5

LARGE-SAMPLE PROPERTIES OF THE LEAST SQUARES AND INSTRUMENTAL VARIABLES ESTIMATORS

5.1 INTRODUCTION

The discussion thus far has concerned **finite-sample properties** of the least squares estimator. We derived its exact mean and variance and the precise distribution of the estimator and several test statistics under the assumptions of normally distributed disturbances and independent observations. These results are independent of the sample size. But the classical regression model with normally distributed disturbances and independent observations is a special case that does not include many of the most common applications, such as panel data and most time series models. This chapter will generalize the classical regression model by relaxing these two important assumptions.[1]

The linear model is one of relatively few settings in which any definite statements can be made about the exact finite sample properties of any estimator. In most cases, the only known properties of the estimators are those that apply to large samples. We can only approximate finite-sample behavior by using what we know about large-sample properties. This chapter will examine the **asymptotic properties** of the parameter estimators in the classical regression model. In addition to the least squares estimator, this chapter will also introduce an alternative technique, the method of instrumental variables. In this case, only the large sample properties are known.

5.2 ASYMPTOTIC PROPERTIES OF THE LEAST SQUARES ESTIMATOR

Using only assumptions A1 through A4 of the classical model (as listed in Table 4.1), we have established that the least squares estimators of the unknown parameters, $\boldsymbol{\beta}$ and σ^2, have the **exact, finite-sample properties** listed in Table 4.3. For this basic model, it is straightforward to derive the large-sample properties of the least squares estimator. The normality assumption, A6, becomes inessential at this point, and will be discarded save for brief discussions of maximum likelihood estimation in Chapters 10 and 17. This section will consider various forms of Assumption A5, the data generating mechanism.

[1] Most of this discussion will use our earlier results on asymptotic distributions. It may be helpful to review Appendix D before proceeding.

5.2.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR OF $\boldsymbol{\beta}$

To begin, we leave the data generating mechanism for $\mathbf{X}$ unspecified—$\mathbf{X}$ may be any mixture of constants and random variables generated independently of the process that generates $\boldsymbol{\varepsilon}$. We do make two crucial assumptions. The first is a modification of Assumption A5 in Table 4.1;

A5a. $(\mathbf{x}_i, \varepsilon_I)$ $i = 1, \ldots, n$ is a sequence of *independent* observations.

The second concerns the behavior of the data in large samples;

$$\operatorname*{plim}_{n\to\infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}, \quad \text{a positive definite matrix.} \tag{5-1}$$

[We will return to (5-1) shortly.] The least squares estimator may be written

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right). \tag{5-2}$$

If $\mathbf{Q}^{-1}$ exists, then

$$\operatorname{plim} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)$$

because the inverse is a continuous function of the original matrix. (We have invoked Theorem D.14.) We require the probability limit of the last term. Let

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \bar{\mathbf{w}}. \tag{5-3}$$

Then

$$\operatorname{plim} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \operatorname{plim} \bar{\mathbf{w}}.$$

From the exogeneity Assumption A3, we have $E[\mathbf{w}_i] = E_{\mathbf{x}}[E[\mathbf{w}_i \mid \mathbf{x}_i]] = E_{\mathbf{x}}[\mathbf{x}_i E[\varepsilon_i \mid \mathbf{x}_i]]$ $= \mathbf{0}$, so the exact expectation is $E[\bar{\mathbf{w}}] = \mathbf{0}$. For any element in $\mathbf{x}_i$ that is nonstochastic, the zero expectations follow from the marginal distribution of ε_i. We now consider the variance. By (B-70), $\operatorname{Var}[\bar{\mathbf{w}}] = E[\operatorname{Var}[\bar{\mathbf{w}} \mid \mathbf{X}]] + \operatorname{Var}[E[\bar{\mathbf{w}} \mid \mathbf{X}]]$. The second term is zero because $E[\varepsilon_i \mid \mathbf{x}_i] = 0$. To obtain the first, we use $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2\mathbf{I}$, so

$$\operatorname{Var}[\bar{\mathbf{w}} \mid \mathbf{X}] = E[\bar{\mathbf{w}}\bar{\mathbf{w}}' \mid \mathbf{X}] = \frac{1}{n}\mathbf{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}]\mathbf{X}\frac{1}{n} = \left(\frac{\sigma^2}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right).$$

Therefore,

$$\operatorname{Var}[\bar{\mathbf{w}}] = \left(\frac{\sigma^2}{n}\right)E\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right).$$

The variance will collapse to zero if the expectation in parentheses is (or converges to) a constant matrix, so that the leading scalar will dominate the product as n increases. Assumption (5-1) should be sufficient. (Theoretically, the expectation could diverge while the probability limit does not, but this case would not be relevant for practical purposes.) It then follows that

$$\lim_{n\to\infty} \operatorname{Var}[\bar{\mathbf{w}}] = 0 \cdot \mathbf{Q} = \mathbf{0}.$$

Since the mean of $\bar{\mathbf{w}}$ is identically zero and its variance converges to zero, $\bar{\mathbf{w}}$ **converges in mean square to zero,** so plim $\bar{\mathbf{w}} = \mathbf{0}$. Therefore,

$$\operatorname{plim} \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} = \mathbf{0}, \tag{5-4}$$

so

$$\operatorname{plim} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}. \tag{5-5}$$

This result establishes that under Assumptions A1–A4 and the additional assumption (5-1), $\mathbf{b}$ is a consistent estimator of $\boldsymbol{\beta}$ in the classical regression model.

Time-series settings that involve time trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about $\mathbf{X}$ that is broad enough to include most of these is the **Grenander conditions** listed in Table 5.1.[2] The conditions ensure that the data matrix is "well behaved" in large samples. The assumptions are very weak and is likely to be satisfied by almost any data set encountered in practice.[3]

5.2.2 ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

To derive the asymptotic distribution of the least squares estimator, we shall use the results of Section D.3. We will make use of some basic central limit theorems, so in addition to Assumption A3 (uncorrelatedness), we will assume that the observations are *independent.* It follows from (5-2) that

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\boldsymbol{\varepsilon}. \tag{5-6}$$

Since the inverse matrix is a continuous function of the original matrix, $\operatorname{plim}(\mathbf{X}'\mathbf{X}/n)^{-1} = \mathbf{Q}^{-1}$. Therefore, if the limiting distribution of the random vector in (5-6) exists, then that limiting distribution is the same as that of

$$\left[\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\right] \left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1} \left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\boldsymbol{\varepsilon}. \tag{5-7}$$

Thus, we must establish the limiting distribution of

$$\left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}(\bar{\mathbf{w}} - E[\bar{\mathbf{w}}]), \tag{5-8}$$

where $E[\bar{\mathbf{w}}] = \mathbf{0}$. [See (5-3).] We can use the multivariate Lindberg–Feller version of the central limit theorem (D.19.A) to obtain the limiting distribution of $\sqrt{n}\bar{\mathbf{w}}$.[4] Using that formulation, $\bar{\mathbf{w}}$ is the average of n independent random vectors $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$, with means $\mathbf{0}$ and variances

$$\operatorname{Var}[\mathbf{x}_i\varepsilon_i] = \sigma^2 E[\mathbf{x}_i\mathbf{x}_i'] = \sigma^2\mathbf{Q}_i. \tag{5-9}$$

[2]Judge et al. (1985, p. 162).

[3]White (2001) continues this line of analysis.

[4]Note that the Lindberg–Levy variant does not apply because $\operatorname{Var}[\mathbf{w}_i]$ is not necessarily constant.

TABLE 5.1 Grenander Conditions for Well Behaved Data

G1. For each column of $\mathbf{X}$, $\mathbf{x}_k$, if $d_{nk}^2 = \mathbf{x}_k'\mathbf{x}_k$, then $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Hence, $\mathbf{x}_k$ does not degenerate to a sequence of zeros. Sums of squares will continue to grow as the sample size increases. No variable will degenerate to a sequence of zeros.

G2. $\text{Lim}_{n\to\infty} x_{ik}^2/d_{nk}^2 = 0$ for all $i = 1, \ldots, n$. This condition implies that no single observation will ever dominate $\mathbf{x}_k'\mathbf{x}_k$, and as $n \to \infty$, individual observations will become less important.

G3. Let $\mathbf{R}_n$ be the sample correlation matrix of the columns of $\mathbf{X}$, excluding the constant term if there is one. Then $\lim_{n\to\infty} \mathbf{R}_n = \mathbf{C}$, a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that $\mathbf{X}$ has full rank in a finite sample, so this assumption ensures that the condition will never be violated.

The variance of $\sqrt{n}\bar{\mathbf{w}}$ is

$$\sigma^2\bar{\mathbf{Q}}_n = \sigma^2\left(\frac{1}{n}\right)[\mathbf{Q}_1 + \mathbf{Q}_2 + \cdots + \mathbf{Q}_n]. \tag{5-10}$$

As long as the sum is not dominated by any particular term and the regressors are well behaved, which in this case means that (5-1) holds,

$$\lim_{n\to\infty} \sigma^2\bar{\mathbf{Q}}_n = \sigma^2\mathbf{Q}. \tag{5-11}$$

Therefore, we may apply the Lindberg–Feller central limit theorem to the vector $\sqrt{n}\,\bar{\mathbf{w}}$, as we did in Section D.3 for the univariate case $\sqrt{n}\bar{x}$. We now have the elements we need for a formal result. If $[\mathbf{x}_i\varepsilon_i]$, $i = 1, \ldots, n$ are independent vectors distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{Q}_i < \infty$, and if (5-1) holds, then

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}]. \tag{5-12}$$

It then follows that

$$\mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}(\sigma^2\mathbf{Q})\mathbf{Q}^{-1}]. \tag{5-13}$$

Combining terms,

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}^{-1}]. \tag{5-14}$$

Using the technique of Section D.3, we obtain the **asymptotic distribution** of $\mathbf{b}$:

THEOREM 5.1 Asymptotic Distribution of b with Independent Observations

If $\{\varepsilon_i\}$ are independently distributed with mean zero and finite variance σ^2 and x_{ik} is such that the Grenander conditions are met, then

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right]. \tag{5-15}$$

In practice, it is necessary to estimate $(1/n)\mathbf{Q}^{-1}$ with $(\mathbf{X}'\mathbf{X})^{-1}$ and σ^2 with $\mathbf{e}'\mathbf{e}/(n - K)$.

If $\boldsymbol{\varepsilon}$ is normally distributed, then Result **FS7** in (Table 4.3, Section 4.8) holds in *every* sample, so it holds asymptotically as well. The important implication of this derivation is that *if the regressors are well behaved and observations are independent,* then the ***asymptotic normality*** *of the least squares estimator does not depend on normality of the disturbances; it is a consequence of the central limit theorem.* We will consider other more general cases in the sections to follow.

5.2.3 CONSISTENCY OF s^2 AND THE ESTIMATOR OF Asy. Var[b]

To complete the derivation of the asymptotic properties of **b**, we will require an estimator of Asy. Var[**b**] $= (\sigma^2/n)\mathbf{Q}^{-1}$.[5] With (5-1), it is sufficient to restrict attention to s^2, so the purpose here is to assess the consistency of s^2 as an estimator of σ^2. Expanding

$$s^2 = \frac{1}{n-K}\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$$

produces

$$s^2 = \frac{1}{n-K}[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] = \frac{n}{n-k}\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \left(\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)\right].$$

The leading constant clearly converges to 1. We can apply (5-1), (5-4) (twice), and the product rule for **probability limits** (Theorem D.14) to assert that the second term in the brackets converges to 0. That leaves

$$\overline{\varepsilon^2} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

This is a narrow case in which the random variables ε_i^2 are independent with the same finite mean σ^2, so not much is required to get the mean to converge almost surely to $\sigma^2 = E[\varepsilon_i^2]$. By the Markov Theorem (D.8), what is needed is for $E[|\varepsilon_i^2|^{1+\delta}]$ to be finite, so the minimal assumption thus far is that ε_i have finite moments up to slightly greater than 2. Indeed, if we further assume that every ε_i has the same distribution, then by the Khinchine Theorem (D.5) or the Corollary to D8, finite moments (of ε_i) up to 2 is sufficient. **Mean square convergence** would require $E[\varepsilon_i^4] = \phi_\varepsilon < \infty$. Then the terms in the sum are independent, with mean σ^2 and variance $\phi_\varepsilon - \sigma^4$. So, under fairly weak condition, the first term in brackets converges in probability to σ^2, which gives our result,

$$\operatorname{plim} s^2 = \sigma^2,$$

and, by the product rule,

$$\operatorname{plim} s^2(\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2\mathbf{Q}^{-1}.$$

The appropriate *estimator* of the asymptotic covariance matrix of **b** is

$$\text{Est.Asy. Var}[\mathbf{b}] = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

[5]See McCallum (1973) for some useful commentary on deriving the asymptotic covariance matrix of the least squares estimator.

5.2.4 ASYMPTOTIC DISTRIBUTION OF A FUNCTION OF b: THE DELTA METHOD

We can extend Theorem D.22 to functions of the least squares estimator. Let $\mathbf{f}(\mathbf{b})$ be a set of J continuous, linear or nonlinear and continuously differentiable functions of the least squares estimator, and let

$$\mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\mathbf{b})}{\partial \mathbf{b}'},$$

where $\mathbf{C}$ is the $J \times K$ matrix whose jth row is the vector of derivatives of the jth function with respect to $\mathbf{b}'$. By the Slutsky Theorem (D.12),

$$\operatorname{plim} \mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$$

and

$$\operatorname{plim} \mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \boldsymbol{\Gamma}.$$

Using our usual linear Taylor series approach, we expand this set of functions in the approximation

$$\mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta}) + \boldsymbol{\Gamma} \times (\mathbf{b} - \boldsymbol{\beta}) + \text{higher-order terms.}$$

The higher-order terms become negligible in large samples if plim $\mathbf{b} = \boldsymbol{\beta}$. Then, the asymptotic distribution of the function on the left-hand side is the same as that on the right. Thus, the mean of the asymptotic distribution is plim $\mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$, and the asymptotic covariance matrix is $\{\boldsymbol{\Gamma}[\text{Asy. Var}(\mathbf{b} - \boldsymbol{\beta})]\boldsymbol{\Gamma}'\}$, which gives us the following theorem:

THEOREM 5.2 Asymptotic Distribution of a Function of b

If $\mathbf{f}(\mathbf{b})$ *is a set of continuous and continuously differentiable functions of* $\mathbf{b}$ *such that* $\boldsymbol{\Gamma} = \partial \mathbf{f}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$ *and if Theorem 5.1 holds, then*

$$\mathbf{f}(\mathbf{b}) \xrightarrow{a} N\left[\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\Gamma}\left(\frac{\sigma^2}{n}\mathbf{Q}^{-1}\right)\boldsymbol{\Gamma}'\right]. \tag{5-16}$$

In practice, the estimator of the asymptotic covariance matrix would be

$$\text{Est.Asy. Var}[\mathbf{f}(\mathbf{b})] = \mathbf{C}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{C}'.$$

If any of the functions are nonlinear, then the property of unbiasedness that holds for $\mathbf{b}$ may not carry over to $\mathbf{f}(\mathbf{b})$. Nonetheless, it follows from (5-4) that $\mathbf{f}(\mathbf{b})$ is a consistent estimator of $\mathbf{f}(\boldsymbol{\beta})$, and the asymptotic covariance matrix is readily available.

5.2.5 ASYMPTOTIC EFFICIENCY

We have not established any large-sample counterpart to the Gauss-Markov theorem. That is, it remains to establish whether the large-sample properties of the least squares

estimator are optimal by any measure. The Gauss-Markov Theorem establishes finite sample conditions under which least squares is optimal. The requirements that the estimator be linear and unbiased limit the theorem's generality, however. One of the main purposes of the analysis in this chapter is to broaden the class of estimators in the classical model to those which might be biased, but which are consistent. Ultimately, we shall also be interested in nonlinear estimators. These cases extend beyond the reach of the Gauss Markov Theorem. To make any progress in this direction, we will require an alternative estimation criterion.

DEFINITION 5.1 Asymptotic Efficiency
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

In Chapter 17, we will show that if the disturbances are normally distributed, then the least squares estimator is also the **maximum likelihood estimator.** Maximum likelihood estimators are asymptotically efficient among consistent and asymptotically normally distributed estimators. This gives us a partial result, albeit a somewhat narrow one since to claim it, we must assume normally distributed disturbances. If some other distribution is specified for $\boldsymbol{\varepsilon}$ and it emerges that $\mathbf{b}$ is not the maximum likelihood estimator, then least squares may not be efficient.

Example 5.1 The Gamma Regression Model
Greene (1980a) considers estimation in a regression model with an asymmetrically distributed disturbance,

$$y = (\alpha - \sigma\sqrt{P}) + \mathbf{x}'\boldsymbol{\beta} - (\varepsilon - \sigma\sqrt{P}) = \alpha^* + \mathbf{x}'\boldsymbol{\beta} + \varepsilon^*,$$

where ε has the gamma distribution in Section B.4.5 [see (B-39)] and $\sigma = \sqrt{P}/\lambda$ is the standard deviation of the disturbance. In this model, the covariance matrix of the least squares estimator of the slope coefficients (not including the constant term) is,

$$\text{Asy. Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1},$$

whereas for the maximum likelihood estimator (which is not the least squares estimator),

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{ML}] \approx [1 - (2/P)]\sigma^2(\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1}.^6$$

But for the asymmetry parameter, this result would be the same as for the least squares estimator. We conclude that the estimator that accounts for the asymmetric disturbance distribution is more efficient asymptotically.

[6]The Matrix $\mathbf{M}^0$ produces data in the form of deviations from sample means. (See Section A.2.8.) In Greene's model, P must be greater than 2.

5.3 MORE GENERAL CASES

The asymptotic properties of the estimators in the classical regression model were established in Section 5.2 under the following assumptions:

A1. ***Linearity:*** $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.
A2. ***Full rank:*** The $n \times K$ sample data matrix, $\boldsymbol{X}$ has full column rank.
A3. ***Exogeneity of the independent variables:*** $E[\varepsilon_i \mid x_{j1}, x_{j2}, \ldots, x_{jK}] = 0$, $i, j = 1, \ldots, n$.
A4. ***Homoscedasticity and nonautocorrelation.***
A5. ***Data generating mechanism-independent observations.***

The following are the crucial results needed: For consistency of **b**, we need (5-1) and (5-4),

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \text{plim}\,\bar{\mathbf{Q}}_n = \mathbf{Q}, \quad \text{a positive definite matrix,}$$
$$\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \text{plim}\,\bar{\mathbf{w}}_n = E[\bar{\mathbf{w}}_n] = \mathbf{0}.$$

(For consistency of s^2, we added a fairly weak assumption about the moments of the disturbances.) To establish asymptotic normality, we will require consistency and (5-12) which is

$$\sqrt{n}\,\bar{\mathbf{w}}_n \xrightarrow{d} N[0, \sigma^2\mathbf{Q}].$$

With these in place, the desired characteristics are then established by the methods of Section 5.2. To analyze other cases, we can merely focus on these three results. It is not necessary to reestablish the consistency or asymptotic normality themselves, since they follow as a consequence.

5.3.1 HETEROGENEITY IN THE DISTRIBUTIONS OF x_i

Exceptions to the assumptions made above are likely to arise in two settings. In a **panel data** set, the sample will consist of multiple observations on each of many observational units. For example, a study might consist of a set of observations made at different points in time on a large number of families. In this case, the **x**s will surely be correlated across observations, at least within observational units. They might even be the same for all the observations on a single family. They are also likely to be a mixture of random variables, such as family income, and nonstochastic regressors, such as a fixed "family effect" represented by a dummy variable. The second case would be a time-series model in which lagged values of the dependent variable appear on the right-hand side of the model.

The panel data set could be treated as follows. Assume for the moment that the data consist of a fixed number of observations, say T, on a set of N families, so that the total number of rows in $\mathbf{X}$ is $n = NT$. The matrix

$$\bar{\mathbf{Q}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i$$

in which n is all the observations in the sample, could be viewed as

$$\bar{\mathbf{Q}}_n = \frac{1}{N}\sum_i \frac{1}{T} \sum_{\substack{\text{observations} \\ \text{for family } i}} \mathbf{Q}_{ij} = \frac{1}{N}\sum_{i=1}^{N} \bar{\mathbf{Q}}_i,$$

where $\bar{\mathbf{Q}}_i$ = average $\mathbf{Q}_{ij}$ for family i. We might then view the set of observations on the ith unit as if they were a single observation and apply our convergence arguments to the number of families increasing without bound. The point is that the conditions that are needed to establish convergence will apply with respect to the number of observational units. The number of observations taken for each observation unit might be fixed and could be quite small.

5.3.2 DEPENDENT OBSERVATIONS

The second difficult case arises when there are lagged dependent variables among the variables on the right-hand side or, more generally, in time series settings in which the observations are no longer independent or even uncorrelated. Suppose that the model may be written

$$y_t = \mathbf{z}_t'\boldsymbol{\theta} + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + \varepsilon_t. \tag{5-17}$$

(Since this model is a time-series setting, we use t instead of i to index the observations.) We continue to assume that the disturbances are uncorrelated across observations. Since y_{t-1} is dependent on y_{t-2} and so on, it is clear that although the disturbances are uncorrelated across observations, the regressor vectors, including the lagged ys, surely are not. Also, although $\text{Cov}[\mathbf{x}_t, \varepsilon_s] = 0$ if $s \geq t\big(\mathbf{x}_t = [\mathbf{z}_t, y_{t-1}, \ldots, y_{t-p}]\big)$, $\text{Cov}[\mathbf{x}_t, \varepsilon_s] \neq 0$ if $s < t$. Every observation y_t is determined by the entire history of the disturbances. Therefore, we have lost the crucial assumption $E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$; $E[\varepsilon_t \mid \text{future } \mathbf{x}\text{s}]$ is not equal to 0. The conditions needed for the finite-sample results we had earlier no longer hold. Without Assumption A3, $E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$, our earlier proof of unbiasedness dissolves, and without unbiasedness, the Gauss–Markov theorem no longer applies. We are left with only asymptotic results for this case.

This case is considerably more general than the ones we have considered thus far. The theorems we invoked previously do not apply when the observations in the sums are correlated. To establish counterparts to the limiting normal distribution of $(1/\sqrt{n})\mathbf{X}'\boldsymbol{\varepsilon}$ and convergence of $(1/n)\mathbf{X}'\mathbf{X}$ to a finite positive definite matrix, it is necessary to make additional assumptions about the regressors. For the disturbances, we replace Assumption A3 following.

AD3. $E[\varepsilon_t \mid \mathbf{x}_{t-s}] = 0$, for all $s \geq 0$.

This assumption states that the disturbance in the period "t" is an innovation; it is new information that enters the process. Thus, it is not correlated with any of the history. It is not uncorrelated with future data, however, since ε_t will be a part of x_{t+r}. Assumptions A1, A2, and A4 are retained (at least for the present). We will also replace Assumption A5 and result (5-1) with two assumptions about the right-hand variables.

First,

$$\text{plim}\frac{1}{T-s}\sum_{t=s+1}^{T}\mathbf{x}_t\mathbf{x}_{t-s}' = \mathbf{Q}(s), \quad \text{a finite matrix, } s \geq 0, \tag{5-18}$$

and $\mathbf{Q}(0)$ is nonsingular if $T \geq K$. [Note that $\mathbf{Q} = \mathbf{Q}(0)$.] This matrix is the sums of cross products of the elements of $\mathbf{x}_t$ with lagged values of $\mathbf{x}_t$. Second, we assume that the roots of the polynomial

$$1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_P z^P = 0 \tag{5-19}$$

are all outside the unit circle. (See Section 20.2 for further details.) Heuristically, these assumptions imply that the dependence between values of the **x**s at different points in time varies only with how far apart in time they are, not specifically with the points in time at which observations are made, and that the correlation between observations made at different points in time fades sufficiently rapidly that sample moments such as $\mathbf{Q}(s)$ above will converge in probability to a population counterpart.[7] Formally, we obtain these results with

AD5. The series on $\mathbf{x}_t$ is **stationary** and **ergodic.**

This assumption also implies that $\mathbf{Q}$(s) becomes a matrix of zeros as s (the separation in time) becomes large. These conditions are sufficient to produce $(1/n)\mathbf{X}'\boldsymbol{\varepsilon} \to \mathbf{0}$ and the consistency of **b**. Further results are needed to establish the asymptotic normality of the estimator, however.[8]

In sum, the important properties of consistency and asymptotic normality of the least squares estimator are preserved under the different assumptions of stochastic regressors, provided that additional assumptions are made. In most cases, these assumptions are quite benign, so we conclude that the two asymptotic properties of least squares considered here, consistency and asymptotic normality, are quite robust to different specifications of the regressors.

5.4 INSTRUMENTAL VARIABLE AND TWO STAGE LEAST SQUARES ESTIMATION

The assumption that $\mathbf{x}_i$ and ε_i are uncorrelated has been crucial in the development thus far. But, there are any number of applications in economics in which this assumption is untenable. Examples include models that contain variables that are measured with error and most dynamic models involving expectations. Without this assumption, none of the

[7] We will examine some cases in later chapters in which this does not occur. To consider a simple example, suppose that **x** contains a constant. Then the assumption requires sample means to converge to population parameters. Suppose that all observations are correlated. Then the variance of $\bar{x}$ is $\text{Var}[(1/T)\Sigma_t x_t] = (1/T^2)\Sigma_t\Sigma_s\text{Cov}[x_t, x_s]$. Since none of the T^2 terms is assumed to be zero, there is no assurance that the double sum converges to zero as $T \to \infty$. But if the correlations diminish sufficiently with distance in time, then the sum may converge to zero.

[8] These appear in Mann and Wald (1943), Billingsley (1979) and Dhrymes (1998).

proofs of consistency given above will hold up, so least squares loses its attractiveness as an estimator. There is an alternative method of estimation called the method of **instrumental variables** (IV). The least squares estimator is a special case, but the IV method is far more general. The method of instrumental variables is developed around the following general extension of the estimation strategy in the classical regression model: Suppose that in the classical model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, the K variables $\mathbf{x}_i$ may be correlated with ε_i. Suppose as well that there exists a set of L variables $\mathbf{z}_i$, where L is at least as large as K, such that $\mathbf{z}_i$ is correlated with $\mathbf{x}_i$ but not with ε_i. We cannot estimate $\boldsymbol{\beta}$ consistently by using the familiar least squares estimator. But we can construct a consistent estimator of $\boldsymbol{\beta}$ by using the assumed relationships among $\mathbf{z}_i$, $\mathbf{x}_i$, and ε_i.

Example 5.2 Models in Which Least Squares is Inconsistent

The following models will appear at various points in this book. In general, least squares will not be a suitable estimator.

Dynamic Panel Data Model In Example 13.6 and Section 18.5, we will examine a model for municipal expenditure of the form $S_{it} = f(S_{it-1}, \ldots) + \varepsilon_{it}$. The disturbances are assumed to be freely correlated across periods, so both $S_{i,t-1}$ and $\varepsilon_{i,t}$ are correlated with $\varepsilon_{i,t-1}$. It follows that they are correlated with each other, which means that this model, even with a linear specification, does not satisfy the assumptions of the classical model. The regressors and disturbances are correlated.

Dynamic Regression In Chapters 19 and 20, we will examine a variety of time series models which are of the form $y_t = f(y_{t-1}, \ldots) + \varepsilon_t$ in which ε_t is (auto-) correlated with its past values. This case is essentially the same as the one we just considered. Since the disturbances are autocorrelated, it follows that the dynamic regression implies correlation between the disturbance and a right hand side variable. Once again, least squares will be inconsistent.

Consumption Function We (and many other authors) have used a macroeconomic version of the consumption function at various points to illustrate least squares estimation of the classical regression model. But, by construction, the model violates the assumptions of the classical regression model. The national income data are assembled around some basic accounting identities, including "$Y = C +$ *investment* $+$ *government spending* $+$ *net exports*." Therefore, although the precise relationship between consumption C, and income Y, $C = f(Y, \varepsilon)$, is ambiguous and is a suitable candidate for modeling, it is clear that consumption (and therefore ε) is one of the main determinants of Y. The model $C_t = \alpha + \beta Y_t + \varepsilon_t$ does not fit our assumptions for the classical model if $\text{Cov}[Y_t, \varepsilon_t] \neq 0$. But it is reasonable to assume (at least for now) that ε_t is uncorrelated with past values of C and Y. Therefore, in this model, we might consider Y_{t-1} and C_{t-1} as suitable instrumental variables.

Measurement Error In Section 5.6, we will examine an application in which an earnings equation $y_{i,t} = f(\textit{Education}_{i,t}, \ldots) + \varepsilon_{i,t}$ is specified for sibling pairs (twins) $t = 1, 2$ for n individuals. Since education is a variable that is measured with error, it will emerge (in a way that will be established below) that this is, once again, a case in which the disturbance and an independent variable are correlated.

None of these models can be consistently estimated by least squares—the method of instrumental variables is the standard approach.

We will now construct an estimator for $\boldsymbol{\beta}$ in this extended model. We will maintain assumption A5 (independent observations with finite moments), though this is only for convenience. These results can all be extended to cases with dependent observations. This will preserve the important result that $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}_{\mathbf{xx}}$. (We use the subscript to differentiate this result from the results given below.) The basic assumptions of the regression model have changed, however. First, A3 (no correlation between $\mathbf{x}$ and ε) is, under our new assumptions,

AI3. $E[\varepsilon_i \mid \mathbf{x}_i] = \eta_i$.

We interpret Assumption AI3 to mean that the regressors now provide information about the expectations of the disturbances. The important implication of AI3 is that the disturbances and the regressors are now correlated. Assumption AI3 implies that

$$E[\mathbf{x}_i\varepsilon_i] = \boldsymbol{\gamma}$$

for some nonzero $\boldsymbol{\gamma}$. If the data are "well behaved," then we can apply Theorem D.5 (Khinchine's theorem) to assert that

$$\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\gamma}.$$

Notice that the original model results if $\eta_i = 0$. Finally, we must characterize the instrumental variables. We assume the following:

AI7. $[\mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, $i = 1, \ldots, n$, are an i.i.d. sequence of random variables.
AI8a. $E[x_{ik}^2] = \mathbf{Q}_{\mathbf{xx},kk} < \infty$, a finite constant, $k = 1, \ldots, K$.
AI8b. $E[z_{il}^2] = \mathbf{Q}_{\mathbf{zz},ll} < \infty$, a finite constant, $l = 1, \ldots, L$.
AI8c. $E[z_{il}x_{ik}] = \mathbf{Q}_{\mathbf{zx},lk} < \infty$, a finite constant, $l = 1, \ldots, L$, $k = 1, \ldots, K$.
AI9. $E[\varepsilon_i \mid \mathbf{z}_i] = 0$.

In later work in time series models, it will be important to relax assumption AI7. Finite means of z_l follows from AI8b. Using the same analysis as in the preceding section, we have

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{zz}}, \text{ a finite, positive definite (assumed) matrix,}$$

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{zx}}, \text{ a finite, } L \times K \text{ matrix with rank } K \text{ (assumed),}$$

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}.$$

In our statement of the classical regression model, we have assumed thus far the special case of $\eta_i = 0$; $\boldsymbol{\gamma} = \mathbf{0}$ follows. There is no need to dispense with Assumption AI7—it may well continue to be true—but in this special case, it becomes irrelevant.

For this more general model, we lose most of the useful results we had for least squares. The estimator $\mathbf{b}$ is no longer unbiased;

$$E[\mathbf{b} \mid \mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta} \neq \boldsymbol{\beta},$$

so the Gauss–Markov theorem no longer holds. It is also inconsistent;

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \text{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{xx}}^{-1}\boldsymbol{\gamma} \neq \boldsymbol{\beta}.$$

(The asymptotic distribution is considered in the exercises.)

We now turn to the instrumental variable estimator. Since $E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$ and all terms have finite variances, we can state that

$$\text{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta} + \text{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = \left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta}.$$

Suppose that $\mathbf{Z}$ has the same number of variables as $\mathbf{X}$. For example, suppose in our consumption function that $\mathbf{x}_t = [1, Y_t]$ when $\mathbf{z}_t = [1, Y_{t-1}]$. We have assumed that the rank of $\mathbf{Z}'\mathbf{X}$ is K, so now $\mathbf{Z}'\mathbf{X}$ is a square matrix. It follows that

$$\left[\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1} \operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \boldsymbol{\beta},$$

which leads us to the **instrumental variable estimator,**

$$\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

We have already proved that $\mathbf{b}_{IV}$ is consistent. We now turn to the asymptotic distribution. We will use the same method as in the previous section. First,

$$\sqrt{n}(\mathbf{b}_{IV} - \boldsymbol{\beta}) = \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon},$$

which has the same limiting distribution as $\mathbf{Q}_{zx}^{-1}[(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}]$. Our analysis of $(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}$ is the same as that of $(1/\sqrt{n})\mathbf{X}'\boldsymbol{\varepsilon}$ in the previous section, so it follows that

$$\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon}\right) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{zz}]$$

and

$$\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon}\right) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{zx}^{-1}\mathbf{Q}_{zz}\mathbf{Q}_{xz}^{-1}].$$

This step completes the derivation for the next theorem.

THEOREM 5.3 Asymptotic Distribution of the Instrumental Variables Estimator

If Assumptions A1, A2, AI3, A4, AS5, AS5a, AI7, AI8a–c and AI9 all hold for $[y_i, \mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, *where* $\mathbf{z}$ *is a valid set of* $L = K$ *instrumental variables, then the asymptotic distribution of the instrumental variables estimator* $\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ *is*

$$\mathbf{b}_{IV} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}_{zx}^{-1}\mathbf{Q}_{zz}\mathbf{Q}_{xz}^{-1}\right]. \tag{5-20}$$

where $\mathbf{Q}_{zx} = \operatorname{plim}(\mathbf{Z}'\mathbf{X}/n)$ *and* $\mathbf{Q}_{zz} = \operatorname{plim}(\mathbf{Z}'\mathbf{Z}/n)$.

To estimate the asymptotic covariance matrix, we will require an estimator of σ^2. The natural estimator is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b}_{IV})^2.$$

A correction for degrees of freedom, as in the development in the previous section, is superfluous, as all results here are asymptotic, and $\hat{\sigma}^2$ would not be unbiased in any event. (Nonetheless, it is standard practice in most software to make the degrees of freedom correction.) Write the vector of residuals as

$$\mathbf{y} - \mathbf{X}\mathbf{b}_{IV} = \mathbf{y} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

Substitute $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and collect terms to obtain $\hat{\boldsymbol{\varepsilon}} = [\mathbf{I} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}']\boldsymbol{\varepsilon}$. Now,

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

$$= \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} + \left(\frac{\boldsymbol{\varepsilon}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) - 2\left(\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right).$$

We found earlier that we could (after a bit of manipulation) apply the product result for probability limits to obtain the probability limit of an expression such as this. Without repeating the derivation, we find that $\hat{\sigma}^2$ is a consistent estimator of σ^2, by virtue of the first term. The second and third product terms converge to zero. To complete the derivation, then, we will estimate Asy. Var[$\mathbf{b}_{IV}$] with

$$\text{Est.Asy. Var}[\mathbf{b}_{IV}] = \frac{1}{n}\left\{\left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1}\right\} \qquad \textbf{(5-21)}$$

$$= \hat{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}.$$

There is a remaining detail. If $\mathbf{Z}$ contains more variables than $\mathbf{X}$, then much of the preceding is unusable, because $\mathbf{Z}'\mathbf{X}$ will be $L \times K$ with rank $K < L$ and will thus not have an inverse. The crucial result in all the preceding is $\text{plim}(\mathbf{Z}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. That is, every column of $\mathbf{Z}$ is asymptotically uncorrelated with $\boldsymbol{\varepsilon}$. That also means that every linear combination of the columns of $\mathbf{Z}$ is also uncorrelated with $\boldsymbol{\varepsilon}$, which suggests that one approach would be to choose K linear combinations of the columns of $\mathbf{Z}$. Which to choose? One obvious possibility is simply to choose K variables among the L in $\mathbf{Z}$. But intuition correctly suggests that throwing away the information contained in the remaining $L - K$ columns is inefficient. A better choice is the projection of the columns of $\mathbf{X}$ in the column space of $\mathbf{Z}$:

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

We will return shortly to the virtues of this choice. With this choice of instrumental variables, $\hat{\mathbf{X}}$ for $\mathbf{Z}$, we have

$$\mathbf{b}_{IV} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

$$= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}. \qquad \textbf{(5-22)}$$

By substituting $\hat{\mathbf{X}}$ in the expression for Est.Asy. Var[$\mathbf{b}_{IV}$] and multiplying it out, we see that the expression is unchanged. The proofs of consistency and asymptotic normality for this estimator are exactly the same as before, because our proof was generic for any valid set of instruments, and $\hat{\mathbf{X}}$ qualifies.

There are two reasons for using this estimator—one practical, one theoretical. If any column of $\mathbf{X}$ also appears in $\mathbf{Z}$, then that column of $\mathbf{X}$ is reproduced exactly in $\hat{\mathbf{X}}$. This is easy to show. In the expression for $\hat{\mathbf{X}}$, if the kth column in $\mathbf{X}$ is one of the columns in $\mathbf{Z}$, say the lth, then the kth column in $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column of an $L \times L$ identity matrix. This result means that the kth column in $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column in $\mathbf{Z}$, which is the kth column in $\mathbf{X}$. This result is important and useful. Consider what is probably the typical application. Suppose that the regression contains K variables, only one of which, say the kth, is correlated with the disturbances. We have one or more instrumental variables in hand, as well as the other $K-1$ variables that certainly qualify as instrumental variables in their own right. Then what we would use is $\mathbf{Z} = [\mathbf{X}_{(k)}, \mathbf{z}_1, \mathbf{z}_2, \ldots]$, where we indicate omission of the kth variable by (k) in the subscript. Another useful interpretation of $\hat{\mathbf{X}}$ is that each column is the set of fitted values when the corresponding column of $\mathbf{X}$ is regressed on all the columns of $\mathbf{Z}$, which is obvious from the definition. It also makes clear why each $\mathbf{x}_k$ that appears in $\mathbf{Z}$ is perfectly replicated. Every $\mathbf{x}_k$ provides a perfect predictor for itself, without any help from the remaining variables in $\mathbf{Z}$. In the example, then, every column of $\mathbf{X}$ except the one that is omitted from $\mathbf{X}_{(k)}$ is replicated exactly, whereas the one that *is* omitted is replaced in $\hat{\mathbf{X}}$ by the predicted values in the regression of this variable on all the $\mathbf{z}$s.

Of all the different linear combinations of $\mathbf{Z}$ that we might choose, $\hat{\mathbf{X}}$ is the most efficient in the sense that the asymptotic covariance matrix of an IV estimator based on a linear combination $\mathbf{ZF}$ is smaller when $\mathbf{F} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ than with any other $\mathbf{F}$ that uses all L columns of $\mathbf{Z}$; a fortiori, this result eliminates linear combinations obtained by dropping any columns of $\mathbf{Z}$. This important result was proved in a seminal paper by Brundy and Jorgenson (1971).

We close this section with some practical considerations in the use of the instrumental variables estimator. By just multiplying out the matrices in the expression, you can show that

$$\begin{aligned}\mathbf{b}_{\mathrm{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'(\mathbf{I} - \mathbf{M}_z)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{M}_z)\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}\end{aligned}$$

since $\mathbf{I} - \mathbf{M}_z$ is idempotent. Thus, when (*and only when*) $\hat{\mathbf{X}}$ is the set of instruments, the IV estimator is computed by least squares regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$. This conclusion suggests (only logically; one need not actually do this in two steps), that $\mathbf{b}_{\mathrm{IV}}$ can be computed in two steps, first by computing $\hat{\mathbf{X}}$, then by the least squares regression. For this reason, this is called the **two-stage least squares** (2SLS) estimator. We will revisit this form of estimator at great length at several points below, particularly in our discussion of simultaneous equations models, under the rubric of "two-stage least squares." One should be careful of this approach, however, in the computation of the asymptotic covariance matrix; $\hat{\sigma}^2$ should not be based on $\hat{\mathbf{X}}$. The estimator

$$s_{\mathrm{IV}}^2 = \frac{(\mathbf{y} - \hat{\mathbf{X}}\mathbf{b}_{\mathrm{IV}})'(\mathbf{y} - \hat{\mathbf{X}}\mathbf{b}_{\mathrm{IV}})}{n}$$

is inconsistent for σ^2, with or without a correction for degrees of freedom.

An obvious question is where one is likely to find a suitable set of instrumental variables. In many time-series settings, lagged values of the variables in the model

provide natural candidates. In other cases, the answer is less than obvious. The asymptotic variance matrix of the IV estimator can be rather large if $\mathbf{Z}$ is not highly correlated with $\mathbf{X}$; the elements of $(\mathbf{Z}'\mathbf{X})^{-1}$ grow large. Unfortunately, there usually is not much choice in the selection of instrumental variables. The choice of $\mathbf{Z}$ is often ad hoc.[9] There is a bit of a dilemma in this result. It would seem to suggest that the best choices of instruments are variables that are highly correlated with $\mathbf{X}$. But the more highly correlated a variable is with the problematic columns of $\mathbf{X}$, the less defensible the claim that these same variables are *uncorrelated* with the disturbances.

5.5 HAUSMAN'S SPECIFICATION TEST AND AN APPLICATION TO INSTRUMENTAL VARIABLE ESTIMATION

It might not be obvious that the regressors in the model are correlated with the disturbances or that the regressors are measured with error. If not, there would be some benefit to using the least squares estimator rather than the IV estimator. Consider a comparison of the two covariance matrices *under the hypothesis that both are consistent, that is, assuming* $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. The difference between the asymptotic covariance matrices of the two estimators is

$$\text{Asy. Var}[\mathbf{b}_{\text{IV}}] - \text{Asy. Var}[\mathbf{b}_{\text{LS}}] = \frac{\sigma^2}{n}\text{plim}\left(\frac{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} - \frac{\sigma^2}{n}\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}$$

$$= \frac{\sigma^2}{n}\,\text{plim}\, n\left[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\right].$$

To compare the two matrices in the brackets, we can compare their inverses. The inverse of the first is $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{X} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}$. Since $\mathbf{M}_{\mathbf{Z}}$ is a nonnegative definite matrix, it follows that $\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}$ is also. So, $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ equals $\mathbf{X}'\mathbf{X}$ minus a nonnegative definite matrix. Since $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ is smaller, in the matrix sense, than $\mathbf{X}'\mathbf{X}$, its inverse is larger. Under the hypothesis, the asymptotic covariance matrix of the LS estimator is never larger than that of the IV estimator, and it will actually be smaller unless all the columns of $\mathbf{X}$ are perfectly predicted by regressions on $\mathbf{Z}$. Thus, we have established that if $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$—that is, if LS is consistent—then it is a preferred estimator. (Of course, we knew that from all our earlier results on the virtues of least squares.)

Our interest in the difference between these two estimators goes beyond the question of efficiency. The null hypothesis of interest will usually be specifically whether $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. Seeking the covariance between $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ through $(1/n)\mathbf{X}'\mathbf{e}$ is fruitless, of course, since the normal equations produce $(1/n)\mathbf{X}'\mathbf{e} = \mathbf{0}$. In a seminal paper, Hausman (1978) suggested an alternative testing strategy. [Earlier work by Wu (1973) and Durbin (1954) produced what turns out to be the same test.] The logic of Hausman's approach is as follows. Under the null hypothesis, we have two consistent estimators of

[9] Results on "optimal instruments" appear in White (2001) and Hansen (1982). In the other direction, there is a contemporary literature on "weak" instruments, such as Staiger and Stock (1997).

$\boldsymbol{\beta}$, $\mathbf{b}_{LS}$ and $\mathbf{b}_{IV}$. Under the alternative hypothesis, only one of these, $\mathbf{b}_{IV}$, is consistent. The suggestion, then, is to examine $\mathbf{d} = \mathbf{b}_{IV} - \mathbf{b}_{LS}$. Under the null hypothesis, plim $\mathbf{d} = \mathbf{0}$, whereas under the alternative, plim $\mathbf{d} \neq \mathbf{0}$. Using a strategy we have used at various points before, we might test this hypothesis with a Wald statistic,

$$H = \mathbf{d}'\{\text{Est.Asy. Var}[\mathbf{d}]\}^{-1}\mathbf{d}.$$

The asymptotic covariance matrix we need for the test is

$$\begin{aligned}\text{Asy. Var}[\mathbf{b}_{IV} - \mathbf{b}_{LS}] = {} & \text{Asy. Var}[\mathbf{b}_{IV}] + \text{Asy. Var}[\mathbf{b}_{LS}] \\ & - \text{Asy. Cov}[\mathbf{b}_{IV}, \mathbf{b}_{LS}] - \text{Asy. Cov}[\mathbf{b}_{LS}, \mathbf{b}_{IV}].\end{aligned}$$

At this point, the test is straightforward, save for the considerable complication that we do not have an expression for the covariance term. Hausman gives a fundamental result that allows us to proceed. Paraphrased slightly,

> *the covariance between an efficient estimator, $\mathbf{b}_E$, of a parameter vector, $\boldsymbol{\beta}$, and its difference from an inefficient estimator, $\mathbf{b}_I$, of the same parameter vector, $\mathbf{b}_E - \mathbf{b}_I$, is zero.*

For our case, $\mathbf{b}_E$ is $\mathbf{b}_{LS}$ and $\mathbf{b}_I$ is $\mathbf{b}_{IV}$. By Hausman's result we have

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_E - \mathbf{b}_I] = \text{Var}[\mathbf{b}_E] - \text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \mathbf{0}$$

or

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \text{Var}[\mathbf{b}_E],$$

so,

$$\text{Asy.Var}[\mathbf{b}_{IV} - \mathbf{b}_{LS}] = \text{Asy. Var}[\mathbf{b}_{IV}] - \text{Asy. Var}[\mathbf{b}_{LS}].$$

Inserting this useful result into our Wald statistic and reverting to our empirical estimates of these quantities, we have

$$H = (\mathbf{b}_{IV} - \mathbf{b}_{LS})'\{\text{Est.Asy. Var}[\mathbf{b}_{IV}] - \text{Est.Asy. Var}[\mathbf{b}_{LS}]\}^{-1}(\mathbf{b}_{IV} - \mathbf{b}_{LS}).$$

Under the null hypothesis, we are using two different, but consistent, estimators of σ^2. If we use s^2 as the common estimator, then the statistic will be

$$H = \frac{\mathbf{d}'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{d}}{s^2}. \tag{5-23}$$

It is tempting to invoke our results for the full rank quadratic form in a normal vector and conclude the degrees of freedom for this chi-squared statistic is K. But that method will usually be incorrect, and worse yet, *unless* $\mathbf{X}$ *and* $\mathbf{Z}$ *have no variables in common, the rank of the matrix in this statistic is less than K, and the ordinary inverse will not even exist.* In most cases, at least some of the variables in $\mathbf{X}$ will also appear in $\mathbf{Z}$. (In almost any application, $\mathbf{X}$ and $\mathbf{Z}$ will both contain the constant term.) That is, some of the variables in $\mathbf{X}$ are known to be uncorrelated with the disturbances. For example, the usual case will involve a single variable that is thought to be problematic or that is measured with error. In this case, our hypothesis, plim$(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$, does not

really involve all K variables, since a subset of the elements in this vector, say K_0, are known to be zero. As such, the quadratic form in the Wald test is being used to test only $K^* = K - K_0$ hypotheses. It is easy (and useful) to show that, in fact, H is a rank K^* quadratic form. Since $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is an idempotent matrix, $(\hat{\mathbf{X}}'\hat{\mathbf{X}}) = \hat{\mathbf{X}}'\mathbf{X}$. Using this result and expanding $\mathbf{d}$, we find

$$\begin{aligned}\mathbf{d} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}[\hat{\mathbf{X}}'\mathbf{y} - (\hat{\mathbf{X}}'\hat{\mathbf{X}})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{e},\end{aligned}$$

where $\mathbf{e}$ is the vector of least squares residuals. Recall that K_0 of the columns in $\hat{\mathbf{X}}$ are the original variables in $\mathbf{X}$. Suppose that these variables are the first K_0. Thus, the first K_0 rows of $\hat{\mathbf{X}}'\mathbf{e}$ are the same as the first K_0 rows of $\mathbf{X}'\mathbf{e}$, which are, of course $\mathbf{0}$. (This statement does not mean that the first K_0 elements of $\mathbf{d}$ are zero.) So, we can write $\mathbf{d}$ as

$$\mathbf{d} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix}\mathbf{0} \\ \hat{\mathbf{X}}^{*\prime}\mathbf{e}\end{bmatrix} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix}\mathbf{0} \\ \mathbf{q}^*\end{bmatrix}.$$

Finally, denote the entire matrix in H by $\mathbf{W}$. (Since that ordinary inverse may not exist, this matrix will have to be a generalized inverse; see Section A.7.12.) Then, denoting the whole matrix product by $\mathbf{P}$, we obtain

$$H = [\mathbf{0}'\ \mathbf{q}^{*\prime}](\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\mathbf{W}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix}\mathbf{0} \\ \mathbf{q}^*\end{bmatrix} = [\mathbf{0}'\ \mathbf{q}^{*\prime}]\mathbf{P}\begin{bmatrix}\mathbf{0} \\ \mathbf{q}^*\end{bmatrix} = \mathbf{q}^{*\prime}\mathbf{P}_{**}\mathbf{q}^*,$$

where $\mathbf{P}_{**}$ is the lower right $K^* \times K^*$ submatrix of $\mathbf{P}$. We now have the end result. Algebraically, H is actually a quadratic form in a K^* vector, so K^* is the degrees of freedom for the test.

Since the preceding Wald test requires a generalized inverse [see Hausman and Taylor (1981)], it is going to be a bit cumbersome. In fact, one need not actually approach the test in this form, and it can be carried out with any regression program. The alternative approach devised by Wu (1973) is simpler. An F statistic with K^* and $n - K - K^*$ degrees of freedom can be used to test the joint significance of the elements of $\boldsymbol{\gamma}$ in the augmented regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{X}}^*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*, \tag{5-24}$$

where $\hat{\mathbf{X}}^*$ are the fitted values in regressions of the variables in $\mathbf{X}^*$ on $\mathbf{Z}$. This result is equivalent to the Hausman test for this model. [Algebraic derivations of this result can be found in the articles and in Davidson and MacKinnon (1993).]

Although most of the results above are specific to this test of correlation between some of the columns of $\mathbf{X}$ and the disturbances, $\boldsymbol{\varepsilon}$, the Hausman test is general. To reiterate, when we have a situation in which we have a pair of estimators, $\hat{\boldsymbol{\theta}}_E$ and $\hat{\boldsymbol{\theta}}_I$, such that under H_0: $\hat{\boldsymbol{\theta}}_E$ and $\hat{\boldsymbol{\theta}}_I$ are both consistent and $\hat{\boldsymbol{\theta}}_E$ is efficient relative to $\hat{\boldsymbol{\theta}}_I$, while under H_1: $\hat{\boldsymbol{\theta}}_I$ remains consistent while $\hat{\boldsymbol{\theta}}_E$ is inconsistent, then we can form a test of the

hypothesis by referring the "Hausman statistic,"

$$H = (\hat{\boldsymbol{\theta}}_I - \hat{\boldsymbol{\theta}}_E)'\{\text{Est.Asy. Var}[\hat{\boldsymbol{\theta}}_I] - \text{Est.Asy. Var}[\hat{\boldsymbol{\theta}}_E]\}^{-1}(\hat{\boldsymbol{\theta}}_I - \hat{\boldsymbol{\theta}}_E) \xrightarrow{d} \chi^2[J],$$

to the appropriate critical value for the chi-squared distribution. The appropriate degrees of freedom for the test, J, will depend on the context. Moreover, some sort of generalized inverse matrix may be needed for the matrix, although in at least one common case, the random effects regression model (see Chapter 13), the appropriate approach is to extract some rows and columns from the matrix instead. The short rank issue is not general. Many applications can be handled directly in this form with a full rank quadratic form. Moreover, the Wu approach *is* specific to this application. The other applications that we will consider, fixed and random effects for panel data and the independence from irrelevant alternatives test for the multinomial logit model, do not lend themselves to the regression approach and are typically handled using the Wald statistic and the full rank quadratic form. As a final note, observe that the short rank of the matrix in the Wald statistic is an algebraic result. The failure of the matrix in the Wald statistic to be positive definite, however, is sometimes a finite sample problem that is not part of the model structure. In such a case, forcing a solution by using a generalized inverse may be misleading. Hausman suggests that in this instance, the appropriate conclusion might be simply to take the result as zero and, by implication, not reject the null hypothesis.

Example 5.3 Hausman Test for a Consumption Function

Quarterly data for 1950.1 to 2000.4 on a number of macroeconomic variables appear in Table F5.1. A consumption function of the form $C_t = \alpha + \beta Y_t + \varepsilon_t$ is estimated using the 204 observations on aggregate U.S. consumption and disposable personal income. In Example 5.2, this model is suggested as a candidate for the possibility of bias due to correlation between Y_t and ε_t. Consider instrumental variables estimation using Y_{t-1} and C_{t-1} as the instruments for Y_t, and, of course, the constant term is its own instrument. One observation is lost because of the lagged values, so the results are based on 203 quarterly observations. The Hausman statistic can be computed in two ways:

1. Use the Wald statistic in (5-23) with the Moore–Penrose generalized inverse. The common s^2 is the one computed by least squares under the null hypothesis of no correlation. With this computation, $H = 22.111$. There is $K^* = 1$ degree of freedom. The 95 percent critical value from the chi-squared table is 3.84. Therefore, we reject the null hypothesis of no correlation between Y_t and ε_t.
2. Using the Wu statistic based on (5-24), we regress C_t on a constant, Y_t, and the predicted value in a regression of Y_t on a constant, Y_{t-1} and C_{t-1}. The t ratio on the prediction is 4.945, so the F statistic with 1 and 201 degrees of freedom is 24.453. The critical value for this F distribution is 4.15, so, again, the null hypothesis is rejected.

5.6 MEASUREMENT ERROR

Thus far, it has been assumed (at least implicitly) that the data used to estimate the parameters of our models are true measurements on their theoretical counterparts. In practice, this situation happens only in the best of circumstances. All sorts of measurement problems creep into the data that must be used in our analyses. Even carefully constructed survey data do not always conform exactly to the variables the analysts have in mind for their regressions. Aggregate statistics such as GDP are only estimates

of their theoretical counterparts, and some variables, such as depreciation, the services of capital, and "the interest rate," do not even exist in an agreed-upon theory. At worst, there may be no physical measure corresponding to the variable in our model; intelligence, education, and permanent income are but a few examples. Nonetheless, they all have appeared in very precisely defined regression models.

5.6.1 LEAST SQUARES ATTENUATION

In this section, we examine some of the received results on regression analysis with badly measured data. The general assessment of the problem is not particularly optimistic. The biases introduced by measurement error can be rather severe. There are almost no known finite-sample results for the models of measurement error; nearly all the results that have been developed are asymptotic.[10] The following presentation will use a few simple asymptotic results for the classical regression model.

The simplest case to analyze is that of a regression model with a single regressor and no constant term. Although this case is admittedly unrealistic, it illustrates the essential concepts, and we shall generalize it presently. Assume that the model

$$y^* = \beta x^* + \varepsilon \tag{5-25}$$

conforms to all the assumptions of the classical normal regression model. If data on y^* and x^* were available, then β would be estimable by least squares. Suppose, however, that the observed data are only imperfectly measured versions of y^* and x^*. In the context of an example, suppose that y^* is ln(output/labor) and x^* is ln(capital/labor). Neither factor input can be measured with precision, so the observed y and x contain errors of measurement. We assume that

$$y = y^* + v \quad \text{with } v \sim N[0, \sigma_v^2], \tag{5-26a}$$

$$x = x^* + u \quad \text{with } u \sim N[0, \sigma_u^2]. \tag{5-26b}$$

Assume, as well, that u and v are independent of each other and of y^* and x^*. (As we shall see, adding these restrictions is not sufficient to rescue a bad situation.)

As a first step, insert (5-26a) into (5-25), assuming for the moment that only y^* is measured with error:

$$y = \beta x^* + \varepsilon + v = \beta x^* + \varepsilon'.$$

This result conforms to the assumptions of the classical regression model. As long as the regressor is measured properly, measurement error on the dependent variable can be absorbed in the disturbance of the regression and ignored. To save some cumbersome notation, therefore, we shall henceforth assume that the measurement error problems concern only the independent variables in the model.

Consider, then, the regression of y on the observed x. By substituting (5-26b) into (5-25), we obtain

$$y = \beta x + [\varepsilon - \beta u] = \beta x + w. \tag{5-27}$$

[10] See, for example, Imbens and Hyslop (2001).

Since x equals $x^* + u$, the regressor in (5-27) is correlated with the disturbance:

$$\text{Cov}[x, w] = \text{Cov}[x^* + u, \varepsilon - \beta u] = -\beta\sigma_u^2. \tag{5-28}$$

This result violates one of the central assumptions of the classical model, so we can expect the least squares estimator

$$b = \frac{(1/n)\sum_{i=1}^n x_i y_i}{(1/n)\sum_{i=1}^n x_i^2}$$

to be inconsistent. To find the probability limits, insert (5-25) and (5-26b) and use the Slutsky theorem:

$$\text{plim}\, b = \frac{\text{plim}(1/n)\sum_{i=1}^n (x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\text{plim}(1/n)\sum_{i=1}^n (x_i^* + u_i)^2}.$$

Since x^*, ε, and u are mutually independent, this equation reduces to

$$\text{plim}\, b = \frac{\beta Q^*}{Q^* + \sigma_u^2} = \frac{\beta}{1 + \sigma_u^2/Q^*}, \tag{5-29}$$

where $Q^* = \text{plim}(1/n)\sum_i x_i^{*2}$. As long as σ_u^2 is positive, b is inconsistent, with a persistent bias toward zero. Clearly, the greater the variability in the measurement error, the worse the bias. The effect of biasing the coefficient toward zero is called **attenuation.**

In a multiple regression model, matters only get worse. Suppose, to begin, we assume that $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, allowing every observation on every variable to be measured with error. The extension of the earlier result is

$$\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) = \mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}, \quad \text{and} \quad \text{plim}\left(\frac{\mathbf{X}'\mathbf{y}}{n}\right) = \mathbf{Q}^*\boldsymbol{\beta}.$$

Hence,

$$\text{plim}\, \mathbf{b} = [\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1}\mathbf{Q}^*\boldsymbol{\beta} = \boldsymbol{\beta} - [\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1}\boldsymbol{\Sigma}_{uu}\boldsymbol{\beta}. \tag{5-30}$$

This probability limit is a mixture of all the parameters in the model. In the same fashion as before, bringing in outside information could lead to **identification.** The amount of information necessary is extremely large, however, and this approach is not particularly promising.

It is common for only a single variable to be measured with error. One might speculate that the problems would be isolated to the single coefficient. Unfortunately, this situation is not the case. For a single bad variable—assume that it is the first—the matrix $\boldsymbol{\Sigma}_{uu}$ is of the form

$$\boldsymbol{\Sigma}_{uu} = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

It can be shown that for this special case,

$$\text{plim}\, b_1 = \frac{\beta_1}{1 + \sigma_u^2 q^{*11}} \tag{5-31a}$$

(note the similarity of this result to the earlier one), and, for $k \neq 1$,

$$\text{plim } b_k = \beta_k - \beta_1 \left[\frac{\sigma_u^2 q^{*k1}}{1 + \sigma_u^2 q^{*11}} \right], \tag{5-31b}$$

where q^{*k1} is the $(k, 1)$th element in $(\mathbf{Q}^*)^{-1}$.[11] This result depends on several unknowns and cannot be estimated. The coefficient on the badly measured variable is still biased toward zero. The other coefficients are all biased as well, although in unknown directions. A badly measured variable contaminates all the least squares estimates.[12] If more than one variable is measured with error, there is very little that can be said.[13] Although expressions can be derived for the biases in a few of these cases, they generally depend on numerous parameters whose signs and magnitudes are unknown and, presumably, unknowable.

5.6.2 INSTRUMENTAL VARIABLES ESTIMATION

An alternative set of results for estimation in this model (and numerous others) is built around the method of instrumental variables. Consider once again the errors in variables model in (5-25) and (5-26a,b). The parameters, β, σ_ε^2, q^*, and σ_u^2 are not identified in terms of the moments of x and y. Suppose, however, that there exists a variable z such that z is correlated with x^* but not with u. For example, in surveys of families, income is notoriously badly reported, partly deliberately and partly because respondents often neglect some minor sources. Suppose, however, that one could determine the total amount of checks written by the head(s) of the household. It is quite likely that this z would be highly correlated with income, but perhaps not significantly correlated with the errors of measurement. If $\text{Cov}[x^*, z]$ is not zero, then the parameters of the model become estimable, as

$$\text{plim } \frac{(1/n)\sum_i y_i z_i}{(1/n)\sum_i x_i z_i} = \frac{\beta \text{ Cov}[x^*, z]}{\text{Cov}[x^*, z]} = \beta. \tag{5-32}$$

In a multiple regression framework, if only a single variable is measured with error, then the preceding can be applied to that variable and the remaining variables can serve as their own instruments. If more than one variable is measured with error, then the first preceding proposal will be cumbersome at best, whereas the second can be applied to each.

For the general case, $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, suppose that there exists a matrix of variables $\mathbf{Z}$ that is not correlated with the disturbances or the measurement error but is correlated with regressors, $\mathbf{X}$. Then the instrumental variables estimator based on $\mathbf{Z}$, $\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$, is consistent and asymptotically normally distributed with asymptotic covariance matrix that is estimated with

$$\text{Est.Asy. Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2[\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{Z}'\mathbf{Z}][\mathbf{X}'\mathbf{Z}]^{-1}. \tag{5-33}$$

For more general cases, Theorem 5.3 and the results in Section 5.4 apply.

[11] Use (A-66) to invert $[\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}] = [\mathbf{Q}^* + (\sigma_u \mathbf{e}_1)(\sigma_u \mathbf{e}_1)']$, where $\mathbf{e}_1$ is the first column of a $K \times K$ identity matrix. The remaining results are then straightforward.

[12] This point is important to remember when the presence of measurement error is suspected.

[13] Some firm analytic results have been obtained by Levi (1973), Theil (1961), Klepper and Leamer (1983), Garber and Klepper (1980), and Griliches (1986) and Cragg (1997).

5.6.3 PROXY VARIABLES

In some situations, a variable in a model simply has no observable counterpart. Education, intelligence, ability, and like factors are perhaps the most common examples. In this instance, unless there is some observable indicator for the variable, the model will have to be treated in the framework of missing variables. Usually, however, such an indicator can be obtained; for the factors just given, years of schooling and test scores of various sorts are familiar examples. The usual treatment of such variables is in the measurement error framework. If, for example,

$$\text{income} = \beta_1 + \beta_2 \text{ education} + \varepsilon$$

and

$$\text{years of schooling} = \text{education} + u,$$

then the model of Section 5.6.1 applies. The only difference here is that the true variable in the model is "latent." No amount of improvement in reporting or measurement would bring the proxy closer to the variable for which it is proxying.

The preceding is a pessimistic assessment, perhaps more so than necessary. Consider a **structural model,**

$$\textit{Earnings} = \beta_1 + \beta_2 \textit{ Experience} + \beta_3 \textit{ Industry} + \beta_4 \textit{ Ability} + \varepsilon$$

Ability is unobserved, but suppose that an indicator, say IQ is. If we suppose that IQ is related to *Ability* through a relationship such as

$$IQ = \alpha_1 + \alpha_2 \textit{ Ability} + v$$

then we may solve the second equation for *Ability* and insert it in the first to obtain the **reduced form equation**

$$\textit{Earnings} = (\beta_1 - \alpha_1/\alpha_2) + \beta_2 \textit{ Experience} + \beta_3 \textit{ Industry} + (\beta_4/\alpha_2)IQ + (\varepsilon - v/\alpha_2).$$

This equation is intrinsically linear and can be estimated by least squares. We do not have a consistent estimator of β_1 or β_4, but we do have one of the coefficients of interest. This would appear to "solve" the problem. We should note the essential ingredients; we require that the **indicator,** IQ, not be related to the other variables in the model, and we also require that v not be correlated with any of the variables. In this instance, some of the parameters of the structural model are identified in terms of observable data. Note, though, that IQ is not a proxy variable, it is an indicator of the latent variable, *Ability*. This form of modeling has figured prominently in the education and educational psychology literature. Consider, in the preceding small model how one might proceed with not just a single indicator, but say with a battery of test scores, all of which are indicators of the same latent ability variable.

It is to be emphasized that a proxy variable is not an instrument (or the reverse). Thus, in the instrumental variables framework, it is implied that we do not regress $\mathbf{y}$ on $\mathbf{Z}$ to obtain the estimates. To take an extreme example, suppose that the full model was

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U},$$

$$\mathbf{Z} = \mathbf{X}^* + \mathbf{W}.$$

That is, we happen to have two badly measured estimates of $\mathbf{X}^*$. The parameters of this model can be estimated without difficulty if $\mathbf{W}$ is uncorrelated with $\mathbf{U}$ and $\mathbf{X}^*$, *but not by regressing* $\mathbf{y}$ on $\mathbf{Z}$. The instrumental variables technique is called for.

When the model contains a variable such as education or ability, the question that naturally arises is, If interest centers on the other coefficients in the model, why not just discard the problem variable?[14] This method produces the familiar problem of an omitted variable, compounded by the least squares estimator in the full model being inconsistent anyway. Which estimator is worse? McCallum (1972) and Wickens (1972) show that the asymptotic bias (actually, degree of inconsistency) is worse if the proxy is omitted, even if it is a bad one (has a high proportion of measurement error). This proposition neglects, however, the precision of the estimates. Aigner (1974) analyzed this aspect of the problem and found, as might be expected, that it could go either way. He concluded, however, that "there is evidence to broadly support use of the proxy."

5.6.4 APPLICATION: INCOME AND EDUCATION AND A STUDY OF TWINS

The traditional model used in labor economics to study the effect of education on income is an equation of the form

$$y_i = \beta_1 + \beta_2 \text{ age}_i + \beta_3 \text{ age}_i^2 + \beta_4 \text{ education}_i + \mathbf{x}_i'\boldsymbol{\beta}_5 + \varepsilon_i,$$

where y_i is typically a wage or yearly income (perhaps in log form) and $\mathbf{x}_i$ contains other variables, such as an indicator for sex, region of the country, and industry. The literature contains discussion of many possible problems in estimation of such an equation by least squares using measured data. Two of them are of interest here:

1. Although "education" is the variable that appears in the equation, the data available to researchers usually include only "years of schooling." This variable is a proxy for education, so an equation fit in this form will be tainted by this problem of measurement error. Perhaps surprisingly so, researchers also find that reported data on years of schooling are themselves subject to error, so there is a second source of measurement error. For the present, we will not consider the first (much more difficult) problem.
2. Other variables, such as "ability"—we denote these μ_i—will also affect income and are surely correlated with education. If the earnings equation is estimated in the form shown above, then the estimates will be further biased by the absence of this "omitted variable." For reasons we will explore in Chapter 22, this bias has been called the selectivity effect in recent studies.

Simple cross-section studies will be considerably hampered by these problems. But, in a recent study, Ashenfelter and Krueger (1994) analyzed a data set that allowed them, with a few simple assumptions, to ameliorate these problems.

Annual "twins festivals" are held at many places in the United States. The largest is held in Twinsburg, Ohio. The authors interviewed about 500 individuals over the age of 18 at the August 1991 festival. Using pairs of twins as their observations enabled them to modify their model as follows: Let (y_{ij}, A_{ij}) denote the earnings and age for

[14] This discussion applies to the measurement error and latent variable problems equally.

twin j, $j = 1, 2$, for pair i. For the education variable, only self-reported "schooling" data, S_{ij}, are available. The authors approached the measurement problem in the schooling variable, S_{ij}, by asking each twin how much schooling they had and how much schooling their sibling had. Denote schooling reported *by* sibling m *of* sibling j by $S_{ij}(m)$. So, the self-reported years of schooling of twin 1 is $S_{i1}(1)$. When asked how much schooling twin 1 has, twin 2 reports $S_{i1}(2)$. The measurement error model for the schooling variable is

$$S_{ij}(m) = S_{ij} + u_{ij}(m), \quad j, m = 1, 2, \text{ where } S_{ij} = \text{"true" schooling for twin } j \text{ of pair } i.$$

We assume that the two sources of measurement error, $u_{ij}(m)$, are uncorrelated and have zero means. Now, consider a simple bivariate model such as the one in (5-25):

$$y_{ij} = \beta S_{ij} + \varepsilon_{ij}.$$

As we saw earlier, a least squares estimate of β using the reported data will be attenuated:

$$\text{plim } b = \frac{\beta \times \text{Var}[S_{ij}]}{\text{Var}[S_{ij}] + \text{Var}[u_{ij}(j)]} = \beta q.$$

(Since there is no natural distinction between twin 1 and twin 2, the assumption that the variances of the two measurement errors are equal is innocuous.) The factor q is sometimes called the **reliability ratio**. In this simple model, if the reliability ratio were known, then β could be consistently estimated. In fact, this construction of this model allows just that. Since the two measurement errors are uncorrelated,

$$\begin{aligned}\text{Corr}[S_{i1}(1), S_{i1}(2)] &= \text{Corr}[S_{i2}(1), S_{i2}(1)] \\ &= \frac{\text{Var}[S_{i1}]}{\{\{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(1)]\} \times \{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(2)]\}\}^{1/2}} = q.\end{aligned}$$

In words, the correlation between the two reported education attainments measures the reliability ratio. The authors obtained values of 0.920 and 0.877 for 298 pairs of identical twins and 0.869 and 0.951 for 92 pairs of fraternal twins, thus providing a quick assessment of the extent of measurement error in their schooling data.

Since the earnings equation is a multiple regression, this result is useful for an overall assessment of the problem, but the numerical values are not sufficient to undo the overall biases in the least squares regression coefficients. An instrumental variables estimator was used for that purpose. The estimating equation for $y_{ij} = \ln \text{Wage}_{ij}$ with the least squares (LS) and instrumental variable (IV) estimates is as follows:

$$y_{ij} = \beta_1 + \beta_2 \text{ age}_i + \beta_3 \text{ age}_i^2 + \beta_4 S_{ij}(j) + \beta_5 S_{im}(m) + \beta_6 \text{ sex}_i + \beta_7 \text{ race}_i + \varepsilon_{ij}$$

LS	(0.088)	(−0.087)	(0.084)		(0.204)	(−0.410)
IV	(0.088)	(−0.087)	(0.116)	(0.037)	(0.206)	(−0.428)

In the equation, $S_{ij}(j)$ is the person's report of his or her own years of schooling and $S_{im}(m)$ is the sibling's report of the sibling's own years of schooling. The problem variable is schooling. To obtain consistent estimates, the method of instrumental variables was used, using each sibling's report of the other sibling's years of schooling as a pair of instrumental variables. The estimates reported by the authors are shown below the equation. (The constant term was not reported, and for reasons not given, the second schooling variable was not included in the equation when estimated by LS.) This

preliminary set of results is presented to give a comparison to other results in the literature. The age, schooling, and gender effects are comparable with other received results, whereas the effect of race is vastly different, −40 percent here compared with a typical value of +9 percent in other studies. The effect of using the instrumental variable estimator on the estimates of β_4 is of particular interest. Recall that the reliability ratio was estimated at about 0.9, which suggests that the IV estimate would be roughly 11 percent higher (1/0.9). Since this result is a multiple regression, that estimate is only a crude guide. The estimated effect shown above is closer to 38 percent.

The authors also used a different estimation approach. Recall the issue of selection bias caused by unmeasured effects. The authors reformulated their model as

$$y_{ij} = \beta_1 + \beta_2 \text{ age}_i + \beta_3 \text{ age}_i^2 + \beta_4 S_{ij}(j) + \beta_6 \text{ sex}_i + \beta_7 \text{ race}_i + \mu_i + \varepsilon_{ij}$$

Unmeasured latent effects, such as "ability," are contained in μ_i. Since μ_i is not observable but is, it is assumed, correlated with other variables in the equation, the least squares regression of y_{ij} on the other variables produces a biased set of coefficient estimates. The difference between the two earnings equations is

$$y_{i1} - y_{i2} = \beta_4[S_{i1}(1) - S_{i2}(2)] + \varepsilon_{i1} - \varepsilon_{i2}.$$

This equation removes the latent effect but, it turns out, worsens the measurement error problem. As before, β_4 can be estimated by instrumental variables. There are two instrumental variables available, $S_{i2}(1)$ and $S_{i1}(2)$. (It is not clear in the paper whether the authors used the two separately or the difference of the two.) The least squares estimate is 0.092, which is comparable to the earlier estimate. The instrumental variable estimate is 0.167, which is nearly 82 percent higher. The two reported standard errors are 0.024 and 0.043, respectively. With these figures, it is possible to carry out Hausman's test;

$$H = \frac{(0.167 - 0.092)^2}{0.043^2 - 0.024^2} = 4.418.$$

The 95 percent critical value from the chi-squared distribution with one degree of freedom is 3.84, so the hypothesis that the LS estimator is consistent would be rejected. (The square root of H, 2.102, would be treated as a value from the standard normal distribution, from which the critical value would be 1.96. The authors reported a t statistic for this regression of 1.97. The source of the difference is unclear.)

5.7 SUMMARY AND CONCLUSIONS

This chapter has completed the description begun in Chapter 4 by obtaining the large sample properties of the least squares estimator. The main result is that in large samples, the estimator behaves according to a normal distribution and converges in probability to the true coefficient vector. We examined several data types, with one of the end results being that consistency and asymptotic normality would persist under a variety of broad assumptions about the data. We then considered a class of estimators, the instrumental variable estimators, which will retain the important large sample properties we found earlier, consistency and asymptotic normality, in cases in which the least squares estima-

tor is inconsistent. Two common applications include dynamic models, including panel data models, and models of measurement error.

Key Terms and Concepts

- Asymptotic distribution
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic covariance matrix
- Asymptotic properties
- Attenuation
- Consistency
- Dynamic regression
- Efficient scale
- Ergodic
- Finite sample properties
- Grenander conditions
- Hausman's specification test
- Identification
- Indicator
- Instrumental variable
- Lindberg–Feller central limit theorem
- Maximum likelihood estimator
- Mean square convergence
- Measurement error
- Panel data
- Probability limit
- Reduced form equation
- Reliability ratio
- Specification test
- Stationary process
- Stochastic regressors
- Structural model
- Two stage least squares

Exercises

1. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, what is plim $F[K, n-K] = \text{plim } \frac{R^2/K}{(1-R^2)/(n-K)}$, assuming that the true value of $\boldsymbol{\beta}$ is zero?
2. Let e_i be the ith residual in the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ in the classical regression model, and let ε_i be the corresponding true disturbance. Prove that $\text{plim}(e_i - \varepsilon_i) = 0$.
3. For the simple regression model $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N[0, \sigma^2]$, prove that the sample mean is consistent and asymptotically normally distributed. Now consider the alternative estimator $\hat{\mu} = \sum_i w_i y_i$, $w_i = \frac{i}{(n(n+1)/2)} = \frac{i}{\sum_i i}$. Note that $\sum_i w_i = 1$. Prove that this is a consistent estimator of μ and obtain its asymptotic variance. [Hint: $\sum_i i^2 = n(n+1)(2n+1)/6$.]
4. In the discussion of the instrumental variables estimator we showed that the least squares estimator $\mathbf{b}$ is biased and inconsistent. Nonetheless, $\mathbf{b}$ does estimate something: $\text{plim}\,\mathbf{b} = \boldsymbol{\theta} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\boldsymbol{\gamma}$. Derive the asymptotic covariance matrix of $\mathbf{b}$, and show that $\mathbf{b}$ is asymptotically normally distributed.
5. For the model in (5-25) and (5-26), prove that when only x^* is measured with error, the squared correlation between y and x is less than that between y^* and x^*. (Note the assumption that $y^* = y$.) Does the same hold true if y^* is also measured with error?
6. Christensen and Greene (1976) estimated a generalized Cobb–Douglas cost function of the form

$$\ln(C/P_f) = \alpha + \beta \ln Q + \gamma(\ln^2 Q)/2 + \delta_k \ln(P_k/P_f) + \delta_l \ln(P_l/P_f) + \varepsilon.$$

P_k, P_l and P_f indicate unit prices of capital, labor, and fuel, respectively, Q is output and C is total cost. The purpose of the generalization was to produce a U-shaped average total cost curve. (See Example 7.3 for discussion of Nerlove's (1963) predecessor to this study.) We are interested in the output at which the cost curve reaches its minimum. That is the point at which $(\partial \ln C/\partial \ln Q)_{|Q=Q^*} = 1$ or $Q^* = \exp[(1-\beta)/\gamma]$. The estimated regression model using the Christensen

and Greene 1970 data are as follows, where estimated standard errors are given in parentheses:

$$\ln(C/P_f) = -7.294 + 0.39091 \ln Q + 0.062413(\ln^2 Q)/2 + 0.07479 \ln(P_k/P_f) + 0.2608 \ln(P_l/P_f) + e.$$
$$\text{(0.34427)} \quad \text{(0.036988)} \quad \text{(0.0051548)} \quad \text{(0.061645)} \quad \text{(0.068109)}$$

The estimated asymptotic covariance of the estimators of β and γ is -0.000187067, $R^2 = 0.991538$ and $\mathbf{e}'\mathbf{e} = 2.443509$. Using the estimates given above, compute the estimate of this **efficient scale.** Compute an estimate of the asymptotic standard error for this estimate, then form a confidence interval for the estimated efficient scale. The data for this study are given in Table F5.2. Examine the raw data and determine where in the sample the efficient scale lies. That is, how many firms in the sample have reached this scale, and is this scale large in relation to the sizes of firms in the sample?

7. The consumption function used in Example 5.3 is a very simple specification. One might wonder if the meager specification of the model could help explain the finding in the Hausman test. The data set used for the example are given in Table F5.1. Use these data to carry out the test in a more elaborate specification

$$c_t = \beta_1 + \beta_2 y_t + \beta_3 i_t + \beta_4 c_{t-1} + \varepsilon_t$$

where c_t is the log of real consumption, y_t is the log of real disposable income, and i_t is the interest rate (90-day T bill rate).

8. Suppose we change the assumptions of the model to **AS5:** $(\mathbf{x}_i, \varepsilon)$ are an independent and identically distributed sequence of random vectors such that $\mathbf{x}_i$ has a finite mean vector, $\boldsymbol{\mu}_{\mathbf{x}}$, finite positive definite covariance matrix $\boldsymbol{\Sigma}_{\mathbf{xx}}$ and finite fourth moments $E[x_j x_k x_l x_m] = \phi_{jklm}$ for all variables. How does the proof of consistency and asymptotic normality of $\mathbf{b}$ change? Are these assumptions weaker or stronger than the ones made in Section 5.2?

9. Now, assume only finite second moments of $\mathbf{x}$; $E[x_i^2]$ is finite. Is this sufficient to establish consistency of $\mathbf{b}$? (Hint: the Cauchy–Schwartz inequality (Theorem D.13), $E[|xy|] \leq \{E[x^2]\}^{1/2}\{E[y^2]\}^{1/2}$ will be helpful.) Is this assumption sufficient to establish asymptotic normality?