STA108 Final Project

Fanny Chow

Spring 2015

Introduction

During the early 1960's, the beginnings of the modern environmental movement raised public awareness of harm to the environment caused by man. At the same time, many Americans began moving from urban areas to suburban areas, resulting in low-density, car-dependent communities. Citizens and scientists alike began questioning the cumulative effects of automobile dependency on air pollution and its negative impacts on human health.

Amongst the many variables that affect air quality and health, there are many confounding and nonconfounding variables, which complicate analysis of the connection between pollution and health. Using 60 U.S. Standard Metropolitan Statistical Areas (SMSA) data obtained form the years 1959-1961, our study focuses on a major epidemiological question: does air pollution have effect on mortality?

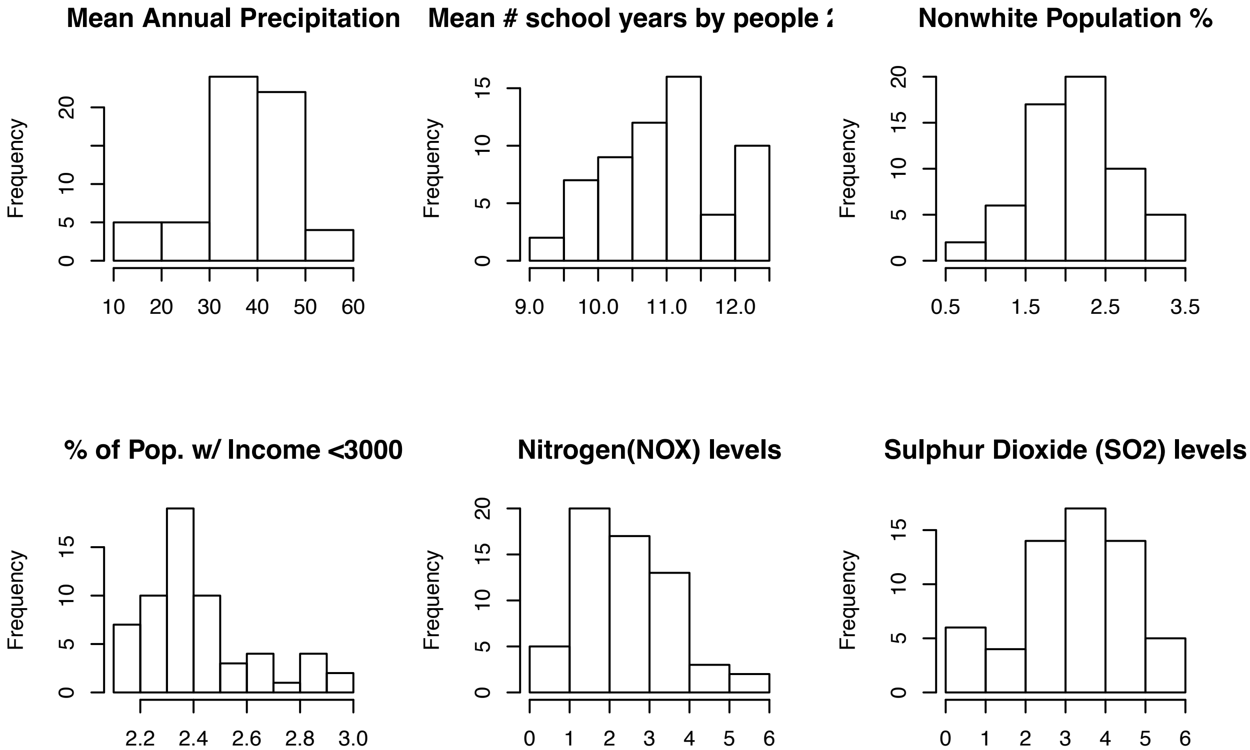
Fitting the Model

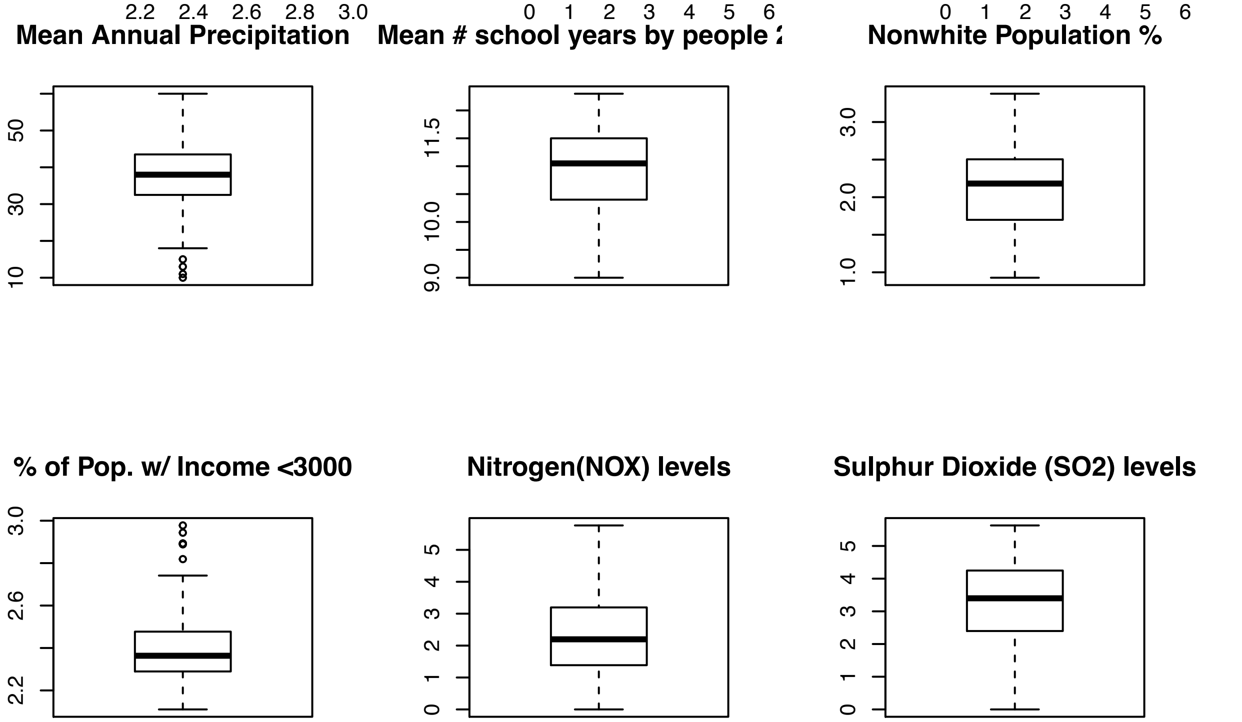
Transformation

Since the variables NOX and SO2 are skewed, we transformed them by using the natural logarithm. And since the variables nonwhite and poor are both skewed, we transformed them using a cube root.

Examining Basic Summary Statistics

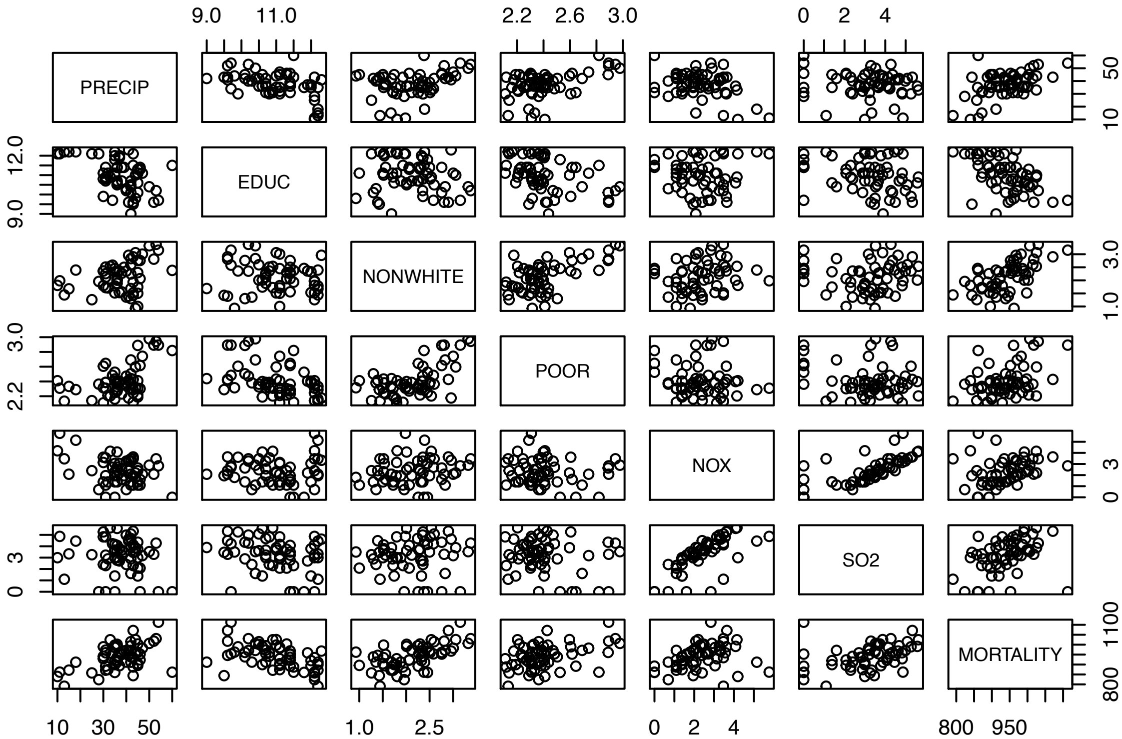
The histograms of the predictors demonstrtate that all the independent variables appear approximately normal with the exception of the mean number of school years by the people 25 and over. The boxplot demonstrates that the predictors precipitation and poverty (the percentage of population with incomes below $3000) contain outliers.





Examining Pairwise Corrlelation Information

Based on the matrix plot of the mortality data, with the exception of NOX levels, it seems that there is an approxmiately linear relationship between the dependent variable (mortality) and the independent variables (precipitation, education, nonwhite, poor, NOX, SO2).



Examining Multicollinearity Issues

Looking at the correlation matrix, there does not appear to be any major problems with multicollinearity since the quantities are not significantly high; they are approximately less than 0.7.

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

	SO2	MORTALITY
PRECIP	-0.1211723	0.5094924
EDUC	-0.2561622	-0.5109813
NONWHITE	0.0592199	0.6063347
POOR	-0.1955220	0.4099867
NOX	0.7328074	0.2919997
SO2	1.0000000	0.4031300
MORTALITY	0.4031300	1.0000000

Estimating Parameters

Model: $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$

Fitted Regression: $\hat{Y} = 980.475 + 2.375x_1 + \text{-}19.100x_2 + 49.905x_3 + \text{-}31.098x_4 + 10.104x_5 + 8.031x_6$

From the basic estimate of the parameters and standard errors, we observe that education and poverty are negatively associated with mortality. The Multiple R-Squared value 0.6985 indicates that about 69.85% of the variability in mortality rates (Y) can be explained by its regression on the predictors: precipitation(x_1), education(x_2), nonwhite(x_3), poor(x_4), NOX(Sx_5), SO2(x_6).

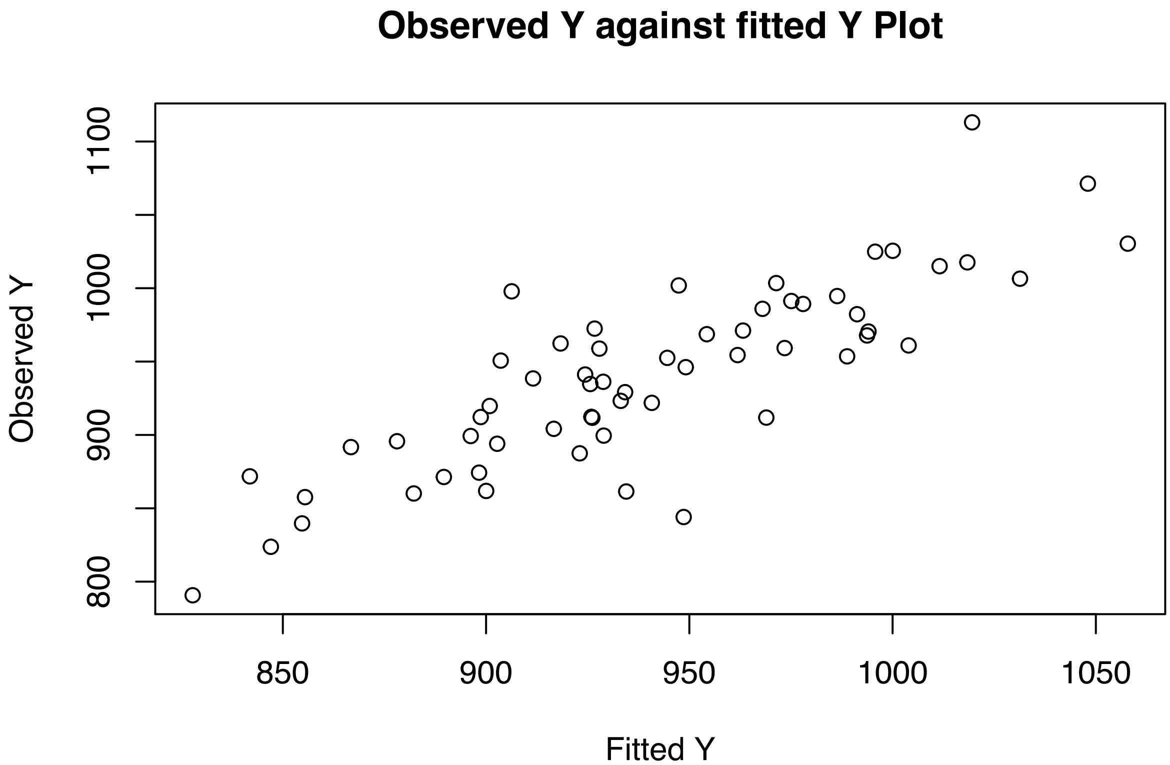
```
Analysis of Variance Table

Response: mortality_transformed$MORTALITY
          Df Sum Sq Mean Sq F value    Pr(>F)
PRECIP     1  59256   59256 45.6291 1.118e-08 ***
EDUC       1  20492   20492 15.7800 0.0002161 ***
NONWHITE   1  51678   51678 39.7940 5.830e-08 ***
POOR       1   7391    7391  5.6911 0.0206571 *
NOX        1  17982   17982 13.8469 0.0004808 ***
SO2        1   2646    2646  2.0377 0.1593045
Residuals 53  68828    1299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = mortality_transformed$MORTALITY ~ ., data = mortality_transformed)

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 ***
PRECIP        2.3748     0.6709   3.540 0.000844 ***
EDUC        -19.1004     7.6787  -2.487 0.016048 *
NONWHITE     49.9051    11.3256   4.406 5.15e-05 ***
POOR        -31.0975    34.5908  -0.899 0.372713
NOX          10.1044     7.1973   1.404 0.166178
SO2           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
```

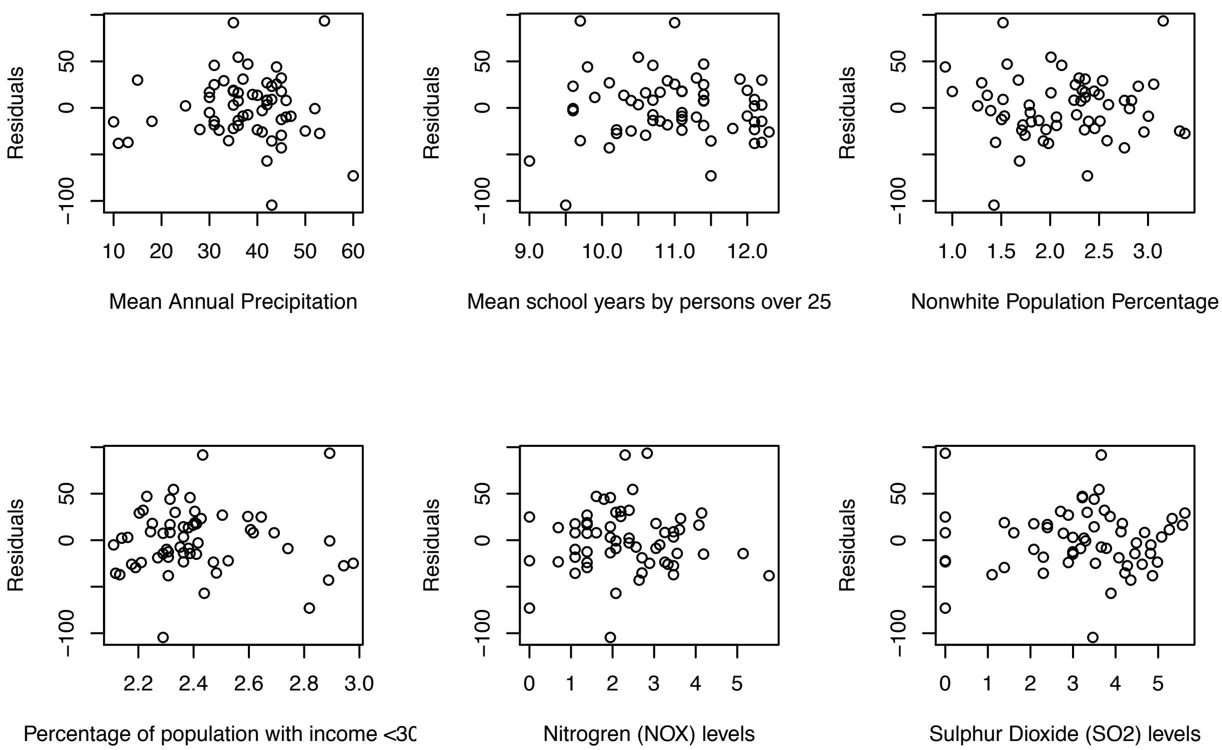
Regression Model Diagnostics

In order to perform multiple linear regression, we must first ensure that the data satisfies basic assumptions of the regression model. The errors must have:
* approximately equal variances
* be normally distributed.

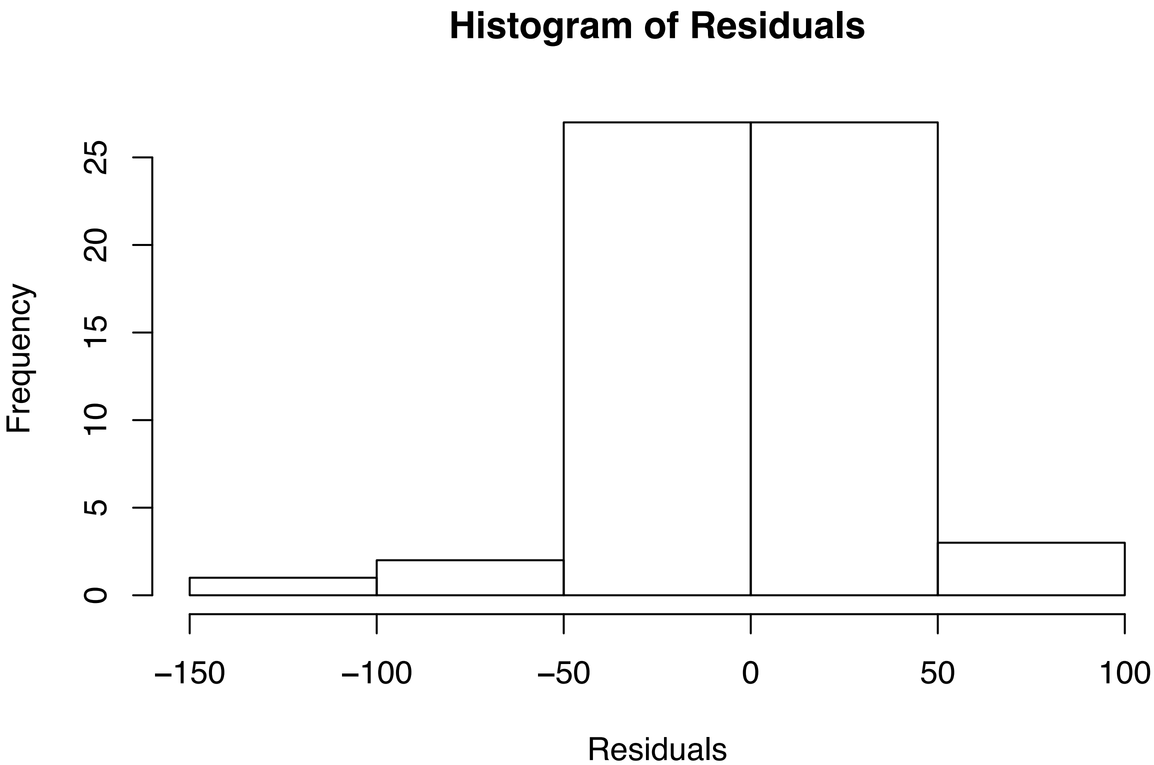
The plot of the residuals against fitted values demonstrates the residuals are approximately normally distributed.



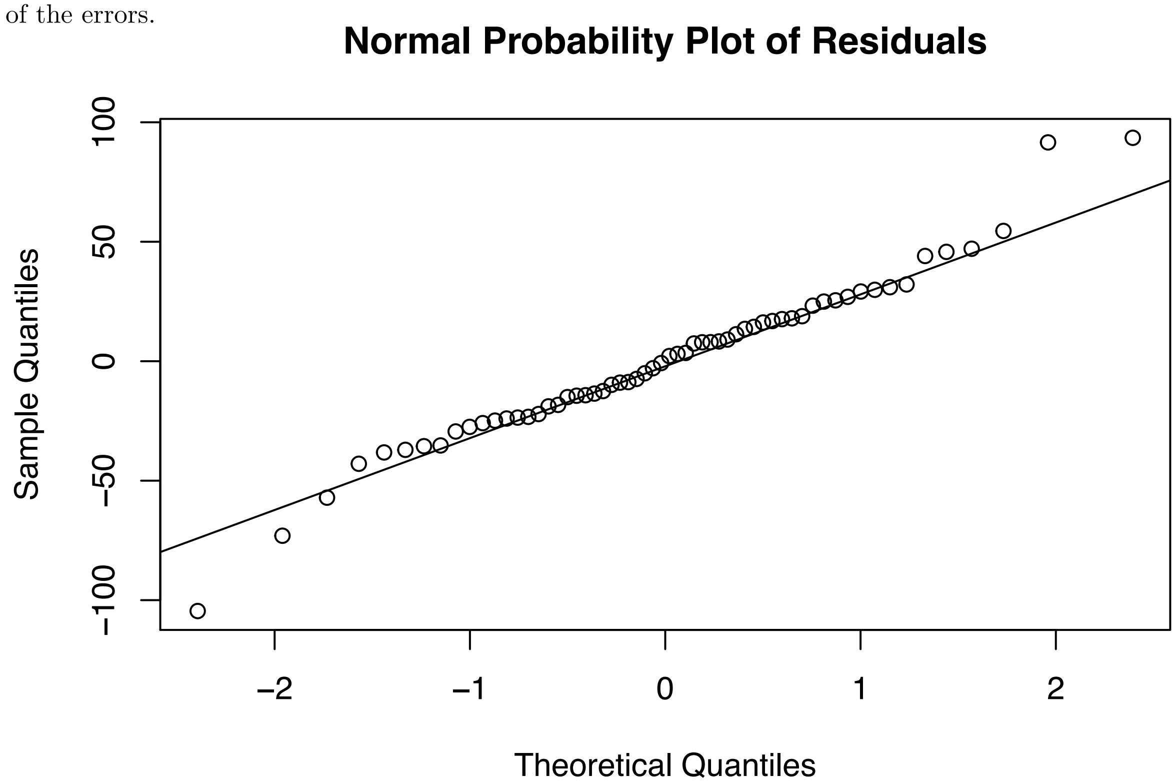
We use the plot of residuals against predictor variables to check the model assumptions: the regression function is linear, the errors have constant variance, and the model fits all but 1 or more outlying observations.



The histogram of residuals resembles the bell curve of a normal distribution.



The normal QQ plot is close to the 45 degree line, which demonstrates the approximate normal distribution of the errors.



Is a Quadratic Model Better?

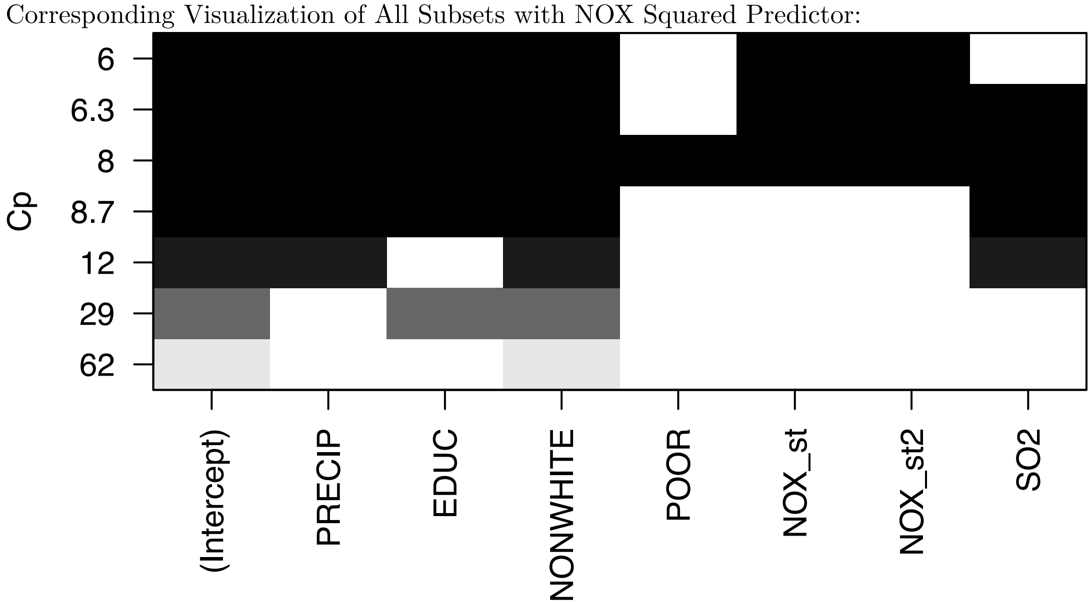
We suspected a nonlinear, quadratic relationship between mortality and NOX levels which led us to fit the model with a square term for NOX. After performing this modification, it seems that the fitting the linear model was still the better model because the Mallow's CP for the best linear model according to all subsets regression is lower than the Mallow's CP value for the best quadratic model. Here are the results from our test:

```
Subset selection object
Call: regsubsets.formula(mortality_transformed2$MORTALITY ~ ., data = mortality_transformed2,
    nbest = 1, nvmax = 7)
7 Variables  (and intercept)
         Forced in Forced out
PRECIP       FALSE      FALSE
EDUC         FALSE      FALSE
NONWHITE     FALSE      FALSE
POOR         FALSE      FALSE
NOX_st       FALSE      FALSE
NOX_st2      FALSE      FALSE
SO2          FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         PRECIP EDUC NONWHITE POOR NOX_st NOX_st2 SO2
1  ( 1 ) " "    " "  "*"      " "  " "    " "     " "
2  ( 1 ) " "    "*"  "*"      " "  " "    " "     " "
3  ( 1 ) "*"    " "  "*"      " "  " "    " "     "*"
4  ( 1 ) "*"    "*"  "*"      " "  " "    " "     "*"
5  ( 1 ) "*"    "*"  "*"      " "  "*"    "*"     " "
6  ( 1 ) "*"    "*"  "*"      " "  "*"    "*"     "*"
7  ( 1 ) "*"    "*"  "*"      "*"  "*"    "*"     "*"
```

Corresponding CP Values:

```
[1] 61.735216 28.894774 11.913117  8.695983  6.029812  6.287593  8.000000
```

Corresponding Visualization of All Subsets with NOX Squared Predictor:



Ommitting Variables with Stepwise Regression

All Subsets Regression

According to all subsets regression, both the variables percentage of the population with income under $3000 and the NOX level should be dropped in order to improve the precision of the model. The models that contain both poverty and NOX variables have the highest Mallows CP values.

```
Subset selection object
Call: regsubsets.formula(mortality_transformed$MORTALITY ~ ., data = mortality_transformed,
    nbest = 1, nvmax = 7)
6 Variables  (and intercept)
         Forced in Forced out
PRECIP       FALSE      FALSE
EDUC         FALSE      FALSE
NONWHITE     FALSE      FALSE
POOR         FALSE      FALSE
NOX          FALSE      FALSE
SO2          FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         PRECIP EDUC NONWHITE POOR NOX SO2
1  ( 1 ) " "    " "  "*"      " "  " " " "
2  ( 1 ) " "    "*"  "*"      " "  " " " "
3  ( 1 ) "*"    " "  "*"      " "  " " "*"
4  ( 1 ) "*"    "*"  "*"      " "  " " "*"
5  ( 1 ) "*"    "*"  "*"      " "  "*" "*"
6  ( 1 ) "*"    "*"  "*"      "*"  "*" "*"
```

The corresponding CP Values:

```
[1] 55.155271 24.261980  8.341163  5.415603  5.808220  7.000000
```

Stepwise Transformation

The results from both stepwise regression and all subsets regression are identical. The stepwise transformation indicate that both poverty and NOX variables could be dropped to improve the model.

```
Call:
lm(formula = mortality_transformed$MORTALITY ~ X3 + X2 + X6 +
    X1, data = mortality_transformed)

Coefficients:
(Intercept)           X3           X2           X6           X1
     883.03        49.40       -15.22        14.95         1.90
```

Examining the Improved Model

Model: $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_6$

Fitted Regression: $\hat{Y} = 883.0325 + 1.8997\ x_1 + \text{-}15.2159 x_2 + 49.4012 x_3 + 14.9480\ x_6$

Fitted Regression: $\hat{Y}$ = 883.0325 + 1.8997(PRECIP) + -15.2159(EDUC) + 49.4012(NONWHITE) + 14.9480(SO2)

```
Call:
lm(formula = mortality_transformed$MORTALITY ~ X3 + X2 + X6 +
    X1, data = mortality_transformed)

Residuals:
    Min      1Q  Median      3Q     Max
-98.369 -19.589  -1.322  17.336 119.182

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 883.0325    93.5624   9.438 4.25e-13 ***
X3           49.4012     8.6557   5.707 4.76e-07 ***
X2          -15.2159     6.8818  -2.211  0.03121 *
X6           14.9480     3.4278   4.361 5.73e-05 ***
X1            1.8997     0.5962   3.186  0.00238 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.17 on 55 degrees of freedom
Multiple R-squared:  0.6847,    Adjusted R-squared:  0.6618
F-statistic: 29.87 on 4 and 55 DF,  p-value: 3.241e-13
```

Summary of Findings

From our analysis, we conclude that pollution affects mortality rates.

Higher SO2 levels are associated with higher mortality rates. Aside from pollution, demographics, such as race and education, also influence mortality rates. It appears that race, the percentage of the population that is nonwhite in 1960, is associated with mortality rates. There is evidence that the predictor education (median number of school years completed by persons of age 25 or over) is negatively associated with mortality. This may be reflective of the fact that well-educated nonwhite individuals in the 1960's are often wealthier and live in suburbs farther from highways and factories, areas subject to less industrial pollution or automobile exhaust.

Since there are so many confounding variables that influence mortality and human health, it is difficult to solely isolate pollution as a leading cause of mortality. Further analysis is needed to expand on the inferences developed from this data set. We recommend seeking Census Bureau data on communities of low-income and people of color and analyzing the corresponding mortality rates and pollution (SO2 and NOX) levels in those regions. With the passing of the Clean Air Act in 1970, it would also be interesting to analyze mortality rates before and after regulations on toxic air pollutants.

Appendix

```
# load the data
library(gdata)
setwd("~/Desktop/STA108/project")
mortality_data = read.xls("mortality.xls")
mortality_data <- cbind(mortality_data, NONWHITE_cuberoot = mortality_data$NONWHITE^(1/3),
                        POOR_cuberoot = mortality_data$POOR^(1/3), lnNOX = log(mortality_data$NOX),
                        lnSO2 = log(mortality_data$SO2))

mortality_transformed <- data.frame(cbind(PRECIP = mortality_data$PRECIP, EDUC = mortality_data$EDUC,
                              NONWHITE = mortality_data$NONWHITE_cuberoot, POOR = mortality_data$POOR_
                              NOX = mortality_data$lnNOX, SO2 = mortality_data$lnSO2, MORTALITY = mort
plot(mortality_transformed) # matrix plot

#Examining Multicollinearity Issues
cor(mortality_transformed) # correlation matrix
mod <- lm(mortality_transformed$MORTALITY~., data = mortality_transformed) # regression model

#Estimating Parameters
anova(mod) #anova table
summary(mod) # estimate of parameters & standard error
###Regression Model Diagnostics
plot(mortality_transformed$MORTALITY~mod$fitted, xlab = "Fitted Y", ylab = "Observed Y", main = "Observe
res <- mod$res
par(mfrow=c(2,3))
plot(res~mortality_transformed$PRECIP, xlab = "Mean Annual Precipitation", ylab = "Residuals")
plot(res~mortality_transformed$EDUC, xlab = "Mean school years by persons over 25", ylab = "Residuals")
plot(res~mortality_transformed$NONWHITE, xlab = "Nonwhite Population Percentage", ylab = "Residuals")
plot(res~mortality_transformed$POOR, xlab = "Percentage of population with income <3000", ylab = "Resid
plot(res~mortality_transformed$NOX, xlab = "Nitrogren (NOX) levels", ylab = "Residuals")
plot(res~mortality_transformed$SO2, xlab = "Sulphur Dioxide (SO2) levels", ylab = "Residuals")

par(mfrow = c(1,1))
hist(mod$res, main = "Histogram of Residuals", xlab = "Residuals") #histogram
qqnorm(mod$res, main = "Normal Probability Plot of Residuals")
qqline(mod$res)


###Is a Quadratic Model Better?
library(leaps)
xbar_NOX <- mean(mortality_transformed$NOX)
NOX_st <- mortality_transformed$NOX - xbar_NOX
mortality_transformed2 <- data.frame(cbind(PRECIP = mortality_data$PRECIP, EDUC = mortality_data$EDUC,
                                          NONWHITE = mortality_data$NONWHITE_cuberoot, POOR = mortality_
                                          NOX_st = NOX_st, NOX_st2 = NOX_st^2,SO2 = mortality_data$lnSO

mod2 <- lm(mortality_transformed2$MORTALITY~., data = mortality_transformed2)
summary(regsubsets(mortality_transformed2$MORTALITY~., data=mortality_transformed2, nbest=1, nvmax=7))
summary(regsubsets(mortality_transformed2$MORTALITY~., data=mortality_transformed2, nbest=1, nvmax=7))$

###Ommitting Variables with Stepwise Regression
######All Subsets Regression
library(leaps)
```

```
summary(regsubsets(mortality_transformed$MORTALITY~., data=mortality_transformed, nbest=1, nvmax=7))
summary(regsubsets(mortality_transformed$MORTALITY~., data=mortality_transformed, nbest=1, nvmax=7))$cp

####Stepwise Transformation
X1 <- mortality_transformed$PRECIP
X2 <- mortality_transformed$EDUC
X3 <- mortality_transformed$NONWHITE
X4 <- mortality_transformed$POOR
X5 <- mortality_transformed$NOX
X6 <- mortality_transformed$SO2

#step
step(object=lm(mortality_transformed$MORTALITY~1,data=mortality_transformed),direction='forward',scope=

###Revisiting the Linear Model
par(mfrow = c(2,3))
hist(mortality_transformed$PRECIP, main = "Mean Annual Precipitation", xlab = "")
hist(mortality_transformed$EDUC, main = "Mean # school years by people 25+", xlab = "")
hist(mortality_transformed$NONWHITE, main = "Nonwhite Population %", xlab = "")
hist(mortality_transformed$POOR, main = "% of Pop. w/ Income <3000", xlab = "")
hist(mortality_transformed$NOX, main = "Nitrogen(NOX) levels", xlab = "")
hist(mortality_transformed$SO2, main = "Sulphur Dioxide (SO2) levels", xlab = "")
boxplot(mortality_transformed$PRECIP, main = "Mean Annual Precipitation")
boxplot(mortality_transformed$EDUC, main = "Mean # school years by people 25+")
boxplot(mortality_transformed$NONWHITE, main = "Nonwhite Population %")
boxplot(mortality_transformed$POOR, main = "% of Pop. w/ Income <3000")
boxplot(mortality_transformed$NOX, main = "Nitrogen(NOX) levels")
boxplot(mortality_transformed$SO2, main = "Sulphur Dioxide (SO2) levels")
```