Regression Discontinuity Analysis

Problem Set

Chuyuan Zhong

2023-04-14

Introduction

You have been hired as a poverty specialist for the RAND Corporation. For your first project, you have been asked to analyze the impacts of the Poverty Assistance Program in the country of Wakanda. This program was tested in 2019. In 2019, the program provided cash benefits ("Poverty Assistance Benefits") to households with incomes below the federal poverty limit, which varies based on household size according to the following schedule:

Household Size	Federal Poverty Limit
1	$12490
2	$16910
3	$21330
4+	$25750

For each household that qualifies for benefits, the cash benefit amount if 20% of the Federal Poverty Limit. For households with incomes above the Federal Poverty Limit, there are no cash benefits.

The goal of this analysis is to estimate the impacts of benefits from 2019 on employment in 2020.

1. Explain how this benefit schedule creates the opportunity to apply the Regression Discontinuity research design to study the impacts of cash benefits from 2019 on employment in 2020. What are the intuitions behind the identifying assumptions in this context?

The benefit schedule creates the discontinuity at the federal poverty limit, where households with incomes below the limit receive poverty assistance benefits, while those above do not. Therefore, we can estimate the impacts of benefits on employment by comparing the employment outcomes of households on both sides of the cutoff.

We further assume that households on both sides of the cutoff have similar and randomly distributed characteristics, such as age, gender, and education level, except for the receipt of cash benefits. This means the assignment of benefits is based on the income cutoff, and we assume that no other discontinuous changes at the cutoff could affect employment outcomes.

Data Processing

Download `rd_problem_set.csv`. This dataset has one observation per household and the following variables:
* **female**: indicator for the gender of the person interviewed in the household

* **age**: age of the person interviewed in the household
* **college**: indicator for the college attendance for the person interviewed in the household
* **nhhld**: number of people in the household
* **inc2019**: household income in 2019
* **pab2019**: amount of poverty assistance benefit
* **emp2020**: indicator for employment in 2020 for person interviewed in household

Here is the preview of the dataset:

```
df<-read.csv('rd_problem_set.csv')
head(df)
```

```
##   female age college nhhld   inc2019 pab2019 emp2020
## 1      0  49       1     1 15554.620       0       0
## 2      0  40       0     1  6747.903    2498       0
## 3      0  18       1     1  4957.522    2498       1
## 4      1  34       0     1 23381.310       0       0
## 5      1  26       0     1  8167.408    2498       0
## 6      0  47       0     1 16063.780       0       1
```

```
cat("The dataset has", nrow(df), "observations and", ncol(df), "variables.")
```

```
## The dataset has 68834 observations and 7 variables.
```

Preliminaries

A. Create a variable "fpl" that has the federal poverty limit for each household.

B. Use this variable to create the running variable "runvar" which captures income relative to the household-specific federal poverty limit.

C. Create a binned version of the running variable (runvarbin) that rounds the values of the running varible to the nearest $100.

D. Create an indicator D equal to 1 for income above the federal poverty limit (given the household's size) and 0 otherwise.

E. Create an indicator T equal to 1 if poverty assistance benefits are positive and 0 otherwise.

F. Unless otherwise stated, use a bandwidth of +/- $5000 around the fpl for the analysis.

```
# fpl variable
df <- df %>%
  mutate(fpl = case_when(nhhld == 1 ~ 12490,
                         nhhld == 2 ~ 16910,
                         nhhld == 3 ~ 21330,
                         nhhld >= 4 ~ 25750,
                         TRUE ~ NA_real_))
# runvar
df$runvar <- df$inc2019 - df$fpl

# runvarbin
df$runvarbin <- floor(df$runvar/100)*100

# D
```

```
df$D <- ifelse(df$runvar > 0, 1, 0)

# T
df$T <- ifelse(df$pab2019 > 0, 1, 0)

# define workdata
workdata <- df[abs(df$runvar) < 5000, ]
```

Data Analysis

Sharp RD First Stage regressions and plots

Estimate the following regressions

$$T_i = \alpha_0 + \alpha_1 runvar_i + \beta D_i + \alpha_2 [runvar_i * D_i] + \epsilon_i$$

$$pab2019_i = \alpha_0 + \alpha_1 runvar_i + \beta D_i + \alpha_2 [runvar_i * D_i] + \epsilon_i$$

Cluster the standard errors based on the binned running variable.

Based on the estimates, how do treatment status and average benefit amounts change for households above and below the 2019 federal poverty limit?

Calculate the means of T and pab2019 within each bin of the binned running variable, and then create 2 first stage plots by plotting each of these outcomes (y-axis) against values of the binned running variable (x-axis). Are these graphs consistent with your regression results?

```
model_T<-lm.cluster(T ~ runvar + D + I(runvar*D), data = workdata, cluster = 'runvarbin')
summary(model_T)
```

```
## R^2= 1
##
##                     Estimate   Std. Error       t value  Pr(>|t|)
## (Intercept)    1.000000e+00 8.520058e-18  1.173701e+17 0.0000000
## runvar        -1.714877e-18 6.004275e-21 -2.856093e+02 0.0000000
## D             -1.000000e+00 6.643970e-16 -1.505124e+15 0.0000000
## I(runvar * D) -7.063667e-19 7.217602e-19 -9.786723e-01 0.3277419
```

This regression model estimates the effects of income relative to the federal poverty limit on receiving cash benefits. The coefficient estimate of `runvar` is close to zero, indicating that it is not statistically significant. The coefficient estimate of `D` is -1, and its p-value of 0 shows statistical significance. This suggests that households that below the federal poverty limit receive cash benefits, but household the above the deral poverty limit do not receive cash benefits.

```
model_p<-lm.cluster(pab2019 ~ runvar + D + I(runvar*D), data = workdata, cluster = 'runvarbin')
summary(model_p)
```

```
## R^2= 0.87409
##
##                     Estimate  Std. Error    t value     Pr(>|t|)
```

```
## (Intercept)    4.080736e+03 16.46009442  247.91695 0.000000e+00
## runvar         6.116361e-02  0.00532692   11.48198 1.625077e-30
## D             -4.080736e+03 16.46009442 -247.91695 0.000000e+00
## I(runvar * D) -6.116361e-02  0.00532692  -11.48198 1.625077e-30
```

The average benefit amount for households at the federal poverty limit (`runvar` = 0) is \$4080.736. The statistically significant results suggest that treatment and income relative to fpl and their mutual effects affect the receipt of average benefit amounts. The negative coefficient estimate for `D` indicates that households below the federal poverty limit receive more benefits than households above the poverty limit since those above the cutoff don't receive benefits.

```
df_means <- workdata %>%
  group_by(runvarbin) %>%
  summarize(mean_T = mean(T), mean_pab2019 = mean(pab2019))
```

```
ggplot(df_means, aes(x=runvarbin, y=mean_T)) +
  geom_point() +
  geom_line(color ='red') +
  labs(x = 'Dist to FPL', y = 'Treatment',
       title = 'First Stage: Sharp RD (Treatment)') +
  theme(plot.title = element_text(hjust = 0.5))
```

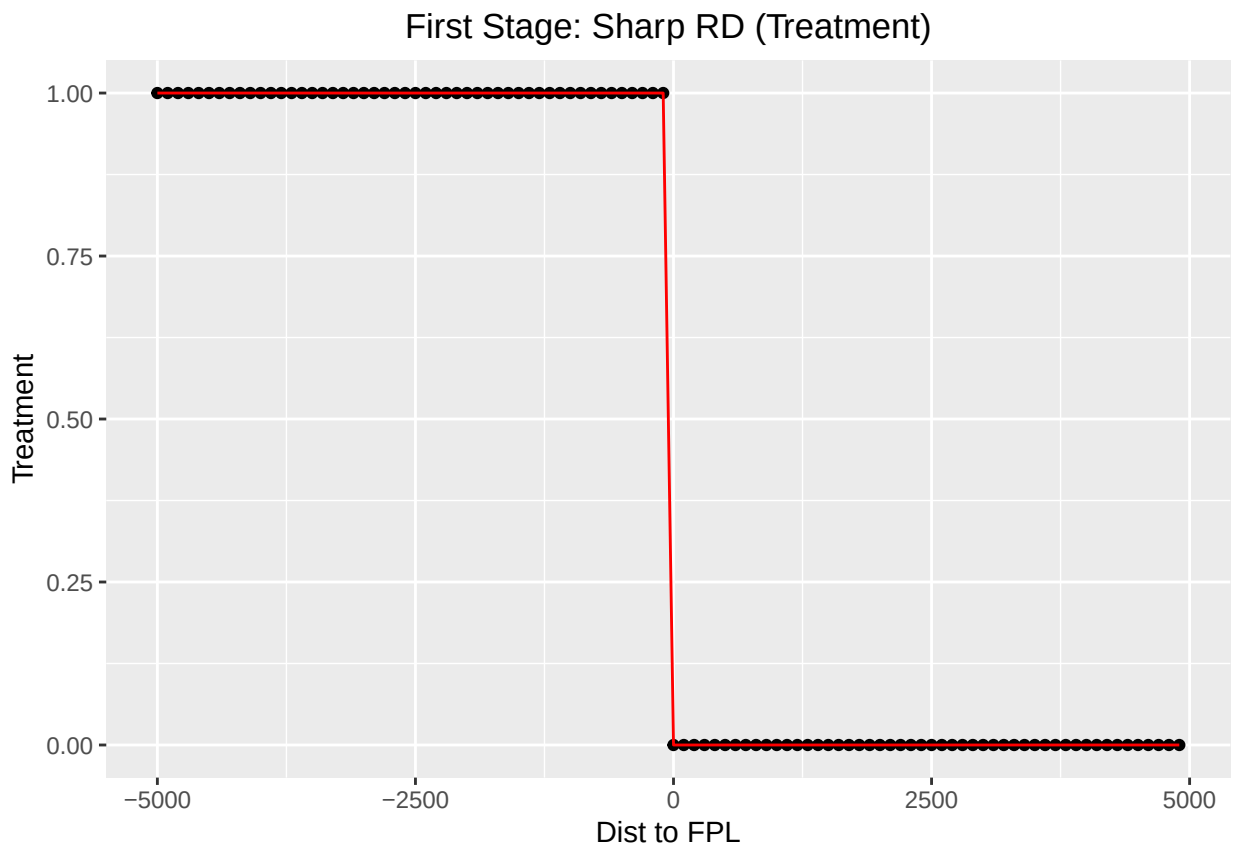


Figure 1: This figure displays the distribution of the mean of treatment outcomes across income levels relative to the federal poverty limit.

```
ggplot(df_means, aes(x=runvarbin, y=mean_pab2019)) +
  geom_point() +
  labs(x = 'Dist to FPL', y = 'amount of benefit',
       title = 'First Stage: Sharp RD (Average Benefit Amount)') +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_vline(xintercept = 0) +
  geom_smooth(formula = y~x*(x<0), color = 'red')
```

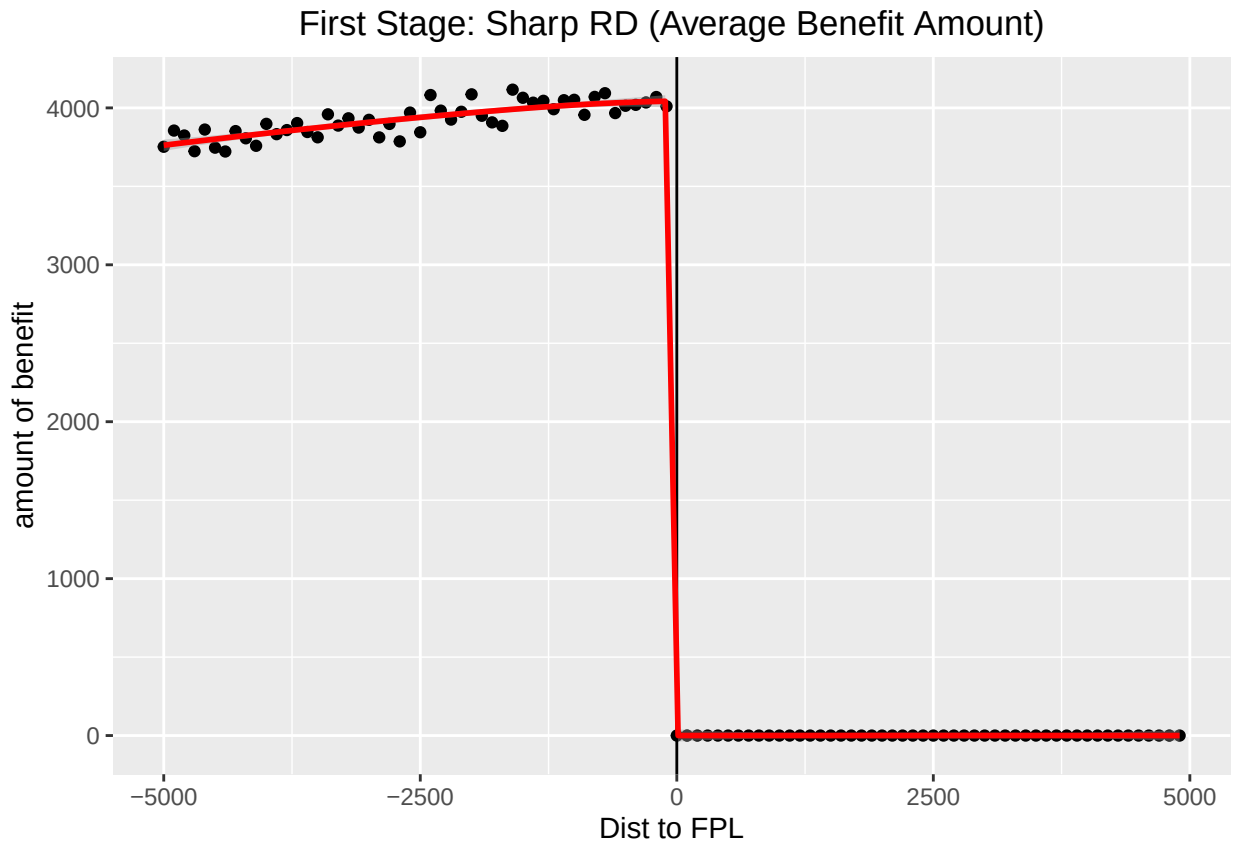


Figure 2: This figure displays the distribution of the mean of the amount of cash benefits across income levels relative to the federal poverty limit.

The two graphs are consistent with the regression results. Both graphs show that households below the federal poverty limit receive cash benefits. The difference is that the first graph examines the treatment variable `D` with income level relative to the federal poverty limit, so the distribution of both sides tends to a straight line. On the other hand, the second graph examines the exact amount of benefit with income relative to the federal poverty limit, and we can see that there is a fluctuation of the amount of benefits for households below the poverty line.

Reduced form regressions and plots:

Estimate the following regressions

$$emp2020_i = \alpha_0 + \alpha_1 runvar_i + \beta D_i + \alpha_2 [runvar_i * D_i] + \epsilon_i$$

Cluster the standard errors based on the binned running variable.

Based on the estimates, how do 2020 employment outcomes change for households above and below the 2019 federal poverty limit?

Calculate the means of emp2020 within each bin of the binned running variable, and then create a reduced form plot by plotting these means (y-axis) against values of the binned running variable (x-axis). Is this graph consistent with your regression results?

Interpret the regression and graph in terms of impacts of poverty assistance benefits on the outcomes.

Interpret the coefficient estimates. Using the ratio of the first stage and reduced form estimates, how much does an additional $1000 of benefits impact the probability of employment?

```
model_e <- lm.cluster(emp2020 ~ runvar + D + I(runvar*D), data = workdata, cluster = 'runvarbin')
summary(model_e)
```

```
## R^2= 0.00185
##
##                    Estimate   Std. Error    t value  Pr(>|t|)
## (Intercept)    4.796557e-01 9.132885e-03 52.5196274 0.0000000
## runvar        -1.785748e-06 3.198637e-06 -0.5582839 0.5766505
## D             -3.114844e-02 1.294466e-02 -2.4062779 0.0161160
## I(runvar * D) -1.112958e-06 4.590990e-06 -0.2424222 0.8084531
```

The estimated coefficient of `D` is -0.03, which indicates that employment outcomes are 3.11% lower for households above the federal poverty limit compared to those above. Households who below the 2019 federal poverty limit have a better employment outcome than households who above the poverty limit.

```
plotdata=aggregate(workdata$emp2020, list(workdata$runvarbin), FUN=mean)
ggplot(plotdata, aes(x=Group.1, y=x))+
  geom_point()+
  labs(x='Dist to FPL', y='Employment Outcome', title = 'Reduced Form (Employment)')+
  geom_line()+
  theme(plot.title = element_text(hjust= 0.5))+
  geom_vline(xintercept = 0)
```

The reduced form graph indicates a lack of a clear discontinuity, which aligns with the regression estimates that suggest there were no significant differences in employment outcomes for households above and below the 2019 federal poverty limit in 2020.

```
library(scales)
ratio <- coef(model_e)[3] / coef(model_p)[3]
percent(ratio*1000, accuracy = 0.01)
```

```
##       D
## "0.76%"
```

To test the effect of an additional $1000 in benefits on employment outcomes, we need to estimate the treatment effect on both employment outcomes and the average benefit amount.

The estimate shows that an increase of $1000 in benefits is associated with a 0.76% increase in the probability of employment.

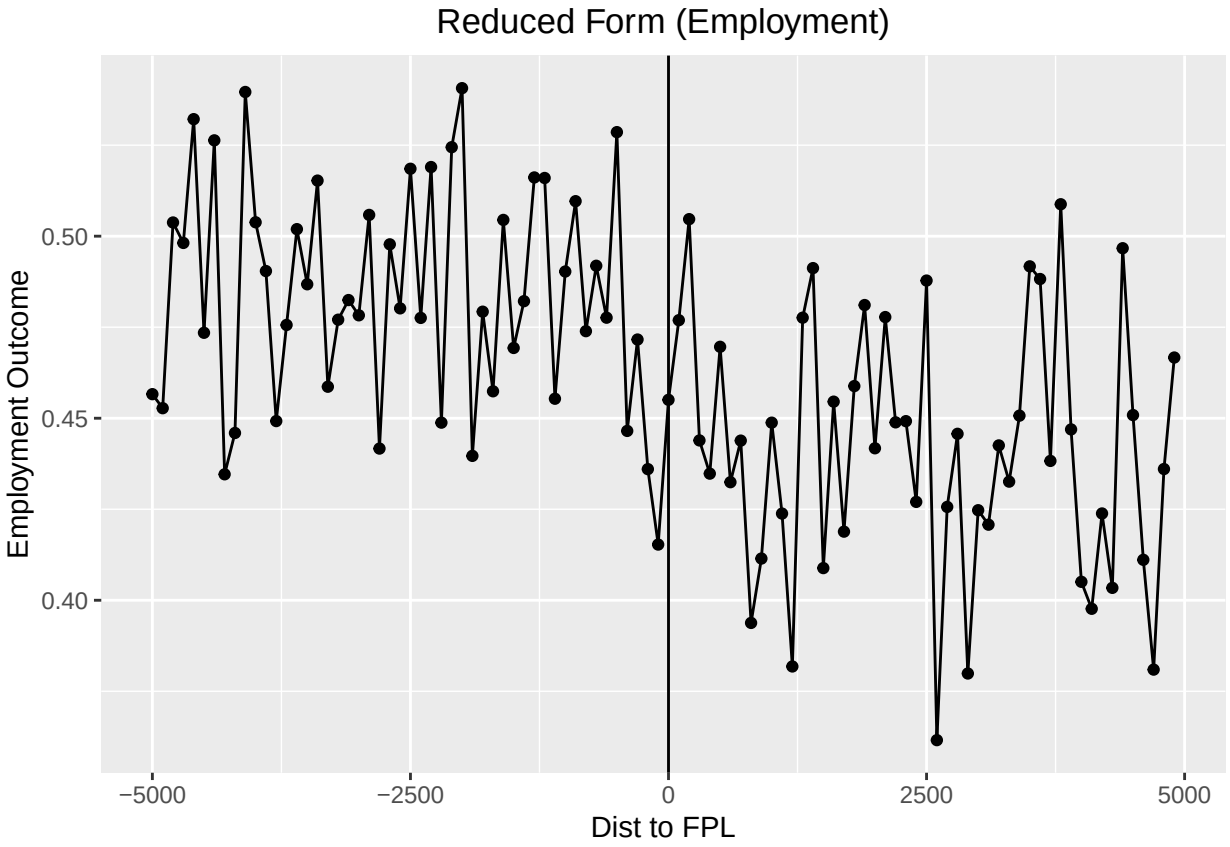


Figure 3: This figure displays the distribution of mean employment outcomes relative to the federal poverty limit across different income levels.

Frequencies plot:

Calculate the counts of the number of observations within each bin of the running variable (Nobs). Using one observation per bin value, estimate the following regression

$$Nobs_b == \alpha_0 + \alpha_1 g(runvar_b) + \beta D_i + \alpha_2 [g(runvar_b) * D_i] + \epsilon_i$$

where b indexes each bin and g(.) is a cubic polynomial of the binned running variable value. Is the coefficient on the indicator variable D significant?

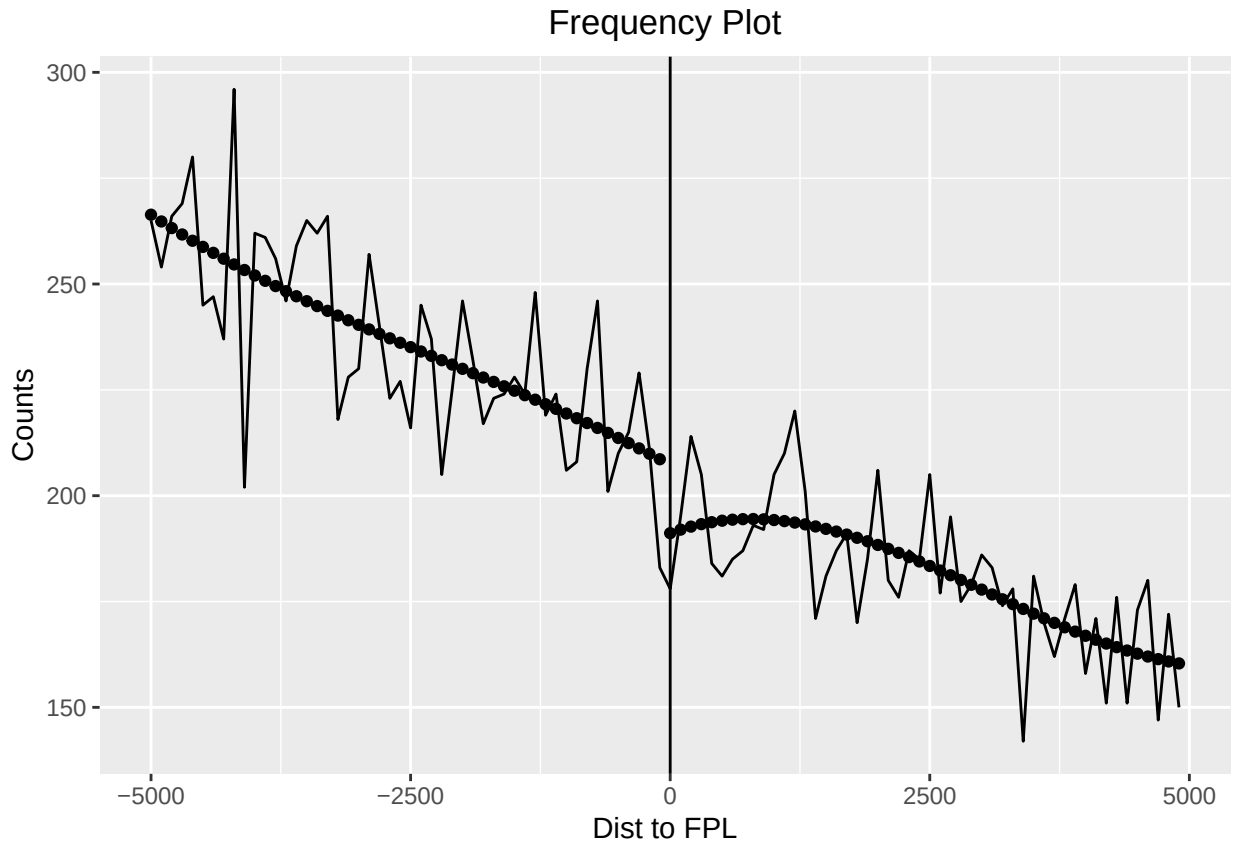
Plot Nobs and the fitted values from this regression. If households could manipulate the running variable to qualify for treatment, what would you expect to see? Is there any evidence that households are able to manipulate the running variable to qualify for treatment?

```
workdata$runvarbin2 = workdata$runvarbin**2
workdata$runvarbin3 = workdata$runvarbin**3

Nobs<-workdata %>%
      group_by(runvarbin,runvarbin2,runvarbin3,D)%>%
      summarise(nobs = length(runvar),.groups = 'drop')

freq <- lm(nobs ~ runvarbin + runvarbin2 + runvarbin3+D +
        I(runvarbin*D)+ I(runvarbin2*D)+I(runvarbin3*D), data = Nobs)

summary(freq)
```

```
##
## Call:
## lm(formula = nobs ~ runvarbin + runvarbin2 + runvarbin3 + D +
##     I(runvarbin * D) + I(runvarbin2 * D) + I(runvarbin3 * D),
##     data = Nobs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.308 -10.355   0.114  10.425  41.373
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.073e+02  9.201e+00  22.528   <2e-16 ***
## runvarbin          -1.339e-02  1.547e-02  -0.865    0.389
## runvarbin2         -1.500e-06  7.011e-06  -0.214    0.831
## runvarbin3         -2.372e-10  9.042e-10  -0.262    0.794
## D                  -1.614e+01  1.214e+01  -1.330    0.187
## I(runvarbin * D)    2.243e-02  2.096e-02   1.070    0.287
## I(runvarbin2 * D) -5.155e-06  9.728e-06  -0.530    0.597
## I(runvarbin3 * D)  9.574e-10  1.279e-09   0.749    0.456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.06 on 92 degrees of freedom
## Multiple R-squared:  0.8218, Adjusted R-squared:  0.8082
## F-statistic: 60.61 on 7 and 92 DF,  p-value: < 2.2e-16
```

The coefficient on the indicator variable D is not significant.

```
library(ggplot2)
Nobs$predicted <- predict(freq, newdata=Nobs)
ggplot(Nobs, aes(x=runvarbin, y=predicted)) +
  geom_point() +
  geom_line(aes(x=runvarbin,y=nobs)) +
  labs(x='Dist to FPL',y='Counts',title='Frequency Plot') +
  theme(plot.title = element_text(hjust=0.5))+
  geom_vline(xintercept = 0)
```



If households could manipulate the running variable to qualify for treatment, we would expect to see that there is a sharp discontinuity at the cutoff, and the number of counts of household above the fpl would be generally larger than the one of household below the fpl around the cutoff. Therefore, there is no strong evidence that households are able to manipulate the running variable to qualify for treatment.

Covariate predicted employment

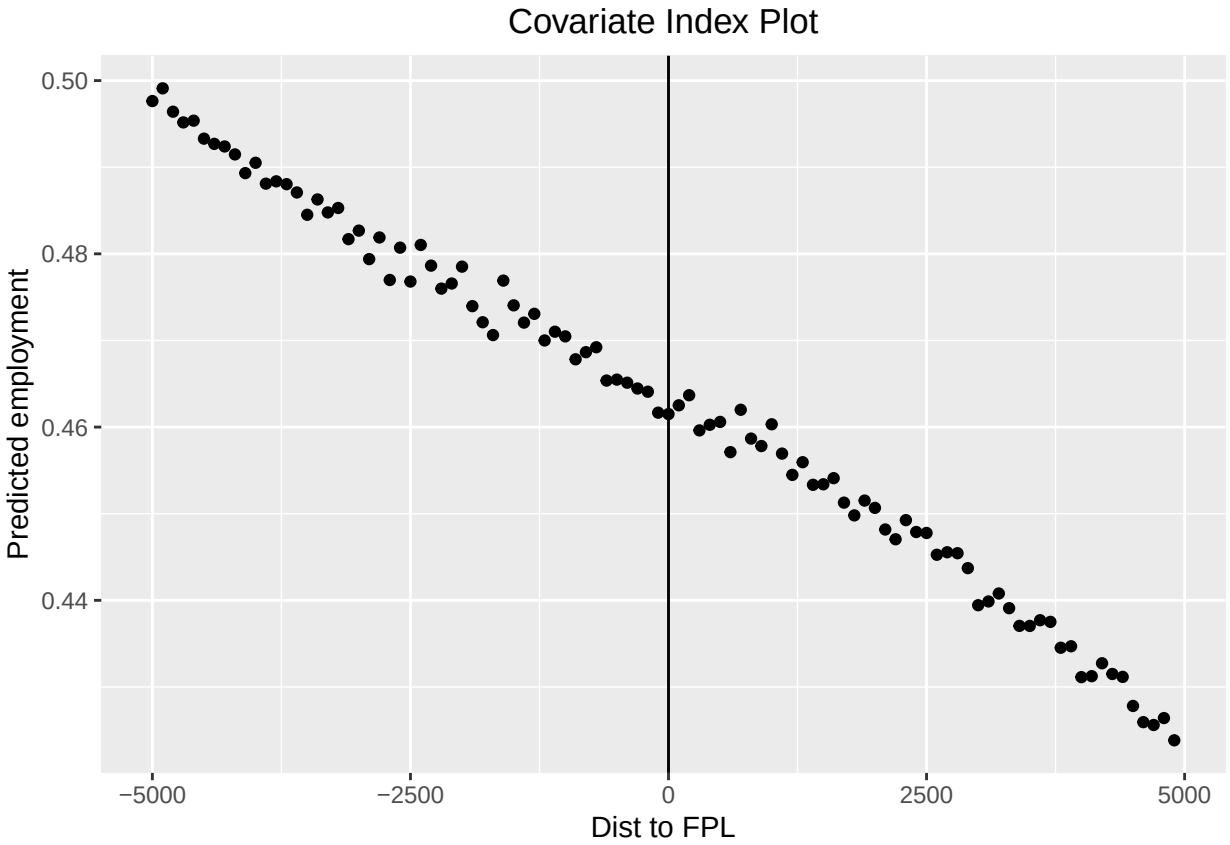
Regress employment in 2020 on a cubic polynomial in age, female, college, dummies for household size, and a cubic polynomial in 2019 household income. Obtain the predicted values and use these predicted values to estimate the same regression as in (4). How do these results compare to the result in (4)? How do these results relate to the RD identifying assumptions and the interpretation of your results from (4)?

```
# create cubic polynomial variables
workdata$age2 = workdata$age**2
workdata$income2 = workdata$inc2019**2
workdata$age3 = workdata$age**3
workdata$income3 = workdata$inc2019**3

# create dummies
workdata <- dummy_cols(workdata, select_columns = "nhhld",
                       remove_first_dummy = FALSE)

# fit the regression
model_cp<-lm(emp2020 ~ female +age+age3+age2+income2+income3+college +
            nhhld + inc2019 +nhhld_1+nhhld_2+nhhld_3+nhhld_4+nhhld_5+
            nhhld_6, data = workdata)
```

```
# predict values
workdata$predict_e <-predict(model_cp, newdata = workdata)

# run regression on predicted values
model_pred<-lm.cluster(predict_e ~ runvar+D+I(D*runvar), data=workdata,
                       cluster='runvarbin')
summary(model_pred)
```

```
## R^2= 0.49698
##
##                     Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)    4.625172e-01 4.104682e-04 1126.804000 0.000000e+00
## runvar        -6.950635e-06 1.220354e-07  -56.955903 0.000000e+00
## D              2.901285e-03 6.921506e-04    4.191696 2.768763e-05
## I(D * runvar) -9.562067e-07 2.227397e-07   -4.292933 1.763281e-05
```

```
plotdata=aggregate(workdata$predict_e, list(workdata$runvarbin), FUN=mean)

ggplot(plotdata, aes(x=Group.1, y=x))+
  geom_point()+
  labs(x='Dist to FPL', y='Predicted employment', title = 'Covariate Index Plot')+
  theme(plot.title = element_text(hjust= 0.5))+
  geom_vline(xintercept = 0)
```



By including a cubic polynomial in age, female, college, income, and dummies for household size, the R-squared value significantly increased from 0.00185 in (4) to 0.49698 in this question. This suggests that the new model can explain much more of the variation in the data. The estimate for the coefficient of `D` is now positive, suggesting that the treatment effect increases employment outcomes for households above the federal poverty limit. We can see that there is no discontinuity now, so comparing to result in (4), there is a stronger evidence to suggest that household who receive cash benefits have higher employment outcomes than those who do not. Regarding to the RD assumption, since there is no discontinuous change observed, we can see that the employment outcomes are only affected by the cutoff.

Sensitivity Analysis

Polynomial specification

So far, we have assumed a linear polynomial specification of the running variable. How do the results change if you use quadratic or cubic polynomial specifications for the running variables?

```
workdata$runvar2 = workdata$runvar**2
workdata$runvar3 = workdata$runvar**3
poly1_reg = lm(emp2020 ~ D + runvar + I(D * runvar), data = workdata)
workdata$pred_poly1 = predict(poly1_reg, newdata = workdata)
poly2_reg = lm(emp2020 ~ D + runvar + runvar2 + I(D * runvar) +
                 I(D * runvar2), data = workdata)
workdata$pred_poly2 = predict(poly2_reg, newdata = workdata)
poly3_reg = lm(emp2020 ~ D + runvar + runvar2 + runvar3 + I(D * runvar) +
                 I(D * runvar2) + I(D * runvar3), data = workdata)
workdata$pred_poly3 = predict(poly3_reg, newdata = workdata)
summary(poly1_reg)
```

```
##
## Call:
## lm(formula = emp2020 ~ D + runvar + I(D * runvar), data = workdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4886 -0.4810 -0.4363  0.5182  0.5660
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.797e-01  9.454e-03  50.737   <2e-16 ***
## D             -3.115e-02  1.395e-02  -2.234   0.0255 *
## runvar        -1.786e-06  3.181e-06  -0.561   0.5745
## I(D * runvar) -1.113e-06  4.848e-06  -0.230   0.8184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4984 on 20864 degrees of freedom
## Multiple R-squared:  0.001851,   Adjusted R-squared:  0.001708
## F-statistic:  12.9 on 3 and 20864 DF,  p-value: 2.049e-08
```

```
summary(poly2_reg)
```

```
##
## Call:
## lm(formula = emp2020 ~ D + runvar + runvar2 + I(D * runvar) +
##     I(D * runvar2), data = workdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4911 -0.4766 -0.4372  0.5208  0.5630
##
## Coefficients:
```

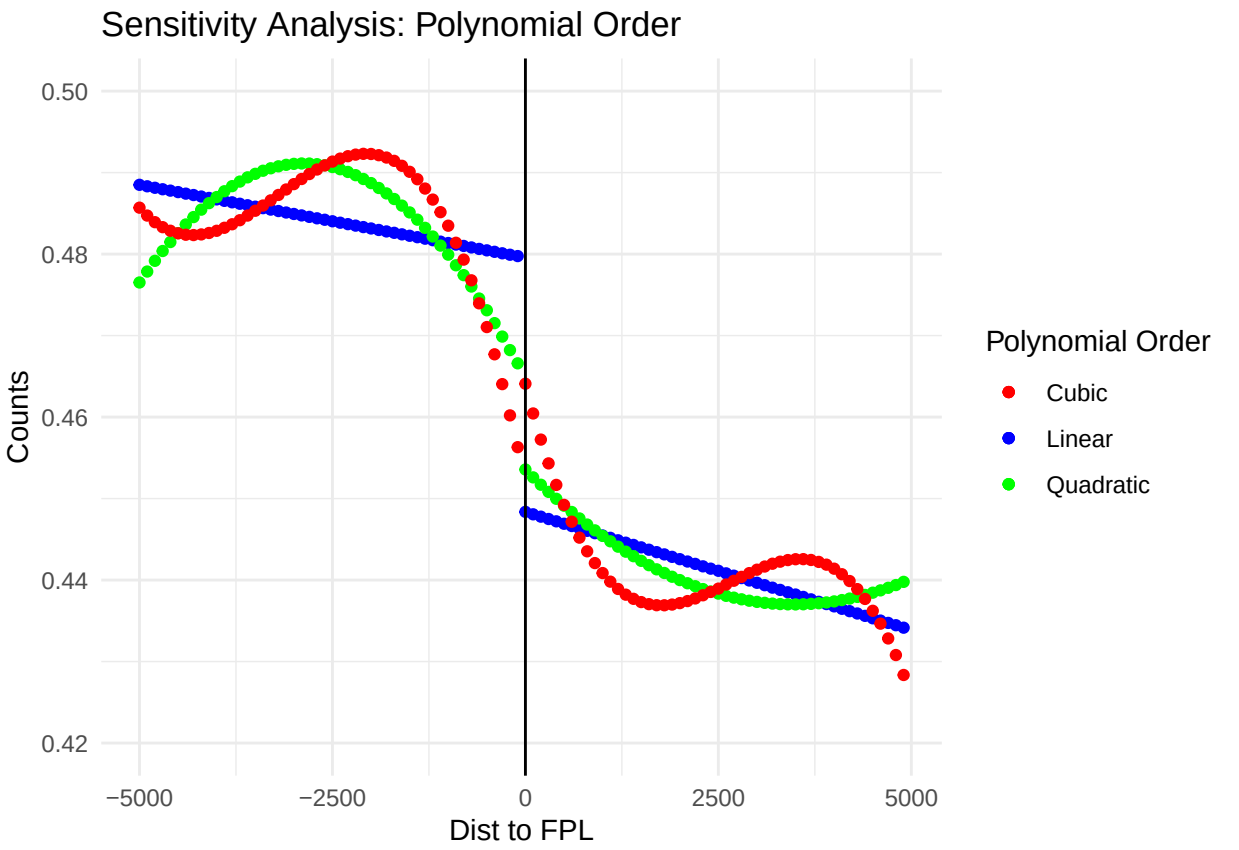
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.656e-01  1.438e-02  32.387   <2e-16 ***
## D              -1.158e-02  2.102e-02  -0.551    0.582
## runvar         -1.811e-05  1.298e-05  -1.395    0.163
## runvar2        -3.211e-09  2.476e-09  -1.297    0.195
## I(D * runvar)   8.461e-06  1.941e-05   0.436    0.663
## I(D * runvar2)  4.578e-09  3.759e-09   1.218    0.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4984 on 20862 degrees of freedom
## Multiple R-squared:  0.001943,   Adjusted R-squared:  0.001704
## F-statistic: 8.122 on 5 and 20862 DF,  p-value: 1.144e-07
```

```
summary(poly3_reg)
```

```
##
## Call:
## lm(formula = emp2020 ~ D + runvar + runvar2 + runvar3 + I(D *
##     runvar) + I(D * runvar2) + I(D * runvar3), data = workdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4923 -0.4797 -0.4370  0.5176  0.5729
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.539e-01  1.942e-02  23.377   <2e-16 ***
## D               1.201e-02  2.821e-02   0.426    0.670
## runvar         -4.527e-05  3.286e-05  -1.378    0.168
## runvar2        -1.653e-08  1.501e-08  -1.101    0.271
## runvar3        -1.755e-12  1.950e-12  -0.900    0.368
## I(D * runvar)   6.855e-06  4.869e-05   0.141    0.888
## I(D * runvar2)  3.245e-08  2.259e-08   1.437    0.151
## I(D * runvar3) -2.031e-13  2.969e-12  -0.068    0.945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4984 on 20860 degrees of freedom
## Multiple R-squared:  0.002018,   Adjusted R-squared:  0.001683
## F-statistic: 6.026 on 7 and 20860 DF,  p-value: 4.874e-07
```

```
plotdata=aggregate(cbind(workdata$pred_poly1,
                         workdata$pred_poly2,
                         workdata$pred_poly3),
                   list(workdata$runvarbin), FUN=mean)

ggplot(plotdata, aes(x = Group.1)) +
  geom_point(aes(y = V1, col = "Linear")) +
  geom_point(aes(y = V2, col = "Quadratic")) +
  geom_point(aes(y = V3, col = "Cubic")) +
  scale_color_manual(values = c("red", "blue", "green")) +
  labs(x = "Dist to FPL", y = "Counts",
```

```
      title = "Sensitivity Analysis: Polynomial Order",
      col = "Polynomial Order") +
ylim(0.42, 0.5) +
theme_minimal() +
geom_vline(xintercept=0)
```

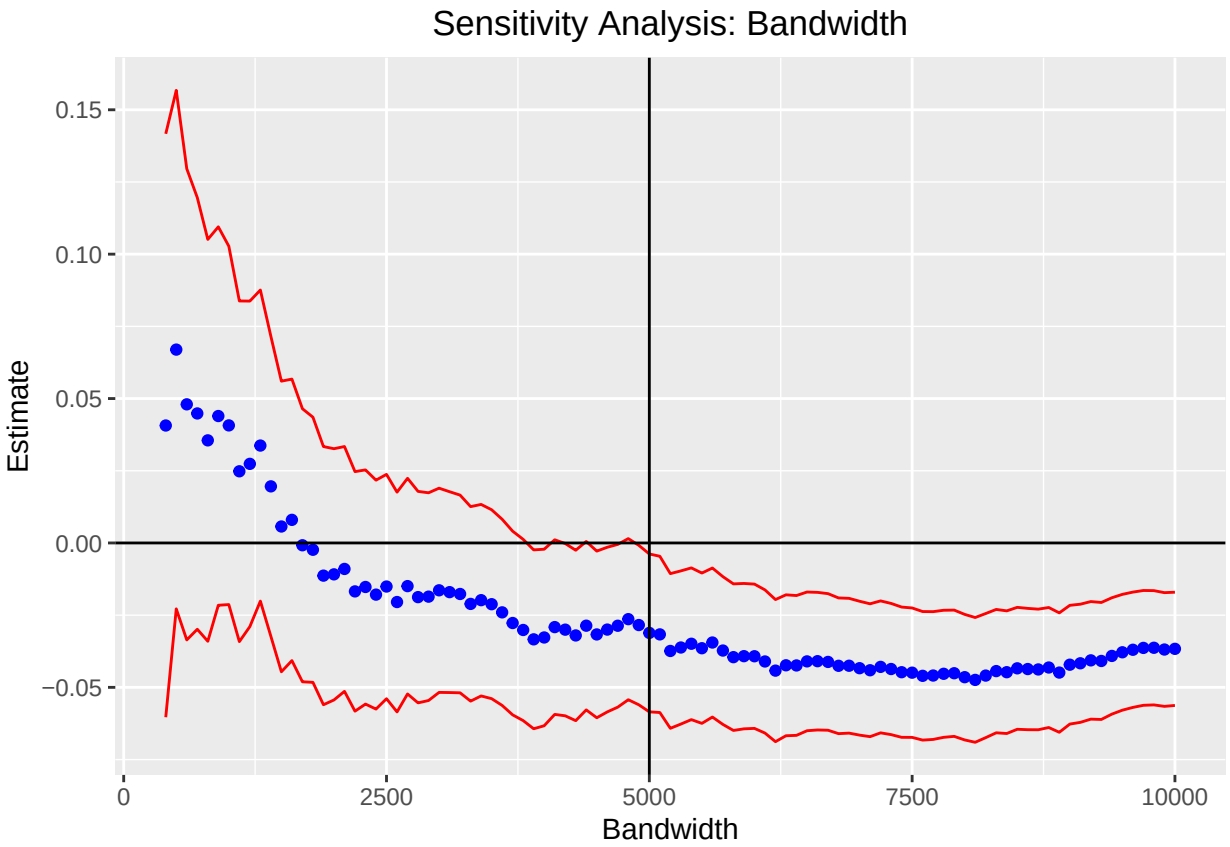


As we move from the linear to the cubic model, the R-squared values are still similar and small. This suggests that there is no significant relationship between these coefficients and employment outcomes, and the performance didn't improved.

Bandwidth

So far, we have used a bandwidth of +/- $5000 around the household-specific federal poverty limit. Vary the bandwidth from +/- $400 to +/- $10000 and plot the estimates and standard errors. How do the estimates vary as the bandwidth increases? What is the minimum bandwidth for which the estimates look stable? What is the minimum bandwidth for which the estimate is statistically significant (different from 0)?

```
CoefMatrix  = matrix(NA, 100, 5)    # Matrix to store our results.
bwidths = seq(from=400, to=10000, by=100)
for(ii in 1:length(bwidths)) {
  bw_reg = lm(emp2020 ~ runvar + D + I(D * runvar),
              data = df[abs(df$runvar) < bwidths[ii],])
  CoefMatrix[ii,1]=bwidths[ii]
  CoefMatrix[ii,2]=coefficients(bw_reg)[3]
  CoefMatrix[ii,3]=coef(summary(bw_reg))[, "Std. Error"][3]
  CoefMatrix[ii,4]= coefficients(bw_reg)[3] - 1.96*CoefMatrix[ii,3]
  CoefMatrix[ii,5]= coefficients(bw_reg)[3] + 1.96*CoefMatrix[ii,3]
}

ggplot(data = data.frame(CoefMatrix),
       aes(x = CoefMatrix[,1])) +
  geom_point(aes(y = CoefMatrix[,2]), color = "blue") +
  geom_line(aes(y = CoefMatrix[,4]), color = "red") +
  geom_line(aes(y = CoefMatrix[,5]), color = "red") +
  geom_vline(xintercept = 5000, color = "black") +
  geom_hline(yintercept = 0, color = "black") +
  labs(x = "Bandwidth", y = "Estimate", title = "Sensitivity Analysis: Bandwidth") + theme(plot.title=el
```




```
sig <- min(CoefMatrix[which(CoefMatrix[,5] < 0), 1])

cat("Minimum bandwidth for statistically significant estimate:", sig, "\n")
```

```
## Minimum bandwidth for statistically significant estimate: 3900
```

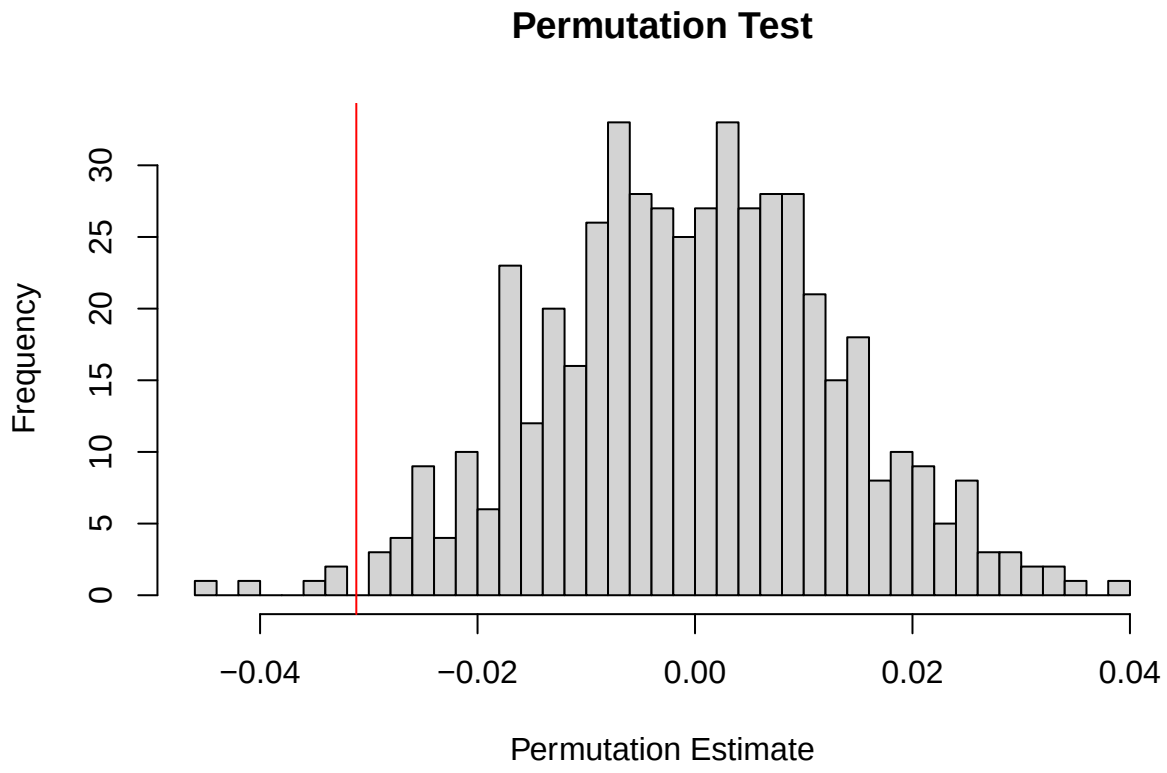
As the bandwidth increases, the estimates gradually become consistent and the confidence intervals become narrow. The minimum bandwidth for statistically significant estimate is at 3900.The minimum bandwidth for stable estimate is at 5000.

Permutation test

One of your colleagues at RAND points out that there may be some special features about the income values that are highlighted by the federal poverty limits. To address this, you implement the following permutation test. You randomly draw household size (1 through 4), assign the federal poverty limit given the above schedule, and then re-run your analysis based on household income relative to the randomly assigned federal poverty limit. You run 500 iterations and compare your estimate based on the actual data to the permutation estimates. Show these results. How do these results address your colleague's concerns?

```
set.seed(911)
reps = 500
CoefMatrix  = matrix(NA, reps, 1)   # Matrix to store our results.
for(ii in 1:reps) {
  df$nhhld = sample(1:4, dim(df)[1], replace = TRUE)
```

```
  df$pfpl = case_when(df$nhhld == 1 ~ 12490,
                      df$nhhld == 2 ~ 16910,
                      df$nhhld == 3 ~ 21330,
                      df$nhhld == 4 ~ 25750)
  df$prunvar = df$inc2019 - df$pfpl
  df$pD = ifelse(df$prunvar > 0, 1, 0)
  ptest_reg = lm(emp2020 ~ prunvar + pD + I(pD * prunvar) , data = df[abs(df$prunvar)<5000, ])
  CoefMatrix[ii,1]=coefficients(ptest_reg)[3]
}
hist(CoefMatrix[,1], breaks = 50, main="Permutation Test",
     xlab="Permutation Estimate")
abline(v = coef(model_e)[3], col="red")
```




```
extreme_estimates<-mean(CoefMatrix[,1] < coef(model_e)[3])
extreme_estimates
```

```
## [1] 0.01
```

Based on the result, 1% estimates are as extreme as the actual data estimates, this provides evidence to against the colleagues' concerns and there are no special features about the income values that are highlighted by the federal poverty limits.