Practicum in Reproducible Research Methods

Homework One
Second Term 2020–21
Professor Miriam Golden
Assignment due: January 19 by 15:00

For this assignment, each research team will establish its initial working practices for the term.

You will be working with one of two datasets made available: the (observational) Italian dataset or the (experimental) Pakistani dataset. Note that whereas the Italian dataset is identical to that used in the published article, the Pakistani dataset contains only a subset of the variables, all of which were collected at the household level (i.e. for individual respondents). This is to allow you to focus on skill-building rather than on figuring out an excessively complex dataset.

Your first task is to *specify a preliminary research question* that the data you select allows you to answer. You should probably skim the relevant article to get initial ideas, and you should certainly read the accompanying codebook to understand the variables I have made available. You will draft your question in a single paragraph, or at any rate, less than a page. You may amend it later in the term, so please do not consider this a firm commitment.

For instance, let's say you decide to focus your efforts this term on learning whether Italian legislators charged with malfeasance who are mentioned more often in the press are less likely to be reelected. This is a purely descriptive question, but it's important and requires good knowledge of the datasets. Later in the term, you may find a way to formulate the analysis so that you can get some identification on whether *press mentions* are casually related to *reelection probabilities*.

Alternatively, let's say you decide to focus on whether there is a statistically significant difference in how much Pakistani respondents like their MPA depending on whether they received a phone call and a question from their MPA compared with not having been treated at all. (That is, you want to compare outcomes for {H0} against those for {H1} + {H2} in Figure 1.) This question involves comparing the difference in the thermometer

scores at baseline and endline for each group against the difference of the other group. Because calls were assigned randomly, this question can be answered causally; that is, you can be sure that a phone call from the MPA did (or did not) have an effect on how much respondents like their MPA. Estimation will require more statistical knowledge than that required for work with the Italian data. You should probably select to work with the randomized Pakistani data only if you have taken a causal inference course.

These questions are merely suggestions. You should explore your dataset and formulate your own.

You will be working as a team to answer your question. So you want to write out your *MOU*: how to divide the work and the credit, the ordering of your names for your eventual written output, and the subsequent ownership rights to your data, among other things. (Don't worry for now that this is all fictitious since you're working with someone else's data. Think it through as if you were going to spend a year of hard work collecting data with your team.)

You also should write our your SOP, which should at a minimum include instructions about how you communicate, how you organize your directories, how you write code, how you generate graphics, and perhaps instructions on certain estimation procedures to adopt as fall-backs.

Please write the three documents in RMarkdown. Output them in .pdf format. You will be working in the Research Practicum on GitHub and will submit your documents to me there.

Due by 15:00 January 19:

1. Summary of your research question
2. *MOU*
3. *SOP*