1. Linear Regression Model

In this note we will consider a simple model in which \( y \) is determined by two explanatory variables \((x_1, x_2)\) plus some random noise, \( u \) as follows.

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u
\]

We will refer to this as the “true” model of \( y \). It is assumed to satisfy the following four conditions.

(1) The model is linear in the parameters \( \beta \).
(2) \( n \) observations drawn randomly from the population of interest.
(3) \( E(u|x_1, x_2) = 0 \). No endogeneity in the true model.
(4) \( x_1 \neq \alpha x_2 \) for any constant \( \alpha \). No perfect collinearity.

These four assumptions guarantee that the OLS estimates of \( \beta_1 \) and \( \beta_2 \) would be unbiased.

\[
E(\beta_j^{OLS}) = \beta_j \quad j = 1, 2
\]

2. Sources of Endogeneity

In general the problem of ‘endogeneity’ refers to anytime there is a violation of the third assumption. In other words, an empirical model for which \( E(u|x) \neq 0 \) is said to suffer from an endogeneity problem. Whenever there is endogeneity, OLS estimates of the \( \beta \)'s will no longer be unbiased. There are at least three generally recognized sources of endogeneity.

(1) Model misspecification or Omitted Variables.
(2) Measurement Error.
(3) Simultaneity.

The classic meaning of endogeneity refers to the simultaneity problem where the flow of causality is not purely from the RHS variables to the LHS variable. In other words, if we think that changes in the LHS variable may cause changes in a RHS variable or that the LHS variable and a RHS variable are being jointly determined, then there is simultaneity and we would not expect the error term to be uncorrelated with the RHS variables. The example discussed in class of regressing crime rates on size of policy fits into this category of endogeneity.
3. Omitted Variables

In this note we focus on the problem of omitted variables. Suppose that in the true linear model above, we simply do not have data for $x_2$. So instead we estimate the following

$$ y = \beta_0 + \beta_1 x_1 + u $$

Let $\hat{\beta}$ be the OLS estimates of $\beta$ from this regression. Is it still true in particular that $E(\hat{\beta}_1) = \beta_1$ - that our estimate of $\beta_1$ will unbiased? In general, unfortunately, no.

To see why, note that the OLS estimate of $\beta_1$ would be

$$ \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) y_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} $$

Substituting in the true model for $y$, we get

$$ \bar{\beta}_1 = \frac{\beta_0 \sum_{i=1}^n (x_{1i} - \bar{x}_1) + \beta_1 \sum_{i=1}^n (x_{1i} - \bar{x}_1) x_{1i} + \beta_2 \sum_{i=1}^n (x_{1i} - \bar{x}_1) x_{2i} + \sum_{i=1}^n (x_{1i} - \bar{x}_1) u_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} $$

Note that $\beta_0 \sum_{i=1}^n (x_{1i} - \bar{x}_1) = 0$ (since $\sum_{i=1}^n (x_{1i} - \bar{x}_1) = 0$) and $\sum_{i=1}^n (x_{1i} - \bar{x}_1) x_{1i} = \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2$ so we have

$$ \bar{\beta}_1 = \frac{\beta_1 \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 + \beta_2 \sum_{i=1}^n (x_{1i} - \bar{x}_1) x_{2i} + \sum_{i=1}^n (x_{1i} - \bar{x}_1) u_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} $$

Taking the expectation of both sides we get

$$ E(\bar{\beta}_1) = \beta_1 + \frac{\beta_2 \sum_{i=1}^n (x_{1i} - \bar{x}_1) x_{2i}}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} $$

which uses the fact that $E(u|x) = 0$ by assumption from the true model. So our estimate of $\beta_1$ when we omit $x_2$ is biased by the second term in that equation. What the hell is that? It turns out that term (without the $\beta_2$) is exactly what you would get if you regress $x_2$ on $x_1$. In other words we can write

$$ E(\hat{\beta}_1) = \beta_1 + \beta_2 \hat{\delta}_1 $$

where $\hat{\delta}_1$ is the OLS estimate from the following regression
Now we have a tidy way to think about the bias introduced into our estimate of $\beta_1$ by the omission of $x_2$ from the model. First note that if $x_1$ and $x_2$ are not correlated at all in the sample, then $\delta_1$ would be equal to 0 and there would be no bias. In other words, the bias is only a problem to the extent that the omitted variable is correlated with other explanatory variables. If it is not, then leaving it out will in general damage the efficiency of our parameter estimates but the extra noise introduced by the omission would be unbiased noise. To get an idea of which direction the bias goes, we will typically have to have a theoretical story about about the sign of $\beta_2$ and $\delta_1$ since we cannot actually estimate either of these coefficients without the omitted variable. (and the problem is that we don’t access to that data)

Example 1. Consider the following true regression model

$$wage = \beta_0 + \beta_1 edu + \beta_2 ability + u$$

Suppose we omit ability from the regression (because we can’t observe it directly). Then the expected value of our OLS estimate would be

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \delta_1$$

where $\beta_2$ is the (hypothetical) coefficient on ability and $\delta_1$ is the hypothetical covariance between ability and education. We would probably expect both of these coefficients to be positive: People with greater ability get paid more. People with greater ability stay in school longer. If we feel these hypotheses about $\beta_2$ and $\delta_1$ are correct, then we would argue that $\hat{\beta}_1$ overstates the effect of education on wages.

Example 2. Consider the following true regression model

$$testScore = \beta_0 + \beta_1 expenditure + \beta_2 stress + u$$

where expenditure is expenditure per pupil and where stress is a measure of hardships due to a poor home environment (poor nutrition, neglect, poor sleep habits, etc). We omit stress from the regression because we cannot observe it. Which direction would that bias $\beta_1$? We might expect $\beta_2$ to be negative and given the political reality of school funding (property tax based), we would probably expect the correlation between stress and expenditure to also be negative. (Wealthier neighborhoods spend more per pupil and also have households with lower stress levels.) Hence, the estimate is biases upward (negative times negative is a positive).

4. References
