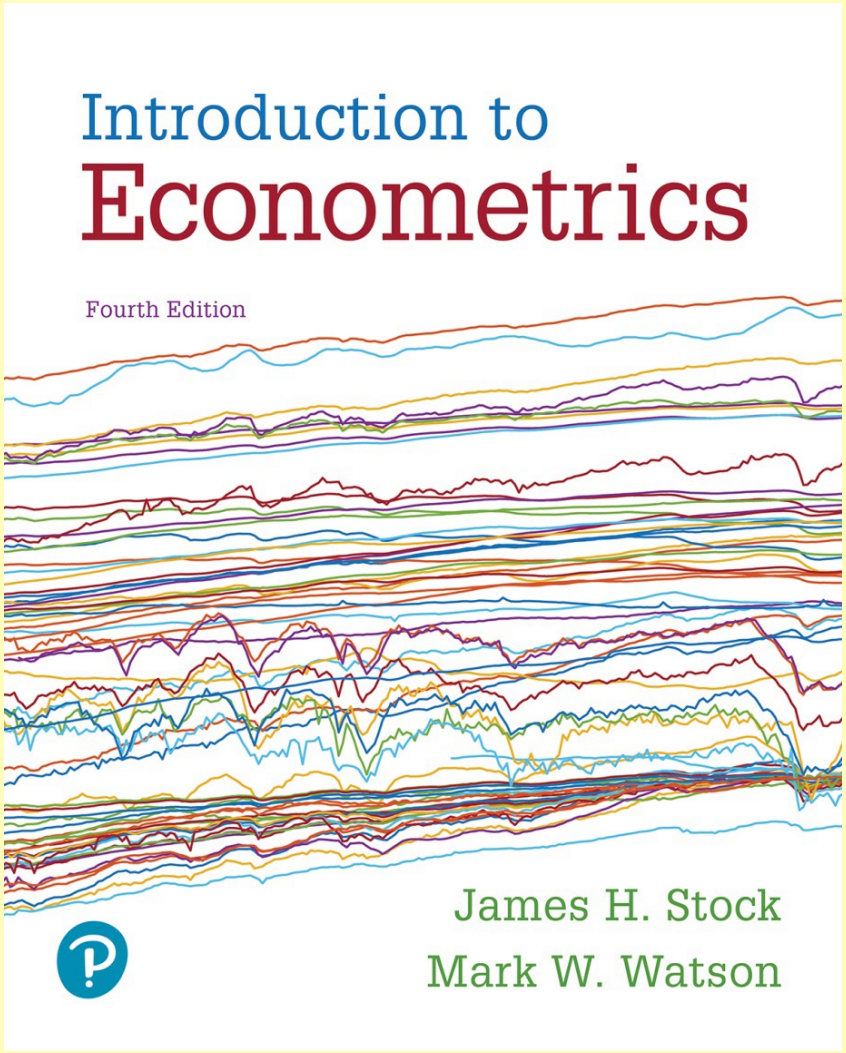
Introduction to Econometrics

Fourth Edition



Chapter 6

Linear Regression with Multiple Regressors

Outline

1. Omitted variable bias
2. Using regression to estimate causal effects
3. Multiple regression and OLS
4. Measures of fit
5. Sampling distribution of the OLS estimator with multiple regressors
6. Control variables

Omitted Variable Bias (SW Section 6.1) (1 of 5)

In the class size example, β_1 is the causal effect on test scores of a change in the *STR* by one student per teacher.

When β_1 is a causal effect, the first least squares assumption for causal inference must hold: $E(u|X) = 0$.

The error u arises because of factors, or variables, that influence Y but are not included in the regression function.

There are always omitted variables!

If the omission of those variables results in $E(u|X) \neq 0$, then the OLS estimator will be biased.

Omitted Variable Bias (SW Section 6.1)
(2 of 5)

The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:

<u>The two conditions for omitted variable bias</u>

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. $corr(Z, X) \neq 0$).

Both *conditions must hold for the omission of* Z *to result in omitted variable bias*.

Omitted Variable Bias (SW Section 6.1) (3 of 5)

In the test score example:

1. English language ability (whether the student has English as a second language) plausibly affects standardized test scores: Z is a determinant of Y.
2. Immigrant communities tend to be less affluent and thus have smaller school budgets and higher STR: Z is correlated with X.

Accordingly, $\hat{\beta}_1$ is biased. What is the direction of this bias?

- *What does common sense suggest?*
- If common sense fails you, there is a formula…

Omitted Variable Bias (SW Section 6.1) (4 of 5)

A formula for omitted variable bias: recall the equation,

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$.

Under Least Squares Assumption #1, $E[(X_i - \mu_X)u_i] = \text{cov}(X_i, u_i) = 0$.

But what if $E[(X_i - \mu_X)u_i] = \text{cov}(X_i, u_i) = \sigma_{Xu} \neq 0$?

Omitted Variable Bias (SW Section 6.1)
(5 of 5)

Let β_1 be the causal effect. Under LSA #2 and #3 (that is, even if LSA #1 does not hold),

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

$$= \left(\frac{\sigma_u}{\sigma_X}\right) \times \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu},$$

Where $\rho_{Xu} = \text{corr}(X,u)$. If assumption #1 is correct, then $\rho_{Xu} = 0,$ but if not we have….

The omitted variable bias formula (1 of 2)

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}$$

- If an omitted variable Z is ***both***:

1. a determinant of Y (that is, it is contained in u); ***and***
2. correlated with X, then $\rho_{Xu} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased and is not consistent.

- For example, districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the effect of having many ESL students factor would result in overstating the class size effect. *Is this is actually going on in the CA data*?

The omitted variable bias formula (2 of 2)

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 Student-Teacher Ratio	
	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)

Using regression to estimate causal effects

- The test score / STR / fraction English Learners 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, $\hat{\beta}_1$ will not be close to β_1.
- We have distinguished between two uses of regression: for prediction, and to estimate **causal effects**.
  - Regression also can be used simply to summarize the data without attaching any meaning to the coefficients or for any other purpose, but we won't focus on this use.
- In the class size application, we clearly are interested in a **causal effect**: what do we expect to happen to test scores if the superintendent reduces the class size?

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.

Comment: shuffle a deck of cards, how many times?

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_i) would be distributed independently of STR_i.
- Thus, $E(u_i|STR_i) = 0$ – that is, LSA #1 holds in a randomized controlled experiment.

How does our observational data differ from this ideal? (1 of 2)

- 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:
  1. a determinant of Y; ***and***
  2. correlated with the regressor X.
- Thus, the "control" and "treatment" groups differ in a systematic way, so corr(STR, $PctEL$) $\neq 0$

How does our observational data differ from this ideal? (2 of 2)

- Randomization implies 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 class (control) and small class (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.

Return to omitted variable bias

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 Population Multiple Regression Model (SW Section 6.2)

- Consider the case of two regressors:

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

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

Interpretation of coefficients in multiple regression (1 of 2)

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

Consider the difference in the expected value of Y for two values of X_1 holding X_2 constant:

Population regression line when $X_1 = X_{1,0}$:

$$Y = \beta_0 + \beta_1 X_{1,0} + \beta_2 X_2$$

Population regression line when $X_1 = X_{1,0} + \Delta X_1$:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_{1,0} + \Delta X_1) + \beta_2 X_2$$

Interpretation of coefficients in multiple regression (2 of 2)

Before : $Y = \beta_0 + \beta_1 X_{1,0} + \beta_2 X_2$

After : $Y + \Delta Y = \beta_0 + \beta_1 (X_{1,0} + \Delta X_1) + \beta_2 X_2$

Difference : $\Delta Y = \beta_1 \Delta X_1$

So :

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \text{ holding } X_2 \text{ constant}$$

$$\beta_2 = \frac{\Delta Y}{\Delta X_2}, \text{ holding } X_1 \text{ constant}$$

$$\beta_0 = \text{predicted value of } Y \text{ when } X_1 = X_2 = 0.$$

The OLS Estimator in Multiple Regression (SW Section 6.3)

- With two regressors, the OLS estimator solves:

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

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **This yields the OLS estimators of β_0 and β_1.**

Example: the California test score data

Regression of $TestScore$ against STR:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

Now include percent English Learners in the district ($PctEL$):

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

- What happens to the coefficient on STR?
- Note: corr(STR, $PctEL$) = 0.19

Multiple regression in STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |   -1.101296   .4328472    -2.54   0.011    -1.95213   -.2504616
       pctel |   -.6497768   .0310318   -20.94   0.000    -.710775   -.5887786
       _cons |    686.0322   8.728224    78.60   0.000    668.8754    703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

More on this printout later…

Measures of Fit for Multiple Regression (SW Section 6.4) (1 of 2)

Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$

SER = std. deviation of $\hat{u}_i$ (with d.f. correction)

$RMSE$ = std. deviation of $\hat{u}_i$ (without d.f. correction)

R^2 = fraction of variance of Y explained by X

$\bar{R}^2$ = "adjusted R^2" = R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$

SER and RMSE

As in regression with a single regressor, the ***SER*** (Standard Error of Regression) and the ***RMSE*** (Root Mean Squared Error) are measures of the spread of the Ys around the regression line:

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

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

R^2 and $\bar{R}^2$ (adjusted R^2) (1 of 2)

The R^2 is the fraction of the variance explained – same definition as in regression with a single regressor:

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

$$\text{where } ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2, \; SSR = \sum_{i=1}^{n} \hat{u}_i^2, \; TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

- The R^2 always increases when you add another regressor (*why?*) – a bit of a problem for a measure of "fit"

R^2 and $\bar{R}^2$ (adjusted R^2) (2 of 2)

The $\bar{R}^2$ (the "adjusted R^2") corrects this problem by "penalizing" you for including another regressor – the $\bar{R}^2$ does not necessarily increase when you add another regressor.

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

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

Measures of Fit for Multiple Regression (SW Section 6.4) (2 of 2)

Test score example:

(1) $\widehat{TestScore} = 698.9 - 2.28 \times STR,$

$$R^2 = .05,\ SER = 18.6$$

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL,$

$$R^2 = .426,\ \bar{R}^2 = .424,\ SER = 14.5$$

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

The Least Squares Assumptions for Causal Inference in Multiple Regression (SW Section 6.5)

Let $\beta_1, \beta_2, \ldots, \beta_k$ be causal effects.

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

1. The conditional distribution of *u* given the *X*'s has mean zero, that is, $E(u_i | X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0$.

2. $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are i.i.d.

3. Large outliers are unlikely: $X_1, \ldots, X_k$, and Y have four moments: $E(X_{1i}^4) < \infty, \ldots, E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$.

4. There is no perfect multicollinearity.

Assumption #1: the conditional mean of *u* given the included *X*s is zero (1 of 2)

$$E(u|X_1 = x_1, \ldots, X_k = x_k) = 0$$

- 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

  1. belongs in the equation (so is in u) ***and***
  2. 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 shortly)

Assumption #1: the conditional mean of *u* given the included *X*s is zero (2 of 2)

Assumption #2: $(X_{1i},\ldots,X_{ki},Y_i)$, $i=1,\ldots,n$, **are i.i.d.**

This is satisfied automatically if the data are collected by simple random sampling.

Assumption #3: large outliers are rare (finite fourth moments)

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).

Assumption #4: There is no perfect multicollinearity

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors

Example: Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
         str |  (dropped)
       _cons |   698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors

- In the previous regression, β_1 is the effect on $TestScore$ of a unit change in STR, holding STR constant (???)
- 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* $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$.

The Sampling Distribution of the OLS Estimator (SW Section 6.6)

Under the four Least Squares Assumptions,

- The sampling distribution of $\hat{\beta}_1$ has mean β_1
- $\text{var}(\hat{\beta}_1)$ is inversely proportional to n.
- Other than its mean and variance, the exact (finite-n) distribution of $\hat{\beta}_1$ is very complicated; but for large n...
  - $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (law of large numbers)
  - $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT)
  - These statements hold for $\hat{\beta}_1, \ldots, \hat{\beta}_k$

Conceptually, there is nothing new here!

Multicollinearity, Perfect and Imperfect (SW Section 6.7)

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

<u>Some more examples of perfect multicollinearity</u>

1. The example from before: you include STR twice,
2. Regress $TestScore$ on a constant, D, and B, where: $D_i = 1$ if $STR \leq 20, = 0$ otherwise; $B_i = 1$ if $STR > 20, = 0$ otherwise, so $B_i = 1 - D_i$ and there is perfect multicollinearity.
3. 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 (1 of 2)

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*?
- *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?*

The dummy variable trap (2 of 2)

- 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 (1 of 2)

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 (2 of 2)

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

Control variables and conditional mean independence (SW Section 6.8) (1 of 2)

We want to get an unbiased estimate of the effect on test scores of changing class size, holding constant factors outside the school committee's control – such as outside learning opportunities (museums, etc), parental involvement in education (reading with mom at home?), etc.

If we could run an experiment, we would randomly assign students (and teachers) to different sized classes.

Then STR_i would be independent of all the other factors that go into u_i, so $E(u_i|STR_i) = 0$ and the OLS slope estimator in the regression of $TestScore_i$ on STR_i would be an unbiased estimator of the desired causal effect.

Control variables and conditional mean independence (SW Section 6.8) (2 of 2)

But with observational data, u_i depends on additional factors (parental involvement, knowledge of English, access in the community to learning opportunities outside school, etc.).

- If you can observe those factors (e.g. $PctEL$), then include them in the regression.
- But usually you can't observe all these omitted causal factors (e.g. parental involvement in homework). ***In this case, you can include "control variables" which are correlated with these omitted causal factors, but which themselves are not causal.***

Control variables in multiple regression

A **control variable** W is a regressor included to hold constant factors that, if neglected, could lead the estimated causal effect of interest to suffer from omitted variable bias.

Control variables: an example from the California test score data (1 of 4)

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 = 0.773$$
$$(5.6)\quad (0.27)\quad (.033)\quad (.024)$$

$PctEL$ = percent English Learners in the school district

$LchPct$ = percent of students receiving a free/subsidized lunch (only students from low-income families are eligible)

- Which variable is the variable of interest?
- Which variables are control variables? Might they have a causal effect themselves? What do they control for?

Control variables: an example from the California test score data (2 of 4)

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 = 0.773$$
$$(5.6)\quad (0.27)\quad (.033)\quad (.024)$$

- STR is the variable of interest
- $PctEL$ probably has a direct causal effect (school is tougher if you are learning English!). But it is also a control variable: immigrant communities tend to be less affluent and often have fewer outside learning opportunities, and $PctEL$ is correlated with those omitted causal variables. *$PctEL$ is both a possible causal variable and a control variable.*
- $LchPct$ might have a causal effect (eating lunch helps learning); it also is correlated with and controls for income-related outside learning opportunities. *$LchPct$ is both a possible causal variable and a control variable.*

Control variables: an example from the California test score data (3 of 4)

1. Three interchangeable statements about what makes for an effective control variable:

i. An effective control variable is one which, when included in the regression, makes the error term uncorrelated with the variable of interest.

ii. Holding constant the control variable(s), the variable of interest is "as if " randomly assigned.

iii. Among individuals (entities) with the same value of the control variable(s), the variable of interest is uncorrelated with the omitted determinants of Y

Control variables: an example from the California test score data (4 of 4)

2. **Control variables need not be causal, and their coefficients generally do not have a causal interpretation.** For example:

$$\widehat{TestScore} = 700.2 - 1.00\,STR - 0.122\,PctEL - 0.547\,LchPct,\ \bar{R}^2 = 0.773$$
$$(5.6)\quad (0.27)\qquad (.033)\qquad\qquad (.024)$$

- Does the coefficient on $LchPct$ have a causal interpretation? If so, then we should be able to boost test scores (by a lot! Do the math!) by simply eliminating the school lunch program, so that $LchPct = 0$! (Eliminating the school lunch program has a well-defined causal effect: we could construct a randomized experiment to measure the causal effect of this intervention.)

The math of control variables: conditional mean independence (1 of 3)

- Because **a control variable is correlated with an omitted causal factor**, LSA #1 ($E(u_i|X_{1i}, \ldots, X_{ki}) = 0$) does not hold. For example, the coefficient on $LchPct$ is correlated with unmeasured determinants of test scores such as outside learning opportunities, so is subject to OV bias. But the fact that $LchPct$ is correlated with these omitted variables is precisely what makes it a good control variable!
- If LSA #1 doesn't hold, then what does?
- We need a mathematical condition for what makes an effective control variable. This condition is **conditional mean independence**: given the control variable, the mean of u_i doesn't depend on the variable of interest

The math of control variables: conditional mean independence (2 of 3)

Let X_i denote the variable of interest and W_i denote the control variable(s).

W is an effective control variable if conditional mean independence holds:

$$E(u_i|X_i, W_i) = E(u_i|W_i) \text{(conditional mean independence)}$$

If W is a control variable, then conditional mean independence replaces LSA #1 – it is the version of LSA #1 that is relevant for control variables.

The math of control variables: conditional mean independence (3 of 3)

Consider the regression model,

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

where X is the variable of interest, β_1 is its causal effect, and W is an effective control variable so that conditional mean independence holds:

$$E(u_i|X_i, W_i) = E(u_i|W_i)$$

In addition, suppose that LSA #2, #3, and #4 hold. Then:

1. β_1 has a causal interpretation.
2. $\hat{\beta}_1$ is unbiased
3. The coefficient on the control variable, $\hat{\beta}_2$, does not in general estimate a causal effect.

The math of conditional mean independence (1 of 5)

(SW Appendix 6.5) *Under conditional mean independence:*

1. β_1 has a causal interpretation.

The math: The expected change in *Y* resulting from a change in *X*, holding (a single) *W* constant, is:

$$
\begin{aligned}
&E(Y|X=x+\Delta x, W=w) - E(Y|X=x, W=w) \\
&\quad = [\beta_0 + \beta_1(x+\Delta x) + \beta_2 w + E(u|X=x+\Delta x, W=w)] \\
&\quad - [\beta_0 + \beta_1 x + \beta_2 w + E(u|X=x, W=w)] \\
&\quad = \beta_1 \Delta x + [E(u|X=x+\Delta x, W=w) - E(u|X=x, W=w)] = \beta_1 \Delta x
\end{aligned}
$$

where the final line follows from conditional mean independence:

$$E(u|X=x+\Delta x, W=w) = E(u|X=x, W=w) = E(u|W=w).$$

The math of conditional mean independence (2 of 5)

Under conditional mean independence:

2. $\hat{\beta}_1$ is unbiased

3. $\hat{\beta}_2$ does not in general estimate a causal effect

The math: Consider the regression model,

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

where u satisfies the conditional mean independence assumption, and where β_1 and β_1 are causal effects.

Suppose that $E(u|W) = \gamma_0 + \gamma_2 W$ (that is, that $E(u|W)$ is linear in W). Then, under conditional mean independence,

The math of conditional mean independence (3 of 5)

$$E(u|X,W)=E(u|W)=\gamma_0+\gamma_2 W. \qquad (*)$$

Let

$$v=u-E(u|X,W) \qquad (**)$$

so that $E(v|X,W)=0$. Combining (*) and (**) yields,

$$\begin{aligned} u &= E(u|X,W)+v \\ &= \gamma_0+\gamma_2 W+v, \text{ where } E(v|X,W)=0 \end{aligned} \qquad (***)$$

Now substitute (***) into the regression,

$$Y=\beta_0+\beta_1 X+\beta_2 W+u \qquad (+)$$

The math of conditional mean independence (4 of 5)

So that

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \beta_2 W + u && (+)\\
&= \beta_0 + \beta_1 X + \beta_2 W + \gamma_0 + \gamma_2 W + v \ \text{ from (***)}\\
&= (\beta_0 + \gamma_0) + \beta_1 X + (\beta_2 + \gamma_2) W + v\\
&= \delta_0 + \beta_1 X + \delta_2 W + v, \text{ where } \delta_0 = \beta_0 + \gamma_0 \text{ and } \delta_2 = \beta_2 + \gamma_2 && (++)
\end{aligned}
$$

- Because $E(v|X, W) = 0$ in equation (++), the OLS estimators of δ_0, β_1, and δ_2 in (++) are unbiased.
- Because the regressors in (+) and (++) are the same, the OLS coefficients in regression (+) satisfy, $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2$. Thus $\hat{\beta}_1$ is an unbiased estimator of the causal effect β_1, but $\hat{\beta}_2$ is not unbiased for β_2.

The math of conditional mean independence (5 of 5)

Under conditional mean independence,

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

and

$$E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2 \neq \beta_2$$

In summary, if W is such that conditional mean independence is satisfied, then:

- The OLS estimator of the effect of interest, $\hat{\beta}_1$, is unbiased.
- The OLS estimator of the coefficient on the control variable, $\hat{\beta}_2$, does not have a causal interpretation. The reason is that the control variable is correlated with omitted variables in the error term, so that $\hat{\beta}_2$ is subject to omitted variable bias.

Next topic: hypothesis tests and confidence intervals…