Individual Coursework Submission Form

Specialist Masters Programme

Surname: Chawla	**First Name: Sarthak**
MSc in: Business Analytics	**Student ID number: 220042213**
Module Code: SMM634	
Module Title: Analytical Methods for Business	
Lecturer: Prof. Radice , Rosailba	**Submission Date: 4 November 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** **%**

Q1

(A) **Justification of the chosen regression model specification.**

Answer

After analyzing the provided dataset, it was discovered that the Price and Parker column contained many NA values. First, imputation was done for effective model fitting, and then standardization was needed because of the variability in the dataset. Finally, the dataset was ideal for analysis and model fitting after imputation and standardization.

The following model was found to be the best fit after testing several models in which price was the dependent variable:

$Y = \beta0 + \beta1x1 + \beta2x2 + \beta3x3 + Ei$

$price = \beta0 + \beta1(year) + \beta2(s.temp) + \beta3(h.temp * w.rain) + Ei$

Where,

y = price (Dependent variable)

Ei = Error component/term

x1 = year (independent variable)

x2 = s.temp (independent variable)

x3 = h.temp*w.rain (interaction variable)

After fitting the model and analyzing the summary table, it was discovered that the fitted model's p value is 1.261e-08, which is less than the significant value (0.05/5%), implying that we reject the null hypothesis. Furthermore, the multiple R2 = 65.69% and adjusted R2 = 61.51% are the highest of any of the tested models. As a result, we can conclude that the fitted model is superior.

(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.**

Answer

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.01058    0.09057  -0.117  0.90758
year         -0.73369    0.11952  -6.138 2.75e-07 ***
s.temp        0.65589    0.12777   5.133 7.30e-06 ***
h.temp        0.29301    0.10507   2.789  0.00799 **
w.rain        0.22996    0.10162   2.263  0.02900 *
h.temp:w.rain 0.32333    0.10779   3.000  0.00458 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6204 on 41 degrees of freedom
Multiple R-squared:  0.6569,    Adjusted R-squared:  0.6151
F-statistic:  15.7 on 5 and 41 DF,  p-value: 1.261e-08
```

The p values of year, s.temp, h.temp, w.rain, and h.temp:w.rain are all less than the significant value (0.05/5%), indicating that all variables are correlated and significant with the dependent variable. Furthermore, the R2 = 65.69% and adjusted R2 = 61.51% indicate that the movement of the independent variable with the dependent variable is fitting for the 65.69%. Furthermore, the fitted model's p value is 1.261e-0.8, which is much lower than the significant value, so we reject the null hypothesis.

The res vs fitted(model) plots indicate that the data is linearly fit along the line. However, there are only a few outliers at the extreme res value.

By analysing the Q-Q plot for plot, it is possible to conclude that the upper and lower end of the QQ plot points deviate. As a result, we can say it has a fat tail at both ends and in the centre that follows a straight line.

(C) **Provide recommendations and limitations of your analysis.**

Answer

Limitations,

1.R^2 value is 65.69% which has a scope to improve for better confidence at the fitted model.

2. Outliers were found during the analysis.

Recommendations,

1. Outliers can be rejected to improve the analysis of the model.

2. High influence points should be kept to a minimum.

(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?**

Answer

I was interested in learning how the price of wine related to the various factors listed and which ones had the greatest influence. After conducting the analysis, I was able to draw the conclusion that the summer harvest's temperature has the greatest influence on prices.

The R2 value is close to 65%, which could be increased to increase confidence in the fitted model and the predicted values, but the model is undoubtedly not the best of all. The QQ plot revealed outliers that could be removed to improve the model's effectiveness.

Q2

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

Answer With reference to the R code it can be assessed that, the fitted statistical model is as follows:

$$\mathbf{CO2 = \beta 0 + \beta 1(income) + \beta 2(fwd) + \beta 3(belief) + Ei}$$

Where,

Y = CO2(Dependent variable)

Ei (Error component/term)

x1 = income (independent variable)

x2 = fwd (independent variable)

x3 = belief (independent variable)

It is assumed that,

1. Errors are distributed normally.
2. No collinearity exists among independent variables.
3. Linear Relationship.

(B) **Comment on the following residual plots.**

Answer

1. Residuals Vs Fitted Plot

By analyzing this plot for plot(co.mod), it is possible to conclude that the fitted model's linearity has been broken and that a quadratic relationship is present. Additionally, there are numerous outliers close to the residual 10 value. Among them are numbers 103, 193, and 140.

2. Normal Q-Q Plot

When plot(co.mod) is used to analyze this plot, it can be deduced that points above the line on the upper end of the QQ plot suggest a small non-normality. It may be claimed that the curve is right- skewed because its right side has a larger tail.

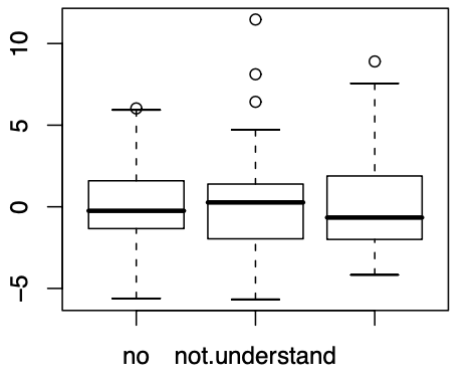
3. Scale Location Plot

The given graph shows that the fitted model is nonlinear, and the residuals are not evenly distributed. Magnitude appears to be highest at fitted values ranging from 0 to 10, medium at 10-20, 20-30, and lowest at 40. Heteroskedasticity is observed.

4. Residuals Vs Leverage Plot

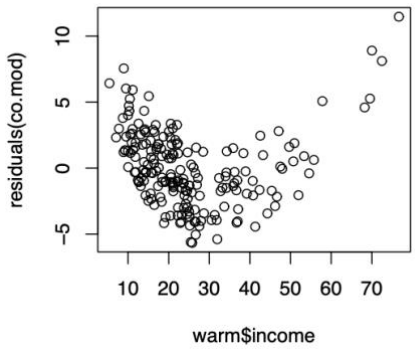
Heteroskedasticity and nonlinearity can be confirmed by analyzing the provided plot for plot(co.mod). Several values, including 103, can have a significant impact on the fitted model, and removing them would significantly alter the model.

5.



This graph depicts the function of a task. The upper and lower margins of the box plot represent the range's upper and lower quartiles. Any points in this box are best suited. The median is represented by the line that runs through the boxes. Outliers are the points on the outskirts.

6.



The plot indicates that there is a quadratic relationship between the income variable and the residuals. Therefore, we can conclude that the fitted model frequently underestimates the income values. The fitted model is therefore incorrect.

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

Answer According to the given R code it can be assessed that, the model fitted is as follows:

$CO2 = \beta0 + \beta1(income) + \beta2(income^2) + \beta3(fwd) + \beta4(belief) + Ei$

Where,

y = CO2(Dependent variable)

Ei (Error component/term)

x1 = income (independent variable)

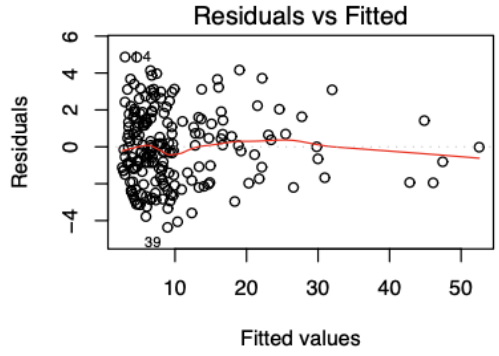
x2 = income^2 (independent variable)

x3 = fwd (independent variable)

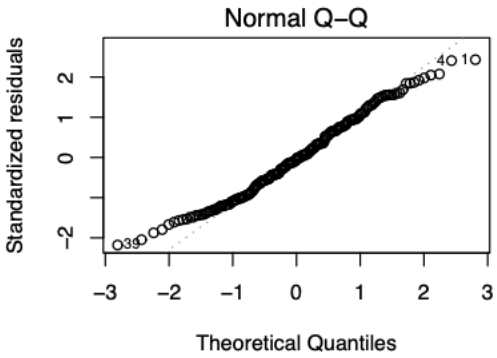
x4 = belief (independent variable)

(D) **Comment on the following residual plots.**

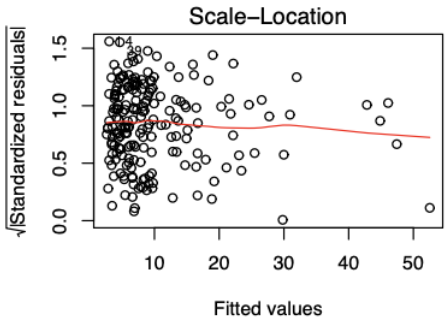
Answer



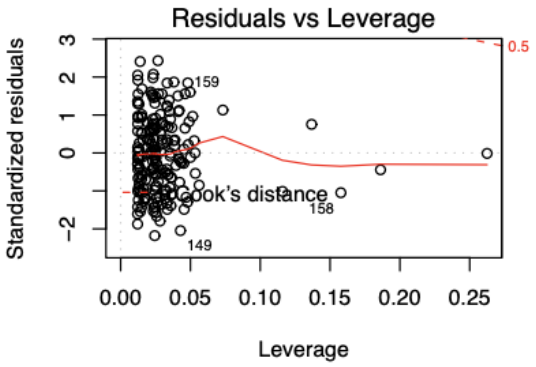
It can be concluded from an analysis of the Residual vs. Fitted plot for plot(co.mod1) that the fitted model maintains linearity. The residuals are not distributed equally across the fitted values, though.



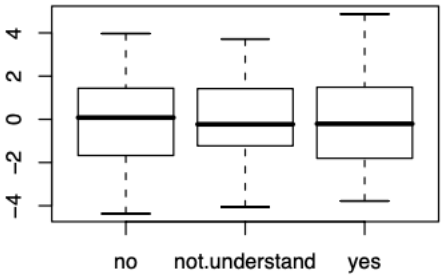
Analysing the Q-Q plot for plot(co.mod1) it can be inferred that the upper and lower end of the QQ plot points deviates within a smaller margin. Hence we can say it is a thin tailed distribution and nearly perfect for normal distribution.



By analyzing the aforementioned graph, it can be seen that the residuals are not evenly distributed, and the fitted model is linear. Magnitude appears to be lowest at fitted values between 40 and 50 and highest at fitted values between 0 and 10. We see heteroskedasticity.



The existence of linearity can be confirmed by analyzing the Standardized residuals vs Leverage plot for plot(co.mod1). Furthermore, no extreme values were observed that would change the fitted model.



This graph depicts the function of a task. The upper and lower margins of the box plot represent the range's upper and lower quartiles. Any points in this box are best suited. The median is represented by the line that runs through the boxes. The points on the outskirts are the outliers.

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

Answer

```
> anova(co.mod1)
Analysis of Variance Table
Response: CO2
             Df  Sum Sq   Mean Sq  F value     Pr(>F)
income        1 13513.8  13513.8  3282.6195  < 2.2e-16 ***
I(income^2)   1   793.7    793.7   192.8083  < 2.2e-16 ***
fwd           1   104.9    104.9    25.4765   1.027e-06 ***
belief        2     7.8      3.9     0.9475   0.3895
Residuals   194   798.7      4.1
```

As can be seen, the given variables' p values are less than 0.05/5%, which means that we reject the null hypothesis for the given variables income, income^2, and fwd. However, since the belief variable's p value exceeds the level of significance, we are unable to reject the null hypothesis and can therefore draw the conclusion that the belief variable has little to no influence on the fitted model or dependent variable.

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

Answer

```
> anova(lm(CO2~belief,data=warm))
Analysis of Variance Table
Response: CO2
             Df   Sum Sq  Mean Sq F value  Pr(>F)
belief        2    16.8    8.4     0.1089  0.8969
Residuals   197 15202.0   77.2

> table(warm$fwd,warm$belief)
   no not.understand yes
0  36    38          87
1  21    13           5
```

The p value for the belief variable is 0.8969, which is significantly higher than the significant value of 0.05/5%, as can be seen. As a result, it can be concluded that belief has no effect on the dependent variable CO2. Additionally, faith has no bearing on the calculation of CO2 emissions. As a result, it can be disregarded from the co.mod1 fitted model.

The two-way table shows that there are 87 households without a four-wheel drive who comprehend the science underlying CO2 emission, and that there are an almost equal number of households who either don't accept the science or don't want to comprehend it. However, only 5 of the car's owners subscribe to the science of emissions. 13 households are unable to comprehend the science of CO2 emission, and 21 households do not believe in emission science.

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

Answer

```
> anova(co.mod2)
Analysis of Variance Table
Response: CO2
             Df  Sum Sq  Mean Sq  F value   Pr(>F)
income        1 13513.8  13513.8  3284.38  < 2.2e-16 ***
I(income^2)   1   793.7    793.7   192.91  < 2.2e-16 ***
fwd           1   104.9    104.9    25.49  1.013e-06 ***
Residuals   196   806.5      4.1
```

According to the given R code , the model fitted is as follows:

$CO2 = \beta0 + \beta1(income) + \beta2(income^2) + \beta3(fwd) + Ei$

Where,

y = CO2 (Dependent variable)

Ei (Error component/term)

x1 = income (independent variable)

x2 = income^2 (independent variable)

x3 = fwd (independent variable)

The variance table reveals that the fitted model is a very sophisticated model. All of the variables significantly affect CO2, the dependent variable. We reject the null hypothesis because the p values for the income, income^2, and fwd are all lower than the significant value (0.05/5%).

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

Answer

```
Analysis of Variance Table
Response: CO2
             Df  Sum Sq  Mean Sq   F value     Pr(>F)
income        1 13513.8 13513.8   3272.8146  < 2.2e-16 ***
I(income^2)   1   793.7   793.7    192.2324  < 2.2e-16 ***
fwd           1   104.9   104.9     25.4004  1.060e-06 ***
income:fwd    1     1.3     1.3      0.3098  0.5784
Residuals   195   805.2     4.1
```

According to the given R code , the statistical model fitted is as follows:

$CO2 = \beta0 + \beta1(income) + \beta2(income^2) + \beta3(fwd*income)+Ei$

Where,

y = CO2 (Dependent variable)

Ei (Error component/term)

x1 = income (independent variable)

x2 = income^2 (independent variable)

x3 = fwd*income (interaction term)

The variance table indicates that the interaction terms of income and forward have no discernible influence on the dependent variable CO2 and are greater than the significant value (0.5/5%). In contrast, the dependent variable CO2 is correlated with income, income^2, and fwd.

(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**

Answer

```
Residuals:
   Min     1Q    Median   3Q    Max
-4.3988 -1.5711 -0.1593 1.5466 4.8989

Coefficients:
            Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)  1.6037309    0.5752177    2.788    0.00583 **
income       0.0860234    0.0387115    2.222    0.02742 *
I(income^2)  0.0075880    0.0005418   14.004    < 2e-16 ***
fwd          1.8310510    0.3626723    5.049    1.01e-06 ***

Residual standard error: 2.028 on 196 degrees of freedom
Multiple R-Squared: 0.947, Adjusted R-squared: 0.9462
F-statistic: 1168 on 3 and 196 DF, p-value: < 2.2e-16
```

```
> anova(co.mod2)
Analysis of Variance Table
Response: CO2
             Df  Sum Sq  Mean Sq  F value   Pr(>F)
income        1 13513.8  13513.8  3284.38  < 2.2e-16 ***
I(income^2)   1   793.7    793.7   192.91  < 2.2e-16 ***
fwd           1   104.9    104.9    25.49  1.013e-06 ***
Residuals   196   806.5      4.1
```

According to the given R code , the model fitted is as follows:

CO2 = β0 + β1(income) + β2(income^2) + β3(fwd) + Ei

Where,

y = CO2 (Dependent variable)

Ei (Error component/term)

x1 = income (independent variable)

x2 = income^2 (independent variable)

x3 = fwd (independent variable

As all the variables in the co.mod2 have p values less than the significant values and are therefore correlated with the dependent variable CO2, it can be said that model **CO2 = β0 + β1(income) + β2(income^2) + β3(fwd) + Ei** is the best fitted model out of all the models fitted above i.e **CO2 = β0 + β1(income) + β2(income^2) + β3(fwd) + β4(belief) + Ei, CO2 = β0 + β1(income) + β2(fwd) + β3(belief) + Ei .** Therefore, we disprove the null hypothesis. Furthermore, the multiple R2 is 94.7%, as can be seen. Additionally, adjusted R^2 is 94.6%, demonstrating that the models accurately depict how independent variables move in relation to the dependent variable.

We can state that when one unit of income increases, CO2 emissions increase by 86 tonnes annually based on our analysis of the p value, which shows that income and CO2 are significantly correlated.

Analysing the p values of fwd similarly, we can conclude that fwd and CO2 are related. Thus, we can conclude that increasing 1 unit of forward motion causes an annual increase in CO2 emissions of 1831 tonnes.