2.5 Expected Values and Variances of the OLS Estimators

In Section 2.1, we defined the population model $y = \beta_0 + \beta_1 x + u$, and we claimed that the key assumption for simple regression analysis to be useful is that the expected value of u given any value of x is zero. In Sections 2.2, 2.3, and 2.4, we discussed the algebraic properties of OLS estimation. We now return to the population model and study the *statistical* properties of OLS. In other words, we now view $\hat{\beta}_0$ and $\hat{\beta}_1$ as *estimators* for the parameters β_0 and β_1 that appear in the population model. This means that we will study properties of the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ over different random samples from the population. (Appendix C contains definitions of estimators and reviews some of their important properties.)

Unbiasedness of OLS

We begin by establishing the unbiasedness of OLS under a simple set of assumptions. For future reference, it is useful to number these assumptions using the prefix "SLR" for simple linear regression. The first assumption defines the population model.

Assumption SLR.1 Linear in Parameters

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

$$y = \beta_0 + \beta_1 x + u, \qquad [2.47]$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

To be realistic, y, x, and u are all viewed as random variables in stating the population model. We discussed the interpretation of this model at some length in Section 2.1 and gave several examples. In the previous section, we learned that equation (2.47) is not as restrictive as it initially seems; by choosing y and x appropriately, we can obtain interesting nonlinear relationships (such as constant elasticity models).

We are interested in using data on y and x to estimate the parameters β_0 and, especially, β_1. We assume that our data were obtained as a random sample. (See Appendix C for a review of random sampling.)

Assumption SLR.2 Random Sampling

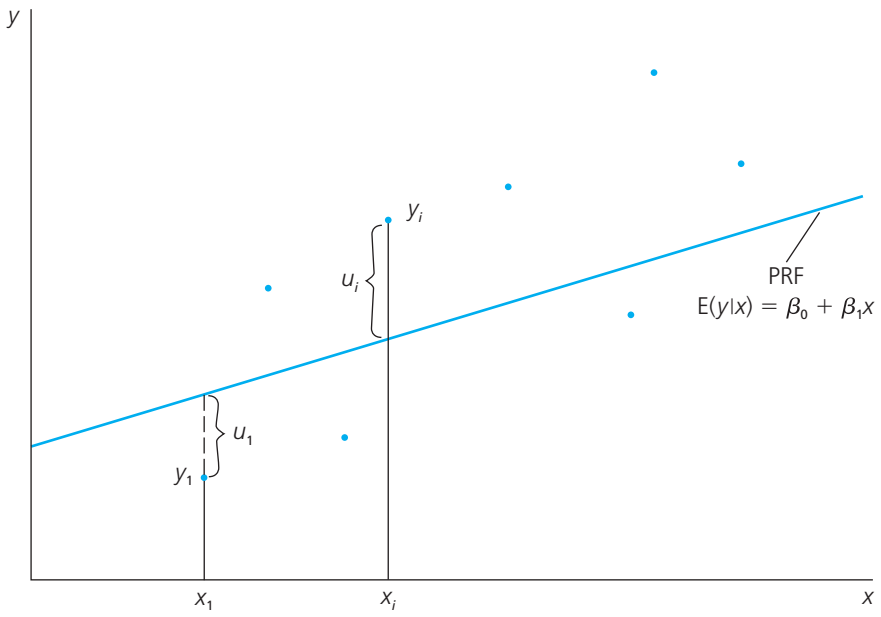
We have a random sample of size n, $\{(x_i, y_i): i = 1, 2, \ldots, n\}$, following the population model in equation (2.47).

We will have to address failure of the random sampling assumption in later chapters that deal with time series analysis and sample selection problems. Not all cross-sectional samples can be viewed as outcomes of random samples, but many can be.

We can write (2.47) in terms of the random sample as

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \ldots, n, \qquad [2.48]$$

FIGURE 2.7 Graph of $y_i = \beta_0 + \beta_1 x_i + u_i$.



where u_i is the error or disturbance for observation i (for example, person i, firm i, city i, and so on). Thus, u_i contains the unobservables for observation i that affect y_i. The u_i should not be confused with the residuals, $\hat{u}_i$, that we defined in Section 2.3. Later on, we will explore the relationship between the errors and the residuals. For interpreting β_0 and β_1 in a particular application, (2.47) is most informative, but (2.48) is also needed for some of the statistical derivations.

The relationship (2.48) can be plotted for a particular outcome of data as shown in Figure 2.7.

As we already saw in Section 2.2, the OLS slope and intercept estimates are not defined unless we have some sample variation in the explanatory variable. We now add variation in the x_i to our list of assumptions.

Assumption SLR.3 **Sample Variation in the Explanatory Variable**

The sample outcomes on x, namely, $\{x_i, i = 1, \ldots, n\}$, are not all the same value.

This is a very weak assumption—certainly not worth emphasizing, but needed nevertheless. If x varies in the population, random samples on x will typically contain variation, unless the population variation is minimal or the sample size is small. Simple inspection of summary statistics on x_i reveals whether Assumption SLR.3 fails: if the sample standard deviation of x_i is zero, then Assumption SLR.3 fails; otherwise, it holds.

Finally, in order to obtain unbiased estimators of β_0 and β_1, we need to impose the zero conditional mean assumption that we discussed in some detail in Section 2.1. We now explicitly add it to our list of assumptions.

Assumption SLR.4 **Zero Conditional Mean**

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$E(u|x) = 0.$$

For a random sample, this assumption implies that $E(u_i|x_i) = 0$, for all $i = 1, 2, \ldots, n$.

In addition to restricting the relationship between u and x in the population, the zero conditional mean assumption—coupled with the random sampling assumption—allows for a convenient technical simplification. In particular, we can derive the statistical properties of the OLS estimators as *conditional* on the values of the x_i in our sample. Technically, in statistical derivations, conditioning on the sample values of the independent variable is the same as treating the x_i as *fixed in repeated samples*, which we think of as follows. We first choose n sample values for $x_1, x_2, \ldots, x_n$. (These can be repeated.) Given these values, we then obtain a sample on y (effectively by obtaining a random sample of the u_i). Next, another sample of y is obtained, using the *same* values for $x_1, x_2, \ldots, x_n$. Then another sample of y is obtained, again using the same $x_1, x_2, \ldots, x_n$. And so on.

The fixed-in-repeated-samples scenario is not very realistic in nonexperimental contexts. For instance, in sampling individuals for the wage-education example, it makes little sense to think of choosing the values of *educ* ahead of time and then sampling individuals with those particular levels of education. Random sampling, where individuals are chosen randomly and their wage and education are both recorded, is representative of how most data sets are obtained for empirical analysis in the social sciences. Once we *assume* that $E(u|x) = 0$, and we have random sampling, nothing is lost in derivations by treating the x_i as nonrandom. The danger is that the fixed-in-repeated-samples assumption *always* implies that u_i and x_i are independent. In deciding when simple regression analysis is going to produce unbiased estimators, it is critical to think in terms of Assumption SLR.4.

Now, we are ready to show that the OLS estimators are unbiased. To this end, we use the fact that $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$ (see Appendix A) to write the OLS slope estimator in equation (2.19) as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad [2.49]$$

Because we are now interested in the behavior of $\hat{\beta}_1$ across all possible samples, $\hat{\beta}_1$ is properly viewed as a random variable.

We can write $\hat{\beta}_1$ in terms of the population coefficients and errors by substituting the right-hand side of (2.48) into (2.49). We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\text{SST}_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\text{SST}_x}, \qquad [2.50]$$

where we have defined the total variation in x_i as $\text{SST}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ to simplify the notation. (This is not quite the sample variance of the x_i because we do not divide by $n - 1$.) Using the algebra of the summation operator, write the numerator of $\hat{\beta}_1$ as

$$\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i$$
$$= \beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i. \qquad [2.51]$$

As shown in Appendix A, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{SST}_x$. Therefore, we can write the numerator of $\hat{\beta}_1$ as $\beta_1\text{SST}_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i$. Putting this over the denominator gives

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\text{SST}_x} = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i u_i, \qquad [2.52]$$

where $d_i = x_i - \bar{x}$. We now see that the estimator $\hat{\beta}_1$ equals the population slope, β_1, plus a term that is a linear combination in the errors $\{u_1, u_2, \ldots, u_n\}$. Conditional on the values of x_i, the randomness in $\hat{\beta}_1$ is due entirely to the errors in the sample. The fact that these errors are generally different from zero is what causes $\hat{\beta}_1$ to differ from β_1.

Using the representation in (2.52), we can prove the first important statistical property of OLS.

THEOREM 2.1

UNBIASEDNESS OF OLS:

Using Assumptions SLR.1 through SLR.4,

$$\text{E}(\hat{\beta}_0) = \beta_0, \text{ and } \text{E}(\hat{\beta}_1) = \beta_1, \qquad [2.53]$$

for any values of β_0 and β_1. In other words, $\hat{\beta}_0$ is unbiased for β_0, and $\hat{\beta}_1$ is unbiased for β_1.

PROOF: In this proof, the expected values are conditional on the sample values of the independent variable. Because SST_x and d_i are functions only of the x_i, they are nonrandom in the conditioning. Therefore, from (2.52), and keeping the conditioning on $\{x_1, x_2, \ldots, x_n\}$ implicit, we have

$$\text{E}(\hat{\beta}_1) = \beta_1 + \text{E}\left[(1/\text{SST}_x)\sum_{i=1}^{n} d_i u_i\right] = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n}\text{E}(d_i u_i)$$
$$= \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i\text{E}(u_i) = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i \cdot 0 = \beta_1,$$

where we have used the fact that the expected value of each u_i (conditional on $\{x_1, x_2, \ldots, x_n\}$) is zero under Assumptions SLR.2 and SLR.4. Since unbiasedness holds for any outcome on $\{x_1, x_2, \ldots, x_n\}$, unbiasedness also holds without conditioning on $\{x_1, x_2, \ldots, x_n\}$.

The proof for $\hat{\beta}_0$ is now straightforward. Average (2.48) across i to get $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$, and plug this into the formula for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}.$$

Then, conditional on the values of the x_i,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)\bar{x}] + \mathrm{E}(\bar{u}) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)]\bar{x},$$

since $\mathrm{E}(\bar{u}) = 0$ by Assumptions SLR.2 and SLR.4. But, we showed that $\mathrm{E}(\hat{\beta}_1) = \beta_1$, which implies that $\mathrm{E}[(\hat{\beta}_1 - \beta_1)] = 0$. Thus, $\mathrm{E}(\hat{\beta}_0) = \beta_0$. Both of these arguments are valid for any values of β_0 and β_1, and so we have established unbiasedness.

Remember that unbiasedness is a feature of the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$, which says nothing about the estimate that we obtain for a given sample. We hope that, if the sample we obtain is somehow "typical," then our estimate should be "near" the population value. Unfortunately, it is always possible that we could obtain an unlucky sample that would give us a point estimate far from β_1, and we can *never* know for sure whether this is the case. You may want to review the material on unbiased estimators in Appendix C, especially the simulation exercise in Table C.1 that illustrates the concept of unbiasedness.

Unbiasedness generally fails if any of our four assumptions fail. This means that it is important to think about the veracity of each assumption for a particular application. Assumption SLR.1 requires that y and x be linearly related, with an additive disturbance. This can certainly fail. But we also know that y and x can be chosen to yield interesting nonlinear relationships. Dealing with the failure of (2.47) requires more advanced methods that are beyond the scope of this text.

Later, we will have to relax Assumption SLR.2, the random sampling assumption, for time series analysis. But what about using it for cross-sectional analysis? Random sampling can fail in a cross section when samples are not representative of the underlying population; in fact, some data sets are constructed by intentionally oversampling different parts of the population. We will discuss problems of nonrandom sampling in Chapters 9 and 17.

As we have already discussed, Assumption SLR.3 almost always holds in interesting regression applications. Without it, we cannot even obtain the OLS estimates.

The assumption we should concentrate on for now is SLR.4. If SLR.4 holds, the OLS estimators are unbiased. Likewise, if SLR.4 fails, the OLS estimators generally will be *biased*. There are ways to determine the likely direction and size of the bias, which we will study in Chapter 3.

The possibility that x is correlated with u is almost always a concern in simple regression analysis with nonexperimental data, as we indicated with several examples in Section 2.1. Using simple regression when u contains factors affecting y that are also

correlated with x can result in *spurious correlation*: that is, we find a relationship between y and x that is really due to other unobserved factors that affect y and also happen to be correlated with x.

EXAMPLE 2.12 **STUDENT MATH PERFORMANCE AND THE SCHOOL LUNCH PROGRAM**

Let *math10* denote the percentage of tenth graders at a high school receiving a passing score on a standardized mathematics exam. Suppose we wish to estimate the effect of the federally funded school lunch program on student performance. If anything, we expect the lunch program to have a positive ceteris paribus effect on performance: all other factors being equal, if a student who is too poor to eat regular meals becomes eligible for the school lunch program, his or her performance should improve. Let *lnchprg* denote the percentage of students who are eligible for the lunch program. Then, a simple regression model is

$$math10 = \beta_0 + \beta_1 lnchprg + u, \qquad [2.54]$$

where u contains school and student characteristics that affect overall school performance. Using the data in MEAP93.RAW on 408 Michigan high schools for the 1992–1993 school year, we obtain

$$\widehat{math10} = 32.14 - 0.319\ lnchprg$$
$$n = 408, R^2 = 0.171.$$

This equation predicts that if student eligibility in the lunch program increases by 10 percentage points, the percentage of students passing the math exam *falls* by about 3.2 percentage points. Do we really believe that higher participation in the lunch program actually *causes* worse performance? Almost certainly not. A better explanation is that the error term u in equation (2.54) is correlated with *lnchprg*. In fact, u contains factors such as the poverty rate of children attending school, which affects student performance and is highly correlated with eligibility in the lunch program. Variables such as school quality and resources are also contained in u, and these are likely correlated with *lnchprg*. It is important to remember that the estimate -0.319 is only for this particular sample, but its sign and magnitude make us suspect that u and x are correlated, so that simple regression is biased.

In addition to omitted variables, there are other reasons for x to be correlated with u in the simple regression model. Because the same issues arise in multiple regression analysis, we will postpone a systematic treatment of the problem until then.

Variances of the OLS Estimators

In addition to knowing that the sampling distribution of $\hat{\beta}_1$ is centered about β_1 ($\hat{\beta}_1$ is unbiased), it is important to know how far we can expect $\hat{\beta}_1$ to be away from β_1 on average.