Statistics Project

IMPLEMENTATION OF VARIOUS FUNCTIONS USING 'MTCARS' DATASET IN R SOFTWARE

MADE BY:
MUKESH SUTAR- 2101048
ADITYA TETE – 2101993
VINIT SHRIWAS-2101334
SAKSHI GAILWAR-2102184

What is **HYPOTHESIS TESTING** ?

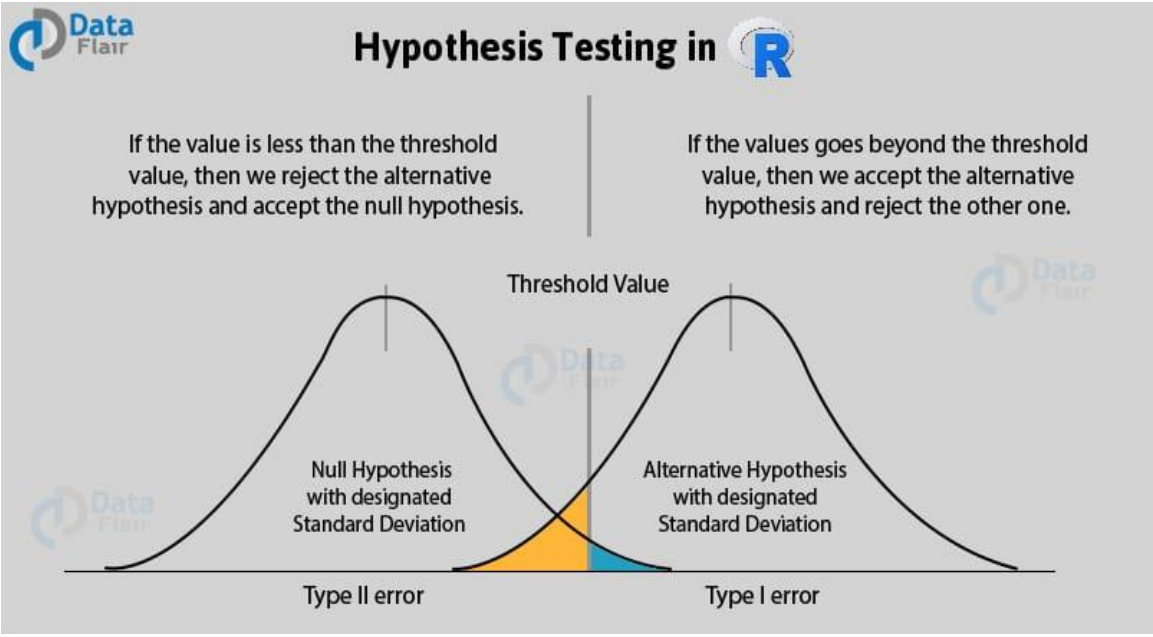
Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population, i.e, it provides a method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population from which the sample was drawn.

The investigator formulates a specific hypothesis, evaluates data from the sample, and uses these data to decide whether they support the specific hypothesis.

STEPS INVOLVED IN HYPOTHESIS TESTING:

There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis and alternate hypothesis (Ho) and (Ha or H1).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.



The Dataset we have used here is 'mtcars'

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2

Showing 1 to 20 of 32 entries, 11 total columns

WE HAVE USED VARIOUS FUNCTIONS OF R SOFTWARE TO ANALYZE THE 'MTCARS' DATASET

SUMMARY()

Initially we have used the command 'summary()' to overview the minimum, maximum, mean, quartiles, etc.

```
> summary(mtcars
+         )
      mpg             cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
      drat             wt             qsec             vs        
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
 Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
       am              gear            carb      
 Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

WHAT IS LINEAR REGESSION MODEL:

A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables). You make this kind of relationship in your head all the time, for example, when you calculate the age of a child based on their height, you are assuming the older they are, the taller they will be.

Linear regression is one of the most basic statistical models out there, its results can be interpreted by almost everyone, and it has been around since the 19th century. This is precisely what makes linear regression so popular.

Next we try to fit Linear Regression Model to predict miles per gallon (mpg) using horsepower (hp)

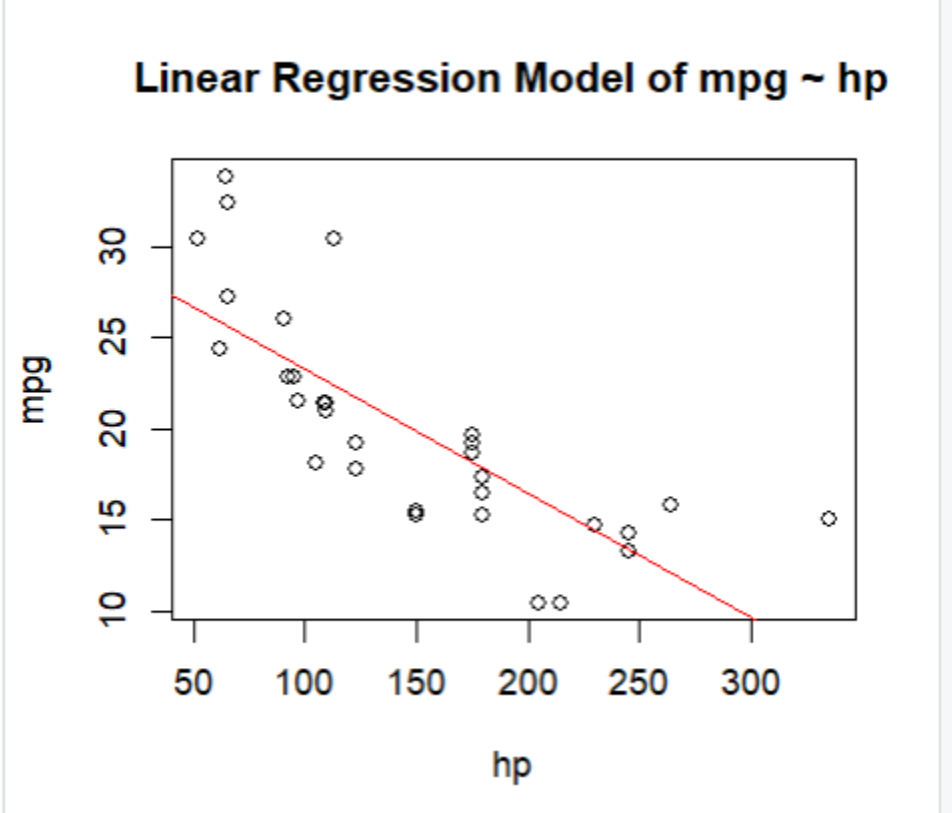
Below is the output

```
# Plot the data points with the regression line
plot(mpg ~ hp, data = mtcars, main = "Linear Regression Model of mpg ~ hp")
abline(model,col="red")
```

```
> # Fit a linear regression model to predict miles per gallon (mpg) using horsepower (hp)
> model <- lm(mpg ~ hp, data = mtcars)
>
> # Print the summary of the model
> summary(model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Using the ‘plot()’ function we have plotted the linear regression model of

miles per gallon (mpg) V/S horsepower (hp)



INTERPRETATION:

As the mpg increases the hp decreases

So, there's a NEGATIVE CORRELATION between mpg and hp

We can verify this using 'cor()' function to find the correlation coefficient between mpg and hp

```
> # Calculate the correlation coefficient between miles per gallon (mpg) and horsepower (hp)
>  cor(mtcars$mpg,mtcars$hp)
[1] -0.7761684
```

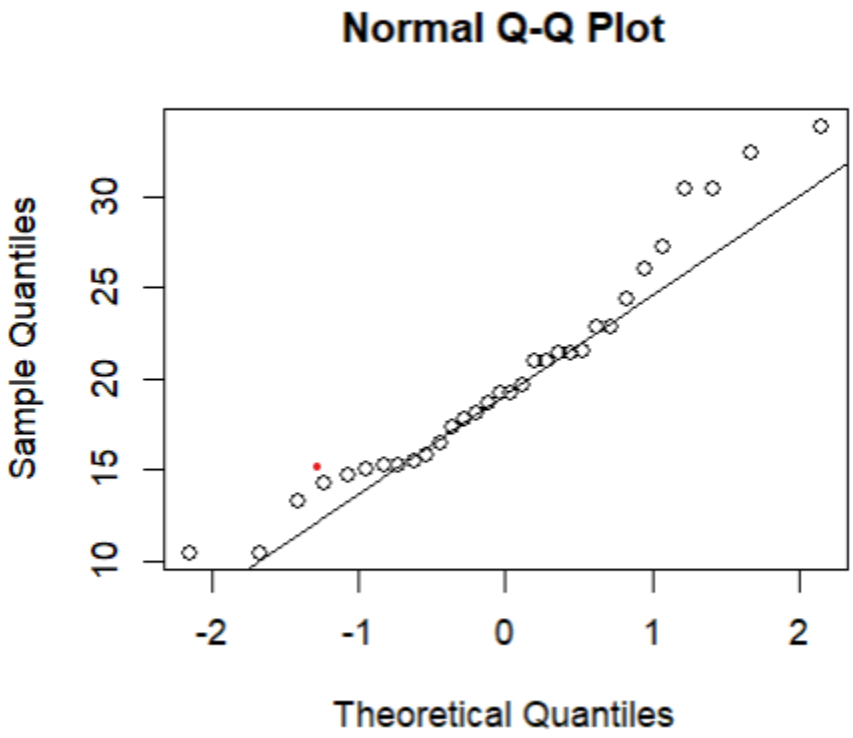
To perform the various hypothesis tests we first have to check whether the dataset is from NORMAL DISTRIBUTION

We can check this using the below methods

1. QQplot
2. Shapiro-Wilk test

```
#Load the mtcars datasheet
 data(mtcars)

#Perform normal probability test using
 qqplot
 qqnorm(mtcars$mpg)
 qqline(mtcars$mpg)
```

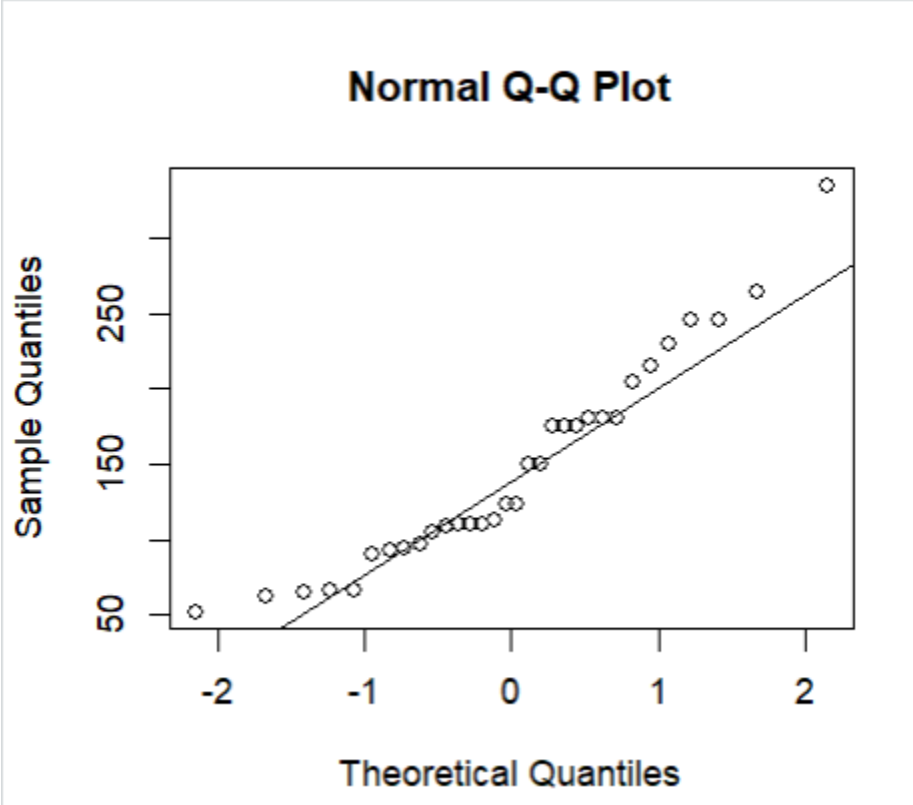
From the above plot we can conclude that the data mpg is from NORMAL DISTRIBUTION. Since all the points lie on the l



ine

Now we check whether the hp data is from Normal distribution

```r
#Load the mtcars datasheet
 data(mtcars)

#Perform normal probability test using
 qqplot
 qqnorm(mtcars$hp)
 qqline(mtcars$hp)
```



From the above plot we can conclude that the data mpg is from NORMAL DISTRIBUTION. Since all the points lie on the line

Both the datasets (mpg and hp) are from NORMAL DISTRIBUTION

NOW WE PERFORM THE FOLLOWING TESTS :

1. Chi-square test

2. t-test

CHI-SQUARE TEST OF INDEPENDENCE

Chi-squared test: You can use a chi-squared test to determine if there is an association between miles per gallon and horsepower. The null hypothesis is that there is no association between the two variables.

```
> # Create a contingency table
> table_mtcars <- table(mtcars$mpg, mtcars$hp)
>
> # Conduct chi-squared test
> chisq.test(table_mtcars)

        Pearson's Chi-squared test

data:  table_mtcars
X-squared = 528, df = 504, p-value = 0.2221
```

INTERPRETATION:

Since the P value is greater than the significance value i.e. alpha=0.05

Therefore we accept the null hypothesis and reject alternative hypothesis

So, there is no association between the miles per gallon and hp.

CONCLUSION:

Using R software we can perform various statistical operations such as Hypothesis Testing , Linear models, Summary, etc and can reach the conclusions with ease an accuracy.