Computer assignment 3a – Econometrics I – due September 19, 8.45am, 2017

THIS IS A GROUP EXERCISE. PLEASE SUBMIT PDF THROUGH BLACKBOARD TIMELY. PART III OF THE ASSIGNMENT ONLY. PDF SHOULD SHOW COMMANDS + OUTPUT + (IF ASKED) EXPLANATION, NO DO-FILES. PLEASE INCLUDE YOUR NAMES.

Goal of the assignment

In this computer assignment, you are asked to test the effect of having to take a resit exam for a course on the grade for the following course in the sequence. Think of Econometrics 1 and Econometrics 2 in the MSc Economics. Eligibility for the treatment, the resit exam, depends on scoring less than 5.5 for Econometrics 1. Let us assume that no other students take the resit exam (for instance, those who score a 7 but want to improve their grade). The data include a number of observable characteristics of the students involved, including gender, age, ethnicity, and special education status. The data are downloadable from Blackboard, look for a file named `ca3a_2017.dta`.[1]

I. Preparing the data file for analysis: creating your do file

(a) Open STATA and then open the STATA do file editor. Give a short description in the first line of the do file, for instance:

```
* Computer Assignment 3a Econometrics, Sep 2017
```

then save the do file under a name or your choosing.

The next line in the do file should tell STATA to open the data from the folder where you saved your data, for instance:

```
use "C:\Users\Nick\Data\ca3a_2017.dta", clear
```

Run this first command of your do file (using the appropriate path) as follows: select the line of code and then click on:  (Execute selection). This loads the data into memory.

II. Descriptive statistics

Before you run any regressions, you should first conduct exploratory data analysis. This gives you the opportunity to check your data for errors, and also to get a feeling for the data (every paper includes a section 'descriptive statistics').

(a) What is the average age of the students in the sample?

(b) Create a histogram of the outcome variable; `grade_ectrics2`.

[1] The assignment is based on simulated data provided by R. Jacob and P. Zhu, 2012, A practical guide to regression discontinuity, MDRC. I modified the data for this assignment.

(c) For how many students do you have grades for both courses? To find this out, run for instance the following command (in STATA ~= stands for 'not equal to' and . stands for 'missing'):

```
summ grade_ectrics2 if grade_ectrics1~=.&grade_ectrics2~=.
```

III. Test of treatment effect

Now you are asked to conduct a test for the presence of a treatment effect.

(a) Produce graphical evidence that the grade for Econometrics 2 is affected by having to take the resit exam for Econometrics 1. As a first step, plot the data around the cutoff point as follows:

```
graph twoway (scatter grade_ectrics2 grade_ectrics1) if
grade_ectrics1>4&grade_ectrics1<7
```

Next, add two linear regression lines, one before and one after the cutoff:

```
graph twoway (scatter grade_ectrics2 grade_ectrics1) (lfit
grade_ectrics2 grade_ectrics1 if grade_ectrics1<5.5) (lfit
grade_ectrics2 grade_ectrics1 if grade_ectrics1>=5.5) if
grade_ectrics1>4&grade_ectrics1<7
```

STATA actually provides a command that can do this for you: `cmogram`. It produces nice plots. If the command `cmogram` is not installed, then type first: `ssc install cmogram` and press enter (if you have trouble installing additional STATA commands on university computers, then consult the footnote on Computer Assignment 2b).

The following shows the average grade for Econometrics 2 by grades for Econometrics 1 (individual observations are actually binned into groups of observations):[2]

```
cmogram grade_ectrics2 grade_ectrics1, scatter lfit line(5.49)
cutpoint(5.49)
```

STATA first spits out all the bins that it created. If STATA displays `[more]`, then simply press any key to continue with the rest of the output. Brief explanation of the options: `scatter` for a scatter plot rather than bar graphs, `lfit` for a linear regression line through the means, `line(5.49)` creates a vertical line at the cutoff, and `cutpoint(5.49)` for the cutoff point.

Whenever you do not fully understand how a STATA command works, then go to Help / STATA Command and simply type the command. You then get a description of the command, including all the options, and at the end some examples of how it is commonly used. Alternatively, in the command line, type the following to get help for `cmogram`, for instance: `help cmogram`

[2] STATA chooses the number of bins on each side of the cutoff for you, you can set that to 10 for instance by including the option: `histopts(bin(10))`.

Is the relationship between grade for Econometrics 1 and Econometrics 2 positive or negative? Why does the observed relationship make sense?

What is the treatment effect in terms of the grade for Econometrics 2? Infer this from the discontinuity at the cutoff point. Derive the percentage change by dividing the absolute change in test scores by the value for the control group at the cutoff.

The plot suggests that assuming a linear relationship between the running variable and the outcome is perfectly fine. At this stage, we limit the bandwidth, i.e. the window around the cutoff. Everything looks linear except for Econometrics 1 grades below 4.5 and above 7. From now on, we will work with this window rather than the whole range of Econometrics 1 grades. This window leaves us plenty of observations around the cutoff point to identify an effect with sufficient statistical power.

(b) Visually check whether the characteristics of students are similar before the cutoff and after it. First, define the treatment variable, as follows:

```
gen treatment=(grade_ectrics1<5.5)
```

The simplest way of comparing the two means for age, for instance, is as follows:

```
graph bar age if grade_ectrics1>=4.5&grade_ectrics1<=7, by(treatment)
```

Also show it for the percentage of children being `white` or not. Obviously, this is not the right falsification test, since the relationship between the running variable and the outcome is not flat, as you discovered under (a). Given the positive relationship, we only need to worry if there is a discontinuity at the cutoff. You can check this by using the `cmogram` command:

```
cmogram white grade_ectrics1 if grade_ectrics1>=4.5&grade_ectrics1<=7,
scatter lfit line(5.49) cutpoint(5.49)
```

Does the graph give you reason to worry? Do this for `age` as well.

(c) Write out the regression equation for estimating the treatment effect in mathematical terms, including all the necessary subscripts. Don't forget the error term. You will be asked to do this on the exam as well. You earn no points if your equation is not exactly right.

Then, estimate the treatment effect on Econometrics 2 grades. Run the following regression:

```
reg grade_ectrics2 treatment grade_ectrics1 if
grade_ectrics1>=4.5&grade_ectrics1<=7
```

Is the estimated coefficient statistically significantly different from zero? Is the estimated effect large or small? Actually, an easy way of getting the treatment effect in percentage terms is to type the following command *right after* running the above regression (this is a 'post-estimation command'):

```
margins, eydx(treatment)
```

(d) Add some covariates to your regression:

```
reg grade_ectrics2 treatment grade_ectrics1 age gender sped if
grade_ectrics1>=4.5&grade_ectrics1<=7
```

Compare the estimated treatment effect from this regression with the estimated treatment effect based on a regression without these covariates (and no missing values for any of the covariates):

```
reg grade_ectrics2 treatment grade_ectrics1 if
age~=.&gender~=.&sped~=.&grade_ectrics1>=4.5&grade_ectrics1<=7
```

Why does it make sense that the two estimated treatment effects are very similar?

(e) As a last step, run a naïve regression of the treatment and the outcome:

```
reg grade_ectrics2 treatment age gender sped if
grade_ectrics1>=4.5&grade_ectrics1<=7
```

Explain the bias in the estimated treatment effect.