7316 - INTRODUCTION TO R

Final Assignment

Teacher: Mickaël Buffart (mickael.buffart@hhs.se)

Deadline: 2022-05-15 23:59 — To submit on Canvas.

1 INSTRUCTIONS:

- This is the final assignment of the course 7316 — Introduction to R for Data Analytics.
- Deadline: 2022-05-15 23:59; submit your `.Rmd` file on Canvas.
- In this final project you are going to replicate some findings from the paper "A Passage to America: University Funding and International Students" (Bound et al., 2020) in the *American Economic Journal: Economic Policy*. You may find more detailed instructions on what to do below.
- The course is pass/fail. In order to pass, you should submit your solutions for the assignment bellow in an *RMarkdown* document (a single `.Rmd` file):
  - The RMarkdown file should generate a Word document containing the code chunks to create all the desired tables and figures, along with the instructions bellow. You can use the `.Rmd` instruction file as a baseline (available on the Github of the course)
  - Attention should be paid both to the quality of the outputs (generate nice tables, nice figures, and nice document) and the quality of the code (respect style guidelines as discussed during the course)
  - Log files, screenshots, or documents in another format than the RMarkdown file will not be accepted.
- This is an individual work and you are required to submit your individual solutions in order to pass.
- **Important:** Your main guide is the paper. It contains all information about the sample construction and empirical strategy. I will guide you through the first steps, but for running the regressions and creating the plots you are on your own. You will probably have to use some unfamiliar functions, but you should be skilled in reading documentations by now.

Bound, J., Braga, B., Khanna, G., & Turner, S. (2020). A Passage to America: University Funding and International Students. *American Economic Journal: Economic Policy*, 12(1), 97-126.

2 Exercise 1: Replicate Figure 4

- Load the dataset named `pub_pvt_scatters.dta` and `univ_names.xls`.
- Merge `pub_pvt_scatters` with `univ_names` using the appropriate merge command (think about which observations you want to keep).
- Create a common deflator *cpi_all* which is just the mean of *cpi* across all universities in the same year.
- Create the real value of the appropriation by dividing *nominal_approp* by *cpi_all*
- Use the `Private` variable to create a categorical variable that differentiate public and private universities.
- Reorder your data by unit and by year.
- Create the log of the variable `real_approp` and `ENROLL_FRESH_NON_RES_ALIEN_DEG`
- Create the difference in log values by university between 2005 and 2012 for these two variables and save it in a new `data.frame`.
- Replicate Figure 4 using *ggplot*, including all the features such as fitted lines, labels and colors. If you want to make it prettier, feel free to be creative.

3 Exercise 2: Replicate Table 2

Table 2 presents the results from the regression of *foreign freshmen enrollment* on the log of *state appropriations*. They run this regression separately for *research*, *AAU* (elite colleges) and *non-research colleges*, in two stages.

The bottom part of the table reports the results from the first-stage regression log state appropriation ~ approp. other univ.. If you are unfamiliar with instrumental variable regressions or panel regressions, take a look at the **Panel and IV Review** document that I uploaded. It contains a very brief summary to give you a basic idea of what these methods accomplish, without diving into the math.

- Load the dataset `univ_data.dta`
- The dependent variable is `l_ENROLL_FRESH_NON_RES_ALIEN_DEG` (*i.e.* ln(foreign first-year enrollment). The explanatory variables are the logs of state appropriation (`l_state_ap`) and the log of the population (`l_population`). All of these variables have already been created.
- Create a variable for the total state appropriation of all other universities within the same state. The variable `nominal_approp` is the total appropriation on the state level, so you can just subtract $nominal\_approp - state\_ap * 100000$ from that (multiplied by 100000 because they are on different scales)
- Find the balanced sample of universities that report foreign, domestic in-state and domestic out-state enrollment in a given year. Drop University-year observations where one or more of those are missing.
- Run the regressions from table 2. Refer to the paper to find the correct specification.
  - Notice that the authors choose a weighted regression. You can set the *weights* option in the regression command to the "weight" variable that is already in the data.
  - *Hint:* I recommend using `felm()` from the lfe package instead of `plm()` because it is not straightforward to calculate cluster robust standard

errors with weighted `plm` regressions. You will have to read a bit the documentation of `felm()` to know how to use it.

- Arrange the regressions in a table following the layout of table 2 in the paper. `stargazer` does not allow to stack regressions on top of each other, so it is ok if you make one table for the OLS and IV regressions and a second one for the first-stage regressions.
- (*optional*[1]). Adjust the standard errors to clustered standard errors. Your standard error might differ a bit, since it depends on how your function estimates the covariance matrix of the errors.

[1] If you skip this part, you will not get penalized. However, this means that you will not get the same errors in your regression table as in the article.