PSTAT 126 Final Assignment

Peiyu Yang

Instructions

This final assignment requires you to do an end-to-end regression analysis, but all prompts given are qualitative questions or requests that one might ask an analyst to answer. Consequently, you are responsible for carrying out model building and checking and determining which calculations to perform and which plots to construct with specific endpoints in mind. Accordingly, it is recommended that you read all prompts first before beginning your work.

You are allowed to consult course materials and classmates as you work on the assignment, but you are expected to prepare your own submission, which means that you should write your own codes and write your answers in your own words. As such, collaboration should be mostly limited to discussion. If you choose to share your work with others, please give your classmates the opportunity to think about how to implement and report relevant analyses and refrain from directly sharing written work in reproducible form. By submitting your work you are acknowledging that you have adhered to these expectations.

Please use this .Rmd file as a template and *modify your own copy of it* to complete the assignment by answering all questions as instructed.

Tip: knit periodically as you go to avoid headaches at the submission stage. Submission instructions follow at the end of the assignment. Enjoy!

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.

```
# read in data and show example rows
pollution <- read_csv('pollution.csv')
head(pollution, 3)
```

```
## # 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
```

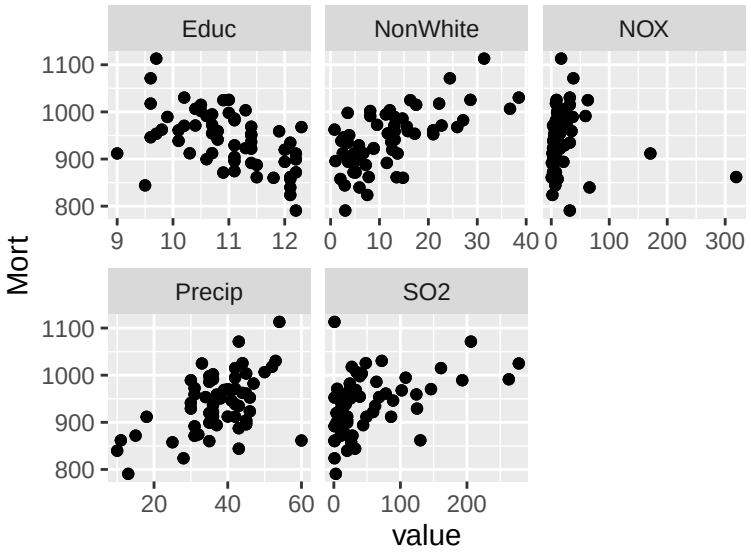
In this data the presence of pollutants is reported as *relative pollution potential*, which is calculated by scaling emissions (tons per day per square kilometer) by a dispersion factor based on local conditions (mixing, wind, area, and the like).

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).



Almost all four plots are close to linear, but for the NOX one, it seems like there are some outliers existing.

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.

```
##  (Intercept)       Precip         Educ     NonWhite          NOX          SO2
## 1000.1026427    1.3792057  -15.0790502    3.1602307   -0.1075735    0.3554170
```

Since it is close to linear, we can fit this with linear model with Mort. Since City is irrelevant, we can drop it here. And the coef can give us the estimates for each predictor.

NOX: For every increase of 1 ton of NOX released per day per km is related to about 0.11 less death per 100k during 1959 to 1961 after accounting for other predictors.

SO2: For every increase of 1 ton of SO2 released per day per km is related to about 0.36 more death per 100k during 1959 to 1961 after accounting for other predictor.

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

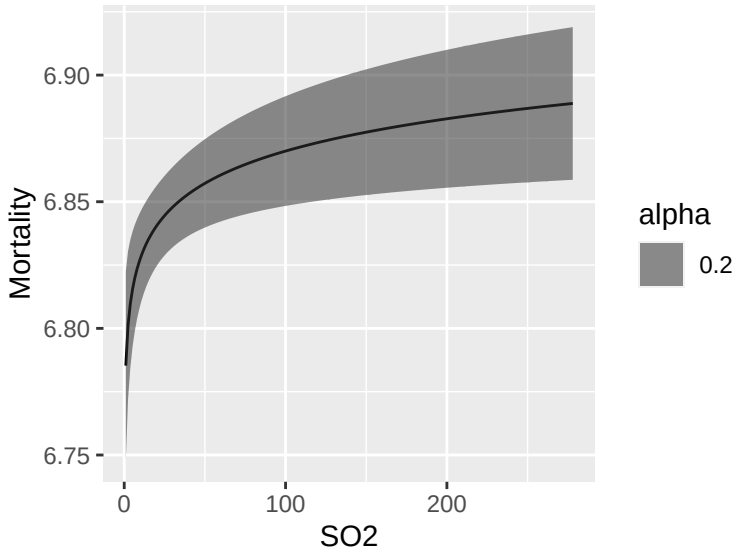
    i. Estimate the reduction in mortality rate associated with a 50% relative decrease in sulfur dioxide emissions.

```
##   log(SO2)
## 0.01839368
```

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

```
##   log(NOX)
## 0.01589271
```

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



There is a fan shape regarding to the plot, which means that as we reduce SO2 more, we will save more lives.

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.

329.5*10^6/(0.1*10^6)/11.358*log(1-0.94)|/40=28.79394971

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

The data are restricted to the time, and all cities are assumed to be equal, which means all citizens are under the same condition of pollution. It is assumed that other variables here are the same.

iii. Do you think these assumptions are reasonable?

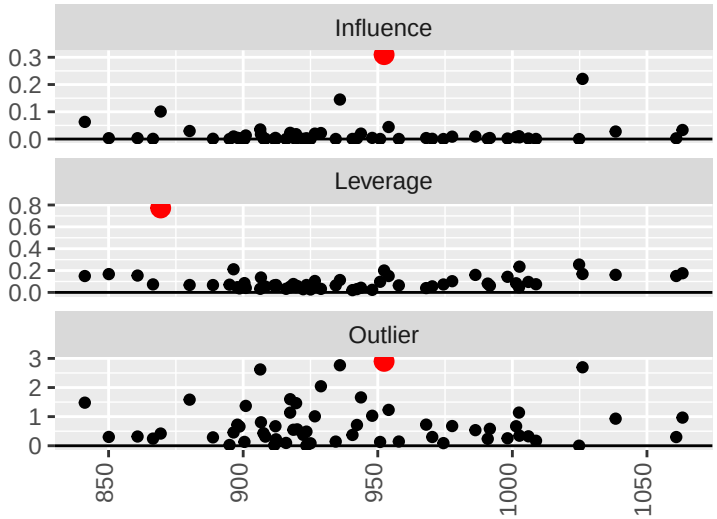
I think it is not reasonable, since it is impossible that pollution in all cities are the same, and there are also many other factors that may differ from cities to cities. It is unlikely that the real situation will be this ideal since many variables will differ to each other.

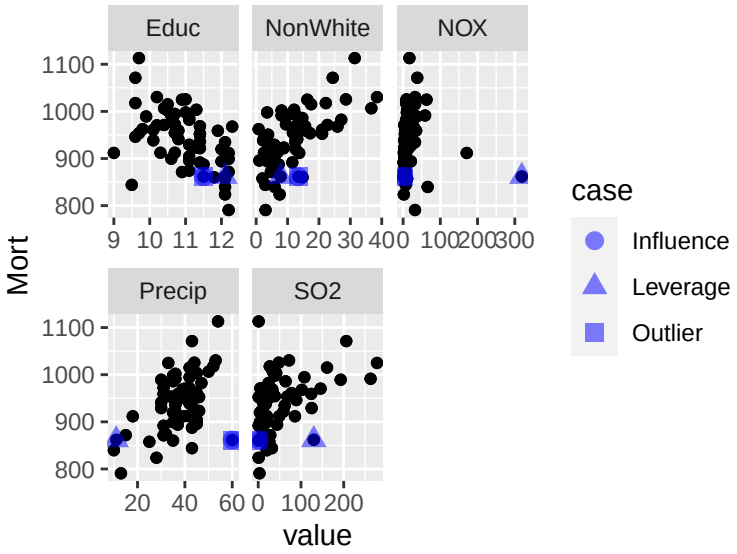
5. Which other variables, if any, seem associated with mortality? Comment briefly on any apparent associations.

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + NOX + SO2, 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 increase of the median education for 25 years old and above is related to the decrease of mortality. The increase of percentage of nonwhite population it related to the increase of 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?





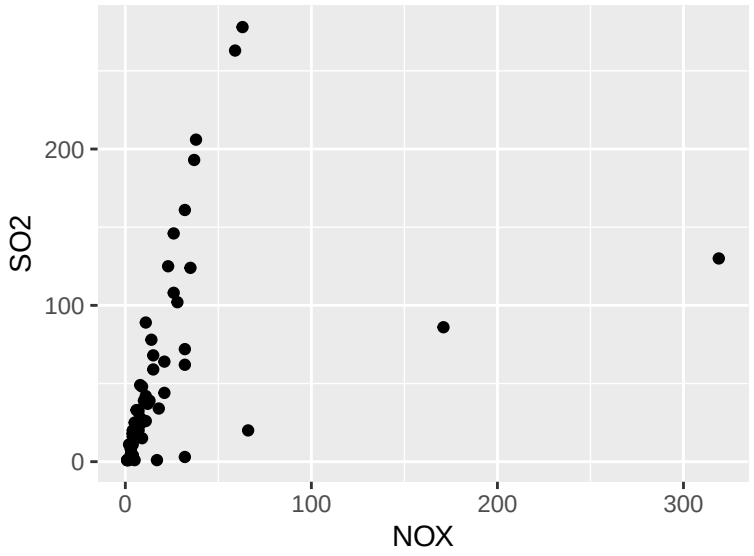

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + NOX + SO2, data = pollution[-unusual_idx,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.653 -20.636  -3.442  18.183  91.690
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1004.65024   96.19626  10.444 2.87e-14 ***
## Precip         1.39894    0.71948   1.944 0.057374 .
## Educ         -15.56126    7.40222  -2.102 0.040489 *
## NonWhite       3.12826    0.65515   4.775 1.56e-05 ***
## NOX           -0.12764    0.14858  -0.859 0.394352
## SO2            0.36098    0.09477   3.809 0.000376 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.33 on 51 degrees of freedom
## Multiple R-squared:  0.6686, Adjusted R-squared:  0.6361
## F-statistic: 20.58 on 5 and 51 DF,  p-value: 3.439e-11

##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + NOX + SO2, 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
```

There are unusual cities, but these unusual cities doesn't effect the conclusion.

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



There is a close relationship between NOX and SO2, which means even though there are outliers, we can still see there is a positive relationship between SO2 and NOX. Although they are related, I think there would be inconsiderable influence on the conclusion.

Submission instructions

1. Clear your environment and run all codes to check for errors. Resolve any if detected.
2. Input your name in the author information, remove the instructions at the beginning and end of the document, and knit to pdf.
3. Inspect the pdf and fix any display issues.
4. Once the pdf looks good, upload a copy to Gradescope.
5. Download a backup copy of your work and store locally.

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(modelr)
library(broom)
library(faraway)
# read in data and show example rows
pollution <- read_csv('pollution.csv')
head(pollution, 3)
p_scatter<-pollution %>%
  pivot_longer(cols=c(Precip, Educ, NonWhite, NOX, SO2)) %>%
  ggplot(aes(x=value, y=Mort))+geom_point()+ facet_wrap(~name, scales='free_x')+geom_point()
p_scatter
fit_pol <- lm(Mort~ Precip+Educ+NonWhite+NOX+SO2, data=pollution)
coef(fit_pol)
fit_log<-lm(log(Mort) ~ log(SO2), data=pollution)
coef(fit_log)[-1]
fit_logD<-lm(log(Mort)~log(NOX), data=pollution)
coef(fit_logD)[-1]
pred_df<-data_grid(data=pollution, SO2=seq_range(SO2,200),.model=fit_log)
pred_df<-pred_df %>%
  cbind(ci=predict(fit_log,pred_df,interval='confidence'))
ggplot(data=pred_df, aes(x=SO2, y=Mortality))+geom_path(aes(y=ci.fit))+
  geom_ribbon(aes(ymin=ci.lwr, ymax=ci.upr, y=ci.fit, alpha=0.2))+
  guides(color='none')
summary(fit_pol)
studentize_fn <- function(resid, n, p){
      resid*sqrt((n - p - 1)/(n - p - resid^2))
      }
    x_matrix <- model.matrix(fit_pol)
    pollution <- augment(fit_pol, pollution) %>%
      mutate(.ext.std.resid = studentize_fn(.std.resid,
                                            n=nrow(x_matrix),
                                            p=ncol(x_matrix)-1))
    p_caseinf <- pollution %>%
      rename(Outlier = .ext.std.resid,
             Leverage = .hat,
             Influence = .cooksd) %>%
      pivot_longer(cols = c(Outlier, Leverage, Influence)) %>%
      ggplot(aes(x = .fitted, y = abs(value))) +
      facet_wrap(~ name, scales = 'free_y', nrow = 3) +geom_point()+
      geom_hline(aes(yintercept=0))+
      theme(axis.text.x=element_text(angle=90, vjust=0.25, hjust=1))+labs(x='',y='')
```

```
    unusual_obs<-pollution %>%
      rename(Outlier=.ext.std.resid,
             Leverage=.hat,
             Influence=.cooksd) %>%
      pivot_longer(cols=c(Outlier, Leverage, Influence)) %>%
      mutate(obs_idx=row_number()) %>%
      group_by(name) %>%
      slice_max(order_by=abs(value),n=1) %>%
      ungroup()
    p_caseinf+geom_point(data=unusual_obs, color='red', size=3)
    unusual_obs_long<-unusual_obs%>%
      rename(case=name) %>%
      select(Mort, Precip, Educ, NonWhite, NOX, SO2, case) %>%
      pivot_longer(cols=c(Precip, Educ, NonWhite, NOX, SO2))
    p_scatter+geom_point(data=unusual_obs_long, aes(shape=case), color="blue",
                                                size=3, alpha=0.5)
    unusual_idx<-pull(unusual_obs, obs_idx)
    fit_dropgroup<-lm(Mort~ Precip+Educ+NonWhite+NOX+SO2, data=pollution[-unusual_idx, ])
    summary(fit_dropgroup)
    summary(fit_pol)
pollution %>%
  ggplot(aes(x=NOX, y=SO2)) +
  geom_point()
```