MAT3375 Project

David D'Souza Jason Lam

December 12, 2020

About the dataset

The Wine Quality dataset was chosen from Kaggle. This dataset describes how much citric acid, residual sugar is in each wine as well as its pH, density, achohol level and the quality (based on a scale from 1 to 10). We created a good predicter for the quality of wine using multiple linear regression.

Assumptions

There are 5 main assumptions that will be tested throughout the report to ensure our model is adequate. These assumptions are:
1. The relationship between the response y and the regressors is linear, at least approximately.
2. The error term has zero mean.
3. The error term has constant variance.
4. The errors are uncorrelated.
5. The errors are normally distributed.

Importing packages

```
library(faraway)
library(MPV)
```

Importing data

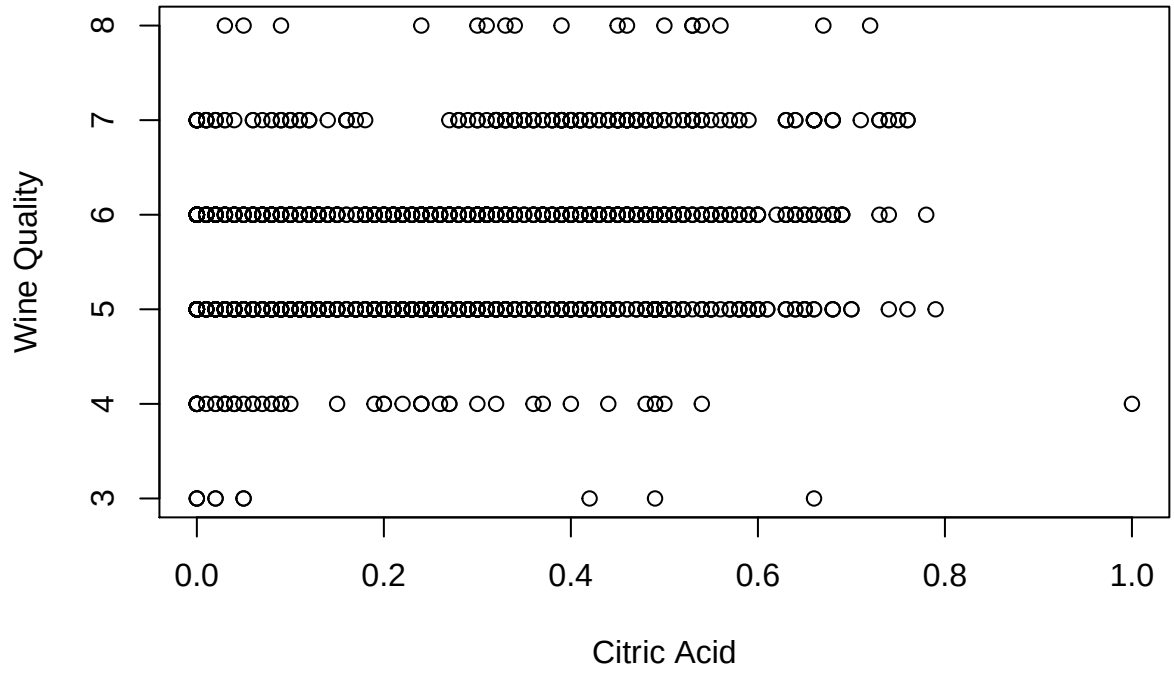
```
m <-  read.csv("./Wine_Quality.csv")
p = 6
k = 5
y = m$quality
n = length(y)
one_vector = rep(1, 1599)
```

Correlation between Regressors and Wine Quality

```
cor(m$citric.acid,y)
```

```
## [1] 0.2263725
```

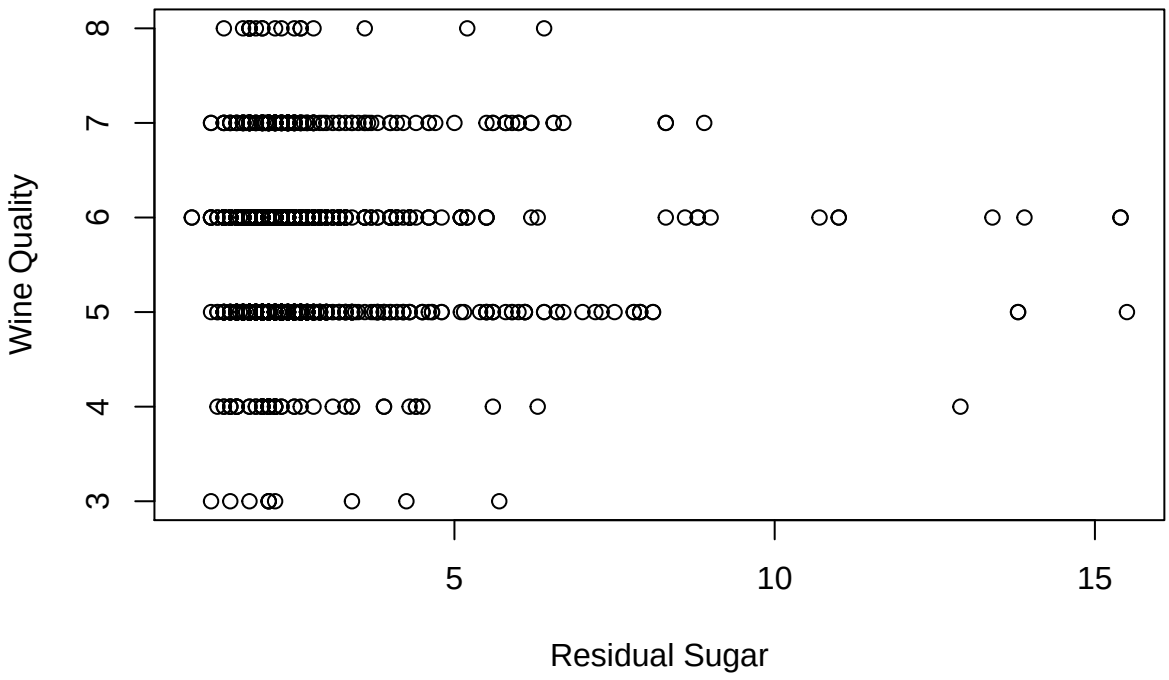
```
plot(m$citric.acid,y,xlab="Citric Acid",ylab="Wine Quality")
```




```
cor(m$residual.sugar,y)
```

```
## [1] 0.01373164
```

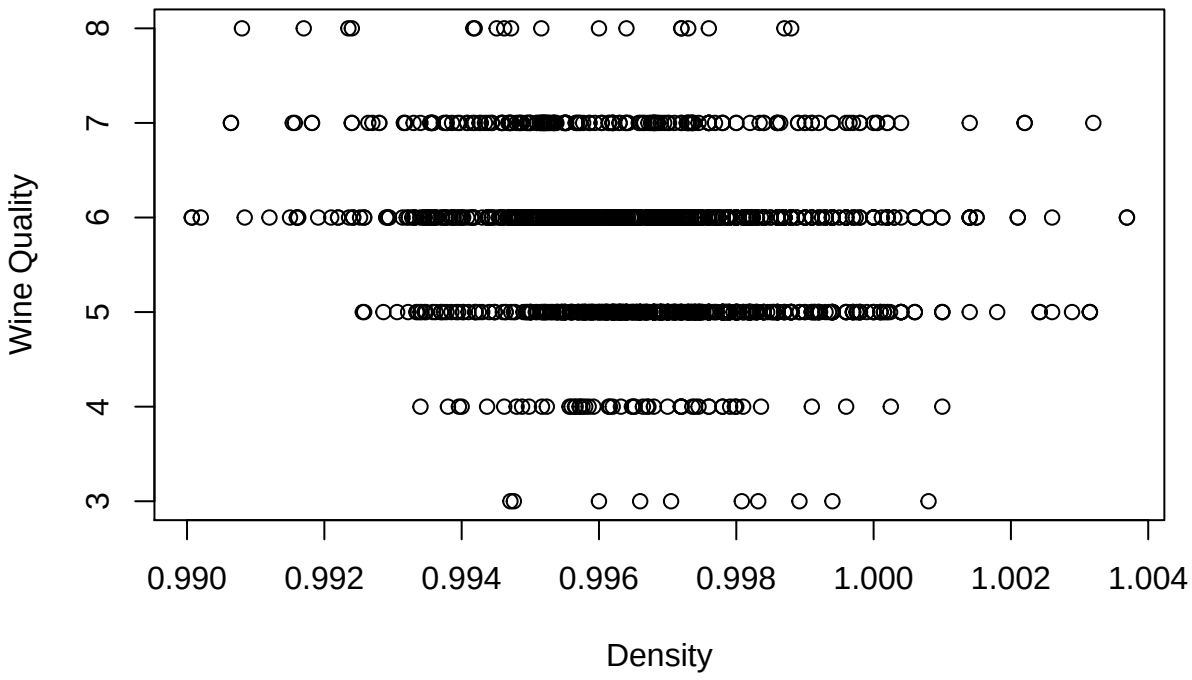
```
plot(m$residual.sugar,y,xlab="Residual Sugar",ylab="Wine Quality")
```




```
cor(m$density,y)
```

```
## [1] -0.1749192
```

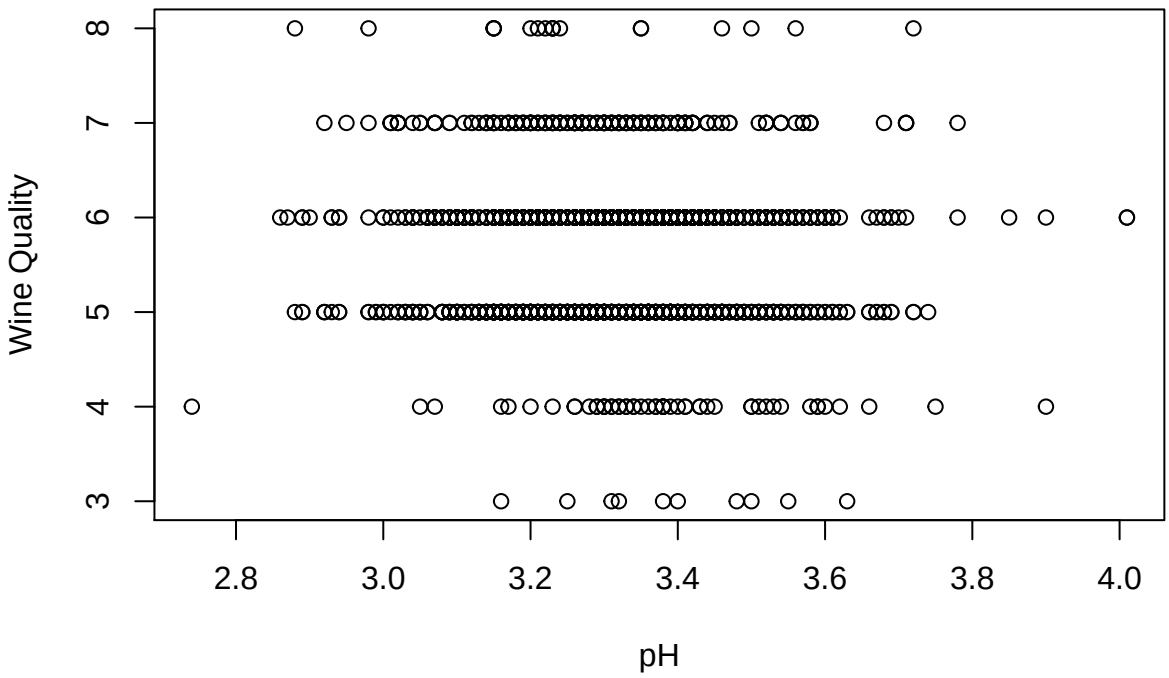
```
plot(m$density,y,xlab="Density",ylab="Wine Quality")
```




```
cor(m$pH,y)
```

```
## [1] -0.05773139
```

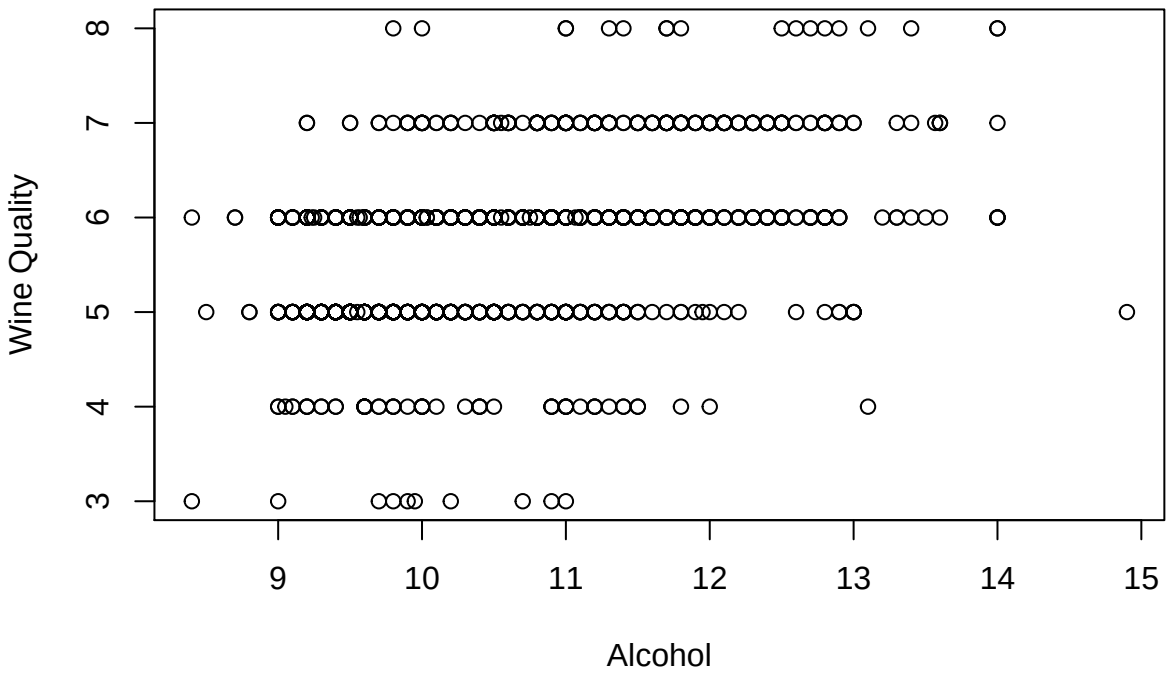
```
plot(m$pH,y,xlab="pH",ylab="Wine Quality")
```




```
cor(m$alcohol,y)
```

```
## [1] 0.4761663
```

```
plot(m$alcohol,y,xlab="Alcohol",ylab="Wine Quality")
```



Creating a Model with all Regressors

```
model <- lm(y~m$citric.acid+m$residual.sugar+m$density+m$pH+m$alcohol)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ m$citric.acid + m$residual.sugar + m$density +
##     m$pH + m$alcohol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6102 -0.4140 -0.1204  0.5173  2.4310
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.22941   13.44880   0.389 0.697448
## m$citric.acid      0.54525    0.12137   4.493 7.55e-06 ***
## m$residual.sugar -0.01795    0.01374  -1.307 0.191538
## m$density         -1.94108   13.38005  -0.145 0.884672
## m$pH              -0.46668    0.14116  -3.306 0.000968 ***
## m$alcohol          0.36308    0.02171  16.726  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6942 on 1593 degrees of freedom
## Multiple R-squared:  0.2634, Adjusted R-squared:  0.2611
## F-statistic: 113.9 on 5 and 1593 DF,  p-value: < 2.2e-16
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##                     Df Sum Sq Mean Sq  F value Pr(>F)
## m$citric.acid        1  53.41  53.405 110.8280 <2e-16 ***
## m$residual.sugar     1   0.37   0.375   0.7780 0.3779
## m$density            1  85.42  85.416 177.2582 <2e-16 ***
## m$pH                 1   0.54   0.537   1.1136 0.2915
## m$alcohol            1 134.80 134.805 279.7504 <2e-16 ***
## Residuals         1593 767.63   0.482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
X = cbind(one_vector, m$citric.acid, m$residual.sugar, m$density, m$pH, m$alcohol)
XX = t(X) %*% X
XX_inverse = solve(XX)
hat_matrix = X %*% XX_inverse %*% t(X)
beta = XX_inverse %*% t(X) %*% y
y_hat = X %*% beta
e = y - y_hat
SS_res = t(y) %*% y - t(beta) %*% t(X) %*% y
```

Testing if there is Multicollinearity between any Regressors

None of the regressors have a Variance Inflation Factor over 10, so there does not seem to be any multicollinearity between any regressors.

```
vif(model)
```

```
##    m$citric.acid m$residual.sugar        m$density             m$pH
##         1.853653         1.244392         2.114731         1.575106
##        m$alcohol
##         1.774716
```

Testing for significance of Regression

Let the null hypothesis be that each coefficient of the regression model is equal to zero. The p-value is very small, so we reject the null hypothesis.

```
SS_r = t(y) %*% (hat_matrix - ((n**(-1)) * one_vector %*% t(one_vector) )) %*% y

F_ob = (SS_r/k) / (SS_res/(n-k-1))
F_ob
```

```
##          [,1]
## [1,] 113.9456
```

```
pf(F_ob, df1 = k, df2 = n-k-1, lower.tail = FALSE)
```

```
##               [,1]
## [1,] 3.965263e-103
```

Removing Regressors and Testing for the Best Model

We now remove regressors that have high p-values and thus are not significant in the model.

We start by removing density because it has the highest p-value among the regressors and intercept coefficients.

```
model2 <- lm(y~m$citric.acid+m$residual.sugar+m$pH+m$alcohol)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ m$citric.acid + m$residual.sugar + m$pH + m$alcohol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6166 -0.4127 -0.1192  0.5173  2.4330
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.27951    0.46124   7.110 1.74e-12 ***
## m$citric.acid      0.53804    0.11069   4.861 1.28e-06 ***
## m$residual.sugar -0.01879    0.01245  -1.510 0.131299
## m$pH             -0.46686    0.14112  -3.308 0.000959 ***
## m$alcohol         0.36499    0.01729  21.107  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.694 on 1594 degrees of freedom
## Multiple R-squared:  0.2634, Adjusted R-squared:  0.2616
## F-statistic: 142.5 on 4 and 1594 DF,  p-value: < 2.2e-16
```

We now remove residual sugar because it has the highest p-value among the remaining regressors and intercept coefficients.

```
model3 <- lm(y~m$citric.acid+m$pH+m$alcohol)
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ m$citric.acid + m$pH + m$alcohol)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.6373 -0.4029 -0.1214  0.5146  2.4497
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.23168    0.46033   7.020 3.26e-12 ***
## m$citric.acid  0.52073    0.11014   4.728 2.47e-06 ***
## m$pH          -0.46283    0.14115  -3.279  0.00106 **
## m$alcohol      0.36417    0.01729  21.062  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6942 on 1595 degrees of freedom
## Multiple R-squared:  0.2624, Adjusted R-squared:  0.261
## F-statistic: 189.1 on 3 and 1595 DF,  p-value: < 2.2e-16
```

We now use PRESS residuals and Akaike Information Criterion to test the power of how good each model is at prediction a new value. Both model2 and model3 seem to be equality good. We will use model3 due to it being simplier than model2 due to it having one less regressors than model2.

```
PRESS(model)
```

```
## [1] 774.475
```

```
PRESS(model2)
```

```
## [1] 773.3838
```

```
PRESS(model3)
```

```
## [1] 773.0936
```

```
AIC(model)
```

```
## [1] 3378.372
```

```
AIC(model2)
```

```
## [1] 3376.393
```

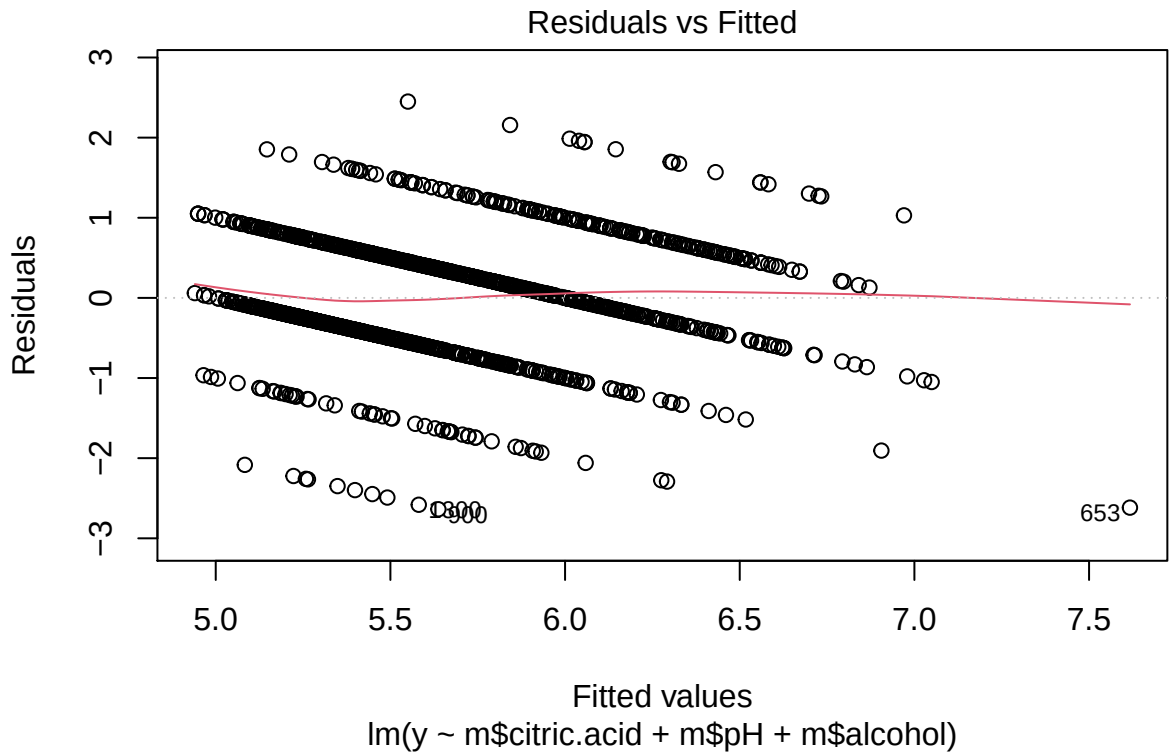
```
AIC(model3)
```

```
## [1] 3376.678
```

Testing Assumptions

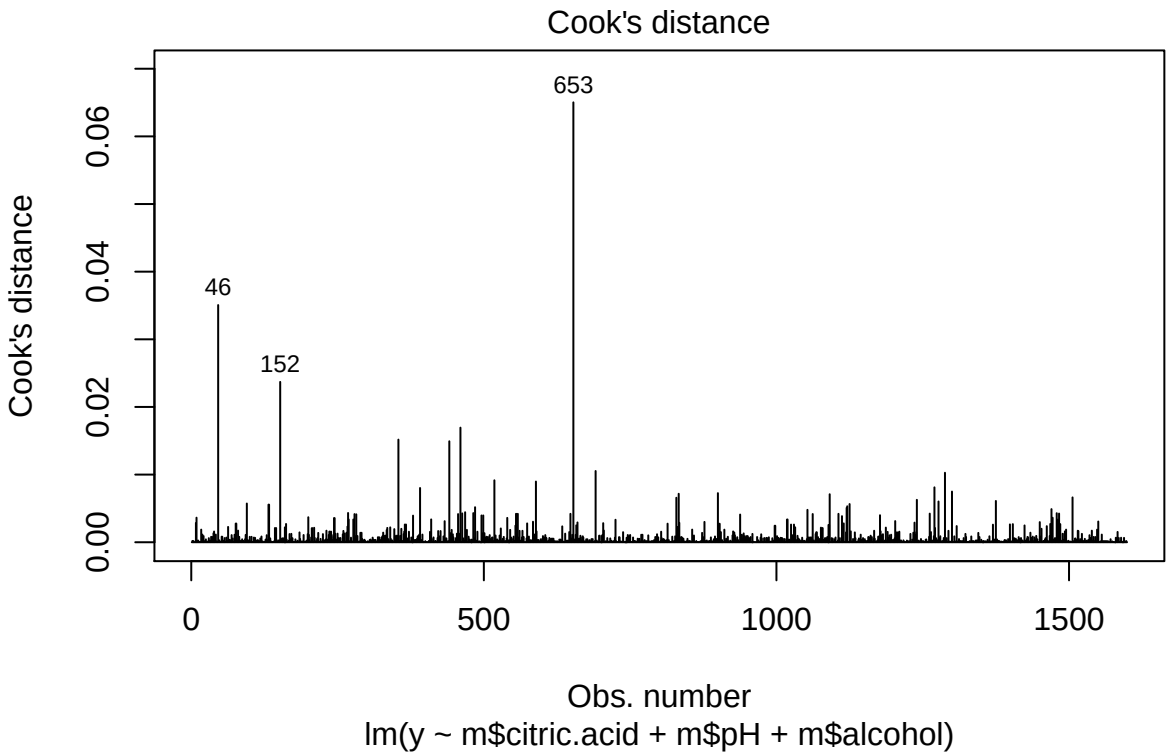
We will show our model does not violate any assumptions.
The response between the response y and the regressors is linear because the Residual vs Fitted plot has no fitted pattern. This also shows the error term has zero mean.

```
plot(model3, 1)
```



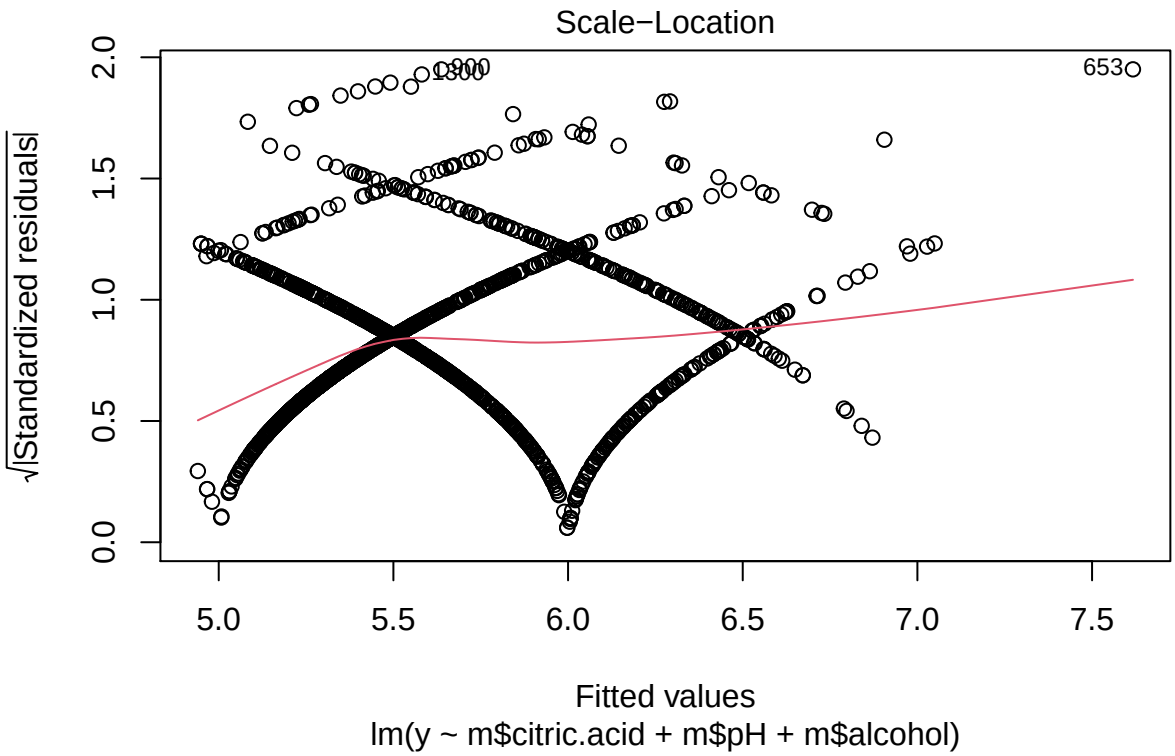
Because the value of the Residual vs Leverage plot is close to zero most of the time, there won't be any extreme values that will affect the regression.

```
plot(model3, 4)
```



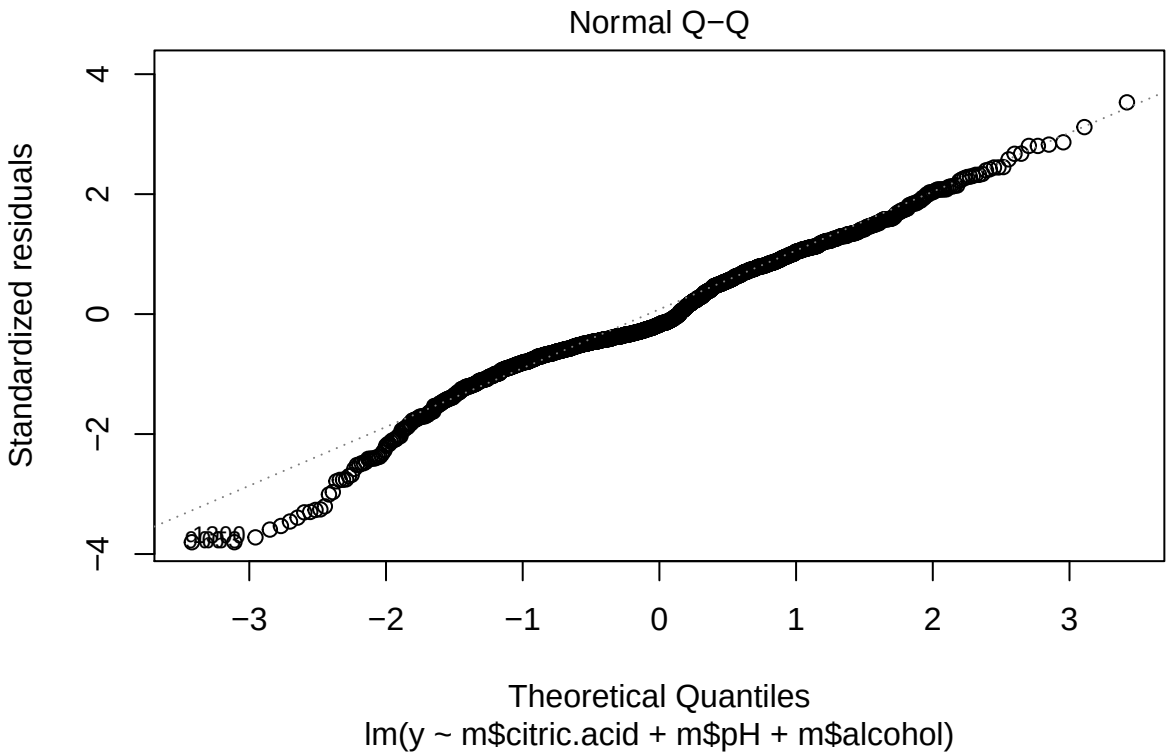
The error term has constant variance because the scale location plot shows the residuals are spread equally along the ranges of predictors.

```
plot(model3, 3)
```



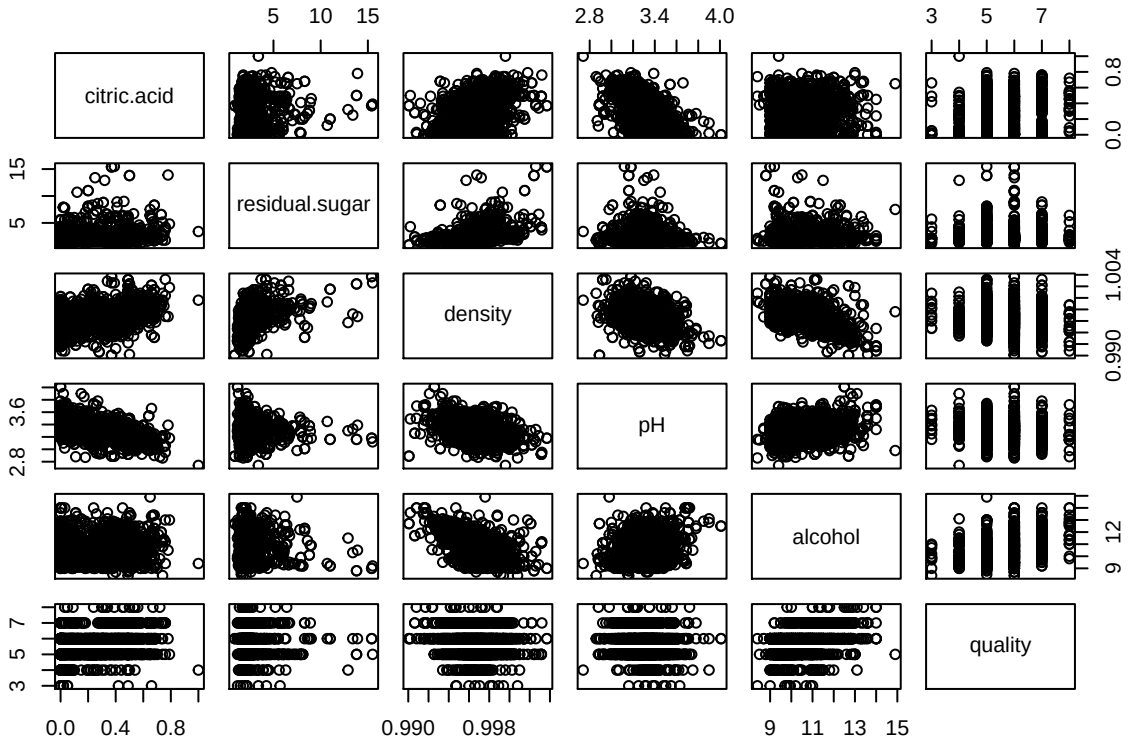
The errors are normally distributed. This is shown in the Q-Q plot showing that the residuals do not stray far from the normal.

```
plot(model3, 2)
```



The correlation matrix shows that there is no significant correlation between each predictor.

```
pairs(m)
```



Therefore, our model does not violate any assumptions.

Finding a Prediction Interval

We first calculate all the values necessary to find a $100(1-\alpha)$ percent prediction interval for an new vector x_0.

```
X = cbind(one_vector, m$citric.acid, m$pH, m$alcohol)
XX = t(X) %*% X
XX_inverse = solve(XX)
beta = XX_inverse %*% t(X) %*% y
SS_res = t(y) %*% y - t(beta) %*% t(X) %*% y
```

Therefore, a $100(1-\alpha)$ percent prediction interval for an new vector x_0 is:
$\hat{y}_0 - t_{\alpha/2,1595}\sqrt{\hat{\sigma}^2(1+x_0'(X'X)^{-1}x_0)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,1595}\sqrt{\hat{\sigma}^2(1+x_0'(X'X)^{-1}x_0)}$
With the following parameters:

$\hat{\beta}$:

```
beta
```

```
##                  [,1]
## one_vector  3.2316762
##             0.5207287
##            -0.4628287
##             0.3641683
```

$\hat{y}_0 = x_0' \hat{\beta}$

$\hat{\sigma}^2$:

```
SS_res
```

```
##          [,1]
## [1,] 768.7351
```

$(X'X)^{-1}$:

```
XX_inverse
```

```
##               one_vector
## one_vector  0.4396738003 -0.057718875 -0.125184861 -0.0008545304
##            -0.0577188753  0.025169517  0.018719215 -0.0010633126
##            -0.1251848612  0.018719215  0.041335705 -0.0016074847
##            -0.0008545304 -0.001063313 -0.001607485  0.0006202856
```