Week 7 Lab

Adrien Bouguen

2025-03-24

Prelab Quiz (5 pts)

1. Download the dataset `covid_death.csv` from Camino
2. Create a binary variable (or logical) urban taking value 1 if urban status is a metropolitan area (large, medium or small) and 0 if the county is labelled either `micropolitan` or `rural`. Attach this variable to the dataset covid_death [1pt]

> Tip
>
> Use `tab1()` from the `epiDisplay` library to best describe your categorical variables.

3. Create a variable that calculate covid death rate by county i.e. the number of covid related death as a percentage of the total number of deaths. Attach this variable to dataset covid_death [1pt]
4. Regress the variable covid death rate on the variable urban (0/1) and present your result in a `modelsummary` table [1pt]
5. Submit your html and qmd file on time (before 4:20 pm) [1pt]

total= 5 pts

Overview

There are large gender differences between men and women in the labor market. In this lab, we propose to investigate some of these differences. For instance, what is the size of the pay gap between men and women? Is this pay gap the consequence of pure discrimination or are there other factors? What is the impact of giving birth of women's labor supply? We will investigate these question using two datasets.

Part 1: CPS data (12 pt)

1. Import the dataset;

https://vincentarelbundock.github.io/Rdatasets/csv/stevedata/gss_wages.csv in R .

You will find the codebook of this dataset here :

https://vincentarelbundock.github.io/Rdatasets/doc/stevedata/gss_wages.html

2. Use `datasummary` to create a table of descriptive statistics that includes at least the mean and the SD of each numerical variable (1pt)

> Tip
>
> `datasummary(All(df)~mean*Arguments(na.rm=TRUE), df)` provides a descriptive stats table of all variables in df and excluding NA values. To add one stats you can add `+ sd`

3. Notice that some variables do not appear in your descriptive tables. This is because these are text variables (we can't average "male" and "female"). To describe these variables you will have to create numerical variables. (1pt)

- Create a binary variable based on the text variable gender, assigning value 1 for female and 0 otherwise. Call it `gender_b`
- By the same token, create a binary variable full_time, Married, college (for junior college, bachelor and graduate students)

> Tip
>
> Use `tab1()` from the `epiDisplay` library to best describe your categorical variables.

- Regenerate a descriptive table for all of these variables

5. Using ggplot, create a graph of income and age. (1pt)
6. What does the relationship between age and income suggest? Is it a linear relationship? (1pt)
7. Regress income and gender. Correct for heteroskedasticity and present your result in a `modelsummary` table. (1pt)
8. Give the interpretation of the constant and its significance level (1pt)
9. Give the interpretation of the coefficient and its significance level (1pt)

10. Does this mean that women are discriminated on the labor market? Why or why not? (1pt)
11. Add the variable full_time into the regression and present your regression in a `modelsummary` table together with the result from question 7 (1pt)
12. What do you notice about the size of the pay gap when controlling for full_time?
13. Why do you think controlling for the variable full time affect the pay gap coefficient ? (1pt)
14. What do these results suggest about the correlation between gender and full_time? (1pt)

> **Tip**
>
> Think Frish & Waugh, Lecture Econ 41 Week 6

14. Verify your intuition about the covariance between gender_l and full_time using R? (1pt)

> **Tip**
>
> Beware that command `cov()` will gives result NA if any value of the variables is NA. To disregard NA, you want to use condition your variable to remove NA or use cor option `use="complete.obs`" (see help section)

Part 2: Fertility (4 pts)

1.Import the fertility dataset in R:

https://vincentarelbundock.github.io/Rdatasets/csv/AER/Fertility.csv. You will find the codebook here: https://vincentarelbundock.github.io/Rdatasets/doc/AER/Fertility.html

2. Morekids in the dataset is a text variable. If you want to use morekids as an outcome variable in a regression you will have to transform it into a binary variable. Generate a new variable morekids_b that takes value 1 when morekids is "yes"and 0 otherwise (1pt)
3. Use morekids_b to show that the probability of having more than 2 kids declines when the first kid is a boy. (1pt)

> **Note**
>
> `use a Linear Probability model and call it regression 3`

4. Interpret your regression 3 (both constant and slope coefficient)

Submit both the Rmd script and HTML report

You are done!!! You are doing excellent!

Total = 18 pts