Linear Regression with Multiple Regressors (SW Ch. 6)

Ercan Karadas

Econometrics (ECON 3112)

Belk College of Business, UNCC

December 3, 2018

Outline

Omitter Variable Bias

Population Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Recall that the error u_i arises because of factors, or variables, that influence Y but are not included in the regression function.
- There are always omitted variables!
- But sometimes, the omission of those variables can lead to a **bias** in the OLS estimator of β_1.
  - Bias in the sense that the model either **always** underestimates or **always** overestimates the true causal effect of X on Y.
- Mathematically, the presence of such a bias lead to one of the following:
  - $\widehat{\beta}_1 > \beta_1$ (Overestimation)
  - $\widehat{\beta}_1 < \beta_1$ (Underestimation)

  for **all** samples.

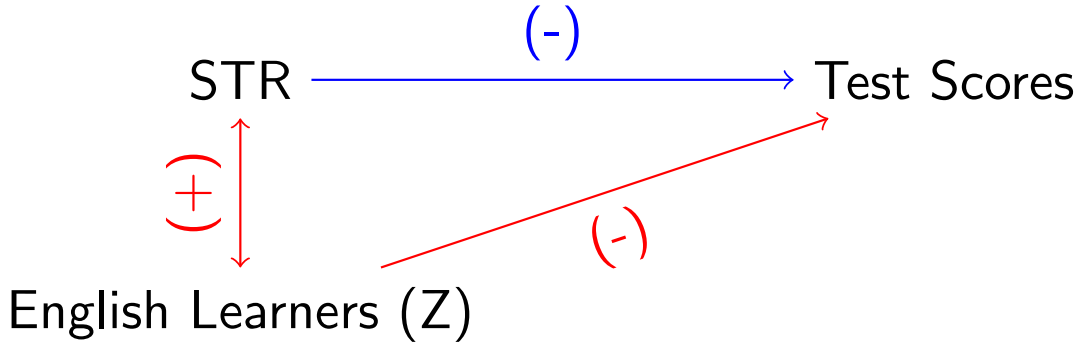
Two Conditions for OV Bias

- The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called **omitted variable bias (OV bias)**
- Suppose Z is a variable that is not already in the model.
- For an omitted variable bias to occur, the omitted variable Z must satisfy two conditions:
  1) Z is a determinant of Y (i.e. Z is part of u); **and**
  2) Z is correlated with the regressor X (i.e. $corr(Z, X) \neq 0$)
- It is important that **both** conditions must hold for the omission of Z to result in an OV bias.

Example: Test Scores Example

- In the test score example, we haven't included English language ability (whether the student has English as a second language) as a regressor.
- Let's define Z as the percentage of English learners in a district. So the variable Z is omitted in the model. Does it lead to an OV bias?
- Let's check the two conditions of OV bias:
  - English language ability plausibly affects standardized test scores: Z is a determinant of Y.
  - Districts with high ratio of immigrant communities (high Z) tend to be less affluent and thus have smaller school budgets and higher STR: Z is correlated with X.
- Both conditions for OV bias are satisfied, therefore $\widehat{\beta}_1$ is biased.
- What is the direction of this bias? Two Approaches:
  - Use an arrow analysis
  - Use a formula

(i) Arrow Analysis



- Want: causal effect of STR on Test Scores, i.e. how much Test Scores will $\uparrow$ in response to a unit $\downarrow$ in STR.
- However,
  - as STR $\downarrow$, on average $Z \downarrow$ as well
  - as $Z \downarrow$, on average Test Scores $\uparrow$
- Therefore, $\downarrow$ in STR has a direct effect on Test Scores (blue arrow) and an indirect effect on Test Scores through Z (red arrows).
- The model that omits Z would overestimate the true causal effect of STR on Test Scores.

(ii) OV Formula

$$\widehat{\beta}_1 \to \beta_1 + \left(\frac{\sigma_u}{\sigma_x}\right)\rho_{xu}, \qquad \rho_{xu} = \text{corr}(X, u)$$

- If LSA # 1 holds:
  - $\rho_{xu} = 0$ and therefore, $\widehat{\beta}_1 \to \beta_1$.
  - $\widehat{\beta}_1$ is a consistent estimator of β_1 (and unbiased)
- If LSA # 1 does NOT hold:
  - $\rho_{xu} \neq 0$ and therefore, $\widehat{\beta}_1 \to \beta_1$.
  - $\widehat{\beta}_1$ is an INconsistent estimator of β_1 (and not unbiased)
- The sign of ρ_{xu} also determines the direction of bias. Suppose $\beta_1 < 0$, then:
  - $\rho_{xu} < 0$: overestimation
  - $\rho_{xu} > 0$: underestimation

Example: Test Scores Example

- $\beta_1 < 0$: STR $\downarrow \Longrightarrow$ Test Scores $\uparrow$
- What about the sign of $\rho_{STR,u}$?
- To determine the sign of $\rho_{STR,u}$, two steps
  - $u = \beta_2 PctEL +$ some other factors... (Z=PctEL)
  - corr$(STR, PctEL) > 0$ and $\beta_2 < 0$
- Therefore, $\rho_{STR,u} < 0$
- Therefore, without PctEL, the model overestimates the true causal effect of STR on Test Scores!
- Let's see whether that is the case...

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent EL (PctEL) have smaller classes
- Among districts with comparable PctEL, the effect of class size is small (recall overall "test score gap" $= 7.4$)

Digression: Causality and Regression Analysis

- The Test Scores example shows that, if an omitted variable satisfies the two conditions for omitted variable bias, then the OLS estimator in the regression omitting that variable is biased and inconsistent.
- So, even if n is large, $\widehat{\beta}_1$ will not be close to β_1
- This raises a deeper question: how do we define β_1?
- That is, what precisely do we want to estimate when we run a regression?

There are (at least) three possible answers to this question:

1) We want to estimate the slope of a line through a scatterplot as a simple summary of the data to which we attach no substantive meaning.
   - This can be useful at times, but isn't very interesting intellectually and isn't what this course is about.

2) We want to make forecasts, or predictions, of the value of Y for an entity not in the data set, for which we know the value of X.
   - Forecasting is an important job for economists, and excellent forecasts are possible using regression methods without needing to know causal effects.

3) We want to estimate the causal effect on Y of a change in X.
   - This is why we are interested in the class size effect. Suppose the school board decided to cut class size by 2 students per class. What would be the effect on test scores? This is a causal question (what is the causal effect on test scores of STR?) so we need to estimate this causal effect. The aim of this course is the estimation of causal effects using regression methods.

What, precisely, is a causal effect?

- "Causality" is a complex concept!
- In this course, we take a practical approach to defining causality.
- A **causal effect** is defined to be the effect measured in an ideal randomized controlled experiment.

Ideal Randomized Controlled Experiment

- **Ideal:** subjects all follow the treatment protocol perfect compliance, no errors in reporting, etc.!
- **Randomized:** subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors)
- **Controlled:** having a control group permits measuring the differential effect of the treatment
- **Experiment:** the treatment is assigned as part of the experiment: the subjects have no choice, so there is no reverse causality in which subjects choose the treatment they think will work best.

Back to class size:

Imagine an ideal randomized controlled experiment for measuring the effect on Test Score of reducing STR...

- In that experiment, students would be randomly assigned to classes, which would have different sizes.
- Because they are randomly assigned, all student characteristics (and thus u) would be distributed independently of STR.
- Thus, $E(u|STR) = 0$, that is, LSA $\#1$ holds in a randomized controlled experiment.

How does our observational data differ from this ideal?

- The treatment is not randomly assigned
- Consider PctEL - percent English learners - in the district. It plausibly satisfies the two criteria for omitted variable bias: Z = PctEL is:
  - a determinant of Y; and
  - correlated with the regressor X.
- Thus, the "control" and "treatment" groups differ in a systematic way, so corr$(STR, PctEL) \neq 0$.
- Randomization + control group means that any differences between the treatment and control groups are random - not systematically related to the treatment

- We can eliminate the difference in PctEL between the large (control) and small (treatment) groups by examining the effect of class size among districts with the same PctEL.
  - If the only systematic difference between the large and small class size groups is in PctEL, then we are back to the randomized controlled experiment - within each PctEL group.
  - This is one way to "control" for the effect of PctEL when estimating the effect of STR.

Three ways to overcome omitted variable bias

1) Run a randomized controlled experiment in which treatment (STR) is randomly assigned: then PctEL is still a determinant of TestScore, but PctEL is uncorrelated with STR. (This solution to OV bias is rarely feasible.)
2) Adopt the "cross tabulation" approach, with finer gradations of STR and PctEL - within each group, all classes have the same PctEL, so we control for PctEL (But soon you will run out of data, and what about other determinants like family income and parental education?)
3) Use a regression in which the omitted variable (PctEL) is no longer omitted: include PctEL as an additional regressor in a multiple regression.

The Multiple Regression Model

Consider the case of two regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Y is the dependent variable
- X_1, X_2 are the two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) denote the ith observation on Y, X_1, and X_2.
- β_0 is the unknown population intercept
- β_1 is the effect on Y of a change in X_1, holding X_2 constant.
- β_2 is the effect on Y of a change in X_2, holding X_1 constant.
- u_i is the regression error (omitted factors)

Interpretation of coefficients in multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Consider changing X_1 by ΔX_1 while holding X_2 constant

- Population regression line before the change:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Population regression line after the change:

$$Y_i + \Delta Y = \beta_0 + \beta_1 (X_{1i} + \Delta X_1) + \beta_2 X_{2i} + u_i$$

- Difference

$$\Delta Y = \beta_1 \Delta X_1$$

- So
  - $\beta_1 = \frac{\Delta Y}{\Delta X_1}$, **holding X_2 constant**
  - $\beta_2 = \frac{\Delta Y}{\Delta X_2}$, **holding X_1 constant**
  - β_0 = predicted value of Y when $X_1 = X_2 = 0$.

The OLS Estimator in Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- How can we estimate β_0, β_1, and β_2 from data?
- We will again focus on the least squares ("ordinary least squares" or "OLS") estimator of the unknown parameters β_0, β_1, and β_2.
- The OLS estimator minimizes the average squared difference between the actual values of Yi and the prediction (predicted value) based on the estimated line:
- The OLS estimator solves,

$$\min_{b_0, b_1, b_2} \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i}) \right]^2$$

- This minimization problem can be solved using calculus.
- The result is the OLS estimators of β_0, β_1, and β_2.

Example: Test Score - Class Size data

- Regression of Test Score against STR:

$$\widehat{\text{Test Score}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR$$

- Now include percentage of English Learners in the district (PctEL):

$$\widehat{\text{Test Score}} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.65} \times PctEL$$

  - The coefficient on STR in the second regression is the effect on Test Scores of a unit change in STR, *holding constant the percentage of English Learners in the district*.
  - Notice that the coefficient on STR falls by one-half. This is because of the omitted variable bias that we discussed in the first section.

R Codes for Estimating Multiple Regression

```
#Prepare variables
library("AER")
data(CASchools)
TestScore <- (CASchools$read + CASchools$math)/2
STR <- CASchools$students/CASchools$teachers
PctEL <- CASchools$english

# Estimate multiple regression and save in myReg:
myReg <- lm(TestScore ~ STR + PctEL)

# An alternative way to estimate multiple regression on R:
myReg <- lm(TestScore ~ STR + english, data = CASchools)
```

R Output (Homoskedasticity-only)

```
> summary(myReg)  # this command produces the output below

Call:
lm(formula = TestScore ~ STR + PctEL)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 686.03224    7.41131  92.566  < 2e-16 ***
STR          -1.10130    0.38028  -2.896  0.00398 **
PctEL        -0.64978    0.03934 -16.516  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

R Output (Heteroskedasticity-robust)

```
# To produce heteroskedasticity-robust output:
> library("car")
> S(myReg, vcov. = hccm) # this replaces "summary" function

Call: lm(formula = TestScore ~ STR + PctEL)
Standard errors computed by hccm

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 686.0322     8.8122   77.85   <2e-16 ***
STR          -1.1013     0.4371   -2.52   0.0121 *
PctEL        -0.6498     0.0313  -20.76   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05

Residual standard deviation: 14.46 on 417 degrees of freedom
Multiple R-squared: 0.4264
F-statistic: 220.1 on 2 and 417 DF,  p-value: < 2.2e-16
```

Measures of Fit in Multiple Regression

As in single regressor case, two regression statistics provide complementary measures of how well the regression line "fits" or "explains" the data:

- The **standard error of the regression (SER)** measures the magnitude of a typical regression residual in the units of Y.
- The **regression $\mathbf{R^2}$** measures the fraction of the variance of Y that is explained by X;
  - it is unitless
  - ranges between zero (no fit) and one (perfect fit)

The Standard Error of the Regression (SER)

- The SER measures the spread of the distribution of $\widehat{u}$:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{u}_i^2} = \sqrt{\frac{1}{n-k-1}SSR}$$

- The SER
  - measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line)
  - has the units of u, which are the units of Y. Therefore, the SER measures the spread of the Y's around the sample regression line.
- The Root Mean Squared Error (RMSE) measures the same thing as the SER without d.f. correction:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2}$$

The Regression $\mathbf{R^2}$ and $\overline{\mathbf{R}}^2$ (adjusted $\mathbf{R^2}$)

- **The Regression $\mathbf{R^2}$** is the fraction of the sample variance of Y_i explained by the regression:

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

$$ESS = \sum_i \left(\widehat{Y}_i - \overline{Y}\right)^2,\ SSR = \sum_i \widehat{u}_i^2,\ TSS = \sum_i \left(Y_i - \overline{Y}\right)^2.$$

- The R^2 always increases when you add another regressor (why?) - a bit of a problem for a measure of "fit".
- The $\overline{R}^2$ (the "adjusted R^2") corrects this problem by "penalizing" you for including another regressor - the R^2 does not necessarily increase when you add another regressor.

$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$$

- Note that $\overline{R}^2 < R^2$, however if n is large the two will be very close.

Example: Test Score - Class Size data

$$(1) \quad \widehat{\text{Test Score}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR$$

$$R^2 = 0.05, \quad SER = 18.6$$

$$(2) \quad \widehat{\text{Test Score}} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.65} \times PctEL$$

$$R^2 = 0.426, \quad \overline{R}^2 = 0.424, \quad SER = 14.5$$

- What - precisely - does this tell you about the fit of regression (2) compared with regression (1)?
- Why are the $\overline{R}^2$ and the R^2 so close in (2)?

The Least Squares Assumptions (LSA) in Multiple Regression

- What, in a precise sense, are the properties of the sampling distribution of the OLS estimators $\widehat{\beta}_i$s?
  - When will $\widehat{\beta}_i$ be unbiased?
  - What is its variance?
  - Etc.
- To answer these questions, we need to make some assumptions about how Y and X are related to each other, and about how they are collected (the sampling scheme)
- These assumptions - there are four - are known as the Least Squares Assumptions.
- The first three of these assumption simply generalizations of the assumptions for the single case, but the fourth one is new.

The Least Squares Assumptions (LSA)

The Population Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, 2, \ldots, n$$

A1. $E[u_i | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k] = 0$: the mean of u is zero for any given values of X's.
   - This implies that $\widehat{\beta}_1$ is unbiased

A2. $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ are i.i.d.
   - This delivers the sampling distribution of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (Why?)

A3. Large outliers in X and/or Y are rare.
   - Technically, X and Y have finite fourth moments

A4. There is no perfect multicollinearity.
   - Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

The LSA # 1: $E[u|X_1 \ldots, X_k] = 0$

- This has the same interpretation as in regression with a single regressor.
- Failure of this condition leads to omitted variable bias, specifically, if an omitted variable
  - belongs in the equation (so is in u) **and**
  - is correlated with an included X

  then this condition fails and there is OV bias.
- The best solution, if possible, is to include the omitted variable in the regression.
- A second, related solution is to include a variable that controls for the omitted variable (discussed in Ch. 7)

The LSA # 2: $(X_{1i}, \ldots, X_{ki}, Y_i)$'s are i.i.d.

- This arises automatically if the entity (individual, district) is sampled by simple random sampling:
- The entities are selected from the same population, so $(X_{1i}, \ldots, X_{ki}, Y_i)$ are identically distributed for all $i = 1, \ldots, n$.
- The entities are selected at random, so the values of $(X_{1i}, \ldots, X_{ki}, Y_i)$ for different entities are independently distributed.
- The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data) - we will deal with that complication in later chapters.
- This assumption is the key to apply LLN and CLT to obtain the sampling distribution of our estimators

The LSA # 3: Outliers are Rare

- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then this assumption is met (Technically, this implies X and Y have finite fourth moments, i.e. $E\left(X^4\right) < \infty$ and $E\left(Y^4\right) < \infty$)
- Standardized test scores automatically satisfy this; STR, family income, etc. satisfy this too.
- The substance of this assumption is that a large outlier can strongly influence the results - so we need to rule out large outliers.
- This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).

The LSA # 4: There is no Perfect Multicollinearity

- **Perfect multicollinearity** arises when one of the regressors is an exact linear function of the other regressors.
- We will return to perfect (and imperfect) multicollinearity shortly, with more examples...
- With these least squares assumptions in hand, we now can derive the sampling distribution of $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$.

The Distribution of the OLS Estimators in Multiple Regression

Under the four Least Squares Assumptions:

- The sampling distribution of $\widehat{\beta}_i$ has mean β_i, i.e. $E(\widehat{\beta}_i) = \beta_i$.
- $var(\widehat{\beta}_i)$ is inversely proportional to n
- Other than its mean and variance, the exact (finite-n) distribution of $\widehat{\beta}_i$ is very complicated; but for large n:
  - $\widehat{\beta}_i$ is consistent: $\widehat{\beta}_i \to \beta_i$
  - CLT applies

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\text{var}\left(\widehat{\beta}_i\right)}} \sim N(0,1), \qquad \text{for all } i$$

- Conceptually, there is nothing new here!

Multicollinearity

- **Perfect multicollinearity** arises when one of the regressors is an exact linear function of the other regressors.
- Mathematically, there is perfect multicollinearity if the following condition holds

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} \ldots + \lambda_k X_{ki} = 0$$

for some constants $\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_k$, not all zero.

Some more examples of perfect multicollinearity

- You include STR twice
- Regress Test Score on a constant, D, and B, where $D_i = 1$ if $STR \leq 20$, and 0 otherwise; $B_i = 1$ if $STR > 20$, and 0 otherwise. $B_i = 1 - D_i$ and therefore there is perfect multicollinearity.
- Would there be perfect multicollinearity if the intercept (constant) were excluded from this regression? This example is a special case of **the dummy variable trap**

The dummy variable trap

- Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive - that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other).
- If you include all these dummy variables and a constant, you will have perfect multicollinearity - this is sometimes called the **dummy variable trap**.
- Why is there perfect multicollinearity here? Use λs to answer formally...
- Solutions to the dummy variable trap:
  (1) Omit one of the groups (e.g. Senior), or
  (2) Omit the intercept
- What are the implications of (1) or (2) for the interpretation of the coefficients?

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know - either by crashing or giving an error message or by "dropping" one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Imperfect Multicollinearity

- Imperfect and perfect multicollinearity are quite different despite the similarity of the names.
- **Imperfect multicollinearity** occurs when two or more regressors are very highly correlated.
- Why the term "multicollinearity"? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line - they are "co-linear" - but unless the correlation is exactly ± 1, that collinearity is imperfect.

Imperfect Multicollinearity

- Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.
- **The idea:** the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant - so the data don't contain much information about what happens when X_1 changes but X_2 doesn't. If so, the variance of the OLS estimator of the coefficient on X_1 will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.
- The math? See SW, App. 6.2