Empirical Project (Resit)

Machine Learning in Econometrics

due on February 1*

1 Data and Goal

The goal of this project is to study whether and to what extent politicians personally benefit from party affiliations. The dataset `politicians` (provided in `RData` and `csv` formats) contains data on individual politicians who have given speeches in parliament. For each politician, we observe the following variables:

variable name	**description**
`speaker`	identifying number of the politician
`party`	party affiliation of the politician; there are only two parties (“A” and “B”)
`age0`	age of the politician when giving the first speech in parliament
`birthplace`	identifying number of the region in which the politician was born
`married`	indicator for marriage (1=married, 0=not married)
`corrindex`	corruption index (larger values indicate larger probability of being corrupt)
`income`	average annual income of the politician (average over entire working life)
`allspeeches`	string containing all speeches the politician has given in parliament

*to be submitted on moodle by February 1, 23:59

There are two outcome variables of interest:

1. The variable `corrindex` contains an index produced by an independent non-governmental institution (NGO), which is is supposed to measure how corrupt a politician is. The index has mean equal to zero, so positive (negative) values indicate that a politician is more (less) corrupt than the average politician. More precisely, the NGO argues that, under some assumptions, larger values of the index indicate a politician who is more likely to engage in corrupt behavior at some point during their political career.
2. The variable `income` measures average annual income of the politician from age 40 to age 50.

We want to use the dataset to study the extent to which party affiliation affects these two outcomes, i.e. whether a party affiliation might make a politician more/less corrupt and whether it might affect their income. Your task is to perform a statistical analysis estimating the causal effects of party affiliation and exploring heterogeneity of the causal effects (i.e., what kind of politician benefits more/less).

You should explore two possibilities for establishing causality:

1. Use the text of the politicians' speeches to control for the "type" of politician. The idea is that the speeches may contain a signal about what kind of politician they are and may proxy for confounding variables (i.e., those that simultaneously affect party affiliation and the outcome).
2. Even after controlling for the type of a politician, there may be a concern that the speeches do not perfectly measure all confounders, so there may be endogeneity even conditional on the type. Explore the possibility of using birthplace as an instrumental variable for party affiliation.

2 Instructions

2.1 General Instructions

It is very important that your report is convincing, precise and concise. So, at every step of your analysis think about how you can justify your choices and assumptions. In particular:

clearly define objects of interest (causal effects), show assumptions under which they are identified, discuss how these assumptions can be justified in this project's context, clearly describe the estimator(s) you use, why you chose the estimator(s), how tuning parameters were chosen, explain assumptions required for the estimator to be consistent for the causal effect, and provide a sensitivity analysis to show that your results are robust to some of your choices. Finally, do not forget to interpret and discuss your results with a clear conclusion.

2.2 Permissible Ressources

You must work on this project on your own, no group work is allowed. You must produce the report as well as code on your own. You may use any "non-human" resources available to you, i.e. books, information on the internet, the lecture materials. However, you are not allowed to use interactions with other humans, i.e. no discussions with your classmates about the project, no discussions with others through the internet, no soliciting of help in online forums etc.

2.3 Format of the Report

- Length: max. 10 pages including graphs, tables, references
- File has to be a pdf document
- Filename: `[lastname]_[firstname]_[studentnumber].pdf`
- Take time to polish the text, so it is easy to read, precise and concise.
- The report must include the following sentence at the end:

  > "I confirm that this report is based on my own work. In preparing this report, I have not received any help from another human nor have I discussed any aspects of the empirical project with others."

  Then sign the report just below this statement.

2.4 Code

You must submit all code used in your statistical analysis in a separate file or multiple files. Make sure your code is readable and well-documented.

2.5 Expected Workload

Expected amount of time spent on the entire project: about 15 hours of empirical analysis + 2-3 hours of carefully writing up the report.