

Core Econometrics III: Problem Set 2

Owen Jetton

05/18/2022

Question 1

```
library(pacman)
p_load(dplyr, tidyverse, stargazer, fixest, magrittr, kableExtra)

data = read.csv("data-002.csv")
```

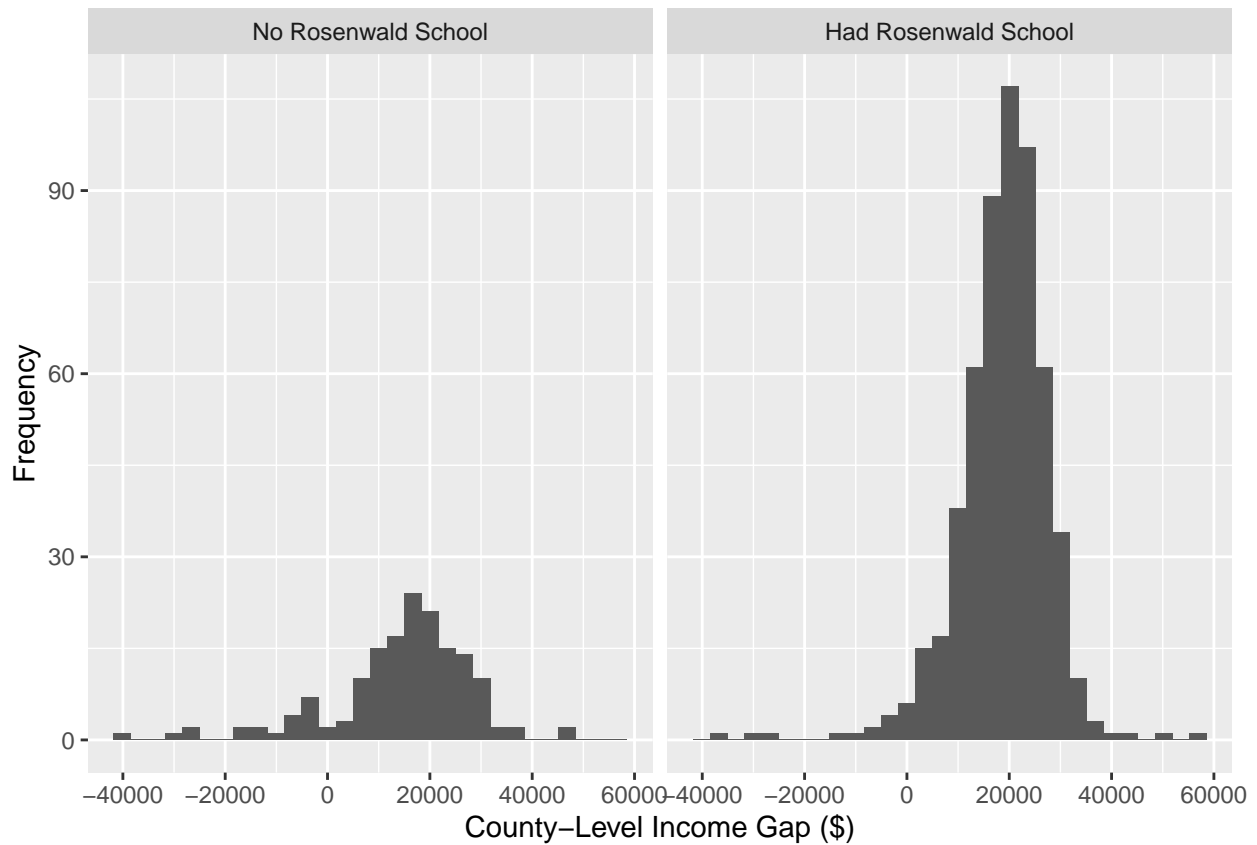
Question 2

```
labels.2.1 = c("No Rosenwald School", "Had Rosenwald School")
names(labels.2.1) = c("0", "1")

ggplot(data) +
  geom_histogram(aes(x = (income_white_2010 - income_black_2010))) +
  facet_wrap(~had_rosenwald_school,
             labeller = labeller(had_rosenwald_school = labels.2.1)) +
  labs(x = "County-Level Income Gap ($)", y = "Frequency")
```

A histogram of the county-level income gap (`income__white__2010 - income__black__2010`) split by whether the county had a Rosenwald school (`had__rosenwald__school`)

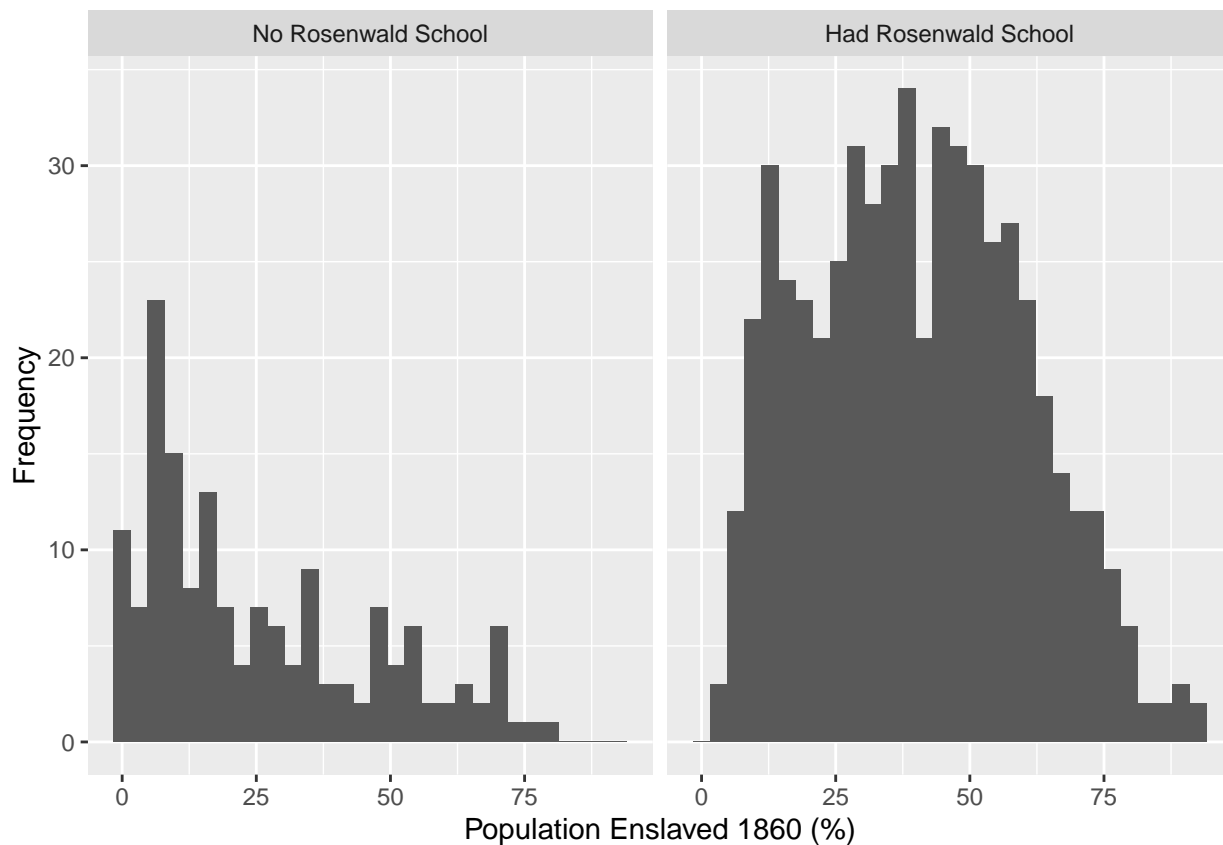
``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



```
ggplot(data) +
  geom_histogram(aes(x = pct_pop_enslaved_1860)) +
  facet_wrap(~had_rosenwald_school,
    labeller = labeller(had_rosenwald_school = labels.2.1)) +
  labs(x = "Population Enslaved 1860 (%)", y = "Frequency")
```

A histogram of the percent of the population in 1860 that was enslaved (pct_pop_enslaved_1860) also split by whether the county had a Rosenwald school (had_rosenwald_school)

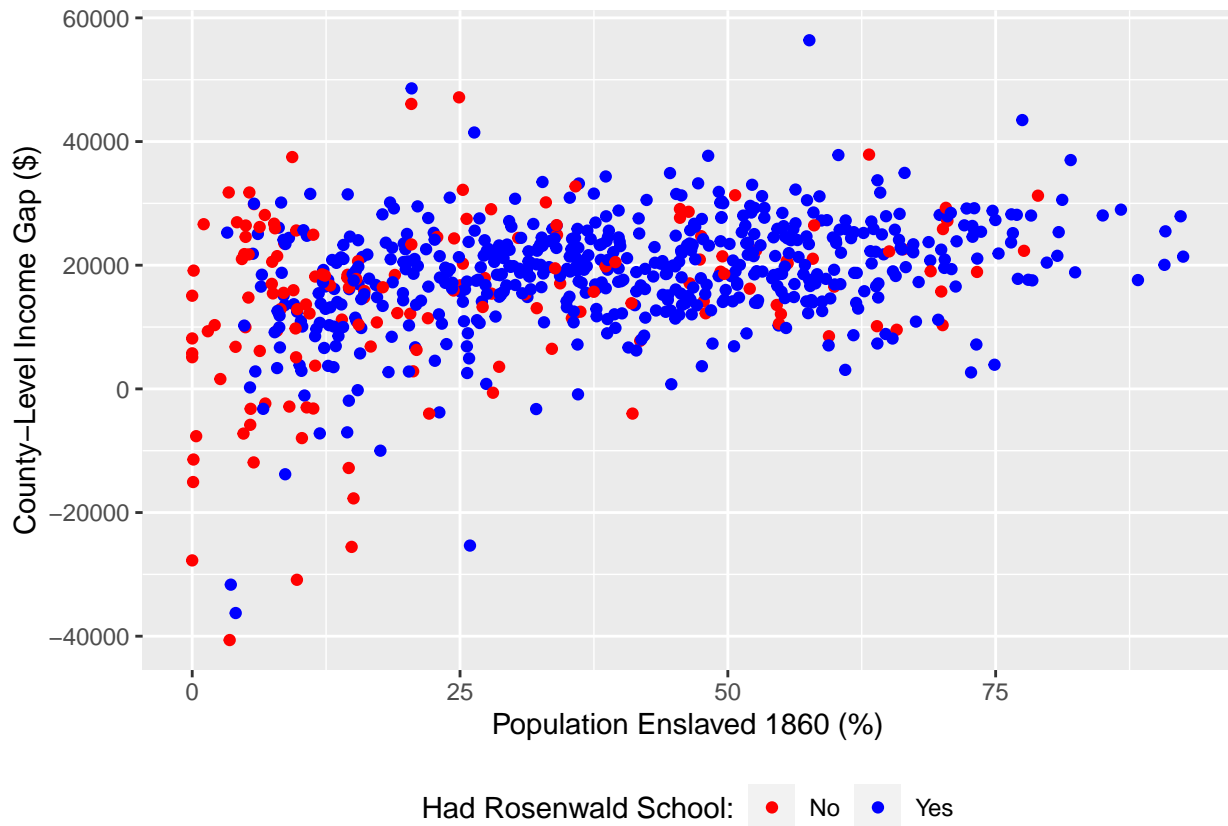
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



```
# making "had_rosenwald_school" a factor so legend isn't continuous
data %<>% mutate(had_rosenwald_school = factor(had_rosenwald_school))

ggplot(data,
  aes(x = pct_pop_enslaved_1860, y = (income_white_2010 - income_black_2010),
    color = had_rosenwald_school)) +
  scale_color_manual(
    breaks = c("0", "1"),
    values = c("red", "blue"),
    labels = c("No", "Yes"),
    name = "Had Rosenwald School:") +
  geom_point() +
  labs(x = "Population Enslaved 1860 (%)", y = "County-Level Income Gap ($)") +
  theme(legend.position = "bottom")
```

A scatter plot with `pct_pop_enslaved_1860` on the x axis and the county income gap on the y axis with the points colored by whether or not the county had a Rosenwald school.



Question 3

```
# create outcome variable
data %<>% mutate(race_income_gap = income_white_2010 - income_black_2010)

# regression of rosenwald school on income gap
reg_3.1 = feols(race_income_gap ~ had_rosenwald_school, data = data)

table3.1 = etable(reg_3.1,
                  style.tex = style.tex("aer"),
                  tex = TRUE)
```

Race Income Gap	
	(1)
(Intercept)	14,537.5*** (802.8)
Had Rosenwald School	4,323.7*** (909.7)
Observations	710
R ²	0.03092
Adjusted R ²	0.02955

The identifying assumptions necessary to interpret these results as causal are that we are estimating the true model, that there are no omitted relevant variables, that the exogenous variables are uncorrelated with the disturbances.

```
reg_4.1 = feols(race_income_gap ~ had_rosenwald_school | state, data = data)
```

```
table3.1 = etable(reg_4.1,
  style.tex = style.tex("aer"),
  tex = TRUE)
```

04 Now add a state fixed effect. Report your findings and discuss the conditional independence assumption required by the regression (for a causal interpretation).

	Race Income Gap
	(1)
Had Rosenwald School	4,285.8*** (1,087.0)
Observations	710
R ²	0.09778
Within R ²	0.02938
State Fixed Effects	✓

Question 5

```
data %<>% mutate(race_income_gap = income_white_2010 - income_black_2010)
```

```
reg_5.1 = feols(race_income_gap ~ had_rosenwald_school + pct_pop_enslaved_1860 + pop_total_1860 | state,
  data = data)
```

```
table5.1 = etable(reg_5.1,
  style.tex = style.tex("aer"),
  tex = TRUE)
```

	Race Income Gap
	(1)
Had Rosenwald School	2,252.9*** (669.1)
% Population Enslaved 1860	129.1*** (20.06)
Total Population 1860	0.1062** (0.0409)
Observations	710
R ²	0.16186
Within R ²	0.09831
State Fixed Effects	✓

06 Does the change in the estimated coefficients on had_rosenwald_school from 04 and 06 suggest match what you would have guessed if these new variables (pct_pop_enslaved_1860 and pop_total_1860) were

causing bias? Explain.

07 What is the required conditional independence assumption required by 06? Do you think it is valid? Explain your answer using a DAG (and discuss which variables should and should not be used as controls).

08 What are your thoughts on adding the county's 2010 population (Black, White, and/or total) as a control in the regression? Will these controls help or hurt our CIA? Explain.

09 Let's try some propensity-score matching. Step 1: Estimate some propensity scores!

To estimate propensity scores, we'll use a logistic regression. Don't worry if you don't know what it is. For now, all you need to know is that it is a nonlinear regression used for binary outcomes. We can use `feglm` from the `fixest` package to estimate a logistic regression, for example

```
feglm(binary_y ~ I(x1^2) + I(x2^2) + x1 + x2 + x1:x2, data = fake_df, family = 'logit')
```

will estimate a logistic regression (notice the argument `family = 'logit'`). The example also shows you how to square variables and take interactions.

You need to estimate a logistic regression as a function of `pct_pop_enslaved_1860` and `pop_total_1860`, their squares, and their interaction.

Finally, you can grab those beautiful (estimated) propensity scores from your saved regression using `saved_reg$fitted.values` (where `saved_reg` is my saved regression object output from `feglm`). You'll want to add those (estimated) propensity scores to your dataset.

10 Estimate a regression where you control for the original variables and for the estimated propensity scores. Does anything change?? Report your results.

11 Wait! How does overlap look? Make a figure to document the overlap.

12 What percent of your observations aren't supported by overlap?

13 Why do we care about enforcing overlap?

14 Repeat the regression with overlap enforced. Report your results. Did anything change?

15 Now weight your regression using the (inverse) propensity scores (as discussed in lecture). Report your results. Did anything change yet??

16 Maybe we need a doubly robust approach. Estimate the "controlling for stuff" regression from 06 for each block, where blocks are based upon observations' propensity scores.

Note: If you're using state fixed effects with small-ish blocks, then some of the blocks may not have variation in `had_rosenwald_school`. For the sake of making your life easier, you can drop the state fixed effects or make the blocks bigger.

Report each block's estimated treatment effect and the ATE.

Q Maybe a selection on unobservables strategy would have been better. Explain your thoughts on the idea of using the county's history of Black enslavement (e.g., `pct_pop_enslaved_1860`) as an instrument for whether the county received a Rosenwald school.