Data Task 2: Discrimination

Please send a document with your results (in any format of your choice, e.g. Word, LaTeX, Slides, R Markdown, Jupyter Notebook...) as well as your code (preferably R, alternatively Stata or Python) until ***June 4, 2023*** *to leonie.bielefeld@econ.lmu.de and emilio.esguerra@econ.lmu.de. You can work in groups of up to three students.*

The objective of this data task is to examine the impact of a discriminatory policy on wages. The policy was implemented from 1913 onward and specifically targeted black workers. Our goal is to analyze how this policy affected the salaries of black workers. You can be access the dataset on our Moodle course page.
In Class 8, we will be discussing a research paper that uses a similar dataset.

1. Load the dataset `wage_panel.csv` and create histograms to visualize the salaries of black and white individuals.

   a) Generate one histogram for the year 1911, displaying salary distributions by group. That is, plot salaries for black and white individuals in one histogram.

   b) Generate one histogram for the year 1921, displaying salary distributions by group. That is, plot salaries for black and white individuals in one histogram.

   c) Describe the histograms from (a) and (b). What conclusions can be drawn from comparing them?

2. Consider the data for the year 1911. Create a table presenting the variable's means and standard deviations separately for black and white individuals. Additionally, include a column indicating the differences in means and whether they are statistically significant. Discuss the key findings.

3. Compute the average $\ln(salary)$:

   a) For black individuals before the policy

   b) For black individuals after the policy

   c) For white individuals before the policy

   d) For white individuals after the policy

   (Hint: *To facilitate the analysis, you can add a dummy variable to the dataset that takes the value of 1 for years when the policy is in place.*)

4. Using your results from Question 3:

   a) Compute the simple difference estimator for black individuals. Discuss whether this estimator is suitable for estimating the causal effect of the policy on black individuals' wages.

b) Compute the difference-in-differences estimator for the wage-effects of the policy on black individuals. Discuss whether this estimator is suitable for estimating the causal effect.

5. To determine the statistical significance of the Difference-in-Differences estimator, estimate the impact of the policy on $\ln(salary)$ for black workers by running the corresponding regression model.

    a) Estimate the most basic Difference-in-Differences regression model.

    b) Provide a complete interpretation of the coefficient quantifying the wage-effect of the policy on black workers.

    (Note: *To perform the regression, your dataset should contain a dummy variable that equals 1 for years in which the discriminatory policy is in place and an interaction term between said dummy variable and a Treatment indicator.*)

6. You could include year fixed effects into your model.

    a) What factors do year fixed effects capture in this setting? Provide an example.

    b) Incorporate year fixed effects into your regression model and estimate the new model.

7. You could additionally include individual fixed effects into your model.

    a) What factors do individual fixed effects capture in this setting? Provide an example.

    b) Now also incorporate individual fixed effects into your regression model and estimate the new model.

8. Consider the following regression model, which estimates year-specific wage effects for black individuals:

$$\begin{aligned}\ln(salary_{it}) =& \beta_{1907} \cdot black_i \times \delta_{1907} + \beta_{1907} \cdot black_i \times \delta_{1909} + \beta_{1913} \cdot black_i \times \delta_{1913} + \\ & \cdots + \beta_{1921} \cdot black_i \times \delta_{1921} + \alpha_i + \delta_t + \epsilon_{it}, \text{ or:}\end{aligned}$$

$$\ln(salary_{it}) = \sum_t \beta_t \cdot black_i \times \delta_t + \alpha_i + \delta_t + \epsilon_{it},\ t \in \{1907, 1909, 1913, 1915, 1917, 1919, 1921\}$$

    a) Estimate the regression model above

    b) Produce a graph that plots the set of β-coefficients

    c) Describe the plot and explain the conclusions that can be drawn from it

9. Produce a nicely formatted regression table with all of your regression results from exercises (5a), (6b), (7b), and (8a).

Addendum: Tips for R

We recommend working with the open source software R to conduct the statistical analyses. To do so, you need to download and install R and R Studio. By installing R, you will install the programming language R on your computer. To work with R in a user-friendly way, you also need RStudio. RStudio is a so-called Integrated Development Environment or a clear interface for working with R. Both R and R Studio are available for free. You can download R here: https://cran.r-project.org. and R Studio here: https://posit.co.

You can find a lot of video tutorials on YouTube that are great for learning to get started with R, but also to find help with specific tasks. The documentation of packages and functions can be found on https://www.rdocumentation.org. The go-to resource for finding solutions to individual problems is Stackoverflow (https://stackoverflow.com/). Googling your coding problem or the received error message will often lead you to this widely-used forum.

You will need a number of functions from add-on packages for this data task (and most other empirical work):
Install the following packages and load them into your R environment:

- `ggplot2` (to visualize data in plots)
- `fixest` (to run regressions with alternative standard errors and fixed effects)
- `modelsummary` (to produce tables with summary statistics)

For example, install and load `ggplot2` by executing the following code:

```
install.packages("ggplot2")
library(ggplot2)
```

Here is a list of functions that might be useful for this data task:

read.csv()	Imports csv-files. Specify the ”sep”-argument if your imported dataset has only one column, e.g. by setting sep = ”;”
load()	Load datasets of type .Rdata
ggplot()	Plots data
ifelse()	Creates new variable based on values of other variable(s)
datasummary_balance()	Produces tables with summary statistics by groups and reports differences in means and their significance
mean()	Calculates means
feols()	OLS Regressions, different SE and inclusion of fixed effects possible
i()	Includes interaction terms to regression models (use with feols)
iplot()	Produces graph that plots the coefficients on interaction terms (use with feols)
etable()	Exports regression tables, e.g. in .tex format (use with feols)