Empirical Project 2: Huai River Experiment

Isaac Wilfong

2025-05-02

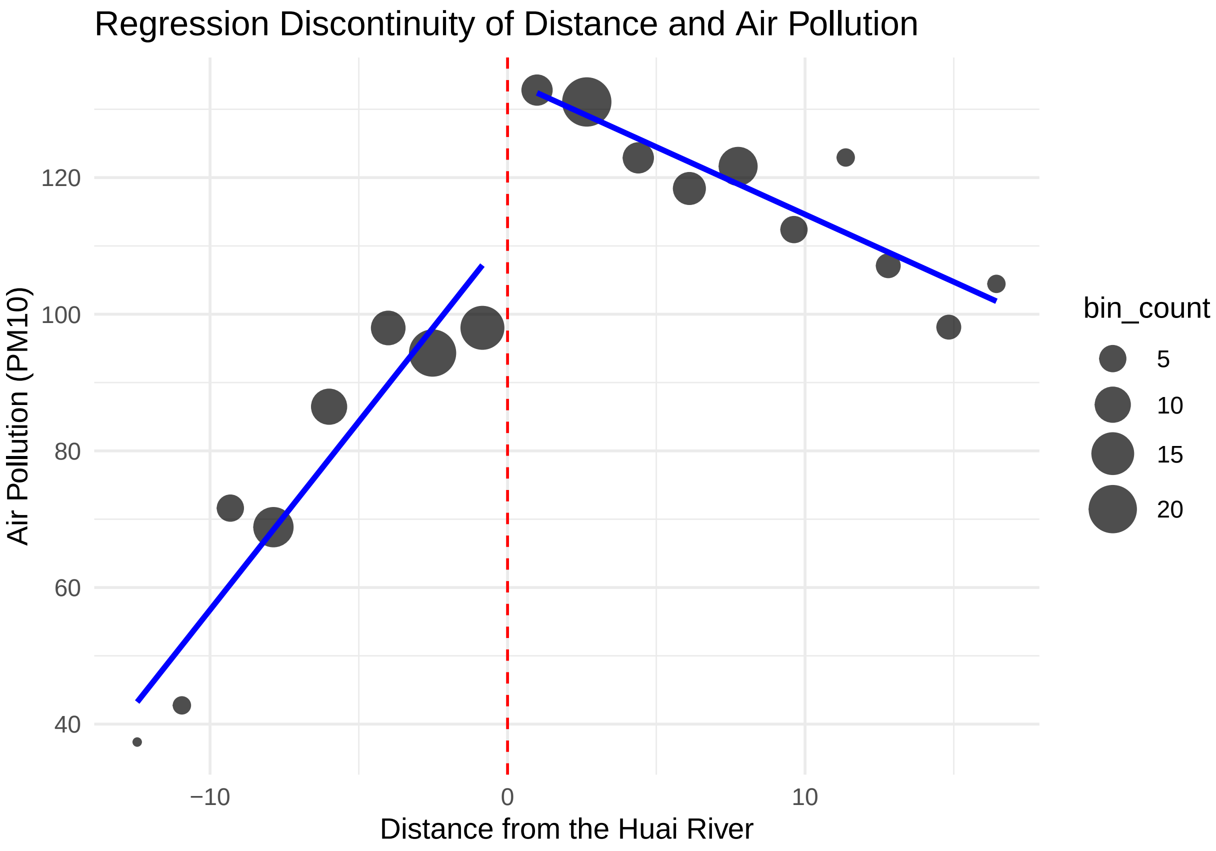
Question 1

The Huai river experiment came from a arbitrary policy divide. The policy was that households below the Huai river were not allowed to burn coal in to house for heating while the north was considered too cold to restrict people from doing this. So, the policy allowed burning of coal in the North. If the mean air pollution north was the Huai river was compared to that mean of air pollution south of the Huai river, that would provide nothing about the causal effect of burning coal and the effect it has on air pollution. This is because of the large size and diversity of landscape in China. While it might seem intuitive to do this, there could be other contributing factors to air pollution that might affect one are differently from the other. The difference between cities that border Mongolia are much different from cities that border North Korea. This would skew results and you would learn little to nothing about the impact of the policy on air pollution.

In a regression discontinuity design, those outside contributing factors are being eliminated due to the experiment being focused on that specific divide. What is being analyzed are towns that are just above the Huai river and just below the Huai river. This is because in theory, towns just above and below this arbitrary line are identical in all the other control variables. Thus you should be able to estimate the casual effect of the policy since just north of the Huai river is identical to just south of the Huai except for the policy. This is known as the continuity assumption, which means that in the absence of the treatment (the Huai River Policy), the outcome variable, air pollution, would change smoothly at the cutoff. In other words, there should be no sudden jump in air pollution at the river unless it is caused by the policy itself This design is exactly how the Ebenstein paper overcame those challenges.

The assignment variable in question is the distance from the Huai river. Specifically, whether the city was located just north of the river versus just south of the river. To reiterate, this assignment variable would yield the casual effect because the cities and towns just north of the river should be identical to cities and towns just south of the river except for the policy. Thus any changes in air pollution can be attributed to the policy.

Question 2

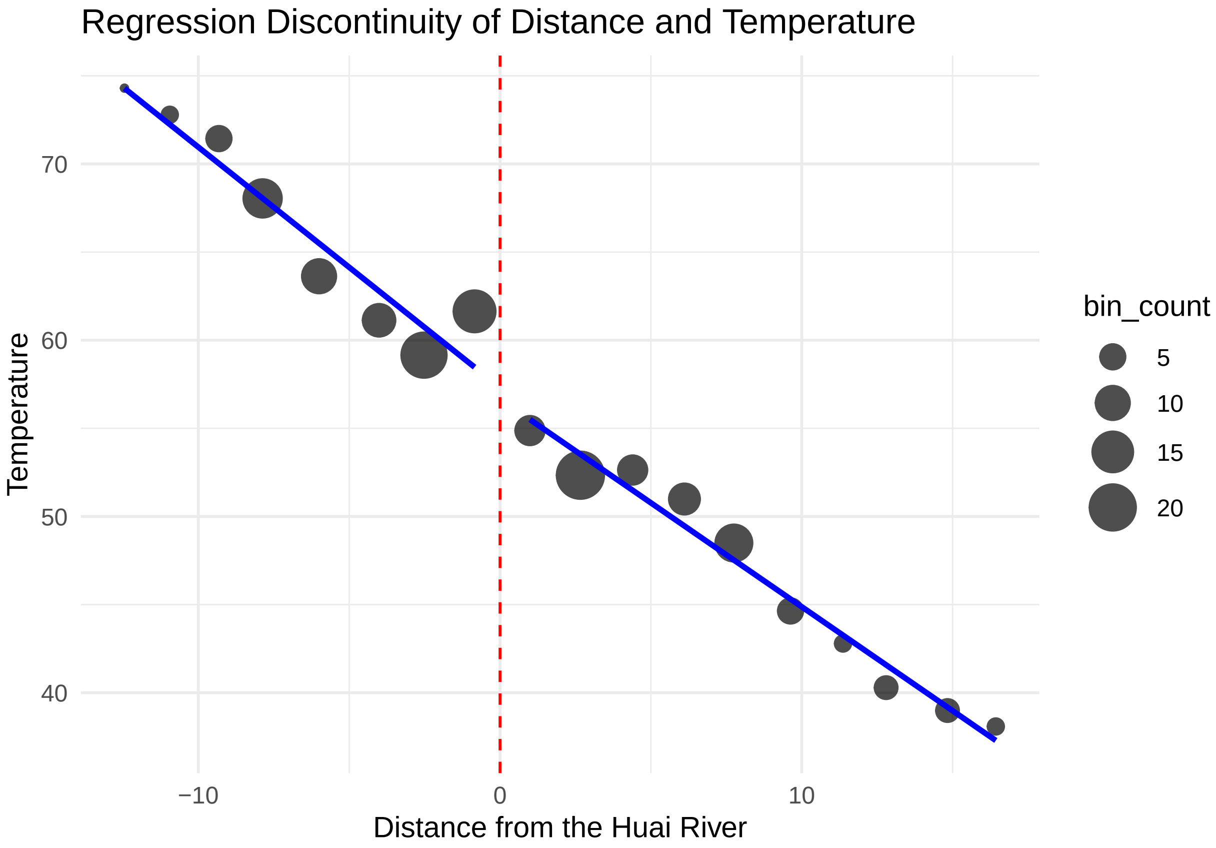


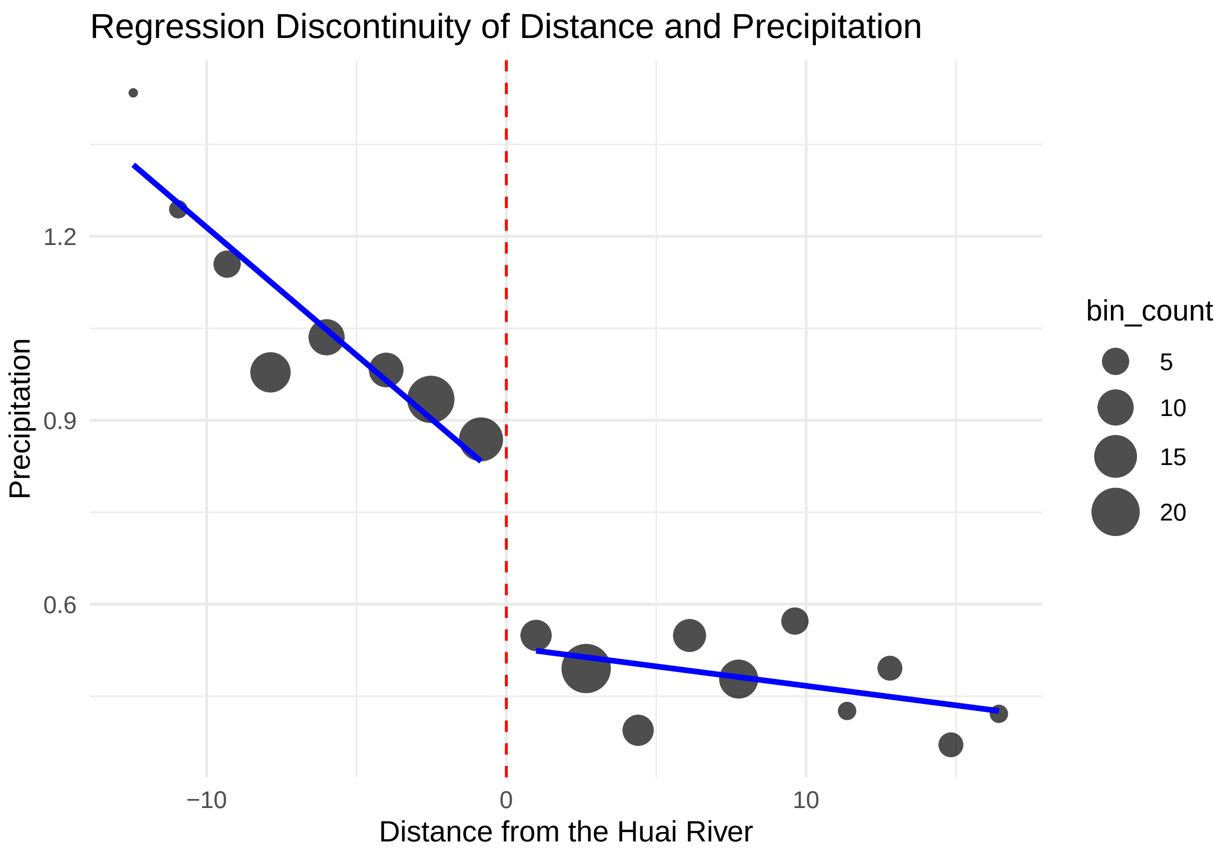
Question 3

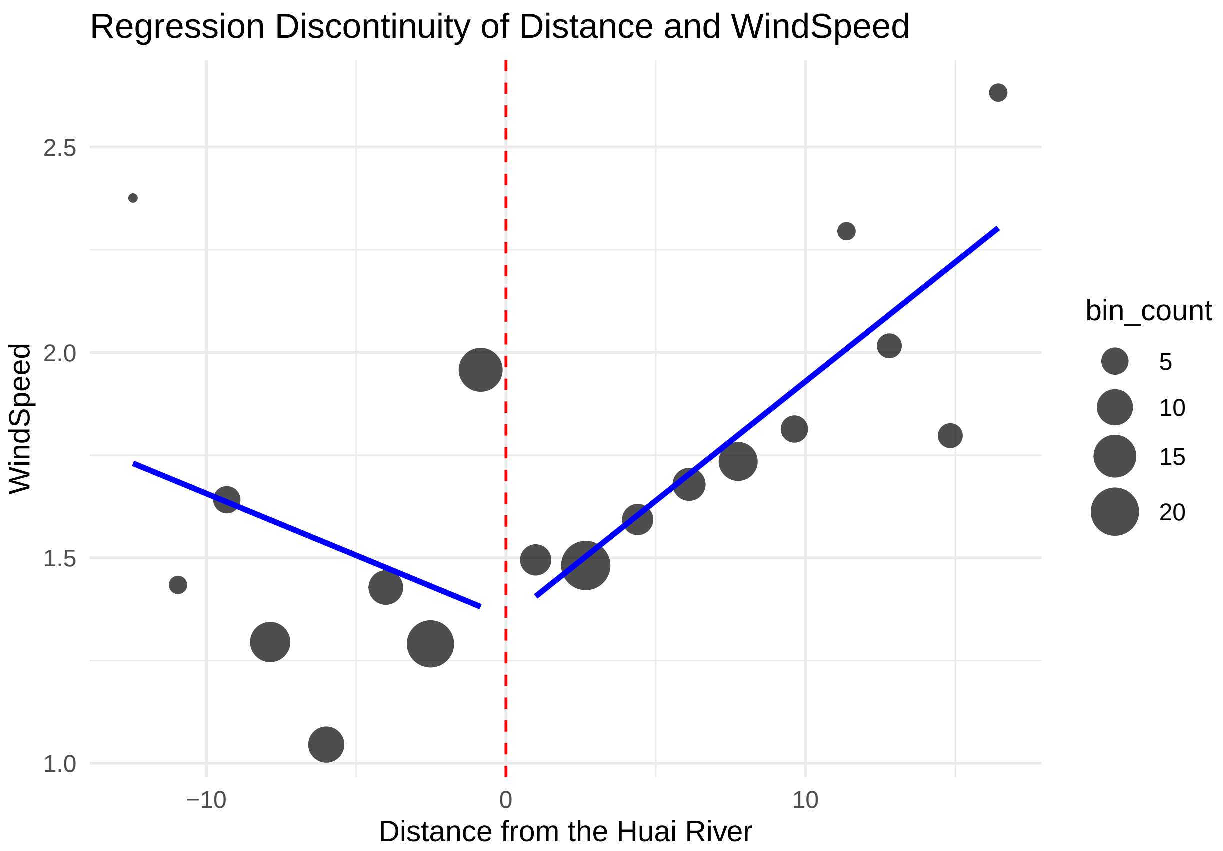
There is no evidence that there is any jump in temperature between either side of the Huai river. The regression has a downward slope across both the areas below the Huai river and the ones above the huai river. This plot can show us regardless of location in proximity to the Huai river, as you move North the temperature tends to drop. This is what we hope for. This infers that temperature is correlated with latitude rather than the Huai river policy itself.

The precipitation binned scatter plot offers a slightly different story. There is a discontinuity at the Huai river. In theory, the towns just above the Huai river and just below the Huai river are identical. This assumption should eliminate any geographical difference between the two. In the western, part of the state it get more mountainous and the river winds heavy. That could explain some of the differences. There is some evidence in other journals that air pollution can affect precipitation. It is doubtful that would affect the drastic change we see in this plot.

Wind speed shows almost no discontinuity at the Huai river. Both regressions show that wind speed in lower closer to the Huai river. That wind speed however, is no different from just south of the Huai river and just north of the Huai river. The farther you get away from the Huai river, the faster the wind speeds get. There is a slight difference in the slopes between the south and the north, however at the boundary line there is no difference.




Regression Discontinuity of Distance and Precipitation
Precipitation
1.2
0.9
0.6
−10
0
10
Distance from the Huai River
bin_count
5
10
15
20



Question 4

Table 1: Comparing RD Robust to OLS

term	RDrobust	OLS
Coefficient Estimate	31.9867191	31.9867191
Standard Error	8.9785723	8.9769092
P-Value	0.0003673	0.0004983

In the table above I have the coefficient estimate on North Huai that both OLS and RDrobust estimated. Then below that I have the robust standard errors for each of the models. As you can see above, these estimates are identical from the RDrobust package and base r's OLS command.

Question 5

Table 2: RD Robust Model Varying Specifications to see Estimates

RobustModel	Estimate	StandardError	PValue
MSE-Optimal (symmetric)	50.165	14.750	0.001
Epanechnikov	48.747	16.487	0.003
Quadratic Polynomial	49.512	12.404	0.000
Difference in Means	42.436	9.753	0.000

For Model B, I used an epanechnikov kernel. In our data, we are measuring observations up to ten degrees of latitude above and below the river. That is a substantial distance when you consider that you could be looking at the differences between two towns that are twenty degrees of latitude away from one another. I chose the epanechnikov kernel to weight observations that are closer to the river heavier than those farther away. This is because observations closer to the river are more likely to be towns of similar stature than those farther away. When measuring discontinuities, I want towns that are identical beside the policy divide. An epanechnikov kernel offers the most statistically optimal and minimizes the mean squared error.

For our coefficient results, it seems that estimates are not affected by the specifications of the model. They do remain stable across different models. Stable in this case meaning a small range of eight or our beta ranging from 42 to 50. There are slight fluctuations in the estimates of the model, but each estimate is well within a range suitable to claim stability. When an estimate relies heavily on model specifications that brings into questions about the robustness of our results. These findings might suggest that the results are more from methodological decisions rather than a stable underlying effect. There is little to no evidence our model suffers from specification choice and rather does have a stable underlying effect.

Question 6

Table 3: Regression Discontinuity Estimates Across Different Model Specifications

Model	Estimate	StandardError	Bandwidth	BiasBandwidth	Effective_N	PValue
Standard OLS	31.987	8.400	NA	NA	147	0.000
MSE-Optimal	50.165	14.750	5.355	7.833	75	0.001
Epanechnikov	48.747	16.487	4.986	7.331	75	0.003
Quadratic Polynomial	49.512	12.404	5.319	9.224	75	0.000
Difference in Means	42.436	9.753	1.000	1.000	75	0.000

Question 7

The identification assumption is violated when our baseline covariates show a discontinuity at the cutoff, indication of systematic differences between treated and untreated groups. In this study, temperature, precipitation, and wind speed were tested for discontinuity. Temperature and wind speed showed no evidence of a discontinuity at the threshold. This supports the identification validity suggesting that the assignment mechanism (distance from the Huai river) is not correlated with pre-treatment variables temperature and wind speed. This suggest our assignment is just as good as random assignment.

However, precipitation does show a discontinuity at the cutoff. This discontinuity would suggest that there are problems with out identification assumption and raises concerns around the validity of there being a casual effect between air pollution and coal burning or at the least the effect estimate might be biased. This is because north of the river might be systemically different than south of the river due to the jump in precipitation at the cutoff line. As a result, estimated causal effects of coal burning on air pollution may be biased, as precipitation could be an unaccounted confounder influencing pollution levels independently.

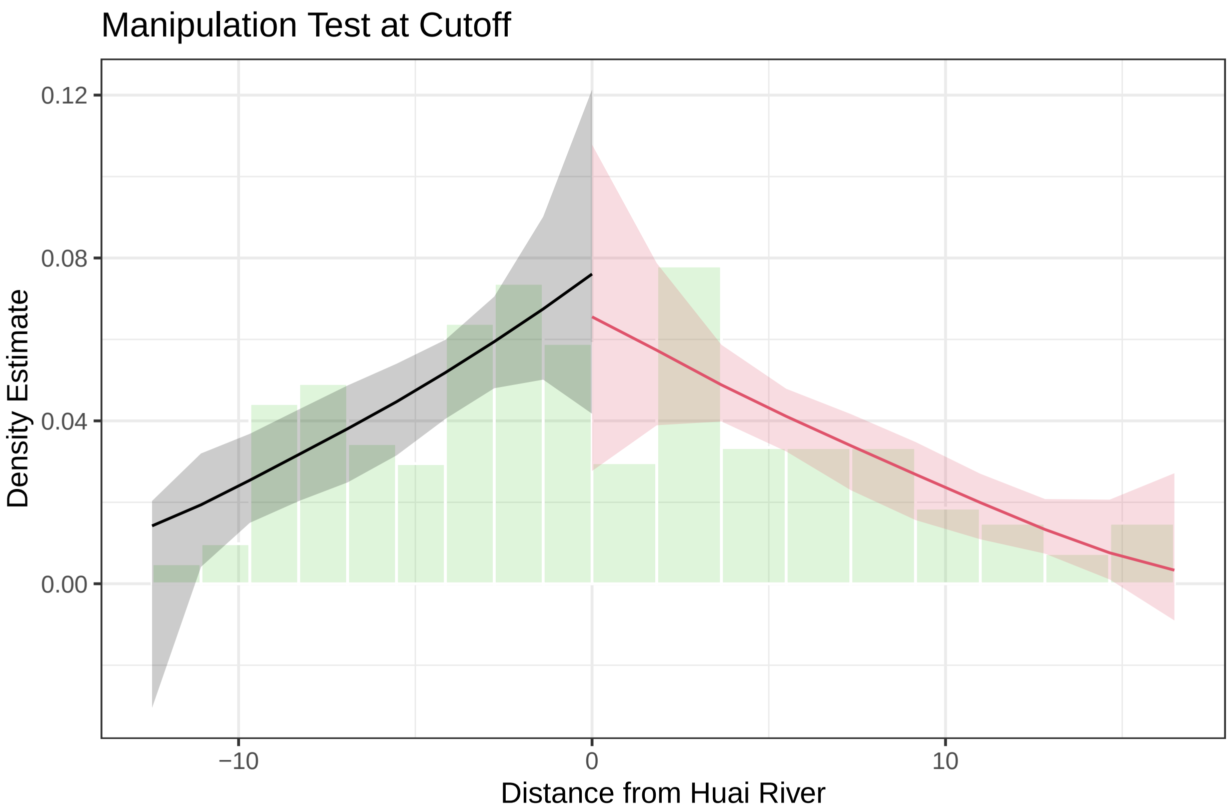
Table 4: Identification Assumption Regression Estimates

Model	Estimate	StandardError	PValue
Temperature	-5.189	4.769	0.277
Wind Speed	-0.450	0.366	0.219
Precipitation	-0.262	0.068	0.000

In the above regression table, temperature and wind speed are not statistically significant. They both have a P-value greater than our acceptable threshold of statistical significance. Precipitation has a non-zero negative coefficient and is statistically significant. This would further lead us to believe that precipitation is violating our identification assumption.

Question 8

In this study there would be no reason to worry about manipulation. Manipulation occurs when someone would give the treatment to an observation just outside of the range of treatment. Usually this happens because the treatment is desirable. An example of this happens in a babies critical weight threshold. Any baby under the 1500 grams receives special neonatal care and any baby over the 1500 gram threshold receive nothing. Since there is no difference between a baby 1499 grams and 1501 grams, a doctor could manipulate the test and give the baby that weighs 1501 grams the treatment. In this study there is no way to manipulate our data. Our assignment variable is distance from the Haui river where the policy boundary is a river. Cities, towns, and rivers are not easily moved and thus it would be difficult for this study to suffer from manipulation.



Data from Regression Discontinuity Analysis

From the above graph the densities appear to have a smooth transition. There is a small jump in density estimates around the cutoff. There are slightly less towns just north of the Huai river. This is small difference and merely a coincidence. As mentioned above, towns and cities can't be moved across the policy divide line. They are rather permanent.

Table 5: Manipulation Test Statistic

Method_Robust	Numerical_Values
Test Statistic	-0.439
P-Value	0.660

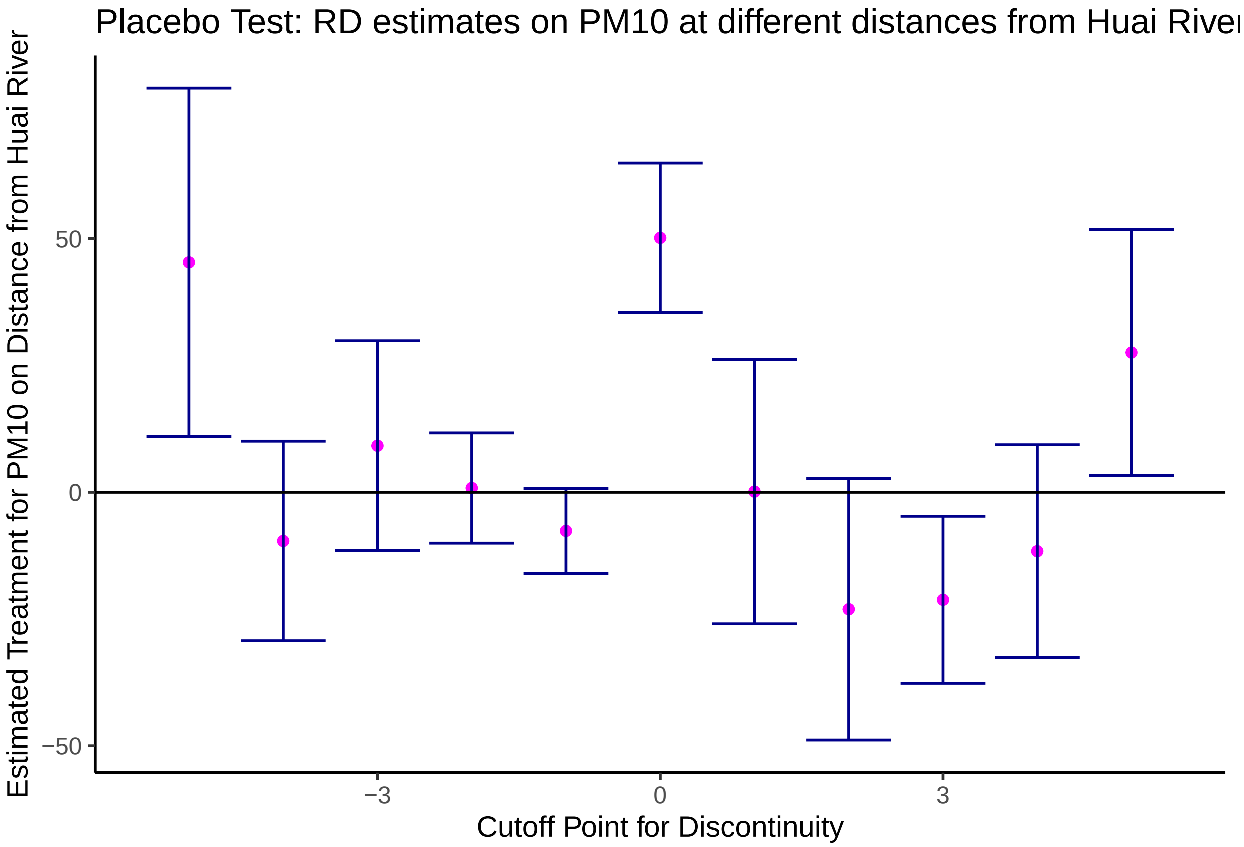
From the above table we have a small test statistic of -0.439 which is less than 1.96 in a two tailed test. This shows that there is no evidence of manipulation. Our P-value is .6 which is way above the threshold for significance. Thus we fail to reject the null hypothesis meaning there is no strong evidence of manipulation at the cutoff.

Question 9

The logic behind a placebo test is to test random locations at different or false cutoffs to try to measure effects. If there are effects at different cutoff points than that would lead us to believe that our estimates are actually from background variation rather than a true underlying effect. If there is no discontinuity at any false cutoffs then that would be strong supporting evidence that there is an underlying effect and background variation is not playing into our overarching results.

The reason the author used false locations is to test for spurious discontinuities. If there were spurious discontinuities than that would support evidence that there is background variation rather than an underlying effect. By using false locations, the author could differentiate whether the effect was caused by this variation or by a true underlying effect.

Figure 4 in the Ebstien paper, shows that there are strong differences in coefficient estimates at the cutoff of zero. This effect isn't shown at any other cutoff. Most effects look almost identical in their coefficient estimates and confidence intervals in that they are zero or no different from being zero. This would pass our placebo test and provide further evidence that there is a true underlying effect rather than background variation.



Question 10

In terms of this study, the effect from the regression discontinuity can be described as a first stage regression estimate. This is similar to the Oregon health experiment in that we are using multi-link casual chain to analyze effects of some treatment. For example, in this study we are analyzing a policy boundary where above the boundary coal is allowed to be burned and below it isn't allowed. Then we are measuring the discontinuity to measure PM10, also known are air pollution. We are then adding life expectancy to that casual chain. Where we get coal burning leads to more air pollution which leads to lower life expectancy.

The estimates in this paper are fuzzy estimates. Although there is a large discontinuity, the discontinuity is not strictly different between above and below the policy cutoff. North and south of the boundary experiences different levels of air pollution. This only affects the probability of exposure to air pollution. Since PM10 levels vary due to other factors this is a fuzzy estimate. The same logic follows for the results of the paper. The author of the paper had to use an instrumental variable approach to estimate casual effects. This instrumental variable approach is not required for sharp estimates. This study was replicated from the author's study and follows the same logic when it comes to different levels of air pollution within the north group and different levels of air pollution within the south group.

This estimate can only be seen as a local average treatment effect and not an average treatment effect. This is because the generality is only local. The effect of PM10 exposure is only for people whom are affected by the policy cutoff. The effect isn't homogeneous of the whole north because not everyone is the north is affected the same amount. The pollution varies even in the north through environmental factors. We should be weary of it because it is not representative of the entire population.

Code for this Project

```r
#reading in libraries
library(tidyverse)
library(ggplot2)
library(rdrobust)
library(rddensity)
library(knitr)
library(sandwich)
library(lmtest)




#reading in Data

River <- read.csv("huairiver.csv")
```

Question 2

```r
# binning our data frame

bin_width <- 1.75  # adjust the bin width
```

```r
binned_River <- River %>% #binning the data set
  mutate(
    bin = case_when(
      dist_huai < 0 ~ floor(dist_huai / bin_width) * bin_width, # everything below zero
      dist_huai >= 0 ~ floor(dist_huai / bin_width) * bin_width # everything above zero
    )
  ) %>%
  group_by(bin) %>%  # grouping by the means
  summarize(
    dist_bin = mean(dist_huai),
    PM10_bin = mean(pm10),
    Temperature = mean(temp),
    Precipitation = mean(prcp),
    WindSpeed = mean(wspd),
    bin_count = n(), # to get the count for the scatter plot
    .groups = "drop"
  )

# getting the slope and coefficient

binned_River_below <- binned_River %>% # filtering below zero
  filter(dist_bin < 0)
binned_River_above <- binned_River %>% # filtering above zero
  filter(dist_bin > 0 )

# used to grab slope and intercept

below_line <- lm(PM10_bin ~ dist_bin, binned_River_below) # used with geom_abline
above_line <- lm(PM10_bin ~ dist_bin, binned_River_above)



#ggplot(binned_River, aes(x = dist_bin, y = PM10_bin)) +
#  geom_point() +
#  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +# line to divide by the policy
  #geom_abline(intercept = coef(below_line)[1], slope = coef(below_line)[2]) + # getting lines
  #geom_abline(intercept = coef(above_line)[1], slope = coef(above_line)[2]) + #getting lines
#  geom_smooth(data = binned_River_below, method = "lm", se=FALSE, color="blue") +
#  geom_smooth(data = binned_River_above, method = "lm", se=FALSE, color="blue") +
#  labs(title = "Regression Discontinuity of #Distance and Air Pollution",
#       x = "Distance from the Huai River",
#       y = "Air Pollution (PM10)")


ggplot(binned_River, aes(x = dist_bin, y = PM10_bin)) +
  geom_point(aes(size = bin_count), alpha = 0.7) + # Adjust point size based on precomputed bin density
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  geom_smooth(data = binned_River_below, method = "lm", se = FALSE, color = "blue") +
  geom_smooth(data = binned_River_above, method = "lm", se = FALSE, color = "blue") +
  labs(title = "Regression Discontinuity of Distance and Air Pollution",
       x = "Distance from the Huai River",
       y = "Air Pollution (PM10)") +
  scale_size_continuous(range = c(1, 8)) + # Adjust size scaling
  theme_minimal()
```

Question 3

```
loop_names = c("Temperature", "Precipitation", "WindSpeed") # getting a list to iterate over

for (n in loop_names){
p <- ggplot(binned_River, aes_string(x = "dist_bin", y = n)) +
  geom_point(aes(size = bin_count),alpha = 0.7) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +# line to divide by the policy
  geom_smooth(data = binned_River_below, method = "lm", se=FALSE, color="blue") +
  geom_smooth(data = binned_River_above, method = "lm", se=FALSE, color="blue") +
  labs(title = paste("Regression Discontinuity of Distance and", n),
       x = "Distance from the Huai River",
       y = n) +
  scale_size_continuous(range = c(1,8)) +
  theme_minimal()
print(p)
}
```

Question 4

```
# creating a variable of treat of whether the observation was above of below the river

binned_River$treat <- ifelse(binned_River$dist_bin > 0, 1, 0)
River$treat <- ifelse(River$dist_huai > 0, 1, 0) # i never used this because it exists in the data

# running our regression

model_one <- lm(pm10 ~ north_huai + dist_huai + dist_huai*north_huai, data=River)
robust_se <- vcovHC(model_one, type = "HC1")



model_robust <- rdrobust(y = River$pm10, x = River$dist_huai, vce = "hc1", p = 1, h = 50,
                         kernel = "uniform", bwselect = "mserd")



robust_results <- data.frame(
  term = c("Coefficient Estimate", "Standard Error", "P-Value"),
  RDrobust = c(model_robust$coef[1], model_robust$se[1], model_robust$pv[1]),
  OLS = c(model_one$coefficients[2], coeftest(model_one, vcov = robust_se)[6],
          coeftest(model_one, vcov = robust_se)[14])
```

```r
)

kable(robust_results, caption = "Comparing RD Robust to OLS")
```

Question 5

```r
# A


model_robust_A <- rdrobust(y = River$pm10, x = River$dist_huai, bwselect = "mserd", vce = "hc1")
# c is zero in our data and that is the default value for the cutoff

# B
model_robust_B <- rdrobust(y = River$pm10, x = River$dist_huai,
                           bwselect = "mserd", kernel = "epanechnikov", vce = "hc1")

# C
model_robust_C <- rdrobust(y = River$pm10, x = River$dist_huai,
                           bwselect = "mserd", kernel = "epanechnikov", p = 2, vce = "hc1")

# E
model_robust_E <- rdrobust(y = River$pm10, x = River$dist_huai,
                           bwselect = "mserd", kernel = "epanechnikov", p = 3, vce = "hc1")

# D

model_robust_D <- rdrobust(y = River$pm10, x = River$dist_huai,
                            bwselect = "mserd", kernel = "epanechnikov", p = 1,
                           h = 2, vce = "hc1" )

#extract results to put in a table
```

```r
coef_A <- model_robust_A$coef[1]
SE_A <- model_robust_A$se[1]
pval_A <- model_robust_A$pv[1]

coef_B <- model_robust_B$coef[1]
SE_B <- model_robust_B$se[1]
pval_B <- model_robust_B$pv[1]

coef_C <- model_robust_C$coef[1]
SE_C <- model_robust_C$se[1]
pval_C <- model_robust_C$pv[1]

coef_D <- model_robust_D$coef[1]
SE_D <- model_robust_D$se[1]
pval_D <- model_robust_D$pv[1]

coef_E <- model_robust_E$coef[1]
SE_E <- model_robust_E$se[1]
pval_E <- model_robust_E$pv[1]

# putting it into separate data frames

rd_table_A <- data.frame(RobustModel = c("MSE-Optimal (symmetric)", "Epanechnikov", "Quadratic Polynomia
                         Estimate = c(coef_A, coef_B,coef_C,coef_D),
                         StandardError = c(SE_A, SE_B,SE_C,SE_D),
                         PValue = c(pval_A,pval_B,pval_C, pval_D))

# round the results

names <- c("RobustModel", "Estimate", "StandardError", "PValue") #creating a vector to iterate over

for (name in names) {
  # if it is a character vector do nothing
  if (is.character(rd_table_A[[name]]) == TRUE){
    rd_table_A[[name]] <- rd_table_A[[name]]}
  # else round it to three digits
  else {
    rd_table_A[[name]] <- round(rd_table_A[[name]], digits = 3)
  }
}

# making the table
 kable(rd_table_A, caption = "RD Robust Model Varying Specifications to see Estimates")

#summary(model_robust_A)
#summary(model_robust_B)
#summary(model_robust_C)
#summary(model_robust_D)
```

Question 6

```
# Creating a data frame with all necessary RD model results

rd_table_B <- data.frame(
  Model = c("Standard OLS", "MSE-Optimal", "Epanechnikov", "Quadratic Polynomial", "Difference in Means
  Estimate = c(model_one$coefficients[2], coef_A, coef_B, coef_C, coef_D),
  StandardError = c(summary(model_one)$coefficients[2, 2], SE_A, SE_B, SE_C, SE_D),
  Bandwidth = c(NA,model_robust_A$bws[[1]], model_robust_B$bws[[1]], model_robust_C$bws[[1]], model_robu
  BiasBandwidth = c(NA,model_robust_A$bws[[2]], model_robust_B$bws[[2]], model_robust_C$bws[[2]], model_
  Effective_N = c(nrow(River), model_robust_A$N[1], model_robust_B$N[1], model_robust_C$N[1], model_robu
  PValue = c(coeftest(model_one, vcov = robust_se)[14],pval_A,pval_B,pval_C,pval_D)
)

# Round numerical results for better readability
```

```
names <- c("Estimate", "StandardError", "Bandwidth","BiasBandwidth", "Effective_N", "PValue")

for (name in names) {
  rd_table_B[[name]] <- round(rd_table_B[[name]], digits = 3)
}

# Generate the table

kable(rd_table_B, caption = "Regression Discontinuity Estimates Across Different Model Specifications")
```

Question 7

```
# creating our models

model_temp <- rdrobust(y = River$temp, x = River$dist_huai, bwselect = "mserd", vce = "hc1")
model_wind <- rdrobust(y = River$wspd, x = River$dist_huai, bwselect = "mserd", vce = "hc1")
model_precip <- rdrobust(y = River$prcp, x = River$dist_huai, bwselect = "mserd", vce = "hc1")

#inputting those models into a data frame

rd_table_C <- data.frame(
  Model = c("Temperature", "Wind Speed", "Precipitation"),
  Estimate = c(model_temp$coef[1], model_wind$coef[1], model_precip$coef[1]),
  StandardError = c(model_temp$se[1],model_wind$se[1],model_precip$se[1]),
  PValue = c(model_temp$pv[1] , model_wind$pv[1], model_precip$pv[1])
)

# Rounding using a for loop

names <- c("Estimate", "StandardError", "PValue") # creating a vector to iterate over
```

```r
for (name in names) {
  rd_table_C[[name]] <- round(rd_table_C[[name]], digits = 3)
}

# plotting our table

kable(rd_table_C, caption = "Identification Assumption Regression Estimates")
```

Question 8

```r
density <- rddensity(River$dist_huai) # using the density function

density_plot <- rdplotdensity(density,River$dist_huai, noPlot = TRUE) # plotting said density function
# no plot is true so that the plot wont be shown

#density_plot gives out 3 values. To access those values I used a name function then $Estplot

density_plot$Estplot +
  ggtitle("Manipulation Test at Cutoff") +
  xlab("Distance from Huai River") +
  ylab("Density Estimate") +
  labs(caption = "Data from Regression Discontinuity Analysis")
```

```r
round_table_7 <- data.frame(Method_Robust = c("Test Statistic", "P-Value"),
                            Numerical_Values = c(density$test$t_jk, density$test$p_jk))

round_table_7$Numerical_Values <- round(round_table_7$Numerical_Values, digits = 3)

kable(round_table_7, caption = "Manipulation Test Statistic")
```

Question 9

```r
# getting a list to iterate over

false_cutoffs <- c(-5,-4,-3,-2,-1,0,1,2,3,4,5)

# iterating over our list to test each false cutoff

## creating empty vectors to save our results
estimate <- c()
errors <- c()
pvalue <- c()
upper <- c() # for confidence interval
lower <- c()

for (i in false_cutoffs) {
  model <- rdrobust(y = River$pm10, x = River$dist_huai, vce = "hc1", c = i)

  estimate <- c(estimate, model$coef[1])
  errors <- c(errors, model$se[1])
  pvalue <- c(pvalue,model$pv[1])
  upper <- c(upper, model$coef[1] + model$se[1]) # confidence interval
  lower <- c(lower, model$coef[1] - model$se[1])


}

# creating a data frame to be able to graph from
placebo <- data.frame(cutoff = c(false_cutoffs),
                      estimate = c(estimate),
                      errors = c(errors),
                      pvalue = c(pvalue),
                      upper = c(upper),
                      lower = c(lower))


ggplot(data=placebo, aes(x=cutoff,y=estimate)) +
  geom_point(color="magenta") +
  geom_errorbar(aes(ymin=lower, ymax=upper), color="darkblue") +
  geom_hline(yintercept = 0) +
  ggtitle("Placebo Test: RD estimates on PM10 at different distances from Huai River") +
  xlab("Cutoff Point for Discontinuity") +
  ylab("Estimated Treatment for PM10 on Distance from Huai River") +
  theme_classic()
```