ØAMET4100 · Spring 2019
Worksheet 5

Instructor: Fenella Carpena

February 7, 2019

1 Internal Validity

Threat to Internal Validity	**Examples/Cases**	**Implications for OLS Estimates**	**Possible Solutions**
OVB	Example: $wages_i = \beta_0 + \beta_1 educ_i + u_i$ where $educ_i$ is educational attainment likely suffers from OVB (what are possible omitted variables here?)		
Functional Form Misspecification	Example: True population regression function is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$, but the regression we run is $Y_i = \beta_0 + \beta_1 X_i + u_i$.		
Measurement Error	Cases: (A) Measurement error in X, (B) Measurement error in Y		

Continued on next page

Continued from previous page

Threat to Internal Validity	**Examples/Cases**	**Implications for OLS Estimates**	**Possible Solutions**
Missing Data and Sample Selection	Cases: (A) Data is missing at random, (B) Missing data is selected based on X, (C) Missing data is selected based on Y.		
Simultaneity	There is a two-way relationship between the independent variable (X_i) and dependent variable (Y_i) in the regression model: $Y_i = \beta_0 + \beta_1 X_i + u_i$ and $X_i = \gamma_0 + \gamma_1 Y_i + v_i$.		
Heteroskedasticity	Problem arises when the regression error is heteroskedastic, but SEs were calculated under homoskedasticity		
Correlation of the error term across observations	Example: when the data are repeated observations of the same entity over time (e.g., panel or time series data)		

2 External Validity

Threat to External Validity	Examples/Cases	Implications for OLS Estimates	Possible Solutions
Differences in population	Example: Lab studies on the toxic effects of chemicals are conducted using animal populations like mice, but the results are used to write health/safety regulations for humans.		
Differences in settings	Example: Examining the effect of student-teacher-ratio on test scores among elementary schools in California, but the results may not apply to elementary schools in Massachusetts.		

3 Exercises

Exercise 3.1 Stock and Watson, Review the Concepts 9.1 Is it possible for an econometric study to have internal validity but not external validity?

Exercise 3.2 Stock and Watson, Exercise 6.6 A researcher plans to study the causal effect of a strong legal system on the economy using data from a sample of countries. The researcher plans to regress national income per capita on whether the country has a strong legal system or not (an indicator variable taking the value 1 or 0, based on expert opinion).

(a) Do you think this regression suffers from omitted variable bias? Which variable would you add to the regression?

(b) Assess whether the regression will likely over- or under-estimate the effect of a strong legal system on income per capita, based on the variable you think is omitted.

Exercise 3.3 Stock and Watson, Review the Concepts 9.2 What is the effect of measurement error in Y? How is this different from the effect of measurement error in X?

Exercise 3.4 Suppose that a state offered voluntary standardized tests to all its third graders and that these data were used in a study of class size on student performance. Explain how sample selection bias might invalidate the results.

Exercise 3.5 Stock and Watson, Exercise 9.3 Labor economists studying the determinants of women's earnings discovered a puzzling empirical result. Using randomly selected employed women, they regressed earnings on the women's number of children and a set of control variables (age, education, occupation, and so forth). They found that women with more children had higher wages, controlling for these other factors. Explain how sample selection might be the cause of this result. (Hint: Notice that women who do not work outside the home are missing from the sample.) [This empirical puzzle motivated James Heckman's research on sample selection that led to his 2000 Nobel Prize in Economics. See Heckman (1974).]

Exercise 3.6 Stock and Watson, Review the Concepts 9.6 A researcher estimates the effect on crime rates of spending on police by using city-level data. Explain how simultaneous causality might invalidate the results.

Exercise 3.7 In microeconomics, you studied the demand and supply of goods in a single market, let the demand (Q_i^D) and supply (Q_i^S) for the i-th good be determined as follows:

$$Q_i^D = \beta_0 + \beta_1 P_i + u_i$$

$$Q_i^S = \gamma_0 + \gamma_1 P_i + v_i$$

where P is the price of the good. In addition, assume that the market clears. Explain how the simultaneous causality bias applies in this situation.

Exercise 3.8 A researcher estimates a regression using two different software packages. The first uses the homoskedasticity-only formula for standard errors. The second uses the heterosekdasticity-robust formula. The standard errors are very different. Which should the researcher use? Why?

Exercise 3.9 Stock and Watson, Exercise 9.1 Suppose that you read a careful statistical study of the effect of improved health of children on their test scores at school. Using data from a project in a West African district, in 2000, the study concluded that students who received multivitamin supplements performed substantially better at school. Use the concept of external validity to determine if these results are likely to apply to India in 2000, the United Kingdom in 2000, and West Africa in 2015.

Exercise 3.10 Stock and Watson, Exercise 9.6) Suppose that $n = 100$ i.i.d. observations for (Y_i, X_i) yield the following regression results:

$$\hat{Y} = \underset{(15.1)}{32.1} + \underset{(12.2)}{66.8} \cdot X,\ SER = 15.1,\ R^2 = 0.81$$

Another researcher is interested in the same regression, but he makes an error when he enters the data into his regression program: He enters each observation twice, so he has 200 observations (with observation 1 entered twice, observation 2 entered twice, and so forth).

(a) Using these 200 observations, what results will be produced by his regression program? (Hint: Write the "incorrect" values of the sample means, variances, and covariance of Y and X as functions of the "correct" values. Use these to determine the regression statistics.)

$$\hat{Y} = \_\_\_\_\_ + \_\_\_\_\_ \cdot X,\ SER = \_\_\_\_\_,\ R^2 = \_\_\_\_\_$$

(b) Which (if any) of the internal validity assumptions are violated?