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

- Using windows explorer, make a new folder called `H:\rproject\clab\clab41`.
- Download the datafile `gt_data.dta` and dofile `progs.do` 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\clab41"     //this is your path
```

- Save the do-file as `clab41.do`.
- In the following, as always, don't forget to apply Rules 1&2: paste the command in your `clab41.do` file, press control-s to save and control-d to run.

Part 1 – Overlap in the assignment of kids to Gifted and Talented (GT) education

In these exercises you are going to replicate results from the lecture and estimate the effect of following a gifted and talented program (`gt`) on students' math and language scores in secondary education. The data are drawn from Booij *et al.* (2016). In Part 1 you will try and figure out if you could apply a Regression Control (RC) design, as you did in previous CLabs 3.1 and 3.2.

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

   ```
   hist XXX, name(hist_XXX)
   ```

   where `XXX` is the name of the variable that you are looking at. Check out all variables (except `id`, of course). What variables do you think are realized prior to, and what are realized post of the GT progam?
2. To estimate the effect of the `gt` we will ultimately have to compare individuals that get the program (`gt=1`) to those that don't (`gt=0`). Type

   ```
   tabstat *, by(gt) format(%6.2f)
   ```

   to see the difference in means between these two groups in terms of all variables in the dataset. Are both groups comparable? Do we have apples to apples or apples to oranges here? If there is selection, is it positive or negative?
3. Observing imbalance (prior differences between the groups) is not a good sign, but not necessarily a problem. Imbalance with respect to gender, for example, only poses a big problem *if it predicts math and/or language scores*. A simple way to check this is to do a regression using prior variables as predictors. For simplicity we will neglect cohort and only look at math:

   ```
   reg resmath male age ist cito, r
   ```

   Is gender predictive? So, is imbalance with respect to gender a problem?

It is not always easy to interpret the coefficients from a regression. The coefficient of age, for example, $\hat{\beta}_{age} = -.1639968$ and significant ($t - stat = -4.15$). What does it mean? It means that when a student's age is 1 year higher, her predicted grade is 0.16 lower. That seems a lot, but is it? That depends how big a change 1 year is.

Recall from **q1** that $SD(age) = 0.48$. A 1 unit change in age is more than $2 \times SD(age)$. If $\pm 2 \times SD(age)$ around the mean contains about 95% of the data, $2 \times SD(age)$ must contain about 47.5%. So if your age is 1 year higher, you "overtake" about 47.5% of students and go from being one of the youngest to one of the oldest. A big change. Also, you can look at the histogram of `age` to see if you think a 1 year age difference is large or not.

4. A handy way to make the size of the coefficients comparable is to scale it by $\frac{SD(age)}{SD(resmath)}\hat{\beta}_{age} \equiv \hat{\beta}^{std}_{age}$. The interpretation then becomes: "if `age` is 1 SD higher, how many SD higher is `resmath` expected to be?" You can get this number quickly by using the `beta` option

```
reg resmath male age ist cito, r beta
```

   What does this reveal? Is the coefficient of `age` really that big, or are some of the others bigger?

5. The philosophy of doing a controlled comparison is to find subgroups of units that are comparable in terms of other observable characteristics (`cohort, male, age, ist, cito`) and compare `gt=0` and `gt=1` people within that subgroup, rather than all units. This requires "overlap": within each subgroup we need treated (`gt=1`) and control (`gt=0`) units.

   From **q4** we know that imbalance with respect to `ist_norm` and `cito` would be most problematic because these are very predictive of math grades. Lets check if we have overlap with respect to these variables:

```
twoway (scatter ist_norm cito if gt==0, ms(Oh) mc(blue) ) ///
  (scatter ist_norm cito if gt==1, ms(X) mc(black)), name(overlap)
```

   Do we have overlap? Can we use the RC design here? Explain.

Part 2 – Regression Discontinuity design: Comparing kids around the cutoff

In part 2 of these exercises you are going to replicate results from the lecture and use a regression discontinuity design to tackle the problem of "no overlap". The previous question showed that there are no control units with `ist_norm=1`, and no treated units with `ist_norm=0` (no overlap). What we could also see from the **q5** scatter plot is that the change is rather abrupt: there are many units that have a value just below the cutoff that do not get `gt==1` but still have a very similar `ist_norm` value as those above. We can use this in a Regression Discontinuity (RD) design.

6. First plot the data to get an idea:

```
twoway (scatter resmath ist_norm if gt==0, ms(Oh) mc(blue) ) ///
       (scatter resmath ist_norm if gt==1, ms(X) mc(black)), name(rd0_ist)
```

By giving different colors and shapes to the scatter points in the sample `gt==0` and `gt==1`, we can clearly see the overlap problem. We can also see, however, that there are many units around `ist_norm=0`, the cutoff value at which treatment status discretely changes from 0 to 1.

7. From the scatter in **q6** it is hard – if not impossible – to see if students score higher after the cutoff. To see we can add regression lines to the plot

```
twoway (scatter resmath ist_norm if gt==0, ms(Oh) mc(blue) ) ///
       (scatter resmath ist_norm if gt==1, ms(X) mc(black)) ///
       (lfit resmath ist_norm if gt==0, lc(red) range(. 0)) ///
       (lfit resmath ist_norm if gt==1, lc(red) range(0 .)), name(rd1_ist)
```

8. The lines in the plot show a jump (discontinuity) at point `ist_norm=0`. From the graph it is difficult to see exactly how large the jump is, and we can not tell if it is significant. For that we need a regression. First we quietly run a regression without controls for future reference and store it as `a`; then we do the RD regression and store it as `b`:
```
qui eststo a: reg resmath gt, r
eststo b: reg resmath gt ist_norm c.gt#c.ist_norm , r
```
What is the estimated effect $\hat{\beta}_1$? And what are $\hat{\beta}_2$ and $\hat{\beta}_3$, and what do they measure?
9. The graph from **q7** does not clearly show how the data fits the line(s). To see that we have plot the data using a smaller number of bins. The program `grrd` does that (check out the code in `progs.do` if you want to learn about it):
```
run progs.do
grrd resmath
```
10. The graph makes is visually clear that there is a discontinuity in average math scores at point `ist_norm=0`. If it is reasonable to assume that students around the cutoff are similar in all aspects *other than* `gt` assignment, we can conclude that the difference that we see is due to the program. We have some additional prior variables (`cohort, male, age, cito`) that we can include to see if the result changes:
```
eststo c: reg resmath gt ist_norm c.gt#c.ist_norm age male cito i.cohort, r
esttab a b c, b(a2) se nogap star(* 0.10 ** 0.05 *** 0.01)
```
Look at columns (1) – (3) and explain the differences and similarities. Do you believe it is credible to assume that students around `ist_norm=0` are similar?
11. Finally, we can repeat the analysis for language grades
```
qui eststo d: reg reslang gt ist_norm c.gt#c.ist_norm age male cito i.cohort,r
grrd reslang
esttab a b c d, b(a2) se nogap star(* 0.10 ** 0.05 *** 0.01)
```
Do we see an effect on language skills as well? How large?

References

Booij, A.S., Plug, E. & Haan, F., 2016. " Enriching Students Pays Off: Evidence from an Individualized Gifted and Talented Program in Secondary Education", *IZA working paper series* 9757.