INDIAN STATISTICAL INSTITUTE, NEW DELHI

DATE: 23/12/2021

REGRESSION TECHNIQUES

Under the guidance of Prof. Dr. Swagata Nandi

Submitted by Group V:
Goda Venkata Adithya Tarun (MD2106)
Mainack Paul (MD2111)
Vicky Gupta (MD2129)

DATA DESCRIPTION

In this project, we are using secondary data (TABLE 1, given at page 59) to analyze the Octane Rating of particular petrol as a function of 3 raw materials and a variable that characterized the manufacturing conditions.

- **Nature of the variables**: Quantitative (Continuous)
- **Response variable**: Octane Rating (OR)
- **Explanatory variables:**
  1. Amount of material 1 (A1)
  2. Amount of material 2 (A2)
  3. Amount of material 3 (A3)
  4. Manufacturing condition rating (A4)

Octane Rating: Octane Rating is defined as a standard measure of the ability of a fuel to withstand compression in an internal combustion engine without detonating. And higher the octane number, the more the fuel can withstand before detonating.

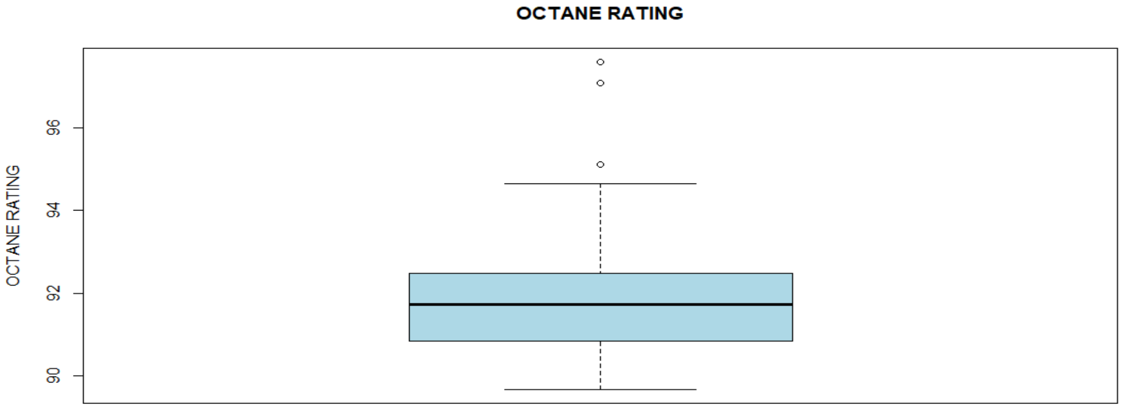
SCRUTINY

Any data needs to be verified for its consistency and homogeneity before starting the analysis. This verification of the data is known as scrutiny. First, we scrutinize the data to check if there are any missing values. We observe that this data set is free from such values. We start the descriptive analysis of the data using boxplot and note some important observations.

DESCRIPTIVE ANALYSIS

Descriptive Analysis helps summarize the data points. This converts the raw data into a form that makes it easier to understand and interpret. We will do our descriptive analysis using a boxplot. We denote by *POP* the potential outlier points and by Q_i the i^{th} quartile for ease of writing.

- ❖ **Analysis of *Octane Rating(OR)***

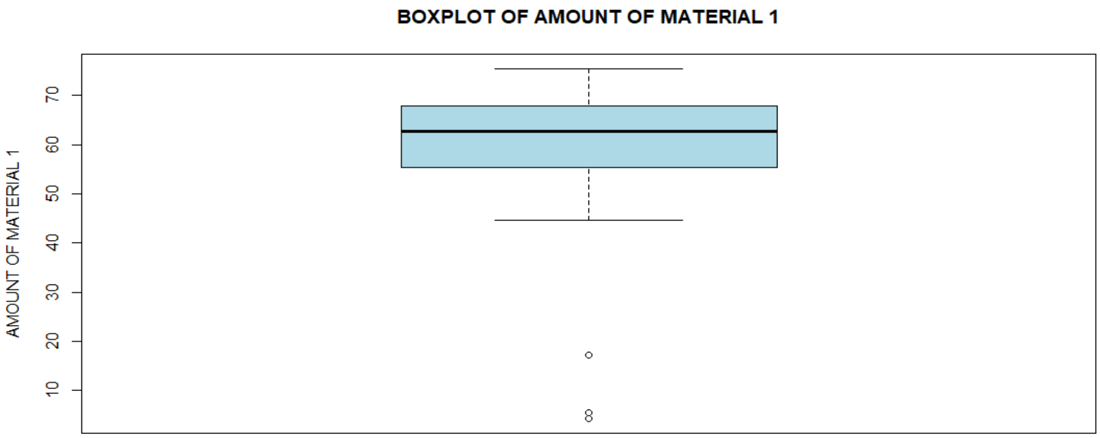


Observations:

Mean	:	91.84988	IQR	:	1.685
Q_1	:	90.8125	Skewness	:	1.598615
Q_2	:	91.735	Kurtosis	:	7.261694
Q_3	:	924975	POP	:	75, 76 & 77

Interpretation: The data of Octane Rating Variable is positively skewed and leptokurtic. There is also a chance of 3 potential outliers.

❖ **Analysis of *Amount of Material 1 (A1)***

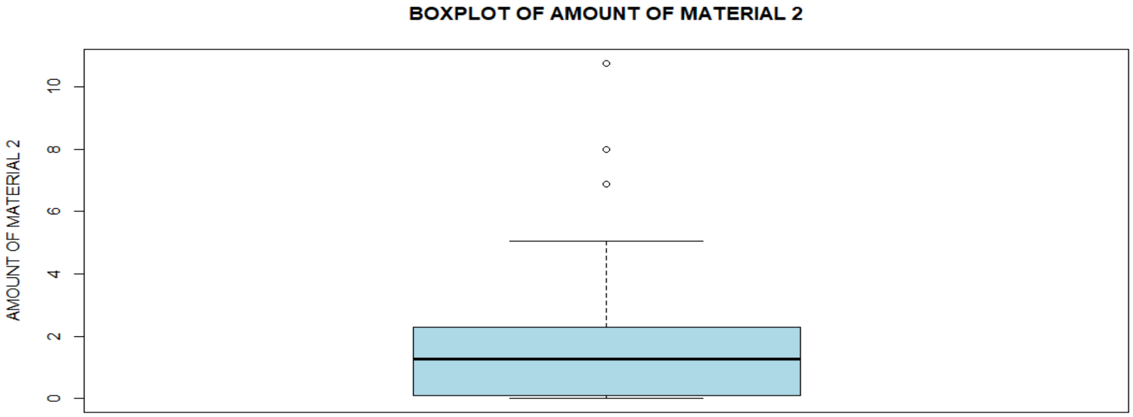


Observations:

Mean	:	60.17061	IQR	:	12.8725
Q_1	:	55.21	Skewness	:	-2.586394
Q_2	:	62.695	Kurtosis	:	11.85312
Q_3	:	68.0825	POP	:	75, 76 & 77

Interpretation: The data of Amount of Material 1 Variable is negatively skewed and leptokurtic. There is also a chance of 3 potential outliers.

❖ Analysis of *Amount of Material 2 (A2)*

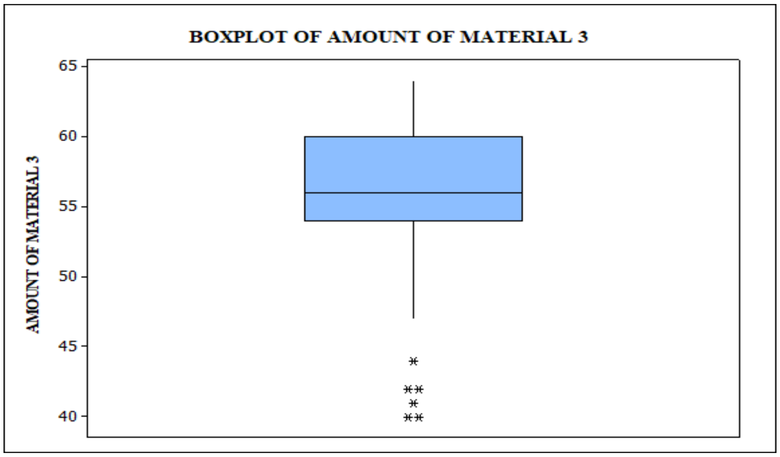


Observations:

Mean	:	1.663659	IQR	:	2.1975
Q_1	:	0.0975	Skewness	:	2.163262
Q_2	:	1.28	Kurtosis	:	9.374097
Q_3	:	2.295	POP	:	44, 75 & 76

Interpretation: The data of Amount of Material 2 Variable is positively skewed and leptokurtic. There is also a chance of 3 potential outliers.

❖ Analysis of *Amount of Material 3 (A3)*

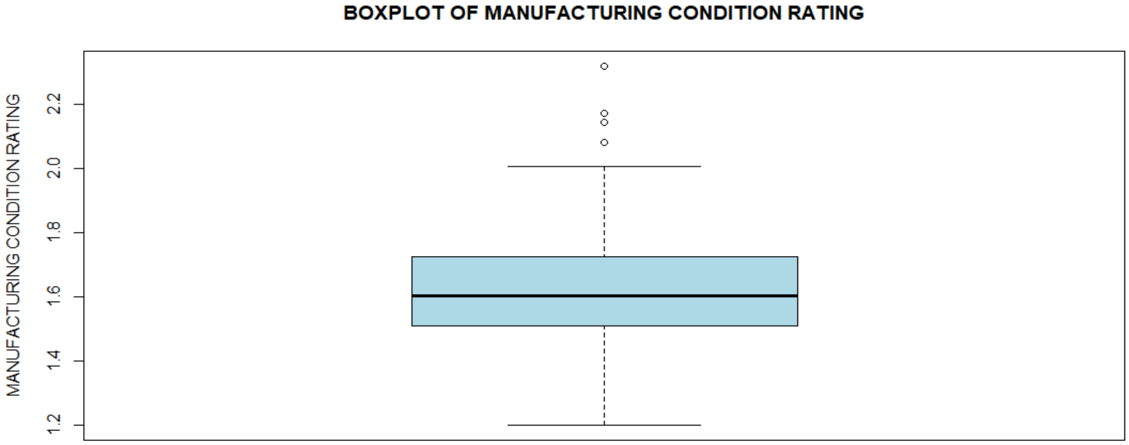


Observations:

Mean	:	55.46341	IQR	:	6
Q_1	:	54	Skewness	:	-1.164489
Q_2	:	56	Kurtosis	:	4.252293
Q_3	:	60	POP	:	71 to 76

Interpretation: The data of Amount of Material 2 Variable is negatively skewed and leptokurtic. There is also a chance of 6 potential outliers.

❖ **Analysis of *Manufacturing Condition Rating (A4)***



Observations:

Mean	:	1.626571	IQR	:	0.216775
Q_1	:	1.50744	Skewness	:	0.7348355
Q_2	:	1.60358	Kurtosis	:	3.977898
Q_3	:	1.72422	POP	:	71 to 74

Interpretation: The data of Amount of Material 2 Variable is positively skewed and leptokurtic. There is also a chance of 4 potential outliers.

Despite we suspect the presence of outliers, we proceed to fit a ‘Least Squares’ model to the given data to see how good the model fit is, and then proceed to check the assumptions of least squares model and apply remedial measures if any assumption is violated. Finally, we can update our model to perform better.

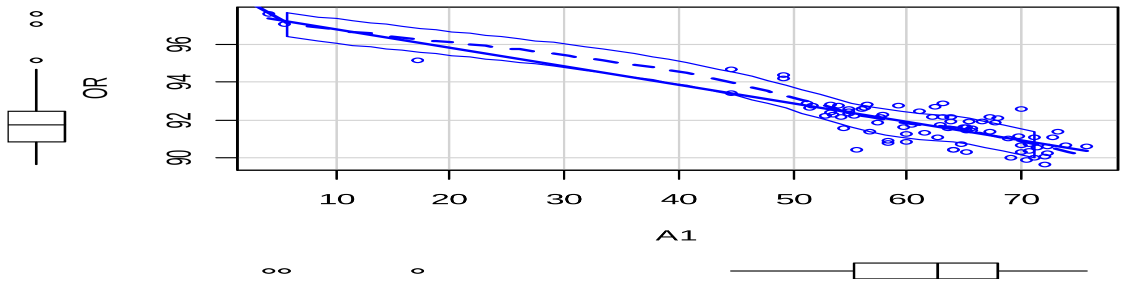
SCATTERPLOTS

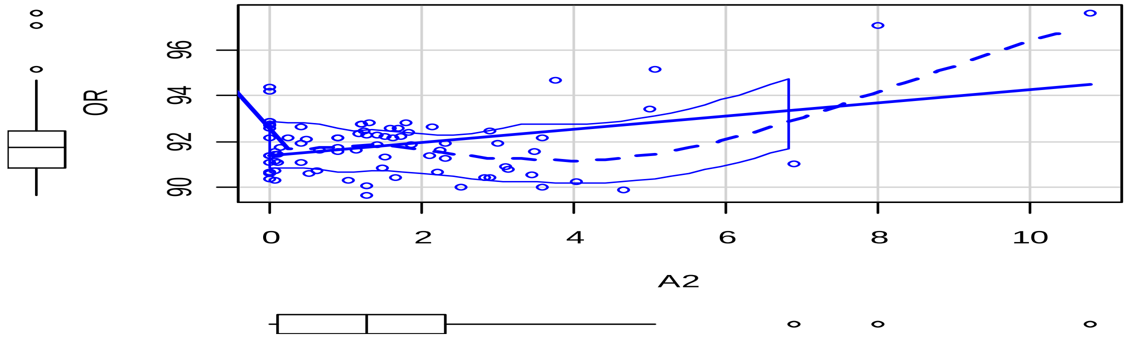
Here, we have plotted the response with the explanatory variables(one at a time).

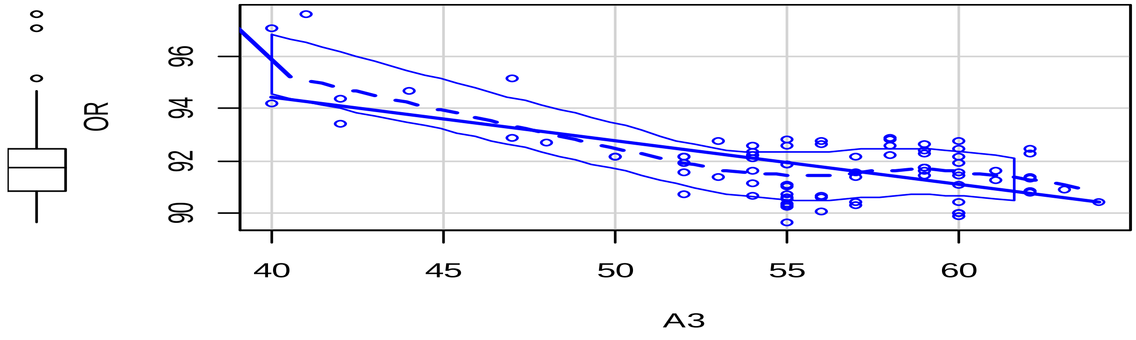
Input:

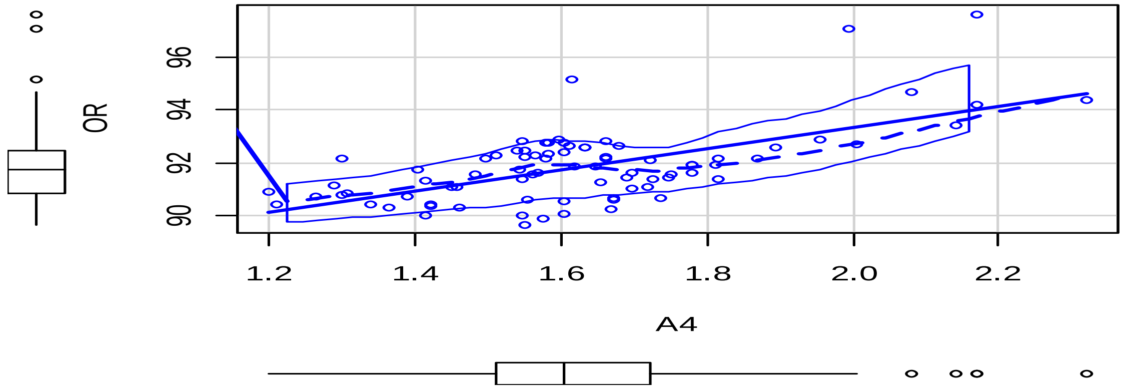
```
library(car)
Data=read.csv(file="E:/ISI/REGRESSION TECHNIQUES/PROJECT.csv")
library(car)
A1=Data1[,2]                    #Extracting the variable
A2=Data1[,3]                    #Extracting the variable
A3=Data1[,4]                    #Extracting the variable
A4=Data1[,5]                    #Extracting the variable
OR=Data1[,6]                    #Extracting the variable
scatterplot(OR~A1)              #Plotting
scatterplot(OR~A2)              #Plotting
scatterplot(OR~A3)              #Plotting
scatterplot(OR~A4)              #Plotting
```

Output:









Interpretation: We can see the scatterplot, linear regression fitted line and a confidence band for each of the explanatory variables against the response. The points are not randomly scattered and nothing specifically can be said about the relationship between the response and the predictors.

LEAST SQUARES FIT

We fit a multiple linear regression model to the given data with an intercept term, including all explanatory variables, that is,

$$OR = \beta_0 + \beta_1 * A1 + \beta_2 * A2 + \beta_3 * A3 + \beta_4 * A4 + \varepsilon$$

under the following assumptions:

- The regression model is linear in parameters.
- The values of regressors are fixed and are independent of the errors ε.
- The mean of error **ε** is **0.**
- The errors ε_i's are homoscedastic.
- There is no autocorrelation between the errors.
- The number of observations is greater than the number of parameters to be estimated.
- There is no correlation between regressors (Absence of Multicollinearity).
- The error **ε** ~ MVN (**0**, $\sigma^2 I_n$).

In the later part of this document, we can find that the assumptions are actually satisfied. If some of the assumptions are violated, we apply necessary remedial measures to get a better model fit. We can see that the number of observations (=82) is greater than the number of parameters to be estimated (= 5), which are the β_is (for i = 0, 1, 2, 3, 4). So, one of the assumptions is satisfied already and we do not need to check for this later.

We can also calculate the values of Residual Sum of Squares (RSS), S^2, R^2, adjusted R^2 as follows:

$RSS = \boldsymbol{e}'\boldsymbol{e}$ where $\boldsymbol{e}$ is the residual vector

$S^2 = \frac{RSS}{n-p}$ where n (=82) is the no. of observations and p (=5) is the number of parameters.

$R^2 = 1 - \frac{RSS}{TSS}$ where $TSS = \sum_{i=1}^{n}(OR_i - \overline{OR})^2$

$Adjusted\ R^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$, which, if close to 1, indicates a decent fit.

The R code to load the data and fit the LS model is given below:

<u>Input</u>:

```
Xtilda = cbind(Data$A1,Data$A2,Data$A3,Data$A4)
Xstar = cbind(
        (Data$A1-mean(Data$A1))/((sum((Data$A1-
        mean(Data$A1))^2))^0.5),(Data$A2-mean(Data$A2))/((sum((Data$A2-
        mean(Data$A2))^2))^0.5),(Data$A3-mean(Data$A3))/((sum((Data$A3-
        mean(Data$A3))^2))^0.5),(Data$A4-mean(Data$A4))/((sum((Data$A4-
        mean(Data$A4))^2))^0.5))
X = cbind(rep(1,82),Xtilda)   #regression matrix
Y = Data$Response             #Response variable
n=82 ; p=5                    #Data size and no. of parameters
model_old = lm(Response~A1+A2+A3+A4,Data)
s = summary(model_old); s     #summary of the fitted regression model
TSS = sum((Y-mean(Y))^2)      #Total sum of squares
RSS = TSS*(1-s$r.squared)     #Residual sum of squares
betacap = s$coefficients[,1]  #fitted regression coefficients
S2 = RSS/(n-p)                #unbiased estimator of variance of errors
Ycap = as.vector(X%*%betacap) #fitted values
e = Y-Ycap                    #residuals

ols_plot_obs_fit(model_old)  #Fitting of actual vs fitted for OR
```

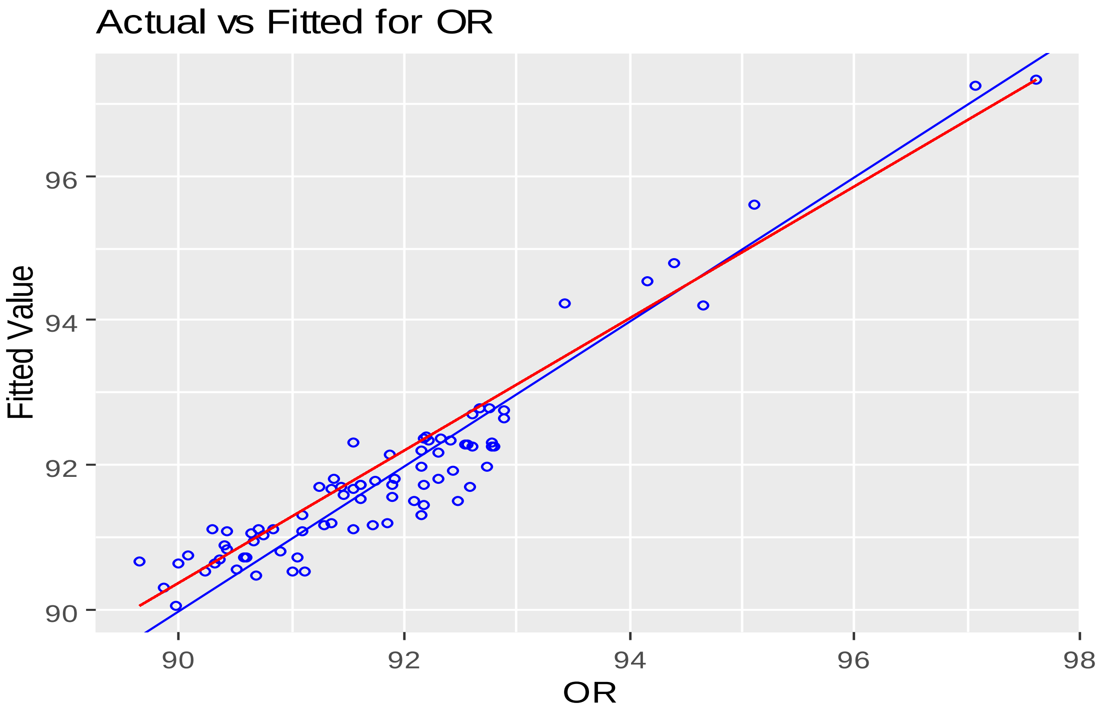
Output:

```
lm(formula = Response ~ A1 + A2 + A3 + A4, data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00612 -0.28588 -0.04679  0.32159  0.98069


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 95.853150   1.224877  78.255  < 2e-16 ***
A1          -0.092821   0.005235 -17.729  < 2e-16 ***
A2          -0.126798   0.032157  -3.943 0.000176 ***
A3          -0.025381   0.013971  -1.817 0.073160 .
A4           1.967603   0.324573   6.062 4.65e-08 ***
Signif. codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4415 on 77 degrees of freedom
Multiple R-squared:  0.9056,  Adjusted R-squared:  0.9007
F-statistic: 184.7 on 4 and 77 DF,  p-value: < 2.2e-16
```

Takeaways from the model:

- The fitted regression model is:

$$OR = 95.8531 - 0.0928 * A1 - 0.1267 * A2 - 0.0253 * A3 + 1.9676 * A4$$



In the y-axis, we have plotted the fitted values of the response obtained using the above stated model and in the x-axis, we have plotted the observed values of the response. We have also plotted a y=x line.

Interpretation: From the plot, it can be seen that most of the points line around the y=x line indicating that the fitted values is pretty near to the observed values. Hence, it meant that the fitted linear regression model is moderately good.

- The data may contain outliers and high leverage points, so the adjusted R^2 value, which is close to 1, should not be thought of as a good measure of goodness of fit yet.

- RSS = 15.00625 ; S^2 = 0.1948 ; R^2 = 0.9056 ; adjusted R^2 = 0.9007
- From the above output, we can see that A3 variable's contribution is insignificant at 5% level of significance. The same can also be observed from the following F-test:

F-test for testing $\boldsymbol{\beta_3} = 0$

We can test the hypothesis H_0: $\beta_3 = 0$ against the alternative H_1: $\beta_3 \neq 0$ by:

F-statistic under H_0 is $F = \dfrac{(A\hat{\boldsymbol{\beta}}-\boldsymbol{c})^T\left(A\left(X^TX\right)^{-1}A^T\right)^{-1}(A\hat{\boldsymbol{\beta}}-\boldsymbol{c})}{qS^2} \sim F_{(q,n-p)}$

where **c** = **0** in this case, A =[0 0 0 1 0] is the q x p matrix of constraints, where q = 1.

<u>Input</u>:

```
A = matrix(c(0,0,0,1,0),nrow=1,ncol=5)
q = nrow(A)
F=((t(A%*%betacap))%*%solve(A%*%solve(t(X)%*%X)%*%(t(A)))%*%(A%*%betacap))/(q*S2)
TabF = qf(0.95,q,n-p)
if(F>TabF)
      {print("A3 variable is significant by F-test")}
else
      {print("A3 variable is insignificant by F-test")}
```

<u>Output</u>:

```
"A3 variable is insignificant by F-test"
```

<u>Interpretation</u>: We have found that A3 variable is insignificant.

We made an assumption that the errors are normally distributed. If this assumption is violated, we could not have performed the above F-test. So, let us check if the assumption is actually satisfied by our data. Later on, we will thoroughly check the presence of outliers and if we find such outlying points, then, we will have to remove those points. We will check for the 'normality' assumption once more to make sure that the deletion of those points has not influenced the normality.

<u>Note:</u> The p-value associated with a test**,** is the probability under the null hypothesis that, the given test statistic takes the observed value and more extreme values in the direction given by the alternative. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis and the chance for the rejection of the null hypothesis increases.

NORMALITY ASSUMPTION BEFORE THE CHECKING OF OUTLIERS

We are often interested in testing hypotheses and constructing confidence intervals about the model parameters. These procedures require that we make the additional assumption that the model errors ε_i are normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean 0 and variance σ^2, that is, homoscedastic. We will check for homoscedasticity later in the following section.

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$$

The need for the Normality assumption

1. ε_i's represent the combined influence (on the dependent variable) of a large number of independent variables that are not explicitly introduced in the regression model. We hope that the influence of these omitted variables is small and random. By Central Limit Theorem, we can say that for large n ε_i's are normally distributed.

2. With the normality assumption, the probability distributions of OLS estimators can be easily derived as any linear function of normally distributed variables is itself normally distributed, which makes our task of hypothesis testing very straightforward.

If the regression assumptions are satisfied then the residuals (e_i's)
have $N_n(0, \sigma^2(I_n - H))$ distribution, which is approximately $N_n(0, \sigma^2)$.

If we find that the fitted values are not normal after visualizing the plots and testing, then, we will have to go for some other method.

Normal Probability Plot

The assumption of normality of disturbances is very much needed for the validity of the results for testing of hypothesis, confidence intervals and prediction intervals. Small departures from normality may not affect the model significantly, but gross non-normality is dangerous.

The normal probability plots help in verifying the assumption of normal distribution. If errors are coming from a distribution with thicker and heavier tails than normal, then the least-squares fit may be sensitive to a small set of data. Heavy tailed error distribution often generates outliers that “pull” the least-squares too much in their direction.

The normal probability plot is a plot of the ordered standardized residuals versus the so-called normal scores. The normal scores are the cumulative probability

$$P_i = \frac{(i-\frac{1}{2})}{n}\ ; i = 1,2, \ldots.,n$$

If the residuals e_1, e_2, ..., e_n are ordered and ranked in increasing order as$e_{(1)}<e_{(2)} <\ldots<e_{(n)}$and the $e_{(i)}$s are plotted against P_i, the plot is called Normal Probability Plot. If the residuals are normally distributed, then the ordered residuals should be approximately the same as the ordered normal scores. So, the resulting points should lay around the straight line with an intercept zero and a slope of one (these are the mean and standard deviation of the standardized residuals).

The Rationale behind plotting $e_{(i)}$ against $P_i = \frac{(i-\frac{1}{2})}{n}$ is as follows:

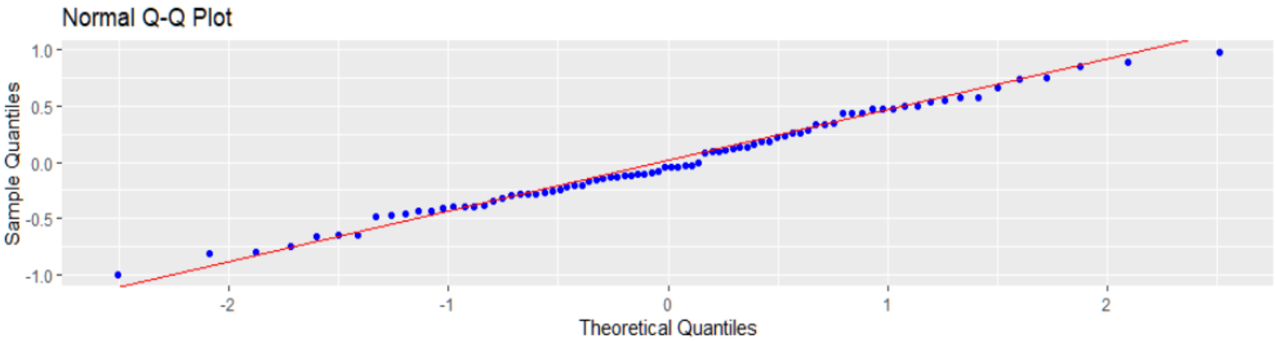
- Divide the whole unit area under the normal curve into ‘n’ equal areas.
- We have a sample of size n data sets.
- We might “accept” that one observation lies is each section, so marked out.
- The first section has one point, so the cumulative probability is P_1 = 1/n. Second section has one point, so cumulative probability up to second section is P_2=(1/n)+(1/n)=2/n and so on.
- Then i_{th} ordered residual observation is plotted against the cumulative area to the middle of i_{th} section, which is.$\frac{(i-\frac{1}{2})}{n}$

Input:

```
library("olsrr")
ols_plot_resid_qq(model_old)          #Q-Q plot of Residuals
```

Output:

```
#The following Normal Q-Q Plot is generated by the above code
```



Interpretation: The quantiles of the residuals almost fall in line with the quantiles of the theoretic normal distribution and hence we can say that the normality assumption is satisfied.

Histogram and Density Plot

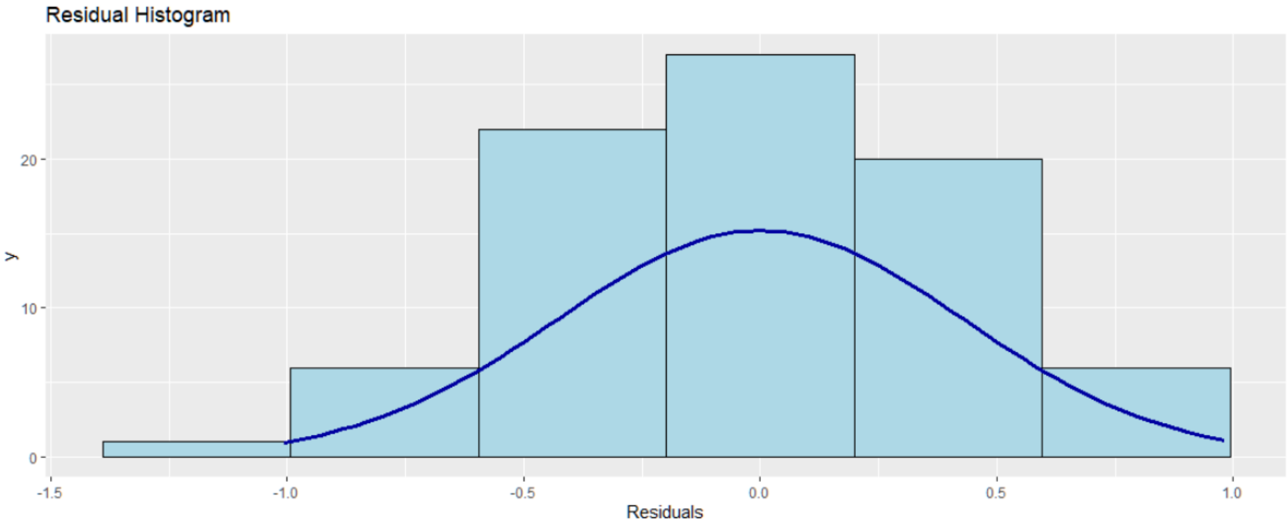
In this method, we plot a histogram of residuals and normal density on the same plot to observe if there is any significant difference between the distribution of the errors and theoretical standard normal distribution. Visualizing the distribution of errors is a graphical method to check the *Normality Assumption.*

These methods of visualization only provide a rough picture of the actual distribution of the errors. They can be used to get an idea about the true distribution of the errors. Also note that we are assuming the residuals are almost similarly distributed as that of the true errors.

Input:

```
library("olsrr")
ols_plot_resid_hist(model_old)       #plotting histogram of residuals
```

Output:



Interpretation: It can be seen from the graph that the errors are almost normal distributed.

However, we have only seen the plots till now and for more accurate conclusion, we will use the 'Shapiro-Wilk' test to check for normality.

Shapiro-Wilk Test

The Shapiro-Wilk provides us a mechanism to check if the random sample actually comes from a normal distribution. The hypothesis we make is:

H_0: The sample comes from a normal distribution against the alternative

H_1: The sample does not come from a normal distribution.

Our chosen level of significance is 0.05.

The test statistic under H_0 for this test is given by:

$$W = \frac{(\sum_{i=1}^{n} \boldsymbol{a_i x_{(i)}})^2}{\sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\bar{x}})^2}$$

where, $x_{(i)}$ is the i^{th} order statistic, $\bar{x}$ is the sample mean and a_is are constants generated from the covariances, variances and means of the sample from a normally distributed sample.

W comes out very small if the sample is not from the normal distribution. Else, it will be large.

The test rejects the null hypothesis if the p-value is less than or equal to 0.05. Failing the normality test allows us to state with 95% confidence that the residuals are not normally distributed and hence the sample did not come from a Normal distribution. Passing the normality test only allows us to state no significant departure from normality was found.

Input:

```
s = shapiro.test(e)
if(s$p.value>0.05)
{print(“We do not reject the null hypothesis at 5% level of significance”)}
else{print(“We reject H0 at 5% level of significance”)}
```

Output:

```
      Shapiro-Wilk normality test

data:  e
W = 0.98902, p-value = 0.7176


[1] "We do not reject the null hypothesis at 5% level of significance"
```

Interpretation: We can conclude that the normality assumption is satisfied by the errors.

Now, the F-test performed earlier has two missing parts, that is, the *homoscedasticity* of the errors and their *independence*. If errors are heteroscedastic, we need to fit a ‘generalized’ least squared model instead of ‘ordinary least squares’ model.

HETEROSCEDASTICITY BEFORE THE CHECKING OF OUTLIERS

Originally, heteroscedasticity means “unequal scatter”. We usually use the term heteroscedasticity in the context of the residuals or error term. Above we have defined all the assumptions of Classical Linear Model, out of which one of them is the homoscedasticity or constant variance: that is, the variance of the error term remains same regardless of the changes in the values of the regressors. If heteroscedasticity is present in the dataset, then, we cannot use ordinary least square (OLS) and we have to modify the model. If we still use the OLS method, the estimators will not be BLUE (Best Linear Unbiased Estimator). Standard errors are biased which leads to bias in test statistics and confidence intervals. OLS gives equal weight to all the observations but observations with larger disturbance term contains less information than observations with smaller variance term.

Assuming that the error variances are unequal and are a function of some known explanatory variables $\mathbf{z}_i$s’ and some unknown quantity $\boldsymbol{\lambda}$,

i.e., $\sigma_i^2 = w(\boldsymbol{z}_i, \boldsymbol{\lambda})$, where for some $\boldsymbol{\lambda}_0$, $w(\mathbf{z},\boldsymbol{\lambda}_0)$ does not depend on $\mathbf{z}$, we proceed to test the heteroscedasticity of errors.

The following are popular and some reliable methods to test the same:

Plot of b_is’ vs Fitted Values

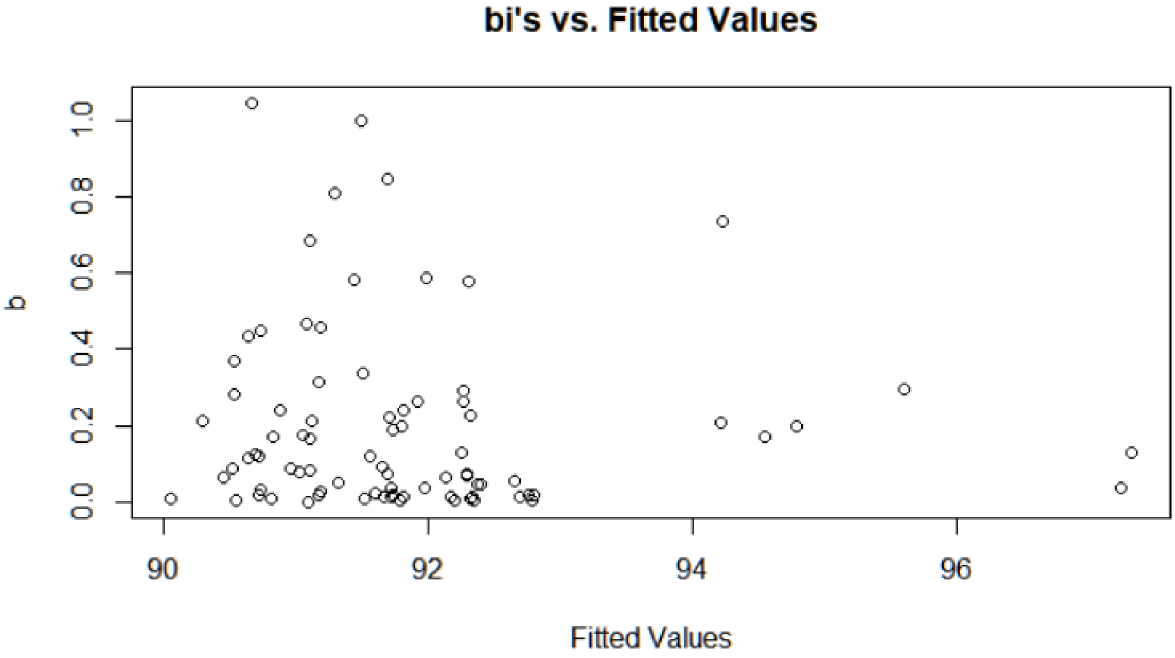
The plot of b_is’ vs fitted values of the response *OR*, if comes out to be wedge shaped or fan shaped, it means that the variances of errors increase with the means of the response. Similar argument works for the decrease of means,

where, $b_i = \frac{e_i^2}{(1-h_i)} \; \forall \; i = 1, 2, \dots, 82$

Input:

```
b = e^2/(1-hii)

plot(Ycap,b,main = "bi's vs.Fitted Values",

           xlab = "Fitted Values",ylab = "b")
```

Output:



Interpretation: We can see that the plot is not wedge shaped, so we can say that there is no heteroscedasticity with respect to means. The points which seem to be far away from the bulk of the data in the above plot must not be misinterpreted to be due to heteroscedastic nature of errors, as they are simply away in the x-direction. It indicates the presence of high-leverage in the X-data.

We will deal with this problem in the following section. This plot is equivalent to plotting r_i^2 vs. the fitted values, where r_is' are internally studentized residuals. This can be seen below.

Plot of r_i^2 vs. the Fitted Values

The internally studentized residuals are defined as follows:

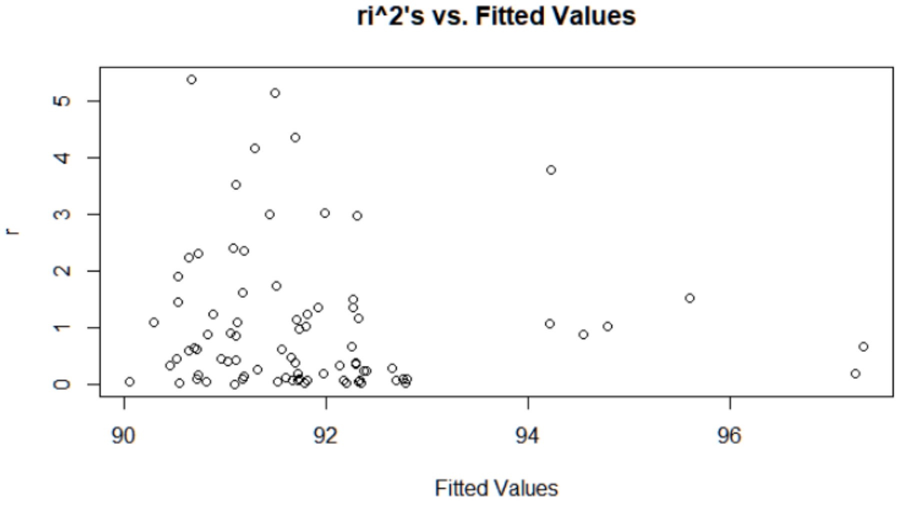
$$r_i = \frac{e_i}{S(1-h_i)^{1/2}} \; \forall \, i = 1, 2, \ldots, 82$$

This result in a similar plot as above as b_i is directly proportional to r_i^2. So, we expect that we will arrive at the same conclusion using this plot as we did in the above, i.e., the plot of b_is' vs fitted values of the response *OR*.

Input:

```
r = e/sqrt(S2*(1-hii))
plot(Ycap,r^2,main="ri^2's vs. Fitted Values",xlab="Fitted Values",ylab="r")
```

Output:



Interpretation: As said above, the two plots differ only in the scale. So, no heteroscedasticity with respect to means are present.

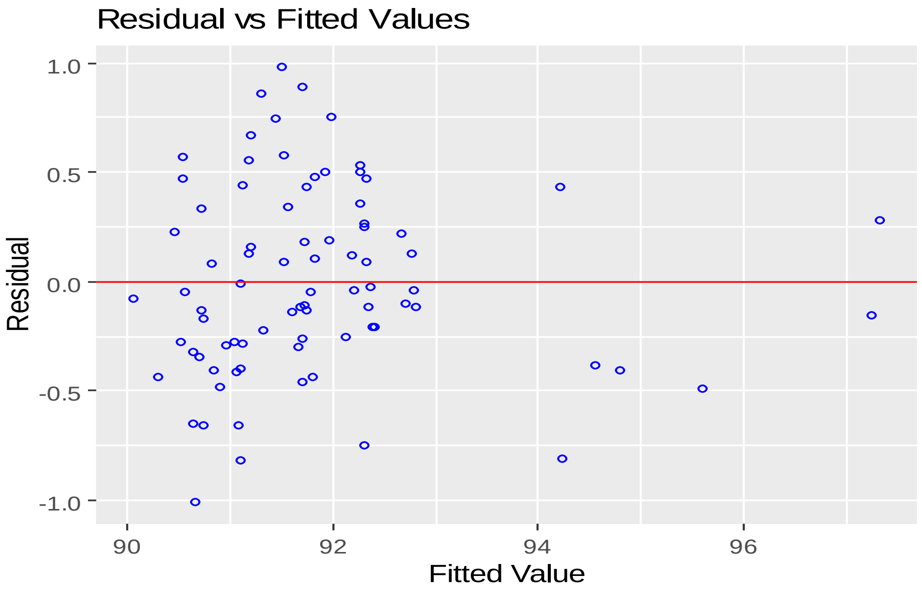
Residuals vs. Fitted Values (Residual Plot)

Instead of observing the b_is' or the r_i^2s', we can directly observe the plot of e_is' vs the fitted values of the response variable. Heteroscedasticity, especially the form with increasing or decreasing variances of errors with the means, can be seen with fan-shaped pattern.

Input:

```
ols_plot_resid_fit(model_old) # Residual Plot
```

Output:



Interpretation: There is no fan shape pattern observed and thus no heteroscedasticity with respect to means is present. We observe that the plot is not totally random. The remote points in the x-direction are probably due to the high-leverage present in the X-data.

So, we could not expect much heteroscedasticity in the errors by the above plots. However, to make a rigid conclusion, we need to perform a formal test for the same.

Breusch-Pagan Test

The Breusch-Pagan test (or BP Test) is a formal test for checking the presence of heteroscedasticity. To test the hypothesis:

H_0: Homoscedasticity is present vs. the alternative H_1: Heteroscedasticity is present,

we employ the BP Test. The procedure of the test is as follows:

- Fit the OLS regression model.
- Calculate the squared residuals of the model.
- Fit a new regression model, using the squared residuals as response variable.
- Calculate the Chi-Square test statistic, χ^2as $n*R^2_{new}$, where, n is the total observations (=82) in this case and R^2_{new} is the R^2 for the new regression model that used the squared residuals as the response values.

The test criterion is to reject the null hypothesis if p-value comes out to be less that 0.05 (5% level of significance), otherwise do not reject H_0.

This test can be easily performed in R using a simple command:

Input:

```
install.packages("lmtest")     #install this library is not installed yet
library(lmtest)                #Load the lmtest library
bptest(model_old)              #Command to run the BP Test
```

Output:

```
      studentized Breusch-Pagan test

data:  LSE
BP = 2.8947, df = 4, p-value = 0.5756
```

As the p-value is greater than 0.05, we fail to reject H_0 and conclude that there is no heteroscedasticity present in the data at 5% level of significance.

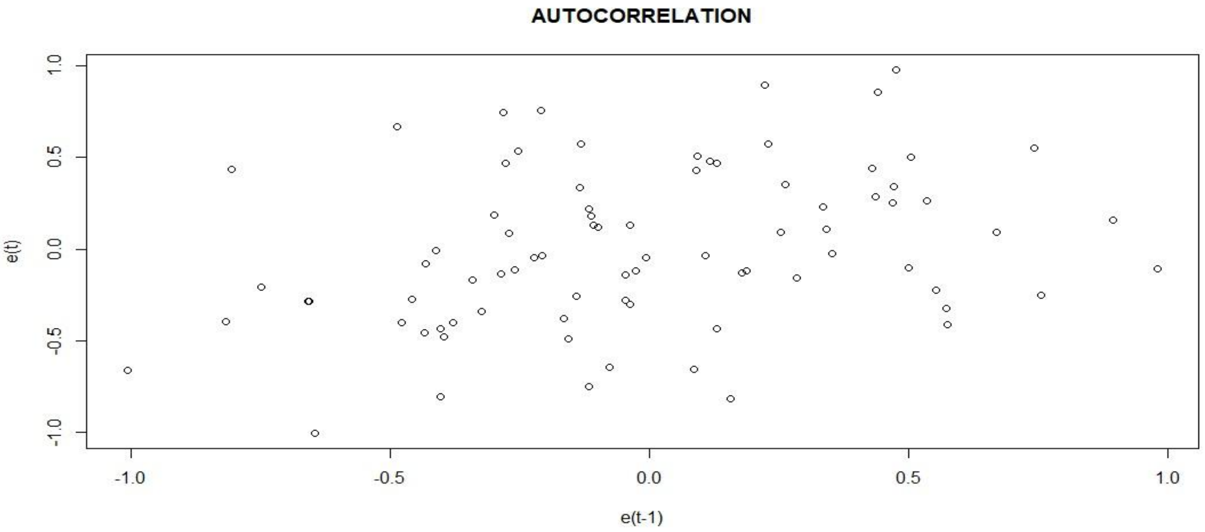
However, since we have stated above that from the plots, it is revealing that there can be some high leverage points present. We check for the presence of such potential high leverage points and if found, we delete them and we again check for the presence heteroscedasticity after the removal of those points as there is a chance that the problem is still well hidden by the leverages. The only missing part now is to check for the independence of the errors.

AUTO-CORRELATION OF ERRORS

The '**auto**' part of autocorrelation comes from the Greek word for '**self**' and autocorrelation means data that is correlated with itself as opposed to being correlated with some other data. Autocorrelation is a mathematical representation which gives us the degree of similarity between a given time-series and a lagged version of itself over successive time intervals. For a given series of numbers if there is a pattern in such a way that the values in the series can be predicted based on the preceding values in the series, the series of numbers is said to exhibit autocorrelation. It may also be termed as 'serial correlation' and 'serial dependence'. An autocorrelation of '+1' represents a perfect positive correlation, while an autocorrelation of '-1' represents a perfect negative correlation.

Above we have defined all the assumptions of Classical Linear Model, out of which one of them is that for the given values of the regressors, there is no autocorrelation between the disturbance terms. If autocorrelation is present in the dataset, then, we cannot use ordinary least square method for estimating the model parameters and we have to modify our method. Those estimators will not be BLUE (Best Linear Unbiased Estimator) if we still use the OLS method. Standard errors are biased which leads to bias in test statistics and confidence intervals. If the errors are correlated, then the variance-covariance matrix will no longer be a diagonal matrix, due to which the need arises to change the whole regression technique, and it also complicates the computations. Hence, it is necessary to check for the presence of autocorrelation in the dataset.

First we will plot a graph of ε_t vs $\varepsilon_{t-1}$and check whether any pattern can be found or not.



Interpretation: No pattern is observed and thus roughly we can say that autocorrelation may not be present.

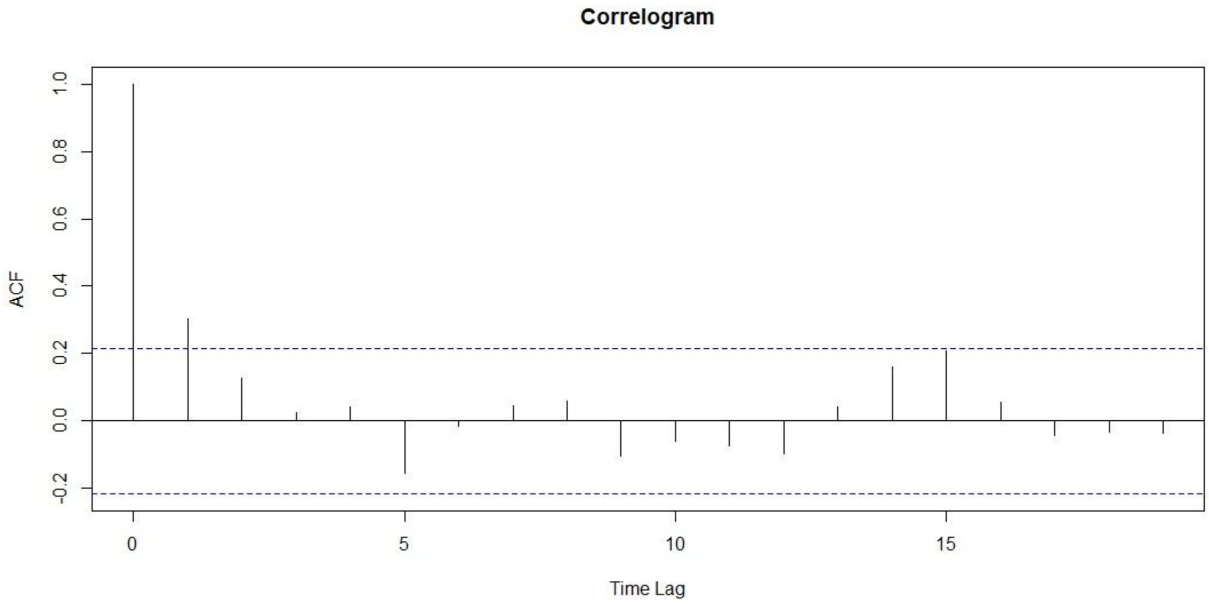
Correlogram

The correlogram is a commonly used tool for checking randomness in a data set. If random, autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

Input:

```
acf(e,xlab="Time Lag",ylab="ACF",main="Correlogram")
```

Output:



Interpretation: From the graph, we can clearly visualize that all the errors are very near to zero. Hence, we can roughly interpret that there is no presence of autocorrelation.

For more conclusive results, we test for the autocorrelation in the model. Assuming that the errors follow first-order autoregressive scheme (i.e., Markov model), we use the Durbin-Watson test.

Durbin -Watson Test

To test H_0 i.e., no autocorrelation vs H_1: Not H_0

After the errors are arranged in time order, we can test for their independence using the Durbin-Watson D statistic, which is given by: $D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$.

Input:

```
num = 0; for (i in 2:n) { num = num+(e[i]-e[i-1])^2 }
d = num/sum(e^2); d
```

Output:

```
[1] 1.342342
```

From the Durbin-Watson tables, we observe that the cutoff values *dl = 1.53* and *du = 1.74*, for n=82 and p-1=4 at 5% level of significance. As the ‘d’ value is less than ‘dl’, we suspect positive autocorrelation in the model.

But note that the given data is not a time series data, so, there is no specific time ordering to the residuals. Had we taken the same data in a different order, we would have got some other d value which might conclude the absence of autocorrelation. So, it makes no sense to find autocorrelation in this data. Hence, we take for granted that the errors are independently distributed.

So, the F-test for insignificance of variable A3 is completely fine now. Although we now suspect that A3 variable might have to be removed from the model. By doing so, there is a chance that A3 can contribute significantly to the model fitted using the data free from the above-mentioned points. But, before actually proceeding to check for presence of outliers and high-leverage points, we observe the multicollinearity. As multicollinearity between the explanatory variables can be influenced by the outliers, we observe the state of multicollinearity before and after the checking of outliers. We suspect this because of the simple correlation matrix obtained as below:

```
cor(Xstar)#R code for correlation matrix

           A1         A2         A3         A4
A1  1.0000000 -0.5894513  0.4487330 -0.3369172
A2 -0.5894513  1.0000000 -0.2983958  0.1611956
A3  0.4487330 -0.2983958  1.0000000 -0.7217482
A4 -0.3369172  0.1611956 -0.7217482  1.0000000
```

High correlations can be seen in the above matrix, which gives suspicion for existence of multicollinearity in the X-data.

MULTICOLLINEARITY BEFORE THE CHECKING OF OUTLIERS

The term ‘multicollinearity’originally meant the presence of a perfect or exact linear relationship among some or all of the explanatory variables in a multiple regression model. It is necessary to check for multicollinearity while fitting a multiple regression model. If the magnitude of correlation between the explanatory variables is high, then the determinant of the X^TX matrix will be close to zero and it may not be invertible. It also causes high variances of regression estimates.

One of the main aims in a regression analysis is to isolate the relationship between the explanatory variables. The interpretation of a regression coefficient is that it represents the amount of change of the dependent variable for a unit change in an explanatory variable, *keeping all the other variables constant***.** But if multicollinearity is present among the explanatory variables, a change in one explanatory variable will be associated with shifts in other explanatory variables as well.

The higher the magnitude of correlation between the explanatory variables, the more difficult it is to change one explanatory variable without changing the other. Also, the estimators will have large variances and covariances making precision difficult and hence the confidence intervals tend to be wider. Another problem is that the t-ratios will be statistically insignificant. Even if R^2 value is high, there is a high chance that some of the regression coefficients will be statistically insignificant. Hence, it is necessary to detect and fix this problem.

DETECTION OF MULTICOLLINEARITY

We detect the presence of multicollinearity using the following methods. However, there is no guarantee that all the following methods may lead to the same conclusion. These methods are not perfect but rather they give a pretty good idea of the presence or absence of multicollinearity. They are not completely to be relied upon but we get a vague picture.

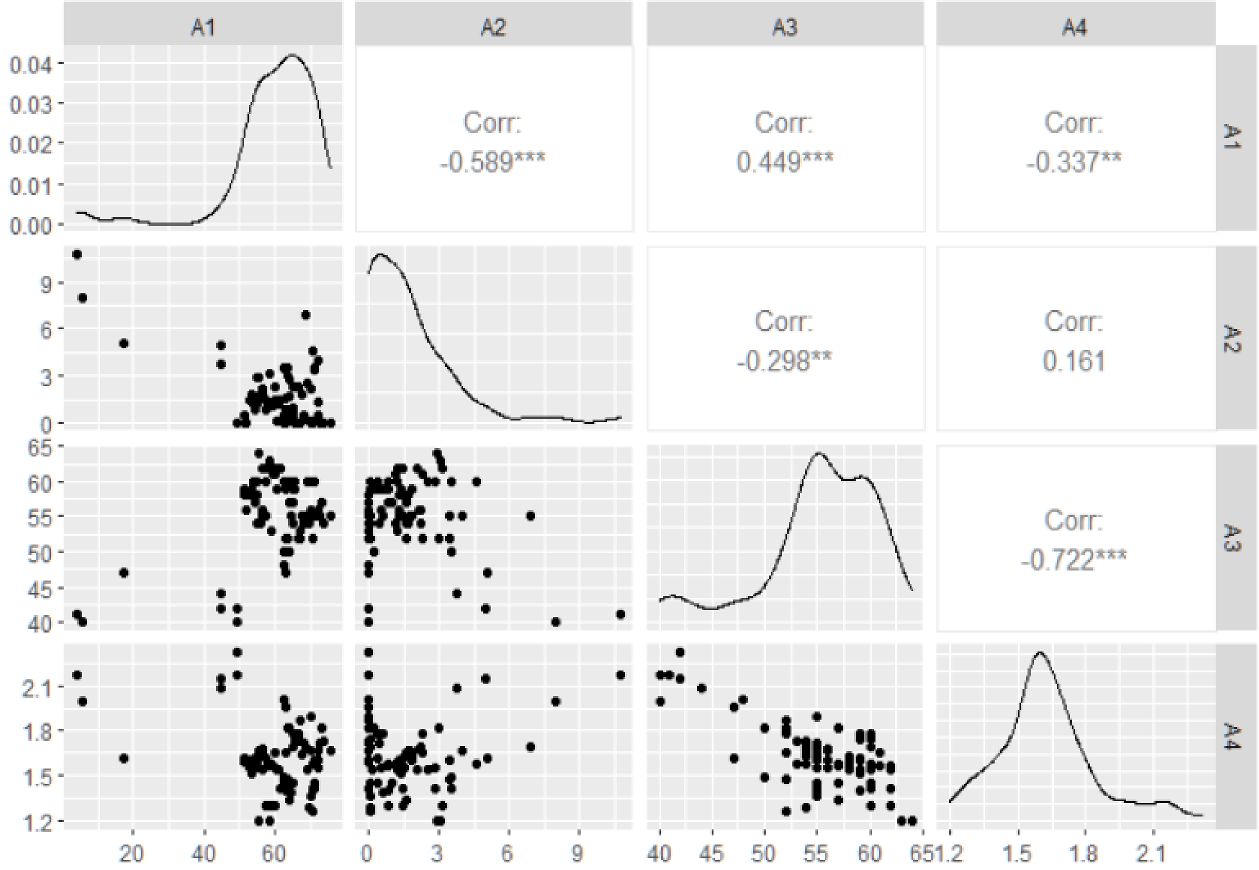
Scatter Plot

Using a scatterplot, we can see a rough picture about the nature of the relationship existing between the independent variables.

Input**:**

```
install.packages("GGally")     #install GGally package if not installed yet
library("GGally")              #Load the GGally library
ggpairs(Data[,-c(1,6)])        #This   generates   scatter   plot   of   all   the
                       variables

#This also generates the partial correlation between the variables
```

Output:



Scatter Plot of all the variables in the regression model

Interpretation: From the above scatterplot, we can visualize the nature of relationship between the explanatory variables. The most observable simple correlation coefficient is between amount of material 3 (A3) and manufacturing condition rating (A4) which is -0.722. Hence, the above-mentioned pair is highly correlated although this fact is unknown if there is any lurking effect present.

Partial Correlation Matrix

For getting more conclusive results, we need a thorough study which is done with the help of partial correlation coefficients.

These coefficients can be obtained as follows:

Input:

```
Install.packages("ppcor")          #install this library if not yet installed
library(ppcor)
pcor(Data[,-c(1,6)],method="pearson")#Calculating Partial Correlation Matrix
```

Output:

```
            A1          A2          A3          A4
A1  1.00000000 -0.5377470  0.2055322 -0.07700686
A2 -0.53774697  1.0000000 -0.1093737 -0.11033963
A3  0.20553218 -0.1093737  1.0000000 -0.68198145
A4 -0.07700686 -0.1103396 -0.6819815  1.00000000
```

From the above matrix, we come to a stronger suspicion for presence of multicollinearity between the variables amount of material 3 (A3) and manufacturing condition rating(A4) as the correlation comes out to be -0.68198145. Another cause of this problem can be due to the presence of outliers.

Hence, we shall use some other measures for detecting the presence of multicollinearity. The measures that we will use here are the *Variance Inflation Factors* and the *Condition Number*.

Variance Inflation Factors

Variation Inflation Factor (VIF) is defined as the ratio of the overall model variance to the variance of a model that includes only that single explanatory variable. It is given by:

$$VIF = \frac{1}{(1 - R_k^2)}$$

where R_k^2 denotes the R^2 value when the k^{th} predictor is regressed over the other predictors.

Thumb rule for VIF is that if its value is less than 5, it indicates the absence of multicollinearity due to that particular variable, and being greater than 5 indicates the presence of the multicollinearity due to that particular variable.

Input:

```
library(car)
vif(model_old)
```

Output:

```
   A1       A2       A3       A4 
1.762293 1.554770 2.346033 2.114003
```

Interpretation: Since each VIF is less than 5, we can say that no serious multicollinearity is present in our model.

Condition Number

As X* denotes the scaled and centered version of the design matrix of our model, the *Condition Number* is defined as $CN = \sqrt{\frac{Largest\ \ EigenValue\ \ of\ X*^{T}X*}{Smallest\ \ EigenValue\ \ of\ X*^{T}X*}}$

Thumb rule for CN is that if its value is less than 10, then, it indicates absence of multicollinearity whereas lying between 10 & 30 indicates that the presence of multicollinearity is moderate; higher than 30 indicates presence of strong multicollinearity.

Input:

```
l = eigen(t(Xstar)%*%Xstar)$values

sqrt(max(l)/min(l))
```

Output:

```
[1] 2.95007
```

Interpretation: The value of CN came out to be 2.95007. Hence, using the CN criterion, we can say that no serious multicollinearity is present.

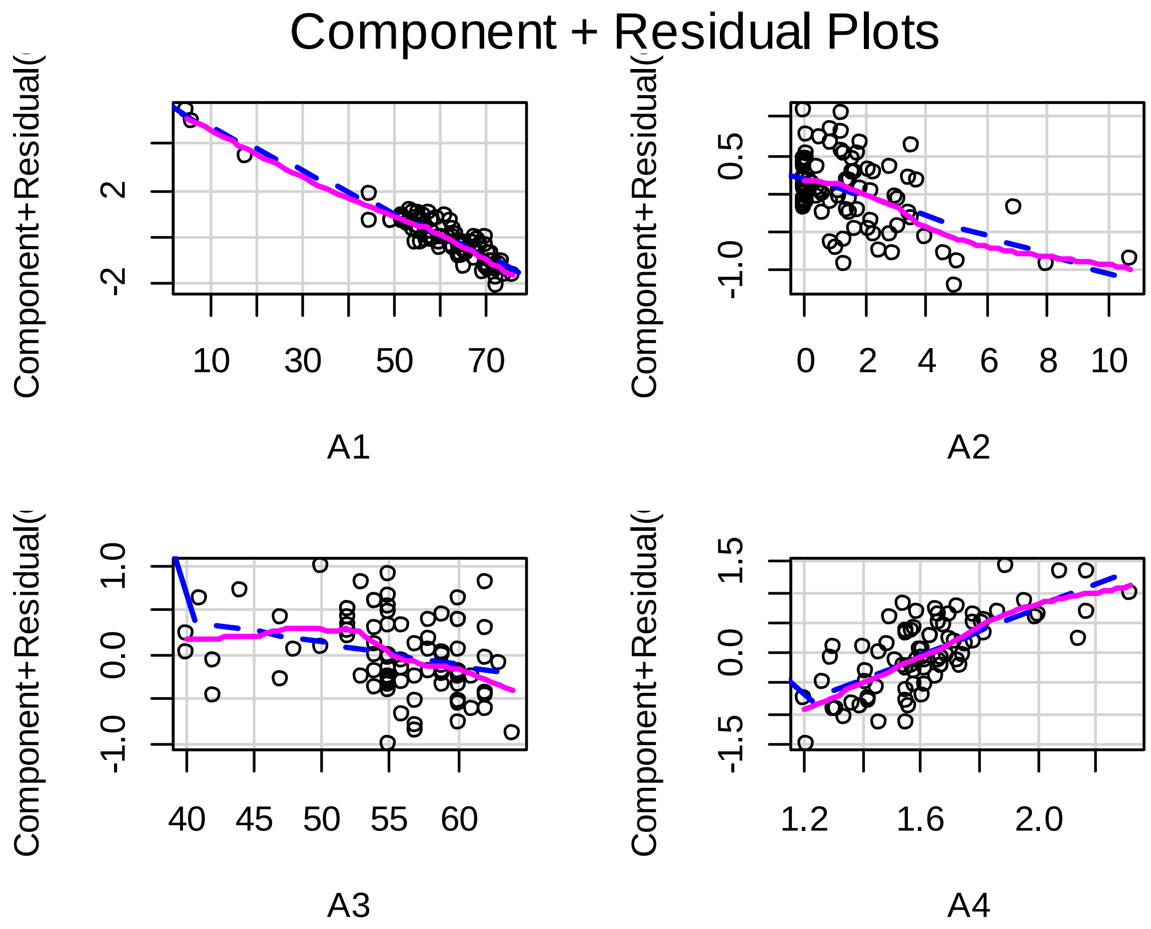
As we have now noted the situation of multicollinearity in our model, we will again check the presence of multicollinearity in our model after checking the presence of outliers, leverages and high influential points.

PARTIAL RESIDUAL PLOT

In applied statistics, a partial residual plot is a graphical technique that attempts to show the relationship between a given explanatory variable and the response variable given that other explanatory variables are also in the model.

Input:

```
library(conf)
crPlots(model_old)
```



Interpretation: There are some significant partial correlations of OR with other Explanatory variables. Also, we can observe in the partial residual plots that Linearity Assumption between the Response and Explanatory variables is more or less evident.

ADDED VARIABLE PLOT

An added-variable plot is a effective way to show the correlation between an independent variable and a dependent variable conditional on other independent variables. For multivariate estimation, a simple scatterplot showing x versus y is not adequate to show the partial correlation of x with y because it ignores the impact of the other covariates. Added-variable plots are also useful for spotting influential outliers in the data which affect the estimated regression parameters.

It is a scatterplot of the transformations of an independent variable (say, x1) and the dependent variable (y) that nets out the influence of all the other independent variables. The fitted regression line through the origin between these transformed variables has the same slope as the coefficient on x1 in the full regression model which includes all the independent variables.

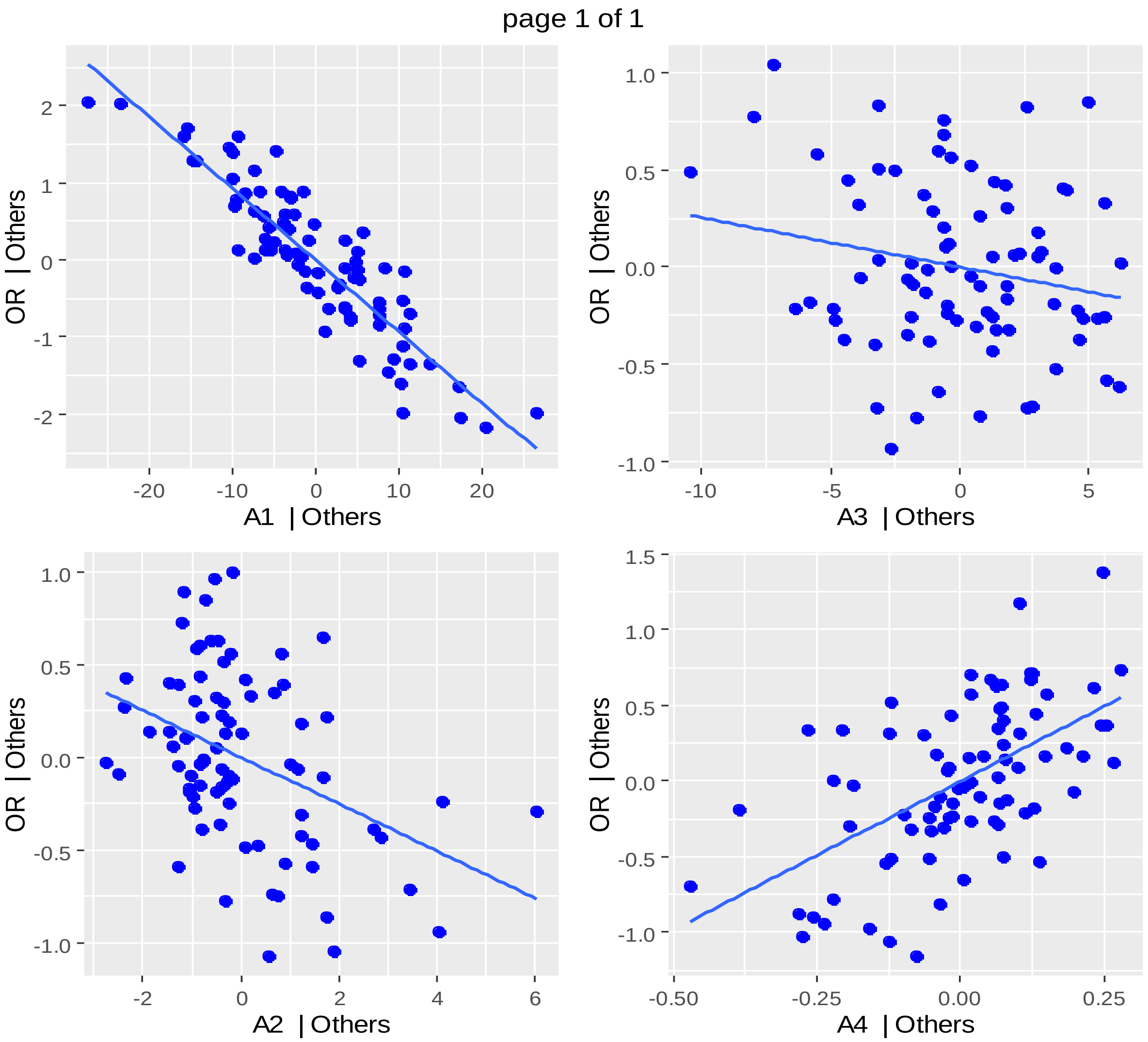
We can see it as the multivariable analogue of using a simple scatterplot with a regression fit when there are no other covariates to show the relationship between a single x variable and a y variable. It is a visually compelling method for showing the nature of the partial correlation between x1 and y as estimated in a multiple regression.

Input:

library(olsrr)

```
>ols_plot_added_variable(model_old)
```

Output:



Interpretation: There are some significant partial correlations of OR with other Explanatory variables. Also, we can observe in the added variable plots that Linearity Assumption between the Response and Explanatory variables is more or less evident.

GOODNESS OF FIT

The goodness of fit of a statistical model is often used to describe how well the statistical model fits a set of observations. The measures of goodness of fit summarize the discrepancy between the observed values and the fitted values obtained from the model that is being fitted. In regression analysis, coefficient of determination (R^2) is used to measure the goodness of fit.

In statistics, R^2 denotes the proportion of the explained variability in the model, that is, it is the proportion of the variance in the dependent variable that can be predicted from the explanatory variables.

The mathematical formula of R^2 is given by

$$R^2 = 1 - \frac{RSS}{TSS}$$

where R^2: Coefficient of Determination
RSS: Residual Sum of Squares
TSS: Total Sum of Squares

Input:

```
model_old = lm(OR~A1+A2+A3+A4)

summary(model_old)
```

Output:

The value of R^2 came out to be **0.9056**

Interpretation:

It implies that approximately 90% of the total variance in the response is explained by the linear regression of OR on A1, A2, A3 and A4.

However, there is a basic limitation in R^2. It assumes that every single explanatory variable that has been included in the model explains the variation of the response variable. The value of R^2 increases with every explanatory variable added to a model. As R^2 always increases and never decreases, it can be a better fit with more terms we add to a model and that can be completely misleading. Hence, we will use here another goodness of fit criteria known as Adjusted R^2. The advantage of Adjusted R^2 is that it only tells us the percentage of variation explained by those explanatory variables that actually affect the response variable.

The mathematical formula of Adjusted R^2 is given by

$$Adjusted\ R^2 = 1 - \frac{(n-1)}{(n-p)} * (1-R^2)$$

where R^2: Coefficient of Determination
n: Number of observations
p: number of parameters

Input:

```
model_old = lm(OR~A1+A2+A3+A4)

summary(model_old)
```

Output:

The value of Adjusted R^2 came out to be **0.9007**

Interpretation:
It implies that approximately 90% of the total variance in the response is explained by those explanatory variables that actually affect the response variable.
Overall, we can interpret that our model is moderately good.

TESTING FOR THE SIGNIFICANCE OF MODEL AND MODEL PARAMETERS

As the residual terms of our linear regression model satisfies the normality assumption**,** two types of hypothesis tests can be carried out. The tests are as follows:-

1). Test for the significance of the regression model: This test is used to check the significance of the regression model.
2). t test: This test is used to check the significance of the individual regression coefficients of the regression model**.**

First we will define them and then we will check them simultaneously.

Test for the significance of the regression model

The test is used to check if there exists a linear statistical relationship between the response variable and at least one of the explanatory variables.
To test
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against H_1: Not H_0.

The test statistic is given by

$$F_0 = \frac{MSR}{MSE}$$

where MSR is the regression mean square and MSE is the error mean square.
Now, MSR = SSR/k and MSE = SSE/ n – (k+1), where SSR and SSE respectively denotes the sum of squares due to regression and error sum of squares. Here, ‘n’ denotes the total number of observations and ‘k’ denotes the total number of explanatory variables.
Under H_0,

$$F_0 \sim F_{k,n-(k+1)}$$

Test for the significance of the individual parameters

The t-test is carried out to check the significance of individual regression coefficients in the multiple linear regression model. A regression model becomes more effective and useful if a significant variable is added to the model while the regression model loses its importance if an insignificant variable is added to it.

To test
H_0: $\beta_j = 0$ against H_1: Not H_0, j = 1(1) 4.

The test statistics is given by

$$t_0 = \hat{\beta}_j \,/\, SE\ (\hat{\beta}_j)$$

where $\hat{\beta}_j$ is the estimated value of β_j and SE ($\hat{\beta}_j$) is the Standard Error of $\hat{\beta}_j$.

Under H_0,

$$t_0 \sim t_{n-(k+1)}$$

where, ‘n’ denotes the total number of observations and ‘k’ denotes the total number of explanatory variables.

Our chosen level of significance is 0.05.

Input:

```
model_old = lm(OR~A1+A2+A3+A4)
summary(model_old)
```

Output:

```
Coefficients:
                  Estimate    Std. Error     t value      Pr(>|t|)
(Intercept)        95.853150    1.224877      78.255       < 2e-16
    A1           -0.092821     0.005235     -17.729       < 2e-16
    A2           -0.126798     0.032157     -3.943         0.000176
    A3           -0.025381     0.013971     -1.817         0.073160
    A4            1.967603     0.324573      6.062       4.65e-08
```

Interpretation:

It is seen that the variables A1, A2 and A4 are significant whereas A3 is not significant. Since at least one of the explanatory variables is significant, hence, our regression model is significant.

INTERPRETATION OF THE PARAMETERS

It is important to interpret the parameters in a model since it reflects the relative importance of the predictors in the definition of the model itself.

i). Interpretation of β_1 (the parameter associated with A1)
We see that the estimated value of **β_1** is -0.092821**.** If A1 increases by 1, then, OR will decrease by the amount -0.092821, keeping all the other explanatory variables fixed. Also, the p-value came out to be very small indicating that the parameter is significant.

ii). Interpretation of β_2 (the parameter associated with A2)
We see that the estimated value of **β_2** is -0.126798**.** If A2 increases by 1, then, OR will decrease by the amount -0.126798, keeping all the other explanatory variables fixed. Also, the p-value came out to be very small indicating that the parameter is significant.
iii). Interpretation of β_3 (the parameter associated with A3)
We see that the estimated value of **β_3** is -0.025381**.** If A3 increases by 1, then, OR will decrease by the amount -0.025381, keeping all the other explanatory variables fixed. Also, the p-value came out to be 0.07 indicating that the parameter is insignificant.

iv). Interpretation of β_4 (the parameter associated with A4)
We see that the estimated value of **β_4** is 1.967603**.** If A4 increases by 1, then, OR will increase by the amount 1.967603, keeping all the other explanatory variables fixed. Also, the p-value came out to be very small indicating that the parameter is significant.

In the following section, we deal with the introduction, identification and remedial measures to be taken for the presence of outliers, high-leverage points and high influential points.

OUTLIERS, HIGH-LEVERAGE AND HIGH-INFLUENTIAL POINTS

Two types of outliers are observed in regression analysis:

First, there may be a significant difference between the explanatory vector $\mathbf{x_i}$ and center of the **x**-data i.e., the point might be remote in p-dimensional space occupied by the rows of X matrix. These points are referred to as *High-Leverage Points.*

Second, there may be a significant difference between the response OR_i and its predicted mean $\mathbf{x_i}^T\boldsymbol{\beta}$, which are referred to as *Outliers*.

Sometimes a small subset of data exerts a disproportionate influence on the model coefficients and properties. In an extreme case, the parameter estimates may depend more on the influential subset of points than on the majority of the data. This is an undesirable situation.

A regression model has to be a representative of all the sample observations and not only of a few. So, we would like to find these influential points and asses their impact on the model.

Our main goal is to identify *High-Influential Points* which can affect our regression line drastically. Mild outliers that are not high leverage points may not be influential points, but if the corresponding point is also of a high-leverage, then they can have a huge impact on the regression quantities. Also, if the point is a high-leverage point it can mask potential outliers in the data. So, it is important to identify high-leverage points and outliers.

Identifying High-Leverage Points

The *Hat matrix* (also known as the *Projection matrix*) diagonal is a standardized measure of the distance of the i^{th}(i = 1, 2, …, 82) observation from the center (or centroid) of the row-space of X (R(X)). Thus, large hat diagonals reveal observations that are potentially influential as they are remote in the R(X) from the rest of the sample.

$$h_{ii} = h_i = \frac{1}{n} + \left(\frac{1}{n-1}\right) MD_i \quad \forall\, i = 1, 2, \ldots, 82$$

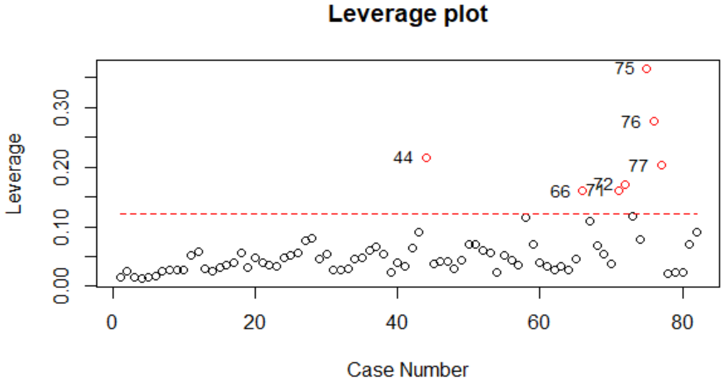
where MD_i is the Mahalanobis Distance, given by, $MD_i = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x})$

If $h_i > {}^{2p}/_{n}$, then the point is far enough from rest of the data to be considered as a high-leverage point.

Input:

```
H = X%*%solve(t(X)%*%X)%*%t(X)        #hat matrix
hii = numeric(length=0)
for (i in 1:n) { hii[i] = H[i,i]}     #hat matrix diagonal
cases_1 = numeric(length=0)           #cases where hi > 2p/n
for (i in 1:n) {  if (hii[i]>2*p/n) {cases_1[length(cases_1)+1] = i} }
plot(hii)
lines(x=1:82,y=rep(2*p/n,82),col="red",lty=2)
points(x=cases_1,hii[cases_1],col="red",text(cases_1,hii[cases_1]),
       labels=as.character(cases_1),pos=2,ces=0.9)
```

Output:



We can see that the samples indexed 44, 66, 71, 72, 75, 76 and 77 are high-leverage points.

Studentized Residuals

Residuals are defined as the difference between the true response variable and it's predicted value.

$$\boldsymbol{e} = (I - H)\boldsymbol{OR}, \qquad with\ E(\boldsymbol{e}) = \boldsymbol{0}\ and\ Var(\boldsymbol{e}) = \sigma^2(I - H)$$

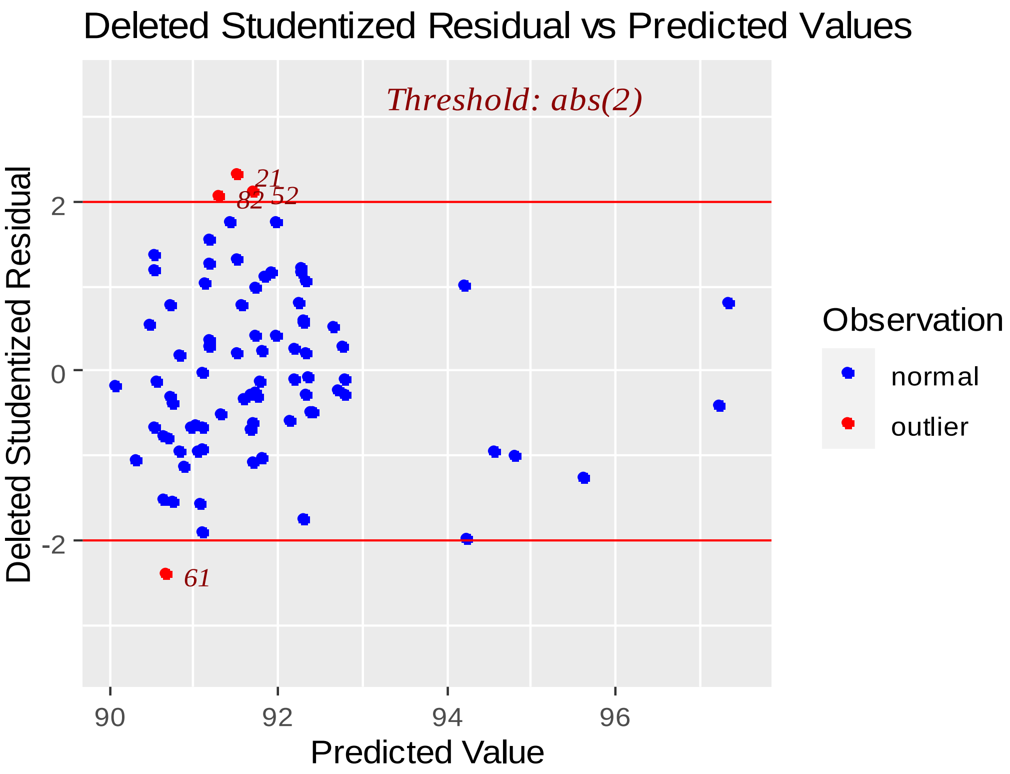
The absolute value of externally studentized residuals $|t_i|$, if greater than 2, implies that the i^{th} point is an outlier.

$$t_i = \frac{e_i}{S(i).\sqrt{1 - h_i}}\ , where\ S(i)^2 = \left(\frac{n-p}{n-p-1}\right)S^2 - \frac{{e_i}^2}{(n-p-1)(1-h_i)}\ \forall\ i = 1, 2, \dots, 82$$

Input:`Si2 = (1/(n-p-1))*((n-p)*S2-(e^2)/(1-hii))`

```
ti = e/((Si2*(1-hii))^0.5)
cases_2 = numeric(length=0)
for (i in 1:n) {  if (abs(ti[i])>2) {cases_2[length(cases_2)+1] = i} }
ols_plot_resid_stud_fit(model_old)
```

Output:



We can see that the samples indexed 21, 52, 61 & 82 are outliers.

Difference in Fitted Values

The change in ithfitted value before and after deletion of the ith indexed data point is called DFFITS.This is called *leave-one-out* diagnostics.

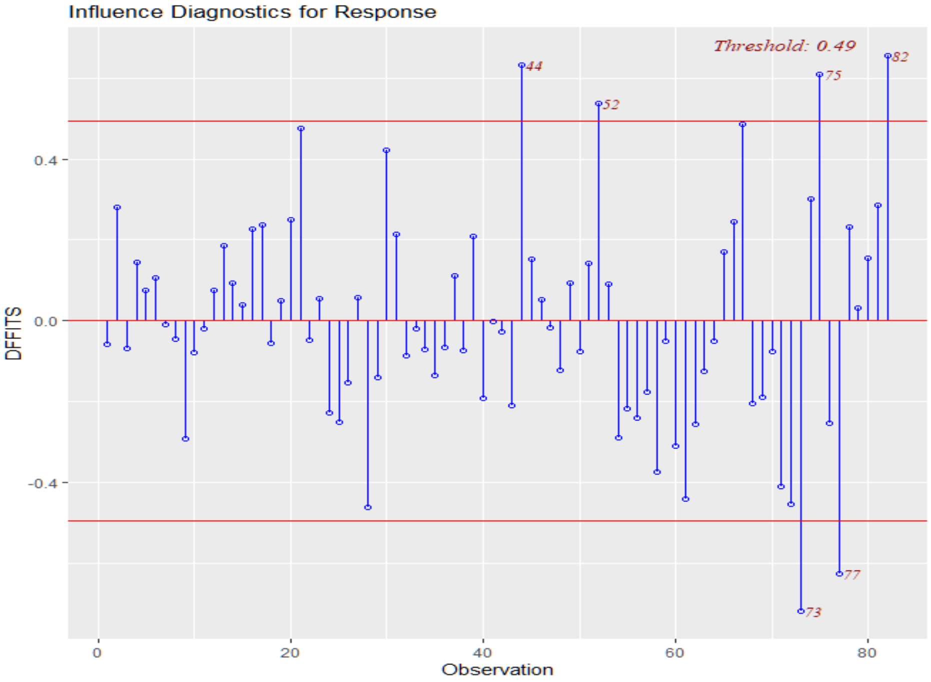
$$DFFITSS_i = t_i \left(\frac{h_i}{1-h_1}\right)^{1/2} \quad \forall\, i = 1, 2, \dots, 82$$

If the ith data point is not an outlier and do not have high-leverage, then $|DFFITSS_i| < 2\sqrt{\frac{p}{n}}$,otherwise, the ith data point may have high influence over regression line.

Input:

```
library("olsrr")        #install olsrr package if it is not yet installed
ols_plot_dfbetas(model_old)
ols_plot_dffits(model_old)
cases_3 = numeric(length=0)
for(i                                                                  in
1:n){if(abs(dffits(model_old)[i])>2*((p/n)^0.5)){cases_3[length(cases_3)+1]=i}}
```

Output:



Interpretation: The samples indexed 44, 52, 73, 75, 77 & 82 have high DFFITS, implying high influential points.

Covariance Ratio

High influential points can be identified using the COVRATIO. It is given by:

$$COVRATIO_i = \frac{\det\{S(i)^2\left(X(i)^T X(i)\right)^{-1}\}}{\det\{S^2(X'X)^{-1}\}} = \left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p}\right)^{-p} (1-h_i)^{-1} \forall i = 1,2,\ldots,82$$

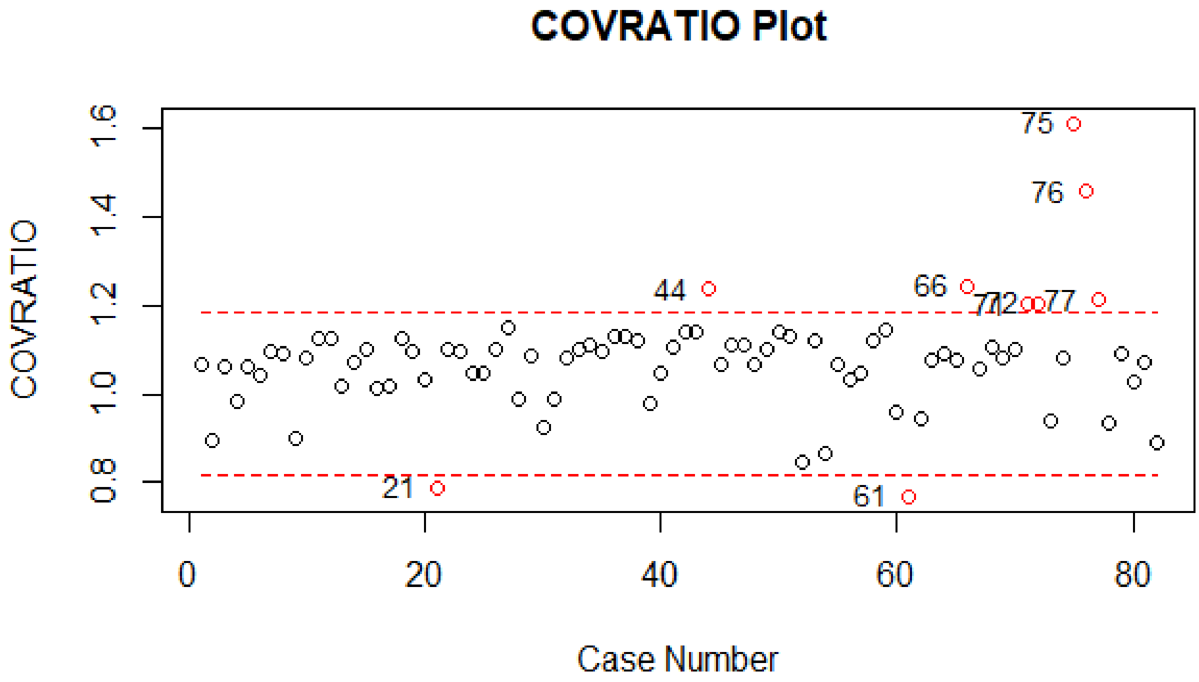
If ith case has$|COVRATIO_i - 1| > \frac{3p}{n}$, then it is considered to have high influence.

Input:

```
cr = covratio(model_old)      #covratio
```

```
cases_4 = numeric(length=0)
for (i in 1:n) {if (abs(cr[i]-1)>3*(p/n)) {cases_4[length(cases_4)+1] = i}}
plot(1:82,cr,xlab = "Case Number",ylab = "COVRATIO",main = "COVRATIO Plot")
lines(x = 1:82,y = rep(1+3*p/n,82),col="red",lty=2)
lines(x = 1:82,y = rep(1-3*p/n,82),col="red",lty=2)
points(cases_4,cr[cases_4],col="red",text(cases_4,cr[cases_4],
       labels=as.character(cases_4),pos=2,cex = 0.9))
```

<u>Output</u>:



<u>Interpretation:</u> Cases indexed 21, 44, 61, 66, 71, 72, 75, 76 & 77 are considered to have high influence for the above plot as they fall outside the threshold limit.

Cook's Distance

The Cook's distance statistics denoted as, Cook's D-statistic is a measure of the distance between the least squares estimate based on all n observations and the estimate obtained by deleting the ith point.

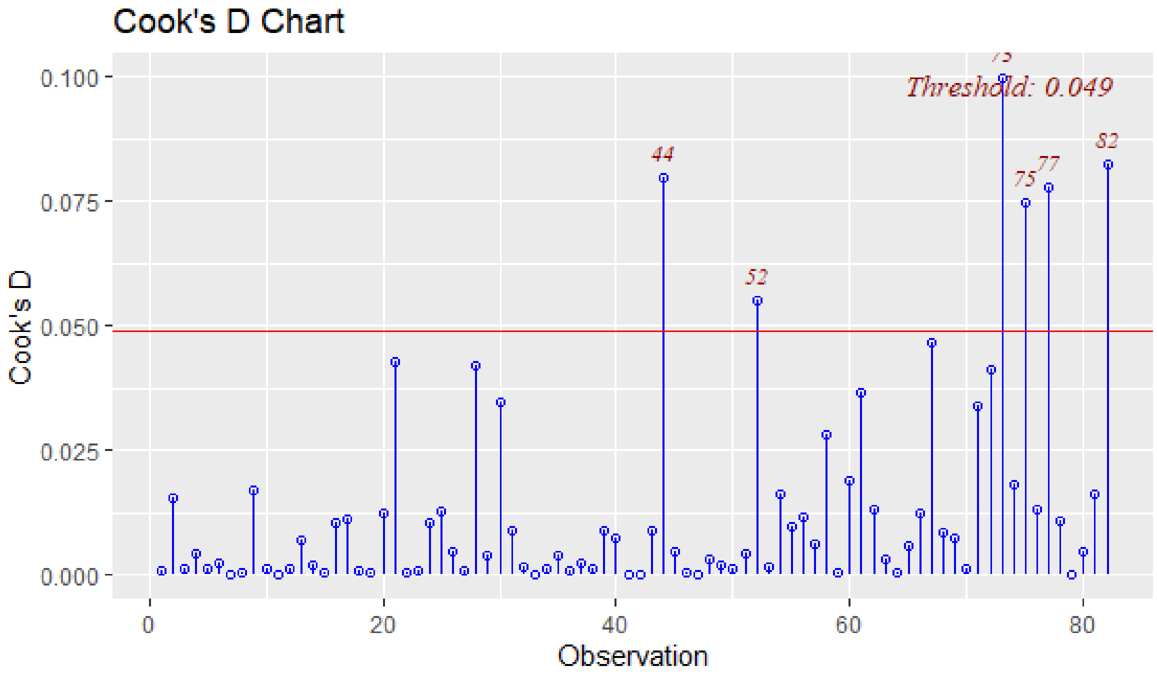
$$D_i = \frac{(\widehat{\boldsymbol{\beta}}(\boldsymbol{i}) - \widehat{\boldsymbol{\beta}})'X'X(\widehat{\boldsymbol{\beta}}(\boldsymbol{i}) - \widehat{\boldsymbol{\beta}})}{p * S^2} \ \forall\, i = 1, 2, \dots, 82$$

A point will have a large Cook's D if it has a large Studentized residual or is a high-leverage point. The cut-off for Cook's D is 4/n if D_i>4/n then it might be high influence point.

Input:

```
Di = cooks.distance(model_old)       #cook's distance
cases_5 = numeric(length=0)
for (i in 1:n) { if (Di[i]>4/n) {cases_5[length(cases_5)+1] = i} }
ols_plot_cooksd_chart(model_old)
```

Output:



Using Cook's-D chart, the cases indexed 44, 52, 73, 75, 77& 82 are highly influential. All cases indices which are suspicious, i.e., that are most likely to have high-leverage or are outliers or highly influential are given in the following table along with the frequency with which they are identified. From this table, we will choose a subset and perform *leave-many-out* diagnostics and run an outlier test.

CaseIndex	21	44	52	61	66	71	72	73	75	76	77	82
Frequency	2	4	3	2	2	2	2	2	4	2	4	3

Outlier Test

We choose a subset of these 12 identified cases and run an outlier F-test. It is only sensible to choose points which occurred with high frequency, and also include some points which have very high leverage or very far from the mean fitted response. After multiple hit and trial method, we chose the best possible subset of potential outliers according to our previous calculations. So, let us test for the subset (44, 52, 71, 72, 73, 75, 76, 77, 82). We will use *outlier shift model* for testing our hypothesis that the above-mentioned points are outliers.

$\boldsymbol{OR} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$ where Z is a matrix of the form $Z = \begin{bmatrix} 0 \\ I_k \end{bmatrix}$ and $\boldsymbol{\gamma}$ is a k-vector containing the shifts of possibly outlying observations.

Let the hat matrix be parted as $H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$, where H_{22} is a matrix of order k x k, and let the residual vector **e** = (I_n-H)**OR** be parted as confirmable with H, as **e** = ($\mathbf{e_1}^T$, $\mathbf{e_2}^T$)^T.

For testing the hypothesis H_0: $\boldsymbol{\gamma}$= **0**, the F statistic is given by:

$$F = \frac{\boldsymbol{e_2}^T (I_k - H_{22})^{-1} \boldsymbol{e_2} / k}{[RSS - \boldsymbol{e_2}^T (I_k - H_{22})^{-1} \boldsymbol{e_2}] / (n-p-k)} \sim F_{(k,n-p-k)} \; under \; H_0$$

Input:

```
all_cases = c(cases_1, cases_2, cases_3, cases_4, cases_5)
suspicious_cases = c(44, 52, 71, 72, 73, 75, 76, 77, 82)
k = length(suspicious_cases)
tmp1 = X[-suspicious_cases,] ;     tmp2 = X[suspicious_cases,]
newX = rbind(tmp1,tmp2)
tmp3 = Y[-suspicious_cases]   ;     tmp4 = Y[suspicious_cases]
newY = c(tmp3,tmp4)
newH = newX%*%solve(t(newX)%*%newX)%*%t(newX)
H22 = newH[(n-k+1):n,(n-k+1):n]
newe = (diag(n)-newH)%*%newY
e2 = as.matrix(newe[(n-k+1):n])
num = (t(e2)%*%solve(diag(k)-H22)%*%e2)
den = (RSS-(t(e2)%*%solve(diag(k)-H22)%*%e2))


outF = ((n-p-k)/k)*(num/den)
newTabF = qf(0.95,k,n-p-k)
if(outF>newTabF){print("The suspicious points are outliers by F-test")}
else{print("The suspicious points are not outliers by F-test")}
```

Output:

```
"The suspicious points are outliers by F-test"
```

Interpretation: As the outlier test, at 5% level of significance, shows evidence that the selected subset is actually influencing the regression model (or at least of high influence), we can remove these points from the original data set and then get a better model fit, that is, better estimates of regression coefficients, a relatively reliable adjusted R^2, less variance of estimates, etc.

The updated dataset is now used (TABLE 2, Pg 63)

Here, AM1: Updated Amount of material 1;
AM2: Updated Amount of material 2;
AM3: Updated Amount of material 3;
MCR: Updated Manufacturing condition rating;
OR_N: Updated octane rating

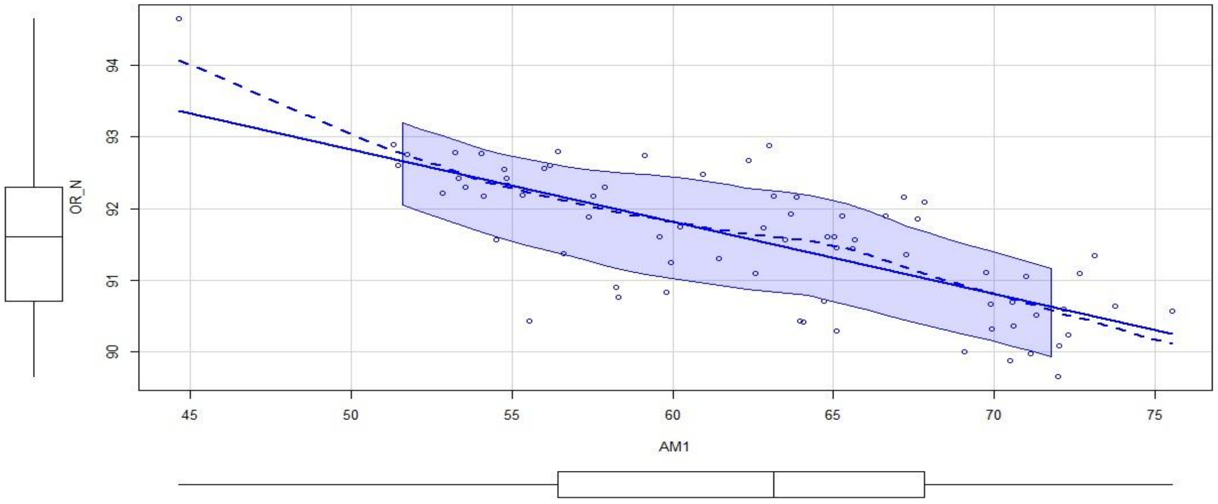
SCATTER PLOTS

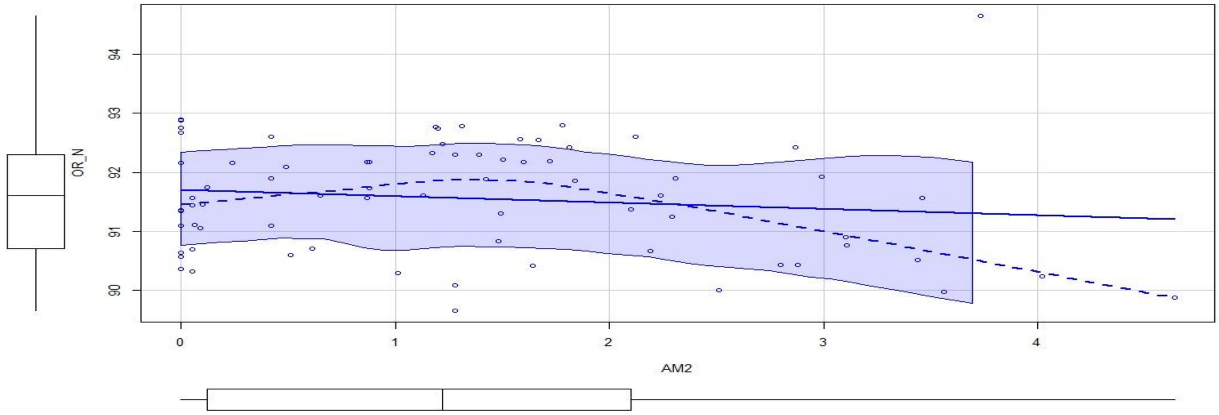
Here, we have plotted the response with the explanatory variables(one at a time).

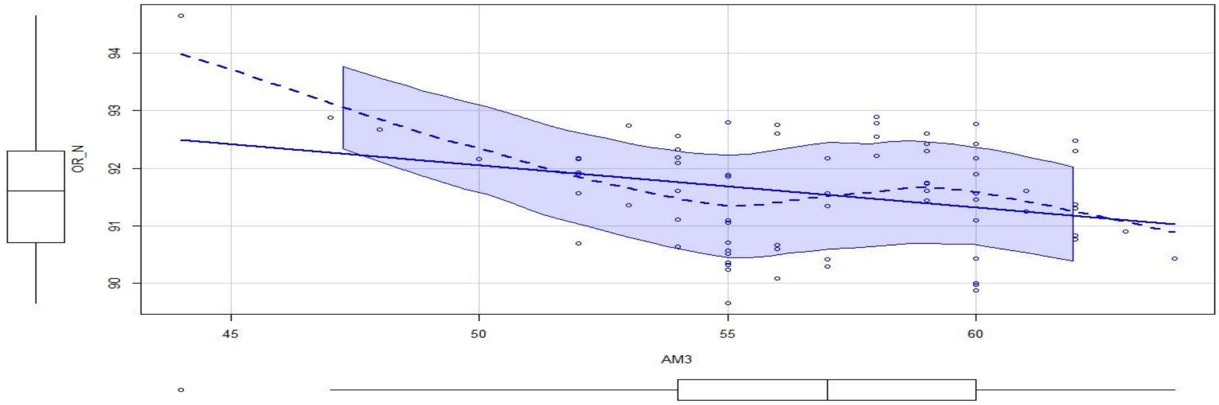
Input:

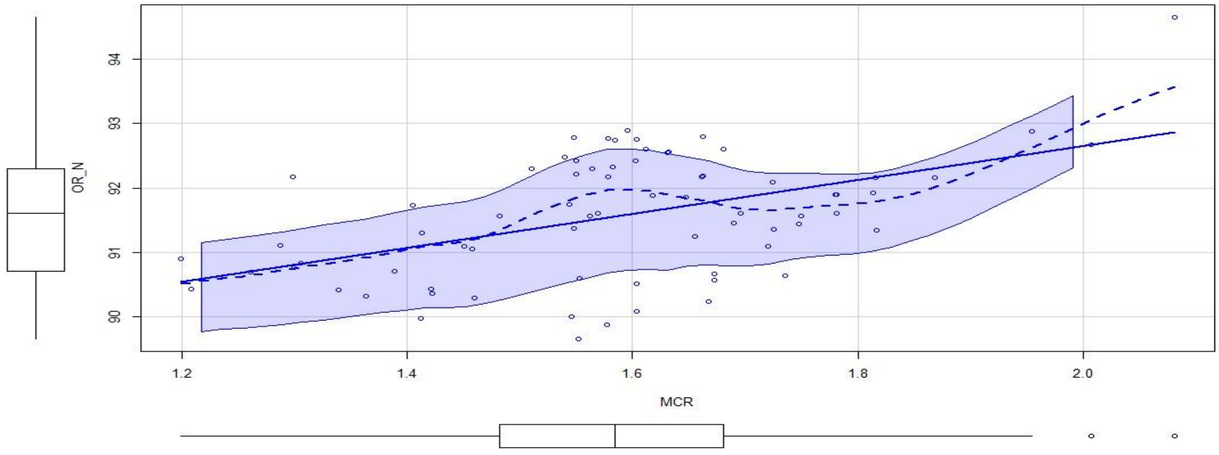
```
library(car)
data=read.csv(file="E:/ISI/REGRESSION TECHNIQUES/PROJECT.csv")
AM1=data1[,2]                    #Extracting the variable
AM2=data1[,3]                    #Extracting the variable
AM3=data1[,4]                    #Extracting the variable
MCR=data1[,5]                    #Extracting the variable
OR_N=data1[,6]                    #Extracting the variable
scatterplot(OR_N~AM1)              #Plotting
scatterplot(OR_N~AM2)              #Plotting
scatterplot(OR_N~AM3)              #Plotting
scatterplot(OR_N~MCR)              #Plotting
```

Output:









Interpretation: From the above plots we can see that after the removal of outliers, the linear relationship between the response and all the explanatory variables(taken one at a time) has become more prominent.

LEAST SQUARES FIT

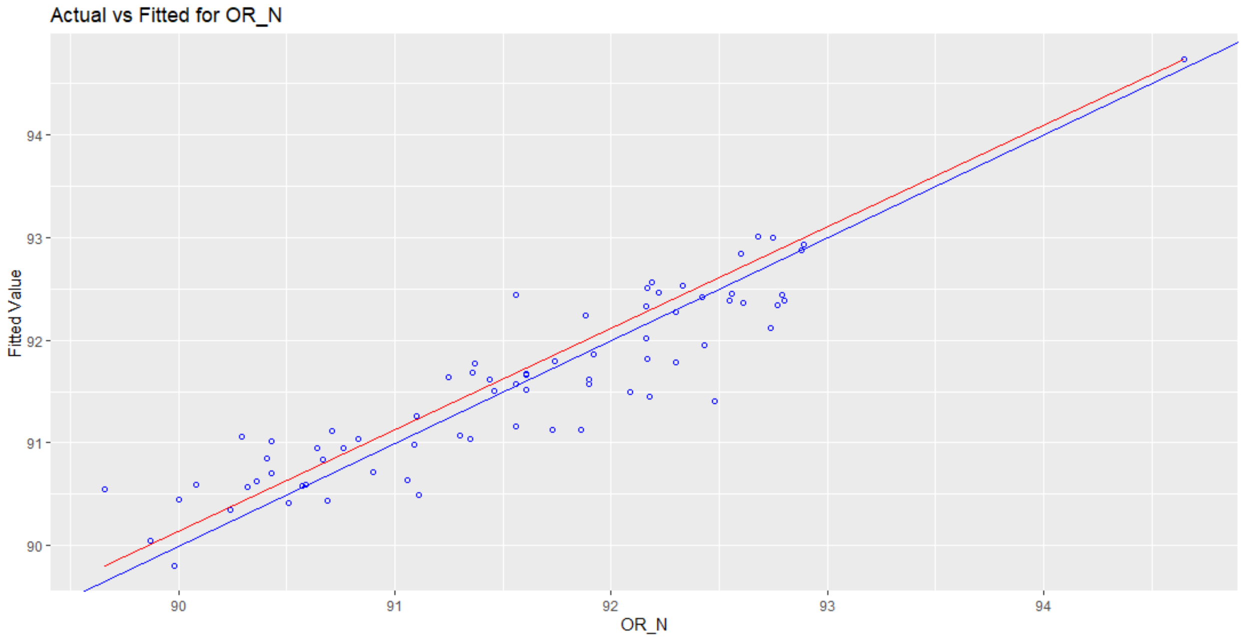
We fit a multiple linear regression model to the given data with an intercept term, including all explanatory variables, that is,

$$OR\_N = \alpha_0 + \alpha_1 * AM1 + \alpha_2 * AM2 + \alpha_3 * AM3 + \alpha_4 * MCR + \varepsilon'$$

The number of observations (=73) is greater than the number of parameters to be estimated (= 5), which are the α_is (for i = 0, 1, 2, 3, 4). So, one of the assumptions is satisfied already and we do not need to check for this later.

Input:

```
model_new=lm(OR_N~AM1+AM2+AM3+MCR)     #New fitted model
summary(model_new)
OR_hat= 98.045624-0.108919*AM1--0.143369*AM2 -
0.046249*AM3+1.976704*MCR                        # Fitted Values
ols_plot_obs_fit(model_new)
```



In the y-axis, we have plotted the predicted values of the response obtained using the above stated model and in the x-axis, we have plotted the observed values of the response. We have also plotted a y=x line.

Interpretation: From the plot, it can be seen that most of the points line around the y=x line indicating that the fitted values is pretty near to the observed values. Hence, it meant that the fitted linear regression model is moderately good.

NORMALITY ASSUMPTION AFTER THE OUTLIERS' REMOVAL

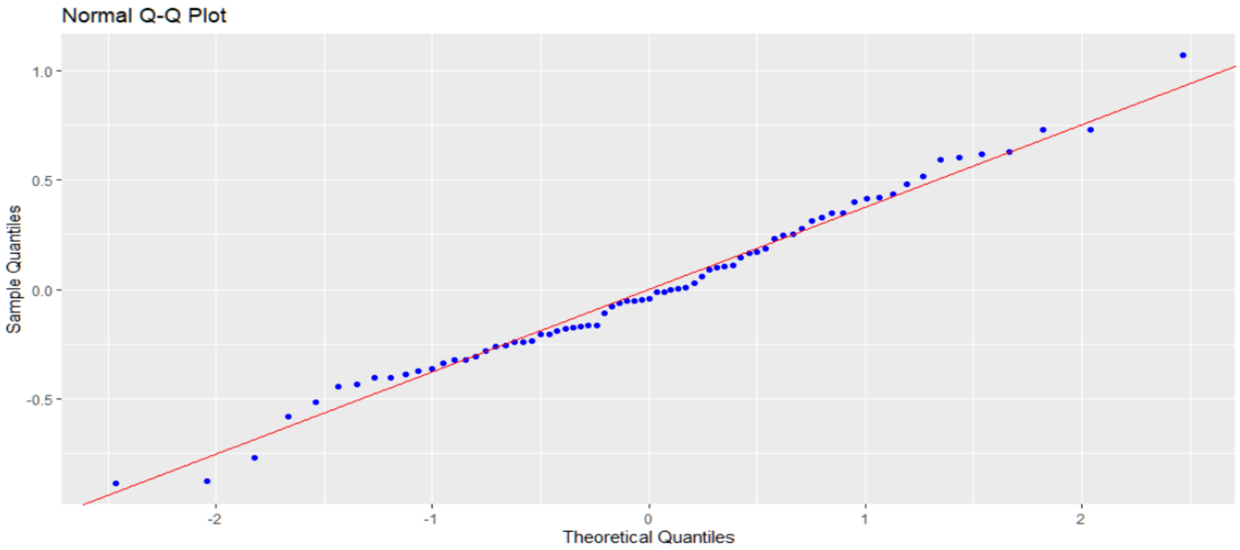
We use the same techniques as above to test the hypothesis that the errors follow normal distribution. Recall that the assumption was satisfied even before the removal of potential outlier points. The assumption is still valid after their removal as is evident from the following methods.

Normal Probability Plot

Input:

```
ols_plot_resid_qq(model_new)#Remember to use the updated data as in the above section
```

Output:



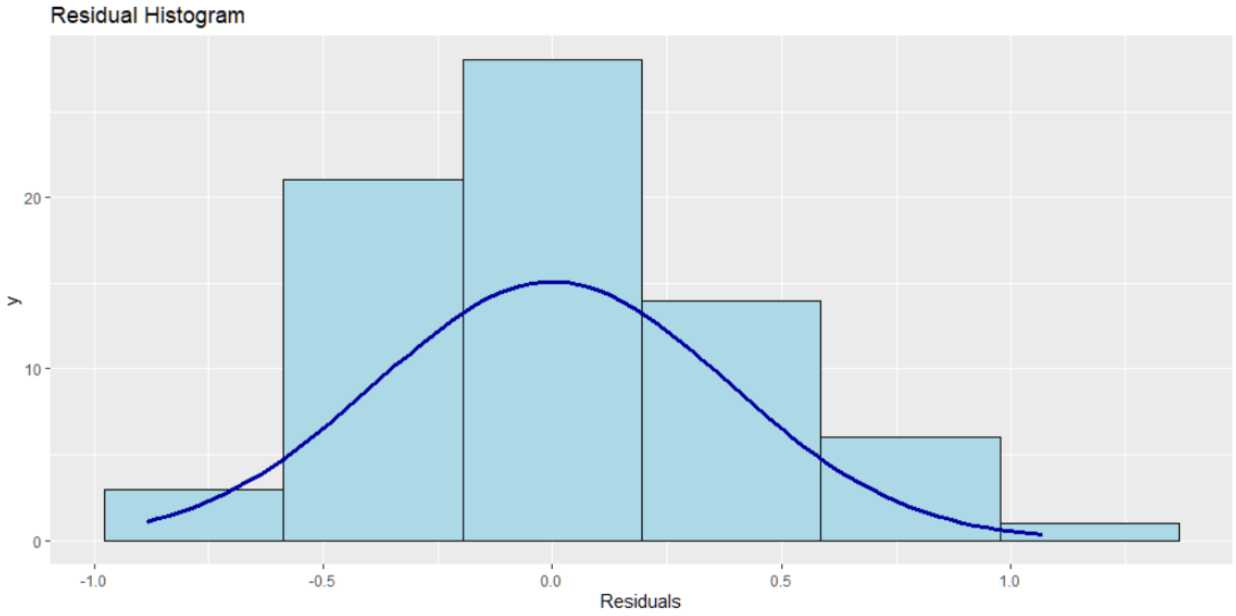
Interpretation: The samples quantiles can still be seen very close to the theoretical quantiles of normal distribution, suggesting that they are still normally distributed. But we will also use the other methods to strengthen our claim.

Histogram and Density Plot

Input:

```
ols_plot_resid_hist(model_new)
```

Output:



Interpretation: This plot shows that the residuals' histogram matches with the density of normal distribution even better than it did before revealing the fact that the errors actually follow normal distribution.

Finally, proceeding to the *Shapiro-Wilk* test, a conclusive statement can be given about the claim.

Shapiro-Wilk Test

Input:

```
s = shapiro.test(e)          #residuals of the updated model
if(s$p.value>0.05)
{print("We do not reject the null hypothesis at 5% level of significance")}
else{print("We reject H0 at 5% level of significance")}
```

Output:

```
      Shapiro-Wilk normality test
data:  e
W = 0.98783, p-value = 0.7107
[1] "We do not reject the null hypothesis at 5% level of significance"
```

Interpretation: This test resulted exactly same as before, which proves our claim that the errors follow normal distribution.

Now the only remaining assumption left to test is the homoscedasticity of errors. If the errors do not have constant variance, we have to make relevant changes to the response variable *OR*, identify if there is any functional relationship between the response variable and the errors.

HETEROSCEDASTICITY AFTER THE REMOVAL OF OUTLIERS

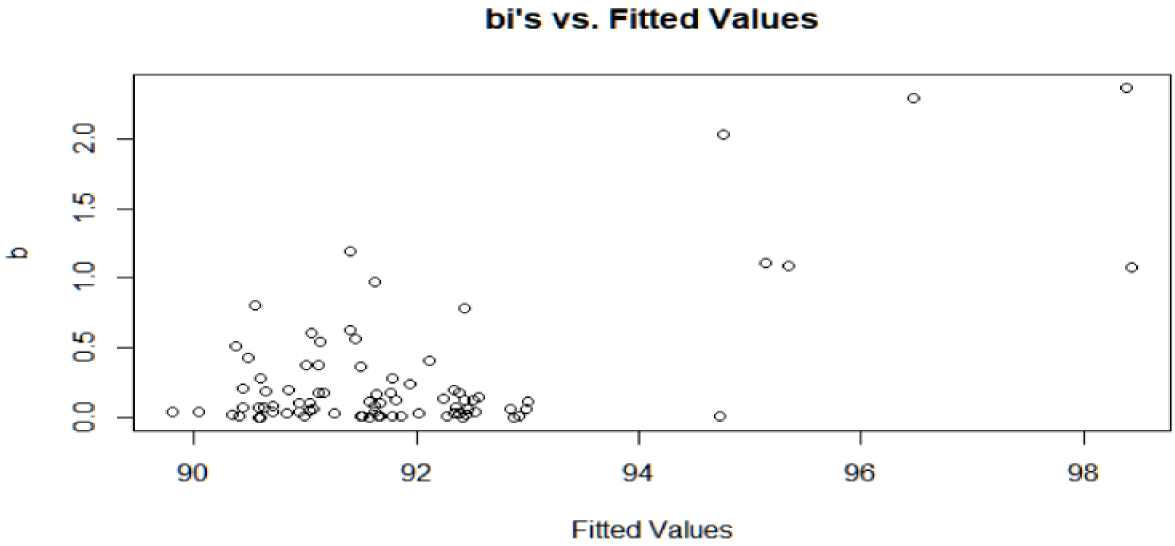
The potential outliers are removed and we have ensured that there is no multicollinearity in the X-data, as well as the errors are normally distributed. We check for the homoscedasticity of the errors in this section after the removal the removal of potential outliers. Recall that the errors were homoscedastic earlier. Using the same techniques as used earlier to detect the heteroscedasticity:

Plot of b_is’ vs. Fitted Values

Input:

```
b = e^2/(1-hii)    #Remember to use the updated data without outliers
plot(Ycap,b,main = "bi's vs. Fitted Values",xlab = "Fitted Values",ylab =
"b")
```

Output:



Interpretation: We can see that the plot above is not wedge shaped, so we can say that there is no heteroscedasticity with respect to means.

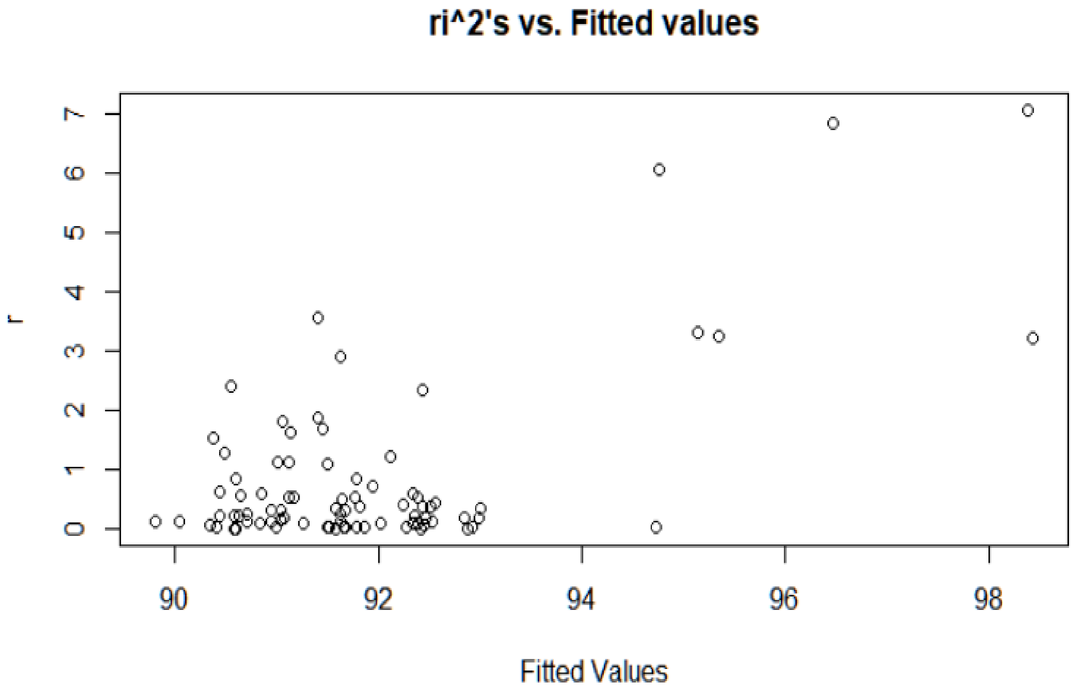
Plot of r_i^2 vs. Fitted values

Input:

```
r = e/sqrt(S2*(1-hii))
```

```
plot(Ycap,r^2,main="ri^2's vs. Fitted values",xlab="Fitted Values",ylab="r")
```

Output:



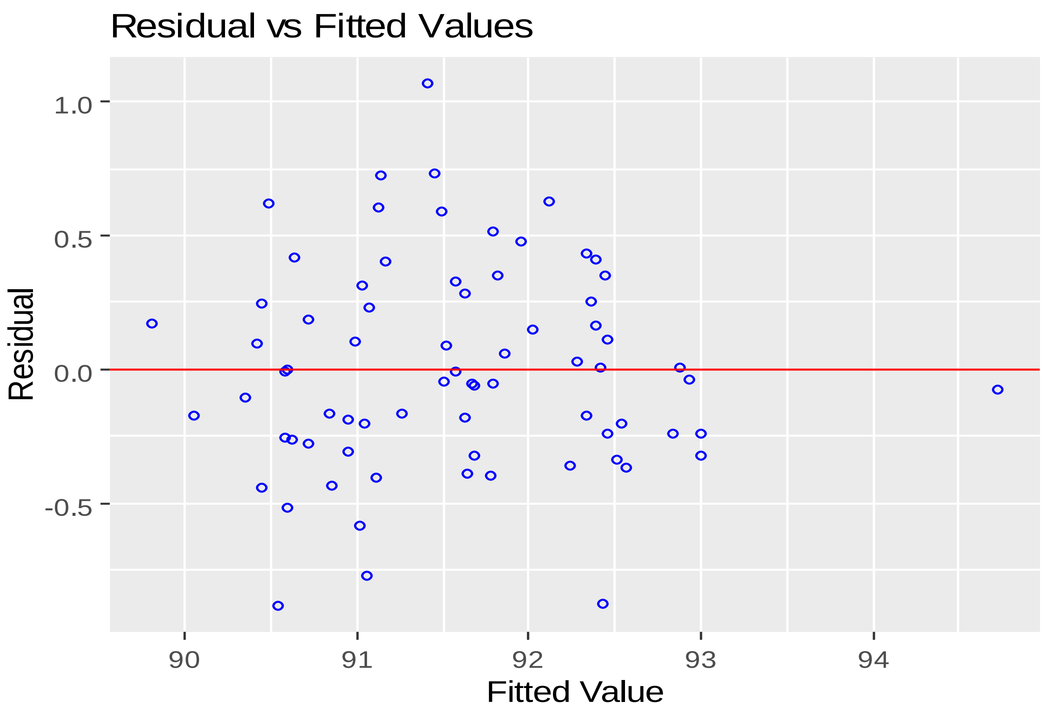
As plotting r_i^2s' instead of b_is' is equivalent to only changing the scale of the plot, so the same comment as the previous plot works here.

Residuals vs. Fitted Values

Input:

```
ols_plot_resid_fit(model_new)
```

Output:



There are no patterns seen in the above plots suggesting that there is no significant heteroscedasticity. But by carrying out the BP test, this statement can be strengthened.

Breusch-Pagan Test

Input:

```
install.packages("lmtest")     #install this library is not installed yet
library(lmtest)                #Load the lmtest library
bptest(LSE)                    #Command to run the BP Test
```

Output:

```
	studentized Breusch-Pagan test

data:  LSE
BP = 4.2423, df = 4, p-value = 0.3742
```

Interpretation: The BP test’s p-value after the removal of potential outlier points is coming out to be 0.3742 which is greater than 0.05. So, we do not reject the null hypothesis that the errors are homoscedastic. Hence, at 5% level of significance, we can conclude that the errors are homoscedastic.

AUTOCORRELATION OF ERRORS AFTER REMOVAL OF OUTLIERS

We use the Durbin Watson test again.

Input:

```
library(DescTools)
d=DurbinWatsonTest(model_new)$statistic
d
```

Output:

```
1.548106
```

Interpretation: Since the value of d lies in between d_u = 1.53 and d_u = 1.74, hence this test is inconclusive.
Hence, we go for another test known as Breusch Godfrey Test.

Breusch Godfrey Test

We assume that the error term 'u_t' follows the p^{th}-order autoregressive AR (q) scheme, that is,

$$u_t = \rho_1 * u_{t-1} + \rho_2 * u_{t-2} + \ldots\ldots\ldots + \rho_q * u_{t-q} + \varepsilon_t$$, where 'ε_t' is the error term.

Basic Test Procedures

Step-1: The first step is to estimate the new parameters using the method of OLS and obtaining the residuals '$\hat{u}_t$'.

Step-2: The second step is to regress 'u_t' on the 4 explanatory variables along with $\hat{u}_{t-1}$, $\hat{u}_{t-2}$,….. $\hat{u}_{t-q}$, where the latter are the lagged values of the residuals that have been estimated in Step-1.

Hence, the required regression is

$\hat{u}_t = \hat{\rho_1} * \hat{u}_{t-1} + \hat{\rho_2} * \hat{u}_{t-2} + \ldots\ldots\ldots + \hat{\rho_q} * \hat{u}_{t-q} + \varepsilon_t$, t = 1(1)27, where '$\hat{\rho_1}$',…..,'$\hat{\rho_q}$' are the constants involved in the regression.

To test H_0: $\rho_1 = \rho_2 = \ldots\ldots = \rho_q = 0$ against H_1: Not H_0, that is, to test

H_0: No Autocorrelation of any order is present against H_1: Not H_0.

The test statistic is given by

$$(n - q) * R^2$$

where R^2: Coefficient of Determination

Under H_0, the distribution of the test statistic is χ^2_q asymptotically, that is Chi-square distribution with degrees of freedom ‘q’, where ‘q’ denotes the number of lags.

Input:

```
library(lmtest)
bgtest(OR_N~AM1+AM2+AM3+MCR)
```

Output:

```
The p-value came out to be 0.06.
```

Interpretation:

No autocorrelation is present in our dataset.

MULTICOLLINEARITY AFTER THE REMOVAL OF OUTLIERS

We have already seen the situation of multicollinearity before the removal of high-influential, high-leverage points and outliers. In the following section, we check for the presence of multicollinearity after the removal of these points from the dataset.

DETECTION OF MULTICOLLINEARITY

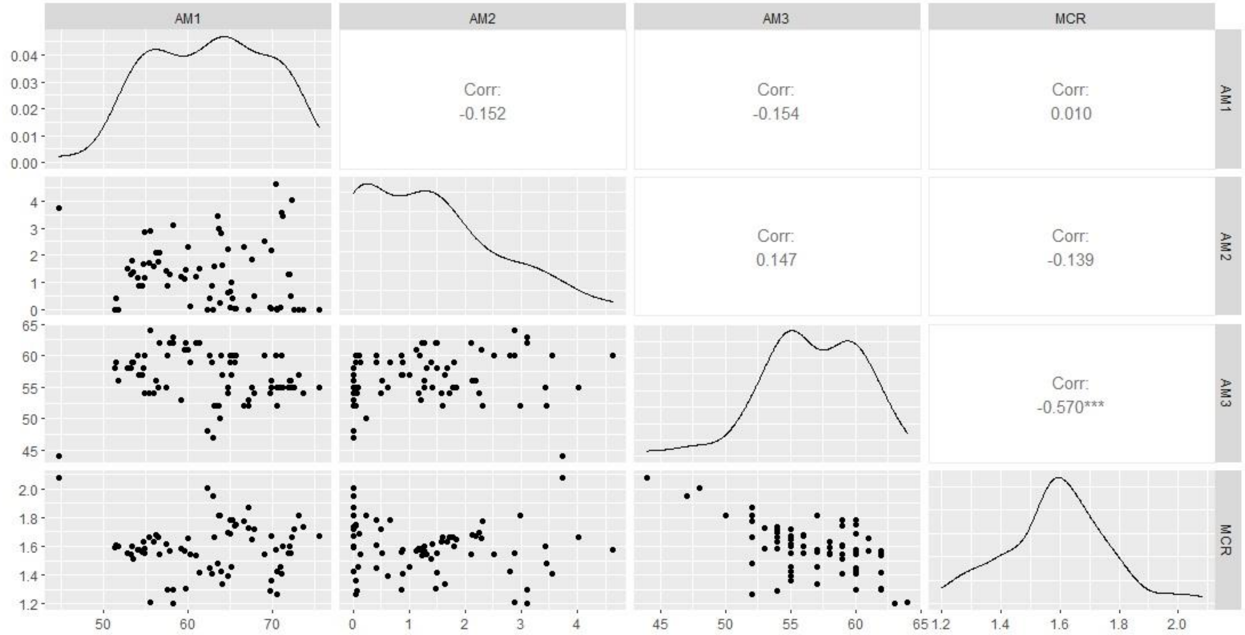
Updating the values of the regression matrix to be the one without the suspicious outlying cases and corresponding OR values, we proceed to check the multicollinearity using the same R codes.

Scatter Plot

Input**:**

```
Data=Data[-c(44,52,71,72,73,75,76,77,82),]   #Removing outliers
from data
ggpairs(Data[,-c(1,6)])       #This generates scatter plot of
all the variables
#This also generates the partial correlation between the
variables
```

Output:



Interpretation: From the above figure, we can observe that there are no high simple correlation coefficients among the explanatory variables. So, it seems that the multicollinearity has a significant decrease after the removal of potential outliers.

Partial Correlation Matrix

Input:

```
pcor(Data[,-c(1,6)],method="pearson")#Calculating Partial Correlation Matrix
```

Output:

```
           AM1         AM2         AM3         AM4
AM1  1.0000000 -0.13932550 -0.17029021 -0.10602218
AM2 -0.1393255  1.00000000  0.05790608 -0.08106639
AM3 -0.1702902  0.05790608  1.00000000 -0.56771904
AM4 -0.1060222 -0.08106639 -0.56771904  1.00000000
```

Interpretation: Now that the partial correlations are significantly low compared to the previous case, we have stronger evidence for the absence of multicollinearity.

Variance Inflation Factors

Input:

```
library(car)
model_new = lm(OR_N~AM1+AM2+AM3+MCR,Data)
vif(model_new)
```

Output:

```
   AM1      AM2      AM3      AM4
1.054207 1.047070 1.535972 1.504650
```

Interpretation: The VIF of each of the explanatory variables is also very less compared to previous computation, which is a good sign providing us an evidence that the multicollinearity is absent.

Condition Number

Input:

```
l = eigen(t(Xstar)%*%Xstar)$values  #use updated Xstar in computation
sqrt(max(l)/min(l))
```

Output:

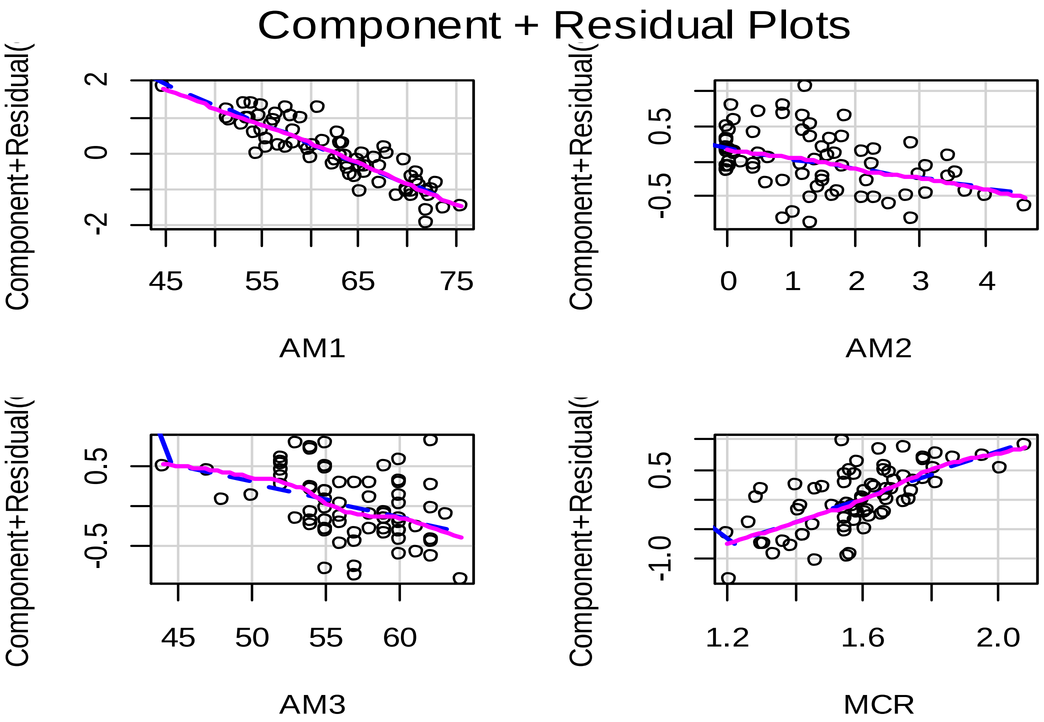
```
[1] 2.015544
```

Interpretation: As the condition number has also lessened, we can conclude that there is no more multicollinearity present in the model after the deletion of potential outliers. So, now our data is clean and free from any potential outliers. We can now proceed to check the assumption of normality, as it is the key-assumption we make to test any hypothesis in linear models.

PARTIAL RESIDUAL PLOTS

Input:

```
>crPlots(model_new)
```



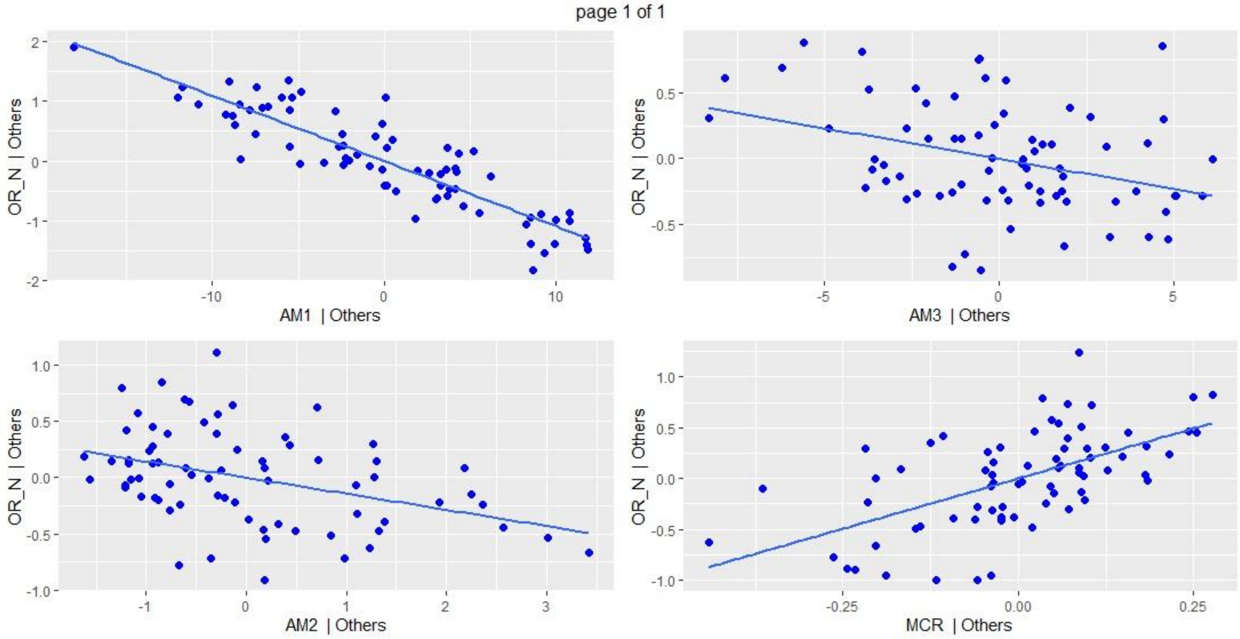
Interpretation: It can be seen that after removing outliers and high influential points, we can observe that the Linearity Assumption have become more clear in the Partial Residual plots.

ADDED VARIABLE PLOT

Input:

```
>ols_plot_added_variable(model_new)
```

Output



Interpretation: It can be seen that after removing outliers and high influential points, we can observe that the Linearity Assumption have become more clear in the Added Variable plots.

Model Selection

We will try to find out the subset of model features which are most appropriate based on several methods.

AIC: The Akaike information criterion(AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.

$$AIC = 2 * (no.of\ estimated\ parameters) - 2 * (Log - Likelihood)$$

BIC: Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to Akaike information criterion (AIC).

$$BIC = (no.of\ estimated\ parameters) * log(size\ of\ sample) - 2 * (Log - Likelihood)$$

Mallow's CP: Mallows' Cp Criterion is a way to assess the fit of a multiple regression model. The technique then compares the full model with a smaller model with "p" parameters and determines how much error is left unexplained by the partial model. Or, more specifically, it estimates the standardized total mean square of estimation for the partial model with the formula

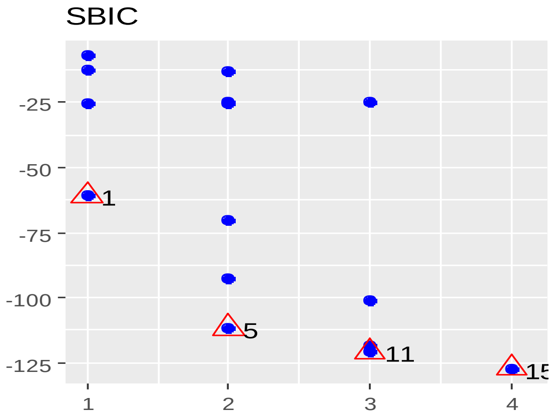
$$C_p = \frac{SS(res)}{p} + 2p - n$$

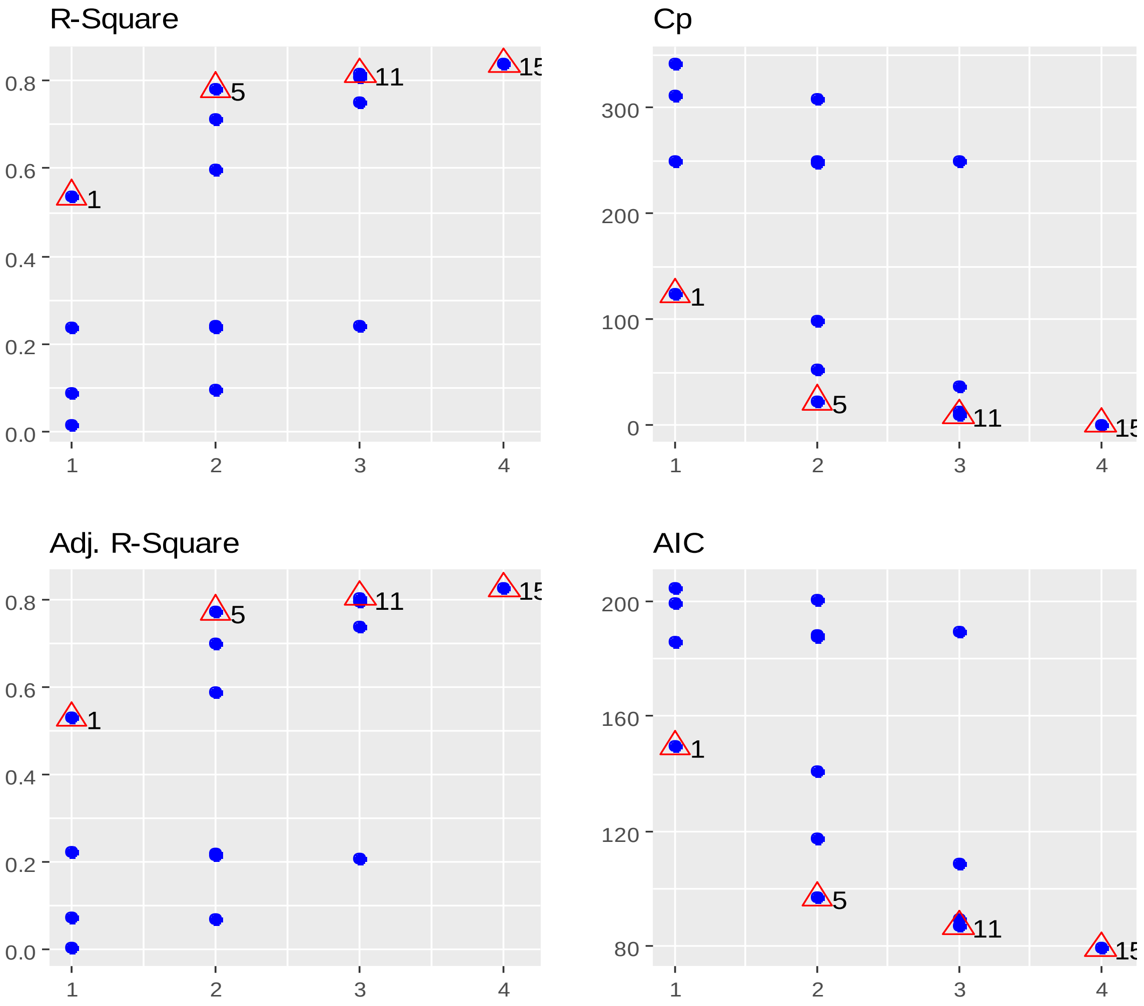
Input:

```
selection=ols_step_all_possible(LSE)
print(selection)
plot(selection)
```

Output:

```
   Index N  Predictors   R-Square Adj. R-Square Mallow's Cp
1      1 1          A1 0.53792132   0.531413172   124.42812
4      2 1          A4 0.23673582   0.225985617   250.50567
3      3 1          A3 0.08902667   0.076196058   312.33736
2      4 1          A2 0.01806726   0.004237222   342.04127
7      5 2       A1 A4 0.78166093   0.775422668    24.39768
6      6 2       A1 A3 0.71106829   0.702813094    53.94806
5      7 2       A1 A2 0.59963311   0.588194057   100.59530
9      8 2       A2 A4 0.24131593   0.219639244   250.58841
10     9 2       A3 A4 0.23739450   0.215605773   252.22994
8     10 2       A2 A3 0.09740702   0.071618650   310.82930
12    11 3    A1 A2 A4 0.81468547   0.806628317    12.57346
13    12 3    A1 A3 A4 0.80774235   0.799383327    15.47988
11    13 3    A1 A2 A3 0.74999630   0.739126577    39.65262
14    14 3    A2 A3 A4 0.24171888   0.208750135   252.41974
15    15 4 A1 A2 A3 A4 0.83755542   0.827999857     5.00000
```





Interpretation: From the above output and plots, we can observe that the Full Model has Lowest AIC, BIC and Mallows CP as well as it has highest R^2 and Adjusted R^2 value. Hence,we conclude that Full model is Better than the other submodels.

The value of R^2 for the final model came out to be **0.8376.**

Interpretation:

It implies that approximately 84% of the total variance in the response is explained by the linear regression of OR_N on AM1, AM2, AM3 and MCR.

Output:

The value of Adjusted R^2 for the final model came out to be **0.828.**

Interpretation:
It implies that approximately 83% of the total variance in the response is explained by the independent variables.

Overall, we can interpret that our model is moderately good.

TESTING FOR THE SIGNIFICANCE OF MODEL AND MODEL PARAMETERS

Test for the significance of the regression model

To test
H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ against H_1: Not H_0, that is, it is equivalent to test

The test statistic is given by

$$F^* = \frac{MSR}{MSE}$$

where MSR is the Regression mean square and MSE is the error mean square of the new model.

Under H_0,

$$F_0 \sim F_{k,n-(k+1)}$$

Test for the significance of the individual parameters

To test
H_0: $\alpha_j = 0$ against H_1: Not $H_{0,}$ j = 1(1) 4.

The test statistics is given by

$$t^* = \hat{\alpha_j} \,/\, SE\ (\hat{\alpha_j}\ )$$

where $\alpha_{\bar{j}}$ is the estimated value of α_j and SE ($\alpha_{\bar{j}}$) is the Standard Error of $\alpha_{\bar{j}}$.

Under H_0,

$$t^* \sim t_{n-(k+1)}$$

where, ‘n’ denotes the total number of observations and ‘k’ denotes the total number of explanatory variables.

Our chosen level of significance is 0.05.
Input:

```
model_new = lm(OR_N~AM1+AM2+AM3+MCR)

summary(model_new)
```

Output:

```
Coefficients:
                Estimate     Std. Error     t value      Pr(>|t|)
(Intercept)      98.045624     1.361097       72.034       < 2e-16
    AM1         -0.108919      0.006897      -15.793       < 2e-16
    AM2         -0.143369      0.040583      -3.533         0.000744
    AM3         -0.046249      0.014947      -3.094         0.002864
    MCR          1.976704      0.326505       6.054         6.84e-08
```

Interpretation:

It is seen that all the variables AM1, AM2, AM3 and AM4 are significant and thus our regression model is also significant.

INTERPRETATION OF THE PARAMETERS

It is important to interpret the parameters in a model since it reflects the relative importance of the predictors in the definition of the model itself.

i). Interpretation of α_1 (the parameter associated with AM1)

We see that the estimated value of α_1 is -0.092821. If AM1 increases by 1, then, OR_N will decrease by the amount -0.108919, keeping all the other explanatory variables fixed. Also, the p-value came out to be very small indicating that the parameter is significant.

ii). Interpretation of α_2 (the parameter associated with AM2)

We see that the estimated value of α_2 is -0.143369. If AM2 increases by 1, then, OR_N will decrease by the amount -0.143369, keeping all the other explanatory variables fixed. Also, the p-value came out to be very small indicating that the parameter is significant.

iii). Interpretation of α_3 (the parameter associated with AM3)

We see that the estimated value of α_3 is -0.046249. If AM3 increases by 1, then, OR_N will decrease by the amount -0.046249, keeping all the other explanatory variables fixed. Also, the p-value came out to be very small indicating that the parameter is significant.

iv). Interpretation of α_4 (the parameter associated with MCR)

We see that the estimated value of α_4 is 1.976704. If MCR increases by 1, then, OR_N will increase by the amount 1.976704, keeping all the other explanatory variables fixed. Also, the p-value came out to be very small indicating that the parameter is significant.

Overall Interpretation: Hence, all the parameters of the model come out to be significant indicating that the removal of the outliers was successful.

FURTHER SCOPE

We can divide the entire dataset into two parts, that is, one is training set and one is test set. In the training set, we try to build a model using it and in the test set, we see how our fitted model performs, which is basically the concept of Machine Learning. Using this, we can check the efficiency of the model and how it may perform in other unknown datasets.

ACKNOWLEDGEMENT

Real learning comes from a practical work. We are really thankful to our supervisor Prof. Swagata Nandi Ma’am for providing us such an opportunity to be engaged in a project work. We are also grateful for her guidance throughout this project.

BIBLIOGRAPHY

➢ Linear Regression Analysis, George A. F. Seber, Alan J. Lee

➢ Introduction to Linear Regression Analysis, Douglas Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining

➢ http://home.iitk.ac.in/~shalab/course5.html

➢ https://cran.r-project.org/

➢ https://www.statology.org/

➢ https://www.r-graph-gallery.com/

TABLE 1: ORIGINAL DATASET

A1	A2	A3	A4	OR
55.33	1.72	54	1.66219	92.19
59.13	1.20	53	1.58399	92.74
57.39	1.42	55	1.61731	91.88
56.43	1.78	55	1.66228	92.80
55.98	1.58	54	1.63195	92.56
56.16	2.12	56	1.68034	92.61
54.85	1.17	54	1.58206	92.33
52.83	1.50	58	1.54998	92.22
54.52	0.87	57	1.56230	91.56
54.12	0.88	57	1.57818	92.17
51.72	0.00	56	1.60401	92.75
51.29	0.00	58	1.59594	92.89
53.22	1.31	58	1.54814	92.79
54.76	1.67	58	1.63134	92.55
53.34	1.81	59	1.60228	92.42
54.84	2.87	60	1.54949	92.43
54.03	1.19	60	1.57841	92.77
51.44	0.42	59	1.61183	92.60
53.54	1.39	59	1.51081	92.30
57.88	1.28	62	1.56443	92.30
60.93	1.22	62	1.53995	92.48

59.59	1.13	61	1.56949	91.61
61.42	1.49	62	1.41330	91.30
56.60	2.10	62	1.54777	91.37
59.94	2.29	61	1.65523	91.25
58.30	3.11	62	1.29994	90.76
58.25	3.10	63	1.19975	90.90
55.53	2.88	64	1.20817	90.43
59.79	1.48	62	1.30621	90.83
57.51	0.87	60	1.29842	92.18
62.82	0.88	59	1.40483	91.73
62.57	0.42	60	1.45056	91.10
60.23	0.12	59	1.54357	91.74
65.08	0.10	60	1.68940	91.46
65.58	0.05	59	1.74695	91.44
65.64	0.05	60	1.74919	91.56
65.28	0.42	60	1.78053	91.90
65.03	0.65	59	1.78104	91.61
67.84	0.49	54	1.72387	92.09
73.74	0.00	54	1.73496	90.64
72.66	0.00	55	1.71966	91.09
71.31	3.44	55	1.60325	90.51
72.30	4.02	55	1.66783	90.24
68.81	6.88	55	1.69836	91.01
66.61	2.31	52	1.77967	91.90

63.66	2.99	52	1.81271	91.92
63.85	0.24	50	1.81485	92.16
67.25	0.00	53	1.72526	91.36
67.19	0.00	52	1.86782	92.16
62.34	0.00	48	2.00677	92.68
62.98	0.00	47	1.95366	92.88
69.89	0.00	55	1.89387	92.59
73.13	0.00	57	1.81651	91.35
65.09	1.01	57	1.45939	90.29
64.71	0.61	55	1.38934	90.71
64.05	1.64	57	1.33945	90.41
63.97	2.80	60	1.42094	90.43
70.48	4.64	60	1.57680	89.87
71.11	3.56	60	1.41229	89.98
69.05	2.51	60	1.54605	90.00
71.99	1.28	55	1.55182	89.66
72.03	1.28	56	1.60390	90.08
69.90	2.19	56	1.67265	90.67
72.16	0.51	56	1.55242	90.59
70.97	0.09	55	1.45728	91.06
70.55	0.05	52	1.26174	90.69
69.73	0.06	54	1.28802	91.11
69.93	0.05	55	1.36399	90.32
70.60	0.00	55	1.42210	90.36

75.54	0.00	55	1.67219	90.57
49.14	0.00	40	2.17140	94.17
49.10	0.00	42	2.31909	94.39
44.66	4.99	42	2.14314	93.42
44.64	3.73	44	2.08081	94.65
4.23	10.76	41	2.17070	97.61
5.53	7.99	40	1.99418	97.08
17.11	5.06	47	1.61437	95.12
67.60	1.84	55	1.64758	91.86
64.81	2.24	54	1.69592	91.61
63.13	1.60	52	1.66118	92.17
63.48	3.46	52	1.48216	91.56
62.25	3.56	50	1.49734	92.16

TABLE 2: UPDATED DATASET

AM1	AM2	AM3	MCR	OR
55.33	1.72	54	1.66219	92.19
59.13	1.2	53	1.58399	92.74
57.39	1.42	55	1.61731	91.88
56.43	1.78	55	1.66228	92.8
55.98	1.58	54	1.63195	92.56
56.16	2.12	56	1.68034	92.61
54.85	1.17	54	1.58206	92.33
52.83	1.5	58	1.54998	92.22
54.52	0.87	57	1.5623	91.56
54.12	0.88	57	1.57818	92.17
51.72	0	56	1.60401	92.75
51.29	0	58	1.59594	92.89
53.22	1.31	58	1.54814	92.79
54.76	1.67	58	1.63134	92.55
53.34	1.81	59	1.60228	92.42
54.84	2.87	60	1.54949	92.43
54.03	1.19	60	1.57841	92.77
51.44	0.42	59	1.61183	92.6
53.54	1.39	59	1.51081	92.3

57.88	1.28	62	1.56443	92.3
60.93	1.22	62	1.53995	92.48
59.59	1.13	61	1.56949	91.61
61.42	1.49	62	1.4133	91.3
56.6	2.1	62	1.54777	91.37
59.94	2.29	61	1.65523	91.25
58.3	3.11	62	1.29994	90.76
58.25	3.1	63	1.19975	90.9
55.53	2.88	64	1.20817	90.43
59.79	1.48	62	1.30621	90.83
57.51	0.87	60	1.29842	92.18
62.82	0.88	59	1.40483	91.73
62.57	0.42	60	1.45056	91.1
60.23	0.12	59	1.54357	91.74
65.08	0.1	60	1.6894	91.46
65.58	0.05	59	1.74695	91.44
65.64	0.05	60	1.74919	91.56
65.28	0.42	60	1.78053	91.9
65.03	0.65	59	1.78104	91.61
67.84	0.49	54	1.72387	92.09
73.74	0	54	1.73496	90.64

72.66	0	55	1.71966	91.09
71.31	3.44	55	1.60325	90.51
72.3	4.02	55	1.66783	90.24
66.61	2.31	52	1.77967	91.9
63.66	2.99	52	1.81271	91.92
63.85	0.24	50	1.81485	92.16
67.25	0	53	1.72526	91.36
67.19	0	52	1.86782	92.16
62.34	0	48	2.00677	92.68
62.98	0	47	1.95366	92.88
73.13	0	57	1.81651	91.35
65.09	1.01	57	1.45939	90.29
64.71	0.61	55	1.38934	90.71
64.05	1.64	57	1.33945	90.41
63.97	2.8	60	1.42094	90.43
70.48	4.64	60	1.5768	89.87
71.11	3.56	60	1.41229	89.98
69.05	2.51	60	1.54605	90
71.99	1.28	55	1.55182	89.66
72.03	1.28	56	1.6039	90.08
69.9	2.19	56	1.67265	90.67

72.16	0.51	56	1.55242	90.59
70.97	0.09	55	1.45728	91.06
70.55	0.05	52	1.26174	90.69
69.73	0.06	54	1.28802	91.11
69.93	0.05	55	1.36399	90.32
70.6	0	55	1.4221	90.36
75.54	0	55	1.67219	90.57
44.64	3.73	44	2.08081	94.65
67.6	1.84	55	1.64758	91.86
64.81	2.24	54	1.69592	91.61
63.13	1.6	52	1.66118	92.17
63.48	3.46	52	1.48216	91.56