Econ 140 - Spring 2016
Section 9

GSI: Fenella Carpena

March 31, 2016

1 Assessing Studies Based on Multiple Regression

1.1 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 only		
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.		

Continued on next page

Continued from previous page

Threat to Internal Validity	**Examples/Cases**	**Implications for OLS Estimates**	**Possible Solutions**
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	Occurs 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)		

1.2 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.		

1.3 Review the Concepts Exercises, Chapter 9

1. What is the difference between external and internal validity? Between the population studied and the population of interest?

2. Key Concept 9.2 describes the problem of variable selection in terms of a trade-off between bias and variance. What is this trade-off? Why could including an additional regressor decrease bias? Increase variance?

3. Economic variables are often measured with error. Does this mean that regression analysis is unreliable? Explain.

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.

5. 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.

6. 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?

2 Panel Data

2.1 Notation: Quick Explanation

Example 2.1.1. Suppose you have the following data set. Fill in the columns for i, t, DItaly, DYear2001 in the following table, where DItaly is a dummy variable equal to 1 if the country is Italy, and DYear2001 is a dummy variable equal to 1 if the year is 2001. What are n and T in this data?

i	t	Country	Year	GDP	Interest rate	Europe	DItaly	DYear2001
		Brazil	2000	100	22	0		
		Brazil	2001	150	18	0		
		Brazil	2002	170	15	0		
		Brazil	2003	220	12	0		
		USA	2000	1000	4	0		
		USA	2001	1150	3	0		
		USA	2002	1400	2	0		
		USA	2003	1900	2	0		
		Italy	2000	60	6	1		
		Italy	2001	130	5	1		
		Italy	2002	150	5	1		
		Italy	2003	190	3	1		

2.2 Regression with Panel Data

Example 2.2.1. Some states in the US have enacted laws that allow citizens to carry concealed weapons. Proponents of argue that if more people carry concealed weapons, crime will decline because criminals will be deterred from attacking other people. Opponents argue that crime will increase because of accidental or spontaneous use of weapons. As a researcher, you would like to examine the effect of concealed weapons laws on violent crime. Suppose that you have a data on 50 US states for the years 1977 to 1999. The data contains violent crime rate (the variable name is *vio*) for the 50 states across these years, and a variable called *guns* which is dummy variable equal to 1 if the state has a law in effect that year which allows citizens to cary concealed weapons.

Type # 1: Regression with Only Entity Fixed Effects. For a regression with only entity fixed effects, we write the population regression as follows

$$vio_{it} = \beta_1 guns_{it} + \alpha_i + u_{it}$$

where, as we've discussed in the previous section of these notes, i denotes the entity, and t denotes time.

(a) What is α_i?

(b) What does α_i capture?

(c) In practice, how do we estimate the above regression?

Type # 2: Regression with Only Time Fixed Effects. For a regression with only time fixed effects, we write the population regression as follows

$$vio_{it} = \beta_1 guns_{it} + \lambda_t + u_{it}$$

where, as we've discussed in the previous section of these notes, i denotes the entity, and t denotes time.

(a) What is λ_t?

(b) What does λ_t capture?

(c) In practice, how do we estimate the above regression?

Type # 3: Regression with Both Entity and Time Fixed Effects. For a regression with both entity and time fixed effects, we write the population regression as follows

$$vio_{it} = \beta_1 guns_{it} + \alpha_i + \lambda_t + u_{it}$$

where, as we've discussed in the previous section of these notes, i denotes the entity, and t denotes time.

(a) What are state and time fixed effects able to control for? What are they **not** able to control for?

(b) What are possible sources of OVB in this regression?

(c) What does serial correlation mean and how does it arise in the above regression? Is serial correlation a threat to internal validity? If so, explain how we can address this issue.

2.3 Review the Concepts Exercises, Chapter 10

1. Why is it necessary to use two subscripts, i and t, to describe panel data? What does i refer to? What does t refer to?

2. A researcher is using a panel data set on $n = 1000$ workers over $T = 10$ years (from 2001 to 2010) that contains the workers' earnings, gender, education, and age. Assume that the education variable is changing over time for at least some individuals in the sample (for example, because some individuals return to school). The researcher is interested in the effect of education on earnings. Give some examples of unobserved person-specific variable that are correlated with both education and earnings. Can you think of examples of time-specific variables that might be correlated with education and earnings? How would you control for these person-specific and time-specific effects in a panel data regression?

3. Can the regression that you suggested in response to Question 2 above be used to estimate the effect of gender on an individual's earnings? Can that regression be used to estimate the effect of the national unemployment rate on an individual's earnings? Explain.

4. In the context of the regression you suggested in Question 2, explain why the regression error for a given individual might be serially correlated.

3 Midterm 2 Practice Problems/Review

MT2, Fall 2013, Question 3. You are someone who loves wine seeking to use your knowledge of econometrics to understand how wine prices are determined. You hypothesize that the current market price of fancy wine is affected by the weather during the growing season just before the grapes are harvested. Accordingly, you obtain data on bottles of wine harvested between 1970 and 2000 and sold in auction *this year*, and hope to estimate the following equation:

$$Price_i = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + \beta_3 RainfallHarv_i + \beta_5 Vintage_i + u_i,$$

where $i = 1, \ldots, n$ indexes bottles of wine sold this year and the variables are defined as follows:

Variable	Description
Age	The age of the bottle of wine (this year minus harvest year)
$Temp$	The average temperature during the growing season in the area where the grapes were grown
$RainfallHarv$	The amount of rainfall that August through September (when the grapes were harvested)
$Vintage$	The year that the grapes were harvested

(a) Are any of the OLS assumptions going to be violated with certainty? If so, which one? Explain.

(b) Unfortunately, the rainfall data you want were not collected in the 1970s for some of the vineyards. You suspect that vineyards producing low-end wines were the last to start collecting weather data. You must decide whether to: (i) run the full regression on a subset of the wines for which rainfall data is available even for the 1970s, or (ii) run a regression without the rainfall variables, but using all of the wines in your sample. Briefly discuss the trade-offs to adopting strategy (i) versus strategy (ii) [Hint: you may want to list the cost(s) and benefit(s) of each strategy.]

(c) You decide to estimate the following, shorter regression on your entire sample:

$$Price_i = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + u_i,$$

but you would like to know by what *percent* the price changes when age changes by a year, all else equal. Select the statement that is *most* correct:

(i) You should regress $Price_i = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + u_i$.

(ii) You should regress $\log(Price_i) = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + u_i$.

(iii) You should regress $Price_i = \beta_0 + \beta_1 \log(Age_i) + \beta_2 Temp_i + u_i$.

(iv) You should regress $\log(Price_i) = \beta_0 + \beta_1 \log(Age_i) + \beta_2 Temp_i + u_i$.

(v) You should run a different regression that is not specified in another option.

(d) Discuss the external validity of your study with respect to California wines.

Final Exam, Fall 2014, Question 4. You are interested in learning how the value of cars evolves as they grow older. You obtain cross sectional data on a large number of cars, and for each vehicle, i, you observe the dollar value (Y_i), the age in years (X_{1i}) and the mileage in miles (X_{2i}). You also observe the make and model of each vehicle, (e.g. "Ford Focus"), and you generate a set of dummy variables, $D_i^1, \ldots, D_i^m$, such that D_i^1 equals 1 if vehicle i is of make and model #1 (e.g. "Ford Focus") and 0 otherwise, and similarly for all makes and models $1, \ldots, m$.

(a) You estimate that $\hat{Y}_i = 30,000 - 3,000 \cdot X_{1i} + 80 \cdot X_{1i}^2$. Based on the estimate, what is the average marginal effect of vehicle age on dollar value when a car is brand new? When it is 10 years old?

(b) Next, you run the regression $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{2i} + u_i$. Interpret the coefficient β_3.

(c) Next, you consider the regression

$$Y_i = \sum_{j=1}^{m} \alpha_j D_i^j + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{2i} + u_i. \tag{1}$$

What is the interpretation of the coefficient α_1?

(d) Assume that for each make and model you observe a large number of vehicles. Write a specification such that the value of cars can evolve *differently* with age for each make and model. Using the regression equation you specified what will be the value of a 5 year old car with 60,000 miles be if it is of make and model #1? And if it is of make and model #2? State your answers as functions of the regression coefficients.

(e) Suppose you obtain panel data, so that you observe each vehicle, its age and its mileage at multiple times, $t = 1, \ldots, T$. With panel data, regression (1) becomes

$$Y_{it} = \sum_{j=1}^{m} \alpha_j D_i^j + \beta_1 X_{1it} + \beta_2 X_{1it}^2 + \beta_3 X_{2it} + u_{it}, \tag{2}$$

where t indexes the time of observation. If you add *vehicle fixed effects* to regression (2), can you still include mileage (X_{2it}) as a regressor? Can you still include $D_i^1, \ldots, D_i^m$ as regressors? Why or why not?