Problem Set 1: OLS Review

EC 421: Introduction to Econometrics

Due *before* midnight on **Saturday, 29 January 2022**

INSTRUCTIONS

DUE Upload your answer on Canvas *before* midnight on Saturday, 29 January 2022.

IMPORTANT You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
2. the R script you used to generate your answers. Each student must turn in her/his own answers.

If you are using RMarkdown, you can turn a single file, but it must be a `html` or `pdf` file with **both** your R code **and** your answers.

If we ask you to create a figure or run a regression, then the figure or the regression results should be in the document that you submit (not just the code—we want the actual figure or regression output with coefficients, standard errors, *etc.*).

INTEGRITY If you are suspected of cheating, then you will receive a zero—for the assignment and possibly for the course. We may report you to the dean. **Cheating includes copying from your classmates, from the internet, and from previous assignments.**

README! The data in this problem set come from three sources—all of which I downloaded from NHGIS.

1. the 1860 US Census
2. the 2010 US Census
3. the American Community Survey (ACS), 2006-2010

The table no the last page of the problem set describes each variable in the data.

OBJECTIVE This problem set has three purposes: (1) reinforce the metrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality within econometrics/regression.

KEY TOPICS In this problem set, you will investigate the link between historical repressive/discriminatry institutions and current outcomes. Specifically, we will look at the modern effects of slavery in the United States—focusing on ten states that unsuccessfully attempted to secede (South Carolina, Mississippi, Florida, Alabama, Georgia, Louisiana, Texas, Virginia, North Carolina, and Tennessee). Each of these states was a "slave state" in 1860, *i.e.*, they allowed Black enslavement. In 1860, the population of enslaved people varied across the counties of these 9 states. In this problem set, you will investigate how the intensity of enslavement in 1860 correlates with (or affects) economic outcomes for Black and White individuals 150 years later (in 2010).

Setup

Q01. Load your R packages. You're probably going to need/want `tidyverse` and `here` (among others). Also: `pacman` and it's `p_load()` function make package management easier.

Q02. Now load the data (stored in `001-data.csv`).

I saved the data as a CSV, so you'll want to use a function that can read CSV files—for example, `read_csv()` in the `readr` package, which is part of the `tidyverse`.

Getting to know your data

Q03. Each row of the dataset represents a different county. How many counties are in the data? *Hint:* Try `dim()`, `nrow()`.

Q04. How many variables are there? How many *numeric* variables are there?

Hint: You have many options here; try `glimpse()` (in the `tidyverse`), `summary()`, or `skim()` (from the `skimr` package).

Q05. How many states are in the data?

Hints:

- The `n_distinct()` function (in the `tidyverse`) will tell you the number of distinct items in a variable.
- There's a variable named `state`.
- You can access a variable in a `data.frame` using the `$`, *e.g.*, `my_data$var1` grabs the variables `var1` from the dataset `my_data`.

Plotting the data

Q06. Plot a histogram of the median income for **Black households** in 2010 (variable: `income_black_2010`). *Note:* Household income is in 2010 dollars.

Don't forget to label your plot's axes. A title would be good too.

Q07. Plot a histogram of the median income for **White households** in 2010 (variable: `income_white_2010`). *Note:* Household income is in 2010 dollars.

Q08. What are the means of the two variables you just plotted?

Q09. Using the histograms from above (**06** and **07**), the means that you just calculated in **08**, and any other figures/summaries you want to produce: Discuss the differences in 2010 household income for Black and White households in these states.

Q10 Create two more histograms

1. for **population** in 1860 population that was enslaved (the variable: `pop_enslaved_1860`).
2. for the **percent** of the 1860 population that was enslaved (the variable: `pct_pop_enslaved_1860`).

What do these histograms tell you?

Q11. Create a scatterplot (AKA: dot plot) with the percent of the population that was enslaved (in 1860) on the `x` axis (`pct_pop_enslaved_1860` on the `x` axis) and median Black household income (`income_black_2010` on the `y` axis.

Q12. Based upon your plot in **Q11**: If we regress the median black household's income on the percent of the population that was enslaved in 1860, would the coefficient on `pct_pop_enslaved_1860` be *positive* or *negative*? **Explain** your answer.

Regression time

Q13. Now regress the median black household's income (`income_black_2010`) on the percent of the population that was enslaved in 1860 (`pct_pop_enslaved_1860`).

Interpret the results of the regression—the meaning of the intercept and the coefficient.

Is the coefficient on `pct_pop_enslaved_1860` statistically significant?

Q14. Now regress the median black household's income (`income_black_2010`) on the **population** that was enslaved in 1860 (`pop_enslaved_1860`).

Interpret the results of the regression—the meaning of the intercept and the coefficient.

Is the coefficient on `pop_enslaved_1860` statistically significant?

Q15. Repeat the regression from **Q14** but now with median White household income in 2010 as the outcome variable (`income_white_2010`).

Does slavery's legacy appear to have the same income effect on White household's income as it did on Black household's income? Explain using the coefficient on `pop_enslaved_1860`.

Q16 Before we can confidently conclude that these coefficients represent **causal** effects, we need to rule out potential sources of omitted-variable bias.

One possible omitted variable is "population".

Explain why population might be an omitted variable (using the two requirements for an omitted variable).

Hint: Would we expect more populated counties to have different levels of income than less populous counties?

Q17 Repeat the regression from **Q14** (regressing median Black household income on the) but now add to more explanatory variables: (1) the county's population in 1860 (`pop_total_1860`) and (2) the county's population in 2010 (`pop_total_2010`).

Does it appear that omitting total county population was causing omitted-variable bias? Explain.

Q18 Given the evidence so far, do you think slavery's appears to affect current economic outcomes (for Black or White households)? Explain your reasoning.

Q19 Suppose we ran the following regression (instead of the regression we ran in **Q14**)

```
lm(log(income_black_2010) ~ pop_enslaved_1860, data = ps_df)
```

How would you interpret the coefficient on `pop_enslaved_1860`?

Q20 What does exogeneity mean in this context?

Description of variables and names

Variable	Description
state	State name
county	County name
gisjoin	Code for GIS joining
pop_enslaved_1860	County population of enslaved persons in 1860 (Census)
pop_total_1860	County population in 1860 (Census)
pop_total_2010	County population in 2010 (Census)
pop_black_2010	County black population in 2010 (Census)
pop_white_2010	County white population in 2010 (Census)
income_black_2010	County median income for Black households 2006-2010 (ACS)
income_white_2010	County median income for White households 2006-2010 (ACS)
pct_pop_enslaved_1860	Percent of county population enslaved in 1860 (Census)
had_rosenwald_school	Indicator variable: Did the county have a Rosenwald school?