Analysis of Mortality

FINAL PROJECT

Haolin Li 913838107 | Haoyue Wang 913079144

STA108 | 08/01/2019

Introduction

Mortality is an essential factor when studying human population development. Therefore, it’s crucial for researchers to find out what correlates and could contribute to mortality. Previous studies have found that race, education, economic, and pollution can possibly be related to mortality. Hence, we will mainly focus on these factors in this project to discover the relationships in between and to predict mortality depending on the results.

In this project we will be using data collected from 60 Standard Metropolitan Statistical Area (SMSA) in the United States in years 1959-1961, from which we will pull several variables in order to find out if pollution has an effect on mortality. We will use total age-adjusted mortality [MORT] from all causes as the response variable, and choose mean annual precipitation [PRECIP], median number of school years completed by person of age 25 or over [EDUC], percentage of population in 1960 that is nonwhite [NONWHITE], percentage of households with annual income under $3000 in 1960 [POOR], relative pollution potential of oxides of nitrogen [NOX], and relative pollution potential of Sulphur dioxide [SO2] as our predictor variables. We will use R command as our tool to organize the data, obtain tables and plots, and perform the calculations to help us build the regression model, analyze the correlation between mortality and all predictor variables, and do the diagnostics.

Data Exploratory & Transformation

As mentioned in the introduction, we will be using data that has a total of 7 variables listed in the table below (Table 1), with their abbreviations that we will be using throughout the whole analysis.

Variable	Abbreviation
Total age-adjusted mortality	MORT
Mean annual precipitation	PRECIP
Median number of school years completed by person of age 25 or over	EDUC
Percentage of population in 1960 that is nonwhite	NONWHITE
Percentage of households with annual income under $3000 in 1960	POOR
Relative pollution potential of oxides of nitrogen	NOX
Relative pollution potential of Sulphur dioxide	SO2

Table 1: Variables

To get a sense of what the dataset looks like, we decided to plot a histogram of each variable. We discovered that the distributions of some variables are skewed according to the histograms (Figure 1), it is reasonable that we will perform transformation on the skewed variables.

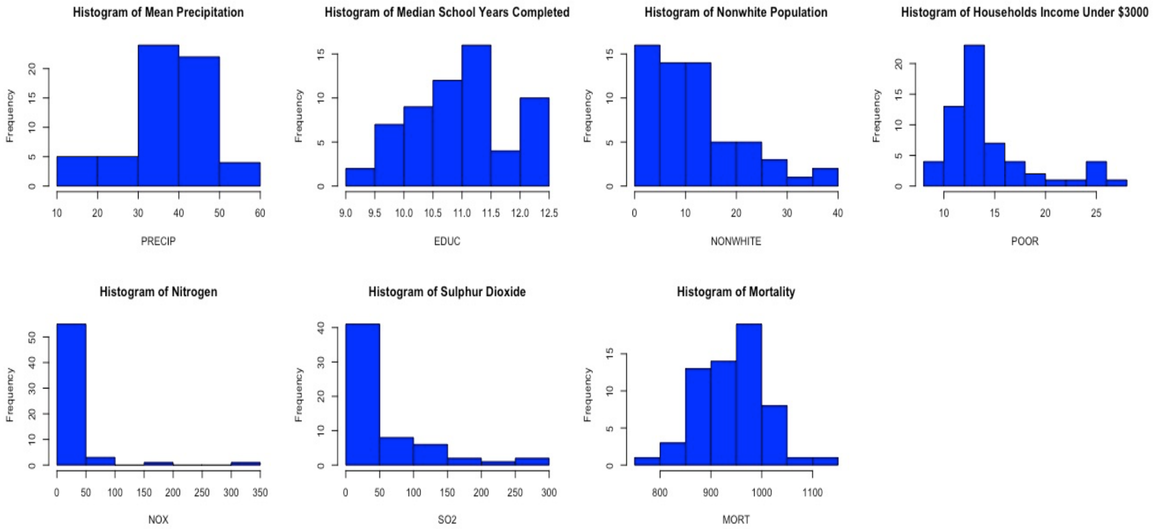
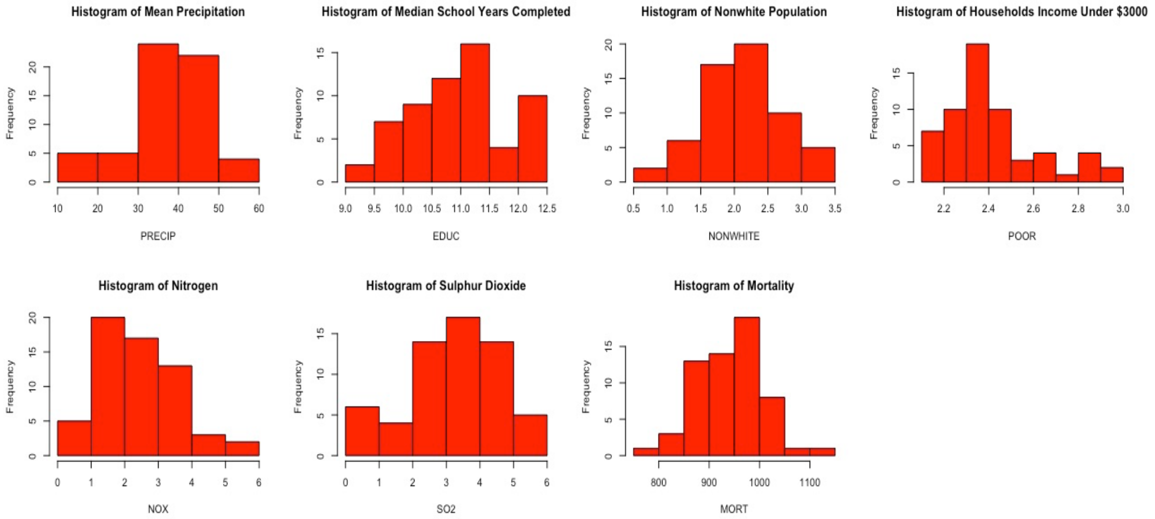


Figure 1: Histograms of Variables

From the histograms of variables, we can clearly observe that variable NOX, SO2, NONWHITE and POOR are highly skewed. Therefore, it is a good idea to transform them by using natural logarithm on NOX and SO2, and using cubic root on NONEWHITE and POOR. Figure below shows the distribution after the transformation. (Figure 2)



It is obvious that after the transformation, each histogram of variable is more symmetric. Therefore, we can now use the transformed data to continue our analysis.

Fitting the Regression

- Model: $Y = \beta0 + \beta X1 + \beta2X2 + \beta3X3 + \beta4X4 + \beta5X5 + \beta6X6 + \varepsilon$
- Notations:

  Y: Total age-adjusted mortality

  X1: Mean precipitation

  X2: Median school years completed

  X3: Cube root of nonwhite population

  X4: Cube root of households with income under $3000

  X5: Natural logarithm of oxides of nitrogen

  X6: Natural logarithm of Sulphur dioxide

Result and Analysis

- *Matrix*

From the matrix plot (Figure 3), we conclude that there is a positive relation between response variable Y[MORT] and predictor variable X1[PRECIP]. Besides, we can see a relatively strong relation not only between X3[NONEWHITE] and X4[POOR], but also between X5[NOX] and X6[SO2]. Comparing the correlation matrix that we obtained (Figure 4), the results from them are consistent with each other.

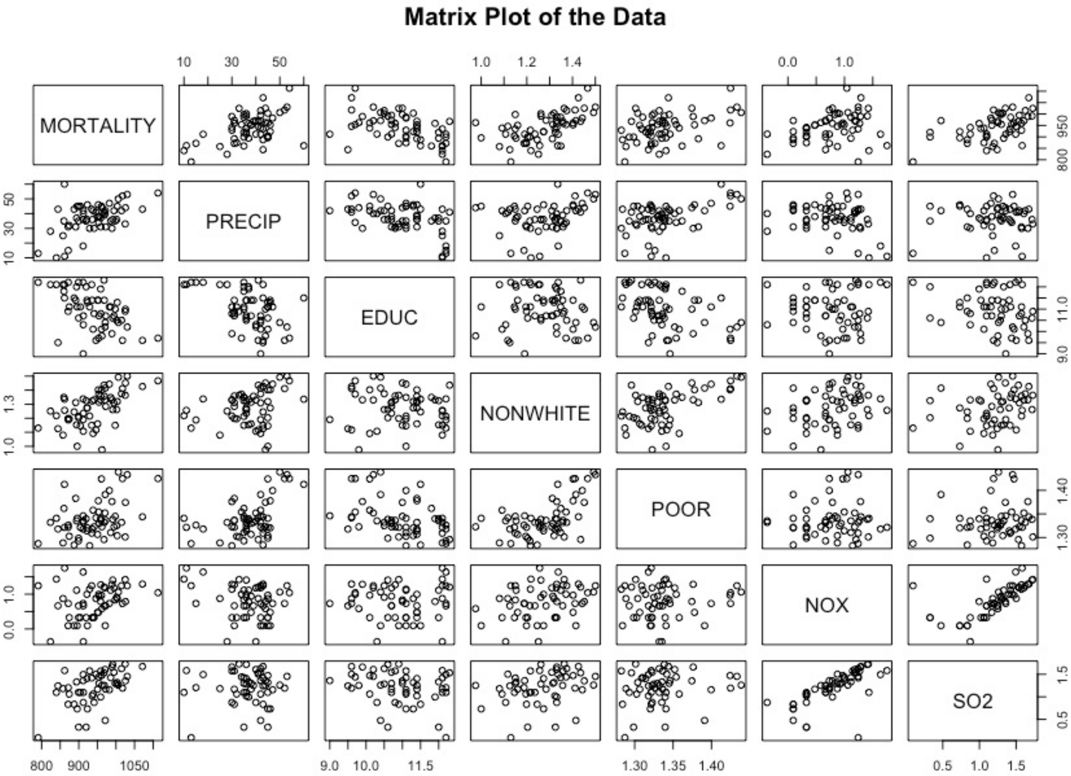
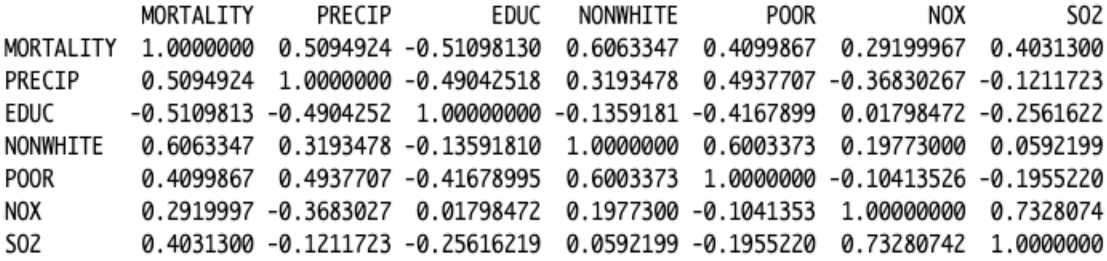


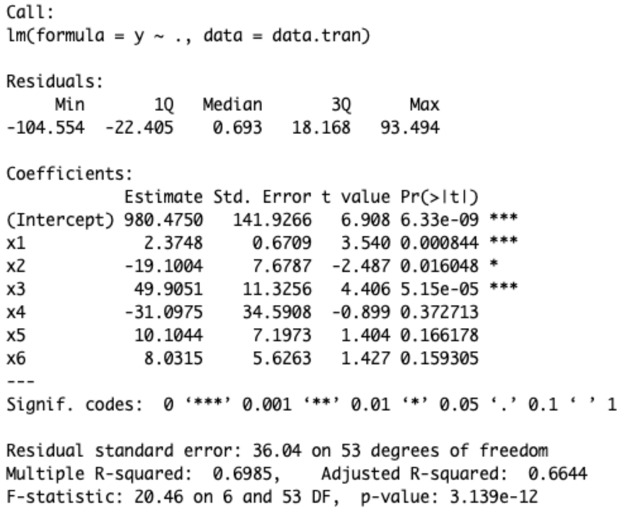
Figure 3. Matrix Plot



```
          MORTALITY     PRECIP        EDUC  NONWHITE       POOR         NOX        SO2
MORTALITY 1.0000000  0.5094924 -0.51098130 0.6063347  0.4099867  0.29199967  0.4031300
PRECIP    0.5094924  1.0000000 -0.49042518 0.3193478  0.4937707 -0.36830267 -0.1211723
EDUC     -0.5109813 -0.4904252  1.00000000 -0.1359181 -0.4167899  0.01798472 -0.2561622
NONWHITE  0.6063347  0.3193478 -0.13591810 1.0000000  0.6003373  0.19773000  0.0592199
POOR      0.4099867  0.4937707 -0.41678995 0.6003373  1.0000000 -0.10413526 -0.1955220
NOX       0.2919997 -0.3683027  0.01798472 0.1977300 -0.1041353  1.00000000  0.7328074
SO2       0.4031300 -0.1211723 -0.25616219 0.0592199 -0.1955220  0.73280742  1.0000000
```

Figure 4. Correlation Matrix

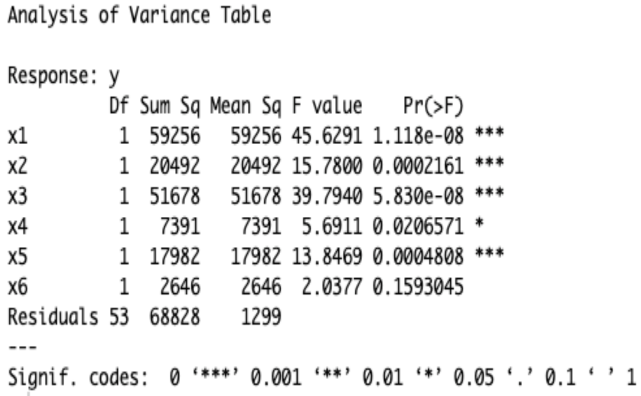
- *Summary of Regression Model*



```
Call:
lm(formula = y ~ ., data = data.tran)

Residuals:
     Min       1Q   Median       3Q      Max 
-104.554  -22.405    0.693   18.168   93.494 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 980.4750   141.9266   6.908 6.33e-09 ***
x1            2.3748     0.6709   3.540 0.000844 ***
x2          -19.1004     7.6787  -2.487 0.016048 *  
x3           49.9051    11.3256   4.406 5.15e-05 ***
x4          -31.0975    34.5908  -0.899 0.372713    
x5           10.1044     7.1973   1.404 0.166178    
x6            8.0315     5.6263   1.427 0.159305    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.04 on 53 degrees of freedom
Multiple R-squared:  0.6985,	Adjusted R-squared:  0.6644 
F-statistic: 20.46 on 6 and 53 DF,  p-value: 3.139e-12
```

-

- *ANOVA Table*



```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x1         1  59256   59256 45.6291 1.118e-08 ***
x2         1  20492   20492 15.7800 0.0002161 ***
x3         1  51678   51678 39.7940 5.830e-08 ***
x4         1   7391    7391  5.6911 0.0206571 *  
x5         1  17982   17982 13.8469 0.0004808 ***
x6         1   2646    2646  2.0377 0.1593045    
Residuals 53  68828    1299                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

-

- *Interpretation*

According to the information from summary of our model, our fitted model is

Y = 980.4750 + 2.3748X1 - 19.1004X2 + 49.9051X3 - 31.0975X4 + 10.1004X5 + 8.0315X6

Since we are testing if there is a multiple linear relationship between variables, we use ANOVA to conduct the F-test. From the summary, the p-value for the F-test is 3.139e-12. Because the p-value is small enough, we can conclude that there is a multiple linear relationship between response Y and our predictor variables.

In addition, R-squared in this model is 0.6985, indicating that 69.85% of the variability in Y can be explained by its regression on all predictor variables. However, since the p-values for X4 ,X5 and X6 are large, this model might not be the most appropriate one.

- *Diagnostics*

From the box-plot of the residuals (Figure 5), we can see there are two outliers which are far away from the median. There outliers are also shown obviously in the histogram of residuals and normal QQ-Plot (Figure 5). From the scatter plot of observed Y vs. Fitted Y, we can see that the fitted Y values are consistent with the observed Y values. On the normal QQ-Plot, points lie almost entirely on a straight line, despite outliers. For the residual plot against each independent variable, we do not discover any patterns in Figure 6.

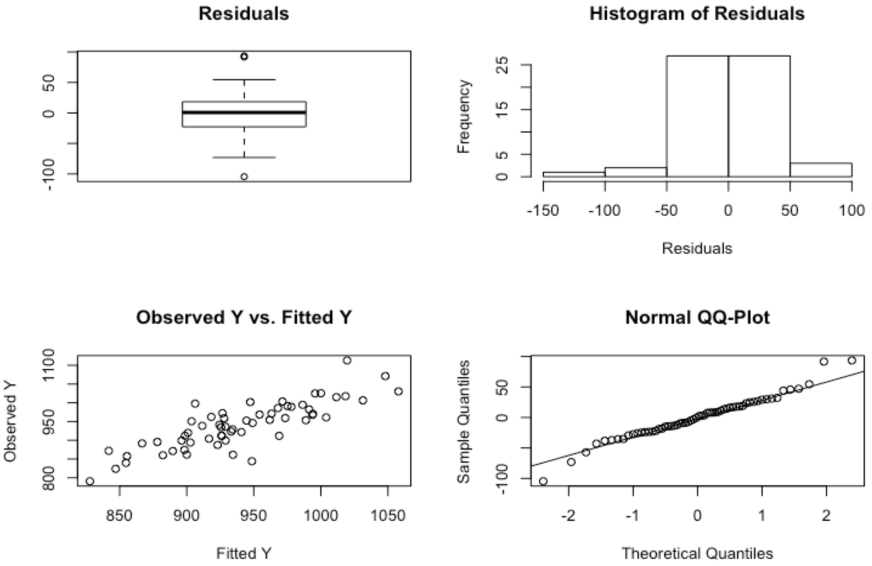


Figure 5. Diagnostics

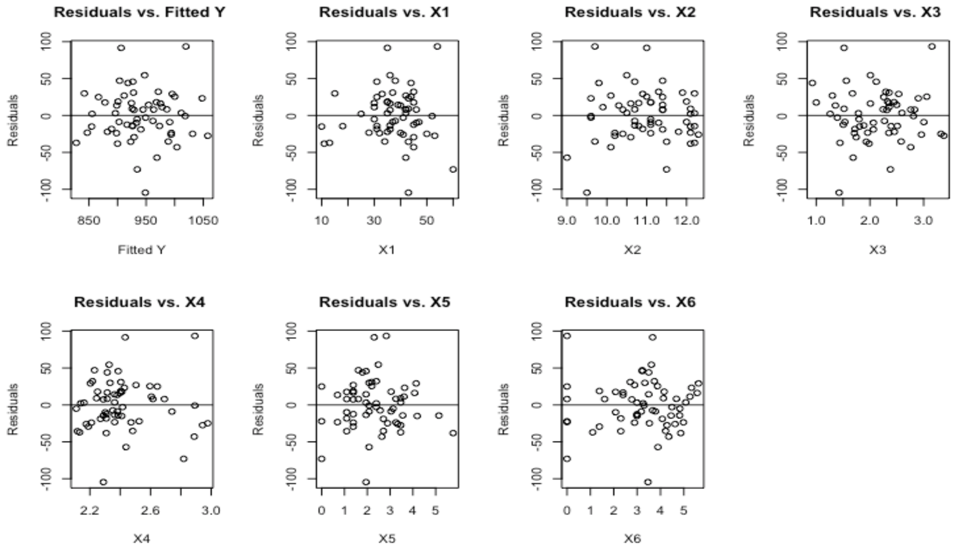


Figure 6. Diagnostics

Conclusion

From the previous results and analysis, we conclude that the fitted model is not the most appropriate for the data because the p-values for X4 and X6 are large. Moreover, although our diagnostics suggests that the majority of our model is reasonable, there are outliers in residuals.

Thus, regardless of whether there is nonlinearity within the data, we need to find a better fit for the data. And then, we will try to add non-linear term into the model.

Regression Transformation

In fitting the regression, we observed that there are outliers in the residuals, that can be discovered in the boxplot of residuals, histogram of residuals and normal QQ-Plot. Therefore, we seek to find a better fit for the data and include the nonlinear terms (such as squares), to obtain a better model.

Previously, we found that the p-values for X4 and X6 are large. We want to know which variable is nonlinear. In Figure 7 we can see that X4 shows an more obvious nonlinear relationship, therefore we will regress Y on X4.

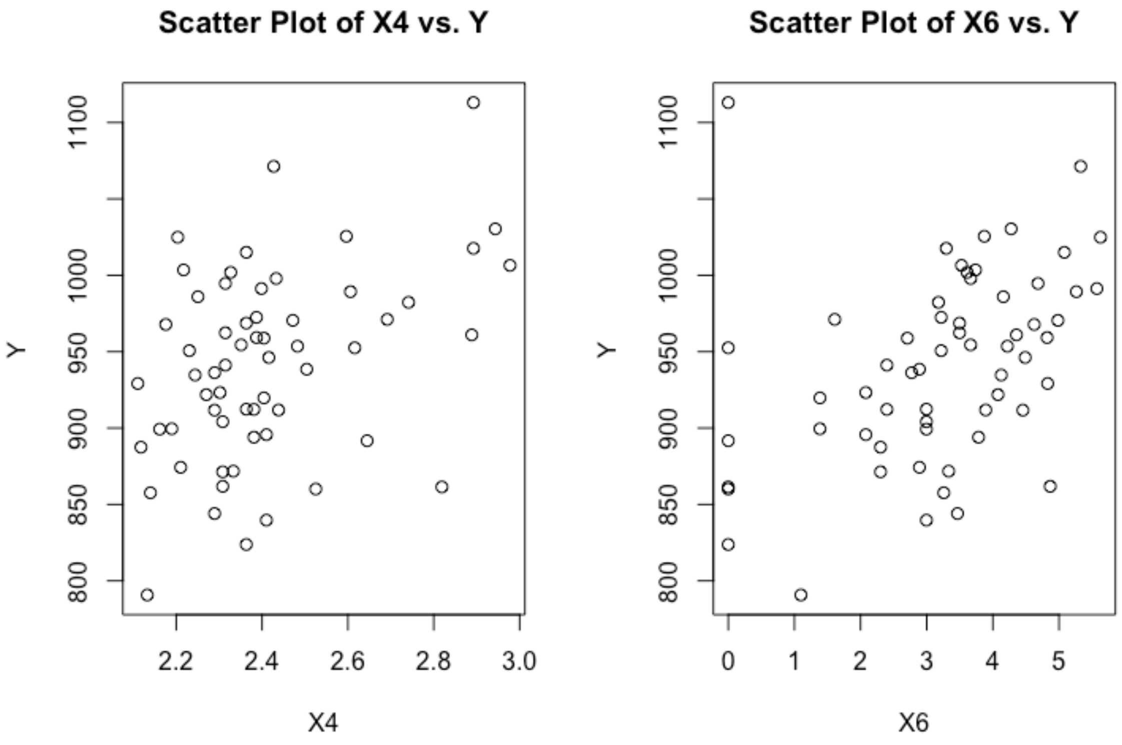


Figure 7. Scatter plots of X4 and X6

- *Model*: Y = β0 + β1x + β11x^2 + ε, x = X – X-hat
- *Notations*:

  Y: Total age-adjusted mortality

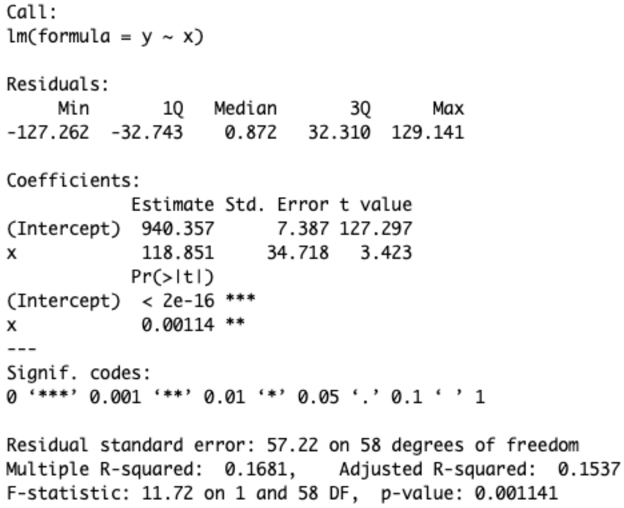
  x: *X – X-hat*

  X: *X4*

  X-hat: Mean of X4

Result and Analysis

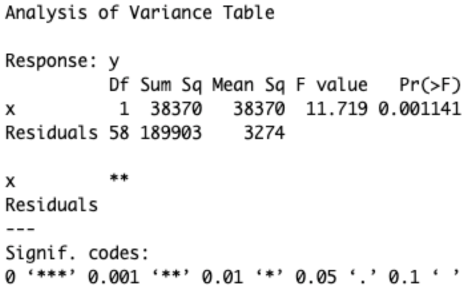
- *Summary of Model*



```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-127.262  -32.743    0.872   32.310  129.141 

Coefficients:
            Estimate Std. Error t value
(Intercept)  940.357      7.387 127.297
x            118.851     34.718   3.423
            Pr(>|t|)    
(Intercept)  < 2e-16 ***
x            0.00114 ** 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.22 on 58 degrees of freedom
Multiple R-squared:  0.1681,	Adjusted R-squared:  0.1537 
F-statistic: 11.72 on 1 and 58 DF,  p-value: 0.001141
```

-

- *ANOVA Table*



```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1  38370   38370  11.719 0.001141
Residuals 58 189903    3274                 
            
x         **
Residuals   
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

-

- *Interpretation*

We obtain ANOVA table to get a summary of F-test to see if there is multiple linear relationship between response variable Y and predictor variables. From the summary, we can get that p-value is 0.001141. It is small enough for us to conclude that the multiple linear relationship does exist. Besides, we can also get R-squared from the summary, which equals 0.1681, indicating that about 16.81% of the variability in Y can be explained by the regression on all predictor variables. However, the p-value of X2 is quite large. Hence, this model is not good enough for the data as well.

- *Diagnostics*

From boxplot of residuals in Figure 8, there is no more outliers in residuals, and it is symmetric about zero. The histogram of residuals is now symmetric and can be considered as normally distributed. Points in normal QQ-Plot are almost on a straight line. There is no obvious pattern on Figure 9, therefore, the model with nonlinear terms added is reasonable.

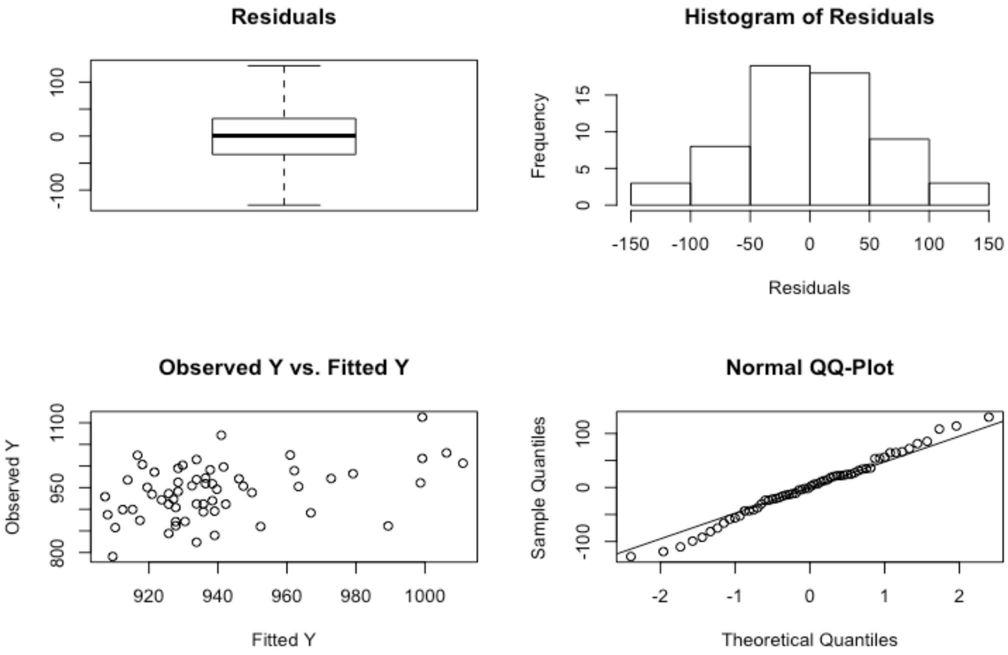


Figure 8. Diagnostics

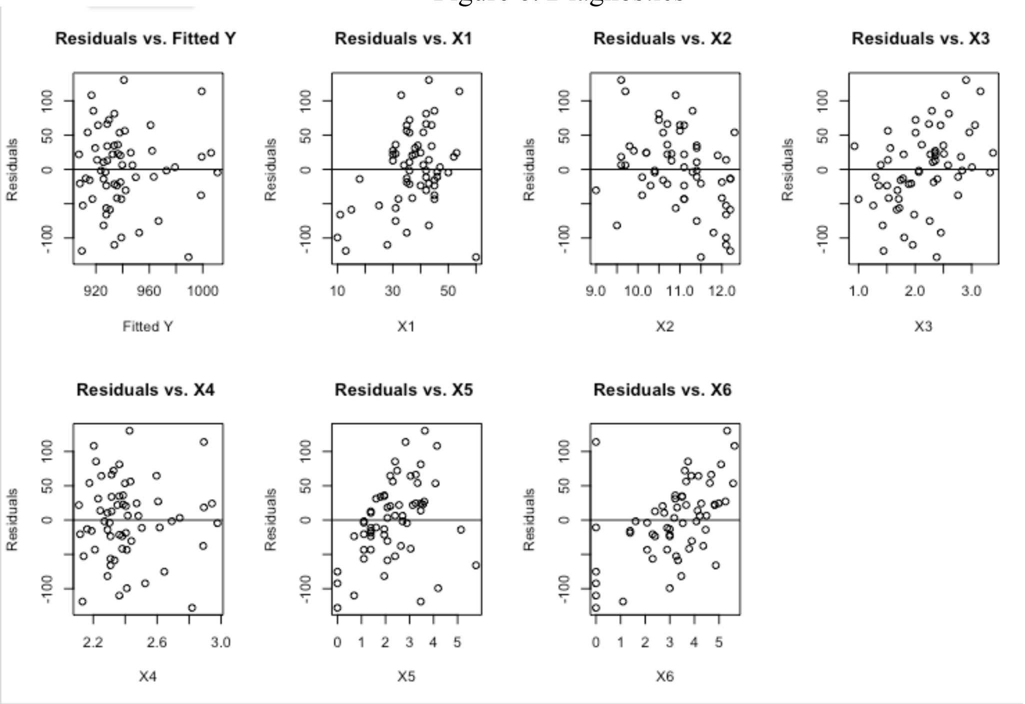


Figure 9. Diagnostics

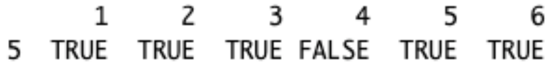
Conclusion

From the results and analysis that we obtained, we cannot conclude that the model is a better fit for the data since it has a lowest R-squared. However, with the small p-value and the diagnostics results showing that no outliers in the residuals, we can say that this model with nonlinear terms added is reasonable.

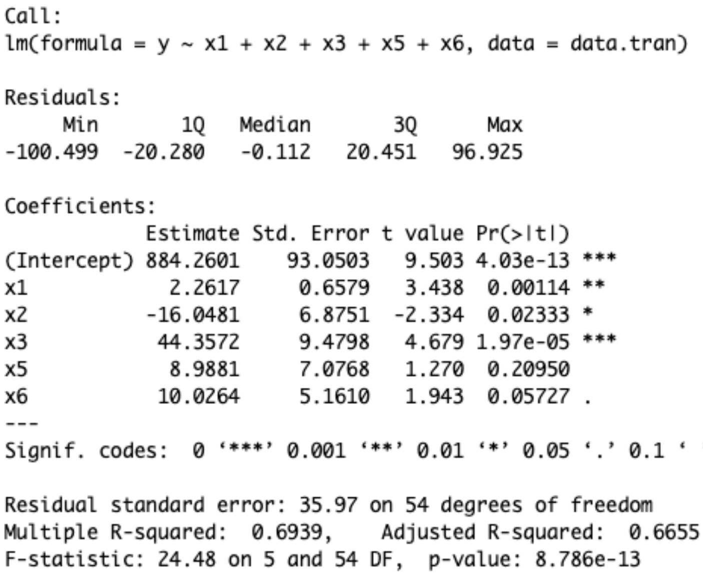
Model Selection

All Subset Regression

The method we are using is “method = ‘adjr2’” in R. The result shows that we should remove X4.



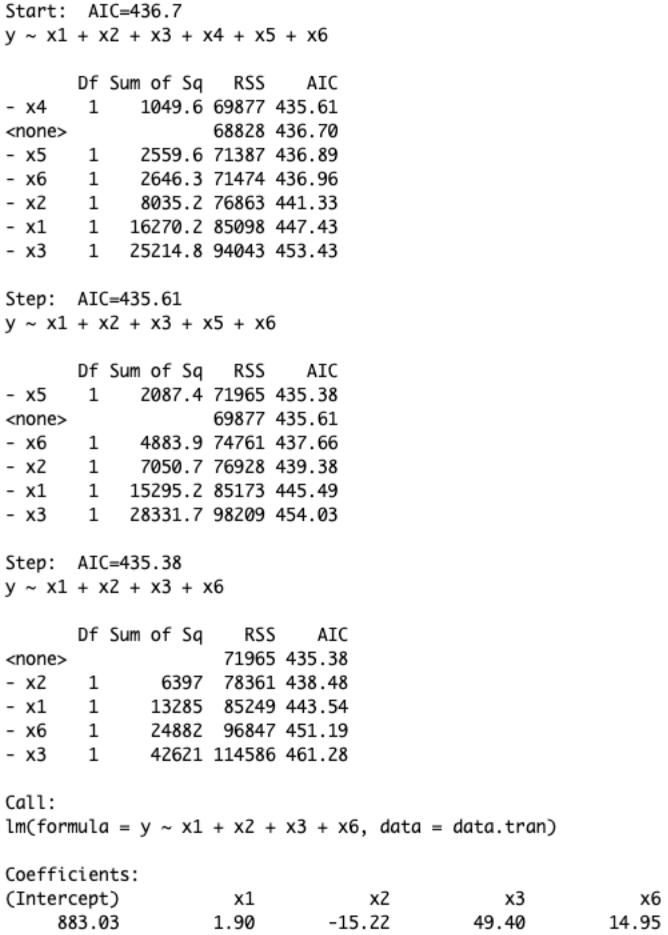
```
     1     2     3     4     5     6
5 TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```



```
Call:
lm(formula = y ~ x1 + x2 + x3 + x5 + x6, data = data.tran)

Residuals:
     Min       1Q   Median       3Q      Max 
-100.499  -20.280   -0.112   20.451   96.925 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 884.2601    93.0503   9.503 4.03e-13 ***
x1            2.2617     0.6579   3.438  0.00114 ** 
x2          -16.0481     6.8751  -2.334  0.02333 *  
x3           44.3572     9.4798   4.679 1.97e-05 ***
x5            8.9881     7.0768   1.270  0.20950    
x6           10.0264     5.1610   1.943  0.05727 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.97 on 54 degrees of freedom
Multiple R-squared:  0.6939,	Adjusted R-squared:  0.6655 
F-statistic: 24.48 on 5 and 54 DF,  p-value: 8.786e-13
```

We can see that the p-value is small and R-squared is relatively high. Thus, this is a good model to use.

Stepwise Method



```
Start:  AIC=436.7
y ~ x1 + x2 + x3 + x4 + x5 + x6

       Df Sum of Sq   RSS    AIC
- x4    1    1049.6 69877 435.61
<none>              68828 436.70
- x5    1    2559.6 71387 436.89
- x6    1    2646.3 71474 436.96
- x2    1    8035.2 76863 441.33
- x1    1   16270.2 85098 447.43
- x3    1   25214.8 94043 453.43

Step:  AIC=435.61
y ~ x1 + x2 + x3 + x5 + x6

       Df Sum of Sq   RSS    AIC
- x5    1    2087.4 71965 435.38
<none>              69877 435.61
- x6    1    4883.9 74761 437.66
- x2    1    7050.7 76928 439.38
- x1    1   15295.2 85173 445.49
- x3    1   28331.7 98209 454.03

Step:  AIC=435.38
y ~ x1 + x2 + x3 + x6

       Df Sum of Sq    RSS    AIC
<none>               71965 435.38
- x2    1      6397  78361 438.48
- x1    1     13285  85249 443.54
- x6    1     24882  96847 451.19
- x3    1     42621 114586 461.28

Call:
lm(formula = y ~ x1 + x2 + x3 + x6, data = data.tran)

Coefficients:
(Intercept)           x1           x2           x3           x6
     883.03         1.90       -15.22        49.40        14.95
```

From this method, the result shows us that we should remove X4 and hence leads to the conclusion that the stepwise method and all subset regression both suggest remove X4. The models that two methods generated are also very similar. Therefore, I will say that the efficiency of both models are similar.

Conclusion

Based on the analysis we conducted above, I think the best model is the first model we fitted because it has a larger value of R-squared and the p-value is a lot smaller than the other model. However, when we compare the residuals of the two models, the second model seems better by having a more normalized histogram of residuals, no outliers shown in the boxplot of residuals and points in normal QQ-Plot almost lie on the straight line.

In addition, from the variable deletion (model selection), the two methods both shows to remove variable X4, as it is also the nonlinear variable we found when we fit our model. The size of the dataset is not large enough here to obtain a more appropriate model. Finally, we find that there is a relationship between the mortality and the predictor variables in this dataset, but there is also other factors that affect the mortality.

Appendix

```
#Read Data
data = read.csv("Desktop/STA108/mortality.csv")
data

par(mfrow = c(2,4))
# Histogram of Predictors (Check which predictor is skewed)
hist(data$PRECIP, col = "blue", main = paste('Histogram of Mean Precipitation'), xlab =
paste('PRECIP'))
hist(data$EDUC, col = "blue", main = paste('Histogram of Median School Years
Completed'), xlab = paste('EDUC'))
hist(data$NONWHITE, col = "blue", main = paste('Histogram of Nonwhite Population'),
xlab = paste('NONWHITE'))
hist(data$POOR, col = "blue", main = paste('Histogram of Households Income Under
$3000'), xlab = paste('POOR'))
hist(data$NOX, col = "blue", main = paste('Histogram of Nitrogen'), xlab =
paste('NOX'))
hist(data$SO2, col = "blue", main = paste('Histogram of Sulphur Dioxide'), xlab =
paste('SO2'))
hist(data$MORTALITY, col = "blue", main = paste('Histogram of Mortality'), xlab =
paste('MORT'))

par(mfrow = c(2,4))
#Transformation (tranformed by natural algorithm and cube root)
hist(data$PRECIP, col = "red", main = paste('Histogram of Mean Precipitation'), xlab =
paste('PRECIP'))
hist(data$EDUC, col = "red", main = paste('Histogram of Median School Years
Completed'), xlab = paste('EDUC'))

data[3] = sign(data[3]) * abs(data[3])^(1/3)
hist(data$NONWHITE, col = "red", main = paste('Histogram of Nonwhite Population'),
xlab = paste('NONWHITE'))
data[4] = sign(data[4]) * abs(data[4])^(1/3)
hist(data$POOR, col = "red", main = paste('Histogram of Households Income Under
$3000'), xlab = paste('POOR'))

data[5] = log(data[5])
data[6] = log(data[6])
hist(data$NOX, col = "red", main = paste('Histogram of Nitrogen'), xlab = paste('NOX'))
hist(data$SO2, col = "red", main = paste('Histogram of Sulphur Dioxide'), xlab =
paste('SO2'))
hist(data$MORTALITY, col = "red", main = paste('Histogram of Mortality'), xlab =
paste('MORT'))
```

```
#2.
# Reorder the data to form model of regression: y=b0+b1x1...
data.tran = data
data.tran =
cbind(data.tran[7],data.tran[1],data.tran[2],data.tran[3],data.tran[4],data.tran[5],data.tran[
6])

plot(data.tran, main=paste('Matrix Plot of the Data')) # matrix plot of the data
cor(data.tran) # correlation matrix

# Fit the regression
names(data.tran) = c('y','x1','x2','x3','x4','x5','x6')
fit = lm(y ~ ., data = data.tran)
summary(fit)
anova(fit) # ANOVA

#3.
# plots
par(mfrow = c(2,2))
boxplot(fit$res, main = 'Residuals') #boxplot of residual
hist(fit$res, main = 'Histogram of Residuals', xlab = 'Residuals') #histogram of residual
plot(fit$fitted, data.tran$y, main = 'Observed Y vs. Fitted Y', xlab = 'Fitted Y', ylab =
'Observed Y') # observed y vs fitted y
qqnorm(fit$res, main = 'Normal QQ-Plot', xlab = 'Theoretical Quantiles', ylab = 'Sample
Quantiles') # normal plot
qqline(fit$res)

# plots of residuals vs independent variables
par(mfrow=c(2,4))
plot(fit$fitted, fit$res,main = 'Residuals vs. Fitted Y ', xlab = 'Fitted Y', ylab =
'Residuals')
abline(h=0)
plot(data.tran$x1, fit$res,main = 'Residuals vs. X1 ', xlab = 'X1', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x2, fit$res,main = 'Residuals vs. X2 ', xlab = 'X2', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x3, fit$res,main = 'Residuals vs. X3 ', xlab = 'X3', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x4, fit$res,main = 'Residuals vs. X4 ', xlab = 'X4', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x5, fit$res,main = 'Residuals vs. X5 ', xlab = 'X5', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x6, fit$res,main = 'Residuals vs. X6 ', xlab = 'X6', ylab = 'Residuals')
abline(h=0)
```

```
# 4.
par(mfrow = c(1,2))
plot(data.tran$x4,data.tran$y, main = 'Scatter Plot of X4 vs. Y', xlab = 'X4', ylab = 'Y') # Shows x4 is nonlinear
plot(data.tran$x6,data.tran$y, main = 'Scatter Plot of X6 vs. Y', xlab = 'X6', ylab = 'Y') # Shows x6 is nonlinear
# polynomial regression
y = data.tran$y
X = data.tran$x4
x = data.tran$x4 - mean(data.tran$x4)
x2 = x ^ 2
#fit/anova/estimations
fit2 = lm(y ~ x)
summary(fit2)
anova(fit2)

#plots
par(mfrow = c(2,2))
boxplot(fit2$res, main = 'Residuals')
hist(fit2$res, main = 'Histogram of Residuals', xlab = 'Residuals')
plot(fit2$fitted, data.tran$y, main = 'Observed Y vs. Fitted Y', xlab = 'Fitted Y', ylab = 'Observed Y')
qqnorm(fit2$res, main = 'Normal QQ-Plot', xlab = 'Theoretical Quantiles', ylab = 'Sample Quantiles')
qqline(fit2$res)

par(mfrow=c(2,4))
plot(fit2$fitted, fit2$res,main = 'Residuals vs. Fitted Y ', xlab = 'Fitted Y', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x1, fit2$res,main = 'Residuals vs. X1 ', xlab = 'X1', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x2, fit2$res,main = 'Residuals vs. X2 ', xlab = 'X2', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x3, fit2$res,main = 'Residuals vs. X3 ', xlab = 'X3', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x4, fit2$res,main = 'Residuals vs. X4 ', xlab = 'X4', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x5, fit2$res,main = 'Residuals vs. X5 ', xlab = 'X5', ylab = 'Residuals')
abline(h=0)
plot(data.tran$x6, fit2$res,main = 'Residuals vs. X6 ', xlab = 'X6', ylab = 'Residuals')
abline(h=0)

#5. Variable Deletion
```

```
# Step
step(lm(y~x1+x2+x3+x4+x5+x6, data = data.tran), ~1, direction = 'backward')
#Subset
library('leap')
install.packages('leaps')
fit3 = leaps(x=data.tran[,-1], y=data.tran[,1], method = 'adjr2')
ind = order(fit3$adjr2, decreasing = TRUE)
lm(y~x1+x2+x3+x5+x6, data = data.tran)
summary(lm(y~x1+x2+x3+x5+x6, data = data.tran))
```