1. Research question

- Analyze and understand the dataset of different collection of cars and explore the relationship between different variable from a set of eleven variables.
- Estimation and comparison between the overall regression model and Stepwise Selection procedures.
- Check all the underlining assumptions for the best fit model.
- Exploratory data analysis for each of the variable.

2. Hypothesis

- Testing the significance of Individual Parameters in the model.
- Testing the significance of Overall Regression of the model.
- The model is a good fit for the given data.
- The explanatory variables are independent.

3. Datasets

We are using "**mtcars"** dataset from R for the purpose of analysis.

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

4. This data set consists of 32 observations on the following 11 variables:

Mpg	Miles/(US) gallon
Cyl	Number of cylinders
Disp	Displacement (cu.in.)
Hp	Gross horsepower
Drat	Rear axle ratio
Wt	Weight (lb/1000)
Qsec	1/4 mile time
Vs	V/S
Am	Transmission (0 = automatic, 1 = manual)
Gear	Number of forward gears
Carb	Number of carburetors

Where,
Mpg is the dependent variable and cyl,disp,…,carb are independent variables.

4. Simple Model Building

a) Fitting a Linear Regression Model

In general the PRF can be any function but for simplicity we restrict ourselves to the class of functions where Y and $X_1, X_2, \ldots, X_p$ are related through a linear function of some unknown parameters which leads to **linear regression analysis**. Let f takes the following form,

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Above equation specifies what we call as multiple linear regression model (MLRM) where $\beta_0, \beta_1, \ldots, \beta_p$ are termed as regression coefficients. We are interested in estimating the PRF

which is equivalent to estimate the unknown parameters $\beta_0, \beta_1, \dots, \beta_p$ on the basis of a random sample from *Y* and given values of the independent variables.

Here, we take in our study, "mpg" as dependent variable an rest all other variables viz. "cyl", "disp", "hp", etc. as independent variables $X_1, X_2, \dots, X_p$.

"mpg" = β_0 + β_1"cyl" + ... + β_p"carb" + ε

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	12.303	18.718	**0.657**	**0.518**
Cyl	-0.111	1.045	**-0.107**	**0.916**
Disp	0.013	0.018	**0.747**	**0.463**
Hp	-0.021	0.022	**-0.987**	**0.335**
Drat	0.787	1.635	**0.481**	**0.635**
Wt	-3.715	1.894	**-1.961**	**0.063**
Qsec	0.821	0.731	**1.123**	**0.274**
Vs	0.318	2.105	**0.151**	**0.881**
Am	2.520	2.057	**1.225**	**0.234**
Gear	0.655	1.493	**0.439**	**0.665**
Carb	-0.199	0.829	**-0.241**	**0.812**

In the above table regression coefficients $\beta_0, \beta_1, \dots, \beta_p$ are represented by column "coefficients". This means, with a unit change in the explanatory variable, there is a coefficient times change in the dependent variable. Ex. With 1 unit change in the velocity (vs) there is a 0.318 unit change in the dependent variable (mpg).

After estimating the unknown parameters involved in a multiple linear regression model, the next task of interest is to test for their statistical significance.

b) Testing the significance of Individual Parameters

Under normality assumption for the error terms, significance of individual parameters can be tested using a t-test. The procedure is as follows:

Step 1: Null Hypothesis is set as $H_0 : \beta_i = 0$ and Alternative Hypothesis is set as $H_1 : \beta_i \neq 0$.

Step 2: Calculate the t-statistic as follows: $$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{ii}}}$$

where $(X^T X)^{-1}_{ii}$ is the ith element of $(X^T X)^{-1}$.

Step 3: Reject H_0 if p value of calculated t statistic is less than $\alpha = 0.05$ and conclude that the β_i is statistically significant at 5% l.o.s. otherwise accept H_0.

For the model under study, we have- refer the above table.

Conclusion

From the above table, since p-value of all the parameters are greater than 0.05, hence we may conclude that all the individual parameters are insignificant at 5% l.o.s. Only "wt" is significant at 10% l.o.s.

c) Testing the significance of Overall Regression

Under normality assumption for the error terms, significance overall regression can be tested using an F-test. The procedure is as follows:

Step 1: Null hypotheses is H0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ and Alternative hypothesis is H1: $\beta_i \neq 0$ for at least one i = 1, 2, ... , p.

Step 2: Calculate the F-statistic as follows:

$$F = \frac{SSReg/p}{SSRes/(n-p-1)} = \frac{MSReg}{MSRes}$$

where MS Reg and MS Res are the mean sum of squares due to regression and residuals.

Step 3: Reject H_0 if p value of calculated F statistic is less than $\alpha = 0.05$ i.e. the overall regression is statistically significant.
For the model under study, we have

ANOVA					
	Df	***SS***	***MS***	***F***	***Significance F***
Regression	10	978.553	97.855	13.933	0.000
Residual	21	147.494	7.024		
Total	31	1126.047			

Conclusion

From the above table, since p-value is less than 0.05, hence we may conclude that the overall regression is significant at 5% l.o.s.

d) R Square and Adjusted R Square

R-Square, also termed as coefficient of determination is a measure of goodness of model. It is defined as follows:

$$R^2 = \frac{SSReg}{SST} = 1 - \frac{SSRes}{SST}$$

R^2 represents the proportion of variability explained by the model. Clearly $0 \leq R^2 \leq 1$. An adequate model is expected to have high R^2.

It can be proved that R^2 is an increasing function of the number of independent variables included in the model and hence it doesn't give true insight about goodness of the model. A refined measure of goodness which is free from this drawback, termed as adjusted R^2 is defined as follows.

$$R^2_{Adj} = 1 - \frac{SSRes/(n-p-1)}{SST/(p-1)}$$

It can be observed that $-\infty \leq R^2_{adj} \leq R^2 \leq 1$.

For the model under study, we have

Regression Statistics	
Multiple R	0.932
R Square	**0.869**
Adjusted R Square	**0.807**
Standard Error	2.650
Observations	32

Conclusion

From the above table, since R^2_{adj} is quite high, hence we may conclude that the selected model is a good fit for the given data.

From the above hypothesis testing we see a high value of overall R^2 but insignificant t-ratios indicating that Multicollinearity may be present among the independent variables.

4.1 Multicollinearity

a) Problem and its Consequences

The existence of near linear relationship among the explanatory variables is termed as **Multicollinearity**. In other words multicollinearity is a situation when one or more explanatory variables can be well expressed as a near linear combination of the other explanatory variables. Multicollinearity can arise due to several reasons like use of too many regressors, faulty data collection etc. In has been seen that presence of multicollinearity seriously weakens the results based on ordinary least squared technique. Following are the common ill consequences of this problem:

1. Variances of the ordinary least squares estimates of the regression coefficients are inflated.
2. Absolute values of the regression coefficients are very high.
3. Regressors which are expected to be important turn out to be insignificant.
4. Regressors have wrong sign as against the a priori belief.

b) Detection and Removal of Multicollinearity

Variance Inflation Factor (VIF) Approach: Define $R^2_{(j)} = R^2$ of the multiple linear regression model where X_j is taken as dependent variable and all other regressors $X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p$. Clearly $R^2_{(j)}$ is expected to be high if there is a strong linear relationship between X_j and the rest of the regressors.

It can be shown that $$V(\hat{\beta}_j) = \frac{\sigma^2}{1 - R^2_{(j)}}$$

Variance Inflation Factor (VIF) of the j^{th} regressor defined as follows.

$$VIF_j = \frac{1}{1 - R^2_{(j)}}$$

Clearly $R^2_{(j)}$ and VIF_j are positively related and hence VIF_j is expected be high if j^{th} is involved in

multicollinearity.

Detection: As a rule-of-thumb if $VIF_j > 10$ then X_j can be taken to have strong linear relationship with the other regressors.

Removal: Deleting the corresponding X_j's will solve the problem.
For the model under study, we have

Model	*VIF*
Cyl	**15.374**
Disp	**21.62**
Hp	9.832
Drat	3.375
Wt	**15.165**
Qsec	7.528
Vs	4.966
Am	4.648
Gear	5.357
Carb	7.909

Conclusion

We see that the highlighted variables are significant (i.e., $VIF_i > 10$). Thus, to counter the problem of multicollinearity we remove "cyl", "disp" and "wt".
For the reduced model, we get,

Regression Statistics	
Multiple R	0.914
R Square	0.836
Adjusted R Square	0.788
Standard Error	2.775
Observations	32

From the above approach used for the removal of multicollinearity, we find that Variance Inflation Factor (VIF) approach is a good fit to the given data.

5. Analysis of single data set.

For visualization, we are now categorizing the data set into continuous and categorical variables.
In the given data set we have the following continuous variables:

1. mpg – Miles/(US) gallon
2. disp – Displacement
3. hp – Gross Horsepower
4. drat – Rear axle ratio
5. wt – Weight (lb/1000)
6. qsec – ¼ mile time.

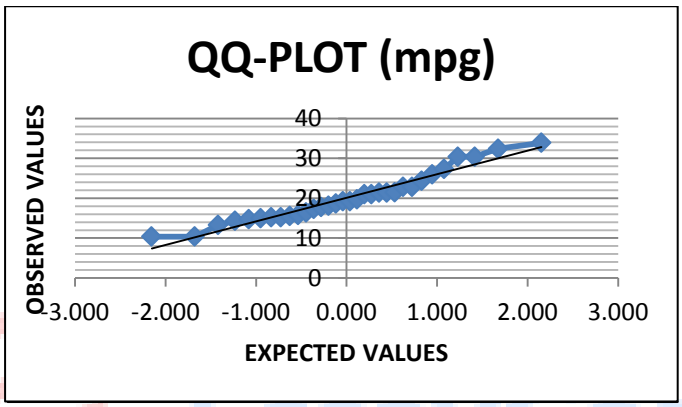
We are using the following measures/ tools for visualization:

1. Q-Q PLOTS:
   To test whether the data is normally distributed or not with the hypothesis:
   - H_o: Sample comes from a normal population.
   - H_1: Sample does not come from the normal population.

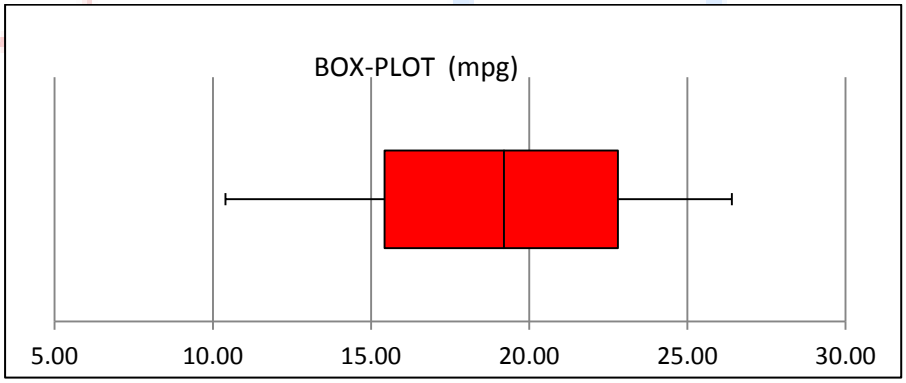
2. BOX-PLOTS:
   - To determine if there are any outliers in the data.

3. DESCRIPTIVE STATISTICS:
   - Like mean, median, mode, etc.

1) **MPG :- Miles/(US) gallon:**



Statistical moments	Values
μ1	20.09
μ2	36.32
μ3	-498495
μ4	15262466

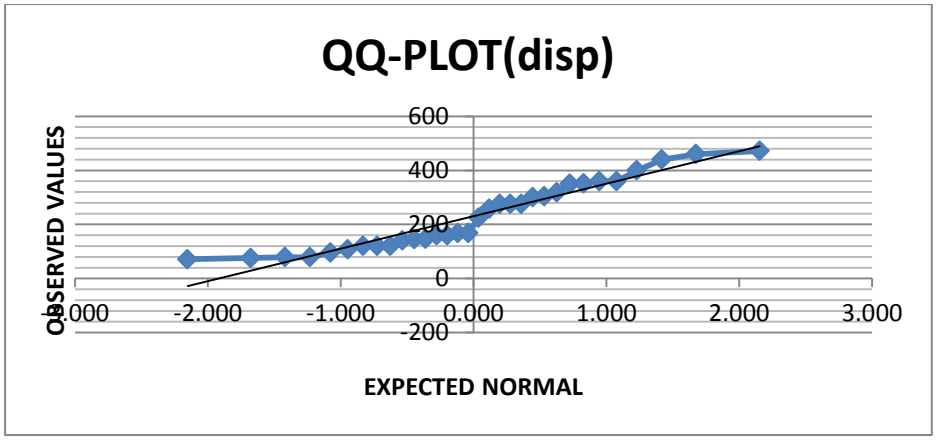


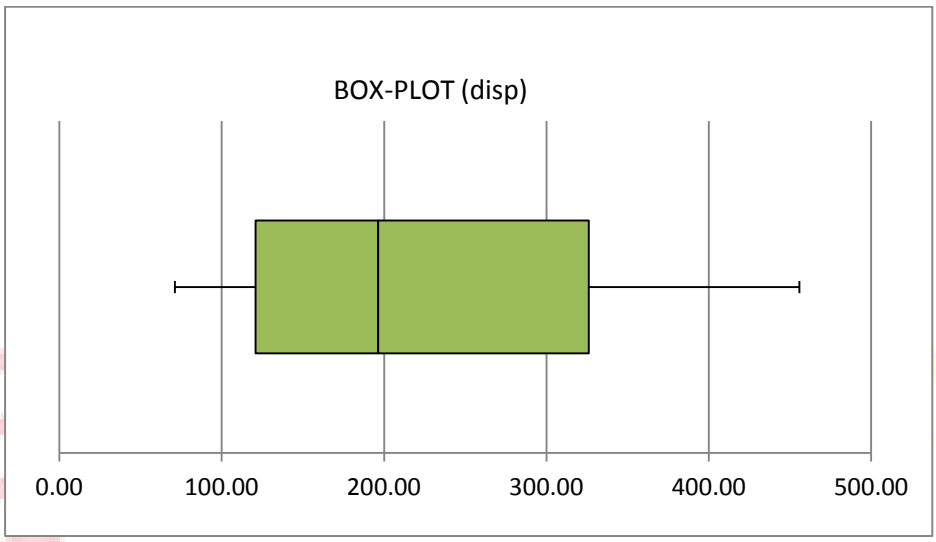
From the normal QQ-Plots we may infer that "mpg" is almost normally distributed.

From Box-Plots we may infer that "mpg" does not contain any outlier.

2) DISP-Displacement:

Statistical moments	Values
μ1	230.7219
μ2	15360.8
μ3	-7.4E+08
μ4	2.77E+11

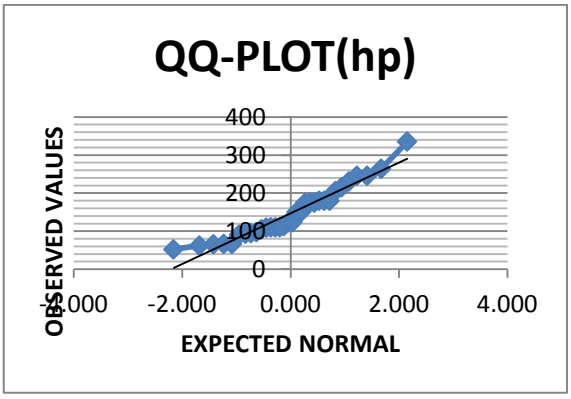




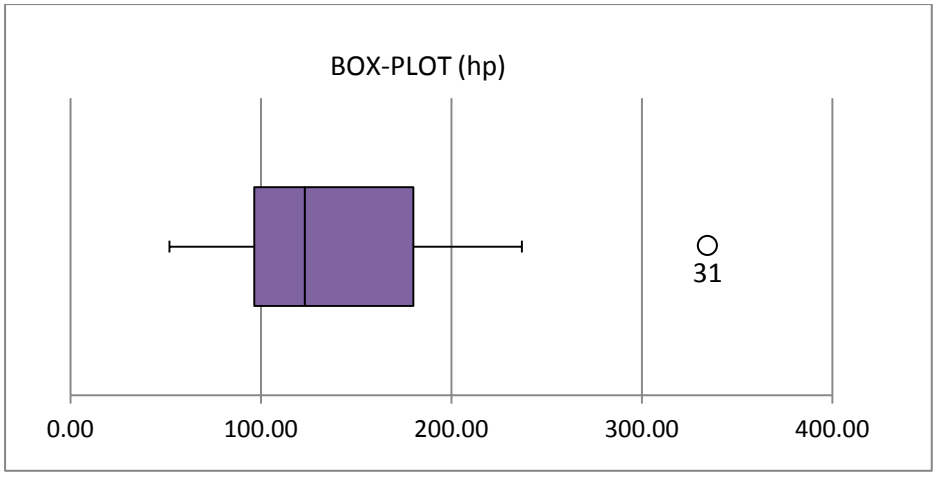
From the QQ-Plot we may infer that “disp” is not normally distributed.

From the box plots we may infer that “disp” does not contain any outlier.

3) <u>HP-GROSS HORSE POWER:</u>



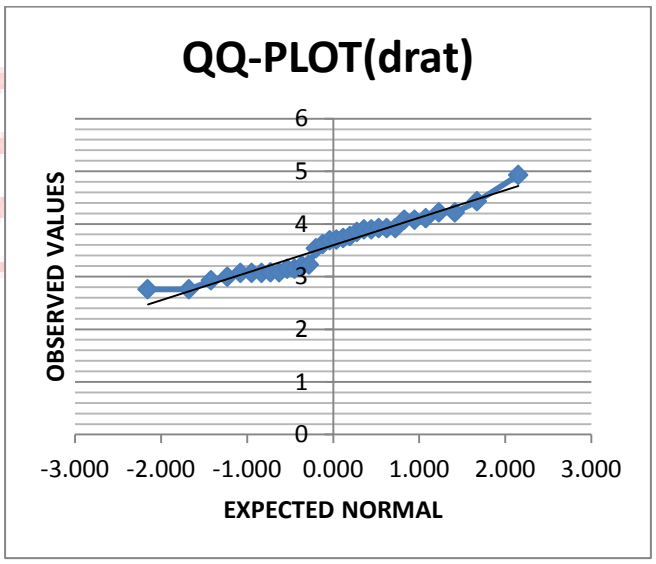
Statistical moments	Values
μ1	146.6875
μ2	4700.867
μ3	-1.9E+08
μ4	4.51E+10



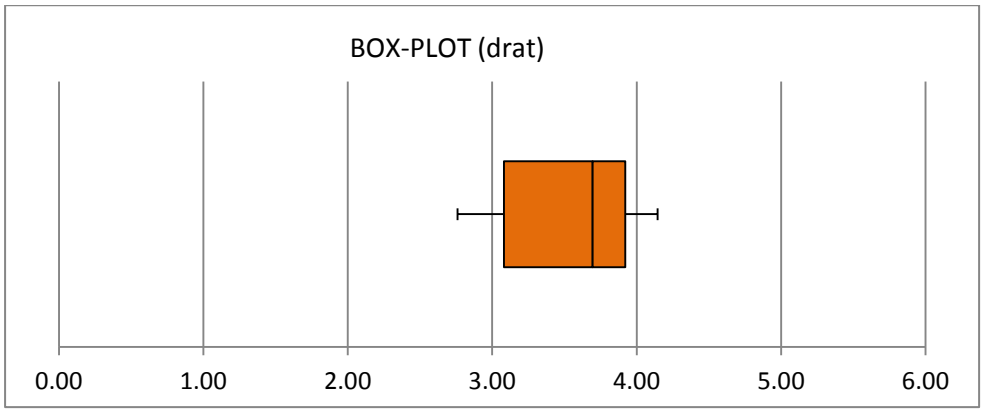
From the above QQ-Plots we may infer that “hp” is not normally distributed.

From the box plots we may infer that “hp” contains an outlier as 31st observation.

4) <u>DRAT-Rear axle ratio:</u>



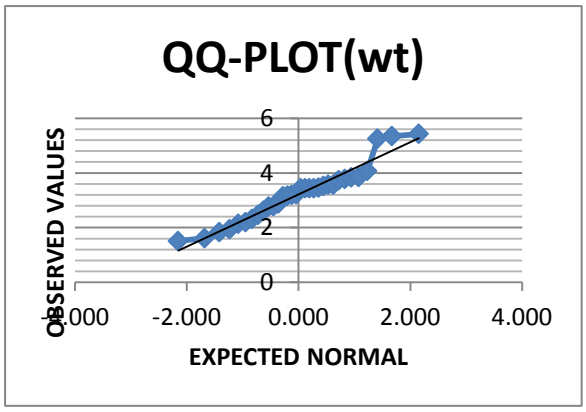
Statistical moments	Values
μ1	3.596563
μ2	0.285881
μ3	-2883.09
μ4	15566.83



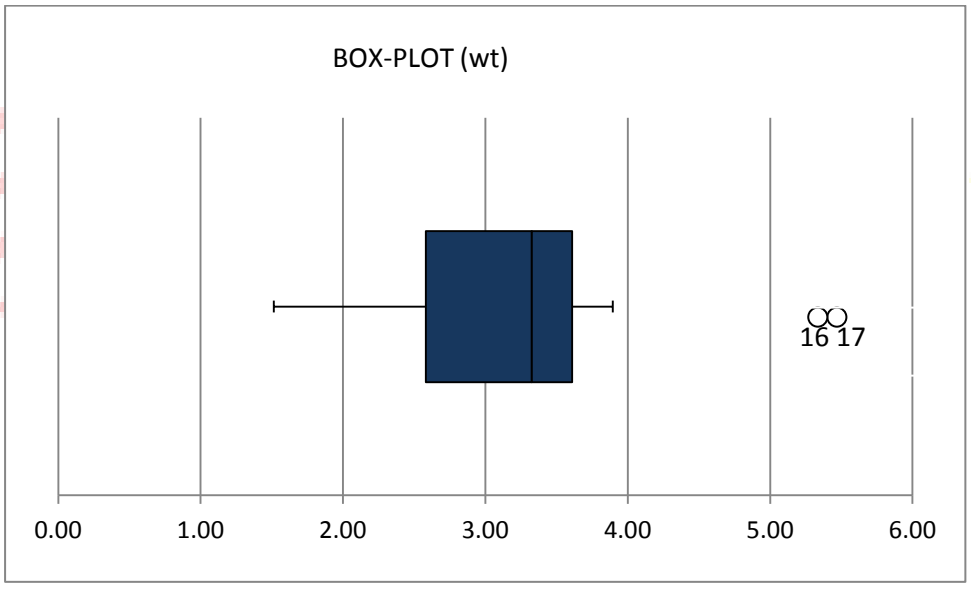
From the above QQ-Plot we may infer that “drat” may be almost normally distributed.

From the box plot we may infer that “drat” does not contain any outlier.

5) WT-weight(lb/1000)



Statistical moments	values
µ1	3.21725
µ2	0.957379
µ3	-2051.96
µ4	10051.06

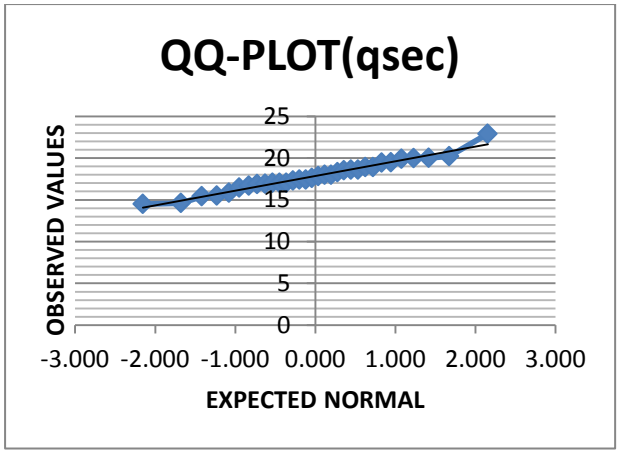


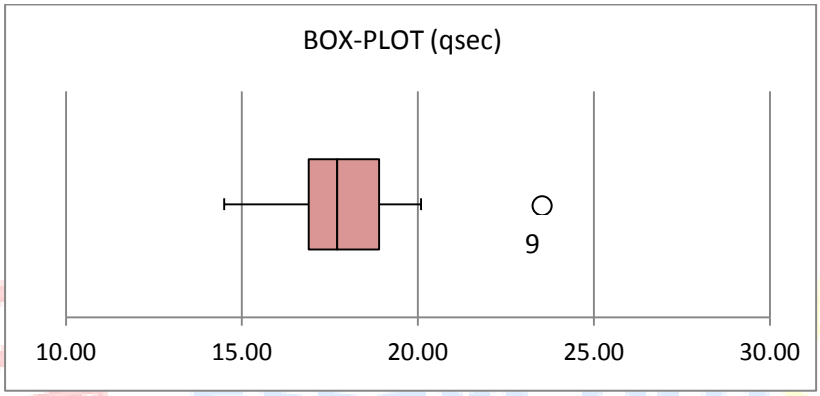
From the QQ-Plot we may infer that “wt” is almost normally distributed.

From box plot we may infer that “wt” has 2 outliers which are 16th and 17th observations.

6) QSEC-1/4 mile time

Statistical moments	Values
µ1	17.84875
µ2	3.193166
µ3	-352478
µ4	9439831





From the QQ-Plot we may infer that "qsec" is almost normally distributed.

From the box plot we may infer that "qsec" contains an outlier as 9th observation.

2. In the given data set we have the following discrete variables:

1. cyl – Number of cylinders
2. vs – V/S
3. am – Transmission (0=automatic, 1=manual)
4. gear – Number of forward gears
5. carb – Number of carburetors

We will use the following measures/tools:

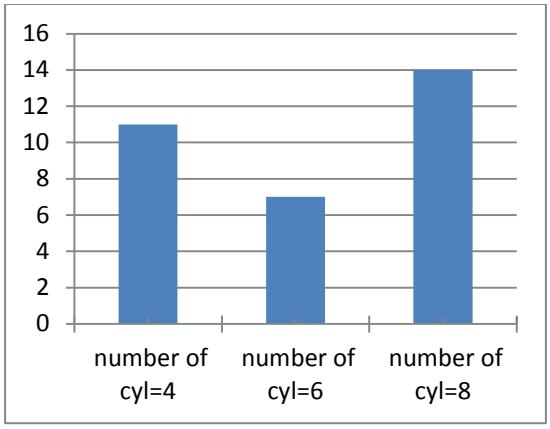
1. Frequency Table
   - To get the frequency of each data point.
2. Bar Plot
   - To represent the frequency distribution of data.

1) CYL-no of cylinders

Frequency table:

valid	Frequency	Cumulative percent
4	11	34.4
6	7	56.2
8	14	100
total	32	

Bar Graph:



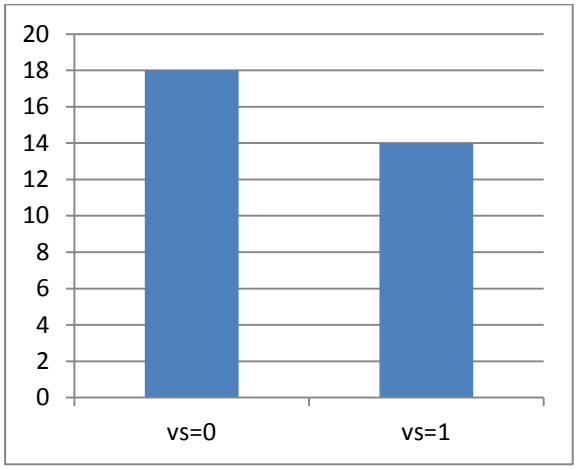
statistical moments	
μ1	6.187
μ2	3.189
μ3	-147
μ4	136

2) VS-V/S

Frequency Table:

Valid	Frequency	Cumulative Frequency
0	18	56.2
1	14	100
Total	32	

Bar graph:



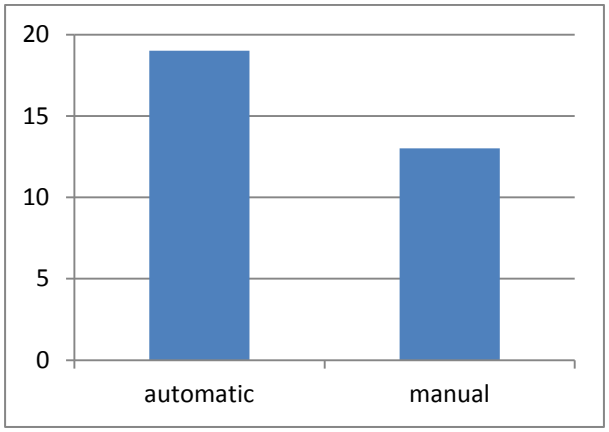
statistical moments	
μ1	0.4375
μ2	0.254032
μ3	-4.20752
μ4	5.468216

3) AM-TRANSMISSION (0=automatic, 1=manual)

Frequency table:

valid	Frequency	Cumulative Frequency
0	19	59.4
1	13	100
total	32	

Bar graph:



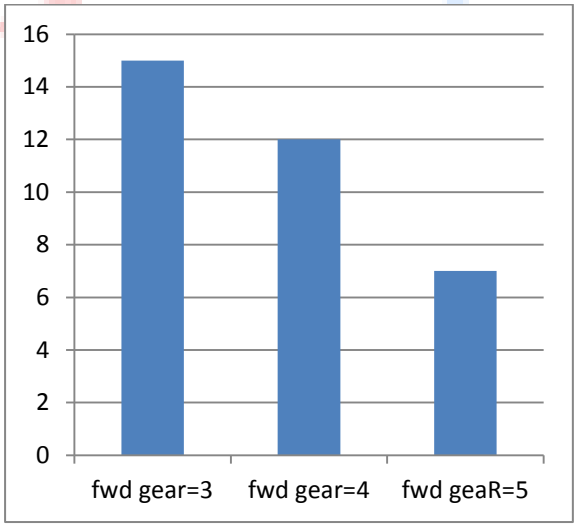
statistical moments	
μ1	0.40625
μ2	0.248992
μ3	-2.70966
μ4	4.666333

4) GEAR-Number of forward gears:

Frequency table:

Valid	Frequency	Cumulative percentage
3	15	46.9
4	12	84.4
5	7	100
Total	32	

Bar graph:



statistical moments	
μ1	3.6875
μ2	0.544355
μ3	-3101.97
μ4	17213.65

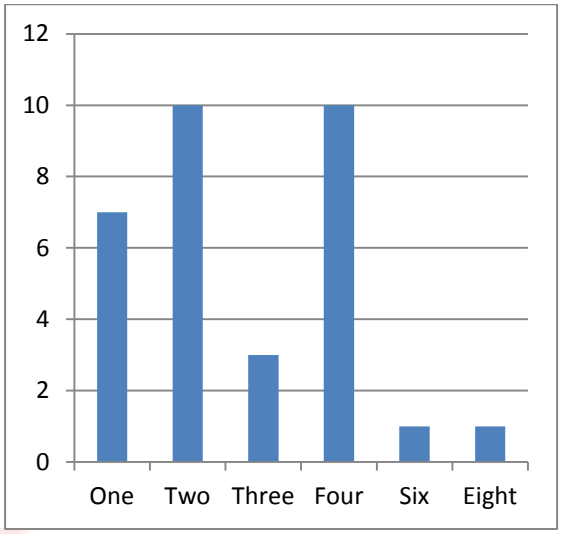
5) CARB- number of carburetors:

Frequency table:

Valid	Frequency	Cumulative percentage
One	7	21.9
Two	10	53.1

Three	3	62.5
Four	10	93.8
Six	1	96.9
Eight	1	100
Total	32	

Bar graph:



statistical moments	
μ1	2.8125
μ2	2.608871
μ3	-1237.63

6. Correlations

- Y and every x data set:

correlation of y (mpg) with :	Values
Cyl	-0.852
Disp	-0.848
Hp	-0.776
Drat	0.681
Wt	-0.868
Qsec	0.419
Vs	0.664
Am	0.599
Gear	0.480
Carb	-0.551

- All combinations of x data sets:

x on x	Cyl	disp	Hp	drat	wt	Qsec	vs	am	Gear	carb
Cyl	1	0.902	0.832	-0.700	0.782	-0.591	-0.811	-0.523	-0.493	0.527
Disp	0.902	1	0.791	-0.710	0.888	-0.434	-0.710	-0.591	-0.556	0.395
Hp	0.832	0.791	1	-0.449	0.659	-0.708	-0.723	-0.243	-0.126	0.750

Drat	-0.700	-0.710	-0.449	1	-0.712	0.091	0.440	0.713	0.700	-0.091
Wt	0.782	0.888	0.659	-0.712	1	-0.175	-0.555	-0.692	-0.583	0.428
Qsec	-0.591	-0.434	-0.708	0.091	-0.175	1	0.745	-0.230	-0.213	-0.656
Vs	-0.811	-0.710	-0.723	0.440	-0.555	0.745	1	0.168	0.206	-0.570
Am	-0.523	-0.591	-0.243	0.713	-0.692	-0.230	0.168	1	0.794	0.058
Gear	-0.493	-0.556	-0.126	0.700	-0.583	-0.213	0.206	0.794	1	0.274
Carb	0.527	0.395	0.750	-0.091	0.428	-0.656	-0.570	0.058	0.274	1

We know, Correlation coefficient measures the degree of linear relationship between two variables. Possible two variable linear relationships between regressors can be detected if some pair wise correlations are significant. Consider the matrix of all pair wise correlations.

As a rule-of-thumb if $|r_{ij}| > 0.5$ then X_i and X_j can be taken to have strong linear relationship. We see that the highlighted correlations are significant (i.e., $| r_{ij} | > 0.5$).

Here we can note that the correlation matrix also resolves the problem of multicollinearity. Thus, to counter the problem of multicollinearity we can remove “disp”, “hp”, “wt”, “cyl”, “vs” and “am”.

7. Regression Y for different selections of X

A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible. To achieve such type of model, we are here using the technique of stepwise selection. *Stepwise selection* is a method that allows moving in either direction, dropping (backward elimination) or adding(forward selection) variables at the various steps. The process is one of alternation between choosing the least significant variable to drop and then re-considering all dropped variables (except the most recently dropped) for re-introduction into the model. For the given data, we run stepwise regression, following results can be obtained:

ANOVA	(predictors:wt,cyl)				
	Df	***SS***	***MS***	***F***	***Significance F***
Regression	2	934.875	467.438	70.908	0.000
Residual	29	191.172	6.592		
Total	31	1126.047			

Regression Statistics	
Multiple R	0.911
R Square	**0.830**
Adjusted R Square	**0.819**
Standard Error	2.568
Observations	32

Conclusion

Thus the model selected by *stepwise* selection method is:

$$\text{“mpg”} = \beta_0 + \beta_1 \text{“wt”} + \beta_2 \text{“cyl”} + \varepsilon$$

From parsimonious modeling, we find that model selected by *stepwise selection* model has less predictors making the model more parsimonious. Hence, we will consider the model

i.e., $$\text{“mpg”} = \beta_0 + \beta_1\text{“wt”} + \beta_2\text{“cyl”} + \varepsilon$$

8. Validation of Assumptions and Residual Analysis

There are **four principal assumptions** which justify the use of linear regression models for purposes of inference or prediction:

a) **linearity and additivity** of the relationship between dependent and independent variables.

b) **statistical independence** of the errors (in particular, no correlation between consecutive errors in the case of time series data)

c) **homoscedasticity** (constant variance) of the errors

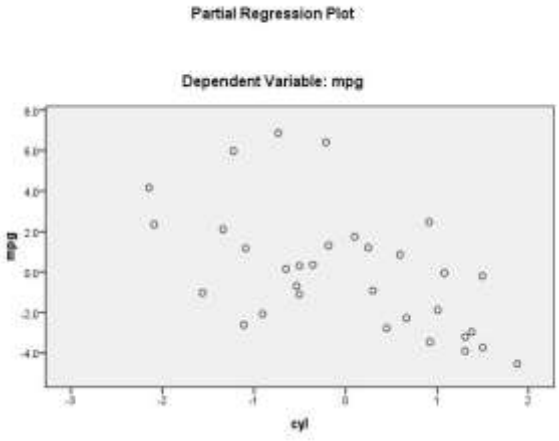
d) **normality** of the error distribution.

a) Linearity of Regression

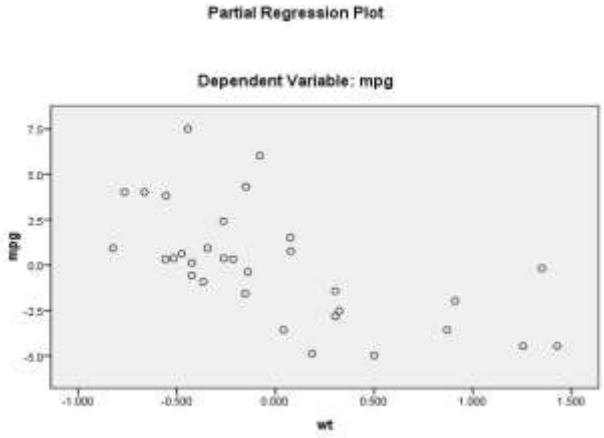
In Classical Multiple Linear Regression Model we assume that the relationship between *y* and $X_1, \ldots, X_p$ is linear in terms of both parameters and independent variables. We are interested in validating the linearity in terms of independent variables.

Linearity of individual regressors in the complete model can be validated using the **partial regression plots**.

For the model under study, we have partial regression plots as:



From the partial regression plot of mpg on cyl we may refer that mpg and cyl are almost linearly related



From the partial regression plot of mpg on wt we may refer that mpg and wt are almost linearly related

b) Autocorrelation

When error terms for different sample observations are correlated to each other, i.e. there is a linear relationship among the error terms the situation is called as autocorrelation or serial correlation. We are interested in detecting if there is an autocorrelation present in the model. This can be done using the **Durbin-Watson Test**.

Durbin Watson Test is based on the assumption that the errors in regression model are generated by a first order autoregressive (AR(1)) process that is,

$$\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i$$

where $\delta_i \sim$ iid $N(0,\sigma_\delta^2)$ and $|\rho| < 1$ is the autocorrelation parameter.

We set up the hypotheses as follows:
Null Hypothesis H_0: $\rho = 0$ Absence of Autocorrelation
Alternative Hypothesis H_1 : $\rho \neq 0$ Presence of Autocorrelation

Durbin Watson test statistic is given by,

$$D = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}(\hat{\epsilon}_i)^2}$$

It can be seen that $D \approx 2(1-r)$ where r is the sample autocorrelation coefficient of the residuals and $0 \leq D \leq 4$. Hence $D = 2$ means no autocorrelation.

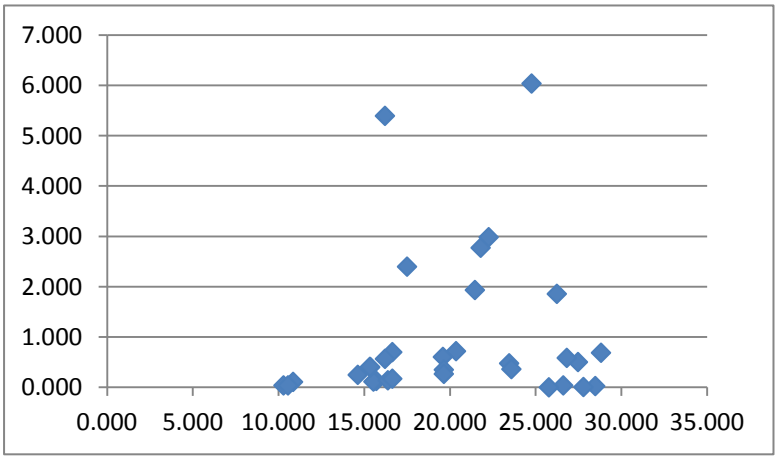
Thus, we have

Durbin watson statistic:	1.617

Conclusion

Thus we have Durbin-Watson statistic $D = 1.671 \approx 2$. Hence we may conclude that the model selected by *stepwise* selection method does not have autocorrelation.

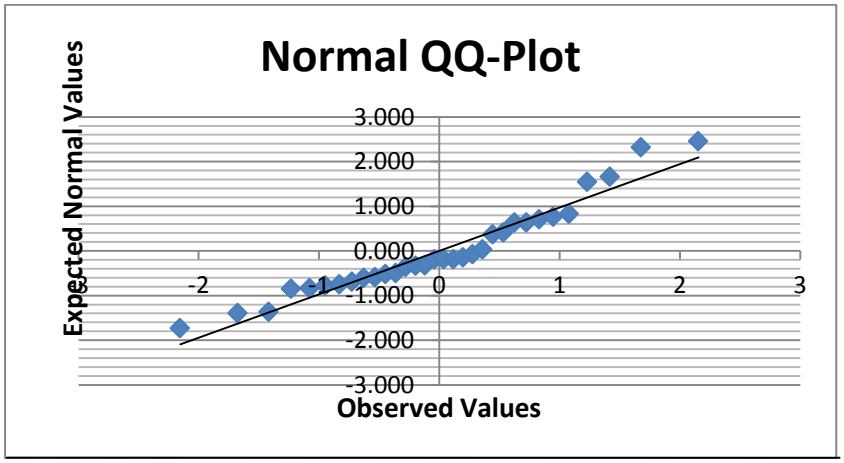
c) Heteroscedasticity



When error terms for different sample observations have same variances the situation is called as homoscedasticity. The opposite of homoscedasticity is called as hetroscedasticity. We are interested in detecting if there is a hetroscedasticity present in the model. We can do this by plotting the predicted dependent variables with the square of the residuals. Since the points do not show any specific pattern, so, from the above scatter plot we may infer that model is homoscedastic. In other words there is no hetroscedasticity in the model.

d) Normality of Errors

We can test the normality of error by means of plotting quantiles of the standardized resiudals against those of a standard normal distribution.



From the normal Q-Q Plot, it can be observed that most of the observations lie around the straight line but few deviate from the line. Thus, we may conclude that the Residuals are almost normally distributed with 0 mean and variance σ^2.

9. Analysis of the results

The tables below represent the regression statistics for the model when all the variables are taken into consideration and the reduced model (stepwise selection).

Regression Statistics(overall)	
Multiple R	0.932
R Square	**0.869**
Adjusted R Square	**0.807**
Standard Error	2.650
Observations	32

Regression Statistics(reduced)	
Multiple R	0.911
R Square	**0.830**
Adjusted R Square	**0.819**
Standard Error	2.568
Observations	32

We can clearly observe that the reduced model is a better fit to the given data as it has a higher adjusted R^2 and involves less number of variables making the model less complex and easy to understand. Moreover, the reduced model not only overcomes the problem of multicollinearity but also validates all the underlining assumptions of regression (refer pt 8).

Sources:

- **Source of Data**
  - Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, **37**, 391–411.
- **Books**
  - Gupta and Kapoor, *Fundamentals Of Mathematical Statistics,* Sultan Chand and Sons, 1970.
  - John W. Tukey., *Exploratory Data Analysis*, Addison-Wesley, 1977.
  - Gupta and Kapoor, *Fundamentals Of Applied Statistics,* Sultan Chand and Sons, 1970.
  - Chatterjee, S. and Price, B. *Regression Analysis by Example*. Wiley, New York, 1977.
  - Darius Singpurwalla, *A Handbook Of Statistics, An overview of statistical methods,* 2013.
  - Mosteller, F., & Tukey, J. W. *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley, 1977.
- **World Wide Web**
  - https://flowingdata.com/2008/02/15/how-to-use-a-box-and-whisker-plot/
  - http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf
  - http://www.123helpme.com/search.asp?text=regression+analysis
  - https://www.youtube.com/watch?v=eYp9QvlDzJA.