Applied Regression and Time Series Analysis (2016 Fall): HW5 - Week 7

Jeffrey Yau and Devesh Tiwari

October 5, 2016

Instructions:

The weekly assignment serves two purposes: (1) Review concepts, techniques, theories, statistical models covered during the week. (2) Extend the materials taught in the asynchronized lectures, assigned readings, and live sessions; some new concepts and/or techniques are introduced in the weekly assignment.

Below are specific instructions:

- **Due: 10/23/2016 (11:59pm PST)**
- You may complete this assignment on your own or in a group of no more than 3 students.
- When working in a group, you are strongly encouraged to complete the assignment on your own before discussing your group mates. Do not use the "division-of-labor" approach to complete the assignment.
- The homework is designed as a quantitative analysis. The instructions and questions are designed to guide you through the analysis of data using regression techniques. As such, you should think of it as a quantitative case study and the result of the study is a report with a set of well-written codes that can be used to reproduce the results in the report.
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. R-script or R markdown file
    2. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    - **SectionNumber_hw02_LastNameFirstInitial.fileExtension**
    - Examples:
      - Section1_hw02_YauJ.Rmd
      - Section1_hw02_YauJ.pdf
      - Section1_hw02_TiwariD_YauJ.Rmd
      - Section1_hw02_TiwariD_YauJ.pdf

Objective:

The key objective of this homework is to practice the use of the difference-in-difference technique to handle potential bias arising from omitted variables.

Description:

The data set in VOUCHER.DTA can be used to estimate the effect of school choice on academic achievement. Attendance at a choice school was paid for by a voucher, which was determined by a lottery among those who applied. The data subset was chosen so that any student in the sample has a valid *1994* math test score. Unfortunately, many students have missing test scores, possibly due to attrition (that is, leaving the Milwaukee public school district). These data include students who applied to the voucher program and were accepted, students who applied and were not accepted, and students who did not apply. Therefore, even though the vouchers were chosen by lottery among those who applied, we do not necessarily have a random sample from a population where being selected for a voucher has been randomly determined. (An important consideration is that students who never applied to the program may be systematically different from those who did, and in ways that we cannot know based on the data.)

This exericse asks you to do a cross-sectional analysis where winning the lottery for a voucher acts as an *instrumental variable* for attending a choice school. Actually, because we have multiple years of data on each student, we construct two variables. The first, *choiceyrs*, is the number of years from 1991 to 1994 that a student attended a choice school; this variable ranges from zero to four. The variable *selectyrs* indicates the number of years a student was selected for a voucher. If the student applied for the program in 1990 and received a voucher then selectyrs = 4; if he or she applied in 1991 and received a voucher then selectyrs = 3; and so on. The outcome of interest is mnce, the student's percentile score on a math test administered in 1994.

Question 1: Of the 990 students in the sample, how many were never awarded a voucher? How many had a voucher available for four years? How many students actually attended a choice school for four years?

Question 2: Run a simple regression of *choiceyrs* on *selectyrs*. Are these variables related in the direction you expected? How strong is the relationship? Is selectyrs a sensible IV candidate for *choiceyrs*?

Question 3 Run a simple regression of mnce on *choiceyrs*. What do you find? Is this what you expected? What happens if you add the variables black, hispanic, and female?

Question 4 Why might choiceyrs be endogenous in an equation such as

$$mnce = \beta_0 + \beta_1 choiceyrs + \beta_2 black + \beta_3 hispanic + \beta_4 female + u_1?$$

Question 5 Estimate the equation in question 4 by instrumental variables, using *selectyrs* as the IV for choiceyrs. Does using IV produce a positive effect of attending a choice school? What do you make of the coefficients on the other explanatory vairables?

Question 6 To control for the possibility that prior achievement affects participating in the lottery (as well as predicting attrition), add *mnce90* — the math score in 1990 — to the equation in Question 4. Estimate the equation by OLS and IV, and compare the results for β_1. For the IV estimate, how much is each year in a choice school worth on the math percentile score? Is this a practically large effect?

Question 7 The variables choiceyrs1, choiceyrs2, and so on are dummy variables indicating the different number of years a student could have been in a choice school (from 1991 to 1994). The dummy variables selectyrs1, selectyrs2, and so on have a similar definition, but for being selected from the lottery. Estimate the equation

$$mnce = \beta_0 + \beta_1 choiceyrs1 + \beta_2 choiceyrs2 + \beta_3 choiceyrs3 + \beta_4 choiceyrs4 + \beta_5 black + \beta_6 hispanic + \beta_7 female + \epsilon?$$

by IV, using as instruments the *four selectyrs dummy variables*. Describe your findings. Do they make sense?