MIT 14.385: Nonlinear Econometric Analysis, Fall 2022
Homework 1 for part 2.

Maximilian Kasy

In this problem, you are asked to replicate some difference-in-difference estimates in R. Your code should run from start to end in one execution, producing all the output. Output and discussion of findings should be integrated in a report generated in R-Markdown (or Quarto). Figures and tables should be clearly labeled and interpretable.

You will replicate some key findings from Derenoncourt and Montialoux (2021), implementing difference in difference estimates in R. The authors provide code in Stata. You should only consult this code to resolve ambiguities, and try to implement the estimates from scratch in R.

1. Read Derenoncourt and Montialoux (2021).
2. Download the source data for this paper from `https://sites.google.com/view/ellora-derenoncourt/us-inequality-data`.
3. Replicate the following figures and tables from the paper:
   Figure II, Figure V, Table I, and Table V.
4. Discuss these estimates in the context of our lectures on causality and identification.

Some R packages that might be useful: *kableExtra* (with the latex booktabs option), *broom* (for cleaning up linear regression estimates), *ggplot* (for data visualization), *fixest* for FE linear regression (*feols*). *stargazer*, *etable*, and *texreg* are other options for creating the output tables.

References

Derenoncourt, E. and Montialoux, C. (2021). Minimum wages and racial inequality. *The Quarterly Journal of Economics*, 136(1):169–228.

Additional hints, thanks to Jaume Vives

- You should work directly with the "cps_master_individual_level.dta" generated by Derenoncourt's pre-processing pipeline.
- To perform the sample selection as in the paper, for most tables and figures you will have to use the following variables: race, age, covered_all, flag_employed, in_sample, year.
- Note that 1962 is *not* used in the analysis. Therefore, you should drop it when estimating the models. The values for 1962 in Figure 1 are the averages of 1961 and 1963, and in Figure V 1962 is manually set to zero.
- The treatment of interest is covered_1966 and the outcome of interest is ln_annual_wage.
- The controls used throughout the paper are the variables "sex", "race", 'schooling', "exp", "exp_square", "exp_cubic", "fullpart", "wkswork2", "ahrsworkt", "marst", "occupation".
- The analysis is sensitive to the choice of reference groups. To identically replicate the paper you will have to use the following reference groups: 1965 for year, 12 for schooling, 6 for wkswork2 and 40 for ahrsworkt. The other variables have the first group (default) as reference. You can use the relevel function in R to change the reference level of the factor variables.
- For Tables I and V note that the variable "Time" is the factor that should be interacted with covered_1966. Time FEs relate to this variable and not year.
- In Table I, state-by-year FEs is actually just state FEs.
- In Table V, for model (3) Black 1961 is also removed.
- Be mindful of memory. rm(.) the dataframes you won't use again in the analysis, otherwise you might run into memory errors.