Individual Coursework Submission Form

Specialist Masters Programme

Surname: Ogoti	**First Name:** Soumya
MSc in: Business Analytics	**Student ID number:** 220045527
Module Code: SMM634	
Module Title: Analytics Methods for Business (PRD1 A 2022/23)	
Lecturer: Dr. Rosalba Radice	**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:** %

Table of Contents

Wine Analysis

Soumya Ogoti

ID: 220045527

Data Frame:

The data frame 'wine' has information on the price and growing characteristics of 25 Bordeaux wines (1952-1998) having 7 columns and 47 rows. The columns in the data are year, price (avg price of the wine as % of 1961 price), h.rain (mm), s.temp (C), w.rain (mm), h.temp (C), parker. Values are missing from the price (19.1%) and parker (38.2%) columns.

Data Cleaning:

From the given dataset, Parker ratings for wines before the year 1970 are missing. The wine guide from (Robert Parker Wine Advocate, 2022) has ratings for wines only after 1970 and it also states that older wines that are past their best are not rated. A regression analysis was performed to regress Parker ratings onto the other regressor variables (year, h.rain, s.temp, w.rain, h.temp). Multiple models were fit through forward-backward selection with the above predictors, however the best fit model with h.rain, s.temp and w.rain had an adjusted R-squared of only 0.35. Based on this, Parker ratings cannot be predicted through regression. Moreover, the factors that influence Parker's ratings are not known, it would be incorrect to impute the missing values with any numerical imputation methods (e.g., mean, most occurring values, etc.). Therefore, rows with missing Parker ratings were dropped in the following analysis. The price column also has missing values which cannot be imputed as the price is the response variable in question. The rows with missing price values were also dropped.

Model selection:

A correlation matrix was drawn to see the correlation among predictor variables. No two variables were strongly correlated.

	year	price	h.rain	s.temp	w.rain	h.temp	parker
year	1.00						
price	-0.21	1.00					
h.rain	-0.31	0.08	1.00				
s.temp	0.41	0.36	0.08	1.00			
w.rain	0.19	0.03	-0.45	-0.27	1.00		
h.temp	0.44	0.33	-0.45	0.48	0.05	1.00	
parker	0.41	0.67	-0.19	0.50	0.28	0.58	1.00

Initially, price is regressed onto all the other predictors as shown below.

$$price = \beta_0 + \beta_1 year + \beta_2 h.rain + \beta_3 s.temp + \beta_4 w.rain + \beta_5 h.temp + \beta_6 parker + \epsilon$$

An adjusted R-squared of 0.66 was observed with an RSE of 4.928. The residual vs. fitted values plot (Fig. 1) exhibits some non-linearity. The Q-Q plot shows that the residuals deviate from normality. However, the residual plots with respect to the independent variables as seen in Fig. 1 do not show any recognizable pattern. There are no influential leverage points

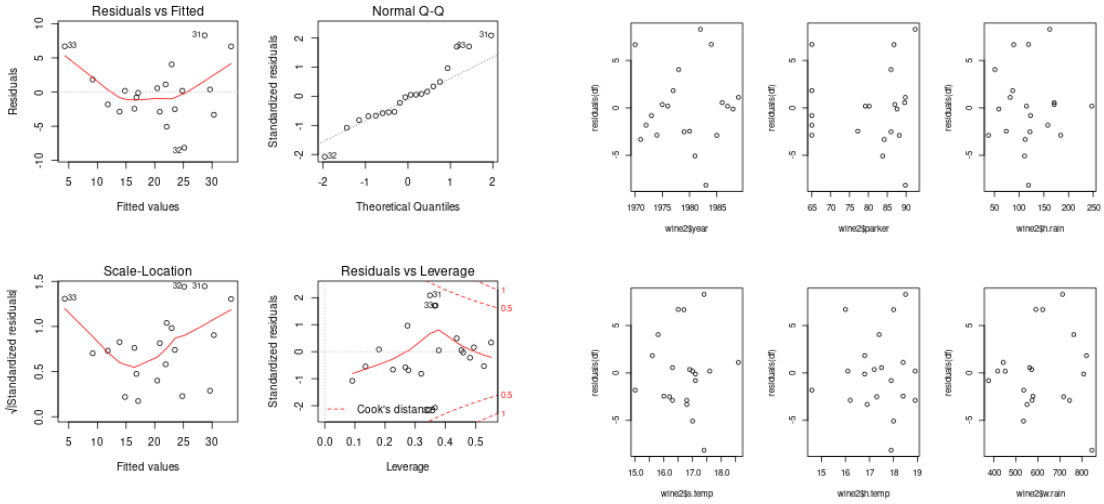


Fig. 1: left: residual plots vs fitted values, right: residual plots vs independent variables

Logarithm and Square-root transformations were applied to the response variable to overcome these issues. It was observed that the logarithm transformation yielded a better fit. The adjusted R-squared value for the model

$\log(price) = \beta_0 + \beta_1 year + \beta_2 h.rain + \beta_3 s.temp + \beta_4 w.rain + \beta_5 h.temp + \beta_6 parker + \epsilon$ was 0.78 and the RSE was 1.2, which is a significant improvement over the previous model.

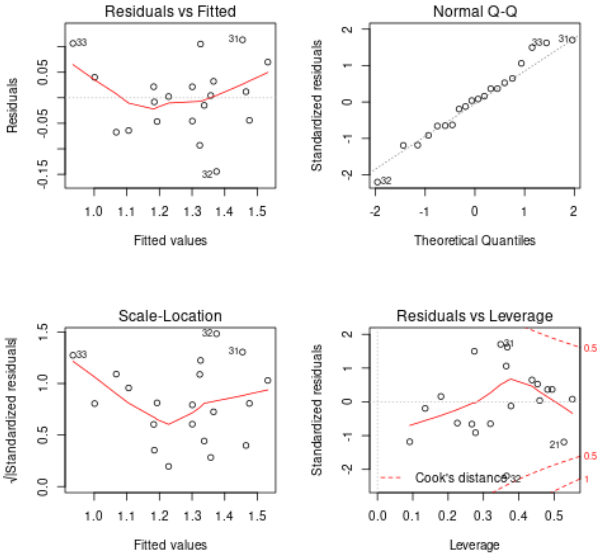


Fig. 2: Residual plots vs fitted values for the above equation. The non-linear trend in the residuals and the deviation from normality have decreased indicating a better fit.

Following this, variables that were of the least significance were removed from the model sequentially. Removing w.rain, h.rain and h.temp improved the model in terms of the variance explained by the model and RSE as seen in Table 1. As s.temp had a p-value close to 0.05, a model without s.temp was fit but it reduced the R^2 value and increased RSE.

Table 1. shows each model and its corresponding adjusted R-squared value and RSE.

Model	Adj R^2	RSE (in log10)
$\log(price) = \beta_0 + \beta_1 year + \beta_2 h.rain + \beta_3 s.temp + \beta_5 h.temp + \beta_6 parker + \epsilon$	0.792	0.080
$\log(price) = \beta_0 + \beta_1 year + \beta_3 s.temp + \beta_5 h.temp + \beta_6 parker + \epsilon$	0.803	0.078
$\log(price) = \beta_0 + \beta_1 year + \beta_3 s.temp + \beta_6 parker + \epsilon$	0.813	0.076
$\log(price) = \beta_0 + \beta_1 year + \beta_6 parker + \epsilon$	0.775	0.083

The analysis shows that the best model is $\log(price) = \beta_0 + \beta_1 year + \beta_3 s.temp + \beta_6 parker + \epsilon$

A stepwise AIC analysis also confirmed the same predictor features account for the greatest amount of variation in the model.

```
price ~ parker + year + s.temp

         Df Sum of Sq    RSS    AIC
<none>                321.52 63.547
- s.temp  1     45.74 367.27 64.207
+ h.temp  1      2.10 319.42 65.416
+ h.rain  1      0.95 320.58 65.488
+ w.rain  1      0.70 320.82 65.503
- year    1    435.55 757.07 78.674
- parker  1    648.16 969.68 83.625

Call:
lm(formula = price ~ parker + year + s.temp, data = wine2)
```

Model Summary:

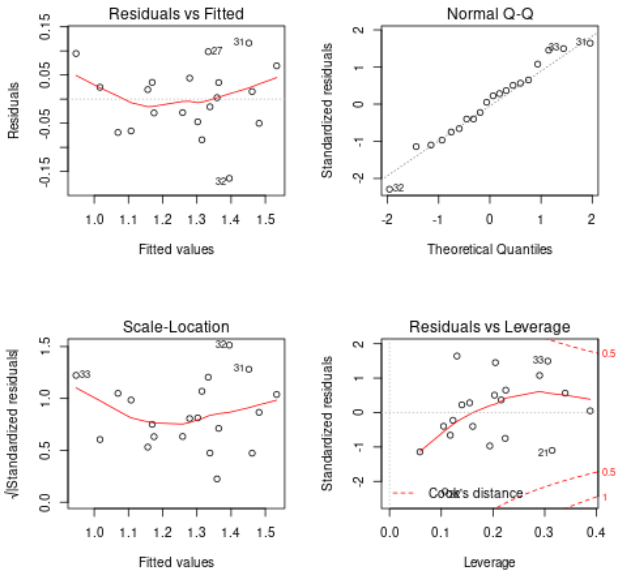
The model is a good fit as all the predictor variables are statistically significant, 81.32% of the variance in the data is explained by the model, and the overall F-statistic value is high

```
Call:
lm(formula = "log10(price) ~ year + parker + s.temp", data = wine2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.164352 -0.047865  0.009362  0.036941  0.116319

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.126266   6.475298   5.116 0.000104 ***
year        -0.017187   0.003346  -5.136 9.96e-05 ***
parker       0.015344   0.002087   7.352 1.63e-06 ***
s.temp       0.055876   0.026266   2.127 0.049292 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07596 on 16 degrees of freedom
Multiple R-squared:  0.8427,    Adjusted R-squared:  0.8132
F-statistic: 28.58 on 3 and 16 DF,  p-value: 1.16e-06
```



with a significantly low p-value (1.1e-6) indicating that the null hypothesis (at least one of the coefficients is zero) can be rejected. The RSE is 10^(0.076) = 1.19. The degrees of freedom is 16 (DOF = n-(p+1), n=20 observations, p=3 predictors).

The coefficients of the model are $\beta_0 = 33.13,\ \beta_1 = -0.017,\ \beta_3 = 0.056,\ \beta_6 = 0.015$. This implies that while keeping the parker rating and s.temp constant, for every 1 year increase in the date of production, the % price scales by a factor of $10^{-0.017}$=0.962. This is derived as follows.

For a year y the avg price p_1 as a % of the 1961 price can be predicted using the above model as $\log(p_1) = 33.12 - 0.017\ y + K$, where K accounts for the parker and s.temp terms. For an increase in the year by 1, the price p_2 can be predicted as $\log(p_2) = 33.12 - 0.017\ (y + 1) + K$. The difference is then

$$\log(p_2) - \log(p_1) = -0.017 \Rightarrow\ p2 = p1 * 10^{-0.017}$$

Similarly, while keeping the year and s.temp constant, an increase in the parker rating by 1 point scales the price by $10^{0.015}$=1.035 and while keeping the year and parker rating constant, an increase in the summer temperature by 1 °C scales the price by $10^{0.056}$=1.137

The residual vs fitted plot shows that the non-linearity has been addressed by the log transformation. The Q-Q plot shows that the normality assumption holds for most of the data points. The heteroscedasticity has been reduced as the standardised residuals are randomly scattered with equal variance. There are no influential leverage points as all points have a Cook's distance < 0.5.

```
Analysis of Variance Table

Response: log10(price)
          Df  Sum Sq Mean Sq F value    Pr(>F)
year       1 0.00827 0.00827  1.4341   0.24853
parker     1 0.46019 0.46019 79.7663 1.292e-07 ***
s.temp     1 0.02611 0.02611  4.5256   0.04929 *
Residuals 16 0.09231 0.00577
```

The ANOVA table shows that a model with year on its own is not statistically significant. When parker and s.temp are added, the SSE reduces than when only year and parker were included in the model.

The VIF for this model showed no collinearity among the predictors (1.290710 1.431326 1.425071).

Recommendation:

The above model was regressed on data for which there was a parker rating and this is a significant variable. Therefore, this model can only be used on wines for which the rating exists. As per (Robert Parker Wine Advocate, 2022) the Parker rating exists only for wines after 1970, hence the model only works for wines produced after 1970.

Limitations:

A limitation of the model is that the dataset is too small to choose the best model that could predict price accurately. This could result in overfitting the model. Missing values in Parker ratings could not be imputed as the factors that influence the ratings are unknown and the model that was regressed was a poor fit. Therefore, filling in these values would introduce bias in the data. Another limitation is the missing values in price which is the target variable. These rows had to be dropped though it was not desirable resulting in even smaller data.

Learnings and Improvement:

Looking at Parker's website, as the ratings were categorised there, a model was fit with categorised Parker values instead of the continuous numeric values. Such a model explained the variance less than the best fit model where ratings were numeric (0.79 as compared to 0.81). This shows that there was loss of information in the model due to the categorisation. The insight is that numerical values should be categorised only when necessary. Another learning was on handling missing values. Multiple statistical methods to impute missing data such as regression, replacing with mean/mode, KNN imputation were investigated. However, a key insight was that data imputation cannot always be performed and missing data should be removed, if necessary, as was the case here. The best fit model explains 81% of the variance in the data. This could possibly be improved if more data was available in terms of years after 1998. This would also enable fitting complex models to predict price better. More data points would make the insights gathered from the model more trustworthy.

References

Robert Parker Wine Advocate, 2022. *The Vine Advocate Vintage Chart.* [Online]
Available at: https://www.robertparker.com/resources/vintage-chart
[Accessed 3 11 2022].

Appendix – Code

```
# load dataset
wine <- read.table("wine.txt", header = TRUE)

# regress parker using the other variables
# ========================================
pp <- lm("parker ~ year + h.rain + s.temp + w.rain + h.temp", data=wine)
summary(pp) # R2 30.11
# remove h.temp and year
pp <- lm("parker ~ h.rain + s.temp + w.rain", data=wine)
summary(pp) # R2 30 -> 35, Error 7.94 > 7.6
# remove h.rain
pp <- lm("parker ~ s.temp + w.rain", data=wine)
summary(pp) # R2 32.2

# Regressing price from other parameters
# ========================================
# remove na rows from price and parker
wine2 <- wine[!is.na(wine$price),]
wine2 <- wine2[!is.na(wine2$parker),]

# correlation matrix
cor(wine2)

# plotting function to save plots
plotter = function(df, name1, name2){
    # plot model fit
    png(name1)
    par(mfrow=c(2,2))
    plot(df)
    dev.off()
    png(name2)
    par(mfrow=c(2, 3))
    plot(wine2$year, residuals(df))
    plot(wine2$parker, residuals(df))
    plot(wine2$h.rain, residuals(df))
    plot(wine2$s.temp, residuals(df))
    plot(wine2$h.temp, residuals(df))
    plot(wine2$w.rain, residuals(df))
    dev.off()}

# regress price on all predictor variables
wp2 <- lm("price ~ year + h.rain + s.temp + w.rain + h.temp + parker",
data=wine2)
```

```
summary(wp2) # An adjusted R2 of 0.66 was observed with a RSE of 4.928.
plotter(wp2, "wp2_plot.png", "wp2_residuals.png")

wp3 <- lm("log10(price) ~ year + h.rain + s.temp + w.rain + h.temp + parker",
data=wine2)
summary(wp3) # adj R2 was 0.78 and the RSE was 1.2
plotter(wp3, "wp3_plot.png", "wp3_residuals.png")

# removing w.rain from the model as it is least significant
wp4 <- lm("log10(price) ~ year + h.rain + s.temp + h.temp + parker",
data=wine2)
summary(wp4)
plotter(wp4, "wp4_plot.png", "wp4_residuals.png")

# removing h.rain from the model as it is least significant
wp5 <- lm("log10(price) ~ year + s.temp + h.temp + parker", data=wine2)
summary(wp5)
plotter(wp5, "wp5_plot.png", "wp5_residuals.png")

# removing h.temp from the model as it is least significant
wp6 <- lm("log10(price) ~ year + s.temp + parker", data=wine2)
summary(wp6)
plotter(wp6, "wp6_plot.png", "wp6_residuals.png")

# removing s.temp from the model as it is least significant
wp7 <- lm("log10(price) ~ year + parker", data=wine2)
summary(wp7) # R2 reduced from wp6. Dont remove s.temp
plotter(wp7, "wp7_plot.png", "wp7_residuals.png")

# we confirm our choice of variables by performing a stepwise AIC analysis
library("MASS")
wp0 <- lm("price ~ 1", data=wine2)
stepAIC(wp0, ~year + h.rain + s.temp + w.rain + h.temp + parker, data=wine2,
direction="both")

anova(wp6)   # ANOVA analysis

library("car")
car::vif(wp6)      # VIF 1.290710 1.431326 1.425071

# prediction analysis
test <- data.frame(price=27, year=1978, parker=86, s.temp=15.8)
test <- data.frame(price=17, year=1988, parker=87.6, s.temp=17.1)
test <- data.frame(price=11, year=1984, parker=65, s.temp=16.5)
confidence <- predict(wp6, test, se.fit=T, interval=c("confidence"))
prediction <- predict(wp6, test, se.fit=T, interval=c("prediction"))
```

Question 2: CO2 Emissions Analysis

2a) $CO2 = \beta_0 + \beta_1 income + \beta_2(fwd = yes) + \gamma_1(belief = yes) + \gamma_2(belief = no) + \epsilon$

A linear regression model is fit with Income, fwd, belief as the predictor variables and CO2 emissions in tonnes per year as the response variable. Since fwd, and belief are categorical variables, I have represented the statistical model using dummy variables for these two.

Assumptions of the model:

- The relationship between response and predictor variables is linear
- That the errors are normally distributed with zero mean
- The errors have constant variance i.e., homoscedasticity of errors
- The observations are independent of each other

2b)

Residual vs Fitted graph: The 'U' shape in the plot indicates a non-linear relationship between predictor variables and response variable. It also shows that the constant variance assumption is not true.

Normal Q-Q: Overall the residuals are normally distributed with deviation towards the end after the second quantile

Scale-Location: Ideally it should show a random spread but we observe an upward trend violating the assumption of homoscedasticity (constant variance). This could be improved by addressing the non-linearity

Residual Vs Leverage: There are a few outlying values but their Cook's distance is <0.5 indicating that they aren't highly influential to the regression fit. Usually, the data point with Cook's distance >0.5 needs to be investigated, >1 is likely to be influential

Residual Vs Belief: Ideally, for categorical variables, the mean of the residuals for each level should be 0 with a similar IQR for a good fit. The plot shows that IQRs are slightly different for the levels and strictly not centred at 0. Overall, it looks okay

Residual Vs Income: There is a clear quadratic distribution of the residuals with respect to income as indicated by the 'U' shape of the plot

2c) $CO2 = \beta_0 + \beta_1 income + \beta_3 income^2 + \beta_2(fwd = yes) + \gamma_1(belief = yes) + \gamma_2(belief = no) + \epsilon$

A quadratic term with respect to income has been added to the previous regression model (from 2a) to address the issues seen on the residual plots above

2d)

Residual vs Fitted graph: The fit looks better now as the residuals are randomly scattered around 0 in comparison to the previous plot.

Normal Q-Q: Here too, the residuals are normally distributed with the deviations reduced from the previous plot

Scale-Location: There is a clear improvement from the previous plot after addressing the non-linearity showing that the spread of residuals is roughly equal for all the fitted values (constant variance of error term)

Residual Vs Leverage: There are no outlying values in this regression fit as the cook's distance is <0.5 for all the data points

Residual Vs Belief: In comparison to the previous plot, the IQR are similar for all levels and closely distributed around 0 with no presence of outliers. This indicates an improvement in the fit

Residual Vs Income: There is no clear pattern, the residuals are randomly scattered around 0 indicating that the quadratic model is a good fit

Residual Vs fwd: The plot shows that IQRs for both the levels of fwd are distributed around 0 with a slight difference in variance between the levels. Overall, it looks like a good fit with respect to fwd

2e) While sequentially increasing the variables of the model, we notice that the Sum of square of residual errors decreases and the corresponding p-values for F- Statistics are significant for each model. However, when belief is added to the model, although SSE drops, its p-value is not significant. This indicates that we cannot reject the null hypothesis that there is no relationship between belief and C02. Hence the variable belief should be dropped from the model.

2f) Fitting a model with only belief as the predictor shows that it is not statistically significant, hence it is safe to exclude this from the model.

The table shows that there is a correlation between fwd and belief. The majority of people who own four-wheel drive do not believe/do not understand the science. The majority of people who do not own four-wheel drive believe in the science.

This could indicate that the belief data is being captured by the fwd variable.

2g) $$CO2 = \beta_0 + \beta_1 income + \beta_3 income^2 + \beta_2(fwd = yes) + \epsilon$$

Belief has been dropped from the model as it did not have significance and was being explained by fwd.
The ANOVA table shows that all three terms when added sequentially (income, quadratic term of income, fwd) have significance with SSE decreasing. Hence all three terms can be included in the model

2h) $$CO2 = \beta_0 + \beta_1 income + \beta_3 income^2 + \beta_2(fw) + \delta(fwd * income) + \epsilon$$

This model incorporates an interaction term between fwd and income.

Drawing the sequential ANOVA table shows that the interaction term is not significant (checking p-value for F statistic) and can be removed from the model

2i) The model is a good fit as all the predictor variables are statistically significant, 94.62% of the variance in the data is being explained by this model, F-stat value is high with less p-value indicating that the null hypothesis (at least one of the coefficients is zero) can be rejected.

The residual standard error is 2.028 i.e., it predicts CO2 with an average error of 2.028 tonnes per year.

Irrespective of income and fwd, the base CO2 emission for a household is 1.603 tonnes per year

Relationship between CO2 and fwd: Keeping income constant, a household with a fwd produces 1.831 tonnes per year of CO2 more than the household without fwd drive

Relationship between CO2 and Income: There exists a quadratic relationship between Income and CO2 emissions. Keeping fwd constant, the increase in CO2 emissions in tonnes per year for an increase in income by x (in thousands of pounds) from an income I (in thousands of pounds) turns out to be –

$$\Delta CO2 = 0.0860234 * x + 2 * (0.0075880) * I * x + 0.0075880 * x^2$$