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

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

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:

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

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 $E(y_t \mid \Omega_t)$. There may be more than one such information set. Thus we might well have both

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

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 $y_t - E(y_t \mid \Omega_t)$ as the error term associated with the t^{th} 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

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

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

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,

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

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

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

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

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

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,

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

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

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

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,

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

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

$$\operatorname*{plim}_{n\to\infty}\left(\frac{a_n}{g(n)}\right) = 0.$$

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

$$\operatorname*{plim}_{n\to\infty}\left(\frac{a_n}{b_n}\right) = 1.$$

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

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

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

$$\begin{aligned}
O(n^p)O(n^q) &= O(n^{p+q});\\
o(n^p)o(n^q) &= o(n^{p+q});\\
O(n^p)o(n^q) &= o(n^{p+q}).
\end{aligned}$$

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 $\text{Var}(\hat{\alpha})$ 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 $\text{Var}(\hat{\alpha})$ 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 $Z\colon \mathbb{R}^{l+1} \to \mathbb{R}$,

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

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

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

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

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

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 $E(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{Z}) = E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$. 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

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

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

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

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 $\boldsymbol{C}\boldsymbol{X} = \mathbf{0}$, which is true because $\boldsymbol{M}_X$ annihilates $\boldsymbol{X}$. Further, we see that

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

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 $\boldsymbol{M}_0 \equiv \mathbf{I} - \boldsymbol{P}_0$:

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

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 $\boldsymbol{r}(\boldsymbol{\beta})$ 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 $\hat{\boldsymbol{r}}$ are **asymptotically nonstochastic**. By this we mean that

$$\hat{\boldsymbol{R}} = \boldsymbol{R}_0 + O(n^{-1/2}) \quad \text{and} \quad \hat{\boldsymbol{X}} = \boldsymbol{X}_0 + O(n^{-1/2}). \tag{5.69}$$

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 $\boldsymbol{R}$ and $\boldsymbol{X}$. This is not a matter of consequence for the $r \times k$ matrix $\boldsymbol{R}$, but it is for the $n \times k$ matrix $\boldsymbol{X}$. We have to be careful because in matrix products like $\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}$ 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 $\hat{\boldsymbol{r}} = O(n^{-1/2})$ to rewrite (5.66) as

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

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

$$\begin{aligned}
n^{-1}(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})_{ij} &= n^{-1}\sum_{t=1}^{n}\hat{X}_{ti}\hat{X}_{tj} \\
&= n^{-1}\sum_{t=1}^{n}\big(X^0_{ti} + O(n^{-1/2})\big)\big(X^0_{tj} + O(n^{-1/2})\big) \\
&= n^{-1}\sum_{t=1}^{n}X^0_{ti}X^0_{tj} + O(n^{-1/2}) \\
&= n^{-1}(\boldsymbol{X}_0^\top\boldsymbol{X}_0)_{ij} + O(n^{-1/2}),
\end{aligned}$$

where X^0_{ti} denotes the ti^{th} element of $\boldsymbol{X}_0$. 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 $n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0$ itself is $O(1)$.

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

$$n^{1/2}\hat{\boldsymbol{r}} = \boldsymbol{R}_0\big(n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0\big)^{-1}n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u} + o(1). \tag{5.71}$$

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

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

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

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

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 $\chi^2(r)$:

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

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

$$\begin{aligned} E(\boldsymbol{\nu}\boldsymbol{\nu}^\top) &= E\big(n^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{X}_2\big) = n^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 (\sigma_0^2 \mathbf{I}) \boldsymbol{M}_1 \boldsymbol{X}_2 \\ &= n^{-1} \sigma_0^2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2) \equiv \sigma_0^2 \boldsymbol{V}. \end{aligned}$$

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

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

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 m^{th} 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.

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

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

The difference between (7.45) and (7.46) is

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

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

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

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

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

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

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

where $\check{\boldsymbol{b}}$ 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 $(\tilde{\boldsymbol{\beta}} - \check{\boldsymbol{\beta}})$ is asymptotically equal to the estimate $\check{\boldsymbol{b}}$ from the GNR (7.38). Thus

$$\boldsymbol{P}_W \left(\boldsymbol{y} - \boldsymbol{x}(\check{\boldsymbol{\beta}}) - \check{\boldsymbol{X}}\check{\boldsymbol{b}} \right) \stackrel{a}{=} \boldsymbol{P}_W \boldsymbol{y} - \boldsymbol{P}_W \boldsymbol{x}(\check{\boldsymbol{\beta}}) - \boldsymbol{P}_W \check{\boldsymbol{X}} (\tilde{\boldsymbol{\beta}} - \check{\boldsymbol{\beta}}). \tag{7.52}$$

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

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

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

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

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 $\check{\boldsymbol{\beta}}$ 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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \boldsymbol{R}^\top(\bar{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} &= \mathbf{0} \\ \boldsymbol{R}(\ddot{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \mathbf{0}, \end{aligned}$$

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

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

If we multiply $\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ by n^{-1}, $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ by $n^{1/2}$, $\boldsymbol{g}(\boldsymbol{\theta}_0)$ 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,

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

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,

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

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,

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

It follows that an appropriate test statistic is

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

where $\hat{\mathcal{J}}$ denotes any consistent estimate of $\mathcal{J}(\boldsymbol{\theta}_0)$ based on the unrestricted estimates $\hat{\boldsymbol{\theta}}$. Different variants of the Wald test will use different estimates of $\mathcal{J}(\boldsymbol{\theta}_0)$. 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.

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

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

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 t^{th} 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 $\tilde{\boldsymbol{\beta}}$ 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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}&\frac{1}{mn-k}\sum_{t=1}^{n}(\boldsymbol{Y}_t-\hat{\boldsymbol{\xi}}_t)\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^\top(\boldsymbol{Y}_t-\hat{\boldsymbol{\xi}}_t)^\top\\ =&\frac{1}{mn-k}\sum_{t=1}^{n}(\boldsymbol{Y}_t-\hat{\boldsymbol{\xi}}_t)\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}_t-\hat{\boldsymbol{\xi}}_t)^\top=\frac{mn}{mn-k}.\end{aligned}\tag{9.70}$$

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

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

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

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

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

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 $\hat{\boldsymbol{X}}^*$ 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

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

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.

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 $\tilde{\boldsymbol{u}}$ 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, $\tilde{\boldsymbol{u}} \stackrel{a}{=} \boldsymbol{M}_0\boldsymbol{u}$, 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).

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 $\boldsymbol{Y}_{t-1}$ (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, $\boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{z}(\boldsymbol{\gamma}) = \boldsymbol{Z}\boldsymbol{\gamma}$. 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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \alpha \boldsymbol{P}_Z \boldsymbol{y} + \text{residuals}, \tag{11.16}$$

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

$$\boldsymbol{M}_X \boldsymbol{y} = \alpha \boldsymbol{M}_X \boldsymbol{P}_Z \boldsymbol{y} + \text{residuals}. \tag{11.17}$$

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

$$\frac{\boldsymbol{y}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{y}}{\acute{s}(\boldsymbol{y}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{P}_Z \boldsymbol{y})^{1/2}}. \tag{11.18}$$

First of all, notice that when only one column of $\boldsymbol{Z}$, say $\boldsymbol{Z}_1$, does not belong to $\mathcal{S}(\boldsymbol{X})$, it must be the case that

$$\mathcal{S}(\boldsymbol{X}, \boldsymbol{P}_Z \boldsymbol{y}) = \mathcal{S}(\boldsymbol{X}, \boldsymbol{Z}) = \mathcal{S}(\boldsymbol{X}, \boldsymbol{Z}_1).$$

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \delta \boldsymbol{Z}_1 + \text{residuals}. \tag{11.19}$$

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

$$\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{u}. \tag{11.20}$$

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

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

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

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\big(\boldsymbol{u}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{u}\big) &= \operatorname*{plim}_{n\to\infty}\big(\boldsymbol{u}^\top \boldsymbol{P}_Z \boldsymbol{u} - \boldsymbol{u}^\top \boldsymbol{P}_Z \boldsymbol{P}_X \boldsymbol{u}\big) \\ &= \sigma_1^2 k_2 - \sigma_1^2 \lim_{n\to\infty}\big(\operatorname{Tr}(\boldsymbol{P}_Z \boldsymbol{P}_X)\big),\end{aligned}$$

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

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

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

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

This regression yields what is called the $\boldsymbol{J_A}$ **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 $\boldsymbol{P_X y}$ contains only the part of $\boldsymbol{y}$ that lies in $\mathcal{S}(\boldsymbol{X})$. It must therefore be independent of $\boldsymbol{M_X y}$, which is what the residuals from (11.22) would be if $\alpha = 0$. Therefore, we can treat $\boldsymbol{P_Z P_X y}$ (or any other regressor that depends on $\boldsymbol{y}$ only through $\boldsymbol{P_X y}$) as if it were a fixed regressor.[4] The $\boldsymbol{P_A}$ **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 $\boldsymbol{u}^\top \boldsymbol{M_X P_Z u}$, subtract it from $\boldsymbol{y}^\top \boldsymbol{M_X P_Z y}$, 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,

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

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.

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

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

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

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

where $\boldsymbol{P}_{M_XAX^*}$ is the matrix that projects onto $\mathcal{S}(\boldsymbol{M}_X\boldsymbol{A}\boldsymbol{X}^*)$, and $\boldsymbol{M}_{X,\,M_XAX^*}$ 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, $\check{\boldsymbol{\beta}}$ might be the OLS estimator for $\boldsymbol{\beta}$ in the model

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

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

$$\begin{aligned}\Lambda_{21} = \frac{\rho_0^2}{\sigma_0^2} &\operatorname*{plim}_{n\to\infty}\Big(\tfrac{1}{n}\boldsymbol{u}_{-1}{}^{\top}\boldsymbol{M}_X(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0+\boldsymbol{u}_{-1})\Big)\\ &\times \operatorname*{plim}_{n\to\infty}\Big(\tfrac{1}{n}(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0+\boldsymbol{u}_{-1})^{\top}\boldsymbol{M}_X(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0+\boldsymbol{u}_{-1})\Big)^{-1}\\ &\times \operatorname*{plim}_{n\to\infty}\Big(\tfrac{1}{n}(\boldsymbol{X}_{-1}\boldsymbol{\beta}_0+\boldsymbol{u}_{-1})^{\top}\boldsymbol{M}_X\boldsymbol{u}_{-1}\Big).\end{aligned}$$

This simplifies to

$$\begin{aligned}&\frac{\rho_0^2}{\sigma_0^2}\sigma_0^2\Big(\sigma_0^2+\operatorname{plim}\tfrac{1}{n}\big\|\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0\big\|^2\Big)^{-1}\sigma_0^2\\ &=\rho_0^2\Big(1+\sigma_0^{-2}\operatorname{plim}\tfrac{1}{n}\big\|\boldsymbol{M}_X\boldsymbol{X}_{-1}\boldsymbol{\beta}_0\big\|^2\Big)^{-1}.\end{aligned}$$

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

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

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

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\alpha}+\gamma\boldsymbol{z}+\text{ residuals}\quad\text{and}\\ \boldsymbol{z} &= \boldsymbol{X}\boldsymbol{\beta}+\delta\boldsymbol{y}+\text{ residuals},\end{aligned}$$

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

$$\begin{aligned}\boldsymbol{M}_X\boldsymbol{y} &= \gamma\boldsymbol{M}_X\boldsymbol{z}+\text{ residuals}\quad\text{and}\\ \boldsymbol{M}_X\boldsymbol{z} &= \delta\boldsymbol{M}_X\boldsymbol{y}+\text{ residuals},\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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.

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 $\boldsymbol{G}(\boldsymbol{\theta})$ 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

$$\boldsymbol{\iota} = \boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{c} + \text{residuals}. \tag{13.81}$$

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 $\boldsymbol{g}(\boldsymbol{\theta}) \equiv \boldsymbol{G}^\top(\boldsymbol{\theta})\boldsymbol{\iota}$. The matrix of sums of squares and cross-products of the regressors, $\boldsymbol{G}^\top(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})$, when divided by n, consistently estimates the information matrix $\mathcal{I}(\boldsymbol{\theta})$. 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 $\hat{\boldsymbol{G}} \equiv \boldsymbol{G}(\hat{\boldsymbol{\theta}})$ and the artificial parameter estimates, which we will denote by $\hat{\boldsymbol{c}}$, are identically zero:

$$\hat{\boldsymbol{c}} = \big(\hat{\boldsymbol{G}}^\top\hat{\boldsymbol{G}}\big)^{-1}\hat{\boldsymbol{G}}^\top\boldsymbol{\iota} = \big(\hat{\boldsymbol{G}}^\top\hat{\boldsymbol{G}}\big)^{-1}\hat{\boldsymbol{g}} = \boldsymbol{0}.$$

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.

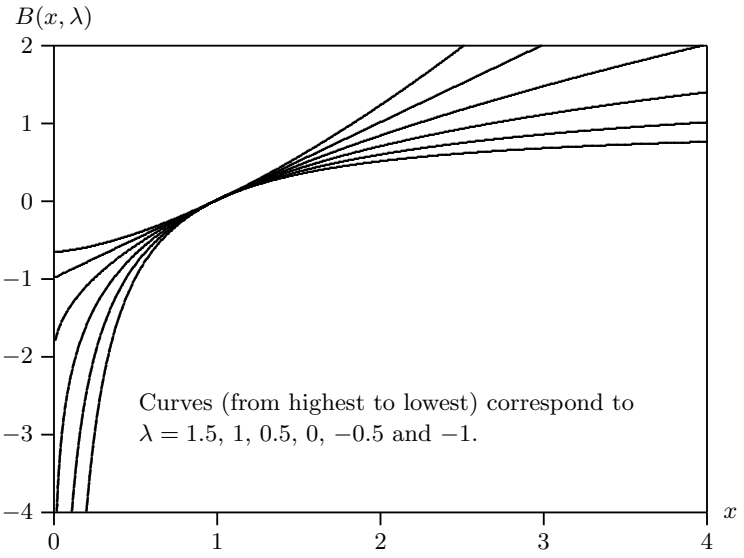

Figure 14.1 Box-Cox transformations for various values of λ

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 $\mathcal{I}(\boldsymbol{\theta})$ satisfies the equality

$$\mathcal{I}(\boldsymbol{\theta}) = \operatorname*{plim}_{n\to\infty}\Big(\tfrac{1}{n}\big(\boldsymbol{F}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) + \boldsymbol{K}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta})\big)\Big) \tag{14.20}$$

and so can be consistently estimated by

$$\tfrac{1}{n}\big(\boldsymbol{F}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\boldsymbol{F}(\boldsymbol{y},\ddot{\boldsymbol{\theta}}) + \boldsymbol{K}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\boldsymbol{K}(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\big), \tag{14.21}$$

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

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{\iota} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix} \boldsymbol{b} + \text{residuals}. \tag{14.22}$$

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 t^{th} 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 $-\frac{1}{2}f_t^2$ 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

$$\boldsymbol{R}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta}) = \rho(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{y},\boldsymbol{\theta}) \quad \text{and} \tag{14.23}$$

$$\operatorname*{plim}_{n\to\infty,\boldsymbol{\theta}}\Big(\tfrac{1}{n}\boldsymbol{R}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\boldsymbol{R}(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\Big) = \rho(\boldsymbol{\theta})\,\mathcal{I}(\boldsymbol{\theta}), \tag{14.24}$$

[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 $\operatorname{plim}_{\boldsymbol{\theta}}$ 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

$$\rho(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n^*}\boldsymbol{r}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta})\Big).$$

Because $\rho(\boldsymbol{\theta})$ is equal to unity for both the OPG regression and the DLR, those two artificial regressions satisfy the simpler conditions

$$\boldsymbol{R}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{y},\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{y},\boldsymbol{\theta}) \quad \text{and} \tag{14.25}$$

$$\operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{R}^\top(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\boldsymbol{R}(\boldsymbol{y},\ddot{\boldsymbol{\theta}})\Big) = \mathcal{I}(\boldsymbol{\theta}), \tag{14.26}$$

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.

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

$$H_C'\colon\ (1-\alpha)\left(\frac{y_t - x_t(\boldsymbol{\beta})}{\sigma_1}\right) + \alpha\left(\frac{\log y_t - z_t(\hat{\boldsymbol{\gamma}})}{\hat{\sigma}_2}\right) = \varepsilon_t.$$

It is straightforward to test H_1 against H_C' by means of the DLR:

$$\begin{bmatrix} \dfrac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{X}}_t & \dfrac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} & \hat{z}_t - \log y_t \\ \mathbf{0} & -1 & \hat{\sigma}_1/y_t \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ s \\ a \end{bmatrix} + \text{residuals}, \qquad (14.45)$$

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

$$H_C''\colon\ (1-\alpha)\left(\frac{y_t - x_t(\hat{\boldsymbol{\beta}})}{\hat{\sigma}_1}\right) + \alpha\left(\frac{\log y_t - z_t(\boldsymbol{\gamma})}{\sigma_2}\right) = \varepsilon_t.$$

It is then straightforward to test H_2 against H_C'' by means of the DLR

$$\begin{bmatrix} \dfrac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{Z}}_t & \dfrac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} & \hat{x}_t - y_t \\ \mathbf{0} & -1 & \hat{\sigma}_2 y_t \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ s \\ a \end{bmatrix} + \text{residuals}. \qquad (14.46)$$

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

can be written as

$$\ell(\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^J) = \sum_{j=1}^{J} \sum_{y_t=j} \boldsymbol{X}_t \boldsymbol{\beta}^j - \sum_{t=1}^{n} \log\Bigl(1 + \sum_{j=1}^{J} \exp(\boldsymbol{X}_t \boldsymbol{\beta}^j)\Bigr).$$

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

$$\frac{\Pr(y_t = l)}{\Pr(y_t = j)} = \frac{\exp(\boldsymbol{X}_t \boldsymbol{\beta}^l)}{\exp(\boldsymbol{X}_t \boldsymbol{\beta}^j)} = \exp\bigl(\boldsymbol{X}_t(\boldsymbol{\beta}^l - \boldsymbol{\beta}^j)\bigr) \tag{15.38}$$

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 i^{th} 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 i^{th} 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

$$\Pr(y_i = l) = \frac{\exp(\boldsymbol{W}_{il} \boldsymbol{\beta})}{\sum_{j=1}^{J} \exp(\boldsymbol{W}_{ij} \boldsymbol{\beta})}. \tag{15.39}$$

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.

they are related to y_t^* and z_t^* as follows:

$$y_t = y_t^* \text{ if } z_t^* > 0; \;\; y_t = 0 \text{ otherwise;}$$
$$z_t = 1 \text{ if } z_t^* > 0; \;\; z_t = 0 \text{ otherwise.}$$

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

$$\sum_{z_t=0} \log\big(\Pr(z_t = 0)\big) + \sum_{z_t=1} \log\big(\Pr(z_t = 1) f(y_t^* \mid z_t = 1)\big), \tag{15.54}$$

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

$$\begin{aligned} &\sum_{z_t=0} \log\big(\Phi(-\boldsymbol{W}_t\boldsymbol{\gamma})\big) + \sum_{z_t=1} \log\Big(\frac{1}{\sigma}\phi\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)\Big) \\ &+ \sum_{z_t=1} \log\left(\Phi\left(\frac{\boldsymbol{W}_t\boldsymbol{\gamma} + \rho\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)}{(1 - \rho^2)^{1/2}}\right)\right). \end{aligned} \tag{15.55}$$

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).

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

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

$$\begin{aligned} \hat{\boldsymbol{M}}_{\boldsymbol{G}}\boldsymbol{\iota} &= b(\hat{\boldsymbol{M}}_{\boldsymbol{G}}\hat{\boldsymbol{m}}) + \text{residuals} \quad \text{and} \\ \hat{\boldsymbol{M}}_{\boldsymbol{G}}\hat{\boldsymbol{m}} &= b^*(\hat{\boldsymbol{M}}_{\boldsymbol{G}}\boldsymbol{\iota}) + \text{residuals}. \end{aligned} \tag{16.64}$$

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

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

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\Big(\operatorname*{plim}_{n\to\infty} \mathcal{H}_0\Big)^{-1} n^{1/2}\boldsymbol{g}(\boldsymbol{\theta}_0). \tag{17.22}$$

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 $N\big(\mathbf{0}, \mathcal{H}(\mu)^{-1}\boldsymbol{V}(\mu)\mathcal{H}(\mu)^{-1}\big)$.

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

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 $\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. 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

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 $\boldsymbol{r}(\acute{\boldsymbol{\theta}}) = \mathbf{0}$. 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

$$\begin{aligned} C(\alpha) = &\tfrac{1}{n}\boldsymbol{\iota}^\top \acute{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}^{-1}\acute{\boldsymbol{D}}\big(\acute{\boldsymbol{D}}^\top\hat{\boldsymbol{\Phi}}^{-1}\acute{\boldsymbol{D}}\big)^{-1}\acute{\boldsymbol{D}}^\top\hat{\boldsymbol{\Phi}}^{-1}\acute{\boldsymbol{F}}^\top\boldsymbol{\iota} \\ &- \tfrac{1}{n}\boldsymbol{\iota}^\top \acute{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}^{-1}\acute{\boldsymbol{D}}_1\big(\acute{\boldsymbol{D}}_1^\top\hat{\boldsymbol{\Phi}}^{-1}\acute{\boldsymbol{D}}_1\big)^{-1}\acute{\boldsymbol{D}}_1^\top\hat{\boldsymbol{\Phi}}^{-1}\acute{\boldsymbol{F}}^\top\boldsymbol{\iota}. \end{aligned} \tag{17.75}$$

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

$$-\frac{ng}{2}\big(\log(2\pi)+1\big)-\frac{n}{2}\log\Big|\frac{1}{n}\big(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\Pi}}\big)^{\top}\big(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\Pi}}\big)\Big|, \tag{18.33}$$

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

$$\big|(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Pi})^{\top}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Pi})\big|. \tag{18.34}$$

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

$$\begin{aligned}
&\big|(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\Pi}}-\boldsymbol{X}\boldsymbol{A})^{\top}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\Pi}}-\boldsymbol{X}\boldsymbol{A})\big| \\
&= \big|(\boldsymbol{M}_X\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A})^{\top}(\boldsymbol{M}_X\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A})\big| \\
&= \big|\boldsymbol{Y}^{\top}\boldsymbol{M}_X\boldsymbol{Y}+\boldsymbol{A}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{A}\big|.
\end{aligned} \tag{18.35}$$

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 $|\boldsymbol{Y}^\top \boldsymbol{M}_X \boldsymbol{Y}|$ for all $\boldsymbol{A} \neq \mathbf{0}$. This implies that $\hat{\boldsymbol{\Pi}}$ 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**

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}\boldsymbol{\eta}_1 + \boldsymbol{e}_1 \\ \boldsymbol{y}_2 &= \boldsymbol{X}\boldsymbol{\eta}_2 + \boldsymbol{y}_1\alpha_1 + \boldsymbol{e}_2 \\ \boldsymbol{y}_3 &= \boldsymbol{X}\boldsymbol{\eta}_3 + [\boldsymbol{y}_1 \quad \boldsymbol{y}_2]\boldsymbol{\alpha}_2 + \boldsymbol{e}_3 \\ \boldsymbol{y}_4 &= \boldsymbol{X}\boldsymbol{\eta}_4 + [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \boldsymbol{y}_3]\boldsymbol{\alpha}_3 + \boldsymbol{e}_4, \end{aligned} \tag{18.36}$$

and so on, where $\boldsymbol{y}_i$ denotes the i^{th} column of $\boldsymbol{Y}$. 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: $\hat{\boldsymbol{e}}_2$ will be orthogonal to $\hat{\boldsymbol{e}}_1$, $\hat{\boldsymbol{e}}_3$ will be orthogonal to $\hat{\boldsymbol{e}}_2$ and $\hat{\boldsymbol{e}}_1$, and so on. According to the URF, all the $\boldsymbol{y}_i$'s are linear combinations of the columns of $\boldsymbol{X}$ 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 $\boldsymbol{e}_i$ 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 $\boldsymbol{\Pi}$ and $(g^2+g)/2$ in the covariance matrix $\boldsymbol{\Omega}$, 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 $\boldsymbol{V}$ 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.

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 i^{th} 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

$$\boldsymbol{W}\left(\boldsymbol{W}^{\top}\acute{\boldsymbol{\Omega}}_{ij}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^{\top},$$

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).

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 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, 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 $\boldsymbol{I(1)}$. 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

$$\begin{aligned} y_t &= \gamma_0 + \gamma_1 t + v_t; \;\; v_t = \alpha v_{t-1} + u_t \\ &= \gamma_0 + \gamma_1 t + \alpha\big(y_{t-1} - \gamma_0 - \gamma_1(t-1)\big) + u_t, \end{aligned} \tag{20.03}$$

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

$$y_t = \beta_0 + \beta_1 t + \alpha y_{t-1} + u_t, \tag{20.04}$$

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

$$\Delta y_t = \beta_0 + \beta_1 t + (\alpha - 1)y_{t-1} + u_t, \tag{20.05}$$

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

Serial correlation is not the only complication that one is likely to encounter 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 performed 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,

$$[\boldsymbol{y}_1 \quad \boldsymbol{y}_2]\boldsymbol{\eta} \equiv \boldsymbol{y}_1 - \eta_2 \boldsymbol{y}_2 = \mathbf{0}. \tag{20.20}$$

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 t^{th} 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 $I(0)$.

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 i^{th} diagonal element. The i^{th} column of $\boldsymbol{A}\boldsymbol{U}$ is $\boldsymbol{A}\boldsymbol{x}_i$, and the i^{th} 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 $\boldsymbol{A}$.

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.