THE CORRELATION BETWEEN POLLUTION AND MORTALITY

Hui Qi, Heming Ma, Yuan Zhang

University of California, Davis
STA 108

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

1.1 Background of study

1.2 Description of the data

1.3 Main objectives of the study

1.4 Main tools of the study

CHAPTER 2: ANALYSIS

2.1 Matrix Plot of Data

2.2 Correlation Matrix

2.3 Linear Regression function

2.4 ANOVA Table

2.5 Estimate Parameter and standard Error

2.6 Fixed Y value against Observed Y

2.7 Residuals against Fixed Y

2.8 Residuals against the Independent Variables

2.9 Histogram of Residuals

2.10 Normal Probability plot of Residuals

2.11 Regarding Question 4

2.12 Forward and Backward stepwise

CHAPTER 3: SUMMARY AND CONCLUSION

3.1 Report Summary

3.2 Appendix

REFERENCES

CHAPTER 1: INTRODUCTION

1.1 Background of study

Backed in 20th Century, some scientists and statisticians were curious about the correlation between pollution and mortality and would like to work in this problem. Therefore, they recorded 3 years' data from 1959 to 1961 on 60 Standard Metropolitan Statistical Area (SMSA) in the United states and they wish that they are able to figure out how these two variables related and are there any other factors affect mortality.

1.2 Description of the data

There are 60 records in total which includes 6 possible variables which would possibility affect this study result. The data for mean annual precipitation (in inches) has an average of **37** and the ranges are from **10** to **60**. The data for Median number of school years completed by persons of age **25** or over [EDUC] has a average of **11** and it ranges from **9** to **12**.The data for Percentage of population in 1960 that is nonwhite [NONWHITE] has an average about **11.87** and the collected data ranges from **0.8** to **38.5**. Percentage of households with annual income under $3000 in 1960 [POOR] has an average of **14** and the data ranges from **9.4** to **26.4**. Relative pollution potential of oxides of nitrogen **(NOX)** at an average of **22.65** and ranges from **1** to **319**. Relative pollution potential of sulphur dioxide **(SO2)** with an average of **53.76** and ranges from **1** to **278** among those 60 cities. After comparing the mean and the ranges, variables NO_X and SO_2 are the most problematic data that we need to pay more attention on.

1.3 Main objectives of the study

We want to build a model that can be used to find how the mortality related to all 6 possible variables, Mean annual precipitation (in inches) [PRECIP], Median number of school years completed by persons 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).

More importantly, we hope after this model is constructed, we are able to lower the percentage of mortality by changing the predictor variables.

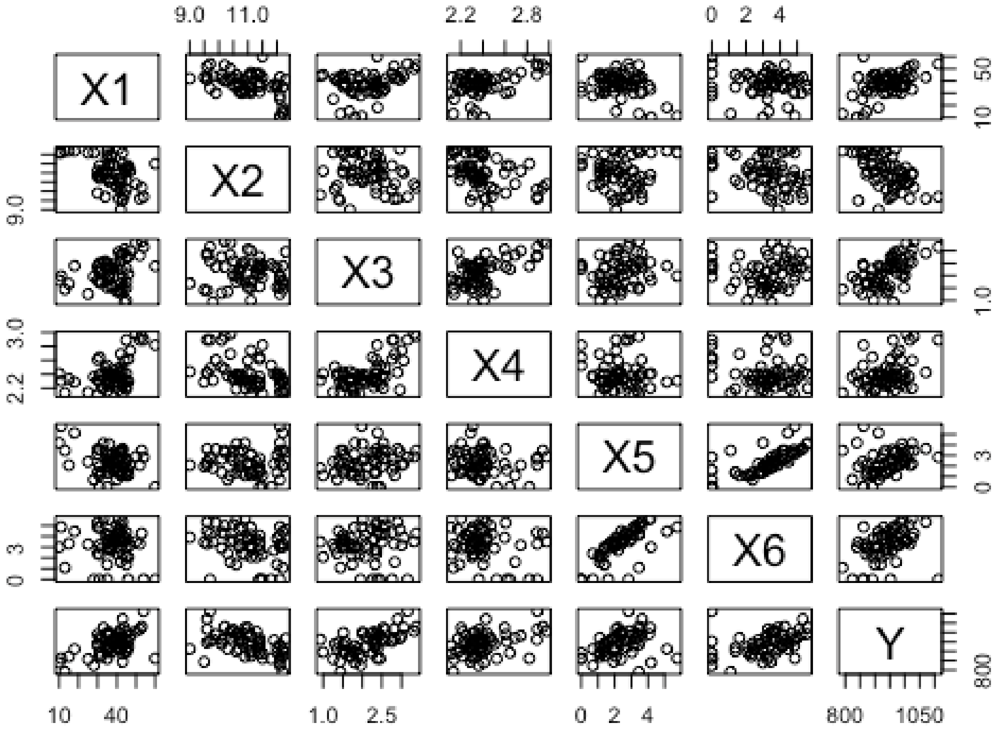
1.4 Main tools of the study

The main tool we used for building and analyzing model is R-Studio package.

CHAPTER 2: ANALYSIS

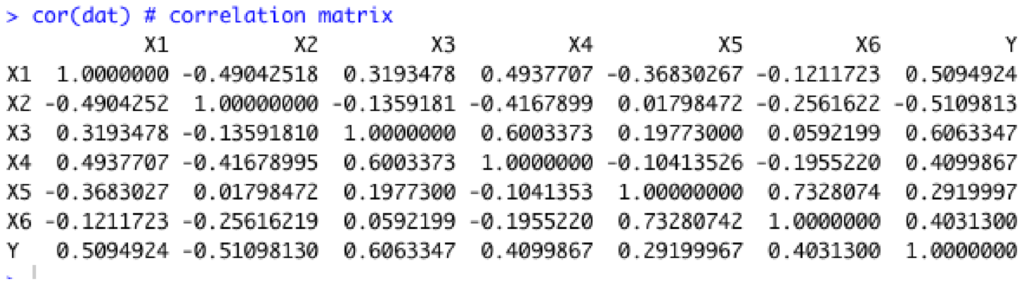
2.1 Matrix Plot of Data

This is a 7*7 Matrix plot which has 6 dependent variables and 1 independent variable. Matrix Plot gives visualization of the bivariate relationships between combinations of variables. In this matrix plot we generated, the stronger the correlation, the more linear and darker the plot is shown. After we constructed the Matrix plot, we need to be care of the multicollinearity. Multicollinearity refers to a situation in which two explanatory variables in a regression model are highly linearly related, which should be avoided. After observing 49 plots, x5 and x6 seem has the strongest multicollinearity.



2.2 Correlation Matrix

The correlation Matrix shows a qualitive aspect of bivariate relationships between combinations of variables. The closer the correlation to 1 or -1, the stronger the correlation two variables have. The following correlation Matrix, x3 and y have a 0.6063 correlation which has the highest correlation among all these dependent variables. The correlation Matrix also warn us about the strong multipolarity between x5 and x6 again, which we mentioned in the Matrix Plot section.

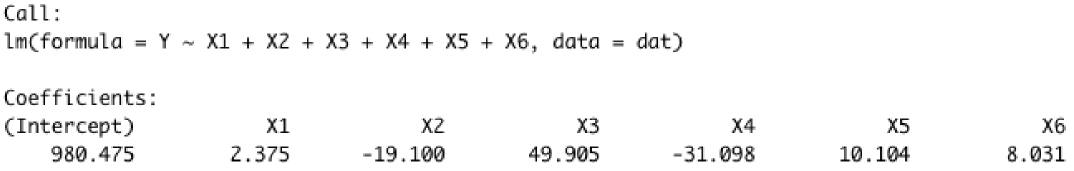


```
> cor(dat) # correlation matrix
           X1          X2         X3         X4          X5         X6          Y
X1  1.0000000 -0.49042518  0.3193478  0.4937707 -0.36830267 -0.1211723  0.5094924
X2 -0.4904252  1.00000000 -0.1359181 -0.4167899  0.01798472 -0.2561622 -0.5109813
X3  0.3193478 -0.13591810  1.0000000  0.6003373  0.19773000  0.0592199  0.6063347
X4  0.4937707 -0.41678995  0.6003373  1.0000000 -0.10413526 -0.1955220  0.4099867
X5 -0.3683027  0.01798472  0.1977300 -0.1041353  1.00000000  0.7328074  0.2919997
X6 -0.1211723 -0.25616219  0.0592199 -0.1955220  0.73280742  1.0000000  0.4031300
Y   0.5094924 -0.51098130  0.6063347  0.4099867  0.29199967  0.4031300  1.0000000
```

]

2.3 Linear Regression function

Function: y hat = 980.475+ 2.375x1-19.1x2+49.905x3-31.098x4+10.104x5+8.031x6

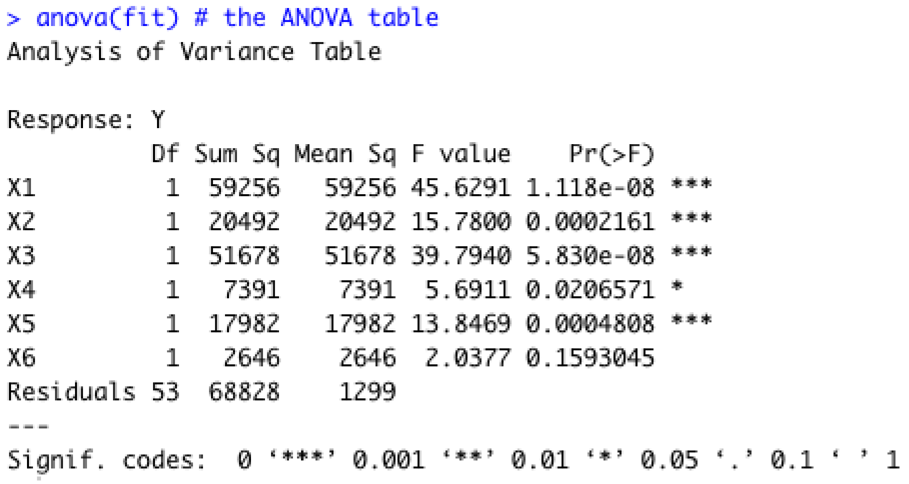


```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = dat)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5           X6  
    980.475        2.375      -19.100       49.905      -31.098       10.104        8.031  
```

2.4 ANOVA Table

Analysis of Variance (ANOVA) is a statistical analysis to test the degree of differences between two or more groups of an experiment. ANOVA table displays the statistics that used to test hypotheses about the population means which includes degree of freedom of the data, the sum of the squares of the data, mean sum of the squares of the data, mean sum of the squares of the data, F-statistic and p-value. From the table, we find that x4 and x6 have relatively bigger p-value, which imply that they might be dropped.

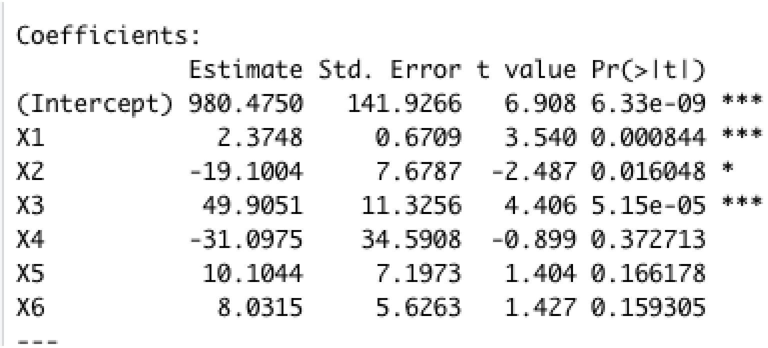


```
> anova(fit) # the ANOVA table
Analysis of Variance Table

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
```

2.5 Estimate Parameter and standard Error

There are large standard Error for b0 and b4.

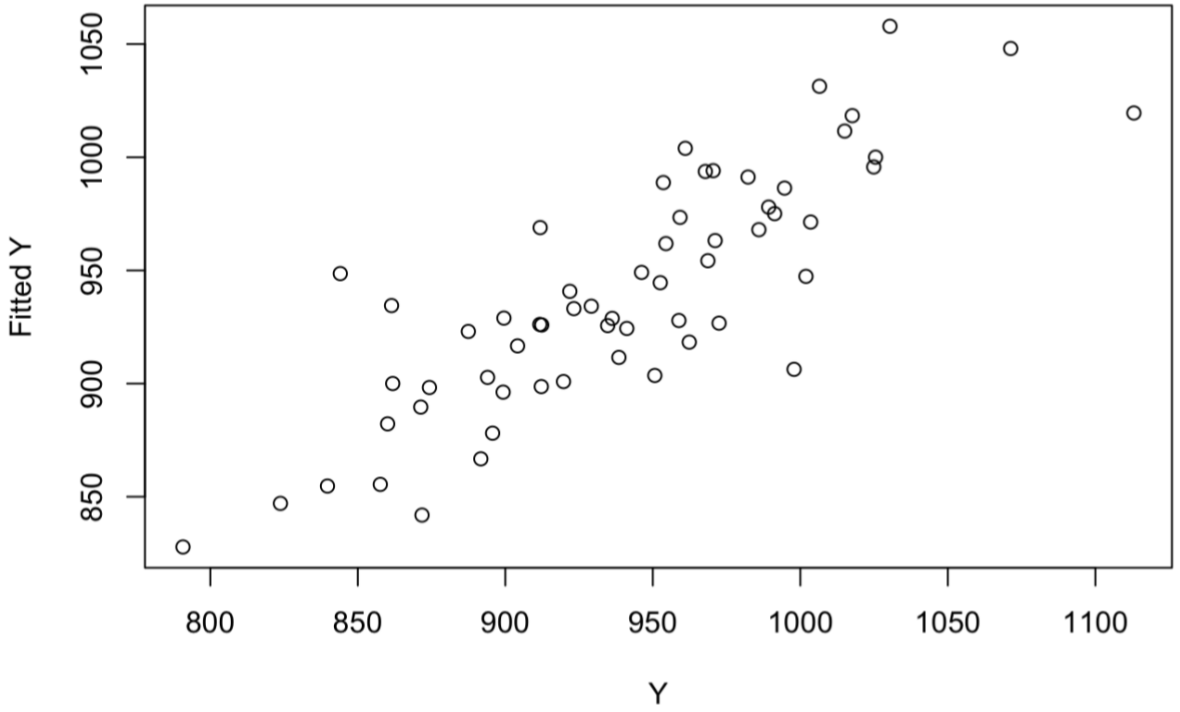


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 980.4750   141.9266   6.908 6.33e-09 ***
X1            2.3748     0.6709   3.540 0.000844 ***
X2          -19.1004     7.6787  -2.487 0.016048 *  
X3           49.9051    11.3256   4.406 5.15e-05 ***
X4          -31.0975    34.5908  -0.899 0.372713    
X5           10.1044     7.1973   1.404 0.166178    
X6            8.0315     5.6263   1.427 0.159305    
---
```

b0 = 980.4750   s(b0) = 141.9266
b1 = 2.3748   s(b1) = 0.6709
b2 = -19.1004   s(b2) = 7.6787
b3 = 49.9051   s(b3) = 11.3256
b4 = -31.0975   s(b4) = 34.5908
b5 = 10.1004   s(b5) = 7.1973
b6 = 8.0315   s(b6) = 5.6263

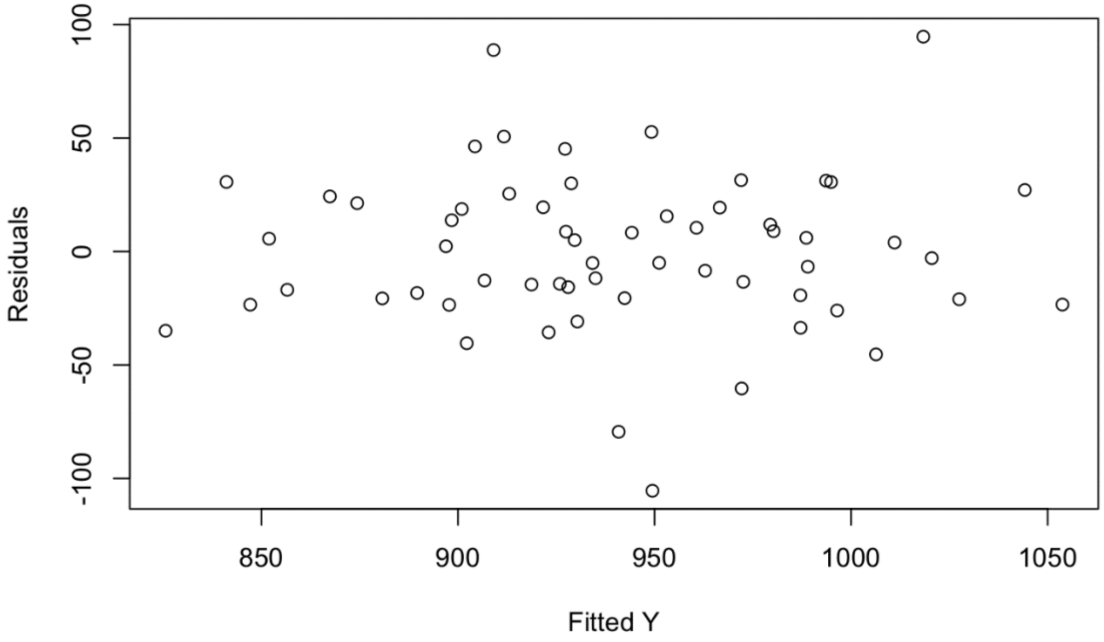
2.6 Fitted Y value against Observed Y

From this plot of Fitted Y against observed Y, it seems that the regression function has a trend of linearity and with equally variance. Although there are outliers but the amount of them are acceptable. Therefore, the outliers would not be a big problem.



2.7 Residuals against Fixed Y

This is a plot about Residuals against Fitted Y. It seems like the points evenly distributed randomly, which do not show specific patterns. From this plot, it seems that the regression function has an equal variance. There are still several outliers exist in this graph. Even though the outliers do not directly affect the picture of residuals against fixed Y, we still want to minimize the possibility of having outliers. There are two main reasons behind outliers. First reason is that the person who entered this data made measurement errors. The second reason is that there are data that are very different to most of the other data because of the regional diffference.



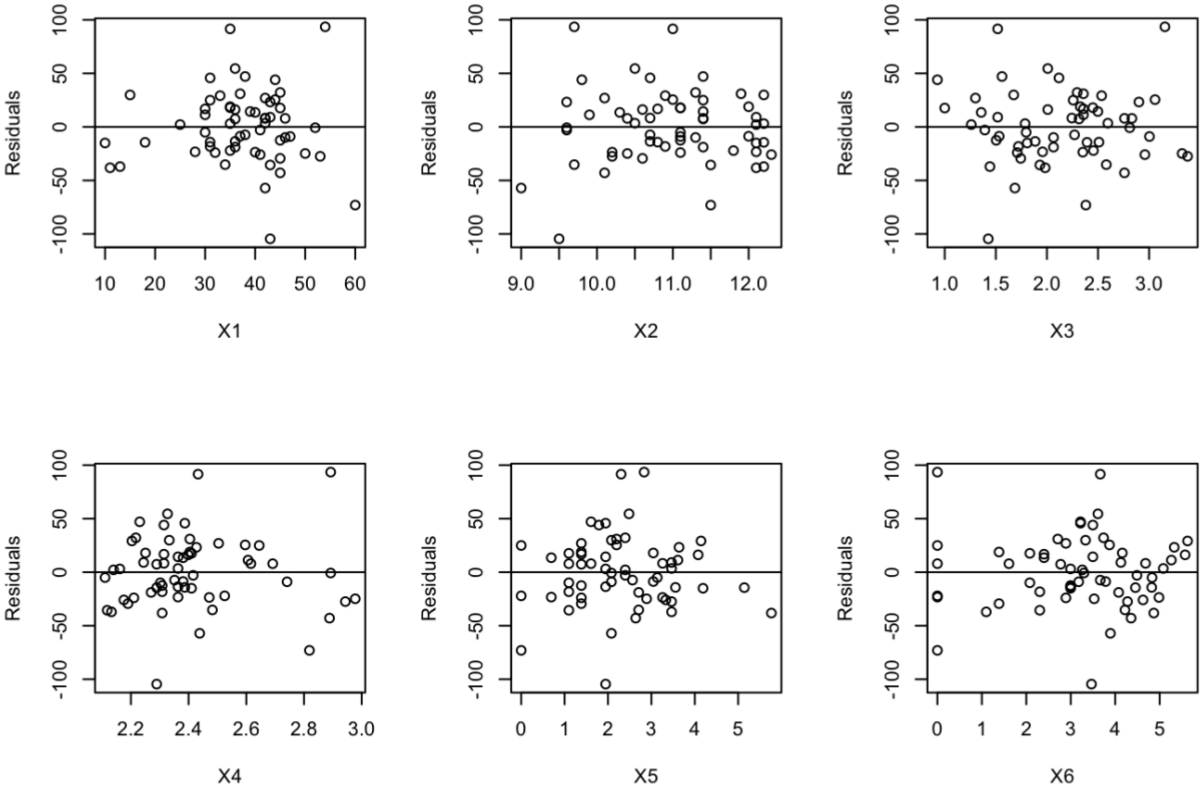
2.8 Residuals against the Independent Variables

Outlier: x1, x2, x3, x4, x5, x6

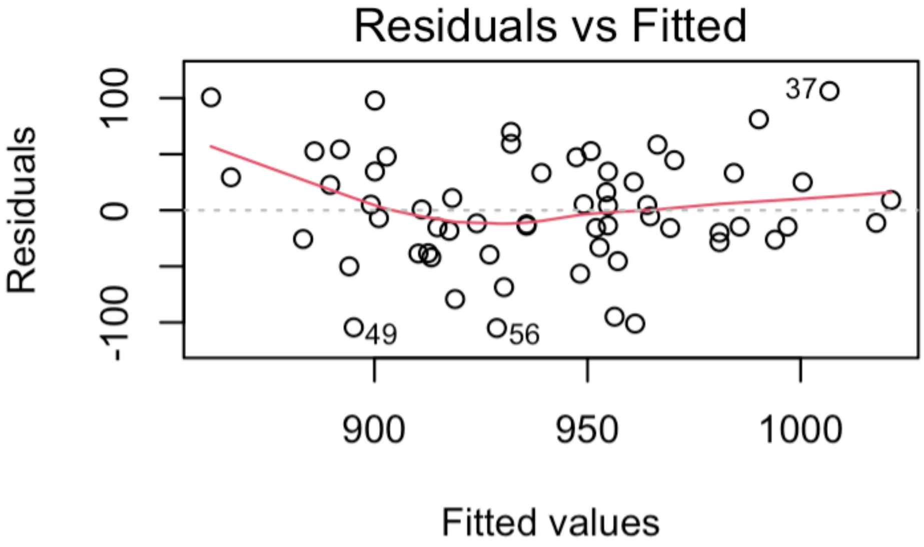
From the observation, every plot has residual points which located around 100 and -100. These points are called the outliers.

Linearity: x1, x2, x3, x4, x5, x6

For x2, x3, x4, x5, and x6, the points in the plot seem random and do not have any pattern. Thus, it implies linearity.



For x1, the point is ambiguous, so we find the linear regression function of the graph of model Y = b0 +b1x1 and plot its residuals against fitted value plot. The plot indicates linearity. The graph is below.



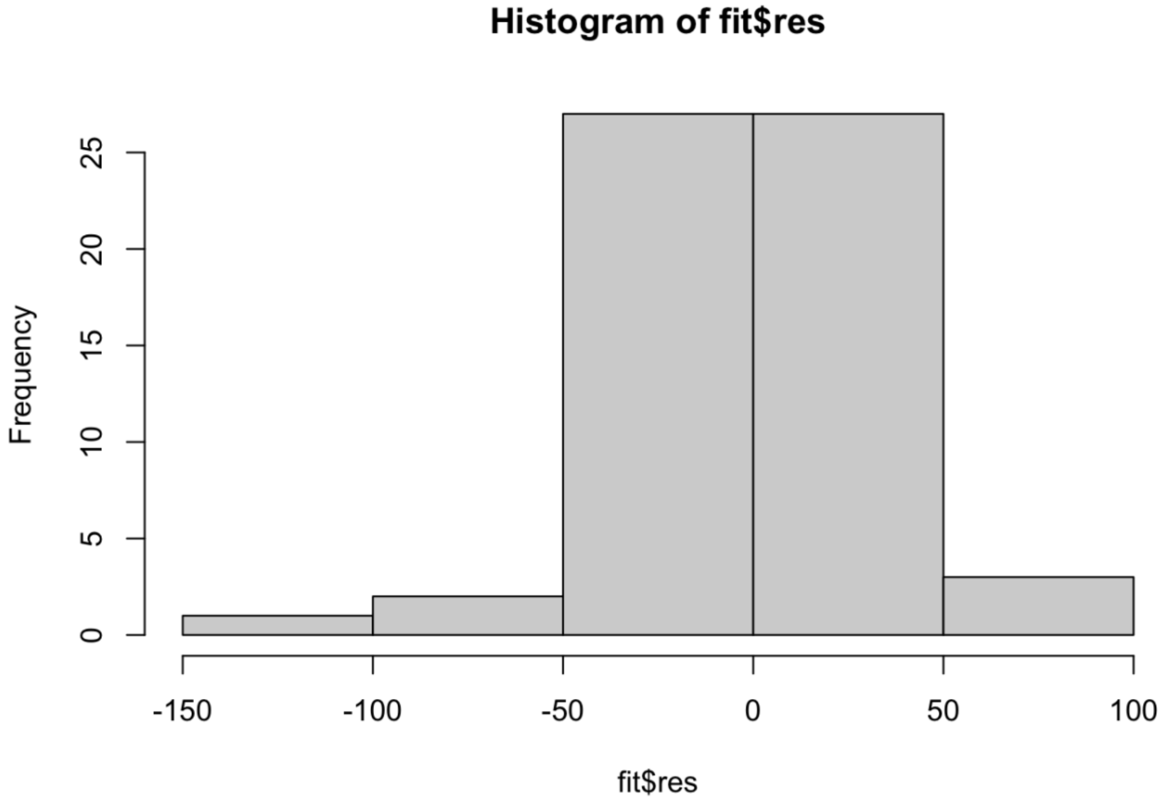
Nonconstant variance: x1, x4

For x4, when x4 is between 2.8 and 3.0, the only positive residual is almost 100, which is an outlier. But there are 5 negative residuals. Thus, the error variance is not constant.

For x1, the error variance is relatively small when x1 is between 10 and 30, but when x1 is between 30 and 60, the error variance is large.

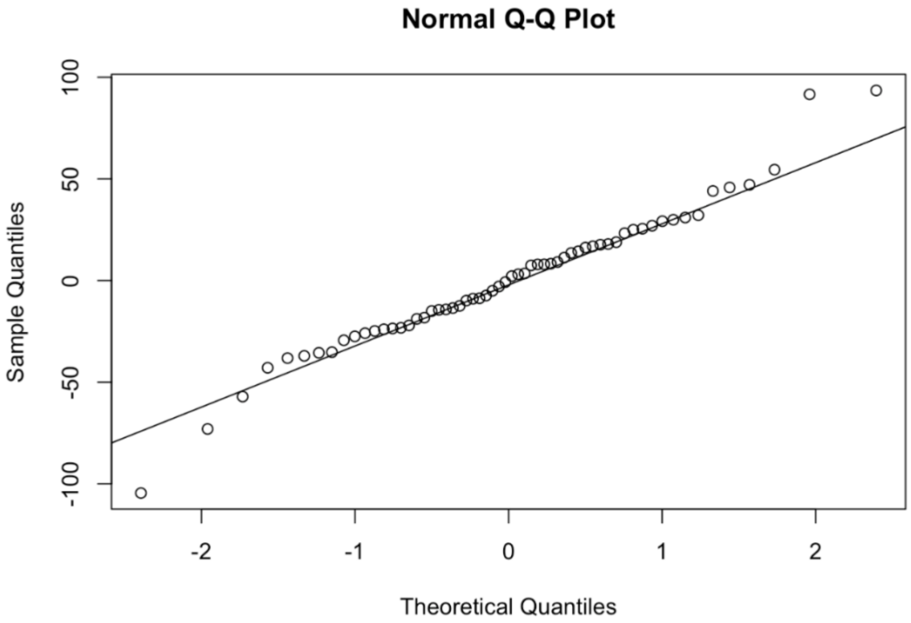
2.9 Histogram of Residuals

It seems that there are outliers that have residual value between -150 and -100. The plot seems symmetric and normal. The normality may result from the large sample size.



2.10 Normal Probability plot of Residuals

Although there are some outliers, most points fall on the line. Therefore, the error is a normal distribution. The normality may result from the large sample size.



2.11 Regarding Question 4

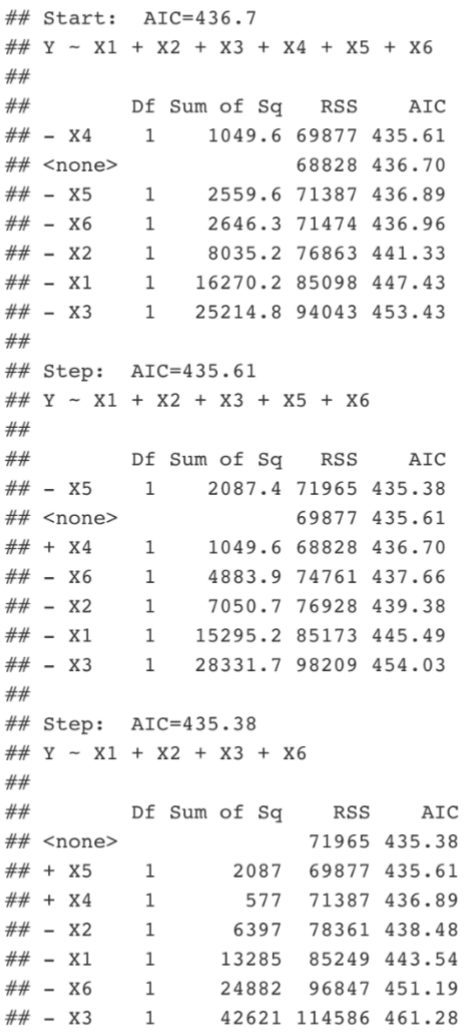
We believe that there is no nonlinearity in the data from our analysis in steps 2 and 3. After graphing the linear regression function, we find that the plot is a good fit since most points fall on the line.

2.12 Forward and Backward stepwise Regression

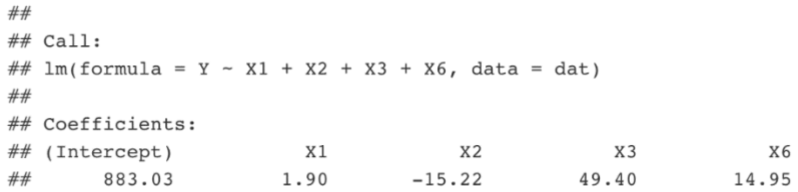
From the correlation matrix in problem 2, we find that the correlation coefficient between X5 and X6 is 0.7328, which is quite large. So, we guess that one of X5 and X6 may be dropped from the linear regression model though the plots in question 3 suggest that X5 against fitted Y and X6 against fitted Y are linear. Therefore, we use both backward and forward stepwise to find which variable we should drop. The backward and forward stepwise procedures for model selection give us the same model.

Backward stepwise: The backward stepwise give the model with

y=883.03+1.9 x1-15.22 x2 +49.40 x3 +14.95 x6

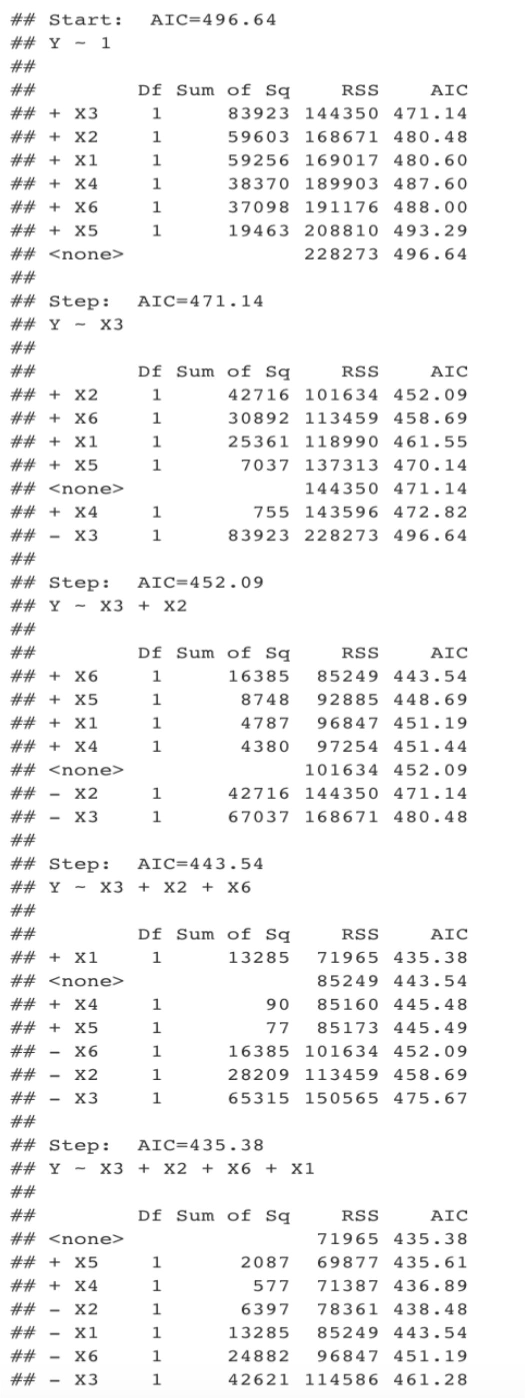


```
## Start:  AIC=436.7
## Y ~ X1 + X2 + X3 + X4 + X5 + X6
##
##        Df Sum of Sq   RSS    AIC
## - X4    1    1049.6 69877 435.61
## <none>               68828 436.70
## - X5    1    2559.6 71387 436.89
## - X6    1    2646.3 71474 436.96
## - X2    1    8035.2 76863 441.33
## - X1    1   16270.2 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
## + X4    1    1049.6 68828 436.70
## - 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
## + X5    1      2087  69877 435.61
## + X4    1       577  71387 436.89
## - 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 = dat)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X6
##      883.03         1.90       -15.22        49.40        14.95
```

Forward stepwise: Forward stepwise also give us the model with y=883.03+1.9 x1-15.22 x2 +49.40 x3 +14.95 x6



```
## Start:  AIC=496.64
## Y ~ 1
##
##         Df Sum of Sq    RSS    AIC
## + X3    1     83923 144350 471.14
## + X2    1     59603 168671 480.48
## + X1    1     59256 169017 480.60
## + X4    1     38370 189903 487.60
## + X6    1     37098 191176 488.00
## + X5    1     19463 208810 493.29
## <none>                228273 496.64
##
## Step:  AIC=471.14
## Y ~ X3
##
##         Df Sum of Sq    RSS    AIC
## + X2    1     42716 101634 452.09
## + X6    1     30892 113459 458.69
## + X1    1     25361 118990 461.55
## + X5    1      7037 137313 470.14
## <none>                144350 471.14
## + X4    1       755 143596 472.82
## - X3    1     83923 228273 496.64
##
## Step:  AIC=452.09
## Y ~ X3 + X2
##
##         Df Sum of Sq    RSS    AIC
## + X6    1     16385  85249 443.54
## + X5    1      8748  92885 448.69
## + X1    1      4787  96847 451.19
## + X4    1      4380  97254 451.44
## <none>                101634 452.09
## - X2    1     42716 144350 471.14
## - X3    1     67037 168671 480.48
##
## Step:  AIC=443.54
## Y ~ X3 + X2 + X6
##
##         Df Sum of Sq    RSS    AIC
## + X1    1     13285  71965 435.38
## <none>                 85249 443.54
## + X4    1        90  85160 445.48
## + X5    1        77  85173 445.49
## - X6    1     16385 101634 452.09
## - X2    1     28209 113459 458.69
## - X3    1     65315 150565 475.67
##
## Step:  AIC=435.38
## Y ~ X3 + X2 + X6 + X1
##
##         Df Sum of Sq    RSS    AIC
## <none>                 71965 435.38
## + X5    1      2087  69877 435.61
## + X4    1       577  71387 436.89
## - 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 ~ X3 + X2 + X6 + X1, data = dat)
##
## Coefficients:
## (Intercept)           X3           X2           X6           X1
##      883.03        49.40       -15.22        14.95         1.90
```

CHAPTER 3: SUMMARY AND CONCLUSION

In the beginning, we transform x5 and x6 by using the natural logarithm since SO2 and NOx are skewed, so it will be a good idea to transform it. Similarly, NONWHITE and POOR are skewed, so we transform x3 and x4 by cube root.

Through diagnosis analysis and the graphics, we determined that all independent variables fit Y linearly. Therefore, the analysis indicates us that mortality related to all the predictor variables and their relation is linear. With the model we built, we could provide some suggestions to the government about how to lower the mortality by changing some variables. The model also demonstrates that mortality is related to pollution. The pollution variables in this case are oxides of nitrogen (NOX) and Sulphur dioxide (SO2). Those two pollution variables have strong correlation with each other; thus, we make a decision on dropping one of them. Otherwise the result of this study is affected negatively. By conducting backward stepwise regression and forward stepwise regression, these two regressions give the same model which is y hat =883.03+1.9 x1-15.22 x2 +49.40 x3 +14.95 x6

All the variables x1, x3 and x6 are relate to y positively, except for x2, which is negatively related to y. Positive correlation means that two variables in which both variables move in the same direction. In contract, negative correlation means that two variables move in the opposite direction. Besides, x2, x3 and x4 have relatively greater influence on y.

We know that mortality is positively related to the SO2. Therefore, we could suggest government to restrict the emission of SO2 from factories or increase the taxes for pollution to lower the mortality. The model also indicates that if the median number of school years completed by persons of age 25 or over is larger, the mortality will decrease. As a result, government could revise policies to encourage people to go to schools, like increasing economic subsidies for education.

APPENDIX

Step 2: have a glimpse of the data

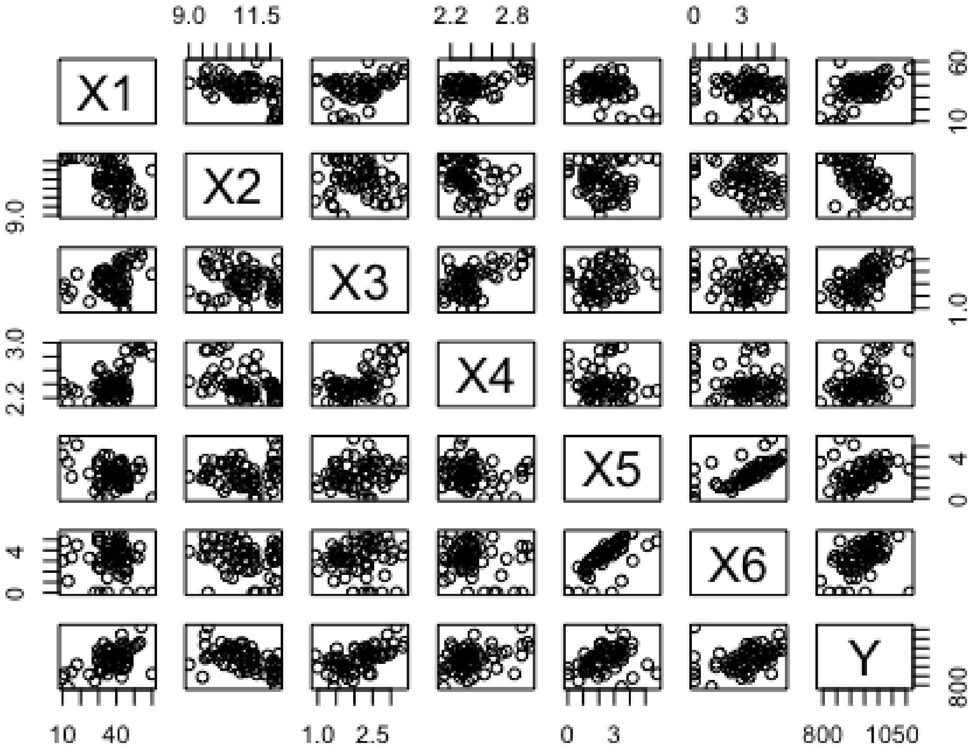
```r
dat <- read.csv("mortality.csv")
dat <- dat[1:7]
names(dat) <- c("X1", "X2", "X3", "X4", "X5", "X6", "Y")
library(kader)

## Warning: package 'kader' was built under R version 4.0.2

dat$X3 <- kader:::cuberoot(dat$X3)
dat$X4 <- kader:::cuberoot(dat$X4)
dat$X5 <- log(dat$X5)
dat$X6 <- log(dat$X6)
head(dat) # have a brief look

##   X1   X2        X3       X4       X5       X6        Y
## 1 36 11.4 2.0645602 2.270189 2.708050 4.077537  921.870
## 2 35 11.0 1.5182945 2.432881 2.302585 3.663562  997.875
## 3 44  9.8 0.9283178 2.314589 1.791759 3.496508  962.354
## 4 47 11.1 3.0036991 2.741295 2.079442 3.178054  982.291
## 5 43  9.6 2.9004359 2.427236 3.637586 5.327876 1071.289
## 6 53 10.2 3.3766567 2.943383 3.465736 4.276666 1030.380

plot(dat) # matrix plot of the data
```



```
cor(dat) # correlation matrix

##             X1          X2         X3         X4          X5         X6
## X1   1.0000000 -0.49042518  0.3193478  0.4937707 -0.36830267 -0.1211723
## X2  -0.4904252  1.00000000 -0.1359181 -0.4167899  0.01798472 -0.2561622
## X3   0.3193478 -0.13591810  1.0000000  0.6003373  0.19773000  0.0592199
## X4   0.4937707 -0.41678995  0.6003373  1.0000000 -0.10413526 -0.1955220
## X5  -0.3683027  0.01798472  0.1977300 -0.1041353  1.00000000  0.7328074
## X6  -0.1211723 -0.25616219  0.0592199 -0.1955220  0.73280742  1.0000000
## Y    0.5094924 -0.51098130  0.6063347  0.4099867  0.29199967  0.4031300
##              Y
## X1   0.5094924
## X2  -0.5109813
## X3   0.6063347
## X4   0.4099867
## X5   0.2919997
## X6   0.4031300
## Y    1.0000000

library(kader)
fit <- lm(Y~X1+X2+X3+X4+X5+X6,data = dat); fit

##
## Call:
```

```
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = dat)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
X5
##     980.475        2.375      -19.100       49.905      -31.098       10.1
04
##          X6
##       8.031
```

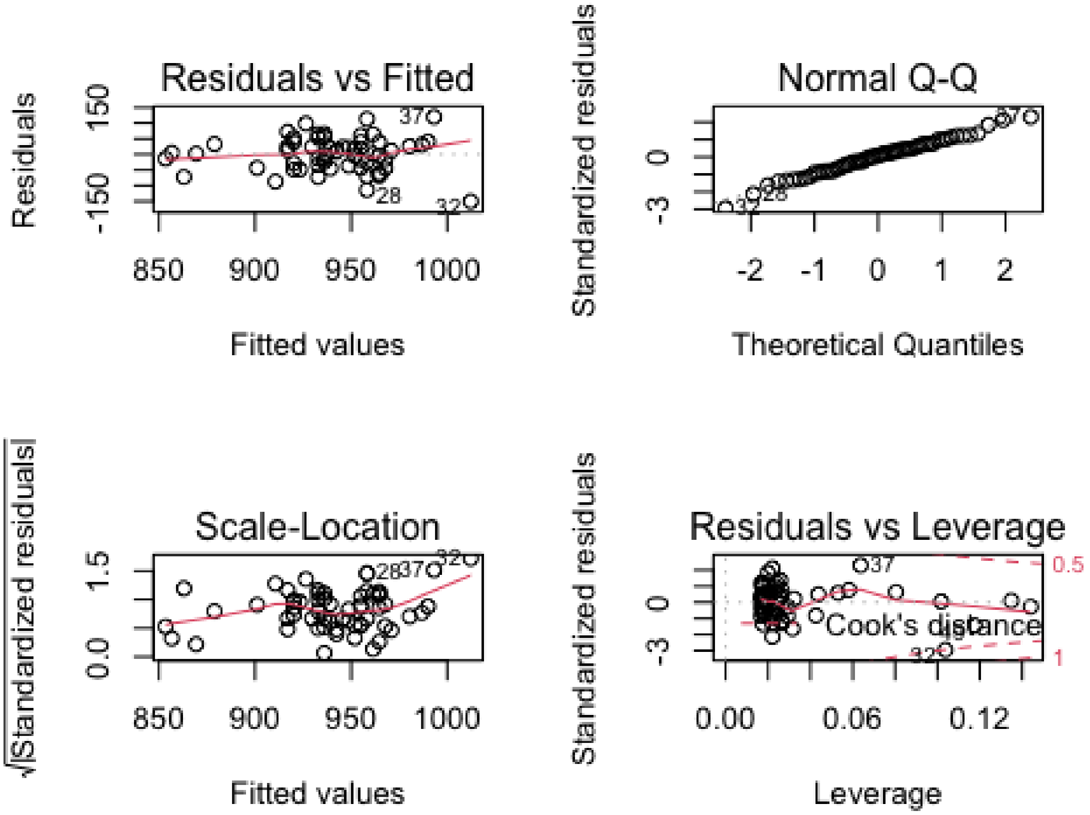
```
anova(fit) # the ANOVA table
```

```
## 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
```

```
summary(fit) # show the standard errors
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = dat)
##
## 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
```

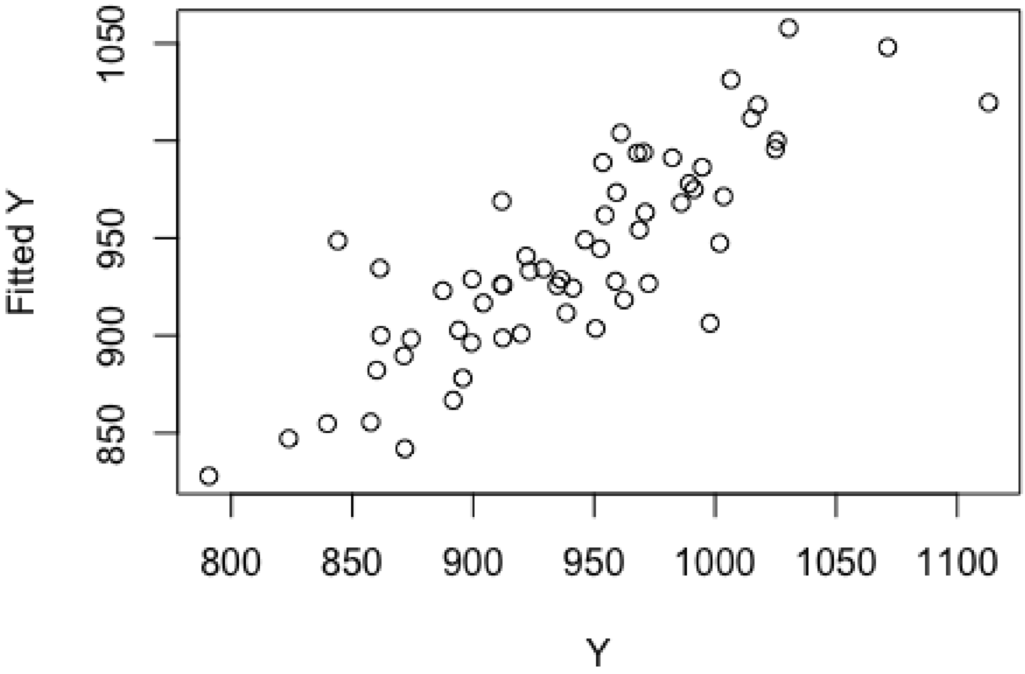
```
par(mfrow=c(2,2))
plot(lm(Y~X1, data = dat))
```



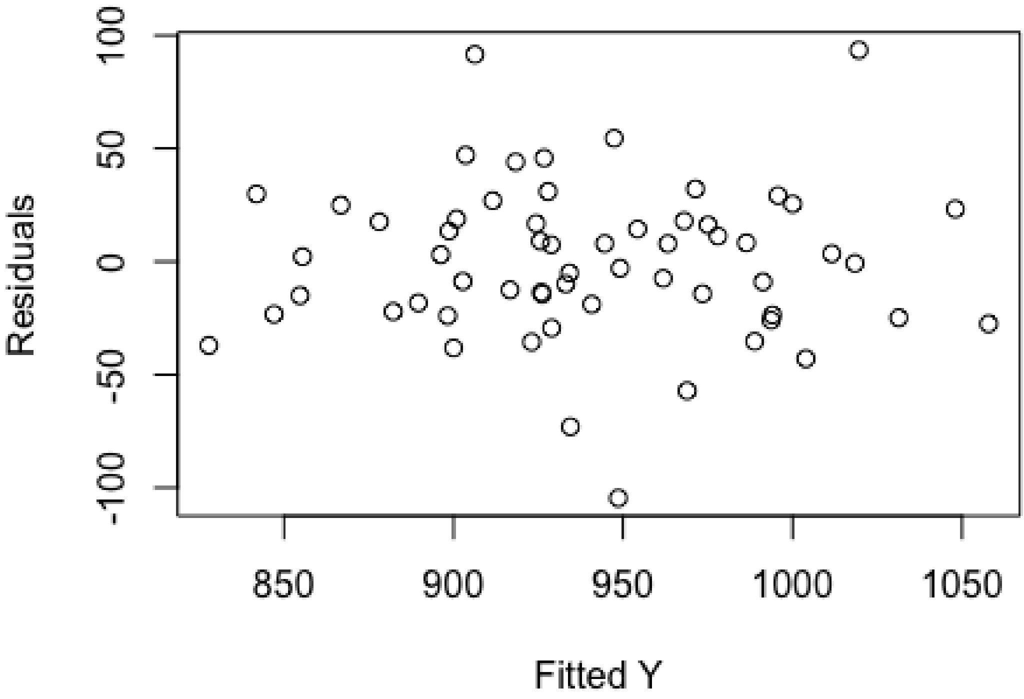
```
par(mfrow=c(1,1))
```

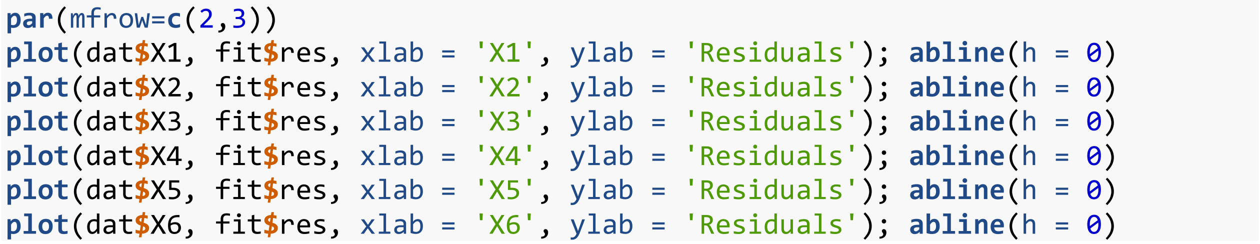
Step 3: Do the diagnostics

```
plot(dat$Y, fit$fitted.values, xlab = "Y", ylab = "Fitted Y") # plot the obse
rved Y values against the fitted Y-values
```

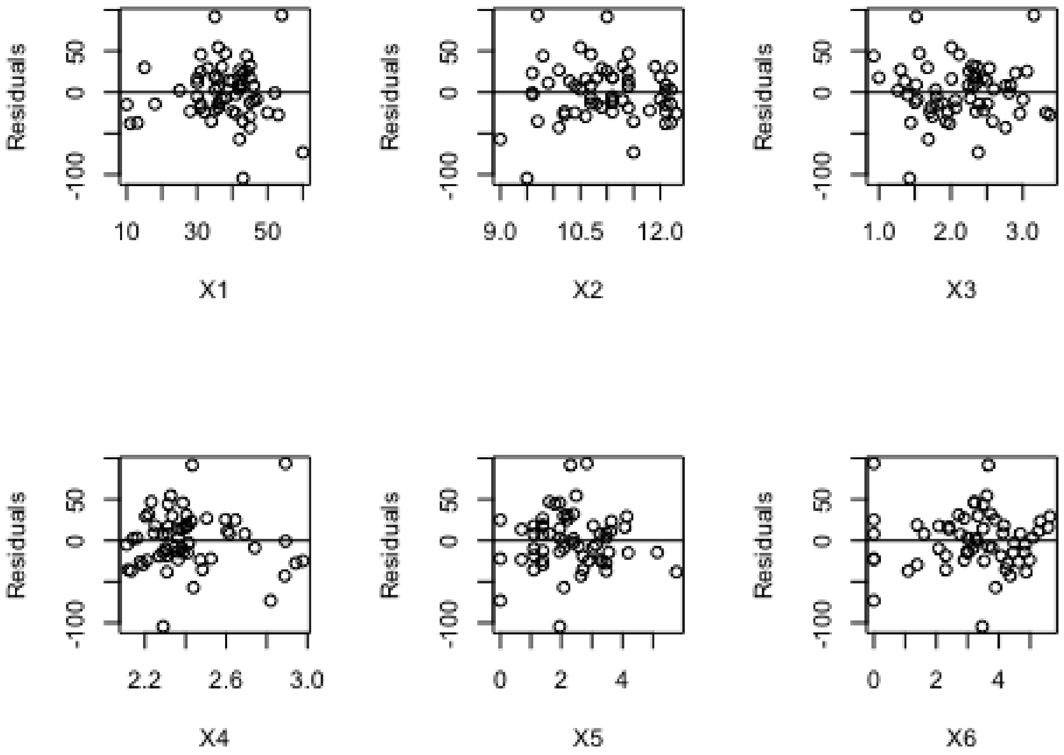


```
# plot the residuals against the independent variables; six graphs
plot(fit$fitted.values, fit$residuals, xlab = "Fitted Y", ylab = "Residuals")
 # plot of fitted Y against residuals
```

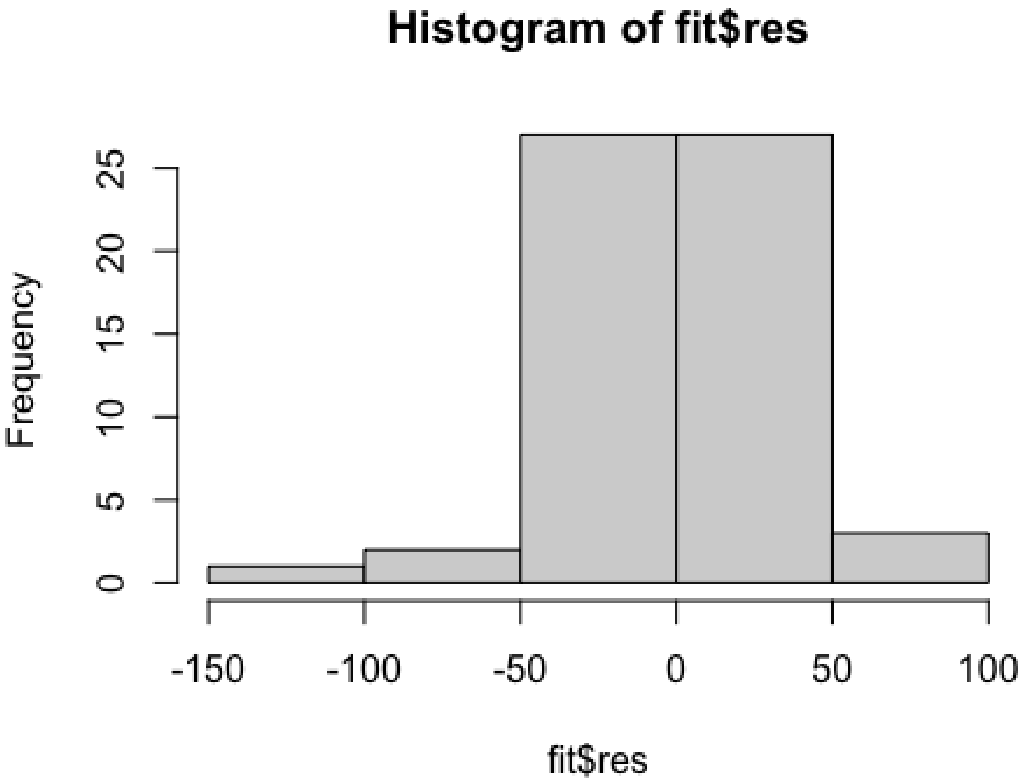




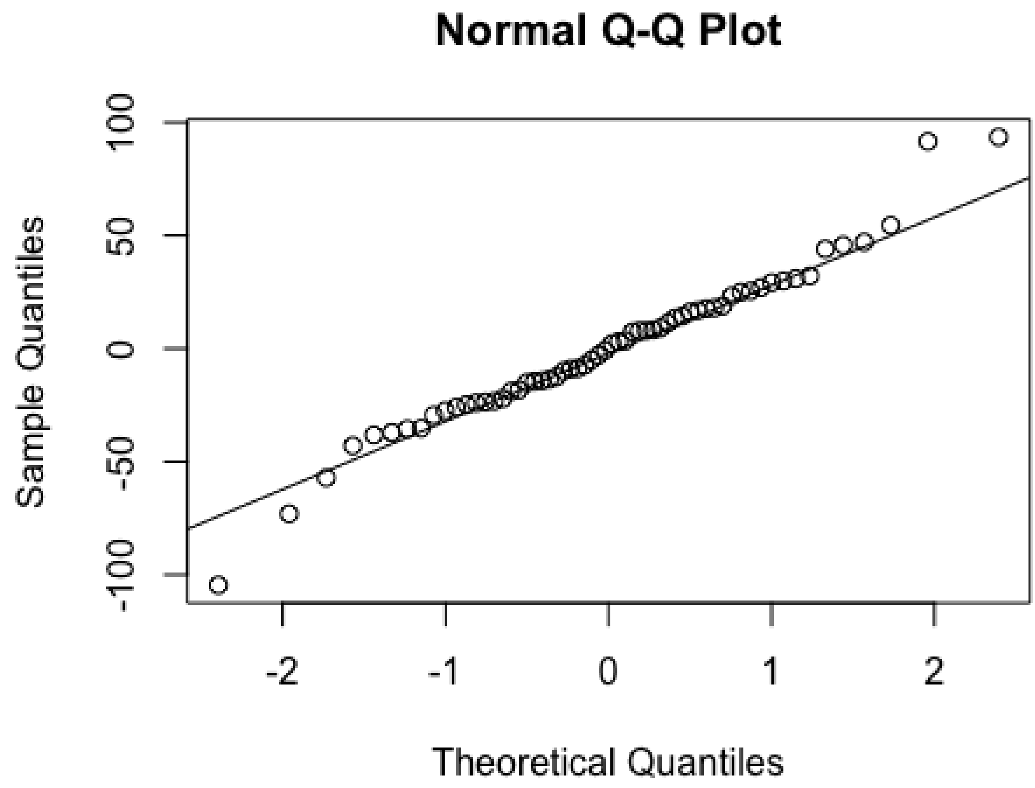
```
par(mfrow=c(2,3))
plot(dat$X1, fit$res, xlab = 'X1', ylab = 'Residuals'); abline(h = 0)
plot(dat$X2, fit$res, xlab = 'X2', ylab = 'Residuals'); abline(h = 0)
plot(dat$X3, fit$res, xlab = 'X3', ylab = 'Residuals'); abline(h = 0)
plot(dat$X4, fit$res, xlab = 'X4', ylab = 'Residuals'); abline(h = 0)
plot(dat$X5, fit$res, xlab = 'X5', ylab = 'Residuals'); abline(h = 0)
plot(dat$X6, fit$res, xlab = 'X6', ylab = 'Residuals'); abline(h = 0)
```



```
par(mfrow=c(1,1))
# histogram of residuals
hist(fit$res)
```



```
# normal probability plot of residuals
qqnorm(fit$res); qqline(fit$res)
```



####

Step 4: Transform the data

```
# The regression is linear, so there is no need to do the transformation.
```

Step 5: Exclude the unnecessary variables

```
model_full=lm(Y~., data=dat)
step(model_full, direction="both") # Backward Stepwise Regreession

## 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
## + X4     1    1049.6 68828 436.70
## - 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
## + X5     1      2087   69877 435.61
## + X4     1       577   71387 436.89
## - 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 = dat)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X6
##      883.03         1.90       -15.22        49.40        14.95
```

```
model_ini=lm(Y~1, data=dat)
step(model_ini, direction="both", scope=list(lower=model_ini, upper=model_ful
l)) # Forward Stepwise Regreession
```

```
## Start:  AIC=496.64
## Y ~ 1
##
##          Df Sum of Sq    RSS    AIC
## + X3     1     83923 144350 471.14
## + X2     1     59603 168671 480.48
## + X1     1     59256 169017 480.60
## + X4     1     38370 189903 487.60
## + X6     1     37098 191176 488.00
## + X5     1     19463 208810 493.29
## <none>               228273 496.64
##
## Step:  AIC=471.14
## Y ~ X3
##
##          Df Sum of Sq    RSS    AIC
## + X2     1     42716 101634 452.09
## + X6     1     30892 113459 458.69
## + X1    1     25361 118990 461.55
## + X5    1      7037 137313 470.14
## <none>              144350 471.14
## + X4    1       755 143596 472.82
## - X3    1     83923 228273 496.64
##
## Step:  AIC=452.09
## Y ~ X3 + X2
##
##        Df Sum of Sq    RSS    AIC
## + X6    1     16385  85249 443.54
## + X5    1      8748  92885 448.69
## + X1    1      4787  96847 451.19
## + X4    1      4380  97254 451.44
## <none>              101634 452.09
## - X2    1     42716 144350 471.14
## - X3    1     67037 168671 480.48
##
## Step:  AIC=443.54
## Y ~ X3 + X2 + X6
##
##        Df Sum of Sq    RSS    AIC
## + X1    1     13285  71965 435.38
## <none>               85249 443.54
## + X4    1        90  85160 445.48
## + X5    1        77  85173 445.49
## - X6    1     16385 101634 452.09
## - X2    1     28209 113459 458.69
## - X3    1     65315 150565 475.67
##
## Step:  AIC=435.38
## Y ~ X3 + X2 + X6 + X1
##
##        Df Sum of Sq    RSS    AIC
## <none>               71965 435.38
## + X5    1      2087  69877 435.61
## + X4    1       577  71387 436.89
## - 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 ~ X3 + X2 + X6 + X1, data = dat)
##
## Coefficients:
## (Intercept)           X3           X2           X6           X1
##      883.03        49.40       -15.22        14.95         1.90
```

```r
# They return the same model: Y ~ X1 + X2 + X3 + X6
bestFit <- lm(Y ~ X1 + X2 + X3 + X6, data = dat); bestFit

##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X6, data = dat)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X6
##      883.03         1.90       -15.22        49.40        14.95
```

REFERENCES

Source: GC McDonald and JS Ayers, “Some applications of the ‘Chernoff Faces’: a technique for graphically representing multivariate data”, in Graphical Representation of Multivariate Data, Academic Press, 1978.