Assignment 2 Submission

Your name(s)

submission date

Statement of Contribution

Include this if pairing.

State what each team member has contributed to the assignment before answering the questions. Each student must contribute towards ALL questions. Your assignment will not be marked without this statement.

Question 1: Air Pollution and Mortality. Does pollution kill people? (30 marks)

(a) (5 marks) Carry out exploratory data analysis (EDA) of this dataset.

Answer:

Climate EDA: The Mortality-Precip plot shows a strong positive linear relationship, indicating that as the precipitation increases, so does the Mortality. The Mortality-Humidity plot shows no clear relationship as datapoints appear randomly scattered. The Mortality-JanTemp plot depicts either no clear relationship or a very weak negative relationship. Lastly the Mortality-JulyTemp plot depicts a moderate positive relationship with a few outliers.

Socioeconomic Variables EDA: From the pairs plot, we can see that the Mortality-Over65 graph appears to have no apparent relationship. The Mortality-House plot shows a weak to moderate positive relationship. The Mortality-Educ plot shows a strong negative relationship. The Mortality-Sound plot shows a moderate to strong negative relationship. The Mortality-Density plot has a moderate positive relationship. The Mortality-NonWhite plot shows a strong positive relationship. The Mortality-WhiteCol plot shows no clear relationship. Lastly the Mortality-Poor plot shows a moderate positive relationship.

Pollution Variables EDA: The Mortality-HC plot shows no correlation between the two variables, same with the Mortality-NOX plot. This is because the datapoints appear randomly scattered. However the Mortality SO2 plot depicts a moderate positive relationship. This means as the SO2 increases so does mortality.

Mortality by Region Analysis: The south region has the highest rates of mortality, the highest median (975) as well as the widest range. The median and range are notably higher than the rest of the regions. The West region has the lowest average mortality rate (875) as well as the most narrow spread. The Midwest and Northeast are similar in average values (~940) and an outlier in the Midwest data. The regional differences stated suggests that the region a person lives in may influence their mortality rate.

Mortality by State Code Analysis: Before we plotted our graph, I first examined the data, using table(pollution$State.code). Many of our states only have a singular datapoint and therefore are too small a sample size. Excluding the states with one datapoint we are left with 10 that we plot. The mean mortality rate differs greatly from each state. The states NY, PA, and OH have higher median mortality rates while CA and TX have significantly lower ones. States such as CA and PA having much wider interquartile range indicating a high variance. This suggests that the state an individual lives is relevant to their mortality rate.

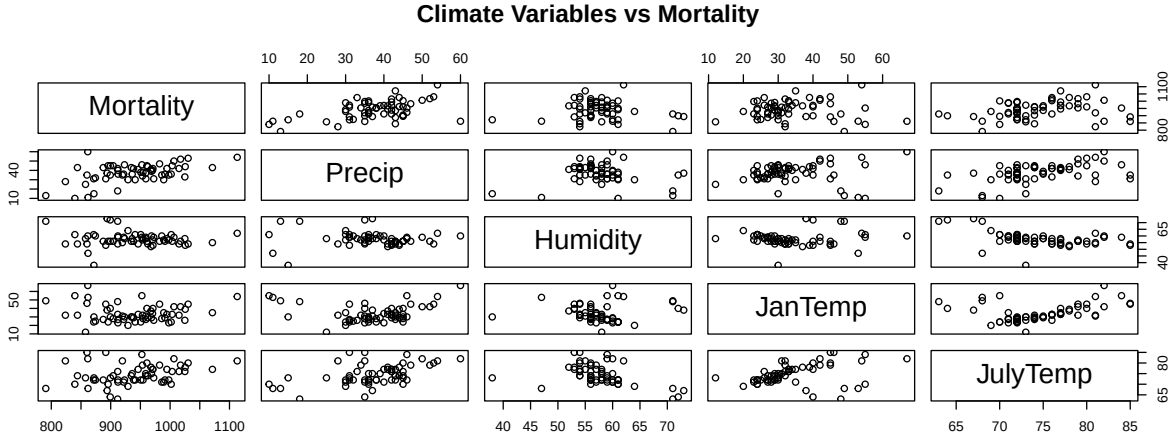


Figure 1: Climate Variables vs Mortality

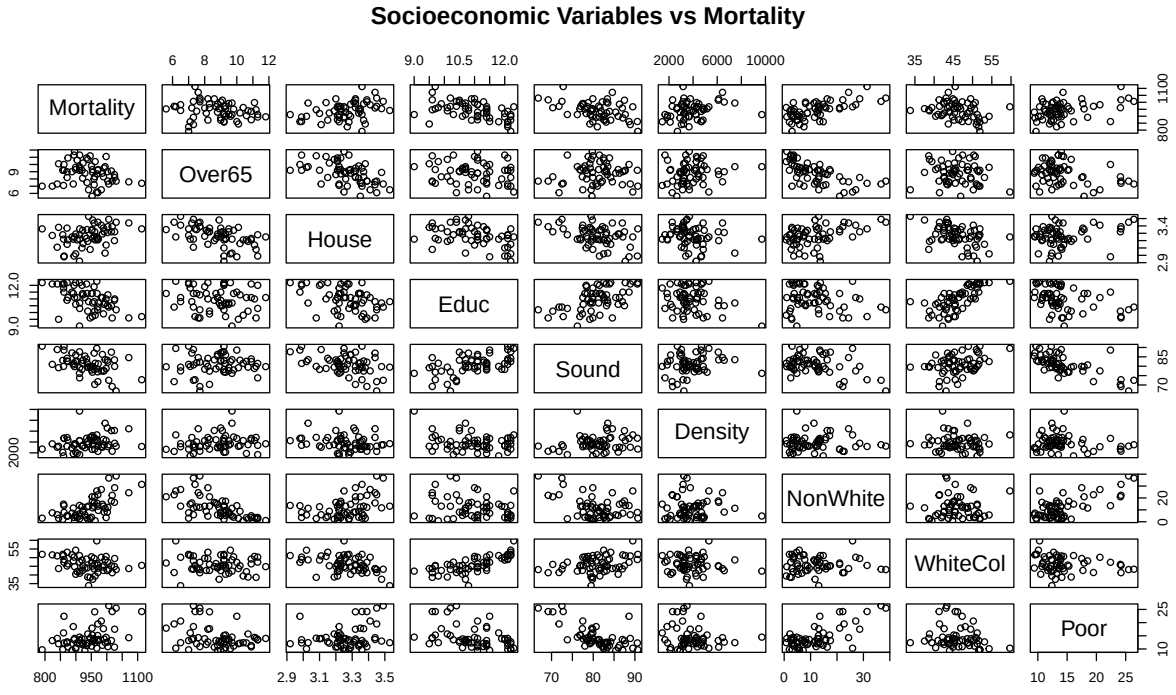


Figure 2: Socioeconomic Variables vs Mortality

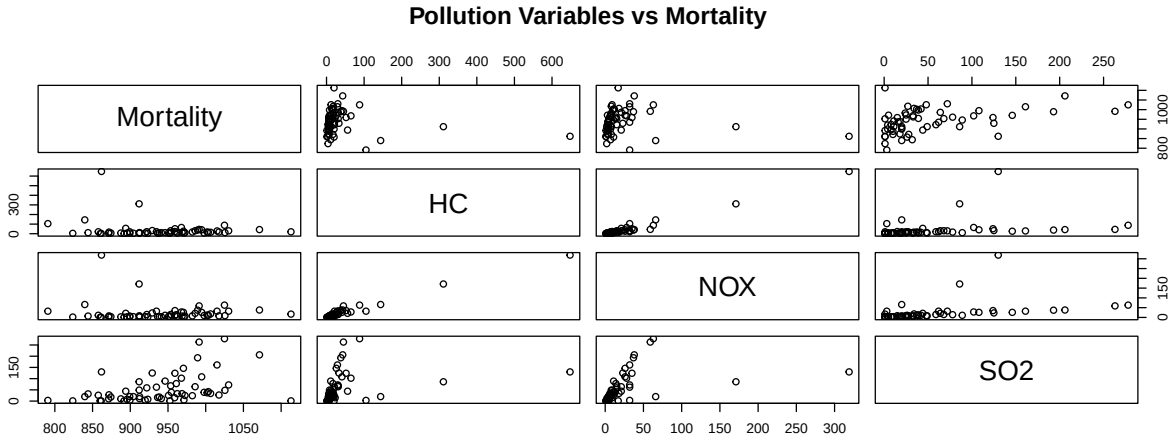


Figure 3: Pollution Variables vs Mortality

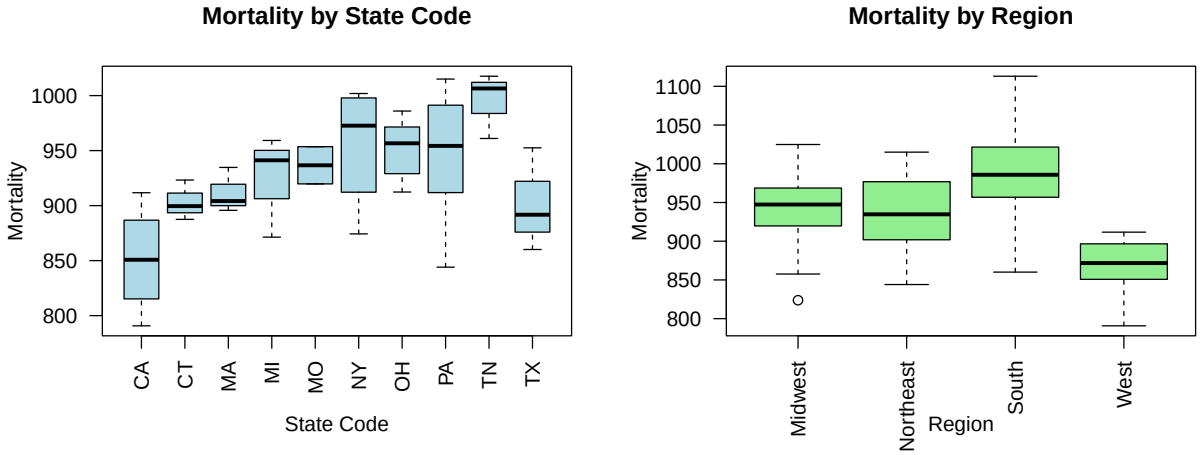


Figure 4: Mortality by Region and State Code

(b) (9 marks) Perform model selection process using the all subset method to arrive at good regression models that account for variation in mortality between the cities that can be attributed to differences in climate and socioeconomic factors. Identify optimal model(s) based on each of the adjusted R^2, BIC and C_p. Use a maximum of 10 variables for the selection. For each criterion, state the number and names of the selected variables.

Answer:

For the adjusted R^2 model, we have 7 variables. Those variables are Precip, JanTemp, JulyTemp, House, Educ, Density and NonWhite.

For the BIC model, we have 4 variables. Those variables are JanTemp, House, Educ and NonWhite.

For the Cp model, we have 6 variables. Those variables are Precip, JanTemp, JulyTemp, Educ, Density and NonWhite.

```
library(leaps)
source("all-subsets-lm.R")

poll_data <- read.csv("Pollution.csv")

# Using the subset of variables
model <- lm(Mortality ~ Precip + Humidity + JanTemp + JulyTemp +
                        Over65 + House + Educ + Sound + Density +
                        NonWhite + WhiteCol + Poor,
            data = poll_data)

allsubsets_adjR2 <- all_subsets_lm(model, criterion = "adjR2", p_max = 10)
allsubsets_bic   <- all_subsets_lm(model, criterion = "BIC", p_max = 10)
allsubsets_c_p   <- all_subsets_lm(model, criterion = "Cp", p_max = 10)

cat("Variables selected by adjusted R²:\n")
```

```
Variables selected by adjusted R²:
```

```
print(allsubsets_adjR2$variables)
```

```
[1] "Precip"   "JanTemp"  "JulyTemp" "House"    "Educ"     "Density"  "NonWhite"
```

```
cat("\nVariables selected by BIC:\n")
```

```

Variables selected by BIC:
```

```
print(allsubsets_bic$variables)
```

```
[1] "JanTemp"  "House"    "Educ"     "NonWhite"
```

```
cat("\nVariables selected by Cp:\n")
```

```

Variables selected by Cp:
```

```r
print(allsubsets_c_p$variables)
```

```
[1] "Precip"   "JanTemp"  "JulyTemp" "Educ"     "Density"  "NonWhite"
```

(c) (4 marks) Fit the model with the lowest C_p as obtained in (b). Write the equation of this fitted model. Interpret this model.

Answer:

Equation: Mortality = 1242.0 + 1.401 × Precip - 1.684 × JanTemp - 2.840 × JulyTemp - 16.16 × Educ + 0.00757 × Density + 5.275 × NonWhite

Interpretation: From viewing the model summary, the R^2 value tells us that the model explains approximately 70.86% of the variation in the data's mortality rate. The model summary also tells us the F-Statistic and the p-value. Those being 21.48 and 1.305e-12 respectively. The large F-statistic (>1) suggests that the model has significant variables chosen. The small p-value (<0.05) also is evidence against the null hypothesis and that our model is statistically significant.

We then interpret the coefficients for our model: - A unit increase in Precip increases the rate by 1.401 deaths per 100,000 - For each degree increase in JanTemp, mortality rate decreases by 1.684 deaths per 100,000 - For each degree increase in JulyTemp, mortality rate decreases by 2.840 deaths per 100,000 - Every year spent in education decreases the mortality rate by 16.16 per 100,000 - Per unit increase in Density, the mortality rate increases by 0.00757 per 100,000 - For every percentage increase in NonWhite, the mortality increases by 5.275 deaths per 100,000

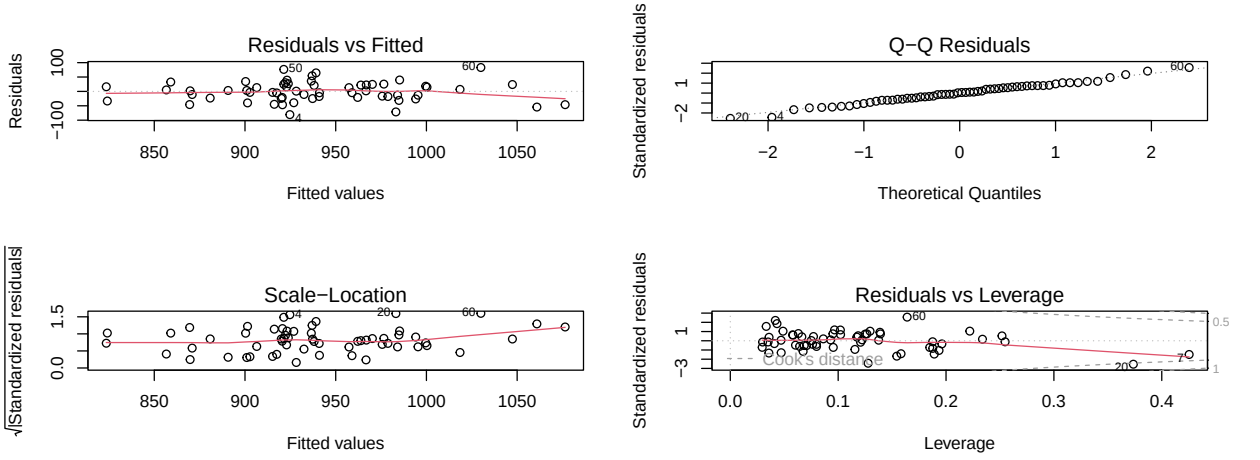
```
Call:
lm(formula = Mortality ~ Precip + JanTemp + JulyTemp + Educ +
    Density + NonWhite, data = poll_data)

Residuals:
    Min      1Q  Median      3Q     Max
-80.685 -21.529   1.422  22.777  83.055

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.242e+03  1.233e+02  10.078 6.41e-14 ***
Precip       1.401e+00  6.074e-01   2.307   0.0250 *
JanTemp     -1.684e+00  5.330e-01  -3.161   0.0026 **
JulyTemp    -2.840e+00  1.289e+00  -2.203   0.0319 *
Educ        -1.616e+01  6.652e+00  -2.429   0.0186 *
Density      7.570e-03  3.316e-03   2.283   0.0265 *
NonWhite     5.275e+00  6.906e-01   7.639 4.24e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.43 on 53 degrees of freedom
Multiple R-squared:  0.7086,    Adjusted R-squared:  0.6756
F-statistic: 21.48 on 6 and 53 DF,  p-value: 1.305e-12
```

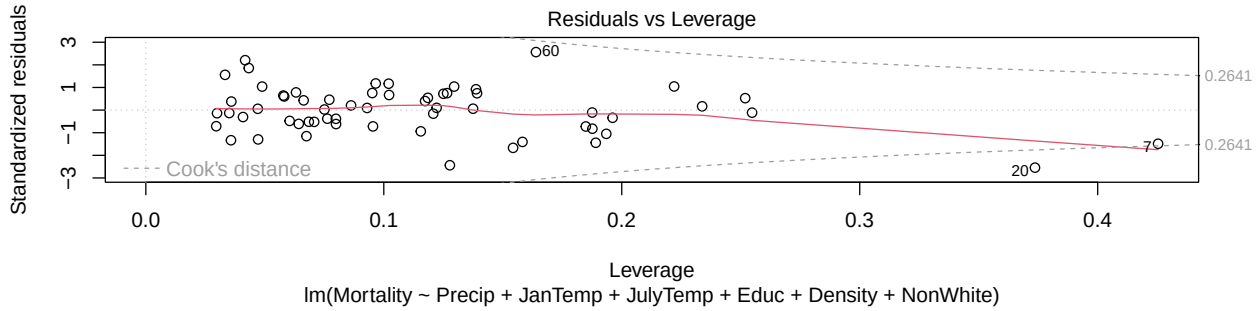
(d) (7 marks) Perform diagnostics checking on the model in (c). Do you think there are influential points in the data? Identify the cities which are influential points using leverage and Cook's distance respectively.

Answer:

We first generate our diagnostic plots, checking the Residuals vs Fitted, Q-Q Residuals, Scale-Location and Residuals vs Leverage. The Residuals vs Fitted has a mostly random scatter of data points. There's minor curvature at the right end but overall this plot suggests the relationship is approximately linear. The Q-Q Residuals follows the diagonal line closely with minor deviations at the ends. Overall, the plot suggests that the residuals are approximately normally distributed. The Scale Location is fairly horizontal however towards the right end it appears to curve upwards. This suggests the variance may be slightly uneven. The Residuals vs Leverage plot depicts data points that require further investigation as they have a high leverage and are far from x=0.

Taking a further look at influential points using leverage and Cook's distance, I use the equations Leverage cutoff = (2(p+1))/n and Cook's cutoff = 2 × (p + 1) / (n - p - 1) to find influential points. Using these equations as well as the linear model we can find the points above the cutoffs. The data point that is above the Cook's cutoff is data point 20, the city being York. The data points above the leverage cutoff are data points 3, 7, 8, 19 and 20. Their cities are San Diego, Miami, Los Angeles, San Francisco and York.






```r
# Find Influential Points using Cook's Distance
cooks_d <- cooks.distance(cp_model)
cook_cutoff_points <- which(cooks_d > cook_cutoff)
cat("Points above Cook's distance cutoff:\n")
```

```
Points above Cook's distance cutoff:
```

```
print(cook_cutoff_points)
```

```
20
20
```

```
# Find high leverage
lev_cutoff <- 2 * (p + 1) / n
high_leverage <- which(hatvalues(cp_model) > lev_cutoff)
cat("\nPoints with high leverage:\n")
```

```
Points with high leverage:
```

```
print(high_leverage)
```

```
 3  7  8 19 20
 3  7  8 19 20
```

```
# Find the Cities
cat("\nCities with high Cook's distance:\n")
```

```
Cities with high Cook's distance:
```

```
print(poll_data[cook_cutoff_points, "City"])
```

```
[1] "York, PA"
```

```
cat("\nCities with high leverage:\n")
```

```
Cities with high leverage:
```

```
print(poll_data[high_leverage, "City"])
```

```
[1] "San Diego, CA"     "Miami, FL"         "Los Angeles, CA"
[4] "San Francisco, CA" "York, PA"
```

(e) (5 marks) Using the model obtained in (c), add the three pollution variables (transformed to their natural logarithm) and obtain the p-value from the extrasum-of-squares F-test due to their addition. Summarise your findings in a few concise sentences.

Answer:

First I created the expanded model by adding the log-transformed variables of HC, NOX and SO2 into the base model. I used the extrasum-of-squares F-test on both the base and expanded linear models, giving me the p-value of 0.008313. This p-value is small, being less than 0.05. This means that the expanded model is statistically significant. This suggests the addition of the logarithm of the three pollution variables improved the model's ability to predict mortality rates. Therefore, we reject the null hypothesis.

```r
poll_data <- read.csv("Pollution.csv")

cp_model <- lm(Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite, data = poll_data)

cp_model_expanded <- lm(Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite +
                          log(HC) + log(NOX) + log(SO2), data = poll_data)

anova(cp_model, cp_model_expanded)
```

```
Analysis of Variance Table

Model 1: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite
Model 2: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite +
    log(HC) + log(NOX) + log(SO2)
  Res.Df   RSS Df Sum of Sq     F   Pr(>F)
1     53 66518
2     50 52712  3     13806 4.365 0.008313 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```