Chapter Five

Multiple Regression Analysis: OLS Asymptotics

In Chapters 3 and 4, we covered what are called *finite sample*, *small sample*, or *exact* properties of the OLS estimators in the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \tag{5.1}$$

For example, the unbiasedness of OLS (derived in Chapter 3) under the first four Gauss-Markov assumptions is a finite sample property because it holds for *any* sample size n (subject to the mild restriction that n must be at least as large as the total number of parameters in the regression model, $k + 1$). Similarly, the fact that OLS is the best linear unbiased estimator under the full set of Gauss-Markov assumptions (MLR.1 through MLR.5) is a finite sample property.

In Chapter 4, we added the classical linear model Assumption MLR.6, which states that the error term u is normally distributed and independent of the explanatory variables. This allowed us to derive the *exact* sampling distributions of the OLS estimators (conditional on the explanatory variables in the sample). In particular, Theorem 4.1 showed that the OLS estimators have normal sampling distributions, which led directly to the t and F distributions for t and F statistics. If the error is not normally distributed, the distribution of a t statistic is not exactly t, and an F statistic does not have an exact F distribution for any sample size.

In addition to finite sample properties, it is important to know the **asymptotic properties** or **large sample properties** of estimators and test statistics. These properties are not defined for a particular sample size; rather, they are defined as the sample size grows without bound. Fortunately, under the assumptions we have made, OLS has satisfactory large sample properties. One practically important finding is that even without the normality assumption (Assumption MLR.6), t and F statistics have *approximately* t and F distributions, at least in large sample sizes. We discuss this in more detail in Section 5.2, after we cover consistency of OLS in Section 5.1.

5.1 CONSISTENCY

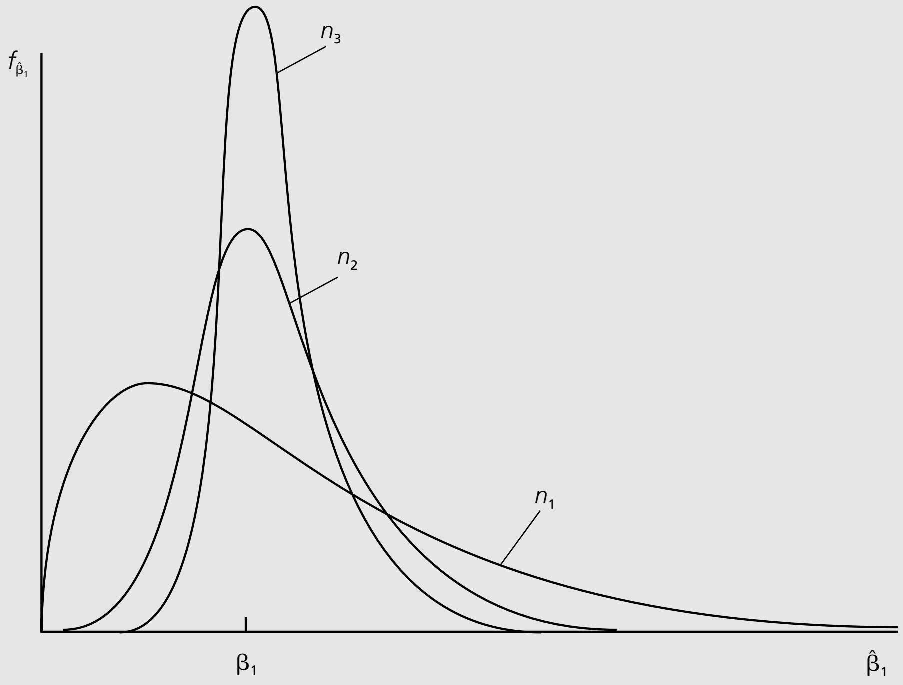
Unbiasedness of estimators, while important, cannot always be achieved. For example, as we discussed in Chapter 3, the standard error of the regression, $\hat{\sigma}$, is not an unbiased estimator for σ, the standard deviation of the error u in a multiple regression model. While the OLS estimators are unbiased under MLR.1 through MLR.4, in Chapter 11 we will find that there are time series regressions where the OLS estimators are not unbiased. Further, in Part 3 of the text, we encounter several other estimators that are biased.

While not all useful estimators are unbiased, virtually all economists agree that **consistency** is a minimal requirement for an estimator. The famous econometrician Clive W.J. Granger once remarked: "If you can't get it right as n goes to infinity, you shouldn't be in this business." The implication is that, if your estimator of a particular population parameter is not consistent, then you are wasting your time.

There are a few different ways to describe consistency. Formal definitions and results are given in Appendix C; here we focus on an intuitive understanding. For concreteness, let $\hat{\beta}_j$ be the OLS estimator of β_j for some j. For each n, $\hat{\beta}_j$ has a probability distribution (representing its possible values in different random samples of size n). Because $\hat{\beta}_j$ is unbiased under assumptions MLR.1 through MLR.4, this distribution has mean value β_j. If this estimator is consistent, then the distribution of $\hat{\beta}_j$ becomes more and more tightly distributed around β_j as the sample size grows. As n tends to infinity, the distribution of $\hat{\beta}_j$ collapses to the single point β_j. In effect, this means that we can make our estimator arbitrarily close to β_j if we can collect as much data as we want. This convergence is illustrated in Figure 5.1.

Figure 5.1

Sampling distributions of $\hat{\beta}_1$ for sample sizes $n_1 < n_2 < n_3$.



Naturally, for any application we have a fixed sample size, which is the reason an asymptotic property such as consistency can be difficult to grasp. Consistency involves a thought experiment about what would happen as the sample size gets large (while at the same time we obtain numerous random samples for each sample size). If obtaining more and more data does not generally get us closer to the parameter value of interest, then we are using a poor estimation procedure.

Conveniently, the same set of assumptions imply both unbiasedness and consistency of OLS. We summarize with a theorem.

THEOREM 5.1 (CONSISTENCY OF OLS)

Under assumptions MLR.1 through MLR.4, the OLS estimator $\hat{\beta}_j$ is consistent for β_j, for all $j = 0,1, \ldots, k$.

A general proof of this result is most easily developed using the matrix algebra methods described in Appendices D and E. But we can prove Theorem 5.1 without difficulty in the case of the simple regression model. We focus on the slope estimator, $\hat{\beta}_1$.

The proof starts out the same as the proof of unbiasedness: we write down the formula for $\hat{\beta}_1$, and then plug in $y_i = \beta_0 + \beta_1 x_{i1} + u_i$:

$$\begin{aligned}\hat{\beta}_1 &= \left(\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)y_i\right)\Big/\left(\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2\right) \\ &= \beta_1 + \left(n^{-1}\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)u_i\right)\Big/\left(n^{-1}\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2\right).\end{aligned} \tag{5.2}$$

We can apply the law of large numbers to the numerator and denominator, which converge in probability to the population quantities, $\text{Cov}(x_1,u)$ and $\text{Var}(x_1)$, respectively. Provided that $\text{Var}(x_1) \neq 0$—which is assumed in MLR.4—we can use the properties of *probability limits* (see Appendix C) to get

$$\begin{aligned}\text{plim } \hat{\beta}_1 &= \beta_1 + \text{Cov}(x_1,u)/\text{Var}(x_1) \\ &= \beta_1, \text{ because } \text{Cov}(x_1,u) = 0.\end{aligned} \tag{5.3}$$

We have used the fact, discussed in Chapters 2 and 3, that $\text{E}(u|x_1) = 0$ implies that x_1 and u are uncorrelated (have zero covariance).

As a technical matter, to ensure that the probability limits exist, we should assume that $\text{Var}(x_1) < \infty$ and $\text{Var}(u) < \infty$ (which means that their probability distributions are not too spread out), but we will not worry about cases where these assumptions might fail.

The previous arguments, and equation (5.3) in particular, show that OLS is consistent in the simple regression case if we assume only zero correlation. This is also true in the general case. We now state this as an assumption.

ASSUMPTION MLR.3′ (ZERO MEAN AND ZERO CORRELATION)

$\text{E}(u) = 0$ and $\text{Cov}(x_j,u) = 0$, for $j = 1,2, \ldots, k$.

In Chapter 3, we discussed why assumption MLR.3 implies MLR.3′, but not vice versa. The fact that OLS is consistent under the weaker assumption MLR.3′ turns out to be useful in Chapter 15 and in other situations. Interestingly, while OLS is unbiased under MLR.3, this is not the case under Assumption MLR.3′. (This was the leading reason we have assumed MLR.3.)

Deriving the Inconsistency in OLS

Just as failure of $\mathrm{E}(u|x_1, \dots, x_k) = 0$ causes bias in the OLS estimators, correlation between u and *any* of $x_1, x_2, \dots, x_k$ generally causes *all* of the OLS estimators to be inconsistent. This simple but important observation is often summarized as: *if the error is correlated with any of the independent variables, then OLS is biased and inconsistent*. This is very unfortunate because it means that any bias persists as the sample size grows.

In the simple regression case, we can obtain the inconsistency from equation (5.3), which holds whether or not u and x_1 are uncorrelated. The **inconsistency** in $\hat{\beta}_1$ (sometimes loosely called the **asymptotic bias**) is

$$\text{plim } \hat{\beta}_1 - \beta_1 = \mathrm{Cov}(x_1,u)/\mathrm{Var}(x_1). \tag{5.4}$$

Because $\mathrm{Var}(x_1) > 0$, the inconsistency in $\hat{\beta}_1$ is positive if x_1 and u are positively correlated, and the inconsistency is negative if x_1 and u are negatively correlated. If the covariance between x_1 and u is small relative to the variance in x_1, the inconsistency can be negligible; unfortunately, we cannot even estimate how big the covariance is because u is unobserved.

We can use (5.4) to derive the asymptotic analog of the omitted variable bias (see Table 3.2 in Chapter 3). Suppose the true model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v,$$

satisfies the first four Gauss-Markov assumptions. Then v has a zero mean and is uncorrelated with x_1 and x_2. If $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ denote the OLS estimators from the regression of y on x_1 and x_2, then Theorem 5.1 implies that these estimators are consistent. If we omit x_2 from the regression and do the simple regression of y on x_1, then $u = \beta_2 x_2 + v$. Let $\tilde{\beta}_1$ denote the simple regression slope estimator. Then

$$\text{plim } \tilde{\beta}_1 = \beta_1 + \beta_2 \delta_1 \tag{5.5}$$

where

$$\delta_1 = \mathrm{Cov}(x_1,x_2)/\mathrm{Var}(x_1). \tag{5.6}$$

Thus, for practical purposes, we can view the inconsistency as being the same as the bias. The difference is that the inconsistency is expressed in terms of the population variance of x_1 and the population covariance between x_1 and x_2, while the bias is based on their sample counterparts (because we condition on the values of x_1 and x_2 in the sample).

If x_1 and x_2 are uncorrelated (in the population), then $\delta_1 = 0$, and $\tilde{\beta}_1$ is a consistent estimator of β_1 (although not necessarily unbiased). If x_2 has a positive partial effect on y, so that $\beta_2 > 0$, and x_1 and x_2 are positively correlated, so that $\delta_1 > 0$, then the inconsistency in $\tilde{\beta}_1$ is positive. And so on. We can obtain the direction of the inconsistency or asymptotic bias from Table 3.2. If the covariance between x_1 and x_2 is small relative to the variance of x_1, the inconsistency can be small.

EXAMPLE 5.1
(Housing Prices and Distance from an Incinerator)

Let y denote the price of a house (*price*), let x_1 denote the distance from the house to a new trash incinerator (*distance*), and let x_2 denote the "quality" of the house (*quality*). The variable *quality* is left vague so that it can include things like size of the house and lot, number of bedrooms and bathrooms, and intangibles such as attractiveness of the neighborhood. If the incinerator depresses house prices, then β_1 should be positive: everything else being equal, a house that is farther away from the incinerator is worth more. By definition, β_2 is positive since higher quality houses sell for more, other factors being equal. If the incinerator was built farther away, on average, from better homes, then *distance* and *quality* are positively correlated, and so $\delta_1 > 0$. A simple regression of *price* on *distance* [or log(*price*) on log(*distance*)] will tend to overestimate the effect of the incinerator: $\beta_1 + \beta_2\delta_1 > \beta_1$.

An important point about inconsistency in OLS estimators is that, by definition, the problem does not go away by adding more observations to the sample. If anything, the problem gets worse with more data: the OLS estimator gets closer and closer to $\beta_1 + \beta_2\delta_1$ as the sample size grows.

Deriving the sign and magnitude of the inconsistency in the general k regressor case is much harder, just as deriving the bias is very difficult. We need to remember that if we have the model in equation (5.1) where, say, x_1 is correlated with u but the other independent variables are uncorrelated with u, *all* of the OLS estimators are generally inconsistent. For example, in the $k = 2$ case,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

suppose that x_2 and u are uncorrelated but x_1 and u are correlated. Then the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ will generally both be inconsistent. (The intercept will also be inconsistent.) The inconsistency in $\hat{\beta}_2$ arises when x_1 and x_2 are correlated, as is usually the case. If x_1 and x_2 are uncorrelated, then any correlation between x_1 and u does *not* result in the inconsistency of $\hat{\beta}_2$: plim $\hat{\beta}_2 = \beta_2$. Further, the inconsistency in $\hat{\beta}_1$ is the same as in (5.4). The same statement holds in the general case: if x_1 is correlated with u, but x_1 and u are uncorrelated with the other independent variables, then only $\hat{\beta}_1$ is inconsistent, and the inconsistency is given by (5.4).

QUESTION 5.1

Suppose that the model

$$score = \beta_0 + \beta_1 skipped + \beta_2 priGPA + u$$

satisfies the first four Gauss-Markov assumptions, where *score* is score on a final exam, *skipped* is number of classes skipped, and *priGPA* is GPA prior to the current semester. If $\tilde{\beta}_1$ is from the simple regression of *score* on *skipped*, what is the direction of the asymptotic bias in $\tilde{\beta}_1$?

5.2 ASYMPTOTIC NORMALITY AND LARGE SAMPLE INFERENCE

Consistency of an estimator is an important property, but it alone does not allow us to perform statistical inference. Simply knowing that the estimator is getting closer to the population value as the sample size grows does not allow us to test hypotheses about the parameters. For testing, we need the sampling distribution of the OLS estimators. Under the classical linear model assumptions MLR.1 through MLR.6, Theorem 4.1 shows that the sampling distributions are normal. This result is the basis for deriving the *t* and *F* distributions that we use so often in applied econometrics.

The exact normality of the OLS estimators hinges crucially on the normality of the distribution of the error, *u*, in the population. If the errors u_1, u_2, …, u_n are random draws from some distribution other than the normal, the $\hat{\beta}_j$ will not be normally distributed, which means that the *t* statistics will not have *t* distributions and the *F* statistics will not have *F* distributions. This is a potentially serious problem because our inference hinges on being able to obtain critical values or *p*-values from the *t* or *F* distributions.

Recall that Assumption MLR.6 is equivalent to saying that the distribution of *y* given $x_1, x_2, \ldots, x_k$ is normal. Since *y* is observed and *u* is not, in a particular application, it is much easier to think about whether the distribution of *y* is likely to be normal. In fact, we have already seen a few examples where *y* definitely cannot have a normal distribution. A normally distributed random variable is symmetrically distributed about its mean, it can take on any positive or negative value (but with zero probability), and more than 95% of the area under the distribution is within two standard deviations.

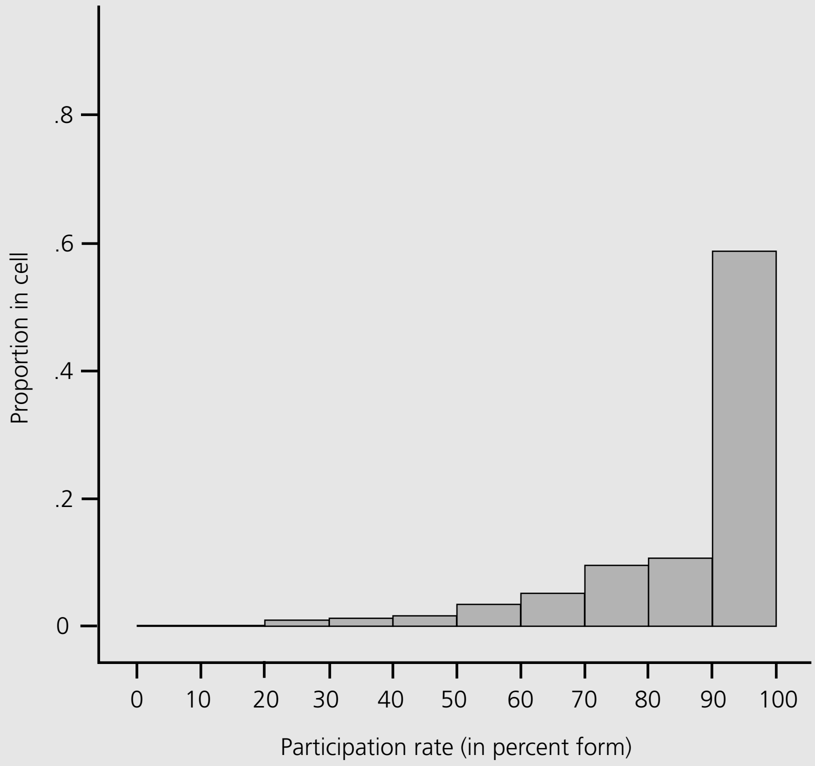
In Example 3.4, we estimated a model explaining the number of arrests of young men during a particular year (*narr86*). In the population, most men are not arrested during the year, and the vast majority are arrested one time at the most. (In the sample of 2,725 men in the data set CRIME1.RAW, fewer than 8% were arrested more than once during 1986.) Because *narr86* takes on only two values for 92% of the sample, it cannot be close to being normally distributed in the population.

In Example 4.6, we estimated a model explaining participation percentages (*prate*) in 401(k) pension plans. The frequency distribution (also called a *histogram*) in Figure 5.2 shows that the distribution of *prate* is heavily skewed to the right, rather than being normally distributed. In fact, over 40% of the observations on *prate* are at the value 100, indicating 100% participation. This violates the normality assumption even conditional on the explanatory variables.

We know that normality plays no role in the unbiasedness of OLS, nor does it affect the conclusion that OLS is the best linear unbiased estimator under the Gauss-Markov assumptions. But exact inference based on *t* and *F* statistics requires MLR.6. Does this mean that, in our analysis of *prate* in Example 4.6, we must abandon the *t* statistics for determining which variables are statistically significant? Fortunately, the answer to this question is *no*. Even though the y_i are not from a normal distribution, we can use the central limit theorem from Appendix C to conclude that the OLS estimators are *approximately* normally distributed, at least in large sample sizes.

Figure 5.2

Histogram of *prate* using the data in 401K.RAW.



THEOREM 5.2 (ASYMPTOTIC NORMALITY OF OLS)

Under the Gauss-Markov assumptions MLR.1 through MLR.5,

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} \text{Normal}(0,\sigma^2/a_j^2)$, where $\sigma^2/a_j^2 > 0$ is the **asymptotic variance** of $\sqrt{n}(\hat{\beta}_j - \beta_j)$; for the slope coefficients, $a_j^2 = \text{plim}\left(n^{-1}\sum_{i=1}^{n} \hat{r}_{ij}^2\right)$, where the $\hat{r}_{ij}$ are the residuals from regressing x_j on the other independent variables. We say that $\hat{\beta}_j$ is *asymptotically normally distributed* (see Appendix C);

(ii) $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 = \text{Var}(u)$;

(iii) For each j,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \overset{a}{\sim} \text{Normal}(0,1), \qquad \textbf{(5.7)}$$

where $\text{se}(\hat{\beta}_j)$ is the usual OLS standard error.

The proof of asymptotic normality is somewhat complicated and is sketched in the appendix for the simple regression case. Part (ii) follows from the law of large numbers, and part (iii) follows from parts (i) and (ii) and the asymptotic properties discussed in Appendix C.

Thorem 5.2 is useful because the normality assumption MLR.6 has been dropped; the only restriction on the distribution of the error is that it has finite variance, something we will always assume. We have also assumed zero conditional mean and homoskedasticity of u.

Notice how the standard normal distribution appears in (5.7), as opposed to the t_{n-k-1} distribution. This is because the distribution is only approximate. By contrast, in Theorem 4.2, the distribution of the ratio in (5.7) was *exactly* t_{n-k-1} for any sample size. From a practical perspective, this difference is irrelevant. In fact, it is just as legitimate to write

$$(\hat{\beta}_j - \beta_j)/\mathrm{se}(\hat{\beta}_j) \overset{a}{\sim} t_{n-k-1}, \tag{5.8}$$

since t_{n-k-1} approaches the standard normal distribution as the degrees of freedom gets large.

Equation (5.8) tells us that t testing and the construction of confidence intervals are carried out *exactly* as under the classical linear model assumptions. This means that our analysis of dependent variables like *prate* and *narr86* does not have to change at all if the Gauss-Markov assumptions hold: in both cases, we have at least 1,500 observations, which is certainly enough to justify the approximation of the central limit theorem.

If the sample size is not very large, then the t distribution can be a poor approximation to the distribution of the t statistics when u is not normally distributed. Unfortunately, there are no general prescriptions on how big the sample size must be before the approximation is good enough. Some econometricians think that $n = 30$ is satisfactory, but this cannot be sufficient for all possible distributions of u. Depending on the distribution of u, more observations may be necessary before the central limit theorem takes effect. Further, the quality of the approximation depends not just on n, but on the *df*, $n - k - 1$: with more independent variables in the model, a larger sample size is usually needed to use the t approximation. Methods for inference with small degrees of freedom and nonnormal errors are outside the scope of this text. We will simply use the t statistics as we always have without worrying about the normality assumption.

It is very important to see that Theorem 5.2 *does* require the homoskedasticity assumption (along with the zero conditional mean assumption). If $\mathrm{Var}(y|\boldsymbol{x})$ is not constant, the usual t statistics and confidence intervals are invalid no matter how large the sample size is; the central limit theorem does not bail us out when it comes to heteroskedasticity. For this reason, we devote all of Chapter 8 to discussing what can be done in the presence of heteroskedasticity.

One conclusion of Theorem 5.2 is that $\hat{\sigma}^2$ is a consistent estimator of σ^2; we already know from Theorem 3.3 that $\hat{\sigma}^2$ is unbiased for σ^2 under the Gauss-Markov assumptions. The consistency implies that $\hat{\sigma}$ is a consistent estimator of σ, which is important in establishing the asymptotic normality result in equation (5.7).

Remember that $\hat{\sigma}$ appears in the standard error for each $\hat{\beta}_j$. In fact, the estimated variance of $\hat{\beta}_j$ is

$$\hat{\text{Var}}\,(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\text{SST}_j(1 - R_j^2)}, \tag{5.9}$$

where SST_j is the total sum of squares of x_j in the sample, and R_j^2 is the R-squared from regressing x_j on all of the other independent variables. In Section 3.4, we studied each component of (5.9), which we will now expound on in the context of asymptotic analysis. As the sample size grows, $\hat{\sigma}^2$ converges in probability to the constant σ^2. Further, R_j^2 approaches a number strictly between zero and unity (so that $1 - R_j^2$ converges to some number between zero and one). The sample variance of x_j is SST_j/n, and so SST_j/n converges to $\text{Var}(x_j)$ as the sample size grows. This means that SST_j grows at approximately the same rate as the sample size: $\text{SST}_j \approx n\sigma_j^2$, where σ_j^2 is the population variance of x_j. When we combine these facts, we find that $\hat{\text{Var}}(\hat{\beta}_j)$ shrinks to zero at the rate of $1/n$; this is why larger sample sizes are better.

Q U E S T I O N 5 . 2

In a regression model with a large sample size, what is an approximate 95% confidence interval for $\hat{\beta}_j$ under MLR.1 through MLR.5? We call this an **asymptotic confidence interval**.

When u is not normally distributed, the square root of (5.9) is sometimes called the **asymptotic standard error**, and t statistics are called **asymptotic *t* statistics**. Because these are the same quantities we dealt with in Chapter 4, we will just call them standard errors and t statistics, with the understanding that sometimes they have only large sample justification.

Using the preceding argument about the estimated variance, we can write

$$\text{se}(\hat{\beta}_j) \approx c_j/\sqrt{n}\,, \tag{5.10}$$

where c_j is a positive constant that does *not* depend on the sample size. Equation (5.10) is only an approximation, but it is a useful rule of thumb: standard errors can be expected to shrink at a rate that is the inverse of the *square root* of the sample size.

EXAMPLE 5.2
(Standard Errors in a Birth Weight Equation)

We use the data in BWGHT.RAW to estimate a relationship where log of birth weight is the dependent variable, and cigarettes smoked per day (*cigs*) and log of family income log(*faminc*) are independent variables. The total number of observations is 1,388. Using the first half of the observations (694), the standard error for $\hat{\beta}_{cigs}$ is about .0013. The standard error using all of the observations is about .00086. The ratio of the latter standard error to the former is $.00086/.0013 \approx .662$. This is pretty close to $\sqrt{694/1{,}388} \approx .707$, the ratio obtained from the approximation in (5.10). In other words, equation (5.10) implies that the standard error using the larger sample size should be about 70.7% of the standard error using the smaller sample. This percentage is pretty close to the 66.2% we actually compute from the ratio of the standard errors.

The asymptotic normality of the OLS estimators also implies that the *F* statistics have approximate *F* distributions in large sample sizes. Thus, for testing exclusion restrictions or other multiple hypotheses, nothing changes from what we have done before.

Other Large Sample Tests: The Lagrange Multiplier Statistic

Once we enter the realm of asymptotic analysis, there are other test statistics that can be used for hypothesis testing. For most purposes, there is little reason to go beyond the usual *t* and *F* statistics: as we just saw, these statistics have large sample justification without the normality assumption. Nevertheless, sometimes it is useful to have other ways to test multiple exclusion restrictions, and we now cover the **Lagrange multiplier LM statistic**, which has achieved some popularity in modern econometrics.

The name "Lagrange multiplier statistic" comes from constrained optimization, a topic beyond the scope of this text. [See Davidson and MacKinnon (1993).] The name **score statistic**—which also comes from optimization using calculus—is used as well. Fortunately, in the linear regression framework, it is simple to motivate the *LM* statistic without delving into complicated mathematics.

The form of the *LM* statistic we derive here relies on the Gauss-Markov assumptions, the same assumptions that justify the *F* statistic in large samples. We do not need the normality assumption.

To derive the *LM* statistic, consider the usual multiple regression model with *k* independent variables:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u. \tag{5.11}$$

We would like to test whether, say, the last *q* of these variables all have zero population parameters: the null hypothesis is

$$\mathrm{H}_0\colon \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \tag{5.12}$$

which puts *q* exclusion restrictions on the model (5.11). As with *F* testing, the alternative to (5.12) is that at least one of the parameters is different from zero.

The *LM* statistic requires estimation of the *restricted* model only. Thus, assume that we have run the regression

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \ldots + \tilde{\beta}_{k-q} x_{k-q} + \tilde{u}, \tag{5.13}$$

where "~" indicates that the estimates are from the restricted model. In particular, $\tilde{u}$ indicates the residuals from the restricted model. (As always, this is just shorthand to indicate that we obtain the restricted residual for each observation in the sample.)

If the omitted variables x_{k-q+1} through x_k truly have zero population coefficients then, at least approximately, $\tilde{u}$ should be uncorrelated with each of these variables in the sample. This suggests running a regression of these residuals on those independent variables excluded under H_0, which is almost what the *LM* test does. However, it turns out that, to get a usable test statistic, we must include *all* of the independent variables in the regression (the reasons for this are technical and unimportant). Thus, we run the regression

$$\tilde{u} \text{ on } x_1, x_2, \dots, x_k. \tag{5.14}$$

This is an example of an **auxiliary regression**, a regression that is used to compute a test statistic but whose coefficients are not of direct interest.

How can we use the regression output from (5.14) to test (5.12)? If (5.12) is true, the *R*-squared from (5.14) should be "close" to zero, subject to sampling error, because $\tilde{u}$ will be approximately uncorrelated with all the independent variables. The question, as always with hypothesis testing, is how to determine when the statistic is large enough to reject the null hypothesis at a chosen significance level. It turns out that, under the null hypothesis, the sample size multiplied by the usual *R*-squared from the auxiliary regression (5.14) is distributed asymptotically as a chi-square random variable with q degrees of freedom. This leads to a simple procedure for testing the joint significance of a set of q independent variables.

THE LAGRANGE MULTIPLIER STATISTIC FOR *q* EXCLUSION RESTRICTIONS:

(i) Regress y on the *restricted* set of independent variables and save the residuals, $\tilde{u}$.

(ii) Regress $\tilde{u}$ on *all* of the independent variables and obtain the *R*-squared, say R_u^2 (to distinguish it from the *R*-squareds obtained with y as the dependent variable).

(iii) Compute $LM = nR_u^2$ [the sample size times the *R*-squared obtained from step (ii)].

(iv) Compare LM to the appropriate critical value, c, in a χ_q^2 distribution; if $LM > c$, the null hypothesis is rejected. Even better, obtain the p-value as the probability that a χ_q^2 random variable exceeds the value of the test statistic. If the p-value is less than the desired significance level, then H_0 is rejected. If not, we fail to reject H_0. The rejection rule is essentially the same as for F testing.

Because of its form, the LM statistic is sometimes referred to as the **n-*R*-squared statistic**. Unlike with the F statistic, the degrees of freedom in the unrestricted model plays no role in carrying out the LM test. All that matters is the number of restrictions being tested (q), the size of the auxiliary *R*-squared (R_u^2), and the sample size (n). The *df* in the unrestricted model plays no role because of the asymptotic nature of the LM statistic. But we must be sure to multiply R_u^2 by the sample size to obtain LM; a seemingly low value of the *R*-squared can still lead to joint significance if n is large.

Before giving an example, a word of caution is in order. If in step (i), we mistakenly regress y on all of the independent variables and obtain the residuals from this unrestricted regression to be used in step (ii), we do not get an interesting statistic: the resulting *R*-squared will be exactly zero! This is because OLS chooses the estimates so that the residuals are uncorrelated in samples with all included independent variables [see equations (3.13)]. Thus, we can only test (5.12) by regressing the restricted residuals on *all* of the independent variables. (Regressing the restricted residuals on the restricted set of independent variables will also produce $R^2 = 0$.)

EXAMPLE 5.3
(Economic Model of Crime)

We illustrate the *LM* test by using a slight extension of the crime model from Example 3.4:

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

where *narr86* is the number of times a man was arrested, *pcnv* is the proportion of prior arrests leading to conviction, *avgsen* is average sentence served from past convictions, *tottime* is total time the man has spent in prison prior to 1986 since reaching the age of 18, *ptime86* is months spent in prison in 1986, and *qemp86* is number of quarters in 1986 during which the man was legally employed. We use the *LM* statistic to test the null hypothesis that *avgsen* and *tottime* have no effect on *narr86* once the other factors have been controlled for.

In step (i), we estimate the restricted model by regressing *narr86* on *pcnv*, *ptime86*, and *qemp86*; the variables *avgsen* and *tottime* are excluded from this regression. We obtain the residuals $\tilde{u}$ from this regression, 2,725 of them. Next, we run the regression

$$\tilde{u} \text{ on } pcnv, ptime86, qemp86, avgsen, \text{ and } tottime; \tag{5.15}$$

as always, the order in which we list the independent variables is irrelevant. This second regression produces $R^2_{\tilde{u}}$, which turns out to be about .0015. This may seem small, but we must multiply it by *n* to get the *LM* statistic: $LM = 2{,}725(.0015) \approx 4.09$. The 10% critical value in a chi-square distribution with two degrees of freedom is about 4.61 (rounded to two decimal places; see Table G.4). Thus, we fail to reject the null hypothesis that $\beta_{avgsen} = 0$ and $\beta_{tottime} = 0$ at the 10% level. The *p*-value is $P(\chi^2_2 > 4.09) \approx .129$, so we would reject H_0 at the 15% level.

As a comparison, the *F* test for joint significance of *avgsen* and *tottime* yields a *p*-value of about .131, which is pretty close to that obtained using the *LM* statistic. This is not surprising since, asymptotically, the two statistics have the same probability of Type I error. (That is, they reject the null hypothesis with the same frequency when the null is true.)

As the previous example suggests, with a large sample, we rarely see important discrepancies between the outcomes of *LM* and *F* tests. We will use the *F* statistic for the most part because it is computed routinely by most regression packages. But you should be aware of the *LM* statistic as it is used in applied work.

One final comment on the *LM* statistic. As with the *F* statistic, we must be sure to use the same observations in steps (i) and (ii). If data are missing for some of the independent variables that are excluded under the null hypothesis, the residuals from step (i) should be obtained from a regression on the reduced data set.

5.3 ASYMPTOTIC EFFICIENCY OF OLS

We know that, under the Gauss-Markov assumptions, the OLS estimators are best linear unbiased. OLS is also **asymptotically efficient** among a certain class of estimators under the Gauss-Markov assumptions. A general treatment is difficult [see Wooldridge (1999, Chapter 4)]. For now, we describe the result in the simple regression case.

In the model

$$y = \beta_0 + \beta_1 x + u, \tag{5.16}$$

u has a zero conditional mean under MLR.3: $\mathrm{E}(u|x) = 0$. This opens up a variety of consistent estimators for β_0 and β_1; as usual, we focus on the slope parameter, β_1. Let $g(x)$ be any function of x; for example, $g(x) = x^2$ or $g(x) = 1/(1 + |x|)$. Then u is uncorrelated with $g(x)$ (see Property CE.5 in Appendix B). Let $z_i = g(x_i)$ for all observations i. Then the estimator

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} (z_i - \bar{z})y_i\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z})x_i\right) \tag{5.17}$$

is consistent for β_1, provided $g(x)$ and x are correlated. (Remember, it is possible that $g(x)$ and x are uncorrelated because correlation measures *linear* dependence.) To see this, we can plug in $y_i = \beta_0 + \beta_1 x_i + u_i$ and write $\tilde{\beta}_1$ as

$$\tilde{\beta}_1 = \beta_1 + \left(n^{-1}\sum_{i=1}^{n} (z_i - \bar{z})u_i\right) \Big/ \left(n^{-1}\sum_{i=1}^{n} (z_i - \bar{z})x_i\right). \tag{5.18}$$

Now, we can apply the law of large numbers to the numerator and denominator, which converge in probability to $\mathrm{Cov}(z,u)$ and $\mathrm{Cov}(z,x)$, respectively. Provided that $\mathrm{Cov}(z,x) \neq 0$—so that z and x are correlated—we have

$$\mathrm{plim}\ \tilde{\beta}_1 = \beta_1 + \mathrm{Cov}(z,u)/\mathrm{Cov}(z,x) = \beta_1,$$

because $\mathrm{Cov}(z,u) = 0$ under MLR.3.

It is more difficult to show that $\tilde{\beta}_1$ is asymptotically normal. Nevertheless, using arguments similar to those in the appendix, it can be shown that $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ is asymptotically normal with mean zero and asymptotic variance $\sigma^2\mathrm{Var}(z)/[\mathrm{Cov}(z,x)]^2$. The asymptotic variance of the OLS estimator is obtained when $z = x$, in which case, $\mathrm{Cov}(z,x) = \mathrm{Cov}(x,x) = \mathrm{Var}(x)$. Therefore, the asymptotic variance of $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$, where $\hat{\beta}_1$ is the OLS estimator, is $\sigma^2\mathrm{Var}(x)/[\mathrm{Var}(x)]^2 = \sigma^2/\mathrm{Var}(x)$. Now, the Cauchy-Schwartz inequality (see Appendix B.4) implies that $[\mathrm{Cov}(z,x)]^2 \leq \mathrm{Var}(z)\mathrm{Var}(x)$, which implies that the asymptotic variance of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is no larger than that of $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$. We have shown in the simple regression case that, under the Gauss-Markov assumptions, the OLS estimator has a smaller asymptotic variance than any estimator of the form (5.17). [The estimator in (5.17) is an example of an *instrumental variables estimator*, which we will study extensively in Chapter 15.] If the homoskedasticity assumption fails, then there are estimators of the form (5.17) that have a smaller asymptotic variance than OLS. We will see this in Chapter 8.

The general case is similar but much more difficult mathematically. In the k regressor case, the class of consistent estimators is obtained by generalizing the OLS first order conditions:

$$\sum_{i=1}^{n} g_j(\boldsymbol{x}_i)(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \dots - \tilde{\beta}_k x_{ik}) = 0, j = 0,1, \dots, k, \tag{5.19}$$

where $g_j(\boldsymbol{x}_i)$ denotes any function of all explanatory variables for observation i. As can be seen by comparing (5.19) with the OLS first order conditions (3.13), we obtain the OLS estimators when $g_0(\boldsymbol{x}_i) = 1$ and $g_j(\boldsymbol{x}_i) = x_{ij}$ for $j = 1,2, \dots, k$. The class of estimators in (5.19) is infinite, because we can use any functions of the x_{ij} that we want.

THEOREM 5.3 (ASYMPTOTIC EFFICIENCY OF OLS)

Under the Gauss-Markov assumptions, let $\tilde{\beta}_j$ denote estimators that solve equations of the form (5.19) and let $\hat{\beta}_j$ denote the OLS estimators. Then for $j = 0,1,2, \dots, k$, the OLS estimators have the smallest asymptotic variances: Avar $\sqrt{n}(\hat{\beta}_j - \beta_j) \leq$ Avar $\sqrt{n}(\tilde{\beta}_j - \beta_j)$.

Proving consistency of the estimators in (5.19), let alone showing they are asymptotically normal, is mathematically difficult. [See Wooldridge (1999, Chapter 5).]

SUMMARY

The claims underlying the material in this chapter are fairly technical, but their practical implications are straightforward. We have shown that the first four Gauss-Markov assumptions imply that OLS is consistent. Furthermore, all of the methods of testing and constructing confidence intervals that we learned in Chapter 4 are approximately valid without assuming that the errors are drawn from a normal distribution (equivalently, the distribution of y given the explanatory variables is not normal). This means that we can apply OLS and use previous methods for an array of applications where the dependent variable is not even approximately normally distributed. We also showed that the *LM* statistic can be used instead of the *F* statistic for testing exclusion restrictions.

Before leaving this chapter, we should note that examples such as Example 5.3 may very well have problems that *do* require special attention. For a variable such as *narr86*, which is zero or one for most men in the population, a linear model may not be able to adequately capture the functional relationship between *narr86* and the explanatory variables. Moreover, even if a linear model does describe the expected value of arrests, heteroskedasticity might be a problem. Problems such as these are not mitigated as the sample size grows, and we will return to them in later chapters.

KEY TERMS

Asymptotic Bias
Asymptotic Confidence Interval
Asymptotic Normality
Asymptotic Properties
Asymptotic Standard Error
Asymptotic *t* Statistics
Asymptotic Variance
Asymptotically Efficient
Auxiliary Regression
Consistency
Inconsistency
Lagrange Multiplier LM Statistic
Large Sample Properties
n-*R*-squared Statistic
Score Statistic

PROBLEMS

5.1 In the simple regression model under MLR.1 through MLR.4, we argued that the slope estimator, $\hat{\beta}_1$, is consistent for β_1. Using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1$, show that plim $\hat{\beta}_0 = \beta_0$. [You need to use the consistency of $\hat{\beta}_1$ and the law of large numbers, along with the fact that $\beta_0 = \mathrm{E}(y) - \beta_1(x_1)$.]

5.2 Suppose that the model

$$pctstck = \beta_0 + \beta_1 funds + \beta_2 risktol + u$$

satisfies the first four Gauss-Markov assumptions, where *pctstck* is the percentage of a worker's pension invested in the stock market, *funds* is the number of mutual funds that the worker can choose from, and *risktol* is some measure of risk tolerance (larger *risktol* means the person has a higher tolerance for risk). If *funds* and *risktol* are positively correlated, what is the inconsistency in $\tilde{\beta}_1$, the slope coefficient in the simple regression of *pctstck* on *funds?*

5.3 The data set SMOKE.RAW contains information on smoking behavior and other variables for a random sample of single adults from the United States. The variable *cigs* is the (average) number of cigarettes smoked per day. Do you think *cigs* has a normal distribution in the U.S. population? Explain.

5.4 In the simple regression model (5.16), under the first four Gauss-Markov assumptions, we showed that estimators of the form (5.17) are consistent for the slope, β_1. Given such an estimator, define an estimator of β_0 by $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x}$. Show that plim $\tilde{\beta}_0 = \beta_0$.

COMPUTER EXERCISES

5.5 Use the data in WAGE1.RAW for this exercise.

(i) Estimate the equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

Save the residuals and plot a histogram.

(ii) Repeat part (i), but with log(*wage*) as the dependent variable.

(iii) Would you say that Assumption MLR.6 is closer to being satisfied for the level-level model or the log-level model?

5.6 Use the data in GPA2.RAW for this exercise.

(i) Using all 4,137 observations, estimate the equation

$$colgpa = \beta_0 + \beta_1 hsperc + \beta_2 sat + u$$

and report the results in standard form.

(ii) Reestimate the equation in part (i), using the first 2,070 observations.

(iii) Find the ratio of the standard errors on *hsperc* from parts (i) and (ii). Compare this with the result from (5.10).

5.7 In equation (4.42) of Chapter 4, compute the *LM* statistic for testing whether *motheduc* and *fatheduc* are jointly significant. In obtaining the residuals for the restricted model, be sure that the restricted model is estimated using only those observations for which all variables in the unrestricted model are available (see Example 4.9).

APPENDIX 5A

We sketch a proof of the asymptotic normality of OLS (Theorem 5.2[i]) in the simple regression case. Write the simple regression model as in equation (5.16). Then, by the usual algebra of simple regression we can write

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/s_x^2)[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i],$$

where we use s_x^2 to denote the sample variance of $\{x_i\colon i = 1,2, \ldots, n\}$. By the law of large numbers (see Appendix C), $s_x^2 \xrightarrow{p} \sigma_x^2 = \text{Var}(x)$. Assumption MLR.4 rules out no perfect collinearity, which means that $\text{Var}(x) > 0$ (x_i varies in the sample, and therefore x is not constant in the population). Next, $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i = n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i + (\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$, where $\mu = \text{E}(x)$ is the population mean of x. Now $\{u_i\}$ is a sequence of i.i.d. random variables with mean zero and variance σ^2, and so $n^{-1/2}\sum_{i=1}^{n}u_i$ converges to the Normal$(0,\sigma^2)$ distribution as $n \rightarrow \infty$; this is just the central limit theorem from Appendix C. By the law of large numbers, $\text{plim}(\mu - \bar{x}) = 0$. A standard result in asymptotic theory is that if $\text{plim}(w_n) = 0$ and z_n has an asymptotic normal distribution, then $\text{plim}(w_n z_n) = 0$. [See Wooldridge (1999, Chapter 3) for more discussion.] This implies that $(\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$ has zero plim. Next, $\{(x_i - \mu)u_i\colon i = 1,2,\ldots\}$ is a sequence of i.i.d. random variables with mean zero—because u and x are uncorrelated under Assumption MLR.3—and variance $\sigma^2\sigma_x^2$ by the homoskedasticity Assumption MLR.5. Therefore, $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ has an asymptotic Normal$(0,\sigma^2\sigma_x^2)$ distribution. We just showed that the difference between $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ and $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ has zero plim. A result in asymptotic theory is that if z_n has an asymptotic normal distribution and $\text{plim}(v_n - z_n) = 0$, then v_n has the same asymptotic normal distribution as z_n. It follows that $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ also has an asymptotic Normal$(0,\sigma^2\sigma_x^2)$ distribution. Putting all of the pieces together gives

$$\begin{aligned}\sqrt{n}(\hat{\beta}_1 - \beta_1) &= (1/\sigma_x^2)[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i] \\ &\quad + [(1/s_x^2) - (1/\sigma_x^2)[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i],\end{aligned}$$

and since $\text{plim}(1/s_x^2) = 1/\sigma_x^2$, the second term has zero plim. Therefore, the asymptotic distribution of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is Normal$(0,\{\sigma^2\sigma_x^2\}/\{\sigma_x^2\}^2)$ = Normal$(0,\sigma^2/\sigma_x^2)$. This completes the proof in the simple regression case, as $a_1^2 = \sigma_x^2$ in this case. See Wooldridge (1999, Chapter 4) for the general case.