# The University of Texas at Dallas
## Naveen Jindal School of Management

### 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.

```
install.packages("readxl")
install.packages("graphics")

library(Hmisc) #Contents and Describe
library(leaps) #Variable selection
library(MASS)
```

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.

```
> setwd("C:/Users/hitpr/Desktop/1st semester/Business Analytics/Homework")
>
> getwd()
[1] "C:/Users/hitpr/Desktop/1st semester/Business Analytics/Homework"
>
```

3.> Loaded data from Mortality.csv.

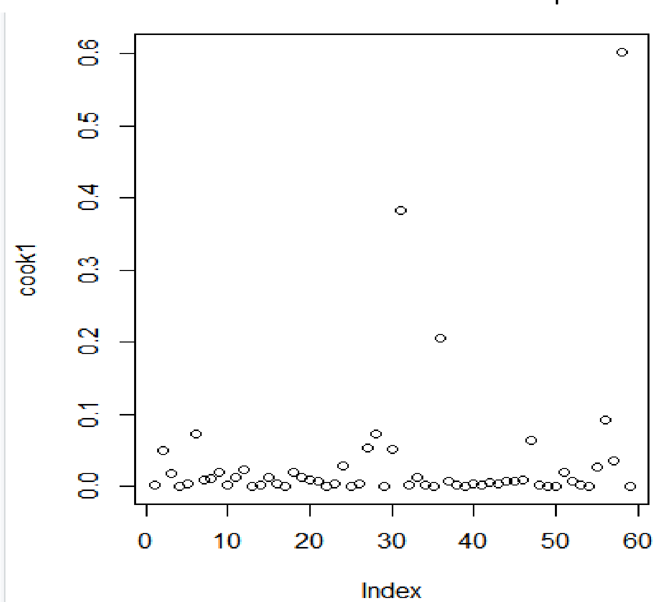
```
> mortality_data <- read.csv("mortality.csv", header=TRUE)
>
> head(mortality_data)
  City JanTemp JulyTemp RelHum Rain Education PopDensity   NW WC     pop HHSiz income HCPot NOxPot SO2Pot Mortality
1    1      27       71     59   36      11.4       3243  8.8 43  660328   3.3  29560    21     15     59       922
2    2      23       72     57   35      11.0       4281  3.5 51  835880   3.1  31458     8     10     39       998
3    3      29       74     54   44       9.8       4260  0.8 39  635481   3.2  31856     6      6     33       962
4    4      45       79     56   47      11.1       3125 27.1 50 2138231   3.4  32452    18      8     24       982
5    5      35       77     55   43       9.6       6441 24.4 44 2199531   3.4  32368    43     38    206      1071
6    6      45       80     54   53      10.2       3325 38.5 43  883946   3.4  27835    30     32     72      1030
```

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

```
> mortality_data$City <- NULL    ##dropping the city variable
>
> head(mortality_data)
  JanTemp JulyTemp RelHum Rain Education PopDensity   NW WC     pop HHSiz income HCPot NOxPot SO2Pot Mortality
1      27       71     59   36      11.4       3243  8.8 43  660328   3.3  29560    21     15     59       922
2      23       72     57   35      11.0       4281  3.5 51  835880   3.1  31458     8     10     39       998
3      29       74     54   44       9.8       4260  0.8 39  635481   3.2  31856     6      6     33       962
4      45       79     56   47      11.1       3125 27.1 50 2138231   3.4  32452    18      8     24       982
5      35       77     55   43       9.6       6441 24.4 44 2199531   3.4  32368    43     38    206      1071
6      45       80     54   53      10.2       3325 38.5 43  883946   3.4  27835    30     32     72      1030
```

4.> 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.

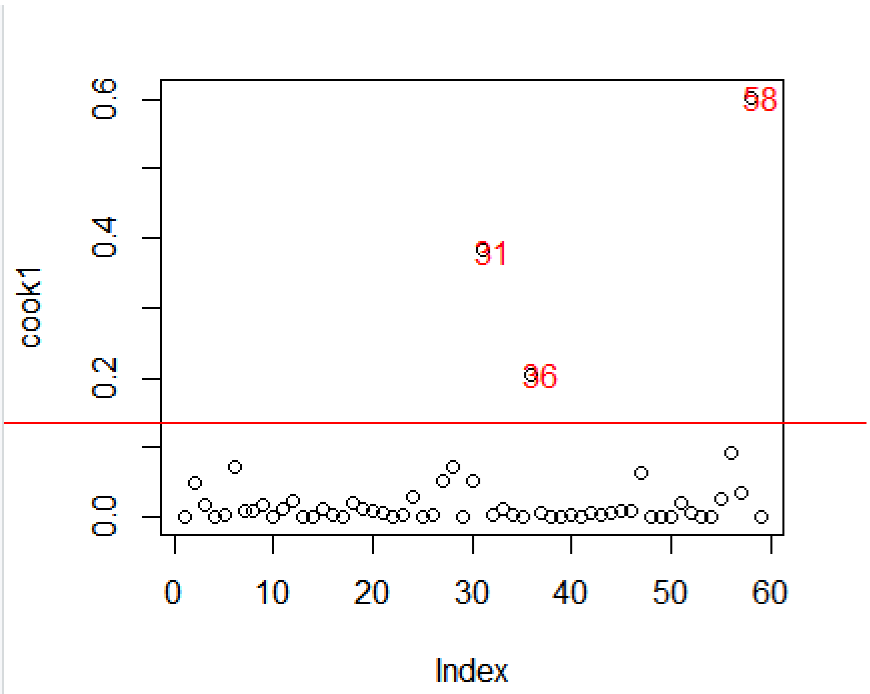
So, creating a dummy regression model using the variables specified in question to calculate cook's distance. Calculate cook's distance and plot them.




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.

```
plot(cook1)
abline(h = 4*mean(cook1, na.rm=T), col="red")

text(x=1:length(cook1)+1, y=cook1, labels=ifelse(cook1>4*mean(cook1, na.rm=T),names(cook1),""), col="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.

```
> outliers0 <- as.numeric(names(cook1)[(cook1 > 4*mean(cook1, na.rm=T))])  # outlier row numbers
> head(mortality_data[outliers0, ])  # outlier observations.
   JanTemp JulyTemp RelHum Rain Education PopDensity   NW WC     pop HHSiz income HCPot NOxPot SO2Pot Mortality
31      67       82     60   60      11.5       4657 13.5 47 1625781   2.6  32808     3      1      1       861
36      54       81     62   54       9.7       3172 31.4 46 1256256   3.4  32704    20     17      1      1113
58      33       76     54   62       9.0       9699  4.8 62  381255   3.2  28985     8      8     49       912
>
> mortality_data1 <-mortality_data[-c(31,36,58), ]
> mortality_data1
   JanTemp JulyTemp RelHum Rain Education PopDensity   NW WC     pop HHSiz income HCPot NOxPot SO2Pot Mortality
1       27       71     59   36      11.4       3243  8.8 43  660328   3.3  29560    21     15     59       922
2       23       72     57   35      11.0       4281  3.5 51  835880   3.1  31458     8     10     39       998
3       29       74     54   44       9.8       4260  0.8 39  635481   3.2  31856     6      6     33       962
4       45       79     56   47      11.1       3125 27.1 50 2138231   3.4  32452    18      8     24       982
5       35       77     55   43       9.6       6441 24.4 44 2199531   3.4  32368    43     38    206      1071
6       45       80     54   53      10.2       3325 38.5 43  883946   3.4  27835    30     32     72      1030
7       30       74     56   43      12.1       4679  3.5 49 2805911   3.2  36644    21     32     62       935
8       30       73     56   45      10.6       2140  5.3 40  438557   3.3  47258     6      4      4       900
```

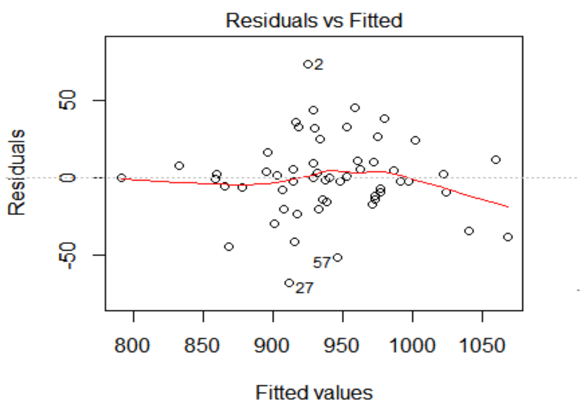
Now we need to run the appropriate regression diagnostics (normality, homoscedasticity) to ensure that the assumptions of OLS are not violated. For that run the regression using the new data frame Mortality_data1 and plot it. Below is the summary of the regression and the plot. As can be seen it is homoskedasticity since the residuals vs fitted is almost linear and horizontal. In

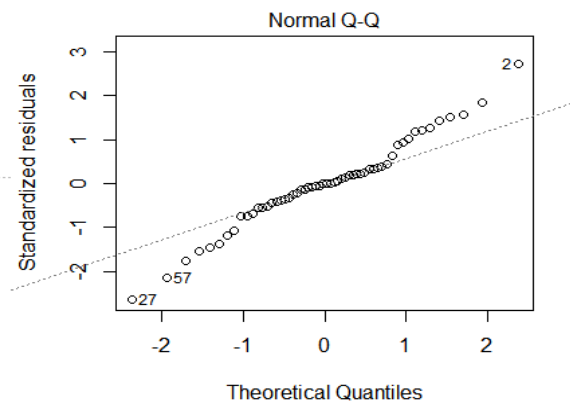
Normal Q-Q graph it is slightly positively skewed however, since most of the observations are on the line, I have considered it as a linear plot, hence normal.

```
Residuals:
   Min     1Q Median     3Q    Max 
-67.44 -12.53   0.01  10.07  72.88 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.34e+03   2.41e+02    5.58  1.7e-06 ***
JanTemp     -7.70e-01   7.11e-01   -1.08   0.2857    
JulyTemp    -2.51e+00   1.63e+00   -1.54   0.1323    
RelHum      -2.53e-01   9.81e-01   -0.26   0.7980    
Rain         1.40e+00   5.05e-01    2.76   0.0085 ** 
Education   -6.01e+00   9.08e+00   -0.66   0.5117    
PopDensity   1.11e-02   4.23e-03    2.62   0.0124 *  
NW           4.31e+00   8.04e-01    5.36  3.5e-06 ***
WC          -1.23e+00   1.40e+00   -0.88   0.3852    
pop          3.71e-07   3.60e-06    0.10   0.9183    
HHSiz       -5.24e+01   3.73e+01   -1.41   0.1674    
income      -1.11e-03   1.09e-03   -1.02   0.3147    
HCPot       -5.39e-01   3.89e-01   -1.39   0.1729    
NOxPot       8.71e-01   7.85e-01    1.11   0.2738    
SO2Pot       1.48e-01   1.22e-01    1.22   0.2302    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29 on 41 degrees of freedom
Multiple R-squared:  0.815,	Adjusted R-squared:  0.751 
F-statistic: 12.9 on 14 and 41 DF,  p-value: 7.96e-11
```




5.> 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)

6.> Find the best model using forward and backward methods.

Run another regression model using the final model shown when we do anova.

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Mortality ~ JanTemp + JulyTemp + RelHum + Rain + Education +
    PopDensity + NW + WC + pop + HHSiz + income + HCPot + NOxPot +
    S02Pot

Final Model:
Mortality ~ JanTemp + JulyTemp + Rain + PopDensity + NW + WC +
    HHSiz + HCPot + S02Pot


         Step Df Deviance Resid. Df Resid. Dev AIC
1                                41      35036 391
2        - pop  1      9.1        42      35046 389
3     - RelHum  1     55.4        43      35101 387
4  - Education  1    414.0        44      35515 385
5     - NOxPot  1    956.3        45      36471 385
6     - income  1   1145.1        46      37616 385
```

```
> summary(model_mortality3)

Call:
lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + PopDensity +
    NW + WC + HHSiz + HCPot + S02Pot, data = mortality_data1)

Residuals:
   Min     1Q Median     3Q    Max
-58.13 -17.42  -2.24  13.07  84.89

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.22e+03   1.69e+02    7.24  4.0e-09 ***
JanTemp     -9.18e-01   6.50e-01   -1.41  0.16470
JulyTemp    -2.05e+00   1.21e+00   -1.69  0.09761 .
Rain         1.65e+00   4.65e-01    3.55  0.00091 ***
PopDensity   1.11e-02   3.79e-03    2.92  0.00534 **
NW           4.37e+00   7.53e-01    5.81  5.6e-07 ***
WC          -2.36e+00   9.51e-01   -2.48  0.01674 *
HHSiz       -4.84e+01   3.59e+01   -1.35  0.18377
HCPot       -1.12e-01   6.83e-02   -1.63  0.10902
S02Pot       2.55e-01   7.57e-02    3.37  0.00151 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29 on 46 degrees of freedom
Multiple R-squared:  0.801,     Adjusted R-squared:  0.762
F-statistic: 20.6 on 9 and 46 DF,  p-value: 2.45e-13
```

a.) This is an important step as leads to increase in Adjusted R squared with reduced number of variables.

b.) 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.

c.) 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)

7.> Running PCA.

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

mortality_data2<- na.omit(mortality_data1)

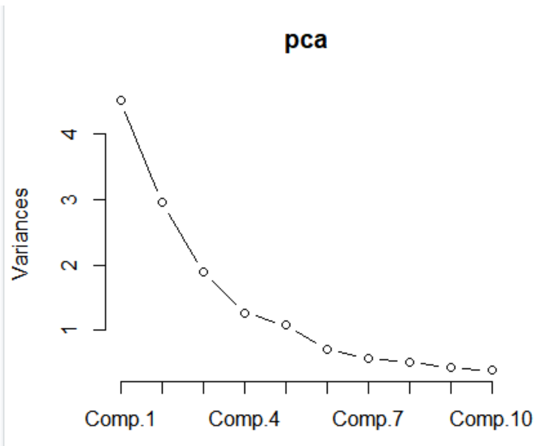
pcamortality<-mortality_data2

Remove dependent variables.

```
> #remove dependent variable
> pcamortality$Mortality <- NULL
> #check for non nummeric variables
> str(pcamortality)
'data.frame':	56 obs. of  14 variables:
 $ JanTemp   : int  27 23 29 45 35 45 30 30 24 27 ...
 $ JulyTemp  : int  71 72 74 79 77 80 74 73 70 72 ...
 $ RelHum    : int  59 57 54 56 55 54 56 56 61 59 ...
 $ Rain      : int  36 35 44 47 43 53 43 45 36 36 ...
 $ Education : num  11.4 11 9.8 11.1 9.6 10.2 12.1 10.6 10.5 10.7 ...
 $ PopDensity: int  3243 4281 4260 3125 6441 3325 4679 2140 6582 4213 ...
 $ NW        : num  8.8 3.5 0.8 27.1 24.4 38.5 3.5 5.3 8.1 6.7 ...
 $ WC        : num  42.6 50.7 39.4 50.2 43.7 43.1 49.2 40.4 42.5 41 ...
 $ pop       : int  660328 835880 635481 2138231 2199531 883946 2805911 438557 1015472 404421 ...
 $ HHSiz     : num  3.34 3.14 3.21 3.41 3.44 3.45 3.23 3.29 3.31 3.36 ...
 $ income    : int  29560 31458 31856 32452 32368 27835 36644 47258 31248 29089 ...
 $ HCPot     : int  21 8 6 18 43 30 21 6 18 12 ...
 $ NOxPot    : int  15 10 6 8 38 32 32 4 12 7 ...
 $ SO2Pot    : int  59 39 33 24 206 72 62 4 37 20 ...
```

Running PCA and checking for loadings and scores.

8.> 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.




```
head(pcamortality)
pca <- princomp(pcamortality, cor = TRUE)
summary(pca) # print variance accounted for
pca$loadings
plot(pca,type="lines")
pca$scores
pca$scores[, 1]
```

9.> Manipulate data to run regression on above PCA.

```
pca_mortality_data<-mortality_data2

pca_mortality_data$pc1<-pca$scores[, 1]
pca_mortality_data$pc2<-pca$scores[, 2]
pca_mortality_data$pc3<-pca$scores[, 3]
pca_mortality_data$pc4<-pca$scores[, 4]
pca_mortality_data$pc5<-pca$scores[, 5]
pca_mortality_data$pc6<-pca$scores[, 6]
pca_mortality_data$pc7<-pca$scores[, 7]
pca_mortality_data$pc8<-pca$scores[, 8]

head(pca_mortality_data)
```

10.> Running regression.

```
> model_mortality4 <- lm(Mortality ~pc1 + pc2 + pc3 + pc4 + pc5 + pc6 + pc7 + pc8 ,data=pca_mortality_data)
> summary(model_mortality4)

Call:
lm(formula = Mortality ~ pc1 + pc2 + pc3 + pc4 + pc5 + pc6 + 
    pc7 + pc8, data = pca_mortality_data)

Residuals:
   Min     1Q Median     3Q    Max 
 -95.2  -15.6   -0.8   18.2   74.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   940.05       4.31  218.12  < 2e-16 ***
pc1           -10.11       2.07   -4.89  1.2e-05 ***
pc2           -23.47       2.73   -8.61  3.2e-11 ***
pc3             7.97       3.16    2.52    0.015 *  
pc4            -4.50       3.84   -1.17    0.247    
pc5           -19.87       4.38   -4.54  3.9e-05 ***
pc6            -9.54       5.17   -1.85    0.071 .  
pc7             4.08       5.74    0.71    0.481    
pc8            13.26       6.19    2.14    0.037 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32 on 47 degrees of freedom
Multiple R-squared:  0.742,	Adjusted R-squared:  0.698 
F-statistic: 16.9 on 8 and 47 DF,  p-value: 1.83e-11
```

a.) 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.

b.) 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)

11.> Found best model using forward and backward methods and use the variables to run another regression.

```
> model_mortality6<- lm(Mortality ~pc1 + pc2 + pc3 + pc5 + pc6 + pc8 ,data=pca_mortality_data)
> summary(model_mortality6)

Call:
lm(formula = Mortality ~ pc1 + pc2 + pc3 + pc5 + pc6 + pc8, data = pca_mortality_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-100.46  -15.34    1.51   16.79   81.50 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   940.05       4.30  218.40  < 2e-16 ***
pc1           -10.11       2.07   -4.89  1.1e-05 ***
pc2           -23.47       2.72   -8.62  2.2e-11 ***
pc3             7.97       3.15    2.53    0.015 *  
pc5           -19.87       4.37   -4.54  3.6e-05 ***
pc6            -9.54       5.16   -1.85    0.070 .  
pc8            13.26       6.18    2.14    0.037 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32 on 49 degrees of freedom
Multiple R-squared:  0.731,	Adjusted R-squared:  0.698 
F-statistic: 22.2 on 6 and 49 DF,  p-value: 1.95e-12
```

a.) 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.

b.) 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)

12.> 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.