Program Eval Final

2023-05-19

```
knitr::opts_chunk$set(echo = TRUE)
```

QUESTION 1 1A We want to answer the question: How does having children impact gender inequality in the labor market?

The naive estimator in this case would be to subtract the mean income of men who have children from women who have children.

t = Yi(Di = 1) - Yi(Di = 0) Where the treatment effect (t) is the comparison of a man's income (Yi) with children (Di = 1) with children to a woman's income (Yi) without children (Di = 0). The difference would hypothetically tell us how much more having kids impacts women then men.

The naive estimator is unlikely to provide an unbiased estimate because of selection bias, some specific example might be:

1) Women might plan differently for careers based on knowing they want to have kids.
2) Women have to take more time off to give birth and recover which leads to a reduction in income.
3) Women might switch to a flexible job after giving birth so they can spend more time with their children– jobs that offer this flexibility may have less income and growth opportunities.

These potential selection issues will lead to an upward bias relative to the truth because we are going to attribute more the of inequality to gender inequality than might actually be the case.

The ideal experiment would be to run a RCT by randomizing men and women with children into jobs and then comparing their incomes to measure the true treatment effect without picking up the confounding variables.

1B Yistg = $\sum_{j \neq -1} \alpha_{j\,g} \cdot \mathbf{I}[j = t] + \sum_k \beta_{k\,g} \cdot \mathbf{I}[k = age_{is}] + \sum_y \gamma_{y\,g} \cdot \mathbf{I}[y = s] + \nu_{ist\,g}$ We run this regression for both men and women to recover the income estimate for each separatley with the outcome of interest for individual i of gender g in year s at event (time) t. We use time, age, and year dummies (the three terms, in that order) to compare incomes before and after childbirth for men and women. Our dummies help us control for external factors besides having your first kid (i.e. ages that men and women have children, inflation, and other time varying things that might bias our estimate).

The assumptions required is that the counter-factual trend is zero, in other words, the only thing changing is treatment status. When measuring the treatment effect it is important we measure only the treatment and not pick up other covariates. By controlling for age, year,

and event time, we try to ensure the only thing changing at this time of having your first kid is having the kid and not also a bunch of external events.

With our controls, the assumption that we have a reliable counterfactual is likely satisfied. With the time dummies, we control for external life-cycle trends (other events that could be different at that time). With the year dummies we control for time trends such as wage inflation and business cycles. By controlling for age we control for the fact that women usually have their first child younger than men (otherwise we could pick up on women having a lower wage at the time of their first kid on their age being lower instead of gender inequality).

1C When we look at figure 2: impact of children in the very long run, we see that female earnings, hours worked, participation rats, and wage rates are all lower in the long run, years after having their first kid.

Firstly, it is important to note that since hours worked and participation rates fall, it follows that earnings logically drop in response. However, earnings fall much more than the other 3 panels observed. This means that although some of the drop in wages is just logistics, there is a chunk of that discrepancy coming from something else. According to this paper, that a part of this discrepancy is due to gender inequality that can be attributed to the dynamic effects of children. The female child penalty is 20% in the long run while men are virtually unaffected. Although there may still be discrimination, it seems like having children explains the gap in wages quite a bit. Family background can shape the type of careers and career trajectories women pursue, but when it comes to equal pay for the same jobs, it seems the disparity between men and women comes down to having kids.

Depending on your view on gender capacities and specialization, this may not be an issue that calls for policy intervention. Women may chose to focus their energy on child-rearing and not put as much into their career and therefore end up with lower wages. However, this disparity may call for intervention if the population believes it is due to pure discrimination and not a difference in quality of work.

QUESTION 2 2A To measure the effect of monsoon predictions on the share of land in cash crops at the farm level, we have the potential outcome framework: $t = Yi(Di = 1) - Yi(Di = 0)$ Where the treatment effect of receiving monsoon predictions is the share of land in cash crops with predictions subtracted by the share of land in cash crops without predictions. The unit of analysis here is a farm.

The ideal experiment to run here is to run an RCT and randomly assign farms to the treatment and control groups where the treatment group gets a monsoon forecast and the control group doesn't get a forecast.

2B Yes, we should be concerned about randomizing at the farm level due to spillover. It's easy for neighbors to share information and so the forecast would likely contaminate our control group. If our control group was also operating off the information (treatment) given to the treatment group, we couldn't recover the true treatment effect of monsoon predictions on farm profits.

To solve this issue we could randomize at a higher level of aggregation, like randomizing at the village level instead of at the farm level. This way there's less chance that the forecast information would be leaked from the treatment to the control group. This way we are more likely to satisfy SUTVA, an assumption required for the RCT to recover the causal effect.

2C The RCT design we would use to measure the spillover is the randomized saturation design. It's a 2 step design to estimate spillovers where we first randomize clusters into treatment intensities (including pure control) to compare high versus low intensity places, and then we randomize units within clusters, which allows us to compare treatment versus control units.

There would be 5 arms: the high intensity cluster with treatment (A1) and control (A2) arms, and the low intensity cluster with treatment (B1) and control (B2) arms, and a pure control arm.

We compare the treatment groups and control groups within both the high intensity and low intensity clusters to estimate the following treatment parameters: a) ITT (Intent-to-Treat): Measure the average treatment effect by comparing the outcomes of farms that received forecasts (Treatment Groups A1 and B1) to those that did not (Control Group). b) SNT (Spillover on the Non-Treated): Measure the average spillover effect on farms that did not receive forecasts (Control Group) by comparing them to farms that received forecasts within the same cluster. c) TCE (Treatment on the Treated): Measure the average treatment effect on farms that received forecasts (Treatment Groups A1 and B1) by comparing them to farms that did not receive forecasts within the same cluster. By implementing this randomized saturation design, we can estimate the ITT, SNT, and TCE treatment parameters, allowing us to analyze the impacts of providing forecasts on cash cropping and farm profits for both treated and untreated farms in different clusters.

2D The regression to estimate these treatment parameters is: Yic = a + sum p=/0 tTRT Dic * 1[pc = p] + sum p=/0 tCTRL Sic * 1[pc = p] + eic Where: Yic represents the outcome variable for unit i in group c. a is the intercept term. tTRT is the treatment effect coefficient, capturing the average effect of receiving the monsoon forecast. Dic is the indicator variable that equals 1 if unit i in group c is in the treatment group (received the monsoon forecast), and 0 otherwise. tCTRL is the control effect coefficient, representing any differences in outcomes among control units (did not receive the forecast). Sic is the indicator variable that equals 1 if unit i in group c is in the control group (did not receive the forecast), and 0 otherwise. eic represents the error term. Every group is compared to pure controls.

We can get the following parameters of interest to estimate the spillover effects: t^ITT (p) = t^trt_p t^SNT(p) = t^ctrl_p t^TCE(p) = pt^trt_p + (1-p)t^ctrl_p

tTRT: This coefficient estimates the average treatment effect of receiving the monsoon forecast. A positive and statistically significant tTRT indicates that the forecast has a beneficial impact on the outcome variable.

tCTRL: This coefficient estimates any differences in outcomes between control units. A significant tCTRL suggests that factors other than the forecast program may be influencing the outcome.

To recommend whether FINALEXAM should scale their forecast program, we would consider the magnitude and statistical significance of the treatment effect coefficient (tTRT). If tTRT is statistically significant and positive, it suggests that the monsoon forecast has a beneficial impact on the outcome variable. We also need to consider the cost-effectiveness and practical implications of scaling up the program.

QUESTION 3 3A Comparing the average number of tax evaders in the 50% tax rate and the 15% tax rate to measure the effect of higher tax rates is using the naive estimator:

t = Yi(Di = 1)- Yi(Di = 0) Where the effect of having a high tax rate is the difference between the number of tax evaders in the 50% bracket and the 15% bracket.

The comparison could estimate the effect if the people in both brackets were identical in all baseline characteristics besides their property tax rate.

However this is likely not true and will lead to selection bias, 2 reasons why this may be problematic is:

1) People who have a higher tax rate might be more inclined to evade taxes since they will lose more.
2) People who have a higher tax bracket probably have more money than those in a lower tax bracket and therefore might have better lawyers and a higher risk tolerance and will therefore be less careful to avoid tax evasion.

3B MAROONS is offering us data to run a selection on observables experiment. By giving us a bunch of potential omitted variables to control for, we can theoretically control for the covariates in the error term that would lead to a biased result and recover the causal effect of having a higher marginal tax rate on tax evasion.

The conditions for this to be true are that these are the only covariates and there are no other observable characteristcs that differ between the 50% bracket group and the 15% bracket group. The other condition is there are no unobservable ways these 2 groups are different. If these 2 conditions are met, than selction on obsetvables would yeild the causal effect.

However, that is very unlikely to be true and thus amkes this experiment design very weak. 2 reasons why this design is problematic for us are:

3C I agree that they can get us a better estimate but they can't fully solve the issue of using selection on observables.

Machine learning can't recover the treatment effect but it can help us choose the covariates (Xs) that are important for the outcome (Y). We predict Y^ as a function of X, predict D^ as a function of X, compute the residuals, and then recover the treatment effect by regressing: Yi^R = a + tD^r + ei

This approach still requires the selection on observables assumptions to hold but we get a better understanding of which covariates to control for. This SOO design is still weak but with ML, we could estimate the treatment effect better than we could without them. It still can't recover the unobservables since those aren't things we can measure and control for which leaves a big gap for OVB.

3D The research design we could use is a regression discontinuity, with 1978 as the cutoff. We would zoom in and measure the difference between tax evasion right at as people crossed that cutoff to see if it's truly the higher tax rate that causes people to evade taxes. We can use this as a test case.

Since there is clear cutoff when treatment turns of (jumping from 15% to 50% means going from 0% probability of treatment to a 100% of treatment) means we run a sharp RD. We run the following regression: t^SRD = E[Yi(1) - Yi(0) | Xi = c] meaning the treatment effect at the cutoff is the expected value of tax evaders at the 50% tax rate minus the expected value of tax evaders at the 15% tax rate at the cutoff.

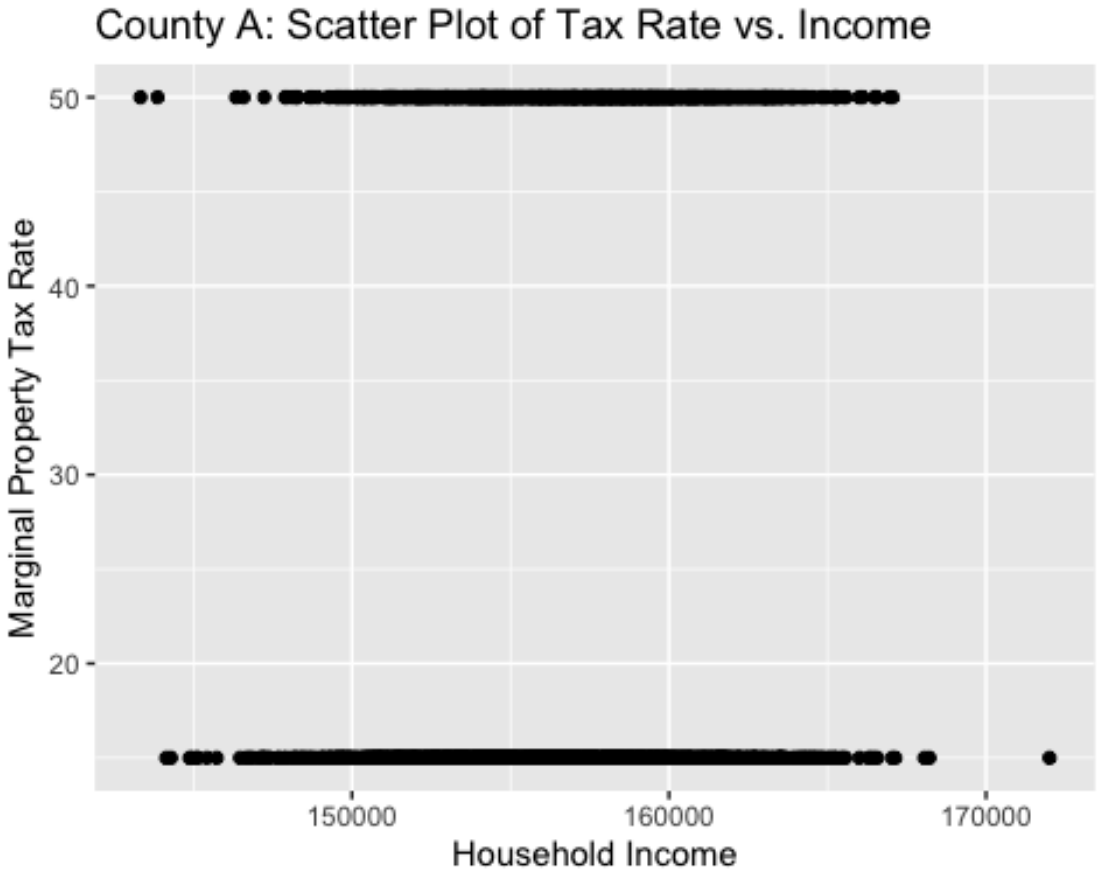
We only need 1 identifying assumption for a sharp RD: all observed and unobserved determinants of tax evasion (besides the tax rate jumping from 15% to 50%) are smooth around the cutoff (i.e. all the covariates are smooth between the pre and post treatment periods). Mathematically: E[Yi(1) | Xi = x] and E[Yi(0) | Xi = x] are continuous.

Our sharp RD gives us the LATE meaning the treatment effect for the crossing that threshold.

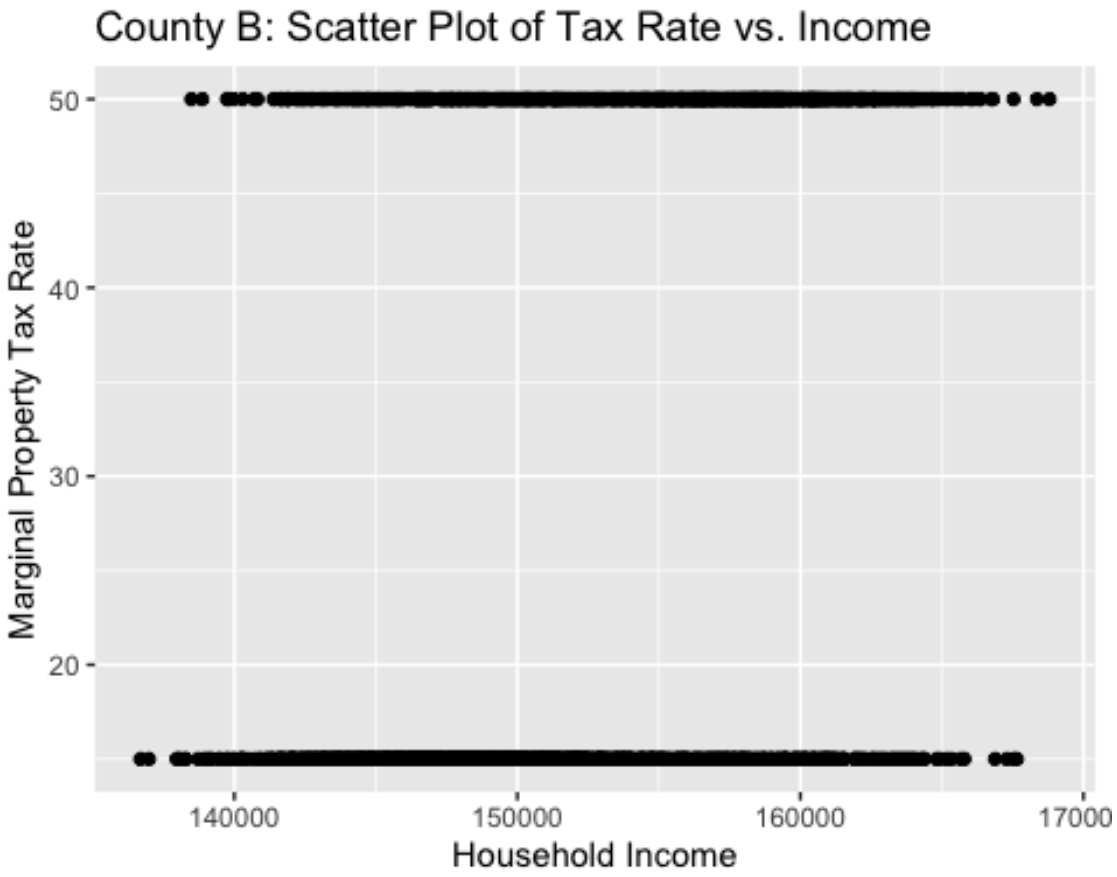
3E County B(santa clara) doesn't have a normal distribution, so estimating the causal effect of having a high marginal property tax rate on tax evasion becomes more complex. We may need to consider subgroup analyses or account for heterogeneity with an instrumental variable.

```
setwd("/Users/misbaharshad/Downloads/")
data <- read.csv("final_exam_2023.csv")

library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

county_A <- data %>% filter(county == "ALAMEDA")
county_B <- data %>% filter(county == "SANTA CLARA")
county_C <- data %>% filter(county == "YOLO")
```

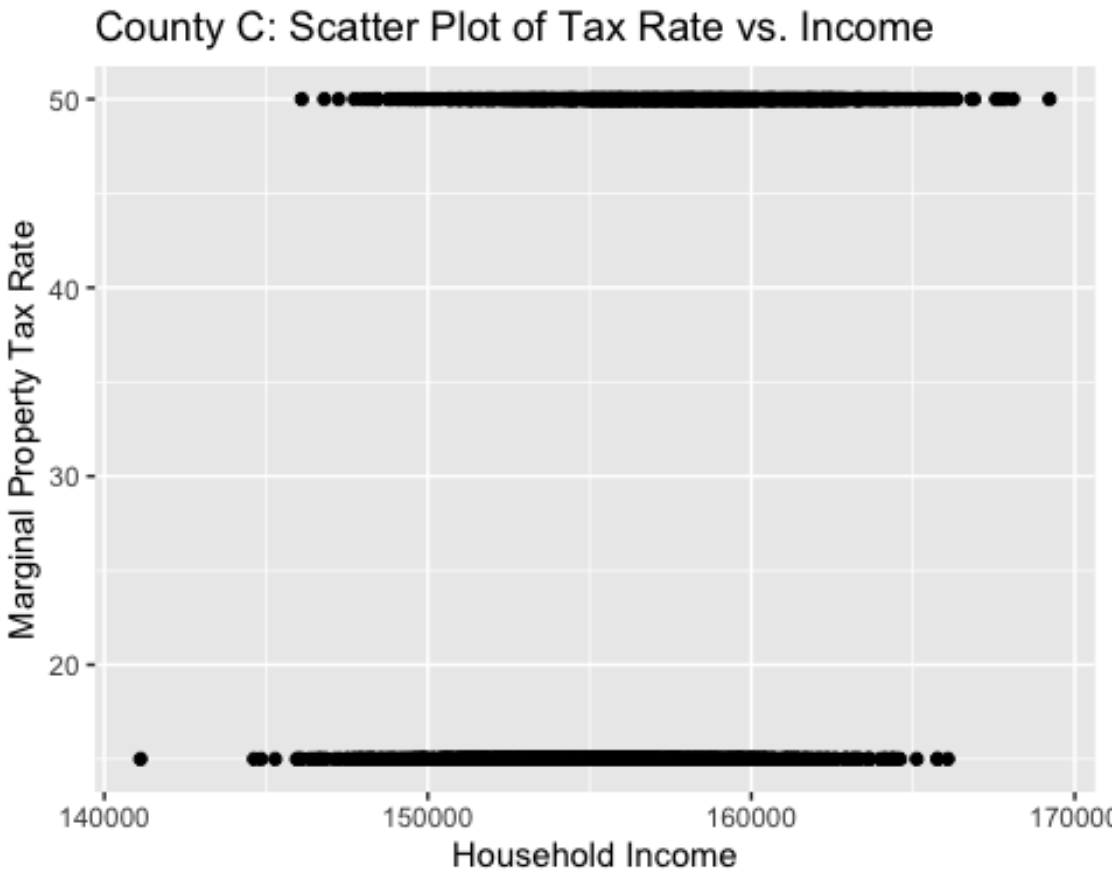
```
ggplot(county_A, aes(x = household_income, y = marginal_property_tax_rate)) +
  geom_point() +
  labs(title = "County A: Scatter Plot of Tax Rate vs. Income",
       x = "Household Income",
       y = "Marginal Property Tax Rate")
```



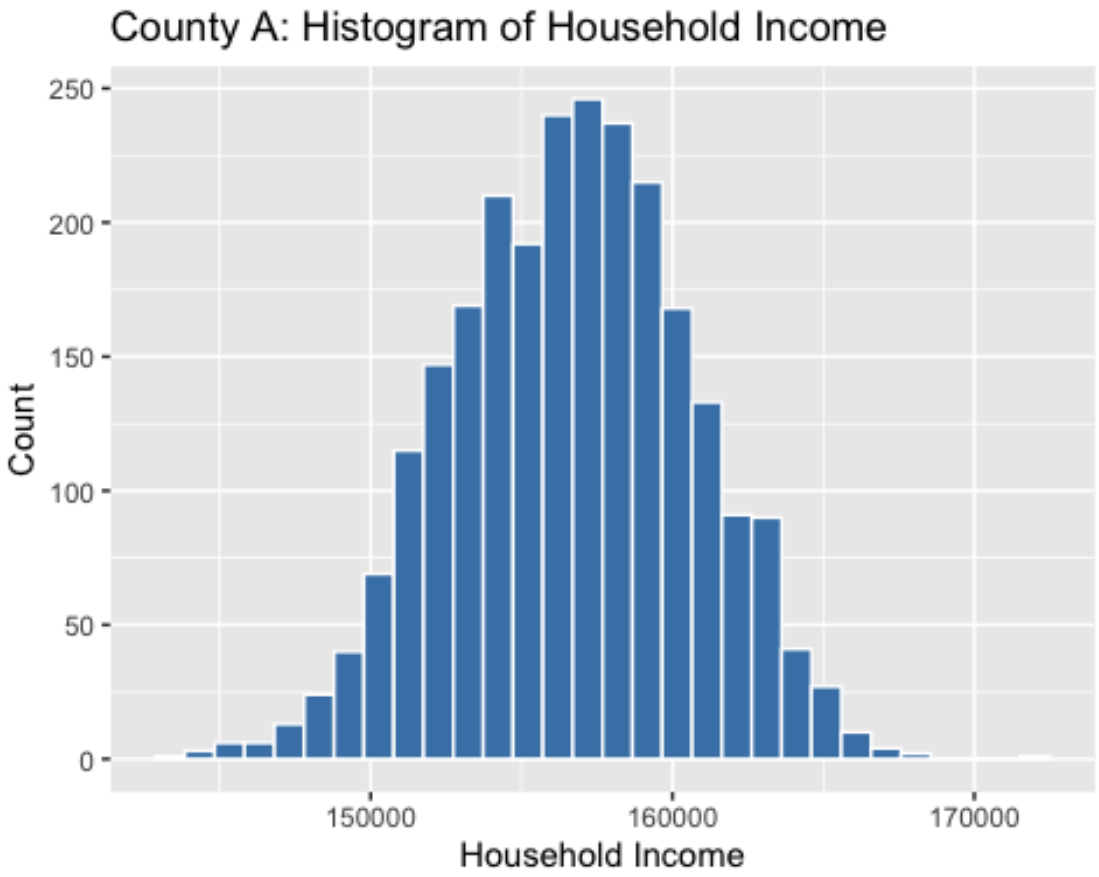

```
ggplot(county_B, aes(x = household_income, y = marginal_property_tax_rate)) +
  geom_point() +
  labs(title = "County B: Scatter Plot of Tax Rate vs. Income",
       x = "Household Income",
       y = "Marginal Property Tax Rate")
```



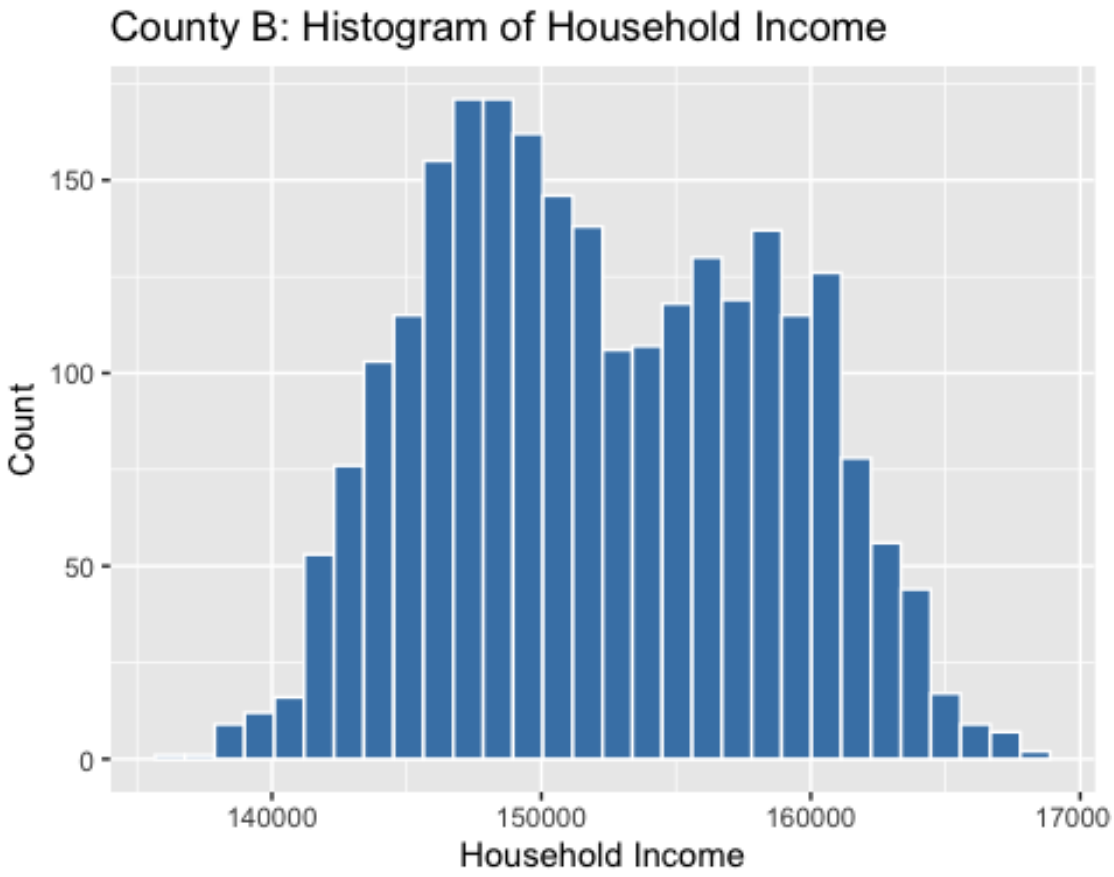

```
ggplot(county_C, aes(x = household_income, y = marginal_property_tax_rate)) +
  geom_point() +
  labs(title = "County C: Scatter Plot of Tax Rate vs. Income",
       x = "Household Income",
       y = "Marginal Property Tax Rate")
```



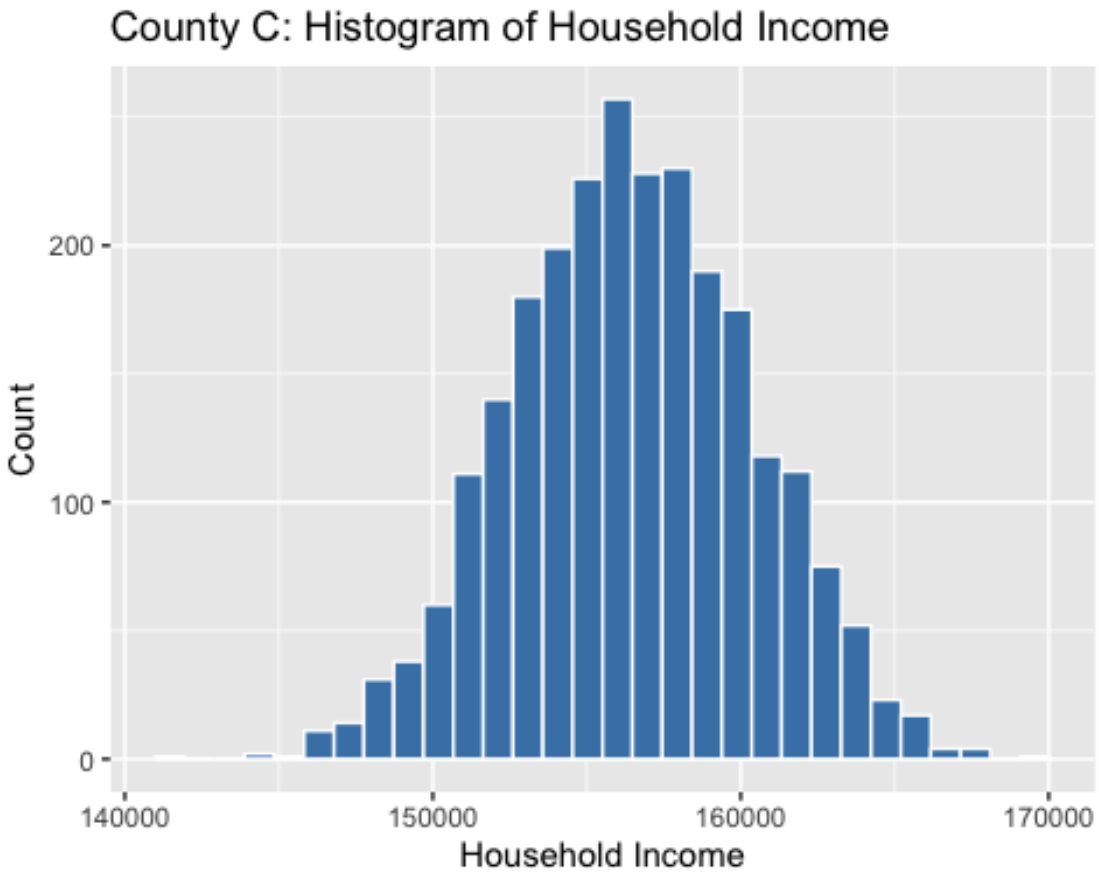

```
ggplot(county_A, aes(x = household_income)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  labs(title = "County A: Histogram of Household Income",
       x = "Household Income",
       y = "Count")
```




```r
ggplot(county_B, aes(x = household_income)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  labs(title = "County B: Histogram of Household Income",
       x = "Household Income",
       y = "Count")
```

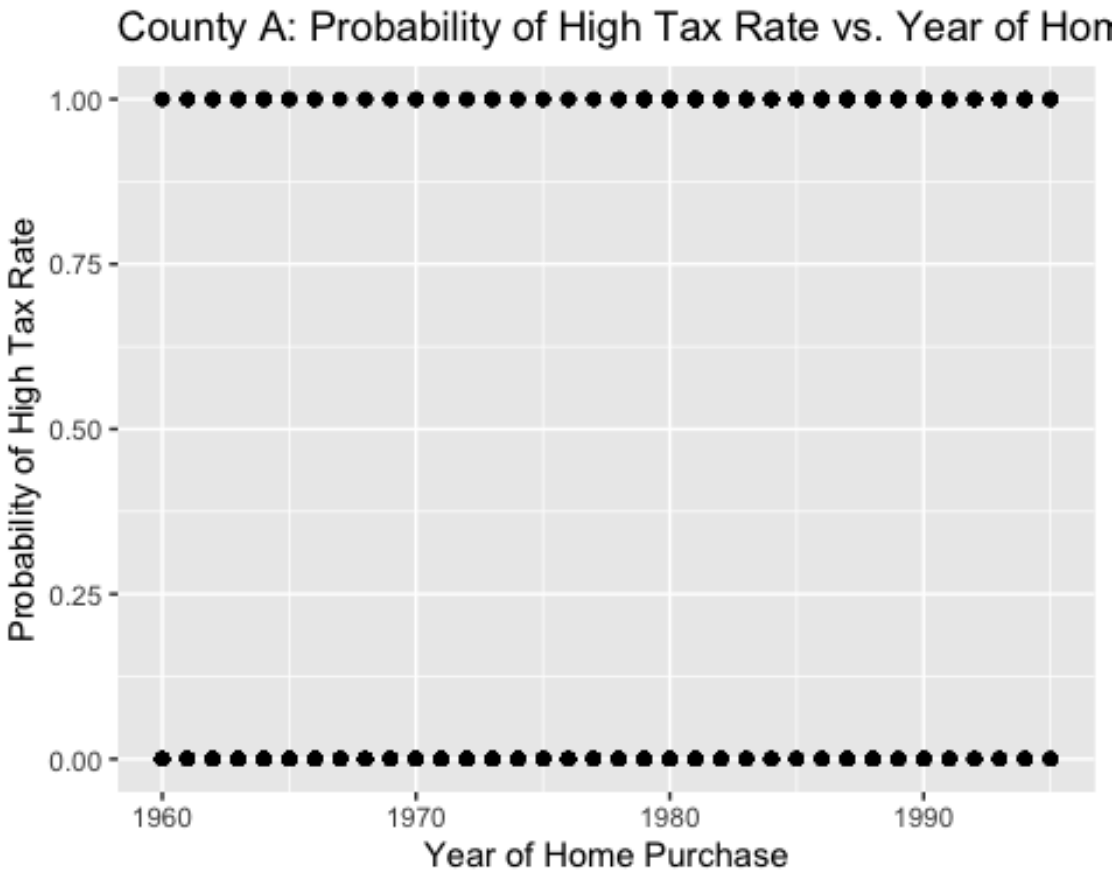



```
ggplot(county_C, aes(x = household_income)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  labs(title = "County C: Histogram of Household Income",
       x = "Household Income",
       y = "Count")
```

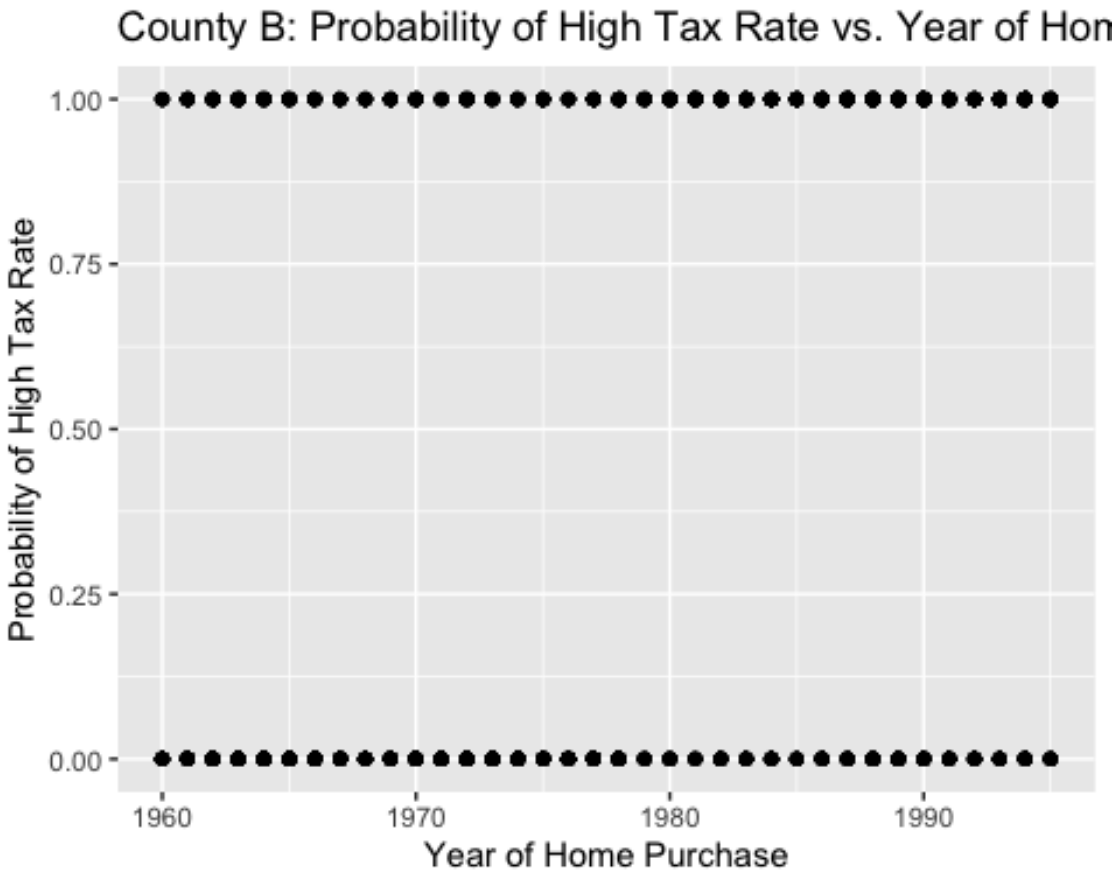


3F

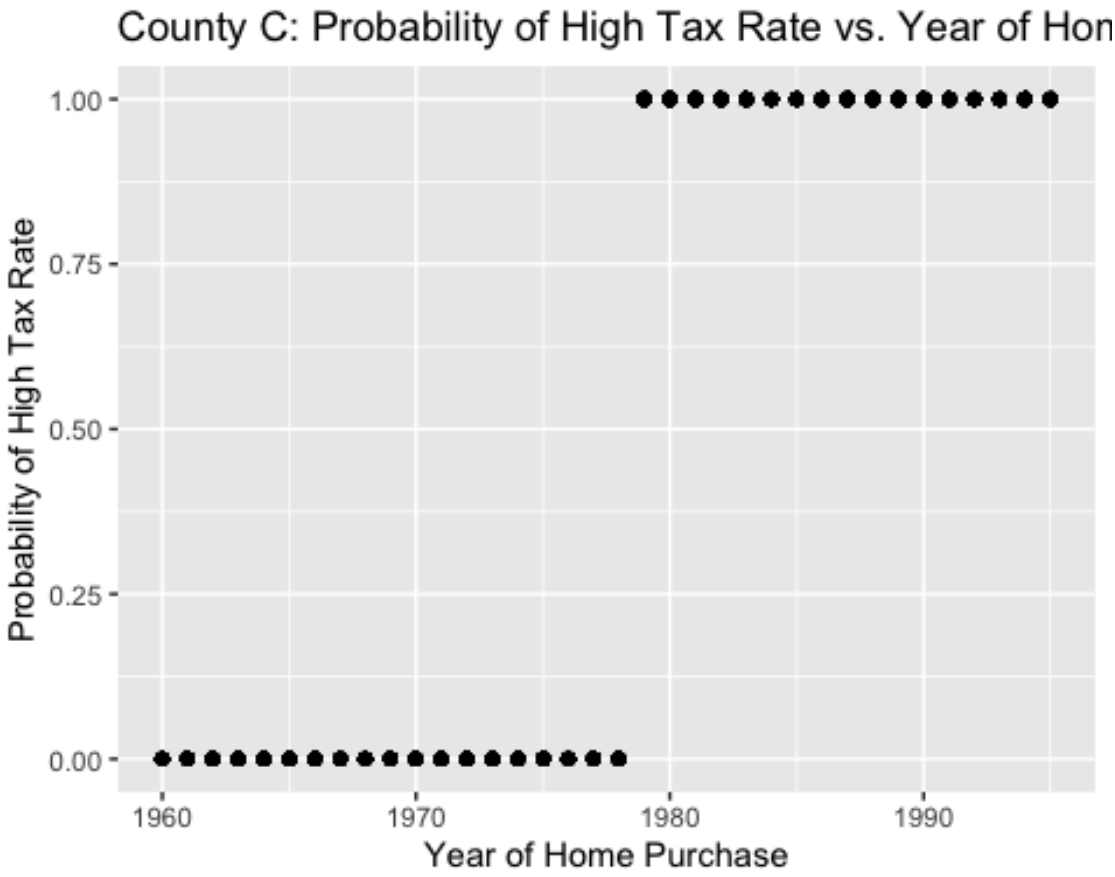
```
county_A <- data %>% filter(county == "ALAMEDA")
county_B <- data %>% filter(county == "SANTA CLARA")
county_C <- data %>% filter(county == "YOLO")

county_A <- county_A %>% mutate(high_tax_dummy =
ifelse(marginal_property_tax_rate >= 50, 1, 0))

county_B <- county_B %>% mutate(high_tax_dummy =
ifelse(marginal_property_tax_rate >= 50, 1, 0))

county_C <- county_C %>% mutate(high_tax_dummy =
ifelse(marginal_property_tax_rate >= 50, 1, 0))

ggplot(county_A, aes(x = year_of_home_purchase, y = high_tax_dummy)) +
  geom_point() +
  labs(title = "County A: Probability of High Tax Rate vs. Year of Home
Purchase",
       x = "Year of Home Purchase",
       y = "Probability of High Tax Rate")
```




```
ggplot(county_B, aes(x = year_of_home_purchase, y = high_tax_dummy)) +
  geom_point() +
  labs(title = "County B: Probability of High Tax Rate vs. Year of Home
Purchase",
       x = "Year of Home Purchase",
       y = "Probability of High Tax Rate")
```




```
ggplot(county_C, aes(x = year_of_home_purchase, y = high_tax_dummy)) +
  geom_point() +
  labs(title = "County C: Probability of High Tax Rate vs. Year of Home
Purchase",
       x = "Year of Home Purchase",
       y = "Probability of High Tax Rate")
```



3G For County A (ALAMEDA): The estimated coefficient for the high_tax_dummy variable is 0.3511, and it is statistically significant ($p < 2e\text{-}16$). This suggests that having a high marginal property tax rate (50%) is associated with an increase in the probability of tax evasion. The positive coefficient indicates that as the tax rate increases, the likelihood of tax evasion also increases.

For County B (SANTA CLARA): The estimated coefficient for the marginal_property_tax_rate variable is 0.0138, and it is statistically significant ($p < 2e\text{-}16$). This implies that there is a positive relationship between the marginal property tax rate and tax evasion. As the tax rate increases, the probability of tax evasion also increases.

For County C (YOLO): The estimated coefficient for the marginal_property_tax_rate variable is 0.0128, and it is statistically significant ($p < 2e\text{-}16$). This suggests that there is a positive association between the marginal property tax rate and tax evasion in County C as well. Higher tax rates are associated with a higher likelihood of tax evasion.

Overall, the results indicate that there is evidence of a positive causal effect of high marginal property tax rates on tax evasion in County A (ALAMEDA), County B (SANTA CLARA), and County C (YOLO). This implies that increasing the tax rate for these counties may lead to an increase in tax evasion behavior.

Considering the results, MAROONS may have some evidence to support their concerns about wealthy households not paying their fair share of property taxes. However, advocating for an increase in the tax rate for all homeowners based solely on these findings may not be recommended. It would be more appropriate for MAROONS to use these results as a starting point for further investigation and policy discussion.

```
county_A_rdd <- lm(evades_taxes_yn ~ high_tax_dummy, data = county_A)

summary(county_A_rdd)

coeftest(county_A_rdd, vcov. = sandwich)


county_B_lm <- lm(evades_taxes_yn ~ marginal_property_tax_rate, data =
county_B)

summary(county_B_lm)

coeftest(county_B_lm, vcov. = sandwich)


county_C_lm <- lm(evades_taxes_yn ~ marginal_property_tax_rate, data =
county_C)

summary(county_C_lm)

coeftest(county_C_lm, vcov. = sandwich)
```

```
Call:
lm(formula = evades_taxes_yn ~ high_tax_dummy, data = county_A)

Residuals:
   Min     1Q  Median     3Q    Max
-0.2288 -0.2288 -0.2288 -0.2288  0.7712

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value
(Intercept)    0.228800   0.008403   27.23
high_tax_dummy      NA         NA      NA
             Pr(>|t|)
(Intercept)     <2e-16 ***
high_tax_dummy      NA
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4201 on 2499 degrees of freedom


t test of coefficients:

             Estimate Std. Error t value
(Intercept) 0.2288000  0.0084012  27.234
```

```
          Pr(>|t|)
(Intercept) < 2.2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Call:
lm(formula = evades_taxes_yn ~ marginal_property_tax_rate, data = county_B)

Residuals:
   Min     1Q Median     3Q    Max
-0.483  0.000  0.000  0.000  0.517

Coefficients:
                       Estimate Std. Error
(Intercept)             -0.2070123  0.0138516
marginal_property_tax_rate  0.0138008  0.0003886
                      t value Pr(>|t|)
(Intercept)             -14.95   <2e-16 ***
marginal_property_tax_rate   35.52   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3389 on 2498 degrees of freedom
Multiple R-squared:  0.3355,             Adjusted R-squared:  0.3352
F-statistic:  1261 on 1 and 2498 DF,  p-value: < 2.2e-16


t test of coefficients:

                       Estimate Std. Error
(Intercept)             -0.2070123  0.0063181
marginal_property_tax_rate  0.0138008  0.0004212
                      t value  Pr(>|t|)
(Intercept)             -32.765 < 2.2e-16 ***
marginal_property_tax_rate  32.765 < 2.2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Call:
lm(formula = evades_taxes_yn ~ marginal_property_tax_rate, data = county_C)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6371 -0.1878 -0.1878  0.3629  0.8122

Coefficients:
                       Estimate Std. Error
(Intercept)             -0.0047980  0.0179304
marginal_property_tax_rate  0.0128387  0.0004986
                      t value Pr(>|t|)
(Intercept)             -0.268    0.789
marginal_property_tax_rate  25.751   <2e-16 ***
---
```

```
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4355 on 2498 degrees of freedom
Multiple R-squared:  0.2098,             Adjusted R-squared:  0.2095
F-statistic: 663.1 on 1 and 2498 DF,  p-value: < 2.2e-16


t test of coefficients:

                       Estimate
(Intercept)            -0.00479799
marginal_property_tax_rate  0.01283872
                     Std. Error t value
(Intercept)            0.01645938 -0.2915
marginal_property_tax_rate  0.00050463 25.4417
                    Pr(>|t|)
(Intercept)             0.7707
marginal_property_tax_rate   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

BONUS QUESTION https://www.washingtonpost.com/wellness/2022/12/21/covid-exercise-hospitalization-mortality/

The study says that regular exercise may protect against covid, or in causal terms, regular exercise causes lower chances of fatal covid. The study says men and women who work out at least 30 minutes were 4X more likely to survive covid than inactive people. The probelem with this study is it uses the naive estimator of comparing exercises to non exercises. There is potential for selection bias here.

People who exercise regularly are also more likely to be eating healthier, be better educated, care more about their overall health, and many other factors. These are potential omitted variables that can cause OVB, meaning by not accounting for them, we overestimate the treatment effect of exercise on covid fatality.