Core Econometrics III: Problem Set 2

Owen Jetton

05/18/2022

Question 1

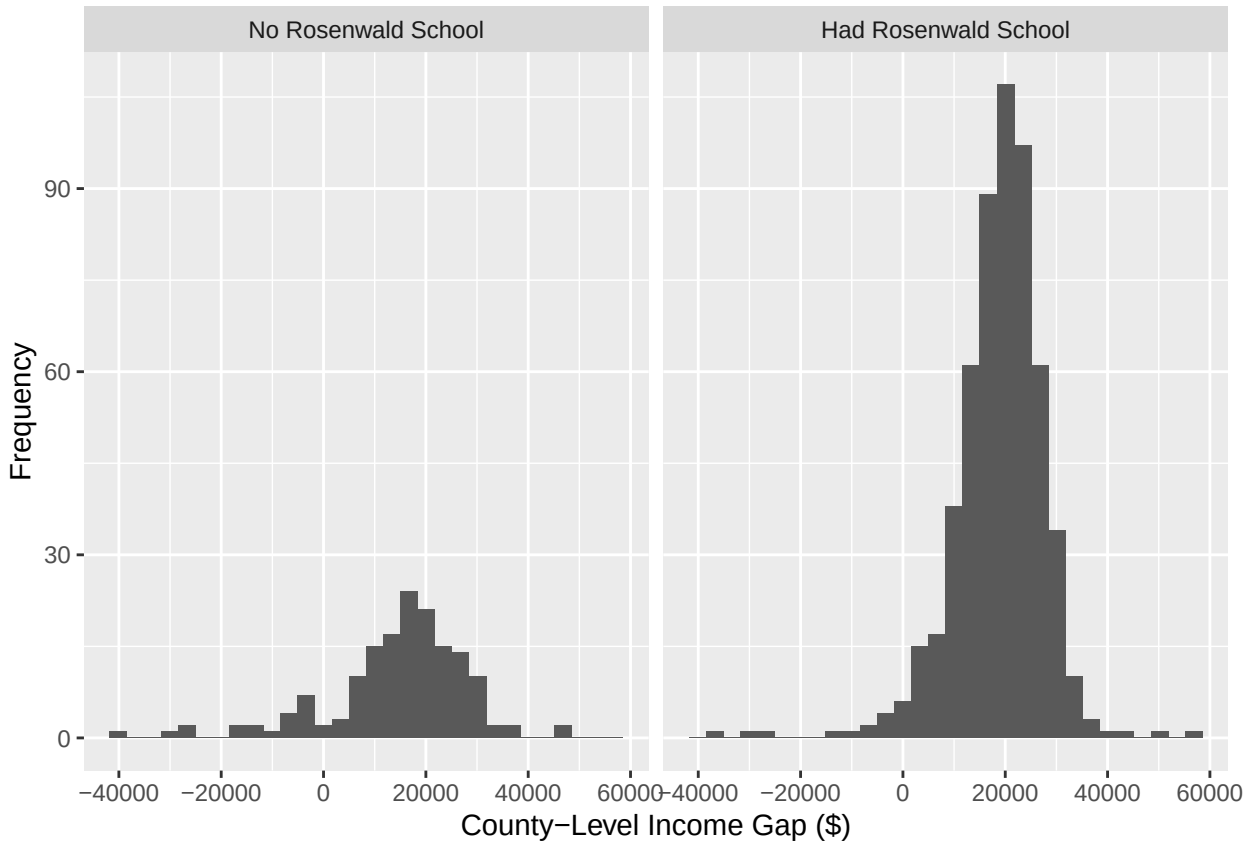
```
library(pacman)
p_load(dplyr, tidyverse, stargazer, fixest, magrittr, kableExtra, scales)

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)

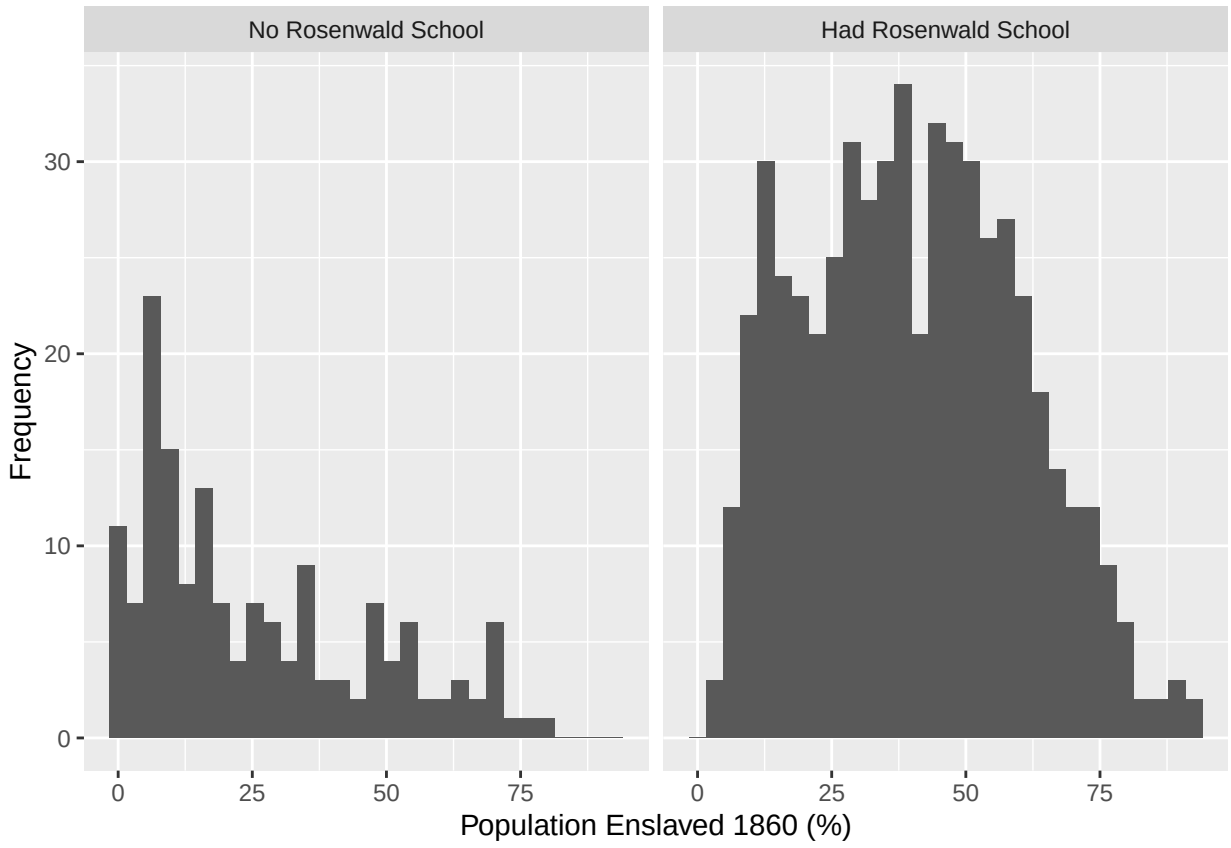
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
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")
```

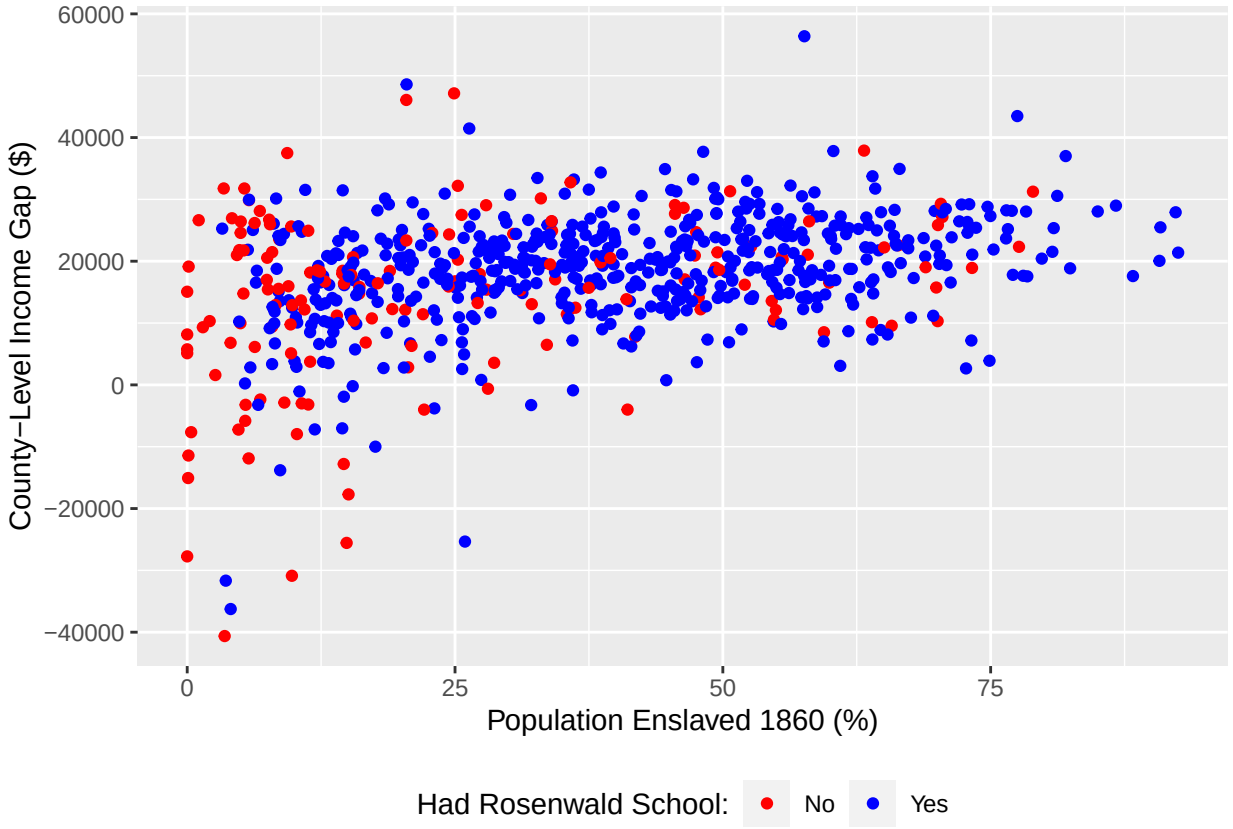
A histogram of the percent of the population in 1860 that was enslaved (pct_pop_enslaved_1860) also split by whether the county had a Rosenwald school (had_rosenwald_school)

```
## `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.

Question 4

```
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)
```

	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	✓

The conditional independence assumption required by this regression for a causal interpretation is that conditional on the state a county is in, potential outcomes of racial income gaps in that county are independent of whether or not that county had a Rosenwald School.

Question 5

```
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	✓

Question 6

If these new variables (pct_pop_enslaved_1860 & pop_total_1860) were causing bias in the regression in question 4, then I would suspect them to be upward biasing factors – since Rosenwald schools were located in

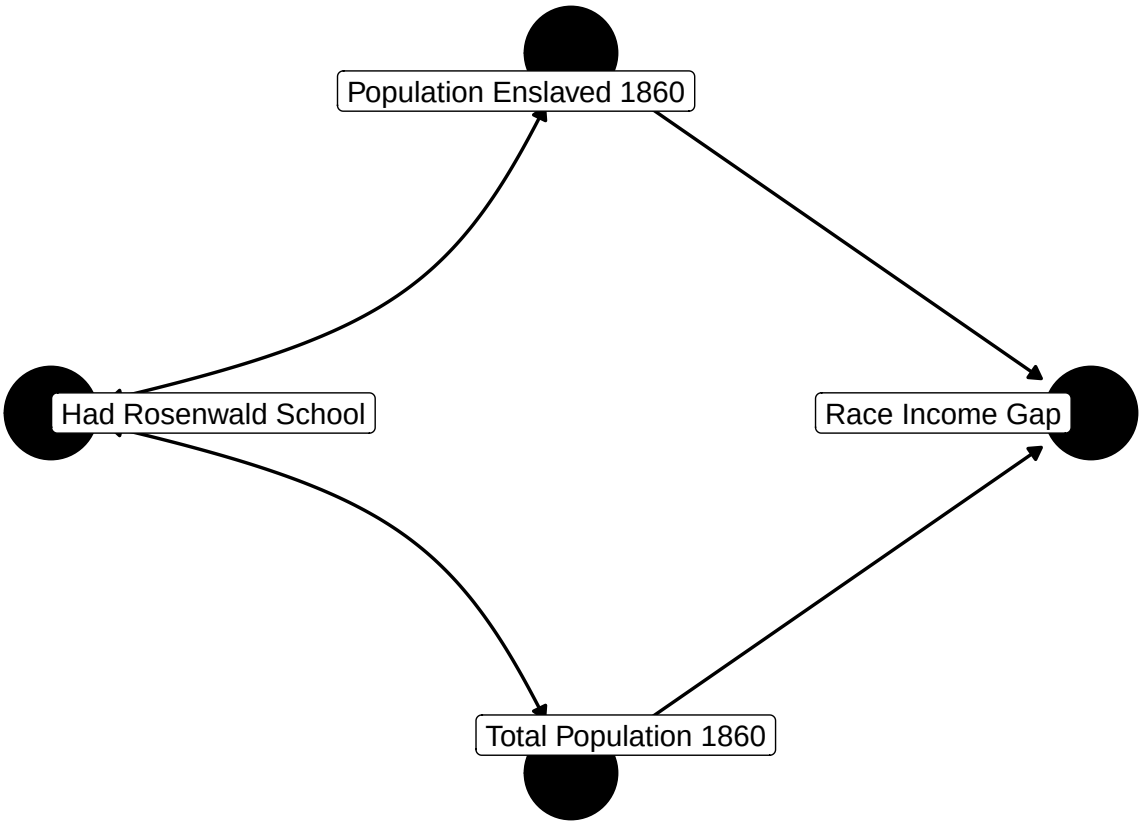
areas with high Black populations (as shown in the histograms above), which will likely also be highly populated areas in general – making both of these new variables positively correlated with had_rosenwald_school. Additionally, it makes sense that they would also correlate positively with the race income gap. Since the estimated coefficient on had_rosenwald_school dropped dramatically after adding the new variables into the regression, the omitted variables were biasing the estimate positively, which matches my intuition.

Question 7

The conditional independence assumption required by the regression in question 5 is that conditional on the percent of population of a county that was in enslaved in 1860 and the total population of that county, potential outcomes of racial income gaps are independent of whether or not that county had a Rosenwald school.

I do not believe this is a valid assumption.

```
p_load(ggdag)

coordinates = list(
  x = c(X = 0, Z = 1, W = 1, Y = 2), # Node name = x coordinate for that node
  y = c(X = 1, Z = 0, W = 2, Y = 1) # Node name = y coordinate for that node
)

dag_7.1 = dagify(W ~~ X,
                 Z ~~ X,
                 Y ~ W + Z,
                 coords = coordinates,
                 labels = c(Y = "Race Income Gap",
                            X = "Had Rosenwald School",
                            W = "Population Enslaved 1860",
                            Z = "Total Population 1860"))


ggdag(dag_7.1,
      text = FALSE,
      use_labels = "label") +
  theme_void()
```



This dag draws out the conditional independence assumption: *Had Rosenwald School* $\perp$ *Race Income gap* | *Population Enslaved 1860, Total Population Enslaved 1860*

By selecting these two variables, we open biasing paths. There is selection bias for which counties got the Rosenwald Schools, mainly those with high black populations, which were counties with high enslavement and total populations.

Question 8

Adding the county's 2010 population (Black, White, and/or total) as a control in the regression is a bad idea because there are numerous unclear causal relationships between these variables and our current controls and treatment. For example, a county with high enslavement in 1860 (therefore a high black population in 1860) is likely to have received a Rosenwald School, and it is likely more black people would have moved to those counties to receive education from those schools, increasing the black population in 2010. Therefore, we have an unclear alternating effect between three variables that will bias our estimates. This hurts our conditional independence assumption.

Question 9

```
# Turn had_rosenwald_school back to numeric (subtract 1 because it adds 1 for some reason)
data %<>% mutate(had_rosenwald_school = as.numeric(as.character(had_rosenwald_school)))



reg_9.1 = feglm(had_rosenwald_school ~ pct_pop_enslaved_1860 + pop_total_1860 + I(pct_pop_enslaved_1860
                data = data, family = 'logit')
```

```
data %<>% mutate(p_score = reg_9.1$fitted.values)
```

Estimate propensity scores

Question 10

```
# Regression using the p_score instead of the variables

reg_10.1 = feols(
  race_income_gap ~ had_rosenwald_school + p_score + pct_pop_enslaved_1860 + pop_total_1860 | state,
  data = data)
```

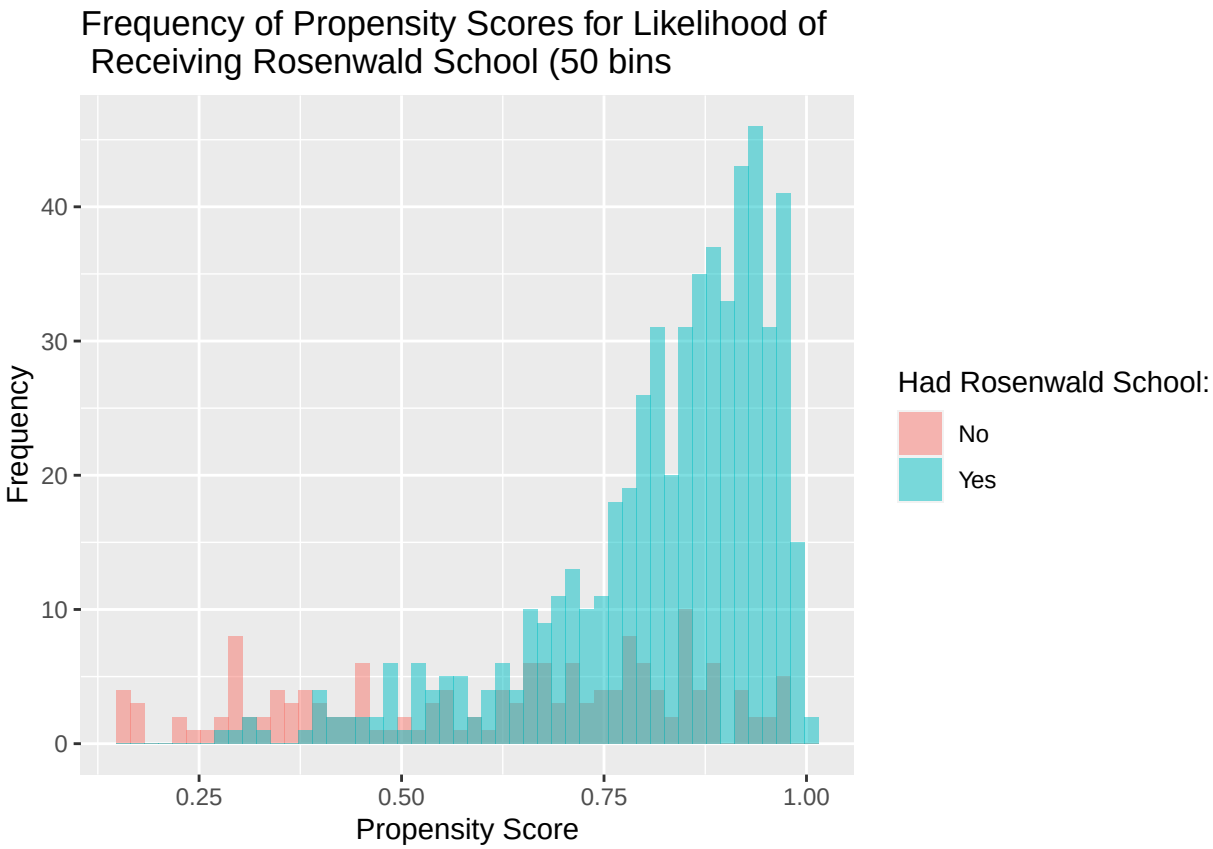
```
table10.1 = etable(reg_10.1,
                   style.tex = style.tex("aer"),
                   tex = TRUE)
```

	Race Income Gap
	(1)
Had Rosenwald School	1,626.8
	(928.2)
Propensity Score	7,464.5
	(4,561.5)
% Population Enslaved 1860	95.69***
	(23.67)
Total Population 1860	0.0619*
	(0.0317)
Observations	710
R^2	0.16806
Within R^2	0.10498
State Fixed Effects	✓

Using the propensity scores instead of the variables has decreased the estimate on the coefficient on had_rosenwald_school by around 600, meaning the effect of a county having Rosenwald schools on that county's race income gap was overestimated in previous regressions. Additionally, the standard errors on the coefficient on had_rosenwald_school has increased, and our estimate is no longer statistically significant.

Question 11

```
data %<>% mutate(color = ifelse(had_rosenwald_school == 1, "Yes", "No"))

ggplot(data) +
  geom_histogram(aes(p_score, fill = color),
                 position="identity", alpha = 0.5, bins = 50) +
  labs(fill = "Had Rosenwald School:",
       x = "Propensity Score",
       y = "Frequency",
       title = "Frequency of Propensity Scores for Likelihood of \n Receiving Rosenwald School (50 bins
```



Question 12

What percent of observations aren't supported by overlap

```
# Create function to assign p_score to a group

  # Create the groups by finding the min, max, and the steps
p_min = min(data$p_score)
p_max = max(data$p_score)
p_step = (p_max - p_min)/50

  # compile these into a handy dataframe
p_data = data.frame(step_max = seq(from = p_min, to = p_max, by = p_step)) %>%
            mutate(step_min = lag(step_max),
                   group = 0:50) %>%
            filter(group >= 1)

  # creating the function itself

find_p_group = function(X) {

  null_vec = logical(length=nrow(p_data))

  for (j in 0:nrow(p_data)) {

    x2_min = p_data$step_min[j]
    x2_max = p_data$step_max[j]
```

```
    if( isTRUE((X >= x2_min) && (X <= x2_max))) {
      null_vec[j] = as.logical("T")} else {
      null_vec[j] = as.logical("F")}
  }

  return(which(null_vec == T))
}



# Use function to create new variable that groups the p_scores into respective boxes
  # I'm sure there are better ways but they weren't working
  # The function works, but started giving me a warning for some curious reason.
  # The try() function supresses that message

try(for (i in 1:(nrow(data))) {
data$p_group[i] = find_p_group(data$p_score[i])
}, silent = T)


# Find the groups that have both had_rosenwald_school of 0 & 1
  # meaning the mean of the groups is between 0 and 1

match_groups = data %>%
          select(had_rosenwald_school, p_group) %>%
          group_by(p_group) %>%
          summarize(avg = mean(had_rosenwald_school)) %>%
          filter(avg > 0 & avg < 1) %>%
          select(p_group)

# Calculate the percent in those groups:

numerator = nrow(data %>% filter(p_group %in%match_groups$p_group))

denominator = nrow(data)

(1 - numerator/denominator) * 100
```

```
## [1] 7.323944
```

Therefore when using 50 groups, about **7.3%** of data points aren't supported by overlap. (When I used 30 groups, it was only 2.1%, and since I believe in erring on the side of caution, I will use this larger number of groups).

Question 13

We care about enforcing overlap because we want there to be a positive probability for a given individual (or county, in this case,) to be placed in either the treatment or control group. This gives us a stronger identification strategy, and rids some selection bias from our sample – as those who would always or never be treated are removed.

Question 14

```
data_overlap = data %>% filter(p_group %in% match_groups$p_group)

# Regression using the p_score instead of the variables

reg_14.1 = feols(
  race_income_gap ~ had_rosenwald_school + p_score + pct_pop_enslaved_1860 + pop_total_1860 | state,
  data = data_overlap)
```

```
table14.1 = etable(reg_14.1,
                   style.tex = style.tex("aer"),
                   tex = TRUE)
```

	Race Income Gap
	(1)
Had Rosenwald School	1,667.6*
	(908.1)
Propensity Score	8,296.3
	(4,958.1)
% Population Enslaved 1860	94.99***
	(21.06)
Total Population 1860	0.0340
	(0.0277)
Observations	658
R^2	0.16565
Within R^2	0.09757
State Fixed Effects	✓

Compared to the regression in question 10, our standard error for the estimate on had_rosenwald_school shrunk, and our estimate increased. The estimate is now statistically significant at the 10% significance level.

Question 15

```
# Running regression weighting by inverse of propensity scores

reg_15.1 = feols(
  race_income_gap ~ had_rosenwald_school + pct_pop_enslaved_1860 + pop_total_1860 | state,
  data = data_overlap,
  weights =  1/data_overlap$p_score)
```

```
table15.1 = etable(reg_15.1,
                   style.tex = style.tex("aer"),
                   tex = TRUE)
```

	Race Income Gap
	(1)
Had Rosenwald School	2,233.0
	(1,349.0)
% Population Enslaved 1860	135.5***
	(21.81)
Total Population 1860	0.0835*
	(0.0425)
Observations	658
R^2	0.15906
Within R^2	0.08687
State Fixed Effects	✓

Our estimate on had_rosenwald_school increased, and its standard error increase. The estimate is once again not statistically significant.

Question 16

```
# Going to use just 20 bins

  # Create the groups by finding the min, max, and the steps
p_min2 = min(data$p_score)
p_max2 = max(data$p_score)
p_step2 = (p_max - p_min)/20

  # compile these into a handy dataframe
p_data2 = data.frame(step_max = seq(from = p_min2, to = p_max2, by = p_step2)) %>%
            mutate(step_min = lag(step_max),
                   group = 0:20) %>%
            filter(group >= 1)

  # creating the function itself

find_p_group2 = function(X) {

  null_vec = logical(length=nrow(p_data2))

  for (j in 0:nrow(p_data2)) {

    x2_min = p_data2$step_min[j]
    x2_max = p_data2$step_max[j]

    if( isTRUE((X >= x2_min) && (X <= x2_max))) {
      null_vec[j] = as.logical("T")} else {
      null_vec[j] = as.logical("F")}
  }

  return(which(null_vec == T))
}


try(for (i in 1:(nrow(data))) {
```

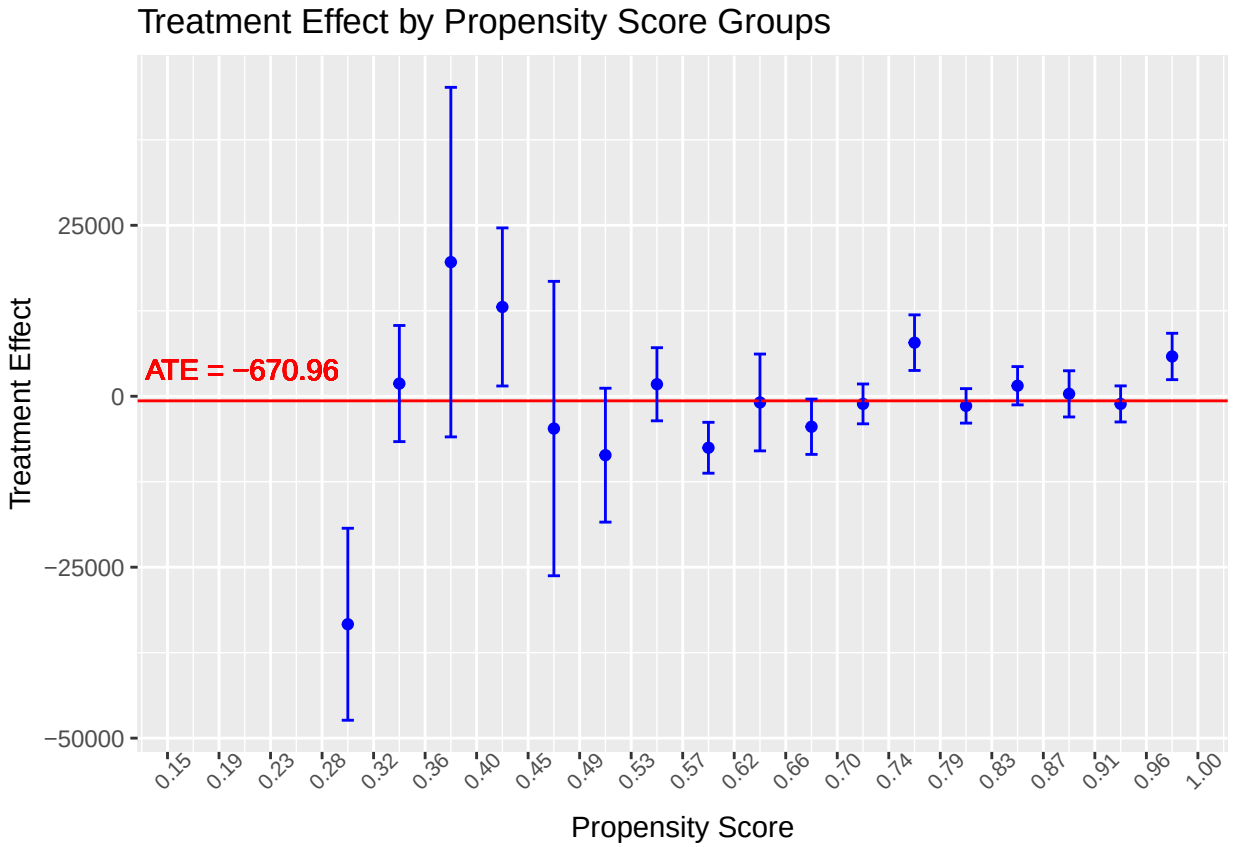
```
data$p_group2[i] = find_p_group2(data$p_score[i])
}, silent = T)


# Find the groups that have both had_rosenwald_school of 0 & 1
  # meaning the mean of the groups is between 0 and 1

match_groups2 = data %>%
          select(had_rosenwald_school, p_group2) %>%
          group_by(p_group2) %>%
          summarize(avg = mean(had_rosenwald_school)) %>%
          filter(avg > 0 & avg < 1) %>%
          select(p_group2)

data_overlap2 = data %>% filter(p_group2 %in% match_groups2$p_group2)
```

```
# Creating data frame for the graph

data_q16 = match_groups2 %>%
            mutate(treatment_effect = NA,
                   standard_error = NA) %>%
            rename(group = p_group2)

# Regressions for each group

for (i in 1:nrow(data_q16)) {

  j = data_q16$group[i]

  data_group = data_overlap2 %>% filter(p_group2 == j)

  reg_16.j = feols(race_income_gap ~ had_rosenwald_school + pct_pop_enslaved_1860 + pop_total_1860 | sta
                data = data_group)

  data_q16$treatment_effect[i] = reg_16.j$coefficients[[1]]
  data_q16$standard_error[i] = reg_16.j$se[[1]]
}

# Compiling and refining data fro the graph

data_q16.2 = right_join(data_q16, p_data2, by = "group")

data_q16.2 %<>% mutate(p_score = (data_q16.2$step_max + data_q16.2$step_min)/2) %>%
                mutate(upper = treatment_effect + standard_error,
                       lower = treatment_effect - standard_error)

# Calculating Average Treatment Effect
ATE.16 = data_q16.2 %>% select(treatment_effect) %>% summarize(ATE = mean(treatment_effect, na.rm = T))

ATE = ATE.16$ATE[1]

# The graph!
```

```
ggplot(data_q16.2, aes(x = p_score, y = treatment_effect)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = ATE, color = "red") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width=.01, color = "blue") +
  scale_x_continuous(breaks = seq(from = p_min2, to = p_max2, by = p_step2),
                     labels = label_number(accuracy = 0.01)) +
  theme(axis.text.x = element_text(size = 8, angle = 45)) +
  labs(x = "Propensity Score", y = "Treatment Effect",
       title = "Treatment Effect by Propensity Score Groups") +
  geom_text(aes(0.21, ATE, label = "ATE = -670.96", vjust = -1), color = "red")
```

Regression Part

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



Q

I think that were we to employ a 'selection on unobservables' strategy by using *pct_pop_enslaved_1860* as an instrument for whether a county received a Rosenwald school (*had_rosenwald_school*), we might run into a few similar issues as seen in this problem set above. We would desire the use of an instrumental variable if the were disturbances affecting both *had_rosenwald_school* and *race_income_gap* were also unassociated with *pct_pop_enslaved_1860*, as that is a requirement for using instrumental variables. However, I find it difficult to believe that these disturbances between our treatment and outcome would be uncorrelated with this instrument, given how closely tied the history of systemic racism is to all of these variables. Therefore, I do not believe this strategy would be a stronger identification strategy than what has been done here.