PSTAT 126 Final Assignment

Preeti Kulkarni

Background

By now it is widely recognized that air quality impacts health, but this was not always the case. The file `pollution.csv` contains data from an early observational study investigating the relationship between specific pollutants and mortality in U.S. cities. Variable descriptions and units are recorded in the metadata file `pollution-metadata.csv`. All measurements were taken for the period 1959 - 1961.

> McDonald, G.C. and Schwing, R.C. (1973). Instabilities of Regression Estimates Relating Air Pollution to Mortality. Technometrics, 15: 463-481.

```
## # A tibble: 3 x 7
##   City           Mort Precip  Educ NonWhite   NOX   SO2
##   <chr>         <dbl>  <dbl> <dbl>    <dbl> <dbl> <dbl>
## 1 San Jose, CA   791.     13  12.2      3      32     3
## 2 Wichita, KS    824.     28  12.1      7.5     2     1
## 3 San Diego, CA  840.     10  12.1      5.9    66    20
```

```
## # A tibble: 60 x 7
##    City              Mort Precip  Educ NonWhite   NOX   SO2
##    <chr>            <dbl>  <dbl> <dbl>    <dbl> <dbl> <dbl>
##  1 San Jose, CA      791.     13  12.2      3      32     3
##  2 Wichita, KS       824.     28  12.1      7.5     2     1
##  3 San Diego, CA     840.     10  12.1      5.9    66    20
##  4 Lancaster, PA     844.     43   9.5      2.9     7    32
##  5 Minneapolis, MN   858.     25  12.1      2      11    26
##  6 Dallas, TX        860.     35  11.8     14.8     1     1
##  7 Miami, FL         861.     60  11.5     13.5     1     1
##  8 Los Angeles, CA   862.     11  12.1      7.8   319   130
##  9 Grand Rapids, MI  871.     31  10.9      5.1     3    10
## 10 Denver, CO        872.     15  12.2      4.7     8    28
## # ... with 50 more rows
```

In this data the presence of pollutants is reported as *relative pollution potential*, which is calculated by scaling emissions (tons per day per square kilometer) by a dispersion factor based on local conditions (mixing, wind, area, and the like).

Questions

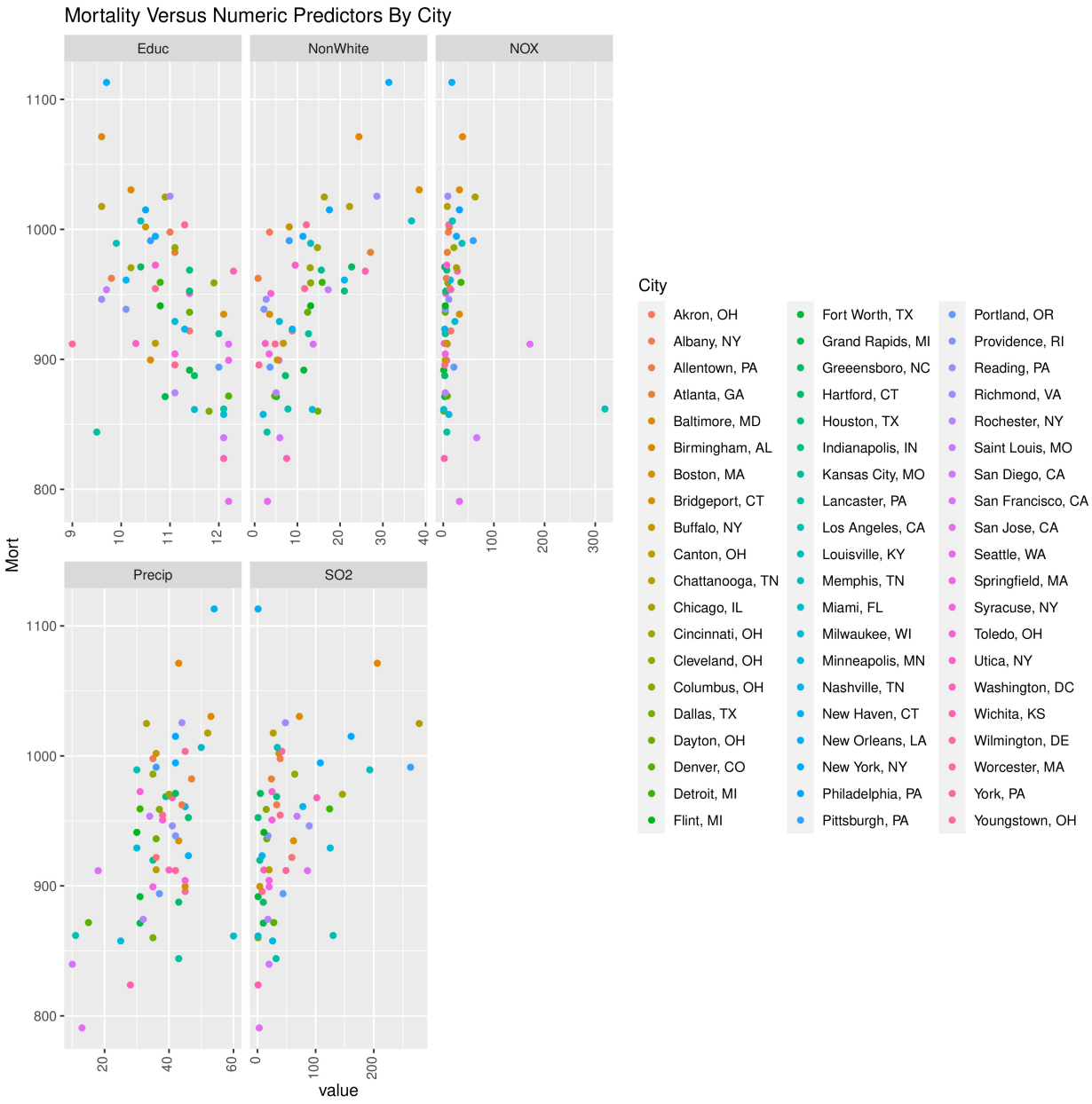
Respond to each question or task immediately below the prompt in a concise manner – aim to give as direct a response as possible. Following this, provide, if appropriate, any supporting information helpful in understanding your answer; please limit such supporting information to a brief paragraph and minimal R output (possibly one table, a few simple calculations, or a plot).

Please include all codes used together with your answer in the .Rmd file (so that they appear in the appendix), but control the code chunks so that *only codes and output that are referenced in your written answers are shown.*

1. Construct a plot of the marginal relationships among the raw data and comment briefly on the plot (identify any notable features).

#create mortality v each predictor for each of the 3 graphs

#Marginal relationship between raw data



Graph1: As education level increases, mortality rate decreases, forming a negative trend. Graph2: As the percent of Nonwhite individuals increase, mortality also increases, forming a positive trend.

Graph3: NOX points are clustered between 0-100 tons per day per square km with some cities having values slightly outside of the trend. We can see that as NOX increases, the mortality rate increases, other than a few outliers, forming a positive trend.

Graph4: As precipitation increases, mortality increases, forming a positive trend.

Graph5: Finally, for SO2, the points are mostly clustered between 0-100 tons per day per square km. As SO2 increases, mortality increases, then plateus when it reaches around 150 tons per day per square km, forming a positive trend.

Also, there is a notably high mortality rate in New Orleans in all graphs despite having a low NOX and SO2 values.

2. Estimate the association between mortality and each of the two pollutants. Describe how you obtained your estimates and be sure to give proper interpretations.

```
##
## Call:
## lm(formula = Mort ~ SO2, data = pollution)
##
## Coefficients:
## (Intercept)          SO2
##     917.884        0.418
```

```
##
## Call:
## lm(formula = Mort ~ NOX, data = newpollution)
##
## Coefficients:
## (Intercept)          NOX
##     925.249        1.132
```

By fitting both NOX and SO2, and removing the outliers from NOX, we are able to see the mean deaths which increase 1.132 and 0.418 per 100,000 people per year respectively. We can see that deaths increase 1.132 units per one hundred thousand people as NOX increases. Similarly, mean deaths increase about 0.418 per 100,000 people per year for SO2.

3. How many lives could be saved each year by curbing emissions? Answer each of the questions below.

    i. Estimate the reduction in mortality rate associated with a 50% relative decrease in sulfur dioxide emissions.

```
##
## Call:
## lm(formula = Mort ~ . - City - SO2 + halfSO2, data = pollution)
##
## Coefficients:
## (Intercept)       Precip         Educ     NonWhite          NOX      halfSO2
##   1000.1026       1.3792     -15.0791       3.1602      -0.1076       0.1777
```

```
##              2.5 %    97.5 %
## halfSO2 0.08609008 0.2693269
```

```
##         2.5 %    97.5 %
## SO2 0.1721802 0.5386538
```

Without reducing 50% of emissions of SO2, we see the confidence interval of 95% is 0.1721802- 0.5386538. After curbing emissions, we see the confidence interval be 0.08609008-0.2693269. Thus, we see that with a 50% relative decrease in SO2, there will also be a 50% decrease in mortality. From 2, we found the original intercept which is 917.884, which we know decreases by 50%, so around 459 lives are saved per 100,000 people per year.

    ii. Estimate the reduction in mortality rate associated with a 50% relative decrease in emissions of oxides of nitrogen.

```
##
## Call:
## lm(formula = Mort ~ . - City - NOX + halfNOX, data = pollution)
##
## Coefficients:
## (Intercept)       Precip         Educ     NonWhite          SO2      halfNOX
##  1000.10264      1.37921    -15.07905      3.16023      0.35542     -0.05379
```

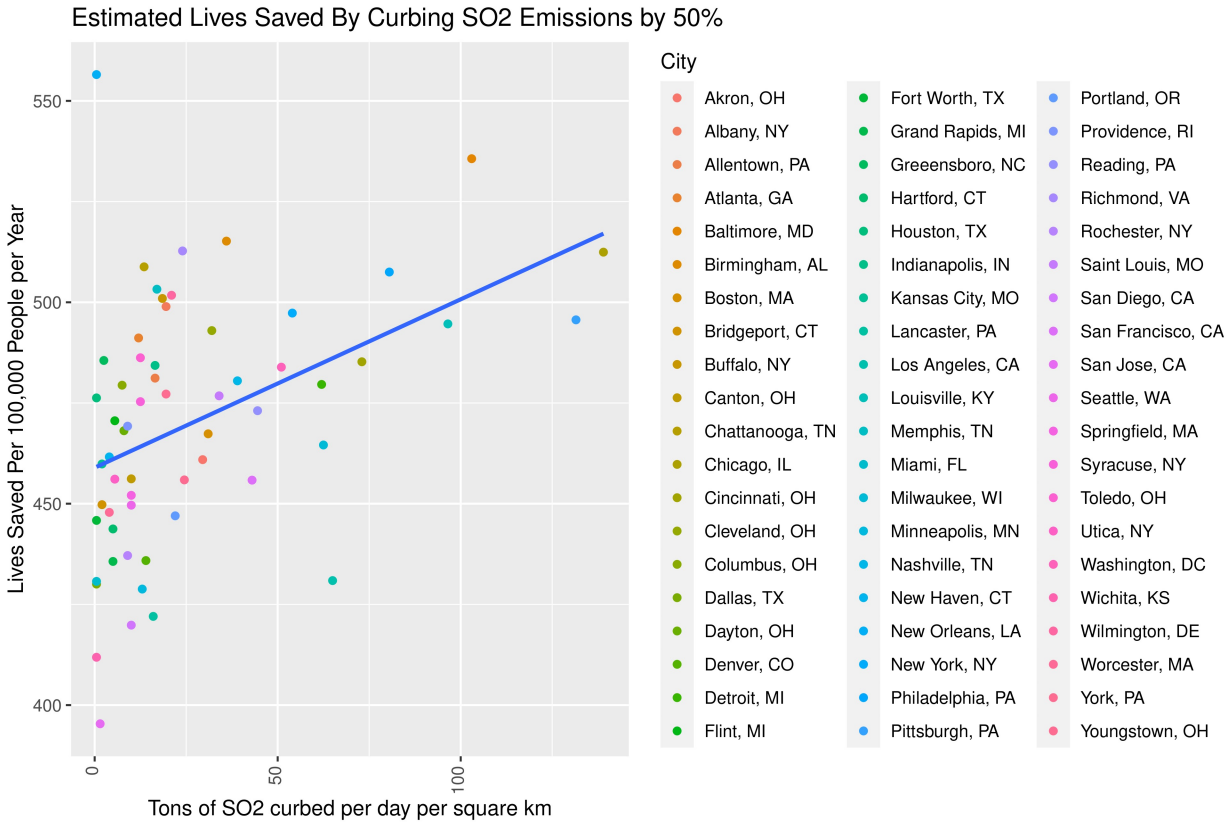
```
##         2.5 %    97.5 %
## NOX -0.380008 0.1648609
```

```
##             2.5 %     97.5 %
## halfNOX -0.190004 0.08243045
```

The reduction in mortality rate associated with a 50% relative decrease in SO2 is around, 50%, because we the original confidence interval of 95% go from -0.380008 0.1648609 to -0.190004 0.08243045,k which we can see is half of the original. By using the original intercept, 925.249, we can halve it to find around 463 lives are saved per 100,000 people per year.

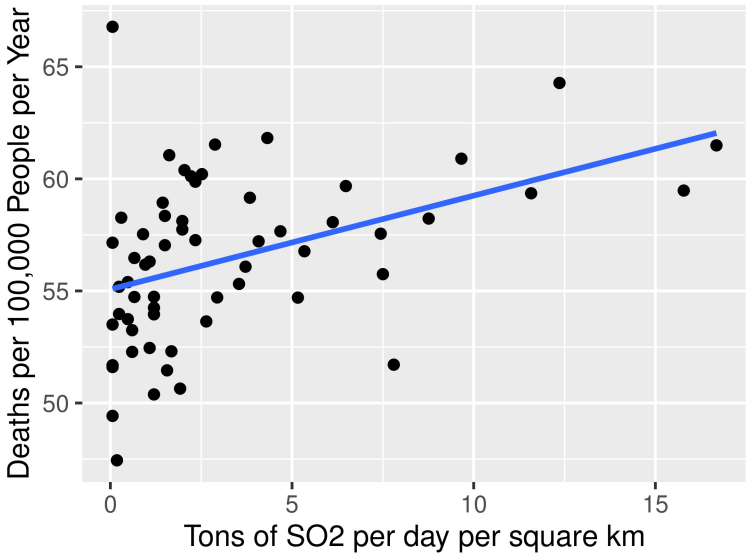
    iii. Construct a visualization that conveys the estimated potential lives saved by reducing SO2 emissions. Provide a brief description of your plot.

The plots shows the number of lives saved by reducing SO2 emissions by 50% per city. The line of best fit shows a positive trend between the amount of deaths increasing as SO2 increases. The x axis is the SO2 emissions curbed, while the y axis represents the lives saved by halving the emissions.


Estimated Lives Saved By Curbing SO2 Emissions by 50%
Lives Saved Per 100,000 People per Year
550
500
450
400
0
50
100
Tons of SO2 curbed per day per square km
City
Akron, OH
Albany, NY
Allentown, PA
Atlanta, GA
Baltimore, MD
Birmingham, AL
Boston, MA
Bridgeport, CT
Buffalo, NY
Canton, OH
Chattanooga, TN
Chicago, IL
Cincinnati, OH
Cleveland, OH
Columbus, OH
Dallas, TX
Dayton, OH
Denver, CO
Detroit, MI
Flint, MI
Fort Worth, TX
Grand Rapids, MI
Greeensboro, NC
Hartford, CT
Houston, TX
Indianapolis, IN
Kansas City, MO
Lancaster, PA
Los Angeles, CA
Louisville, KY
Memphis, TN
Miami, FL
Milwaukee, WI
Minneapolis, MN
Nashville, TN
New Haven, CT
New Orleans, LA
New York, NY
Philadelphia, PA
Pittsburgh, PA
Portland, OR
Providence, RI
Reading, PA
Richmond, VA
Rochester, NY
Saint Louis, MO
San Diego, CA
San Francisco, CA
San Jose, CA
Seattle, WA
Springfield, MA
Syracuse, NY
Toledo, OH
Utica, NY
Washington, DC
Wichita, KS
Wilmington, DE
Worcester, MA
York, PA
Youngstown, OH

4. The EPA reports a 94% decrease in the national average sulfur dioxide concentration between 1980 and 2020.

i. Estimate the number of lives saved each year among the current population by this reduction, all else being equal.




```
##             2.5 %     97.5 %
## SO2nf 0.01033081 0.03231923
```

```
##          2.5 %    97.5 %
## SO2 0.1721802 0.5386538
```

We can see that through the confidence intervals, with 95% confidence, the mortality rates are between 0.1721802-0.5386538, and when the emissions are curbed 94%, the interval is 0.01033081-0.03231923, which we see is a 94% decrease. Therefore, by using the intercept in problem 2 and multiply it by 0.94, (917.884*0.94) which means around 863 lives are saved per 100,000 people per year.

ii. What implicit assumptions are made by using metropolitan-level data from 1959-1961 to calculate this estimate?

The implicit assumption made about the metropolitan data from 1959-1961 is that people in 1980-2020 have a mortality at the same rate as people in 1959-1961, so people in 1980-2020 would die from other causes at the same rate as they did in 1959-1961.

iii. Do you think these assumptions are reasonable?

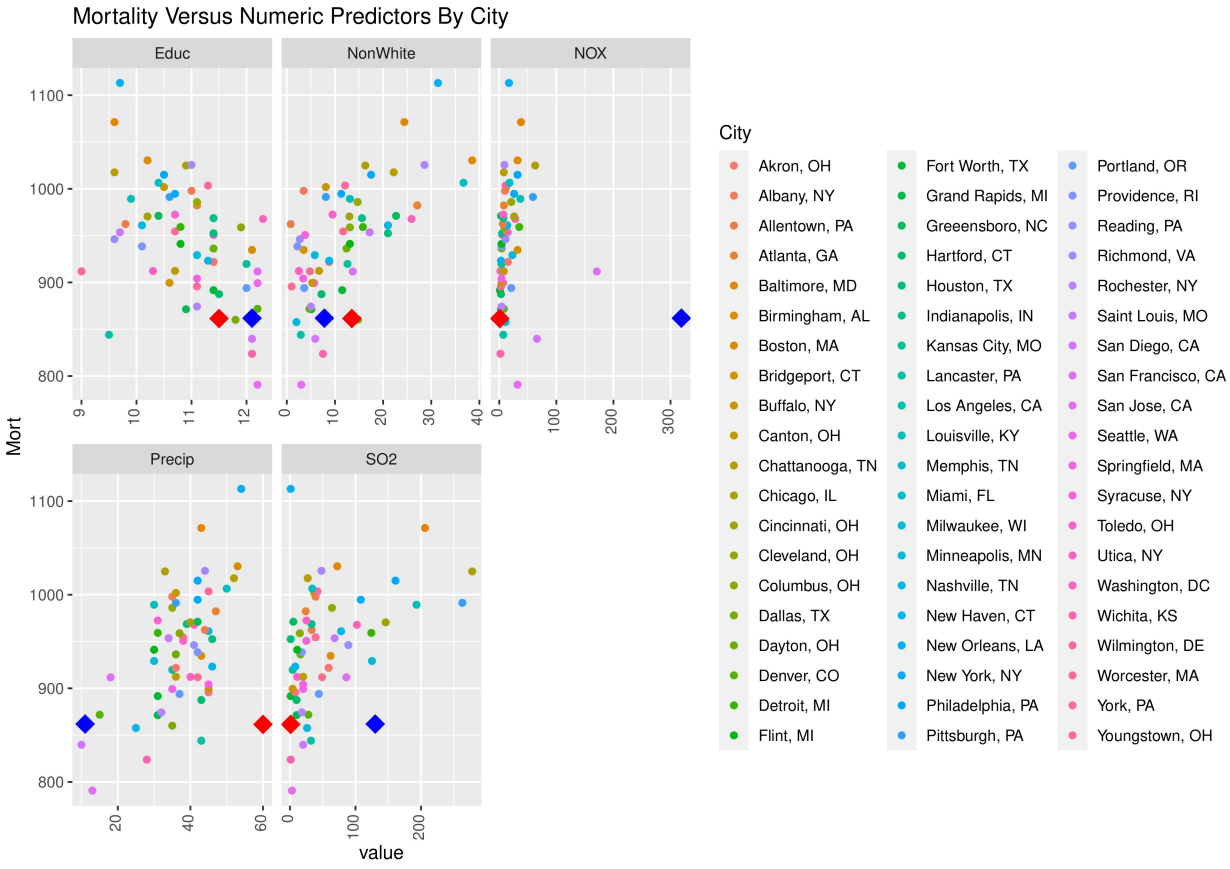
No, it is not reasonable to assume that people in 1959-1961 died at the same rate as people in 1980-2020 because of advancements in medicine and technology. However, recently with the pandemic, the mortality rate has increased, even with medicine and technology, so there are specific factors in certain years that cannot be assumed by previous data.

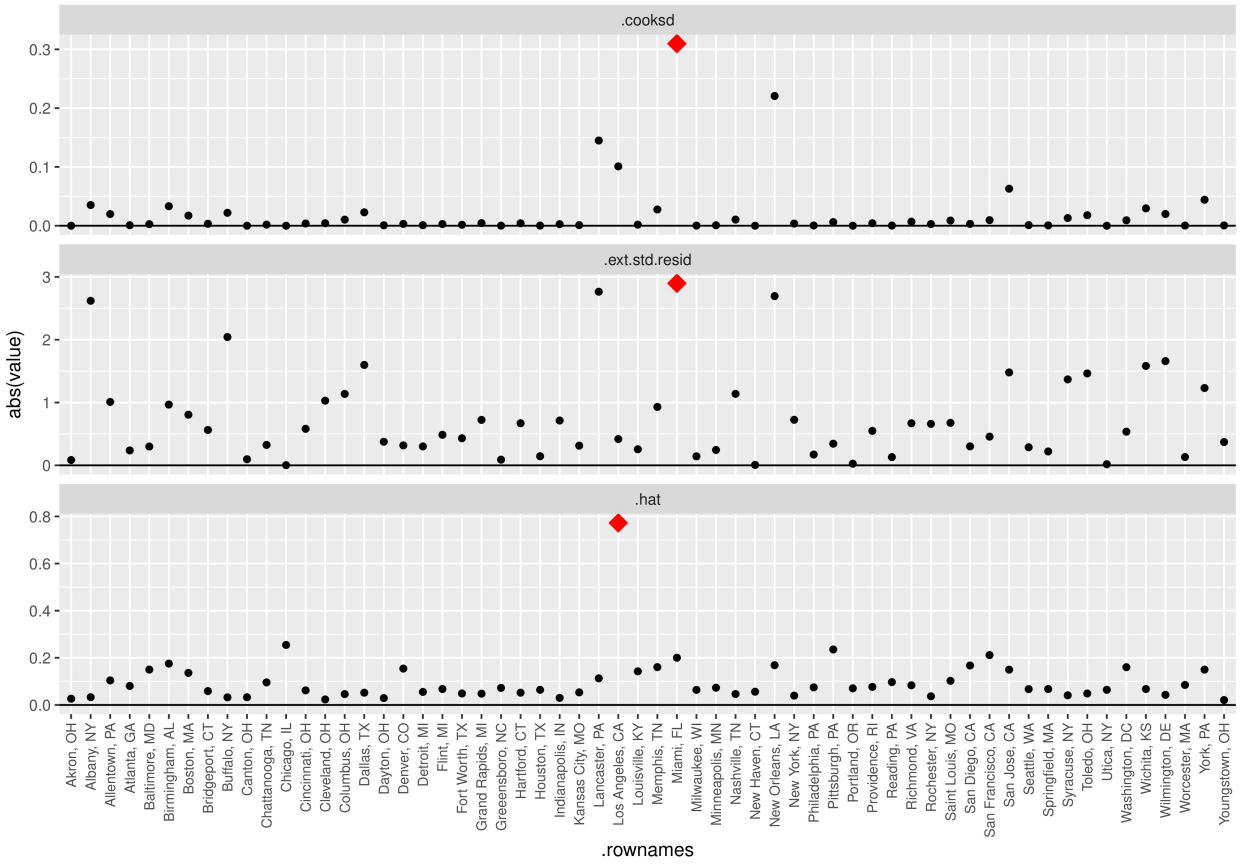
5. Which other variables, if any, seem associated with mortality? Comment briefly on any apparent associations.

```
##
## Call:
## lm(formula = Mort ~ . - City, data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.893 -18.986  -3.433  15.872  91.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1000.1026    92.3982  10.824 3.85e-15 ***
## Precip         1.3792     0.7000   1.970 0.053943 .
## Educ         -15.0791     7.0706  -2.133 0.037518 *
## NonWhite       3.1602     0.6287   5.026 5.84e-06 ***
## NOX           -0.1076     0.1359  -0.792 0.432030
## SO2            0.3554     0.0914   3.889 0.000278 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.36 on 54 degrees of freedom
## Multiple R-squared:  0.6698, Adjusted R-squared:  0.6392
## F-statistic:  21.9 on 5 and 54 DF,  p-value: 6.478e-12
```

The variables most closely associated with mortality is Non White, with a p value of 5.84e-06. Precip has a p value of 0.053943, which is close to 0.05, still showing an association. NOX has a p value of 0.432030, which leads us to conclude it does not have an association to mortality. SO2 and Educ have p values less than 0.05, meaning they are also closely associated with mortality.

6. Are any of the cities in the dataset unusual relative to the others? If so, in what way, and do these cities affect your conclusions?





We can see that Los Angeles and Miami are picked as unusual cities. Los Angeles is a leverage point in our graph, while Miami is an outlier and influential point. By looking at the data, Los Angeles has an unusually high NOX amount, while Miami has an unusually high precipitation amount. The leverage point does not affect our conclusions because it is one point, and the outlier "pulls" the fit toward itself, but the change is not substantial because it is only one point, so it does not affect our conclusions.

7. Are any of the variables besides mortality closely related with one another? How might this affect your analysis (if at all)?

Education is most closely related to Precip, with a p value of 0.0202. The evidence to support this claim is moderate, since not all p values are close to this number. Precip is closely related to NOX with a p value of 0.0045. The evidence to support this claim is strong, because there are no other p values close to this. NonWhite has the most significant relationship with NOX, but there is low-moderate evidence to support this claim because other p values are also similarly small. NOX has a significant relationship with SO2 with a p value of 0.002739. The evidence to support this claim is strong, as the other p values are not in a similar range.

Since our analysis focuses on the relationship between mortality, and the relationship between Education and Precip is not strong, we can assume that it does not affect our conclusions. This makes sense logically, because Education and Precip would not have a relationship in the US. In our analysis, the relationship between NonWhite and NOX could represent the minority workers that live near or around commerical factories or power plants that emit gases. This could also relate to lack of wealth, where poorer people live near more pollution, compared to wealthy people who have trees and parks around them. Also, the strong relationship between NOS and SO2 and NOX and Precip could suggest that areas with precipitation emit more of these gases, which should be discussed in the analysis.

Code appendix

```
# knit options
knitr::opts_chunk$set(echo = F,
                      results = 'markup',
                      fig.width = 4,
                      fig.height = 3,
                      fig.align = 'center',
                      message = F,
                      warning = F)

# packages
library(tidyverse)
library(tidymodels)
library(modelr)
library(gridExtra)
library(grid)
library(ggplot2)
library(broom)
# read in data and show example rows
pollution <- read_csv('pollution.csv')
head(pollution, 3)
pollution

mortvpred_scatterplot<- pollution %>%
  pivot_longer(cols= c(Precip, Educ, NonWhite,
                       NOX, SO2)) %>%
  ggplot(aes(x = value, y= Mort)) +
  labs(title= 'Mortality Versus Numeric Predictors By City') +
  facet_wrap(~ name, scales= 'free_x', nrow= 2) +
  geom_point(aes(color= City)) +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.25,
                                   hjust = 1))
mortvpred_scatterplot


newpollution <- pollution[-c(8, 19),]

fit_MortSO2 <-lm(Mort ~ SO2, data= pollution)
fit_MortSO2
#0.418
fit_MortNOX <-lm(Mort ~  NOX, data= newpollution)
fit_MortNOX
#1.132

fit_Mort <- lm(Mort ~ . - City, data = pollution)
halfSO2 <- pollution$SO2*2

fit_MortSO2half <-lm(Mort ~ . -City - SO2 + halfSO2, data= pollution)
fit_MortSO2half

confint(fit_MortSO2half, 'halfSO2', level = .95)
```

```r
confint(fit_Mort, 'SO2', level = .95)

halfNOX <- pollution$NOX/0.5

fit_MortNOXhalf <- lm(Mort ~ . - City - NOX + halfNOX, data = pollution)
fit_MortNOXhalf

confint(fit_Mort, 'NOX', level = .95)
confint(fit_MortNOXhalf, 'halfNOX', level = .95)

SO2reducedhalf_scatterplot<- pollution %>%
  ggplot(aes(x = SO2/2, y= Mort-Mort/2)) +
  labs(title= 'Estimated Lives Saved By Curbing SO2 Emissions by 50%', x= 'Tons of SO2 curbed per day p
  geom_point(aes(color= City)) +
  geom_smooth(method='lm', formula= y~x, se= F) +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.25,
                                   hjust = 1))
SO2reducedhalf_scatterplot

#Graphing the estimate
pollution %>%
  ggplot(aes(x = SO2*.06, y = Mort*.06)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x, se = F) +
  labs(x = 'Tons of SO2 per day per square km', y = 'Deaths per 100,000 People per Year')


#confidence intervals
SO2nf <- pollution$SO2/.06
SO2nfFit <- lm(Mort ~ . - City - SO2 + SO2nf, data = pollution)
confint(SO2nfFit, 'SO2nf', level = .95)
confint(fit_Mort, 'SO2', level = .95)
summary(fit_Mort)

fitMort<- lm(Mort ~ ., data= pollution)
fitMort_df <- augment(fit_Mort)

#studentize function

studentize <- function(resid, n, p){
  resid*sqrt((n - p - 1)/(n - p - resid^2))
}

n <- nrow(model.matrix(fit_Mort))
p <- ncol(model.matrix(fit_Mort)) - 1



fitMort_dfscatter<- fitMort_df %>%
  mutate(.ext.std.resid = studentize(.std.resid,
                                     n, p),
         .rownames = City) %>%
```

```
  pivot_longer(c(.hat,
                 .cooksd,
                 .ext.std.resid)) %>%
  ggplot(aes(x = .rownames, y = abs(value))) +
  facet_wrap(~ name,
             ncol = 1,
             scales = 'free_y') +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1,
                                   vjust = 0.5,
                                   size = 8))

#unusual observations
unusualobs_cooksd <- fitMort_df %>% slice_max(c(.cooksd), n = 1)
unusualobs_cooksdg <- unusualobs_cooksd %>%
  pivot_longer(c(Precip,
                 Educ,
                 NonWhite,
                 NOX,
                 SO2))

unusualobs_hat <- fitMort_df %>% slice_max(c(.hat), n = 1)
unusualobs_hatg <- unusualobs_hat %>%
  pivot_longer(c(Precip,
                 Educ,
                 NonWhite,
                 NOX,
                 SO2))


mortvpred_scatterplot + geom_point(data = unusualobs_cooksdg,
                                   color = 'red',
                                   shape = 'diamond',
                                   size = 5) +
  geom_point(data = unusualobs_hatg,
                                   color = 'blue',
                                   shape = 'diamond',
                                   size = 5)

#cooks and hat graph
unusual_obs <- fitMort_df %>%
  mutate(.ext.std.resid = studentize(.std.resid,
                                     n, p),
         .rownames = City) %>%
  pivot_longer(c(.hat,
                 .cooksd,
                 .ext.std.resid)) %>%
  group_by(name) %>%
  slice_max(order_by = abs(value), n = 1) %>%
  ungroup()
```

```
fitMort_dfscatter + geom_point(data = unusual_obs,
                                    color = 'red',
                                    shape = 'diamond',
                                    size = 5)
fitEduc<- lm(Educ ~ . - City, data= pollution)
fitEduc%>%summary

fitNW<- lm(NonWhite ~ . - City, data= pollution)
fitNW%>%summary

fitNOX<- lm(NOX ~ .  - City, data= pollution)
fitEduc%>%summary

fitPrecip<- lm(Precip ~ . - City, data= pollution)
fitPrecip%>%summary

fitSO2<- lm(SO2 ~ . - City, data= pollution)
fitSO2%>%summary
```