HW2

Erik Andersen

2022-05-20

Contents

Setup

```r
# Load packages

pacman::p_load(tidyverse, broom, haven, data.table, here, magrittr, stargazer, DT)

# Load data

huairiver_df = read_dta(here("data", "huairiver.dta"))
```

Question 1

A simple comparison of air pollution across northern and southern cities would not measure the causal effect of the policy because of confounding variables. It is unlikely that people to the north and south of the Huai River would use indoor coal furnaces equally even if they had them. The area to the north of the river is likely colder than the area to the south, so they would have more need for heat. So, the coal furnaces are likely uses more in the north than they would be in the south. A simple comparison of the levels of air pollution would miss this, and so over estimate the effect of the policy.

The regression discontinuity overcomes this problem because it measures the discontinuous jump right at the river. It is unlikely that the areas separated by just the river have wildly different micro climates, so we would expect (in the absence of the policy) that they would have highly comparable heating needs. So, if there is a discontinuous jump in pollution as the authors found, we can meaningfully attribute that to the policy. The RD design means we are comparing very similar populations in which we would expect to find minimal differences, so when we find a difference, we can interpret it as causal.

Question 2

The outcome variable in figure 2 is the PM_{10} concentration in the relevant areas. The assignment variable is the number of degrees north of the Huai River. The outcome measures how polluted the various areas are, while the assignment variable measures whether each city was affected by the policy or not. Cities where the assignment variable is greater than 0 received the treatment because they are north of the river; cites where it is less than 0 are south of the river and so did were not allowed to use indoor coal heating.

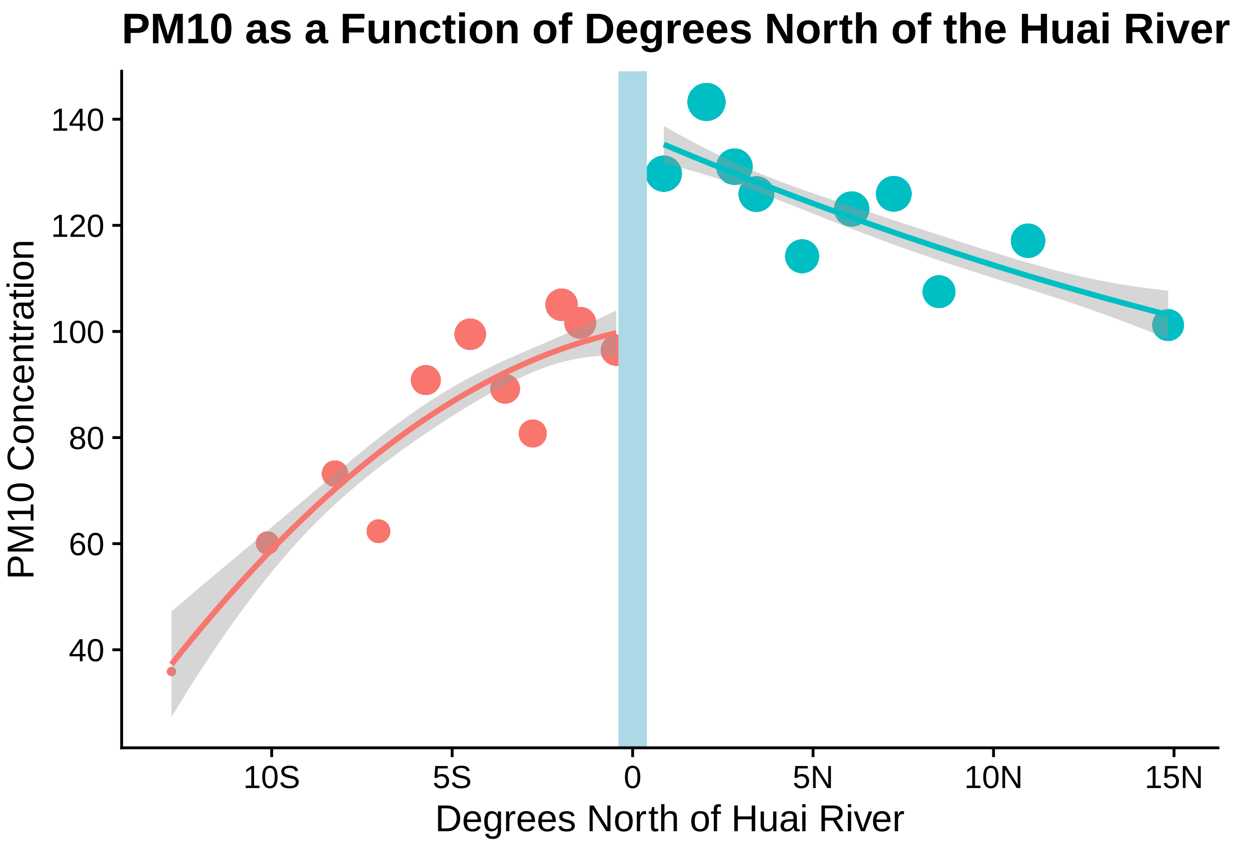
Question 3

A binned scatter plot is a scatter plot that divides the data into a certain amount of groups, takes the mean of each group and plots the means. This is in contrast to a normal scatter plot where each point on the graph represents one point of data. In the binscatter, each point is the mean of say a 10th of the data.

It is straightforward to construct a binscatter. First, you choose a number of "bins" to divide your data in to. 10-20 is usually a good number. Next, take the average of each of those 10-20 bins. Now you are left with only 10 data points. Now use your favorite plotting function to make a scatter plot of the new data set, and there you go: a binscatter.

Question 4

```
# Add bins to data

huairiver_df %<>% mutate(bin = cut(dist_huai, breaks = quantile(dist_huai, probs = seq(0,1, by = 0.05),

# Make the summarized data frame

huairiver_sum_df = huairiver_df |>
  group_by(bin) |>
  summarise(dist = mean(dist_huai, na.rm = T),
            pm10 = mean(pm10, na.rm = T),
            north_huai = north_huai)

# Now we can make the binscatter more easily with the collapsed data frame

huairiver_sum_df |>
  ggplot(aes(x = dist, y = pm10, size = pm10, color = as_factor(north_huai))) +
  geom_point() +
  geom_smooth(data = filter(huairiver_sum_df, dist <0), method = lm, formula = y~poly(x, 2)) + # Differ
  geom_smooth(data = filter(huairiver_sum_df, dist >= 0), method = lm, formula = y~poly(x, 2)) +
  geom_vline(xintercept = 0, size = 5, color = 'light blue') +
  labs(x = "Degrees North of Huai River",
       y = "PM10 Concentration",
       title = "PM10 as a Function of Degrees North of the Huai River") +
  scale_x_continuous(labels = c('15S', "10S", "5S", "0", "5N", "10N", "15N")) +
  scale_y_continuous(breaks = seq(40, 160, by = 20)) +
  cowplot::theme_cowplot() +
  theme(legend.position = 'none')
```

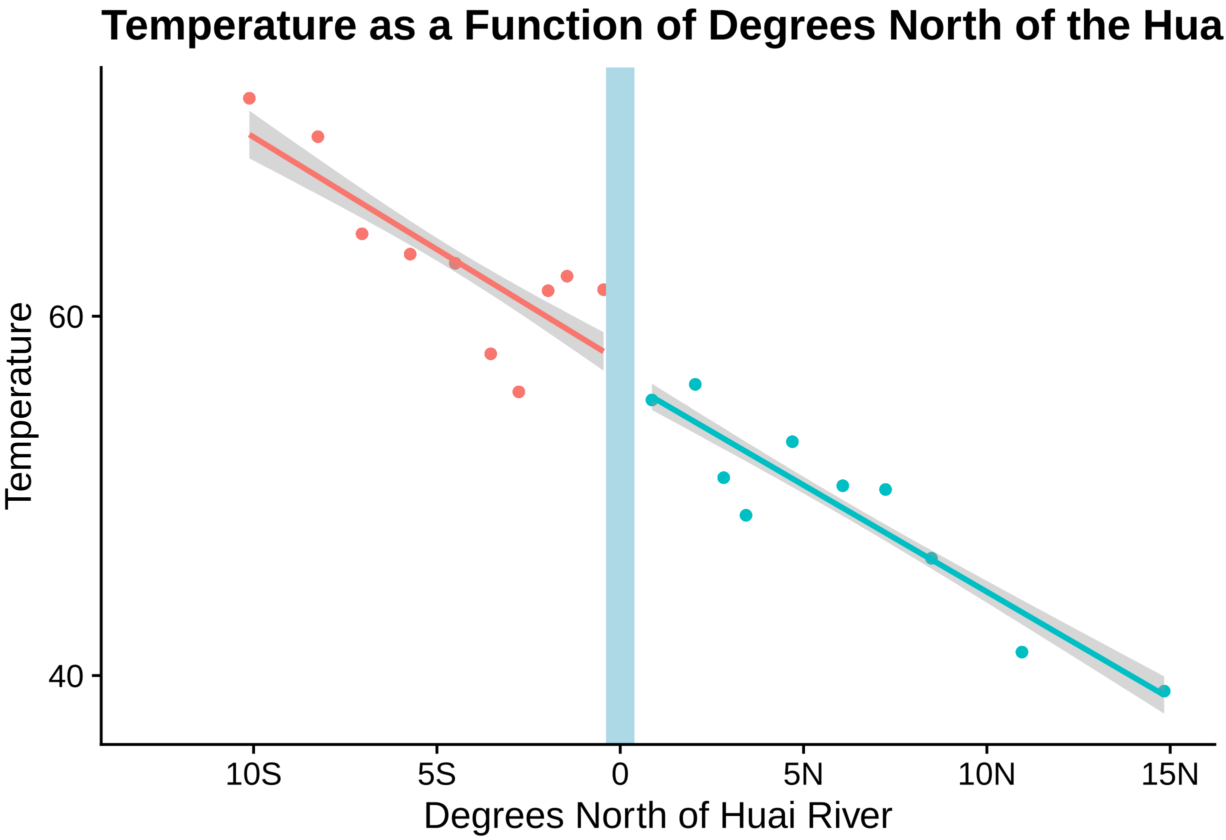


Part a

Part b

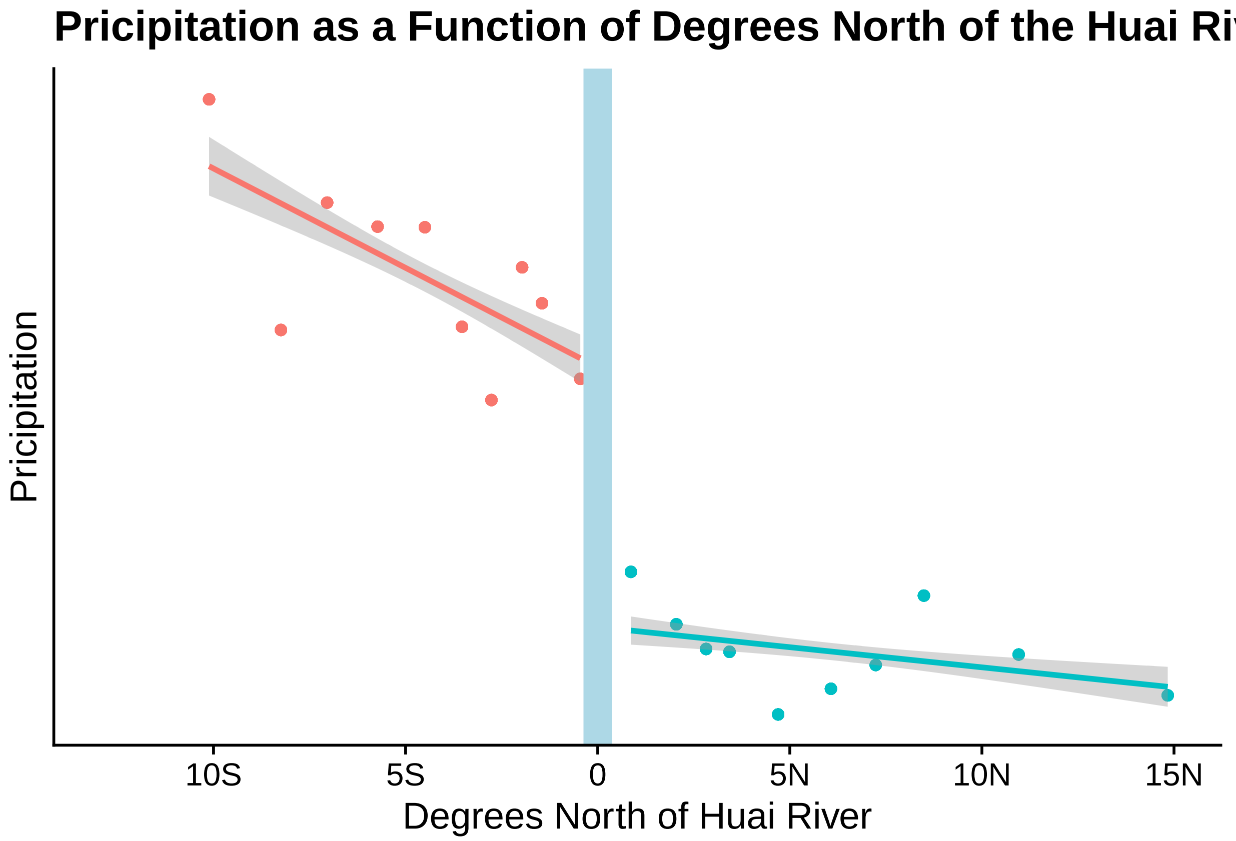
```
# Add temp, precipitation, and wind speed to dataframe

huairiver_sum_df = huairiver_df |>
  group_by(bin) |>
  summarise(dist = mean(dist_huai, na.rm = T),
            pm10 = mean(pm10, na.rm = T),
            temp = mean(temp, na.rm = T),
            pricip = mean(prcp, na.rm = T),
            wind = mean(wspd, na.rm =T),
            north_huai = north_huai)
```

```
# Now we can make the binscatter more easily with the collapsed data frame

huairiver_sum_df |>
  ggplot(aes(x = dist, y = temp, color = as_factor(north_huai))) +
  geom_point() +
  geom_smooth(data = filter(huairiver_sum_df, dist <0), method = lm, formula = y~poly(x)) + # Different
  geom_smooth(data = filter(huairiver_sum_df, dist >= 0), method = lm, formula = y~poly(x)) +
  geom_vline(xintercept = 0, size = 5, color = 'light blue') +
  labs(x = "Degrees North of Huai River",
       y = "Temperature",
       title = "Temperature as a Function of Degrees North of the Huai River") +
  scale_x_continuous(labels = c('15S', "10S", "5S", "0", "5N", "10N", "15N")) +
  scale_y_continuous(breaks = seq(40, 160, by = 20)) +
  cowplot::theme_cowplot() +
  theme(legend.position = 'none')
```



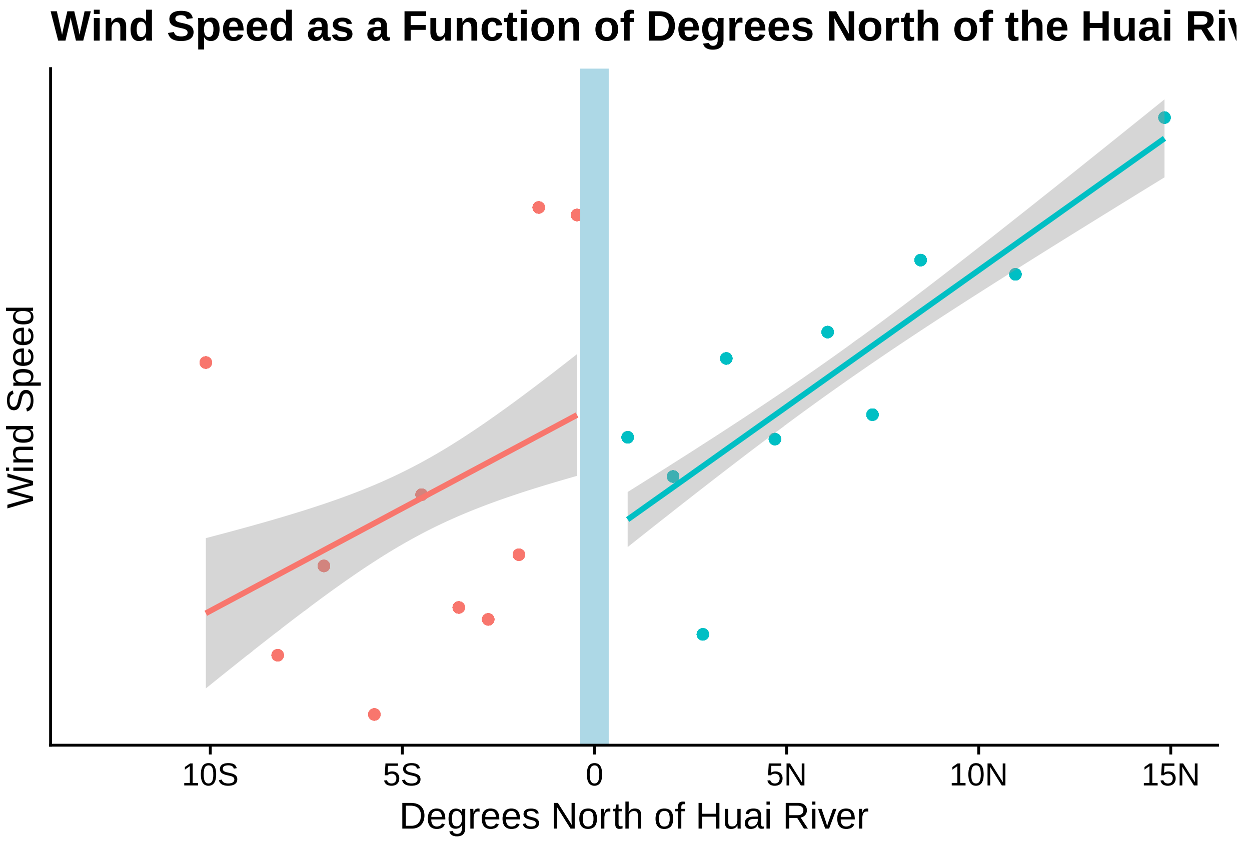
Temperature

```r
huairiver_sum_df |>
  ggplot(aes(x = dist, y = pricip, color = as_factor(north_huai))) +
  geom_point() +
  geom_smooth(data = filter(huairiver_sum_df, dist <0), method = lm, formula = y~poly(x)) + # Different
  geom_smooth(data = filter(huairiver_sum_df, dist >= 0), method = lm, formula = y~poly(x)) +
  geom_vline(xintercept = 0, size = 5, color = 'light blue') +
  labs(x = "Degrees North of Huai River",
       y = "Pricipitation",
       title = "Pricipitation as a Function of Degrees North of the Huai River") +
  scale_x_continuous(labels = c('15S', "10S", "5S", "0", "5N", "10N", "15N")) +
  scale_y_continuous(breaks = seq(40, 160, by = 20)) +
  cowplot::theme_cowplot() +
  theme(legend.position = 'none')
```



Pricipitation

```
huairiver_sum_df |>
  ggplot(aes(x = dist, y = wind, color = as_factor(north_huai))) +
  geom_point() +
  geom_smooth(data = filter(huairiver_sum_df, dist <0), method = lm, formula = y~poly(x)) + # Different
  geom_smooth(data = filter(huairiver_sum_df, dist >= 0), method = lm, formula = y~poly(x)) +
  geom_vline(xintercept = 0, size = 5, color = 'light blue') +
  labs(x = "Degrees North of Huai River",
       y = "Wind Speed",
       title = "Wind Speed as a Function of Degrees North of the Huai River") +
  scale_x_continuous(labels = c('15S', "10S", "5S", "0", "5N", "10N", "15N")) +
  scale_y_continuous(breaks = seq(40, 160, by = 20)) +
  cowplot::theme_cowplot() +
  theme(legend.position = 'none')
```



Wind Speed

Question 5

Question 6

The identification assumption for a regression discontinuity design is that the treatment, in this case the Huai River policy, is the only reason for discrete jumps in the outcome variable, in this case PM^{10} concentration around the cutoff.

The plots from 4b are mostly consistant with that, but not entirely. If the identifying assumption holds, we would expect to