Individual Coursework Submission Form

Specialist Masters Programme

Surname: Joshi	**First Name: Rutva**
MSc in: Business Analytics	**Student ID number: 220052555**
Module Code: SMM634	
Module Title: Analytics Methods for business	
Lecturer: Prof Radice, Rosailba	**Submission Date: 04/11/2022**

Declaration:
By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: | **Final Mark:** **%**

1. (a) Justification of the chosen regression model specification.

The following model :

$$Y = \beta 0 + \beta 1(price) + \beta 2(h.rain) + \beta 3(s.temp) + \beta 4(w.rain) + \beta(h.temp) + \varepsilon i$$

Here X0 is constant that is year , X1 is price, X2 is h.rain , X3 is s.temp , X4 is w.rain , X5 is h.temp and Ei is Error term

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how changes vary as the independent variable(s) change. Multiple linear regression **is** used to estimate the relationship between **two or** more independent variables **and** one dependent variable**.**

- The data follows normal distribution .
- Here y is the dependent variable whereas the other X component are the independent variable
- For linearity the line of best fit through all the data points are straight line , rather tha curve or some sort of other factors showing in graph .

There are no unobserved variances among the variables because the observations in the dataset were gathered using statistically sound techniques.

Prior to creating the regression model, it is crucial to determine whether any of the independent variables in multiple linear regression are genuinely connected with one another. Only one independent variable should be utilised in the regression model if the correlation between the other two is too high (r2 > 0.6).

Across the range of values for the independent variable, the magnitude of the inaccuracy in our estimate remains rather constant.

The model fit is a key factor to think about when choosing a model for multiple linear regression analysis. In a multiple linear regression model, including independent variables will always increase the percentage of the dependent variable's variance that is explained (usually denoted as R2). Therefore, an over-fit model may occur from the addition of too many independent variables without any supporting hypothesis.

1. (b) Using the final model, provide a summary (e.g., using tables and figures) of the empirical findings as well as interpretation of the estimated model parameters.

```
2. 'data.frame':    25 obs. of  7 variables:
3. ##  $ Year       : int  1952 1953 1955 1957 1958 1959 1960 1961 1962 1963 ...
4. ##  $ Price      : num  7.5 8.04 7.69 6.98 6.78 ...
5. ##  $ WinterRain : int  600 690 502 420 582 485 763 830 697 608 ...
6. ##  $ Avg temp      : num  17.1 16.7 17.1 16.1 16.4 ...
7. ##  $ HarvestRain: int  160 80 130 110 187 187 290 38 52 155 ...
```

The Given table shows the data frame which has year , price , rain in the harvest month , average temperature , rain in the winter preceding harvest , harvest temperature .

```
model1 <- lm(Price ~ AGST, data=Wine)

summary(model1)
```

The above equation shows the lm plot for price and average temperature . Here it shows that the year is constant where the avg. temp differs in every year .

The residual variable for the above equation is 0.57 which is the smaller value according to the analysis of variance that was performed. variables and (factor). Despite the fact that the optimism of p values is greater than 0.05, the null hypothesis for these variables cannot be disregarded.

The R2 (adj) value (5%) is a change in accordance with R2 dependent on the number of x-variables in the model . With just a single x-variable, the charged R2 isn't significant.

When the p value of the variable is less than 0.05 that is 5% we cannot neglect the null hypothesis . Here in the above equation y is dependent variable but as B0 is constant the answer doesn’t change with the regression line .

```
model3 <- lm(Price ~ AGST + WinterRain + HarvestRain + data=Wine)

summary(model3)
```

It shows that when the mean of price , average temperature , rain winter and harvest rain for the dataset of wine is collect the p value of the variance of all the variable is 1.73 . Here we can see that the p value is more than 1 so this model cant be fiited .

Hence the Null hypothesis is not neglected for the the function . Therefore it also shows that this model is not best fitted model .

```
WineTest <- read.txt("wine_test.txt")
str(WineTest)
```

The above equation shows the interpretation of the dataset in which the value of all the given set is given . It shows that the value by the mean and median is not the exact value for the model by which we can get the relationship between the model linear . Hence this model is not fitted model .

1. (c) Provide recommendations and limitations of your analysis.

   Typically, regression analysis is used in research to establish that a relationship between variables exists. Correlation does not imply causation, though; just because two variables are connected doesn't indicate they are responsible for each other's occurrence. In fact, even a line in a simple linear regression that closely matches the information emphasis may not guarantee a relationship between the variables and logical outcomes.

   The given dataset has many more errors and so the residual fitted plot has vary residual values in the graph set .

   Linear Regression performs well when the dataset is linearly separable. We can use it to find the nature of the relationship among the variables.

   Over-fitting is a common problem with linear regression, however it may be readily prevented by using cross-validation, regularisation (L1 and L2) techniques, and some dimensionality reduction approaches.

   The limitation :

   1. Main limitation of regression is the assumption of linearity between the dependent variable and the independent variables.

2. There is no thrash value needed or Null value needed .
3. A straight line can be created that touches the majority of plots when the training datasets are projected.
4. If the number of observations are lesser than the number of features, Linear Regression should not be used, otherwise it may lead to overfit because is starts considering noise in this scenario while building the model.
5. Linear regression is very sensitive to outliers .So, outliers should be analysed and removed before applying Linear Regression to the dataset.
6. Before applying Linear regression, multicollinearity should be removed (using dimensionality reduction techniques) because it assumes that there is no relationship among independent variables.

1. (d) What did you learn from the analysis? What is the answer, if any, to the questions you set out to address? How can the analysis be improved?

Main limitation of Linear Regression is the **assumption of linearity** between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.

The necessary point learned during fitting this model was the dataset should be clear . If the p value of the plot is near to the 1 then there is linear relationship between the model . Hence we can say that the we can neglect the Null hypothesis or Null value and can prove that the model is fitted .

Linear Regression performs well when the dataset is **linearly separable**. We can use it to find the nature of the relationship among the variables.

During the extracting the value from data set we have found that there is many N/A values in the dataset and that creates the problem or error in the data set .

Correlation of error terms.

The actual relation between response and the predictor is not linear, then all the conclusion we draw becomes null and void. Also, the accuracy of the model may drop significantly.

Solution for the correlation error is that plot the some residual points on the residual plot and in this way we can define the correlation between two points .

In situations when there are non-linear relationships, linear regression performs poorly. They are not naturally flexible enough to capture more complicated patterns, and it can be difficult and time-consuming to include the proper interaction terms or polynomials. This was the problem which I faced during this model .

2 (A) By **looking at the R code below write down the statistical model that has been fitted and the model assumptions?**

$CO2 = \beta0 + \beta1(income) + \beta2(fwd) + \beta3(belief) + Ei$

Here y = CO2 , x1 = income , x2 = factor , x3 = belief , Ei = component error

The model's fundamental assumptions are as follows:

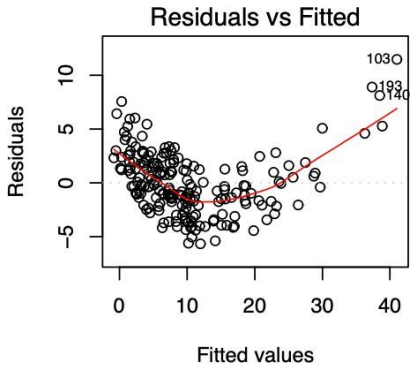
1. The independent variable x and the dependent variable y have a linear relationship.

2. The following factors do not correlate.

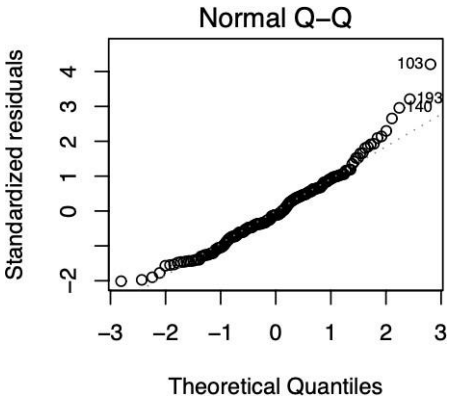
3. The variance of component error is constant.

4. The normality variance of the subsequent variables.

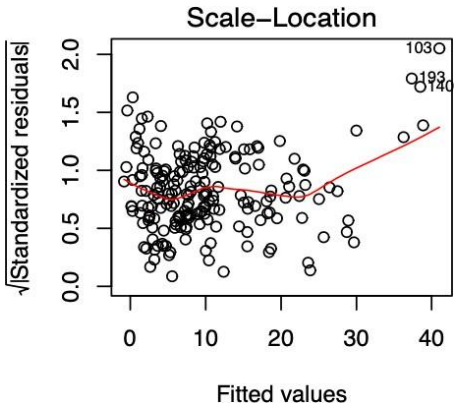
(B) Comment on the following residual plots.



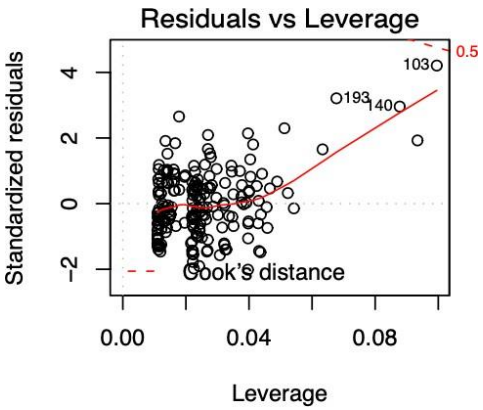
As the red line is not quite as close to the dashed line, we can observe that linearity doesn't seem to hold quite well in this case. The range of the residuals appears to be expanding as we move further to the right on the x-axis. The majority of the high points are between 5 and 15 while the lowest point is 40.



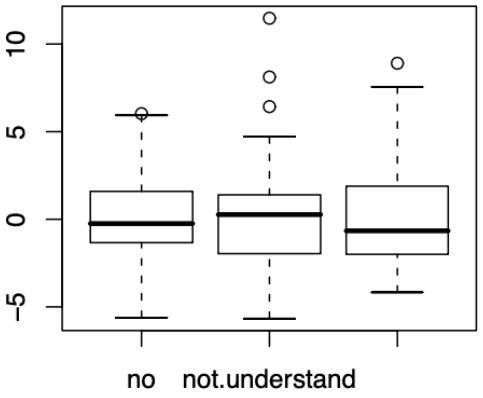
The above normal Q-Q plot demonstrates that some points are above the dashed line, suggesting that the assumption of normalization is not always valid.

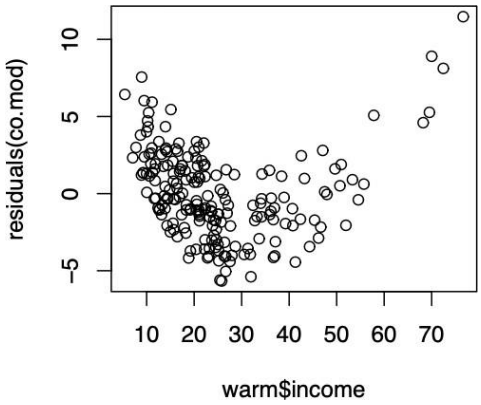


The above scale- location plot shows non linearity because it is not in the straight line . There are several outliers, with residuals between 0 – 10



The nonlinearity is depicted in the Residual vs. Leverage graphic above. There are several residuals in the range of 0 and 0.4. Numerous locations, including 103 and 140, deviate from the model's constant variance.





The above plots show the quadratic relationship. Whether there is homoscedastic or not is less obvious. As the above plot doesn't shows the linear relationship ,the model seems to be inappropriate .

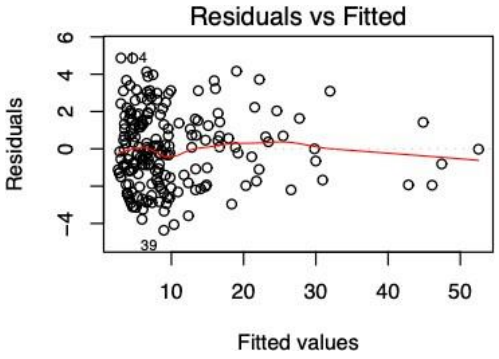
(C) Using the R code below, explain what model has been fitted.

The following code shows the model :

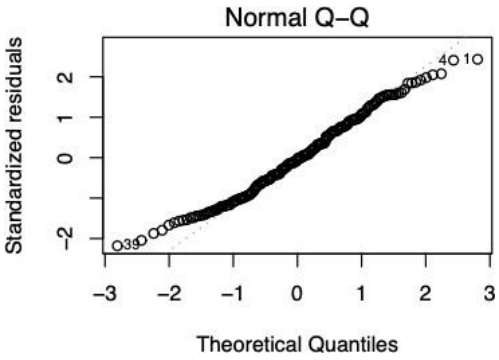
$CO2 = \beta0 + \beta1(income) + \beta2(fwd) + \beta3(belief) + Ei$

Here y = CO2 , x1 = income , x2 = factor , x3 = belief , Ei = component error

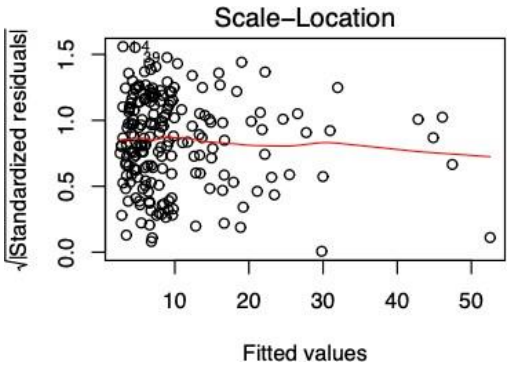
.



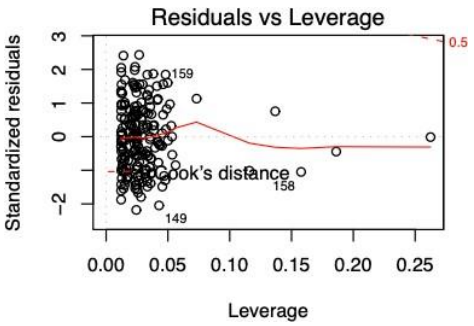
The model is deemed to be fitted because the Residual vs. Fitted plot shown above exhibits linearity in plot. As a result, the residual value is more between 0 and 10 and lowest at 50.



The above plot normal Q-Q , we can said that the lower end and upper end of Q-Q slightly differs from the dashed line but it shows the linear relationship in the plot .



The linear relation between them is also shown by the plot scale and location shown above. However, when we contrast the plot with the two mentioned graphs, it becomes clear that the last end is a little bit in the highest to the lowest. The residual variables do not have a uniform distribution.



The above plot residuals vs leverage shows us that the model is fitted perfectly as it shows that the linear relationship exists among the residual and leverage points.

(E) What conclusions can be drawn from the following analysis of variance?

We can ignore the null value of income, fwd because the p values of income are equal to 0.05, which is very little, according to the analysis of variance that was performed. variables and (factor). Despite the fact that the optimism of p values is greater than 0.05, the null hypothesis for these variables cannot be disregarded.

(F) Explain the results of the following analysis of variance. Also comment on the two- way table below.

From the given table we can analyis the belief variable of p-value is 0.89 which is higher than compare to the value 0.5 . Therefore we cannot neglect the null hypothesis . The given table shows that we have the.

(G) Looking at the R code below, explain which model has been fitted and comment on the analysis of variance table.

The following R code is

$CO2 = \beta0 + \beta1(income) + \beta2(fwd) + \beta3(belief) + Ei$

Here y = CO2 , x1 = income , x2 = factor , x3 = belief , Ei = component error

We can observe from the given variance table that the association between the independent and dependent variables is exceptionally strong. All of the variables' P-values are less than 0.05. So in this context, we can disregard the null hypothesis.

(H) By looking at the R code below which model has been fitted? What are the conclusions from the analysis of variance table?

The following R code is

$CO2 = \beta0 + \beta1(income) + \beta2(fwd) + \beta3(belief) + Ei$

Here y = CO2 , x1 = income , x2 = factor , x3 = belief , Ei = component error

From the given variance table we can see that there is exceptional relationship between the independent variables and the dependent variables . This means all the p values are less than 0.05. Here we can neglect the NULL hypothesis.

2(I) Do you think that the following model fitted in R is reasonable as compared to all models fitted above? Comment on the relationships between CO2 and income, and CO2 and fwd.

The following R code is

$CO2 = \beta0 + \beta1(income) + \beta2(fwd) + \beta3(belief) + Ei$

Here y = CO2 , x1 = income , x2 = factor , x3 = belief , Ei = component error

This example demonstrates the extraordinary relation between CO2 and x1, x2, and x3. Therefore, we can conclude that it is the model that fits the data the best overall. As a result, we can reject the null hypothesis. By examining the preceding table, we can determine that the dependent variable CO2 and the independent variable x1, which is income, have a specific relationship. This allows us to speak about every variable. As a result, this model is the best fit.