3

Multiple Regression Analysis: Estimation

In Chapter 2, we learned how to use simple regression analysis to explain a dependent variable, y, as a function of a single independent variable, x. The primary drawback in using simple regression analysis for empirical work is that it is very difficult to draw ceteris paribus conclusions about how x affects y: the key assumption, SLR.4—that all other factors affecting y are uncorrelated with x—is often unrealistic.

Multiple regression analysis is more amenable to ceteris paribus analysis because it allows us to *explicitly* control for many other factors that simultaneously affect the dependent variable. This is important both for testing economic theories and for evaluating policy effects when we must rely on nonexperimental data. Because multiple regression models can accommodate many explanatory variables that may be correlated, we can hope to infer causality in cases where simple regression analysis would be misleading.

Naturally, if we add more factors to our model that are useful for explaining y, then more of the variation in y can be explained. Thus, multiple regression analysis can be used to build better models for predicting the dependent variable.

An additional advantage of multiple regression analysis is that it can incorporate fairly general functional form relationships. In the simple regression model, only one function of a single explanatory variable can appear in the equation. As we will see, the multiple regression model allows for much more flexibility.

Section 3.1 formally introduces the multiple regression model and further discusses the advantages of multiple regression over simple regression. In Section 3.2, we demonstrate how to estimate the parameters in the multiple regression model using the method of ordinary least squares. In Sections 3.3, 3.4, and 3.5, we describe various statistical properties of the OLS estimators, including unbiasedness and efficiency.

The multiple regression model is still the most widely used vehicle for empirical analysis in economics and other social sciences. Likewise, the method of ordinary least squares is popularly used for estimating the parameters of the multiple regression model.

3.1 Motivation for Multiple Regression

The Model with Two Independent Variables

We begin with some simple examples to show how multiple regression analysis can be used to solve problems that cannot be solved by simple regression.

The first example is a simple variation of the wage equation introduced in Chapter 2 for obtaining the effect of education on hourly wage:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u, \tag{3.1}$$

where *exper* is years of labor market experience. Thus, *wage* is determined by the two explanatory or independent variables, education and experience, and by other unobserved factors, which are contained in u. We are still primarily interested in the effect of *educ* on *wage*, holding fixed all other factors affecting *wage*; that is, we are interested in the parameter β_1.

Compared with a simple regression analysis relating *wage* to *educ*, equation (3.1) effectively takes *exper* out of the error term and puts it explicitly in the equation. Because *exper* appears in the equation, its coefficient, β_2, measures the ceteris paribus effect of *exper* on *wage*, which is also of some interest.

Not surprisingly, just as with simple regression, we will have to make assumptions about how u in (3.1) is related to the independent variables, *educ* and *exper*. However, as we will see in Section 3.2, there is one thing of which we can be confident: because (3.1) contains experience explicitly, we will be able to measure the effect of education on wage, holding experience fixed. In a simple regression analysis—which puts *exper* in the error term—we would have to assume that experience is uncorrelated with education, a tenuous assumption.

As a second example, consider the problem of explaining the effect of per student spending (*expend*) on the average standardized test score (*avgscore*) at the high school level. Suppose that the average test score depends on funding, average family income (*avginc*), and other unobservables:

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u. \tag{3.2}$$

The coefficient of interest for policy purposes is β_1, the ceteris paribus effect of *expend* on *avgscore*. By including *avginc* explicitly in the model, we are able to control for its effect on *avgscore*. This is likely to be important because average family income tends to be correlated with per student spending: spending levels are often determined by both property and local income taxes. In simple regression analysis, *avginc* would be included in the error term, which would likely be correlated with *expend*, causing the OLS estimator of β_1 in the two-variable model to be biased.

In the two previous similar examples, we have shown how observable factors other than the variable of primary interest [*educ* in equation (3.1) and *expend* in equation (3.2)] can be included in a regression model. Generally, we can write a model with two independent variables as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{3.3}$$

where β_0 is the intercept, β_1 measures the change in y with respect to x_1, holding other factors fixed, and β_2 measures the change in y with respect to x_2, holding other factors fixed.

Multiple regression analysis is also useful for generalizing functional relationships between variables. As an example, suppose family consumption (*cons*) is a quadratic function of family income (*inc*):

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u, \tag{3.4}$$

where u contains other factors affecting consumption. In this model, consumption depends on only one observed factor, income; so it might seem that it can be handled in a simple regression framework. But the model falls outside simple regression because it contains two functions of income, *inc* and inc^2 (and therefore three parameters, β_0, β_1, and β_2). Nevertheless, the consumption function is easily written as a regression model with two independent variables by letting $x_1 = inc$ and $x_2 = inc^2$.

Mechanically, there will be *no* difference in using the method of ordinary least squares (introduced in Section 3.2) to estimate equations as different as (3.1) and (3.4). Each equation can be written as (3.3), which is all that matters for computation. There is, however, an important difference in how one *interprets* the parameters. In equation (3.1), β_1 is the ceteris paribus effect of *educ* on *wage*. The parameter β_1 has no such interpretation in (3.4). In other words, it makes no sense to measure the effect of *inc* on *cons* while holding inc^2 fixed, because if *inc* changes, then so must inc^2! Instead, the change in consumption with respect to the change in income—the marginal propensity to consume—is approximated by

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc.$$

See Appendix A for the calculus needed to derive this equation. In other words, the marginal effect of income on consumption depends on β_2 as well as on β_1 and the level of income. This example shows that, in any particular application, the definitions of the independent variables are crucial. But for the theoretical development of multiple regression, we can be vague about such details. We will study examples like this more completely in Chapter 6.

In the model with two independent variables, the key assumption about how u is related to x_1 and x_2 is

$$\mathrm{E}(u|x_1,x_2) = 0. \tag{3.5}$$

The interpretation of condition (3.5) is similar to the interpretation of Assumption SLR.4 for simple regression analysis. It means that, for any values of x_1 and x_2 in the population, the average unobservable is equal to zero. As with simple regression, the important part of the assumption is that the expected value of u is the same for all combinations of x_1 and x_2; that this common value is zero is no assumption at all as long as the intercept β_0 is included in the model (see Section 2.1).

How can we interpret the zero conditional mean assumption in the previous examples? In equation (3.1), the assumption is $\mathrm{E}(u|educ,exper) = 0$. This implies that other factors affecting *wage* are not related on average to *educ* and *exper*. Therefore, if we think innate ability is part of u, then we will need average ability levels to be the same across all

combinations of education and experience in the working population. This may or may not be true, but, as we will see in Section 3.3, this is the question we need to ask in order to determine whether the method of ordinary least squares produces unbiased estimators.

The example measuring student performance [equation (3.2)] is similar to the wage equation. The zero conditional mean assumption is $E(u|expend,avginc) = 0$, which means that other factors affecting test scores—school or student characteristics—are, on average, unrelated to per student funding and average family income.

QUESTION 3.1

A simple model to explain city murder rates (*murdrate*) in terms of the probability of conviction (*prbconv*) and average sentence length (*avgsen*) is

$$murdrate = \beta_0 + \beta_1 prbconv + \beta_2 avgsen + u.$$

What are some factors contained in u? Do you think the key assumption (3.5) is likely to hold?

When applied to the quadratic consumption function in (3.4), the zero conditional mean assumption has a slightly different interpretation. Written literally, equation (3.5) becomes $E(u|inc,inc^2) = 0$. Since inc^2 is known when inc is known, including inc^2 in the expectation is redundant: $E(u|inc,inc^2) = 0$ is the same as $E(u|inc) = 0$. Nothing is wrong with putting inc^2 along with inc in the expectation when stating the assumption, but $E(u|inc) = 0$ is more concise.

The Model with k Independent Variables

Once we are in the context of multiple regression, there is no need to stop with two independent variables. Multiple regression analysis allows many observed factors to affect y. In the wage example, we might also include amount of job training, years of tenure with the current employer, measures of ability, and even demographic variables like number of siblings or mother's education. In the school funding example, additional variables might include measures of teacher quality and school size.

The general **multiple linear regression model** (also called the *multiple regression model*) can be written in the population as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u, \qquad (3.6)$$

where β_0 is the **intercept**, β_1 is the parameter associated with x_1, β_2 is the parameter associated with x_2, and so on. Since there are k independent variables and an intercept, equation (3.6) contains $k + 1$ (unknown) population parameters. For shorthand purposes, we will sometimes refer to the parameters other than the intercept as **slope parameters**, even though this is not always literally what they are. [See equation (3.4), where neither β_1 nor β_2 is itself a slope, but together they determine the slope of the relationship between consumption and income.]

The terminology for multiple regression is similar to that for simple regression and is given in Table 3.1. Just as in simple regression, the variable u is the **error term** or **disturbance**. It contains factors other than $x_1, x_2, \ldots, x_k$ that affect y. No matter how many explanatory variables we include in our model, there will always be factors we cannot include, and these are collectively contained in u.

When applying the general multiple regression model, we must know how to interpret the parameters. We will get plenty of practice now and in subsequent chapters, but it is useful at

TABLE 3.1
Terminology for Multiple Regression

y	$x_1, x_2, \ldots, x_k$
Dependent Variable	Independent Variables
Explained Variable	Explanatory Variables
Response Variable	Control Variables
Predicted Variable	Predictor Variables
Regressand	Regressors

this point to be reminded of some things we already know. Suppose that CEO salary (*salary*) is related to firm sales (*sales*) and CEO tenure (*ceoten*) with the firm by

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u. \tag{3.7}$$

This fits into the multiple regression model (with $k = 3$) by defining $y = \log(salary)$, $x_1 = \log(sales)$, $x_2 = ceoten$, and $x_3 = ceoten^2$. As we know from Chapter 2, the parameter β_1 is the (ceteris paribus) *elasticity* of *salary* with respect to *sales*. If $\beta_3 = 0$, then $100\beta_2$ is approximately the ceteris paribus percentage increase in *salary* when *ceoten* increases by one year. When $\beta_3 \neq 0$, the effect of *ceoten* on *salary* is more complicated. We will postpone a detailed treatment of general models with quadratics until Chapter 6.

Equation (3.7) provides an important reminder about multiple regression analysis. The term "linear" in multiple linear regression model means that equation (3.6) is linear in the *parameters*, β_j. Equation (3.7) is an example of a multiple regression model that, while linear in the β_j, is a nonlinear relationship between *salary* and the variables *sales* and *ceoten*. Many applications of multiple linear regression involve nonlinear relationships among the underlying variables.

The key assumption for the general multiple regression model is easy to state in terms of a conditional expectation:

$$E(u|x_1, x_2, \ldots, x_k) = 0. \tag{3.8}$$

At a minimum, equation (3.8) requires that all factors in the unobserved error term be uncorrelated with the explanatory variables. It also means that we have correctly accounted for the functional relationships between the explained and explanatory variables. Any problem that causes u to be correlated with any of the independent variables causes (3.8) to fail. In Section 3.3, we will show that assumption (3.8) implies that OLS is unbiased and will derive the bias that arises when a key variable has been omitted from

the equation. In Chapters 15 and 16, we will study other reasons that might cause (3.8) to fail and show what can be done in cases where it does fail.

3.2 Mechanics and Interpretation of Ordinary Least Squares

We now summarize some computational and algebraic features of the method of ordinary least squares as it applies to a particular set of data. We also discuss how to interpret the estimated equation.

Obtaining the OLS Estimates

We first consider estimating the model with two independent variables. The estimated OLS equation is written in a form similar to the simple regression case:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \tag{3.9}$$

where $\hat{\beta}_0$ is the estimate of β_0, $\hat{\beta}_1$ is the estimate of β_1, and $\hat{\beta}_2$ is the estimate of β_2. But how do we obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? The method of **ordinary least squares** chooses the estimates to minimize the sum of squared residuals. That is, given n observations on y, x_1, and x_2, $\{(x_{i1}, x_{i2}, y_i): i = 1, 2, \ldots, n\}$, the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are chosen simultaneously to make

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \tag{3.10}$$

as small as possible.

In order to understand what OLS is doing, it is important to master the meaning of the indexing of the independent variables in (3.10). The independent variables have two subscripts here, i followed by either 1 or 2. The i subscript refers to the observation number. Thus, the sum in (3.10) is over all $i = 1$ to n observations. The second index is simply a method of distinguishing between different independent variables. In the example relating *wage* to *educ* and *exper*, $x_{i1} = educ_i$ is education for person i in the sample, and $x_{i2} = exper_i$ is experience for person i. The sum of squared residuals in equation (3.10) is $\sum_{i=1}^{n} (wage_i - \hat{\beta}_0 - \hat{\beta}_1 educ_i - \hat{\beta}_2 exper_i)^2$. In what follows, the i subscript is reserved for indexing the observation number. If we write x_{ij}, then this means the i^{th} observation on the j^{th} independent variable. (Some authors prefer to switch the order of the observation number and the variable number, so that x_{1i} is observation i on variable one. But this is just a matter of notational taste.)

In the general case with k independent variables, we seek estimates $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ in the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{3.11}$$

The OLS estimates, $k + 1$ of them, are chosen to minimize the sum of squared residuals:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2. \tag{3.12}$$

This minimization problem can be solved using multivariable calculus (see Appendix 3A). This leads to $k + 1$ linear equations in $k + 1$ unknowns $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$:

$$\begin{aligned}
&\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\quad\vdots \\
&\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0.
\end{aligned} \tag{3.13}$$

These are often called the OLS **first order conditions**. As with the simple regression model in Section 2.2, the OLS first order conditions can be obtained by the method of moments: under assumption (3.8), $\mathrm{E}(u) = 0$ and $\mathrm{E}(x_j u) = 0$, where $j = 1, 2, \ldots, k$. The equations in (3.13) are the sample counterparts of these population moments, although we have omitted the division by the sample size n.

For even moderately sized n and k, solving the equations in (3.13) by hand calculations is tedious. Nevertheless, modern computers running standard statistics and econometrics software can solve these equations with large n and k very quickly.

There is only one slight caveat: we must assume that the equations in (3.13) can be solved *uniquely* for the $\hat{\beta}_j$. For now, we just assume this, as it is usually the case in well-specified models. In Section 3.3, we state the assumption needed for unique OLS estimates to exist (see Assumption MLR.3).

As in simple regression analysis, equation (3.11) is called the **OLS regression line** or the **sample regression function (SRF)**. We will call $\hat{\beta}_0$ the **OLS intercept estimate** and $\hat{\beta}_1, \ldots, \hat{\beta}_k$ the **OLS slope estimates** (corresponding to the independent variables $x_1, x_2, \ldots, x_k$).

In order to indicate that an OLS regression has been run, we will either write out equation (3.11) with y and $x_1, \ldots, x_k$ replaced by their variable names (such as *wage*, *educ*, and *exper*), or we will say that "we ran an OLS regression of y on $x_1, x_2, \ldots, x_k$" or that "we regressed y on $x_1, x_2, \ldots, x_k$." These are shorthand for saying that the method of ordinary least squares was used to obtain the OLS equation (3.11). Unless explicitly stated otherwise, we always estimate an intercept along with the slopes.

Interpreting the OLS Regression Equation

More important than the details underlying the computation of the $\hat{\beta}_j$ is the *interpretation* of the estimated equation. We begin with the case of two independent variables:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \tag{3.14}$$

The intercept $\hat{\beta}_0$ in equation (3.14) is the predicted value of y when $x_1 = 0$ and $x_2 = 0$. Sometimes, setting x_1 and x_2 both equal to zero is an interesting scenario; in other cases, it will not make sense. Nevertheless, the intercept is always needed to obtain a prediction of y from the OLS regression line, as (3.14) makes clear.

The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have **partial effect**, or **ceteris paribus**, interpretations. From equation (3.14), we have

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2,$$

so we can obtain the predicted change in y given the changes in x_1 and x_2. (Note how the intercept has nothing to do with the changes in y.) In particular, when x_2 is held fixed, so that $\Delta x_2 = 0$, then

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1,$$

holding x_2 fixed. The key point is that, by including x_2 in our model, we obtain a coefficient on x_1 with a ceteris paribus interpretation. This is why multiple regression analysis is so useful. Similarly,

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2,$$

holding x_1 fixed.

EXAMPLE 3.1

(Determinants of College GPA)

The variables in GPA1.RAW include college grade point average (*colGPA*), high school GPA (*hsGPA*), and achievement test score (*ACT*) for a sample of 141 students from a large university; both college and high school GPAs are on a four-point scale. We obtain the following OLS regression line to predict college GPA from high school GPA and achievement test score:

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT. \tag{3.15}$$

How do we interpret this equation? First, the intercept 1.29 is the predicted college GPA if *hsGPA* and *ACT* are both set as zero. Since no one who attends college has either a zero high school GPA or a zero on the achievement test, the intercept in this equation is not, by itself, meaningful.

More interesting estimates are the slope coefficients on *hsGPA* and *ACT*. As expected, there is a positive partial relationship between *colGPA* and *hsGPA*: holding *ACT* fixed, another point on *hsGPA* is associated with .453 of a point on the college GPA, or almost half a point. In other words, if we choose two students, A and B, and these students have the same ACT score, but the high school GPA of Student A is one point higher than the high school GPA of

Student B, then we predict Student A to have a college GPA .453 higher than that of Student B. (This says nothing about any two actual people, but it is our best prediction.)

The sign on *ACT* implies that, while holding *hsGPA* fixed, a change in the ACT score of 10 points—a very large change, since the average score in the sample is about 24 with a standard deviation less than three—affects *colGPA* by less than one-tenth of a point. This is a small effect, and it suggests that, once high school GPA is accounted for, the ACT score is not a strong predictor of college GPA. (Naturally, there are many other factors that contribute to GPA, but here we focus on statistics available for high school students.) Later, after we discuss statistical inference, we will show that not only is the coefficient on ACT practically small, it is also statistically insignificant.

If we focus on a simple regression analysis relating *colGPA* to *ACT* only, we obtain

$$\widehat{colGPA} = 2.40 + .0271\, ACT;$$

thus, the coefficient on *ACT* is almost three times as large as the estimate in (3.15). But this equation does *not* allow us to compare two people with the same high school GPA; it corresponds to a different experiment. We say more about the differences between multiple and simple regression later.

The case with more than two independent variables is similar. The OLS regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{3.16}$$

Written in terms of changes,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k. \tag{3.17}$$

The coefficient on x_1 measures the change in $\hat{y}$ due to a one-unit increase in x_1, holding all other independent variables fixed. That is,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1, \tag{3.18}$$

holding $x_2, x_3, \ldots, x_k$ fixed. Thus, we have *controlled for* the variables $x_2, x_3, \ldots, x_k$ when estimating the effect of x_1 on y. The other coefficients have a similar interpretation.

The following is an example with three independent variables.

EXAMPLE 3.2

(Hourly Wage Equation)

Using the 526 observations on workers in WAGE1.RAW, we include *educ* (years of education), *exper* (years of labor market experience), and *tenure* (years with the current employer) in an equation explaining log(*wage*). The estimated equation is

$$\widehat{\log(wage)} = .284 + .092\, educ + .0041\, exper + .022\, tenure. \tag{3.19}$$

As in the simple regression case, the coefficients have a percentage interpretation. The only difference here is that they also have a ceteris paribus interpretation. The coefficient .092 means that, holding *exper* and *tenure* fixed, another year of education is predicted to increase log(*wage*) by .092, which translates into an approximate 9.2 percent [100(.092)] increase in *wage*. Alternatively, if we take two people with the same levels of experience and job tenure, the coefficient on *educ* is the proportionate difference in predicted wage when their education levels differ by one year. This measure of the return to education at least keeps two important productivity factors fixed; whether it is a good estimate of the ceteris paribus return to another year of education requires us to study the statistical properties of OLS (see Section 3.3).

On the Meaning of "Holding Other Factors Fixed" in Multiple Regression

The partial effect interpretation of slope coefficients in multiple regression analysis can cause some confusion, so we attempt to prevent that problem now.

In Example 3.1, we observed that the coefficient on *ACT* measures the predicted difference in *colGPA*, holding *hsGPA* fixed. The power of multiple regression analysis is that it provides this ceteris paribus interpretation even though the data have *not* been collected in a ceteris paribus fashion. In giving the coefficient on *ACT* a partial effect interpretation, it may seem that we actually went out and sampled people with the same high school GPA but possibly with different ACT scores. This is not the case. The data are a random sample from a large university: there were no restrictions placed on the sample values of *hsGPA* or *ACT* in obtaining the data. Rarely do we have the luxury of holding certain variables fixed in obtaining our sample. *If* we could collect a sample of individuals with the same high school GPA, then we could perform a simple regression analysis relating *colGPA* to *ACT*. Multiple regression effectively allows us to mimic this situation without restricting the values of any independent variables.

The power of multiple regression analysis is that it allows us to do in nonexperimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed.

Changing More than One Independent Variable Simultaneously

Sometimes, we want to change more than one independent variable at the same time to find the resulting effect on the dependent variable. This is easily done using equation (3.17). For example, in equation (3.19), we can obtain the estimated effect on *wage* when an individual stays at the same firm for another year: *exper* (general workforce experience) and *tenure* both increase by one year. The total effect (holding *educ* fixed) is

$$\widehat{\Delta \log(wage)} = .0041\,\Delta exper + .022\,\Delta tenure = .0041 + .022 = .0261,$$

or about 2.6 percent. Since *exper* and *tenure* each increase by one year, we just add the coefficients on *exper* and *tenure* and multiply by 100 to turn the effect into a percent.

OLS Fitted Values and Residuals

After obtaining the OLS regression line (3.11), we can obtain a *fitted* or *predicted value* for each observation. For observation i, the fitted value is simply

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}, \tag{3.20}$$

which is just the predicted value obtained by plugging the values of the independent variables for observation i into equation (3.11). We should not forget about the intercept in obtaining the fitted values; otherwise, the answer can be very misleading. As an example, if in (3.15), $hsGPA_i = 3.5$ and $ACT_i = 24$, $\widehat{colGPA}_i = 1.29 + .453(3.5) + .0094(24) = 3.101$ (rounded to three places after the decimal).

Normally, the actual value y_i for any observation i will not equal the predicted value, $\hat{y}_i$: OLS minimizes the *average* squared prediction error, which says nothing about the prediction error for any particular observation. The **residual** for observation i is defined just as in the simple regression case,

$$\hat{u}_i = y_i - \hat{y}_i. \tag{3.21}$$

There is a residual for each observation. If $\hat{u}_i > 0$, then $\hat{y}_i$ is below y_i, which means that, for this observation, y_i is underpredicted. If $\hat{u}_i < 0$, then $y_i < \hat{y}_i$, and y_i is overpredicted.

The OLS fitted values and residuals have some important properties that are immediate extensions from the single variable case:

1. The sample average of the residuals is zero and so $\bar{y} = \bar{\hat{y}}$.
2. The sample covariance between each independent variable and the OLS residuals is zero. Consequently, the sample covariance between the OLS fitted values and the OLS residuals is zero.
3. The point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ is always on the OLS regression line: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \ldots + \hat{\beta}_k \bar{x}_k$.

QUESTION 3.2

In Example 3.1, the OLS fitted line explaining college GPA in terms of high school GPA and ACT score is

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT.$$

If the average high school GPA is about 3.4 and the average ACT score is about 24.2, what is the average college GPA in the sample?

The first two properties are immediate consequences of the set of equations used to obtain the OLS estimates. The first equation in (3.13) says that the sum of the residuals is zero. The remaining equations are of the form $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, which implies that each independent variable has zero sample covariance with $\hat{u}_i$. Property (3) follows immediately from property (1).

A "Partialling Out" Interpretation of Multiple Regression

When applying OLS, we do not need to know explicit formulas for the $\hat{\beta}_j$ that solve the system of equations in (3.13). Nevertheless, for certain derivations, we do need explicit formulas for the $\hat{\beta}_j$. These formulas also shed further light on the workings of OLS.

Consider again the case with $k = 2$ independent variables, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. For concreteness, we focus on $\hat{\beta}_1$. One way to express $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} \hat{r}_{i1} y_i\right) / \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right), \tag{3.22}$$

where the $\hat{r}_{i1}$ are the OLS residuals from a simple regression of x_1 on x_2, using the sample at hand. We regress our first independent variable, x_1, on our second independent variable, x_2, and then obtain the residuals (y plays no role here). Equation (3.22) shows that we can then do a simple regression of y on $\hat{r}_1$ to obtain $\hat{\beta}_1$. (Note that the residuals $\hat{r}_{i1}$ have a zero sample average, and so $\hat{\beta}_1$ is the usual slope estimate from simple regression.)

The representation in equation (3.22) gives another demonstration of $\hat{\beta}_1$'s partial effect interpretation. The residuals $\hat{r}_{i1}$ are the part of x_{i1} that is uncorrelated with x_{i2}. Another way of saying this is that $\hat{r}_{i1}$ is x_{i1} after the effects of x_{i2} have been *partialled out*, or *netted out*. Thus, $\hat{\beta}_1$ measures the sample relationship between y and x_1 after x_2 has been partialled out.

In simple regression analysis, there is no partialling out of other variables because no other variables are included in the regression. Computer Exercise C3.5 steps you through the partialling out process using the wage data from Example 3.2. For practical purposes, the important thing is that $\hat{\beta}_1$ in the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ measures the change in y given a one-unit increase in x_1, holding x_2 fixed.

In the general model with k explanatory variables, $\hat{\beta}_1$ can still be written as in equation (3.22), but the residuals $\hat{r}_{i1}$ come from the regression of x_1 on $x_2, \ldots, x_k$. Thus, $\hat{\beta}_1$ measures the effect of x_1 on y after $x_2, \ldots, x_k$ have been partialled or netted out.

Comparison of Simple and Multiple Regression Estimates

Two special cases exist in which the simple regression of y on x_1 will produce the *same* OLS estimate on x_1 as the regression of y on x_1 and x_2. To be more precise, write the simple regression of y on x_1 as $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, and write the multiple regression as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. We know that the simple regression coefficient $\tilde{\beta}_1$ does not usually equal the multiple regression coefficient $\hat{\beta}_1$. It turns out there is a simple relationship between $\tilde{\beta}_1$ and $\hat{\beta}_1$, which allows for interesting comparisons between simple and multiple regression:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1, \tag{3.23}$$

where $\tilde{\delta}_1$ is the slope coefficient from the simple regression of x_{i2} on x_{i1}, $i = 1, \ldots, n$. This equation shows how $\tilde{\beta}_1$ differs from the partial effect of x_1 on $\hat{y}$. The confounding term is the partial effect of x_2 on $\hat{y}$ times the slope in the sample regression of x_2 on x_1. (See Section 3A.4 in the chapter appendix for a more general verification.)

The relationship between $\tilde{\beta}_1$ and $\hat{\beta}_1$ also shows there are two distinct cases where they are equal:

1. The partial effect of x_2 on $\hat{y}$ is zero in the sample. That is, $\hat{\beta}_2 = 0$.
2. x_1 and x_2 are uncorrelated in the sample. That is, $\tilde{\delta}_1 = 0$.

Even though simple and multiple regression estimates are almost never identical, we can use the above formula to characterize why they might be either very different or quite similar. For example, if $\hat{\beta}_2$ is small, we might expect the multiple and simple regression estimates of β_1 to be similar. In Example 3.1, the sample correlation between *hsGPA* and *ACT* is about 0.346, which is a nontrivial correlation. But the coefficient on *ACT* is fairly little. It is not surprising to find that the simple regression of *colGPA* on *hsGPA* produces a slope estimate of .482, which is not much different from the estimate .453 in (3.15).

EXAMPLE 3.3

[Participation in 401(k) Pension Plans]

We use the data in 401K.RAW to estimate the effect of a plan's match rate (*mrate*) on the participation rate (*prate*) in its 401(k) pension plan. The match rate is the amount the firm contributes to a worker's fund for each dollar the worker contributes (up to some limit); thus, *mrate* = .75 means that the firm contributes 75 cents for each dollar contributed by the worker. The participation rate is the percentage of eligible workers having a 401(k) account. The variable *age* is the age of the 401(k) plan. There are 1,534 plans in the data set, the average *prate* is 87.36, the average *mrate* is .732, and the average *age* is 13.2.

Regressing *prate* on *mrate*, *age* gives

$$\widehat{prate} = 80.12 + 5.52\, mrate + .243\, age.$$

Thus, both *mrate* and *age* have the expected effects. What happens if we do not control for *age*? The estimated effect of *age* is not trivial, and so we might expect a large change in the estimated effect of *mrate* if *age* is dropped from the regression. However, the simple regression of *prate* on *mrate* yields $\widehat{prate} = 83.08 + 5.86\, mrate$. The simple regression estimate of the effect of *mrate* on *prate* is clearly different from the multiple regression estimate, but the difference is not very big. (The simple regression estimate is only about 6.2 percent larger than the multiple regression estimate.) This can be explained by the fact that the sample correlation between *mrate* and *age* is only .12.

In the case with k independent variables, the simple regression of y on x_1 and the multiple regression of y on $x_1, x_2, \ldots, x_k$ produce an identical estimate of x_1 only if (1) the OLS coefficients on x_2 through x_k are all zero *or* (2) x_1 is uncorrelated with *each of* $x_2, \ldots, x_k$. Neither of these is very likely in practice. But if the coefficients on x_2 through x_k are small, or the sample correlations between x_1 and the other independent variables are insubstantial, then the simple and multiple regression estimates of the effect of x_1 on y can be similar.

Goodness-of-Fit

As with simple regression, we can define the **total sum of squares (SST)**, the **explained sum of squares (SSE)**, and the **residual sum of squares** or **sum of squared residuals (SSR)** as

$$\text{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.24}$$

$$\text{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.25}$$

$$\text{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \tag{3.26}$$

Using the same argument as in the simple regression case, we can show that

$$\text{SST} = \text{SSE} + \text{SSR}. \tag{3.27}$$

In other words, the total variation in $\{y_i\}$ is the sum of the total variations in $\{\hat{y}_i\}$ and in $\{\hat{u}_i\}$.

Assuming that the total variation in y is nonzero, as is the case unless y_i is constant in the sample, we can divide (3.27) by SST to get

$$\text{SSR/SST} + \text{SSE/SST} = 1.$$

Just as in the simple regression case, the R-squared is defined to be

$$R^2 \equiv \text{SSE/SST} = 1 - \text{SSR/SST}, \tag{3.28}$$

and it is interpreted as the proportion of the sample variation in y_i that is explained by the OLS regression line. By definition, R^2 is a number between zero and one.

R^2 can also be shown to equal the squared correlation coefficient between the actual y_i and the fitted values $\hat{y}_i$. That is,

$$R^2 = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2\right)}. \tag{3.29}$$

[We have put the average of the $\hat{y}_i$ in (3.29) to be true to the formula for a correlation coefficient; we know that this average equals $\bar{y}$ because the sample average of the residuals is zero and $y_i = \hat{y}_i + \hat{u}_i$.]

An important fact about R^2 is that it never decreases, and it usually increases when another independent variable is added to a regression. This algebraic fact follows because, by definition, the sum of squared residuals never increases when additional regressors are added to the model. For example, the last digit of one's social security number has nothing to do with one's hourly wage, but adding this digit to a wage equation will increase the R^2 (by a little, at least).

The fact that R^2 never decreases when *any* variable is added to a regression makes it a poor tool for deciding whether one variable or several variables should be added to a model. The factor that should determine whether an explanatory variable belongs in a model is whether the explanatory variable has a nonzero partial effect on y in the *population*. We

will show how to test this hypothesis in Chapter 4 when we cover statistical inference. We will also see that, when used properly, R^2 allows us to *test* a group of variables to see if it is important for explaining y. For now, we use it as a goodness-of-fit measure for a given model.

EXAMPLE 3.4

(Determinants of College GPA)

From the grade point average regression that we did earlier, the equation with R^2 is

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$
$$n = 141, R^2 = .176.$$

This means that *hsGPA* and *ACT* together explain about 17.6 percent of the variation in college GPA for this sample of students. This may not seem like a high percentage, but we must remember that there are many other factors—including family background, personality, quality of high school education, affinity for college—that contribute to a student's college performance. If *hsGPA* and *ACT* explained almost all of the variation in *colGPA*, then performance in college would be preordained by high school performance!

EXAMPLE 3.5

(Explaining Arrest Records)

CRIME1.RAW contains data on arrests during the year 1986 and other information on 2,725 men born in either 1960 or 1961 in California. Each man in the sample was arrested at least once prior to 1986. The variable *narr86* is the number of times the man was arrested during 1986: it is zero for most men in the sample (72.29 percent), and it varies from 0 to 12. (The percentage of men arrested once during 1986 was 20.51.) The variable *pcnv* is the proportion (not percentage) of arrests prior to 1986 that led to conviction, *avgsen* is average sentence length served for prior convictions (zero for most people), *ptime86* is months spent in prison in 1986, and *qemp86* is the number of quarters during which the man was employed in 1986 (from zero to four).

A linear model explaining arrests is

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

where *pcnv* is a proxy for the likelihood for being convicted of a crime and *avgsen* is a measure of expected severity of punishment, if convicted. The variable *ptime86* captures the incarcerative effects of crime: if an individual is in prison, he cannot be arrested for a crime outside of prison. Labor market opportunities are crudely captured by *qemp86*.

First, we estimate the model without the variable *avgsen*. We obtain

$$\widehat{narr86} = .712 - .150\, pcnv - .034\, ptime86 - .104\, qemp86$$
$$n = 2{,}725, R^2 = .0413.$$

This equation says that, as a group, the three variables *pcnv*, *ptime86*, and *qemp86* explain about 4.1 percent of the variation in *narr86*.

Each of the OLS slope coefficients has the anticipated sign. An increase in the proportion of convictions lowers the predicted number of arrests. If we increase *pcnv* by .50 (a large increase in the probability of conviction), then, holding the other factors fixed, $\Delta\widehat{narr86} = -.150(.50) = -.075$. This may seem unusual because an arrest cannot change by a fraction. But we can use this value to obtain the predicted change in expected arrests for a large group of men. For example, among 100 men, the predicted fall in arrests when *pcnv* increases by .50 is −7.5.

Similarly, a longer prison term leads to a lower predicted number of arrests. In fact, if *ptime86* increases from 0 to 12, predicted arrests for a particular man fall by .034(12) = .408. Another quarter in which legal employment is reported lowers predicted arrests by .104, which would be 10.4 arrests among 100 men.

If *avgsen* is added to the model, we know that R^2 will increase. The estimated equation is

$$\widehat{narr86} = .707 - .151\, pcnv + .0074\, avgsen - .037\, ptime86 - .103\, qemp86$$
$$n = 2{,}725,\ R^2 = .0422.$$

Thus, adding the average sentence variable increases R^2 from .0413 to .0422, a practically small effect. The sign of the coefficient on *avgsen* is also unexpected: it says that a longer average sentence length increases criminal activity.

Example 3.5 deserves a final word of caution. The fact that the four explanatory variables included in the second regression explain only about 4.2 percent of the variation in *narr86* does not necessarily mean that the equation is useless. Even though these variables collectively do not explain much of the variation in arrests, it is still possible that the OLS estimates are reliable estimates of the ceteris paribus effects of each independent variable on *narr86*. As we will see, whether this is the case does not directly depend on the size of R^2. Generally, a low R^2 indicates that it is hard to predict individual outcomes on y with much accuracy, something we study in more detail in Chapter 6. In the arrest example, the small R^2 reflects what we already suspect in the social sciences: it is generally very difficult to predict individual behavior.

Regression through the Origin

Sometimes, an economic theory or common sense suggests that β_0 should be zero, and so we should briefly mention OLS estimation when the intercept is zero. Specifically, we now seek an equation of the form

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \ldots + \tilde{\beta}_k x_k, \tag{3.30}$$

where the symbol "~" over the estimates is used to distinguish them from the OLS estimates obtained along with the intercept [as in (3.11)]. In (3.30), when $x_1 = 0, x_2 = 0, \ldots, x_k = 0$, the predicted value is zero. In this case, $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ are said to be the OLS estimates from the regression of y on $x_1, x_2, \ldots, x_k$ *through the origin.*

The OLS estimates in (3.30), as always, minimize the sum of squared residuals, but with the intercept set at zero. You should be warned that the properties of OLS that we derived earlier no longer hold for regression through the origin. In particular, the

OLS residuals no longer have a zero sample average. Further, if R^2 is defined as 1 − SSR/SST, where SST is given in (3.24) and SSR is now $\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_k x_{ik})^2$, then R^2 can actually be negative. This means that the sample average, $\bar{y}$, "explains" more of the variation in the y_i than the explanatory variables. Either we should include an intercept in the regression or conclude that the explanatory variables poorly explain y. In order to always have a nonnegative R-squared, some economists prefer to calculate R^2 as the squared correlation coefficient between the actual and fitted values of y, as in (3.29). (In this case, the average fitted value must be computed directly since it no longer equals $\bar{y}$.) However, there is no set rule on computing R-squared for regression through the origin.

One serious drawback with regression through the origin is that, if the intercept β_0 in the population model is different from zero, then the OLS estimators of the slope parameters will be biased. The bias can be severe in some cases. The cost of estimating an intercept when β_0 is truly zero is that the variances of the OLS slope estimators are larger.

3.3 The Expected Value of the OLS Estimators

We now turn to the statistical properties of OLS for estimating the parameters in an underlying population model. In this section, we derive the expected value of the OLS estimators. In particular, we state and discuss four assumptions, which are direct extensions of the simple regression model assumptions, under which the OLS estimators are unbiased for the population parameters. We also explicitly obtain the bias in OLS when an important variable has been omitted from the regression.

You should remember that statistical properties have nothing to do with a particular sample, but rather with the property of estimators when random sampling is done repeatedly. Thus, Sections 3.3, 3.4, and 3.5 are somewhat abstract. Although we give examples of deriving bias for particular models, it is not meaningful to talk about the statistical properties of a set of estimates obtained from a single sample.

The first assumption we make simply defines the multiple linear regression (MLR) model.

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u, \tag{3.31}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobservable random error or disturbance term.

Equation (3.31) formally states the **population model**, sometimes called the **true model**, to allow for the possibility that we might estimate a model that differs from (3.31). The key feature is that the model is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_k$. As we know, (3.31) is quite flexible because y and the independent variables can be arbitrary functions of the

underlying variables of interest, such as natural logarithms and squares [see, for example, equation (3.7)].

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $\{(x_{i1},x_{i2},\ldots,x_{ik},y_i): i = 1,2,\ldots,n\}$, following the population model in Assumption MLR.1.

Sometimes, we need to write the equation for a particular observation i: for a randomly drawn observation from the population, we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i. \tag{3.32}$$

Remember that i refers to the observation, and the second subscript on x is the variable number. For example, we can write a CEO salary equation for a particular CEO i as

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ceoten_i + \beta_3 ceoten_i^2 + u_i. \tag{3.33}$$

The term u_i contains the unobserved factors for CEO i that affect his or her salary. For applications, it is usually easiest to write the model in population form, as in (3.31). It contains less clutter and emphasizes the fact that we are interested in estimating a population relationship.

In light of model (3.31), the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ from the regression of y on $x_1, \ldots, x_k$ are now considered to be estimators of $\beta_0, \beta_1, \ldots, \beta_k$. We saw, in Section 3.2, that OLS chooses the estimates for a particular sample so that the residuals average out to zero and the sample correlation between each independent variable and the residuals is zero. Still, we need an assumption that ensures the OLS estimators are well defined.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.3 is more complicated than its counterpart for simple regression because we must now look at relationships between all independent variables. If an independent variable in (3.31) is an exact linear combination of the other independent variables, then we say the model suffers from **perfect collinearity**, and it cannot be estimated by OLS.

It is important to note that Assumption MLR.3 *does* allow the independent variables to be correlated; they just cannot be *perfectly* correlated. If we did not allow for any correlation among the independent variables, then multiple regression would be of very limited use for econometric analysis. For example, in the model relating test scores to educational expenditures and average family income,

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u,$$

we fully expect *expend* and *avginc* to be correlated: school districts with high average family incomes tend to spend more per student on education. In fact, the primary motivation for including *avginc* in the equation is that we suspect it is correlated with *expend*, and so we would like to hold it fixed in the analysis. Assumption MLR.3 only rules out *perfect* correlation between *expend* and *avginc* in our sample. We would be very unlucky to obtain a sample where per student expenditures are perfectly correlated with average family income. But some correlation, perhaps a substantial amount, is expected and certainly allowed.

The simplest way that two independent variables can be perfectly correlated is when one variable is a constant multiple of another. This can happen when a researcher inadvertently puts the same variable measured in different units into a regression equation. For example, in estimating a relationship between consumption and income, it makes no sense to include as independent variables income measured in dollars as well as income measured in thousands of dollars. One of these is redundant. What sense would it make to hold income measured in dollars fixed while changing income measured in thousands of dollars?

We already know that different nonlinear functions of the same variable *can* appear among the regressors. For example, the model $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ does not violate Assumption MLR.3: even though $x_2 = inc^2$ is an exact function of $x_1 = inc$, inc^2 is not an exact *linear* function of *inc*. Including inc^2 in the model is a useful way to generalize functional form, unlike including income measured in dollars and in thousands of dollars.

Common sense tells us not to include the same explanatory variable measured in different units in the same regression equation. There are also more subtle ways that one independent variable can be a multiple of another. Suppose we would like to estimate an extension of a constant elasticity consumption function. It might seem natural to specify a model such as

$$\log(cons) = \beta_0 + \beta_1 \log(inc) + \beta_2 \log(inc^2) + u, \tag{3.34}$$

where $x_1 = \log(inc)$ and $x_2 = \log(inc^2)$. Using the basic properties of the natural log (see Appendix A), $\log(inc^2) = 2 \cdot \log(inc)$. That is, $x_2 = 2x_1$, and naturally this holds for all observations in the sample. This violates Assumption MLR.3. What we should do instead is include $[\log(inc)]^2$, not $\log(inc^2)$, along with $\log(inc)$. This is a sensible extension of the constant elasticity model, and we will see how to interpret such models in Chapter 6.

Another way that independent variables can be perfectly collinear is when one independent variable can be expressed as an exact linear function of two or more of the other independent variables. For example, suppose we want to estimate the effect of campaign spending on campaign outcomes. For simplicity, assume that each election has two candidates. Let *voteA* be the percentage of the vote for Candidate A, let *expendA* be campaign expenditures by Candidate A, let *expendB* be campaign expenditures by Candidate B, and let *totexpend* be total campaign expenditures; the latter three variables are all measured in dollars. It may seem natural to specify the model as

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u, \tag{3.35}$$

in order to isolate the effects of spending by each candidate and the total amount of spending. But this model violates Assumption MLR.3 because $x_3 = x_1 + x_2$ by definition. Trying to interpret this equation in a ceteris paribus fashion reveals the problem. The parameter of β_1 in equation (3.35) is supposed to measure the effect of increasing expenditures by Candidate A by one dollar on Candidate A's vote, holding Candidate B's spending *and* total spending fixed. This is nonsense, because if *expendB* and *totexpend* are held fixed, then we cannot increase *expendA*.

The solution to the perfect collinearity in (3.35) is simple: drop any one of the three variables from the model. We would probably drop *totexpend*, and then the coefficient on *expendA* would measure the effect of increasing expenditures by A on the percentage of the vote received by A, holding the spending by B fixed.

The prior examples show that Assumption MLR.3 can fail if we are not careful in specifying our model. Assumption MLR.3 also fails if the sample size, n, is too small in relation to the number of parameters being estimated. In the general regression model in equation (3.31), there are $k + 1$ parameters, and MLR.3 fails if $n < k + 1$. Intuitively, this makes sense: to estimate $k + 1$ parameters, we need at least $k + 1$ observations. Not surprisingly, it is better to have as many observations as possible, something we will see with our variance calculations in Section 3.4.

QUESTION 3.3

In the previous example, if we use as explanatory variables *expendA*, *expendB*, and *shareA*, where *shareA* = 100·(*expendA*/*totexpend*) is the percentage share of total campaign expenditures made by Candidate A, does this violate Assumption MLR.3?

If the model is carefully specified and $n \geq k + 1$, Assumption MLR.3 can fail in rare cases due to bad luck in collecting the sample. For example, in a wage equation with education and experience as variables, it is possible that we could obtain a random sample where each individual has exactly twice as much education as years of experience. This scenario would cause Assumption MLR.3 to fail, but it can be considered very unlikely unless we have an extremely small sample size.

The final, and most important, assumption needed for unbiasedness is a direct extension of Assumption SLR.4.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0. \tag{3.36}$$

One way that Assumption MLR.4 can fail is if the functional relationship between the explained and explanatory variables is misspecified in equation (3.31): for example, if we forget to include the quadratic term inc^2 in the consumption function $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ when we estimate the model. Another functional form misspecification occurs when we use the level of a variable when the log of the variable is what actually shows up in the population model, or vice versa. For example, if the true model has

log(*wage*) as the dependent variable but we use *wage* as the dependent variable in our regression analysis, then the estimators will be biased. Intuitively, this should be pretty clear. We will discuss ways of detecting functional form misspecification in Chapter 9.

Omitting an important factor that is correlated with any of x_1, x_2, ..., x_k causes Assumption MLR.4 to fail also. With multiple regression analysis, we are able to include many factors among the explanatory variables, and omitted variables are less likely to be a problem in multiple regression analysis than in simple regression analysis. Nevertheless, in any application, there are always factors that, due to data limitations or ignorance, we will not be able to include. If we think these factors should be controlled for and they are correlated with one or more of the independent variables, then Assumption MLR.4 will be violated. We will derive this bias later.

There are other ways that u can be correlated with an explanatory variable. In Chapter 15, we will discuss the problem of measurement error in an explanatory variable. In Chapter 16, we cover the conceptually more difficult problem in which one or more of the explanatory variables is determined jointly with y. We must postpone our study of these problems until we have a firm grasp of multiple regression analysis under an ideal set of assumptions.

When Assumption MLR.4 holds, we often say that we have **exogenous explanatory variables**. If x_j is correlated with u for any reason, then x_j is said to be an **endogenous explanatory variable**. The terms "exogenous" and "endogenous" originated in simultaneous equations analysis (see Chapter 16), but the term "endogenous explanatory variable" has evolved to cover any case in which an explanatory variable may be correlated with the error term.

Before we show the unbiasedness of the OLS estimators under MLR.1 to MLR.4, a word of caution. Beginning students of econometrics sometimes confuse Assumptions MLR.3 and MLR.4, but they are quite different. Assumption MLR.3 rules out certain relationships among the independent or explanatory variables and has *nothing* to do with the error, u. You will know immediately when carrying out OLS estimation whether or not Assumption MLR.3 holds. On the other hand, Assumption MLR.4—the much more important of the two—restricts the relationship between the unobservables in u and the explanatory variables. Unfortunately, we will never know for sure whether the average value of the unobservables is unrelated to the explanatory variables. But this is the critical assumption.

We are now ready to show unbiasedness of OLS under the first four multiple regression assumptions. As in the simple regression case, the expectations are conditional on the values of the explanatory variables in the sample, something we show explicitly in Appendix 3A but not in the text.

Theorem 3.1 (Unbiasedness of OLS)

Under Assumptions MLR.1 through MLR.4,

$$\mathrm{E}(\hat{\beta}_j) = \beta_j, \; j = 0, 1, \ldots, k, \tag{3.37}$$

for any values of the population parameter β_j. In other words, the OLS estimators are unbiased estimators of the population parameters.

In our previous empirical examples, Assumption MLR.3 has been satisfied (because we have been able to compute the OLS estimates). Furthermore, for the most part, the samples are randomly chosen from a well-defined population. If we believe that the specified models are correct under the key Assumption MLR.4, then we can conclude that OLS is unbiased in these examples.

Since we are approaching the point where we can use multiple regression in serious empirical work, it is useful to remember the meaning of unbiasedness. It is tempting, in examples such as the wage equation in (3.19), to say something like "9.2 percent is an unbiased estimate of the return to education." As we know, an estimate cannot be unbiased: an estimate is a fixed number, obtained from a particular sample, which usually is not equal to the population parameter. When we say that OLS is unbiased under Assumptions MLR.1 through MLR.4, we mean that the *procedure* by which the OLS estimates are obtained is unbiased when we view the procedure as being applied across all possible random samples. We hope that we have obtained a sample that gives us an estimate close to the population value, but, unfortunately, this cannot be assured. What is assured is that we have no reason to believe our estimate is more likely to be too big or more likely to be too small.

Including Irrelevant Variables in a Regression Model

One issue that we can dispense with fairly quickly is that of **inclusion of an irrelevant variable** or **overspecifying the model** in multiple regression analysis. This means that one (or more) of the independent variables is included in the model even though it has no partial effect on y in the population. (That is, its population coefficient is zero.)

To illustrate the issue, suppose we specify the model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \tag{3.38}$$

and this model satisfies Assumptions MLR.1 through MLR.4. However, x_3 has no effect on y after x_1 and x_2 have been controlled for, which means that $\beta_3 = 0$. The variable x_3 may or may not be correlated with x_1 or x_2; all that matters is that, once x_1 and x_2 are controlled for, x_3 has no effect on y. In terms of conditional expectations, $E(y|x_1,x_2,x_3) = E(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Because we do not know that $\beta_3 = 0$, we are inclined to estimate the equation including x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3. \tag{3.39}$$

We have included the irrelevant variable, x_3, in our regression. What is the effect of including x_3 in (3.39) when its coefficient in the population model (3.38) is zero? In terms of the unbiasedness of $\hat{\beta}_1$ and $\hat{\beta}_2$, there is *no effect*. This conclusion requires no special derivation, as it follows immediately from Theorem 3.1. Remember, unbiasedness means $E(\hat{\beta}_j) = \beta_j$ for *any* value of β_j, including $\beta_j = 0$. Thus, we can conclude that $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$, and $E(\hat{\beta}_3) = 0$ (for any values of β_0, β_1, and β_2). Even though $\hat{\beta}_3$ itself will never be exactly zero, its average value across all random samples will be zero.

The conclusion of the preceding example is much more general: including one or more irrelevant variables in a multiple regression model, or overspecifying the model, does not affect the unbiasedness of the OLS estimators. Does this mean it is harmless to include irrelevant variables? No. As we will see in Section 3.4, including irrelevant variables can have undesirable effects on the *variances* of the OLS estimators.

Omitted Variable Bias: The Simple Case

Now suppose that, rather than including an irrelevant variable, we omit a variable that actually belongs in the true (or population) model. This is often called the problem of **excluding a relevant variable** or **underspecifying the model**. We claimed in Chapter 2 and earlier in this chapter that this problem generally causes the OLS estimators to be biased. It is time to show this explicitly and, just as importantly, to derive the direction and size of the bias.

Deriving the bias caused by omitting an important variable is an example of **misspecification analysis**. We begin with the case where the true population model has two explanatory variables and an error term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{3.40}$$

and we assume that this model satisfies Assumptions MLR.1 through MLR.4.

Suppose that our primary interest is in β_1, the partial effect of x_1 on y. For example, y is hourly wage (or log of hourly wage), x_1 is education, and x_2 is a measure of innate ability. In order to get an unbiased estimator of β_1, we *should* run a regression of y on x_1 and x_2 (which gives unbiased estimators of β_0, β_1, and β_2). However, due to our ignorance or data unavailability, we estimate the model by *excluding* x_2. In other words, we perform a simple regression of y on x_1 only, obtaining the equation

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.41}$$

We use the symbol "~" rather than "^" to emphasize that $\tilde{\beta}_1$ comes from an underspecified model.

When first learning about the omitted variable problem, it can be difficult to distinguish between the underlying true model, (3.40) in this case, and the model that we actually estimate, which is captured by the regression in (3.41). It may seem silly to omit the variable x_2 if it belongs in the model, but often we have no choice. For example, suppose that *wage* is determined by

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u. \tag{3.42}$$

Since ability is not observed, we instead estimate the model

$$wage = \beta_0 + \beta_1 educ + v,$$

where $v = \beta_2 abil + u$. The estimator of β_1 from the simple regression of *wage* on *educ* is what we are calling $\tilde{\beta}_1$.

We derive the expected value of $\tilde{\beta}_1$ conditional on the sample values of x_1 and x_2. Deriving this expectation is not difficult because $\tilde{\beta}_1$ is just the OLS slope estimator from a simple regression, and we have already studied this estimator extensively in Chapter 2. The difference here is that we must analyze its properties when the simple regression model is misspecified due to an omitted variable.

As it turns out, we have done almost all of the work to derive the bias in the simple regression estimator of $\tilde{\beta}_1$. From equation (3.23) we have the algebraic relationship $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the slope estimators (if we could have them) from the multiple regression

$$y_i \text{ on } x_{i1}, x_{i2}, i = 1, \ldots, n \tag{3.43}$$

and $\tilde{\delta}_1$ is the slope from the simple regression

$$x_{i2} \text{ on } x_{i1}, i = 1, \ldots, n. \tag{3.44}$$

Because $\tilde{\delta}_1$ depends only on the independent variables in the sample, we treat it as fixed (nonrandom) when computing $\mathrm{E}(\tilde{\beta}_1)$. Further, since the model in (3.40) satisfies Assumptions MLR.1 to MLR.4, we know that $\hat{\beta}_1$ and $\hat{\beta}_2$ would be unbiased for β_1 and β_2, respectively. Therefore,

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_1) &= \mathrm{E}(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = \mathrm{E}(\hat{\beta}_1) + \mathrm{E}(\hat{\beta}_2)\tilde{\delta}_1 \\ &= \beta_1 + \beta_2\tilde{\delta}_1, \end{aligned} \tag{3.45}$$

which implies the bias in $\tilde{\beta}_1$ is

$$\mathrm{Bias}(\tilde{\beta}_1) = \mathrm{E}(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1. \tag{3.46}$$

Because the bias in this case arises from omitting the explanatory variable x_2, the term on the right-hand side of equation (3.46) is often called the **omitted variable bias.**

From equation (3.46), we see that there are two cases where $\tilde{\beta}_1$ is unbiased. The first is pretty obvious: if $\beta_2 = 0$—so that x_2 does not appear in the true model (3.40)—then $\tilde{\beta}_1$ is unbiased. We already know this from the simple regression analysis in Chapter 2. The second case is more interesting. If $\tilde{\delta}_1 = 0$, then $\tilde{\beta}_1$ is unbiased for β_1, even if $\beta_2 \neq 0$.

Because $\tilde{\delta}_1$ is the sample covariance between x_1 and x_2 over the sample variance of x_1, $\tilde{\delta}_1 = 0$ if, and only if, x_1 and x_2 are uncorrelated in the sample. Thus, we have the important conclusion that, if x_1 and x_2 are uncorrelated in the sample, then $\tilde{\beta}_1$ is unbiased. This is not surprising: in Section 3.2, we showed that the simple regression estimator $\tilde{\beta}_1$ and the multiple regression estimator $\hat{\beta}_1$ are the same when x_1 and x_2 are uncorrelated in the sample. [We can also show that $\tilde{\beta}_1$ is unbiased without conditioning on the x_{i2} if $\mathrm{E}(x_2|x_1) = \mathrm{E}(x_2)$; then, for estimating β_1, leaving x_2 in the error term does not violate the zero conditional mean assumption for the error, once we adjust the intercept.]

When x_1 and x_2 are correlated, $\tilde{\delta}_1$ has the same sign as the correlation between x_1 and x_2: $\tilde{\delta}_1 > 0$ if x_1 and x_2 are positively correlated and $\tilde{\delta}_1 < 0$ if x_1 and x_2 are negatively correlated. The sign of the bias in $\tilde{\beta}_1$ depends on the signs of both β_2 and $\tilde{\delta}_1$ and is summarized

TABLE 3.2

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

in Table 3.2 for the four possible cases when there is bias. Table 3.2 warrants careful study. For example, the bias in $\tilde{\beta}_1$ is positive if $\beta_2 > 0$ (x_2 has a positive effect on y) and x_1 and x_2 are positively correlated, the bias is negative if $\beta_2 > 0$ and x_1 and x_2 are negatively correlated, and so on.

Table 3.2 summarizes the direction of the bias, but the size of the bias is also very important. A small bias of either sign need not be a cause for concern. For example, if the return to education in the population is 8.6 percent and the bias in the OLS estimator is 0.1 percent (a tenth of one percentage point), then we would not be very concerned. On the other hand, a bias on the order of three percentage points would be much more serious. The size of the bias is determined by the sizes of β_2 and $\tilde{\delta}_1$.

In practice, since β_2 is an unknown population parameter, we cannot be certain whether β_2 is positive or negative. Nevertheless, we usually have a pretty good idea about the direction of the partial effect of x_2 on y. Further, even though the sign of the correlation between x_1 and x_2 cannot be known if x_2 is not observed, in many cases, we can make an educated guess about whether x_1 and x_2 are positively or negatively correlated.

In the wage equation (3.42), by definition, more ability leads to higher productivity and therefore higher wages: $\beta_2 > 0$. Also, there are reasons to believe that *educ* and *abil* are positively correlated: on average, individuals with more innate ability choose higher levels of education. Thus, the OLS estimates from the simple regression equation $wage = \beta_0 + \beta_1 educ + v$ are *on average* too large. This does not mean that the estimate obtained from our sample is too big. We can only say that if we collect many random samples and obtain the simple regression estimates each time, then the average of these estimates will be greater than β_1.

EXAMPLE 3.6

(Hourly Wage Equation)

Suppose the model $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u$ satisfies Assumptions MLR.1 through MLR.4. The data set in WAGE1.RAW does not contain data on ability, so we estimate β_1 from the simple regression

$$\widehat{\log(wage)} = .584 + .083\, educ \qquad (3.47)$$
$$n = 526,\ R^2 = .186.$$

This is the result from only a single sample, so we cannot say that .083 is greater than β_1; the true return to education could be lower or higher than 8.3 percent (and we will never know for sure). Nevertheless, we know that the average of the estimates across all random samples would be too large.

As a second example, suppose that, at the elementary school level, the average score for students on a standardized exam is determined by

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 povrate + u, \tag{3.48}$$

where *expend* is expenditure per student and *povrate* is the poverty rate of the children in the school. Using school district data, we only have observations on the percentage of students with a passing grade and per student expenditures; we do not have information on poverty rates. Thus, we estimate β_1 from the simple regression of *avgscore* on *expend*.

We can again obtain the likely bias in $\tilde{\beta}_1$. First, β_2 is probably negative: there is ample evidence that children living in poverty score lower, on average, on standardized tests. Second, the average expenditure per student is probably negatively correlated with the poverty rate: the higher the poverty rate, the lower the average per student spending, so that $\text{Corr}(x_1, x_2) < 0$. From Table 3.2, $\tilde{\beta}_1$ will have a positive bias. This observation has important implications. It could be that the true effect of spending is zero; that is, $\beta_1 = 0$. However, the simple regression estimate of β_1 will usually be greater than zero, and this could lead us to conclude that expenditures are important when they are not.

When reading and performing empirical work in economics, it is important to master the terminology associated with biased estimators. In the context of omitting a variable from model (3.40), if $E(\tilde{\beta}_1) > \beta_1$, then we say that $\tilde{\beta}_1$ has an **upward bias**. When $E(\tilde{\beta}_1) < \beta_1$, $\tilde{\beta}_1$ has a **downward bias**. These definitions are the same whether β_1 is positive or negative. The phrase **biased towards zero** refers to cases where $E(\tilde{\beta}_1)$ is closer to zero than β_1. Therefore, if β_1 is positive, then $\tilde{\beta}_1$ is biased towards zero if it has a downward bias. On the other hand, if $\beta_1 < 0$, then $\tilde{\beta}_1$ is biased towards zero if it has an upward bias.

Omitted Variable Bias: More General Cases

Deriving the sign of omitted variable bias when there are multiple regressors in the estimated model is more difficult. We must remember that correlation between a single explanatory variable and the error generally results in *all* OLS estimators being biased. For example, suppose the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \tag{3.49}$$

satisfies Assumptions MLR.1 through MLR.4. But we omit x_3 and estimate the model as

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2. \tag{3.50}$$

Now, suppose that x_2 and x_3 are uncorrelated, but that x_1 is correlated with x_3. In other words, x_1 is correlated with the omitted variable, but x_2 is not. It is tempting to think that, while $\tilde{\beta}_1$ is probably biased based on the derivation in the previous subsection, $\tilde{\beta}_2$ is unbiased because x_2 is uncorrelated with x_3. Unfortunately, this is *not* generally the case: both $\tilde{\beta}_1$ and $\tilde{\beta}_2$ will normally be biased. The only exception to this is when x_1 and x_2 are also uncorrelated.

Even in the fairly simple model above, it can be difficult to obtain the direction of bias in $\tilde{\beta}_1$ and $\tilde{\beta}_2$. This is because x_1, x_2, and x_3 can all be pairwise correlated. Nevertheless, an approximation is often practically useful. If we assume that x_1 and x_2 are uncorrelated, then we can study the bias in $\tilde{\beta}_1$ as if x_2 were absent from both the population and the estimated models. In fact, when x_1 and x_2 are uncorrelated, it can be shown that

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)x_{i3}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}.$$

This is just like equation (3.45), but β_3 replaces β_2, and x_3 replaces x_2 in regression (3.44). Therefore, the bias in $\tilde{\beta}_1$ is obtained by replacing β_2 with β_3 and x_2 with x_3 in Table 3.2. If $\beta_3 > 0$ and $\mathrm{Corr}(x_1, x_3) > 0$, the bias in $\tilde{\beta}_1$ is positive, and so on.

As an example, suppose we add *exper* to the wage model:

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

If *abil* is omitted from the model, the estimators of both β_1 and β_2 are biased, even if we assume *exper* is uncorrelated with *abil*. We are mostly interested in the return to education, so it would be nice if we could conclude that $\tilde{\beta}_1$ has an upward or a downward bias due to omitted ability. This conclusion is not possible without further assumptions. As an *approximation*, let us suppose that, in addition to *exper* and *abil* being uncorrelated, *educ* and *exper* are also uncorrelated. (In reality, they are somewhat negatively correlated.) Since $\beta_3 > 0$ and *educ* and *abil* are positively correlated, $\tilde{\beta}_1$ would have an upward bias, just as if *exper* were not in the model.

The reasoning used in the previous example is often followed as a rough guide for obtaining the likely bias in estimators in more complicated models. Usually, the focus is on the relationship between a particular explanatory variable, say, x_1, and the key omitted factor. Strictly speaking, ignoring all other explanatory variables is a valid practice only when each one is uncorrelated with x_1, but it is still a useful guide. Appendix 3A contains a more careful analysis of omitted variable bias with multiple explanatory variables.

3.4 The Variance of the OLS Estimators

We now obtain the variance of the OLS estimators so that, in addition to knowing the central tendencies of the $\hat{\beta}_j$, we also have a measure of the spread in its sampling distribution. Before finding the variances, we add a homoskedasticity assumption, as in Chapter 2. We do this for two reasons. First, the formulas are simplified by imposing the constant error

variance assumption. Second, in Section 3.5, we will see that OLS has an important efficiency property if we add the homoskedasticity assumption.

In the multiple regression framework, homoskedasticity is stated as follows:

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any values of the explanatory variables. In other words, $\text{Var}(u|x_1,\ldots,x_k) = \sigma^2$.

Assumption MLR.5 means that the variance in the error term, u, conditional on the explanatory variables, is the *same* for all combinations of outcomes of the explanatory variables. If this assumption fails, then the model exhibits heteroskedasticity, just as in the two-variable case.

In the equation

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

homoskedasticity requires that the variance of the unobserved error u does not depend on the levels of education, experience, or tenure. That is,

$$\text{Var}(u|educ, exper, tenure) = \sigma^2.$$

If this variance changes with any of the three explanatory variables, then heteroskedasticity is present.

Assumptions MLR.1 through MLR.5 are collectively known as the **Gauss-Markov assumptions** (for cross-sectional regression). So far, our statements of the assumptions are suitable only when applied to cross-sectional analysis with random sampling. As we will see, the Gauss-Markov assumptions for time series analysis, and for other situations such as panel data analysis, are more difficult to state, although there are many similarities.

In the discussion that follows, we will use the symbol $\mathbf{x}$ to denote the set of all independent variables, $(x_1, \ldots, x_k)$. Thus, in the wage regression with *educ*, *exper*, and *tenure* as independent variables, $\mathbf{x} = (educ, exper, tenure)$. Then we can write Assumptions MLR.1 and MLR.4 as

$$\text{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

and Assumption MLR.5 is the same as $\text{Var}(y|\mathbf{x}) = \sigma^2$. Stating the assumptions in this way clearly illustrates how Assumption MLR.5 differs greatly from Assumption MLR.4. Assumption MLR.4 says that the expected value of y, given $\mathbf{x}$, is linear in the parameters, but it certainly depends on $x_1, x_2, \ldots, x_k$. Assumption MLR.5 says that the variance of y, given $\mathbf{x}$, does *not* depend on the values of the independent variables.

We can now obtain the variances of the $\hat{\beta}_j$, where we again condition on the sample values of the independent variables. The proof is in the appendix to this chapter.

Theorem 3.2 (Sampling Variances of the OLS Slope Estimators)
Under Assumptions MLR.1 through MLR.5, conditional on the sample values of the independent variables,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)}, \tag{3.51}$$

for $j = 1, 2, \ldots, k$, where $\text{SST}_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

Before we study equation (3.51) in more detail, it is important to know that all of the Gauss-Markov assumptions are used in obtaining this formula. Whereas we did not need the homoskedasticity assumption to conclude that OLS is unbiased, we do need it to validate equation (3.51).

The size of $\text{Var}(\hat{\beta}_j)$ is practically important. A larger variance means a less precise estimator, and this translates into larger confidence intervals and less accurate hypotheses tests (as we will see in Chapter 4). In the next subsection, we discuss the elements comprising (3.51).

The Components of the OLS Variances: Multicollinearity

Equation (3.51) shows that the variance of $\hat{\beta}_j$ depends on three factors: σ^2, SST_j, and R_j^2. Remember that the index j simply denotes any one of the independent variables (such as education or poverty rate). We now consider each of the factors affecting $\text{Var}(\hat{\beta}_j)$ in turn.

THE ERROR VARIANCE, σ^2. From equation (3.51), a larger σ^2 means larger variances for the OLS estimators. This is not at all surprising: more "noise" in the equation (a larger σ^2) makes it more difficult to estimate the partial effect of any of the independent variables on y, and this is reflected in higher variances for the OLS slope estimators. Because σ^2 is a feature of the population, it has nothing to do with the sample size. It is the one component of (3.51) that is unknown. We will see later how to obtain an unbiased estimator of σ^2.

For a given dependent variable y, there is really only one way to reduce the error variance, and that is to add more explanatory variables to the equation (take some factors out of the error term). Unfortunately, it is not always possible to find additional legitimate factors that affect y.

THE TOTAL SAMPLE VARIATION IN x_j, SST_j. From equation (3.51), we see that the larger the total variation in x_j is, the smaller is $\text{Var}(\hat{\beta}_j)$. Thus, everything else being equal, for estimating β_j, we prefer to have as much sample variation in x_j as possible. We already discovered this in the simple regression case in Chapter 2. Although it is rarely

possible for us to choose the sample values of the independent variables, there *is* a way to increase the sample variation in each of the independent variables: increase the sample size. In fact, when sampling randomly from a population, SST_j increases without bound as the sample size gets larger and larger. This is the component of the variance that systematically depends on the sample size.

When SST_j is small, $\text{Var}(\hat{\beta}_j)$ can get very large, but a small SST_j is not a violation of Assumption MLR.3. Technically, as SST_j goes to zero, $\text{Var}(\hat{\beta}_j)$ approaches infinity. The extreme case of no sample variation in x_j, $SST_j = 0$, is not allowed by Assumption MLR.3.

THE LINEAR RELATIONSHIPS AMONG THE INDEPENDENT VARIABLES, R_j^2. The term R_j^2 in equation (3.51) is the most difficult of the three components to understand. This term does not appear in simple regression analysis because there is only one independent variable in such cases. It is important to see that this R-squared is distinct from the R-squared in the regression of y on $x_1, x_2, \ldots, x_k$: R_j^2 is obtained from a regression involving only the independent variables in the original model, where x_j plays the role of a dependent variable.

Consider first the $k = 2$ case: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Then, $\text{Var}(\hat{\beta}_1) = \sigma^2/[SST_1(1 - R_1^2)]$, where R_1^2 is the R-squared from the simple regression of x_1 on x_2 (and an intercept, as always). Because the R-squared measures goodness-of-fit, a value of R_1^2 close to one indicates that x_2 explains much of the variation in x_1 in the sample. This means that x_1 and x_2 are highly correlated.

As R_1^2 increases to one, $\text{Var}(\hat{\beta}_1)$ gets larger and larger. Thus, a high degree of linear relationship between x_1 and x_2 can lead to large variances for the OLS slope estimators. (A similar argument applies to $\hat{\beta}_2$.) See Figure 3.1 for the relationship between $\text{Var}(\hat{\beta}_1)$ and the R-squared from the regression of x_1 on x_2.

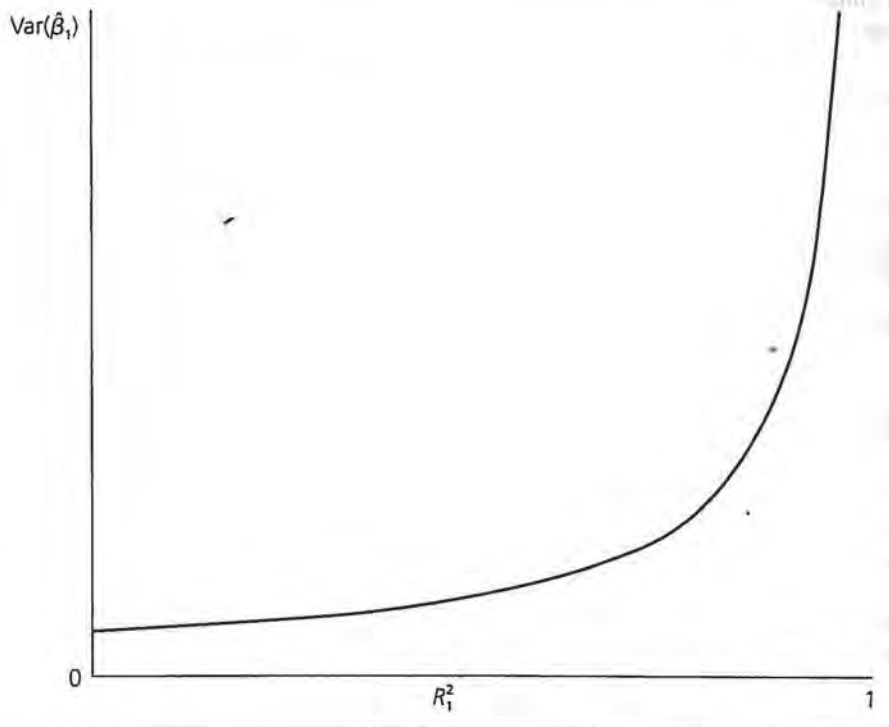
In the general case, R_j^2 is the proportion of the total variation in x_j that can be explained by the *other* independent variables appearing in the equation. For a given σ^2 and SST_j, the smallest $\text{Var}(\hat{\beta}_j)$ is obtained when $R_j^2 = 0$, which happens if, and only if, x_j has zero sample correlation with *every other* independent variable. This is the best case for estimating β_j, but it is rarely encountered.

The other extreme case, $R_j^2 = 1$, is ruled out by Assumption MLR.3, because $R_j^2 = 1$ means that, in the sample, x_j is a *perfect* linear combination of some of the other independent variables in the regression. A more relevant case is when R_j^2 is "close" to one. From equation (3.51) and Figure 3.1, we see that this can cause $\text{Var}(\hat{\beta}_j)$ to be large: $\text{Var}(\hat{\beta}_j) \to \infty$ as $R_j^2 \to 1$. High (but not perfect) correlation between two or more independent variables is called **multicollinearity**.

Before we discuss the multicollinearity issue further, it is important to be very clear on one thing: a case where R_j^2 is close to one is *not* a violation of Assumption MLR.3.

Since multicollinearity violates none of our assumptions, the "problem" of multicollinearity is not really well defined. When we say that multicollinearity arises for estimating β_j when R_j^2 is "close" to one, we put "close" in quotation marks because there is no absolute number that we can cite to conclude that multicollinearity is a problem. For example, $R_j^2 = .9$ means that 90 percent of the sample variation in x_j can be explained by the other independent variables in the regression model. Unquestionably, this means that x_j has a strong linear relationship to the other independent variables. But whether this trans-

FIGURE 3.1
$\text{Var}(\hat{\beta}_1)$ as a function of R_1^2.



lates into a $\text{Var}(\hat{\beta}_j)$ that is too large to be useful depends on the sizes of σ^2 and SST_j. As we will see in Chapter 4, for statistical inference, what ultimately matters is how big $\hat{\beta}_j$ is in relation to its standard deviation.

Just as a large value of R_j^2 can cause a large $\text{Var}(\hat{\beta}_j)$, so can a small value of SST_j. Therefore, a small sample size can lead to large sampling variances, too. Worrying about high degrees of correlation among the independent variables in the sample is really no different from worrying about a small sample size: both work to increase $\text{Var}(\hat{\beta}_j)$. The famous University of Wisconsin econometrician Arthur Goldberger, reacting to econometricians' obsession with multicollinearity, has (tongue in cheek) coined the term **micronumerosity**, which he defines as the "problem of small sample size." [For an engaging discussion of multicollinearity and micronumerosity, see Goldberger (1991).]

Although the problem of multicollinearity cannot be clearly defined, one thing is clear: everything else being equal, for estimating β_j, it is better to have less correlation between x_j and the other independent variables. This observation often leads to a discussion of how

to "solve" the multicollinearity problem. In the social sciences, where we are usually passive collectors of data, there is no good way to reduce variances of unbiased estimators other than to collect more data. For a given data set, we can try dropping other independent variables from the model in an effort to reduce multicollinearity. Unfortunately, dropping a variable that belongs in the population model can lead to bias, as we saw in Section 3.3.

Perhaps an example at this point will help clarify some of the issues raised concerning multicollinearity. Suppose we are interested in estimating the effect of various school expenditure categories on student performance. It is likely that expenditures on teacher salaries, instructional materials, athletics, and so on, are highly correlated: wealthier schools tend to spend more on everything, and poorer schools spend less on everything. Not surprisingly, it can be difficult to estimate the effect of any particular expenditure category on student performance when there is little variation in one category that cannot largely be explained by variations in the other expenditure categories (this leads to high R_j^2 for each of the expenditure variables). Such multicollinearity problems can be mitigated by collecting more data, but in a sense we have imposed the problem on ourselves: we are asking questions that may be too subtle for the available data to answer with any precision. We can probably do much better by changing the scope of the analysis and lumping all expenditure categories together, since we would no longer be trying to estimate the partial effect of each separate category.

Another important point is that a high degree of correlation between certain independent variables can be irrelevant as to how well we can estimate other parameters in the model. For example, consider a model with three independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

where x_2 and x_3 are highly correlated. Then $\text{Var}(\hat{\beta}_2)$ and $\text{Var}(\hat{\beta}_3)$ may be large. But the amount of correlation between x_2 and x_3 has no direct effect on $\text{Var}(\hat{\beta}_1)$. In fact, if x_1 is uncorrelated with x_2 and x_3, then $R_1^2 = 0$ and $\text{Var}(\hat{\beta}_1) = \sigma^2/\text{SST}_1$, regardless of how much correlation there is between x_2 and x_3. If β_1 is the parameter of interest, we do not really care about the amount of correlation between x_2 and x_3.

QUESTION 3.4

Suppose you postulate a model explaining final exam score in terms of class attendance. Thus, the dependent variable is final exam score, and the key explanatory variable is number of classes attended. To control for student abilities and efforts outside the classroom, you include among the explanatory variables cumulative GPA, SAT score, and measures of high school performance. Someone says, "You cannot hope to learn anything from this exercise because cumulative GPA, SAT score, and high school performance are likely to be highly collinear." What should be your response?

The previous observation is important because economists often include many control variables in order to isolate the causal effect of a particular variable. For example, in looking at the relationship between loan approval rates and percent of minorities in a neighborhood, we might include variables like average income, average housing value, measures of creditworthiness, and so on, because these factors need to be accounted for in order to draw causal conclusions about discrimination. Income, housing prices, and creditworthiness are generally highly correlated with each other. But high correlations among these controls do not make it more difficult to determine the effects of discrimination.

Variances in Misspecified Models

The choice of whether or not to include a particular variable in a regression model can be made by analyzing the tradeoff between bias and variance. In Section 3.3, we derived the bias induced by leaving out a relevant variable when the true model contains two explanatory variables. We continue the analysis of this model by comparing the variances of the OLS estimators.

Write the true population model, which satisfies the Gauss-Markov assumptions, as

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

We consider two estimators of β_1. The estimator $\hat{\beta}_1$ comes from the multiple regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \tag{3.52}$$

In other words, we include x_2, along with x_1, in the regression model. The estimator $\tilde{\beta}_1$ is obtained by omitting x_2 from the model and running a simple regression of y on x_1:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.53}$$

When $\beta_2 \neq 0$, equation (3.53) excludes a relevant variable from the model and, as we saw in Section 3.3, this induces a bias in $\tilde{\beta}_1$ unless x_1 and x_2 are uncorrelated. On the other hand, $\hat{\beta}_1$ is unbiased for β_1 for any value of β_2, including $\beta_2 = 0$. It follows that, if bias is used as the only criterion, $\hat{\beta}_1$ is preferred to $\tilde{\beta}_1$.

The conclusion that $\hat{\beta}_1$ is always preferred to $\tilde{\beta}_1$ does not carry over when we bring variance into the picture. Conditioning on the values of x_1 and x_2 in the sample, we have, from (3.51),

$$\text{Var}(\hat{\beta}_1) = \sigma^2/[\text{SST}_1(1 - R_1^2)], \tag{3.54}$$

where SST_1 is the total variation in x_1, and R_1^2 is the R-squared from the regression of x_1 on x_2. Further, a simple modification of the proof in Chapter 2 for two-variable regression shows that

$$\text{Var}(\tilde{\beta}_1) = \sigma^2/\text{SST}_1. \tag{3.55}$$

Comparing (3.55) to (3.54) shows that $\text{Var}(\tilde{\beta}_1)$ is always *smaller* than $\text{Var}(\hat{\beta}_1)$, unless x_1 and x_2 are uncorrelated in the sample, in which case the two estimators $\tilde{\beta}_1$ and $\hat{\beta}_1$ are the same. Assuming that x_1 and x_2 are not uncorrelated, we can draw the following conclusions:

1. When $\beta_2 \neq 0$, $\tilde{\beta}_1$ is biased, $\hat{\beta}_1$ is unbiased, and $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.
2. When $\beta_2 = 0$, $\tilde{\beta}_1$ and $\hat{\beta}_1$ are both unbiased, and $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.

From the second conclusion, it is clear that $\tilde{\beta}_1$ is preferred if $\beta_2 = 0$. Intuitively, if x_2 does not have a partial effect on y, then including it in the model can only exacerbate the multicollinearity problem, which leads to a less efficient estimator of β_1. A higher variance for the estimator of β_1 is the cost of including an irrelevant variable in a model.

The case where $\beta_2 \neq 0$ is more difficult. Leaving x_2 out of the model results in a biased estimator of β_1. Traditionally, econometricians have suggested comparing the likely size of the bias due to omitting x_2 with the reduction in the variance—summarized in the size of R_1^2—to decide whether x_2 should be included. However, when $\beta_2 \neq 0$, there are two favorable reasons for including x_2 in the model. The most important of these is that any bias in $\tilde{\beta}_1$ does not shrink as the sample size grows; in fact, the bias does not necessarily follow any pattern. Therefore, we can usefully think of the bias as being roughly the same for any sample size. On the other hand, $\text{Var}(\tilde{\beta}_1)$ and $\text{Var}(\hat{\beta}_1)$ both shrink to zero as n gets large, which means that the multicollinearity induced by adding x_2 becomes less important as the sample size grows. In large samples, we would prefer $\hat{\beta}_1$.

The other reason for favoring $\hat{\beta}_1$ is more subtle. The variance formula in (3.55) is conditional on the values of x_{i1} and x_{i2} in the sample, which provides the best scenario for $\tilde{\beta}_1$. When $\beta_2 \neq 0$, the variance of $\tilde{\beta}_1$ conditional only on x_1 is larger than that presented in (3.55). Intuitively, when $\beta_2 \neq 0$ and x_2 is excluded from the model, the error variance increases because the error effectively contains part of x_2. But (3.55) ignores the error variance increase because it treats both regressors as nonrandom. A full discussion of which independent variables to condition on would lead us too far astray. It is sufficient to say that (3.55) is too generous when it comes to measuring the precision in $\tilde{\beta}_1$.

Estimating σ^2: Standard Errors of the OLS Estimators

We now show how to choose an unbiased estimator of σ^2, which then allows us to obtain unbiased estimators of $\text{Var}(\hat{\beta}_j)$.

Because $\sigma^2 = \text{E}(u^2)$, an unbiased "estimator" of σ^2 is the sample average of the squared errors: $n^{-1}\sum_{i=1}^{n} u_i^2$. Unfortunately, this is not a true estimator because we do not observe the u_i. Nevertheless, recall that the errors can be written as $u_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik}$, and so the reason we do not observe the u_i is that we do not know the β_j. When we replace each β_j with its OLS estimator, we get the OLS residuals:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}.$$

It seems natural to estimate σ^2 by replacing u_i with the $\hat{u}_i$. In the simple regression case, we saw that this leads to a biased estimator. The unbiased estimator of σ^2 in the general multiple regression case is

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2\right) \Big/ (n - k - 1) = \text{SSR}/(n - k - 1). \tag{3.56}$$

We already encountered this estimator in the $k = 1$ case in simple regression.

The term $n - k - 1$ in (3.56) is the **degrees of freedom** (*df*) for the general OLS problem with n observations and k independent variables. Since there are $k + 1$ parameters in a regression model with k independent variables and an intercept, we can write

$$df = n - (k + 1)$$
$$= (\text{number of observations}) - (\text{number of estimated parameters}). \quad (3.57)$$

This is the easiest way to compute the degrees of freedom in a particular application: count the number of parameters, including the intercept, and subtract this amount from the number of observations. (In the rare case that an intercept is not estimated, the number of parameters decreases by one.)

Technically, the division by $n - k - 1$ in (3.56) comes from the fact that the expected value of the sum of squared residuals is $E(SSR) = (n - k - 1)\sigma^2$. Intuitively, we can figure out why the degrees of freedom adjustment is necessary by returning to the first order conditions for the OLS estimators. These can be written as $\sum_{i=1}^{n} \hat{u}_i = 0$ and $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, where $j = 1, 2, \ldots, k$. Thus, in obtaining the OLS estimates, $k + 1$ restrictions are imposed on the OLS residuals. This means that, given $n - (k + 1)$ of the residuals, the remaining $k + 1$ residuals are known: there are only $n - (k + 1)$ degrees of freedom in the residuals. (This can be contrasted with the *errors* u_i, which have n degrees of freedom in the sample.)

For reference, we summarize this discussion with Theorem 3.3. We proved this theorem for the case of simple regression analysis in Chapter 2 (see Theorem 2.3). (A general proof that requires matrix algebra is provided in Appendix E.)

Theorem 3.3 (Unbiased Estimation of σ^2)

Under the Gauss-Markov Assumptions MLR.1 through MLR.5, $E(\hat{\sigma}^2) = \sigma^2$.

The positive square root of $\hat{\sigma}^2$, denoted $\hat{\sigma}$, is called the **standard error of the regression** (**SER**). The SER is an estimator of the standard deviation of the error term. This estimate is usually reported by regression packages, although it is called different things by different packages. (In addition to SER, $\hat{\sigma}$ is also called the *standard error of the estimate* and the *root mean squared error*.)

Note that $\hat{\sigma}$ can either decrease or increase when another independent variable is added to a regression (for a given sample). This is because, although SSR must fall when another explanatory variable is added, the degrees of freedom also falls by one. Because SSR is in the numerator and df is in the denominator, we cannot tell beforehand which effect will dominate.

For constructing confidence intervals and conducting tests in Chapter 4, we will need to estimate the **standard deviation of $\hat{\beta}_j$**, which is just the square root of the variance:

$$sd(\hat{\beta}_j) = \sigma/[SST_j(1 - R_j^2)]^{1/2}.$$

Since σ is unknown, we replace it with its estimator, $\hat{\sigma}$. This gives us the **standard error of $\hat{\beta}_j$**:

$$se(\hat{\beta}_j) = \hat{\sigma}/[SST_j(1 - R_j^2)]^{1/2}. \quad (3.58)$$

Just as the OLS estimates can be obtained for any given sample, so can the standard errors. Since $\text{se}(\hat{\beta}_j)$ depends on $\hat{\sigma}$, the standard error has a sampling distribution, which will play a role in Chapter 4.

We should emphasize one thing about standard errors. Because (3.58) is obtained directly from the variance formula in (3.51), and because (3.51) relies on the homoskedasticity Assumption MLR.5, it follows that the standard error formula in (3.58) is *not* a valid estimator of $\text{sd}(\hat{\beta}_j)$ if the errors exhibit heteroskedasticity. Thus, while the presence of heteroskedasticity does not cause bias in the $\hat{\beta}_j$, it does lead to bias in the usual formula for $\text{Var}(\hat{\beta}_j)$, which then invalidates the standard errors. This is important because any regression package computes (3.58) as the default standard error for each coefficient (with a somewhat different representation for the intercept). If we suspect heteroskedasticity, then the "usual" OLS standard errors are invalid, and some corrective action should be taken. We will see in Chapter 8 what methods are available for dealing with heteroskedasticity.

3.5 Efficiency of OLS: The Gauss-Markov Theorem

In this section, we state and discuss the important **Gauss-Markov Theorem**, which justifies the use of the OLS method rather than using a variety of competing estimators. We know one justification for OLS already: under Assumptions MLR.1 through MLR.4, OLS is unbiased. However, there are *many* unbiased estimators of the β_j under these assumptions (for example, see Problem 3.13). Might there be other unbiased estimators with variances smaller than the OLS estimators?

If we limit the class of competing estimators appropriately, then we can show that OLS *is* best within this class. Specifically, we will argue that, under Assumptions MLR.1 through MLR.5, the OLS estimator $\hat{\beta}_j$ for β_j is the **best linear unbiased estimator (BLUE)**. In order to state the theorem, we need to understand each component of the acronym "BLUE." First, we know what an estimator is: it is a rule that can be applied to any sample of data to produce an estimate. We also know what an unbiased estimator is: in the current context, an estimator, say, $\tilde{\beta}_j$, of β_j is an unbiased estimator of β_j if $\text{E}(\tilde{\beta}_j) = \beta_j$ for any $\beta_0, \beta_1, \ldots, \beta_k$.

What about the meaning of the term "linear"? In the current context, an estimator $\tilde{\beta}_j$ of β_j is linear if, and only if, it can be expressed as a linear function of the data on the dependent variable:

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i, \tag{3.59}$$

where each w_{ij} can be a function of the sample values of all the independent variables. The OLS estimators are linear, as can be seen from equation (3.22).

Finally, how do we define "best"? For the current theorem, best is defined as *smallest variance*. Given two unbiased estimators, it is logical to prefer the one with the smallest variance (see Appendix C).

Now, let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ denote the OLS estimators in model (3.31) under Assumptions MLR.1 through MLR.5. The Gauss-Markov Theorem says that, for any estimator $\tilde{\beta}_j$ that is *linear* and *unbiased*, $\text{Var}(\hat{\beta}_j) \leq \text{Var}(\tilde{\beta}_j)$, and the inequality is usually strict. In other words, in the class of linear unbiased estimators, OLS has the smallest variance (under the five Gauss-Markov assumptions). Actually, the theorem says more than this. If we want to estimate any linear function of the β_j, then the corresponding linear combination of the OLS estimators achieves the smallest variance among all linear unbiased estimators. We conclude with a theorem, which is proven in Appendix 3A.

Theorem 3.4 (Gauss-Markov Theorem)

Under Assumptions MLR.1 through MLR.5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the best linear unbiased estimators (BLUEs) of $\beta_0, \beta_1, \ldots, \beta_k$, respectively.

It is because of this theorem that Assumptions MLR.1 through MLR.5 are known as the Gauss-Markov assumptions (for cross-sectional analysis).

The importance of the Gauss-Markov Theorem is that, when the standard set of assumptions holds, we need not look for alternative unbiased estimators of the form in (3.59): none will be better than OLS. Equivalently, if we are presented with an estimator that is both linear and unbiased, then we know that the variance of this estimator is at least as large as the OLS variance; no additional calculation is needed to show this.

For our purposes, Theorem 3.4 justifies the use of OLS to estimate multiple regression models. If any of the Gauss-Markov assumptions fail, then this theorem no longer holds. We already know that failure of the zero conditional mean assumption (Assumption MLR.4) causes OLS to be biased, so Theorem 3.4 also fails. We also know that heteroskedasticity (failure of Assumption MLR.5) does not cause OLS to be biased. However, OLS no longer has the smallest variance among linear unbiased estimators in the presence of heteroskedasticity. In Chapter 8, we analyze an estimator that improves upon OLS when we know the brand of heteroskedasticity.

SUMMARY

1. The multiple regression model allows us to effectively hold other factors fixed while examining the effects of a particular independent variable on the dependent variable. It explicitly allows the independent variables to be correlated.

2. Although the model is linear in its *parameters*, it can be used to model nonlinear relationships by appropriately choosing the dependent and independent variables.

3. The method of ordinary least squares is easily applied to estimate the multiple regression model. Each slope estimate measures the partial effect of the corresponding independent variable on the dependent variable, holding all other independent variables fixed.

4. R^2 is the proportion of the sample variation in the dependent variable explained by the independent variables, and it serves as a goodness-of-fit measure. It is important not to put too much weight on the value of R^2 when evaluating econometric models.

5. Under the first four Gauss-Markov assumptions (MLR.1 through MLR.4), the OLS estimators are unbiased. This implies that including an irrelevant variable in a model has no effect on the unbiasedness of the intercept and other slope estimators. On the other hand, omitting a relevant variable causes OLS to be biased. In many circumstances, the direction of the bias can be determined.

6. Under the five Gauss-Markov assumptions, the variance of an OLS slope estimator is given by $\text{Var}(\hat{\beta}_j) = \sigma^2/[\text{SST}_j(1 - R_j^2)]$. As the error variance σ^2 increases, so does $\text{Var}(\hat{\beta}_j)$, while $\text{Var}(\hat{\beta}_j)$ decreases as the sample variation in x_j, SST_j, increases. The term R_j^2 measures the amount of collinearity between x_j and the other explanatory variables. As R_j^2 approaches one, $\text{Var}(\hat{\beta}_j)$ is unbounded.

7. Adding an irrelevant variable to an equation generally increases the variances of the remaining OLS estimators because of multicollinearity.

8. Under the Gauss-Markov assumptions (MLR.1 through MLR.5), the OLS estimators are best linear unbiased estimators (BLUEs).

The Gauss-Markov Assumptions

The following is a summary of the five Gauss-Markov assumptions that we used in this chapter. Remember, the first four were used to establish unbiasedness of OLS, while the fifth was added to derive the usual variance formulas and to conclude that OLS is best linear unbiased.

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

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

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobservable random error or disturbance term.

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0.$$

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, \ldots, x_k) = \sigma^2.$$

KEY TERMS

Best Linear Unbiased Estimator (BLUE)
Biased Towards Zero
Ceteris Paribus
Degrees of Freedom (*df*)
Disturbance
Downward Bias
Endogenous Explanatory Variable
Error Term
Excluding a Relevant Variable
Exogenous Explanatory Variable
Explained Sum of Squares (SSE)
First Order Conditions
Gauss-Markov Assumptions
Gauss-Markov Theorem
Inclusion of an Irrelevant Variable
Intercept
Micronumerosity
Misspecification Analysis
Multicollinearity
Multiple Linear Regression Model
Multiple Regression Analysis
OLS Intercept Estimate
OLS Regression Line
OLS Slope Estimate
Omitted Variable Bias
Ordinary Least Squares
Overspecifying the Model
Partial Effect
Perfect Collinearity
Population Model
Residual
Residual Sum of Squares
Sample Regression Function (SRF)
Slope Parameter
Standard Deviation of $\hat{\beta}_j$
Standard Error of $\hat{\beta}_j$
Standard Error of the Regression (SER)
Sum of Squared Residuals (SSR)
Total Sum of Squares (SST)
True Model
Underspecifying the Model
Upward Bias

PROBLEMS

3.1 Using the data in GPA2.RAW on 4,137 college students, the following equation was estimated by OLS:

$$\widehat{colgpa} = 1.392 - .0135\ hsperc + .00148\ sat$$
$$n = 4{,}137, R^2 = .273,$$

where *colgpa* is measured on a four-point scale, *hsperc* is the percentile in the high school graduating class (defined so that, for example, *hsperc* = 5 means the *top* five percent of the class), and *sat* is the combined math and verbal scores on the student achievement test.

(i) Why does it make sense for the coefficient on *hsperc* to be negative?

(ii) What is the predicted college GPA when *hsperc* = 20 and *sat* = 1050?

(iii) Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but Student A's SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?

(iv) Holding *hsperc* fixed, what difference in SAT scores leads to a predicted *colgpa* difference of .50, or one-half of a grade point? Comment on your answer.

3.2 The data in WAGE2.RAW on working men was used to estimate the following equation:

$$\widehat{educ} = 10.36 - .094\, sibs + .131\, meduc + .210\, feduc$$
$$n = 722, R^2 = .214,$$

where *educ* is years of schooling, *sibs* is number of siblings, *meduc* is mother's years of schooling, and *feduc* is father's years of schooling.

(i) Does *sibs* have the expected effect? Explain. Holding *meduc* and *feduc* fixed, by how much does *sibs* have to increase to reduce predicted years of education by one year? (A noninteger answer is acceptable here.)

(ii) Discuss the interpretation of the coefficient on *meduc*.

(iii) Suppose that Man A has no siblings, and his mother and father each have 12 years of education. Man B has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between B and A?

3.3 The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years. (See also Computer Exercise C2.3.)

(i) If adults trade off sleep for work, what is the sign of β_1?

(ii) What signs do you think β_2 and β_3 will have?

(iii) Using the data in SLEEP75.RAW, the estimated equation is

$$\widehat{sleep} = 3638.25 - .148\, totwrk - 11.13\, educ + 2.20\, age$$
$$n = 706, R^2 = .113.$$

If someone works five more hours per week, by how many minutes is *sleep* predicted to fall? Is this a large tradeoff?

(iv) Discuss the sign and magnitude of the estimated coefficient on *educ*.

(v) Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*? What other factors might affect the time spent sleeping? Are these likely to be correlated with *totwrk*?

3.4 The median starting salary for new law school graduates is determined by

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

where *LSAT* is the median LSAT score for the graduating class, *GPA* is the median college GPA for the class, *libvol* is the number of volumes in the law school library, *cost* is

the annual cost of attending law school, and *rank* is a law school ranking (with *rank* = 1 being the best).

(i) Explain why we expect $\beta_5 \leq 0$.

(ii) What signs do you expect for the other slope parameters? Justify your answers.

(iii) Using the data in LAWSCH85.RAW, the estimated equation is

$$\widehat{\log(salary)} = 8.34 + .0047\, LSAT + .248\, GPA + .095 \log(libvol) + .038 \log(cost) - .0033\, rank$$
$$n = 136, R^2 = .842.$$

What is the predicted ceteris paribus difference in salary for schools with a median GPA different by one point? (Report your answer as a percentage.)

(iv) Interpret the coefficient on the variable log(*libvol*).

(v) Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

3.5 In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

(i) In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

does it make sense to hold *sleep*, *work*, and *leisure* fixed, while changing *study*?

(ii) Explain why this model violates Assumption MLR.3.

(iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?

3.6 Consider the multiple regression model containing three independent variables, under Assumptions MLR.1 through MLR.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You are interested in estimating the sum of the parameters on x_1 and x_2; call this $\theta_1 = \beta_1 + \beta_2$.

(i) Show that $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ is an unbiased estimator of θ_1.

(ii) Find $\text{Var}(\hat{\theta}_1)$ in terms of $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$, and $\text{Corr}(\hat{\beta}_1, \hat{\beta}_2)$.

3.7 Which of the following can cause OLS estimators to be biased?

(i) Heteroskedasticity.

(ii) Omitting an important variable.

(iii) A sample correlation coefficient of .95 between two independent variables both included in the model.

3.8 Suppose that average worker productivity at manufacturing firms (*avgprod*) depends on two factors, average hours of training (*avgtrain*) and average worker ability (*avgabil*):

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u.$$

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that *avgtrain* and *avgabil* are negatively correlated, what is the likely bias in $\tilde{\beta}_1$ obtained from the simple regression of *avgprod* on *avgtrain*?

3.9 The following equation describes the median housing price in a community in terms of amount of pollution (*nox* for nitrous oxide) and the average number of rooms in houses in the community (*rooms*):

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u.$$

(i) What are the probable signs of β_1 and β_2? What is the interpretation of β_1? Explain.

(ii) Why might *nox* [or more precisely, log(*nox*)] and *rooms* be negatively correlated? If this is the case, does the simple regression of log(*price*) on log(*nox*) produce an upward or a downward biased estimator of β_1?

(iii) Using the data in HPRICE2.RAW, the following equations were estimated:

$$\widehat{\log(price)} = 11.71 - 1.043 \log(nox), n = 506, R^2 = .264.$$

$$\widehat{\log(price)} = 9.23 - .718 \log(nox) + .306 rooms, n = 506, R^2 = .514.$$

Is the relationship between the simple and multiple regression estimates of the elasticity of *price* with respect to *nox* what you would have predicted, given your answer in part (ii)? Does this mean that $-.718$ is definitely closer to the true elasticity than -1.043?

3.10 Suppose that you are interested in estimating the ceteris paribus relationship between y and x_1. For this purpose, you can collect data on two control variables, x_2 and x_3. (For concreteness, you might think of y as final exam score, x_1 as class attendance, x_2 as GPA up through the previous semester, and x_3 as SAT or ACT score.) Let $\tilde{\beta}_1$ be the simple regression estimate from y on x_1 and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, x_2, x_3.

(i) If x_1 is highly correlated with x_2 and x_3 in the sample, and x_2 and x_3 have large partial effects on y, would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or very different? Explain.

(ii) If x_1 is almost uncorrelated with x_2 and x_3, but x_2 and x_3 are highly correlated, will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or very different? Explain.

(iii) If x_1 is highly correlated with x_2 and x_3, and x_2 and x_3 have small partial effects on y, would you expect $\text{se}(\tilde{\beta}_1)$ or $\text{se}(\hat{\beta}_1)$ to be smaller? Explain.

(iv) If x_1 is almost uncorrelated with x_2 and x_3, x_2 and x_3 have large partial effects on y, and x_2 and x_3 are highly correlated, would you expect $\text{se}(\tilde{\beta}_1)$ or $\text{se}(\hat{\beta}_1)$ to be smaller? Explain.

3.11 Suppose that the population model determining y is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

and this model satisifies Assumptions MLR.1 through MLR.4. However, we estimate the model that omits x_3. Let $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\beta}_2$ be the OLS estimators from the regression of y on x_1 and x_2. Show that the expected value of $\tilde{\beta}_1$ (given the values of the independent variables in the sample) is

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

where the $\hat{r}_{i1}$ are the OLS residuals from the regression of x_1 on x_2. [*Hint:* The formula for $\tilde{\beta}_1$ comes from equation (3.22). Plug $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$ into this equation. After some algebra, take the expectation treating x_{i3} and $\hat{r}_{i1}$ as nonrandom.]

3.12 The following equation represents the effects of tax revenue mix on subsequent employment growth for the population of counties in the United States:

$$growth = \beta_0 + \beta_1 share_{\mathrm{P}} + \beta_2 share_{\mathrm{I}} + \beta_3 share_{\mathrm{S}} + other\ factors,$$

where *growth* is the percentage change in employment from 1980 to 1990, $share_{\mathrm{P}}$ is the share of property taxes in total tax revenue, $share_{\mathrm{I}}$ is the share of income tax revenues, and $share_{\mathrm{S}}$ is the share of sales tax revenues. All of these variables are measured in 1980. The omitted share, $share_{\mathrm{F}}$, includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

(i) Why must we omit one of the tax share variables from the equation?

(ii) Give a careful interpretation of β_1.

3.13 (i) Consider the simple regression model $y = \beta_0 + \beta_1 x + u$ under the first four Gauss-Markov assumptions. For some function $g(x)$, for example $g(x) = x^2$ or $g(x) = \log(1 + x^2)$, define $z_i = g(x_i)$. Define a slope estimator as

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

Show that $\tilde{\beta}_1$ is linear and unbiased. Remember, because $\mathrm{E}(u|x) = 0$, you can treat both x_i and z_i as nonrandom in your derivation.

(ii) Add the homoskedasticity assumption, MLR.5. Show that

$$\mathrm{Var}(\tilde{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} (z_i - \bar{z})^2\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right)^2.$$

(iii) Show directly that, under the Gauss-Markov assumptions, $\mathrm{Var}(\hat{\beta}_1) \leq \mathrm{Var}(\tilde{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator. [*Hint:* The Cauchy-Schwartz inequality in Appendix B implies that

and report the equation in the usual form, including the sample size and R-squared. Are the signs of the slope coefficients what you expected? Explain.

(ii) What do you make of the intercept you estimated in part (i)? In particular, does it make sense to set the two explanatory variables to zero? [*Hint*: Recall that log(1)=0.]

(iii) Now run the simple regression of *math10* on log(*expend*), and compare the slope coefficient with the estimate obtained in part (i). Is the estimated spending effect now larger or smaller than in part (i)?

(iv) Find the correlation between *lexpend* = log(*expend*) and *lnchprg*. Does its sign make sense to you?

(v) Use part (iv) to explain your findings in part (iii).

C3.8 Use the data in DISCRIM.RAW to answer this question. These are zip code–level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

(i) Find the average values of *prpblck* and *income* in the sample, along with their standard deviations. What are the units of measurement of *prpblck* and *income*?

(ii) Consider a model to explain the price of soda, *psoda*, in terms of the proportion of the population that is black and median income:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u.$$

Estimate this model by OLS and report the results in equation form, including the sample size and R-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on *prpblck*. Do you think it is economically large?

(iii) Compare the estimate from part (ii) with the simple regression estimate from *psoda* on *prpblck*. Is the discrimination effect larger or smaller when you control for income?

(iv) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 income + u.$$

If *prpblck* increases by .20 (20 percentage points), what is the estimated percentage change in *psoda*? (*Hint*: The answer is 2.*xx*, where you fill in the "*xx*.")

(v) Now add the variable *prppov* to the regression in part (iv). What happens to $\hat{\beta}_{prpblck}$?

(vi) Find the correlation between log(*income*) and *prppov*. Is it roughly what you expected?

(vii) Evaluate the following statement: "Because log(*income*) and *prppov* are so highly correlated, they have no business being in the same regression."

APPENDIX 3A

3A.1 Derivation of the First Order Conditions in Equation (3.13)

The analysis is very similar to the simple regression case. We must characterize the solutions to the problem

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik})^2.$$

Taking the partial derivatives with respect to each of the b_j (see Appendix A), evaluating them at the solutions, and setting them equal to zero gives

$$-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$-2 \sum_{i=1}^{n} x_{ij}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0, \quad \text{for all } j = 1, \ldots, k.$$

Canceling the -2 gives the first order conditions in (3.13).

3A.2 Derivation of Equation (3.22)

To derive (3.22), write x_{i1} in terms of its fitted value and its residual from the regression of x_1 on $x_2, \ldots, x_k$: $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$, for all $i = 1, \ldots, n$. Now, plug this into the second equation in (3.13):

$$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0. \tag{3.60}$$

By the definition of the OLS residual $\hat{u}_i$, since $\hat{x}_{i1}$ is just a linear function of the explanatory variables $x_{i2}, \ldots, x_{ik}$, it follows that $\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = 0$. Therefore, equation (3.60) can be expressed as

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0. \tag{3.61}$$

Since the $\hat{r}_{i1}$ are the residuals from regressing x_1 on $x_2, \ldots, x_k$, $\sum_{i=1}^{n} x_{ij}\hat{r}_{i1} = 0$, for all $j = 2, \ldots, k$. Therefore, (3.61) is equivalent to $\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 x_{i1}) = 0$. Finally, we use the fact that $\sum_{i=1}^{n} \hat{x}_{i1}\hat{r}_{i1} = 0$, which means that $\hat{\beta}_1$ solves

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0.$$

Now, straightforward algebra gives (3.22), provided, of course, that $\sum_{i=1}^{n} \hat{r}_{i1}^2 > 0$; this is ensured by Assumption MLR.3.

3A.3 Proof of Theorem 3.1

We prove Theorem 3.1 for $\hat{\beta}_1$; the proof for the other slope parameters is virtually identical. (See Appendix E for a more succinct proof using matrices.) Under Assumption MLR.3, the OLS estimators exist, and we can write $\hat{\beta}_1$ as in (3.22). Under Assumption MLR.1, we can write y_i as in (3.32); substitute this for y_i in (3.22). Then, using $\sum_{i=1}^n \hat{r}_{i1} = 0$, $\sum_{i=1}^n x_{ij}\hat{r}_{i1} = 0$, for all $j = 2, \ldots, k$, and $\sum_{i=1}^n x_{i1}\hat{r}_{i1} = \sum_{i=1}^n \hat{r}_{i1}^2$, we have

$$\hat{\beta}_1 = \beta_1 + \left(\sum_{i=1}^n \hat{r}_{i1}u_i\right)\Big/\left(\sum_{i=1}^n \hat{r}_{i1}^2\right). \tag{3.62}$$

Now, under Assumptions MLR.2 and MLR.4, the expected value of each u_i, given all independent variables in the sample, is zero. Since the $\hat{r}_{i1}$ are just functions of the sample independent variables, it follows that

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1|\mathbf{X}) &= \beta_1 + \left(\sum_{i=1}^n \hat{r}_{i1}\mathrm{E}(u_i|\mathbf{X})\right)\Big/\left(\sum_{i=1}^n \hat{r}_{i1}^2\right) \\ &= \beta_1 + \left(\sum_{i=1}^n \hat{r}_{i1}\cdot 0\right)\Big/\left(\sum_{i=1}^n \hat{r}_{i1}^2\right) = \beta_1, \end{aligned}$$

where $\mathbf{X}$ denotes the data on all independent variables and $\mathrm{E}(\hat{\beta}_1|\mathbf{X})$ is the expected value of $\hat{\beta}_1$, given $x_{i1}, \ldots, x_{ik}$, for all $i = 1, \ldots, n$. This completes the proof.

3A.4 General Omitted Variable Bias

We can derive the omitted variable bias in the general model in equation (3.31) under the first four Gauss-Markov assumptions. In particular, let the $\hat{\beta}_j, j = 0, 1, \ldots, k$ be the OLS estimators from the regression using the full set of explanatory variables. Let the $\tilde{\beta}_j, j = 0,1, \ldots, k-1$ be the OLS estimators from the regression that leaves out x_k. Let $\tilde{\delta}_j, j = 1, \ldots, k-1$ be the slope coefficient on x_j in the auxiliary regression of x_{ik} on $x_{i1}, x_{i2}, \ldots x_{i,k-1}, i = 1, \ldots, n$. A useful fact is that

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k\tilde{\delta}_j. \tag{3.63}$$

This shows explicitly that, when we do not control for x_k in the regression, the estimated partial effect of x_j equals the partial effect when we include x_k plus the partial effect of x_k on $\hat{y}$ times the partial relationship between the omitted variable, x_k, and $x_j, j < k$. Conditional on the entire set of explanatory variables, $\mathbf{X}$, we know that the $\hat{\beta}_j$ are all unbiased for the corresponding β_j, $j = 1, \ldots, k$. Further, since $\tilde{\delta}_j$ is just a function of $\mathbf{X}$, we have

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_j|\mathbf{X}) &= \mathrm{E}(\hat{\beta}_j|\mathbf{X}) + \mathrm{E}(\hat{\beta}_k|\mathbf{X})\tilde{\delta}_j \\ &= \beta_j + \beta_k\tilde{\delta}_j. \end{aligned} \tag{3.64}$$

Equation (3.64) shows that $\tilde{\beta}_j$ is biased for β_j unless $\beta_k = 0$—in which case x_k has no partial effect in the population—or $\tilde{\delta}_j$ equals zero, which means that x_{ik} and x_{ij} are

partially uncorrelated in the sample. The key to obtaining equation (3.64) is equation (3.63). To show equation (3.63), we can use equation (3.22) a couple of times. For simplicity, we look at $j = 1$. Now, $\tilde{\beta}_1$ is the slope coefficient in the simple regression of y_i on $\tilde{r}_{i1}$, $i = 1, \ldots, n$, where the $\tilde{r}_{i1}$ are the OLS residuals from the regression of x_{i1} on x_{i2}, $x_{i3}, \ldots, x_{i,k-1}$. Consider the numerator of the expression for $\tilde{\beta}_1$: $\sum_{i=1}^{n} \tilde{r}_{i1} y_i$. But for each i, we can write $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$ and plug in for y_i. Now, by properties of the OLS residuals, the $\tilde{r}_{i1}$ have zero sample average and are uncorrelated with $x_{i2}, x_{i3}, \ldots, x_{i,k-1}$ in the sample. Similarly, the $\hat{u}_i$ have zero sample average and zero sample correlation with $x_{i1}, x_{i2}, \ldots, x_{ik}$. It follows that the $\tilde{r}_{i1}$ and $\hat{u}_i$ are uncorrelated in the sample (since the $\tilde{r}_{i1}$ are just linear combinations of $x_{i1}, x_{i2}, \ldots, x_{i,k-1}$). So

$$\sum_{i=1}^{n} \tilde{r}_{i1} y_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} \right) + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right). \tag{3.65}$$

Now, $\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} = \sum_{i=1}^{n} \tilde{r}_{i1}^2$, which is also the denominator of $\tilde{\beta}_1$. Therefore, we have shown that

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right) \Big/ \left(\sum_{i=1}^{n} \tilde{r}_{i1}^2, \right) \\ &= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1. \end{aligned}$$

This is the relationship we wanted to show.

3A.5 Proof of Theorem 3.2

Again, we prove this for $j = 1$. Write $\hat{\beta}_1$ as in equation (3.62). Now, under MLR.5, $\text{Var}(u_i|\mathbf{X}) = \sigma^2$, for all $i = 1, \ldots, n$. Under random sampling, the u_i are independent, even conditional on $\mathbf{X}$, and the $\hat{r}_{i1}$ are nonrandom conditional on $\mathbf{X}$. Therefore,

$$\begin{aligned} \text{Var}(\hat{\beta}_1|\mathbf{X}) &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \, \text{Var}(u_i|\mathbf{X}) \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2 \\ &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \sigma^2 \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2 = \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \end{aligned}$$

Now, since $\sum_{i=1}^{n} \hat{r}_{i1}^2$ is the sum of squared residuals from regressing x_1 on $x_2, \ldots, x_k$, $\sum_{i=1}^{n} \hat{r}_{i1}^2 = \text{SST}_1(1 - R_1^2)$. This completes the proof.

3A.6 Proof of Theorem 3.4

We show that, for any other linear unbiased estimator $\tilde{\beta}_1$ of β_1, $\text{Var}(\tilde{\beta}_1) \geq \text{Var}(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator. The focus on $j = 1$ is without loss of generality.

For $\tilde{\beta}_1$ as in equation (3.59), we can plug in for y_i to obtain

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} u_i.$$

Now, since the w_{i1} are functions of the x_{ij},

$$\mathrm{E}(\tilde{\beta}_1|\mathbf{X}) = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1}x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1}x_{ik} + \sum_{i=1}^{n} w_{i1}\mathrm{E}(u_i|\mathbf{X})$$

$$= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1}x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1}x_{ik}$$

because $\mathrm{E}(u_i|\mathbf{X}) = 0$, for all $i = 1, \ldots, n$ under MLR.2 and MLR.4. Therefore, for $\mathrm{E}(\tilde{\beta}_1|\mathbf{X})$ to equal β_1 for any values of the parameters, we must have

$$\sum_{i=1}^{n} w_{i1} = 0, \quad \sum_{i=1}^{n} w_{i1}x_{i1} = 1, \quad \sum_{i=1}^{n} w_{i1}x_{ij} = 0, \quad j = 2, \ldots, k. \tag{3.66}$$

Now, let $\hat{r}_{i1}$ be the residuals from the regression of x_{i1} on $x_{i2}, \ldots, x_{ik}$. Then, from (3.66), it follows that

$$\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} = 1 \tag{3.67}$$

because $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ and $\sum_{i=1}^{n} w_{i1}\hat{x}_{i1} = 0$. Now, consider the difference between $\mathrm{Var}(\tilde{\beta}_1|\mathbf{X})$ and $\mathrm{Var}(\hat{\beta}_1|\mathbf{X})$ under MLR.1 through MLR.5:

$$\sigma^2 \sum_{i=1}^{n} w_{i1}^2 - \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \tag{3.68}$$

Because of (3.67), we can write the difference in (3.68), without σ^2, as

$$\sum_{i=1}^{n} w_{i1}^2 - \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} \right)^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \tag{3.69}$$

But (3.69) is simply

$$\sum_{i=1}^{n} (w_{i1} - \hat{\gamma}_1\hat{r}_{i1})^2, \tag{3.70}$$

where $\hat{\gamma}_1 = \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)$, as can be seen by squaring each term in (3.70), summing, and then canceling terms. Because (3.70) is just the sum of squared residuals from the simple regression of w_{i1} on $\hat{r}_{i1}$—remember that the sample average of $\hat{r}_{i1}$ is zero—(3.70) must be nonnegative. This completes the proof.