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 $a_n = o_p\big(g(n)\big)$ 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

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

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 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.

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).

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 $\boldsymbol{A}$.

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.