IMSE 514 – MULTIVARIATE STATISTICS
HOMEWORK 1

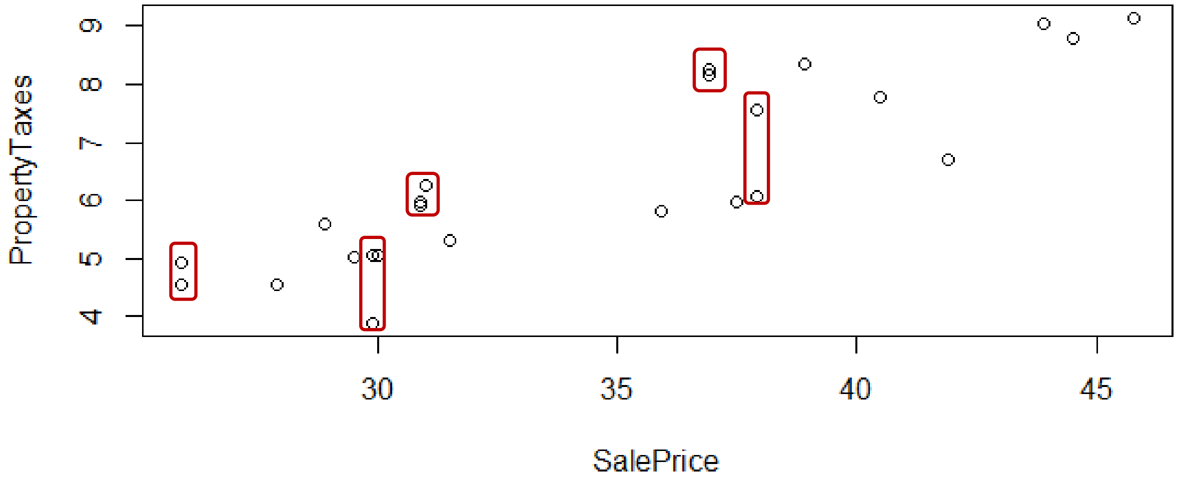
SURESH OOTY

Question 1:

a) Draw the scatter plot (let X = property taxes, and Y = selling price) and discuss what you see from this plot.

Using

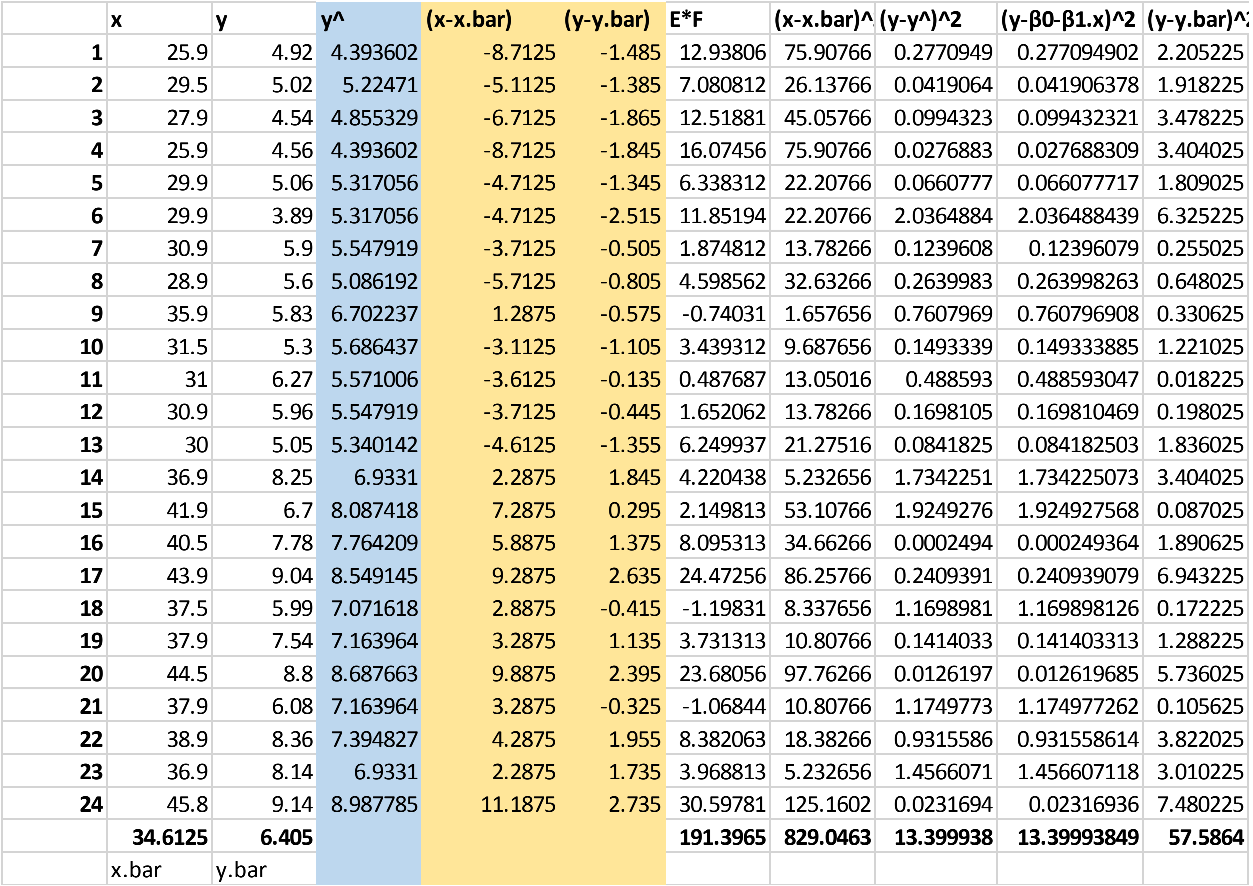
```
> plot(x<-Q1data$SalePrice..K.,y<-Q1data$PropertyTaxes..K.,xlab="SalePrice",y
lab="PropertyTaxes")
```



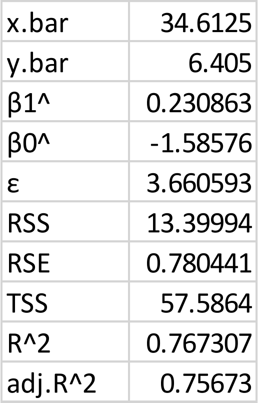
The plot indicates that the Property taxes are directly proportional to the Sale Price. Houses sold for Higher Sale Prices have higher property taxes. But it is also observed that some of houses that have same sale prices, have different property taxes. (Marked in boxes))

b) Conduct the simple regression analysis and comments on the results. (Use both equations and Statistical software)

Using Equations:



	x	y	y^	(x-x.bar)	(y-y.bar)	E*F	(x-x.bar)^	(y-y^)^2	(y-β0-β1.x)^2	(y-y.bar)^
1	25.9	4.92	4.393602	-8.7125	-1.485	12.93806	75.90766	0.2770949	0.277094902	2.205225
2	29.5	5.02	5.22471	-5.1125	-1.385	7.080812	26.13766	0.0419064	0.041906378	1.918225
3	27.9	4.54	4.855329	-6.7125	-1.865	12.51881	45.05766	0.0994323	0.099432321	3.478225
4	25.9	4.56	4.393602	-8.7125	-1.845	16.07456	75.90766	0.0276883	0.027688309	3.404025
5	29.9	5.06	5.317056	-4.7125	-1.345	6.338312	22.20766	0.0660777	0.066077717	1.809025
6	29.9	3.89	5.317056	-4.7125	-2.515	11.85194	22.20766	2.0364884	2.036488439	6.325225
7	30.9	5.9	5.547919	-3.7125	-0.505	1.874812	13.78266	0.1239608	0.12396079	0.255025
8	28.9	5.6	5.086192	-5.7125	-0.805	4.598562	32.63266	0.2639983	0.263998263	0.648025
9	35.9	5.83	6.702237	1.2875	-0.575	-0.74031	1.657656	0.7607969	0.760796908	0.330625
10	31.5	5.3	5.686437	-3.1125	-1.105	3.439312	9.687656	0.1493339	0.149333885	1.221025
11	31	6.27	5.571006	-3.6125	-0.135	0.487687	13.05016	0.488593	0.488593047	0.018225
12	30.9	5.96	5.547919	-3.7125	-0.445	1.652062	13.78266	0.1698105	0.169810469	0.198025
13	30	5.05	5.340142	-4.6125	-1.355	6.249937	21.27516	0.0841825	0.084182503	1.836025
14	36.9	8.25	6.9331	2.2875	1.845	4.220438	5.232656	1.7342251	1.734225073	3.404025
15	41.9	6.7	8.087418	7.2875	0.295	2.149813	53.10766	1.9249276	1.924927568	0.087025
16	40.5	7.78	7.764209	5.8875	1.375	8.095313	34.66266	0.0002494	0.000249364	1.890625
17	43.9	9.04	8.549145	9.2875	2.635	24.47256	86.25766	0.2409391	0.240939079	6.943225
18	37.5	5.99	7.071618	2.8875	-0.415	-1.19831	8.337656	1.1698981	1.169898126	0.172225
19	37.9	7.54	7.163964	3.2875	1.135	3.731313	10.80766	0.1414033	0.141403313	1.288225
20	44.5	8.8	8.687663	9.8875	2.395	23.68056	97.76266	0.0126197	0.012619685	5.736025
21	37.9	6.08	7.163964	3.2875	-0.325	-1.06844	10.80766	1.1749773	1.174977262	0.105625
22	38.9	8.36	7.394827	4.2875	1.955	8.382063	18.38266	0.9315586	0.931558614	3.822025
23	36.9	8.14	6.9331	2.2875	1.735	3.968813	5.232656	1.4566071	1.456607118	3.010225
24	45.8	9.14	8.987785	11.1875	2.735	30.59781	125.1602	0.0231694	0.02316936	7.480225
	34.6125	**6.405**				**191.3965**	**829.0463**	**13.399938**	**13.39993849**	**57.5864**
	x.bar	y.bar								



x.bar	34.6125
y.bar	6.405
β1^	0.230863
β0^	-1.58576
ε	3.660593
RSS	13.39994
RSE	0.780441
TSS	57.5864
R^2	0.767307
adj.R^2	0.75673

The regression equation from above XL calculation

Property Price = -1.58576 + 0.230863 (Sale Price)

Using R:

```
> srm<-lm(Q1data$PropertyTaxes..K.~Q1data$SalePrice..K.)
> summary(srm)

Call:
lm(formula = Q1data$PropertyTaxes..K. ~ Q1data$SalePrice..K.)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4271 -0.3331  0.1323  0.4966  1.3169 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -1.58576    0.95160  -1.666     0.11    
Q1data$SalePrice..K.  0.23086    0.02711   8.517 2.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7804 on 22 degrees of freedom
Multiple R-squared:  0.7673,	Adjusted R-squared:  0.7567 
F-statistic: 72.55 on 1 and 22 DF,  p-value: 2.054e-08
```

The above simple regression analysis show that the Sale price is significant in influencing the Property Tax.

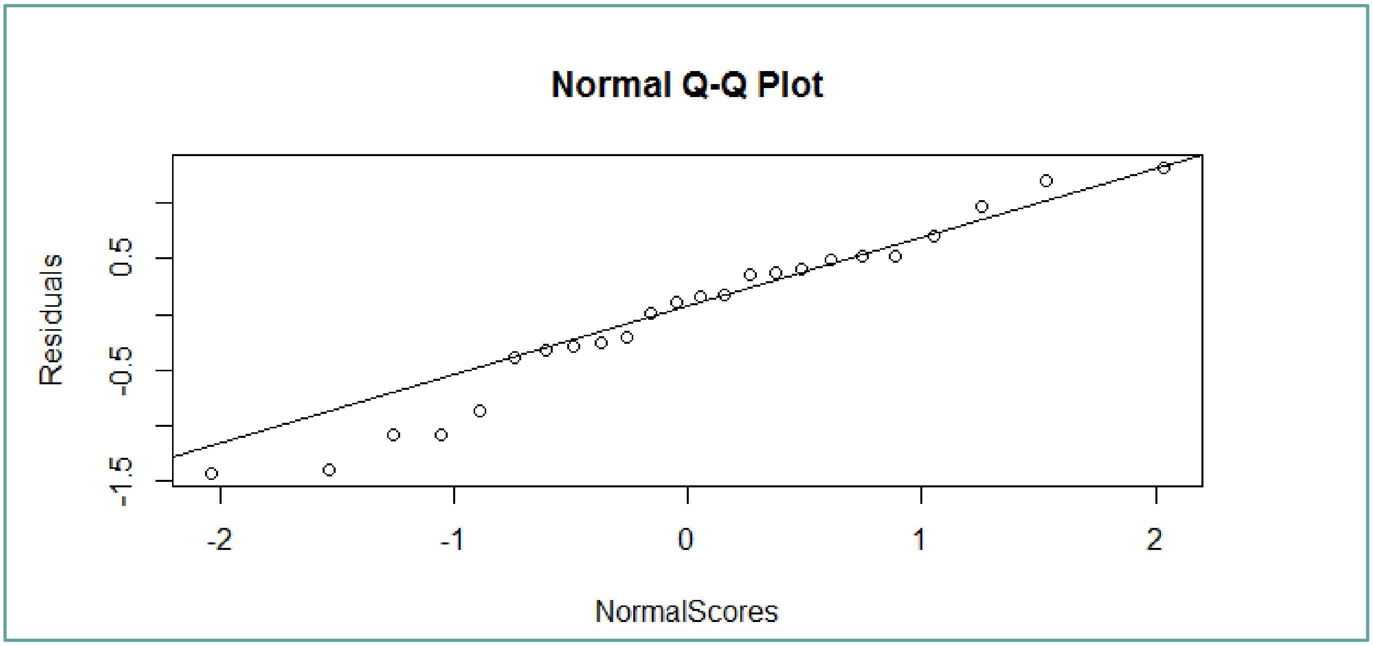
And the regression equation is

Property Tax = -1.58576 + 0.23086 (Sale Price)

c) Conduct residual analysis (including all the residuals plot, and normal probability plots).

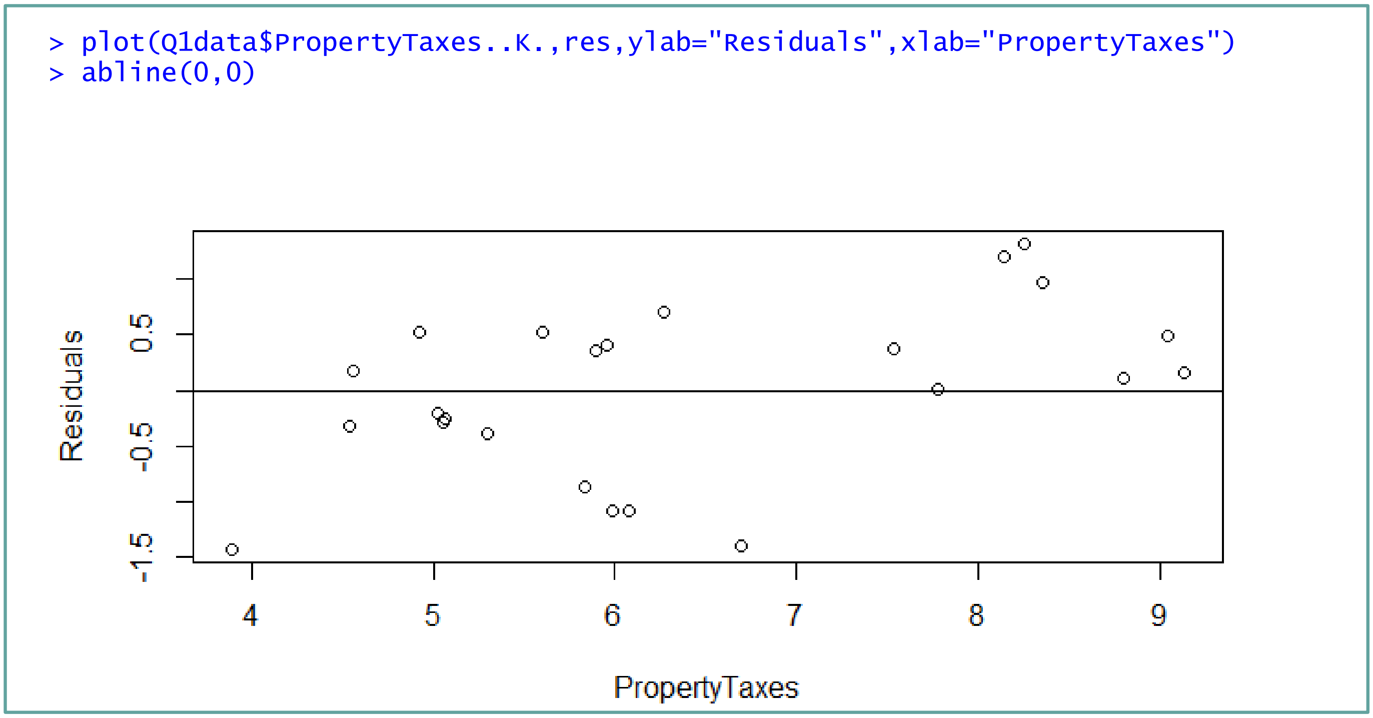
Normality Check:

```
> res<-resid(srm)
> res
          1           2           3           4           5           6 
 0.52639804 -0.20471047 -0.31532891  0.16639804 -0.25705586 -1.42705586 
          7           8           9          10          11          12 
 0.35208066  0.51380761 -0.87223673 -0.38643743  0.69899431  0.41208066 
         13          14          15          16          17          18 
-0.29014221  1.31689980 -1.38741759  0.01579128  0.49085546 -1.08161829 
         19          20          21          22          23          24 
 0.37603632  0.11233737 -1.08396368  0.96517284  1.20689980  0.15221485 
> qqnorm(res,ylab="Residuals",xlab="NormalScores")
> qqline(res)
```



The data seem to be normal.

Independence check:

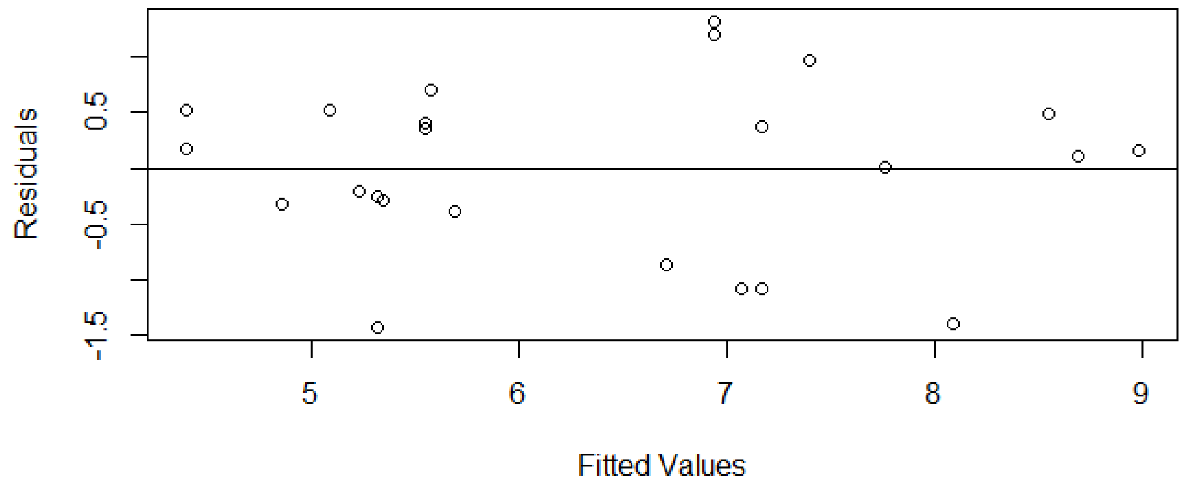


```
> plot(Q1data$PropertyTaxes..K.,res,ylab="Residuals",xlab="PropertyTaxes")
> abline(0,0)
```

The plot show the data is reasonably independent.

Homoscedasticity (Equal Variance) Check:

```
> fittedY<-fitted.values(srm)
> fittedY
       1        2        3        4        5        6        7        8 
4.393602 5.224710 4.855329 4.393602 5.317056 5.317056 5.547919 5.086192 
       9       10       11       12       13       14       15       16 
6.702237 5.686437 5.571006 5.547919 5.340142 6.933100 8.087418 7.764209 
      17       18       19       20       21       22       23       24 
8.549145 7.071618 7.163964 8.687663 7.163964 7.394827 6.933100 8.987785 
> plot(fittedY,res,ylab="Residuals",xlab="Fitted Values")
> abline(0,0)
```



The plot shows that homoscedasticity assumption is not violated.

d) If someone pays $6K annual taxes, what is the predicted selling price (based on your regression model)

From the equation :> Property Tax = -1.58576 + 0.23086 (Sale Price)

Sale Price = (6 + 1.58576) / 0.23086 = $32.85822K

2. Variation in gasoline mileage among makes and models of automobiles is influenced substantially by several factors such as weight of the car, horsepower and etc. We measured the MPG (miles per gallon) and some other information of 82 different cars which were made by different auto companies from the US, Japan, and Europe. We are interested to know what might be the important factors for the MPG of a car. The candidate predictor variables are:

- □ VOL: Cubic feet of cab space
- □ HP: Engine horsepower
- □ SP: Top speed (mph)
- □ WT: Vehicle weight (100 lb)
- □ AREA: where the car was made (1: US, 2: Europe, 3: Japan)

Please conduct possible thorough analysis based on what we have discussed in the first two lectures. In addition to using the statistical software, please apply the equations from the lecture notes as much as you could. Don't forget to discuss your analytical results.

Using R with library (leaps):

```
> mil<-read.csv("HW1-Q2-data.csv",header=T)
> null<-lm(MPG~1,data=mil)
> AREA<-factor(AREA)
> full<-lm(MPG~VOL+HP+SP+WT+AREA)
```

The data was carefully loaded, such that AREA is made as factor. And the variables VOL, HP, SP, WT & AREA were chosen for full.

```
> step(full,data=mil,direction="both")
> step(null,scope=list(lower=null,upper=full),direction="forward")
> step(full,data=mil,direction="backward")
```

For all the above three trials, the final regression equation was exactly same as below.

```
Start:  AIC=220.93
MPG ~ VOL + HP + SP + WT + AREA

       Df Sum of Sq    RSS    AIC
- AREA  2      4.69 1027.4 217.30
- VOL   1      4.81 1027.5 219.31
<none>              1022.7 220.93
- HP    1    262.96 1285.7 237.69
- SP    1    318.87 1341.6 241.18
- WT    1    920.03 1942.7 271.54

Step:  AIC=217.3
MPG ~ VOL + HP + SP + WT

       Df Sum of Sq    RSS    AIC
- VOL   1      6.27 1033.7 215.80
<none>              1027.4 217.30
- HP    1    309.67 1337.0 236.90
- SP    1    373.36 1400.7 240.72
- WT    1   1013.76 2041.2 271.59

Step:  AIC=215.8
MPG ~ HP + SP + WT

       Df Sum of Sq    RSS    AIC
<none>              1033.7 215.80
- HP    1    349.37 1383.0 237.68
- SP    1    396.97 1430.6 240.45
- WT    1   1322.87 2356.5 281.37

Call:
lm(formula = MPG ~ HP + SP + WT)

Coefficients:
(Intercept)           HP           SP           WT  
   194.1296       0.4052      -1.3200      -1.9221  
```

```
Regression equation:
```

MPH = 194.1296 + 0.4052 (HP) – 1.32 (SP) - 1.9221 (WT)

```
> summary(lm(MPG ~ WT + SP + HP))

Call:
lm(formula = MPG ~ WT + SP + HP)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1633 -2.8387  0.2464  1.7889 12.5566 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 194.12962   23.32213   8.324 2.22e-12 ***
WT           -1.92210    0.19238  -9.991 1.31e-15 ***
SP           -1.32000    0.24118  -5.473 5.19e-07 ***
HP            0.40518    0.07891   5.135 2.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.64 on 78 degrees of freedom
Multiple R-squared:  0.8725,	Adjusted R-squared:  0.8676 
F-statistic: 177.9 on 3 and 78 DF,  p-value: < 2.2e-16
```

The adjusted R2 value is satisfactory and the P values of the variables show that they are significant.

The equation shows that “weight of the vehicle” & “top speed” tend to reduce the MPG achieved, whereas the “manufactured country” & “volume in cabin space” does not have any influence to the mileage achieved.

Further observation:
When the **MAKE.MODEL** was tried as a factor, the regression model showed that none of the variables could have significant influence. Hence, it was a wise decision to keep the **MAKE.MODEL** away from the list of predictors.