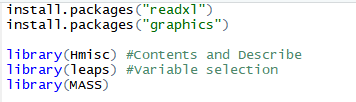
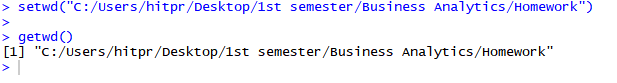
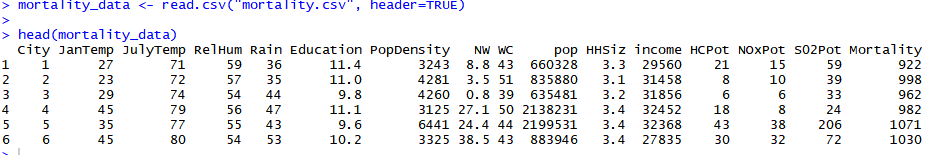
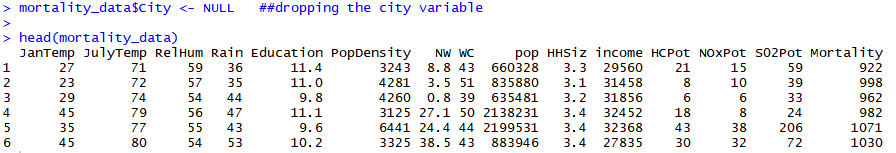
**SMSA air pollution Data Analysis**

Researchers at General Motors collected data on some U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality. The dependent variable for analysis is age adjusted mortality ("Mortality"). The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. As a simple initial test, we will use regression models to determine whether air pollution is significantly related to mortality.

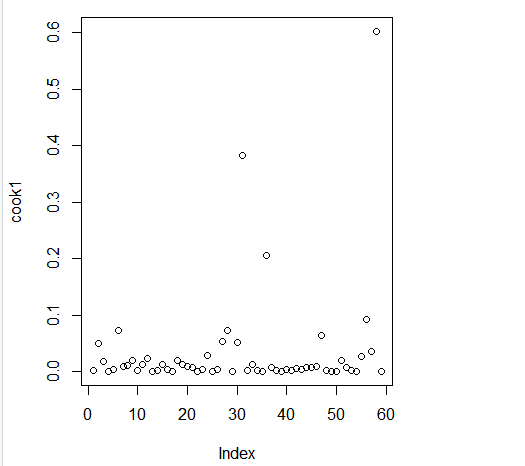
|  |  |  |
| --- | --- | --- |
| **#** | **Variable** | **Description** |
| 1 | city | City ID |
| 2 | JanTemp | Mean January temperature (F) |
| 3 | JulyTemp | Mean July temperature (F) |
| 4 | RelHum | Relative Humidity |
| 5 | Rain | Annual rainfall (inches) |
| 6 | Education | Median education |
| 7 | PopDensity | Population density |
| 8 | NW | Percentage of non-whites |
| 9 | WC | Percentage of white collar workers |
| 10 | pop | Population |
| 11 | HHSiz | Average household size |
| 12 | income | Median income |
| 13 | HCPot | HC pollution potential |
| 14 | NOxPot | Nitrous Oxide pollution potential |
| 15 | SO2Pot | Sulfur Dioxide pollution potential |
| 16 | Mortality | Age adjusted mortality |

1. Started R. Installed some packages and loaded some libraries.  
   
2. Set up my working repository where .csv files are saved for both Mortality and Transactions data. Used the functions setwd() and getwd() as specified.  
   
3. Loaded data from Mortality.csv. 

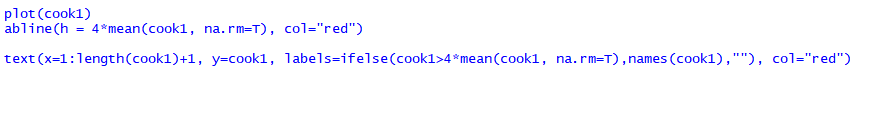
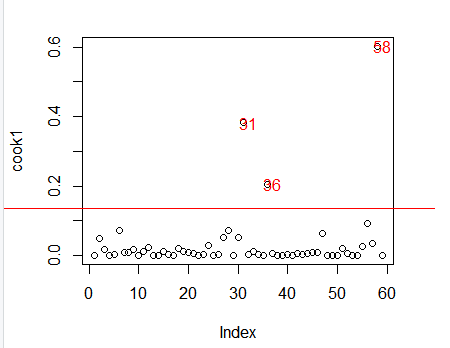
Removed City by assigning null as it is an identifier column.

1. Removing outliers: Used cook’s distance to remove the outliers. It is used to find the influence of the data points on the regression model. Data points with large residuals may distort the outcome and accuracy of a regression. Cook’s distance depends on the error values y minus y hat. Larger the cook’s distance more is the effect of that observation on the regression that leads to skewness.

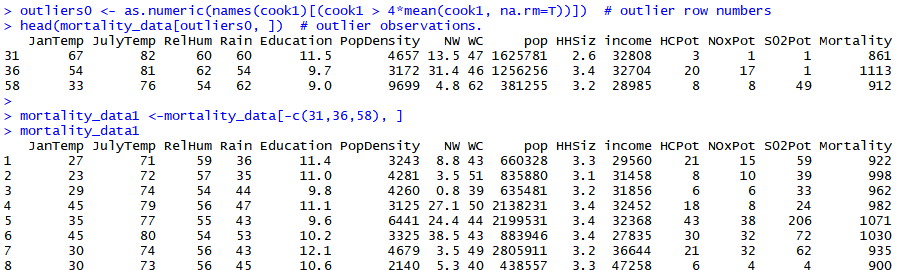
So, creating a dummy regression model using the variables specified in question to calculate cook’s distance. Calculate cook’s distance and plot them.

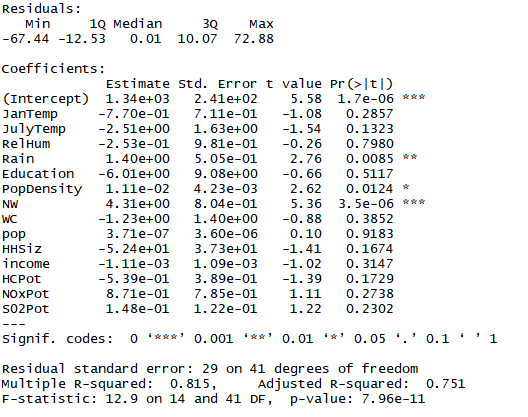


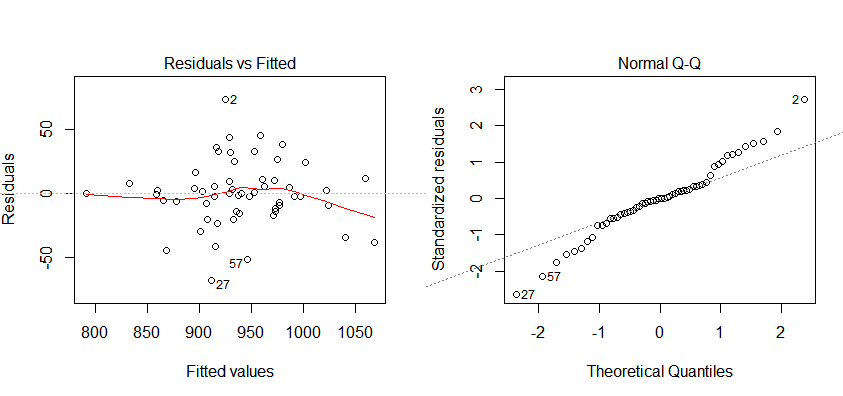
Now for better visualization and understanding drawing a threshold line (using abline()) in red in the plot to show what I am considering as outliers. The threshold selected is 4 times the mean of cook’s distance. Also labeling the outliers (using text()) with the row number or the observation number in red.

Select the outlier row numbers. Check the exact outlier rows and remove the outliers by taking compliment of the rows from the actual data i.e. mortality\_data and store it into mortality\_data1. Now if we check the data in Mortality\_data1, we can see that the rows 31, 36 and 58 are missing. Therefore, the outliers are removed. We now need to use Mortality\_data1 for further steps and not Mortality\_data.

  
Now we need to run the appropriate regression diagnostics (normality, homoscedasticity) to ensure that the assumptions of OLS are not violated. For that run the regression using the new data frame Mortality\_data1 and plot it. Below is the summary of the regression and the plot. As can be seen it is homoskedasticity since the residuals vs fitted is almost linear and horizontal. In Normal Q-Q graph it is slightly positively skewed however, since most of the observations are on the line, I have considered it as a linear plot, hence normal.





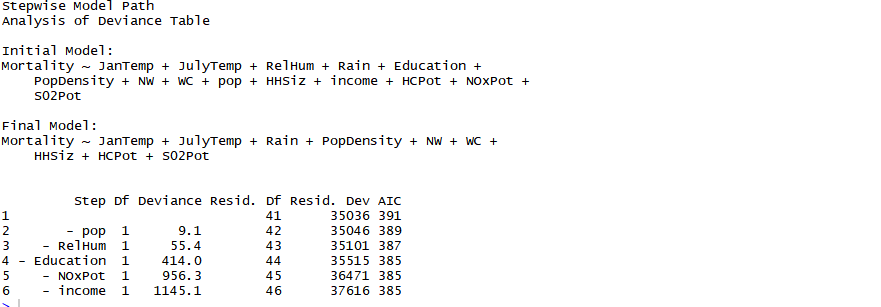
1. a.) R squared value for above model: 0.815

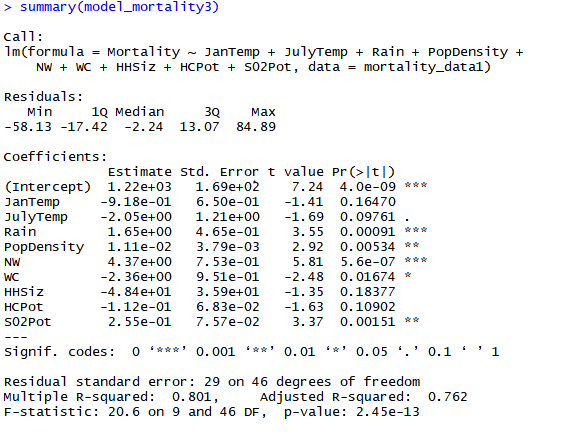
Adjusted R squared for above model: 0.751

R squared means that 81.5 % of the variance is explained by the regression model.  
Adjusted R squared is the R squared value that has been adjusted for the number of predictors in the model. Comparison between 2 different models for a dataset is done using the Adjusted R squared value.

b.) There are three significant variables in this model namely Rain (t value=2.76, Pr(>|t|)= 0.0085), PopDensity(t value=2.62, Pr(>|t|)= 0.0124), NW(t value=5.36, Pr(>|t|)= 3.5e^-06)

1. Find the best model using forward and backward methods.  
   Run another regression model using the final model shown when we do anova.





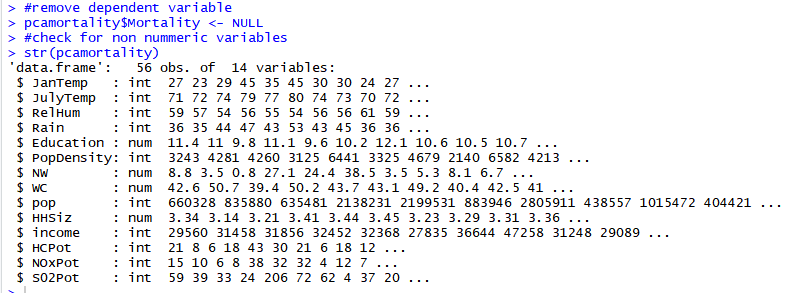
1. This is an important step as leads to increase in Adjusted R squared with reduced number of variables.
2. R squared means that 80.1 % of the variance is explained by the regression model.  
   Adjusted R squared is the R squared value that has been adjusted for the number of predictors in the model. Comparison between 2 different models for a dataset is done using the Adjusted R squared value.
3. There are five significant variables in this model namely Rain (t value=3.55, Pr(>|t|)= 0.00091), PopDensity(t value=2.92, Pr(>|t|)= 0.00534), NW(t value=5.81, Pr(>|t|)= 5.6e^-07), WC(t value=-2.48, Pr(>|t|)= 5.6e^-07), S02Pot(t value=3.37, Pr(>|t|)= 0.00151)
4. Running PCA.

Store data in another data frame after removing any missing values.

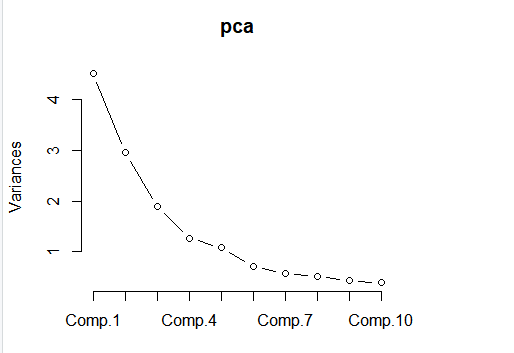
mortality\_data2<- na.omit(mortality\_data1)

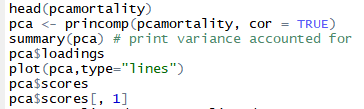
pcamortality<-mortality\_data2

Remove dependent variables.

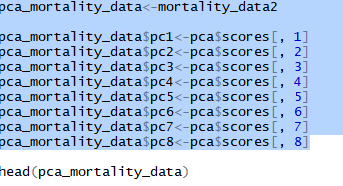
Running PCA and checking for loadings and scores.

1. Draw the scree plot for PCA. I have selected 8 components as the variance almost stabilizes after Comp8. Also the Cumulative proportion of Variance reaches ~ 0.9 till Comp8.

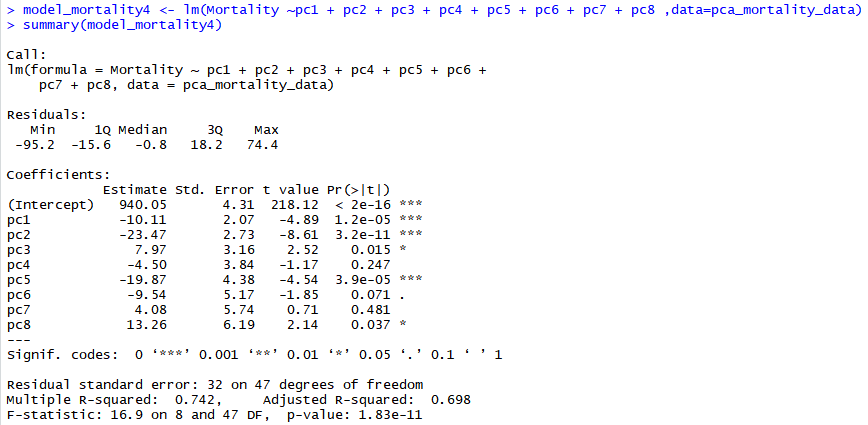




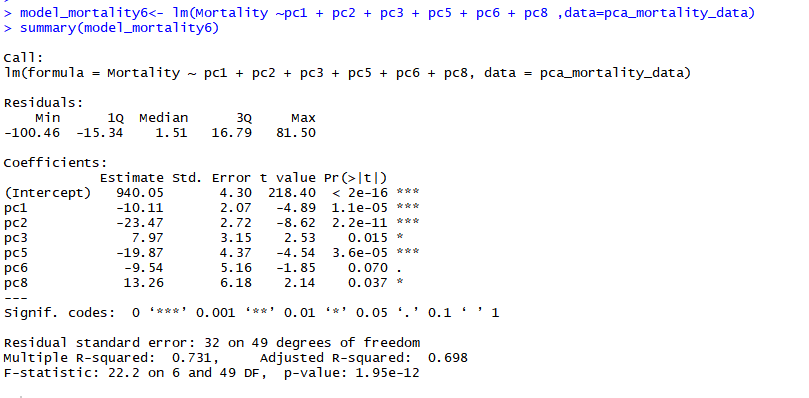
1. Manipulate data to run regression on above PCA.



1. Running regression.



1. R squared means that 74.2% of the variance is explained by the regression model.  
   Adjusted R squared is the R squared value that has been adjusted for the number of predictors in the model. Comparison between 2 different models for a dataset is done using the Adjusted R squared value.
2. There are five significant variables in this model namely pc1(t value=-4.89, Pr(>|t|)= <1.2e-05), pc2(t value=-8.61, Pr(>|t|)= 3.2e-11), pc3(t value=2.52, Pr(>|t|)= 0.015), pc5(t value=-4.54, Pr(>|t|)= 3.9e-05), pc8(t value=2.14, Pr(>|t|)= 0.037)
3. Found best model using forward and backward methods and use the variables to run another regression.



1. R squared means that 73.1% of the variance is explained by the regression model.  
   Adjusted R squared is the R squared value that has been adjusted for the number of predictors in the model. Comparison between 2 different models for a dataset is done using the Adjusted R squared value.
2. There are five significant variables in this model namely pc1(t value=-4.89, Pr(>|t|)= <1.1e-05), pc2(t value=-8.62, Pr(>|t|)= 2.2e-11), pc3(t value=2.53, Pr(>|t|)= 0.015), pc5(t value=-4.54, Pr(>|t|)= 3.6e-05), pc8(t value=2.14, Pr(>|t|)= 0.037)
3. In the above regression models the model in step 6 has the maximum Adjusted R squared value and the model in step 8 has the minimum number of variables. I will personally select model in step 6 as it has decrease number of variables (although not minimum) and has the max adjusted R squared value.