MODULE 2: PLAYING WITH DATA! – PRACTICE

7316 - INTRODUCTION TO DATA ANALYSIS WITH R

Mickaël Buffart (mickael.buffart@hhs.se)

In this assignment, you will work on some tables from David Card and Alan Krueger's seminal paper from 1994 on the minimum wage introduction in New Jersey (Card and Krueger 1994). The paper is available on the course webpage.

1 Basic Setup: Tidyverse and Rio

1. Install the packages `tidyverse` and `rio`, if you have not yet installed them.
2. Create an RStudio project for this assignment. Create a `data` folder in the project repository.
    1. *Optional*: make the project a git repository and add the `data` folder to `.gitignore`
    2. What is the point of adding the `data` folder to `.gitignore`?
3. Download the dataset `card_krueger_public.dta` (available in 7316_ module_2_data.zip on canvas) from the course website. Copy it into your project data folder. Load the data.

2 Take a first look

4. Describe your dataset, its structure, how many variables, types of variables, etc.

You will notice that the data does not contain any variable names. We, therefore, refer to the codebook to find the necessary variables. I have prepared a CSV file with the variable names and labels called `card_krueger_variable_names.csv`.

5. Load this list and assign each column the appropriate variable name.

The dataset is still large, given that we only want to replicate two tables.

6. Drop all variables except `SHEET`, `CHAIN`, `STATE`, `EMPFT`, `EMPPT`, `NMGRS`, `EMPFT2`, `EMPPT2`, `NMGRS2`, `STATUS2`

3 Summarizing the data

7. Check if `EMPFT` contains missing values.

We now want to get a feeling for the observations we are dealing with. Card and Krueger sample restaurants of different US fast food chains (Burger King, KFC, Roy Rogers, Wendy's). We would

like to know the distribution of the different chains across New Jersey and Pennsylvania (table 2 in the paper)

8. Create a separate dummy variable for each chain that equals 1 (or `TRUE`) if the store belongs to this chain and 0 (`FALSE`) otherwise
9. Tabulate the mean of each of these 4 variables by State
10. Save the tabulated values into a `matrix`
11. Transform the `CHAIN` variable into a factor variable with properly labeled categories.
12. Remove the `STATE` dummy from the dataframe.
13. Transpose the matrix, rename the columns to correspond to the *Distributon of Store Types* section of Table 2 and turn it into a data.frame.
14. Print the table

4 Tyding up the dataset

If you look at the data, you will realize that the values for a single store are spread across several columns. The number of full-time employees is recorded in the variable `EMPFT` for the first year and `EMPFT2` for the second year. This violates the *tidy* principle that each observation has its own row. To make the tyding easier, we first reduce the number of variables by aggregating full-time employment, part-time employment and managers into one variable for full-time equivalents (`FTE`).

15. Aggregate the employment for each store and period into two new variables called `FTE1` and `FTE2`. Follow the paper and use the formula $FTE = EMPFT + 0.5 * EMPPT + NMGRS$
16. Order the data in `FTE1` ascending order and `FTE2` descending order.
17. Gather the data into a dataframe object, such that for each store, you have two observations of FTE, one for each year. Name the object `data_tidy`.
18. Save `data_tidy` into an Excel file named "`data_tidy.xlsx`", placed in a `derived_data` folder, inside your `data` folder.

5 References

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84 (4): 772–93. https://doi.org/10.3386/w4509.