ECON 340
Economic Research Methods

Div Bhagia

Lecture 18
Omitted Variable Bias, Multiple Regression Model

Test Scores and Class Size

	Dependent variable:
	testscr
str	-2.280***
	(0.480)
Constant	698.933***
	(9.467)
Observations	420
R2	0.051
Adjusted R2	0.049
Note:	*p<0.1; **p<0.05; ***p<0.01

Omitted Variable Bias

- School districts with lower student–teacher ratios tend to have higher test scores
- However, students from districts with small classes may have other advantages that help them perform well
- Omitted factors (e.g. student characteristics) can make the OLS estimator biased
- Today's lecture: *omitted variable bias* and *multiple regression*, a method that can eliminate this bias

Omitted Variable Bias

Consider the following linear regression model:

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

Omitted variable bias occurs when both conditions are true:

(1) The omitted variable is correlated with X

(2) The omitted variable $\rightarrow$ Y

Omitted Variable Bias

In our example:

$$TestScore = \beta_0 + \beta_1 \cdot STR + u$$

Which of these omitted factors will lead to bias?

(a) percentage of English learners

(b) time of day when tests were conducted

(c) parking lot space per pupil

(d) computers per student

Omitted Variable Bias

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

- Remember u represents all factors, other than X, that are determinants of Y.
- Omitted Variable Bias means that the exogeneity assumption $E(u|X) = 0$ doesn't hold.
- If $E(u|X) \neq 0$, OLS estimator is biased.

Omitted Variable Bias

When $E(u|X) \neq 0$,

$$\hat{\beta}_1 = \beta_1 + \frac{Cov(X, u)}{Var(X)}$$

Direction and strength of bias depends on the correlation between u and X.

Omitted Variable Bias

In our example:

$$TestScore = \beta_0 + \beta_1 \cdot STR + u$$

What should be the direction of bias due to the following omitted variables?

(a) percentage of English learners

(b) computers per student

Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

- Again can be used for both purposes, causal inference and prediction
- As before we need the data to come from a random sample and no large outliers, but now in addition we also need that X_1 and X_2 are not perfectly multi collinear.
- Moreover, we can modify the mean independence to:

$$E(u|X_1, X_2) = 0$$

Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

- Assumptions: (1) random sample, (2) no large outliers, (3) no perfect multicollinearity, (4) $E(u|X_1, X_2) = 0$
- Under these assumptions, β_1 captures the causal effect of X_1 keeping X_2 constant, and β_2 captures the causal effect of X_2 keeping X_1 constant.

Control Variables

- While there are cases where we might want to evaluate the effect of both the variables, it is hard to find exogenous variables

- A really good use of the multiple regression model is to instead *control* for omitted variable W while trying to estimate the causal effect of X

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

Control Variables

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

- So instead of assumption (4), we can assume *conditional mean independence*

$$E(u|X, W) = E(u|W)$$

- The idea is that once you control for the W, X becomes independent of u

- Under this modified assumption, we can interpret β_1 as the causal effect of X while *controlling* for W

In Summary

$$TestScore = \beta_0 + \beta_1 \cdot STR + \beta_2 \cdot comp\_stu + u$$

- Under assumption:

$$E(u|STR, comp\_stu) = 0$$

β_1 causal impact of STR, and β_2 causal impact of $comp\_stu$

- Under conditional independence:

$$E(u|STR, comp\_stu) = E(u|comp\_stu)$$

β_1 causal impact of STR, and β_2 could still be biased

Test Scores and Class Size

	Dependent variable:	
	testscr	
	(1)	(2)
str	-2.280***	-1.593***
	(0.480)	(0.493)
comp_stu		65.160***
		(14.351)
Observations	420	420
R2	0.051	0.096
Adjusted R2	0.049	0.092

Goodness of Fit: The R^2

Total Sum of Squares: $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Explained Sum of Squares: $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

Residual Sum of Squares: $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2$

$$TSS = ESS + RSS$$

A measure of goodness of fit:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Adjusted R^2

R^2 never decreases when an explanatory variable is added

An alternative measure called Adjusted R^2

$$AdjustedR^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

where k is the number of variables.

$AdjustedR^2$ only rises if RSS declines by a larger percentage than the degrees of freedom.