DataSci 271: Homework 4

Paul Laskowski

January 1, 2016

Instructions:

The weekly assignment serves two purposes: (1) Extend the materials taught in the asynchronized materials; some new concepts or techniques are introduced in the weekly assignment. (2) Ensure that you have learned the concepts, techniques, theories, statistical models covered in a specific week. Below are some guidelines:

- Submit 2 files. Missing one of the two files will result in a 50% reduction in grade.
    1. A report (in pdf format) detailing your answers and all the steps to arrive at your answers
    2. A well-documented R-script, jupyter notebook, or Rmd file detailing all of the codes used to arrive at your answers.
- Late submission will not receive any credit.
- All the steps used to arrive at your final answers need to be shown clearly. These steps are as important as the final answer.
- The final answer of each question needs to be very easy identified; the use of bold fonts, highlights, or circling will help.
- This is a group project. Form a group with 3 or 4 people.
- Although this is a group project, we encourage you to attempt all of the exercises before discussing with your teammates. Do not use the "division-of-labor" approach. Each of the students in a group is expected to make sufficient contribution to the lab. If any of your teammate does not make sufficient contribution, please contact your instructor.
- DO NOT copy and paste or even leverage on the solutions we gave to the students in previous semesters. Violation will be reported to the Director of the MIDS program and the Office that oversees UC Berkeley Academic Integrity. In any case, the lab has various subtle changes that make those answers not directly applicable.

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.

Exercises

Question 1:

Examine and summarize the dataset. Note that the actual data is found in the data object, while descriptions can be found in the desc object. How many observations and variables are there?

Examine the variables of key 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, cperf, cbball, and cbowl to represent 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.

Which of these variables has the highest variance?

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 Estimate the first-difference model given above. Using the best practices descibed in class, interpret the slope coefficients and comment on their statistical significance and practical significance.

Question 6 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?