Problem 4 : Linear Regression

Installing ISLR package:
install.packages("ISLR")
Loading the library to use the datasets
library(ISLR)
Loading data:
data("Auto")
Viewing the first 10 examples using the head function
head(Auto, 10)

```
> head(Auto, 10)
   mpg cylinders displacement horsepower weight acceleration year origin                      name
1   18         8          307        130   3504         12.0   70      1 chevrolet chevelle malibu
2   15         8          350        165   3693         11.5   70      1         buick skylark 320
3   18         8          318        150   3436         11.0   70      1        plymouth satellite
4   16         8          304        150   3433         12.0   70      1             amc rebel sst
5   17         8          302        140   3449         10.5   70      1               ford torino
6   15         8          429        198   4341         10.0   70      1          ford galaxie 500
7   14         8          454        220   4354          9.0   70      1          chevrolet impala
8   14         8          440        215   4312          8.5   70      1         plymouth fury iii
9   14         8          455        225   4425         10.0   70      1          pontiac catalina
10  15         8          390        190   3850          8.5   70      1        amc ambassador dpl
```

(a) dim(Auto)

1. There are 392 training observations and 8 features (including the 'name' feature). m= 392, n=8
2. Considering $X \in R^{m \times n}$: $X \in R^{392 \times 8}$
   X is a skinny/tall matrix as it has a lot more observations than features (m>>n)

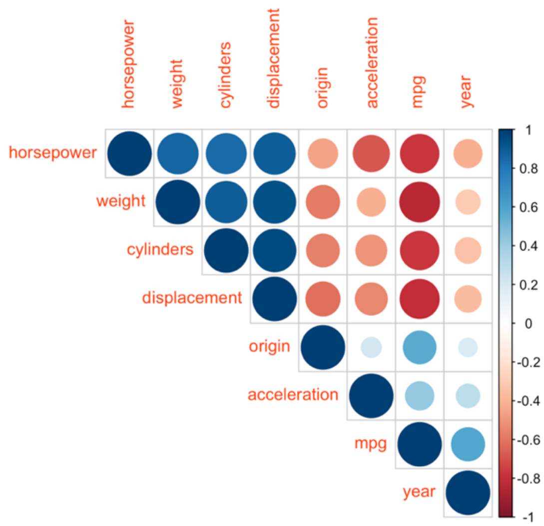
(b) Basic exploratory data analysis by performing correlations:

1. Data without 'name' variable:
   num_data <- Auto %>% select(-name)
   library(corrplot)
   Calculation correlation of each variable with another in the numeric dataset
   correlation <- cor(num_data)
   correlation

```
> correlation
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

Plotting correlations:

corrplot(correlation, method="circle", type="upper", order = "hclust")



2. Interpretation of highly correlated features:
   Cylinders and displacement have the maximum absolute correlation (0.95), followed by weight and displacement (0.93), followed by horsepower and displacement (0.89); i.e. As the number of cylinders increase, the displacement also increases significantly, As the displacement increases, weight increases etc.
   Miles per gallon (mpg) has an inverse strong correlation with weight, displacement, horsepower and cylinders (-0.83, -0.80, -0.77, -0.77 respectively).
   This indicates that the features are not independent of one another. (The correlation is the proof that there is a significant change in one variable with respect to another)

(c) Linear Regression:

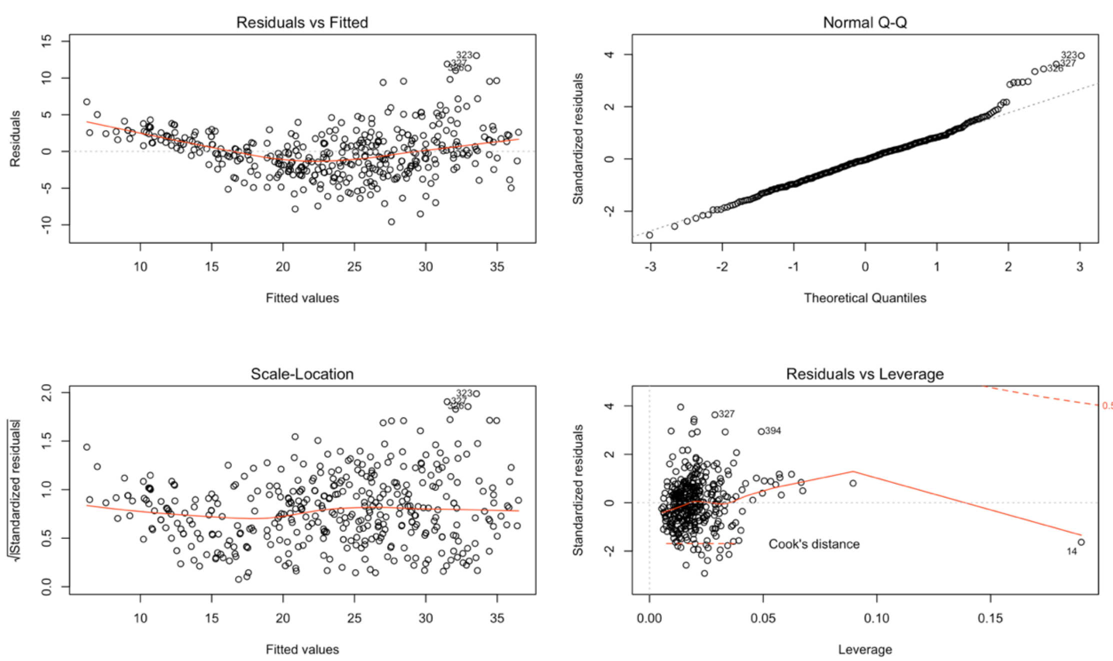
mod <- lm(mpg~., data=num_data)
summary(mod)

```
> summary(mod)

Call:
lm(formula = mpg ~ ., data = num_data)

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
```

1. Yes, there exists a relationship between the output response and the input features. The adjusted R square of the model is 0.8182, indicating 81.82% of the variance in the mpg variable can be explained by the given variables.
2. The p-value of respective variables indicate the relationship with mpg. If p-value < 0.05: we reject the null hypothesis (there exists no relationship between mpg and the respective variable) which means that there exists a relationship between the variables. Therefore, weight and year are highly related to mpg, followed by origin, followed by displacement. On the other hand: acceleration, horsepower and cylinders are not related to mpg.
3. Feature 'year' is related to the output response 'mpg' as the p-value < 0.05, indicating the rejection of null hypothesis. The coefficient of the year model is 0.75 which indicates that a unit increase in year variable, increases mpg by 0.75 given this model's fit.

(d) Diagnostic plots:

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



Inferring plot to identify problems and presence of outliers:

- The residual v/s fitted plot shows that there exists a slight pattern such that when the fitted values are large, the residuals are more scattered and when the fitted values are small, the residuals are concentrated and very less. This indicates an inconsistent variance in the dataset.

- When normal Q-Q plot has a perfect linear relationship, it indicates normal distribution of residuals. The plot here shows the deviation of standardized residuals from the linear relationship in higher quantiles, indicating the presence of outliers. The outliers in the dataset are observations: 326, 323, 327.
- Observation 14 is an influential point in the dataset indicated from standardized residuals v/s leverage graph. Another outlier: observation 394 can be spotted here.

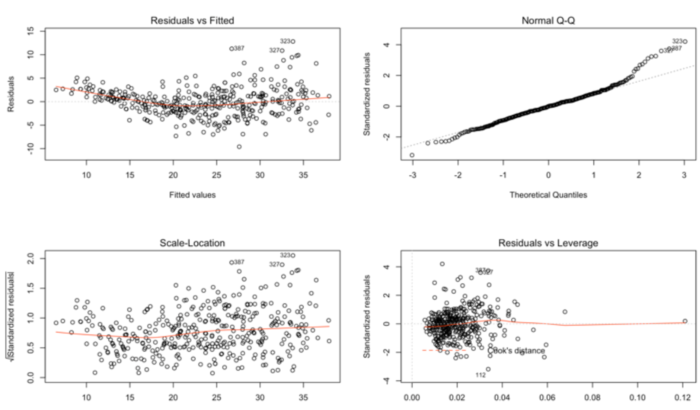
(e)

Transforming independent variables to log (x_j):

```
> summary(mod_loge)

Call:
lm(formula = num_data$mpg ~ ., data = logedata)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5987 -1.8172 -0.0181  1.5906 12.8132 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -66.5643    17.5053  -3.803 0.000167 ***
cylinders      1.4818     1.6589   0.893 0.372273    
displacement  -1.0551     1.5385  -0.686 0.493230    
horsepower    -6.9657     1.5569  -4.474 1.01e-05 ***
weight       -12.5728     2.2251  -5.650 3.12e-08 ***
acceleration  -4.9831     1.6078  -3.099 0.002082 ** 
year          54.9857     3.5555  15.465  < 2e-16 ***
origin         1.5822     0.5083   3.113 0.001991 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.069 on 384 degrees of freedom
Multiple R-squared:  0.8482,	Adjusted R-squared:  0.8454 
F-statistic: 306.5 on 7 and 384 DF,  p-value: < 2.2e-16
```

The adjusted R square increases to 84.54% from 81.52%

Residual plots of log transformation model:



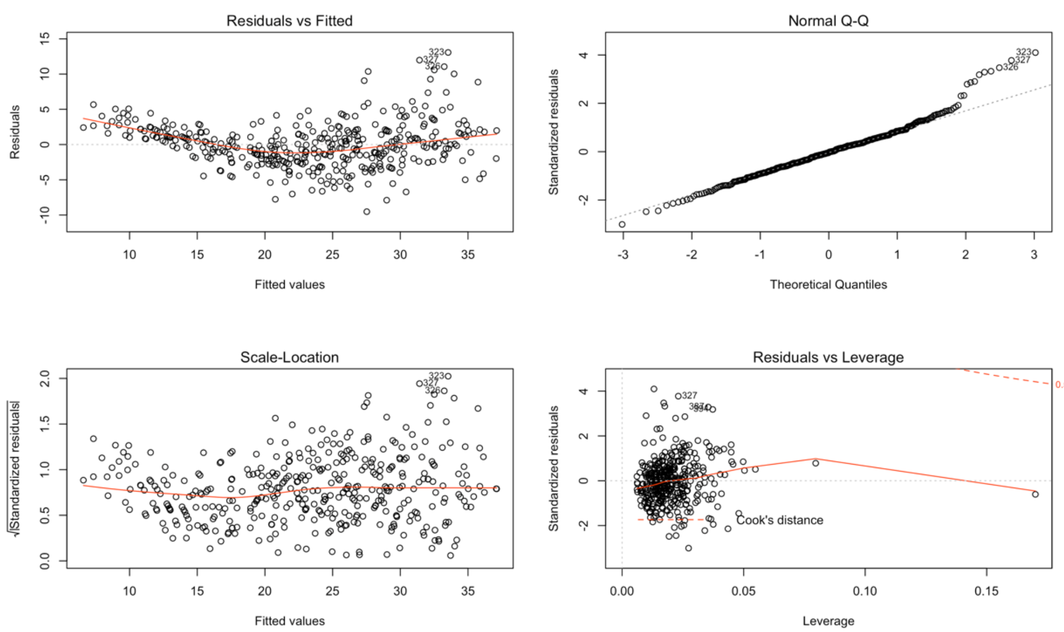
A better fit was established with log transformation of x_j (as the adjusted R square increased). This was because the influential point was normalized (see the Residual v/s Leverage plot), a slightly uniform variance can be seen from residual v/s fitted plot. The residuals are more normalized than before, hence the result.

Transforming independent variables to $(x_j)^{-1/2}$:

```
> summary(mod_sqrt)

Call:
lm(formula = num_data$mpg ~ ., data = sqrtdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5250 -1.9822 -0.1111  1.7347 13.0681 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -49.79814    9.17832  -5.426 1.02e-07 ***
cylinders    -0.23699    1.53753  -0.154   0.8776    
displacement  0.22580    0.22940   0.984   0.3256    
horsepower   -0.77976    0.30788  -2.533   0.0117 *  
weight       -0.62172    0.07898  -7.872 3.59e-14 ***
acceleration -0.82529    0.83443  -0.989   0.3233    
year         12.79030    0.85891  14.891  < 2e-16 ***
origin        3.26036    0.76767   4.247 2.72e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.21 on 384 degrees of freedom
Multiple R-squared:  0.8338,	Adjusted R-squared:  0.8308 
F-statistic: 275.3 on 7 and 384 DF,  p-value: < 2.2e-16
```

The adjusted R square increases to 83.08% from 81.52%

Residual plots of $(x_j)^{-1/2}$ transformation model:



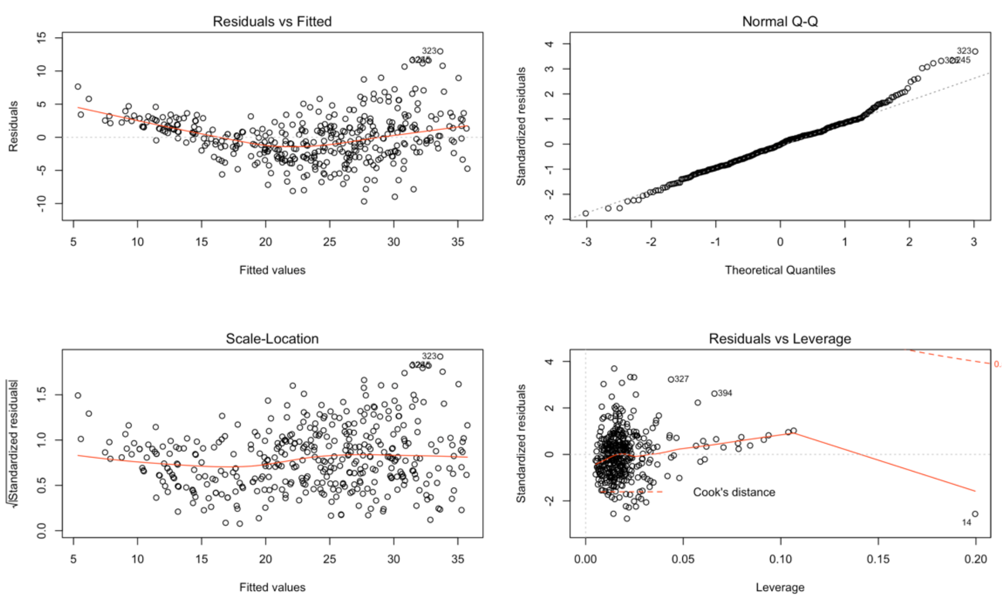
This case is similar to the previous case where the influential point gets removed and the variance of residuals becomes slightly more uniform. Hence, the adjusted R square increases, indicating a better fit.

Transforming independent variables to $(x_j)^2$:

```
> summary(mod_sqr)

Call:
lm(formula = num_data$mpg ~ ., data = sqrdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6786 -2.3227 -0.0582  1.9073 12.9807 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.208e+00  2.356e+00   0.513 0.608382    
cylinders    -8.829e-02  2.521e-02  -3.502 0.000515 ***
displacement  5.680e-05  1.382e-05   4.109 4.87e-05 ***
horsepower   -3.621e-05  4.975e-05  -0.728 0.467201    
weight       -9.351e-07  8.978e-08 -10.416  < 2e-16 ***
acceleration  6.278e-03  2.690e-03   2.334 0.020130 *  
year          4.999e-03  3.530e-04  14.160  < 2e-16 ***
origin        4.129e-01  6.914e-02   5.971 5.37e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.539 on 384 degrees of freedom
Multiple R-squared:  0.7981,	Adjusted R-squared:  0.7944 
F-statistic: 216.8 on 7 and 384 DF,  p-value: < 2.2e-16
```

The adjusted R square decreases to 79.44 % from 81.52%

Residual plots of $(x_j)^2$ transformation model:



This plot is worse than the residual plot of x_j since the values were squared, the residuals increased with higher fitted values and decreased with smaller fitted values, making the pattern in residual v/s fitted value more prominent.