# The effect of eligibility on wealth

October 2, 2020

It is the year 1991. You are given a task by the US government to estimate the causal effect of eligibility for pension saving plans on household wealth. For that purpose, you receive data on households that contains a variety of household characteristics, including eligibility status and several wealth measures. At the end, you write a short scientific report (5-7) pages to describe your conclusions. Please proceed as follows.

1. Read the paper "The Impact of 401 (K) participation on the wealth distribution: an instrumental quantile regression analysis" by Chernozhukov and Hansen (Restat 2004). Focus on chapters 1 (intro), 3 (data description) and avoid section 2 (the method).

2. In your introduction, state briefly the purpose of your study.

3. Load the dataset pension from the R-package hdm (if you do not have the package on your computer, you have to install it). Write the code for this (and all subsequent) task(s) in a file pension.txt (with proper comments for each command).

4. Save the dataset under a name "mydata", so I can follow all your steps.

5. Check which variables your dataset has.

6. Write a section "Data and descriptives" in which you describe your dataset: which variables, summary statistics for your variables, some nice associations (correlations, correlation matrices, densities, etc.). You can use the lecture "descriptive statistics.pdf" that I have uploaded in Canvas. Which associations do you find interesting (on a premature, i.e. non-causal level)? Which assertions do you have after exploring the dataset? This section should be around a page, a page and a half. You might want to check some empirical papers to see how such a section looks like.

7. Write a section "Empirical strategy and results" to describe well... your empirical results and strategy.

(a) Define the ATE of interest and say what is its interpretation.

(b) Describe the simple comparison of means as an estimator and what the assumption behind is.

(c) Which variables have been omitted (i.e. are unobserved) that would invalidate the simple strategy? Describe the endogeneity problem behind the simple comparison of means.

(d) Describe an estimator that uses a conditional independence assumption. Do you find this assumption plausible?

(e) To estimate the conditional means, start by using a very simple parametric model (i.e. a linear model).

(f) Now increase the flexibility of this model using a more complex nonlinear squares model. Compare the assumptions behind these two first models.

(g) Now estimate the conditional means with a nonparametric and a semiparametric estimator. Which are the assumptions behind those strategies?

(h) Finally, estimate the conditional means using a Lasso estimator.

(i) Compare the ATE estimates under all strategies - are they similar? What would be the conclusions based on these estimates? Would they be different?

8. Summarize your conclusions in a way the policy maker can understand them.