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

- Using windows explorer, make a new folder called `H:\rproject\clab\clab32`.
- Download the datafile `glass.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\clab32"        //this is your path
```

- Save the do-file as `clab32.do`.
- In the following, as always, don't forget to apply Rules 1&2: paste the command in your `clab32.do` file, press control-s to save and control-d to run.

Part 1 – Lazear's incentive effect estimation by comparison

In a seminal paper by Edward Lazear (2000), the author finds that the productivity of car glass mechanics increases by about 20% if they are paid by a piece-rate of about $20 *per installed glass* rather than a fixed wage per hour. The piece-rate, which is specified in the Performance Pay Plan (`ppp`), guarantees a minimum pay of about $11 per hour to shield workers from downside risk. In this exercise you are going to estimate the effect of `ppp` on the monthly average number of glass units that a worker is able to fit per day (`glass`) using simulated data in spirit of Lazear's.

1. Open the data and familiarize yourself using `desc` and `sum`. Visualize the data using

```
hist glass, name(post)
hist glass_pre, name(pre)
scatter glass glass_pre, name(scatter)
```

   where the name options make sure the graphs remain in memory so you can look at them again later. Do you understand what all the variables are?

2. To be able to estimate the effect of `ppp`, the first thing you need is variation: some people with `ppp=0` and others with `ppp=1`. Without variation, no comparison is possible. Type `tab ppp` to see if there is variation and explain. Then copy-paste-run

```
twoway (kdensity glass if ppp==0, lc(black)) ///
  (kdensity glass if ppp==1, lc(blue)), ///
  legend(order(1 "No PPP" 2 "PPP")) name(kden)
```

   to visualize the difference in the number of glass units installed between both groups. What does the graph show? Can you tell whether `ppp` increases productivity?

3. Simply eyeballing the distributions is useful, but it does not tell us whether the observed differences are simply driven by chance (i.e. sampling uncertainty). To test whether $\mathcal{H}_0: \mu_0 = \mu_1$ vs. $\mathcal{H}_1: \mu_0 \neq \mu_1$, you can do the t-test:

```
ttest glass, by(ppp) uneq
```

   Interpret the result. Does it surprise you?

4. In the previous computer lab session (3.1) you saw that a t-test is the same as a regression with one dummy variable. Type

```
eststo a: reg glass ppp, r
```

   to verify this in this context. Explain by comparing to question 3.

5. Given the findings in Lazear's paper, your results may be surprising. After you do an analysis, however, you should always ask yourself the question: "Is the comparison that I have made valid (fruitful), or am I (partly) comparing apples to oranges here?" When you have additional information in your dataset you may try to see if the groups are comparable in that respect. This dataset has prior productivity `glass_pre` that we can look at by typing:
```
tabstat glass_pre_glass, by(ppp) f(%6.2f)
```
Interpret the results. Do you have evidence of selection? Negative/positive?

Part 2 – Lazear's incentive effect estimation by controlled comparison

Sometimes, a non-fruitful comparison can be changed into a fruitful one by controlling for observable differences. For this to be credible, you *need to know something about the selection mechanism*. In this case, suppose we know that the board of the car glass company decided to select half of the worst performing shops in the previous month to get `ppp`, and keep the remaining shops at regular pay. Then, clearly, comparing shops with `ppp=1` to those with `ppp=0` does not make any sense because the bad shops - with on average worse workers - have been selected to get `ppp=1`. A case of negative selection.

Comparing bad to good shops is clearly not a fruitful way to learn about the effect of `ppp`. When you consider the individual worker, however, this may be different. Consider two workers that both produced `glass_pre=3`. Both these workers are likely to be of similar quality (including skill, motivation, and effort). Still, one worker gets `ppp=0` while the other gets `ppp=1`. Why? Because the board selects the bad shops on the basis of *average* quality of the workers in the shop. The first worker has good colleagues on average, while the other does not. Assuming colleague quality is random, we can compare `ppp=0` to `ppp=1` for workers that have the same `glass_pre`. Or, in other words, we can get the causal effect of `ppp` by controlling for `glass_pre`, which prevents comparing apples to oranges (people with different values of `glass_pre`). In this part, we will see that the easiest way of doing so is by linear regression.

6. In the example above we compared two individuals with the same value of `glass_pre`. In practice we want to do this comparison for all values of `glass_pre`. To make it simpler, divide the `glass_pre` in three equally sized groups or *bins*, and do the comparison *per bin*. The code below achieves this
```
egen glass_pre_bin3  = cut(glass_pre), gr(3)
tab glass_pre glass_pre_bin3
tab  glass_pre_bin3 ppp
table  glass_pre_bin3 ppp, c(mean glass) format(%6.2f)
eststo b: reg glass ppp i.glass_pre_bin3, r
```
Look carefully at the output. How are the bins formed? Is there variation in each bin (people with `ppp=0` and `ppp=1`)? What seems to be the effect in each bin? What does the regression give you? Note that the `i.` operator tells Stata to make a dummy variable for each bin.

7. Repeat question 6 using 10 bins rather than 3. Store the regression as `c` rather than `b`.

8. It is insightful to put the (controlled) analyses into a graph. For that we need some code. Copy-paste-run the following:
```
bys glass_pre_bin10: egen my0=mean(glass) if ppp==0
bys glass_pre_bin10: egen my1=mean(glass) if ppp==1
bys glass_pre_bin10: egen mx=mean(glass_pre)
```

```
bys glass_pre_bin10 ppp: egen ny=count(glass)
bys glass_pre_bin10 ppp: gen absmx = abs(glass_pre-mx)
bys glass_pre_bin10 ppp (absmx): gen t = _n
twoway (scatter my0 my1 mx if t==1 [fweight=ny], ms(Oh X) ///
 mc(blue black)), legend(order(1 "no PPP" 2 "PPP"))  ///
 xtitle(Average #glass installed last month (10 bins)) ///
 ytitle(Average #glass installed this month) name(control10)
```

Look at the data and the newly created variables to understand the code. Also, remove the weights `[fweight=ny]` to see what that does. The selection `if t==1` is not needed but prevents projecting multiple points onto themselves. Explain the graph and how it relates to the previous question.

9. Creating 10 bins of the `glass_pre` variable nicely illustrates how controlling works. It may not be entirely valid, however, because within a bin the `ppp=0` group may still have slightly lower values of `glass_pre` than the `ppp==1` group. To completely prevent apple-to-orange comparisons (in terms of `glass_pre`) we have to make a bin for each value of `glass_pre`. You can do that using the code below

```
egen glass_pre_binall = group(glass_pre)
tab  glass_pre_binall ppp
table  glass_pre_binall ppp, c(mean glass) format(%6.2f)
eststo d: reg glass ppp i.glass_pre_binall , r
```

Look carefully at the output. Is there variation in each bin (people with `ppp=0` and `ppp=1`)? What seems to be the effect in each bin? What does the regression give you?

10. The simplest way to control for `glass_pre` is to simply include it into the regression linearly, so without making a bin for each value. You get this including the original `glass_pre` variable without the `i.` operator. Run this regression and put all in a nice table by[1]

```
eststo e: reg glass ppp glass_pre , r
esttab a b c d e, b(a3) se drop(*.glass*) ///
 mti(nocontrol control3 control10 controlall controllinreg)
```

The table summarizes all your estimates so far. From columns (3) to (5) it seems that controlling for `glass_pre` using 10 bins, all bins, or linearly, does not change the results much: you always get an estimate of about 0.6. Mean productivity is about 3, so the increase in productivity due to `ppp` is about 0.6/3 = 20%. Exactly what the Lazear (2000) paper finds.

11. The conclusion from question 10 is that if you want to control for a variable, you can simply add it to the regression; you do not have to create all possible bins. This is usually true, if the pattern in the data are not too non-linear. Lets make it visual

```
predict yhat
twoway (scatter my0 my1 mx [fweight=ny], ms(Oh X) mc(blue black)) ///
 (line yhat glass_pre if ppp==0, lc(blue)) (line yhat glass_pre ///
 if ppp==1, lc(black)) if t==1, legend(order(1 "no PPP" 2 "PPP" ///
 3 "Linear fit" 4 "Linear fit") row(1))   title(Control glass_pre) ///
 xtitle(Average #glass installed last month (10 bins)) ///
 ytitle(Average #glass installed this month) name(regfit)
```

What does this graph show? Are you surprised to find that the results in column (4) and (5) from question 9 are very similar after seeing this graph?

12. Finally, we would like to visually compare what happens when you do control for `glass_pre` using a regression, compared to simply looking at the (uncontrolled) difference. Copy-paste-run the following

[1] First type `ssc install estout` if you get an error message.

```
drop my0 my1 ny mx
egen my0=mean(glass) if ppp==0
egen my1=mean(glass) if ppp==1
egen mx=mean(glass_pre)
bys ppp: egen ny=count(glass)
twoway (scatter my0 my1 mx [fweight=ny], msize(large large) ms(Oh X) ///
 mc(blue black)) if t==1, legend(order(1 "no PPP" 2 "PPP")) ///
 xlabel(1(1)6) title(No control) ///
 xtitle(Average #glass installed last month (1 bin)) ///
 ytitle(Average #glass installed this month) name(nocontrol) nodraw
graph combine nocontrol regfit, ycommon
```

What does the graph show? Explain.

References

Lazear, E. P., 2000. "Performance Pay and Productivity." *American Economic Review* 90 (5): 1346-1361.