Applied Regression and Time Series Analysis (2016 Fall): HW4 - Week 6

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: 10/16/2016 (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 of the Data

The file *athletics.RData* contains a two-year panel of data on 59 universities. Some variables relate to admissions, while others related to atheletic performance. You will use this dataset to investigate whether athletic success causes more students to apply to a university.

This data was made available by Wooldridge, and collected by Patrick Tulloch, then an economics student at MSU. It may have been further modified to test your proficiency. Sources are as follows:

Peterson's Guide to Four Year Colleges, 1994 and 1995 (24th and 25th editions). Princeton University Press. Princeton, NJ.

The Official 1995 College Basketball Records Book, 1994, NCAA.

1995 Information Please Sports Almanac (6th edition). Houghton Mifflin. New York, NY.

Question 1:

Conduct a quick examination and EDA of the dataset.

Examine the variables of interest: *apps* represents the number of applications for admission. *bowl, btitle*, and *finfour* are indicators of athletic success. The three athletic performance variables are all lagged by one year. Intuitively, this is because we expect a school's athletic success in the previous year to affect how many applications it recieves in the current year.

Question 2:

Note that the dataset is in long format, with a separate row for each year for each school. To prepare for a difference-in-difference analysis, transfer the dataset to wide-format. Each school should have a single row of data, with separate variables for *1992* and *1993*. For example, you should have an apps.1992 variable and an *apps.1993* variable to record the number of applications in either year.

Create a new variable, *clapps* to represent *the change in the log of the number of applications from 1992 to 1993*. Examine this variable and its distribution.

Which schools had the greatest increase and the greatest decrease in number of log applications?

Question 3 Similarly to above, create three variables, *cbowl*, *cbtitle*, and *cfinfour*, where each of these variables represents *the changes in the three athletic success variables*. Since these variables are lagged by one year, you are actually computing the change in athletic success from 1991 to 1992.

Question 4 We are interested in a population model,

$$lapps_i = \gamma_0 + \beta_0 I_{1993} + \beta_1 bowl_i + \beta_2 btitle_i + \beta_3 finfour_i + a_i + u_{it}$$

Here, I_{1993} is an indicator variable for the year *1993*. a_i is the time-constant effect of school i. u_{it} is the idiosyncratic effect of school i at time t. The athletic success indicators are all lagged by one year as discussed above.

At this point, we assume that (1) all data points are independent random draws from this population model (2) there is no perfect multicollinearity (3) $E(a_i) = E(u_{it}) = 0$

You will estimate the first-difference equation,

$$clapps_i = \beta_0 + \beta_1 cbowl_i + \beta_2 cbtitle_i + \beta_3 cfinfour_i + a_i + cu_i$$

where $cu_i = u_{i1993} - u_{i1992}$ is the change in the idiosyncratic term from 1992 to 1993.

a) What additional assumption is needed for this population model to be causal? Write this in mathematical notation and also explain it intuitively in English.

b) What additional assumption is needed for OLS to consistently estimate the first-difference model? Write this in mathematical notation and also explain it intuitively in English. Comment on whether this assumption is plausible in this setting.

Question 5 Test the joint signifance of the three indicator variables. This is the test of the overall model. What impact does the result have on your conclusions?