Live Session 6: Difference in difference

Devesh Tiwari

June 15, 2015

Introduction

Recall the five steps in model building:

(1) Question Formation

(2) Exploratory Data Analysis

(3) Model Execution and Evaluation

(4) Model Diagnosics

(5) Model Scrutiny

We have spent a great deal of time on steps 1 - 4 and we now move on to step 5: Model Scrutniy. At this stage of the process, we examine the results from one or more of our models and critically evaluate them with respect to the question we posed at the beginning of the process. At this step, we ask: To what extent have we satisfied the CLM assumptions? Have we been able to answer the question at hand? Are there any weaknesses to our analysis or data that might shed doubt on the veracity of the answer? Do we have any solutions in mind?

Once you have exhausted your modeling energies, you will likely find that your model has visible weaknesses. A very common weakness lies in our ability to draw an unbiased estimate between two or more variables and in being able to draw a causal estimate. Sometimes, the solutions to these problems require us to gather more data, while at other times it requires us to adopt different empirical (or identification) strategies. Two such strategies are difference in difference estimation and instrumental variable analysis.

Breakout Session: Motivating Questions

(1) Consider the following population model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where you only have the neccessary information to conduct the following regresssion:

$$Y = \alpha_0 + \alpha_1 + \mu$$

Under what circumstances would the OLS estimator for α_1 be unbiased? Biased? How do you know?

(2) Suppose you were deciding between two models: One with nearly every relevant variable you could think of included in the model and a shorter model with fewer variables. Compare and contrast these two models in light of the CLM assumptions we have discussed thus far. Please state which assumptions have been violated and how you would spot them during the model building process.

(3) Returning to a situation where you have a regression that manages to satisfy all of the CLM assumptions. What does this model tell you? Can you claim you have identified an unbiased relationship between Y and X? Can you claim you have identified a causal relationship?

Data Exercise

The file crime3.Rdata contains a two-year panel of crime data taken from E. Eide (1994) Economics of Crime: Deterence of the Rational Offender. Amsterdam: North Holland. The data are for 53 police districts in Norway and include a variable for the clear-up rate, or fraction of recorded crimes solved by the police. The clear-up rate can be interpreted as measuring the probability that a criminal is captured, which is different than the severity of punishment if the capture takes place. You will use a difference-in-difference analysis to estimate the causal effect that the clear-up rate has on the amount of crime. The differenced model you will estimate is:

$$clcrime_i = \beta_0 + \beta_1 cclrprc1_i + \beta_2 cclrprc2_i + \epsilon_i$$

Here, the dependent variable is the change in the log of number of crimes, while the predictors are change in the clear-up rate, lagged by 1 year and by 2 years.

Breakout Session

Your task is to explore the relationship between clearance rate and crime rate using these data. One model should use an undifferenced approach and another model should use a difference in difference model. For the time being, feel free to ignore the EDA but do conduct some post regression diagnostics.

a) Compare and contrast the relationship between clearance rates and crime. Do you get a different answer depending on which model you chose?

b) For concreteness, list at least two potential threats, that either exist in the data or the data generation process, that might impair your ability to estimate a causal and unbiased effect from the undifferenced model?

c) What does the difference in difference model do to the omitted variable problem that exists in the undifferenced model?

d) What assumptions are needed for this model to be causal? Do you feel confident that the difference in difference model is producing an unbiased causal estimate?

e) A politician reads your report and decides to make substantial investments into police departments in order to increase the clearance rate of crimes. She argues that this investment will cause crime to drop. Write down an empirical model that you can use to test his claim. Under what circumstances would her causal claim be valid? Invalid?