Omitted Variable Bias and Bad Controls

Chungsang Tom Lam[1]

[1]Department of Economics
Clemson University

February 6, 2020

Recall the Partitioned Regression

Consider the true model:

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon \tag{1}$$

Recall:

pure effect of X1

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}(X_1'X_2)\hat{\beta}_2$$
$$\hat{\beta}_2 = (X_2'X_2)^{-1}X_2'y - (X_2'X_2)^{-1}(X_2'X_1)\hat{\beta}_1$$

Define:

$$\tilde{\beta}_1 = (X_1'X_1)^{-1}X_1'y$$
$$\tilde{\beta}_2 = (X_2'X_2)^{-1}X_2'y$$

Omitted Variable Bias

- $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimates for regressing y on X_1 and X_2 in one regression.
- $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are the estimates for regressing y on X_1 only and regressing y on X_2 only.
- We already know $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased and consistent under the OLS assumptions. For $\tilde{\beta}_1$ and $\tilde{\beta}_2$ to be an unbiased and consistent estimator of β_1 and β_2, we consider the expectation and probability limit of the terms $(X_1'X_1)^{-1}(X_1'X_2)\hat{\beta}_2$ and $(X_2'X_2)^{-1}(X_2'X_1)\hat{\beta}_1$.

Omitted Variable Bias

- When X_1 and X_2 are uncorrelated **OR** $\beta_2 = 0$, $\tilde{\beta}_1$ is an unbiased and consistent estimator of β_1
- When X_1 and X_2 are uncorrelated **OR** $\beta_1 = 0$, $\tilde{\beta}_2$ is an unbiased and consistent estimator of β_2
- We can also determine the sign of the bias of $\tilde{\beta}_1$ if we have an idea of how X_1 and X_2 are correlated and the sign of β_2.

$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}(X_1'X_2)\hat{\beta}_2$ $\tilde{\beta}_1 = (X_1'X_1)^{-1}X_1'y$ $\tilde{\beta}_2$

if X1 and X2 positive correlated: $(X_1'X_1)^{-1}(X_1'X_2)$ >0

then, if $\hat{\beta}_2$ >0, $(X_1'X_1)^{-1}(X_1'X_2)\hat{\beta}_2$ >0, $\hat{\beta}_1 = \tilde{\beta}_1$ - a positive number,,,if we use $\tilde{\beta}_1$ to estimate β_1,

β_1 is overesitimated

	positively correlated X_1,X_2	negatively correlated X_1,X_2
positive β_2	β_1 overestimated	β_1 underestimated
negative β_2	β_1 underestimated	β_1 overestimated

Bad Controls

- So if omitted variable bias is so bad, we should just add everything we have into the regression?
- What exactly is $(X_1'X_1)^{-1}(X_1'X_2)\hat{\beta}_2$?
- The above proof only tell us that if X_2 is omitted, (under certain conditions), $\hat{\beta}_1$ is not an unbiased and consistent estimator of β_1. It does not tell us whether β_1 is the parameter we want to estimate.
- If X_1 can indirectly affect y via affecting X_2, we need to consider if we want to take into account this indirect effect or not.
- Including X_2 in the regression essentially taking the indirect effect away from the estimation (aka controlling for X_2), we may NOT want to do that.

Bad Controls