Econ 140 - Spring 2016
Section 6

GSI: Fenella Carpena

March 3, 2016

1 Omitted Variable Bias

Exercise 1.1. (Stock & Watson, Review the Concepts, Exercise 6.1) A researcher is interested in the effect on test scores of computer usage. Using school district data like that used in this chapter, she regresses district average test scores on the number of computers per student. Will $\widehat{\beta}_1$ be an unbiased estimator of the effect on test scores of increasing the number of computers per student? Why or why not? If you think $\widehat{\beta}_1$ is biased, is it biased up or down? Why?

The regression given is $testscore = \beta_0 + \beta_1 comp + u$, where *comp* is the number of computers per student. We are asked whether $\widehat{\beta}_1$ is unbiased. In this case, $\widehat{\beta}_1$ likely suffers from omitted variable bias (OVB), since there are other factors (apart from computers per students) that affect test scores and are correlated with computers per student. One example is the teacher-student-ratio (call it TSR for short). Note that TSR satisfies the two conditions for OVB: (1) computers per student and TSR are likely positively correlated; this is because schools that spend more on computers probably are wealthy schools that also hire more teachers, so they also have more teachers per student; (2) TSR is a determinant of test score outcomes, since if there are more teachers per student, there are more resources that a student can rely on to get better grades.

We are also asked if $\widehat{\beta}_1$ is biased up or down. Considering TSR as the omitted variable, it will be biased upward. Why? The sign of the bias is determined by the sign of β_2, which is the coefficient on the omitted variable (here, TSR), multiplied by the sign of $cov(TSR, comp)$. Note that: (1) it is likely that TSR has a positive effect on test scores, so $\beta_2 > 0$, and (2) it is likely that $cov(TSR, comp) > 0$, for the same reason mentioned in the previous paragraph. Taking these two points together, and because multiplying two positives is still positive, we have upward bias.

2 Multiple Regression Model

Exercise 2.1. (Stock & Watson, Review the Concepts, Exercise 6.2) A multiple regression includes two regressors: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$.

(a) What is the expected change in Y if X_1 increases by 3 units and X_2 is unchanged?
$3\beta_1$

(b) What is the expected change in Y if X_2 decreases by 5 units and X_1 is unchanged?
$-5\beta_2$

(c) What is the expected change in Y if X_1 increases by 3 units and X_2 decreases by 5 units?
$3\beta_1 - 5\beta_2$

Exercise 2.2. (Stock & Watson, Exercise 6.5) Data were collected from a random sample of 220 home sales from a community in 2013. Let $Price$ denote the selling price (in \$1000), BDR denote the number of bedrooms, $Bath$ denote the number of bathrooms, $Hsize$ denote the size of the house (in square feet), $Lsize$ denote the lot size (in square feet), Age denote the age of the house (in years), and $Poor$ denote a binary variable that is equal to 1 if the condition of the house is reported as "poor." An estimated regression yields

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize + 0.090Age - 48.8Poor$$

(a) Suppose that a homeowner converts part of an existing family room in her house into a new bathroom. What is the expected increase in the value of the house?
23.4*1 = 23.4. Note that price is measured in thousands of dollars, so the expected increase is 23,400.

(b) Suppose that a homeowner adds a new bathroom to her house, which increases the size of the house by 100 square feet. What is the expected increase in the value of the house?
23.4*1 + 0.156*100 = 39. So the expected increase is 39,000.

(c) What is the loss in value if a homeowner lets his house run down so that its condition becomes "poor"?
-48.8*1 = -48.8. So the expected loss in value is 48,800

3 $\overline{R}^2$ (Adjusted R^2)

Exercise 3.1. (Stock & Watson, Review the Concepts, Exercise 6.3) How does $\overline{R}^2$ differ from R^2? Why is $\overline{R}^2$ useful in a regression model with multiple regressors?

One of the "problems" with R^2 as a measure of fit is that it weakly increases (mechanically) as more regressors are added. So a model with more terms will appear to have a better fit simply because it has more terms, even though those terms offer no explanatory power for Y.

$\overline{R}^2$ is different from R^2 because $\overline{R}^2$ takes into account the number of explanatory variables in the regression. $\overline{R}^2$ is useful because it gives us a better sense of the "goodness of fit," whereas R^2 weakly increases (mechanically) when adding a regressor.

4 Perfect and Imperfect Multicollinearity

Exercise 4.1. (Stock & Watson, Review the Concepts, Exercise 6.4) Explain why two perfectly multicollinear regressors cannot be included in a linear multiple regression. Give two examples of a pair of perfectly multicollinear regressors.

If two regressors X_1 and X_2 are perfectly multicollinear, this means that one regressor can be written as a linear function of the other, e.g., we can write $X_1 = a + bX_2$ for some constants a and b. In this case, both X_1 and X_2 cannot be included in the regression because we will not be able to estimate the OLS coefficients on X_1 and X_2 respectively. Specifically, the regression cannot determine the effect of a change in X_1 holding X_2 constant, precisely because X_1 is a linear function of X_2. If X_2 is held constant, then so is X_1. Hence, there will be no variation in X_1.

Two examples of perfectly multicollinear regressors:

- Consider the regression $wage = \beta_0 + \beta_1 female + \beta_2 male + u$, where $female$ and $male$ are dummy variables for women and men, respectively. In this regression, $female$ and $male$ are perfectly multicollinear, because $female + male = 1$. This case of putting both the $female$ and $male$ variables in the same regression is an example of the **dummy variable trap.**
- Consider the regression of a student's college GPA on his/her gender, e.g. $GPA = \beta_0 + \beta_1 female + u$, but the data contains students from all-female (i.e. not co-ed) schools. Since all the students in the data are

female, we know that the $female$ variable is constant, taking on the value 1 for all observations. Hence, the constant term β_0 and $female$ will be perfectly multicollinear (since they are both constants).

Exercise 4.2 (Stock & Watson, Review the Concepts, Exericse 6.5) Explain why it is difficult to estimate precisely the partial effect of X_1 holding X_2 constant, if X_1 and X_2 are highly correlated.

If X_1 and X_2 are highly correlated, this means that X_1 and X_2 are imperfectly multicollinear. Hence, most of the variation in X_1 coincides with the variation in X_2. Let's consider the regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$. How do we interpret β_1? A 1 unit change in X_1 is associated with a β_1 change in Y, *holding X_2 constant.* So, if we hold X_2 constant, because X_1 and X_2 are highly correlated, there is not enough variation in X_1 that can be used to estimate the effect of X_1 on Y. This leads to imprecise estimates of our OLS coefficients.

Another way to think about this is that if we hold X_2 constant, it is "as if" we are holding X_1 constant as well (again, because X_1 and X_2 are highly correlated). Thus, the variation in X_1 will be low. Recall that $SE(\widehat{\beta}_1)$ is inversely related to the variance of X_1, so low variation in X_1 means that $SE(\widehat{\beta}_1)$ is large, which means that our estimate of $\widehat{\beta}_1$ is imprecise.

5 Additional Exercises

Question 1. Consider the following regression model to explain city crime rates ($crimerate$) in terms of the probability of conviction ($prbconv$) and average sentence length ($avgsen$).

$$crimerate_i = \beta_0 + \beta_1 * prbconv_i + \beta_2 * avgsen_i + u_i.$$

What are some factors contained in u_i? Do you think that the first least squares assumption in this model is likely to hold?

Factors contained in u_i are those that affect crime rate but are not captured in $prbconv$ and $avgsen$. One example might be the size of the police force. The first least squares assumption is unlikely to hold since the size of the police force is probably correlated with $prbconv$: cities that have a larger police force might be the ones that really care about crime prevention and public safety, and if so, they likely also care about enforcement so that the probability of conviction would also be higher. Another possible factor that is contained in u is wealth of the city.

Wealth levels likely affect crime rates since in poor cities, there are fewer job opportunities so some residents might resort to theft, etc. At the same time, poor cities likely also have fewer resources for fighting crime and enforcing laws, so that the probability of conviction might also be lower. This would violate the first least squares assumption since u (which contains wealth) is correlated with $prbconv$.

Question 2. Consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Does a higher R^2 mean that it is more likely that X_i causes Y_i? What about a higher $\overline{R}^2$?

No (for both R^2 and $\bar{R}^2$). R^2 and $\overline{R}^2$ are measures of fit of a regression and they do not say anything about causality.

Question 3. Suppose we want to estimate the effect campaign spending on campaign outcomes. Let $voteA_i$ denote the percent of the vote for candidate A in county i, let $expendA_i$ denote the campaign expenditures for candidate A, let $expendB_i$ denote the campaign expenditures for candidate B. Let $totexp_i$ be the total expenditures across both candidates, and let $shareA_i$ denote the percentage of total campaign expenditures made by candidate A (in other words $shareA_i = 100 \cdot (expendA_i/totexp_i)$. Consider the regression model

$$voteA_i = \beta_0 + \beta_1 expendA_i + \beta_2 expendB_i + \beta_3 shareA_i + u_i.$$

Does this regression violate the fourth least squares assumption? Explain why or why not.

No. None of the variables in the regression have a linear relationship with each other. In particular, note that while $shareA$ is a function of both $expendA$ and $expendB$, it is not linear in those variables since $shareA = 100 \cdot \frac{expendA}{totexpend} = 100 \cdot \frac{expendA}{expendA+expendB}$. Also note that while $expendA_i + expendB_i = totexp_i$, $totexp_i$ (which varies across cities i) is not included in the regression, so $expendA$ and $expendB$ are not perfectly collinear in the regression. However, if $totexp_i$ was constant across cities, then the regression would violate the fourth least squares assumption.

Question 4. Which of the following can cause OLS estimators to be biased? Select all that apply.
(a) Heteroskedasticity
(b) Omitting an important variable
(c) X_1 and X_2 are both included in the regression, but their sample correlation is very close to 1

Only B. A and C are not related to unbiasedness of the estimators. It is possible for the estimator to be unbiased regardless of the whether the errors are homoskedastic or heteroskedastic. As for C, this situation only affects the precision (i.e., standard errors) of our estimates, but not the bias.

Question 5. In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours the spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

(a) In the model $GPA_i = \beta_0 + \beta_1 study_i + \beta_2 sleep_i + \beta_3 work_i + \beta_4 leisure_i + u_i$, does it make sense to hold $sleep$, $work$ and $leisure$ fixed, while changing $study$?
No because all four activities sum to 168. So it is not possible to increase one variable while holding others constant.

(b) Explain why this model violates the fourth least squares assumption.
Note that $study + sleep + work + leisure = 168$ so these variables exhibit perfect multicollinearity.

(c) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies the fourth least squares assumption? Provide an interpretation of one of the parameters in your proposed model.
We can omit one of the categories, e.g. $GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + u_i$. In this case, the interpretation of β_1 is that it is the effect of 1 more hour of study on GPA, holding $sleep$ and $work$ constant. If we are holding $sleep$ and $work$ constant, then $leisure$ will necessarily fall by 1 hour. So β_1 represents the effect of on GPA of spending 1 less hour on $leisure$ and 1 more hour on $studying$. Note that I have omitted $leisure$ in the regression, but it is also possible to omit any one of the other categories.