Proj2_EC425

Nathan Bracken

5/9/2022

```r
huai_df = here("data", "huairiver.dta") |> read_dta()
```

Specific questions to address in your summary

1. Explain why a simple comparison of air pollution in northern cities versus southern cities would not measure the causal effect of the Huai River Policy. Explain how did the Ebenstein et al. paper overcome this problem by using a regression discontinuity design.

A simple comparison of air pollution in northern cities versus southern cities would not measure the causal effect of the Huai River Policy because as you move further south and north of the Huai river, the surrounding areas begin to differ more and more, weakening the required assumption that all else must be held constant. There are likely properties of the environemnt that will effect the spread of particulate matter in the area. The Ebenstein et al. paper overcame this problem using a regression discontinuity design by identifying the Huai river as the cutoff point between the treated and untreated populations for coal heating and then compared the groups closest to the river on either side. By making the comparison of the two groups, the assumption can hold that the groups on either side of the river are very similar and the difference in life expectancy and particulate matter can be estimated.

2. Explain what is the outcome variable and what is the assignment variable in Fig.2 of the Ebenstein et al. paper?

The outcome variable in figure 2 is PM_{10} micro grams per meters cubed. The assignment variable in figure 2 is Degrees north of the Huai river boundary. The outcome variable is a measure of the density of particulate matter in a given area and the assignment variable is the distance north and south of the river. The researchers are testing if there is an increase in particulate matter in the group on the north side of the river (with a positive count on the assignment variable) as compared to the group on the south side of the river (a negative count on the assignment variable).

3. What is a binned scatter plot? Explain how it is constructed. Hint: Binscatter plot is covered in week 1's lecture.

A binned scatter plot is a scatter plot that consolidates the scatter points into groups based on a mean reporting metric rather than plotting each individual point. By consolidating each individual point into a group, the readers of the paper can more easily see what points are driving the trends in the charts.

4. Graphical regression discontinuity analysis.

a. Draw a binned scatter plot to visualize how PM10 changes at the Huai River line. Display fitted lines (linear, or quadratic, or whatever functional form you see fit) based on what you see in the data.

```r
# dist bin
huai_df = huai_df %>%
  mutate(dist_bin = cut(dist_huai,
                        breaks = 31,
                        na.rm = TRUE))
```

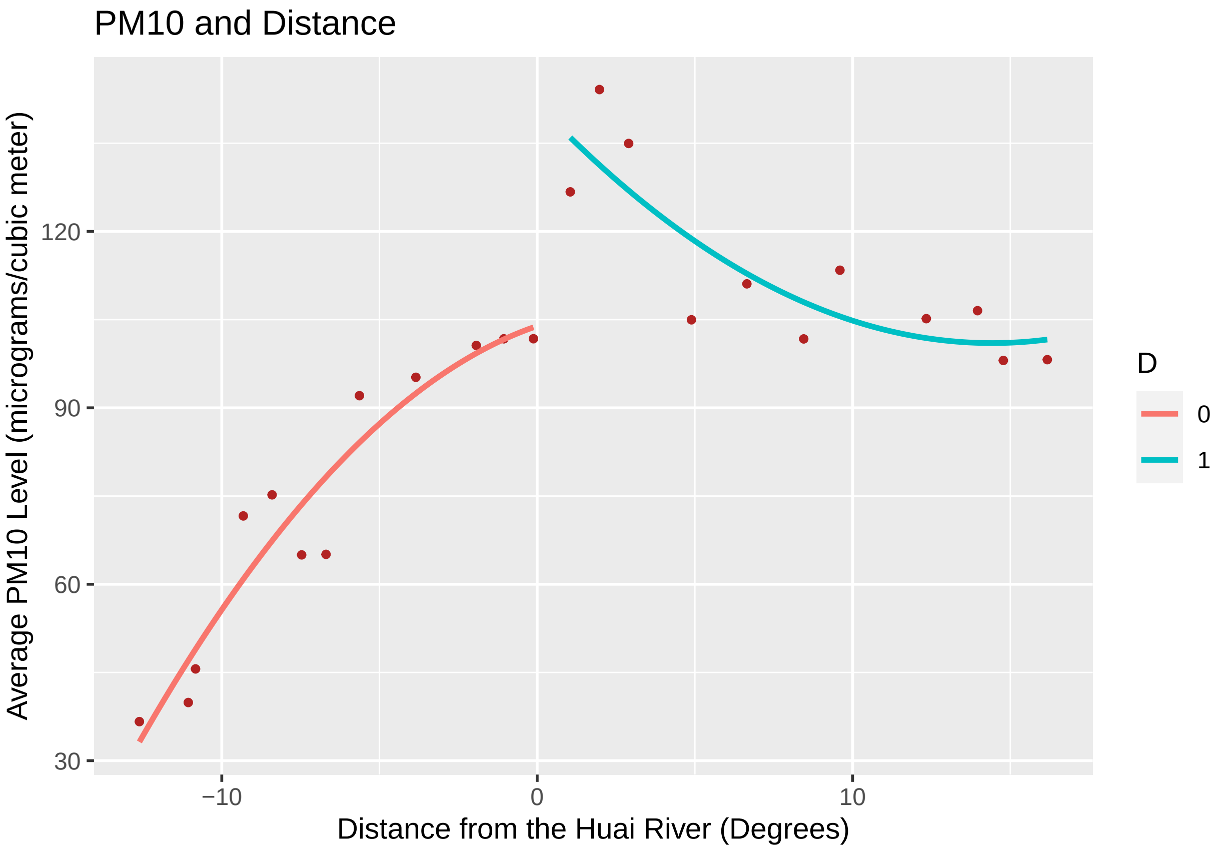
```
# Checking to see if the bins are factors
is.factor(huai_df$dist_bin)
```

```
## [1] TRUE
```

```
# Checking # bin obs, we do not waant them to be equal because not all of the locations are same
table(huai_df$dist_bin)
```

```
## 
##  (-12.8,-11.8]  (-11.8,-10.9]  (-10.9,-9.95]     (-9.95,-9]     (-9,-8.06] 
##              2              1              1              5              6 
##  (-8.06,-7.11]  (-7.11,-6.17]  (-6.17,-5.23]  (-5.23,-4.28]  (-4.28,-3.34] 
##              6              5              7              4             10 
##   (-3.34,-2.4]   (-2.4,-1.45] (-1.45,-0.509] (-0.509,0.434]   (0.434,1.38] 
##              8             14              7              7              5 
##    (1.38,2.32]    (2.32,3.26]    (3.26,4.21]    (4.21,5.15]     (5.15,6.1] 
##              9             10              8              5              5 
##     (6.1,7.04]    (7.04,7.98]    (7.98,8.93]    (8.93,9.87]    (9.87,10.8] 
##              6              7              6              3              3 
##    (10.8,11.8]    (11.8,12.7]    (12.7,13.6]    (13.6,14.6]    (14.6,15.5] 
##              0              4              0              2              2 
##    (15.5,16.5] 
##              3
```

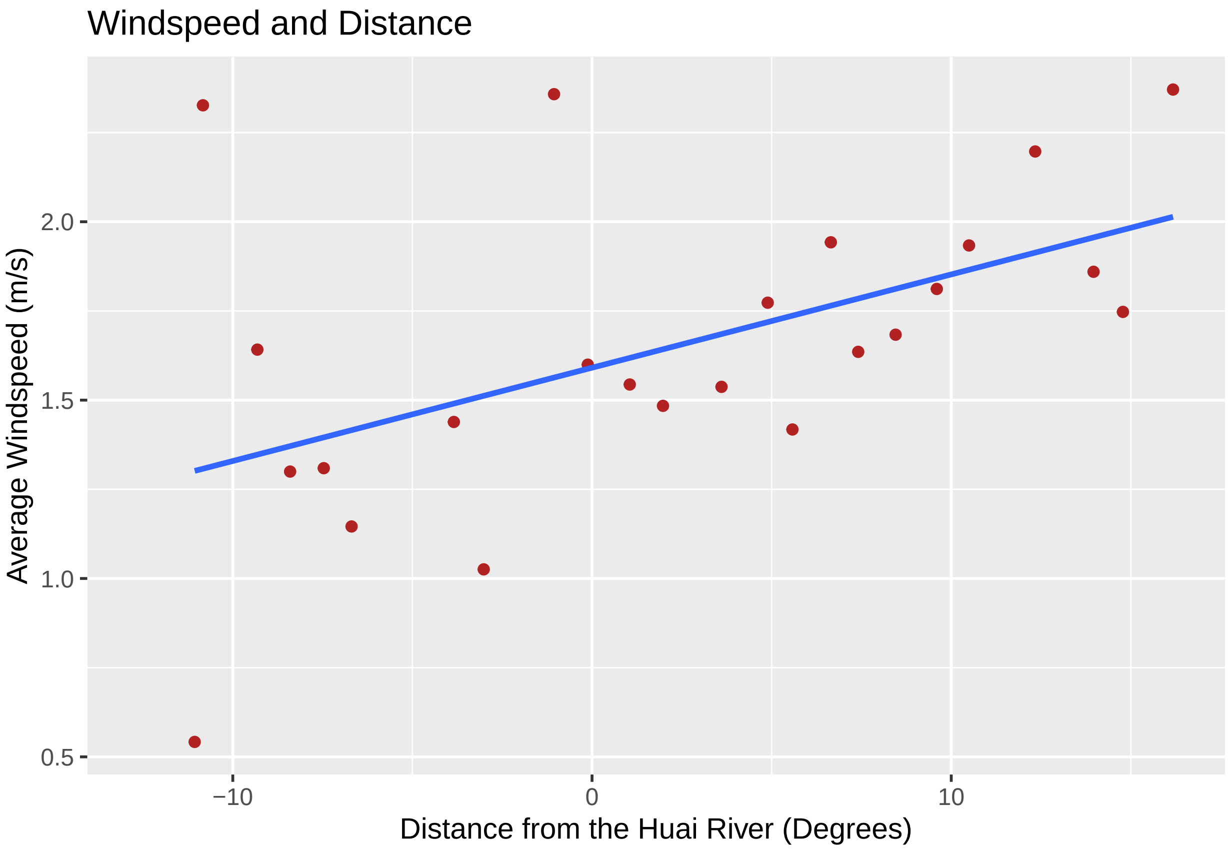
```
huai_df %>%
group_by(dist_bin) %>%
summarise(dist_huai = mean(dist_huai), pm10 = mean(pm10)) %>%
mutate(D = as.factor(ifelse(dist_huai >= 0, 1, 0))) %>%
ggplot(aes(x = dist_huai, y = pm10, color = D)) +
geom_point(colour = "firebrick", size = 1, alpha = 1) +
stat_smooth(method = "lm", formula = y ~ x + I(x^2), se = FALSE) +
xlab("Distance from the Huai River (Degrees)") +
ylab("Average PM10 Level (micrograms/cubic meter)") +
ggtitle("PM10 and Distance")
```



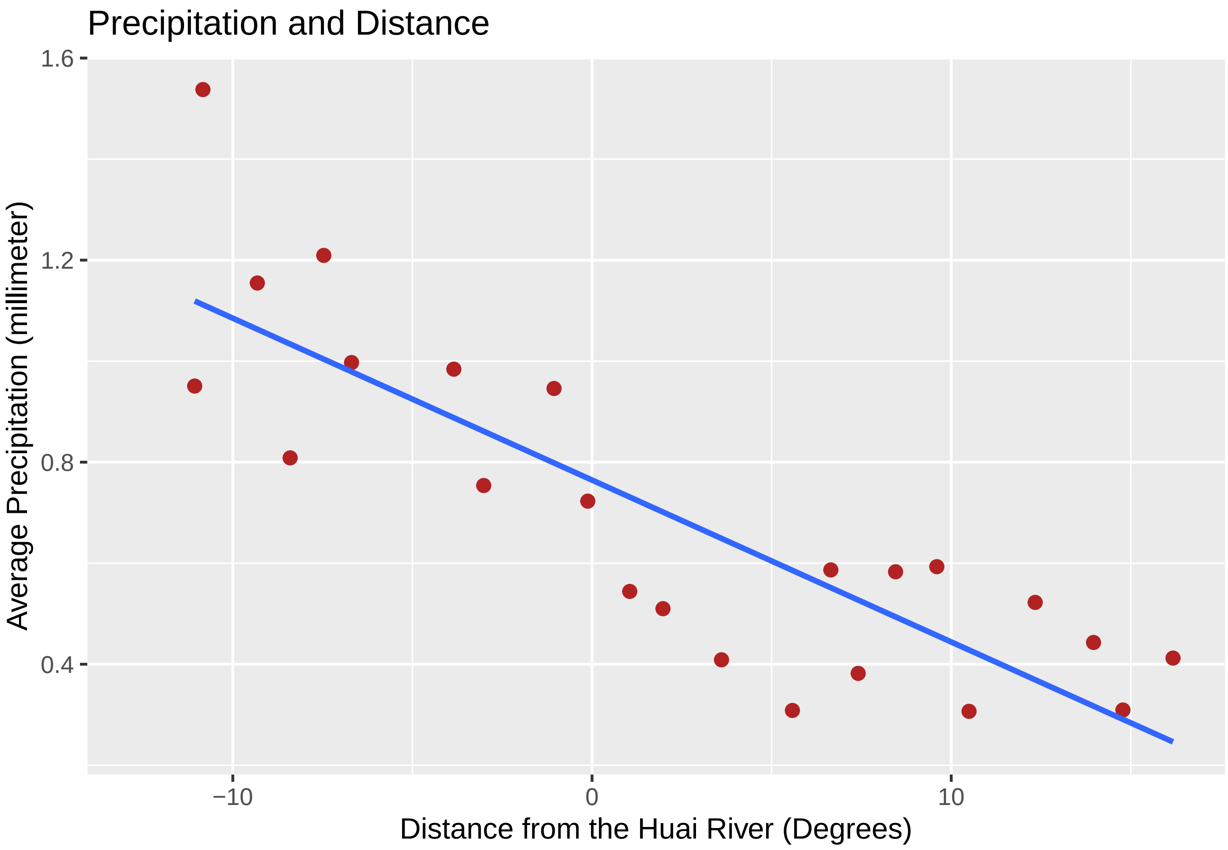
b. Draw binned scatter plots to test whether (i) temperature, (ii) precipitation, and (iii) wind speed changes at the Huai River line. Display fitted lines (linear, or quadratic, or whatever functional form you see fit) based on what you see in the data.

I selected linear forms for each of the binned scatter plots because I feel that these charts most clearly represent the relationship between the variables. There seems to be a direct linear relationship between distance and each of the environmental variables.

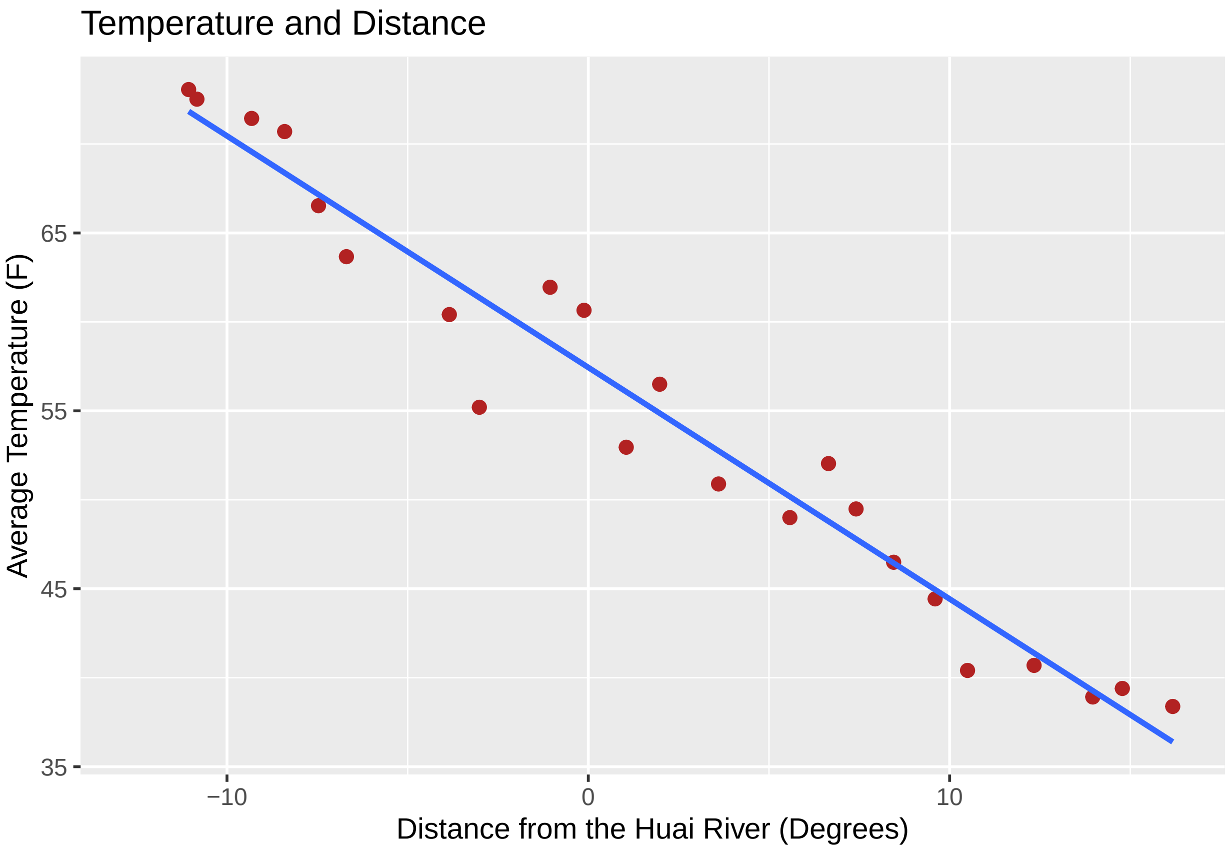
```
huai_df %>%
group_by(dist_bin) %>%
summarise(dist_huai = mean(dist_huai), wspd = mean(wspd)) %>%
ggplot(aes(x = dist_huai, y = wspd)) +
geom_point(colour = "firebrick", alpha = 1) +
stat_smooth(method = "lm", se = FALSE) +
xlab("Distance from the Huai River (Degrees)") +
ylab("Average Windspeed (m/s)") +
ggtitle("Windspeed and Distance")
```



```
huai_df %>%
group_by(dist_bin) %>%
summarise(dist_huai = mean(dist_huai), precip = mean(prcp)) %>%
ggplot(aes(x = dist_huai, y = precip)) +
geom_point(colour = "firebrick", size = 2, alpha = 1) +
geom_smooth(method = "lm", se = FALSE) +
xlab("Distance from the Huai River (Degrees)") +
ylab("Average Precipitation (millimeter)") +
ggtitle("Precipitation and Distance")
```



```
huai_df %>%
group_by(dist_bin) %>%
summarise(dist_huai = mean(dist_huai), temp = mean(temp)) %>%
ggplot(aes(x = dist_huai, y = temp)) +
geom_point(colour = "firebrick", size = 2, alpha = 1) +
geom_smooth(method = "lm", se = FALSE) +
xlab("Distance from the Huai River (Degrees)") +
ylab("Average Temperature (F)") +
ggtitle("Temperature and Distance")
```



5. Regression analysis. Run the regressions that correspond to your three graphs in 4a and 4b to quantify the discontinuities that you see in the data. Report a 95% confidence interval for each of these estimates. Note: Part of this question is to get you to think about how to do bandwidth selection in a regression discontinuity design. Carefully read through how Ebenstein et al. choose 2 bandwidth (e.g., page 4 of the paper) and the associated material in the SI Appendix. Try your best to replicate their bandwidth selection in your work.

```
# Specifying regression models
# eric said that for the scope of this class the bwselect() automatic selection is appropriate
lm1 = rdrobust(y = huai_df$pm10, x = huai_df$dist_huai, all = TRUE)
lm2 = rdrobust(y = huai_df$wspd, x = huai_df$dist_huai, all = TRUE)
lm3 = rdrobust(y = huai_df$temp, x = huai_df$dist_huai, all = TRUE)
lm4 = rdrobust(y = huai_df$prcp, x = huai_df$dist_huai, all = TRUE)

# printing each of the called models
summary(lm1)
```

```
## Call: rdrobust
##
## Number of Obs.                  154
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   80           74
## Eff. Number of Obs.              47           37
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   5.268        5.268
## BW bias (b)                   7.995        7.995
## rho (h/b)                     0.659        0.659
## Unique Obs.                      80           74
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    50.197    14.492     3.464     0.001    [21.792 , 78.602]
## Bias-Corrected    54.412    14.492     3.754     0.000    [26.007 , 82.816]
##         Robust    54.412    15.702     3.465     0.001    [23.637 , 85.187]
## =============================================================================
```

```
summary(lm2)
```

```
## Call: rdrobust
##
## Number of Obs.                  156
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   78           78
## Eff. Number of Obs.              31           20
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   3.125        3.125
## BW bias (b)                   4.655        4.655
## rho (h/b)                     0.671        0.671
## Unique Obs.                      78           78
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    -0.447     0.377    -1.185     0.236    [-1.187 , 0.292]
## Bias-Corrected    -0.416     0.377    -1.103     0.270    [-1.156 , 0.323]
##         Robust    -0.416     0.444    -0.937     0.349    [-1.287 , 0.455]
## =============================================================================
```

```
summary(lm3)
```

```
## Call: rdrobust
##
## Number of Obs.                  153
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   76           77
## Eff. Number of Obs.              38           29
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   3.699        3.699
## BW bias (b)                   5.970        5.970
## rho (h/b)                     0.620        0.620
## Unique Obs.                     76          77
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    -5.447     5.663    -0.962     0.336   [-16.547 , 5.653]
## Bias-Corrected    -5.409     5.663    -0.955     0.340   [-16.509 , 5.691]
##         Robust    -5.409     7.215    -0.750     0.453   [-19.550 , 8.732]
## =============================================================================
```

```
summary(lm4)
```

```
## Call: rdrobust
##
## Number of Obs.                153
## BW type                     mserd
## Kernel                 Triangular
## VCE method                     NN
##
## Number of Obs.                 76          77
## Eff. Number of Obs.            40          29
## Order est. (p)                  1           1
## Order bias  (q)                 2           2
## BW est. (h)                 3.861       3.861
## BW bias (b)                 6.204       6.204
## rho (h/b)                   0.622       0.622
## Unique Obs.                    76          77
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    -0.271     0.070    -3.855     0.000    [-0.408 , -0.133]
## Bias-Corrected    -0.269     0.070    -3.825     0.000    [-0.406 , -0.131]
##         Robust    -0.269     0.083    -3.238     0.001    [-0.431 , -0.106]
## =============================================================================
```

6. Recall that any quasi experiment requires an identification assumption to make it as good as an experiment. What is the identification assumption for regression discontinuity design? Explain whether your graphs in 4b are consistent with that assumption.

The identification assumption for regression discontinuity design is the covariate smoothness test. My graphs in 4b are consistent with the assumption that the cutoff does not contribute to some change in the outcome variable. We do not see a large jump in the data when looking at the environment variables and the distance compared to the Huai river. Temperature and precipitation are very smooth, while windspeed is slightly less smooth with lots of variation south of the river and low variation above the river.

7. Another type of validity test for regression discontinuity design is the manipulation test. Do we need to worry about manipulation in this study context? Explain why or why not. If you believe a manipulation test should be done, report such a test.

We do not need to worry about manipulation in this context because the treatment group was assigned treatment based upon their location in reference to the river and people were not allowed to move to receive coal. A manipulation test would tell us if people are further from the cutoff, they have a smaller chance of receiving the treatment. The line between treatment and not treatment is very clear with it being the river. The environment that the individuals face is generally shared over an area so any manipulation in the data would have to be across a large section of the sample, which is unlikely.

8. Consider the “placebo test” in Fig. 4 of the Ebenstein et al. paper.

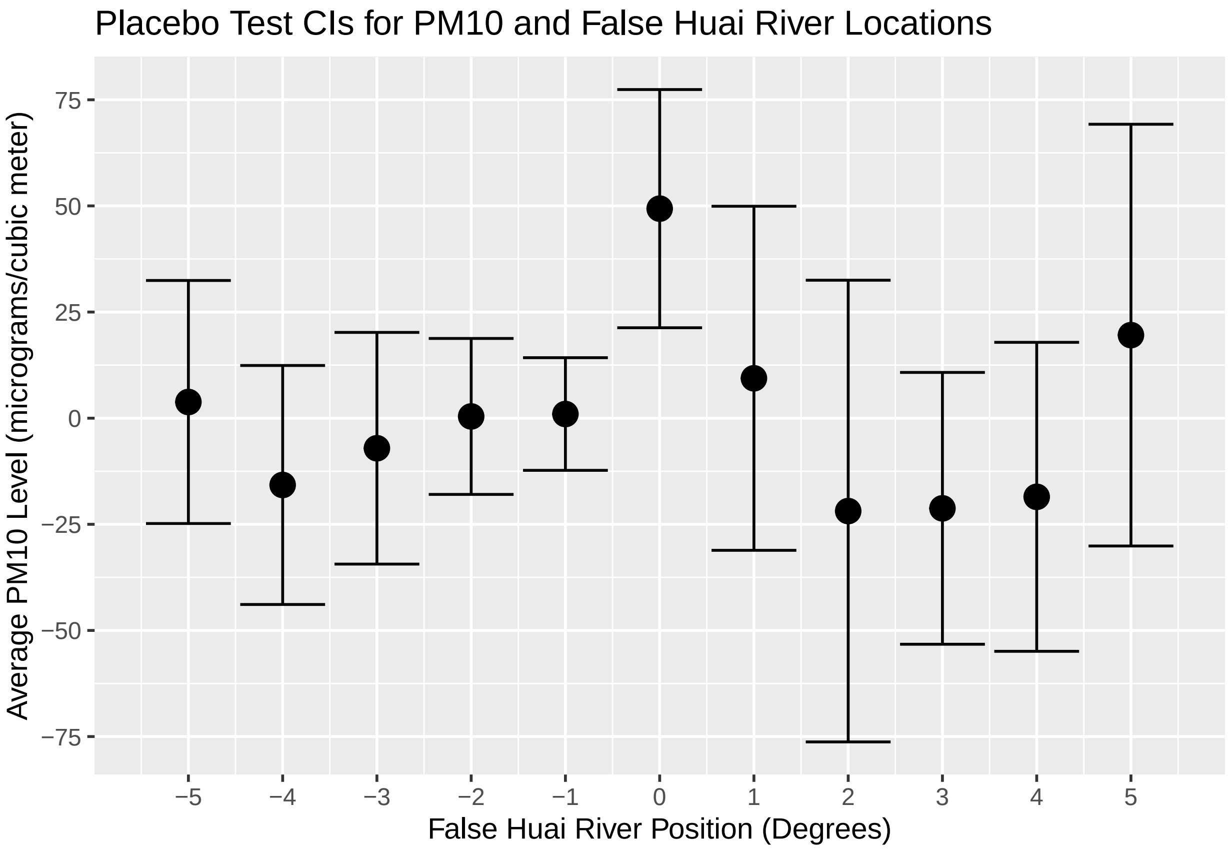
a. Explain the logic of the “placebo test” underlying Fig. 4. Why did the authors estimate regression discontinuity using false locations of the Huai River? What do the results of this test tell us?

For the placebo test, the authors estimated regressions by falsely placing the huai river at different distances north and south of the actual Huai river. The authors wanted to test to see if their were statistically significant life expectancy outcomes in other places if the river would have been there to estimate some sort of “spurious” relationship between locations. By doing so, they can see if the Huai river is actually the cutoff between the two groups with the coal being the treatment that makes the difference in the life expectancy.

b. Replicate Fig. 4 of Ebenstein et al. (2017). Hint: To obtain cities’ distance to a “placebo” Huai River that is 1-degree North of the true Huai River, simply add 1 to the “dist_huai” variable.

```
# writing a function to iterate the various placebo haui river locations
p_tib = map_dfr(.x = -5:5, .f = function(i){
  # c is the cutoff which we have set to i
  # h is the bandwidth size which I selected to be 5 from above PM10 regression
   rdrob = rdrobust(y = huai_df$pm10, x = huai_df$dist_huai, c = i, h = 5)
      tibble(
      cutoff = i,
      estimate = rdrob$coef[1],
      std.error = rdrob$se[1],
      ci_low = estimate - (1.96 * std.error),
      ci_high = estimate + (1.96 * std.error)
    )
})
```

```
ggplot(data = p_tib, aes(x = cutoff, y = estimate)) +
    geom_point(size = 4) +
    geom_errorbar(aes(ymax = ci_high, ymin = ci_low)) +
    scale_y_continuous(breaks=c(-75,-50,-25,0,25,50,75)) +
    scale_x_continuous(breaks=c(-5,-4,-3,-2,-1,0,1,2,3,4,5))+
  xlab("False Huai River Position (Degrees)") +
  ylab("Average PM10 Level (micrograms/cubic meter)") +
  ggtitle("Placebo Test CIs for PM10 and False Huai River Locations")
```


Placebo Test CIs for PM10 and False Huai River Locations
Average PM10 Level (micrograms/cubic meter)
75
50
25
0
−25
−50
−75
−5
−4
−3
−2
−1
0
1
2
3
4
5
False Huai River Position (Degrees)