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:

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

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 $a_n = o_p\bigl(g(n)\bigr)$ means that

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

Similarly, the notation $a_n = O_p\bigl(g(n)\bigr)$ 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

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

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

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

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

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

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

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

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

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

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

It follows that an appropriate test statistic is

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

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

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

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

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

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

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

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 $\boldsymbol{X\beta} + \boldsymbol{u}$. Since $\boldsymbol{M}_X \boldsymbol{X\beta} = \boldsymbol{0}$, that numerator becomes

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

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

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

This simplifies to

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

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.

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

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

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

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

and so can be consistently estimated by

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

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

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

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

[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 $\text{plim}_{\boldsymbol{\theta}}$ indicates, as usual, that the probability limit is being taken under the DGP characterized by the parameter vector $\boldsymbol{\theta}$, and $\rho(\boldsymbol{\theta})$ is a scalar defined as

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

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

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

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

as well as the original conditions (14.23) and (14.24). However, these simpler conditions are not satisfied by the GNR and are thus evidently too simple in general.

It is now easy to see that the DLR (14.21) satisfies conditions (14.25) and (14.26). For the first of these, simple calculation shows that

$$\begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{f}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{\iota} \end{bmatrix} = -\boldsymbol{F}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{y},\boldsymbol{\theta}) + \boldsymbol{K}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{\iota},$$

which by (14.19) is equal to the gradient $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{\theta})$. For the second, we see that

$$\begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}) \end{bmatrix} = \boldsymbol{F}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta}) + \boldsymbol{K}^\top(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta}).$$

The right-hand side here is just the expression that appears in the fundamental result (14.20). Hence it is clear that the DLR must satisfy (14.26). All this discussion assumes, of course, that the matrices $\boldsymbol{F}(\boldsymbol{y},\boldsymbol{\theta})$ and $\boldsymbol{K}(\boldsymbol{y},\boldsymbol{\theta})$ satisfy appropriate regularity conditions, which may not always be easy to verify in practice; see Davidson and MacKinnon (1984a).

The DLR can be used in all the same ways that the GNR and the OPG regression can be used. In particular, it can be used

(i) to verify that the first-order conditions for a maximum of the log-likelihood function are satisfied sufficiently accurately,

(ii) to calculate estimated covariance matrices,

(iii) to calculate test statistics,

(iv) to calculate one-step efficient estimates, and

(v) as a key part of procedures for finding ML estimates.

can be written as

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

This function is a sum of contributions from each observation. Each contribution has two terms: The first is $\boldsymbol{X}_t\boldsymbol{\beta}^j$, where the index j is that for which $y_t = j$ (or zero if $j = 0$), and the second is minus the logarithm of the denominator that appears in (15.35) and (15.36).

One important property of the multinomial logit model is that

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

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.

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).

The estimator (17.63) was proposed by Hansen (1982) and White and Domowitz (1984), and was used in some of the earlier published work that employed GMM estimation, such as Hansen and Singleton (1982). From the point of view of theory, it is necessary to let the truncation parameter p, usually referred to as the **lag truncation parameter**, go to infinity at some suitable rate. A typical rate would be $n^{1/4}$, in which case $p = o(n^{1/4})$. This ensures that, for large enough n, all the nonzero $\boldsymbol{\Gamma}(j)$'s are estimated consistently. Unfortunately, this type of result is not of much use in practice, where one typically faces a given, finite n. We will return to this point a little later, and for the meantime suppose simply that we have somehow selected an appropriate value for p.

A much more serious difficulty associated with (17.63) is that, in finite samples, it need not be positive definite or even positive semidefinite. If one is unlucky enough to be working with a data set that yields a nondefinite $\hat{\boldsymbol{\Phi}}$, then (17.63) is unusable. There are numerous ways out of this difficulty. The most widely used was suggested by Newey and West (1987a). It is simply to multiply the $\hat{\boldsymbol{\Gamma}}(j)$'s by a sequence of weights that decrease as $|j|$ increases. Specifically, the estimator that they propose is

$$\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \left(1 - \frac{j}{p+1}\right)\left(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}(j)^{\top}\right). \tag{17.64}$$

It can be seen that the weights $1 - j/(p+1)$ decrease linearly with j from a value of 1 for $\hat{\boldsymbol{\Gamma}}(0)$ by steps of $1/(p+1)$ down to a value of $1/(p+1)$ for $|j| = p$. The use of such a set of weights is clearly compatible with the idea that the impact of the autocovariance of order j diminishes with $|j|$.

We will not attempt even to sketch a proof of the consistency of the Newey-West or similar estimators. We have alluded to the sort of regularity conditions needed for consistency to hold: Basically, the autocovariance matrices of the empirical moments must tend to zero quickly enough as p increases. It would also go well beyond the scope of this book to provide a theoretical justification for the Newey-West estimator. It rests on considerations of the so-called "frequency domain representation" of the $\boldsymbol{F}_t$'s and also of a number of notions associated with nonparametric estimation procedures. Interested readers are referred to Andrews (1991b) for a rather complete treatment of many of the issues. This paper suggests some alternatives to the Newey-West estimator and shows that in some circumstances they are preferable. However, the performance of the Newey-West estimator is never greatly inferior to that of the alternatives. Consequently, its simplicity is much in its favor.

Let us now return to the linear IV model with empirical moments given by $\boldsymbol{W}^{\top}(\boldsymbol{y} - \boldsymbol{X\beta})$. In order to be able to use (17.64), we suppose that the true error terms $u_t \equiv y_t - \boldsymbol{X}_t\boldsymbol{\beta}_0$ satisfy an appropriate mixing condition. Then the sample autocovariance matrices $\hat{\boldsymbol{\Gamma}}(j)$ for $j = 0, \ldots, p$, for some given p, are calculated as follows. A preliminary consistent estimate of $\boldsymbol{\beta}$ is first obtained 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).

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 $\boldsymbol{A} \neq \boldsymbol{0}$. 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.

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 conthat 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 $\boldsymbol{A}$.

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.