# ST3131 Regression Analysis Assignment

*Douglas Wei Jing Allwood (A0183939L)*

*April 2020*

## 1. Introduction

Our goal is to propose a model that accurately describes the Quality of wine using information pertaining to its clarity, aroma, body, flavor, oakiness, and the region it was from.

Specifically, we want to estimate a model of the following form:

$$Quality = \beta_0+\beta_1(Clarity)+\beta_2(Aroma)+\beta_3(Body)+\beta_4(Flavor)+\beta_5(Oakiness)+\beta_6(Quality)+\beta_7(Region)$$

$$\beta_i \in \mathbb{R}, \forall i$$

Note that the form of the model shown above could be expanded to include interaction terms in the form of $\beta_8 * Clarity * Body$ or even higher order terms such as $\beta_9 * Clarity^2$. Such additions will be considered when necessary. For example, when trying to correct for linearity or constant variance between regressors and response.

To produce a model that can go beyond just accurately describing the given (training) data, to also accurately make new predictions or mild extrapolations, the model produced must be adequate for the given problem. This is especially important given that the dataset is relatively small, only having 38 samples. Hence, we will decide on a model by prioritising model adequacy over data-specific metrics such as the $R^2$ coefficient or hypothesis testing performed on model parameters. Though these will still be used to refined the adequate model once it is found.

## 2. Data exploration before modelling

This is what a sample of the given data looks like:

```
##   Clarity Aroma Body Flavor Oakiness Quality Region
## 1       1   3.3  2.8    3.1      4.1     9.8      1
## 2       1   4.4  4.9    3.5      3.9    12.6      1
## 3       1   3.9  5.3    4.8      4.7    11.9      1
## 4       1   3.9  2.6    3.1      3.6    11.1      1
## 5       1   5.6  5.1    5.5      5.1    13.3      1
```

We notice that *Clarity* and *Region* take on a limited range of values:

*Region*: 1, 2, 3
Almost certainly a categorical regressor as it describes the location where the wine originated from.

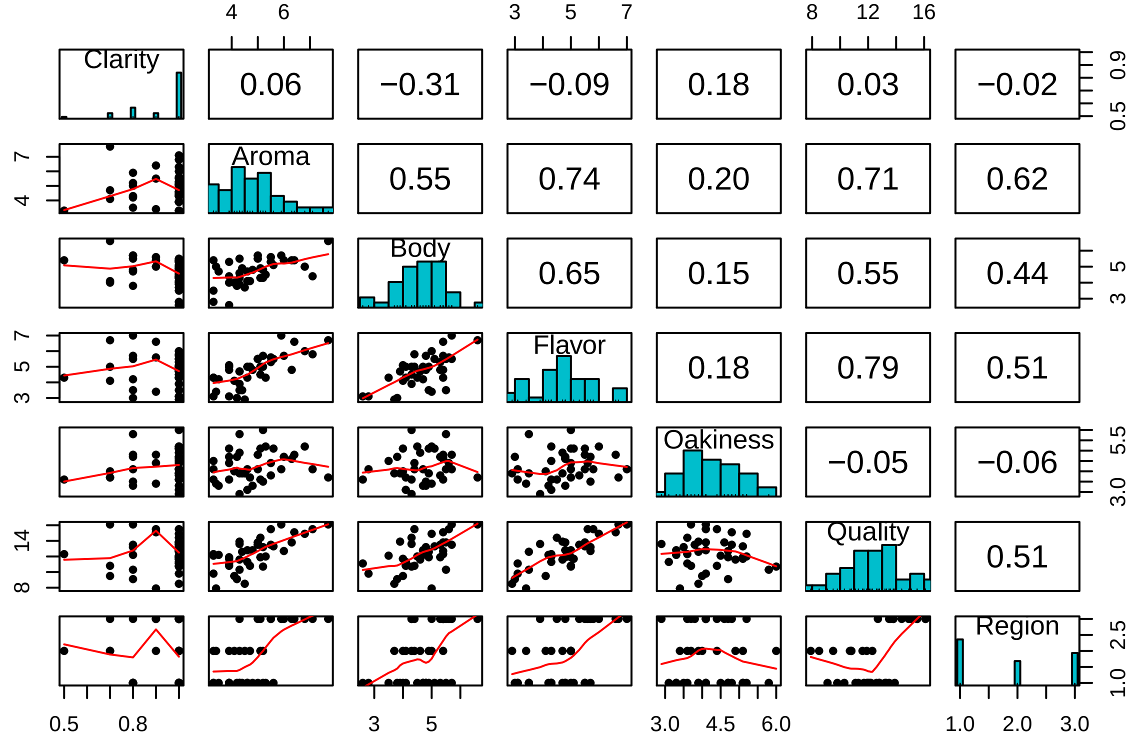
*Clarity*: 0.5, 0.7, 0.8, 0.9, 1
Likely possess an ordering where a wine with a lower clarity would be assigned a lower value, vice versa. Since an ordering could exist, this variable is unlikely to be categorical.

Hence, we transform *Region* into a categorical variable.

### 2.1 Understanding relationships in the data

We now want to understand the relation between all the variables (regressors and response) provided.

The following graph shows a grid of the correlations and scatter plots between variables, and the histogram of each individual variable.



The histogram of *Quality* appears symmetric, almost Normally distributed.

We observe that *Quality* appears to be positively linearly related to *Aroma* and *Flavor* as seen in both the scatter plots and their high correlation values with *Quality*. 0.71 and 0.79, respectively. This indicates that *Aroma* and *Flavor* may be useful in predicting the *Quality* of a wine sample. However, *Aroma* and *flavor* are also highly correlated (0.74), this reveals the potential for mutlicollinearity which could increase the variance of fitted parameters and hence decrease the effectiveness of a fitted model when performing mild extrapolation. Hence, we must also try to detect for mutlicolinearity in the data exploration phase and address it in the model building if it exists.

The relationship between *Quality* and the other regressors is harder to determine and may exist with 3 or more variables, interaction terms, or perhaps not exist at all.

Generally, if we consider the scatter plots along the *Quality* row, it does not appear necessary to transform the regressors as the scatter plots either already show some (weak) linear relationship with Quality, or no relationship at all. We will revisit the need to transform variables during residual analysis.

Note that the scatter plots above can only help us to detect if a 2D relationship exists between each variableand will fail to detect if relationships exist in 3D or higher. Hence, we cannot drop variables yet or be certain about the extent of multicollinearity in the data.

## 2.2 Detecting multicollinearity

**The code for detecting multicollinearity is in the Appendix**

The condition indices are: 1, 1.911732, 2.998716, 6.89035, 9.729912

The VIF values are: 1.26639, 2.381143, 2.056492, 2.682277, 1.096731

The small condition number (9.73) and small VIF values (< 10) above suggest that multicollinearity either does not exist in the given data or is very weak. Either way, this suggests that we do not have to worry about multicollinearity in the building of the model.

Hence, it appears that the correlation between regressors does not actually seem to be causing multicollinearity problems as previously feared.

# 3. Perform initial Linear Regression and check residuals

We now apply the standard Linear Regression model in R with the aim of studying the residuals and correcting any violations of our assumptions by transforming the variables.

```
model1 <- lm(Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region, data = wine)
summary(model1)
```

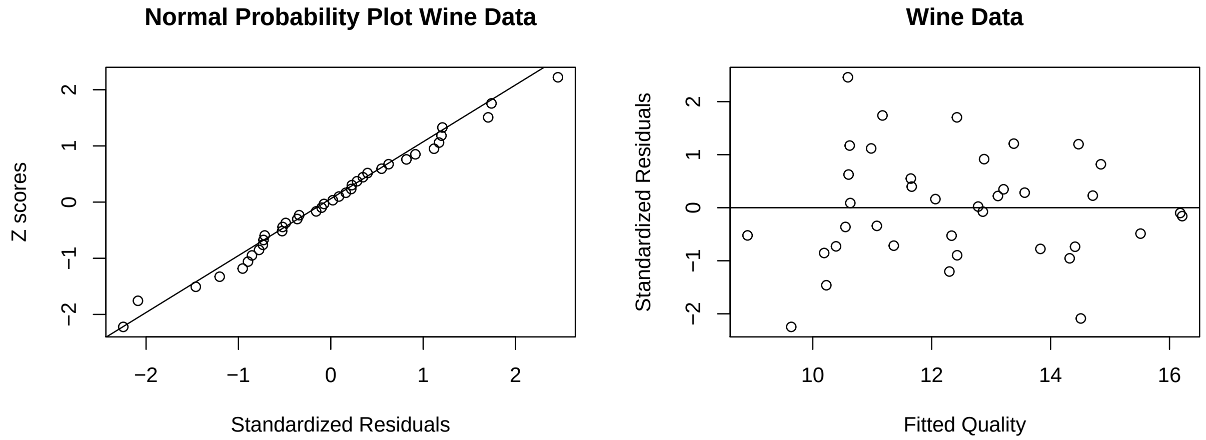
```
##
## Call:
## lm(formula = Quality ~ Clarity + Aroma + Body + Flavor + Oakiness +
##     Region, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80824 -0.58413 -0.02081  0.48627  1.70909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.81437    1.96944   3.968 0.000417 ***
## Clarity      0.01705    1.45627   0.012 0.990736
## Aroma        0.08901    0.25250   0.353 0.726908
## Body         0.07967    0.26772   0.298 0.768062
## Flavor       1.11723    0.24026   4.650 6.25e-05 ***
## Oakiness    -0.34644    0.23301  -1.487 0.147503
## Region2     -1.51285    0.39227  -3.857 0.000565 ***
## Region3      0.97259    0.51017   1.906 0.066218 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9154 on 30 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.7997
## F-statistic:  22.1 on 7 and 30 DF,  p-value: 3.295e-10
```

**Please find the Anova table of Model1 in the Appendix if needed**

Referencing the Summary table of model1 there are several important observations:

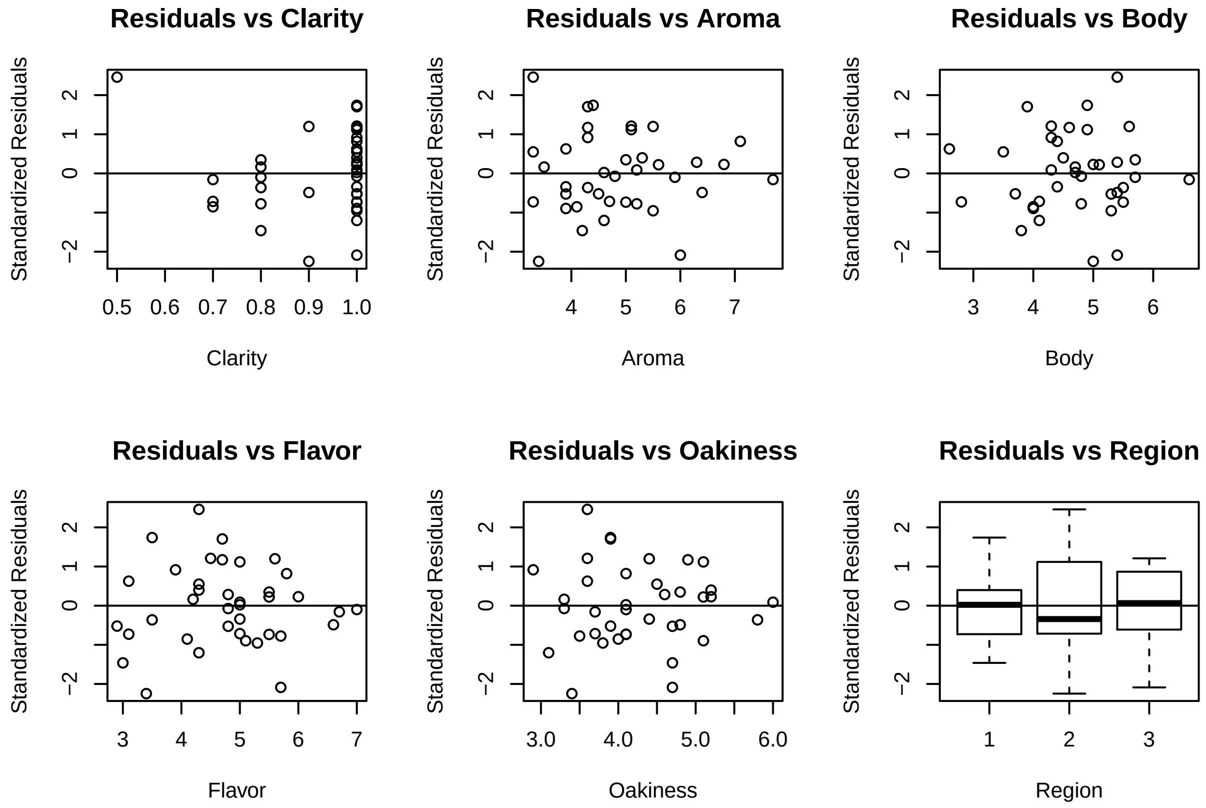
1. The F-test on the significance of the model shows that model1 is significant
2. As previously speculated, Flavor does help in predicting wine Quality as seen in its statistically significant non-zero estimated coefficient, $\hat{\beta_4}$. (P-value $6.25 * 10^{-5}$ for the t-test)
3. Contrary to what was expected, the coefficient for Aroma has a p-value of 0.72 for the t-test of significance which suggests that this variable could be dropped from the model. This may warrant further transformations of the Aroma variable.
4. There are several variables that have t-test results suggesting that they are insignificant. We could consider variable selection techniques later on as filtering out variables would improve the AIC|BIC of our model by producing a smaller AIC|BIC value.

## 3.1 Analysis of model graphs



In the Normal Probability QQ plot we see that the 38 data points closely follow the line $y = x$. Hence, our assumption that the errors are normally distributed seems to hold.

Additionally, the standardized residuals seem to be randomly distributed about the $y = 0$ line when plotted against the fitted *Quality* values. Majority of the points fall within the range $-2 \leq e_i \leq 2$, with the exception of one point greater than 2 and two points less than -2. If we consider these three points to be outliers, this proportion of outliers $\frac{3}{38} \approx 0.079$ is reasonable for a normal distribution. As we would expect approximately 95% of the data to fall within 2 standard deviations and 5% to fall outside.



From the individual Residual against Regressor plots we see that the variance of the residuals appear to be related to Clarity in an widening funnel shape, and Aroma in a narrowing funnel shape. Hence, we can consider transforming these regressors or dropping them entirely.

The rest of the plots show a random distribution of residuals on both sides of the $y = 0$ line, which agrees with our assumption that residuals are independent of these regressors.

## 4. Transformation of variables

To decide on an appropriate transformation of the variables **Clarity** and **Aroma** we can use the Box-Tidwell procedure.

**Please see the appendix for the code for the Box-Tidwell procedure**

The results of the Box-Tidwell procedure are as follows:

- **Clarity**: $Clarity^{1853.6}$
- **Aroma**: $Power^{1.0070}$

As *Clarity* takes on a limited range of values between 0 and 1, raising it to a large power would send the values to 0 if they are less than 1. This suggests that we either remove the *Clarity* variable or transform it into an indicator: 1 if *Clarity* is 1, and 0 otherwise. However, this transformation to a indicator would not make sense in the context of what the variable is supposed to be - an ordered numerical value indicating the clarity of a wine sample. On, the other hand, due to the large p-value of *Clarity* in the t-test of the model1 summary (p-value of 0.99), we can choose to drop *Clarity*.

For *Aroma*, the result of the Box-Tidwell procedure suggests that no transformation is necessary. Hence, we only have to decide whether to keep it or drop it.

To analytically decide whether to drop these variables, we can use the AIC and BIC metrics via step regression.

## 5. AIC and BIC step regression

```
model4 <- step(model1, direction = c("both"))
summary(model4)
```

```
##
## Call:
## lm(formula = Quality ~ Flavor + Oakiness + Region, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81290 -0.59794  0.03423  0.42452  1.71484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.1208     1.0164   7.990 3.23e-09 ***
## Flavor        1.1920     0.1772   6.727 1.15e-07 ***
## Oakiness     -0.3183     0.2039  -1.561 0.128060
## Region2      -1.5155     0.3614  -4.193 0.000194 ***
## Region3       1.0935     0.4009   2.728 0.010130 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8763 on 33 degrees of freedom
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.8164
## F-statistic: 42.14 on 4 and 33 DF,  p-value: 1.595e-12
```

**Please see the appendix for the intermediate results of step regression of model4**

The model produced by step regression dropped both *Clarity* and *Aroma*. Additonally, it dropped the *Body* variable.

The final model produced by step regression has an approximately equal $R^2$ value as model1, the full model. (*model1* $R^2 = 0.8376$, *model4* $R^2 = 0.8363$)

# 6. Recommendation

Overall, we saw in the Residual against Regressor graphs that *Clarity* and *Aroma* violate the assumption that residuals are independent of the regressors. Additionally, the t-test and step regression both suggest that we drop these variables. Hence, we will drop them from the final model.

For the *Body* variable, though the t-test and step regression suggest that we drop the variable, as the Residual against *Body* graph does not show any violation of our assumptions, we will keep the variable. The reason for keeping the variable is that it shows a weak linear relationship with *Quality* with a correlation of 0.55. Hence, the variable might be a weak but useful indicator of *Quality*. If we were to collect more data, we could reevaluate keeping this variable.

# Final Model

```
model5 <- lm(Quality ~ Body + Flavor + Oakiness + Region, data = wine)
summary(model5)
```

```
##
## Call:
## lm(formula = Quality ~ Body + Flavor + Oakiness + Region, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8197 -0.5784  0.0102  0.4969  1.6320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.94075    1.15130   6.897 8.31e-08 ***
## Body         0.08283    0.23648   0.350 0.728431
## Flavor       1.15607    0.20689   5.588 3.59e-06 ***
## Oakiness    -0.32450    0.20744  -1.564 0.127570
## Region2     -1.52294    0.36695  -4.150 0.000229 ***
## Region3      1.06728    0.41315   2.583 0.014562 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8882 on 32 degrees of freedom
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8114
## F-statistic: 32.84 on 5 and 32 DF,  p-value: 1.052e-11
```

The final model we produce is:

$$\hat{Quality} = 7.94 + 0.08(Body) + 1.16(Flavor) - 0.32(Oakiness) - 1.52(I(Region == 2)) + 1.07(I(Region == 3))$$

Where $I(Region == i) = 1$ if Region = i and 0 otherwise.
Note that the final model is statistically significant according to the F-test as it has p-value $= 1.52 * 10^{-11}$.

# Appendix

All code and auxilary tables (e.g. Anova tables that were not referenced but may be of interest)

## Import data and libraries

```
library(dplyr)
library(ggplot2)
library(psych)

wine <- read.csv("../data/wine.csv")
attach(wine)

wine$Region <- as.factor(wine$Region)
attach(wine)
```

## Multicollinearity

```
# Centering data: Subtract the mean, divide by sqrt(S_xx)
# sqrt(S_xx) = (n-p) * MS_res
# n = 38, p = 7, n-p = 31
q <- Quality
C <- (Clarity-mean(Clarity))/(sqrt(var(Clarity))*37)
A <- (Aroma - mean(Aroma))/(sqrt(var(Aroma))*37)
B <- (Body - mean(Body))/(sqrt(var(Body))*37999)
FL <- (Flavor - mean(Flavor))/(sqrt(var(Flavor))*37)
O <- (Oakiness - mean(Oakiness))/(sqrt(var(Oakiness))*37)

x <- cbind(C, A, B, FL, O)
x <- cor(x) # Correlation matrix of x, X'X

# The condition number is:
cond <- max(eigen(x)$values)/min(eigen(x)$values)

#condition indices:
cat("The condition indices are: ", max(eigen(x)$values)/eigen(x)$values, "\n")
```

```
## The condition indices are:  1 1.911732 2.998716 6.89035 9.729912
```

```
#### To find VIFs:
C <- solve(x)  #this is (X'X)^(-1)
VIF <- diag(C)
cat("The VIF values are: ", VIF, "\n")
```

```
## The VIF values are:  1.26639 2.381143 2.056492 2.682277 1.096731
```

## Model1

```
model1 <- lm(Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region,
             data = wine)
anova(model1)
```

```
## Analysis of Variance Table
##
```

```
## Response: Quality
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Clarity    1  0.125   0.125  0.1494 0.7018243
## Aroma      1 77.353  77.353 92.3064 1.152e-10 ***
## Body       1  6.414   6.414  7.6544 0.0096032 **
## Flavor     1 19.050  19.050 22.7324 4.484e-05 ***
## Oakiness   1  8.598   8.598 10.2598 0.0032129 **
## Region     2 18.108   9.054 10.8042 0.0002924 ***
## Residuals 30 25.140   0.838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Box-Tidwell on Clarity variable

```
# For the Clarity variable
model2 <- lm(Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region,
             data = wine)
model3 <- lm(Quality ~ Clarity + I(Clarity * log(Clarity)) + Aroma + Body +
             Flavor + Oakiness + Region, data = wine)
gamma1 <- model3$coefficients["I(Clarity * log(Clarity))"]
beta1 <- model2$coefficients["Clarity"]
power1 <- as.numeric((gamma1/beta1) + 1)
#power1
```

### Box-Tidwell on Aroma variable

```
# For the Aroma variable
model2 <- lm(Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region,
             data = wine)
model3 <- lm(Quality ~ Clarity + Aroma + I(Aroma * log(Aroma)) + Body +
             Flavor + Oakiness + Region, data = wine)
gamma2 <- model3$coefficients["I(Aroma * log(Aroma))"]
beta2 <- model2$coefficients["Aroma"]
power2 <- as.numeric((gamma2/beta2) + 1)
#power2
```

### AIC and BIC step regression

```
model4 <- step(model1, direction = c("both"))
```

```
## Start:  AIC=0.3
## Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region
##
##            Df Sum of Sq    RSS     AIC
## - Clarity   1    0.0001 25.140 -1.6984
## - Body      1    0.0742 25.214 -1.5866
## - Aroma     1    0.1041 25.244 -1.5415
## <none>                  25.140  0.3014
## - Oakiness  1    1.8525 26.993  1.0031
## - Region    2   18.1079 43.248 16.9159
## - Flavor    1   18.1210 43.261 18.9274
##
## Step:  AIC=-1.7
## Quality ~ Aroma + Body + Flavor + Oakiness + Region
##
##            Df Sum of Sq    RSS     AIC
## - Body      1    0.0864 25.227 -3.5680
## - Aroma     1    0.1048 25.245 -3.5404
## <none>                  25.140 -1.6984
## - Oakiness  1    2.0316 27.172 -0.7454
## + Clarity   1    0.0001 25.140  0.3014
## - Flavor    1   18.1527 43.293 16.9554
## - Region    2   20.5655 45.706 17.0162
##
## Step:  AIC=-3.57
## Quality ~ Aroma + Flavor + Oakiness + Region
##
##            Df Sum of Sq    RSS     AIC
## - Aroma     1    0.1151 25.342 -5.3949
## <none>                  25.227 -3.5680
## - Oakiness  1    1.9841 27.211 -2.6909
## + Body      1    0.0864 25.140 -1.6984
## + Clarity   1    0.0123 25.214 -1.5866
## - Region    2   20.6267 45.853 15.1388
## - Flavor    1   23.2503 48.477 19.2531
##
## Step:  AIC=-5.39
## Quality ~ Flavor + Oakiness + Region
##
##            Df Sum of Sq    RSS     AIC
## <none>                  25.342 -5.3949
## - Oakiness  1     1.871 27.213 -4.6877
## + Aroma     1     0.115 25.227 -3.5680
## + Body      1     0.097 25.245 -3.5404
## + Clarity   1     0.010 25.331 -3.4107
## - Region    2    27.114 52.456 18.2508
## - Flavor    1    34.753 60.095 25.4169
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = Quality ~ Flavor + Oakiness + Region, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81290 -0.59794  0.03423  0.42452  1.71484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.1208     1.0164   7.990 3.23e-09 ***
## Flavor        1.1920     0.1772   6.727 1.15e-07 ***
## Oakiness     -0.3183     0.2039  -1.561 0.128060
## Region2      -1.5155     0.3614  -4.193 0.000194 ***
## Region3       1.0935     0.4009   2.728 0.010130 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.8763 on 33 degrees of freedom
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.8164
## F-statistic: 42.14 on 4 and 33 DF,  p-value: 1.595e-12
```

### Final model

```
model5 <- lm(Quality ~ Body + Flavor + Oakiness + Region, data = wine)
anova(model5)
```

```
## Analysis of Variance Table
##
## Response: Quality
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Body       1 46.603  46.603 59.0726 9.236e-09 ***
## Flavor     1 50.393  50.393 63.8764 4.008e-09 ***
## Oakiness   1  5.873   5.873  7.4441   0.01025 *
## Region     2 26.675  13.338 16.9065 9.759e-06 ***
## Residuals 32 25.245   0.789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```