Greatest Contribution to Mortality

By:

James Hizon

Haofeng Zhu

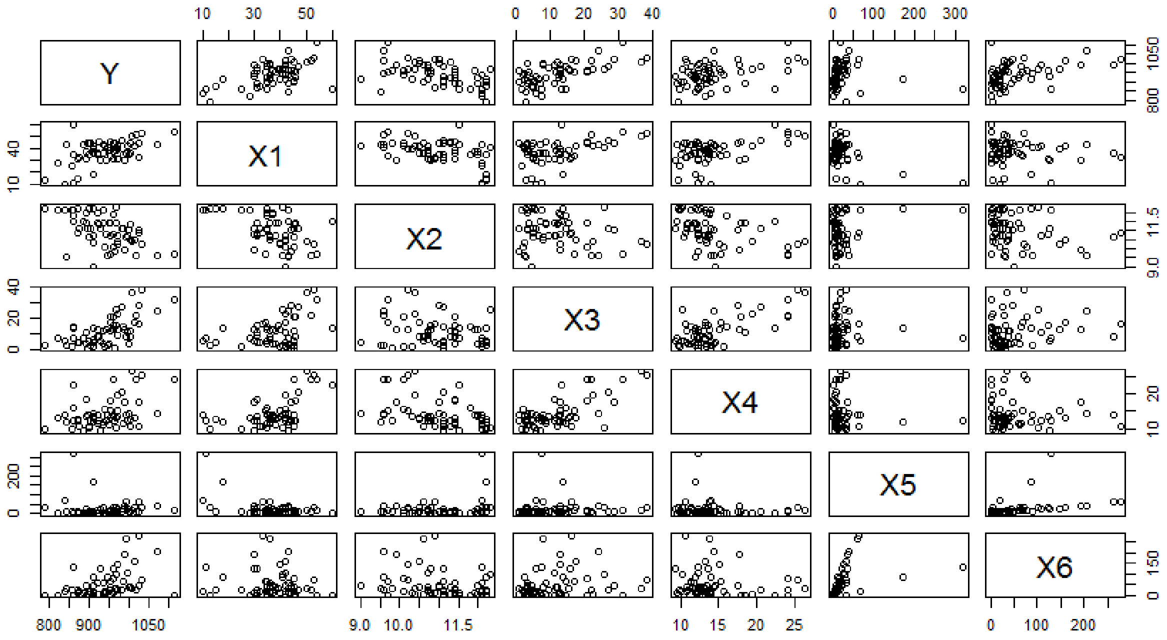
Yan Liang

Introduction:

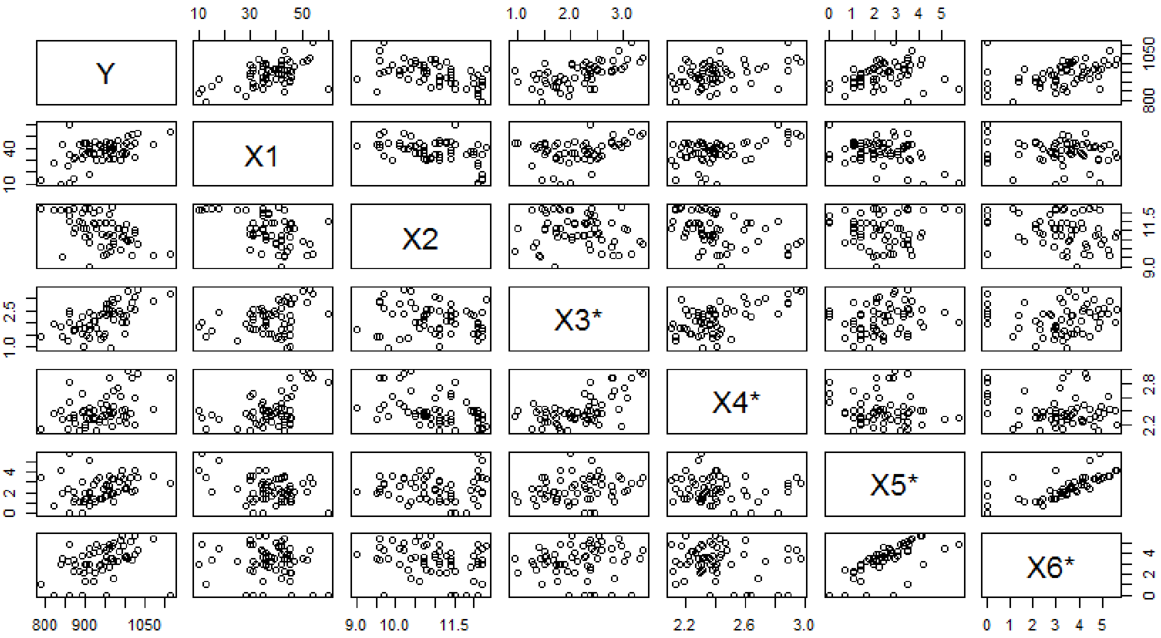
In this project, we will make observations on how certain variables are all connected to increasing, faster mortality rates. Our goal is to test if there is linearity between mortality rates and each of the given independent variables, such as precipitation. Additionally, we want to see if any of the given variables show multicollinearity with other variables. Furthermore, we want to compare which of the factors from the data: precipitation, higher education, ethnicity, socioeconomic status, or pollution, have a higher impact on mortality.

Matrix Plot

Before Transformation:



After Transformation:



Summary/Comment:
In the above matrix, we find that the inclusion of the qualitative variable “City” in the data creates complications beyond our current skills, hence we removed the variable. For further convenience, we replaced all the variables “Mortality”, “Precipitation”, “Education”, “Nonwhite”, “Poor”, “NO_x”, “SO_2” with “Y”, “X_1”, “X_2”, “X_3”, “X_4”, “X_5”, “X_6.”

From the original matrix of scatter plot before any transformation, we find that the linearity between “Y” and each of the “X” variables are not as evident, the scatterplots of “X_5” and “X_6” are skewed since most our observations of “X_5” and “X_6” are inside of the interval relative to the other observed units, which indicates that we may need to do transformations on the “X” variables. Thus, we transform X_3 and X_4 into the cube root, as well as X_5 and X_6 into the natural logarithm. And then we get a new matrix of the scatter plot, from which we can see an evident linearity exists between the Y and the X variables and apparent multicollinearity exist between X_5 and X_6. From now on, we will denote the transformed X_3, X_4, X_5 and X_6 as X_3*, X_4*, X_5* and X_6*.

Correlation Matrix

```
> cor(dat)
           Y          X1          X2        X3*       X4*        X5*        X6*
Y   1.0000000  0.5094924 -0.51098130  0.6063347  0.4099867  0.29199967  0.4031300
X1   0.5094924  1.0000000 -0.49042518  0.3193478  0.4937707 -0.36830267 -0.1211723
```

```
X2  -0.5109813 -0.4904252  1.00000000 -0.1359181 -0.4167899  0.01798472 -0.2561622
X3*  0.6063347  0.3193478 -0.13591810  1.0000000  0.6003373  0.19773000  0.0592199
X4*  0.4099867  0.4937707 -0.41678995  0.6003373  1.0000000 -0.10413526 -0.1955220
X5*  0.2919997 -0.3683027  0.01798472  0.1977300 -0.1041353  1.00000000  0.7328074
X6*  0.4031300 -0.1211723 -0.25616219  0.0592199 -0.1955220  0.73280742  1.0000000
```

Summary/Comment:
In the correlation matrix, the occurrence of multicollinearity is slightly more evident between X_3* and X_4* since the correlation coefficient between them is 0.6003373, which is slightly greater than 0.5. The occurrence of multicollinearity is even higher between X_5* and X_6* since the correlation coefficient is 0.7328074. In other words, X_5* and X_6* are highly related to each other. This makes sense because X_5* (NO_X) and X_6* (SO_2) are both dependent on oxygen, so it affects the air in which people breathe. Toxic chemicals would then lead to shortened life span and increase the mortality rate by which people die.

The correlation coefficients between Y and X_1, X_2 and X_3* variables are relatively larger than the correlation coefficients between X_4*, X_5* and X_6* respectively. This shows that mortality rates seem to be more dependent on whether we are making an observation in areas where we expect to see higher precipitation, have a higher level of education, and are non-white in the 1960s, vs. factors such as being poor or pollution. Poor income families or the condition of living in areas where the air is polluted may have some effect on mortality rates, but just not as much the following: the effect in areas with higher precipitation leading to more crop production implying less people dying from starvation, having the right education on why murder is wrong or keeping people off the streets as well as understanding how to prevent people dying from diabetes, obesity or diseases like sexually transmitted diseases like STDs and HIVs worldwide, teen suicide prevention, etc.

Summary fit:

```
> summary(fit)

Call:
lm(formula = Y ~ ., 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
```

Summary/Comment:

From the summary fit, we achieve the multiple regression function as:

Residuals = Y-ŷ

$\hat{y} = 980.4750+2.3748X_1+(-19.1004)X_2+49.9051X_3*+(-31.0975)X_4*+10.1044X_5*+8.0315X_6*$

The estimates of β_0, β_1, β_2, ${}_3$, B_4, B_5, and B_6 are 980.4750, 2.3748,-19.1004, 49.9051, -31.0975,10.1044, and 8.0315 and their respective standard errors, p-values, and their t-values. We can also see that the p-values for “X_1”, “X_2”, “X_3*” are within our alpha level 0.05, and “X_4*” , “X_5*” and ”X_6*” exceed our alpha level of 0.05. We can conclude that the coefficients of “X_1”, “X_2”, “X_3*” and are not equal to zero and “X_4*” ,“X_5*” and “X_6*”are equal to zero individually(not at the same time equal to zero or not equal to zero) under the alpha level of 0.05.

The p-value corresponding to the F-statistic is 3.139e-12, which indicates that not all of the coefficients of the X variables are equal to zero.

The R-squared in our summary fit is 0.6985 and the adjusted R-squared is 0.6644, these two values are relatively low, which indicates that we may need to select a better fitting model.

ANOVA Table:

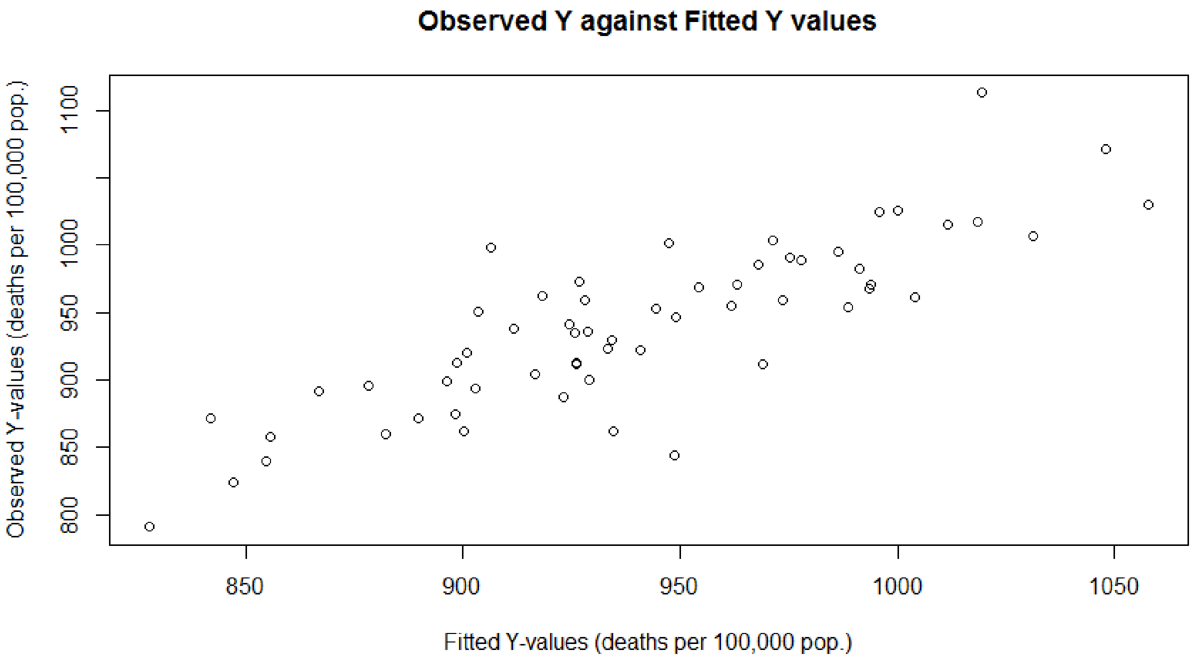
```
> anova(fit)
Analysis of Variance Table

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/Comment:

From the ANOVA table, we obtained sequential sum of squares of “X_1”, “X_2”, “X_3*”, “X_4*”, “X_5*”, “X_6*” and the F-values and p-values. The F-values help us decide whether the coefficient of the current X variable is equal to zero based on the assumptions of the previous existing variable. We see that sequential inclusion of the variables “X_1”,”X_2”,”X_3*”, “X_4*”, “X_5*” is acceptable since their p-value is within alpha level of 0.1, while inclusion of the variable X_6* does not help truly improve the quality of the fit given “X_1”...“X_5*” already exist in the model. This is also consistent with our findings in matrix of scatter plot and correlation matrix that there exist apparent multicollearity between “X_5*” and “X_6*.”

<u>Observed Y-values vs. Fitted Y-values:</u>



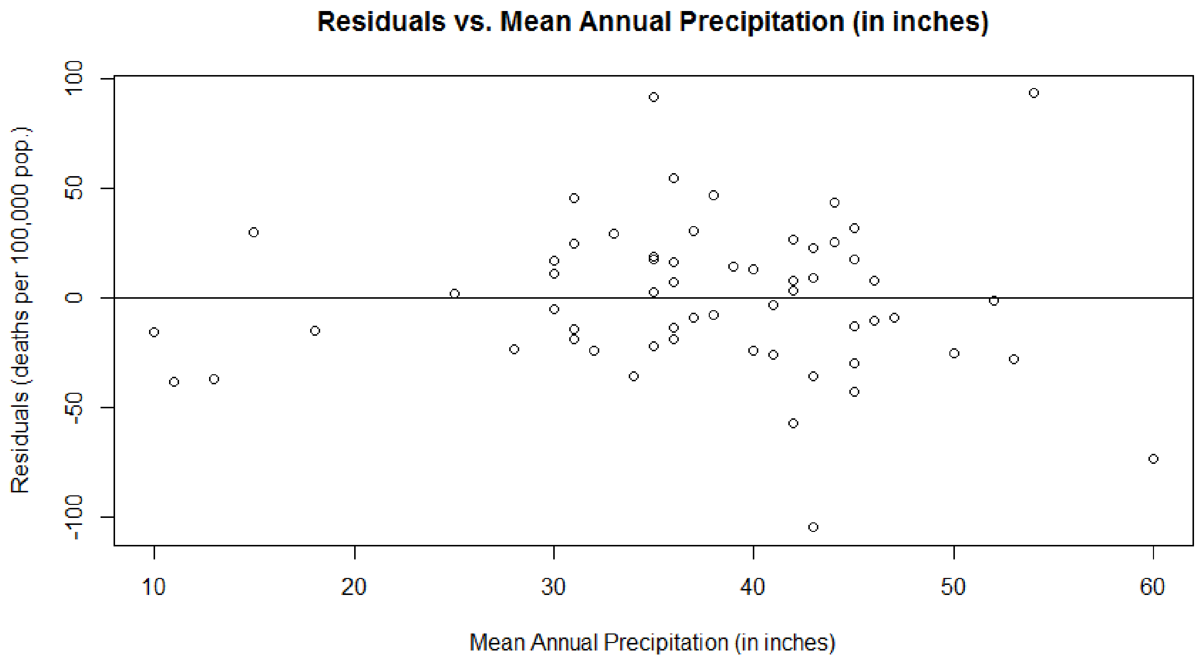
Summary/Comment:

From the fitted regression above, we see that the observed residuals can be fitted into a linear regression indicating the high positive correlation between “Y” and fitted “Y”. In terms of mortality rates, since it does not lie perfectly close to a perfect line Y=X, we should expect a few errors in our model since if our model is perfectly right, the only thing that will cause the dots in the plot deviate from the line Y=X is the existence of error terms.

The few errors expressed would imply that there may be several other factors thrown in the bag that will lessen lifespan that we should include as variables.

Residuals against Each Independent X Variable

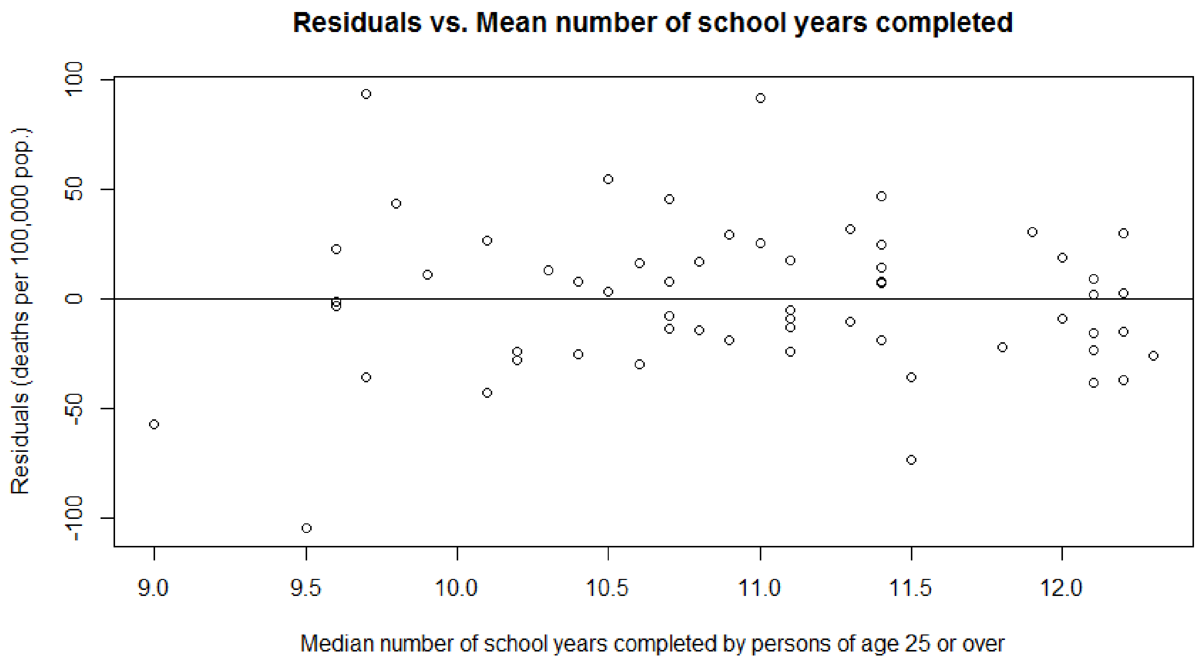
X1:



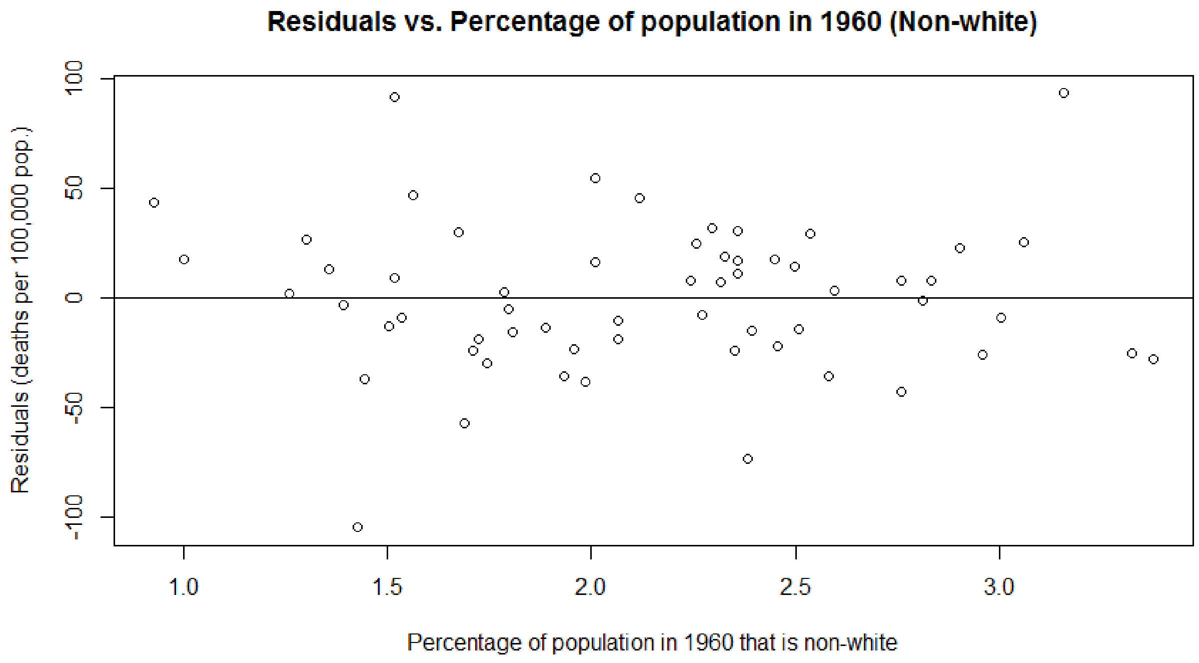
Summary/Comment:

In the above scatterplot, we see that the plotted residuals against the "Mean Annual Precipitation" are scattered, showing little to no correlations between residuals and the mean annual precipitation.

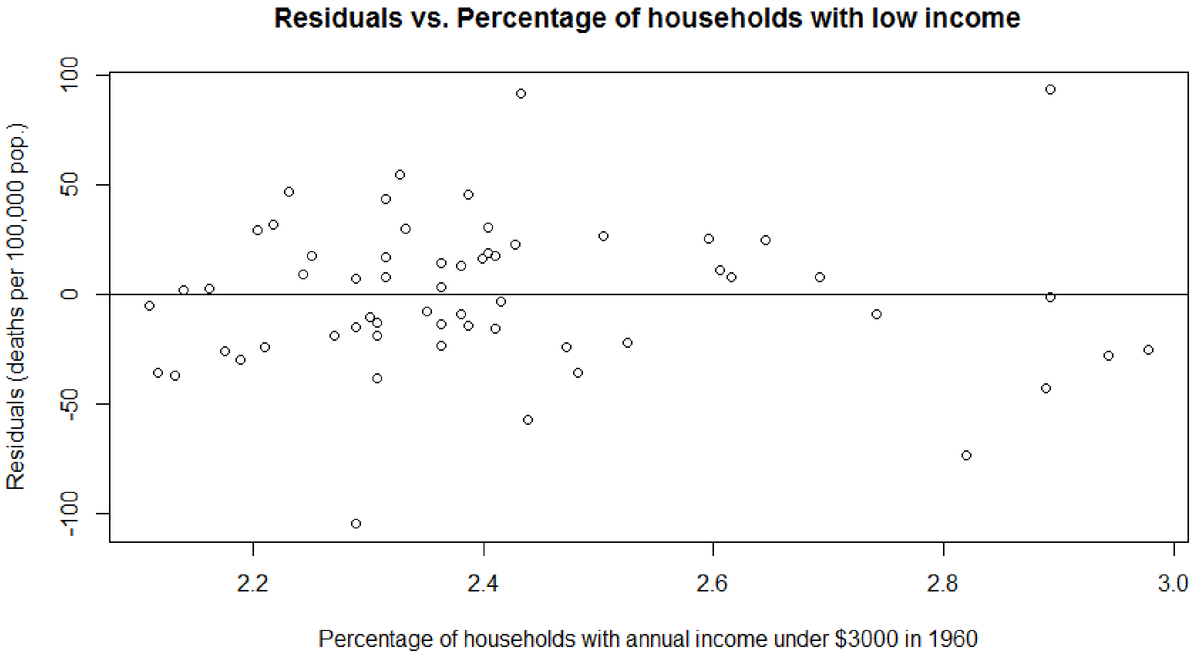
X2:


Residuals vs. Mean number of school years completed
Residuals (deaths per 100,000 pop.)
100
50
0
-50
-100
9.0
9.5
10.0
10.5
11.0
11.5
12.0
Median number of school years completed by persons of age 25 or over

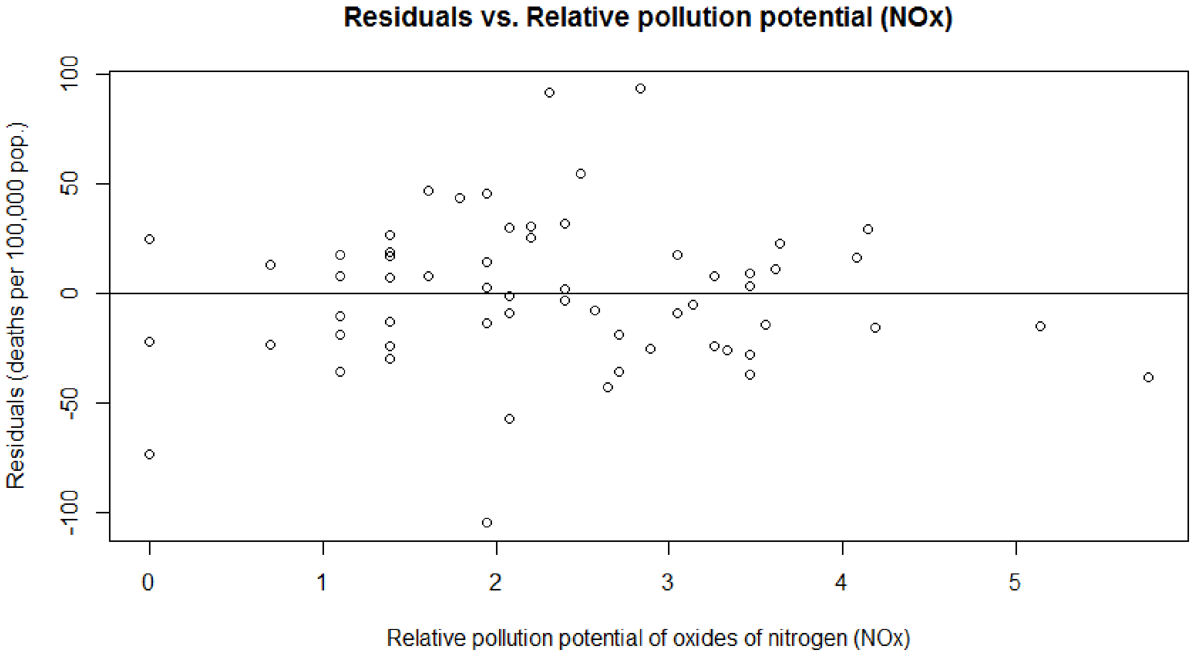
X3:


Residuals vs. Percentage of population in 1960 (Non-white)
Residuals (deaths per 100,000 pop.)
100
50
0
-50
-100
1.0
1.5
2.0
2.5
3.0
Percentage of population in 1960 that is non-white

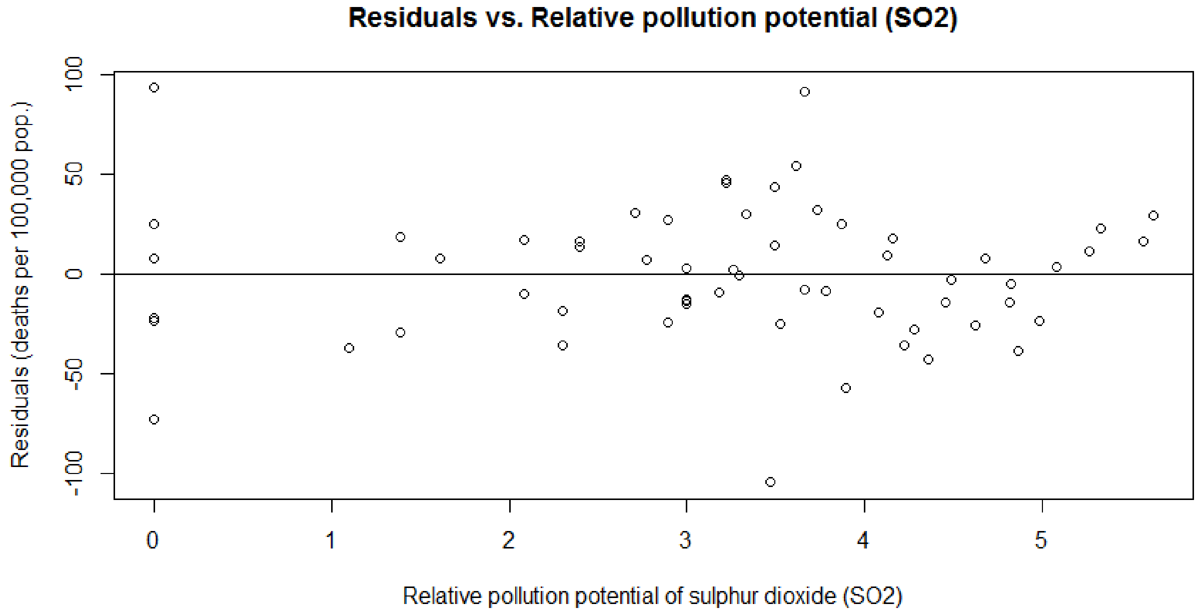
X4:


Residuals vs. Percentage of households with low income
Residuals (deaths per 100,000 pop.)
100
50
0
-50
-100
2.2
2.4
2.6
2.8
3.0
Percentage of households with annual income under $3000 in 1960

X5:


Residuals vs. Relative pollution potential (NOx)
Residuals (deaths per 100,000 pop.)
100
50
0
-50
-100
0
1
2
3
4
5
Relative pollution potential of oxides of nitrogen (NOx)

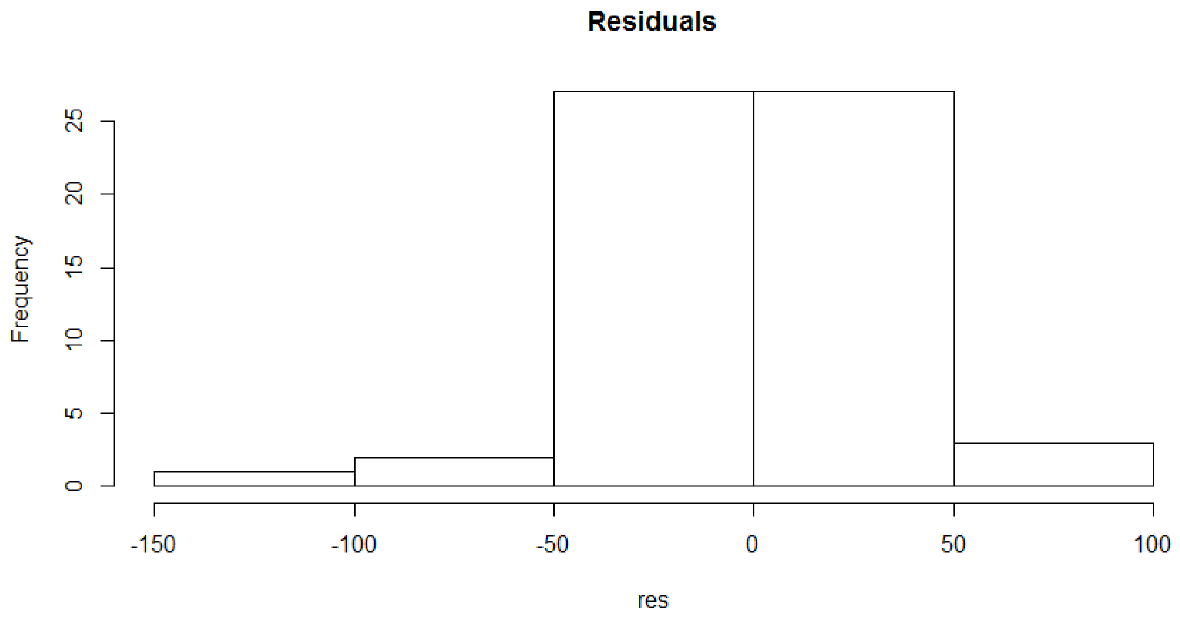
X6:



Summary/Comment:

From all the plots above, we see that the model is not so satisfactory since there are quite a few apparent outliers in each of the plot. The assumption of equal variances of error terms in relation to each individual X variable is acceptable here.

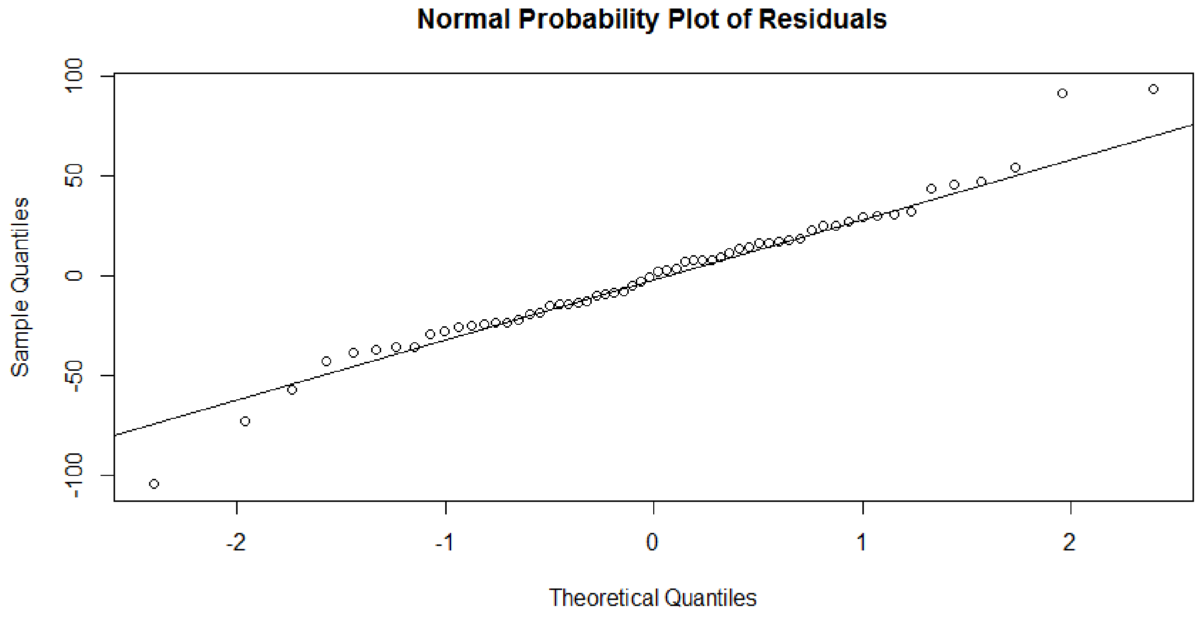
Histogram of Residuals:



Summary/Comment:

From the histogram of residuals above, we see that the assumptions of a normal distribution in error terms seem to be invalid. The histogram does not look bell-shaped or symmetrical.

<u>Normal Probability Plot</u>



Summary/Comment:

From the normal probability plot of residuals above, see that most of the dots line up perfectly close to a straight line, which indicates that our assumption of normality of error terms seems acceptable.

Note that the conclusions drawn from the histogram of residuals and normal probability plot of residuals are in conflict with each other. The reason may be that the sample size is not sufficiently large here and there may exist some error in the data which causes the difference of the two conclusions.

Including Non-linear terms:

From the matrix of scatter plot we obtained in Step 2, we see that there seems to be one bump in the relationship between Y and X_2, Y and X_3^*, which indicates that we may add quadratic terms of X_2 and X_3^* in our model. Moreover, from the correlation matrix we obtained in Step 2, we see that the correlation between Y and X_5^* is quite small, which indicates we may add quadratic terms of X_5^* as well.

When including quadratic terms of X_2,X_3^* and X_5^*, we need to centralize them into x_2, x_3^* and x_5^* and also use $x_2^{2,}x_3^*$^2 and x_5^*^2 in order to prevent potential computational risks.

Summary Fit:

```
> summary(fit)

Call:
lm(formula = Y ~ ., data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-87.502 -23.177   2.079  18.685  82.074

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 936.0585    91.6299  10.216  7.8e-14 ***
X1            1.7341     0.6819   2.543 0.014129 *
x2          -21.8477     7.5461  -2.895 0.005604 **
x2.square   -15.8961     6.2963  -2.525 0.014800 *
`x3*`        46.2933    11.5794   3.998 0.000211 ***
`x3*square`  11.3636    14.4690   0.785 0.435936
```

```
`X4*`        -23.2883    35.8572  -0.649 0.519005
`x5*`         17.6020     7.5352   2.336 0.023544 *
`x5*square`  -5.1631      2.5485  -2.026 0.048128 *
`X6*`          3.1981     5.5020   0.581 0.563677
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error: 33.53 on 50 degrees of freedom
Multiple R-squared:  0.7537,  Adjusted R-squared:  0.7094
F-statistic:    17 on 9 and 50 DF,  p-value: 2.148e-12
```

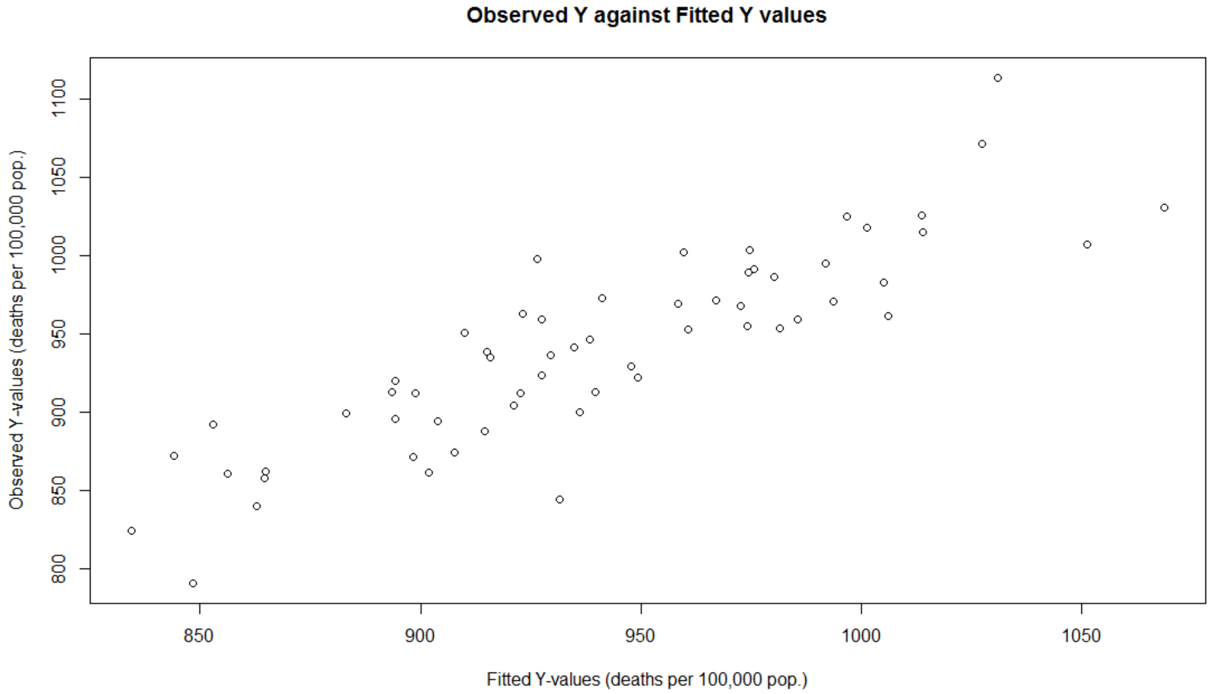
Summary/Comment:
We see that the Adjusted R-squared has increased from 0.6644 to 0.7094, which indicates that our inclusion of quadratic terms of X_2,X_3^* and X_5^* have improved the quality of the fit significantly. Note that the p-value of x_3^*^2, X_4^* and X_6^* is much larger the the common alpha level 0.05, which indicates that these three variables are most likely to be excluded from our model. We will see that happening in our Best Subsets Method and Stepwise Regression Method in Step 5.

ANOVA Model:

```
> anova(fit)
Analysis of Variance Table

Response: Y
            Df Sum Sq Mean Sq F value    Pr(>F)
X1           1  59256   59256 52.7021 2.357e-09 ***
x2           1  20492   20492 18.2261 8.747e-05 ***
x2.square    1   7272    7272  6.4677  0.014126 *
`x3*`        1  48956   48956 43.5412 2.549e-08 ***
`x3*square`  1   2371    2371  2.1089  0.152691
`X4*`        1  10286   10286  9.1481  0.003924 **
`x5*`        1  17593   17593 15.6473  0.000241 ***
`x5*square`  1   5450    5450  4.8474  0.032330 *
`X6*`        1    380     380  0.3379  0.563677
Residuals   50  56218    1124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Summary/Comment:
From the ANOVA table, we obtained sequential sum of squares of “X_1”, “x_2”, “x_2^2”,“x_3*”,”x_3*^2” “X_4*”, “x_5*”,“x_5*^2”, “X_6*” and the F-values and P-values. The F-values help us decide whether the coefficient of the current X variable is equal to zero based on the assumptions of the previous existing variable. We can conclude that the coefficient of x_3*^2 is equal to zero in the existence of “X_1”, “x_2”, “x_2^2” and “x_3*” under alpha level 0.1. We can also conclude that the coefficient of X_6* is equal to zero given that “X_1”, “x_2”, “x_2^2”,“x_3*”,”x_3*^2” “X_4*”, “x_5*” and “x_5*^2” are already in the model under alpha level 0.1.

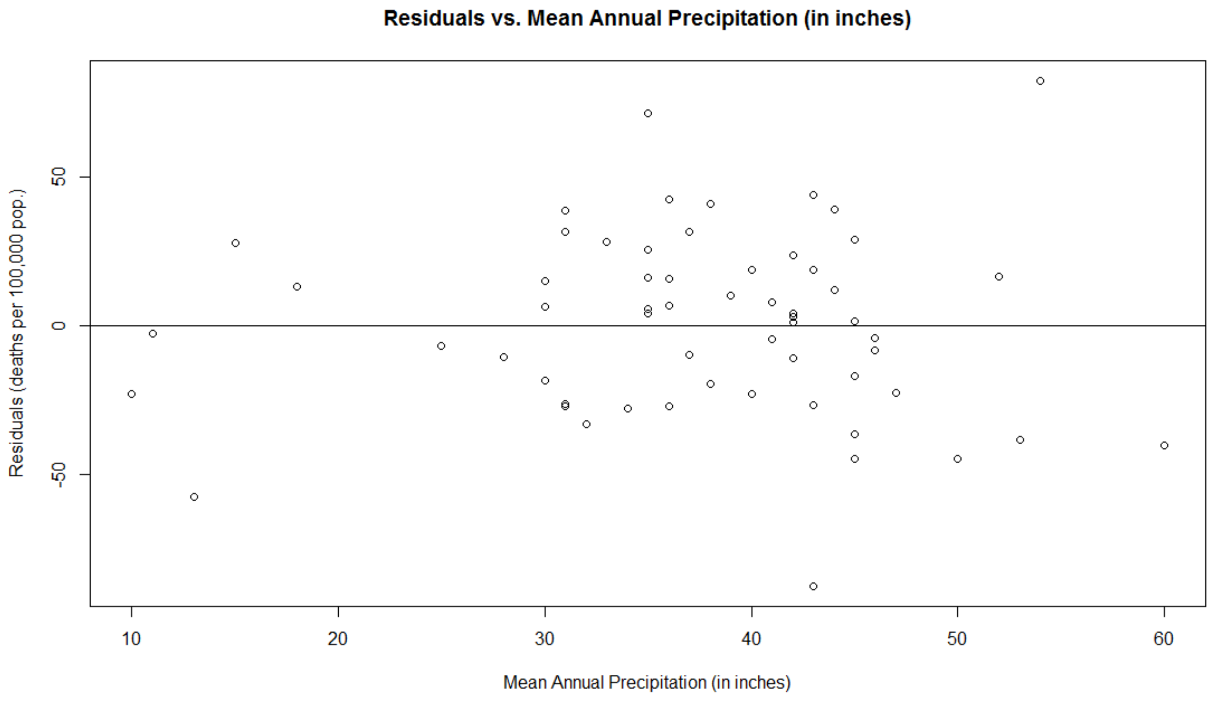
Observed Y against Fitted Y Values:



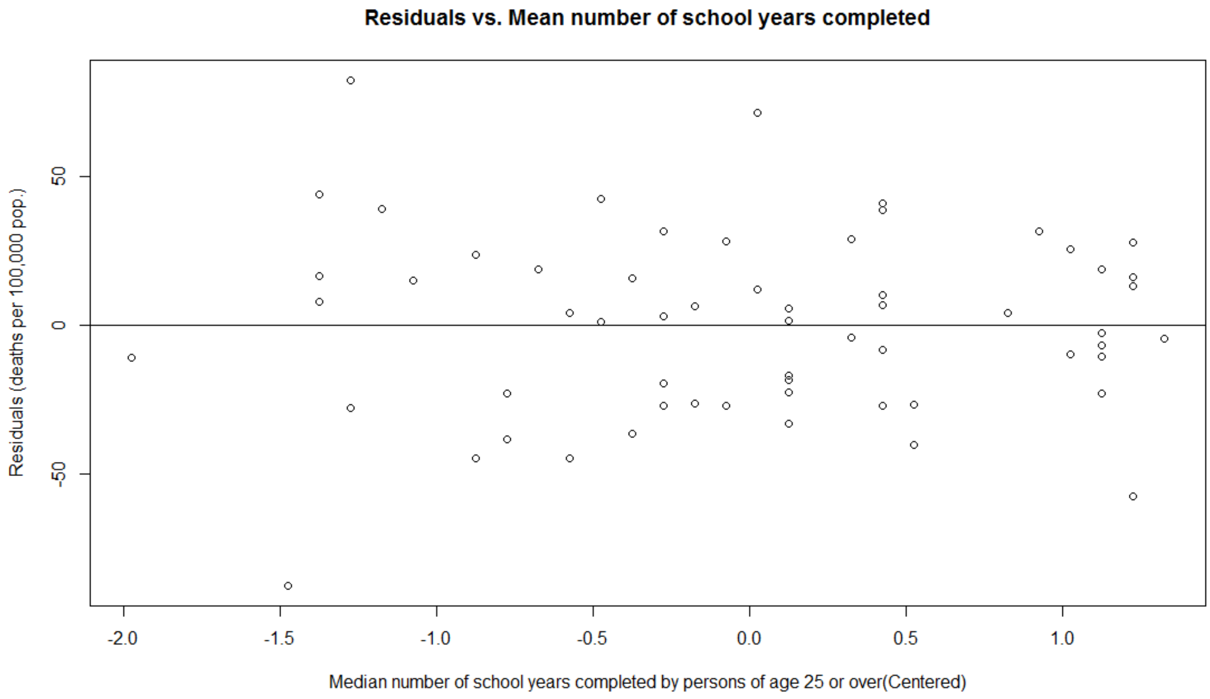
Summary/Comment:
This time, we see that the observed Y values against fitted Y values fit closer to the line Y=X, which indicates that the quality of the fit has been improved. However, it seems there are still one or two outliers.

Plot of Residuals vs. X variables:

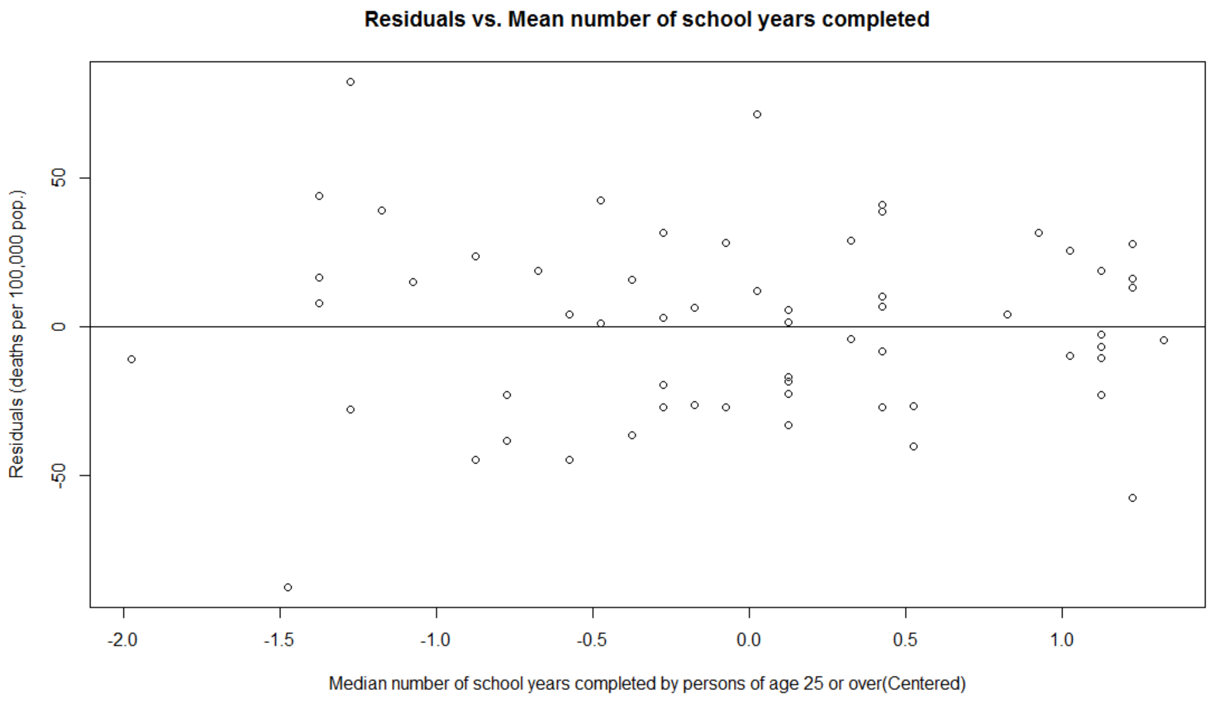
X1:


Residuals vs. Mean Annual Precipitation (in inches)
Residuals (deaths per 100,000 pop.)
50
0
-50
10
20
30
40
50
60
Mean Annual Precipitation (in inches)

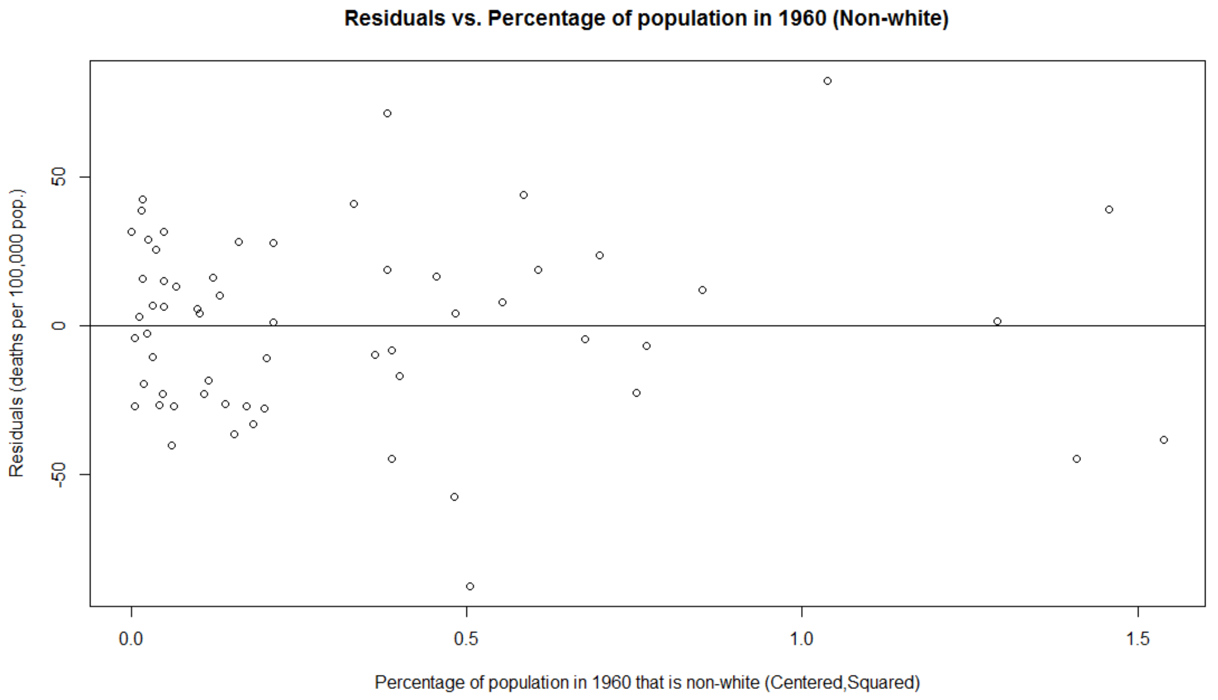
X2:


Residuals vs. Mean number of school years completed
Residuals (deaths per 100,000 pop.)
50
0
-50
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
Median number of school years completed by persons of age 25 or over(Centered)

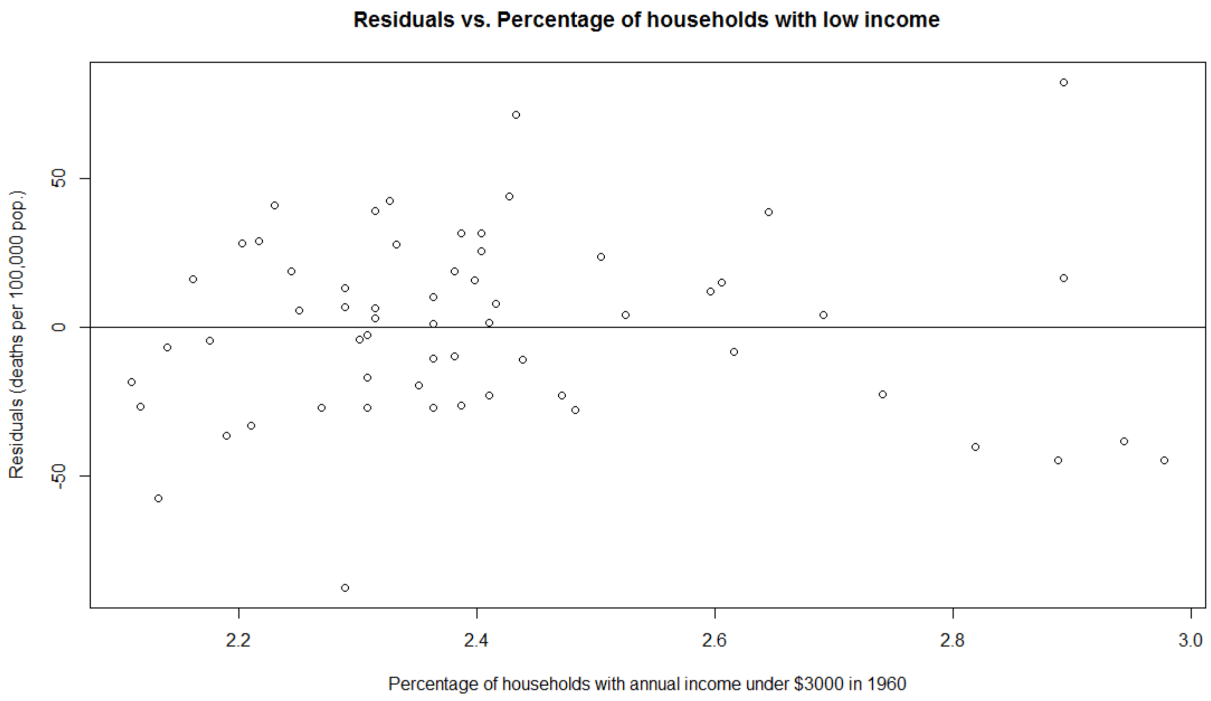
X3:


Residuals vs. Mean number of school years completed
Residuals (deaths per 100,000 pop.)
50
0
-50
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
Median number of school years completed by persons of age 25 or over(Centered)

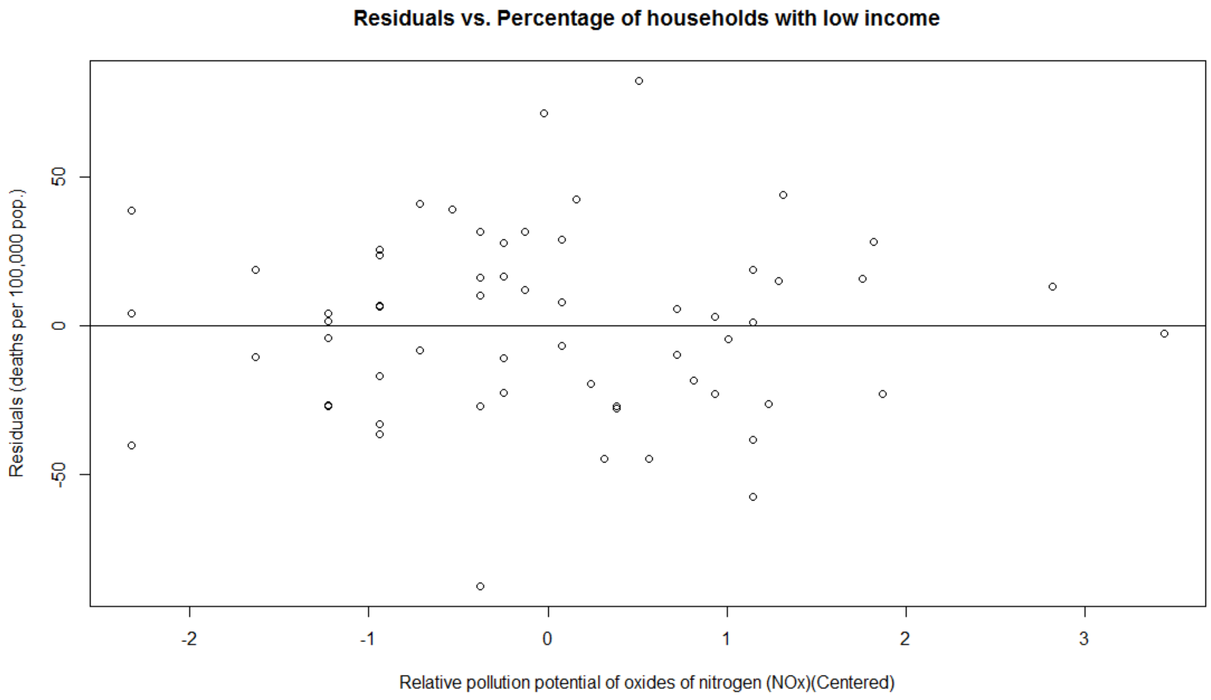
X3^2:


Residuals vs. Percentage of population in 1960 (Non-white)
Residuals (deaths per 100,000 pop.)
50
0
-50
0.0
0.5
1.0
1.5
Percentage of population in 1960 that is non-white (Centered,Squared)

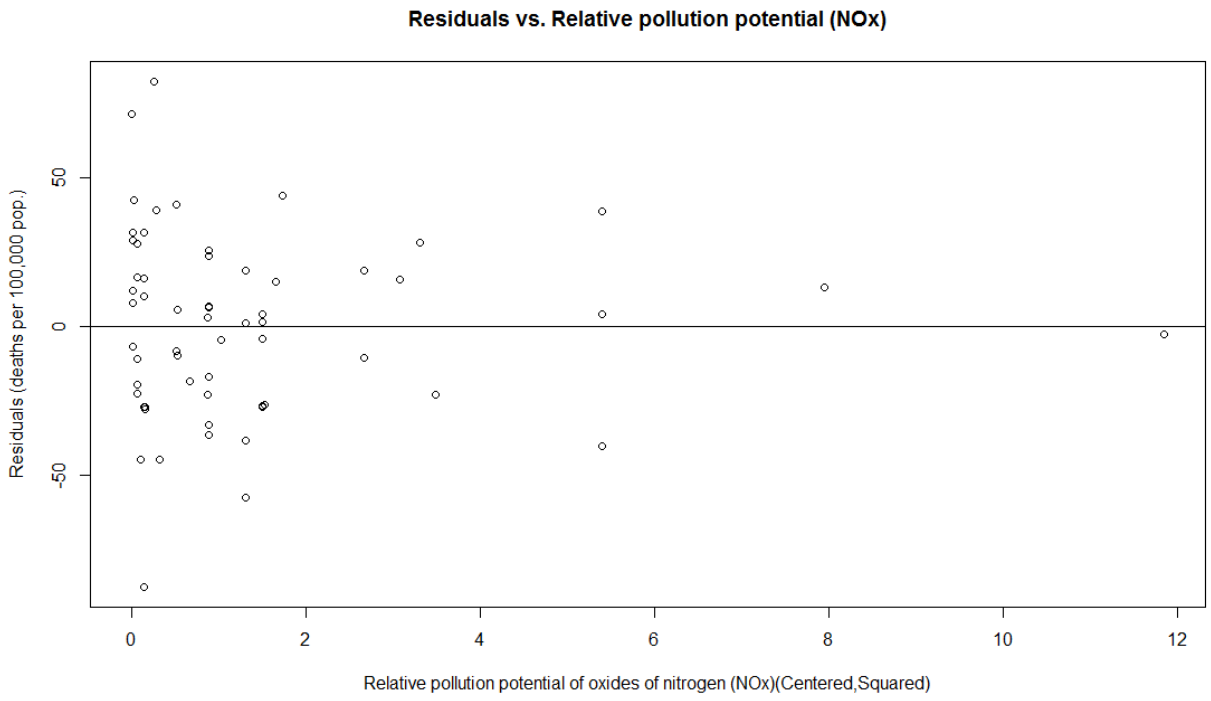
X4:


Residuals vs. Percentage of households with low income
Residuals (deaths per 100,000 pop.)
50
0
-50
2.2
2.4
2.6
2.8
3.0
Percentage of households with annual income under $3000 in 1960

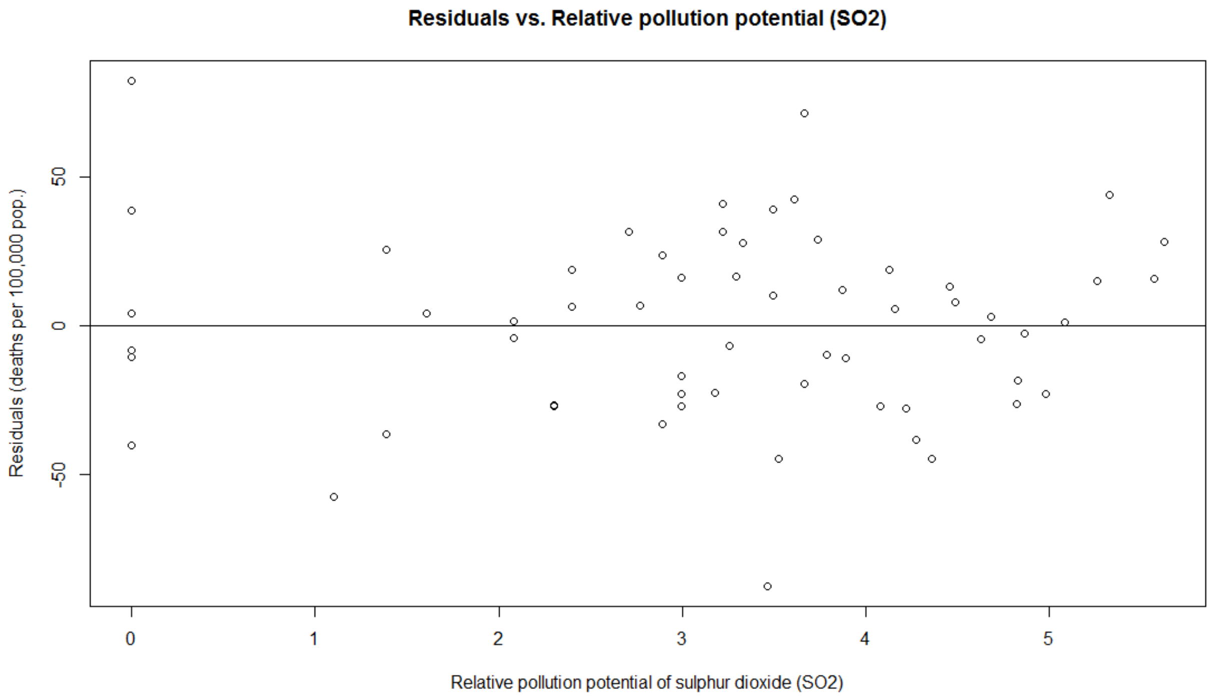
X5:


Residuals vs. Percentage of households with low income
Residuals (deaths per 100,000 pop.)
50
0
-50
-2
-1
0
1
2
3
Relative pollution potential of oxides of nitrogen (NOx)(Centered)

X5^2:


Residuals vs. Relative pollution potential (NOx)
Residuals (deaths per 100,000 pop.)
50
0
-50
0
2
4
6
8
10
12
Relative pollution potential of oxides of nitrogen (NOx)(Centered,Squared)

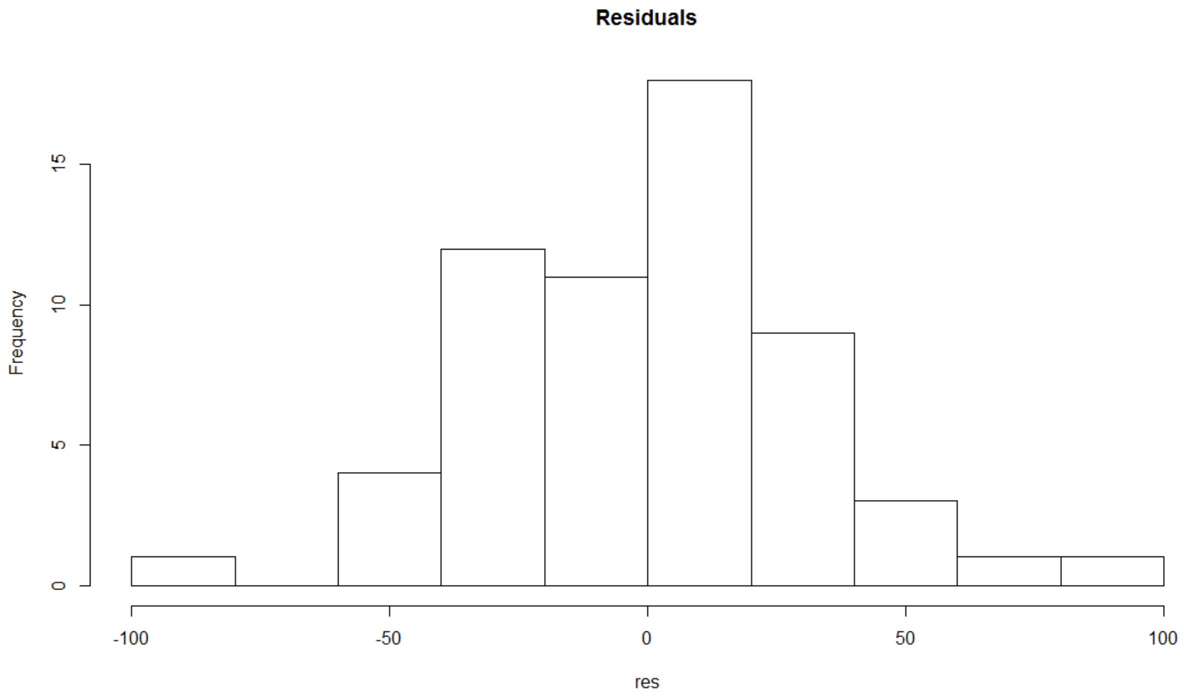
X6:


Residuals vs. Relative pollution potential (SO2)
Residuals (deaths per 100,000 pop.)
50
0
-50
0
1
2
3
4
5
Relative pollution potential of sulphur dioxide (SO2)

Summary/Comment:

We see that all of the plots of the residuals against each individual X variable have improved and there are no apparent violations of the linearity between Y and each individual X variables. Furthermore, the assumption of equal variances of error terms in relation to each individual X variable appears to be acceptable. However, it seems there are still quite a few outliers in the plot related to $X_3^*^{2,}X_5^*$^2 and $X_{6.}$

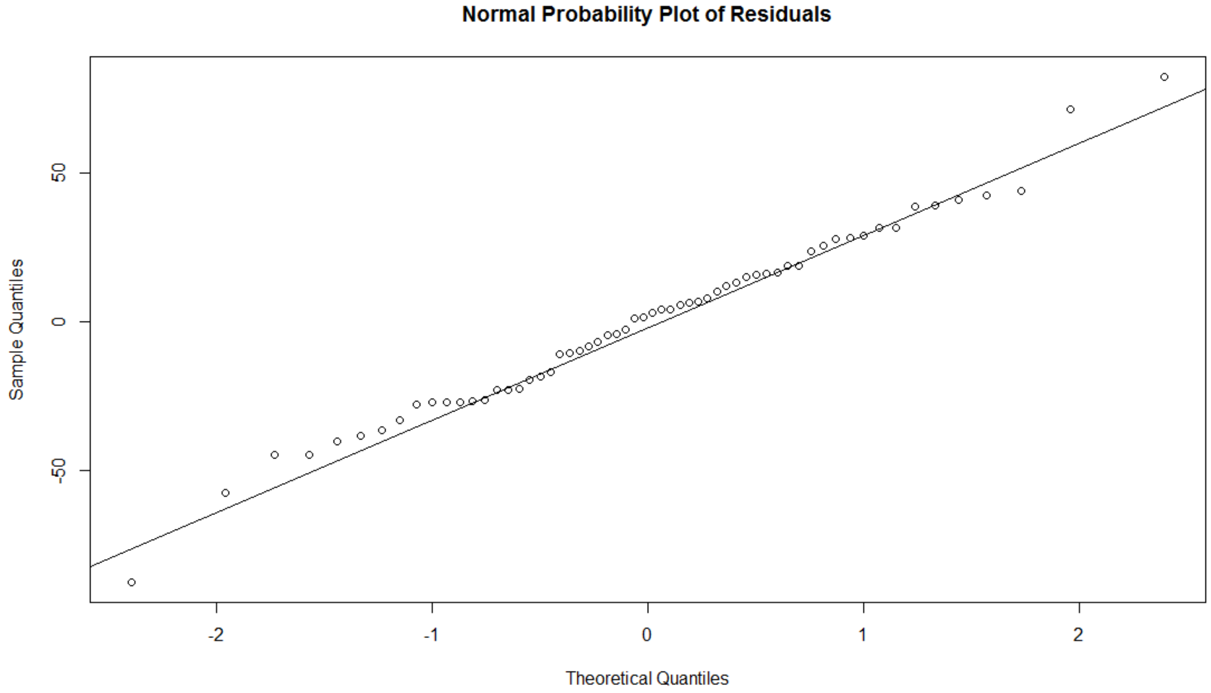
Histogram of Residuals:



Summary/Comment:

It appears that the histogram of residuals looks much more bell-shaped and relatively more symmetrical than the previous histogram. Thus, we can conclude our inclusion of quadratic terms have improved the quality of the fit by a substantial amount.

Normal Probability Plot:



Summary/Comment:
From the normal probability plot of residuals above, we can see that most of the dots lie perfectly close to a straight line, which indicates that our assumption of normality of error terms appears to be acceptable.

Best Subset Model:

```
> fit$which[ind,]
     1      2      3           4    5          6       7       8       9
6  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  FALSE
```

From using Mallows' Cp as our criterion for our best subset model, we can delete X_3*squared,X_4^* and X_6* from our model.

Backwise Stepwise Regression:

```
Start:  AIC=430.56
Y ~ X1 + x2 + x2.square + `x3*` + `x3*square` + `X4*` + `x5*` +
    `x5*square` + `X6*`

              Df Sum of Sq   RSS    AIC F value    Pr(>F)
- `X6*`        1     379.9 56597 428.96  0.3379 0.5636766
- `X4*`        1     474.3 56692 429.06  0.4218 0.5190048
- `x3*square`  1     693.5 56911 429.29  0.6168 0.4359363
<none>                     56218 430.56
- `x5*square`  1    4614.6 60832 433.29  4.1042 0.0481284 *
- `x5*`        1    6135.3 62353 434.77  5.4567 0.0235437 *
- x2.square    1    7166.6 63384 435.76  6.3740 0.0147997 *
- X1           1    7271.4 63489 435.86  6.4672 0.0141290 *
- x2           1    9424.6 65642 437.86  8.3823 0.0056042 **
- `x3*`        1   17970.9 74188 445.20 15.9834 0.0002106 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Step:  AIC=428.96
Y ~ X1 + x2 + x2.square + `x3*` + `x3*square` + `X4*` + `x5*` +
    `x5*square`

              Df Sum of Sq   RSS    AIC F value    Pr(>F)
- `x3*square`  1     605.3 57203 427.60  0.5454 0.4635832
- `X4*`        1     760.0 57357 427.76  0.6849 0.4117731
<none>                     56597 428.96
+ `X6*`        1     379.9 56218 430.56  0.3379 0.5636766
- `x5*square`  1    5450.2 62048 432.48  4.9111 0.0311723 *
- X1           1    7918.0 64515 434.82  7.1349 0.0101205 *
- x2.square    1    8297.7 64895 435.17  7.4770 0.0085718 **
- x2           1   13262.6 69860 439.59 11.9509 0.0011102 **
- `x3*`        1   17712.6 74310 443.30 15.9608 0.0002084 ***
- `x5*`        1   21070.0 77667 445.95 18.9862 6.379e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Step:  AIC=427.6
```

```
Y ~ X1 + x2 + x2.square + `x3*` + `X4*` + `x5*` + `x5*square`

             Df Sum of Sq   RSS    AIC F value    Pr(>F)
- `X4*`        1     346.9 57550 425.96  0.3153 0.5768431
<none>                     57203 427.60
+ `x3*square`  1     605.3 56597 428.96  0.5454 0.4635832
+ `X6*`        1     291.6 56911 429.29  0.2613 0.6114101
- `x5*square`  1    6546.8 63750 432.10  5.9514 0.0181531 *
- x2.square    1    7912.8 65115 433.37  7.1931 0.0097861 **
- X1           1   10413.7 67616 435.64  9.4665 0.0033342 **
- x2           1   12712.7 69915 437.64 11.5565 0.0013036 **
- `x3*`        1   17705.0 74908 441.78 16.0947 0.0001937 ***
- `x5*`        1   25749.6 82952 447.90 23.4076  1.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Step:  AIC=425.96
Y ~ X1 + x2 + x2.square + `x3*` + `x5*` + `x5*square`

             Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>                     57550 425.96
+ `X6*`        1     488.7 57061 427.45  0.4454  0.507484
+ `X4*`        1     346.9 57203 427.60  0.3153  0.576843
+ `x3*square`  1     192.1 57357 427.76  0.1742  0.678146
- `x5*square`  1    7961.8 65511 431.74  7.3323  0.009097 **
- x2.square    1    8694.2 66244 432.41  8.0069  0.006567 **
- X1           1   10070.7 67620 433.64  9.2746  0.003615 **
- x2           1   12779.2 70329 436.00 11.7690  0.001174 **
- `x3*`        1   22787.2 80337 443.98 20.9858 2.845e-05 ***
- `x5*`        1   29288.4 86838 448.65 26.9730 3.347e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Call:
lm(formula = Y ~ X1 + x2 + x2.square + `x3*` + `x5*` + `x5*square`,
    data = dat2)

Coefficients:
(Intercept)           X1           x2    x2.square        `x3*`        `x5*`
```

```
  890.620      1.876     -21.621     -16.760      39.319      22.712
`x5*square`
   -6.233
```

Summary:

Here we performed stepwise regression on backwards elimination. In backwards elimination, we observed the F-to-remove values less than 4 and the least value using AIC Criterion. And, we checked that this gives the same best model that we did using best subset regression. So, we find confirmation that the given values portray the least correlation vs. the other given variables.

Possible Further Analysis:

We may also want to include the qualitative variable "City" in our model. Also, the interaction terms between X variables have not been included in our model yet. That is really important because precipitation may have effects on the toxicity of oxides of nitrogen and sulfur dioxide.

In cloud formation, precipitation may play a huge role, due to acid rain, and may even affect vegetation and pesticides in the food that people eat. There is a huge connection, where sickness may be found in the mist to lead to disease and ultimately, face death.

Appendix

```
#Project

setwd("C:/Users/James/Desktop/STA 108")

dat=read.csv('mortality2.csv', header = TRUE)

View(dat)


#Y=dat
names(dat) = c('Y', 'X1', 'X2', 'X3', 'X4','X5','X6')
head(dat)

dat$X3 = (dat$X3)^(1/3)

dat$X4 = (dat$X4)^(1/3)

View(dat)

dat$X5 = log(dat$X5)

dat$X6 = log(dat$X6)




names(dat) = c('Y', 'X1', 'X2', 'X3*', 'X4*','X5*','X6*')


plot(dat)

cor(dat)

#fit = lm(Y ~ ., data = dat)
fit=lm(Y ~ ., data = dat)
```

```
summary(fit)
anova(fit)

fit$coef

fit$fitted

fit$res

res = fit$res

#Observed Y against Fitted Y values


plot(fit$fitted.values, dat$Y, main = 'Observed Y against Fitted Y values',
    xlab = 'Fitted Y-values (deaths per 100,000 pop.)',
    ylab = 'Observed Y-values (deaths per 100,000 pop.)');

#Residuals against independent each X variable

plot(dat$X1, res, main = 'Residuals vs. Mean Annual Precipitation (in inches)',
    xlab = 'Mean Annual Precipitation (in inches)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline(h=0)

plot(dat$X2, res, main = 'Residuals vs. Mean number of school years completed',
    xlab = 'Median number of school years completed by persons of age 25 or over',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

plot(dat$X3, res, main = 'Residuals vs. Percentage of population in 1960 (Non-white)',
    xlab = 'Percentage of population in 1960 that is non-white',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

plot(dat$X4, res, main = 'Residuals vs. Percentage of households with low income',
    xlab = 'Percentage of households with annual income under $3000 in 1960',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

plot(dat$X5, res, main = 'Residuals vs. Relative pollution potential (NOx)',
    xlab = 'Relative pollution potential of oxides of nitrogen (NOx)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

plot(dat$X6, res, main = 'Residuals vs. Relative pollution potential (SO2)',
    xlab = 'Relative pollution potential of sulphur dioxide (SO2)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

#Histogram

hist(res, main = 'Residuals')

qqnorm(fit$res, main = 'Normal Probability Plot of Residuals')
qqline(fit$res)

#Add non-linear terms to model (don't need interaction terms).

x2=dat$X2-mean(dat$X2)
x2.square=x2^2
x3=dat$X3-mean(dat$X3)
x3.square=x3^2
x5=dat$X5-mean(dat$X5)
x5.square=x5^2

dat2=cbind(dat[,1:2],x2,x2.square,x3,x3.square,dat[,5],x5,x5.square,dat[,7])
names(dat2)=c('Y','X1','x2','x2.square','x3*','x3*square', 'X4*', 'x5*','x5*square' ,'X6*')
View(dat2)


#Y.quad=Y+dat$X1^(1/3)+dat$X2^2
#View(Y.quad)
#plot(Y.quad)

fit=lm(Y ~ ., data = dat2)
#X5.square=(dat[,6])^2
#dat2=data.frame(dat,X5.square)

summary(fit)
anova(fit)
```

```
fit$coef

fit$fitted

fit$res

res = fit$res

#Observed Y against Fitted Y values


plot(fit$fitted.values, dat2$Y, main = 'Observed Y against Fitted Y values',
    xlab = 'Fitted Y-values (deaths per 100,000 pop.)',
    ylab = 'Observed Y-values (deaths per 100,000 pop.)');



#Residuals against independent each X variable



plot(dat2$X1, res, main = 'Residuals vs. Mean Annual Precipitation (in inches)',
    xlab = 'Mean Annual Precipitation (in inches)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline(h=0)

plot(dat2$x2, res, main = 'Residuals vs. Mean number of school years completed',
    xlab = 'Median number of school years completed by persons of age 25 or over(Centered)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

plot(dat2$x2.square, res, main = 'Residuals vs. Mean number of school years completed',
    xlab = 'Median number of school years completed by persons of age 25 or over (Centered and
Squared)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)


plot(dat2$'x3*', res, main = 'Residuals vs. Percentage of population in 1960 (Non-white)',
    xlab = 'Percentage of population in 1960 that is non-white(Centered)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)
```

```
plot(dat2$'x3*square', res, main = 'Residuals vs. Percentage of population in 1960 (Non-white)',
    xlab = 'Percentage of population in 1960 that is non-white (Centered,Squared)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

plot(dat2$'X4*', res, main = 'Residuals vs. Percentage of households with low income',
    xlab = 'Percentage of households with annual income under $3000 in 1960',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)

plot(dat2$'x5*', res, main = 'Residuals vs. Percentage of households with low income',
    xlab = 'Relative pollution potential of oxides of nitrogen (NOx)(Centered)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)


plot(dat2$'x5*square', res, main = 'Residuals vs. Relative pollution potential (NOx)',
    xlab = 'Relative pollution potential of oxides of nitrogen (NOx)(Centered,Squared)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)


plot(dat2$'X6*', res, main = 'Residuals vs. Relative pollution potential (SO2)',
    xlab = 'Relative pollution potential of sulphur dioxide (SO2)',
    ylab = 'Residuals (deaths per 100,000 pop.)'); abline (h=0)


#Histogram

hist(res, main = 'Residuals')

qqnorm(fit$res, main = 'Normal Probability Plot of Residuals')
qqline(fit$res)

library('leaps')

fit = leaps(dat2[,-1], dat2[,1], method = 'Cp')
fit



ind = order(fit$Cp, decreasing = FALSE)
fit$Cp
```

```
ind
fit$Cp[ind]
fit$which[ind,]
#Stepwise Backward
#library(MASS)
fit = lm(Y ~., data = dat2)
fit
dropterm(fit, scope = terms(fit), test = 'F')
fit = update(fit, Y ~ . - `X6*`)
fit
dropterm(fit, scope = terms(fit), test = 'F')
fit = update(fit, Y ~ . - `x3*square`)
fit
dropterm(fit, scope = terms(fit), test = 'F')
fit = update(fit, Y ~ . - `X4*`)
fit
dropterm(fit, scope = terms(fit), test = 'F')
#Stop here. Conclude that we only delete X6*, x3*square and X4*.

Step (fit, direction = “both”, test= “F”)
```