Experiment 8:

Write R program to implement linear and multiple regression on ‘mtcars’ dataset to estimate the value of ‘mpg’ variable, with best R^2 and plot the original values in ‘green’ and predicted values in ‘red’.

Solution:

The built-in **mtcars** data frame contains information including their weight, fuel efficiency (in miles-per-gallon), speed, etc. of 32 models of automobile from 1973-74 as reported in Motor Trend Magazine. In analyzing the dataset of different collection of cars, we will explore the relationship between a set of eleven variables, and miles per gallon (MPG). Dataset Motor Trend has been used to find out that,

- Is an automatic or manual transmission better for miles per gallon?
- How different is the MPG between automatic and manual transmission?

Read MTCARS dataframe

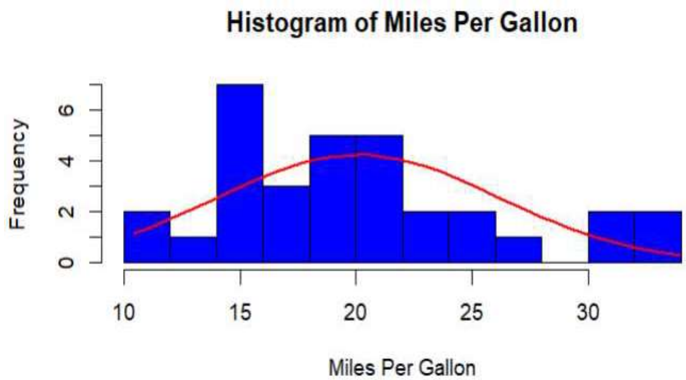
> data (“mtcars”)

> head (mtcars)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

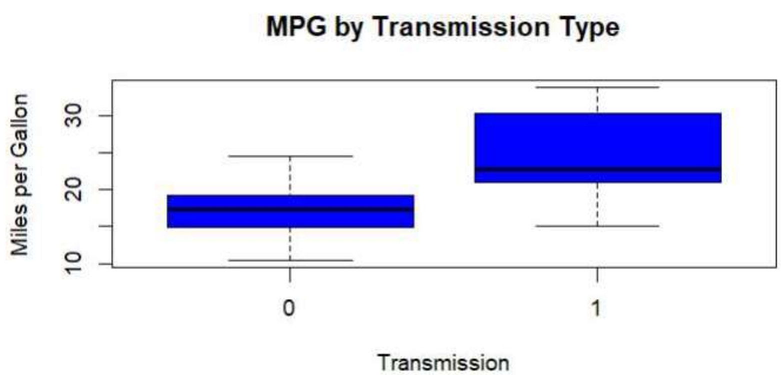
We have to explore the relationship between a set of variables and miles per gallon (mpg), so mpg is our ***dependent variable***. Plot dependent variable to check its distribution.

```
x <- mtcars$mpg
h<- hist (x, breaks=10, col="blue", xlab="Miles Per Gallon", main="Histogram of Miles Per Gallon")
xfit <- seq (min(x),max(x),length=40)
yfit <- dnorm (xfit, mean=mean(x), sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines (xfit, yfit, type="l", col="red", lwd=2)
```



The distribution of mpg, showing in the graphs is approximately normal and does not contain any outliers. Now we check how mpg can be changed by automatic or manual transmission, by plotting a box plot. From this boxplot, it seems that automatic cars have a lower miles per gallon, and so a lower fuel potency, than manual cars do.

```
> boxplot(mpg~am, data = mtcars, col = c("blue", "blue"), xlab = "Transmission",
      ylab = "Miles per Gallon", main = "MPG by Transmission Type")
```



Correlation analysis

A correlation test is performed to determine the relationship between the variables, and to find out which variables should be included in our model to answer the questions. The correlation matrix is

```
> cor (mtcars, use="complete.obs", method="pearson")
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.0000000	-0.8521620	-0.8475514	-0.7761684	0.68117191	-0.8676594	0.41868403	0.6640389	0.59983243	0.4802848	-0.55092507
cyl	-0.8521620	1.0000000	0.9020329	0.8324475	-0.69993811	0.7824958	-0.59124207	-0.8108118	-0.52260705	-0.4926866	0.52698829
disp	-0.8475514	0.9020329	1.0000000	0.7909486	-0.71021393	0.8879799	-0.43369788	-0.7104159	-0.59122704	-0.5555692	0.39497686
hp	-0.7761684	0.8324475	0.7909486	1.0000000	-0.44875912	0.6587479	-0.70822339	-0.7230967	-0.24320426	-0.1257043	0.74981247
drat	0.6811719	-0.6999381	-0.7102139	-0.4487591	1.00000000	-0.7124406	0.09120476	0.4402785	0.71271113	0.6996101	-0.09078980
wt	-0.8676594	0.7824958	0.8879799	0.6587479	-0.71244065	1.0000000	-0.17471588	-0.5549157	-0.69249526	-0.5832870	0.42760594
qsec	0.4186840	-0.5912421	-0.4336979	-0.7082234	0.09120476	-0.1747159	1.00000000	0.7445354	-0.22986086	-0.2126822	-0.65624923
vs	0.6640389	-0.8108118	-0.7104159	-0.7230967	0.44027846	-0.5549157	0.74453544	1.0000000	0.16834512	0.2060233	-0.56960714
am	0.5998324	-0.5226070	-0.5912270	-0.2432043	0.71271113	-0.6924953	-0.22986086	0.1683451	1.00000000	0.7940588	0.05753435
gear	0.4802848	-0.4926866	-0.5555692	-0.1257043	0.69961013	-0.5832870	-0.21268223	0.2060233	0.79405876	1.0000000	0.27407284
carb	-0.5509251	0.5269883	0.3949769	0.7498125	-0.09078980	0.4276059	-0.65624923	-0.5696071	0.05753435	0.2740728	1.00000000

The values in correlation matrix shows that variables such as ***wt, cyl, disp,*** and ***hp*** are highly correlated with the dependent variable ***mpg***. Hence, they should be included in the regression model. From the correlation matrix, it can be also be observed that ***cyl*** and ***disp*** are highly correlated with each other. In order to avoid the problem of collinearity only one variable from these two will be included in the model.

Simple Regression model

```
> fit <-lm(mpg~am, data=mtcars)
> summary(fit)
```

```
Call:
lm(formula = mpg ~ am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.3923 -3.0923 -0.2974  3.2439  9.5077 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   17.147      1.125  15.247 1.13e-15 ***
am             7.245      1.764   4.106 0.000285 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom
Multiple R-squared:  0.3598,	Adjusted R-squared:  0.3385 
F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

From the above summary, there exists a linear relation between the predictor variable MPG and AM. Intercepts and Coefficient can be explained that, on average automatic transmission cars has 17.147 MPG and manual transmission cars has 24.39(17.147 + 7.24).The value of R^2 is 0.3385, which means this model only explain 33.85% of the variance.

Multiple Regression model

In the correlation analysis, it is observed that variables such as ***wt, cyl, am,*** and ***hp*** are highly correlated with the dependent variable ***mpg***. So, we apply a multi variant regression for ***mpg*** on ***am, wt, cyl,*** and ***hp***.

```
> mfit <- lm (mpg ~ am + cyl + wt + hp, data = mtcars)
> anova (fit, mfit)
```

```
Analysis of Variance Table

Model 1: mpg ~ am
Model 2: mpg ~ am + cyl + wt + hp
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     30 720.9
2     27 170.0  3     550.9 29.166 1.274e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value f 1.274e-08, it is clear that the multivariate model of regression is different from that of above simple model.

> summary (mfit)

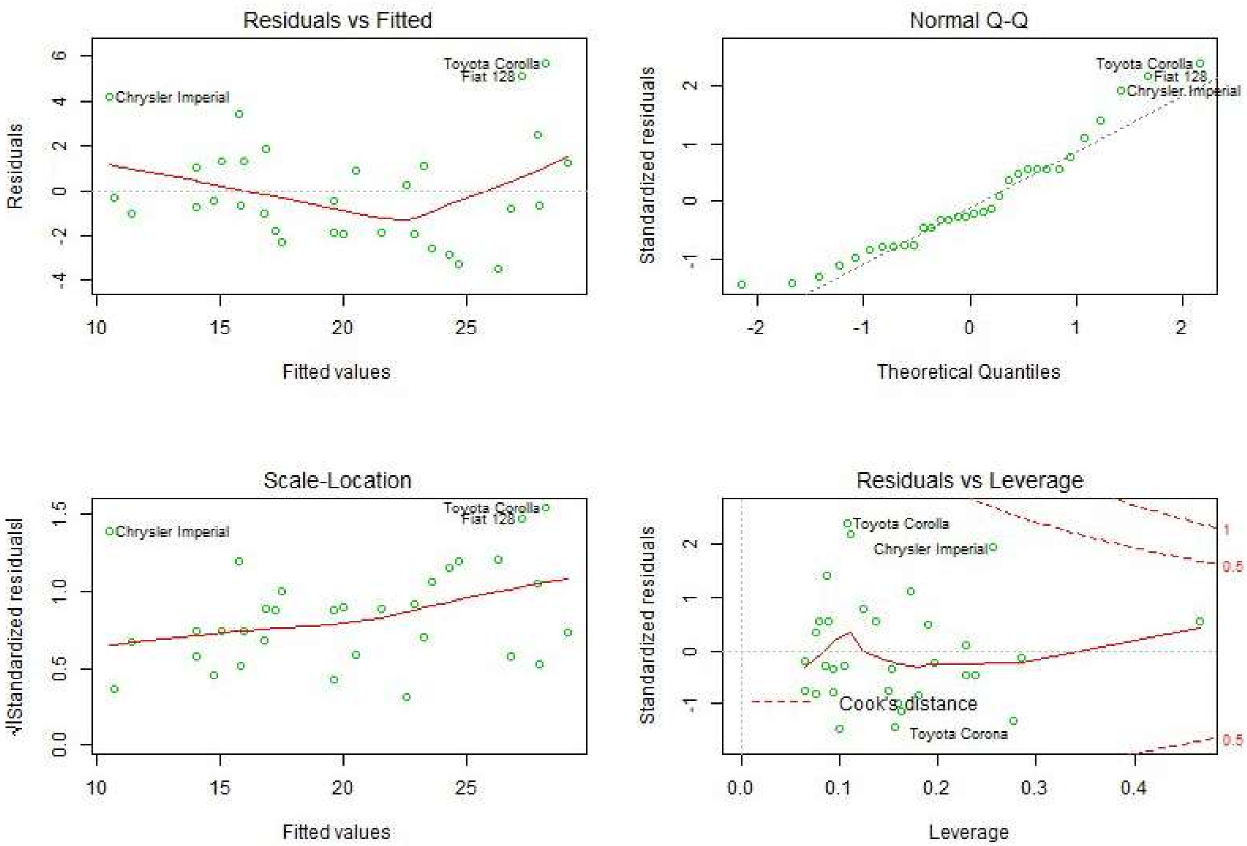
```
Call:
lm(formula = mpg ~ am + cyl + wt + hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4765 -1.8471 -0.5544  1.2758  5.6608

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
am           1.47805    1.44115   1.026   0.3142
cyl         -0.74516    0.58279  -1.279   0.2119
wt          -2.60648    0.91984  -2.834   0.0086 **
hp          -0.02495    0.01365  -1.828   0.0786 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.509 on 27 degrees of freedom
Multiple R-squared:  0.849,     Adjusted R-squared:  0.8267
F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Multivariate regression model explain 84.9% variance. It can be seen that ***wt*** and up to some extent ***hp*** confound the relationship between ***am*** and ***mpg***.

Result plots

> par (mfrow = c(2,2))

> plot (mfit, col=3)


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Toyota Corona
Leverage