MODULE 4: MODELING DATA. – PRACTICE.

7316 - INTRODUCTION TO DATA ANALYSIS WITH R

Mickaël Buffart (mickael.buffart@hhs.se)

In module 3, we replicated figures from two papers and created a descriptive statistics table. In this module, we will replicate some of the results of the same two papers.

1. Exercise 1. Replicate Table 2 (Washington, 2008)

1.1 Data Preparation

- Load your dataset from Exercise 1 of Module 3. Create all extra variables that you need for the regression. Don't forget the two fixed effects for *region* and *number of children* (Fixed effects in this context means simply coding these variables as factors).
- Check the data type of the variables that you will need. Convert them if necessary.
- Recode the Religion as a factor variable to align with the table's regressors. Set *protestant* as the reference group.

1.2 Regressions

Table 2 shows the coefficients from the model $Score = \beta_0 + \beta_1 ngirls + \beta_2 female + \beta_3 White + \ldots + \beta_{13} Dem.vote.share$ where the score is either the NOW for the 105^{th} Congress or the AAUW score for each congress from the 105^{th} to the 108^{th}.

- Run the regression from the first column where the dependent variable is the total NOW score using the `lm()` function.
  - Hint: if you selected only the relevant variables beforehand, you can just type `nowtot ~ .` as the formula.
- Save the regression output as an object.
- Run the regression with the AAUW score as the dependent variable for each congress separately (In the next module, we will learn loops. Loops are useful to avoid repeating the same operation multiple times; for now, you can do it by hand).
- Combine the NOW score regression and the AAUW regressions into one table using `stargazer`.

  - omit the region and number of children fixed effects from the output
  - keep the number of observations as the only statistic
  - set the type to HTML

2. Exercise 2: Replicate Table 2 (Bound et al., 2020)

Table 2, from Bound *et al.* 2020, 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 $log(stateappropriation) \sim approp.otheruniv.$. 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 brief summary to give you a basic idea of what these methods accomplish, without diving into the math.

- Load the dataset `univ_data.dta` available in the data from module 4.
  - **Note:** the data file is the one from the author, which you can find on the article page, and the corresponding Stata code (the authors run the models in Stata).
- 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 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 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 of 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 stacking regressions on top of each other, so it is ok if you make one table for the OLS and IV regressions and a second for the first-stage regressions.

  - Adjust the standard errors to clustered standard errors. Your standard error will differ a bit since it depends on how your function estimates the covariance matrix of the errors.

- In this assignment, you may feel that you have done the same things three times. In the next module, we will see how to write functions, so that you can systematize your process.

3. References

- 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.
- Washington, E. L. (2008). Female socialization: how daughters affect their legislator fathers. *American Economic Review*, *98*(1), 311-32.