Assignment 1(Problem-2.4)

Objective:-Investigate the degree to which engine displacement contributes in prediction for gasoline mileage by fitting a simple linear Regression model i.e.

$$Y= \beta\ 0+ \beta\ 1*X1+\varepsilon$$

Where ,

Y- Response variable , β0- Intercept , ε-Random Error[follows N(0,σ2)]

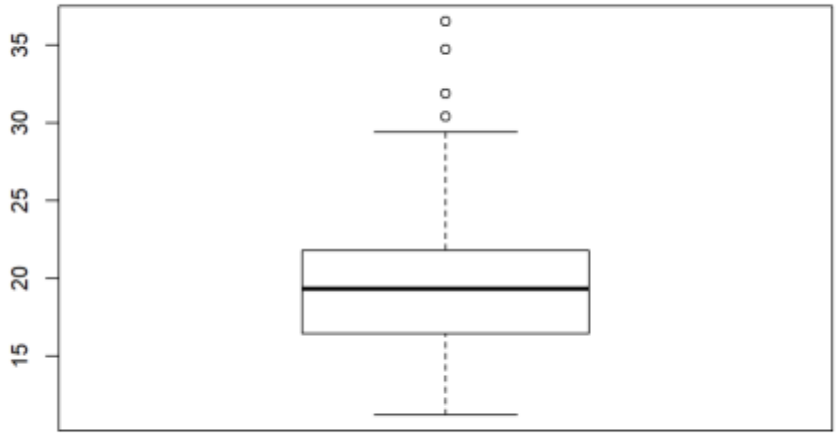
X1-Predictor , β1- coefficient of X1

Here our predictor variable (x1) is engine displacement and response variable(y) is gasoline mileage.

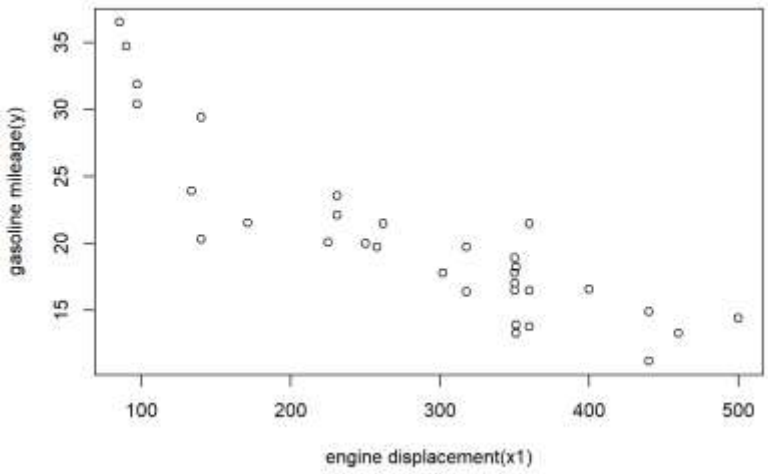
Summary:-

y	x1
## Min. :11.20	Min. : 85.3
## 1st Qu.:16.48	1st Qu.:211.5
## Median :19.30	Median :318.0
## Mean :20.22	Mean :285.0
## 3rd Qu.:21.66	3rd Qu.:353.2
## Max. :36.50	Max. :500.0

Boxplot of Y:-



ScatterPlot of data:-

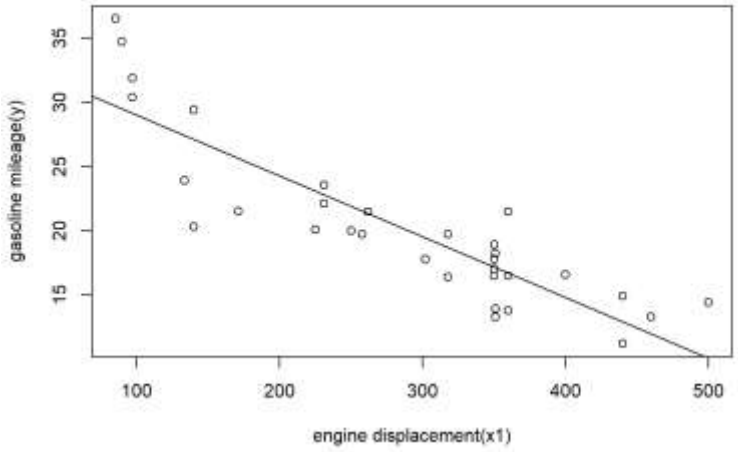


Exploratory data analysis:-

- *Dataset has 32 rows and 12 variables and all variables are quantitative variables in which both the variables of interests are continuous variables.*
- *note that data set has 0 null values means there is no null values in our variables of interest.*
- *There are 20 unique values for x1 and 28 unique values for y.*

- *Minimum and maximum values for x1 are 85.3 cubic inches and 500 cubic inches along with mean value 285 cubic inches.*
- *For y the minimum and maximum values are 11.20 miles per gallon and 36.50 miles per gallon with average 20.22 miles per gallon.*
- *Boxplot of y seems symmetric and have 4 outliers(observations outside of Inter quartile range).*
- *From the scatter plot of y versus x1 shows a linear relationship on average with a negative slope. As engine displacement (x1) is increasing, the gasoline mileage (y) tends to decrease. So we should think of fitting a simple linear regression model and test the significance of the model and check whether there is any violation of assumption or not.*

Model Fitting:-



Summary table:- ##

Call:

lm(formula = y ~ x1, data = data)

##

Residuals:

Min 1Q Median 3Q Max

-6.7923 -1.9752 0.0044 1.7677 6.8171

##

Coefficients:

Estimate Std. Error t value Pr(>|t|)

```
## (Intercept) 33.722677   1.443903   23.36   < 2e-16 ***
## x1          -0.047360   0.004695       -10.09  3.74e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.065 on 30 degrees of freedom
## Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
## F-statistic: 101.7 on 1 and 30 DF,  p-value: 3.743e-11
```

Analysis of Variance Table

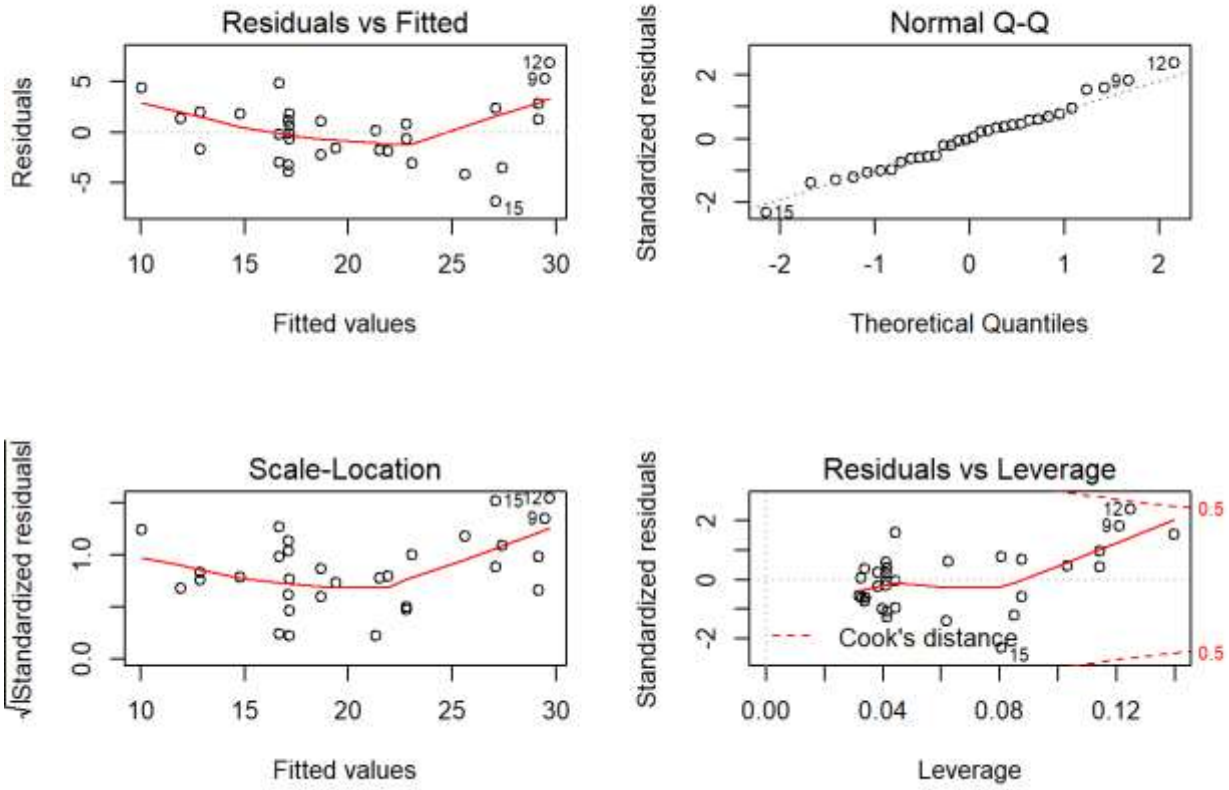
```
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 955.72  955.72  101.74 3.743e-11 ***
## Residuals 30 281.82    9.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions from summary and Anova table of model:-

- *Fitted line:- y = 33.722677-0.047360*x1*
- *Estimates of β0 andβ1 are 33.72267669,-0.04735958 respectively.*
- *The test statistic for H0: β0 = 0 is reported as t0 = 23.36 with p-value=2e-16 ,which is very small in compare to a significance level=0.05.Hence we reject H0 and feel very confident in claiming that the intercept of this model is not zero.*
- *The test statistic for H0: β1 = 0 is reported as t0 = -10.09 with p-value=3.74e-11,which is very small in compare to a significance level=0.05.Hence we fail to accept H0 and feel confident in claiming that there is a significant linear relationship between y and x1.*
- *By above two points we can say that our model is significant and captured 77.23 variability of y since our coefficient of determination R2 is 0.7723.*
- *From the anova table we can observe that SSregression=955.72 with degree of freedom 1,SSresidual=281.82 with degree of freedom (32-2=30) and SSt can be find from the sum of SSregression and SSresidual which is 1237.54 with degree of freedom 31.*

- *Unbiased Estimate of σ2 is MSresidual which is 9.39.*

Output plots of the fitted model:-



Observations from the above plots:-
These plots comes handy to check the assumptions of the model.

- *Linearity assumption can be check by the scatter plot and also by the Residuals versus fitted plot of the data.We can observe there is some discrepancy in the plot 2 to 3 observation have high residuals and causing the discrepancy otherwise all other observations seems to follow the horizontal line indicating residuals are around zero.Also mean of residuals are zero approximately.*
- *From the normal qqplot we can observe that residuals after standardizing seems to fall on the straight line of theoretical normal quantiles and does not show any distinct pattern.Which implies residuals are approximately normally distributed. We can prove it use Shapiro test also.*
- *Homogenity of variance for errors can be check through scale-location plot. It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. In this case residuals appears equally spread for range (0,23) of fitted values but seems to increase afterwards. So errors have uniform variance till fitted values upto 23 afterwards errors showed Heteroskedasticity.*

- *The last plot Residual vs leverage helps to understand the values which can influence our model if excluded. In our case observation having value 12 has high leverage and can influence our model if excluded.*
- *From the Shapiro test of y we got the p-value=0.004439 which is less than 0.05 significance level, hence we need to reject the hypothesis that our yi's follows normal distribution. But if we change significance level from 0.05 to 0.001 them we will fail to reject null hypothesis.*
- *So from these plots we have verified our assumptions for this model and all the assumptions seems to follow except the 5^{th} point.*

Summary of the Assignment :-

- *1. Minimum and maximum values for x1 are 85.3 cubic inches and 500 cubic inches along with mean value 285 cubic inches.*
- *For y the minimum and maximum values are 11.20 miles per gallon and 36.50 miles per gallon with average 20.22 miles per gallon.*
- *Simple regression model seems significant.*
- *Fitted regression model is y = 33.722677-0.047360*x1. And Estimates of β0 andβ1 are 33.72267669,-0.04735958 respectively.*
- *Mean gasoline mileage when engine displacement =0 is 33.72 miles per gallon.*
- *An increase in 1cubic inches engine displacement will decrease mean gasoline mileage by -0.047 miles per gallon.*
- *77 percent of the total variability in gasoline mileage is accounted by the linear relationship with engine displacement.*
- *95 %confident interval on the mean gasoline mileage for 275 cubic inches engine displacement is(19.58,21.80)miles per gallon with the point estimate 20.7 miles per gallon.*
- *95% prediction interval on the mileage is (17.94,22.50) miles per gallon.*
- *We can observe that in the above two confidence intervals first one is smaller and 2^{nd} one is wider because prediction interval for the response variable accounts for both the uncertainty in estimating the population mean with the random variation of individual values. So prediction interval is wider than a confidence interval for mean response.*

References:-

1. *http://analyticspro.org/2016/03/07/r-tutorial-how-to-use-diagnostic-plots-for-regression-models/*
2. *Introduction to Linear Regression Analysis by Douglas C. Montogomery , chapter 1.*