PSTAT 126 Final Assignment

Andrew Pascual

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
```

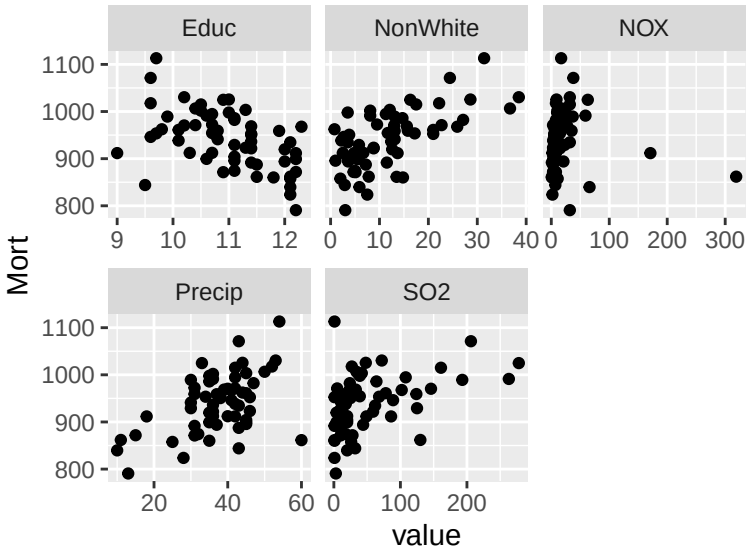
In this data the presence of pollutants is reported as *relative pollution potential*, which is calculated by scaling emissions (tons per day per square kilometer) by a dispersion factor based on local conditions (mixing, wind, area, and the like).

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).



It seems as though in nonwhite, high precipitation, and high SO2 emissions areas have higher mortality rates. Also, higher educated areas have lower mortality rates, and NOX emissions does not seem to have a big impact on mortality, however it does seem to have some outliers.

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 ~ NOX, data = pollution)
##
## Coefficients:
## (Intercept)          NOX
##    942.7095      -0.1039

##
## Call:
## lm(formula = Mort ~ log(SO2), data = pollution)
##
## Coefficients:
## (Intercept)     log(SO2)
##      886.82        16.74
```

The association between mortality and NOX emissions is slightly negative but doesn't seem significant. The association between mortality and SO2 emissions is positive and seems to have a significant affect on mortality.

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.

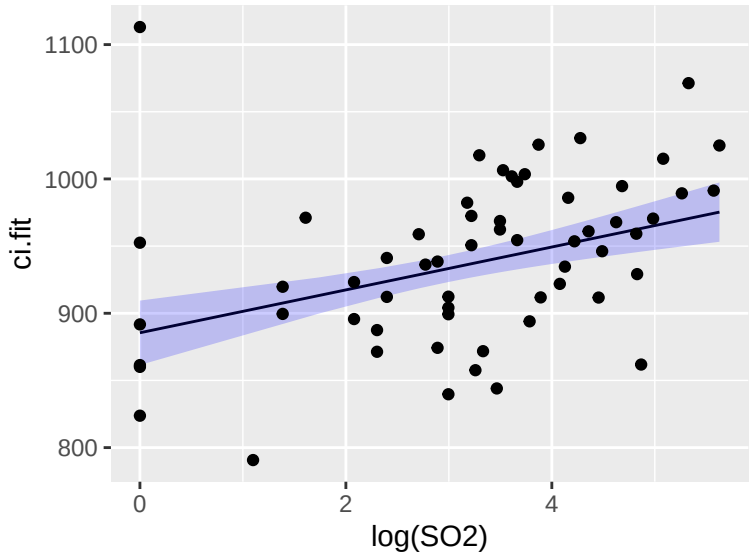
```
##                2.5 %    97.5 %
## log(SO2) -0.5135647 -15.23173
```

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

```
##        2.5 %      97.5 %
## NOX 0.178597 -0.08182466
```

```
iii. Construct a visualization that conveys the estimated potential lives saved by reducing SO2 emissio
```

```
##     2.5 % 97.5 %
## SO2    NA     NA
```



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 %
## log(SO2) 6.347976 25.12995
```

The amount of lives per year is about 6.34-25.13

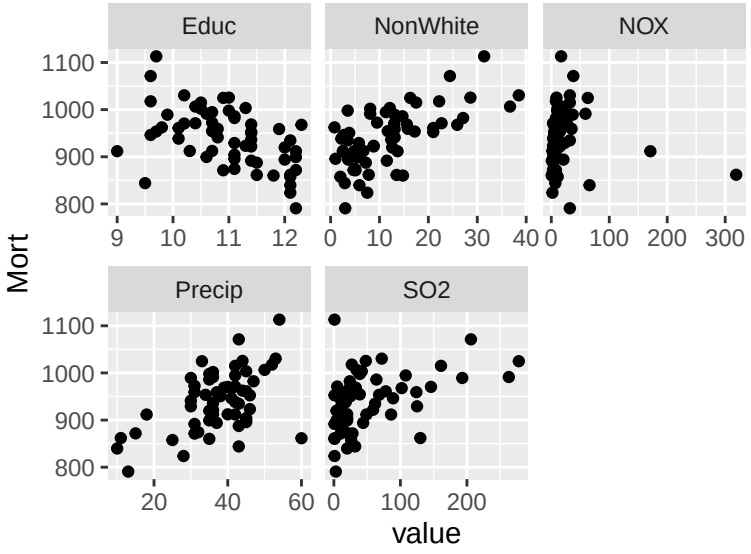
```
ii. What implicit assumptions are made by using metropolitan-level data from 1959-1961 to calculate this
The assumptions are that the data is comparable and that lower SO2 levels will influence mortality rate

iii. Do you think these assumptions are reasonable?
These assumptions may not be reasonable, as accoring to the graph from epa.gov, the air quality has dec
```

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

Education, Precipitation, and Nonwhite all seem to have somewhat of an association according to the marginal relationships from Q1. It seems like a higher percentage of nonwhite population and a higher rate of precipitation leads to a higher mortality rate. On the other hand, cities with a higher median education seem to have lower mortality rates.

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?



New Orleans is the only city that is unusual compared to the others. I believe that this does not affect the conclusions significantly.

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

```
##                Mort     Precip       Educ   NonWhite         NOX         SO2
## Mort       1.00000000  0.5094923 -0.5109838  0.6437334 -0.07738348  0.4259528
## Precip     0.50949235  1.0000000 -0.4904252  0.4132045 -0.48732074 -0.1069239
## Educ      -0.51098382 -0.4904252  1.0000000 -0.2087739  0.22440191 -0.2343459
## NonWhite   0.64373344  0.4132045 -0.2087739  1.0000000  0.01838530  0.1592930
## NOX       -0.07738348 -0.4873207  0.2244019  0.0183853  1.00000000  0.4093936
## SO2        0.42595282 -0.1069239 -0.2343459  0.1592930  0.40939361  1.0000000
```

The variables that seem to have correlation with one another are education and precipitation and NOX and precipitation. These relationships would not have a significant impact on the analysis that I have done.

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(tidymodels)
library(modelr)
library(faraway)
# read in data and show example rows
pollution <- read_csv('pollution.csv')
head(pollution, 3)
pollution %>%
  pivot_longer(-c(Mort, City)) %>%
  ggplot(aes(x = value, y = Mort)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point()
fit <- lm(Mort ~ Educ + NonWhite + NOX + Precip + log(SO2), data = pollution)

fit_NOX <- lm(Mort ~ NOX, data = pollution)
fit_SO2 <- lm(Mort ~ log(SO2), data = pollution)
fit_NOX
fit_SO2
fit2 <- lm(Mort ~ Educ + NonWhite + Precip +log(NOX) + log(SO2), data = pollution)
confint(fit2, 'log(SO2)', level = 0.95) * log(0.5)
confint(fit, 'NOX')*(-0.5)
maxdiff <- range(pollution$SO2) %>% diff
confint(fit_SO2, 'SO2') * maxdiff

add_preds <- function(.data, .model){
  .data %>% cbind(ci = predict(.model, .data, interval = 'confidence')) %>%
    as_tibble()
}

pollution %>%
  data_grid(SO2 = seq_range(SO2, 60), .model = pollution) %>%
  add_preds(fit) %>%
  ggplot(aes(x = log(SO2), y = ci.fit)) +
  geom_path() +
  geom_ribbon(aes(ymin = ci.lwr, ymax = ci.upr), alpha = 0.2, fill = 'blue') +
  geom_point(aes(y = Mort), data = pollution)

confint(fit_SO2, 'log(SO2)') * (0.94)
fit <- lm(Mort ~ ., data = pollution)
fit_df <- augment(fit)
n <- nrow(model.matrix(fit))
```

```r
p <- ncol(model.matrix(fit)) - 1
studentize <- function(resid, n, p){
  resid*sqrt((n - p - 1)/(n - p - resid^2))
}

fig1 <- pollution %>%
  pivot_longer(-c(Mort, City)) %>%
  ggplot(aes(x = value, y = Mort)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point(aes())
unusual_obs <- fit_df %>%
  slice_max(.cooksd)

unusual_obs_long <- unusual_obs %>%
  pivot_longer(c(Precip,
                 Educ,
                 NonWhite,
                 SO2,
                 NOX))

fig1 + geom_point(data = unusual_obs_long,
                  color = 'red')
pollutionCorr <- pollution[c(2:7)]
cor(pollutionCorr)
```