Unbiasedness of OLS in the Linear Regression Model

Aaron Smith

April 28, 2020

I wrote this document using R Markdown. Click here to download the markdown file including R code.

In the texbook (pg 73), we made a mathematical error when proving unbiasedness of OLS. We specify a population model in which the X variables are uncorrelated with the errors. However, to prove unbiasedness in general, we need an additional condition. Here, we provide the technical details.

The two main ways to prove unbiasedness are to add an assumption that the model is correctly specified or to use a large sample. The former implies exact unbiasedness, and the latter implies approximate unbiasedness. However, there are many examples in which the model is incorrectly specified and the sample is small, yet OLS is unbiased (see Example B). There are other examples in which OLS is biased (see Example C).

The population regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, cov[X_i, \varepsilon_i] = 0$$

As we describe in Ch 4 of the book, the condition that $cov[X_i, \varepsilon_i] = 0$ means that the model is the best linear prediction of Y using X. The model leaves no information about X in the error because it uses all the information in X to predict Y. Sometimes, an econometrician's main goal is not to get the best prediction of Y given X. We study that case in Chapter 11.

The sample model is:

$$Y_i = b_0 + b_1 X_i + e_i$$

and the OLS estimate of β_1 is given by:

$$b_1 = \frac{\sum_{i=1}^N x_i Y_i}{\sum_{i=1}^N x_i^2},$$

where lower-case notation indicates that a variable has been demeaned, i.e., $x_i = X_i - \bar{X}$.

Proving Unbiasedness

Plugging the population regression equation into the OLS formula gives:

$$b_1 = \beta_1 + \frac{\sum_{i=1}^N x_i \varepsilon_i}{\sum_{i=1}^N x_i^2} = \beta_1 + \frac{N^{-1}\sum_{i=1}^N x_i \varepsilon_i}{N^{-1}\sum_{i=1}^N x_i^2}$$

In the textbook, we define $w_i = \frac{x_i}{\sum_{i=1}^N x_i^2}$ and write $b_1 = \beta_1 + \sum_{i=1}^N w_i \varepsilon_i$, but I don't use that shorthand here.

We make assumption CR1, that the sample is representative of the population.

Now, take expectations of b_1:

$$E[b_1] = \beta_1 + E\left[\frac{N^{-1}\sum_{i=1}^N x_i \varepsilon_i}{N^{-1}\sum_{i=1}^N x_i^2}\right]$$

The difficulty is that

$$E\left[\frac{N^{-1}\sum_{i=1}^N x_i \varepsilon_i}{N^{-1}\sum_{i=1}^N x_i^2}\right] \neq \frac{E\left[N^{-1}\sum_{i=1}^N x_i \varepsilon_i\right]}{E\left[N^{-1}\sum_{i=1}^N x_i^2\right]}$$

Thus, even though the mean of the numerator is zero, the mean of the ratio may not be zero. To see this point more clearly, we can write

$$E\left[\frac{N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i}{N^{-1}\sum_{i=1}^{N} x_i^2}\right] = E\left[N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i\right] E\left[\frac{1}{N^{-1}\sum_{i=1}^{N} x_i^2}\right] + cov\left[N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i, \frac{1}{N^{-1}\sum_{i=1}^{N} x_i^2}\right]$$

Because, $E\left[N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i\right] = 0$, we have

$$E[b_1] = \beta_1 + cov\left[N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i, \frac{1}{N^{-1}\sum_{i=1}^{N} x_i^2}\right]$$

Thus, any bias in the OLS estimator comes from non-zero covariance between the numerator and denominator. Getting this covariance to zero requires some additional conditions.

Here are the two main ways to show unbiasedness.

1. Assume Correct Specification for the Mean

Correct specification means that the population relationship between Y and X really is a straight line. Mathematically, it implies the mean independence condition: $E[\varepsilon_i|X_1, X_2, ..., X_N] = 0$. With this condition, you can apply the law of iterated expectations to prove unbiasedness as follows:

$$E[b_1] = \beta_1 + E\left[E\left[\frac{N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i}{N^{-1}\sum_{i=1}^{N} x_i^2}\middle| X_1, X_2, ..., X_N\right]\right] = \beta_1 + E\left[\frac{N^{-1}\sum_{i=1}^{N} x_i E\left[\varepsilon_i|X_1, X_2, ..., X_N\right]}{N^{-1}\sum_{i=1}^{N} x_i^2}\right] = \beta_1$$

Uncorrelatedness of X_i and ε_i means (loosely) that larger values X_i don't imply larger or smaller values of ε_i. The mean independence condition implies that, if I know the values of all of the X's, then the expected value of ε_i is zero and *this is true for all i.* OLS can be unbiased if the mean independence assumption fails (see Example B below), but it is not necessarily so.

2. Use Large Sample Theory

In a large sample, the law of large numbers implies that the denominator in the OLS formula ($N^{-1}\sum_{i=1}^{N} x_i^2$) is essentially a constant equal to the variance of X. In a small sample, the value of $N^{-1}\sum_{i=1}^{N} x_i^2$ may differ from the variance of X due to randomness. But, in large samples the randomness averages out.

If we treat the denominator as a constant, then we don't have any correlation between the numerator and denominator in the OLS formula. Mathematically, we have:

$$E[b_1] = \beta_1 + E\left[\frac{N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i}{N^{-1}\sum_{i=1}^{N} x_i^2}\right] \approx \beta_1 + E\left[\frac{N^{-1}\sum_{i=1}^{N} x_i\varepsilon_i}{var(X_i)}\right] = \beta_1 + \frac{N^{-1}\sum_{i=1}^{N} E\left[x_i\varepsilon_i\right]}{var(X_i)} = \beta_1$$

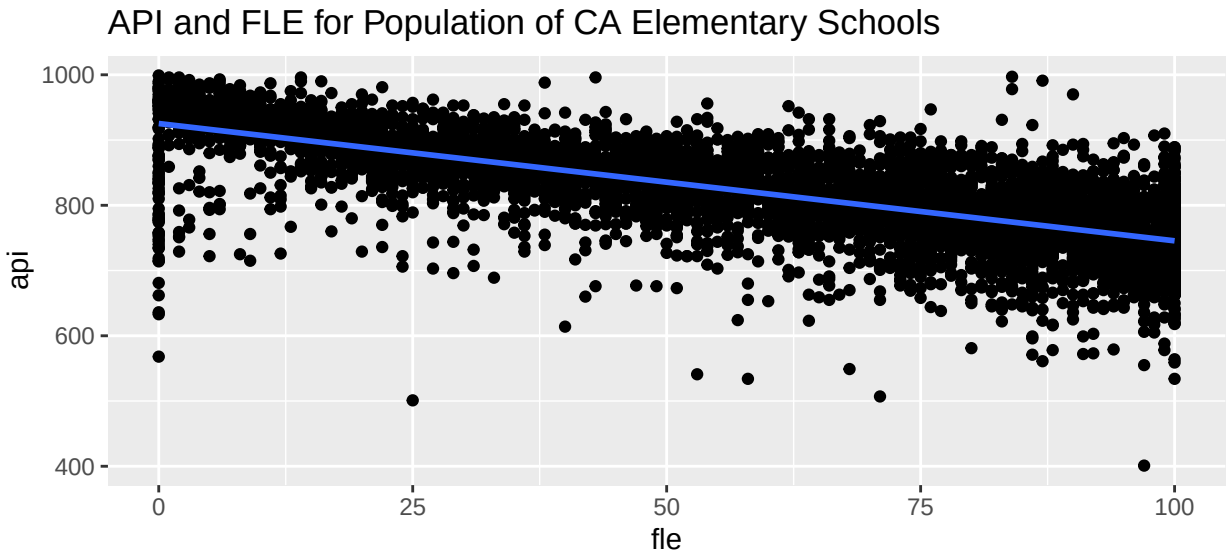
The last equality follows from the zero covariance between X_i and ε_i, which implies $E[x_i\varepsilon_i] = 0$.

Example A: California Schools

This is the example in the textbook.

The population consists of 5,765 elementary schools in 2013. We observe a test score (api) and the propportion of students eligible for free or reduced price lunch (fle) for each school. The scatter plot below shows the data and the population regression line. The slope of the population regression line is $\beta_1 = -1.80$.

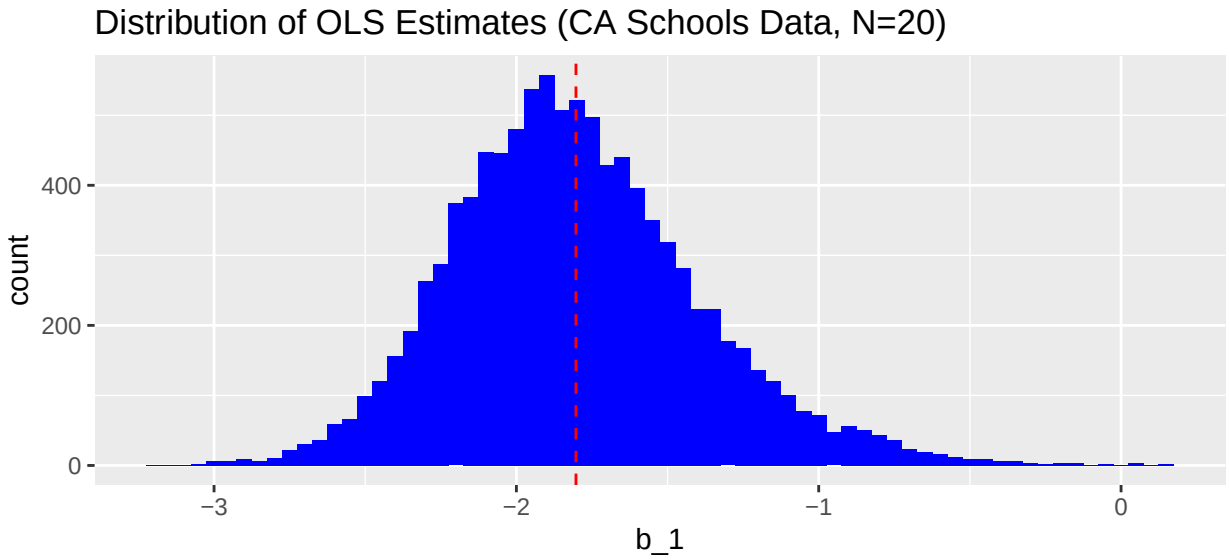
The linear regression looks like a pretty good fit to these data. It does not look (to me) like the true population relationship is a curve.



Next, we take a random sample of 20 schools from the population and run the regression. We don't get exactly -1.80 because we observe only 20 randomly chosen shcools and not all 5,765. We repeat this process 10,000 times, generating 10,000 different samples, running a regression for each one and saving the estimated slope values b_1.

OLS is unbiased if the average across these 10,000 b_1 values equals -1.80.

The histogram below shows the distribution of b_1 values from this exercise. The mean is -1.80.



So, in this example, the linear fit looks to be appropriate and OLS is unbiased even though the sample is small.

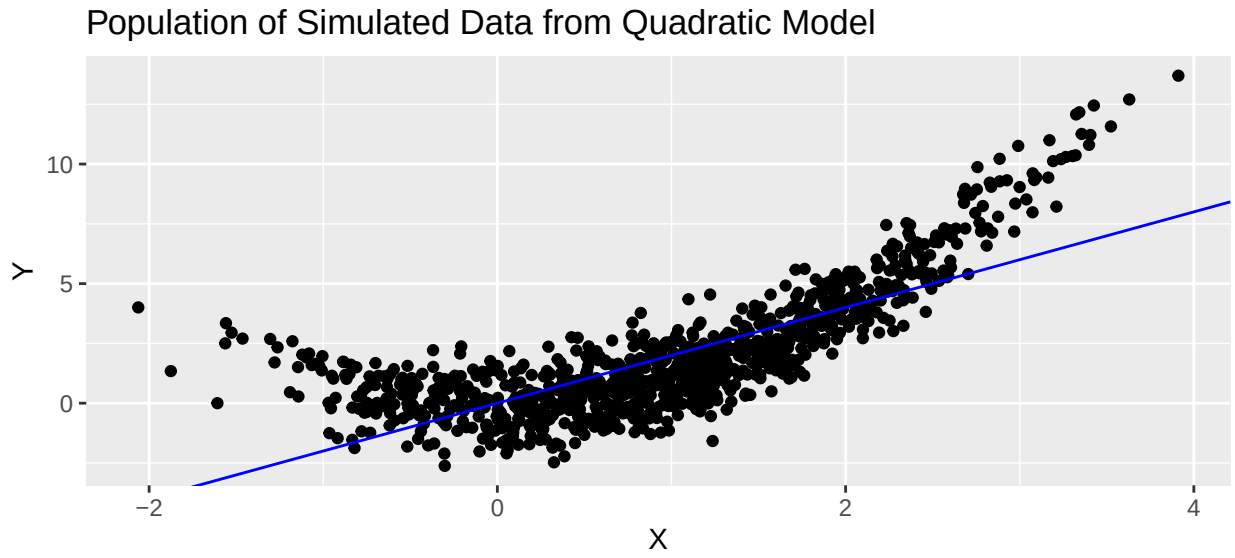
Example B: Misspecified Model

This is a made-up example. Suppose the true relationship between Y and X is quadratic. Specifcally, the data are generated by the equation

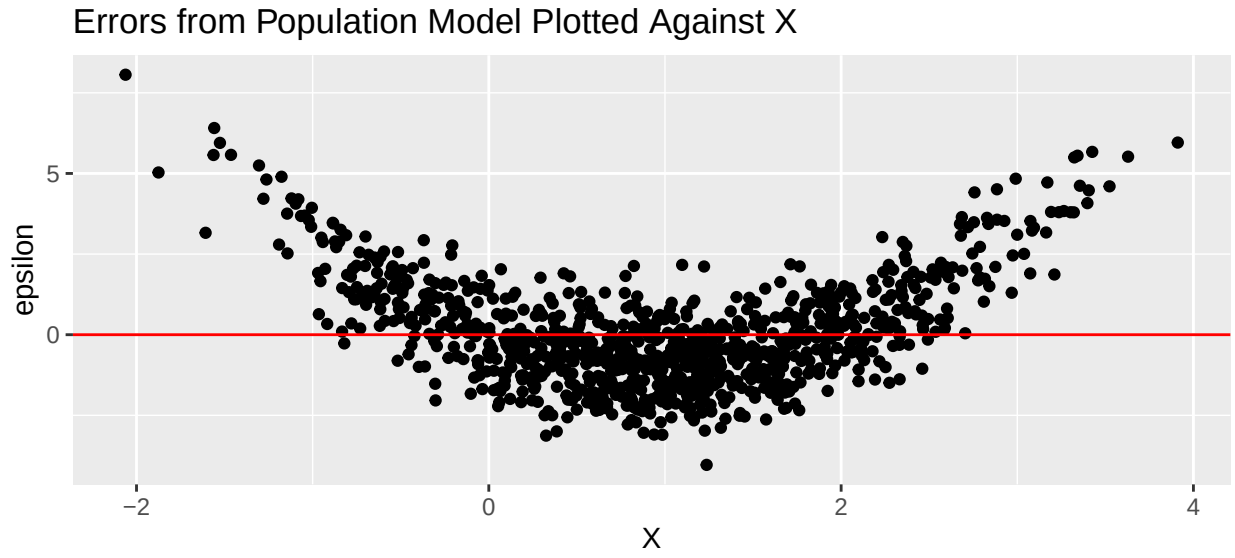
$$Y_i = X_i^2 + u_i,$$

where $u_i \sim iidN(0,1)$ and $X_i \sim iidN(1,1)$. However, you as an econometric analyst don't know this, so you fit a linear regression model. You are fitting a straight line through a curve.

The figure below plots a population of 1000 observations generated from the quadratic equation. The blue line is the population regression line through these 1000 dots. It is obvious that the linear regression is not the right model.



Next, I plot the errors from the regression given by the blue line. The errors are uncorrelated with X because the best-fit straight line is flat (the red line below). However, the mean independence condition fails. If I know $X_i = 1$, then I predict $\varepsilon_i < 0$. If I know $X_i = 3$, then I predict $\varepsilon_i > 0$. So, the expected value of ε_i conditional on X_i is NOT zero for all i.



The slope of the population regression line is 2, i.e., $\beta_1 = 2$. However, the mean independence condition fails because we are fitting a straight line to a curved relationship. This means OLS is biased, right? This means that, if we apply OLS to a linear model using a sample from this population, we'll get an average slope different from 2, right?

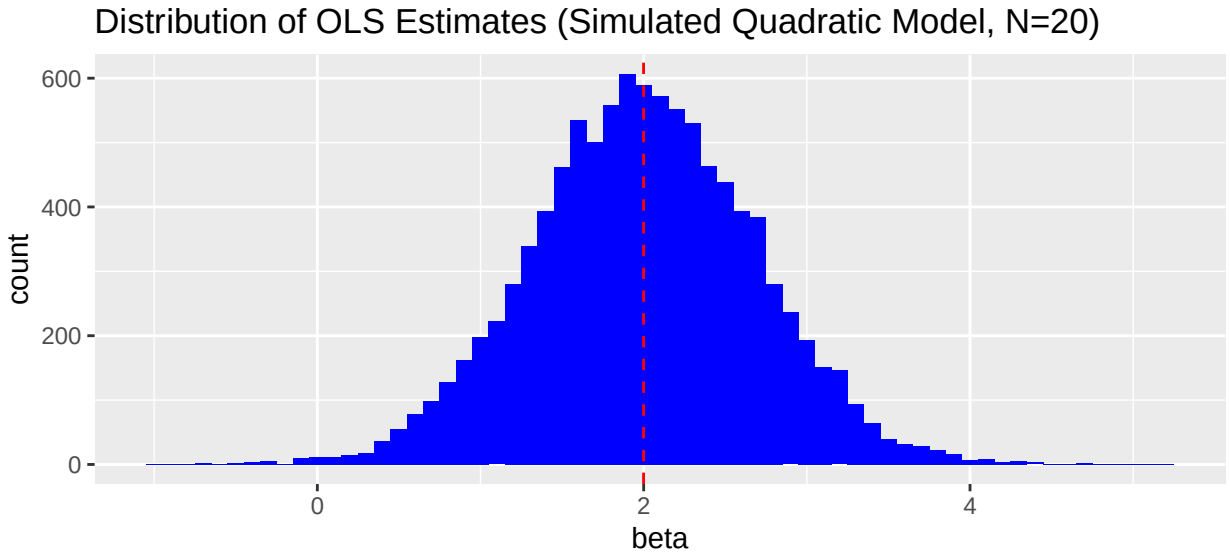
Wrong.

It's hard to prove mathematically that OLS is unbiased, but in this particular case, it does turn out to be so.

To demonstrate, I drew 10,000 random samples of size 20 from this population. For each sample, I fit the regression equation and saved the value of b_1. This is the same as the simulation exercise we do in Chapter 5 of the textbook. The population value is $\beta_1 = 2$, which is the slope of the blue line in the first figure above.

Below is a histogram of the b_1 values. Most of them are between 1 and 3. The mean equals 2, so OLS is unbiased at estimating the slope of the blue line, even though the mean independence assumption fails.

In other cases, OLS may be biased (e.g., if I repeat this experiment except with the X variables drawn from a skewed distribution like Chi square, I get a bias in the OLS estimator when the sample size is small).



Example C: Autoregression

Sometimes in time series data, we use last month's value of the variable to predict this month's value. This produces the following regression equation, known as an autoregression:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t, E[Y_{t-1}\varepsilon_t] = 0$$

I am using t for time rather than i for individual to denote each observation here.

The mean independence assumption does not hold for this model because the same variable appears on both sides of the equation. Mathematically, the expected value of the error at time t conditional on all the right-hand-side observations is

$$E[\varepsilon_t | Y_0, Y_1, ..., Y_{T-1}] = Y_t - \beta_0 - \beta_1 Y_{t-1} \neq 0$$

In this case, OLS is biased, but the bias diminishes as the sample size grows.

To illustrate, I generated data for four different sample sizes, ran the regression and saved the slope estimate b_1. I repeated this process 1000 times for each sample size.

The histogram below shows the results. The black vertical line is the mean of b_1. The red vertical line is the population value $\beta_1 = 0.95$. The bias is quite large for a sample of 25, as the average OLS estimate is less than 0.8. For a sample size of 400, the bias is almost zero and it disappears by the time we get to a sample of 1,600.