UNIVERSITY OF AMSTERDAM

Research Project

Computer ex. 3.1: Controlled comparison

Part 0 – Open Stata, and make your own do-file

- Using windows explorer, make a new folder called `H:\rproject\clab\clab31`.
- Download the datafile `docsim.dta` from **Bb** and put it in this folder.
- Start Stata through the start menu button
- In de white command window type `doedit` to start de do-file editor. Place the Stata screen on the left and the do file editor on the right such that you can easily switch between the two.
- In the first two lines of the do-file type

```
cls                             //this clears the screen
clear all                       //this clears the memory
cd “H:\rproject\clab\clab31”    //this is your path
```

- Save the do-file as `clab31.do`.
- In the following, as always, don’t forget to apply Rules 1&2: paste the command in your `clab31.do` file, press control-s to save and control-d to run.

Part 1 – Controlled comparison by *t*-test per subgroup

In part 1 of these exercises you are going to replicate results from the lecture and estimate the effect of becoming a doctor on income at age 35. To be more precise, we want to test the hypothesis $\mathcal{H}_0\colon E[Y_i(0)] = E[Y_i(1)]$ vs. $\mathcal{H}_1\colon E[Y_i(0)] \neq E[Y_i(1)]$ where $Y_i(0)/Y_i(1)$ are individual i earnings at age 35 without/with attending medschool. Comparing students that went to medschool (`x=1`) to those that did not (`x=0`) is not a “fruitful” comparison because of selection: students that choose to apply to medschool are different from those that do not. A natural experiment in the Netherlands helps to solve this selection problem. There, medschool applicants are randomly admitted conditional on grades in high school (`hsgpa`).

1. Open the file `docsim` and then type `desc` and `sum` to familiarize yourself. Also make histograms of `x`, `hsgpa`, and `y` to get more understanding of the data. Do you see strange things?
2. Type `tab group apply` and `tab group hsgpa` to understand what groups there are in the data.
3. A naïve estimate of the effect of becoming a doctor is to compare all doctors to all non-doctors in our sample of students. You get this estimate and statistical significance by doing a *t-test*: `ttest y, by(x)`. What would the conclusion be on the basis of this? Explain what lecture slide this relates to.
4. To visualize your analysis copy the following into your do-file and run it

```
gen x_shft = x + 0.01*(group==1)+ 0.02*(group==2)
gen y_1 =  y   if group==1
gen y_2 =  y   if group==2
gen y_34 =  y  if inlist(group,3,4)
twoway (scatter y_34 y_1 y_2 x_shft , ms(o o o) mc(gs8 red orange)) ///
       (lfit y x), xtitle("Medschool") xlabel(0 "No" 1 "Yes") name(all) ///
       ytitle(Yearly income age 35 (1000 euro)) ///
       legend(order(1 "Groups 3/4" 2 "Group 1" 3 "Group 2" 4 ///
       "Comparison  uncontrolled") cols(3))
```

The first few statements create the `y` variable per subgroup and shifted `x` such that you can see the dots of the separate groups. Explain the colored dots and the line.

5. We know that comparing all doctors to non-doctors is not a ″fruitful″ comparison because then you are also comparing applicants to non-applicants who may be very different. To illustrate this point type

```
tab x apply
tabstat y hsgpa if x==0, by(apply)
drop if apply == 0
count
```

   The summary statistics table shows the mean wage and also high school GPA (group) by applicant status. What does this reveal to us? Using the final statements we remove all the non-applicants. How many observations remain?

6. By removing the non-applicants from the sample we have effectively controlled for it. That is to say, we will no longer compare applicants to non-applicants using this sample. Type

```
ttest y, by(x)
tabstat y hsgpa , by(x)
```

   and interpret the results. Is this a fruitful comparison? Why/why not? What slide do these results relate to?

7. You can again visualize the previous comparison using a graph by copy-pasting the code below in your do-file. (If you have time, try to experiment with the colors and other options)

```
twoway (scatter y_1 y_2 x_shft , ms(o o) mc(red orange)) ///
       (lfit y x , lc(gs8)) , xlabel(0 "No" 1 "Yes") ///
       xtitle("Medschool") ytitle(Yearly income age 35 (1000 euro))  ///
       legend(order(1 "Group 1" 2 "Group 2" 3 ///
       "Comparison (uncontrolled)") rows(1)) name(applall)
```

   Can you explain the results from the previous question using the graph?

8. The theme this week is “controlled comparison”. You can do a “controlled” *t*-test by using the `bysort` command which can be abbreviated

```
bys hsgpa: ttest y , by(x)
```

   Interpret the result, and explain what slides it relates to.

9. The controlled comparison can be visualized by

```
twoway (scatter y_1 y_2 x_shft , ms(o o) mc(red orange)) ///
       (lfit y_1 x , lc(red)) (lfit y_2 x , lc(orange)), ///
       xtitle("Medschool") ytitle(Yearly income age 35 ///
       (1000 euro)) xlabel(0 "No" 1 "Yes") ///
       legend(order(3 "Comparison within Group 1" 4 ///
       "Comparison within Group 2") rows(1)) name(applby)
```

   Explain what you see, and how controlling works.

Part 2 – Controlled comparison by regression

In the previous part you did a controlled comparison by doing two separate *t-tests* ‘s. In principle this works well and gives you two separate estimates, one for group `hsgpa=0` and one for `hsgpa=1`. Both estimates are unbiased for the Conditional Average Treatment Effect of each subgroup.

The disadvantage advantage of having two estimates is that each of them uses only a part of the data. The way to combine both estimates, while controlling for `hsgpa,` is regression.

10. Type
    ```
    eststo r1: reg y x, r
    eststo r2: reg y x hsgpa, r
    ```
    to do two regressions stored as `r1` and `r2` respectively. Explain what both results are. Which one do you trust?
11. To show you the relationship between a *t-test* and regression, you can do the regression on both subgroups separately (as with the *t-test*).
    ```
    eststo r3: reg y x if hsgpa==0, r
    eststo r4: reg y x if hsgpa==1, r
    ```
    what do you get? Compare it to question 8.
12. With the `esttab` command we can display multiple regressions that are stored in memory in one table. First we need to install the `estout` package which does not come built-in with Stata.
    ```
    ssc install estout //need only run once on given PC
    esttab r1 r3 r4 r2, b(a2) se
    ```
    The sub option `b(a2)` tells Stata to use two digits, and the `se` option gives a table with standard errors rather than t-statistics. This is the default layout of a regression table that you will see in most empirical papers written in economics. Comment on what you see in the table. In particular focus on the coeffients, the standard errors, and *N*.