CS 5565

Introduction to statistical learning

Linear Regression in R

1.

R is installed on a windows machine and data sets from http://www-bcf.usc.edu/˜ gareth/ISL/ also being used.

2. a)

```
> save.image("C:\\Users\\Shubh\\Documents\\myauto.RData")
> attach(auto)
> lm(mpg~horsepower)

Call:
lm(formula = mpg ~ horsepower)

Coefficients:
(Intercept)   horsepower
    39.9359      -0.1578

> regcar = lm(mpg~horsepower)
> summary(regcar)

Call:
lm(formula = mpg ~ horsepower)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i) Yes, there is relation between predictor and response.

ii) The relationship between mpg & horsepower is strong because the standard error for coefficient is very low (.006446). Also, the p value is very small (<2.2e – 16).

iii) The relationship between predictor and response is negative because, the coefficient value is negative (-.157845).

iv)

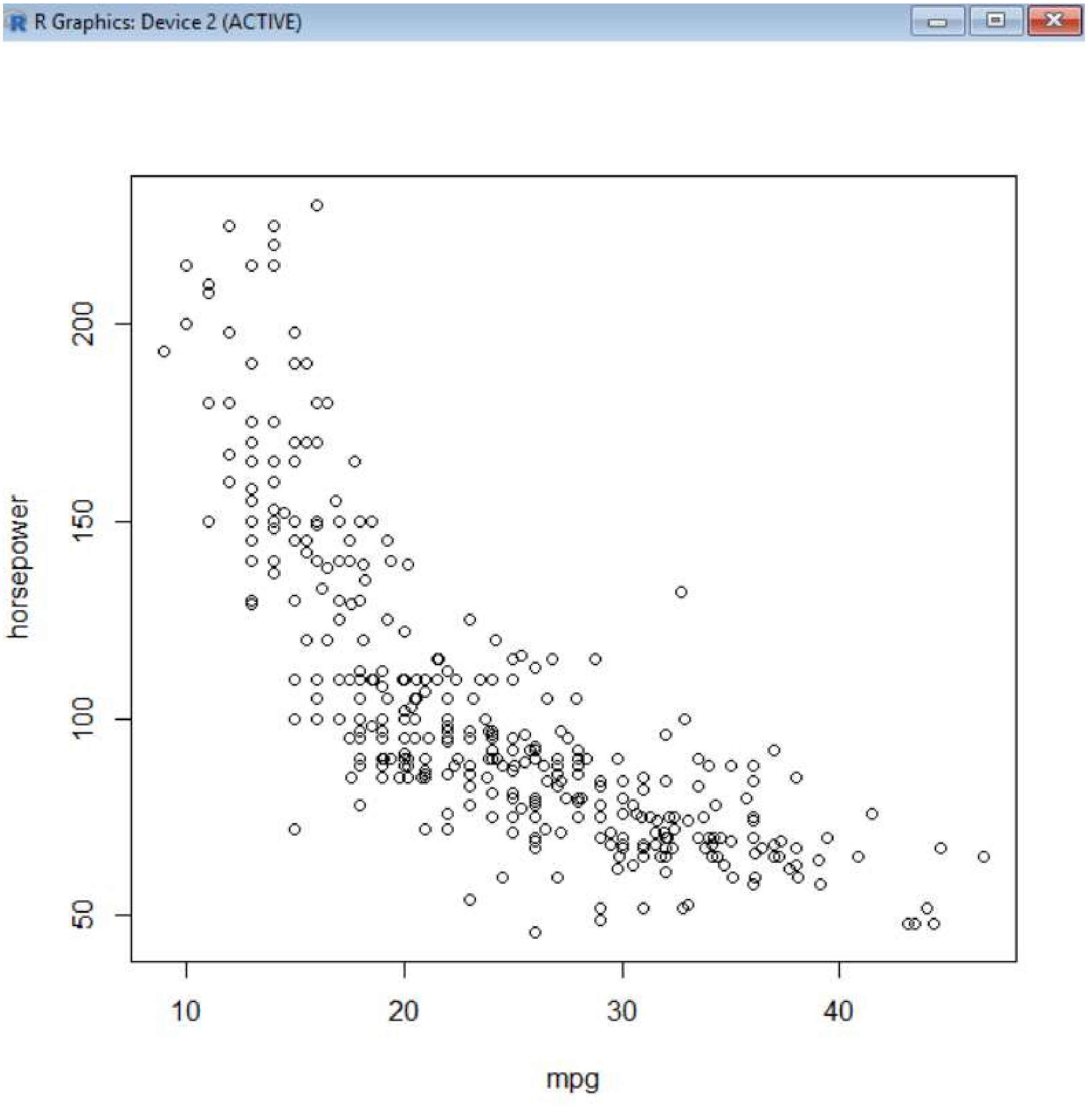
```
> predict(regcar,data.frame(horsepower=98),interval="confidence",level=.95)
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

And

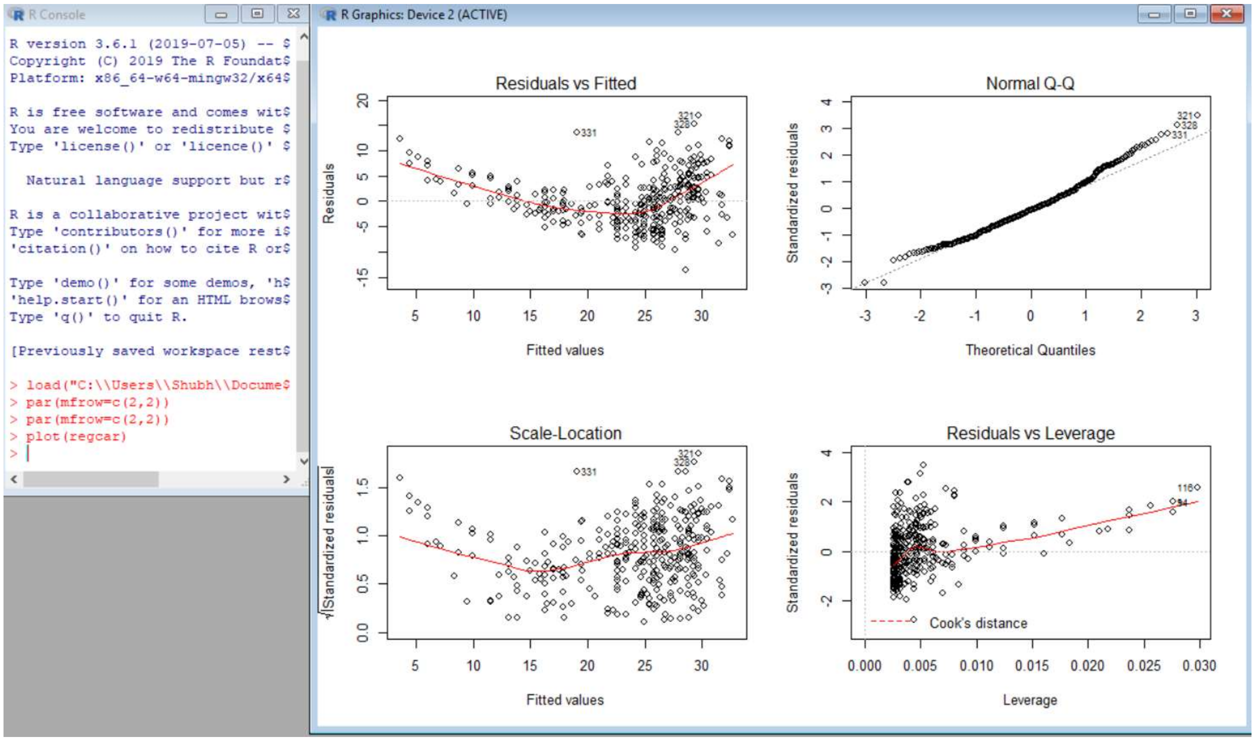
```
> predict(regcar,data.frame(horsepower=98),interval="prediction",level=.95)
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

b)

```
> plot(~mpg+horsepower,auto)
```



c) The diagnostic plot is shown as below:

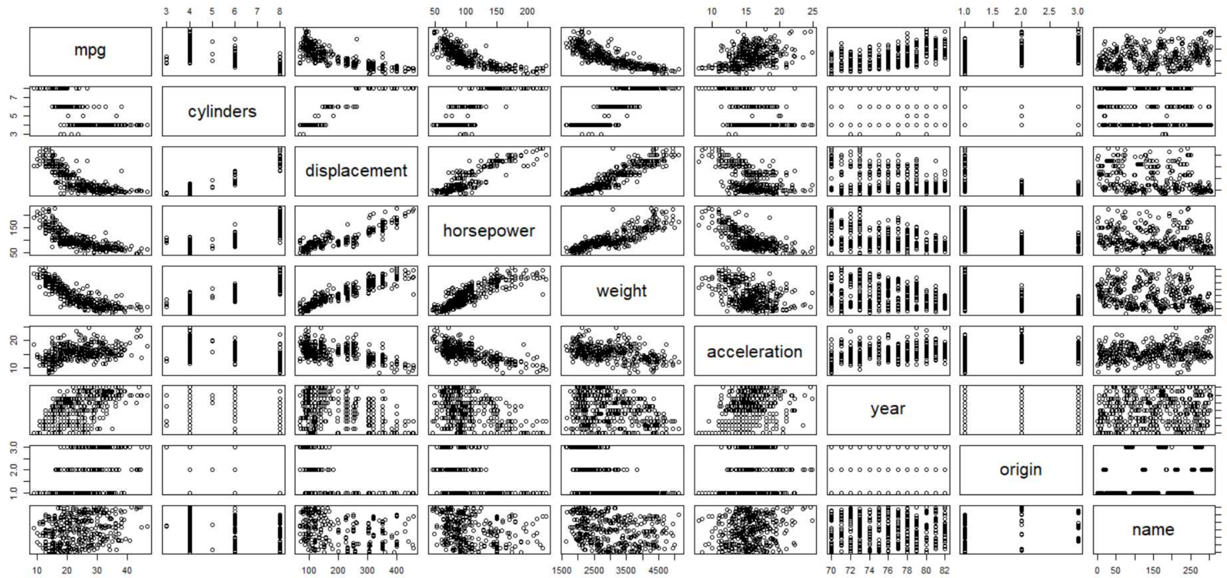


The problems we can see from these plots:

1. Few data points are observed as outliners like 310, 328, 331
2. A strong pattern in residuals indicate indicates non-linearity in the data.
3. From Residuals V fitted value plot shows a non-constant variance in error terms.
4. In Residual V leverage plot shows high leverage point.

3) The scatterplot is shown below:

```
> load("C:\\Users\\Shubh\\Documents\\myauto_lab2.RData")
> pairs(auto)
```



b)

```
> load("C:\\Users\\Shubh\\Documents\\myauto_lab2.RData")
> fix(auto)
> newauto=auto[,-9]
> fix(newauto)
> cor(newauto)
                    mpg  cylinders displacement horsepower     weight
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
             acceleration       year     origin
mpg             0.4233285  0.5805410  0.5652088
cylinders      -0.5046834 -0.3456474 -0.5689316
displacement   -0.5438005 -0.3698552 -0.6145351
horsepower     -0.6891955 -0.4163615 -0.4551715
weight         -0.4168392 -0.3091199 -0.5850054
acceleration    1.0000000  0.2903161  0.2127458
year            0.2903161  1.0000000  0.1815277
origin          0.2127458  0.1815277  1.0000000
> save.image("C:\\Users\\Shubh\\Documents\\myauto_lab2_new")
```

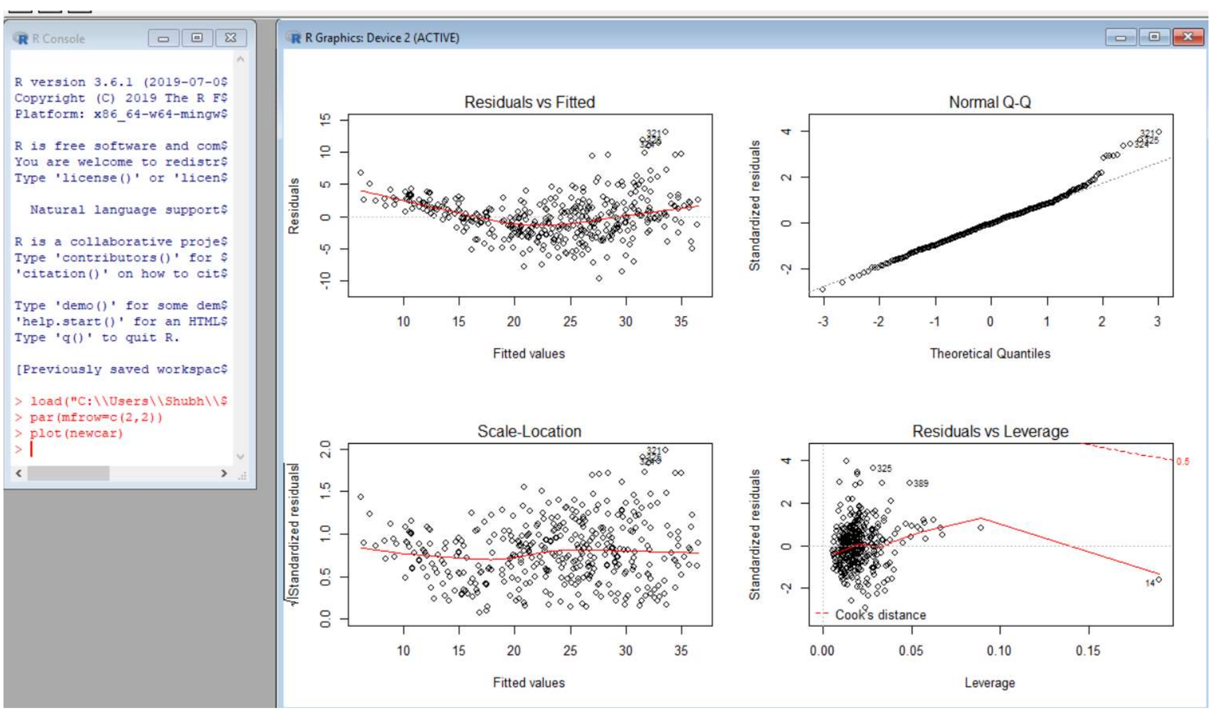
c) i) Yes, there is relationship between predictors & the response.

```
> attach(newauto)
> newcar=lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)
> summary(newcar)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

ii) We can see that the p value is lowest for weight (< 2e -16), year (< 2e -16) and origin (4.67e-07). Hence, these predictors appear to have a statistically significant relationship to the response.

iii) From the coefficient of year we can see that, estimate is non zero, standard error has a low value, the t-value is high and also p-value is very low (< 2e -16), these implies that year has a strong predictor relationship with response mpg.

d) The diagnostic plot is shown as below:



So, the problems we can see in this plot are:

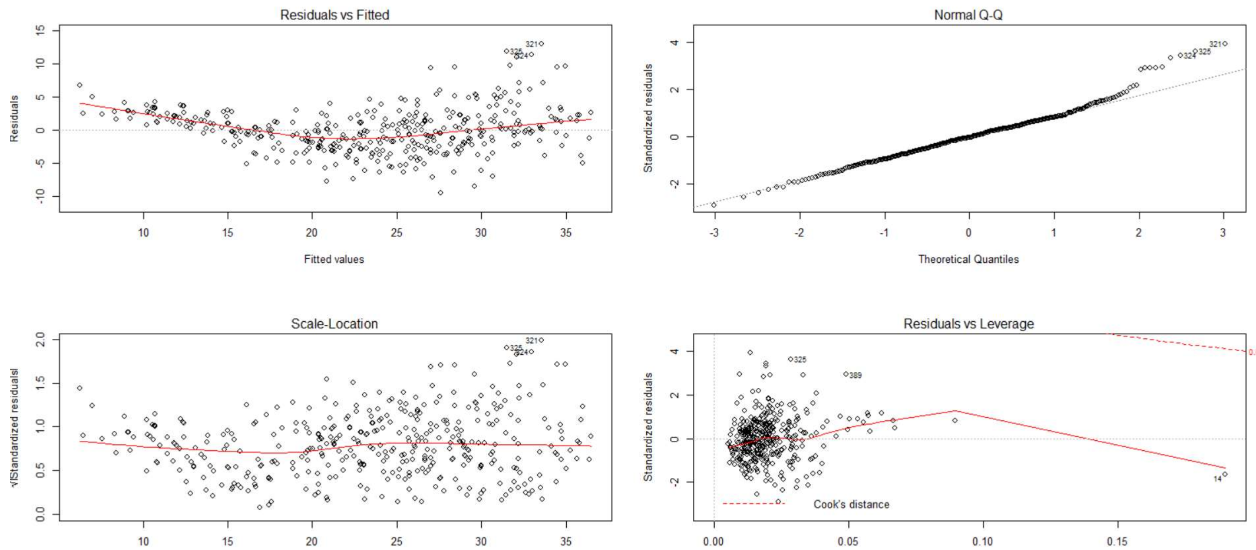
- From normal Q-Q plot, we can see there are outliers in 321,324, 325
- From Residuals V Leverage plot, we see high leverage point found in 14.
- Residual vs fitted plot shows a pattern that indicates non-linearity in the data set. But the pattern is not evident as we saw in case of simple linear regression plots.

e)

<u>Observation – 1 : Without interaction</u>

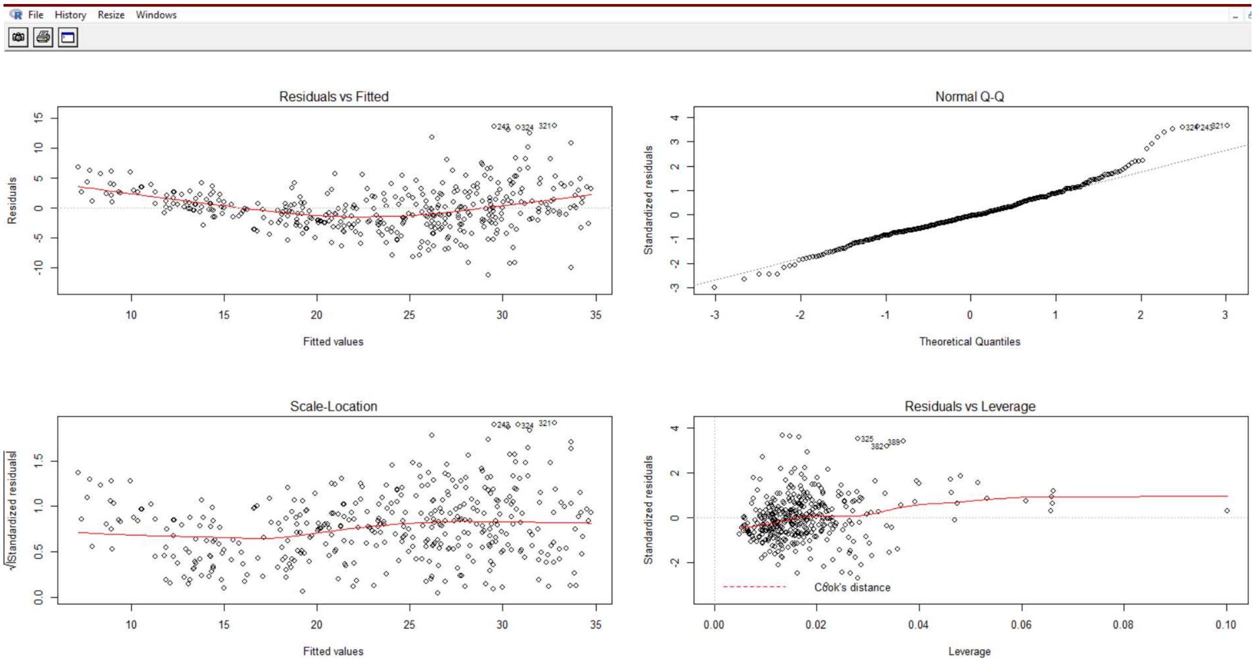
```
> summary(newcar)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(newcar)
```



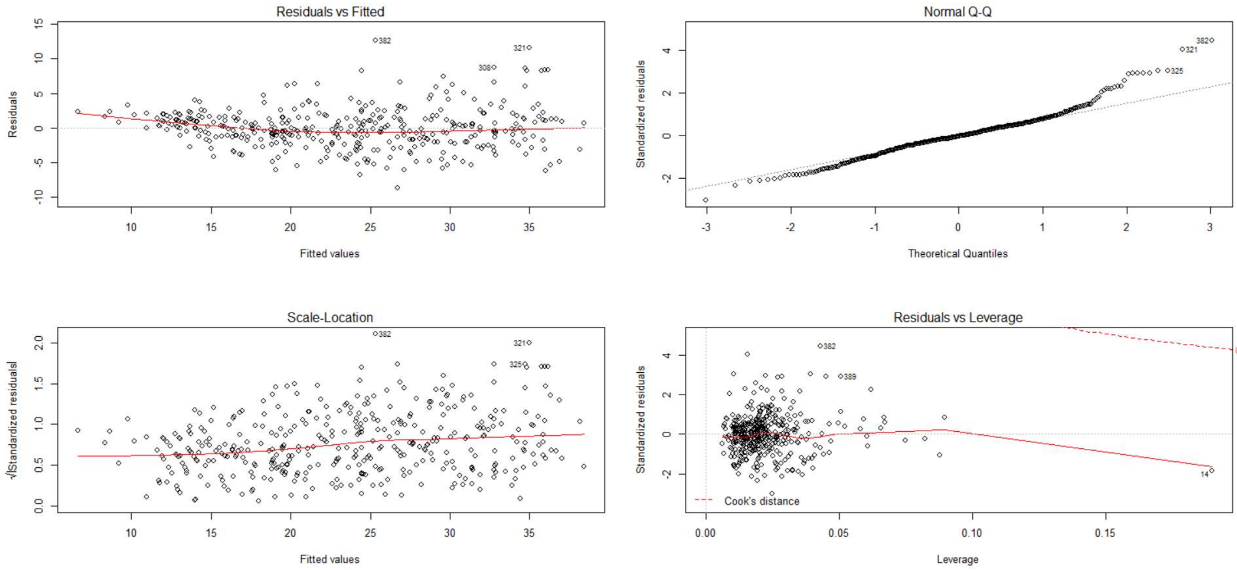
Observation – 2 : interaction type A:

```
> attach(newauto)
>  newcar1 = lm(mpg~cylinders+displacement+horsepower:weight+acceleration+year+origin)
> summary(newcar1)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower:weight +
    acceleration + year + origin)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2355  -2.2876  -0.1861   2.2236  13.7952

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.809e+01  4.782e+00  -3.784 0.000179 ***
cylinders         -8.878e-01  3.654e-01  -2.430 0.015558 *
displacement      -9.767e-03  8.550e-03  -1.142 0.254062
acceleration      -1.933e-01  8.705e-02  -2.221 0.026963 *
year               6.921e-01  5.675e-02  12.195  < 2e-16 ***
origin             1.651e+00  3.175e-01   5.200 3.25e-07 ***
horsepower:weight -1.146e-05  2.596e-06  -4.416 1.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.788 on 385 degrees of freedom
Multiple R-squared:  0.7681,    Adjusted R-squared:  0.7645
F-statistic: 212.5 on 6 and 385 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(newcar1)
>
```



Observation – 3 : interaction type B:

```
> newcar2 = lm(mpg~cylinders+displacement*horsepower+weight+acceleration+year+origin)
> summary(newcar2)

Call:
lm(formula = mpg ~ cylinders + displacement * horsepower + weight +
    acceleration + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7010 -1.6009 -0.0967  1.4119 12.6734

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.894e+00  4.302e+00  -0.440  0.66007
cylinders                6.466e-01  3.017e-01   2.143  0.03275 *
displacement            -7.487e-02  1.092e-02  -6.859 2.80e-11 ***
horsepower              -1.975e-01  2.052e-02  -9.624  < 2e-16 ***
weight                  -3.147e-03  6.475e-04  -4.861 1.71e-06 ***
acceleration            -2.131e-01  9.062e-02  -2.351  0.01921 *
year                     7.379e-01  4.463e-02  16.534  < 2e-16 ***
origin                   6.891e-01  2.527e-01   2.727  0.00668 **
displacement:horsepower  5.236e-04  4.813e-05  10.878  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.912 on 383 degrees of freedom
Multiple R-squared:  0.8636,    Adjusted R-squared:  0.8608
F-statistic: 303.1 on 8 and 383 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(newcar2)
```

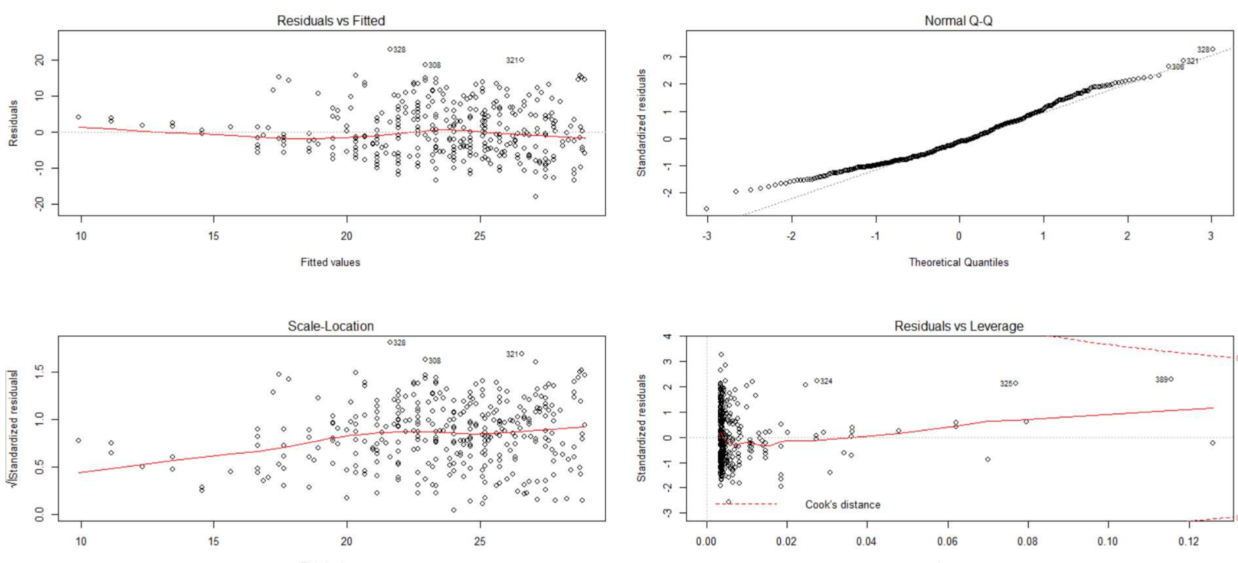


Observations:

We observe that, observation 1 & observation 2 do not show much of the difference in residual plots, but observation 3 shows some improvement in Residuals V Fitted plot, as the residuals are much linear in nature, which describes a better fit.

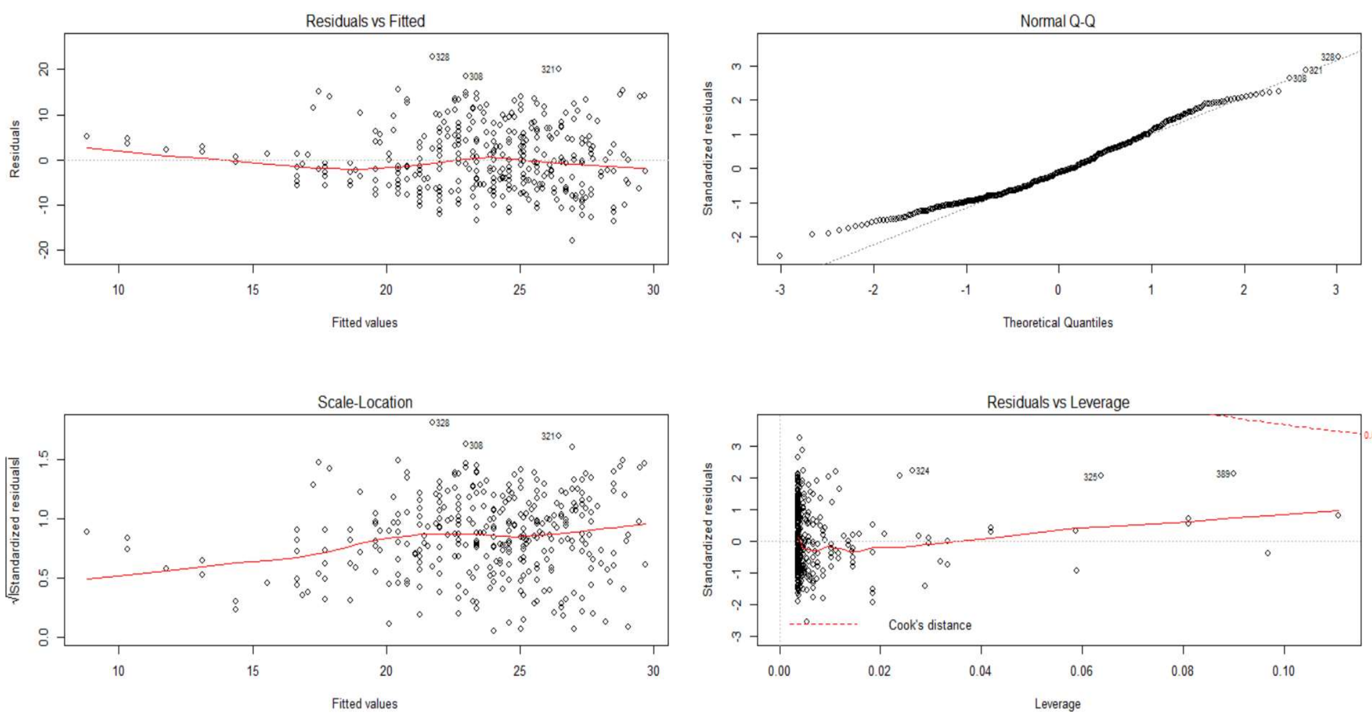
f) Below are few transformations of the variables:

Observation -1 : transformation as X^2:



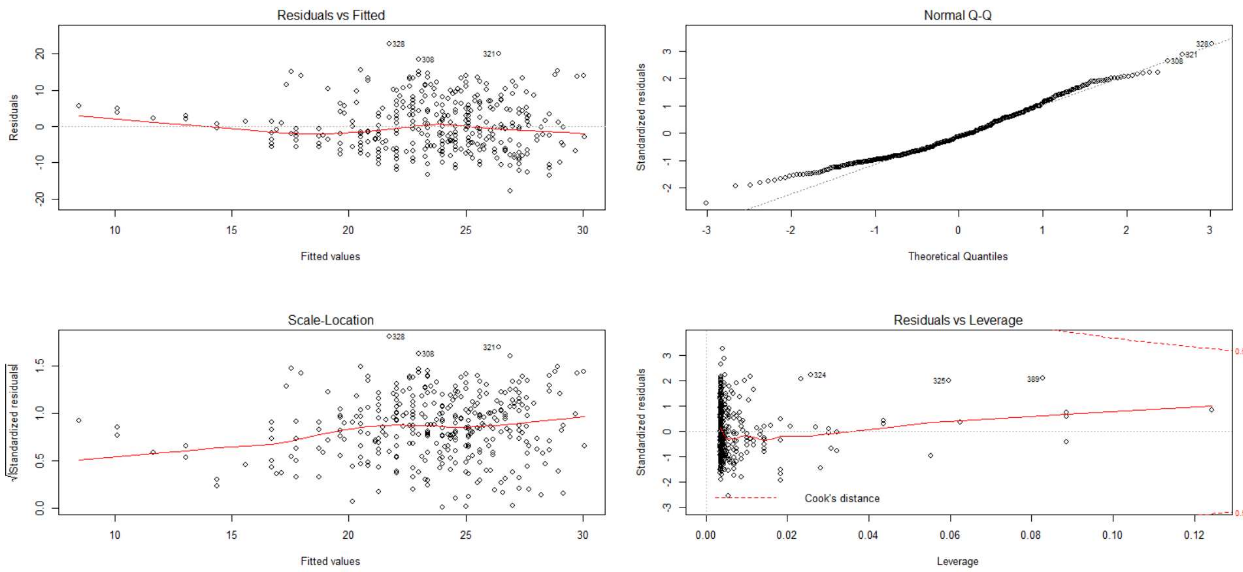
```
> load("C:\\Users\\Shubh\\Documents\\myauto_new.RData")
> attach(auto)
> newcarx=lm(mpg~acceleration+I(acceleration^2))
> par(mfrow=c(2,2))
> plot(newcarx)
```

Observation -2 : transformation as X^.5:




```
> newcary=lm(mpg~acceleration+I(acceleration^.5))
> par(mfrow=c(2,2))
> plot(newcary)
```

Observation -3 : transformation as log(X):




```
> newcarz=lm(mpg~acceleration+I(log(acceleration)))
> par(mfrow=c(2,2))
> plot(newcarz)
```

Final observations:

So, we can see a common thing for all observations 1, 2 & 3 that, Residuals plots are much more flat & hardly any pattern can be seen there.

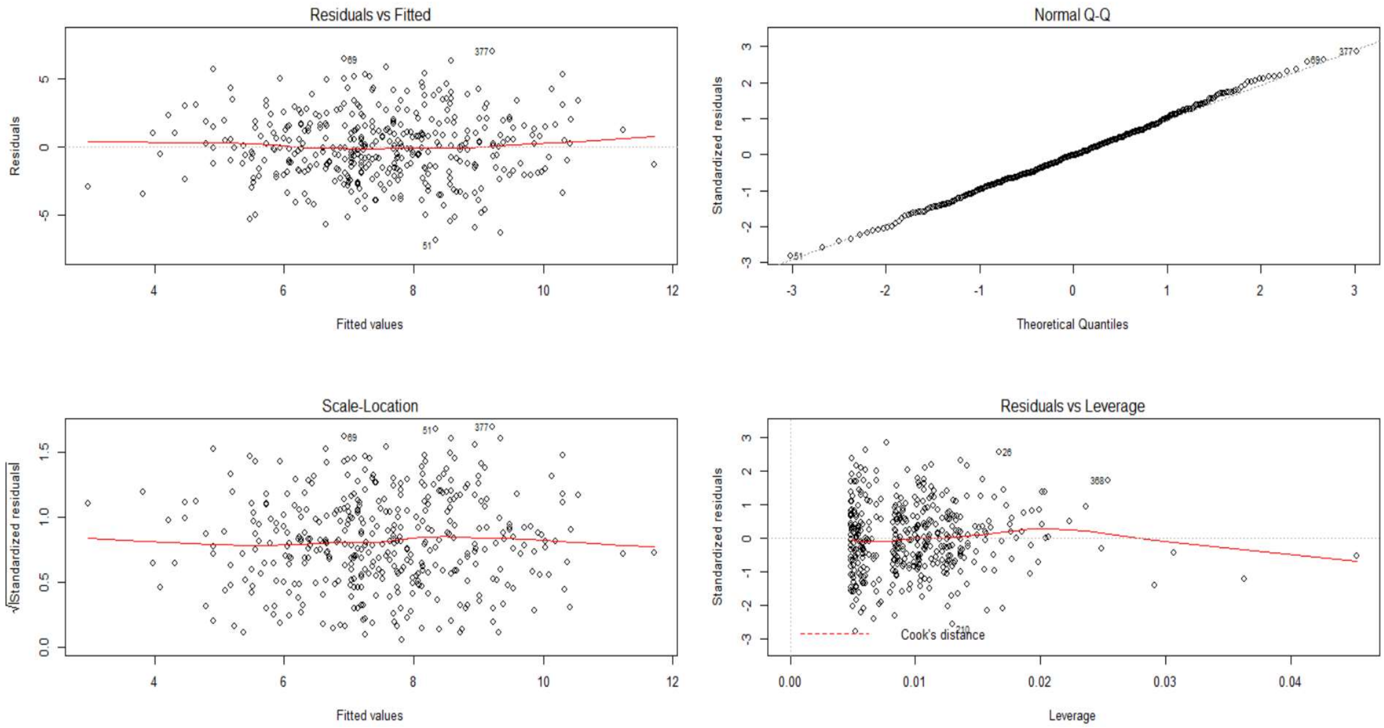
Also, there is no high leverage point found in these plots.

Hence, we can infer that, these are better fit compared to the predecessors (with X).

4) a) Below is the regression model for Carseat:

```
> load("C:\\Users\\Shubh\\Documents\\seat.RData")
> attach(Carseats)
> lm.fit=lm(Sales~Price+Urban+US)
> save.image("C:\\Users\\Shubh\\Documents\\seat.RData")
> summary(lm.fit)

Call:
lm(formula = Sales ~ Price + Urban + US)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16

> contrasts(Urban)
    Yes
No    0
Yes   1
> contrasts(US)
    Yes
No    0
Yes   1
> par(mfrow=c(2,2))
> plot(lm.fit)
```

Now when we plot this model we get below:



b) **<u>Interpretation of the coefficients are as below:</u>**

Residuals red line is almost flat in Fitted value plot, hence residuals do not show any strong pattern.

Price has a negative coefficient (-0.054459), that means if price increases the sales decreases.

First dummy variable UrbanYes has negative coefficient (-0.021916) , that means when we go to Urban area the sales decreases compared to Rural area.

Second dummy variable USYes has positive coefficient (1.200573) , that means US has higher sales than non-US countries.

c) **<u>The model in equation form:</u>**

Sales = (13.043469) + (– 0.054459) *Price + (- 0.021916)*Urban+ (1.200573)* US

From Contrast command we can see the interpretation of the dummy variables, hence:

If Urban = Yes, it's value is 1, If Urban = No, it's value is 0

If US = Yes, it's value is 1, If US = No, it's value is 0

d) To predict for which of the predictors can you reject the null hypothesis, we construct a multiple-regression model as below:

```
> summary(Carseats)
     Sales          CompPrice       Income        Advertising       Population        Price        ShelveLoc         Age          Education     Urban       US
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0   No :118   No :142
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0   Yes:282   Yes:258
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
> lm.fit1=lm(Sales~CompPrice+Income+Advertising+Population+Price+ShelveLoc+Age+Education+Urban+US)
> summary(lm.fit1)

Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Population +
    Price + ShelveLoc + Age + Education + Urban + US)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8692 -0.6908  0.0211  0.6636  3.4115

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
Income           0.0158028  0.0018451   8.565 2.58e-16 ***
Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
Population       0.0002079  0.0003705   0.561    0.575
Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
Education       -0.0211018  0.0197205  -1.070    0.285
UrbanYes         0.1228864  0.1129761   1.088    0.277
USYes           -0.1840928  0.1498423  -1.229    0.220
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8698
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

From this model we can see that, for Population, Education, UrbanYes & USYes P value is high (> .01), Hence we **cannot** reject null hypothesis for these 4 predictors.

Other than that, all the predictors have non-zero coefficients, as well as much lower p value (< <.01), so for these variables we can reject null hypothesis for these predictors. These are:

CompPrice   Income   Advertising   Price   ShelveLoc   Age.

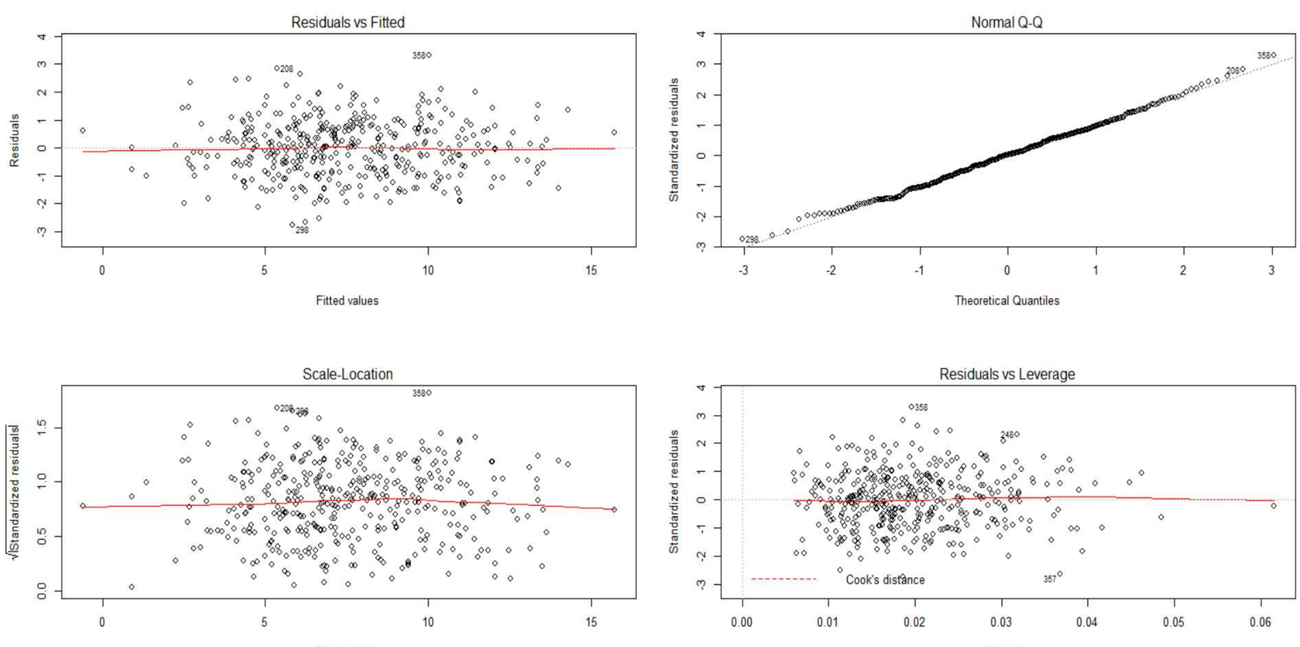
e)

Now if we want to fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. We need to construct a multiple regression model using

CompPrice   Income   Advertising   Price   ShelveLoc   Age.

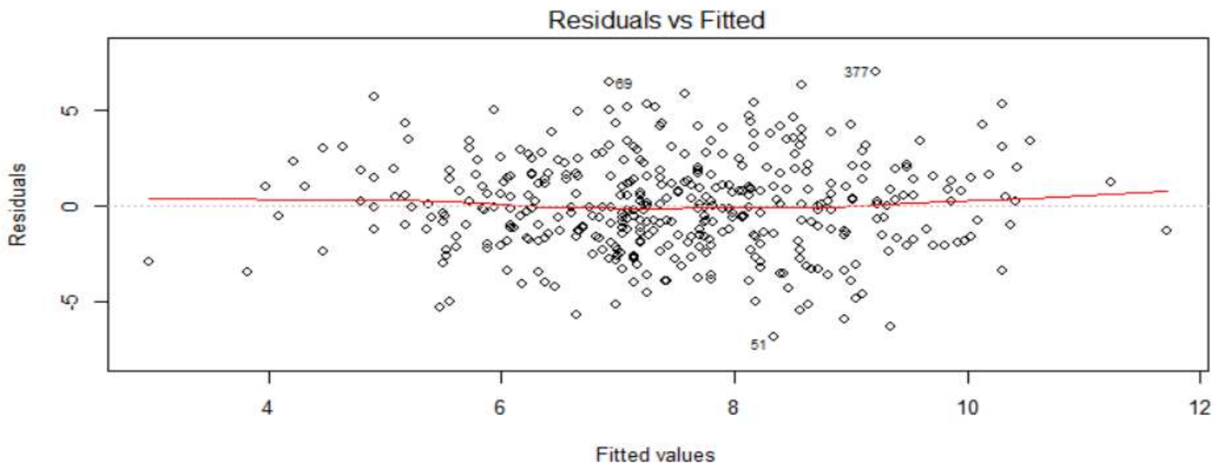
The model is shown as below:

```
> load("C:\\Users\\Shubh\\Documents\\seat.RData")
> attach(Carseats)
> lm.fit1=lm(Sales~CompPrice+Income+Advertising+Price+ShelveLoc+Age)
> summary(lm.fit1)

Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
    ShelveLoc + Age)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7728 -0.6954  0.0282  0.6732  3.3292

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.475226   0.505005   10.84   <2e-16 ***
CompPrice        0.092571   0.004123   22.45   <2e-16 ***
Income           0.015785   0.001838    8.59   <2e-16 ***
Advertising      0.115903   0.007724   15.01   <2e-16 ***
Price           -0.095319   0.002670  -35.70   <2e-16 ***
ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
Age             -0.046128   0.003177  -14.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.019 on 392 degrees of freedom
Multiple R-squared:  0.872,     Adjusted R-squared:  0.8697
F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(lm.fit1)
```

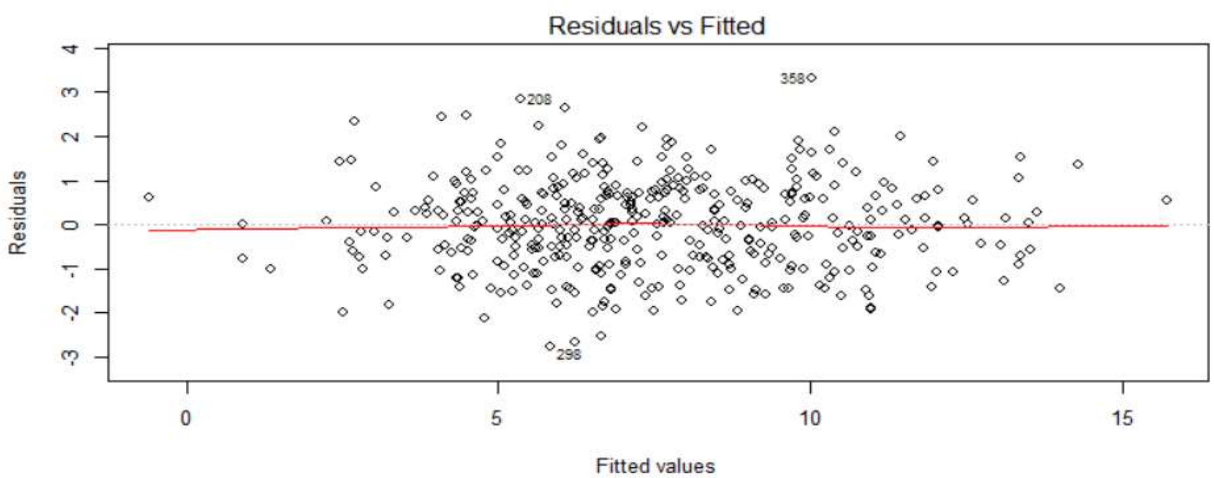


f) Now if we compare (a) & (e) to see how well they fit the data come as below:

<u>for (a) :</u>



<u>for (e) :</u>



So, we can see that model (e) slightly fitted better compared to (a).

g)

95% confidence intervals for the coefficient(s) are below:

```
> confint(lm.fit1)
                      2.5 %      97.5 %
(Intercept)      4.48236820  6.46808427
CompPrice        0.08446498  0.10067795
Income           0.01217210  0.01939784
Advertising      0.10071856  0.13108825
Price           -0.10056844 -0.09006946
ShelveLocGood    4.53585700  5.13549250
ShelveLocMedium  1.70550103  2.19848429
Age             -0.05237301 -0.03988204
```

h)

We observe are outliers at (208, 358, 298).

But we do not observe any high leverage observations in the model from (e).

5)

By performing a simple linear regression of y onto x, without an intercept we get:

```
> set.seed(1)
> x = rnorm(100)
> y=2*x+rnorm(100)
> lm.regfit=lm(y~x+0)
> summary(lm.regfit)

Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9154 -0.6472 -0.1771  0.5056  2.3109 

Coefficients:
  Estimate Std. Error t value Pr(>|t|)    
x   1.9939     0.1065   18.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared:  0.7798,	Adjusted R-squared:  0.7776 
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

Observations:

The estimated value for x is a non-zero one. Also, Standard error is 0.1065, the t value is big and the p – value is <2e-16, which is much less compared to .01. All these results imply that, H0 or null hypothesis can be strongly rejected.

b)

By performing a simple linear regression of y onto x, without an intercept we get:

```
> lm.regfitnew=lm(x~y+0)
> summary(lm.regfitnew)

Call:
lm(formula = x ~ y + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2368  0.1030  0.2858  0.8938

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y  0.39111    0.02089   18.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

Observations:

The estimated value for y is a non-zero one. Also, Standard error is 0.02089, the t value is big and the p –value is <2e-16, which is much less compared to .01. All these results imply that, H0 or null hypothesis can be strongly rejected.

c) By comparing (a) & (b) we get below relation between them:

Both has same t-statistics and p-value.

From (a) we can estimate the equation is like:

Y = 2*x + error …………………….   β = 1.9939 ~ 2

From (a) we can estimate the equation is like:

x = (1/2) *y + error …………………….   β = .39111 ~ .5

So, basically both are indicating the same linear equation, Y = 2x

d)

5) d) As mentioned, for the regression Y onto x, without an intercept, the t-statistic for $H_0: \beta = 0$ takes below form ⟶

$$t = \hat{\beta} / SE(\hat{\beta}) \quad \text{------} \quad (1)$$

Now given, $SE(\hat{\beta}) = \sqrt{\dfrac{\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2}{(n-1)\sum_{i'=1}^{n} x_{i'}^2}}$ ------ (2)

from, equation 3.38 we get ⟶

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i Y_i\right) \Big/ \left(\sum_{i'=1}^{n} x_{i'}^2\right) \quad \text{------} \quad (3)$$

So, substituting $\hat{\beta}$ from (3) to equation (1) we get ⟶

$$t = \frac{\left(\sum_{i=1}^{n} x_i Y_i\right) \Big/ \left(\sum_{i'=1}^{n} x_{i'}^2\right)}{SE(\hat{\beta})} \quad \text{------} \quad (4)$$

Now, substituting $SE(\hat{\beta})$ from (2) to (4) we get ⟶

$$t = \frac{\left(\sum_{i=1}^{n}(x_i y_i) \Big/ \sum_{i'=1}^{n}(x_{i'})^2\right)}{\sqrt{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 \Big/ (n-1)\sum_{i'=1}^{n} x_{i'}^2}}$$

$1 \quad 1/a$

$$= \frac{(\sqrt{n-1})\left(\sum_{i=1}^{n}(x_i y_i)\right) \times \sqrt{a}}{\sqrt{\sum_{i=1}^{n}(y_i^2 - 2x_i y_i\hat{\beta} + x_i^2\hat{\beta}^2) \times \cancel{a}}\,(\sqrt{a})}$$

$\left[\text{say } \sum_{i'=1}^{n} x_{i'}^2 = a\right]$

$$= \frac{(\sqrt{n-1})\left(\sum_{i=1}^{n}(x_i y_i)\right)}{\sqrt{\left(\sum_{i'=1}^{n}x_{i'}^2\right)\left(\sum_{i=1}^{n}y_i^2\right) - \left(\sum_{i'=1}^{n}x_{i'}^2\right)\left(\sum_{i=1}^{n}2x_i y_i\right)\left(\sum_{i=1}^{n}(x_i y_i) \Big/ \cancel{\sum_{i'=1}^{n}x_{i'}^2}\right) + \sum_{i'=1}^{n}(x_{i'}^2)\sum_{i=1}^{n}(x_i^2)\left(\sum_{i=1}^{n}x_i y_i \Big/ \sum_{i'=1}^{n}x_i^2\right)^2}}$$

$$t = \frac{(\sqrt{n-1})\left(\sum_{i=1}^{n}(x_i y_i)\right)}{\sqrt{\left(\sum_{i=1}^{n}x_i^2\right)\left(\sum_{i=1}^{n}y_i^2\right) - \left(\sum_{i=1}^{n}x_i y_i\right)^2}}$$

Hence proved, t can be written the way mentioned in question.

So to confirm numerically from R, we put the formula in R we get results as below:

```
> load("C:\\Users\\Shubh\\Documents\\last.RData")
> n=length(x)
> t=sqrt(n-1)*(x%*%y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
> t
         [,1]
[1,] 18.72593
```

Hence this proves that t statistic can be written as

$$\frac{(\sqrt{n-1})\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i'=1}^{n} y_{i'}^2) - (\sum_{i'=1}^{n} x_{i'} y_{i'})^2}}$$

e) By performing a simple linear regression of y onto x, without intercept we get:

```
> set.seed(1)
> x = rnorm(100)
> y=2*x+rnorm(100)
> lm.regfit=lm(y~x+0)
> summary(lm.regfit)

Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9154 -0.6472 -0.1771  0.5056  2.3109 

Coefficients:
  Estimate Std. Error t value Pr(>|t|)    
x   1.9939     0.1065   18.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared:  0.7798,	Adjusted R-squared:  0.7776 
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

By performing a simple linear regression of x onto y, without intercept we get:

```
> lm.regfitnew=lm(x~y+0)
> summary(lm.regfitnew)

Call:
lm(formula = x ~ y + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2368  0.1030  0.2858  0.8938

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y  0.39111    0.02089   18.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

So, we see that when regression is performed **without an intercept**, the t-statistic for H0: β = 0. is the same for the regression of y onto x as it is for the regression, which is ~ 18.73

f)

By performing a simple linear regression of y onto x, with intercept we get:

```
> lm.regfit1=lm(y~x)
> summary(lm.regfit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8768 -0.6138 -0.1395  0.5394  2.3462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03769    0.09699  -0.389    0.698
x            1.99894    0.10773  18.556   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

By performing a simple linear regression of x onto y, with intercept we get:

```
> lm.regfitnew1=lm(x~y)
> summary(lm.regfitnew1)

Call:
lm(formula = x ~ y)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90848 -0.28101  0.06274  0.24570  0.85736 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.03880    0.04266    0.91    0.365    
y            0.38942    0.02099   18.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4249 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762 
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

So, we see that when regression is performed with an intercept, the t-statistic for H0: β = 0. is the same for the regression of y onto x as it is for the regression, which is ~ 18.56