Understanding Consequences of Higher Marginal Property Tax Rate

Misbah Arshad

SUMMARY

Problem: An NGO is concerned that wealthy households are not paying their fair share in taxes.

Method: To determine whether higher tax rates is likely to lead to higher tax evasion rates, I used Regression Discontinuity (RD).

Data: I used a fictitious data set with 3 counties: Alameda, Santa Clara, and YOLO, where a high marginal rate is 50% and a low marginal rate is 15%. Home purchased before 1978 get a low marginal rate and homes purchased after 1978 get a high marginal rate (based on California's Proposition 13).

METHOD: REGRESSION DISCONTINUITY TESTS - BUNCHING & SMOOTHNESS

We can not simply compare the difference in tax evasion rates with a high marginal versus a low marginal tax rate to determine if higher marginal tax rates cause higher tax evasion rates because there may be an omitted variable that we are not accounting for in the difference.

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 is 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) - we run a sharp RD.

The identifying assumption for a sharp RD is that 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]andE[Yi(0)|Xi=x] are continuous at x = c

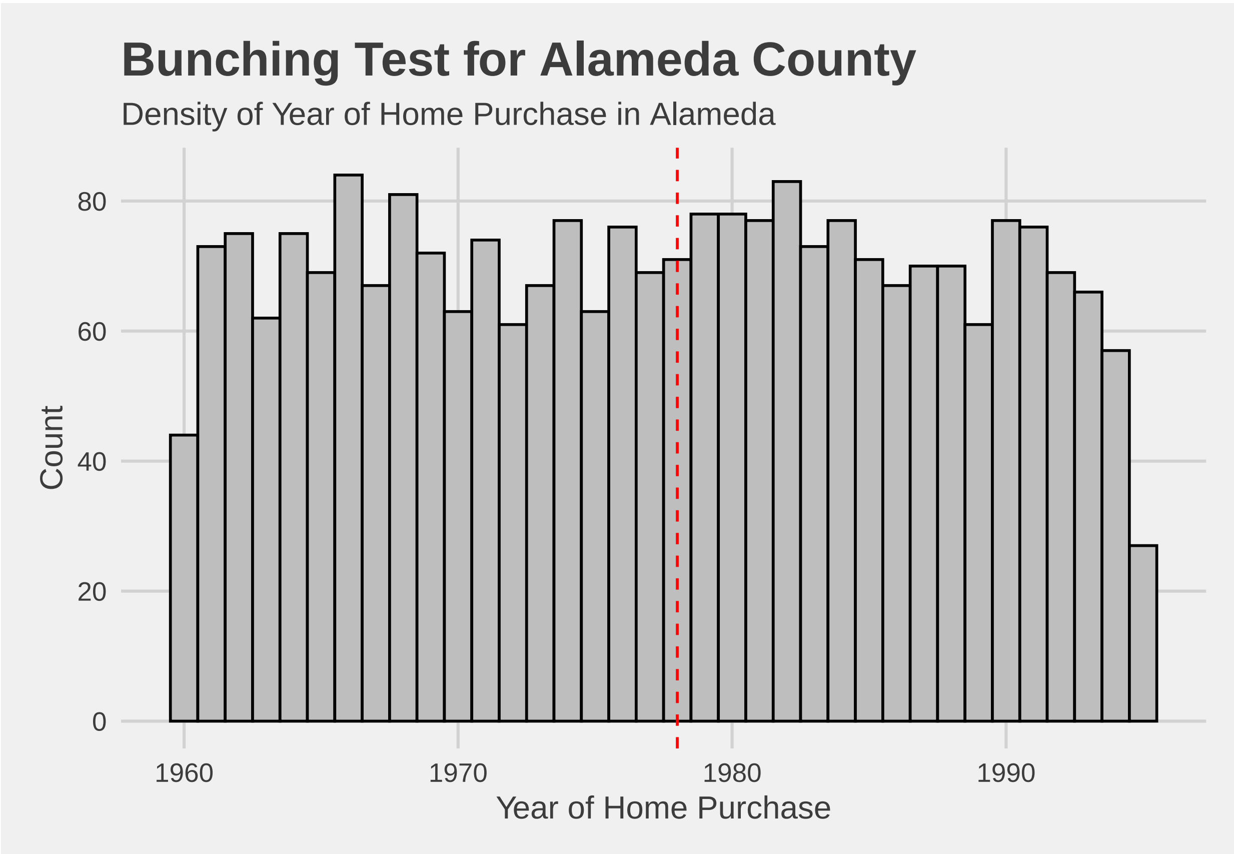
We will separate the dataset by county to evaluate each case.

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

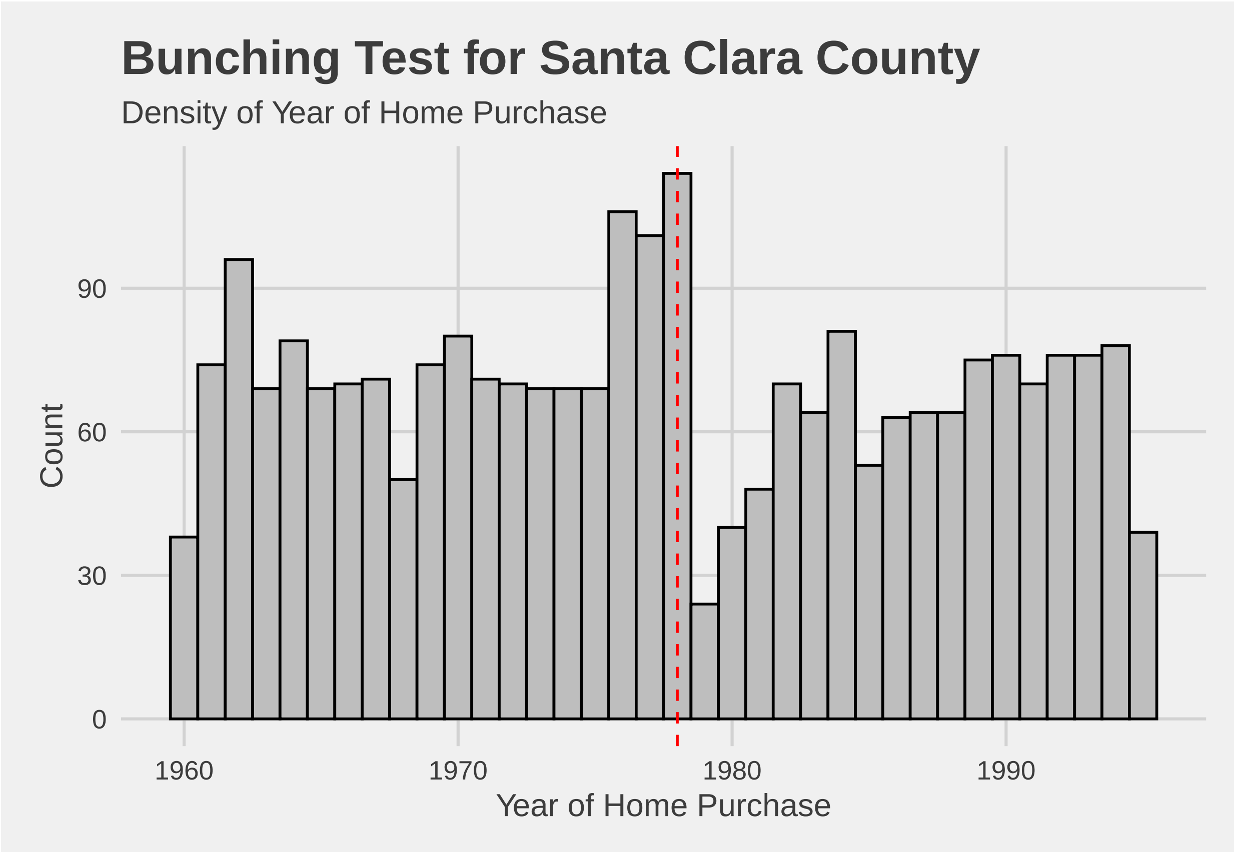
Bunching Tests: Alameda and YOLO County pass the bunching test for RD to be valid because there is no bunching around the cutoff. However, Santa Clara County does not because we can see the sharp drop off in number of home purchased after the cutoff.

This makes Santa Clara (county B dataset) invalid for a RD, however it helps us to know that in this case, people responded to the policy of changing the tax structure. The fact that people bought houses at a much lower rate when the tax rate went up is valuable information because we can dig into how Santa Clara differed from the other counties. A possible explanation is that in the other counties, residents were not told about the tax change so they could not alter there behavior around the cutoff (thus making their data sets valid) but residents in Santa Clara had advanced warning.

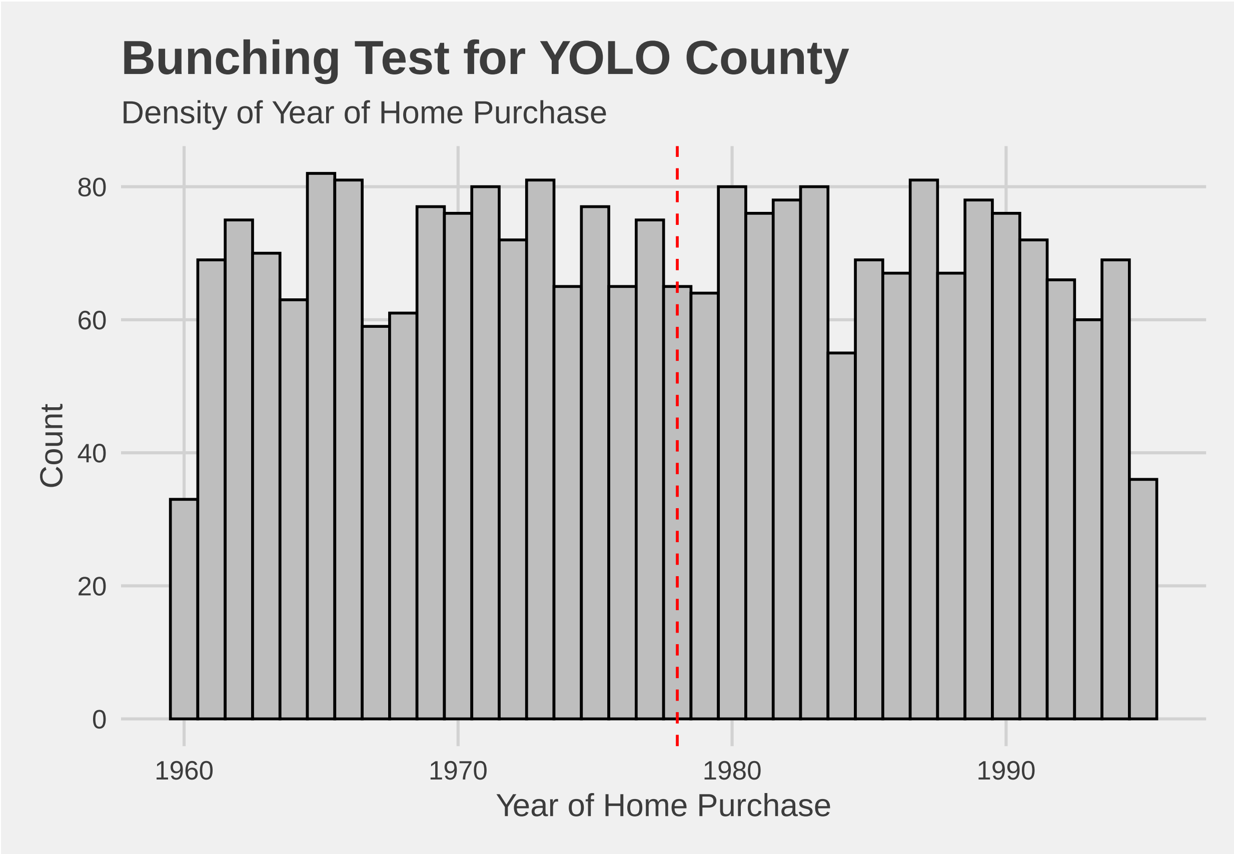
```r
cutoff_year = 1978

ggplot(county_A, aes(x = year_of_home_purchase)) +
  geom_histogram(binwidth = 1, color = "black", fill = "gray") +
  geom_vline(xintercept = cutoff_year, linetype = "dashed", color = "red") +
  labs(title = "Bunching Test for Alameda County",
      subtitle = "Density of Year of Home Purchase in Alameda",
       x = "Year of Home Purchase", y = "Count") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```




```r
ggplot(county_B, aes(x = year_of_home_purchase)) +
  geom_histogram(binwidth = 1, color = "black", fill = "gray") +
  geom_vline(xintercept = cutoff_year, linetype = "dashed", color = "red") +
  labs(title = "Bunching Test for Santa Clara County",
      subtitle = "Density of Year of Home Purchase",
       x = "Year of Home Purchase", y = "Count") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```




```
ggplot(county_C, aes(x = year_of_home_purchase)) +
  geom_histogram(binwidth = 1, color = "black", fill = "gray") +
  geom_vline(xintercept = cutoff_year, linetype = "dashed", color = "red") +
  labs(title = "Bunching Test for YOLO County",
      subtitle = "Density of Year of Home Purchase",
       x = "Year of Home Purchase", y = "Count") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```



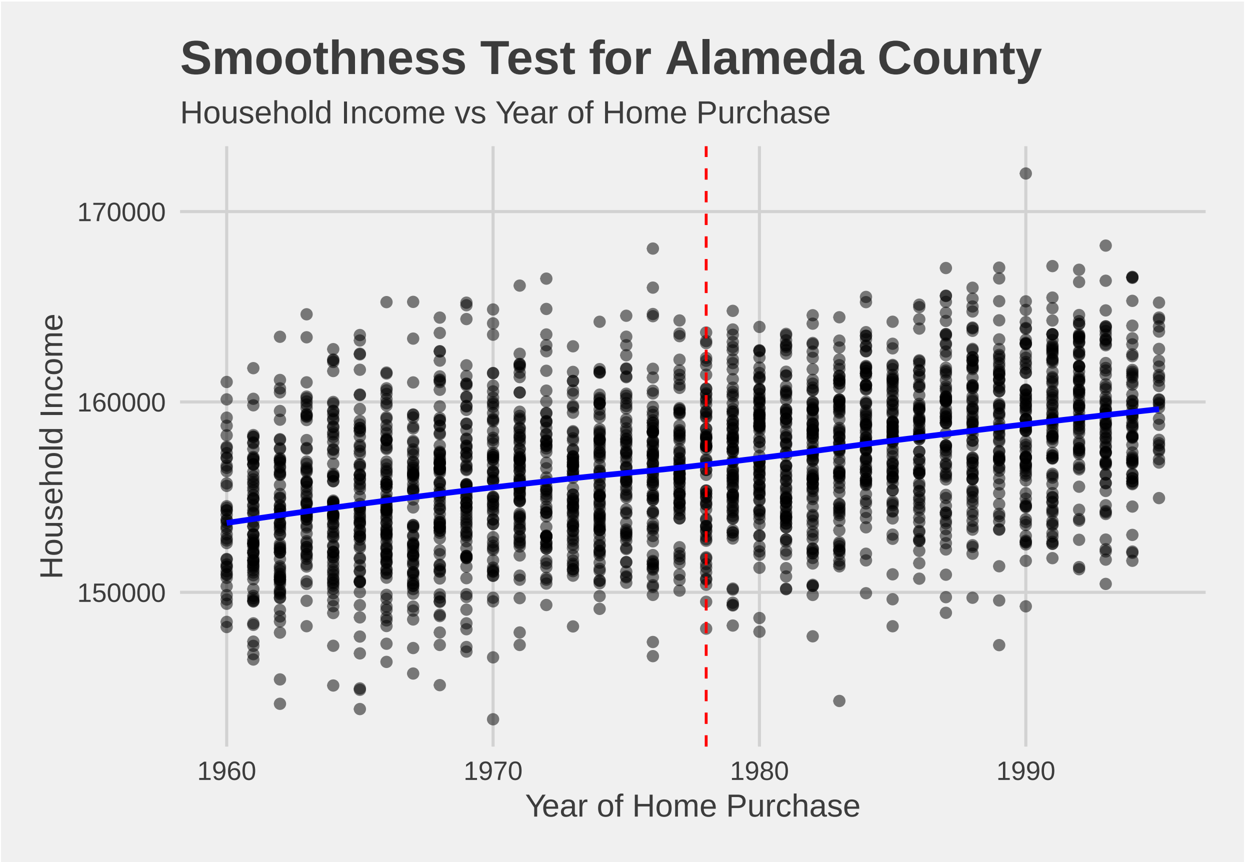
Smoothness Tests: The other variable in our dataset is household income, so we conduct a smoothness test to ensure it remains smooth around the cutoff for our RD to be valid.

Alameda and YOLO county are both smooth around the cutoff - there is no jump or drop in household income at 1978 that might interfere with our ability to infer causation from a RD. However, as with bunching, Santa Clara does not pass the smoothness test either. There is a jump in hh income after the tax change cutoff which may affect the results in a way that will not allow us to attribute the change in tax evasion solely to the marginal tax rate.

Since Santa Clara does not pass the smoothness or bunching test to be valid for an RD, we will not use it for the RD to estimate if a higher tax rate leads to higher tax evasion rates. The difference in home purchases and household income around the cutoff may lend insight to how behavior changes with the change in tax policy. Further research can be conducted.

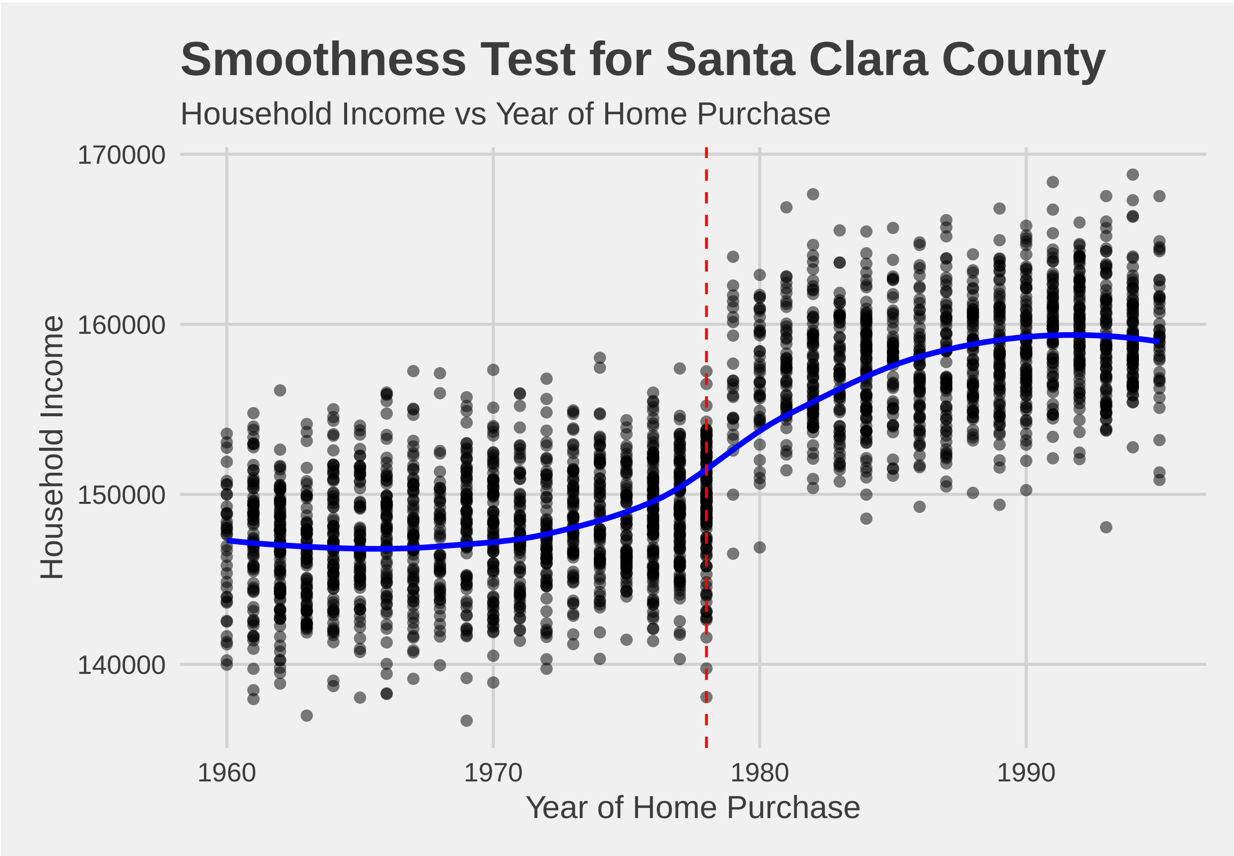
```
ggplot(county_A, aes(x = year_of_home_purchase, y = household_income)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(title = "Smoothness Test for Alameda County",
  subtitle = "Household Income vs Year of Home Purchase",
       x = "Year of Home Purchase", y = "Household Income") +
  geom_vline(xintercept = 1978, linetype = "dashed", color = "red") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



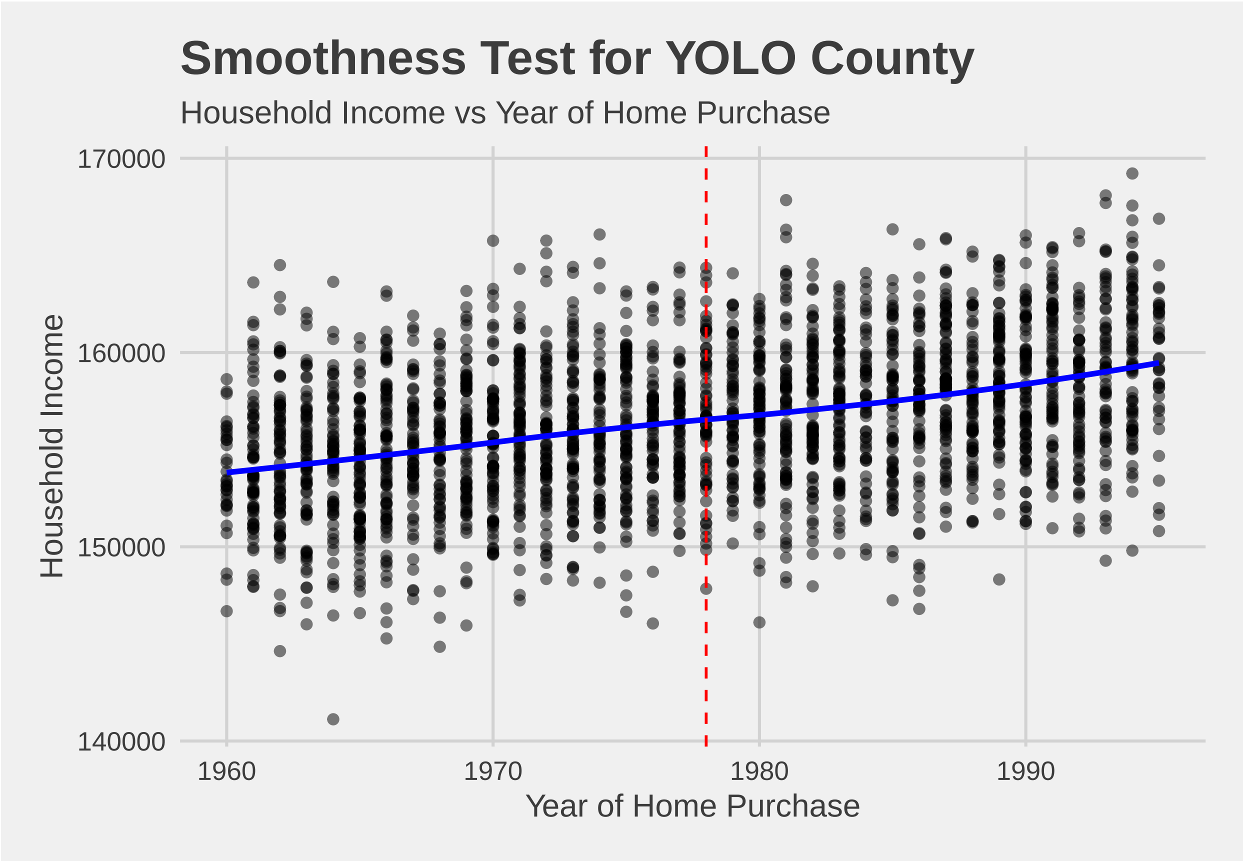

```
ggplot(county_B, aes(x = year_of_home_purchase, y = household_income)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(title = "Smoothness Test for Santa Clara County",
  subtitle = "Household Income vs Year of Home Purchase",
       x = "Year of Home Purchase", y = "Household Income") +
  geom_vline(xintercept = 1978, linetype = "dashed", color = "red") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
ggplot(county_C, aes(x = year_of_home_purchase, y = household_income)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(title = "Smoothness Test for YOLO County",
       subtitle = "Household Income vs Year of Home Purchase",
       x = "Year of Home Purchase", y = "Household Income") +
  geom_vline(xintercept = 1978, linetype = "dashed", color = "red") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



REGRESSION DISCONTINUITY: FUZZY AND SHARP

First, we create a dummy variable for the high marginal tax rate to convert all households having a 50% or higher marginal tax rate to 1 and all households below 50% to 0.

```
county_A <- county_A %>%
  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))
```

Check the data for the regression to verify that 1978 is the correct cutoff.

```
tax_rate_by_year_A <- county_A %>%
  group_by(year_of_home_purchase) %>%
  summarise(prob_high_tax = mean(high_tax_dummy, na.rm = TRUE))

tax_rate_by_year_C <- county_C %>%
  group_by(year_of_home_purchase) %>%
  summarise(prob_high_tax = mean(high_tax_dummy, na.rm = TRUE))
```

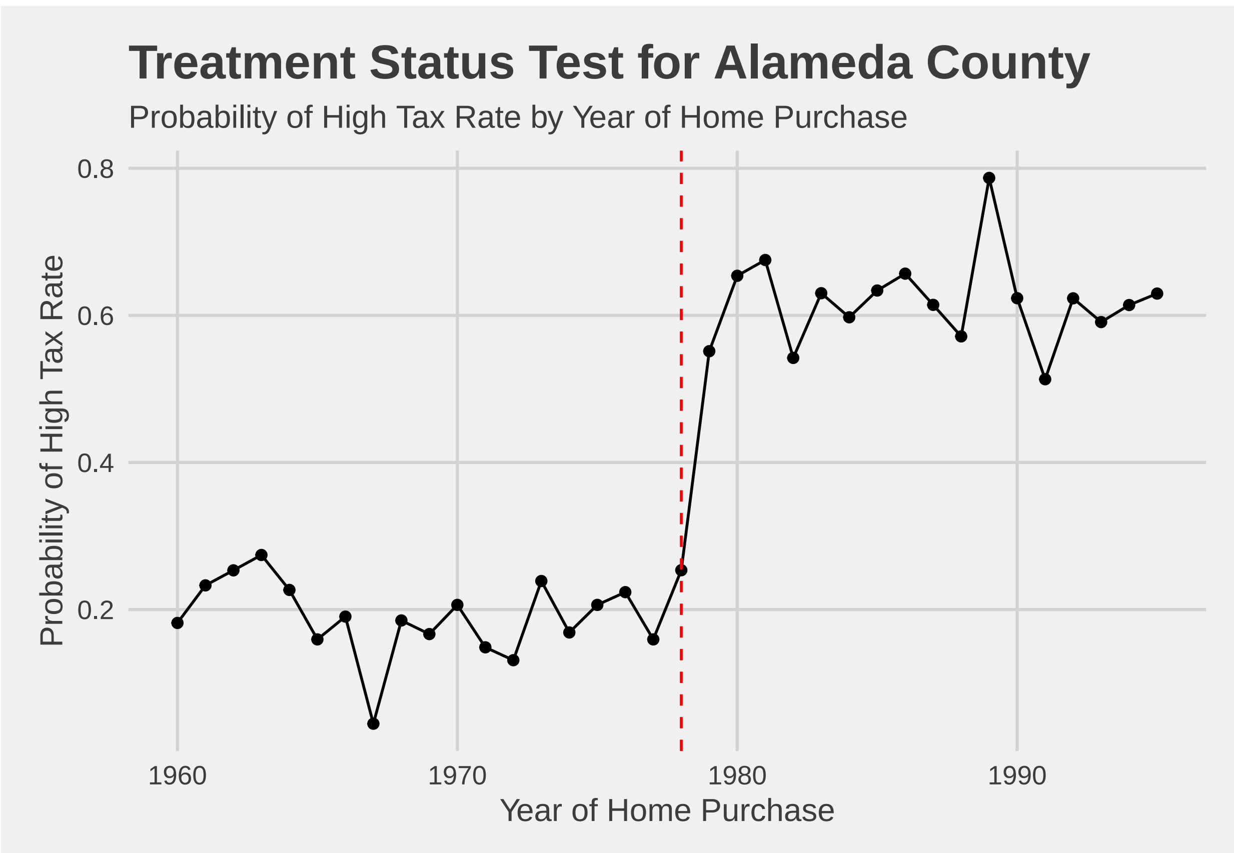
Aggregate:

Plots To conduct a sharp RD, the plot should show a distinct jump from 0% to 100% without variation. We can plot the relationship between the probability of a high tax rate and year of home purchase to test the treatment status.

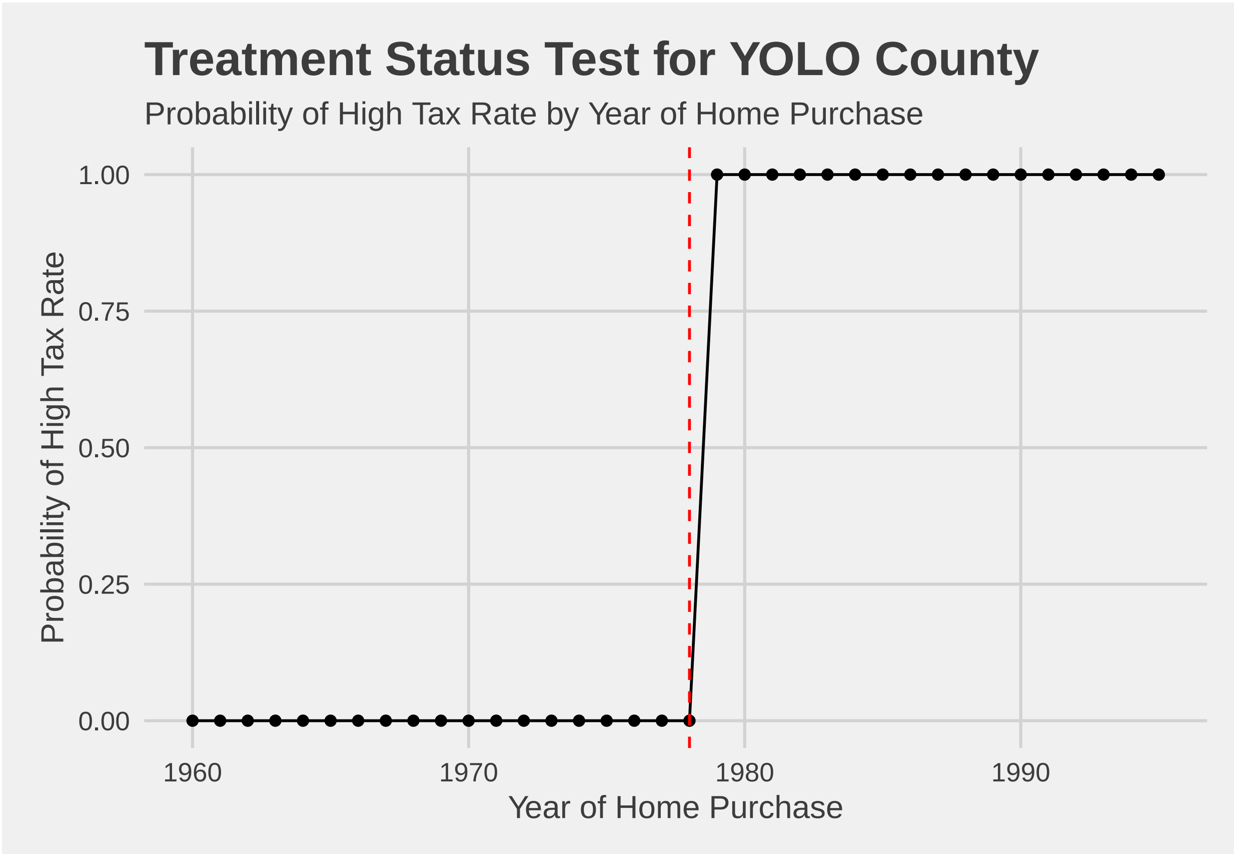
In Alameda County, the cutoff affects the probability of receiving treatment but does not guarantee it. Since there is some variation in the tax rates before and after the cutoff and therefore we need to run a fuzzy RD instead of a sharp RD.

In YOLO County, the cutoff completely determines whether treatment is received (i.e. the treatment status jumps from 0% to 100%) and therefore we can run a sharp RD on the data set.

```
ggplot(tax_rate_by_year_A, aes(x = year_of_home_purchase, y = prob_high_tax)) +
  geom_point() +
  geom_line() +
  labs(title = "Treatment Status Test for Alameda County",
      subtitle = "Probability of High Tax Rate by Year of Home Purchase",
       x = "Year of Home Purchase", y = "Probability of High Tax Rate") +
  geom_vline(xintercept = 1978, linetype = "dashed", color = "red") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```




```
ggplot(tax_rate_by_year_C, aes(x = year_of_home_purchase, y = prob_high_tax)) +
  geom_point() +
  geom_line() +
  labs(title = "Treatment Status Test for YOLO County",
      subtitle = "Probability of High Tax Rate by Year of Home Purchase",
       x = "Year of Home Purchase", y = "Probability of High Tax Rate") +
  geom_vline(xintercept = 1978, linetype = "dashed", color = "red") +
  theme_fivethirtyeight() +
  theme(axis.title = element_text())
```



Both counties have a positive coefficient that was statistically significant (p-value < 2e-16) indicating that a higher marginal tax rate increases the likelihood of tax evasion.

Alameda County (Fuzzy Regression Discontinuity):

1) First Stage: regress treatment on the instrument (year of home purchase). This first stage regression demonstrates that year of home purchase is a strong instrument for the treatment variable (high tax rate dummy), as it significantly predicts whether a household in Alameda County faces a high tax rate. This can be seen by the high f-statistic that is significant at the 1% level and the p-value that is also significant at 1%.

```
first_stage <- lm(high_tax_dummy ~ year_of_home_purchase, data = county_A)
stargazer(first_stage, type = "text", title = "First Stage Regression Results",
          dep.var.labels = "High Tax Rate (Dummy)",
          covariate.labels = c("Instrument: Year of Home Purchase"),
          out = "first_stage_output.txt")
```

```
##
## First Stage Regression Results
## ==============================================================
##                                         Dependent variable:
##                                     ---------------------------
##                                        High Tax Rate (Dummy)
## --------------------------------------------------------------
## Instrument: Year of Home Purchase             0.018***
##                                                (0.001)
##
## Constant                                     -35.191***
```

```
##                                                     (1.791)
##
## -------------------------------------------------------------
## Observations                                         2,500
## R2                                                   0.136
## Adjusted R2                                          0.136
## Residual Std. Error                            0.454 (df = 2498)
## F Statistic                               394.507*** (df = 1; 2498)
## =============================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

2) Reduced Form: regress the outcome on the instrument (year of home purchase) This reduced form regression demonstrates that the year of home purchase significantly influences the likelihood of tax evasion in Alameda County. The positive and statistically significant coefficient for year of home purchase indicates that as the year of home purchase increases, the probability of tax evasion also increases. This suggests that the timing of when a household purchased their home is a significant predictor of tax evasion.

```r
reduced_form <- lm(evades_taxes_yn ~ year_of_home_purchase, data = county_A)
stargazer(reduced_form,
          type = "text",
          title = "Reduced Form Regression Results",
          dep.var.labels = "Tax Evasion (Yes/No)",
          covariate.labels = c("Year of Home Purchase"),
          out = "reduced_form_output.txt")
```

```
##
## Reduced Form Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                         Tax Evasion (Yes/No)
## ------------------------------------------------
## Year of Home Purchase          0.017***
##                                 (0.001)
##
## Constant                      -32.797***
##                                 (1.522)
##
## ------------------------------------------------
## Observations                     2,500
## R2                               0.159
## Adjusted R2                      0.158
## Residual Std. Error        0.385 (df = 2498)
## F Statistic            471.025*** (df = 1; 2498)
## ================================================
## Note:                   *p<0.1; **p<0.05; ***p<0.01
```

3) Second Stage: Isolates the causal effect of the treatment (high marginal tax rate) on the outcome (tax evasion rate), using the instrument to correct for the fuzziness.

The coefficient for high tax rate (dummy) is 0.928 and is statistically significant at the 1% level, which suggests households that have o a high tax rate are 92% more likely to engage in tax evasion compared to households with the low tax rate.

The model explains about 30% of the variation in tax evasion (R-squared = 0.301). This suggests a moderately strong fit of the model. The results are robust with highly significant coefficients, and the standard errors are relatively small, further emphasizing the reliability of the estimates.

```r
Alameda_Fuzzy_RD <- ivreg(evades_taxes_yn ~ high_tax_dummy | year_of_home_purchase, data = county_A)
stargazer(Alameda_Fuzzy_RD, type = "text", title = "Fuzzy RD Results",
          dep.var.labels = "Tax Evasion (Yes/No)",
          covariate.labels = c("High Tax Rate (Dummy)", "Instrument: Year of Home Purchase"),
          out = "fuzzy_rd_output.txt")
```

```
##
## Fuzzy RD Results
## ===============================================================
##                                       Dependent variable:
##                                   ---------------------------
##                                      Tax Evasion (Yes/No)
## ---------------------------------------------------------------
## High Tax Rate (Dummy)                      0.928***
##                                             (0.039)
##
## Instrument: Year of Home Purchase          -0.135***
##                                             (0.017)
##
## ---------------------------------------------------------------
## Observations                                 2,500
## R2                                           0.301
## Adjusted R2                                  0.301
## Residual Std. Error                    0.351 (df = 2498)
## ===============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```

```r
summary(Alameda_Fuzzy_RD)
```

```
##
## Call:
## ivreg(formula = evades_taxes_yn ~ high_tax_dummy | year_of_home_purchase,
##     data = county_A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7935  0.1347  0.1347  0.1347  0.2065
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.13467    0.01680  -8.015 1.67e-15 ***
## high_tax_dummy  0.92816    0.03897  23.815  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3513 on 2498 degrees of freedom
## Multiple R-Squared: 0.3013,  Adjusted R-squared: 0.301
## Wald test: 567.2 on 1 and 2498 DF,  p-value: < 2.2e-16
```

YOLO County (Sharp Regression Discontinuity): Yi = a + T * Di + f(Xi) + Ei

Yi = the probability of tax evasion (outcome variable) a (alpha) = intercept T (tau) = treatment effect at the cutoff (the difference in tax evasion rates between those who are just below and just above 1978 (cutoff year) Di = treatment indicator (1 if year of home purchase is 1978 or greater and 0 otherwise) f(Xi) = function of the year of home purchase (running variable) Ei = error term

The estimated coefficient for the high tax rate is 0.0128, and it is statistically significant ($p < 2e\text{-}16$). This means that households with a high tax rate are 44.9 percentage points more likely to evade taxes than those with a low tax rate, holding other factors constant. The model explains 21% of the variance in tax evasion behavior.

```
YOLO_Sharp_RD <- lm(evades_taxes_yn ~ high_tax_dummy, data = county_C)
summary(YOLO_Sharp_RD)
```

```
##
## Call:
## lm(formula = evades_taxes_yn ~ high_tax_dummy, data = county_C)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6371 -0.1878 -0.1878  0.3629  0.8122
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.18778    0.01196   15.70   <2e-16 ***
## high_tax_dummy  0.44936    0.01745   25.75   <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
```

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

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    0.187783   0.010725  17.509 < 2.2e-16 ***
## high_tax_dummy 0.449355   0.017662  25.442 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CONCLUSION

The regression discontinuity analysis reveals that increasing the tax rate from 15% to 50% leads to a measurable increase in tax evasion behavior. This suggests that the current tax structure may incentivize non-compliance, particularly among wealthy households.

Based on the cursory analysis on the limited datasets of 3 counties, I recommend that the NGO concerned with ensuring wealthy households pay their fair share of taxes advocate for policies designed to minimize tax evasion incentives while maintaining progressive taxation goals.