

Empirical Project 2

Devin Bunch

5/6/2021

The Effect of China's Huai River Policy on Air Pollution

In this empirical project, you will use a regression discontinuity design to estimate the causal effect of Huai River Policy on air pollution.

The Stata data file huairiver.dta consists of geographic, weather, and air quality information for 161 cities in China. These data were originally used in Ebenstein et al. (2017). The life expectancy variables used in the original paper are confidential and not available. For this empirical project, we will focus on the impact of the Huai River Policy on PM10 air pollution.

```
#Load packages and libraries
library(pacman)
p_load(readr,dplyr, tidyverse, ggplot2, skimr, haven, stargazer, tidymodels, skimr, janitor, magrittr, c

#more specifically,
library(tidyverse) # for ggplot(), %>%, mutate(), and friends
library(broom) # Convert models to data frames
library(rdrobust) # For robust nonparametric regression discontinuity
library(rddensity) # For nonparametric regression discontinuity density tests
library(modelsummary) #to create side-by-side regression tables

#load data from downloads since Rconsole cannot download ".dta" files
river_data <- read_dta("~/Downloads/huairiver.dta")

#load the extra data given in the article for the specific recreation of figures
#pm10_data <- read_dta("~/Desktop/DSP_PM10.dta")

#nevermind on the extra dataset, but it is a nice touch.

#look at our data
#skim(river_data)
#skim(pm10_data)

#glimpse(river_data)
#glimpse(pm10_data)

#and the names of our variables(columns)
names(river_data)
```

```
## [1] "dsp6"      "wspd"      "temp"      "prcp"      "north_huai"
## [6] "dist_huai" "pm10"
```

```
#names(pm10_data)
```

Summary Questions

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 China would not measure the causal effect of the Huai River Policy. This is due to the fact that the new policy created immediate effects in geographical location wherein the Chinese government held high restrictions on migration. At least three limitations have plagued the existing evidence linking health to air pollution, especially at the concentrations that prevail in many of today's developing countries. First, the literature is almost entirely composed of observational studies, comparing populations across locations with varying exposure to pollution. These studies are likely to confound air pollution with unobserved determinants of health that are correlated with pollution exposure (e.g., income, hospital quality, water pollution, etc.),

Second, the available evidence is largely based on examinations of populations exposed to the modest levels of PM that are commonly observed in developed countries, where reliable pollution and health data are more readily available (6). PM concentrations in many developing countries (e.g., India and China) are 5–10 times higher than in developed countries; consequently, the existing evidence has little empirical relevance for these countries if there is a nonlinear relationship between health and pollution. Third, the most important questions about pollution center on the impacts of sustained exposure. However, there have been few opportunities to measure long-run exposure to air pollution. However, these studies have at least one of the following limitations: they (i) exploit observational variation in PM, (ii) use small samples, (iii) assume no selective migration, or (iv) focus on low levels of pollution found in the United States.] As a consequence, the existing literature focuses on shorter-run variation in PM exposure and often examines outcomes (e.g., hospitalization, infant outcomes) that are only indirectly related to longer-run outcomes, like life expectancy. This paper estimates the effect of sustained exposure to particulate matter smaller than 10 μ m (PM₁₀) on life expectancy with recent data from China and, in so doing, addresses each of the previous literature's limitations. First, the quasiexperimental research design is based on China's Huai River Policy. The policy was instituted during the 1950s when economic resources were allocated through central planning and dictated that areas to the north of the Huai River received free or highly subsidized coal for indoor heating. This led to the construction of a coal-powered centralized heating infrastructure only in cities north of the Huai River, and no equivalent system is in cities to the south; the legacy of that policy is evident even today, with very different rates of indoor heating north and south of the Huai River. Consequently, the findings are derived from a regression discontinuity (RD) design based on distance from the Huai River.

The amount of air pollution in the north versus the south would give us approximate levels of the air pollution there. It would yield two mean, aggregated numbers divided between the north and the south. However, since there are numerous confounders for air pollution, we cannot relate air pollution levels to a policy's enactment directly. The policy banned heating in the south during the winter because it was too expensive for the Chinese government, so if we were to base our causal analysis of this policy just on air pollution levels, we would probably think that the south decreased air pollution.

This is because there are bias variables X_i that we must control for.

Aggregating the mean of the north and aggregating the mean of the south, separately, and then finding the distance between would be a poor representation of

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

Figure 2 given in the Ebenstein et al. paper was constructed using a Regression Discontinuity Design from a data extract from the Chinese Disease Surveillance Points (DSP) system recorded throughout the 1990s and beyond. We are assuming that the Huai River Policy only influences

treatment where the probability of treatment rises at some threshold, but being above or below the threshold does not fully determine treatment status. With this running assumption, we have met the required Regression Discontinuity's Identification assumption. We separately estimate the following parametric equations to test for the impacts of the Huai River Policy on PM10 concentrations, and on life expectancy in China (by local linear regression using the parametric RD approach) :

$$PM_j = \alpha_0 + \alpha_1 N_j + f(L_j) + N_j f(L_j) + X_j + u_j$$

$$Y_j = \beta_0 + \beta_1 N_j + f(L_j) + N_j f(L_j) + X_j + E_j$$

where

j references a DSP location in China

PM_j is the average annual ambient concentration of PM10 in location j over the period 2004–2012

Y_j is a measure of location j 's mortality rate (or) Y_j is a yearly measure of life expectancy at birth
 N_j is an indicator variable equal to one for locations that are north of the Huai River line
 $f(L_j)$ is a polynomial in degrees north of the Huai River that is interacted with N_j (chosen based on goodness of fit criteria) such that L_j is within h latitude degrees of the Huai River.

X_j is a vector of the demographic and city characteristics other than air quality that are associated with mortality rates

The outcome variable of interest is life expectancy and pm10 concentration levels. The assignment variable is northern location in relation to the Huai River, grouped into 154 DSL locations.

3. What is a binned scatter plot? Explain how it is constructed.

A binned scatterplot displays the relationship between the x axis variable and the y axis outcome variable by grouping along the x axis. It's where the data points are grouped into bins, and an aggregate statistic is used to summarize each bin. A binned scatter plot is a more scalable alternative to the standard scatter plot. The data points are grouped into bins, and an aggregate statistic is used to summarize each bin. When working on a large dataset, scatterplots might become difficult to read, especially if you are looking for concentrations (densities) to judge the relationship and you have many observations that have the same or very similar values (overplotted points). A standard plot, displays a marker for every data point. To produce a binned plot, imagine a grid be placed on the scatterplot; then count the number of data points in each grid cell and display a marker with a size (or a colour intensity) that reflects the number of points in each grid cell.

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.

Replicating Figure 2 of the paper, we depict how exposure to PM10 increases significantly at the Huai River, but pollution exists south and north of the river, making our context naturally analogous to a fuzzy RD. A fuzzy RD is where the ratio is estimated as the ratio of two “sharp” discontinuities; in practice, we use the optimal bandwidth for life expectancy as the bandwidth for both life expectancy and PM10 concentration level.

```
#Goal: we want to create bins in our data set, where we go from the minimum distance, to the max, by 1
#let's make bins in our data frame and then save it as another variable to avoid confusion
#first, let us see what the minimum and maximum x=dist_huai values are in order to choose the number ra
summary(river_data)
```

dsp6	wspd	temp	prcp
Min. :110101	Min. :0.5287	Min. :32.92	Min. :0.04486
1st Qu.:230826	1st Qu.:1.0027	1st Qu.:48.56	1st Qu.:0.46903
Median :410102	Median :1.5091	Median :57.48	Median :0.65719
Mean :384234	Mean :1.5577	Mean :56.09	Mean :0.71555
3rd Qu.:511025	3rd Qu.:1.9861	3rd Qu.:62.64	3rd Qu.:0.95082
Max. :654025	Max. :3.2519	Max. :76.38	Max. :1.53767
NA's :5	NA's :8	NA's :8	NA's :8

north_huai dist_huai pm10

north_huai	dist_huai	pm10
Min. :0.0000	Min. :-12.77608	Min. : 27.28
1st Qu.:0.0000	1st Qu.: -4.10249	1st Qu.: 79.25
Median :0.0000	Median : -0.04709	Median :101.06
Mean :0.4907	Mean : 0.69795	Mean :103.09
3rd Qu.:1.0000	3rd Qu.: 5.29008	3rd Qu.:124.77
Max. :1.0000	Max. : 16.47519	Max. :307.31
NA's :7		

```
#now that we see our x variable limits, we have to round our number range to include each observation,
river_data <- river_data %>% mutate(bin_dist = cut(dist_huai, breaks=seq(-13,17, by = 1), na.rm = TRUE))

#let's check to make sure our new variable we just made is a factor
is.factor(river_data$bin_dist)
```

```
[1] TRUE
```

```
#let's look at the values we've got
table(river_data$bin_dist)
```

```
(-13,-12] (-12,-11] (-11,-10] (-10,-9] (-9,-8] (-8,-7] (-7,-6] (-6,-5] 2 1 1 5 6 7 5 8 (-5,-4] (-4,-3] (-3,-2] (-2,-1] (-1,0]
(0,1] (1,2] (2,3] 6 10 11 11 9 3 8 10 (3,4] (4,5] (5,6] (6,7] (7,8] (8,9] (9,10] (10,11] 12 2 8 5 8 6 3 3 (11,12]
(12,13] (13,14] (14,15] (15,16] (16,17] 1 3 1 3 1 2
```

```
#Check, looks like we have our bins set perfect with one output for each bin, and it goes from -13 to 1
```

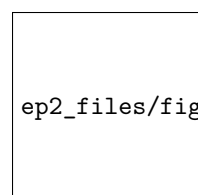
```
#now let us group bins together, and take the mean of each bin, and for each variable
river_bin <- river_data %>% group_by(river_data$bin_dist) %>% summarise(dist_huai = mean(dist_huai), pm10 = mean(pm10))
```

```
#now we have binned our data apart by 1 degree latitude, displaying the average output for each singular bin
```

```
#now we want to plot the average concentration of PM10 against the average distance away one is from the river
```

```
river_bin %>% ggplot(aes(x=dist_huai, y = pm10)) + geom_point(color = "pink3", size = 2, alpha = 1) +
  geom_smooth(method="lm", se = FALSE)
```

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



```
#Great, now let us build off this graph and make it look more similar to figure 2 of the Ebenstein et al
#The graph prints, but let's add {r, results="markup"} to keep working on the plot
```

```
#ok, let's make a ggplot and store them in variables to try to replicate Figure 2 in the paper
```

```
g <- ggplot(data = river_data, aes(x = dist_huai, y = pm10, color = pm10)) + labs(x= "Degrees North of
                                                                    Fitted values from a
                                                                    from the Huai River c
```

```
#print out our base graph
#g
```

```
#make it look nicer, add to base graph
#g1 <- g + geom_point(color = "orange", aplha = 0.5) + scale_color_manual(values= c(""))
g1 <- g + geom_smooth(data = river_data %>% filter(north_huai==0), se = FALSE, method = 'lm', color = "
```

```
#keep going
```

```
g2 <- g1 + geom_smooth(data = river_data %>% filter(north_huai==1), se = FALSE, method = 'lm', color = "
```

```
gg2 <- g2 + geom_point(data = river_data %>% filter(north_huai==1), color = "forestgreen", size =2)
```

```
#add the vertical line at distance zero
```

```
g3 <- gg2 + geom_vline(data = river_data, xintercept = 0, aplha = 0.7)
```

```
#some fails
```

```
##+ scale_color_manual(aes(x = "North", color = "green"))
```

```
##+ geom_point(aes(color="slategrey"))
```

```
##+ stat_summary_bin(fun='mean', bins=10, color='black', geom='point') + theme(legend.position = "bottom
```

```
#good layout
```

```
g4 <- g3 + geom_point(data = river_data %>% filter(north_huai==0), color = "dodgerblue", size =2)
```

```
g4
```

ep2_files/figure-latex/trial one-1.pdf

>**Graphical Interpretation** First off, graphically, we can see that we do not have a sharp regression discontinuity design—we have a fuzzy RD design.

We begin the analysis graphically with an assessment of the Huai River Policy's impact on pollution. Fig. 2 plots PM10 at DSP locations against their degrees north of the Huai River line. The circles in Fig. 2 represent the average PM10 concentration across locations within 1°-latitude distance bins from the Huai River; each circle's size is proportional to the population at the DSP locations within the relevant bin. The plotted line in Fig. 2 is generated by using a kernel-weighted local linear regression on either side of the river, which is similar to a nonparametric RD approach. This is estimated with a triangular kernel and bandwidth chosen by the method prescribed by Imbens and Kalyanaraman (14). Fig. 2 reveals a discontinuous change in ambient PM10 concentrations to the north of the river; it indicates that the Huai River Policy increased PM10 concentrations by about 42 g/m³. The plot in Fig. 3 is almost a mirror image of Fig. 2.

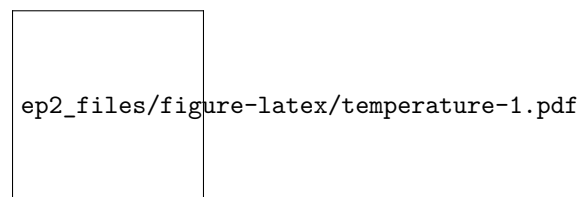
b. Draw binned scatter plots to test whether (i) temperature, (ii) precipitation, and (iii) wind speed changes at the Huai River line. Display fitted lines (linear, or quadratic, or whatever

functional form you see fit) based on what you see in the data.

we separately test whether the Huai River Policy caused discontinuous changes in PM10 and life expectancy to the north of the river. how to determine the bandwidth .. derik got bandwidth output of 5.6 for cct, 9.29 for ik. using rdg bandwidthth function...by the constant treatment effect assumption, “this” regression is linearly approximated, and parallel.

Binscatter is thus known as a “non-parametric” way of getting $E[Y|X]$. Also allows us to assess the functional form assumption: Let the data tell you if you should fit a linear model or something else.

```
#regression on temperature
t1 <- ggplot(data = river_data, aes(x = dist_huai, y = temp)) + geom_smooth(data = river_data %>% filter(dist_huai > 0), aes(x = dist_huai, y = temp))
#graphically,
t1
```



```
#looks like there's no obvious discontinuity
```

```
#regression on Precipitation
p1 <- ggplot(data = river_data, aes(x = dist_huai, y = prcp)) + geom_smooth(data = river_data %>% filter(dist_huai > 0), aes(x = dist_huai, y = prcp))
```

```
## Warning: geom_vline(): Ignoring 'data' because 'xintercept' was provided.
```

```
## Warning: Ignoring unknown parameters: aplha
```

```
#graphically,
p1
```

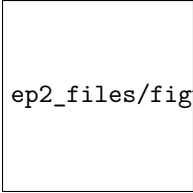
```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



ep2_files/figure-latex/unnamed-chunk-4-1.pdf

#looks like no promising correlation nor discontinuity

#regression on wind speeds

```
p1 <- ggplot(data = river_data, aes(x = dist_huai, y = wspd)) + geom_smooth(data = river_data %>% filter(dist_huai > 0))
```

Fitted values from a
from the Huai River

```
## Warning: geom_vline(): Ignoring 'data' because 'xintercept' was provided.
```

```
## Warning: Ignoring unknown parameters: aplha
```

#graphically,
p1

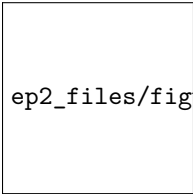
```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 4 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



ep2_files/figure-latex/unnamed-chunk-5-1.pdf

#looks like we could possibly have a fuzzy regression discontinuity

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. (more written about this in the assignment)

rdrobust—this command now allows for covariate-adjusted point estimation and covariate-adjusted robust bias-corrected inference. In addition, this command now allows for different heteroskedasticity-robust (heteroskedasticity-consistent k class or HCk class) and cluster-robust variance estimation methods. When mean squared error (MSE)–optimal bandwidths are used, the resulting point estimator for the RD treatment effect is MSE optimal. When coverage error-rate (CER) optimal bandwidths are used, the resulting confidence intervals (CIs) for the RD treatment effect are CER-optimal.

The Regression in Regression Discontinuity

$$= \lim_{x \rightarrow c^+} E[Y | x] - \lim_{x \rightarrow c^-} E[Y | x]$$

Our goal is to estimate the diff in outcomes for just-treated vs. just-untreated.

We implement this by comparing average outcomes just below and above the cutoff

$= 1; + = 0;$

Where $+$ is the bandwidth which we're "close" to c . the variable name is h .

This leads to the regression-based RD:

$Y_{c \pm h}$ for $+$ where $= 1$

This means were only looking at observations that are very close to our cutoff of $c = 0$.

```
#load new package needed for RD regressions, although it is outdated.
rdbwselect_2014(y = river_bin$pm10, x = river_bin$dist_huai, bwselect = "IK")
```

```
## Call:
## rdbwselect_2014(y = river_bin$pm10, x = river_bin$dist_huai,
##   bwselect = "IK")
##
## BW Selector    IK
## Number of Obs 24
## NN Matches    3
## Kernel Type   Triangular
##
##               Left Right
## Number of Obs    11   13
## Order Loc Poly (p) 1    1
## Order Bias (q)    2    2
##
##               h      b
## [1,] 7.079376 10.49187
```

#From our simplistic old method,

```
#No controls:
names(river_bin)
```

```
## [1] "river_data$bin_dist" "dist_huai"          "pm10"
```

```
#for temperature,
temp_coefficient = lm(temp~dist_huai, data = river_data)
temp_coefficient
```

```
##
## Call:
## lm(formula = temp ~ dist_huai, data = river_data)
##
## Coefficients:
## (Intercept)    dist_huai
##      57.260      -1.275
```



```
#for precipitation,
rain_coefficient = lm(prcp~dist_huai, data = river_data)
rain_coefficient
```

```
##
## Call:
## lm(formula = prcp ~ dist_huai, data = river_data)
##
## Coefficients:
## (Intercept)    dist_huai
##      0.74829      -0.03582
```

```
#for windspeed,
wind_coefficient = lm(wspd~dist_huai, data = river_data)
wind_coefficient
```

```
##
## Call:
## lm(formula = wspd ~ dist_huai, data = river_data)
##
## Coefficients:
## (Intercept)    dist_huai
##      1.53381      0.02769
```

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.

answer goes here I have taken notes on this in class.

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.

For a manipulation test, we are assuming that bandwidth(how far apart an observation is from the cutoff) is as good as randomly assigned in the neighborhood of $c = 0$.

By looking at the distribution of X_i , we gather a “TRUE” or “FALSE” result from our Manipulation test. If the result comes back “TRUE” (i.e., we can assume that $X_i - c$ replicates randomization in the bandwidth of c . If we assume this is true, then units of observation cannot be resorted in the bandwidth around $c = 0$.

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?

answer goes here

the purpose of the regression using the false location was to prove the exact cutoff threshold was sound and immobile and strict and in direct response of the huai river policy.

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.

```

#use mutate to make that plus 1 go through the whole column of distance

#week 1 lecture notes
# Estimate the linear model
#lm_bin_1 <-lm(int_rt~ 1 + fico_bin, data = loan_data)# Summarize the results of the linear regression

#model for -5
minus5_df <- river_data %>%
  mutate(north_minus5 = ifelse(dist_huai - 5 > 0, 1, 0), dist_huai_minus5 := dist_huai - 5)

model_minus5 <- lm(pm10 ~ dist_huai + north_minus5 + dist_huai:north_minus5, minus5_df %>%
  filter(between(dist_huai_minus5, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_minus5") %>% mutate(x = -5)
model_minus5

```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	x
north_minus5	-14.9	39.6	-0.375	0.709	-94.3	64.5	-5

```
minus5_df
```

```

#model for -4
minus4_df <- river_data %>%
  mutate(north_minus4 = ifelse(dist_huai - 4 > 0, 1, 0), dist_huai_minus4 := dist_huai - 4)

model_minus4 <- lm(pm10 ~ dist_huai + north_minus4 + dist_huai:north_minus4, minus4_df %>%
  filter(between(dist_huai_minus4, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_minus4") %>% mutate(x = -4)

#model for -3
minus3_df <- river_data %>%
  mutate(north_minus3 = ifelse(dist_huai - 3 > 0, 1, 0), dist_huai_minus3 := dist_huai - 3)

model_minus3 <- lm(pm10 ~ dist_huai + north_minus3 + dist_huai:north_minus3, minus3_df %>%
  filter(between(dist_huai_minus3, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_minus3") %>% mutate(x = -3)

#model for -2
minus2_df <- river_data %>%
  mutate(north_minus2 = ifelse(dist_huai - 2 > 0, 1, 0), dist_huai_minus2 := dist_huai - 2)

model_minus2 <- lm(pm10 ~ dist_huai + north_minus2 + dist_huai:north_minus2, minus2_df %>%
  filter(between(dist_huai_minus2, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_minus2") %>% mutate(x = -2)

#model for -1

```

```

minus1_df <- river_data %>%
  mutate(north_minus1 = ifelse(dist_huai - 1 > 0, 1, 0), dist_huai_minus1 := dist_huai - 1)

model_minus1 <- lm(pm10 ~ dist_huai + north_minus1 + dist_huai:north_minus1, minus1_df %>%
  filter(between(dist_huai_minus1, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_minus1") %>% mutate(x = -1)

#model for 0
minus0_df <- river_data %>%
  mutate(north_minus0 = ifelse(dist_huai - 0 > 0, 1, 0), dist_huai_minus0 := dist_huai - 0)

model_minus0 <- lm(pm10 ~ dist_huai + north_minus0 + dist_huai:north_minus0, minus0_df %>%
  filter(between(dist_huai_minus0, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_minus0") %>% mutate(x = 0)

#model for +1
plus1_df <- river_data %>%
  mutate(north_plus1 = ifelse(dist_huai + 1 > 0, 1, 0), dist_huai_plus1 := dist_huai + 1)

model_plus1 <- lm(pm10 ~ dist_huai + north_plus1 + dist_huai:north_plus1, plus1_df %>%
  filter(between(dist_huai_plus1, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_plus1") %>% mutate(x = 1)

#model for +2
plus2_df <- river_data %>%
  mutate(north_plus2 = ifelse(dist_huai + 2 > 0, 1, 0), dist_huai_plus2 := dist_huai + 2)

model_plus2 <- lm(pm10 ~ dist_huai + north_plus2 + dist_huai:north_plus2, plus2_df %>%
  filter(between(dist_huai_plus2, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_plus2") %>% mutate(x = 2)

#model for +3
plus3_df <- river_data %>%
  mutate(north_plus3 = ifelse(dist_huai + 3 > 0, 1, 0), dist_huai_plus3 := dist_huai + 3)

model_plus3 <- lm(pm10 ~ dist_huai + north_plus3 + dist_huai:north_plus3, plus3_df %>%
  filter(between(dist_huai_plus3, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_plus3") %>% mutate(x = 3)

#model for +4
plus4_df <- river_data %>%
  mutate(north_plus4 = ifelse(dist_huai + 4 > 0, 1, 0), dist_huai_plus4 := dist_huai + 4)

model_plus4 <- lm(pm10 ~ dist_huai + north_plus4 + dist_huai:north_plus4, plus4_df %>%
  filter(between(dist_huai_plus4, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_plus4") %>% mutate(x = 4)

```

```

#model for +5
plus5_df <- river_data %>%
  mutate(north_plus5 = ifelse(dist_huai + 5 > 0, 1, 0), dist_huai_plus5 := dist_huai + 5)

model_plus5 <- lm(pm10 ~ dist_huai + north_plus5 + dist_huai:north_plus5, plus5_df %>%
  filter(between(dist_huai_plus5, -5, 5))) %>%
  broom::tidy(conf.int = TRUE) %>% filter(term == "north_plus5") %>% mutate(x = 5)

#Now let's create one data frame with all of our models we just made
models <- bind_rows(model_minus5, model_minus4, model_minus3, model_minus2, model_minus1, model_minus0,
  model_plus1, model_plus2, model_plus3, model_plus4, model_plus5)

#print it out
models

#Great--we have each x value, and have variables for each cutoff of the confidence intervals for the y
ggplot(data = models, aes(x = x, y = estimate)) +
  geom_point(size = 2) +
  geom_point(aes(y= conf.high), shape = 95, size = 2) +
  geom_point(aes(y = conf.low), shape = 95, size = 2) +
  geom_linerange(aes(ymin = conf.low, ymax = conf.high), alpha = 0.85, color = "lightpink3") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "pink3") +
  scale_x_continuous(breaks = seq(-5, 5, 1)) +
  scale_y_continuous(breaks = seq(-100, 100, 25)) +
  labs(title = "Hi", x = "Distance from the Huai River",
  caption = "RD estimates of the change in PM10 and life expectancy at the Huai River and discontinuities")

```

ep2_files/figure-latex/unnamed-chunk-9-1.pdf

Footnote for the figure: RD estimates of the change in PM10 and life expectancy at the Huai River and discontinuities estimated at 1°-latitude displacements from the actual Huai River.

conclusion

The analysis indicates that PM10 exposure causes people to live substantially shorter and sicker lives at the concentrations present today in China and other developing countries.

the elevated mortality rates are concentrated among cardiorespiratory causes of death, whereas there is little evidence of a difference in mortality rates for causes that are not plausibly related to air pollution.

Table 1: Data for the first 100 rows (rows 1-100)									
dsp6	wspd	temp	prcp	north_huai	dist_huai	pm10	bin_dist	north_minus5	dist_huai_minus5
1.1e+05	2.09	53.1	0.534	1	7.06	143	(7,8]	1	2.0
1.1e+05	2.09	53.1	0.534	1	7.01	143	(7,8]	1	2.0
1.2e+05	2.29	54.7	0.526	1	6.2	114	(6,7]	1	1.2
1.2e+05	2.29	54.7	0.526	1	7.05	120	(7,8]	1	2.0
1.3e+05	1.59	52.1	0.657	1	6.63	107	(6,7]	1	1.6
1.3e+05	1.59	52.1	0.657	1	7.1	94.8	(7,8]	1	2.1
1.3e+05	1.25	51	0.667	1	6.93	73.6	(6,7]	1	1.9
1.3e+05	1.14	57.4	0.469	1	3.92	139	(3,4]	0	-1.0
1.3e+05	1.14	57.4	0.469	1	4.24	119	(4,5]	0	-0.7
1.31e+05	2.07	47.7	0.282	1	8.15	65	(8,9]	1	3.1
1.31e+05	2.07	47.7	0.282	1	7.92	138	(7,8]	1	2.9
1.31e+05	1.72	44.3	0.403	1	8.27	140	(8,9]	1	3.2
1.4e+05	1.36	50.1	0.42	1	5.4	150	(5,6]	1	0.4
1.4e+05	1.52	48.6	0.527	1	5.29	135	(5,6]	1	0.2
1.4e+05	1.18	47.5	0.481	1	3.61	143	(3,4]	0	-1.3
1.41e+05	1.36	50.1	0.42	1	6.83	148	(6,7]	1	1.8
1.41e+05				1	2.95	143	(2,3]	0	-2.0
1.41e+05	0.983	50.2	0.458	1	5.33	139	(5,6]	1	0.3
1.5e+05	1.49	44	0.333	1	8.2	95.4	(8,9]	1	3.2
1.5e+05	1.69	42	0.337	1	10.5	125	(10,11]	1	5.4
1.51e+05	2	44.3	0.329	1	10.8	151	(10,11]	1	5.7
1.51e+05	1.67	44.2	0.198	1	7.53	180	(7,8]	1	2.5
1.53e+05	2.12	34.9	0.254	1	10.3		(10,11]	1	5.2
2.1e+05	2.08	45.7	0.514	1	9.77	126	(9,10]	1	4.7
2.1e+05	3.07	51.1	0.717	1	6.29	82.3	(6,7]	1	1.2
2.1e+05	1.95	47.9	0.746	1	8.7	123	(8,9]	1	3.7
2.11e+05	0.923	47.2	0.989	1	8.52	90.6	(8,9]	1	3.5
2.11e+05	2.73	45.6	0.561	1	9.29	111	(9,10]	1	4.2
2.11e+05	1.95	47.9	0.746	1	8.87	97	(8,9]	1	3.8
2.2e+05	2.59	41.3	0.523	1	12	94.7	(11,12]	1	6.9
2.2e+05	2.59	41.3	0.523	1	12.73	106	(12,13]	1	7.6
2.2e+05	1.43	40	0.563	1	12.4	127	(12,13]	1	7.3
2.31e+05	0.628	42	0.704	1	9.74	103	(9,10]	1	3.9

term	estimate	std.error	statistic	p.value	conf.low	conf.high	x
north_minus5	-14.9	39.6	-0.375	0.709	-94.3	64.5	-5
north_minus4	7.72	38.3	0.202	0.841	-68.8	84.2	-4
north_minus3	1.11	21.6	0.0512	0.959	-42	44.2	-3
north_minus2	19.7	17.7	1.11	0.271	-15.7	55	-2
north_minus1	28.2	17.3	1.63	0.107	-6.18	62.5	-1
north_minus0	55.1	17.1	3.23	0.00183	21.2	89.1	0
north_plus1	9.11	14.3	0.637	0.526	-19.3	37.5	1
north_plus2	14.4	16.6	0.867	0.389	-18.6	47.4	2
north_plus3	1.92	21.4	0.0895	0.929	-40.8	44.6	3
north_plus4	-14.5	21.9	-0.66	0.511	-58.2	29.3	4
north_plus5	-10.3	23.6	-0.436	0.664	-57.3	36.8	5