The Study of Factors That Affect Death Rate in Cities

Yuwei Chen, David Liu, Peijun Xiao

March 16, 2015

Abstract
Our project focuses on the death rates in various cities in the United States and the variables that might be associated. In this study, we ask four main questions. First, we are interested in observing which variables are most useful in explaining trends in the death rate in a city to construct the best linear regression model given our data set. We are also concerned with analyzing the statistical relationships of both pollution and schooling with death rate.

Problem and Motivation
The data gathered reflects the death rate of 60 metropolitan cities in the United States during the 1980s and the various factors that might be associated including demographics, characteristics of the cities, characteristics of households, and data on three air pollutants, climate and weather. Cities including New York, Los Angeles, Miami, and New Orleans stand out as outliers in our data. They should be noted for their extreme characteristics and their influence on the rest of our statistical data.

Statistical data might not always allow us to make causal statements, but at least they help us observe trends and build upon our knowledge database of the world. The death rate in a city could be a sorrowful fact for some people, which is why observing various variables and their effects on different cities will help us develop a greater understanding of these morbid realities. With the various variables brought up during our report, such as schooling, demographics, pollution, and etc. we hope to shed light on possibly overlooked factors that could affect death rate.

Questions of Interest
-1. What collection of variables has the highest association to the death rate in the city?

-2. Is there a strong relationship between the average death rate in a city and the average number of years of schooling for people over 22?

-3. How does air pollution relate to the death rate in a city?

-4. How well can the model that we obtained from problem 1 explain the death rate of San Francisco?

Data
- Name of Date Set: Death Rate Dataset
- Source: Researchers collected data on 60 standard metropolitan areas in the the United States in a study of possible factors that attribute to mortality.
- Related variables: The data include measuring demographic characteristics of the cities, characteristics of households, variables recording the pollution potential of four different air pollutants and the death data (deaths per 100000).
- Reference:
    - R F Gunst and R L Mason, Regression Analysis and Its Applications, Dekker, 1980, pages 370-371.
    - Helmut Spaeth, Mathematical Algorithms for Linear Regression, Academic Press, 1991, ISBN 0-12-656460-4.

Statistics Methods
- To find and analyze the collection of variables that has the highest association to the death rate in the city, we use stepwise selection and Cp and SBC criterions as reference to choose a multiple linear regression model with death rate as the response variable and the best selection of explanatory variables included in the data set. ”
- We do t-test and find the extra sum of square to find whether there is a strong relationship between the death rate in a city and the average number of years of schooling for people over 22. We also find the confidence interval of the parameter of this variable.
- We construct a multiple regression model and obtain confidence interval, R^2 to assess the relation between death rate and variables related to air pollution.

- To answer how well the model that we obtained from question 1 explains the death rate of San Francisco, we obtain the 95% confidence interval of the data and decide if the real average death rate of San Francisco is in the confidence interval.

Statistical Analysis, Result and Interpretation

Find the Collection of Useful Variables

Step 1: Model Selection

Using stepwise selection, we obtained a model with death rate as response variables, and the explanatory variables are the size of the nonwhite population, the number of years of schooling for persons over 22, the average January temperature, the population per square mile, the average annual precipitation, the sulfur dioxide pollution index and the average July temperature.

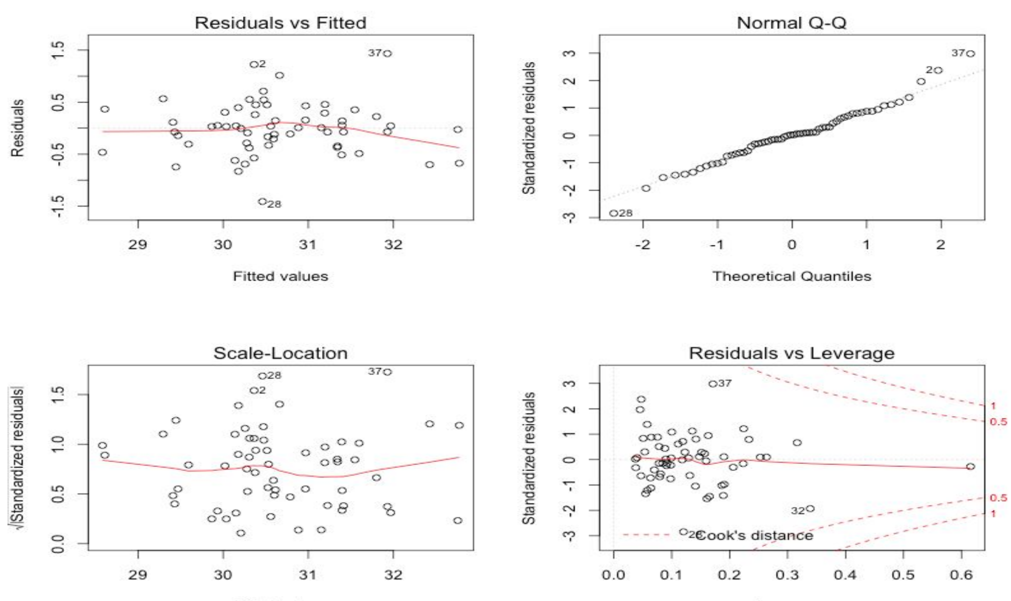
Step 2: Transformation on Response Variable

Using boxcox command twice, we found out the best transformation on Y is to do the square root. We observed that SOx_index does not have an obvious linear relationship to death_rate. However, if we transfer use square root to transform SOx_index, the parameter of July temperature will have p-value 0.14, which means it is not significant. Since the diagnosis plot of the model with transformation on response variable is good enough, we will not do transformation on parameters.

Step 3: Model Comparison

(See Appendix for the detailed comparison of models) We observed that the model we selected from the stepwise selection is not the best model obtained from Cp and SBC, but it is considered as the one of the best three, since the Cp and SBC criterions of the original model is close to the best model.

Step 4: Final Model Diagnosis



The residual v.s fitted value plot shows that the assumption of equal variance and linearity are mostly satisfied. The three notable outliers might cause the distortion of data. The qq plot shows that assumption of normality is in question due to the skewed tails ends of the plot. Looking at the residual v.s leverage plot, once again, we see that variance is is relatively equal. We can see that observation 28,38 and 37 are outliers in the data set but not extreme enough to have strong effect on the value of parameters.

Step 5: Model Interpretation

Final model:

sqrt(death rate)=33.79 + 0.07333*(size of the nonwhite population)-0.2531*(number of years of schooling for persons over 22) -0.02467*(average January temperature)+ 0.0001337* (population per square mile)+ 0.02815*(average annual precipitation)+ 0.002817*(sulfur dioxide pollution index)-0.03523*(average July temperature)

•Intercept: The average natural death rate in the city with the measure of other variables included in the model is 33.79^2 =1141.76 per 100000 persons, and we are 95% confident that natural death rate is between 929.1087 to 1456.088 deaths per 100000 persons, eliminating effects of other factors.

•Nonwhite Population: For every 1% increase of nonwhite residence in the city, the the square root of expected average deaths per 100000 persons increases 0.0733, and we are 95% confident that the square root of mean deaths per 100000 persons increases 0.0522 to 0.0944, holding other explanatory factor constant.

•School: For every extra one year of school for people over 22 in the city, the square root of expected average square root of deaths per 100000 persons decrease 0.2531, and we are 95% confident that on average the square root of deaths per 100000 persons will decrease 0.05123 to 0.455, holding other explanatory factor constant.

•January Temperature: For every one Fahrenheit degree increase of average January temperature, the the square root of expected average deaths per 100000 persons decreases 0.02406, and we are 95% confident that on average the square root of deaths per 100000 persons will decrease 0.000114 to 0.03796, holding other explanatory factor constant.

•Population Density: For every one population per square mile increase in the city, the average square root of deaths per 100000 persons increase 0.0001337, and we are 95% confident that on average the square root of deaths per 100000 persons increase 0.00002 to 0.0002, holding other explanatory factor constant.

•Precipitation: For every one unit increase in the average annual precipitation, the average square root of deaths per 100000 persons increase 0.02815, and we are 95% confident that on average the square root of deaths per 100000 persons increase 0.0093 to 0.047, holding other explanatory factor constant. deaths per 100000 persons.

• SO_2 index: For every one unit of the square root of sulfur dioxide pollution index, the average square root of deaths per 100000 persons increase 0.002817, and we are 95% confident that on average the square root of deaths per 100000 persons increase 0.000078 to 0.0056, holding other explanatory factor constant. deaths per 100000 persons

•July Temperature: For every one Fahrenheit degree increase in average July temperature, the average square root of expected deaths per 100000 persons decrease 0.03523, and we are 95% confident that on average the square root of deaths per 100000 persons will change -0.0744 to 0.0039 holding other explanatory factor constant. The interval includes zero, which implies that July temperature might not be statistically significant.

Step 6: Final Model Analysis

<u>Level of Significance</u>: The p-value for July average temperature is greater than 0.05, while all other variables are are significant at 0.05 level of significance. However, with our cp comparison (see Appendix), we determined that July average temperature is still a significant variable in our model. Seeing that the r-squared is 0.7604 is a good indicator that these variables approximately estimate 76.04% of the variability in death rate.

<u>Correlation coefficients</u>: From the correlation coefficient matrix, the problem of multicollinearity does not exist. The number of years of schooling for people over 22 and the average temperature in January and July have negative relationships with the average death rate of the city per 100,000, while the factors we included in the model have positive association. Only the percentage of nonwhite population, the number of years of schooling for people over 22 and the average precipitation have a strong relationship to the average death rate of the city per 100,000 persons. Factor such as population per square mile and sulfur dioxide pollution index has economically insignificant relationship with the change of death rate.

Find Relationship Between Education and Death Rate

Definitions, notation and assumptions:

In this question, we examine the relationship between the school variable and death rate in a city. The model that we just obtained includes the variable of the number of years of schooling for persons over 22. By conducting summary command, we get statistics about this variable and its connection to death rate.

Analysis/ Interpretation:

In order to test whether a relationship between the number of years of schooling for persons over 22 and death rate exists, we ran a hypothesis test.

$H_0 : \beta_{school} = 0$ $H_1 : \beta_{school} \neq 0$ P(t < -4.29)=0.018631< $\alpha = 0.05$

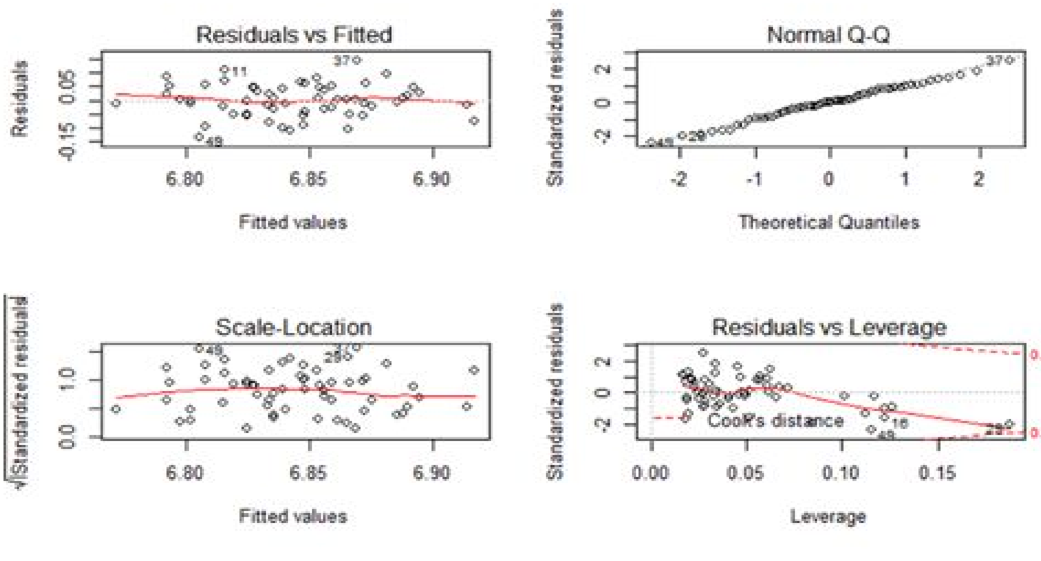
We reject the H_0. This suggests that school is statistically significant in explaining death rate, which means we can assume there is a linear relationship between the two variables. From the model we see that the parameter of this variable is -0.2486, and the correlation coefficient between school year and death rate is -0.5109475, This suggests a fairly good correlation between the variables, and indicates a negative relationship between school year and death rate. This means that for every one more years of school, we have lower average lower death rate. The coefficient of partial determination is $R^2_{school|nonwhite,\,jan\_temp,\,population\ density,\,precip,\,SOx\_index,\,jul\_temp} = \frac{1.7679}{16,2999} = 0.1085$. Approximately 10.85% reduction in error sum of squares is due to adding the school variable to the model in question 1 with all other explanatory variable.We are 95% confident that one extra year of schooling for person over 22 will have a change in square root of deaths per 100000 persons falling in the interval [-0.45389, -0.04332003]. It is statistically significant, and economically, its variation is related lager change in death rate compared to other factors.

Study on the Association Between Air Pollution and Death Rate

In the dataset, variables such as the hydrocarbon (X_1), nitric oxide (X_2) and sulfur dioxide (X_3) pollution index are related to air pollution. In order to examine the relation between death rate(Y) in a city and air pollution, we constructed a multiple regression model using death rate as response variable, and three air pollution indexes as explanatory variables. $H_0 : \beta_h = 0, \beta_n = 0, \beta_s = 0$; $H_1 : \beta_h \neq 0, \beta_n \neq 0, or\ \ \beta_s \neq 0$; P(F>9.152)=5.05e-05< $\alpha = 0.01$. Therefore, we reject H_0 and conclude that the hydrocarbon, nitric oxide and sulfur dioxide pollution index are jointly significant at 0.01 level of significance by F- test.

Model Construction/Model Transformation:

$H_0 : \beta_i = 0, i = h, n, s$; $H_1 : \beta_i \neq 0, i = h, n, s$; $p(t_h < -2.408) = 0.0193 < \alpha = 0.05$; $p(t_n > 2.075) = 0.0426 < \alpha = 0.05$; $p(t_s > 1.257) = 0.214 > \alpha = 0.05$. Using t test, we found that unlike the hydrocarbon, nitric oxide pollution index, sulfur dioxide pollution index solely is not statistically significant at 0.05 level of significance. Therefore, the multiple regression model will only include hydrocarbon (X_1), nitric oxide (X_2) pollution index. Based on correlation matrix, we decided to conducted logarithm transformation on X_1 ,and X_2. Then we produced diagnosis plots of this model.



Residual-fitted value plot shows linearity, generally equal variance and no pattern. In normal QQ plot, most observation points lie on the line with few outliers, implying normality of variance. No observation has Cook's distance larger than 0.5. General linear model assumptions are fulfilled. Observation 29, Los Angeles, Long Beach, CA is an outlier. It has extremely high hydrocarbon pollution index, and it is known for air pollution.

Model:
death rate=942.08-63.97*log(hydrocarbon pollution index)+75.35*log(nitric oxide pollution index)
Analysis and Interpretation:
Using "summary" command, we confirmed that both parameters of log(hydrocarbon pollution index) and log(nitric oxide pollution index) are statistically significant. These two factor explain 22.38% of variation of death rate in this model. We found that 95% confidence interval of b_1 is [-103.4386,-24.50141], b_2 is [36.08165, 114.6183]. Since we had logarithm transformation on both explanatory variables, we interpreted the confidence interval differently. On average, a one percent increase of hydrocarbon pollution index will cause deaths per 100000 people decrease by any amount from 0.2450141 to 1.034386 persons. A one percent increase in nitric oxide pollution index will cause deaths per 100000 people to increase by any amount from 0.3608165 to 1.146183 persons. This model states that air pollution is related to the death rate in a city. However, this model is used to examine the relation between air pollution and death rate. It cannot be used in general prediction case. Noted that r between two explanatory variable is 0.98, and their vif value are 31.56791. these statistics implies that hydrocarbon pollution index and nitric oxide pollution index are highly correlated. Multicollinearity can cause biased estimation of parameters.

Check Mean Response of Death Rate to Test the Fit of the Model
We used statistics of San Francisco in the data set and calculated the fitted death rate using the model stated in question 1. We expect the average death rate in San Francisco is 910.6695 per 100000 persons. We also constricted a confidence interval of mean death rate. With 95% confidence, the average death rate of San Francisco is from 877.9477 to 943.3913 per 100,000 persons. our observation shows that the actual death rate in San Francisco is 911 per 100,000 persons, which is in our confidence interval. Thus, we conclude model explains the average death rate of San Francisco very well.

Criticism and Possible Extensions
- Some models we constructed did not strictly satisfy assumptions of general linear regression model. For example, the final model in the first question did not satisfy the assumption of normality of residuals. However, as long as the model is useful in explaining and predicting data and does not strongly deviate from assumptions, it should be considered a good model. In fact, this model explains the average death rate of San Francisco fairly well.

- The model we chose does not have the lowest AIC value. However, it has good performance in Cp value and SBC value. Choosing this model, we can avoid the issue of multilinearity, which can lead to biased estimation of parameter of factors and standard errors.
- The R^2 stated in the model summary is 0.7605, which means it leaves approximately 25% of the variability of death rate unexplained. This could be a result of excluding relevant variables in our data set. However, this could also be a sign that some variables outside the data set makes our estimation of parameters biased.
- The model that we obtain may not be able to explain and predict the death rate in cities nowadays, since the data set is from sometime earlier than 1980s. A survey has to be conducted to obtain latest statistics increase accuracy and usefulness. Within the data measured, there is also a lot of error and variability that could minimized to improve the quality of the data set.
- Being that our data set reflects only cities in United States, our report generalizes to only metropolitan areas in the United States. Our data does not apply to other areas for example third world countries or non metropolitan areas. As with all datasets, there are many related variables that were not accounted for, which also leaves room for the dataset to improve.

Conclusions

Our data shows that variables involving demographics, characteristics of the cities, characteristics of households, climate related statistics, and data on three air pollutants have some degree of statistical significance in the variability of average death rates in a city. Certain variables such as years of schooling, monthly temperatures of January and July have a negative covariance with death rate while other variables including certain pollutants have a positive covariance with death rate. Jointly, our linear regression model shows that they could potentially explain 75% of the variability in death rate. We conduct models and found that education is positively associate with death rates in cities, and Air pollutants have mixed relationship with death rates. OUr model is not perfect, but it is useful to assess to the death in some metropolitans. The research gathered from our data set is not enough to make general predictive statements about present day death rates because our constructed linear regression model was based on data earlier than the 1980s and much of that data is out-dated.

Appendices A

R code for Question 1:

```
> deathrate = read.table("~/Desktop/STA 108/deathrate_dataset.txt",
header=TRUE)
> reg=lm(death_rate~1, data=deathrate)
>
step(reg,scope=death_rate~precip+jan_temp+jul_temp+age_65_plus+househ
old+school+kitchen+pop_density+nonwhite+office+inc_30k+HC_index+NOx_i
ndex+SOx_index+atmos, direction="both")
Start:  AIC=496.64
death_rate ~ 1

              Df Sum of Sq    RSS    AIC
+ nonwhite     1     94617 133659 466.52
+ school       1     59595 168680 480.48
+ precip       1     59257 169019 480.61
+ kitchen      1     41576 186699 486.57
+ SOx_index    1     41389 186886 486.63
+ inc_30k      1     38468 189808 487.57
+ household    1     29143 199133 490.44
+ office       1     18391 209884 493.60
+ jul_temp     1     17706 210569 493.79
+ pop_density  1     16087 212189 494.25
<none>                     228276 496.64
+ HC_index     1      7167 221109 496.72
+ age_65_plus  1      6966 221309 496.78
+ NOx_index    1      1476 226800 498.25
+ jan_temp     1       764 227511 498.44
+ atmos        1       667 227609 498.46

Step:  AIC=466.52
death_rate ~ nonwhite

              Df Sum of Sq    RSS    AIC
+ school       1     33850  99809 451.00
+ SOx_index    1     24490 109169 456.38
+ age_65_plus  1     21363 112296 458.07
+ jan_temp     1     18679 114980 459.49
+ office       1     18212 115446 459.73
+ pop_density  1     16521 117138 460.61
+ precip       1     16287 117372 460.73
```

```
+ kitchen      1      7264 126395 465.17
+ HC_index     1      5881 127778 465.82
<none>                     133659 466.52
+ jul_temp     1      2943 130716 467.19
+ household    1      2125 131534 467.56
+ NOx_index    1      1724 131935 467.74
+ inc_30k      1       865 132794 468.13
+ atmos        1         2 133657 468.52
- nonwhite     1     94617 228276 496.64

…

Step:  AIC=426.5
death_rate ~ nonwhite + school + jan_temp + pop_density + precip +
    SOx_index

              Df Sum of Sq    RSS    AIC
+ jul_temp     1      3369  54695 424.91
+ atmos        1      1906  56158 426.49
<none>                      58064 426.50
+ household    1      1129  56936 427.32
+ kitchen      1       757  57307 427.71
+ NOx_index    1       549  57515 427.93
+ HC_index     1       221  57843 428.27
+ age_65_plus  1       187  57878 428.30
+ office       1        26  58038 428.47
+ inc_30k      1        12  58052 428.49
- SOx_index    1      5550  63615 429.98
- pop_density  1      6142  64206 430.53
- school       1      6440  64504 430.81
- precip       1      6645  64709 431.00
- jan_temp     1     16679  74743 439.65
- nonwhite     1     50277 108342 461.92

Step:  AIC=424.91
death_rate ~ nonwhite + school + jan_temp + pop_density + precip +
    SOx_index + jul_temp

              Df Sum of Sq    RSS    AIC
<none>                      54695 424.91
+ kitchen      1      1151  53544 425.63
+ household    1      1039  53656 425.76
- jul_temp     1      3369  58064 426.50
```

```
+ inc_30k      1      242  54453 426.65
+ atmos        1      190  54505 426.70
+ HC_index     1       32  54663 426.88
+ age_65_plus  1       29  54666 426.88
+ NOx_index    1       13  54682 426.90
+ office       1        5  54690 426.91
- SOx_index    1     4452  59147 427.61
- pop_density  1     5695  60390 428.85
- school       1     6826  61521 429.97
- precip       1     8920  63616 431.98
- jan_temp     1    13970  68665 436.56
- nonwhite     1    51875 106570 462.93

Call:
lm(formula = death_rate ~ nonwhite + school + jan_temp + pop_density
+
    precip + SOx_index + jul_temp, data = deathrate)

Coefficients:
(Intercept)     nonwhite       school     jan_temp  pop_density
precip    SOx_index
  1.167e+03    4.528e+00   -1.572e+01   -1.481e+00    8.076e-03
1.681e+00    1.723e-01
   jul_temp
 -2.143e+00

> best=lm(formula = death_rate ~ nonwhite + school + jan_temp +
pop_density + precip + SOx_index + jul_temp, data = deathrate)
> #Use other criterion to check
> library(leaps)
>
subsets=regsubsets(death_rate~precip+jan_temp+age_65_plus+household+s
chool+kitchen+pop_density+nonwhite+office+inc_30k+HC_index+NOx_index+
SOx_index+atmos+jul_temp, data=deathrate, nbest=1)
> c1=summary(subsets)$which   # Get the subsets used
> c2=summary(subsets)$cp      # Get Cp for Best Subsets
> c3=summary(subsets)$bic     # Get SBC for Best Subsets
> p = rowSums(summary(subsets)$which)   # Retrieve p = # betas
> summary(subsets)$bic  +  2*log(nrow(mtcars))*p - 2*p
        1          2          3          4          5          6
7          8
-14.064140 -22.560128 -25.873124 -28.118507 -22.507569 -18.959319
-13.660773  -8.155062
```

```
> #Get AIC for Best Subsets, by SBC w/different penalty
> cbind("p"=p, "cp"=c2,"sbc"=c3,"adjRsp"=summary(subsets)$adjr2,c1)
```

The pink and yellow highlights indicates the best three models selected by Cp or SBC criterions

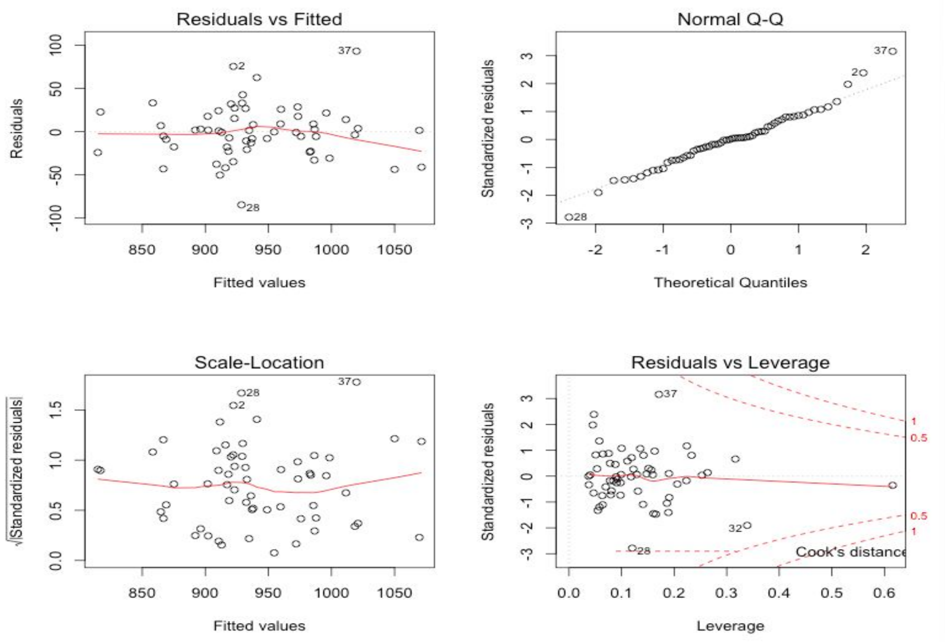
The green highlighted model is our selected model from stepwise function, which can also be considered as a good model.

```
  p         cp        sbc    adjRsp (Intercept) precip jan_temp jul_temp age_65_plus
household school kitchen
1 2  71.815664 -23.927084 0.4043896           1      0        0        0           0
0        0      0
1 2 105.306044  -9.964223 0.2483272           1      0        0        0           0
0        1      0
1 2 105.630003  -9.843842 0.2468175           1      1        0        0           0
0        0      0
2 3  41.445845 -37.354544 0.5474273           1      0        0        0           0
0        1      0
2 3  50.395979 -31.976615 0.5049887           1      0        0        0           0
0        0      0
2 3  53.386485 -30.282025 0.4908087           1      0        0        0           1
0        0      0
3 4  25.704421 -45.599011 0.6249718           1      0        1        0           0
0        1      0
3 4  27.703608 -44.074849 0.6153230           1      1        0        0           0
0        0      0
3 4  29.487332 -42.746886 0.6067142           1      0        0        0           0
0        1      0
4 5  14.396471 -52.775866 0.6835496           1      0        1        0           0
0        1      0
4 5  18.156730 -49.370802 0.6650714           1      0        1        0           0
0        1      0
4 5  19.501570 -48.198436 0.6584627           1      1        1        0           0
0        0      0
5 6  12.833718 -52.096399 0.6955214           1      1        1        0           0
0        1      0
5 6  13.399519 -51.540931 0.6926895           1      1        1        0           0
0        1      0
5 6  13.432830 -51.508389 0.6925228           1      0        1        0           0
1        1      0
6 7   9.526002 -53.479622 0.7168434           1      1        1        0           0
0        1      0
6 7  10.052725 -52.913141 0.7141573           1      0        1        0           0
0        1      0
6 7  10.561381 -52.371120 0.7115634           1      1        1        1           0
0        1      0
7 8   8.181616 -53.112547 0.7287808           1      1        1        0           0
0        1      0
7 8   8.303929 -52.972072 0.7281451           1      1        1        1           0
0        1      0
7 8   9.453967 -51.667111 0.7221676           1      0        1        0           0
0        1      0
8 9   7.208276 -52.538309 0.7392201           1      1        1        1           0
0        1      0
```

```
8 9   7.646816 -52.005962 0.7368960          1      1        1        0           0
0       1       0
8 9   9.085182 -50.292350 0.7292734          1      0        1        1           0
1       1       0
  pop_density nonwhite office inc_30k HC_index NOx_index SOx_index atmos
1           0        1      0       0        0         0         0     0
1           0        0      0       0        0         0         0     0
1           0        0      0       0        0         0         0     0
2           0        1      0       0        0         0         0     0
2           0        1      0       0        0         0         1     0
2           0        1      0       0        0         0         0     0
```

```
3           0        1      0       0        0         0         0     0
3           0        1      0       0        0         0         1     0
3           0        1      0       0        0         0         1     0
4           1        1      0       0        0         0         0     0
4           0        1      0       0        0         0         1     0
4           0        1      0       0        0         0         1     0
5           1        1      0       0        0         0         0     0
5           0        1      0       0        0         0         1     0
5           1        1      0       0        0         0         0     0
6           1        1      0       0        0         0         1     0
6           1        1      0       0        1         1         0     0
6           1        1      0       0        0         0         0     0
7           1        1      0       0        1         1         0     0
7           1        1      0       0        0         0         1     0
7           1        1      0       0        1         1         0     1
8           1        1      0       0        1         1         0     0
8           1        1      0       0        1         1         0     1
8           1        1      0       0        1         1         0     0
```

```
> #Focus back on our model obtained from Stepwise function
> plot(best)
```




```
> #Found non-normality, do Y transformation
> library(MASS)
> boxcox(best)
> #lamda is 0.5, consider to use square root
> besty=lm(sqrt(death_rate) ~ nonwhite + school + jan_temp +
pop_density + precip + SOx_index + jul_temp, data = deathrate)
> plot(besty)
> boxcox(besty)
> #use pairs to the linearity between death_rate and variables
> #However, the model is still not normal distributed, consider
transformation on X
> pairs(death_rate~nonwhite + school + jan_temp + pop_density +
precip + sqrt(SOx_index) + jul_temp, data = deathrate)
> #It is obvious that Y~Sox_index
> bestx=lm(sqrt(death_rate)~ nonwhite + school + jan_temp +
pop_density + precip + sqrt(SOx_index) + jul_temp, data = deathrate)

>#Final model analysis
> summary(bestx)

Call:
lm(formula = sqrt(death_rate) ~ nonwhite + school + jan_temp +
    pop_density + precip + SOx_index + jul_temp, data = deathrate)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.4105 -0.3140  0.0097  0.2940  1.4342

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.432e+01  1.913e+00  17.936  < 2e-16 ***
nonwhite     7.333e-02  1.051e-02   6.977 5.39e-09 ***
school      -2.531e-01  1.006e-01  -2.515 0.015026 *
jan_temp    -2.467e-02  6.624e-03  -3.724 0.000484 ***
pop_density  1.337e-04  5.658e-05   2.363 0.021905 *
precip       2.815e-02  9.411e-03   2.991 0.004242 **
SOx_index    2.817e-03  1.365e-03   2.064 0.044035 *
jul_temp    -3.523e-02  1.952e-02  -1.805 0.076933 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5286 on 52 degrees of freedom
Multiple R-squared:  0.7607,     Adjusted R-squared:  0.7285
F-statistic: 23.62 on 7 and 52 DF,  p-value: 4.507e-14

> plot(bestx)
> cor(cbind(deathrate$nonwhite,deathrate$school ,deathrate$jan_temp,
deathrate$pop_density ,deathrate$precip ,
deathrate$SOx_index,deathrate$jul_temp), y=deathrate$death_rate)
           [,1]
[1,]  0.64380482
[2,] -0.51094752
[3,] -0.05785809
[4,]  0.26546331
[5,]  0.50949321
[6,]  0.42580750
[7,]  0.27850652
> #Confidence Interval

> aa=cbind(lower=3.432e+01-qt(0.975,52)*1.913e+00,
upper=3.432e+01+qt(0.975,52)*1.913e+00)
> aa^2
        lower    upper
[1,] 929.1087 1456.088
> ab=cbind(lower=7.333e-02-qt(0.975,52)*1.051e-02,
upper=7.333e-02+qt(0.975,52)*1.051e-02)
> ab^2
           lower      upper
[1,] 0.002729032 0.00891511
```

```
> ac=cbind(lower=-2.531e-01 -qt(0.975,52)*1.006e-01, upper=-2.531e-01
+qt(0.975,52)*1.006e-01)
> ac^2
         lower       upper
[1,] 0.2069965 0.002624649
> ad=cbind(lower=-2.467e-02-qt(0.975,52)*6.624e-03,
upper=-2.467e-02+qt(0.975,52)*6.624e-03)
> ad^2
           lower        upper
[1,] 0.001441116 0.0001294582
> ae=cbind(lower=1.337e-04 -qt(0.975,52)*5.658e-05, upper=1.337e-04
+qt(0.975,52)*5.658e-05)
> ae^2
            lower        upper
[1,] 4.065838e-10 6.112568e-08
> af=cbind(lower= 2.815e-02-qt(0.975,52)*9.411e-03 , upper=
2.815e-02+qt(0.975,52)*9.411e-03 )
> af^2
            lower       upper
[1,] 8.584851e-05 0.002212249
> ag=cbind(lower=2.817e-03-qt(0.975,52)*1.365e-03   ,
upper=2.817e-03+qt(0.975,52)*1.365e-03   )
> ag^2
            lower        upper
[1,] 6.072635e-09 3.086995e-05
>
ah=cbind(lower=-3.523e-02-qt(0.975,52)*1.952e-02,upper=-3.523e-02+qt(
0.975,52)* 1.952e-02 )
> ah^2
           lower       upper
[1,] 0.005535322 1.55216e-05
```

```
> cbind(lower=3.432e+01-qt(0.95,52)*1.913e+00,
upper=3.432e+01+qt(0.95,52)*1.913e+00)
        lower    upper
[1,] 31.11632 37.52368
> cbind(lower=7.333e-02-qt(0.95,52)*1.051e-02,
upper=7.333e-02+qt(0.95,52)*1.051e-02)
          lower      upper
[1,] 0.05572902 0.09093098
```

```
> cbind(lower=-2.531e-01 -qt(0.95,52)*1.006e-01, upper=-2.531e-01
+qt(0.95,52)*1.006e-01)
          lower       upper
[1,] -0.4215737 -0.08462627
> cbind(lower=-2.467e-02-qt(0.95,52)*6.624e-03,
upper=-2.467e-02+qt(0.95,52)*6.624e-03)
           lower       upper
[1,] -0.03576314 -0.01357686
> cbind(lower=1.337e-04 -qt(0.95,52)*5.658e-05, upper=1.337e-04
+qt(0.95,52)*5.658e-05)
            lower        upper
[1,] 3.894609e-05 0.0002284539
> cbind(lower= 2.815e-02-qt(0.95,52)*9.411e-03 , upper=
2.815e-02+qt(0.95,52)*9.411e-03 )
         lower     upper
[1,] 0.0123895 0.0439105
> cbind(lower=2.817e-03-qt(0.95,52)*1.365e-03   ,
upper=2.817e-03+qt(0.95,52)*1.365e-03   )
            lower       upper
[1,] 0.0005310493 0.005102951
> cbind(lower=-3.523e-02-qt(0.95,52)* 1.952e-02  ,
upper=-3.523e-02-03+qt(0.95,52)* 1.952e-02   )
           lower    upper
[1,] -0.06791993 -3.00254
> rbind(aa,ab,ac,ad,ae,af,ag,ah)
             lower         upper
[1,]  3.048128e+01 38.1587153381
[2,]  5.224014e-02  0.0944198579
[3,] -4.549687e-01 -0.0512313314
[4,] -3.796203e-02 -0.0113779716
[5,]  2.016392e-05  0.0002472361
[6,]  9.265447e-03  0.0470345531
[7,]  7.792711e-05  0.0055560729
[8,] -7.439975e-02  0.0039397456
> rbind(aa,ab,ac,ad,ae,af,ag,ah)
             lower         upper
[1,]  3.048128e+01 38.1587153381
[2,]  5.224014e-02  0.0944198579
[3,] -4.549687e-01 -0.0512313314
[4,] -3.796203e-02 -0.0113779716
[5,]  2.016392e-05  0.0002472361
[6,]  9.265447e-03  0.0470345531
[7,]  7.792711e-05  0.0055560729
[8,] -7.439975e-02  0.0039397456
```

R Code for Question 2:

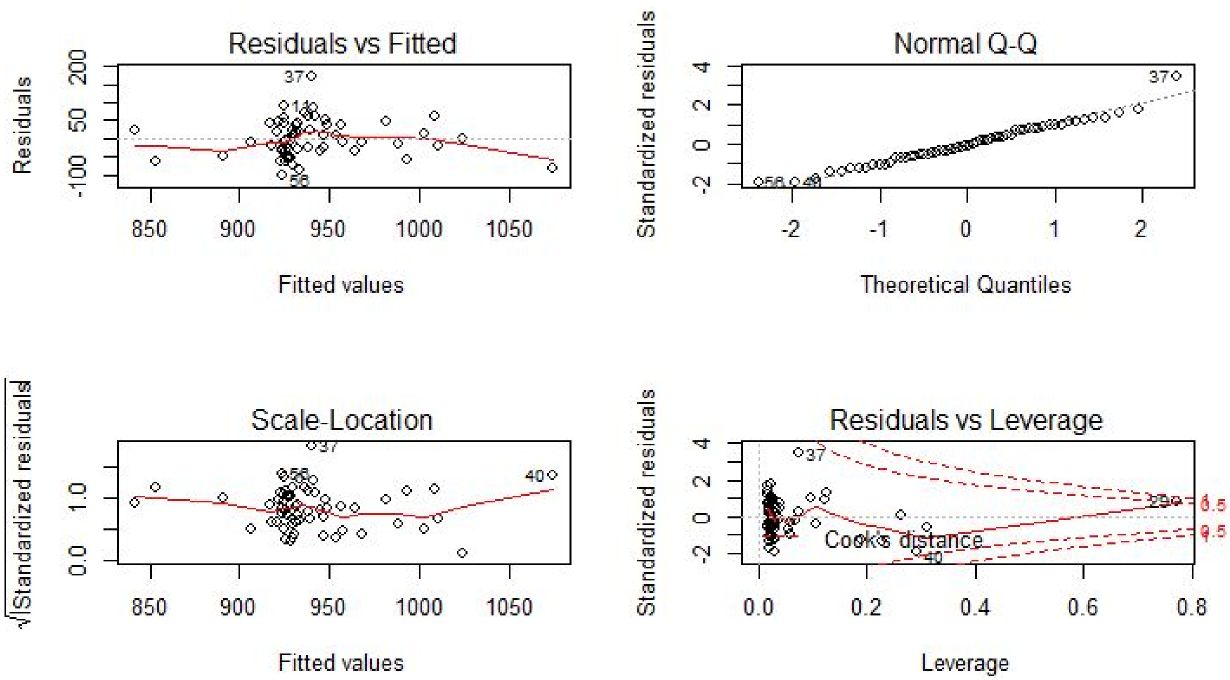
```
> anova(lm(sqrt(death_rate)~nonwhite +jan_temp + pop_density + precip
+ (SOx_index) +jul_temp +school, data=deathrate_dataset))
Analysis of Variance Table
Response: sqrt(death_rate)
            Df  Sum Sq Mean Sq F value    Pr(>F)
nonwhite     1 24.7851 24.7851 88.6889 7.921e-13 ***
jan_temp     1  5.1628  5.1628 18.4742 7.578e-05 ***
pop_density  1  7.0102  7.0102 25.0847 6.686e-06 ***
precip       1  4.1664  4.1664 14.9086 0.0003139 ***
SOx_index    1  2.4977  2.4977  8.9377 0.0042595 **
jul_temp     1  0.8080  0.8080  2.8914 0.0950284 .
school       1  1.7679  1.7679  6.3262 0.0150261 *
Residuals   52 14.5320  0.2795
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(lm(sqrt(death_rate)~nonwhite +jan_temp + pop_density + precip
+ (SOx_index) +jul_temp, data=deathrate_dataset))
Analysis of Variance Table
Response: sqrt(death_rate)
            Df  Sum Sq Mean Sq F value    Pr(>F)
nonwhite     1 24.7851 24.7851 80.5900 3.182e-12 ***
jan_temp     1  5.1628  5.1628 16.7872 0.0001440 ***
pop_density  1  7.0102  7.0102 22.7940 1.462e-05 ***
precip       1  4.1664  4.1664 13.5472 0.0005457 ***
SOx_index    1  2.4977  2.4977  8.1215 0.0062173 **
jul_temp     1  0.8080  0.8080  2.6274 0.1109716
Residuals   53 16.2999  0.3075
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Code for Question 3:

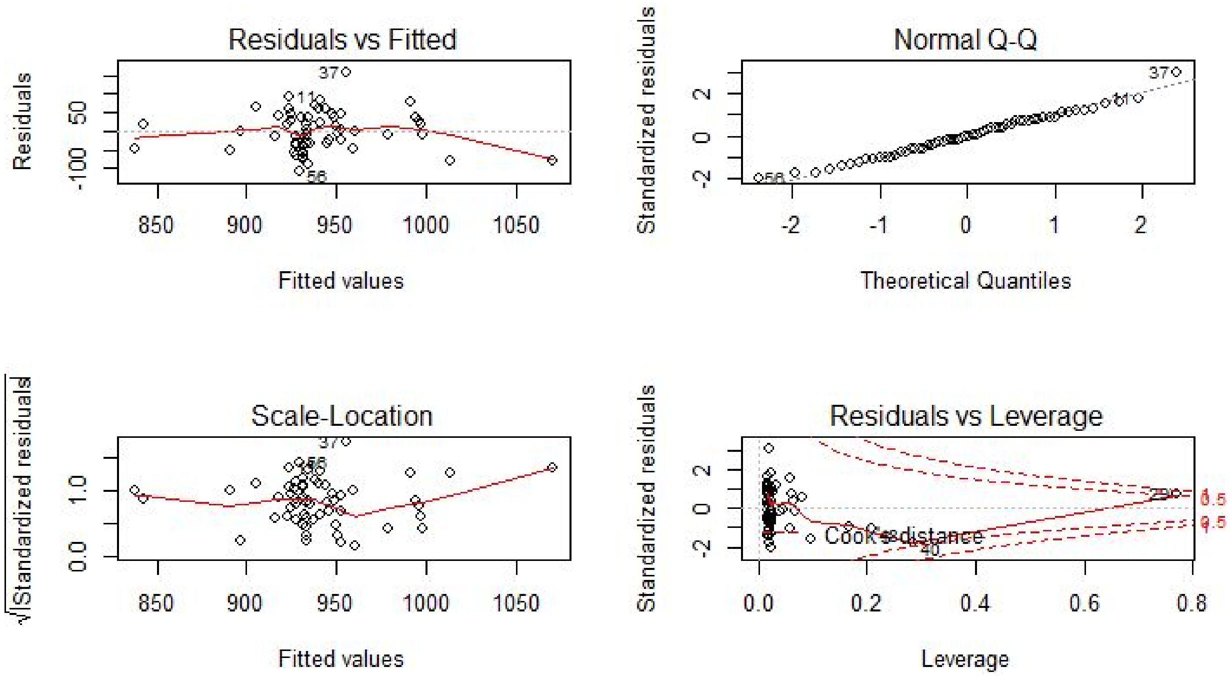
```
> #model with three pollution index variables
> model3=lm(formula = death_rate ~ HC_index + NOx_index + SOx_index,
data = deathrate_dataset)
> summary(model3)
Call:
lm(formula = death_rate ~ HC_index + NOx_index + SOx_index, data =
deathrate_dataset)
Residuals:
```

```
     Min        1Q    Median        3Q       Max
-100.292  -33.352   -5.458   37.506  172.586
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 924.3782     9.0897 101.695   <2e-16 ***
HC_index     -1.5064     0.6255  -2.408   0.0193 *
NOx_index     2.7082     1.3053   2.075   0.0426 *
SOx_index     0.2226     0.1771   1.257   0.2140
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 52.3 on 56 degrees of freedom
Multiple R-squared:  0.329,      Adjusted R-squared:  0.293
F-statistic: 9.152 on 3 and 56 DF,  p-value: 5.05e-05
> plot(model3)
```

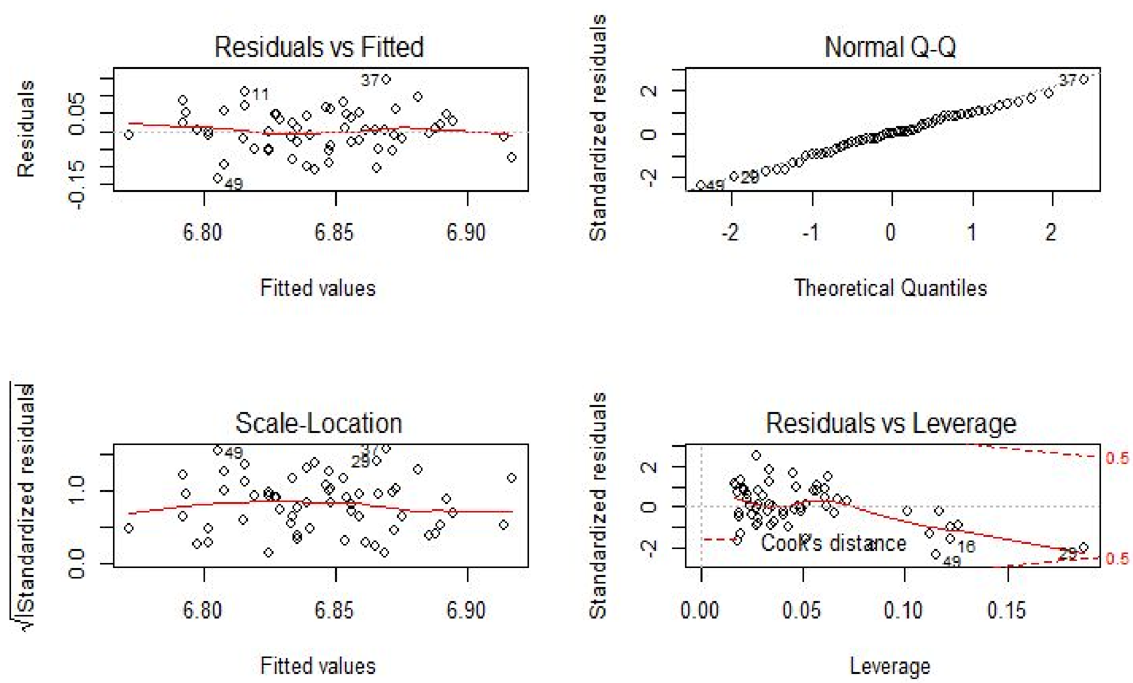



```
> #standard error of parameter of the hydrocarbon pollution index
> sehydrocarbon=0.6255
> #standard error of parameter of the nitric oxide pollution index
> senitric=1.3053
> #value of parameter of the hydrocarbon pollution index
> bHC_index=-1.5064
> #value of parameter of the nitric oxide pollution index
> bNOx_index=2.7082
# construct a model without SOx_index
```

```
> model3.1=lm(death_rate~HC_index+NOx_index, data=deathrate_dataset)
> summary(model3.1)
Call:
lm(formula = death_rate ~ HC_index + NOx_index, data =
deathrate_dataset)
Residuals:
     Min       1Q   Median       3Q      Max
-105.833  -35.312   -1.773   37.367  157.370
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 929.9486     7.9766 116.585  < 2e-16 ***
HC_index     -2.0936     0.4180  -5.008 5.63e-06 ***
NOx_index     3.9796     0.8294   4.798 1.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 52.57 on 57 degrees of freedom
Multiple R-squared:  0.3101,     Adjusted R-squared:  0.2859
F-statistic: 12.81 on 2 and 57 DF,  p-value: 2.548e-05
> library(leaps)
ll=regsubsets(death_rate~HC_index+NOx_index+SOx_index,data=deathrate_
dataset, nbest=2)
> ci=summary(ll)$which
> c1=summary(ll)$which
> c2=summary(ll)$cp
> c2=summary(ll)$bic
> c2=summary(ll)$cp
> c3=summary(ll)$bic
> p = rowSums(summary(ll)$which)
> summary(ll)$bic + 2*log(nrow(deathrate_dataset))*p - 2*p
        1         1         2         2         3
 8.562932 18.652065  8.580070 11.353960 17.193609
> c4=summary(ll)$bic + 2*log(nrow(deathrate_dataset))*p - 2*p
> cbind("p"=p, "cp"=c2,"SBC"=c3,"adjRsq" =
summary(ll)$adjr2,"AIC"=c4,c1)
  p        cp       SBC     adjRsq       AIC (Intercept) HC_index NOx_index SOx_index
1 2 12.324941 -3.814446 0.16719672  8.562932           1        0         0         1
1 2 24.836378  6.274687 0.01469654 18.652065           1        1         0         0
2 3  3.580076 -9.985998 0.28585149  8.580070           1        1         1         0
2 3  6.304583 -7.212108 0.25206024 11.353960           1        1         0         1
3 4  4.000000 -7.561148 0.29304600 17.193609           1        1         1         1
#pick p=3, explanatory variables include HC_index and NOx_index
plot(model3.1)
```



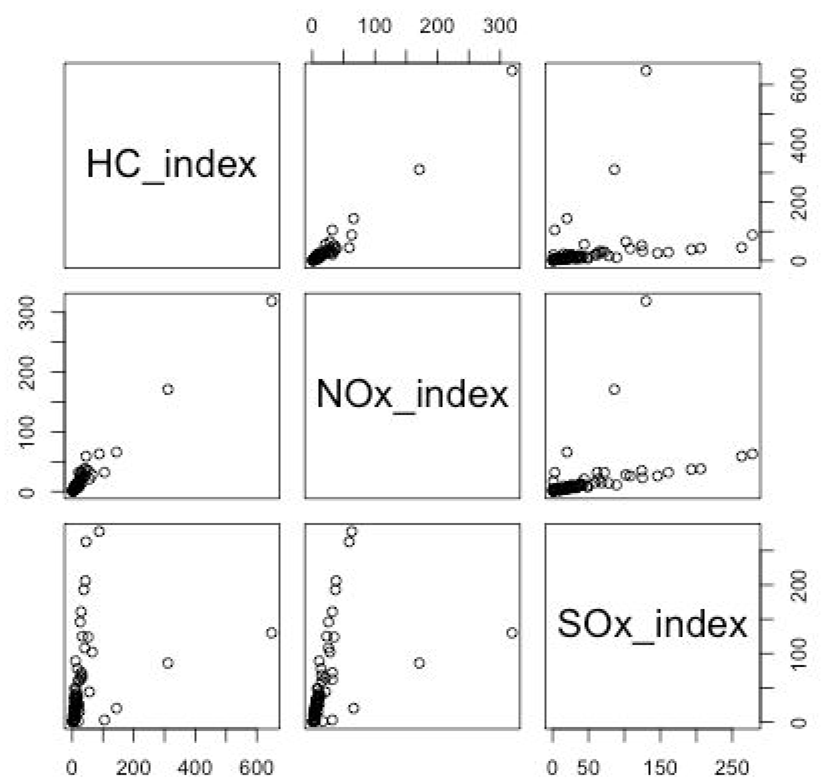

```
> model3.3=lm(log(death_rate)~log(HC_index) + log(NOx_index),
data=deathrate_dataset)
> plot(model3.3)
```



```
> summary(model3.3)
Call:
lm(formula = death_rate ~ log(HC_index) + log(NOx_index), data =
deathrate_dataset)
Residuals:
     Min       1Q   Median       3Q      Max
-114.832  -34.193    0.556   42.230  149.158
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      942.08      19.46  48.407  < 2e-16 ***
log(HC_index)    -63.97      19.71  -3.245 0.001966 **
log(NOx_index)    75.35      19.61   3.843 0.000308 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 55.75 on 57 degrees of freedom
Multiple R-squared:  0.2238,    Adjusted R-squared:  0.1966
F-statistic: 8.219 on 2 and 57 DF,  p-value: 0.0007304
```

#Make a **scatter plot** of the data.

```
pairs(cbind(HC_index,NOx_index,SOx_index))
```



```
>cor(cbind(HC_index, NOx_index, SOx_index), y=NULL)
           HC_index NOx_index SOx_index
HC_index  1.0000000 0.9840337 0.2822963
NOx_index 0.9840337 1.0000000 0.4100839
SOx_index 0.2822963 0.4100839 1.0000000
> #Check the Variance Inflation Factor for the test of
multicollinearity
> vif(model3)
 HC_index NOx_index SOx_index
71.393408 78.987159  2.718799
> vif(lm(formula = death_rate ~ NOx_index + HC_index, data =
deathrate_dataset))
NOx_index  HC_index
 31.56791  31.56791
```

R Code for Question 4:

```
model1.1=lm(sqrt(death_rate)~ nonwhite + school + jan_temp +
pop_density + precip +SOx_index+jul_temp  , data = deathrate_dataset)
> dataSF=data.frame("nonwhite"=13.7,"school"=12.2, "jan_temp"= 48,
"pop_density"=4253, "precip"=18 , "SOx_index"=86, "jul_temp"= 63)
> predict(model1, dataSF, interval="confidence")
       fit      lwr      upr
1 910.6695 877.9477 943.3913
```