Expectation

- What is the expectation of $\hat{\beta}_0$, $\hat{\beta}_1$?
- We want

$$
\begin{aligned}
E(\hat{\beta}_0) &= \beta_0 \\
E\left(\hat{\beta}_1\right) &= \beta_1
\end{aligned}
$$

that is, on average $\hat{\beta}_0$ and $\hat{\beta}_1$ are "correct"! This is called the **unbiasedness** property of the estimators.

Unbiasedness

- The OLS estimators are unbiased under four assumptions.
- This set of assumptions is often referred to as assumptions for the Classical Linear Regression Model

Assumptions for Unbiasedness

First we need to define the basic 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

$$Y = \beta_0 + \beta_1 X + U$$

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

This assumption alone is not restrictive at all.

Assumptions for Unbiasedness

Now we need to assume something about the sample.

- **Assumption (SLR.2-Random Sampling)** We have a random sample of size n, (X_i, Y_i), i $= 1$, ..., n, following the population model defined in SLR.1.
- This now defines the basic environment.

Assumptions for Unbiasedness

Next we need an assumption that allows us to estimate the model.

- **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.
- Without this assumption we would have real trouble. Practically the denominator $\hat{\beta}_1$ is $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. This would be zero if there is no variation in X.

Assumptions for Unbiasedness

- **Assumptions (SLR.4-mean independence)** The error U has an expected value of zero given any value of the explanatory variable. In other words

$$E\left(U|X\right)=0$$

- This and the linearity assumption really work together.

Unbiasedness

- **Theorem**: under Assumptions SLR.1-4, the OLS estimators are unbiased, that is, $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$.

The proof uses the law of iterated expectation: for any random variable W and Z,

$$E\left[W\right] = E\left[E\left[W|Z\right]\right].$$

Unbiasedness

Proof of unbiasedness of $\hat{\beta}_1$:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)\, y_i}{\sum_{i=1}^{n}(x_i - \bar{x}_n) x_i} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)\,(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n}(x_i - \bar{x}_n) x_i} \\
&= 0 + \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)\, u_i}{\sum_{i=1}^{n}(x_i - \bar{x}_n) x_i}
\end{aligned}$$

Unbiasedness

Proof of unbiasedness of $\hat{\beta}_1$ (cont.):

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) u_i}{\sum_{i=1}^{n} (x_i - \bar{x}_n) x_i}$$

Take conditional expectation on both sides

$$\begin{aligned} E\left[\hat{\beta}_1 | X\right] &= \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) E\left[U_i | X_i = x_i\right]}{\sum_{i=1}^{n} (x_i - \bar{x}_n) x_i} \\ &= \beta_1 \end{aligned}$$

By the law of iterated expectation,

$$E\left[\hat{\beta}_1\right] = E\left[E\left[\hat{\beta}_1 | X\right]\right] = E\left(\beta_1\right) = \beta_1$$

Outline

- Unbiasedness of the Estimator
- **Omitted Variable Bias**
- Multiple Regression

Omitted Variable Bias

- Now we focus on causal analysis with $E[U|X] = 0$.
- In causal analysis, SLR.4 $E[U|X] = 0$ is often a suspicious assumption in a simple regression model.
- U include variables other than X that affects Y. In other words, it includes the variables that are omitted.
- Let's pick one omitted variable Z:

$$U = \beta_2 Z + \varepsilon$$

- What happens if X is correlated with Z?

Omitted Variable Bias

- For simplicity, let's assume $Z = \delta X + \xi$, also assume $E(\varepsilon|X) = 0$, $E(\xi|X) = 0$.
- Then $U = \beta_2 Z + \varepsilon = \beta_2 (\delta X + \xi) + \varepsilon = \beta_2 \delta X + \beta_2 \xi + \varepsilon$ and

$$E[U|X] = \beta_2 \delta X \neq 0$$

- But let $U^* = \beta_2 \xi + \varepsilon$, then

$$Y = \beta_0 + (\beta_1 + \beta_2 \delta) X + U^*.$$

  This new error term U^* satisfies

$$E[U^*|X] = E[\beta_2 \xi + \varepsilon | X] = \beta_2 E[\xi|X] + E[\varepsilon|X] = 0$$

  .

- Thus, when regressing Y on X, the OLS estimator satisfies

$$E\left[\hat{\beta}_1\right] = \beta_1 + \beta_2 \delta \neq \beta_1$$

Omitted Variable Bias

- Bias of $\hat{\beta}_1$: $E\left(\hat{\beta}_1\right) - \beta_1 = \beta_2\delta$
- Positive bias if β_2 and δ have the same signs
- i.e. the omitted variable has positive direct effect on Y and is positively correlated with X or the omitted variable has negative direct effect on Y and is negatively correlated with X
- Negative bias if they have opposite signs

Omitted Variable Bias

- Example: X : education, Y : wage, Z : talent
- Then, $\delta > 0$ if Education and talent are positively correlated.
- $\beta_2 > 0$ if talent affects income directly or through channels other than education.
- If those are true, the OLS estimator for return on education is upward biased.

Omitted Variable Bias

- How to deal with "Omitted Variable Bias"?
- The first answer: do not omit variables!
- We thus need regressions that have more than one regressors.