiv B(y_t, \lambda) - x_t(\boldsymbol{\beta}).
$$

If the DLR (14.33) is evaluated at unrestricted ML estimates $\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma})$, all the estimated coefficients will be zero. Since the first-order conditions for σ imply that

$$
\hat{\sigma} = \left(\frac{1}{n} \sum_{t=1}^n \hat{u}_t^2 \right)^{1/2},
$$

the total sum of squares from the artificial regression will be $2n$. Thus the OLS covariance matrix estimate will simply be $\big(2n/(2n - k - 2)\big)(\hat{\boldsymbol{R}}^\top \hat{\boldsymbol{R}})^{-1}$, where $\hat{\boldsymbol{R}}$ denotes the matrix of regressors that appears in (14.33), evaluated at the ML estimates. By the fundamental result (14.20), this OLS covariance matrix provides a valid estimate of the asymptotic covariance matrix of the ML estimator $\hat{\boldsymbol{\theta}}$.

It is clear from (14.33) that this asymptotic covariance matrix is not block-diagonal between $\boldsymbol{\beta}$ and the other parameters. Forming the matrix $\boldsymbol{R}^\top \boldsymbol{R}$, dividing by n, and taking probability limits, we see that the $(\boldsymbol{\beta}, \boldsymbol{\beta})$ block of the information matrix $\mathfrak{I}(\boldsymbol{\theta})$ is simply

$$
\sigma^{-2} \operatorname*{plim}_{n \to \infty} \Big(\frac{1}{n} \boldsymbol{X}^\top(\boldsymbol{\beta}) \boldsymbol{X}(\boldsymbol{\beta}) \Big),
\tag{14.34}
$$

as it would be if this were a nonlinear regression model. The (σ, σ) element is simply $2/\sigma^2$, which again is what it would be if this were a nonlinear regression model. But $\mathfrak{I}(\boldsymbol{\theta})$ also contains a (λ, λ) element, a (λ, σ) element, and a $(\boldsymbol{\beta}, \lambda)$ row and column, all of which are clearly nonzero. For example, the element corresponding to β_i and λ is

$$
- \operatorname*{plim}_{n \to \infty} \Big(\frac{1}{n\sigma^2 \lambda^2} \sum_{t=1}^n X_{ti}(\boldsymbol{\beta}) \big(\lambda y_t^\lambda \log y_t - y_t^\lambda + 1 \big) \Big).
$$

The (λ, λ) and (λ, σ) elements can also be obtained in a straightforward fashion and are easily seen to be nonzero.

unidentified. However, following the procedure used to obtain the J and P tests, we can replace the parameters of the model that is *not* being tested by estimates. Thus, if we wish to test H_1, we can replace $\boldsymbol{\gamma}$ and σ_2 by ML estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}_2$ so that H_C becomes

$$H_C': \quad (1-\alpha)\left(\frac{y_t - x_t(\boldsymbol{\beta})}{\sigma_1}\right) + \alpha\left(\frac{\log y_t - z_t(\hat{\boldsymbol{\gamma}})}{\hat{\sigma}_2}\right) = \varepsilon_t.$$

It is straightforward to test H_1 against H_C' by means of the DLR:

$$\begin{bmatrix} \dfrac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} \\[2mm] 1 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{X}}_t & \dfrac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} & \hat{z}_t - \log y_t \\[2mm] \boldsymbol{0} & -1 & \hat{\sigma}_1/y_t \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ s \\ a \end{bmatrix} + \text{residuals}, \quad (14.45)$$

where $\hat{x}_t \equiv x_t(\hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{X}}_t \equiv \boldsymbol{X}_t(\hat{\boldsymbol{\beta}})$, and $\hat{z}_t \equiv z_t(\hat{\boldsymbol{\gamma}})$. The DLR (14.45) is actually a simplified version of the DLR that one obtains initially. First, $\hat{\sigma}_1$ times the original regressor for σ_1 has been subtracted from the original regressor for α. Then the regressors corresponding to $\boldsymbol{\beta}$ and σ_1 have been multiplied by $\hat{\sigma}_1$, and the regressor corresponding to α has been multiplied by $\hat{\sigma}_2$. None of these modifications affects the subspace spanned by the columns of the regressor, and hence none of them affects the test statistic(s) one obtains. The last column of the regressor matrix in (14.45) is the one that corresponds to α. The other columns should be orthogonal to the regressand by construction.

Similarly, if we wish to test H_2, we can replace $\boldsymbol{\beta}$ and σ_1 by ML estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_1$ so that H_C becomes

$$H_C'': \quad (1-\alpha)\left(\frac{y_t - x_t(\hat{\boldsymbol{\beta}})}{\hat{\sigma}_1}\right) + \alpha\left(\frac{\log y_t - z_t(\boldsymbol{\gamma})}{\sigma_2}\right) = \varepsilon_t.$$

It is then straightforward to test H_2 against H_C'' by means of the DLR

$$\begin{bmatrix} \dfrac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} \\[2mm] 1 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{Z}}_t & \dfrac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} & \hat{x}_t - y_t \\[2mm] \boldsymbol{0} & -1 & \hat{\sigma}_2 y_t \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ s \\ a \end{bmatrix} + \text{residuals}. \quad (14.46)$$

Once again, this is a simplified version of the DLR that one obtains initially, and the last column of the regressor matrix is the one that corresponds to α.

The tests we have just discussed evidently generalize very easily to models involving any sort of transformation of the dependent variable, including Box-Cox models and other models in which the transformation depends on one or more unknown parameters. For more details, see Davidson and MacKinnon (1984a). It should be stressed that the artificial compound model (14.44) is quite arbitrary. Unlike the similar-looking model for regression models that was employed in Section 11.3, it does not yield tests asymptotically equivalent

the regularity conditions needed for the ML estimates $\hat{\boldsymbol{\beta}}$ to be consistent and asymptotically normal, with asymptotic covariance matrix given by the inverse of the information matrix in the usual way. See, for example, Gouriéroux and Monfort (1981). In the case of the logit model, the first-order conditions (15.10) simplify to

$$\sum_{t=1}^{n} \big(y_t - \Lambda(\boldsymbol{X}_t\hat{\boldsymbol{\beta}})\big)X_{ti} = 0, \;\; i = 1, \ldots, k,$$

because $\lambda(x) = \Lambda(x)\big(1 - \Lambda(x)\big)$. Notice that conditions (15.10) look just like the first-order conditions for weighted least squares estimation of the nonlinear regression model

$$y_t = F(\boldsymbol{X}_t\boldsymbol{\beta}) + e_t, \tag{15.11}$$

with weights given by

$$\Big(F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)\Big)^{-1/2}.$$

This makes sense, since the variance of the error term in (15.11) is

$$\begin{aligned}
E(e_t^2) &= E\big(y_t - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)^2 \\
&= F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)^2 + \big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)\big(F(\boldsymbol{X}_t\boldsymbol{\beta})\big)^2 \\
&= F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big).
\end{aligned}$$

Thus one way to obtain ML estimates of any binary response model is to apply iteratively reweighted nonlinear least squares to (15.11) or to whatever nonlinear regression model is appropriate if the index function is not $\boldsymbol{X}_t\boldsymbol{\beta}$. For most models, however, this is generally not the best approach, and a better one is discussed in the next section.

Using the fact that ML is equivalent to a form of weighted NLS for binary response models, it is obvious that the asymptotic covariance matrix for $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ must be

$$\Big(\tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Psi}(\boldsymbol{\beta}_0)\boldsymbol{X}\Big)^{-1},$$

where $\boldsymbol{X}$ is an $n \times k$ matrix with typical row $\boldsymbol{X}_t$ and typical element X_{ti}, and $\boldsymbol{\Psi}(\boldsymbol{\beta})$ is a diagonal matrix with typical diagonal element

$$\Psi(\boldsymbol{X}_t\boldsymbol{\beta}) = \frac{f^2(\boldsymbol{X}_t\boldsymbol{\beta})}{F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)}. \tag{15.12}$$

The numerator reflects the fact that the derivative of $F(\boldsymbol{X}_t\boldsymbol{\beta})$ with respect to β_i is $f(\boldsymbol{X}_t\boldsymbol{\beta})X_{ti}$, and the denominator is simply the variance of e_t in (15.11). In the logit case, $\Psi(\boldsymbol{X}_t\boldsymbol{\beta})$ simplifies to $\lambda(\boldsymbol{X}_t\boldsymbol{\beta})$.

15.4 An Artificial Regression

There exists a very simple and very useful artificial regression for binary response models. Like other artificial regressions, it can be used for a variety of purposes, including parameter estimation, covariance matrix estimation, and hypothesis testing. This artificial regression was suggested by Engle (1984) and Davidson and MacKinnon (1984b). It can be derived in several ways, of which the easiest is to treat it as a modified version of the Gauss-Newton regression.

As we have seen, the binary response model (15.03) can be written in the form of the nonlinear regression model (15.11), that is, as $y_t = F(\boldsymbol{X}_t\boldsymbol{\beta}) + e_t$. We have also seen that the error term e_t has variance

$$V(\boldsymbol{X}_t\boldsymbol{\beta}) \equiv F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big), \tag{15.19}$$

which implies that (15.11) must be estimated by GNLS. The ordinary GNR corresponding to (15.11) would be

$$y_t - F(\boldsymbol{X}_t\boldsymbol{\beta}) = f(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t\boldsymbol{b} + \text{residual}, \tag{15.20}$$

but this is clearly inappropriate because of the heteroskedasticity of the e_t's. Instead, we must multiply both sides of (15.20) by the square root of the inverse of (15.19). This yields the artificial regression

$$\big(V(\boldsymbol{X}_t\boldsymbol{\beta})\big)^{-1/2}\big(y_t - F(\boldsymbol{X}_t\boldsymbol{\beta})\big) = \big(V(\boldsymbol{X}_t\boldsymbol{\beta})\big)^{-1/2}f(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t\boldsymbol{b} + \text{residual}, \tag{15.21}$$

which looks like the GNR for a nonlinear regression model estimated by weighted least squares (see Section 9.4). Regression (15.21) is a special case of what we will call the **binary response model regression**, or **BRMR**. This form of the BRMR is valid for any binary response model of the form (15.03).[4] In the case of the logit model, it simplifies to

$$\big(\lambda(\boldsymbol{X}_t\boldsymbol{\beta})\big)^{-1/2}\big(y_t - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})\big) = \big(\lambda(\boldsymbol{X}_t\boldsymbol{\beta})\big)^{1/2}\boldsymbol{X}_t\boldsymbol{b} + \text{residual}.$$

The BRMR satisfies the general properties of artificial regressions that we discussed in Section 14.4. In particular, it is closely related both to the gradient of the loglikelihood function (15.09) and to the information matrix. The

---

[4] Some authors write the BRMR in other ways. For example, in Davidson and MacKinnon (1984b), the regressand was defined as

$$y_t\left(\frac{1 - F(\boldsymbol{X}_t\boldsymbol{\beta})}{F(\boldsymbol{X}_t\boldsymbol{\beta})}\right)^{1/2} + (y_t - 1)\left(\frac{F(\boldsymbol{X}_t\boldsymbol{\beta})}{1 - F(\boldsymbol{X}_t\boldsymbol{\beta})}\right)^{1/2}.$$

It is a good exercise to verify that this is just another way of writing the regressand of (15.21).

so and is probably to be preferred, since the factor of s^2 in (15.23) simply introduces additional randomness into the estimate of the covariance matrix.

As usual, the covariance matrix of $\hat{\boldsymbol{\beta}}$ can also be estimated as minus the inverse of the numerical Hessian or as the inverse of the outer product of the CG matrix, $\hat{\boldsymbol{G}}^\top \hat{\boldsymbol{G}}$. In the case of the logit model, minus the numerical Hessian is actually equal to the estimated information matrix $\boldsymbol{X}^\top \hat{\boldsymbol{\Psi}} \boldsymbol{X}$, because

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} = \frac{\partial}{\partial \beta_j} \left(\sum_{t=1}^n \big(y_t - \Lambda(\boldsymbol{X}_t \boldsymbol{\beta}) \big) X_{ti} \right) = - \sum_{t=1}^n \lambda(\boldsymbol{X}_t \boldsymbol{\beta}) X_{ti} X_{tj}.$$

However, in the case of most other binary response models, including the probit model, minus the Hessian will differ from, and generally be more complicated than, the information matrix.

Like all artificial regressions, the BRMR is particularly useful for hypothesis testing. Suppose that $\boldsymbol{\beta}$ is partitioned as $[\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2]$, where $\boldsymbol{\beta}_1$ is a $(k-r)$–vector and $\boldsymbol{\beta}_2$ is an r–vector. If $\tilde{\boldsymbol{\beta}}$ denotes the vector of ML estimates subject to the restriction that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, we can test that restriction by running the BRMR

$$\tilde{V}_t^{-1/2}(y_t - \tilde{F}_t) = \tilde{V}_t^{-1/2} \tilde{f}_t \boldsymbol{X}_{t1} \boldsymbol{b}_1 + \tilde{V}_t^{-1/2} \tilde{f}_t \boldsymbol{X}_{t2} \boldsymbol{b}_2 + \text{residual}, \qquad (15.24)$$

where $\tilde{F}_t \equiv F(\boldsymbol{X}_t \tilde{\boldsymbol{\beta}})$, $\tilde{f}_t \equiv f(\boldsymbol{X}_t \tilde{\boldsymbol{\beta}})$, and $\tilde{V}_t \equiv V(\boldsymbol{X}_t \tilde{\boldsymbol{\beta}})$. Here $\boldsymbol{X}_t$ has been partitioned into two vectors, $\boldsymbol{X}_{t1}$ and $\boldsymbol{X}_{t2}$, corresponding to the partitioning of $\boldsymbol{\beta}$. The regressors that correspond to $\boldsymbol{\beta}_1$ are orthogonal to the regressand, while those that correspond to $\boldsymbol{\beta}_2$ are not. All the usual test statistics for $\boldsymbol{b}_2 = \boldsymbol{0}$ are valid. However, in contrast to the case of the Gauss-Newton regression, there is no particular reason to use an F test, because there is no variance parameter to estimate. The best test statistic to use in finite samples, according to Monte Carlo results obtained by Davidson and MacKinnon (1984b), is probably the explained sum of squares from regression (15.24). It will be asymptotically distributed as $\chi^2(r)$ under the null hypothesis. Note that nR^2 will not be equal to the explained sum of squares in this case, because the total sum of squares will not be equal to n.

In one very special case, the BRMR (15.24) becomes extremely simple. Suppose the null hypothesis is that all the slope coefficients are zero. In this case, $\boldsymbol{X}_{t1}$ is just unity, $\boldsymbol{X}_t \tilde{\boldsymbol{\beta}} = \tilde{\beta}_1 = F^{-1}(\bar{y})$, and, in obvious notation, regression (15.24) becomes

$$\bar{V}^{-1/2}(y_t - \bar{F}) = \bar{V}^{-1/2} \bar{f} b_1 + \bar{V}^{-1/2} \bar{f} \boldsymbol{X}_{t2} \boldsymbol{b}_2 + \text{residual}.$$

Neither subtracting a constant from the regressand nor multiplying the regressand and regressors by a constant has any effect on the F statistic for $\boldsymbol{b}_2 = \boldsymbol{0}$. Thus it is clear that we can test the all-slopes-zero hypothesis simply by calculating an F statistic for $\boldsymbol{c}_2 = \boldsymbol{0}$ in the linear regression

$$\boldsymbol{y} = c_1 + \boldsymbol{X}_2 \boldsymbol{c}_2 + \text{residuals}.$$

can be written as

$$\ell(\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^J) = \sum_{j=1}^{J}\sum_{y_t=j} \boldsymbol{X}_t\boldsymbol{\beta}^j - \sum_{t=1}^{n}\log\left(1+\sum_{j=1}^{J}\exp(\boldsymbol{X}_t\boldsymbol{\beta}^j)\right).$$

This function is a sum of contributions from each observation. Each contribution has two terms: The first is $\boldsymbol{X}_t\boldsymbol{\beta}^j$, where the index j is that for which $y_t = j$ (or zero if $j = 0$), and the second is minus the logarithm of the denominator that appears in (15.35) and (15.36).

One important property of the multinomial logit model is that

$$\frac{\Pr(y_t = l)}{\Pr(y_t = j)} = \frac{\exp(\boldsymbol{X}_t\boldsymbol{\beta}^l)}{\exp(\boldsymbol{X}_t\boldsymbol{\beta}^j)} = \exp\big(\boldsymbol{X}_t(\boldsymbol{\beta}^l - \boldsymbol{\beta}^j)\big) \qquad (15.38)$$

for any two responses l and j (including response zero if we interpret $\boldsymbol{\beta}^0$ as a vector of zeros). Thus the odds between any two responses depend solely on $\boldsymbol{X}_t$ and on the parameter vectors associated with those two responses. They do not depend on the parameter vectors associated with any of the other responses. In fact, we see from (15.38) that the log of the odds between responses l and j is simply $\boldsymbol{X}_t\boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^* \equiv (\boldsymbol{\beta}^l - \boldsymbol{\beta}^j)$. Thus, conditional on either j or l being chosen, the choice between them is determined by an ordinary logit model with parameter vector $\boldsymbol{\beta}^*$.

Closely related to the multinomial logit model is the **conditional logit model** pioneered by McFadden (1974a, 1974b). See Domencich and McFadden (1975), McFadden (1984), and Greene (1990a, Chapter 20) for detailed treatments. The conditional logit model is designed to handle consumer choice among J (*not* $J + 1$) discrete alternatives, where one and only one of the alternatives can be chosen. Suppose that when the ith consumer chooses alternative j, he or she obtains utility

$$U_{ij} = \boldsymbol{W}_{ij}\boldsymbol{\beta} + \varepsilon_{ij},$$

where $\boldsymbol{W}_{ij}$ is a row vector of characteristics of alternative j as they apply to consumer i. Let y_i denote the choice made by the ith consumer. Presumably $y_i = l$ if U_{il} is at least as great as U_{ij} for all $j \neq l$. Then if the disturbances ε_{ij} for $j = 1,\ldots, J$ are independent and identically distributed according to the Weibull distribution, it can be shown that

$$\Pr(y_i = l) = \frac{\exp(\boldsymbol{W}_{il}\boldsymbol{\beta})}{\sum_{j=1}^{J}\exp(\boldsymbol{W}_{ij}\boldsymbol{\beta})}. \qquad (15.39)$$

This closely resembles (15.37), and it is easy to see that the probabilities must add to unity.

There are two key differences between the multinomial logit and conditional logit models. In the former, there is a single vector of independent variables for each observation, and there are J different vectors of parameters.

In the latter, the values of the independent variables vary across alternatives, but there is just a single parameter vector $\boldsymbol{\beta}$. The multinomial logit model is a straightforward generalization of the logit model that can be used to deal with any situation involving three or more unordered qualitative responses. In contrast, the conditional logit model is specifically designed to handle consumer choices among discrete alternatives based on the characteristics of those alternatives.

Depending on the nature of the explanatory variables, there can be a number of subtleties associated with the specification and interpretation of conditional logit models. There is not enough space in this book to treat these adequately, and so readers who intend to estimate such models are urged to consult the references mentioned above. One important property of conditional logit models is the analog of (15.38):

$$\frac{\Pr(y_i = l)}{\Pr(y_i = j)} = \frac{\exp(\boldsymbol{W}_{il}\boldsymbol{\beta})}{\exp(\boldsymbol{W}_{ij}\boldsymbol{\beta})}. \tag{15.40}$$

This property is called the **independence of irrelevant alternatives**, or **IIA**, property. It implies that adding another alternative to the model, or changing the characteristics of another alternative that is already included, will not change the odds between alternatives l and j.

The IIA property can be extremely implausible in certain circumstances. Suppose that there are initially two alternatives for traveling between two cities: flying Monopoly Airways and driving. Suppose further that half of all travelers fly and the other half drive. Then Upstart Airways enters the market and creates a third alternative. If Upstart offers a service identical to that of Monopoly, it must gain the same market share. Thus, according to the IIA property, one third of the travelers must take each of the airlines and one third must drive. So the automobile has lost just as much market share from the entry of Upstart Airways as Monopoly Airways has! This seems very implausible.[6] As a result, a number of papers have been devoted to the problem of testing the independence of irrelevant alternatives property and finding tractable models that do not embody it. See, in particular, Hausman and Wise (1978), Manski and McFadden (1981), Hausman and McFadden (1984), and McFadden (1987).

This concludes our discussion of qualitative response models. More detailed treatments may be found in surveys by Maddala (1983), McFadden (1984), Amemiya (1981; 1985, Chapter 9), and Greene (1990a, Chapter 20), among others. In the next three sections, we turn to the subject of limited dependent variables.

---

[6] One might object that a price war between Monopoly and Upstart would convince some drivers to fly instead. So it would. But if the two airlines offered lower prices, that would change one or more elements of the $\boldsymbol{W}_{ij}$'s associated with them. The above analysis assumes that all the $\boldsymbol{W}_{ij}$'s remain unchanged.

$h \equiv 1/\sigma$, and the loglikelihood function can be shown to be globally concave in the latter parametrization. This implies that it must have a unique maximum no matter how it is parametrized. The $(k+1) \times (k+1)$ covariance matrix of the ML estimates may as usual be estimated in several ways. Unfortunately, as with the truncated regression model discussed in the previous section, the only artificial regression that is presently known to be applicable to this model is the OPG regression.

There is an interesting relationship among the tobit, truncated regression, and probit models. Suppose, for simplicity, that $x_t(\boldsymbol{\beta}) = \boldsymbol{X}_t\boldsymbol{\beta}$. Then the tobit loglikelihood function can be rewritten as

$$\sum_{y_t>0} \log\left(\frac{1}{\sigma}\phi\left(\frac{1}{\sigma}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})\right)\right) + \sum_{y_t=0} \log\left(\Phi\left(-\frac{1}{\sigma}\boldsymbol{X}_t\boldsymbol{\beta}\right)\right). \qquad (15.49)$$

Now let us both add and subtract the term $\sum_{y_t>0} \log\left(\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)\right)$ in (15.49), which then becomes

$$\begin{aligned} &\sum_{y_t>0} \log\left(\frac{1}{\sigma}\phi\left(\frac{1}{\sigma}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})\right)\right) - \sum_{y_t>0} \log\left(\Phi\left(\frac{1}{\sigma}\boldsymbol{X}_t\boldsymbol{\beta}\right)\right) \\ &+ \sum_{y_t=0} \log\left(\Phi\left(-\frac{1}{\sigma}\boldsymbol{X}_t\boldsymbol{\beta}\right)\right) + \sum_{y_t>0} \log\left(\Phi\left(\frac{1}{\sigma}\boldsymbol{X}_t\boldsymbol{\beta}\right)\right). \end{aligned} \qquad (15.50)$$

The first line here is the loglikelihood function for a truncated regression model; it is just (15.43) with $y^l = 0$ and $x_t(\boldsymbol{\beta}) = \boldsymbol{X}_t\boldsymbol{\beta}$ and with the set of observations to which the summations apply adjusted appropriately. The second line is the loglikelihood function for a probit model with index function $\boldsymbol{X}_t\boldsymbol{\beta}/\sigma$. Of course, if all we had was the second line here, we could not identify $\boldsymbol{\beta}$ and σ separately, but since we also have the first line, that is not a problem.

Expression (15.50) makes it clear that the tobit model is like a truncated regression model combined with a probit model, with the coefficient vectors in the latter two models restricted to be proportional to each other. Cragg (1971) argued that this restriction may sometimes be unreasonable and proposed several more general models as plausible alternatives to the tobit model. It may sometimes be desirable to test the tobit model against one or more of these more general models; see Lin and Schmidt (1984) and Greene (1990a, Chapter 21).

As we mentioned earlier, it is easy to modify the tobit model to handle different types of censoring. For example, one possibility is a model with **double censoring**. Suppose that

$$y_t^* = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$

$$y_t = y_t^* \text{ if } y_t^l \leq y_t^* \leq y_t^u; \ y_t = y_t^l \text{ if } y_t^* < y_t^l; \ y_t = y_t^u \text{ if } y_t^* > y_t^u.$$

they are related to y_t^* and z_t^* as follows:

$$y_t = y_t^* \text{ if } z_t^* > 0; \quad y_t = 0 \text{ otherwise;}$$

$$z_t = 1 \text{ if } z_t^* > 0; \quad z_t = 0 \text{ otherwise.}$$

There are two types of observations: ones for which both y_t and z_t are observed to be zero and ones for which $z_t = 1$ and y_t is equal to y_t^*. The loglikelihood function for this model is thus

$$\sum_{z_t=0} \log\big(\Pr(z_t = 0)\big) + \sum_{z_t=1} \log\big(\Pr(z_t = 1) f(y_t^* \mid z_t = 1)\big), \qquad (15.54)$$

where $f(y_t^* \mid z_t = 1)$ denotes the density of y_t^* conditional on $z_t = 1$. The first term of (15.54) is the summation over all observations for which $z_t = 0$ of the logarithms of the probability that $z_t = 0$. It is exactly the same as the corresponding term in a probit model for z_t by itself. The second term is the summation over all observations for which $z_t = 1$ of the probability that $z_t = 1$ times the density of y_t conditional on $z_t = 1$. Using the fact that we can factor a joint density any way we like, this second term can also be written as

$$\sum_{z_t=1} \log\big(\Pr(z_t = 1 \mid y_t^*) f(y_t^*)\big),$$

where $f(y_t^*)$ is the unconditional density of y_t^*, which is just a normal density with conditional mean $\boldsymbol{X}_t\boldsymbol{\beta}$ and variance σ^2.

The only difficulty in writing out the loglikelihood function (15.54) explicitly is to calculate $\Pr(z_t = 1 \mid y_t^*)$. Since u_t and v_t are bivariate normal, we can write

$$z_t^* = \boldsymbol{W}_t\boldsymbol{\gamma} + \rho\Big(\frac{1}{\sigma}\big(y_t^* - \boldsymbol{X}_t\boldsymbol{\beta}\big)\Big) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}\big(0, (1 - \rho^2)\big).$$

It follows that

$$\Pr(z_t = 1 \mid y_t^*) = \Phi\left(\frac{\boldsymbol{W}_t\boldsymbol{\gamma} + \rho\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)}{(1 - \rho^2)^{1/2}}\right),$$

since $y_t = y_t^*$ when $z_t = 1$. Thus the loglikelihood function (15.54) becomes

$$\sum_{z_t=0} \log\big(\Phi(-\boldsymbol{W}_t\boldsymbol{\gamma})\big) + \sum_{z_t=1} \log\Big(\frac{1}{\sigma}\phi\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)\Big)$$

$$+ \sum_{z_t=1} \log\left(\Phi\left(\frac{\boldsymbol{W}_t\boldsymbol{\gamma} + \rho\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)}{(1 - \rho^2)^{1/2}}\right)\right). \qquad (15.55)$$

The first term looks like the corresponding term for a probit model. The

second term looks like the loglikelihood function for a linear regression model with normal errors. The third term is one that we have not seen before.

Maximum likelihood estimates can be obtained in the usual way by maximizing (15.55). However, this maximization is relatively burdensome, and so instead of ML estimation a computationally simpler technique proposed by Heckman (1976) is often used. **Heckman's two-step method** is based on the fact that the first equation of (15.53) can be rewritten as

$$y_t^* = \boldsymbol{X}_t\boldsymbol{\beta} + \rho\sigma v_t + e_t. \tag{15.56}$$

The idea is to replace y_t^* by y_t and v_t by its mean conditional on $z_t = 1$ and on the realized value of $\boldsymbol{W}_t\boldsymbol{\gamma}$. As can be seen from (15.42), this conditional mean is $\phi(\boldsymbol{W}_t\boldsymbol{\gamma})/\Phi(\boldsymbol{W}_t\boldsymbol{\gamma})$, a quantity that is sometimes referred to as the **inverse Mills ratio**. Hence regression (15.56) becomes

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \rho\sigma\,\frac{\phi(\boldsymbol{W}_t\boldsymbol{\gamma})}{\Phi(\boldsymbol{W}_t\boldsymbol{\gamma})} + \text{residual}. \tag{15.57}$$

It is now easy to see how Heckman's two-step method works. In the first step, an ordinary probit model is used to obtain consistent estimates $\hat{\boldsymbol{\gamma}}$ of the parameters of the selection equation. In the second step, the **selectivity regressor** $\phi(\boldsymbol{W}_t\boldsymbol{\gamma})/\Phi(\boldsymbol{W}_t\boldsymbol{\gamma})$ is evaluated at $\hat{\boldsymbol{\gamma}}$, and regression (15.57) is estimated by OLS for the observations with $z_t = 1$ only. This regression provides a test for sample selectivity as well as an estimation technique. The coefficient on the selectivity regressor is $\rho\sigma$. Since $\sigma \neq 0$, the ordinary t statistic for this coefficient to be zero can be used to test the hypothesis that $\rho = 0$; it will be asymptotically distributed as $N(0,1)$ under the null hypothesis. Thus, if this coefficient is not significantly different from zero, the investigator may reasonably decide that selectivity is not a problem for this data set and proceed to use least squares as usual.

Even when the hypothesis that $\rho = 0$ cannot be accepted, OLS estimation of regression (15.57) yields consistent estimates of $\boldsymbol{\beta}$. However, the OLS covariance matrix is valid only when $\rho = 0$. In this respect, the situation is very similar to the one encountered at the end of the previous section, when we were testing for possible simultaneity bias in models with truncated or censored dependent variables. There are actually two problems. First of all, the residuals in (15.57) will be heteroskedastic, since a typical residual is equal to

$$u_t - \rho\sigma\,\frac{\phi(\boldsymbol{W}_t\boldsymbol{\gamma})}{\Phi(\boldsymbol{W}_t\boldsymbol{\gamma})}.$$

Secondly, the selectivity regressor is being treated like any other regressor, when it is in fact part of the error term. One could solve the first problem by using a heteroskedasticity-consistent covariance matrix estimator (see Chapter 16), but that would not solve the second one. It is possible to obtain a

valid covariance matrix estimate to go along with the two-step estimates of $\boldsymbol{\beta}$ from (15.57). However, the calculation is cumbersome, and the estimated covariance matrix is not always positive definite. See Greene (1981b) and Lee (1982) for more details.

It should be stressed that the consistency of this two-step estimator, like that of the ML estimator, depends critically on the assumption of normality. This can be seen from the specification of the selectivity regressor as the inverse Mills ratio $\phi(\boldsymbol{W}_t\boldsymbol{\gamma})/\Phi(\boldsymbol{W}_t\boldsymbol{\gamma})$. When the elements of $\boldsymbol{W}_t$ are the same as, or a subset of, the elements of $\boldsymbol{X}_t$, as is often the case in practice, it is only the nonlinearity of $\phi(\boldsymbol{W}_t\boldsymbol{\gamma})/\Phi(\boldsymbol{W}_t\boldsymbol{\gamma})$ as a function of $\boldsymbol{W}_t\boldsymbol{\gamma}$ that makes the parameters of the second-step regression identifiable. The exact form of the nonlinear relationship depends critically on the normality assumption. Pagan and Vella (1989), Smith (1989), and Peters and Smith (1991) discuss various ways to test this crucial assumption. Many of the tests suggested by these authors are applications of the OPG regression.

Although the two-step method for dealing with sample selectivity is widely used, our recommendation would be to use regression (15.57) only as a procedure for testing the null hypothesis that selectivity bias is not present. When that hypothesis is rejected, ML estimation based on (15.55) should probably be used in preference to the two-step method, unless it is computationally prohibitive.

15.9 CONCLUSION

Our treatment of binary response models in Sections 15.2 to 15.4 was reasonably detailed, but the discussions of more general qualitative response models and limited dependent variable models were necessarily quite superficial. Anyone who intends to do empirical work that employs this type of model will wish to consult some of the more detailed surveys referred to above. All of the methods that we have discussed for handling limited dependent variables rely heavily on the assumptions of normality and homoskedasticity. These assumptions should always be tested. A number of methods for doing so have been proposed; see, among others, Bera, Jarque, and Lee (1984), Lee and Maddala (1985), Blundell (1987), Chesher and Irish (1987), Pagan and Vella (1989), Smith (1989), and Peters and Smith (1991).

16.3 COVARIANCE MATRIX ESTIMATION

At first glance, the generalized OLS covariance matrix estimator and its NLS analog (16.08) do not seem to be very useful. To compute them we need to know $\boldsymbol{\Omega}$, but if we knew $\boldsymbol{\Omega}$, we could use GLS or GNLS and obtain more efficient estimates. This was the conventional wisdom among econometricians until a decade ago. But an extremely influential paper by White (1980) showed that it is in fact possible to obtain an estimator of the covariance matrix of least squares estimates that is asymptotically valid when there is heteroskedasticity of unknown form.[2] Such an estimator is a called a **heteroskedasticity-consistent covariance matrix estimator**, or **HCCME**.

The key to obtaining an HCCME is to recognize that we do *not* have to estimate $\boldsymbol{\Omega}$ consistently. That would indeed be an impossible task, since $\boldsymbol{\Omega}$ has n diagonal elements to estimate. The asymptotic covariance matrix of a vector of NLS estimates, under heteroskedasticity, is given by expression (16.08), which can be rewritten as

$$\plim_{n\to\infty}\left(\tfrac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0\right)^{-1}\plim_{n\to\infty}\left(\tfrac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{\Omega}\boldsymbol{X}_0\right)\plim_{n\to\infty}\left(\tfrac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0\right)^{-1}. \qquad (16.09)$$

The first and third factors here are identical, and we can easily estimate them in the usual way. A consistent estimator is

$$\tfrac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}},$$

where $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}})$. The only tricky thing, then, is to estimate the second factor. White showed that this second factor can be estimated consistently by

$$\tfrac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{X}}, \qquad (16.10)$$

where $\hat{\boldsymbol{\Omega}}$ may be any of several different *inconsistent* estimators of $\boldsymbol{\Omega}$. The simplest version of $\hat{\boldsymbol{\Omega}}$, and the one that White proposed in the context of linear regression models, has t^{th} diagonal element equal to $\hat{u}_t^2$, the tth squared least squares residual.

Unlike $\boldsymbol{\Omega}$, the middle factor of (16.09) has only $\tfrac{1}{2}(k^2 + k)$ distinct elements, whatever the sample size. That is why it is possible to estimate it consistently. A typical element of this matrix is

$$\plim_{n\to\infty}\left(\tfrac{1}{n}\sum_{t=1}^{n}\omega_t^2 X_{ti}X_{tj}\right), \qquad (16.11)$$

---

[2] Precursors of White's paper in the statistics literature include Eicker (1963, 1967) and Hinkley (1977), as well as some of the early papers on bootstrapping (see Chapter 21).

where $X_{ti} \equiv X_{ti}(\boldsymbol{\beta}_0)$. On the other hand, a typical element of (16.10) is

$$\frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2 \hat{X}_{ti}\hat{X}_{tj}. \tag{16.12}$$

Because $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$, $\hat{u}_t$ is consistent for u_t, $\hat{u}_t^2$ is consistent for u_t^2, and $\hat{X}_{ti}$ is consistent for X_{ti}. Thus expression (16.12) is asymptotically equal to

$$\frac{1}{n}\sum_{t=1}^{n} u_t^2 X_{ti}X_{tj}. \tag{16.13}$$

Under our assumptions, we can apply a law of large numbers to (16.13); see White (1980, 1984) and Nicholls and Pagan (1983) for some technical details. It follows immediately that (16.13), and so also (16.12), tends in probability to (16.11). Consequently, the matrix

$$\bigl(n^{-1}\hat{\boldsymbol{X}}^{\top}\hat{\boldsymbol{X}}\bigr)^{-1}\bigl(n^{-1}\hat{\boldsymbol{X}}^{\top}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{X}}\bigr)\bigl(n^{-1}\hat{\boldsymbol{X}}^{\top}\hat{\boldsymbol{X}}\bigr)^{-1} \tag{16.14}$$

consistently estimates (16.09). Of course, in practice one ignores the factors of n^{-1} and uses the matrix

$$\bigl(\hat{\boldsymbol{X}}^{\top}\hat{\boldsymbol{X}}\bigr)^{-1}\hat{\boldsymbol{X}}^{\top}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{X}}\bigl(\hat{\boldsymbol{X}}^{\top}\hat{\boldsymbol{X}}\bigr)^{-1} \tag{16.15}$$

to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$.

Asymptotically valid inferences about $\boldsymbol{\beta}$ may be based on the HCCME (16.15) in the usual way. However, one must be cautious when n is not large. There is a good deal of evidence that this HCCME is somewhat unreliable in finite samples. After all, the fact that (16.14) estimates (16.09) consistently does not imply that the former always estimates the latter very well in finite samples.

It is possible to modify the HCCME (16.15) so that it has better finite-sample properties. The major problem is that the squared least squares residuals $\hat{u}_t^2$ are not unbiased estimates of the squared error terms u_t^2. The easiest way to improve the HCCME is simply to multiply (16.15) by $n/(n-k)$. This is analogous to dividing the sum of squared residuals by $n-k$ rather than n to obtain the OLS variance estimator s^2. A second, and better, approach is to define the tth diagonal element of $\hat{\boldsymbol{\Omega}}$ as $\hat{u}_t^2/(1-\hat{h}_t)$, where $\hat{h}_t \equiv \hat{\boldsymbol{X}}_t(\hat{\boldsymbol{X}}^{\top}\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}_t^{\top}$ is the tth diagonal element of the "hat" matrix $\hat{\boldsymbol{P}}_X$ that projects orthogonally onto the space spanned by the columns of $\hat{\boldsymbol{X}}$. Recall from Section 3.2 that, in the OLS case with constant variance σ^2, the expectation of $\hat{u}_t^2$ is $\sigma^2(1-h_t)$. Thus, in the linear case, dividing $\hat{u}_t^2$ by $1-h_t$ would yield an unbiased estimate of σ^2 if the error terms were actually homoskedastic.

A third possibility is to use a technique called the "jackknife" that we will not attempt to discuss here; see MacKinnon and White (1985). The resulting

or, in more compact notation, as

$$\sigma_t^2 = \alpha + A(L, \boldsymbol{\gamma}) u_t^2 + B(L, \boldsymbol{\delta}) \sigma_t^2,$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are parameter vectors with typical elements γ_i and δ_j, respectively, and $A(L, \boldsymbol{\gamma})$ and $B(L, \boldsymbol{\delta})$ are polynomials in the lag operator L. In the GARCH model, the conditional variance σ_t^2 depends on its own past values as well as on lagged values of u_t^2. This means that σ_t^2 effectively depends on all past values of u_t^2. In practice, a GARCH model with very few parameters often performs as well as an ARCH model with many parameters. In particular, one simple model that often works very well is the **GARCH(1, 1)** model,

$$\sigma_t^2 = \alpha + \gamma_1 u_{t-1}^2 + \delta_1 \sigma_{t-1}^2. \tag{16.21}$$

In practice, one must solve a GARCH model to eliminate the σ_{t-j}^2 terms from the right-hand side before one can estimate it. The problem is essentially the same as estimating a moving average model or an ARMA model with a moving average component; see Section 10.7. For example, the GARCH(1, 1) model (16.21) can be solved recursively to yield

$$\sigma_t^2 = \frac{\alpha}{1 - \delta_1} + \gamma_1 \big(u_{t-1}^2 + \delta_1 u_{t-2}^2 + \delta_1^2 u_{t-3}^2 + \delta_1^3 u_{t-4}^2 + \cdots \big). \tag{16.22}$$

Various assumptions can be made about the presample error terms. The simplest is to assume that they are zero, but it is more realistic to assume that they are equal to their unconditional expectation.

It is interesting to observe that, when γ_1 and δ_1 are both near zero, the solved GARCH(1, 1) model (16.22) looks like an ARCH(1) model. Because of this, it turns out that an appropriate test for GARCH(1, 1) errors is simply to regress the squared residuals on a constant term and on the squared residuals lagged once. In general, an LM test against GARCH(p, q) errors is the same as an LM test against ARCH($\max(p, q)$) errors. These results are completely analogous to the results for testing against ARMA(p, q) errors that we discussed in Section 10.8.

There are three principal ways to estimate regression models with ARCH and GARCH errors: feasible GLS, one-step efficient estimation, and maximum likelihood. In the simplest approach, which is feasible GLS, one first estimates the regression model by ordinary or nonlinear least squares, then uses the squared residuals to estimate the parameters of the ARCH or GARCH process, and finally uses weighted least squares to estimate the parameters of the regression function. This procedure can run into difficulties if the conditional variances predicted by the fitted ARCH process are not all positive, and various ad hoc methods may then be used to ensure that they are all positive.

The estimates of the ARCH parameters obtained by this sort of feasible GLS procedure will not be asymptotically efficient. Engle (1982b) therefore

Next, we derive a useful and general result that will allow us to replace the vector of derivatives $\boldsymbol{\mu}_0$ in (16.54) by something more manageable. The moment condition under test is given by (16.48). The moment can be written out explicitly as

$$E_{\boldsymbol{\theta}}\big(m_t(y_t, \boldsymbol{\theta})\big) = \int_{-\infty}^{\infty} m_t(y_t, \boldsymbol{\theta}) L_t(y_t, \boldsymbol{\theta}) dy_t. \qquad (16.55)$$

Differentiating the right-hand side of (16.55) with respect to the components of $\boldsymbol{\theta}$, we obtain, by the same sort of reasoning as led to the information matrix equality (8.44),

$$E_{\boldsymbol{\theta}}\big(m_t(\boldsymbol{\theta}) \boldsymbol{G}_t(\boldsymbol{\theta})\big) = -E_{\boldsymbol{\theta}}\big(\boldsymbol{N}_t(\boldsymbol{\theta})\big). \qquad (16.56)$$

Here $\boldsymbol{G}_t(\boldsymbol{\theta})$ is the contribution made by observation t to the gradient of the loglikelihood function, and the $1 \times k$ row vector $\boldsymbol{N}_t(\boldsymbol{\theta})$ has typical element $\partial m_t(\boldsymbol{\theta})/\partial \theta_i$.[5] The most useful form of our result is obtained by summing (16.56) over t. Let $\boldsymbol{m}(\boldsymbol{\theta})$ be an n–vector with typical element $m_t(\boldsymbol{\theta})$, and let $\boldsymbol{N}(\boldsymbol{\theta})$ be an $n \times k$ matrix with typical row $\boldsymbol{N}_t(\boldsymbol{\theta})$. Then

$$\frac{1}{n} E_{\boldsymbol{\theta}}\big(\boldsymbol{G}^{\top}(\boldsymbol{\theta}) \boldsymbol{m}(\boldsymbol{\theta})\big) = -\frac{1}{n} E_{\boldsymbol{\theta}}\big(\boldsymbol{N}^{\top}(\boldsymbol{\theta}) \boldsymbol{\iota}\big), \qquad (16.57)$$

where, as usual, $\boldsymbol{G}(\boldsymbol{\theta})$ denotes the CG matrix. In (16.54), $\boldsymbol{\mu}_0 = n^{-1} \boldsymbol{N}_0^{\top} \boldsymbol{\iota}$, where $\boldsymbol{N}_0 \equiv \boldsymbol{N}(\boldsymbol{\theta}_0)$. By the law of large numbers, this will converge to the limit of the right-hand side of (16.57), and so also to the limit of the left-hand side. Thus, if $\boldsymbol{G}_0 \equiv \boldsymbol{G}(\boldsymbol{\theta}_0)$, we can assert that

$$\boldsymbol{\mu}_0 = \frac{1}{n} \boldsymbol{N}_0^{\top} \boldsymbol{\iota} \stackrel{a}{=} -\frac{1}{n} \boldsymbol{G}_0^{\top} \boldsymbol{m}_0. \qquad (16.58)$$

We next make use of the very well-known result (13.18) on the relationship between ML estimates, the information matrix, and the score vector:

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \mathfrak{I}_0^{-1} n^{-1/2} \boldsymbol{g}_0. \qquad (16.59)$$

Since the information matrix $\mathfrak{I}_0$ is asymptotically equal to $n^{-1} \boldsymbol{G}_0^{\top} \boldsymbol{G}_0$ (see Section 8.6), and $\boldsymbol{g}_0 = \boldsymbol{G}_0^{\top} \boldsymbol{\iota}$, (16.59) becomes

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \big(n^{-1} \boldsymbol{G}_0^{\top} \boldsymbol{G}_0\big)^{-1} n^{-1/2} \boldsymbol{G}_0^{\top} \boldsymbol{\iota}.$$

This result, combined with (16.58), allows us to replace the right-hand side of (16.54) by

$$n^{-1/2} \boldsymbol{m}_0^{\top} \boldsymbol{\iota} - n^{-1} \boldsymbol{m}_0^{\top} \boldsymbol{G}_0 \big(n^{-1} \boldsymbol{G}_0^{\top} \boldsymbol{G}_0\big)^{-1} n^{-1/2} \boldsymbol{G}_0^{\top} \boldsymbol{\iota} = n^{-1/2} \boldsymbol{m}_0^{\top} \boldsymbol{M}_G \boldsymbol{\iota}, \quad (16.60)$$

where $\boldsymbol{M}_G$ denotes the matrix that projects orthogonally onto $\mathcal{S}^{\perp}(\boldsymbol{G}_0)$.

---

[5] Our usual notation would have been $\boldsymbol{M}_t(\boldsymbol{\theta})$ instead of $\boldsymbol{N}_t(\boldsymbol{\theta})$, but this would conflict with the standard notation for complementary orthogonal projections.

The result (16.60) makes clear just what the difference is between the empirical moment evaluated at the unknown $\boldsymbol{\theta}_0$ and evaluated at the ML estimates $\hat{\boldsymbol{\theta}}$, that is, between $n^{-1/2}\boldsymbol{m}_0^\top\boldsymbol{\iota}$ and $n^{-1/2}\hat{\boldsymbol{m}}^\top\boldsymbol{\iota}$. The effect of using the estimates is an implicit orthogonal projection of the vector $\boldsymbol{m}_0$ onto the orthogonal complement of the space $\mathcal{S}(\boldsymbol{G}_0)$ associated with the model parameters. This projection is what causes the variance of the expression that we can actually calculate to be smaller than the variance of the corresponding expression based on the true parameters. The variances used in the skewness and kurtosis tests discussed in the last section can also be computed using (16.60).

We are now ready to obtain an appropriate expression for the asymptotic variance of $n^{-1/2}\hat{\boldsymbol{m}}^\top\boldsymbol{\iota}$. We require, as we suggested earlier, that $n^{-1/2}\boldsymbol{m}_0^\top\boldsymbol{\iota}$ should satisfy CLT and that, in a neighborhood of $\boldsymbol{\theta}_0$, $n^{-1}\boldsymbol{m}^\top(\boldsymbol{\theta})\boldsymbol{G}_i(\boldsymbol{\theta})$ should satisfy WULLN (Definition 4.17) for all $i = 1, \ldots, k$. The asymptotic variance is then clearly $\text{plim}(n^{-1}\boldsymbol{m}_0^\top\boldsymbol{M}_G\boldsymbol{m}_0)$, which can be consistently estimated by $n^{-1}\hat{\boldsymbol{m}}^\top\hat{\boldsymbol{M}}_G\hat{\boldsymbol{m}}$. This suggests using the test statistic

$$\frac{n^{-1/2}\hat{\boldsymbol{m}}^\top\boldsymbol{\iota}}{\left(n^{-1}\hat{\boldsymbol{m}}^\top\hat{\boldsymbol{M}}_G\hat{\boldsymbol{m}}\right)^{1/2}} = \frac{\hat{\boldsymbol{m}}^\top\boldsymbol{\iota}}{\left(\hat{\boldsymbol{m}}^\top\hat{\boldsymbol{M}}_G\hat{\boldsymbol{m}}\right)^{1/2}}, \tag{16.61}$$

which will be asymptotically distributed as $N(0, 1)$.

The connection with the OPG regression is now evident. The test statistic (16.61) is *almost* the t statistic on the coefficient b from the following OPG regression:

$$\boldsymbol{\iota} = \hat{\boldsymbol{G}}\boldsymbol{c} + b\hat{\boldsymbol{m}} + \text{residuals}. \tag{16.62}$$

Asymptotically, the statistic (16.61) and the t statistic from (16.62) are equivalent, because the sum of squared residuals from (16.62) tends to n for large sample sizes under the null hypothesis: The regressors $\hat{\boldsymbol{G}}$ are always orthogonal to $\boldsymbol{\iota}$, and $\hat{\boldsymbol{m}}$ is orthogonal to $\boldsymbol{\iota}$ if the moment condition is satisfied. This result is very satisfactory. Without the regressor $\hat{\boldsymbol{m}}$, which is the vector that serves to define the empirical moment, regression (16.62) would be just the OPG regression associated with the original model, and the SSR would always be equal to n. Thus the OPG version of the CM test, like all the other tests we have discussed that are implemented by artificial regressions, is just a test for the significance of the coefficients on one or more test regressors.

It is now plain how to extend CM tests to a set of two or more moment conditions. One simply creates a test regressor for each of the empirical moments so as to produce an $n \times r$ matrix $\hat{\boldsymbol{R}} \equiv \boldsymbol{R}(\hat{\boldsymbol{\theta}})$, where r is the number of moment conditions. One then uses the explained sum of squares from the OPG regression

$$\boldsymbol{\iota} = \hat{\boldsymbol{G}}\boldsymbol{c} + \hat{\boldsymbol{R}}\boldsymbol{b} + \text{residuals}$$

or any other asymptotically equivalent test of the artificial hypothesis $\boldsymbol{b} = \boldsymbol{0}$. It is now clear that, as we suggested above, any test capable of being carried out by means of an OPG regression can be interpreted as a CM test.

One simply has to interpret the test columns in the regression as empirical moments.

An interesting variant of the test regression (16.62) was suggested by Tauchen (1985). In effect, he interchanged the regressand $\boldsymbol{\iota}$ and the test regressor $\hat{\boldsymbol{m}}$ so as to obtain the regression

$$\hat{\boldsymbol{m}} = \hat{\boldsymbol{G}}\boldsymbol{c}^* + b^*\boldsymbol{\iota} + \text{residuals}. \tag{16.63}$$

The test statistic is the ordinary t statistic for $b^* = 0$. It is numerically identical to the t statistic on b in (16.62). This fact follows from a result we obtained in section 12.7, of which we now give a different, geometrical, proof. Apply the FWL Theorem to both (16.62) and (16.63) so as to obtain the two regressions

$$\hat{\boldsymbol{M}}_{\boldsymbol{G}}\boldsymbol{\iota} = b(\hat{\boldsymbol{M}}_{\boldsymbol{G}}\hat{\boldsymbol{m}}) + \text{residuals} \quad \text{and}$$
$$\hat{\boldsymbol{M}}_{\boldsymbol{G}}\hat{\boldsymbol{m}} = b^*(\hat{\boldsymbol{M}}_{\boldsymbol{G}}\boldsymbol{\iota}) + \text{residuals}. \tag{16.64}$$

These are both univariate regressions with n observations. The single t statistic from each of them is given by the product of the same scalar factor, $(n-1)^{1/2}$, and the cotangent of the angle between the regressand and the regressor (see Appendix A). Since this angle is unchanged when the regressor and regressand are interchanged, so is the t statistic. The FWL Theorem implies that the t statistics from the first and second rows of (16.64) are equal to those from the OPG regression (16.62) and Tauchen's regression (16.63), respectively, times the same degrees of freedom correction. Thus we conclude that the t statistics based on the latter two regressions are numerically identical.

Since the first-order conditions for $\hat{\boldsymbol{\theta}}$ imply that $\boldsymbol{\iota}$ is orthogonal to all of the columns of $\hat{\boldsymbol{G}}$, the OLS estimate of b^* in (16.63) will be equal to the sample mean of the elements of $\hat{\boldsymbol{m}}$. This would be so even if the regressors $\hat{\boldsymbol{G}}$ were omitted from the regression. However, because $\boldsymbol{\theta}$ has been estimated, those regressors must be included if we are to obtain a valid estimate of the variance of the sample mean. As is the case with all the other artificial regressions we have studied, omitting the regressors that correspond to parameters estimated under the null hypothesis results in a test statistic that is too small, asymptotically.

Let us reiterate our earlier warnings about the OPG regression. As we stressed when we introduced it in Section 13.7, test statistics based on it often have poor finite-sample properties. They tend to reject the null hypothesis too often when it is true. This is just as true for CM tests as for LM tests or $C(\alpha)$ tests. If possible, one should therefore use alternative tests that have better finite-sample properties, such as tests based on the GNR, the HRGNR, the DLR (Section 14.4), or the BRMR (Section 15.4), when these procedures are applicable. Of course, they will be applicable in general only if the CM test can be reformulated as an ordinary test, with an explicit alternative

and evaluate it at the ML estimates $\hat{\boldsymbol{\theta}}$ to obtain $\hat{\boldsymbol{Z}}$. Then one performs an OPG regression, with regressors $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{Z}}$, and uses n minus the SSR as the test statistic. Provided the matrix $[\hat{\boldsymbol{G}}\ \ \hat{\boldsymbol{Z}}]^{\top}[\hat{\boldsymbol{G}}\ \ \hat{\boldsymbol{Z}}]$ has full rank asymptotically, the test statistic will be asymptotically distributed as $\chi^2\left(\frac{1}{2}k(k+1)\right)$. When some of the columns of $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{Z}}$ are perfectly collinear, as quite often happens, the number of degrees of freedom for the test must of course be reduced accordingly.

It is illuminating to consider as an example the univariate nonlinear regression model

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2),$$

where $x_t(\boldsymbol{\beta})$ is a twice continuously differentiable function that depends on $\boldsymbol{\beta}$, a p–vector of parameters, and also on exogenous and predetermined variables which vary across observations. Thus the total number of parameters is $k = p + 1$. For this model, the contribution to the loglikelihood function from the tth observation is

$$\ell_t(\boldsymbol{\beta}, \sigma) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}\left(y_t - x_t(\boldsymbol{\beta})\right)^2.$$

Thus the contribution from the tth observation to the regressor corresponding to the ith element of $\boldsymbol{\beta}$ is

$$G_{ti}(\boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^2}\left(y_t - x_t(\boldsymbol{\beta})\right)X_{ti}(\boldsymbol{\beta}), \tag{16.66}$$

where, as usual, $X_{ti}(\boldsymbol{\beta})$ denotes the derivative of $x_t(\boldsymbol{\beta})$ with respect to β_i. Similarly, the contribution from the tth observation to the regressor corresponding to σ is

$$G_{t,k}(\boldsymbol{\beta}, \sigma) = -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left(y_t - x_t(\boldsymbol{\beta})\right)^2. \tag{16.67}$$

Using (16.66) and (16.67), it is easy to work out the regressors for the OPG version of the IM test. We make the definitions

$$\hat{e}_t \equiv \frac{1}{\hat{\sigma}}\left(y_t - x_t(\hat{\boldsymbol{\beta}})\right), \ \ \hat{X}_{ti} \equiv X_{ti}(\hat{\boldsymbol{\beta}}), \ \ \text{and} \ \ X_{tij}^*(\boldsymbol{\beta}) \equiv \frac{\partial X_{ti}(\boldsymbol{\beta})}{\partial \beta_j}.$$

Then, up to multiplicative factors that can have no effect on the fit of the regression, and hence no effect on the value of the IM test statistic, the regressors for the test regression are

$$\text{for } \beta_i : \qquad \hat{e}_t \hat{X}_{ti}; \tag{16.68}$$

$$\text{for } \sigma : \qquad \hat{e}_t^2 - 1; \tag{16.69}$$

$$\text{for } \beta_i \times \beta_j : \quad (\hat{e}_t^2 - 1)\hat{X}_{ti}\hat{X}_{tj} + \hat{\sigma}\hat{e}_t\hat{X}_{tij}^*; \tag{16.70}$$

$$\text{for } \sigma \times \beta_i : \quad (\hat{e}_t^3 - 3\hat{e}_t)\hat{X}_{ti}; \tag{16.71}$$

$$\text{for } \sigma \times \sigma : \quad \hat{e}_t^4 - 5\hat{e}_t^2 + 2. \tag{16.72}$$

Expressions (16.68) and (16.69) give the elements of each row of $\hat{\mathbf{G}}$, while expressions (16.70)–(16.72) give the elements of each row of $\hat{\mathbf{Z}}$. When the original regression contains a constant term, (16.69) will be perfectly collinear with (16.70) when i and j both index the constant. Therefore, the latter will have to be dropped and the degrees of freedom for the test reduced by one to $\frac{1}{2}(p+2)(p+1) - 1$.

Expressions (16.68)–(16.72) show what forms of misspecification the IM test is testing for in the nonlinear regression context. It is evident from (16.71) that the (β_i, σ) regressors are those corresponding to skewness interacting with the $\hat{X}_{ti}$'s. It appears that such skewness, if present, would bias the estimates of the covariances of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$. If we add five times (16.69) to (16.72), the result is $\hat{e}_t^4 - 3$, from which we see that the linearly independent part of the (σ, σ) regressor is testing in the kurtosis direction. Either platykurtosis or leptokurtosis would lead to bias in the estimate of the variance of $\hat{\sigma}$. It is evident from (16.70) that if $x_t(\boldsymbol{\beta})$ were linear, the (β_i, β_j) regressors would be testing for heteroskedasticity of exactly the type that White's (1980) test is designed to detect; see Section 16.5. In the nonlinear regression case considered here, however, these regressors are testing at the same time for misspecification of the regression function. For more details on the special case of linear regression models, see Hall (1987).

The above analysis suggests that, in the case of regression models, it is probably more attractive to test directly for heteroskedasticity, skewness, kurtosis, and misspecification of the regression function than to use an IM test. We have already seen how to test for each of these types of misspecification individually. Individual tests may well be more powerful and more informative than an IM test, especially if only a few things are actually wrong with the model. If one is primarily interested in inferences about $\boldsymbol{\beta}$, then testing for skewness and kurtosis may be optional.

There is one very serious problem with IM tests based on the OPG regression. In finite samples, they tend to reject the null hypothesis much too often when it is true. In this respect, IM tests seem to be even worse than other specification tests based on the OPG regression. Monte Carlo results demonstrating the dreadful finite-sample performance of the OPG version of the IM test may be found in Taylor (1987), Kennan and Neumann (1988), Orme (1990a), Hall (1990), Chesher and Spady (1991), and Davidson and MacKinnon (1992a). In some of these papers, there are cases in which OPG IM tests reject correct null hypotheses virtually all the time. The problem seems to grow worse as the number of degrees of freedom increases, and it does not go away quickly as the sample size increases. One extreme example, given in Davidson and MacKinnon (1992a), is a linear regression model with 10 regressors, and thus 65 degrees of freedom, for which the OPG form of the IM test rejects the true null hypothesis at the nominal 5% level an amazing 99.9% of the time when $n = 200$ and 92.7% of the time even when $n = 1000$.

where $\boldsymbol{g}(\boldsymbol{\theta})$ denotes the gradient of Q, that is, the k–vector with typical component $\partial Q(\boldsymbol{\theta})/\partial \theta_j$. As usual, $\mathcal{H}^*$ denotes a matrix of which the elements are evaluated at the appropriate $\boldsymbol{\theta}_j^*$.

If we are to be able to deduce the asymptotic normality of $\hat{\boldsymbol{\theta}}$ from (17.21), it must be possible to apply a law of large numbers to $\mathcal{H}^*$ and a central limit theorem to $n^{1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$. We would then obtain the result that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\Big(\plim_{n\to\infty} \mathcal{H}_0\Big)^{-1} n^{1/2}\boldsymbol{g}(\boldsymbol{\theta}_0). \tag{17.22}$$

What regularity conditions do we need for (17.22)? First, in order to justify the short Taylor expansion in (17.20), it is necessary that Q be at least twice continuously differentiable with respect to $\boldsymbol{\theta}$. If so, then it follows that the Hessian of Q is $O(1)$ as $n \to \infty$. Because of this, we denote it by $\mathcal{H}_0$ rather than $\boldsymbol{H}$; see Section 8.2. Then we need conditions that allow the application of a law of large numbers and a central limit theorem. Rather formally, we may state a theorem based closely on Theorem 8.3 as follows:

Theorem 17.2. Asymptotic Normality of M-Estimators

The *M*-estimator derived from the sequence of criterion functions Q is asymptotically normal if it satisfies the conditions of Theorem 17.1 and if in addition

(i) for all n and for all $\boldsymbol{\theta} \in \Theta$, $Q^n(\boldsymbol{y}^n, \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ for almost all $\boldsymbol{y}$, and the limit function $\bar{Q}(\mu, \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \Theta$ and for all $\mu \in \mathbb{M}$;

(ii) for all DGPs $\mu \in \mathbb{M}$ and for all sequences $\{\boldsymbol{\theta}^n\}$ that tend in probability to $\boldsymbol{\theta}(\mu)$ as $n \to \infty$, the Hessian matrix $\mathcal{H}^n(\boldsymbol{y}^n, \boldsymbol{\theta}^n)$ of Q^n with respect to $\boldsymbol{\theta}$ tends uniformly in probability to a positive definite, finite, nonrandom matrix $\mathcal{H}(\mu)$; and

(iii) for all DGPs $\mu \in \mathbb{M}$, $n^{1/2}$ times the gradient of $Q^n(\boldsymbol{y}^n, \boldsymbol{\theta})$, or $n^{1/2}\boldsymbol{g}\big(\boldsymbol{y}^n, \boldsymbol{\theta}(\mu)\big)$, converges in distribution as $n \to \infty$ to a multivariate normal distribution with mean zero and finite covariance matrix $\boldsymbol{V}(\mu)$.

Under these conditions, the distribution of $n^{1/2}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\mu)\big)$ tends to $N\big(\boldsymbol{0},\ \mathcal{H}(\mu)^{-1}\boldsymbol{V}(\mu)\mathcal{H}(\mu)^{-1}\big)$.

It is not worth spending any time on the proof of Theorem 17.2. What we must do, instead, is to return to the GMM case and investigate the conditions under which the criterion function (17.13), suitably divided by n^2, satisfies the requirements of the theorem. Without further ado, we assume that all of the contributions $f_{ti}(y_t, \boldsymbol{\theta})$ are at least twice continuously differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \Theta$, for all y_t, and for all allowed values of any predetermined or exogenous variables on which they may depend. Next, we

$\boldsymbol{A}_0$ by $\boldsymbol{A}(\boldsymbol{y})$, and $\boldsymbol{\Phi}_{ij}$ by expression (17.29) without the probability limit. Although this yields a consistent estimate of (17.30), it is often a very noisy one. We will discuss this issue further in Section 17.5, but it is still far from being completely resolved.

It is interesting to illustrate (17.31) for the case of the IV estimator defined by (17.08). The result will enable us to construct a heteroskedasticity-consistent estimate of the covariance matrix of the latter. We merely have to establish some notational equivalences between the IV case and the more general case discussed above. In the IV case, the elements of the matrix $\boldsymbol{F}$ become $f_{ti} = W_{ti}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})$. Therefore,

$$\boldsymbol{D} = -\plim_{n\to\infty}\left(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{X}\right) \tag{17.33}$$

and

$$\boldsymbol{A}_0 = \plim_{n\to\infty}\left(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}\right)^{-1}. \tag{17.34}$$

The matrix $\boldsymbol{\Phi}$ is obtained from (17.29):

$$\boldsymbol{\Phi} = \plim_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2\boldsymbol{W}_t^\top\boldsymbol{W}_t\right) = \plim_{n\to\infty}\left(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}\right), \tag{17.35}$$

where $\boldsymbol{\Omega}$ is the diagonal matrix with typical element $E(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2$. By substituting (17.33), (17.34), and (17.35) into (17.31), we obtain the following expression for the asymptotic covariance matrix of the IV estimator:

$$\plim_{n\to\infty}\left(\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{P}_W\boldsymbol{X}\right)^{-1}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{P}_W\boldsymbol{\Omega}\boldsymbol{P}_W\boldsymbol{X}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{P}_W\boldsymbol{X}\right)^{-1}\right). \tag{17.36}$$

The matrix (17.36) is clearly analogous for IV estimation to (16.08) for NLS estimation: It provides the asymptotic covariance matrix in the presence of heteroskedasticity of unknown form. Thus we see that HCCMEs of the sort discussed in Section 16.3 are available for the IV estimator. One can use any of the inconsistent estimators $\hat{\boldsymbol{\Omega}}$ suggested there in order to obtain a consistent estimator of $\plim\left(n^{-1}\boldsymbol{X}^\top\boldsymbol{P}_W\boldsymbol{\Omega}\boldsymbol{P}_W\boldsymbol{X}\right)$.

Readers may reasonably wonder why we have obtained a covariance matrix robust *only* to heteroskedasticity and not also to serial correlation of the error terms. The answer is that the covariance matrix $\boldsymbol{V}$ of (17.30) is valid only if condition CLT is satisfied by the contributions to the empirical moments. That condition will *not* be satisfied if the error terms have an arbitrary pattern of correlation among themselves. In Section 17.5, we will discuss methods for dealing with serial correlation, but these will take us out of the asymptotic framework we have used up to now.

17.3 EFFICIENT GMM ESTIMATORS

It is not completely straightforward to answer the question of whether GMM estimators are asymptotically efficient, since a number of separate issues are involved. The first issue was raised at the beginning of the last section, in connection with estimation by instrumental variables. We saw there that, for a given set of empirical moments $\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, a whole family of estimators can be generated by different choices of the weighting matrix $\boldsymbol{A}(\boldsymbol{y})$ used to construct a quadratic form from the moments. Asymptotically, the most efficient of these estimators is obtained by choosing $\boldsymbol{A}(\boldsymbol{y})$ such that it tends to a nonrandom probability limit proportional to the inverse of the limiting covariance matrix of the empirical moments, suitably weighted by an appropriate power of the sample size n. This turns out to be true quite generally, as we now show.

Theorem 17.3. A Necessary Condition for Efficiency

A necessary condition for the estimator obtained by minimizing the quadratic form (17.13) to be asymptotically efficient is that it should be asymptotically equal to the estimator defined by minimizing (17.13) with $\boldsymbol{A}(\boldsymbol{y})$ independent of $\boldsymbol{y}$ and equal to the inverse of the asymptotic covariance matrix of the empirical moments $n^{-1/2}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{\iota}$.

Note that, when the necessary condition holds, the form of the asymptotic covariance matrix of the GMM estimator $\hat{\boldsymbol{\theta}}$ becomes much simpler. For arbitrary limiting weighting matrix $\boldsymbol{A}_0$, that matrix was given by (17.31). If the necessary condition is satisfied, then $\boldsymbol{A}_0$ in (17.31) may be replaced by the inverse of $\boldsymbol{\Phi}$, which, according to its definition (17.29), is the asymptotic covariance of the empirical moments. Substituting $\boldsymbol{A}_0 = \boldsymbol{\Phi}^{-1}$ into (17.31) gives the simple result that

$$\boldsymbol{V}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = \big(\boldsymbol{D}^\top\boldsymbol{\Phi}^{-1}\boldsymbol{D}\big)^{-1}.$$

Theorem 17.3 will be proved if we can show that, for all symmetric, positive definite matrices $\boldsymbol{A}_0$, the difference

$$\big(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D}\big)^{-1}\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{\Phi}\boldsymbol{A}_0\boldsymbol{D}\big(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D}\big)^{-1} - \big(\boldsymbol{D}^\top\boldsymbol{\Phi}^{-1}\boldsymbol{D}\big)^{-1} \qquad (17.37)$$

is positive semidefinite. To show this, we rewrite (17.37) as

$$\big(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D}\big)^{-1}\boldsymbol{D}^\top\boldsymbol{A}_0\Big(\boldsymbol{\Phi} - \boldsymbol{D}\big(\boldsymbol{D}^\top\boldsymbol{\Phi}^{-1}\boldsymbol{D}\big)^{-1}\boldsymbol{D}^\top\Big)\boldsymbol{A}_0\boldsymbol{D}\big(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D}\big)^{-1}. \qquad (17.38)$$

Since the matrix $\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D}$ is nonsingular, (17.38) is positive definite if the matrix in large parentheses is. Since $\boldsymbol{\Phi}$ is a positive definite, symmetric $l \times l$ matrix, we can find another positive definite, symmetric $l \times l$ matrix $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi}^2 = \boldsymbol{\Phi}^{-1}$. In terms of $\boldsymbol{\Psi}$, the matrix in large parentheses becomes

$$\boldsymbol{\Psi}^{-1}\big(\mathbf{I} - \boldsymbol{P}_{\Psi D}\big)\boldsymbol{\Psi}^{-1} = \boldsymbol{\Psi}^{-1}\boldsymbol{M}_{\Psi D}\boldsymbol{\Psi}^{-1}, \qquad (17.39)$$

require of course that $l \geq k$, where the parameter vector $\hat{\boldsymbol{\theta}}$ has k elements. The empirical moment conditions that we use for estimation can be expressed as

$$\boldsymbol{W}^{\top}\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}, \tag{17.47}$$

where $\boldsymbol{f}$ is an n–vector with typical component f_t. If $l = k$, the estimator $\hat{\boldsymbol{\theta}}$ is obtained by solving the k equations (17.47). If $l > k$, it is obtained by minimizing the quadratic form constructed from the components of the left-hand side of (17.47) and an estimate of their covariance matrix. Let $\boldsymbol{\Omega}$ denote the covariance matrix of the f_t's. Thus, if the DGP is denoted by μ and the true parameter vector by $\boldsymbol{\theta}_0$,

$$\Omega_{ts} = E_{\mu}\big(f_t(\boldsymbol{\theta}_0)f_s(\boldsymbol{\theta}_0) \,|\, \Omega_t\big) \quad \text{for all } t \leq s.$$

Then the conditional covariance matrix of the empirical moments in (17.47) is $\boldsymbol{\Phi} \equiv \boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}$.

In the usual case, with $l > k$, the criterion function used for obtaining parameter estimates is

$$\boldsymbol{f}(\boldsymbol{\theta})^{\top}\boldsymbol{W}\big(\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}\big)^{-1}\boldsymbol{W}^{\top}\boldsymbol{f}(\boldsymbol{\theta}).$$

The asymptotic covariance matrix of this estimator is given by the probability limit of $(\boldsymbol{D}^{\top}\boldsymbol{\Phi}^{-1}\boldsymbol{D})^{-1}$, where

$$\boldsymbol{D}_{ij} = \plim_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n} W_{ti}\frac{\partial f_t}{\partial \theta_j}\right). \tag{17.48}$$

Let $\boldsymbol{J}(\boldsymbol{y},\boldsymbol{\theta})$ denote the $n \times k$ matrix with typical element $\partial f_t(y_t, \boldsymbol{\theta})/\partial \theta_j$.[1] Then the right-hand side of (17.48) is the limit of $n^{-1}\boldsymbol{W}^{\top}\boldsymbol{J}$. Thus the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ reduces to the limit of

$$\left(\left(\tfrac{1}{n}\boldsymbol{J}^{\top}\boldsymbol{W}\right)\left(\tfrac{1}{n}\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{W}^{\top}\boldsymbol{J}\right)\right)^{-1}. \tag{17.49}$$

The first result about how to choose the instruments $\boldsymbol{W}$ optimally is simple and intuitive. It is that if we increase the number of instruments, the limiting covariance matrix (17.49) cannot increase. Imagine that instead of the empirical moment conditions (17.47) we use a set of linear combinations of them. That is, we replace (17.47) by

$$\boldsymbol{B}^{\top}\boldsymbol{W}^{\top}\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0},$$

---

[1] The notation $\boldsymbol{J}$ was chosen because the matrix is the *Jacobian* of $\boldsymbol{f}$ with respect to $\boldsymbol{\theta}$ and because $\boldsymbol{F}$ was previously used to denote something else.

for some $l \times p$ matrix $\boldsymbol{B}$, where $p \leq l$. It is easy to see that this corresponds to replacing $\boldsymbol{D}$ by $\boldsymbol{B}^{\top}\boldsymbol{D}$ and $\boldsymbol{\Phi}$ by $\boldsymbol{B}^{\top}\boldsymbol{\Phi}\boldsymbol{B}$. Consider the difference

$$\boldsymbol{D}^{\top}\boldsymbol{\Phi}^{-1}\boldsymbol{D} - \boldsymbol{D}^{\top}\boldsymbol{B}\big(\boldsymbol{B}^{\top}\boldsymbol{\Phi}\boldsymbol{B}\big)^{-1}\boldsymbol{B}^{\top}\boldsymbol{D}$$

between the inverses of the $k \times k$ asymptotic covariance matrices corresponding to the instruments $\boldsymbol{W}$ and $\boldsymbol{W}\boldsymbol{B}$, respectively. If, as before, we denote by $\boldsymbol{\Psi}$ a symmetric $l \times l$ matrix such that $\boldsymbol{\Psi}^2 = \boldsymbol{\Phi}^{-1}$, this difference is

$$\boldsymbol{D}^{\top}\boldsymbol{\Psi}\Big(\mathbf{I} - \boldsymbol{\Psi}^{-1}\boldsymbol{B}\big(\boldsymbol{B}^{\top}\boldsymbol{\Psi}^{-2}\boldsymbol{B}\big)^{-1}\boldsymbol{B}^{\top}\boldsymbol{\Psi}^{-1}\Big)\boldsymbol{\Psi}\boldsymbol{D}. \qquad (17.50)$$

This matrix is clearly positive semidefinite, because the matrix in large parentheses is the orthogonal projection off the columns of $\boldsymbol{\Psi}^{-1}\boldsymbol{B}$. For any two symmetric, positive definite matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ of the same dimension, $\boldsymbol{P} - \boldsymbol{Q}$ is positive semidefinite if and only if $\boldsymbol{Q}^{-1} - \boldsymbol{P}^{-1}$ is positive semidefinite (see Appendix A). Thus the fact that (17.50) is positive semidefinite establishes our first result.

This result might seem to suggest that one should always use as many instruments as possible in order to get as efficient estimates as possible. Such a conclusion is generally wrong, however. Recall the discussion in Section 7.5, illustrated by Figure 7.1. There we saw that, in the ordinary IV context, there is a trade-off between asymptotic efficiency and bias in finite samples. The same trade-off arises in the GMM case as well. Using a large number of overidentifying restrictions may lead to a smaller asymptotic covariance matrix, but the estimates may be seriously biased. Another argument against the use of too many instruments is simply that there are inevitably diminishing returns, on account of the existence of the GMM bound.

The second result shows how to choose the instruments $\boldsymbol{W}$ optimally. It says that if we set $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}\boldsymbol{J}$ in (17.47), then the asymptotic covariance matrix that results is smaller than the one given by any other choice. From (17.49) it then follows that the GMM bound for the asymptotic covariance matrix is plim $(n^{-1}\boldsymbol{J}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{J})^{-1}$. Unfortunately, as we will see, this result is not always useful in practice.

The proof is very simple. As with the first result, it is easiest to work with the inverses of the relevant covariance matrices. Let the symmetric $n \times n$ matrix $\boldsymbol{\Upsilon}$ be defined so that $\boldsymbol{\Upsilon}^2 \equiv \boldsymbol{\Omega}$. Then, suppressing limits and factors of n for the moment, we see that

$$\begin{aligned}
&\boldsymbol{J}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{J} - \boldsymbol{J}^{\top}\boldsymbol{W}\big(\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}\big)^{-1}\boldsymbol{W}^{\top}\boldsymbol{J} \\
&= \boldsymbol{J}^{\top}\boldsymbol{\Upsilon}^{-1}\Big(\mathbf{I} - \boldsymbol{\Upsilon}\boldsymbol{W}\big(\boldsymbol{W}^{\top}\boldsymbol{\Upsilon}^2\boldsymbol{W}\big)^{-1}\boldsymbol{W}^{\top}\boldsymbol{\Upsilon}\Big)\boldsymbol{\Upsilon}^{-1}\boldsymbol{J}.
\end{aligned} \qquad (17.51)$$

Since the matrix in large parentheses is the orthogonal projection off the columns of $\boldsymbol{\Upsilon}\boldsymbol{W}$, this expression is positive semidefinite, and the second result is established.

This equation defines an ordinary IV estimator in terms of the transformed variables $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$ and the transformed instruments $\boldsymbol{Z}$. Thus the estimator defined by (17.53) can be calculated with no more difficulty than the GLS estimator. It is appropriate to use it when GLS or feasible GLS would have been appropriate except for possible correlation of the error terms with the regressors.

The estimator defined by (17.53) bears a close resemblance to the H2SLS estimator (17.44) defined in the last section. In fact, replacing $\boldsymbol{W}$ in the latter by $\boldsymbol{\Omega}^{-1}\boldsymbol{W}$ yields the former. The theory developed in this section shows that if it is possible to choose $\boldsymbol{W}$ as the conditional expectations of the regressors $\boldsymbol{X}$ (or linear combinations of them), then the estimator defined by (17.53) is asymptotically efficient, and the H2SLS estimator is not. The advantage of H2SLS is that it can be calculated in the presence of heteroskedasticity of unknown form, since $n^{-1}\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}$ can be estimated consistently by use of inconsistent estimators of $\boldsymbol{\Omega}$. (17.53), on the other hand, can be formulated only if $\boldsymbol{\Omega}$ itself can be consistently estimated, because expressions like $n^{-1}\boldsymbol{W}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{W}$ and $n^{-1}\boldsymbol{W}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{y}$ *cannot* be estimated consistently without a consistent estimate of $\boldsymbol{\Omega}$. Thus both estimators are useful, but in different circumstances.

The concept of the GMM bound was introduced, not under that name, by Hansen (1985), who also provided conditions for optimal instruments. The arguments used in order to derive the bound have a longer history, however, and Hansen traces the history of the search for efficient instruments back as far as Basmann (1957) and Sargan (1958).

17.5 COVARIANCE MATRIX ESTIMATION

In previous sections, we mentioned the difficulties that can arise in estimating covariance matrices in the GMM context. In fact, problems occur at two distinct points: once for the choice of the weighting matrix to be used in constructing a criterion function and again for estimating the asymptotic covariance matrix of the estimates. Fortunately, similar considerations apply to both problems, and so we can consider them together.

Recall from (17.31) that the asymptotic covariance matrix of a GMM estimator computed using a weighting matrix $\boldsymbol{A}_0$ is

$$\left(\boldsymbol{D}^{\top}\boldsymbol{A}_0\boldsymbol{D}\right)^{-1}\boldsymbol{D}^{\top}\boldsymbol{A}_0\boldsymbol{\Phi}\boldsymbol{A}_0\boldsymbol{D}\left(\boldsymbol{D}^{\top}\boldsymbol{A}_0\boldsymbol{D}\right)^{-1},$$

in the notation of Section 17.2. If the necessary condition for efficiency of Theorem 17.3 is to be satisfied, it is required that $\boldsymbol{A}_0 \stackrel{a}{=} \boldsymbol{\Phi}^{-1}$, where $\boldsymbol{\Phi}$ is the $l \times l$ asymptotic covariance matrix of the empirical moments $n^{-1/2}\boldsymbol{F}^{\top}(\boldsymbol{\theta})\boldsymbol{\iota}$ with typical element

$$n^{-1/2}\sum_{t=1}^{n} f_{ti}(y_t, \boldsymbol{\theta}).$$

The estimator (17.63) was proposed by Hansen (1982) and White and Domowitz (1984), and was used in some of the earlier published work that employed GMM estimation, such as Hansen and Singleton (1982). From the point of view of theory, it is necessary to let the truncation parameter p, usually referred to as the **lag truncation parameter**, go to infinity at some suitable rate. A typical rate would be $n^{1/4}$, in which case $p = o(n^{1/4})$. This ensures that, for large enough n, all the nonzero $\boldsymbol{\Gamma}(j)$'s are estimated consistently. Unfortunately, this type of result is not of much use in practice, where one typically faces a given, finite n. We will return to this point a little later, and for the meantime suppose simply that we have somehow selected an appropriate value for p.

A much more serious difficulty associated with (17.63) is that, in finite samples, it need not be positive definite or even positive semidefinite. If one is unlucky enough to be working with a data set that yields a nondefinite $\hat{\boldsymbol{\Phi}}$, then (17.63) is unusable. There are numerous ways out of this difficulty. The most widely used was suggested by Newey and West (1987a). It is simply to multiply the $\hat{\boldsymbol{\Gamma}}(j)$'s by a sequence of weights that decrease as $|j|$ increases. Specifically, the estimator that they propose is

$$\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \left(1 - \frac{j}{p+1}\right)\left(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}(j)^{\top}\right). \tag{17.64}$$

It can be seen that the weights $1 - j/(p+1)$ decrease linearly with j from a value of 1 for $\hat{\boldsymbol{\Gamma}}(0)$ by steps of $1/(p+1)$ down to a value of $1/(p+1)$ for $|j| = p$. The use of such a set of weights is clearly compatible with the idea that the impact of the autocovariance of order j diminishes with $|j|$.

We will not attempt even to sketch a proof of the consistency of the Newey-West or similar estimators. We have alluded to the sort of regularity conditions needed for consistency to hold: Basically, the autocovariance matrices of the empirical moments must tend to zero quickly enough as p increases. It would also go well beyond the scope of this book to provide a theoretical justification for the Newey-West estimator. It rests on considerations of the so-called "frequency domain representation" of the $\boldsymbol{F}_t$'s and also of a number of notions associated with nonparametric estimation procedures. Interested readers are referred to Andrews (1991b) for a rather complete treatment of many of the issues. This paper suggests some alternatives to the Newey-West estimator and shows that in some circumstances they are preferable. However, the performance of the Newey-West estimator is never greatly inferior to that of the alternatives. Consequently, its simplicity is much in its favor.

Let us now return to the linear IV model with empirical moments given by $\boldsymbol{W}^{\top}(\boldsymbol{y} - \boldsymbol{X\beta})$. In order to be able to use (17.64), we suppose that the true error terms $u_t \equiv y_t - \boldsymbol{X}_t\boldsymbol{\beta}_0$ satisfy an appropriate mixing condition. Then the sample autocovariance matrices $\hat{\boldsymbol{\Gamma}}(j)$ for $j = 0, \ldots, p$, for some given p, are calculated as follows. A preliminary consistent estimate of $\boldsymbol{\beta}$ is first obtained

by the ordinary IV procedure. Next, the residuals $\hat{u}_t$ are combined with the instruments in the direct product $\hat{\boldsymbol{V}} \equiv \hat{\boldsymbol{u}} * \boldsymbol{W}$. Then $\hat{\boldsymbol{\Gamma}}(j)$ is n^{-1} times the $l \times l$ matrix of inner products of the columns of $\hat{\boldsymbol{V}}$ with these same columns lagged j times, the initial unobserved elements being replaced by zeros. As we saw above, $\hat{\boldsymbol{\Gamma}}(0)$ is just $n^{-1} \boldsymbol{W}^{\top} \hat{\boldsymbol{\Omega}} \boldsymbol{W}$, where $\hat{\boldsymbol{\Omega}} = \text{diag}(\hat{u}_t^2)$. Finally, $\hat{\boldsymbol{\Phi}}$ is formed by use of (17.64).

As before, the $\hat{\boldsymbol{\Phi}}$ thus obtained can be used for two purposes. One is to form what is called a **heteroskedasticity and autocorrelation consistent**, or **HAC**, covariance matrix estimator for the ordinary IV estimator. Since the IV estimator is based on the empirical moments $\boldsymbol{W}^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ and the weighting matrix $(\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}$, as can be seen from (17.09), the HAC covariance matrix estimator is found by applying the formula (17.31) to the present case and using (17.33) and (17.34). We obtain

$$\big(\boldsymbol{X}^{\top}\boldsymbol{P}_W\boldsymbol{X}\big)^{-1}\boldsymbol{X}^{\top}\boldsymbol{W}\big(\boldsymbol{W}^{\top}\boldsymbol{W}\big)^{-1}n\hat{\boldsymbol{\Phi}}\big(\boldsymbol{W}^{\top}\boldsymbol{W}\big)^{-1}\boldsymbol{W}^{\top}\boldsymbol{X}\big(\boldsymbol{X}^{\top}\boldsymbol{P}_W\boldsymbol{X}\big)^{-1}. \quad (17.65)$$

In the simple case in which $\boldsymbol{W} = \boldsymbol{X}$, this rather complicated formula becomes

$$\big(\boldsymbol{X}^{\top}\boldsymbol{X}\big)^{-1}n\hat{\boldsymbol{\Phi}}\big(\boldsymbol{X}^{\top}\boldsymbol{X}\big)^{-1}.$$

When there is no serial correlation, implying that $n\hat{\boldsymbol{\Phi}} = \boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W}$, this simplifies to the familiar HCCME (16.15), specialized to the case of a linear regression model. It is a good exercise to see what (17.65) reduces to when there is no serial correlation and $\boldsymbol{W} \neq \boldsymbol{X}$.

More interesting than the HAC covariance matrix estimator is the estimator analogous to the H2SLS estimator, (17.44). For this, instead of using $(\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}$ as weighting matrix, we use the inverse of $\hat{\boldsymbol{\Phi}}$, calculated in the manner described above by use of the ordinary IV estimator as the preliminary consistent estimator. The criterion function becomes

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\top}\boldsymbol{W}\hat{\boldsymbol{\Phi}}^{-1}\boldsymbol{W}^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

and the estimator, which is sometimes called **two-step two-stage least squares**, is therefore

$$\hat{\boldsymbol{\beta}} = \big(\boldsymbol{X}^{\top}\boldsymbol{W}\hat{\boldsymbol{\Phi}}^{-1}\boldsymbol{W}^{\top}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^{\top}\boldsymbol{W}\hat{\boldsymbol{\Phi}}^{-1}\boldsymbol{W}^{\top}\boldsymbol{y}. \quad (17.66)$$

This is very similar to (17.44), in which the matrix $\hat{\boldsymbol{\Phi}}$ is replaced by $\boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W}$. Indeed, in the absence of autocorrelation, $n^{-1}\boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W}$ is the appropriate estimator of $\boldsymbol{\Phi}$. It is easier to obtain an estimate of the asymptotic covariance matrix of (17.66) than of the ordinary IV estimator. It is simply

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = \big(\boldsymbol{X}^{\top}\boldsymbol{W}\hat{\boldsymbol{\Phi}}^{-1}\boldsymbol{W}^{\top}\boldsymbol{X}\big)^{-1}.$$

So far, there is very little practical experience of the estimator (17.66). One reason for this is that econometricians often prefer to model dynamics explicitly (see Chapter 19) rather than leaving all the dynamics in the error term

and employing a specification-consistent estimator. Even if the latter provides consistent estimates of some parameters, it may say nothing about the most interesting ones and may allow serious specification errors to go undetected. Another reason is that there is little evidence concerning the properties of (17.66) in finite samples. The results of Cragg (1983) and Tauchen (1986) for related estimators suggest that these may sometimes be poor.

One important practical problem is how to choose the lag truncation parameter p. Theory is signally unhelpful here. As we mentioned earlier, there are results establishing rates at which p may tend to infinity as the sample size tends to infinity. But if an investigator has a sample of precisely 136 observations, what value of p should be chosen? Andrews (1991b) confronts this problem directly and provides data-dependent methods for choosing p, based on the estimation of an optimal value of a parameter he defines. It is fair to say that none of his methods is elementary, and we cannot discuss them here. Perhaps the most encouraging outcome of his investigations is that, in the neighborhood of the optimal value of p, variations in p have little influence on the performance of the HAC estimator.

Andrews (1991b) also provides valuable evidence about HAC covariance matrix estimators, (17.64) and others, from Monte Carlo experiments. Perhaps the most important finding is that *none* of the HAC estimators he considers is at all reliable for sample sizes up to 250 or so if the errors follow an AR(1) process with autocorrelation parameter greater than 0.9. This disappointing result is related to the fact that AR(1) processes with parameters near unity are close to having what is called a **unit root**. This phenomenon is studied in Chapter 20, and we will see that unit roots throw most conventional econometric theory into confusion.

If we stay away from unit roots and near-unit roots, things are more orderly. We saw in Chapter 16 that it is possible to use HCCMEs even in the presence of homoskedasticity with little loss of accuracy, provided that one of the better HCCMEs is used. It appears that much the same is true for HAC estimators. With an ordinary regression model with serially uncorrelated, homoskedastic errors, the loss of precision due to the use of the Newey-West estimator, say, as opposed to the usual OLS estimator, $\hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, is small. With some of the other HAC estimators considered by Andrews, the loss is smaller still, which implies that the Newey-West estimator is generally not the best available. Similarly, if the errors are heteroskedastic but still serially uncorrelated, then an HCCME is much better than the OLS estimator but only very slightly better than the HAC estimator.

If the errors are autocorrelated at order one and homoskedastic, both the OLS estimator and the HCCME are dominated not only by the HAC estimator, as one would expect, but also by the straightforward estimator computed by estimating the autocorrelation parameter ρ and using the covariance matrix estimator of a feasible GLS procedure. This last estimator is in these circumstances preferable to the HAC ones. In fact, it is only when the errors are

both heteroskedastic and serially correlated that the HAC estimators really come into their own. Even in these circumstances, it is possible, with some patterns of heteroskedasticity, that the feasible GLS estimator, which takes no account of possible heteroskedasticity, can outperform the HAC estimators. But that is probably the exception rather than the rule, for Andrews finds other patterns of heteroskedasticity, which, in combination with serial correlation, require the use of HAC estimators for reasonably accurate inference.

Clearly, the last word on HAC estimators has by no means been said. For instance, in the usual implementation of the Newey-West estimator for linear IV models, we have that $\hat{\boldsymbol{\Gamma}}(0)$ is just $n^{-1}\boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W}$, with $\hat{\boldsymbol{\Omega}}$ the rather poor estimator associated with the HC_0 form of the HCCME. It would seem reasonable to suppose that it would be better to use other forms of $\boldsymbol{\Omega}$ in the Newey-West estimator, just as it is in HCCMEs, and to find similar ways of improving the estimators $\hat{\boldsymbol{\Gamma}}(j)$ for $j \neq 0$. At the time of writing, however, no evidence is available on whether these conjectures are justified. A quite different approach, which we do not have space to discuss, was recently suggested by Andrews and Monahan (1992).

In the next section, we will leave behind the "grubby details" of covariance matrix estimation, assume that a suitable covariance matrix estimator is available, and turn our attention to asymptotic tests of overidentifying restrictions and other aspects of specification testing in GMM models.

17.6 INFERENCE WITH GMM MODELS

In this section, we undertake an investigation of how hypotheses may be tested in the context of GMM models. We begin by looking at tests of overidentifying restrictions and then move on to develop procedures akin to the classical tests studied in Chapter 13 for models estimated by maximum likelihood. The similarities to procedures we have already studied are striking. There is one important difference, however: We will not be able to make any great use of artificial linear regressions in order to implement the tests we discuss. The reason is simply that such artificial regressions have not yet been adequately developed. They exist only for some special cases, and their finite-sample properties are almost entirely unknown. However, there is every reason to hope and expect that in a few years it will be possible to perform inference on GMM models by means of artificial regressions still to be invented.

In the meantime, there are several testing procedures for GMM models that are not difficult to perform. The most important of these is a test of the overidentifying restrictions that are usually imposed. Suppose that we have estimated a vector $\boldsymbol{\theta}$ of k parameters by minimizing the criterion function

$$\boldsymbol{\iota}^{\top}\boldsymbol{F}(\boldsymbol{\theta})\hat{\boldsymbol{\Phi}}^{-1}\boldsymbol{F}^{\top}(\boldsymbol{\theta})\boldsymbol{\iota}, \tag{17.67}$$

in which the empirical moment matrix $\boldsymbol{F}(\boldsymbol{\theta})$ has $l > k$ columns. Observe that

where $\boldsymbol{\Psi}^2 = \boldsymbol{\Phi}^{-1}$, and $\boldsymbol{M}_{\Psi D}$ is the $l \times l$ orthogonal projection matrix onto the orthogonal complement of the k columns of $\boldsymbol{\Psi D}$. By construction, the l–vector $n^{-1/2}\boldsymbol{\Psi F}_0^\top \boldsymbol{\iota}$ has the $N(\mathbf{0}, \mathbf{I})$ distribution asymptotically. It follows, then, that (17.68) is asymptotically distributed as chi-squared with number of degrees of freedom equal to the rank of $\boldsymbol{M}_{\Psi D}$, that is, $l - k$, the number of overidentifying restrictions.

Hansen's test of overidentifying restrictions is completely analogous, in the present more general context, to the one for IV estimation discussed in Section 7.8, based on the criterion function (7.56). It is a good exercise to work through the derivation given above for the simple case of a linear regression model with homoskedastic, serially uncorrelated errors, in order to see how closely the general case mimics the simple one.[2]

Hansen's test of overidentifying restrictions is perhaps as close as one can come in econometrics to a portmanteau specification test. Because models estimated by GMM are subject to so few restrictions, their "specification" is not very demanding. In particular, if nothing more is required than the existence of the moments used to identify the parameters, then only two things are left to test. One is the set of any overidentifying restrictions used, and the other is parameter constancy.[3] Because Hansen's test of overidentifying restrictions has as many degrees of freedom as there are overidentifying restrictions, it may be possible to achieve more power by reducing the number of degrees of freedom. However, if Hansen's test statistic is small enough numerically, no such test can reject, for the simple reason that Hansen's statistic provides an upper bound for all possible test statistics for which the null hypothesis is the estimated model. This last fact follows from the observation that no criterion function of the form (17.67) can be less than zero.

Tests for which the null hypothesis is not the estimated model are not subject to the bound provided by Hansen's statistic. This is just as well, of course, since otherwise it would be impossible to reject a just identified model at all. A test for parameter constancy is not subject to the bound either, although at first glance the null hypothesis would appear to be precisely the estimated model. The reason was discussed in Section 11.2 in connection with tests for parameter constancy in nonlinear regression models estimated by means of instrumental variables. Essentially, in order to avoid problems of identification, it is necessary to double the number of instruments used, by splitting the original ones up as in (11.09). Exactly the same considerations apply for GMM models, of course, especially those that are just identified or have few overidentifying restrictions. But if one uses twice as many instruments, the null model has effectively been changed, and for that reason,

---

[2] Hansen's test statistic, (17.68), is sometimes referred to as the J statistic. For obvious reasons (see Chapter 11) we prefer not to give it that name.

[3] Tests of parameter constancy in models estimated by GMM are discussed by Hoffman and Pagan (1989) and Ghysels and Hall (1990).

and that, as we have seen, is both Hansen's statistic and the LM statistic in these circumstances.

Finally, we consider $C(\alpha)$ tests. Let $\acute{\boldsymbol{\theta}}$ be a parameter vector satisfying the restrictions $\boldsymbol{r}(\acute{\boldsymbol{\theta}}) = \boldsymbol{0}$. Then the test statistic can be formed as though it were the difference of two LM statistics, one for the restricted and one for the unrestricted model, both evaluated at $\acute{\boldsymbol{\theta}}$. Suppose, for simplicity, that the parameter vector $\boldsymbol{\theta}$ can be partitioned as $[\boldsymbol{\theta}_1 \vdots \boldsymbol{\theta}_2]$ and that the restrictions can be written as $\boldsymbol{\theta}_2 = 0$. The first term of the $C(\alpha)$ statistic has the form (17.72) but is evaluated at $\acute{\boldsymbol{\theta}}$ rather than the genuine constrained estimator $\tilde{\boldsymbol{\theta}}$. The second term should take the form of an LM statistic appropriate to the constrained model, for which only $\boldsymbol{\theta}_1$ may vary. This corresponds to replacing the matrix $\tilde{\boldsymbol{D}}$ in (17.72) by $\acute{\boldsymbol{D}}_1$, where the partition of $\boldsymbol{D}$ as $[\boldsymbol{D}_1 \; \boldsymbol{D}_2]$ corresponds to the partition of $\boldsymbol{\theta}$. The $C(\alpha)$ test statistic is therefore

$$
\begin{aligned}
C(\alpha) = &\frac{1}{n} \boldsymbol{\iota}^\top \acute{\boldsymbol{F}} \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}} \big(\acute{\boldsymbol{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}} \big)^{-1} \acute{\boldsymbol{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{F}}^\top \boldsymbol{\iota} \\
&- \frac{1}{n} \boldsymbol{\iota}^\top \acute{\boldsymbol{F}} \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}}_1 \big(\acute{\boldsymbol{D}}_1^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}}_1 \big)^{-1} \acute{\boldsymbol{D}}_1^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{F}}^\top \boldsymbol{\iota}.
\end{aligned}
\tag{17.75}
$$

Here, as before, $\hat{\boldsymbol{\Phi}}$ is a suitable estimate of $\boldsymbol{\Phi}$. To show that (17.75) is asymptotically equivalent to the true LM statistic, it is enough to modify the details of the proof of the corresponding asymptotic equivalence in Section 13.7.

In the general case in which the restrictions are expressed as $\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{0}$, another form of the $C(\alpha)$ test may be more convenient, since forming a matrix to correspond to $\boldsymbol{D}_1$ may not be simple. This other form is

$$
\boldsymbol{\iota}^\top \acute{\boldsymbol{F}} \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}} \big(\acute{\boldsymbol{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}} \big)^{-1} \acute{\boldsymbol{R}}^\top \Big(\acute{\boldsymbol{R}} \big(\acute{\boldsymbol{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}} \big)^{-1} \acute{\boldsymbol{R}}^\top \Big)^{-1} \acute{\boldsymbol{R}} \big(\acute{\boldsymbol{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{D}} \big)^{-1} \acute{\boldsymbol{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \acute{\boldsymbol{F}}^\top \boldsymbol{\iota}.
$$

For this statistic to be useful, the difficulty of computing the actual constrained estimate $\tilde{\boldsymbol{\theta}}$ must outweigh the complication of the above formula. The formula itself can be established, at the cost of some tedious algebra, by adapting the methods of Section 8.9. We leave the details to the interested reader.

The treatment we have given of LM, LR, and Wald tests has largely followed that of Newey and West (1987b). This article may be consulted for more details of regularity conditions sufficient for the results merely asserted here to hold. Another paper on testing models estimated by GMM is Newey (1985b). Nonnested hypothesis tests for models estimated by GMM are discussed by Smith (1992). These papers do not deal with $C(\alpha)$ tests, however.

An interesting question is whether the conditional moment tests discussed in the last chapter in the context of models estimated by maximum likelihood have any counterpart for models estimated by GMM. For simplicity, suppose that there is a single conditional moment of which the expectation is zero if the model is correctly specified. If the corresponding empirical moment is used as an overidentifying restriction, then it can be tested in the same way

and k exogenous or predetermined variables. Then the model can be written in matrix form as

$$\boldsymbol{Y\Gamma} = \boldsymbol{XB} + \boldsymbol{U}. \tag{18.01}$$

Here $\boldsymbol{Y}$ denotes an $n \times g$ matrix of endogenous variables, $\boldsymbol{X}$ denotes an $n \times k$ matrix of exogenous or predetermined variables, $\boldsymbol{\Gamma}$ denotes a $g \times g$ matrix of coefficients, $\boldsymbol{B}$ denotes a $k \times g$ matrix of coefficients, and $\boldsymbol{U}$ denotes an $n \times g$ matrix of error terms.

It is at once clear that the model (18.01) contains too many coefficients to estimate. A typical observation for the lth equation can be written as

$$\sum_{i=1}^{g} \Gamma_{il} Y_{ti} = \sum_{j=1}^{k} B_{jl} X_{tj} + u_{tl}.$$

Multiplying all of the Γ_{il}'s and B_{jl}'s by any nonzero constant would simply have the effect of multiplying u_{tl} by that same constant for all t, but would not change the pattern of the error terms across observations at all. Thus it is necessary to impose some sort of normalization on each of the equations of the model. The obvious one is to set $\Gamma_{ii} = 1$ for all i; each endogenous variable, y_1 through y_g, would then have a coefficient of unity in one and only one equation. However, as we saw in Section 7.3, many other normalizations could be used. We could, for example, set $\Gamma_{1l} = 1$ for all l; the coefficient on the first endogenous variable would then be unity in every equation.

The model (18.01) makes no sense if the matrix $\boldsymbol{\Gamma}$ cannot be inverted, since otherwise it would be impossible to determine $\boldsymbol{Y}$ uniquely as a function of $\boldsymbol{X}$ and $\boldsymbol{U}$. We may therefore postmultiply both sides of (18.01) by $\boldsymbol{\Gamma}^{-1}$ to obtain

$$\boldsymbol{Y} = \boldsymbol{XB\Gamma}^{-1} + \boldsymbol{U\Gamma}^{-1} \tag{18.02}$$

$$= \boldsymbol{X\Pi} + \boldsymbol{V}. \tag{18.03}$$

Expression (18.02) is the **restricted reduced form**, or **RRF**, and expression (18.03) is the **unrestricted reduced form**, or **URF**. The restrictions are that $\boldsymbol{\Pi} = \boldsymbol{B\Gamma}^{-1}$. Notice that, even in the unlikely event that the columns of $\boldsymbol{U}$ were independent, the columns of $\boldsymbol{V}$ would not be. Thus the various equations of the reduced form are almost certain to have correlated errors.

The imposition of normalization restrictions is necessary but not sufficient to obtain estimates of $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$. The problem is that, unless we impose some restrictions on it, the model (18.01) has too many coefficients to estimate. The matrix $\boldsymbol{\Gamma}$ contains $g^2 - g$ coefficients, because of the g normalization restrictions, while the matrix $\boldsymbol{B}$ contains gk. There are thus $g^2 + gk - g$ structural coefficients in total. But the matrix $\boldsymbol{\Pi}$ in the unrestricted reduced form contains only gk coefficients. It is obviously impossible to determine the $g^2 + gk - g$ structural coefficients uniquely from the gk coefficients of the

they restrict themselves to fully specified parametric models capable of being estimated by maximum likelihood. We will, however, make use of one of their specific examples as a concrete illustration of a number of points.

Let the $1 \times g$ vector $\boldsymbol{Y}_t$ denote the tth observation on a set of variables that we wish to model as a simultaneous process, and let the $1 \times k$ vector $\boldsymbol{X}_t$ be the tth observation on a set of explanatory variables, some or all of which may be lagged $\boldsymbol{Y}_t$'s. We may write an, in general nonlinear, simultaneous equations model as

$$\boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta}) = \boldsymbol{U}_t, \qquad (18.04)$$

where $\boldsymbol{h}_t$ is a $1 \times g$ vector of functions, somewhat analogous to the regression function of a univariate model, $\boldsymbol{\theta}$ is a p–vector of parameters, and $\boldsymbol{U}_t$ is a $1 \times g$ vector of error terms. The linear model (18.01) is seen to be a special case of (18.04) if we rewrite it as

$$\boldsymbol{Y}_t \boldsymbol{\Gamma} = \boldsymbol{X}_t \boldsymbol{B} + \boldsymbol{U}_t$$

and define $\boldsymbol{\theta}$ so that it consists of all the elements of $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ which have to be estimated. Here $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ are the tth rows of the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. A set of (conditional) moment conditions could be based on (18.04), by writing

$$E\big(\boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta})\big) = \boldsymbol{0},$$

where the expectation could be interpreted as being conditional on some appropriate information set.

Definition 18.1.

The explanatory variables $\boldsymbol{X}_t$ are **predetermined** in equation i of the model (18.04), for $i = 1, \ldots, g$, if, for all $t = 1, \ldots, n$,

$$\boldsymbol{X}_t \perp\!\!\!\perp u_{i,t+s} \quad \text{for all } s \geq 0.$$

Here the symbol $\perp\!\!\!\perp$ is used to express statistical independence. The definition applies to any context, such as the time-series one, in which there is a natural ordering of the observations. The next concept does not require this.

Definition 18.2.

The explanatory variables $\boldsymbol{X}_t$ are **strictly exogenous** in equation i of (18.04) if, for all $t = 1, \ldots, n$,

$$\boldsymbol{X}_t \perp\!\!\!\perp \boldsymbol{U}_s \quad \text{for all } s = 1, \ldots, n.$$

If (18.04) represents a structural form, then either predeterminedness or strict exogeneity allows us to treat this form as a characterization of the process generating $\boldsymbol{Y}_t$ conditional on $\boldsymbol{X}_t$. Thus we may, for example, write down a loglikelihood function based on (18.04), which can be maximized in

order to provide consistent estimates of the parameters $\boldsymbol{\theta}$; see Section 18.4. If (18.04) is thought of as providing conditional moment conditions, then either predeterminedness or strict exogeneity allows us to use the columns of the matrix $\boldsymbol{X}$ as instruments for the estimation of $\boldsymbol{\theta}$ by some sort of IV procedure, such as 2SLS, 3SLS, or GMM. In claiming this, we assume of course that there are enough instruments in $\boldsymbol{X}$ to *identify* all of the parameters in $\boldsymbol{\theta}$.

Unfortunately, the concept of strict exogeneity is much too restrictive, at least for time-series applications. In this context, very few variables are strictly exogenous, although many are predetermined. However, as we now show, a variable can be predetermined or not in one and the same model depending on how the model is parametrized. Furthermore, predeterminedness is not always necessary for consistent estimation. Thus predeterminedness is not a very satisfactory concept.

Consider the following simultaneous model, taken from Engle, Hendry, and Richard (1983):

$$y_t = \beta x_t + \varepsilon_{1t} \tag{18.05}$$

$$x_t = \delta_1 x_{t-1} + \delta_2 y_{t-1} + \varepsilon_{2t}, \tag{18.06}$$

where the error terms are normally, independently, and identically distributed for each t, with covariance matrix

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

If $\sigma_{12} \neq 0$, x_t is correlated with ε_{1t} and estimation of (18.05) by OLS will not be consistent because x_t is not predetermined in (18.05).

Now let us consider the expectation of y_t conditional on x_t and all lagged y_t's and x_t's. We have

$$E(y_t \mid x_t, y_{t-1}, x_{t-1} \cdots) = \beta x_t + E(\varepsilon_{1t} \mid x_t, y_{t-1}, x_{t-1} \cdots). \tag{18.07}$$

Notice that ε_{2t} is defined by (18.06) as a linear combination of the conditioning variables. Thus the conditional expectation of ε_{1t} in (18.07) is

$$E(\varepsilon_{1t} \mid \varepsilon_{2t}) = \frac{\sigma_{12}}{\sigma_{22}} \varepsilon_{2t} = \frac{\sigma_{12}}{\sigma_{22}} (x_t - \delta_1 x_{t-1} - \delta_2 y_{t-1}).$$

We may therefore write

$$y_t = b x_t + c_1 x_{t-1} + c_2 y_{t-1} + v_t, \tag{18.08}$$

with

$$b = \beta + \frac{\sigma_{12}}{\sigma_{22}}, \quad c_1 = -\delta_1 \frac{\sigma_{12}}{\sigma_{22}}, \quad c_2 = -\delta_2 \frac{\sigma_{12}}{\sigma_{22}}, \tag{18.09}$$

and with v_t independent of x_t. Thus x_t is predetermined in (18.08), whatever the value of σ_{12}, even though it is not predetermined in (18.05) when $\sigma_{12} \neq 0$.

on asymptotic theory, and one cannot hope to obtain consistent parameter estimates if the parameters are not asymptotically identified.

In this section, we will discuss the asymptotic identifiability of a linear simultaneous equations model by the two-stage least squares estimator introduced in Section 7.5. This may seem a very limited topic, and in a certain sense it is indeed limited. However, it is a topic that has given rise to a truly vast literature, to which we can in no way do justice here; see Fisher (1976) and Hsiao (1983). There exist models that are not identified by the 2SLS estimator but are identified by other estimators, such as the FIML estimator, and we will briefly touch on such cases later. It is not a simple task to extend the theory we will present in this section to the context of nonlinear models, for which it is usually better to return to the general theory expounded in Section 5.2.

We begin with the linear simultaneous equations model, (18.01). This model consists of DGPs that generate samples for which each observation is a g–vector $\boldsymbol{Y}_t$ of dependent variables, conditional on a set of exogenous and lagged dependent variables $\boldsymbol{X}_t$. Since the exogenous variables in $\boldsymbol{X}_t$ are assumed to be weakly exogenous, their generating mechanism can be ignored. In order to discuss identification, little needs to be assumed about the error terms $\boldsymbol{U}_t$. They must evidently satisfy the condition that $E(\boldsymbol{U}_t) = 0$, and it seems reasonable to assume that they are serially independent and that $E(\boldsymbol{U}_t^\top \boldsymbol{U}_t) = \boldsymbol{\Sigma}_t$, where $\boldsymbol{\Sigma}_t$ is a positive definite matrix for all t. If inferences are to be based on the usual 2SLS covariance matrix, it will be necessary to make the further assumption that the error terms are homoskedastic, that is, $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}$ for all t.

It is convenient to treat the identification of the parameters of a simultaneous equations model equation by equation, since it is entirely possible that the parameters of some equations may be identified while the parameters of others are not. In order to simplify notation, we will consider, without loss of generality, only the parameters of the first equation of the system, that is, the elements of the first columns of the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$. As we remarked in Section 18.1, restrictions must be imposed on the elements of these matrices for identification to be possible. It is usual to assume that these restrictions all take the form of zero restrictions on some elements. A variable is said to be **excluded** from an equation if the coefficient corresponding to that variable for that equation is restricted to be zero; otherwise, it is said to be **included** in the equation. As discussed in Section 6.4, it is always possible in the context of a single equation to perform a reparametrization such that all restrictions take the form of zero restrictions. But in the context of a simultaneous equations model, such reparametrizations exist in general only if there are no **cross-equation restrictions**, that is, restrictions which involve the parameters of more than one equation of the system. If there are cross-equation restrictions, then to all intents and purposes we leave the context of linear systems. We would in any case have to abandon the 2SLS estimator if we wished to impose cross-equation restrictions.

and (18.20), be expressed as

$$\boldsymbol{\pi}_1 - \boldsymbol{\Pi}_{11}\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1$$

$$\boldsymbol{\pi}_2 - \boldsymbol{\Pi}_{21}\boldsymbol{\gamma}_1 = \mathbf{0}.$$

The first of these two equations serves to define $\boldsymbol{\beta}_1$ in terms of $\boldsymbol{\Pi}$ and $\boldsymbol{\gamma}_1$, and allows us to see that $\boldsymbol{\beta}_1$ can be identified if $\boldsymbol{\gamma}_1$ can be. The second equation shows that $\boldsymbol{\gamma}_1$ is determined uniquely if and only if the submatrix $\boldsymbol{\Pi}_{21}$ has full column rank, that is, if the rank of the matrix is equal to the number of columns (see Appendix A). The submatrix $\boldsymbol{\Pi}_{21}$ has $k - k_1$ rows and g_1 columns. Therefore, if the order condition is satisfied, there are at least as many rows as columns. The condition for the identifiability of $\boldsymbol{\gamma}_1$, and so also of $\boldsymbol{\beta}_1$, is thus simply that the columns of $\boldsymbol{\Pi}_{21}$ in the DGP should be linearly independent.

It is instructive to show why this last condition is equivalent to the rank condition in terms of $\operatorname{plim}(n^{-1}\boldsymbol{Z}^\top\boldsymbol{P}_X\boldsymbol{Z})$. If, as we have tacitly assumed throughout this discussion, the exogenous variables $\boldsymbol{X}$ satisfy the condition that $\operatorname{plim}(n^{-1}\boldsymbol{X}^\top\boldsymbol{X})$ is positive definite, then $\operatorname{plim}(n^{-1}\boldsymbol{Z}^\top\boldsymbol{P}_X\boldsymbol{Z})$ can fail to have full rank only if $\operatorname{plim}(n^{-1}\boldsymbol{X}^\top\boldsymbol{Z})$ has rank less than $g_1 + k_1$, the number of columns of $\boldsymbol{Z}$. The probability limit of the matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{Z}$ follows from (18.22), with $\boldsymbol{X}$ replacing $\boldsymbol{W}$. If, for notational simplicity, we drop the probability limit and the factor of n^{-1}, which are not essential to the discussion, the matrix of interest can be written as

$$\begin{bmatrix} \boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_1\boldsymbol{\Pi}_{11} + \boldsymbol{X}_1^\top\boldsymbol{X}_2\boldsymbol{\Pi}_{21} \\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_1\boldsymbol{\Pi}_{11} + \boldsymbol{X}_2^\top\boldsymbol{X}_2\boldsymbol{\Pi}_{21} \end{bmatrix}. \tag{18.23}$$

This matrix does not have full column rank of $g_1 + k_1$ if and only if there exists a nonzero $(g_1 + k_1)$–vector $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1 \vdots \boldsymbol{\theta}_2]$ such that postmultiplying (18.23) by $\boldsymbol{\theta}$ gives zero. If we write this condition out and rearrange slightly, we obtain

$$\begin{bmatrix} \boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_1 + \boldsymbol{\Pi}_{11}\boldsymbol{\theta}_2 \\ \boldsymbol{\Pi}_{21}\boldsymbol{\theta}_2 \end{bmatrix} = \mathbf{0}. \tag{18.24}$$

The first matrix on the left-hand side here is just $\boldsymbol{X}^\top\boldsymbol{X}$ and is therefore nonsingular. The condition reduces to the two vector equations

$$\boldsymbol{\theta}_1 + \boldsymbol{\Pi}_{11}\boldsymbol{\theta}_2 = \mathbf{0} \tag{18.25}$$

$$\boldsymbol{\Pi}_{21}\boldsymbol{\theta}_2 = \mathbf{0}. \tag{18.26}$$

If these equations hold for some nonzero $\boldsymbol{\theta}$, it is clear that $\boldsymbol{\theta}_2$ cannot be zero. Consequently, the second of these equations can hold only if $\boldsymbol{\Pi}_{21}$ has less than full column rank. It follows that if the rank condition in terms of $\boldsymbol{Z}^\top\boldsymbol{P}_X\boldsymbol{Z}$ does not hold, then it does not hold in terms of $\boldsymbol{\Pi}_{21}$ either. Conversely, suppose that (18.26) holds for some nonzero g_1–vector $\boldsymbol{\theta}_2$. Then $\boldsymbol{\Pi}_{21}$ does not have full column rank. Define $\boldsymbol{\theta}_1$ in terms of this $\boldsymbol{\theta}_2$ and $\boldsymbol{\Pi}$ by means

By the same token, if the parameters of the structural model are not constant over the entire sample, then the parameters of the URF will not be constant either. Since the equations of the URF are estimated by ordinary least squares, it is very easy to test them for evidence of misspecification such as serial correlation, heteroskedasticity, and nonconstant coefficients. If they fail any of these tests, then one may reasonably conclude that the structural model is misspecified, even if one has not actually estimated it. The converse is not true, however, since these tests may well lack power, especially if only one of the structural equations is misspecified.

One additional misspecification test that should always be performed is a test of any **overidentifying restrictions**. In Section 7.8, we discussed how to test overidentifying restrictions for a single equation estimated by IV or 2SLS. Here we are interested in all of the overidentifying restrictions for the entire system. The number of degrees of freedom for the test is equal to the number of elements in the $\boldsymbol{\Pi}$ matrix of the URF, gk, minus the number of free parameters in $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ jointly. In most cases there will be some overidentifying restrictions, and in many cases there will be a large number of them. The most natural way to test these is probably to use an LR test. The restricted value of the loglikelihood function is the value of (18.30) at the FIML estimates $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{\Gamma}}$, and the unrestricted value is

$$-\frac{ng}{2}\big(\log(2\pi) + 1\big) - \frac{n}{2}\log\left|\frac{1}{n}\big(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Pi}}\big)^{\top}\big(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Pi}}\big)\right|, \qquad (18.33)$$

where $\hat{\boldsymbol{\Pi}}$ denotes the OLS estimates of the parameters of the URF. As usual, twice the difference between the unrestricted and restricted values of the loglikelihood function will be asymptotically distributed as χ^2 with as many degrees of freedom as there are overidentifying restrictions. If one suspects that the overidentifying restrictions are violated and therefore does not want to bother estimating the structural model, one could instead use a Wald test, as suggested by Byron (1974).

We have not yet explained why the OLS estimates $\hat{\boldsymbol{\Pi}}$ are also the ML estimates. It can easily be seen from (18.33) that, in order to obtain ML estimates of $\boldsymbol{\Pi}$, we need to minimize the determinant

$$\left|(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Pi})^{\top}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Pi})\right|. \qquad (18.34)$$

Suppose that we evaluate this determinant at any set of estimates $\acute{\boldsymbol{\Pi}}$ not equal to $\hat{\boldsymbol{\Pi}}$. Since we can always write $\acute{\boldsymbol{\Pi}} = \hat{\boldsymbol{\Pi}} + \boldsymbol{A}$ for some matrix $\boldsymbol{A}$, (18.34) becomes

$$\begin{aligned}
&\left|(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Pi}} - \boldsymbol{X}\boldsymbol{A})^{\top}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Pi}} - \boldsymbol{X}\boldsymbol{A})\right| \\
={}& \left|(\boldsymbol{M}_X\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A})^{\top}(\boldsymbol{M}_X\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A})\right| \\
={}& \left|\boldsymbol{Y}^{\top}\boldsymbol{M}_X\boldsymbol{Y} + \boldsymbol{A}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{A}\right|.
\end{aligned} \qquad (18.35)$$

Because the determinant of the sum of two positive definite matrices is always greater than the determinants of either of those matrices (see Appendix A), it follows from (18.35) that (18.34) will exceed $|\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|$ for all $\mathbf{A} \neq \mathbf{0}$. This implies that $\hat{\mathbf{\Pi}}$ minimizes (18.34), and so we have proved that equation-by-equation OLS estimates of the URF are also ML estimates for the entire system.

If one does not have access to a regression package that calculates (18.33) easily, there is another way to do so. Consider the **recursive system**

$$
\begin{aligned}
\mathbf{y}_1 &= \mathbf{X}\boldsymbol{\eta}_1 + \mathbf{e}_1 \\
\mathbf{y}_2 &= \mathbf{X}\boldsymbol{\eta}_2 + \mathbf{y}_1 \alpha_1 + \mathbf{e}_2 \\
\mathbf{y}_3 &= \mathbf{X}\boldsymbol{\eta}_3 + [\mathbf{y}_1 \quad \mathbf{y}_2]\boldsymbol{\alpha}_2 + \mathbf{e}_3 \\
\mathbf{y}_4 &= \mathbf{X}\boldsymbol{\eta}_4 + [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \mathbf{y}_3]\boldsymbol{\alpha}_3 + \mathbf{e}_4,
\end{aligned}
\tag{18.36}
$$

and so on, where $\mathbf{y}_i$ denotes the ith column of $\mathbf{Y}$. This system of equations can be interpreted as simply a reparametrization of the URF (18.03). It is easy to see that if one estimates these equations by OLS, all the residual vectors will be mutually orthogonal: $\hat{\mathbf{e}}_2$ will be orthogonal to $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_3$ will be orthogonal to $\hat{\mathbf{e}}_2$ and $\hat{\mathbf{e}}_1$, and so on. According to the URF, all the $\mathbf{y}_i$'s are linear combinations of the columns of $\mathbf{X}$ plus random errors. Therefore, the equations of (18.36) are correct for any arbitrary choice of the α parameters: The $\boldsymbol{\eta}_i$'s simply adjust to whatever choice is made. If, however, we *require* that the error terms $\mathbf{e}_i$ should be orthogonal, then this serves to identify a particular unique choice of the α's. In fact, the recursive system (18.36) has exactly the same number of parameters as the URF (18.03): g vectors $\boldsymbol{\eta}_i$, each with k elements, $g-1$ vectors $\boldsymbol{\alpha}_i$, with a total of $g(g-1)/2$, and g variance parameters, for a total of $gk + (g^2 + g)/2$. The URF has gk parameters in $\mathbf{\Pi}$ and $(g^2 + g)/2$ in the covariance matrix $\mathbf{\Omega}$, for the same total. What has happened is that the α parameters in (18.36) have replaced the off-diagonal elements of the covariance matrix of $\mathbf{V}$ in the URF.

Since the recursive system (18.36) is simply a reparametrization of the URF (18.03), it should come as no surprise that the loglikelihood function for the former is equal to (18.33). Because the residuals of the various equations in (18.36) are orthogonal, the value of the loglikelihood function for (18.36) is simply the sum of the values of the loglikelihood functions from OLS estimation of the individual equations. This result, which readers can easily verify numerically, sometimes provides a convenient way to compute the loglikelihood function for the URF. Except for this purpose, recursive systems are not generally of much interest. They do not convey any information that is not already provided by the URF, and the parametrization depends on an arbitrary ordering of the equations.

Solving for $\hat{\boldsymbol{\gamma}}_1$ then yields

$$\hat{\boldsymbol{\gamma}}_1 = \big(\boldsymbol{Y}_1^\top(\boldsymbol{M}_1 - \hat{\kappa}\boldsymbol{M}_X)\boldsymbol{Y}_1\big)^{-1}\boldsymbol{Y}_1^\top(\boldsymbol{M}_1 - \hat{\kappa}\boldsymbol{M}_X)\boldsymbol{y}.$$

Since $\boldsymbol{X}_1 \in \mathcal{S}(\boldsymbol{X})$, $\boldsymbol{M}_1 - \hat{\kappa}\boldsymbol{M}_X = \boldsymbol{M}_1(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_X)$. Using this fact and a little algebra, which we leave as an exercise, it can be shown that $\hat{\boldsymbol{\gamma}}_1$ can also be computed using the formula

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\gamma}}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{Y}_1 \\ \boldsymbol{Y}_1^\top\boldsymbol{X}_1 & \boldsymbol{Y}_1^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_X)\boldsymbol{Y}_1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^\top\boldsymbol{y} \\ \boldsymbol{Y}_1^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_X)\boldsymbol{y} \end{bmatrix}, \qquad (18.53)$$

which yields $\hat{\boldsymbol{\beta}}_1$ as well. Then if we define $\boldsymbol{Z}$ as $[\boldsymbol{X}_1 \quad \boldsymbol{Y}_1]$ and $\boldsymbol{\delta}$ as $[\boldsymbol{\beta}_1 \vdots \boldsymbol{\gamma}_1]$, as in (18.18), (18.53) can be written in the very simple form

$$\hat{\boldsymbol{\delta}} = \big(\boldsymbol{Z}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_X)\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_X)\boldsymbol{y}. \qquad (18.54)$$

Equation (18.53) is one way of writing LIML as a member of what is called the **K-class** of estimators; see Theil (1961) and Nagar (1959). Equation (18.54) is a simpler way of doing the same thing. The K-class consists of all estimators that can be written in either of these two forms, but with an arbitrary scalar K replacing $\hat{\kappa}$. We use K rather than the more traditional k to denote this scalar in order to avoid confusion with the number of exogenous variables in the system. The LIML estimator is thus a K-class estimator with $K = \hat{\kappa}$. Similarly, as is evident from (18.54), the 2SLS estimator is a K-class estimator with $K = 1$, and the OLS estimator is a K-class estimator with $K = 0$. Since $\hat{\kappa} = 1$ for a structural equation that is just identified, it follows immediately from (18.54) that the LIML and 2SLS estimators coincide in this special case.

It can be shown that K-class estimators are consistent whenever K tends to 1 asymptotically at a rate faster than $n^{-1/2}$; see Schmidt (1976), among others. Even though the consistency of LIML follows from general results for ML estimators, it is interesting to see how this result for the K-class applies to it. We have already seen that $n\log(\hat{\kappa})$ is the LR test statistic for the null hypothesis that the overidentifying restrictions on the structural equation being estimated are valid. If we Taylor expand the logarithm, we find that $n\log(\hat{\kappa}) \cong n(\hat{\kappa} - 1)$. Since this test statistic has an asymptotic χ^2 distribution, it must be $O(1)$, and so $\hat{\kappa} - 1$ must be $O(n^{-1})$. This then establishes the consistency of LIML.

There are many other K-class estimators. For example, Sawa (1973) has suggested a way of modifying the 2SLS estimator to reduce bias, and Fuller (1977) and Morimune (1978, 1983) have suggested modified versions of the LIML estimator. Fuller's estimator, which is the simplest of these, uses $K = \hat{\kappa} - \alpha/(n - k)$, where α is a positive constant that must be chosen by the investigator. One good choice is $\alpha = 1$, since it yields estimates that

In order to make (18.61) operational, we need to estimate the covariance matrix $\boldsymbol{\Sigma}$ of the error terms. In the case of an SUR model, we could use OLS on each equation individually. Since OLS is inconsistent for simultaneous equations models, we use 2SLS on each equation instead. Thus the first two "stages" of 3SLS are simply the two stages of 2SLS, applied to each separate equation of (18.01). The covariances of the error terms are then estimated from the 2SLS residuals:

$$\tilde{\sigma}_{ij} = \frac{1}{n} \sum_{t=1}^{n} \tilde{u}_{ti} \tilde{u}_{tj}. \tag{18.62}$$

Of course, these residuals must be the genuine 2SLS residuals, not the residuals from OLS estimation of the second-stage regressions; see Section 7.5. Thus we see that the 3SLS estimators $\tilde{\boldsymbol{\delta}}_1$ through $\tilde{\boldsymbol{\delta}}_g$ must jointly solve the first-order conditions

$$\sum_{j=1}^{g} \tilde{\sigma}^{ij} \boldsymbol{Z}_i^\top \boldsymbol{P}_X \big(\boldsymbol{y}_j - \boldsymbol{Z}_j \tilde{\boldsymbol{\delta}}_j\big) = \boldsymbol{0}. \tag{18.63}$$

The solution is easy to write down. If $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_1 \vdots \cdots \vdots \boldsymbol{\delta}_g]$ and matrices enclosed in square brackets $[\cdot]$ denote partitioned matrices characterized by a typical block, then the 3SLS estimator $\tilde{\boldsymbol{\delta}}$ can be written very compactly as

$$\tilde{\boldsymbol{\delta}} = \big[\tilde{\sigma}^{ij} \boldsymbol{Z}_i^\top \boldsymbol{P}_X \boldsymbol{Z}_j\big]^{-1} \left[\sum_{j=1}^{g} \tilde{\sigma}^{ij} \boldsymbol{Z}_i^\top \boldsymbol{P}_X \boldsymbol{y}_j\right]. \tag{18.64}$$

It is more common to see the 3SLS estimator written using an alternative notation that involves Kronecker products; see almost any econometrics textbook. Although Kronecker products can sometimes be useful (Magnus and Neudecker, 1988), we prefer the compact notation of (18.64).

The 3SLS estimator is closely related both to the 2SLS estimator and to the GLS estimator for multivariate SUR models in which the explanatory variables are all exogenous or predetermined. If we assume that $\boldsymbol{\Sigma}$ is diagonal, conditions (18.63) become simply

$$\tilde{\sigma}^{ii} \boldsymbol{Z}_i^\top \boldsymbol{P}_X \big(\boldsymbol{y}_i - \boldsymbol{Z}_i \boldsymbol{\delta}_i\big) = \boldsymbol{0},$$

which are equivalent to the conditions for equation-by-equation 2SLS. Thus 3SLS and 2SLS will be asymptotically (but not numerically) equivalent when the structural form errors are not contemporaneously correlated. It is also easy to see that the SUR estimator for linear models is just a special case of the 3SLS estimator. Since all regressors can be used as instruments in the SUR case, it is no longer necessary to use 2SLS in the preliminary stage. Equivalently, the fact that each regressor matrix $\boldsymbol{Z}_i$ is just a submatrix of the full regressor matrix, $\boldsymbol{X}$, implies that $\boldsymbol{P}_X \boldsymbol{Z}_i = \boldsymbol{Z}_i$. Thus (18.63) simplifies to

$$\sum_{j=1}^{g} \tilde{\sigma}^{ij} \boldsymbol{Z}_i^\top \big(\boldsymbol{y}_j - \boldsymbol{Z}_j \boldsymbol{\delta}_j\big) = \boldsymbol{0},$$

Consequently, the matrix (18.69), evaluated at the ML estimates, becomes

$$-\hat{\boldsymbol{Y}}^{\top}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}} - \boldsymbol{X}\hat{\boldsymbol{B}})\hat{\boldsymbol{\Sigma}}^{-1}.$$

Now at last we can select the elements of the two partial derivative matrices which are actually zero when evaluated at the ML estimates. The parameters that appear in the ith equation are found in the ith columns of the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$, and so the appropriate partial derivatives are found in the ith columns of the partial derivative matrices. For the matrix corresponding to $\boldsymbol{B}$, this column is $\boldsymbol{X}^{\top}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}} - \boldsymbol{X}\hat{\boldsymbol{B}})(\hat{\boldsymbol{\Sigma}}^{-1})_i$. From this column we wish to select only those rows for which the corresponding element of the column $\boldsymbol{B}_i$ is unrestricted, that is, the elements corresponding to the $n \times k_i$ matrix $\boldsymbol{X}_i$. Since in order to select rows of a matrix product, we need only select the corresponding rows of the left-most factor, the zero elements are those of the k_i–vector $\boldsymbol{X}_i^{\top}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}} - \boldsymbol{X}\hat{\boldsymbol{B}})(\hat{\boldsymbol{\Sigma}}^{-1})_i$.

By exactly similar reasoning, we find that, for each $i = 1, \ldots, g$, the g_i–vector $\hat{\boldsymbol{Y}}_i^{\top}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}} - \boldsymbol{X}\hat{\boldsymbol{B}})(\hat{\boldsymbol{\Sigma}}^{-1})_i$ is zero, where $\hat{\boldsymbol{Y}}_i$ contains only those columns of $\hat{\boldsymbol{Y}}$ that correspond to the matrix $\boldsymbol{Y}_i$ of endogenous variables included as regressors in the ith equation. If we write $\hat{\boldsymbol{Z}}_i \equiv [\boldsymbol{X}_i \quad \hat{\boldsymbol{Y}}_i]$, then all the first-order conditions corresponding to the parameters of the ith equation can be written as

$$\hat{\boldsymbol{Z}}_i^{\top}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}} - \boldsymbol{X}\hat{\boldsymbol{B}})(\boldsymbol{\Sigma}^{-1})_i = \boldsymbol{0}.$$

These conditions can be further simplified. Note that

$$(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}} - \boldsymbol{X}\hat{\boldsymbol{B}})(\hat{\boldsymbol{\Sigma}}^{-1})_i = \sum_{j=1}^{g} \hat{\sigma}^{ij}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}}_j - \boldsymbol{X}\hat{\boldsymbol{B}}_j)$$

$$= \sum_{j=1}^{g} \hat{\sigma}^{ij}(\boldsymbol{y}_j - \boldsymbol{Z}_j\hat{\boldsymbol{\delta}}_j).$$

The full set of first-order conditions defining the FIML estimates can thus be written as

$$\sum_{j=1}^{g} \hat{\sigma}^{ij}\hat{\boldsymbol{Z}}_i^{\top}(\boldsymbol{y}_j - \boldsymbol{Z}_j\hat{\boldsymbol{\delta}}_j) = \boldsymbol{0}, \quad \text{for } i = 1, \ldots, g. \tag{18.72}$$

The conditions (18.72) are now in a form very similar indeed to that of the conditions (18.63) that define the 3SLS estimator. In fact, if we let $\bar{\boldsymbol{Y}}_i$ denote the $n \times g_i$ matrix of fitted values from the *unrestricted* reduced form, so that $\bar{\boldsymbol{Y}}_i = \boldsymbol{P}_X\boldsymbol{Y}_i$ for $i = 1, \ldots, g$, then

$$\boldsymbol{P}_X\boldsymbol{Z}_i = \boldsymbol{P}_X[\boldsymbol{X}_i \quad \boldsymbol{Y}_i] = [\boldsymbol{X}_i \quad \bar{\boldsymbol{Y}}_i] \equiv \bar{\boldsymbol{Z}}_i.$$

Thus the conditions (18.63) that define the 3SLS estimator can be written as

$$\sum_{j=1}^{g} \tilde{\sigma}^{ij}\bar{\boldsymbol{Z}}_i^{\top}(y_j - \boldsymbol{Z}_j\tilde{\boldsymbol{\delta}}_j) = \boldsymbol{0}. \tag{18.73}$$

where $\boldsymbol{f}_i(\cdot)$ is an n–vector of nonlinear functions, $\boldsymbol{u}_i$ is an n–vector of error terms, and $\boldsymbol{\theta}$ is a p–vector of parameters to be estimated. In general, subject to whatever restrictions need to be imposed for the system to be identified, all the endogenous and exogenous variables and all the parameters may appear in any equation.

The first step in any sort of IV procedure is to choose the instruments to be used. If the model is nonlinear only in the parameters, the matrix of optimal instruments is $\boldsymbol{X}$. As we have seen, however, there is no simple way to choose the instruments for models that are nonlinear in one or more of the endogenous variables. The theory of Section 17.4 can be applied, of course, but the result that it yields is not very practical. Under the usual assumptions about the error terms, namely, that they are homoskedastic and independent across observations but correlated across equations for each observation, one finds that a matrix of instruments $\boldsymbol{W}$ will be optimal if $\mathcal{S}(\boldsymbol{W})$ is equal to the subspace spanned by the union of the columns of the $E(\partial \boldsymbol{f}_i/\partial\boldsymbol{\theta})$. This result was originally derived by Amemiya (1977). It makes sense but is generally not very useful in practice. For now, we simply assume that *some* valid $n \times m$ matrix of instruments $\boldsymbol{W}$ is available, with $m \geq p$.

A nonlinear IV procedure for full-system estimation, similar in spirit to the single-equation NL2SLS procedure based on minimizing (18.78), was first proposed by Jorgenson and Laffont (1974) and called **nonlinear three-stage least squares**, or **NL3SLS**. The name is somewhat misleading, for the same reason that the name "NL2SLS" is misleading. By analogy with (18.60), the criterion function we would really like to minimize is

$$\sum_{i=1}^{g}\sum_{j=1}^{g} \sigma^{ij} \boldsymbol{f}_i^{\top}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})\boldsymbol{P}_W \boldsymbol{f}_j(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta}). \tag{18.80}$$

In practice, however, the elements σ^{ij} of the inverse of the contemporaneous covariance matrix $\boldsymbol{\Sigma}$ will not be known and will have to be estimated. This may be done in several ways. One possibility is to use NL2SLS for each equation separately. This will generally be easy, but it may not be possible if some parameters are identified only by cross-equation restrictions. Another approach which will work in that case is to minimize the criterion function

$$\sum_{i=1}^{g} \boldsymbol{f}_i^{\top}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})\boldsymbol{P}_W \boldsymbol{f}_i(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta}), \tag{18.81}$$

in which the unknown covariance matrix $\boldsymbol{\Sigma}$ is replaced by the identity matrix. The estimator obtained by minimizing (18.81) will evidently be a valid GMM estimator and thus will be consistent even though it is inefficient. Whichever inefficient estimator is used initially, it will yield g vectors of residuals $\acute{\boldsymbol{u}}_i$ from which the matrix $\boldsymbol{\Sigma}$ may be estimated consistently in exactly the same way as for linear models; see (18.62). Replacing the unknown σ^{ij}'s in (18.80) by

the elements $\acute{\sigma}^{ij}$ of the inverse of the estimate of $\boldsymbol{\Sigma}$ then yields the criterion function

$$\sum_{i=1}^{g}\sum_{j=1}^{g}\acute{\sigma}^{ij}\boldsymbol{f}_i^\top(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})\boldsymbol{P}_W\boldsymbol{f}_j(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta}), \tag{18.82}$$

which can actually be minimized in practice.

As usual, the minimized value of the criterion function (18.82) provides a test statistic for overidentifying restrictions; see Sections 7.8 and 17.6. If the model and instruments are correctly specified, this test statistic will be asymptotically distributed as $\chi^2(m-p)$; recall that m is the number of instruments and p is the number of free parameters. Moreover, if the model is estimated unrestrictedly and subject to r distinct restrictions, the difference between the two values of the criterion function will be asymptotically distributed as $\chi^2(r)$. If the latter test statistic is to be employed, it is important that the same estimate of $\boldsymbol{\Sigma}$ be used for both estimations, since otherwise the test statistic may not even be positive in finite samples.

When the sample size is large, it may be less computationally demanding to obtain one-step efficient estimates rather than actually to minimize (18.82). Suppose the initial consistent estimates, which may be either NL2SLS estimates or systems estimates based on (18.81), are denoted $\acute{\boldsymbol{\theta}}$. Then a first-order Taylor-series approximation to $\boldsymbol{f}_i(\boldsymbol{\theta}) \equiv \boldsymbol{f}_i(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})$ around $\acute{\boldsymbol{\theta}}$ is

$$\boldsymbol{f}_i(\acute{\boldsymbol{\theta}}) + \boldsymbol{F}_i(\acute{\boldsymbol{\theta}})(\boldsymbol{\theta} - \acute{\boldsymbol{\theta}}),$$

where $\boldsymbol{F}_i$ is an $n \times p$ matrix of the derivatives of $\boldsymbol{f}_i(\boldsymbol{\theta})$ with respect to the p elements of $\boldsymbol{\theta}$. If certain parameters do not appear in the ith equation, the corresponding columns of $\boldsymbol{F}_i$ will be identically zero. The one-step estimates, which will be asymptotically equivalent to NL3SLS estimates, are simply $\grave{\boldsymbol{\theta}} = \acute{\boldsymbol{\theta}} - \acute{\boldsymbol{t}}$, where $\acute{\boldsymbol{t}}$ denotes the vector of *linear* 3SLS estimates

$$\acute{\boldsymbol{t}} = \left[\sum_{i=1}^{g}\sum_{j=1}^{g}\acute{\sigma}^{ij}\acute{\boldsymbol{F}}_i^\top\boldsymbol{P}_W\acute{\boldsymbol{F}}_j\right]^{-1}\left[\sum_{i=1}^{g}\sum_{j=1}^{g}\acute{\sigma}^{ij}\acute{\boldsymbol{F}}_i^\top\boldsymbol{P}_W\acute{\boldsymbol{f}}_j\right]. \tag{18.83}$$

Compare expression (18.64), for the case with no cross-equation restrictions.

It is clear that NL3SLS can be generalized to handle heteroskedasticity of unknown form, serial correlation of unknown form, or both. For example, to handle heteroskedasticity one would simply replace the matrix $\boldsymbol{P}_W$ in (18.82) and (18.83) by the matrix

$$\boldsymbol{W}(\boldsymbol{W}^\top\acute{\boldsymbol{\Omega}}_{ij}\boldsymbol{W})^{-1}\boldsymbol{W}^\top,$$

where, by analogy with (18.76), $\acute{\boldsymbol{\Omega}}_{ij} = \text{diag}(\acute{u}_{ti}\acute{u}_{tj})$ for $i,j = 1,\ldots,g$. The initial estimates $\acute{\boldsymbol{\theta}}$ need not take account of heteroskedasticity. For a more detailed discussion of this sort of procedure, and of NL3SLS in general, see Gallant (1987, Chapter 6).

The other full-systems estimation method that is widely used is **nonlinear FIML**. For this, it is convenient to write the equation system to be estimated not as (18.79) but rather as

$$\boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta}) = \boldsymbol{U}_t, \quad \boldsymbol{U}_t \sim \text{NID}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{18.84}$$

where $\boldsymbol{\theta}$ is still a p-vector of parameters, $\boldsymbol{h}_t$ is a $1 \times g$ vector of nonlinear functions, and $\boldsymbol{U}_t$ is a $1 \times g$ vector of error terms. There need be no conflict between (18.79) and (18.84) if we think of the ith element of $\boldsymbol{h}_t(\cdot)$ as being the same as the tth element of $\boldsymbol{f}_i(\cdot)$.

The density of the vector $\boldsymbol{U}_t$ is

$$(2\pi)^{-g/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\tfrac{1}{2} \boldsymbol{U}_t \boldsymbol{\Sigma}^{-1} \boldsymbol{U}_t^\top\right).$$

To obtain the density of $\boldsymbol{Y}_t$, we must replace $\boldsymbol{U}_t$ by $\boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta})$ and multiply by the Jacobian factor $|\det \boldsymbol{J}_t|$, where $\boldsymbol{J}_t \equiv \partial \boldsymbol{h}_t(\boldsymbol{\theta}) / \partial \boldsymbol{Y}_t$, that is, the $g \times g$ matrix of derivatives of $\boldsymbol{h}_t$ with respect to the elements of $\boldsymbol{Y}_t$. The result is

$$(2\pi)^{-g/2} |\det \boldsymbol{J}_t| |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\tfrac{1}{2} \boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta}) \boldsymbol{\Sigma}^{-1} \boldsymbol{h}_t^\top(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta})\right).$$

It follows immediately that the loglikelihood function is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = -\frac{ng}{2} \log(2\pi) + \sum_{t=1}^{n} \log|\det \boldsymbol{J}_t| - \frac{n}{2} \log|\boldsymbol{\Sigma}|$$
$$- \frac{1}{2} \sum_{t=1}^{n} \boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta}) \boldsymbol{\Sigma}^{-1} \boldsymbol{h}_t^\top(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta}). \tag{18.85}$$

This may then be maximized with respect to $\boldsymbol{\Sigma}$ and the result substituted back in to yield the concentrated loglikelihood function

$$\ell^c(\boldsymbol{\theta}) = -\frac{ng}{2}\big(\log(2\pi) + 1\big) + \sum_{t=1}^{n} \log|\det \boldsymbol{J}_t|$$
$$- \frac{n}{2} \log\left|\frac{1}{n} \sum_{t=1}^{n} \boldsymbol{h}_t^\top(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta}) \boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\theta})\right|. \tag{18.86}$$

Inevitably, there is a strong resemblance between (18.85) and (18.86) and their counterparts (18.28) and (18.30) for the linear case. The major difference is that the Jacobian term in (18.85) and (18.86) is the sum of the logs of n different determinants. Thus every time one evaluates one of these loglikelihood functions, one has to calculate n different determinants. This can be very expensive if g or n is large. Of course, the problem goes away if the model is linear in the endogenous variables, since $\boldsymbol{J}_t$ will then be the same for all t.

Chapter 20

Unit Roots and Cointegration

20.1 INTRODUCTION

As we saw in the last chapter, the usual asymptotic results cannot be expected to apply if any of the variables in a regression model is generated by a nonstationary process. For example, in the case of the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, the usual results depend on the assumption that the matrix $n^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}$ tends to a finite, positive definite matrix as the sample size n tends to infinity. When this assumption is violated, some very strange things can happen, as we saw when we discussed "spurious" regressions between totally unrelated variables in Section 19.2. This is a serious practical problem, because a great many economic time series trend upward over time and therefore seem to violate this assumption.

Two obvious ways to keep standard assumptions from being violated when using such series are to detrend or difference them prior to use. But detrending and differencing are very different operations; if the former is appropriate, the latter will not be, and vice versa. Detrending a time series y_t will be appropriate if it is trend-stationary, which means that the DGP for y_t can be written as

$$y_t = \gamma_0 + \gamma_1 t + u_t, \tag{20.01}$$

where t is a time trend and u_t follows a stationary ARMA process. On the other hand, differencing will be appropriate if the DGP for y_t can be written as

$$y_t = \gamma_1 + y_{t-1} + u_t, \tag{20.02}$$

where again u_t follows a stationary ARMA process. If the u_t's were serially independent, (20.02) would be a random walk with drift, the drift parameter being γ_1. They will generally not be serially independent, however. As we will see shortly, it is no accident that the same parameter γ_1 appears in both (20.01) and (20.02).

The choice between detrending and differencing comes down to a choice between (20.01) and (20.02). The main techniques for choosing between them are various tests for what are called **unit roots**. The terminology comes from the literature on time-series processes. Recall from Section 10.5 that for an AR

process $A(L)u_t = \varepsilon_t$, where $A(L)$ denotes a polynomial in the lag operator, the stationarity of the process depends on the roots of the polynomial equation $A(z) = 0$. If all roots are outside the unit circle, the process is stationary. If any root is equal to or less than 1 in absolute value, the process is not stationary. A root that is equal to 1 in absolute value is called a **unit root**. When a process has a unit root, as (20.02) does, it is said to be **integrated of order one** or $I(1)$. A series that is $I(1)$ must be differenced once in order to make it stationary.

The obvious way to choose between (20.01) and (20.02) is to nest them both within a more general model. There is more than one way to do so. The most plausible model that includes both (20.01) and (20.02) as special cases is arguably

$$y_t = \gamma_0 + \gamma_1 t + v_t; \quad v_t = \alpha v_{t-1} + u_t$$
$$= \gamma_0 + \gamma_1 t + \alpha\big(y_{t-1} - \gamma_0 - \gamma_1(t-1)\big) + u_t, \qquad (20.03)$$

where u_t follows a stationary process. This model was advocated by Bhargava (1986). When $|\alpha| < 1$, (20.03) is equivalent to the trend-stationary model (20.01); when $\alpha = 1$, it reduces to (20.02).

Because (20.03) is nonlinear in the parameters, it is convenient to reparametrize it as

$$y_t = \beta_0 + \beta_1 t + \alpha y_{t-1} + u_t, \qquad (20.04)$$

where

$$\beta_0 \equiv \gamma_0(1 - \alpha) + \gamma_1\alpha \quad \text{and} \quad \beta_1 \equiv \gamma_1(1 - \alpha).$$

It is easy to verify that the estimates of α from least squares estimation of (20.03) and (20.04) will be identical, as will the estimated standard errors of those estimates if, in the case of (20.03), the latter are based on the Gauss-Newton regression. The only problem with the reparametrization (20.04) is that it hides the important fact that $\beta_1 = 0$ when $\alpha = 1$.

If y_{t-1} is subtracted from both sides, equation (20.04) becomes

$$\Delta y_t = \beta_0 + \beta_1 t + (\alpha - 1)y_{t-1} + u_t, \qquad (20.05)$$

where Δ is the first-difference operator. If $\alpha < 1$, (20.05) is equivalent to the model (20.01), whereas, if $\alpha = 1$, it is equivalent to (20.02). Thus it is conventional to test the null hypothesis that $\alpha = 1$ against the one-sided alternative that $\alpha < 1$. Since this is a test of the null hypothesis that there is a unit root in the stochastic process which generates y_t, such tests are commonly called **unit root tests**.

At first glance, it might appear that a unit root test could be accomplished simply by using the ordinary t statistic for $\alpha - 1 = 0$ in (20.05), but this is not so. When $\alpha = 1$, the process generating y_t is integrated of order one. This means that y_{t-1} will not satisfy the standard assumptions needed

These are often called **augmented Dickey-Fuller tests**, or **ADF tests**. They were proposed originally by Dickey and Fuller (1979) under the assumption that the error terms follow an AR process of known order. Subsequent work by Said and Dickey (1984) and Phillips and Perron (1988) showed that they are asymptotically valid under much less restrictive assumptions. Consider the test regressions (20.05), (20.06), (20.07), or (20.11). We can write any of these regressions as

$$\Delta y_t = \boldsymbol{X}_t \boldsymbol{\beta} + (\alpha - 1)y_{t-1} + u_t, \tag{20.14}$$

where $\boldsymbol{X}_t$ consists of whatever set of nonstochastic regressors is included in the test regression: nothing at all for (20.06), a constant for (20.07), a constant and a linear trend for (20.05), and so on.

Now suppose, for simplicity, that the error term u_t in (20.14) follows the stationary AR(1) process $u_t = \rho u_{t-1} + \varepsilon_t$. Then (20.14) would become

$$\Delta y_t = \boldsymbol{X}_t \boldsymbol{\beta} - \rho \boldsymbol{X}_{t-1} \boldsymbol{\beta} + (\rho + \alpha - 1)y_{t-1} - \alpha\rho y_{t-2} + \varepsilon_t$$

$$= \boldsymbol{X}_t \boldsymbol{\beta}^* + (\rho + \alpha - 1 - \alpha\rho)y_{t-1} + \alpha\rho(y_{t-1} - y_{t-2}) + \varepsilon_t \tag{20.15}$$

$$= \boldsymbol{X}_t \boldsymbol{\beta}^* + (\alpha - 1)(1 - \rho)y_{t-1} + \alpha\rho\Delta y_{t-1} + \varepsilon_t. \tag{20.16}$$

We are able to replace $\boldsymbol{X}_t \boldsymbol{\beta} - \rho \boldsymbol{X}_{t-1} \boldsymbol{\beta}$ by $\boldsymbol{X}_t \boldsymbol{\beta}^*$ in (20.15), for some choice of $\boldsymbol{\beta}^*$, because every column of $\boldsymbol{X}_{t-1}$ lies in $\mathcal{S}(\boldsymbol{X})$. This is a consequence of the fact that $\boldsymbol{X}_t$ can include only such deterministic variables as a constant, a linear trend, and so on (see Section 10.9). Thus each element of $\boldsymbol{\beta}^*$ is a linear combination of the elements of $\boldsymbol{\beta}$.

Equation (20.16) is a linear regression of Δy_t on $\boldsymbol{X}_t$, y_{t-1}, and Δy_{t-1}. This is just the original regression (20.14), with one additional regressor, Δy_{t-1}. Adding this regressor has caused the serially dependent error term u_t to be replaced by the serially independent error term ε_t. The ADF version of the τ statistic, which we will refer to as the τ' statistic, is simply the ordinary t statistic for the coefficient on y_{t-1} in (20.16) to be zero. If the serial correlation in the error terms of (20.14) were fully accounted for by an AR(1) process, this τ' statistic would have exactly the same asymptotic distribution as the ordinary DF τ statistic for the same specification of $\boldsymbol{X}_t$. The fact that the coefficient on y_{t-1} is $(\alpha-1)(1-\rho)$ rather than $\alpha-1$ does not matter. Because it is assumed that $|\rho| < 1$, this coefficient can be zero only if $\alpha = 1$. Thus a test for the coefficient on y_{t-1} to be zero is equivalent to a test for $\alpha = 1$.

It is evidently very easy to compute τ' statistics using regressions like (20.16), but it is not so easy to compute the corresponding z' statistics. If the coefficient of y_{t-1} were multiplied by n, the result would be $n(\hat{\alpha} - 1)(1 - \hat{\rho})$ rather than $n(\hat{\alpha} - 1)$. This test statistic clearly would not have the same asymptotic distribution as z. Thus, in order to compute a valid z' statistic from regression (20.16), it is necessary to divide the coefficient of y_{t-1} by $1 - \hat{\rho}$; see Dickey, Bell, and Miller (1986).

   Serial correlation is not the only complication that one is likely to en-
counter when trying to compute unit root test statistics. One very serious
problem is that these statistics are severely biased against rejecting the null
hypothesis when they are used with data that have been seasonally adjusted
by means of a linear filter or by the methods used by government statistical
agencies. In Section 19.6, we discussed the tendency of the OLS estimate of
α in the regression $y_t = \beta_0 + \alpha y_{t-1} + u_t$ to be biased toward 1 when y_t is
a seasonally adjusted series. This bias is present for all the test regressions
we have discussed. Even when $\hat{\alpha}$ is not actually biased *toward* 1, it will be
less biased *away* from 1 than the corresponding estimate using an unfiltered
series. Since the tabulated distributions of the test statistics are based on
the behavior of $\hat{\alpha}$ for the latter case, it is likely that test statistics computed
using seasonally adjusted data will reject the null hypothesis substantially less
often than they should according to the critical values in Table 20.1. That
is exactly what Ghysels and Perron (1993) found in a series of Monte Carlo
experiments.

   If possible, one should therefore avoid using seasonally adjusted data to
compute unit root tests. One possibility is to use annual data. This may
cause the sample size to be quite small, but the consequences of that are not
as severe as one might fear. As Shiller and Perron (1985) point out, the power
of these tests depends more on the **span** of the data (i.e., the number of years
the sample covers) than on the number of observations. The reason for this is
that if α is in fact positive but less than 1, it will be closer to 1 when the data
are observed more frequently. Thus a test based on n annual observations may
have only slightly less power than a test based on $4n$ quarterly observations
that have not been seasonally adjusted and may have more power than a test
based on $4n$ seasonally adjusted observations.

   If quarterly or monthly data are to be used, they should if possible not be
seasonally adjusted. Unfortunately, as we remarked in Chapter 19, seasonally
unadjusted data for many time series are not available in many countries.
Moreover, the use of seasonally unadjusted data may make it necessary to
add seasonal dummy variables to the regression and to account for fourth-
order or twelfth-order serial correlation.

   A second major problem with unit root tests is that they are very sensitive
to the assumption that the process generating the data has been stable over
the entire sample period. Perron (1989) showed that the power of unit root
tests is dramatically reduced if the level or the trend of a series has changed
exogenously at any time during the sample period. Even though the series
may actually be stationary in each of the two parts of the sample, it can be
almost impossible to reject the null that it is $I(1)$ in such cases.

   Perron therefore proposed techniques that can be used to test for unit
roots conditional on exogenous changes in level or trend. His tests are per-
formed by first regressing y_t on a constant, a time trend, and one or two
dummy variables that allow either the constant, the trend, or both the con-

employed. We know that variables which are $I(1)$ tend to diverge as $n \to \infty$, because their unconditional variances are proportional to n. Thus it might seem that such variables could never be expected to obey any sort of long-run equilibrium relationship. But in fact it is possible for two or more variables to be $I(1)$ and yet for certain linear combinations of those variables to be $I(0)$. If that is the case, the variables are said to be **cointegrated**. If two or more variables are cointegrated, they must obey an equilibrium relationship in the long run, although they may diverge substantially from equilibrium in the short run. The concept of cointegration is fundamental to the understanding of long-run relationships among economic time series. It is also quite recent. The earliest reference is probably Granger (1981), the best-known paper is Engle and Granger (1987), and two relatively accessible articles are Hendry (1986) and Stock and Watson (1988a).

Suppose, to keep matters simple, that we are concerned with just two variables, y_{t1} and y_{t2}, each of which is known to be $I(1)$. Then, in the simplest case, y_{t1} and y_{t2} would be cointegrated if there exists a vector $\boldsymbol{\eta} \equiv [1 \quad -\eta_2]^\top$ such that, when the two variables are in equilibrium,

$$[\boldsymbol{y}_1 \quad \boldsymbol{y}_2]\boldsymbol{\eta} \equiv \boldsymbol{y}_1 - \eta_2\boldsymbol{y}_2 = \boldsymbol{0}. \tag{20.20}$$

Here $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ denote n–vectors with typical elements y_{t1} and y_{t2}, respectively. The 2–vector $\boldsymbol{\eta}$ is called a **cointegrating vector**. It is clearly not unique, since it could be multiplied by any nonzero scalar without affecting the equality in (20.20).

Realistically, one might well expect y_{t1} and y_{t2} to be changing systematically as well as stochastically over time. Thus one might expect (20.20) to contain a constant term and perhaps one or more trend terms as well. If we write $\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2]$, (20.20) can be rewritten to allow for this possibility as

$$\boldsymbol{Y}\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}, \tag{20.21}$$

where, as in (20.14), $\boldsymbol{X}$ denotes a nonstochastic matrix that may or may not have any elements. If it does, the first column will be a constant, the second, if it exists, will be a linear time trend, the third, if it exists, will be a quadratic time trend, and so on. Since $\boldsymbol{Y}$ could contain more than two variables, (20.21) is actually a very general way of writing a cointegrating relationship among any number of variables.

At any particular time t, of course, an equality like (20.20) or (20.21) cannot be expected to hold exactly. We may therefore define the **equilibrium error** ν_t as

$$\nu_t = \boldsymbol{Y}_t\boldsymbol{\eta} - \boldsymbol{X}_t\boldsymbol{\beta}, \tag{20.22}$$

where $\boldsymbol{Y}_t$ and $\boldsymbol{X}_t$ denote the tth rows of $\boldsymbol{Y}$ and $\boldsymbol{X}$, respectively. In the special case of (20.20), this equilibrium error would simply be $y_{t1} - \eta_2 y_{t2}$. The m variables y_{t1} through y_{tm} are said to be cointegrated if there exists a vector $\boldsymbol{\eta}$ such that ν_t in (20.22) is $I(0)$.

determinant to be minimized can be expressed as a function of $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ alone, as follows:

$$\left|\left(\Delta\boldsymbol{Y} - \boldsymbol{Y}_{-p}\boldsymbol{\eta}\boldsymbol{\alpha}^\top\right)^\top\boldsymbol{M}_\Delta\left(\Delta\boldsymbol{Y} - \boldsymbol{Y}_{-p}\boldsymbol{\eta}\boldsymbol{\alpha}^\top\right)\right|. \tag{20.38}$$

Let us write $\boldsymbol{Y}_{-p}^*$ for $\boldsymbol{M}_\Delta\boldsymbol{Y}_{-p}$ and $\Delta\boldsymbol{Y}^*$ for $\boldsymbol{M}_\Delta\Delta\boldsymbol{Y}$. Then (20.38) can be expressed as

$$\left|\left(\Delta\boldsymbol{Y}^* - \boldsymbol{Y}_{-p}^*\boldsymbol{\eta}\boldsymbol{\alpha}^\top\right)^\top\left(\Delta\boldsymbol{Y}^* - \boldsymbol{Y}_{-p}^*\boldsymbol{\eta}\boldsymbol{\alpha}^\top\right)\right|. \tag{20.39}$$

It is now easy to concentrate this expression with respect to $\boldsymbol{\alpha}$, for, if we hold $\boldsymbol{\eta}$ fixed, the residuals in (20.39) depend linearly on $\boldsymbol{\alpha}$. If $\boldsymbol{V} \equiv \boldsymbol{Y}_{-p}^*\boldsymbol{\eta}$, we obtain the determinant

$$\left|(\Delta\boldsymbol{Y}^*)^\top\boldsymbol{M}_V\Delta\boldsymbol{Y}^*\right|. \tag{20.40}$$

By use of the same trick we had recourse to in Section 18.5, we can treat (20.40) as one factor in the decomposition of the determinant of a larger matrix. Consider

$$\left|\begin{matrix}(\Delta\boldsymbol{Y}^*)^\top\Delta\boldsymbol{Y}^* & (\Delta\boldsymbol{Y}^*)^\top\boldsymbol{V} \\ \boldsymbol{V}^\top\Delta\boldsymbol{Y}^* & \boldsymbol{V}^\top\boldsymbol{V}\end{matrix}\right|.$$

By the result (A.26) of Appendix A, this matrix can be factorized either as

$$\left|\boldsymbol{V}^\top\boldsymbol{V}\right|\left|(\Delta\boldsymbol{Y}^*)^\top\boldsymbol{M}_V\Delta\boldsymbol{Y}^*\right|$$

or as

$$\left|(\Delta\boldsymbol{Y}^*)^\top\Delta\boldsymbol{Y}^*\right|\left|\boldsymbol{V}^\top\boldsymbol{M}^*\boldsymbol{V}\right|,$$

where $\boldsymbol{M}^*$ projects orthogonally onto $\mathcal{S}^\perp(\Delta\boldsymbol{Y}^*)$. Since $\left|(\Delta\boldsymbol{Y}^*)^\top\Delta\boldsymbol{Y}^*\right|$ does not depend on $\boldsymbol{\eta}$, we see that minimizing (20.40) is equivalent to minimizing the ratio

$$\frac{\left|\boldsymbol{V}^\top\boldsymbol{M}^*\boldsymbol{V}\right|}{\left|\boldsymbol{V}^\top\boldsymbol{V}\right|} = \frac{\left|\boldsymbol{\eta}^\top(\boldsymbol{Y}_{-p}^*)^\top\boldsymbol{M}^*\boldsymbol{Y}_{-p}^*\boldsymbol{\eta}\right|}{\left|\boldsymbol{\eta}^\top(\boldsymbol{Y}_{-p}^*)^\top\boldsymbol{Y}_{-p}^*\boldsymbol{\eta}\right|} \tag{20.41}$$

with respect to $\boldsymbol{\eta}$. The minimum of (20.40) is then the minimum of (20.41) times $\left|(\Delta\boldsymbol{Y}^*)^\top\Delta\boldsymbol{Y}^*\right|$.

The least variance ratio problem that had to be solved in the LIML context (see (18.49)) involved a ratio of quadratic forms rather than the determinants that appear in (20.41). Even so, the present problem can be solved by the same technique as (18.49), namely, by converting the problem into an eigenvalue-eigenvector problem. Before we go into details, notice that (20.41) is invariant if $\boldsymbol{\eta}$ is replaced by $\boldsymbol{\eta}\boldsymbol{B}$, for any nonsingular $r \times r$ matrix $\boldsymbol{B}$. This is precisely what we noted earlier in speaking of the nonuniqueness of (20.36). We therefore cannot expect to obtain a unique minimizing $\boldsymbol{\eta}$ but only an r–dimensional subspace.

For the actual minimization, it is convenient to work with a transformation of $\boldsymbol{\eta}$. Let $\boldsymbol{S}$ denote any $m \times m$ matrix with the property that $\boldsymbol{S}^{\top}\boldsymbol{S} = (\boldsymbol{Y}^*_{-p})^{\top}\boldsymbol{Y}^*_{-p}$, and define the $m \times r$ matrix $\boldsymbol{\zeta}$ as $\boldsymbol{S}\boldsymbol{\eta}$. The ratio (20.41) becomes

$$\frac{\left|\boldsymbol{\zeta}^{\top}(\boldsymbol{S}^{-1})^{\top}(\boldsymbol{Y}^*_{-p})^{\top}\boldsymbol{M}^*\boldsymbol{Y}^*_{-p}\boldsymbol{S}^{-1}\boldsymbol{\zeta}\right|}{\left|\boldsymbol{\zeta}^{\top}\boldsymbol{\zeta}\right|}. \tag{20.42}$$

Since all that matters is the subspace spanned by the r columns of $\boldsymbol{\zeta}$, we may without loss of generality choose $\boldsymbol{\zeta}$ such that $\boldsymbol{\zeta}^{\top}\boldsymbol{\zeta} = \mathbf{I}_r$. Let us define the $m \times m$ positive definite matrix $\boldsymbol{A}$ to be the matrix that appears in the numerator of (20.42). Then we have to minimize $|\boldsymbol{\zeta}^{\top}\boldsymbol{A}\boldsymbol{\zeta}|$ with respect to $\boldsymbol{\zeta}$ subject to the constraint that $\boldsymbol{\zeta}^{\top}\boldsymbol{\zeta} = \mathbf{I}$.

In order to perform this minimization, it turns out to be enough to consider the eigenvalue-eigenvector problem associated with $\boldsymbol{A}$. If we solve this problem, we will obtain an orthogonal matrix $\boldsymbol{Z}$, the columns of which are orthonormalized eigenvectors of $\boldsymbol{A}$, and a diagonal matrix $\boldsymbol{\Lambda}$, the diagonal elements of which are the eigenvalues of $\boldsymbol{A}$, which must evidently lie between zero and unity. Then $\boldsymbol{A}\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{\Lambda}$. If the columns of $\boldsymbol{Z}$ and $\boldsymbol{\Lambda}$ are arranged in increasing order of the eigenvalues $\lambda_1, \ldots, \lambda_m$, we may choose the ML estimate $\hat{\boldsymbol{\zeta}}$ to be the first r columns of $\boldsymbol{Z}$. Geometrically, the columns of $\hat{\boldsymbol{\zeta}}$ span the space spanned by the eigenvectors of $\boldsymbol{A}$ that correspond to the r smallest eigenvalues. The fact that $\boldsymbol{Z}$ is orthogonal means that $\hat{\boldsymbol{\zeta}}$ satisfies the constraint, and the choice of the *smallest* eigenvalues serves to minimize the determinant $|\boldsymbol{\zeta}^{\top}\boldsymbol{A}\boldsymbol{\zeta}|$.

The ML estimate of the space of cointegrating vectors $\mathcal{S}(\boldsymbol{\eta})$ can now be recovered from $\hat{\boldsymbol{\zeta}}$ by the formula $\hat{\boldsymbol{\eta}} = \boldsymbol{S}^{-1}\hat{\boldsymbol{\zeta}}$. The matrix $\hat{\boldsymbol{\alpha}}$ needed in order to obtain ML estimates of the parameters contained in the matrix $\boldsymbol{\Pi}$ can then be obtained as the OLS estimates from the multivariate regression of $\Delta\boldsymbol{Y}^*$ on $\boldsymbol{Y}^*_{-p}\hat{\boldsymbol{\eta}}$. Subsequently, estimates of the matrices $\boldsymbol{\Gamma}_i$, $i = 1, \ldots, p-1$, can also be obtained by OLS.

Often, we are not especially interested in the parameters of the VAR (20.35). The focus of our interest is more likely to be testing the hypothesis of noncointegration against an alternative of cointegration of some chosen order. Should the null hypothesis that $r = 0$ be rejected, we may then wish to test the hypothesis that $r = 1$ against the alternative that $r = 2$, and so forth. The eigenvalues λ_i, $i = 1, \ldots, m$, provide a very convenient way to do this, in terms of a likelihood ratio test. It is clear that if we select some value of r, the minimized determinant $|\boldsymbol{\zeta}^{\top}\boldsymbol{A}\boldsymbol{\zeta}|$ is just the product of the r smallest eigenvalues, $\lambda_1 \cdots \lambda_r$. The minimum of (20.40) is this product multiplied by $|(\Delta\boldsymbol{Y}^*)^{\top}\Delta\boldsymbol{Y}^*|$. If $r = 0$, then the minimum of (20.40) is simply this last determinant. Likelihood ratios for different values of r are therefore just products of some of the eigenvalues, raised to the power $n/2$; recall (9.65). If we take logs and multiply by 2 in order to obtain an LR statistic, we obtain $-n$ times the sum of the logs of the appropriate eigenvalues.

results mean that it is usually practical to perform only a small number of experiments. These must be designed to shed as much light as possible on the issues of interest.

The first thing to recognize is that results from Monte Carlo experiments are necessarily random. At a minimum, this means that results must be reported in a way which allows readers to appreciate the extent of experimental randomness. Moreover, it is essential to perform enough replications so that the results are sufficiently accurate for the purpose at hand. The number of replications that is needed can sometimes be substantially reduced by using variance reduction techniques, which will be discussed in the next two sections. Such techniques are by no means always readily available, however. In this section, we consider various other aspects of the design of Monte Carlo experiments.

We first consider the problem of determining how many replications to perform. As an example, suppose that the investigator is interested in calculating the size of a certain test statistic (i.e., the probability of rejecting the null hypothesis when it is true) at, say, the nominal .05 level. Let us denote this unknown quantity by p. Each replication will generate a test statistic that either exceeds or does not exceed the nominal critical value. These can be thought of as independent Bernoulli trials. Suppose N replications are performed and R rejections are obtained. Then the obvious estimator of p, which is also the ML estimator, is R/N. The variance of this estimator is $N^{-1}p(1-p)$, which can be estimated by $R(N-R)/N^3$.

Now suppose that one wants the length of a 95% confidence interval on the estimate of p to be approximately .01. Using the normal approximation to the binomial, which is surely valid here since N will be a large number, we see that the confidence interval must cover $2 \times 1.96 = 3.92$ standard errors. Hence we require that

$$3.92 \left(\frac{p(1-p)}{N} \right)^{1/2} = .01. \qquad (21.02)$$

Assuming that p is .05, the nominal level of the test being investigated, we can solve (21.02) for N. The result is $N \cong 7299$. To be on the safe side (since p may well exceed .05, implying that R/N may have a larger variance) the investigator would probably choose $N = 8000$. This is a rather large number of replications and may be expensive to compute. If one were willing to let the 95% confidence interval on p have a length of .02, one could make do with a sample one-quarter as large, or roughly 2000 replications.

If the objective of an experiment is to compare two or more estimators or two or more test statistics, fewer replications may be needed to obtain a given level of accuracy than would be needed to estimate the properties of either of them with the same level of accuracy. Suppose, for example, that we are interested in comparing the biases of two estimators, say $\hat{\theta}$ and $\tilde{\theta}$, of a

parameter the true value of which is θ_0. On each replication, say the jth, we obtain realizations $\hat{\theta}_j$ and $\tilde{\theta}_j$ of each of the two estimators. The biases of the two estimators are

$$B(\hat{\theta}) \equiv E(\hat{\theta} - \theta_0) \quad \text{and} \quad B(\tilde{\theta}) \equiv E(\tilde{\theta} - \theta_0),$$

which may be estimated by

$$\hat{B}(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta_0) \quad \text{and} \quad \tilde{B}(\tilde{\theta}) = \frac{1}{N}\sum_{j=1}^{N}(\tilde{\theta}_j - \theta_0).$$

The difference between $B(\hat{\theta})$ and $B(\tilde{\theta})$ is

$$E(\hat{\theta} - \theta_0) - E(\tilde{\theta} - \theta_0) = E(\hat{\theta} - \tilde{\theta}), \tag{21.03}$$

which may be estimated by

$$\frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \tilde{\theta}_j). \tag{21.04}$$

It is possible and indeed likely that the variance of (21.04) will be substantially smaller than the variance of either $\hat{B}(\hat{\theta})$ or $\tilde{B}(\tilde{\theta})$, because both $\hat{\theta}_j$ and $\tilde{\theta}_j$ depend on the same pseudo-random vector $\boldsymbol{u}^j$. The variance of (21.04) is

$$\frac{1}{N}V(\hat{\theta}) + \frac{1}{N}V(\tilde{\theta}) - \frac{2}{N}\text{Cov}(\hat{\theta}, \tilde{\theta}),$$

which will be smaller than the variance of either $\hat{B}(\hat{\theta})$ or $\tilde{B}(\tilde{\theta})$ whenever $\text{Cov}(\hat{\theta}, \tilde{\theta})$ is positive and large enough. This will very often be the case, since it is likely that $\hat{\theta}_j$ and $\tilde{\theta}_j$ will be strongly positively correlated. Thus it may require far fewer replications to estimate (21.03) than to estimate $B(\hat{\theta})$ and $B(\tilde{\theta})$ with the same level of accuracy. Of course, this assumes that $\hat{\theta}_j$ and $\tilde{\theta}_j$ are obtained using the same set of pseudo-random variates, but that is how the Monte Carlo experiment would normally be designed. We will encounter an idea similar to this one when we discuss the method of antithetic variates in the next section.

The second important thing to keep in mind when designing Monte Carlo experiments is that the results will often be highly sensitive to certain aspects of the experimental design and largely or totally insensitive to other aspects. Obviously, one will want to vary the former across the experiments while fixing the latter in a more or less arbitrary fashion. For example, many test statistics related to regression models are invariant to the variance of the error terms. Consider the ordinary t statistic for $\alpha = 0$ in the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \alpha\boldsymbol{z} + \boldsymbol{u}. \tag{21.05}$$

repeated samples, but it may yield results that depend on the idiosyncratic characteristics of the particular set of X_t's which was drawn.

Another possibility is to use genuine economic data for the X_t's. If these data are chosen with care, this approach can ensure that the X_t's are indeed typical of those which appear in econometric models. However, it raises a problem of how to vary the sample size. If one uses either genuine data or a single set of generated data, the matrix $n^{-1}X^\top X$ will change as the sample size n changes. This may make it difficult to separate the effects of changes in n from the effects of changes in $n^{-1}X^\top X$. One solution to this problem is to pick, or generate, a single set of X_t's for a sample of size m and then repeat these as many times as necessary to create X_t's for samples of larger sizes. This requires that $n = cm$, where c is an integer. Obvious choices for m are 50 and 100; n could then be any integer multiple of 50 or 100. The problem with this approach, of course, is that no matter how many replications are performed, all the results will depend on the choice of the initial set of X_t's.

In many cases, how the X_t's are chosen will not matter much. However, there are cases for which it can have a substantial impact on the results. For example, MacKinnon and White (1985) used Monte Carlo experiments to examine the finite-sample performance of various heteroskedasticity-consistent covariance matrix estimators (HCCMEs; see Section 16.3). They used 50 observations on genuine economic data for the X_t's, repeating these 50 observations as many times as necessary for each sample size. As Chesher and Jewitt (1987) subsequently showed, the performance of the estimators depends critically on the h_t's, that is, the diagonal elements of the matrix P_X; the larger are the largest h_t's, the worse will be the finite-sample performance of tests based on all HCCMEs. When the X matrix is generated the way MacKinnon and White generated it, with $n = 50c$, all of the h_t's must approach zero at a rate proportional to $1/c$ (and hence also to $1/n$). Thus MacKinnon and White were guaranteed to find that results improved rapidly as the sample size was increased. In contrast, Cragg (1983), doing Monte Carlo experiments on a related issue (see Section 17.3), generated the X_t's randomly from the lognormal distribution. This distribution has a long right-hand tail and thus occasionally throws up large values of certain X_t's. These produce relatively large values of h_t, and as a result the largest values of h_t tend to zero at a rate very much slower than $1/n$. Thus, as the Chesher-Jewitt analysis would have predicted, Cragg found that finite-sample performance improved only slowly as the sample size was increased.

More recently, Chesher and Peters (1994) have shown that the distributions of many estimators of interest to econometricians depend crucially on the way the regressors are distributed. If the regressors are symmetrically distributed about their medians, these estimators will have special properties that do not hold in general. Since regressors used in Monte Carlo experiments might well be symmetrically distributed, there is a risk that the results of such experiments could be seriously misleading.

be impossible to compute $\ddot{\boldsymbol{\beta}}$ from a real data set, but in the context of a Monte Carlo experiment, it is perfectly easy to do so. We know $\boldsymbol{\beta}_0$ and hence $\boldsymbol{X}_0 \equiv \boldsymbol{X}(\boldsymbol{\beta}_0)$. Using these and the error vector $\boldsymbol{u}^j$ that we generate at each replication, we can easily compute $\ddot{\boldsymbol{\beta}}^j$.

Suppose that $\theta \equiv \theta(\hat{\boldsymbol{\beta}})$ is some scalar quantity of which we wish to calculate the mean using the results of the Monte Carlo experiment. For example, if we were interested in the bias of $\hat{\beta}_2$, θ would be $\hat{\beta}_2 - \beta_{20}$; if we were interested in the mean squared error of $\hat{\beta}_3$, θ would be $(\hat{\beta}_3 - \beta_{30})^2$; if we were interested in the size of a test, θ would be 1 if the test rejected and 0 otherwise; and so on. On each replication, we obtain t_j, a realization of θ, which is equal to $\theta(\hat{\boldsymbol{\beta}}^j)$. We also obtain a control variate τ_j, which would normally be some function of $\ddot{\boldsymbol{\beta}}$. The τ_j's must be known to have mean zero and finite variance, which need not be known. If we were interested in the bias of $\hat{\beta}_2$, for example, the natural choice for τ would be $\ddot{\beta}_2$. In some other cases, it is not so obvious how to choose τ, however, and there may be several possible choices.

If the control variate τ were not available, we would estimate θ by

$$\bar{\theta} \equiv \frac{1}{N} \sum_{j=1}^{N} t_j,$$

and this naive estimator would have variance $V(\bar{\theta}) = N^{-1} V(t)$, which could be estimated by

$$\hat{V}(\bar{\theta}) = \frac{1}{N(N-1)} \sum_{j=1}^{N} (t_j - \bar{\theta})^2.$$

When the control variate τ is available, $\bar{\theta}$ will in most cases no longer be optimal. Consider instead the control variate (CV) estimator

$$\ddot{\theta}(\lambda) \equiv \bar{\theta} - \lambda \bar{\tau}, \tag{21.10}$$

where $\bar{\tau}$ is the sample mean of the τ_j's. This estimator involves subtracting from $\bar{\theta}$ some multiple λ of the sample mean of the control variates; how λ may be chosen will be discussed in the next paragraph. On average, what is subtracted will be zero, since τ_j has population mean zero. This implies that $\ddot{\theta}(\lambda)$ must have the same population mean as $\bar{\theta}$. But, in any given sample, the mean of the τ_j's will be nonzero. If, for example, it is positive, and if τ_j and t_j are strongly positively correlated, it is very likely that $\bar{\theta}$ will also exceed its population mean. Thus, by subtracting from $\bar{\theta}$ a multiple of the mean of the τ_j's, we are likely to obtain a better estimate of θ.

The variance of the CV estimator (21.10) is

$$V\big(\ddot{\theta}(\lambda)\big) = V(\bar{\theta}) + \lambda^2 V(\bar{\tau}) - 2\lambda \text{Cov}(\bar{\theta}, \bar{\tau}). \tag{21.11}$$

tion corresponds to one experiment, the dependent variable is some quantity that was estimated in the experiments, and the independent variables are functions of the various parameter values, chosen by the experimenter, which characterize each experiment. Response surfaces have been used by Hendry (1979), Mizon and Hendry (1980), Engle, Hendry, and Trumble (1985), Ericsson (1991), and MacKinnon (1991), among others; they are discussed at length in Hendry (1984). For criticisms of this approach, see Maasoumi and Phillips (1982), along with the reply by Hendry (1982).

If a response surface that adequately explains the experimental results can be found, this approach to summarizing Monte Carlo results has much to recommend it. First of all, it may be a good deal easier to understand the behavior of the estimator or test statistic of interest from the parameters of a response surface than from several tables full of numbers. Secondly, if the response surface is correctly specified, it eliminates, or at least greatly reduces, what Hendry (1984) refers to as the problem of **specificity**. What this means is that each individual experiment gives results for a single assumed DGP only, and any set of Monte Carlo experiments gives results for a finite set of assumed DGPs only. For other parameter values or values of n, the reader must interpolate from the results in the tables, which is often difficult to do. In contrast, a correctly specified response surface gives results for whole families of DGPs rather than solely for the parameter values chosen by the experimenter. The catch, of course, is that the response surface must be correctly specified, and this is not always an easy task.

One of the most interesting features of response surfaces, which distinguishes them from most other applications of regression models in economics, is that the data are generated by the experimenter. Thus, if the data are not sufficiently informative, there is always an easy solution: Simply run more experiments and obtain more data. In most cases, each data point for the response surface corresponds to a single Monte Carlo experiment. The dependent variable is then some quantity estimated by the experiment, such as the mean or mean squared error of the estimates of a certain parameter or the estimated size of a test. Because such estimates are normally accompanied by estimates of their standard errors, estimates which should be very accurate if the experiments involve a sufficient number of replications, the investigator is in the unique position of being able to use GLS with a fully specified covariance matrix. If every experiment used a different set of random numbers, each observation for the response surface would be independent, and this covariance matrix would therefore be diagonal. If the same random numbers were used across several experiments, perhaps to increase the precision with which differences across parameter values were estimated, the covariance matrix would of course be nondiagonal, but the form of the nondiagonality would be known, and the covariance matrix could easily be estimated.

To make the above remarks more concrete, let us denote the quantity of interest by ψ. It must be a function of the sample size n and of the

that this determinant is a polynomial in λ, of degree n if $\boldsymbol{A}$ is $n \times n$. The fundamental theorem of algebra tells us that such a polynomial has n complex roots, say $\lambda_1, \ldots, \lambda_n$. To each λ_i there must correspond an eigenvector $\boldsymbol{x}_i$. This eigenvector is determined only up to a scale factor, because if $\boldsymbol{x}_i$ is an eigenvector corresponding to λ_i, then so is $\alpha \boldsymbol{x}_i$ for any nonzero scalar α. The eigenvector $\boldsymbol{x}_i$ does not necessarily have real elements if λ_i itself is not real.

If $\boldsymbol{A}$ is a real symmetric matrix, it can be shown that the eigenvalues λ_i are in fact all real and that the eigenvectors can be chosen to be real as well. If $\boldsymbol{A}$ is a positive definite matrix, then all its eigenvalues are positive. This follows from the facts that

$$\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} = \lambda \boldsymbol{x}^\top \boldsymbol{x}$$

and that both $\boldsymbol{x}^\top \boldsymbol{x}$ and $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$ are positive. The eigenvectors of a real symmetric matrix can be chosen to be mutually orthogonal. If one looks at two eigenvectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, corresponding to two distinct eigenvalues λ_i and λ_j, then $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are necessarily orthogonal:

$$\lambda_i \boldsymbol{x}_j^\top \boldsymbol{x}_i = \boldsymbol{x}_j^\top \boldsymbol{A} \boldsymbol{x}_i = (\boldsymbol{A} \boldsymbol{x}_j)^\top \boldsymbol{x}_i = \lambda_j \boldsymbol{x}_j^\top \boldsymbol{x}_i,$$

which is impossible unless $\boldsymbol{x}_j^\top \boldsymbol{x}_i = 0$. If not all the eigenvalues are distinct, then two (or more) eigenvectors may correspond to one and the same eigenvalue. When that happens, these two eigenvectors span a space that is orthogonal to all other eigenvalues by the reasoning just given. Since any linear combination of the two eigenvectors will also be an eigenvector corresponding to the one eigenvalue, one may choose an orthogonal set of them. Thus, whether or not all the eigenvalues are distinct, eigenvectors may be chosen to be **orthonormal**, by which we mean that they are mutually orthogonal and each has norm equal to 1. Thus the eigenvectors of a real symmetric matrix provide an orthonormal basis.

Let $\boldsymbol{U} \equiv [\, \boldsymbol{x}_1 \;\; \cdots \;\; \boldsymbol{x}_n \,]$ be a matrix the columns of which are an orthonormal set of eigenvectors of $\boldsymbol{A}$, corresponding to the eigenvalues λ_i, $i = 1, \ldots, n$. Then we can write the eigenvalue relationship (A.28) for all the eigenvalues at once as

$$\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}, \tag{A.30}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with λ_i as its ith diagonal element. The ith column of $\boldsymbol{A}\boldsymbol{U}$ is $\boldsymbol{A}\boldsymbol{x}_i$, and the ith column of $\boldsymbol{U}\boldsymbol{\Lambda}$ is $\lambda_i \boldsymbol{x}_i$. Since the columns of $\boldsymbol{U}$ are orthonormal, we find that $\boldsymbol{U}^\top \boldsymbol{U} = \mathbf{I}$, which implies that $\boldsymbol{U}^\top = \boldsymbol{U}^{-1}$. A matrix with this property is said to be an **orthogonal matrix**. Postmultiplying (A.30) by $\boldsymbol{U}^\top$ gives

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top. \tag{A.31}$$

This equation expresses the **diagonalization** of $\boldsymbol{A}$.

Berndt, E. R., and N. E. Savin (1975). "Estimation and hypothesis testing in singular equation systems with autoregressive disturbances," *Econometrica*, **43**, 937–57.

Berndt, E. R., and N. E. Savin (1977). "Conflict among criteria for testing hypotheses in the multivariate linear regression model," *Econometrica*, **45**, 1263–78.

Betancourt, R., and H. Kelejian (1981). "Lagged endogenous variables and the Cochrane-Orcutt procedure," *Econometrica*, **49**, 1073–78.

Bhargava, A. (1986). "On the theory of testing for unit roots in observed time series," *Review of Economic Studies*, **53**, 369–84.

Bickel, P. J., and K. A. Doksum (1981). "An analysis of transformations revisited," *Journal of the American Statistical Association*, **76**, 296–311.

Bickel, P. J., and D. A. Freedman (1981). "Some asymptotic theory for the bootstrap," *Annals of Statistics*, **9**, 1196–1217.

Billingsley, P. (1968). *Convergence of Probability Measures*, New York, John Wiley & Sons.

Billingsley, P. (1979). *Probability and Measure*, New York, John Wiley & Sons.

Blundell, R. W. (ed.) (1987). Special Issue on Specification Testing in Limited and Discrete Dependent Variable Models, *Journal of Econometrics*, **34**, No. 1/2.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, **31**, 307–27.

Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992). "ARCH modeling in finance: a review of the theory and empirical evidence," *Journal of Econometrics*, **52**, 5–59.

Bollerslev, T., R. F. Engle, and J. M. Wooldridge (1988). "A capital asset pricing model with time varying covariances," *Journal of Political Economy*, **96**, 116–31.

Boothe, P., and J. G. MacKinnon (1986). "A specification test for models estimated by GLS," *Review of Economics and Statistics*, **68**, 711–14.

Bowden, R. J., and D. A. Turkington (1984). *Instrumental Variables*, Cambridge, Cambridge University Press.

Bowen, W. G., and T. A. Finegan (1969). *The Economics of Labor Force Participation*, Princeton, Princeton University Press.

Box, G. E. P., and D. R. Cox (1964). "An analysis of transformations," *Journal of the Royal Statistical Society*, Series B, **26**, 211–52.

Box, G. E. P., and G. M. Jenkins (1976). *Time Series Analysis, Forecasting and Control*, Revised edition, San Francisco, Holden Day.

Box, G. E. P., and M. E. Muller (1958). "A note on the generation of random normal deviates," *Annals of Mathematical Statistics*, **29**, 610–11.

Box, G. E. P., and D. A. Pierce (1970). "The distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *Journal of the American Statistical Association*, **65**, 1509–26.

Breusch, T. S. (1978). "Testing for autocorrelation in dynamic linear models," *Australian Economic Papers*, **17**, 334–55.

Campbell, J. Y., and P. Perron (1991). "Pitfalls and opportunities: What macroeconomists should know about unit roots," in *NBER Macroeconomics Annual*, Vol. 6, eds. O. J. Blanchard and S. Fisher, Cambridge, Mass., MIT Press.

Campbell, J. Y., and R. J. Shiller (1987). "Cointegration and tests of present value models," *Journal of Political Economy*, **95**, 1062–88.

Carroll, R. J., and D. Ruppert (1982). "Robust estimation in heteroskedastic linear models," *Annals of Statistics*, **10**, 429–41.

Carroll, R. J., and D. Ruppert (1984). "Power transformations when fitting theoretical models to data," *Journal of the American Statistical Association*, **79**, 321–28.

Carroll, R. J., and D. Ruppert (1988). *Transformation and Weighting in Regression*, London, Chapman and Hall.

Casella, G., and R. L. Berger (1990). *Statistical Inference*, Pacific Grove, Calif., Wadsworth and Brooks/Cole.

Chamberlain, G. (1982). "The general equivalence of Granger and Sims causality," *Econometrica*, **50**, 569–82.

Chamberlain, G. (1984). "Panel Data," Ch. 22 in *Handbook of Econometrics*, Vol. II, eds. Z. Griliches and M. D. Intriligator, Amsterdam, North-Holland.

Chamberlain, G. (1986). "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, **32**, 189–218.

Chamberlain, G. (1987). "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, **34**, 305–34.

Chambers, J. M. (1977). *Computational Methods for Data Analysis*, New York, John Wiley & Sons.

Chesher, A. (1983). "The information matrix test: simplified calculation via a score test interpretation," *Economics Letters*, **13**, 45–48.

Chesher, A. (1984). "Testing for neglected heterogeneity," *Econometrica*, **52**, 865–72.

Chesher, A. (1989). "Hájek inequalities, measures of leverage and the size of heteroskedasticity robust tests," *Econometrica*, **57**, 971–77.

Chesher, A., and G. Austin (1991). "The finite-sample distributions of heteroskedasticity robust Wald statistics," *Journal of Econometrics*, **47**, 153–73.

Chesher, A., and M. Irish (1987). "Residual analysis in the grouped and censored normal linear model," *Journal of Econometrics*, **34**, 33–61.

Chesher A., and I. Jewitt (1987). "The bias of a heteroskedasticity consistent covariance matrix estimator," *Econometrica*, **55**, 1217–22.

Chesher, A., and S. Peters (1994). "Symmetry, regression design, and sampling distributions," *Econometric Theory*, **10**, 116–29.

Chesher, A., and R. Spady (1991). "Asymptotic expansions of the information matrix test statistic," *Econometrica*, **59**, 787–815.

Chipman, J. S. (1979). "Efficiency of least squares estimation of linear trend when residuals are autocorrelated," *Econometrica*, **47**, 115–28.

Chow, G. C. (1960). "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, **28**, 591–605.

Davidson, R., and J. G. MacKinnon (1985a). "The interpretation of test statistics," *Canadian Journal of Economics*, **18**, 38–57.

Davidson, R., and J. G. MacKinnon (1985b). "Heteroskedasticity-robust tests in regression directions," *Annales de l'INSÉÉ*, **59/60**, 183–218.

Davidson, R., and J. G. MacKinnon (1985c). "Testing linear and loglinear regressions against Box-Cox alternatives," *Canadian Journal of Economics*, **18**, 499–517.

Davidson, R., and J. G. MacKinnon (1987). "Implicit alternatives and the local power of test statistics," *Econometrica*, **55**, 1305–29.

Davidson, R., and J. G. MacKinnon (1988). "Double-length artificial regressions," *Oxford Bulletin of Economics and Statistics*, **50**, 203–17.

Davidson, R., and J. G. MacKinnon (1989). "Testing for consistency using artificial regressions," *Econometric Theory*, **5**, 363–84.

Davidson, R., and J. G. MacKinnon (1990). "Specification tests based on artificial regressions," *Journal of the American Statistical Association*, **85**, 220–27.

Davidson, R., and J. G. MacKinnon (1992a). "A new form of the information matrix test," *Econometrica*, **60**, 145–57.

Davidson, R., and J. G. MacKinnon (1992b). "Regression-based methods for using control variates in Monte Carlo experiments," *Journal of Econometrics*, **54**, 203–22.

Deaton, A. S. (1974). "The analysis of consumer demand in the United Kingdom, 1900-1970," *Econometrica*, **42**, 341–67.

Deaton, A. S. (1978). "Specification and testing in applied demand analysis," *Economic Journal*, **88**, 524–36.

Deaton, A. S., and J. Muellbauer (1980). *Economics and Consumer Behaviour*, Cambridge, Cambridge University Press.

DeJong, D. N., and C. H. Whiteman (1991). "The temporal stability of dividends and stock prices: evidence from the likelihood function," *American Economic Review*, **81**, 600–617.

Dent, W. (1977). "Computation of the exact likelihood function of an ARIMA process," *Journal of Statistical Computation and Simulation*, **5**, 193–206.

Dhrymes, P. J. (1971). *Distributed Lags: Problems of Estimation and Formulation*, San Francisco, Holden-Day.

Dhrymes, P. (1986). "Limited dependent variables," Ch. 27 in *Handbook of Econometrics*, Vol. III, eds. Z. Griliches and M. D. Intriligator, Amsterdam, North-Holland.

Dhrymes, P. J., R. Berner, and D. Cummins (1974). "A comparison of some limited information estimators for dynamic simultaneous equations models with autocorrelated errors," *Econometrica*, **42**, 311–32.

Dickey, D. A., W. R. Bell, and R. B. Miller (1986). "Unit roots in time series models: tests and implications," *The American Statistician*, **40**, 12–26.

Dickey, D. A., and W. A. Fuller (1979). "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, **74**, 427–31.

Leamer, E. E. (1987). "Errors in variables in linear systems," *Econometrica*, **55**, 893–909.

L'Ecuyer, P. (1988). "Efficient and portable combined random number generators," *Communications of the ACM*, **31**, 742–49 and 774.

Lee, L.-F. (1978). "Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables," *International Economic Review*, **19**, 415–33.

Lee, L.-F. (1981). "Simultaneous equations models with discrete and censored variables," in *Structural Analysis of Discrete Data with Econometric Applications*, eds. C. F. Manski and D. McFadden, Cambridge, Mass., MIT Press.

Lee, L.-F. (1982). "Some approaches to the correction of selectivity bias," *Review of Economic Studies*, **49**, 355–72.

Lee, L.-F. (1992). "Semiparametric nonlinear least-squares estimation of truncated regression models," *Econometric Theory*, **8**, 52–94.

Lee, L.-F., and G. S. Maddala (1985). "The common structure of tests for selectivity bias, serial correlation, heteroskedasticity and non-normality in the Tobit model," *International Economic Review*, **26**, 1–20.

Leech, D. (1975). "Testing the error specification in nonlinear regression," *Econometrica*, **43**, 719–25.

Lewis, P. A. W., and E. J. Orav (1989). *Simulation Methodology for Statisticians, Operations Analysts and Engineers*, Pacific Grove, Calif., Wadsworth and Brooks/Cole.

Lin, T.-F., and P. Schmidt (1984). "A test of the tobit specification against an alternative suggested by Cragg," *Review of Economics and Statistics*, **66**, 174–77.

Lindley, D. V. (1957). "A statistical paradox," *Biometrika*, **44**, 187–92.

Litterman, R. B. (1979). "Techniques of forecasting using vector autoregressions," Federal Reserve Bank of Minneapolis, Working Paper No. 15.

Litterman, R. B. (1986). "Forecasting with Bayesian vector autoregressions–five years of experience," *Journal of Business and Economic Statistics*, **4**, 25–38.

Litterman, R. B., and L. Weiss (1985). "Money, real interest rates, and output: A reinterpretation of postwar U. S. data," *Econometrica*, **53**, 129–56.

Ljung, G. M., and G. E. P. Box (1978). "On a measure of lack of fit in time-series models," *Biometrika*, **65**, 297–303.

Lovell, M. C. (1963). "Seasonal adjustment of economic time series and multiple regression analysis," *Journal of the American Statistical Association*, **58**, 993–1010.

Lukacs, E. (1975). *Stochastic Convergence*, Second edition, New York, Academic Press.

Maasoumi, E., and P. C. B. Phillips (1982). "On the behavior of inconsistent instrumental variable estimators," *Journal of Econometrics*, **19**, 183–201.

MacDonald, G. M., and J. G. MacKinnon (1985). "Convenient methods for estimation of linear regression models with MA(1) errors," *Canadian Journal of Economics*, **18**, 106–16.

MacKinnon, J. G. (1979). "Convenient singularities and maximum likelihood estimation," *Economics Letters*, **3**, 41–44.

Phillips, P. C. B. (1986). "Understanding spurious regressions in econometrics," *Journal of Econometrics*, **33**, 311–40.

Phillips, P. C. B. (1987). "Time series regression with a unit root," *Econometrica*, **55**, 277–301.

Phillips, P. C. B. (1991a). "Optimal inference in cointegrated systems," *Econometrica*, **59**, 283–306.

Phillips, P. C. B. (1991b). "To criticize the critics: an objective Bayesian analysis of stochastic trends," *Journal of Applied Econometrics*, **6**, 333–64.

Phillips, P. C. B. (1991c). "Bayesian routes and unit roots: de rebus prioribus semper est disputandum," *Journal of Applied Econometrics*, **6**, 435–73.

Phillips, P. C. B., and B. E. Hansen (1990). "Statistical inference in instrumental variables regression with $I(1)$ processes," *Review of Economic Studies*, **57**, 99–125.

Phillips, P. C. B., and S. Ouliaris (1990). "Asymptotic properties of residual based tests for cointegration," *Econometrica*, **58**, 165–93.

Phillips, P. C. B., and J. Y. Park (1988). "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, **56**, 1065–83.

Phillips, P. C. B., and P. Perron (1988). "Testing for a unit root in time series regression," *Biometrika*, **75**, 335–46.

Pitman, E. J. G. (1949). "Notes on non-parametric statistical inference," Columbia University, New York, mimeo.

Plosser, C. I. (1979a). "Short-term forecasting and seasonal adjustment," *Journal of the American Statistical Association*, **74**, 15–24.

Plosser, C. I. (1979b). "The analysis of seasonal economic models," *Journal of Econometrics*, **10**, 147–63.

Plosser, C. I., G. W. Schwert, and H. White (1982). "Differencing as a test of specification," *International Economic Review*, **23**, 535–52.

Poirier, D. J. (1978a). "The effect of the first observation in regression models with first-order autoregressive disturbances," *Applied Statistics*, **27**, 67–68.

Poirier, D. J. (1978b). "The use of the Box-Cox transformation in limited dependent variable models," *Journal of the American Statistical Association*, **73**, 284–87.

Poirier, D. J., and P. A. Ruud (1979). "A simple Lagrange Multiplier test for lognormal regression," *Economics Letters*, **4**, 251–55.

Pollak, R. A., and T. J. Wales (1969). "Estimation of the linear expenditure system," *Econometrica*, **37**, 611–28.

Pollak, R. A., and T. J. Wales (1978). "Estimation of complete demand systems from household budget data: the linear and quadratic expenditure systems," *American Economic Review*, **68**, 349–59.

Pollak, R. A., and T. J. Wales (1981). "Demographic variables in demand analysis," *Econometrica*, **49**, 1533–51.

Pollak, R. A., and T. J. Wales (1987). "Pooling international consumption data," *Review of Economics and Statistics*, **69**, 90–99.

Pollak, R. A., and T. J. Wales (1991). "The likelihood dominance criterion: a new approach to model selection," *Journal of Econometrics*, **47**, 227–42.