Generalised Linear Models (CW2)

Ezra Nwobodo

2023-12-04

This dataset is related to red Vinho Verde wine samples from the north of Portugal. The goal is to model wine quality based on physicochemical tests and determine which of these variables contribute the most to wine quality (which is determined using sensory data).

There are 11 variables plus 1 response variable and 1599 observations. Every variable is continuous except for the quality score, which is an integer between 0 and 10 with 10 being the best.

Despite a non continuous and bounded response, we will use a simple linear model to analyse this data and also test the model's effectiveness. This is what some of the data looks like:

```r
wine <- read.csv("winequality-red.csv", header=TRUE, sep = ";")
head(wine)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

Before fitting our model, we centre the data in each variable. This allows us to make meaningful inferences from the intercept. For example, the estimated quality score is 5.63602 for wine with the average acidity, average alcohol content etc.

```r
# Center each column except "quality" in the "wine" database
c.wine <- as.data.frame(scale(wine[, -which(names(wine) == "quality")]))

# Add the "quality" column back to the centered data
c.wine$quality <- wine$quality
```

```r
# Fit the basic linear model
fit0 <- lm(quality ~ ., data = c.wine )
summary(fit0)
```

```
##
## Call:
## lm(formula = quality ~ ., data = c.wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.63602    0.01621 347.788  < 2e-16 ***
## fixed.acidity         0.04351    0.04518   0.963   0.3357
## volatile.acidity     -0.19403    0.02168  -8.948  < 2e-16 ***
## citric.acid          -0.03556    0.02867  -1.240   0.2150
## residual.sugar        0.02303    0.02115   1.089   0.2765
## chlorides            -0.08821    0.01973  -4.470 8.37e-06 ***
## free.sulfur.dioxide   0.04562    0.02271   2.009   0.0447 *
## total.sulfur.dioxide -0.10739    0.02397  -4.480 8.00e-06 ***
## density              -0.03375    0.04083  -0.827   0.4086
## pH                   -0.06386    0.02958  -2.159   0.0310 *
## sulphates             0.15533    0.01938   8.014 2.13e-15 ***
## alcohol               0.29433    0.02822  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

All following observations are made assuming the model is valid. Firstly we see that $R^2_{adj} = 0.3561$ which suggests that the model doesn't explain variability in the data very well. However, an F-Statistic of 81.35 on 11 and 1587 degrees of freedom with p-value `< 2.2e-16` suggests that there is strong evidence that at least one of the variables is associated with wine quality.

We can also say that: for any given value of other variables, increasing the `alcohol` content by one point will see the the quality score increase by 0.29433; and under the same conditions, increasing `volatile acidity` by one point will see the quality score decrease by 0.19403.

We can be very confident in these two observations as they display p-values of `< 2e-16` which is essentially 0. This is based on the Wald test which compares

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0, \text{ for individual } j.$$

So if the null hypothesis were true and we were collect another set of observations, then for each variable we'd see a more extreme value of $|t|$ (|10.429| and |-8.498| in our case) with probability 0. So we reject the null hypothesis and say that both `alcohol` content and `volatile acidity` are part of the model.

Now with some of the other variables, we don't have enough evidence to suggest that they contribute significantly to wine quality. For example, `density` has a p-value of $0.406 > 0.05$. In order to improve the interpretability and predictive power of our model, we want to select an appropriate subset of our 11 variables. We can't, however, just remove variables that are not significant since removing 1 variable will change the significance of all the others (due to factors such as correlation between variables).

A methodical way to choose which variables are part of the model is the forward selection algorithm which is based on minimising the Akaike Information Criterion (AIC). The AIC, along with other information criteria such as Mallow's C_p, is used as a measure to compare different models. It has a goodness-of-fit component (residual sum of squares (RSS) in the context of linear modelling) and a penalty on the number of variables included. The penalty is due to the RSS monotonically decreasing with the addition of variables (as otherwise the full model would always be chosen).

Now, we use the forward selection algorithm starting with a null model and ending with a full model containing all of the interaction terms. We include interaction terms in our search so that we can potentially find a more optimal model and we choose forward rather than backwards selection as we know that some variables are highly significant which means that the algorithm won't terminate prematurely (also the backward selection algorithm would take up even more pages in the document than this one).

```
library(MASS)
null <- lm(quality ~ 1, data=c.wine)
full <- lm(quality ~ alcohol*volatile.acidity*sulphates*citric.acid*total.sulfur.dioxide*
                     density*chlorides*fixed.acidity*pH*free.sulfur.dioxide*
                     residual.sugar, data=c.wine)
step(null, scope =list(lower =null, upper = full), direction = "forward")
```

```
## Start:  AIC=-682.5
## quality ~ 1
##
##                        Df Sum of Sq     RSS      AIC
## + alcohol               1   236.295  805.87 -1091.65
## + volatile.acidity      1   158.967  883.20  -945.14
## + sulphates             1    65.865  976.30  -784.89
## + citric.acid           1    53.405  988.76  -764.61
## + total.sulfur.dioxide  1    35.707 1006.46  -736.24
## + density               1    31.887 1010.28  -730.19
## + chlorides             1    17.318 1024.85  -707.29
## + fixed.acidity         1    16.038 1026.13  -705.29
## + pH                    1     3.473 1038.69  -685.84
## + free.sulfur.dioxide   1     2.674 1039.49  -684.61
## <none>                              1042.17  -682.50
## + residual.sugar        1     0.197 1041.97  -680.80
##
## Step:  AIC=-1091.65
## quality ~ alcohol
##
##                        Df Sum of Sq    RSS     AIC
## + volatile.acidity      1    94.074 711.80 -1288.1
## + sulphates             1    44.977 760.89 -1181.5
## + citric.acid           1    31.953 773.92 -1154.3
## + pH                    1    26.362 779.51 -1142.8
## + fixed.acidity         1    24.623 781.25 -1139.3
## + total.sulfur.dioxide  1     8.270 797.60 -1106.2
## + density               1     5.203 800.67 -1100.0
## <none>                              805.87 -1091.7
## + chlorides             1     0.611 805.26 -1090.9
## + free.sulfur.dioxide   1     0.325 805.55 -1090.3
## + residual.sugar        1     0.041 805.83 -1089.7
##
## Step:  AIC=-1288.14
## quality ~ alcohol + volatile.acidity
```

```
##
##                            Df Sum of Sq    RSS     AIC
## + sulphates                 1   19.6916 692.10 -1331.0
## + total.sulfur.dioxide      1    6.3730 705.42 -1300.5
## + pH                        1    5.9515 705.84 -1299.6
## + fixed.acidity             1    5.7061 706.09 -1299.0
## + density                   1    1.9410 709.86 -1290.5
## <none>                                 711.80 -1288.1
## + free.sulfur.dioxide       1    0.6621 711.13 -1287.6
## + alcohol:volatile.acidity  1    0.4568 711.34 -1287.2
## + chlorides                 1    0.3762 711.42 -1287.0
## + citric.acid               1    0.1936 711.60 -1286.6
## + residual.sugar            1    0.0101 711.79 -1286.2
##
## Step:  AIC=-1331
## quality ~ alcohol + volatile.acidity + sulphates
##
##                              Df Sum of Sq    RSS     AIC
## + alcohol:sulphates           1   13.3466 678.76 -1360.1
## + total.sulfur.dioxide        1    8.2176 683.89 -1348.1
## + chlorides                   1    7.4925 684.61 -1346.4
## + fixed.acidity               1    3.3282 688.78 -1336.7
## + pH                          1    3.0454 689.06 -1336.0
## + free.sulfur.dioxide         1    1.1129 690.99 -1331.6
## <none>                                   692.10 -1331.0
## + alcohol:volatile.acidity    1    0.6578 691.45 -1330.5
## + volatile.acidity:sulphates  1    0.2921 691.81 -1329.7
## + citric.acid                 1    0.2522 691.85 -1329.6
## + density                     1    0.2222 691.88 -1329.5
## + residual.sugar              1    0.0143 692.09 -1329.0
##
## Step:  AIC=-1360.13
## quality ~ alcohol + volatile.acidity + sulphates + alcohol:sulphates
##
##                              Df Sum of Sq    RSS     AIC
## + total.sulfur.dioxide        1    9.0169 669.74 -1379.5
## + pH                          1    3.5584 675.20 -1366.5
## + chlorides                   1    3.0289 675.73 -1365.3
## + fixed.acidity               1    2.4299 676.33 -1363.9
## + free.sulfur.dioxide         1    1.5762 677.18 -1361.8
## <none>                                   678.76 -1360.1
## + citric.acid                 1    0.2538 678.50 -1358.7
## + alcohol:volatile.acidity    1    0.0335 678.72 -1358.2
## + residual.sugar              1    0.0126 678.75 -1358.2
## + volatile.acidity:sulphates  1    0.0070 678.75 -1358.2
## + density                     1    0.0045 678.75 -1358.1
##
## Step:  AIC=-1379.52
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     alcohol:sulphates
##
##                                         Df Sum of Sq    RSS     AIC
## + sulphates:total.sulfur.dioxide         1   12.7774 656.96 -1408.3
## + volatile.acidity:total.sulfur.dioxide  1    6.2841 663.46 -1392.6
## + pH                                     1    3.8745 665.87 -1386.8
## + chlorides                              1    3.2846 666.46 -1385.4
## + fixed.acidity                          1    1.3334 668.41 -1380.7
## + free.sulfur.dioxide                    1    1.0286 668.71 -1380.0
## <none>                                               669.74 -1379.5
## + residual.sugar                         1    0.3045 669.44 -1378.2
## + volatile.acidity:sulphates             1    0.1890 669.55 -1378.0
## + alcohol:volatile.acidity               1    0.1688 669.57 -1377.9
## + citric.acid                            1    0.0679 669.67 -1377.7
## + alcohol:total.sulfur.dioxide           1    0.0431 669.70 -1377.6
## + density                                1    0.0094 669.73 -1377.5
##
## Step:  AIC=-1408.32
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     alcohol:sulphates + sulphates:total.sulfur.dioxide
##
##                                           Df Sum of Sq    RSS     AIC
## + chlorides                              1    5.2303 651.73 -1419.1
## + volatile.acidity:total.sulfur.dioxide  1    4.4312 652.53 -1417.1
## + pH                                     1    4.2543 652.71 -1416.7
## + free.sulfur.dioxide                    1    0.9459 656.02 -1408.6
## + fixed.acidity                          1    0.8776 656.09 -1408.5
## <none>                                               656.96 -1408.3
## + citric.acid                            1    0.3841 656.58 -1407.2
## + volatile.acidity:sulphates             1    0.2888 656.67 -1407.0
## + residual.sugar                         1    0.2047 656.76 -1406.8
## + alcohol:volatile.acidity               1    0.1399 656.82 -1406.7
## + density                                1    0.0545 656.91 -1406.5
## + alcohol:total.sulfur.dioxide           1    0.0004 656.96 -1406.3
##
## Step:  AIC=-1419.1
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + alcohol:sulphates + sulphates:total.sulfur.dioxide
##
##                                           Df Sum of Sq    RSS     AIC
## + pH                                     1    6.6385 645.09 -1433.5
## + volatile.acidity:total.sulfur.dioxide  1    4.6982 647.04 -1428.7
## + sulphates:chlorides                    1    1.5476 650.19 -1420.9
## + fixed.acidity                          1    1.1283 650.61 -1419.9
## + free.sulfur.dioxide                    1    0.9107 650.82 -1419.3
## <none>                                               651.73 -1419.1
## + volatile.acidity:sulphates             1    0.4995 651.23 -1418.3
## + residual.sugar                         1    0.4007 651.33 -1418.1
## + alcohol:volatile.acidity               1    0.3840 651.35 -1418.0
## + volatile.acidity:chlorides             1    0.3299 651.40 -1417.9
## + alcohol:chlorides                      1    0.0937 651.64 -1417.3
## + alcohol:total.sulfur.dioxide           1    0.0229 651.71 -1417.2
## + density                                1    0.0124 651.72 -1417.1
## + citric.acid                            1    0.0078 651.73 -1417.1
## + total.sulfur.dioxide:chlorides         1    0.0035 651.73 -1417.1
##
## Step:  AIC=-1433.47
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + pH + alcohol:sulphates + sulphates:total.sulfur.dioxide
```

```
##
##                                         Df Sum of Sq    RSS     AIC
## + volatile.acidity:total.sulfur.dioxide  1    3.7897 641.31 -1440.9
## + citric.acid                            1    2.4365 642.66 -1437.5
## + free.sulfur.dioxide                    1    1.9926 643.10 -1436.4
## + sulphates:pH                           1    1.6503 643.44 -1435.6
## + sulphates:chlorides                    1    1.6315 643.46 -1435.5
## + chlorides:pH                           1    1.1174 643.98 -1434.2
## + fixed.acidity                          1    0.8398 644.26 -1433.5
## <none>                                               645.09 -1433.5
## + density                                1    0.5921 644.50 -1432.9
## + volatile.acidity:sulphates             1    0.5268 644.57 -1432.8
## + alcohol:chlorides                      1    0.2556 644.84 -1432.1
## + residual.sugar                         1    0.1657 644.93 -1431.9
## + volatile.acidity:chlorides             1    0.1494 644.95 -1431.8
## + volatile.acidity:pH                    1    0.0892 645.01 -1431.7
## + alcohol:volatile.acidity               1    0.0547 645.04 -1431.6
## + total.sulfur.dioxide:chlorides         1    0.0510 645.04 -1431.6
## + alcohol:pH                             1    0.0214 645.07 -1431.5
## + total.sulfur.dioxide:pH                1    0.0054 645.09 -1431.5
## + alcohol:total.sulfur.dioxide           1    0.0004 645.09 -1431.5
##
## Step:  AIC=-1440.89
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + pH + alcohol:sulphates + sulphates:total.sulfur.dioxide +
##     volatile.acidity:total.sulfur.dioxide
##
##                                  Df Sum of Sq    RSS     AIC
## + citric.acid                     1    3.1947 638.11 -1446.9
## + free.sulfur.dioxide             1    2.4813 638.82 -1445.1
## + sulphates:pH                    1    1.6494 639.66 -1443.0
## + sulphates:chlorides             1    1.5283 639.78 -1442.7
## + chlorides:pH                    1    1.2621 640.04 -1442.0
## + fixed.acidity                   1    1.1706 640.13 -1441.8
## <none>                                        641.31 -1440.9
## + density                         1    0.7907 640.51 -1440.9
## + alcohol:total.sulfur.dioxide    1    0.6450 640.66 -1440.5
## + total.sulfur.dioxide:chlorides  1    0.5171 640.79 -1440.2
## + volatile.acidity:sulphates      1    0.3780 640.93 -1439.8
## + total.sulfur.dioxide:pH         1    0.2877 641.02 -1439.6
## + alcohol:chlorides               1    0.2849 641.02 -1439.6
## + residual.sugar                  1    0.2757 641.03 -1439.6
## + volatile.acidity:chlorides      1    0.2219 641.08 -1439.5
## + alcohol:pH                      1    0.1146 641.19 -1439.2
## + alcohol:volatile.acidity        1    0.0019 641.30 -1438.9
## + volatile.acidity:pH             1    0.0001 641.31 -1438.9
##
## Step:  AIC=-1446.88
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + pH + citric.acid + alcohol:sulphates + sulphates:total.sulfur.dioxide +
##     volatile.acidity:total.sulfur.dioxide
##
##                                    Df Sum of Sq    RSS     AIC
## + free.sulfur.dioxide               1   1.80889 636.30 -1449.4
## + sulphates:pH                      1   1.63822 636.47 -1449.0
## + sulphates:chlorides               1   1.59890 636.51 -1448.9
## + sulphates:citric.acid             1   1.59412 636.52 -1448.9
## + chlorides:pH                      1   1.32499 636.79 -1448.2
## <none>                                          638.11 -1446.9
## + citric.acid:chlorides             1   0.78105 637.33 -1446.8
## + citric.acid:total.sulfur.dioxide  1   0.73523 637.38 -1446.7
## + total.sulfur.dioxide:chlorides    1   0.58241 637.53 -1446.3
## + residual.sugar                    1   0.57829 637.53 -1446.3
## + alcohol:total.sulfur.dioxide      1   0.53497 637.58 -1446.2
## + alcohol:pH                        1   0.24283 637.87 -1445.5
## + volatile.acidity:sulphates        1   0.20964 637.90 -1445.4
## + volatile.acidity:chlorides        1   0.17459 637.94 -1445.3
## + alcohol:chlorides                 1   0.14576 637.96 -1445.2
## + total.sulfur.dioxide:pH           1   0.12133 637.99 -1445.2
## + citric.acid:pH                    1   0.09349 638.02 -1445.1
## + volatile.acidity:citric.acid      1   0.05801 638.05 -1445.0
## + volatile.acidity:pH               1   0.04332 638.07 -1445.0
## + alcohol:citric.acid               1   0.03117 638.08 -1445.0
## + alcohol:volatile.acidity          1   0.02630 638.08 -1444.9
## + fixed.acidity                     1   0.01864 638.09 -1444.9
## + density                           1   0.00589 638.10 -1444.9
##
## Step:  AIC=-1449.42
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + pH + citric.acid + free.sulfur.dioxide + alcohol:sulphates +
##     sulphates:total.sulfur.dioxide + volatile.acidity:total.sulfur.dioxide
##
##                                            Df Sum of Sq    RSS     AIC
## + sulphates:citric.acid                     1   1.90413 634.40 -1452.2
## + sulphates:chlorides                       1   1.66266 634.64 -1451.6
## + sulphates:pH                              1   1.45611 634.85 -1451.1
## + chlorides:pH                              1   1.40518 634.90 -1451.0
## + chlorides:free.sulfur.dioxide             1   1.37779 634.92 -1450.9
## + total.sulfur.dioxide:chlorides            1   1.01399 635.29 -1450.0
## + citric.acid:total.sulfur.dioxide          1   0.94270 635.36 -1449.8
## + citric.acid:chlorides                     1   0.88569 635.42 -1449.6
## + alcohol:free.sulfur.dioxide               1   0.81108 635.49 -1449.5
## <none>                                                  636.30 -1449.4
## + alcohol:total.sulfur.dioxide              1   0.65353 635.65 -1449.1
## + sulphates:free.sulfur.dioxide             1   0.59197 635.71 -1448.9
## + residual.sugar                            1   0.40324 635.90 -1448.4
## + volatile.acidity:free.sulfur.dioxide      1   0.37338 635.93 -1448.3
## + citric.acid:free.sulfur.dioxide           1   0.32753 635.97 -1448.2
## + volatile.acidity:sulphates                1   0.26855 636.03 -1448.1
## + total.sulfur.dioxide:pH                   1   0.25644 636.05 -1448.1
## + alcohol:pH                                1   0.19334 636.11 -1447.9
## + volatile.acidity:chlorides                1   0.17273 636.13 -1447.8
## + alcohol:chlorides                         1   0.14382 636.16 -1447.8
## + citric.acid:pH                            1   0.09055 636.21 -1447.6
## + volatile.acidity:citric.acid              1   0.07869 636.22 -1447.6
## + fixed.acidity                             1   0.05613 636.25 -1447.6
## + alcohol:citric.acid                       1   0.05534 636.25 -1447.6
## + pH:free.sulfur.dioxide                    1   0.04894 636.25 -1447.5
## + volatile.acidity:pH                       1   0.04751 636.25 -1447.5
## + density                                   1   0.01841 636.28 -1447.5
## + alcohol:volatile.acidity                  1   0.00726 636.29 -1447.4
## + total.sulfur.dioxide:free.sulfur.dioxide  1   0.00052 636.30 -1447.4
##
## Step:  AIC=-1452.21
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + pH + citric.acid + free.sulfur.dioxide + alcohol:sulphates +
##     sulphates:total.sulfur.dioxide + volatile.acidity:total.sulfur.dioxide +
##     sulphates:citric.acid
##
##                                             Df Sum of Sq    RSS     AIC
## + citric.acid:total.sulfur.dioxide          1   1.07213 633.33 -1452.9
## <none>                                                  634.40 -1452.2
## + alcohol:free.sulfur.dioxide               1   0.73904 633.66 -1452.1
## + chlorides:free.sulfur.dioxide             1   0.69555 633.70 -1452.0
## + citric.acid:free.sulfur.dioxide           1   0.66192 633.74 -1451.9
## + alcohol:total.sulfur.dioxide              1   0.57327 633.82 -1451.7
## + total.sulfur.dioxide:chlorides            1   0.55957 633.84 -1451.6
## + volatile.acidity:free.sulfur.dioxide      1   0.40654 633.99 -1451.2
## + citric.acid:pH                            1   0.37693 634.02 -1451.2
## + sulphates:chlorides                       1   0.36809 634.03 -1451.1
## + chlorides:pH                              1   0.36386 634.03 -1451.1
## + residual.sugar                            1   0.36322 634.03 -1451.1
## + alcohol:chlorides                         1   0.31362 634.08 -1451.0
## + sulphates:pH                              1   0.29189 634.11 -1450.9
## + sulphates:free.sulfur.dioxide             1   0.28677 634.11 -1450.9
## + total.sulfur.dioxide:pH                   1   0.19426 634.20 -1450.7
## + alcohol:pH                                1   0.16180 634.24 -1450.6
## + volatile.acidity:pH                       1   0.11253 634.28 -1450.5
## + volatile.acidity:chlorides                1   0.09768 634.30 -1450.5
## + fixed.acidity                             1   0.09312 634.30 -1450.4
## + density                                   1   0.06739 634.33 -1450.4
## + alcohol:citric.acid                       1   0.05871 634.34 -1450.4
## + citric.acid:chlorides                     1   0.04101 634.36 -1450.3
## + alcohol:volatile.acidity                  1   0.01669 634.38 -1450.2
## + volatile.acidity:citric.acid              1   0.00754 634.39 -1450.2
## + pH:free.sulfur.dioxide                    1   0.00661 634.39 -1450.2
## + volatile.acidity:sulphates                1   0.00168 634.40 -1450.2
## + total.sulfur.dioxide:free.sulfur.dioxide  1   0.00146 634.40 -1450.2
##
## Step:  AIC=-1452.91
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + pH + citric.acid + free.sulfur.dioxide + alcohol:sulphates +
##     sulphates:total.sulfur.dioxide + volatile.acidity:total.sulfur.dioxide +
##     sulphates:citric.acid + total.sulfur.dioxide:citric.acid
##
##                                               Df Sum of Sq    RSS     AIC
## + chlorides:free.sulfur.dioxide                1   0.91133 632.41 -1453.2
## <none>                                                    633.33 -1452.9
## + total.sulfur.dioxide:chlorides               1   0.75711 632.57 -1452.8
## + sulphates:citric.acid:total.sulfur.dioxide  1   0.74468 632.58 -1452.8
## + alcohol:free.sulfur.dioxide                  1   0.59100 632.73 -1452.4
## + sulphates:free.sulfur.dioxide                1   0.49609 632.83 -1452.2
## + sulphates:chlorides                         1   0.44426 632.88 -1452.0
## + volatile.acidity:free.sulfur.dioxide        1   0.42123 632.90 -1452.0
## + chlorides:pH                                1   0.40139 632.92 -1451.9
## + citric.acid:pH                              1   0.38607 632.94 -1451.9
## + alcohol:total.sulfur.dioxide                1   0.38459 632.94 -1451.9
## + residual.sugar                              1   0.33519 632.99 -1451.8
## + alcohol:chlorides                           1   0.26859 633.06 -1451.6
## + sulphates:pH                                1   0.18644 633.14 -1451.4
## + pH:free.sulfur.dioxide                      1   0.16813 633.16 -1451.3
## + volatile.acidity:pH                         1   0.16665 633.16 -1451.3
## + alcohol:pH                                  1   0.12007 633.21 -1451.2
## + volatile.acidity:chlorides                  1   0.07804 633.25 -1451.1
## + alcohol:citric.acid                         1   0.07235 633.25 -1451.1
## + citric.acid:chlorides                       1   0.05798 633.27 -1451.1
## + alcohol:volatile.acidity                    1   0.05696 633.27 -1451.1
## + citric.acid:free.sulfur.dioxide             1   0.04482 633.28 -1451.0
## + total.sulfur.dioxide:free.sulfur.dioxide    1   0.03836 633.29 -1451.0
## + fixed.acidity                               1   0.03108 633.29 -1451.0
## + density                                     1   0.02506 633.30 -1451.0
## + total.sulfur.dioxide:pH                     1   0.00726 633.32 -1450.9
## + volatile.acidity:sulphates                  1   0.00483 633.32 -1450.9
## + volatile.acidity:citric.acid                1   0.00066 633.32 -1450.9
##
## Step:  AIC=-1453.22
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + pH + citric.acid + free.sulfur.dioxide + alcohol:sulphates +
##     sulphates:total.sulfur.dioxide + volatile.acidity:total.sulfur.dioxide +
##     sulphates:citric.acid + total.sulfur.dioxide:citric.acid +
##     chlorides:free.sulfur.dioxide
##
##                                              Df Sum of Sq    RSS     AIC
## <none>                                                    632.41 -1453.2
## + residual.sugar                              1   0.75427 631.66 -1453.1
## + citric.acid:pH                              1   0.48915 631.92 -1452.5
## + sulphates:citric.acid:total.sulfur.dioxide  1   0.44478 631.97 -1452.3
## + alcohol:chlorides                           1   0.41712 632.00 -1452.3
## + sulphates:free.sulfur.dioxide               1   0.27169 632.14 -1451.9
## + volatile.acidity:free.sulfur.dioxide        1   0.26819 632.15 -1451.9
## + sulphates:chlorides                         1   0.25224 632.16 -1451.8
## + alcohol:total.sulfur.dioxide                1   0.22758 632.19 -1451.8
## + chlorides:pH                                1   0.22104 632.19 -1451.8
## + alcohol:free.sulfur.dioxide                 1   0.21744 632.20 -1451.8
## + sulphates:pH                                1   0.17856 632.24 -1451.7
## + volatile.acidity:pH                         1   0.14024 632.27 -1451.6
## + citric.acid:free.sulfur.dioxide             1   0.13791 632.28 -1451.6
## + volatile.acidity:chlorides                  1   0.13386 632.28 -1451.5
## + total.sulfur.dioxide:chlorides              1   0.08836 632.33 -1451.4
## + alcohol:pH                                  1   0.08203 632.33 -1451.4
## + alcohol:citric.acid                         1   0.06154 632.35 -1451.4
## + alcohol:volatile.acidity                    1   0.04878 632.37 -1451.3
## + citric.acid:chlorides                       1   0.03544 632.38 -1451.3
## + pH:free.sulfur.dioxide                      1   0.02941 632.38 -1451.3
## + fixed.acidity                               1   0.01788 632.40 -1451.3
## + volatile.acidity:sulphates                  1   0.00684 632.41 -1451.2
## + volatile.acidity:citric.acid                1   0.00082 632.41 -1451.2
## + total.sulfur.dioxide:pH                     1   0.00051 632.41 -1451.2
## + total.sulfur.dioxide:free.sulfur.dioxide    1   0.00034 632.41 -1451.2
## + density                                     1   0.00003 632.41 -1451.2
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + pH + citric.acid + free.sulfur.dioxide +
##     alcohol:sulphates + sulphates:total.sulfur.dioxide + volatile.acidity:total.sulfur.dioxide +
##     sulphates:citric.acid + total.sulfur.dioxide:citric.acid +
##     chlorides:free.sulfur.dioxide, data = c.wine)
##
## Coefficients:
##                           (Intercept)                                alcohol
##                               5.63684                                0.31459
##                      volatile.acidity                              sulphates
##                              -0.17930                                0.21596
##                  total.sulfur.dioxide                              chlorides
##                              -0.13430                               -0.06532
##                                    pH                            citric.acid
##                              -0.10285                               -0.05630
##                   free.sulfur.dioxide                      alcohol:sulphates
##                               0.05830                                0.05486
##        sulphates:total.sulfur.dioxide  volatile.acidity:total.sulfur.dioxide
##                              -0.07848                                0.07081
##                 sulphates:citric.acid       total.sulfur.dioxide:citric.acid
##                              -0.02908                                0.03475
##         chlorides:free.sulfur.dioxide
##                              -0.02534
```

So, the best model to use is the one that only contains these 14 variables (alcohol, volatile.acidity, sulphates, total.sulfur.dioxide, chlorides, pH, citric.acid, free.sulfur.dioxide, alcohol:sulphates, sulphates:total.sulfur.dioxide, volatile.acidity:total.sulfur.dioxide, sulphates:citric.acid, total.sulfur.dioxide:citric.acid, chlorides:free.sulfur.dioxide) as it minimises the AIC.

```
fit1 <- lm(quality ~ alcohol + volatile.acidity + sulphates +
    total.sulfur.dioxide + chlorides + pH + citric.acid + free.sulfur.dioxide +
    alcohol:sulphates + sulphates:total.sulfur.dioxide + volatile.acidity:total.sulfur.dioxide +
    sulphates:citric.acid + total.sulfur.dioxide:citric.acid +
    chlorides:free.sulfur.dioxide,
    data = c.wine)
summary(fit1)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + pH + citric.acid + free.sulfur.dioxide +
##     alcohol:sulphates + sulphates:total.sulfur.dioxide + volatile.acidity:total.sulfur.dioxide +
##     sulphates:citric.acid + total.sulfur.dioxide:citric.acid +
##     chlorides:free.sulfur.dioxide, data = c.wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63657 -0.36712 -0.07204  0.42939  1.98705
##
```

```
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           5.63684    0.01688 333.969  < 2e-16 ***
## alcohol                               0.31459    0.01788  17.599  < 2e-16 ***
## volatile.acidity                     -0.17930    0.02053  -8.733  < 2e-16 ***
## sulphates                             0.21596    0.02039  10.589  < 2e-16 ***
## total.sulfur.dioxide                 -0.13430    0.02334  -5.753 1.05e-08 ***
## chlorides                            -0.06532    0.02111  -3.094  0.00201 **
## pH                                   -0.10285    0.02023  -5.083 4.15e-07 ***
## citric.acid                          -0.05630    0.02350  -2.396  0.01671 *
## free.sulfur.dioxide                   0.05830    0.02233   2.611  0.00911 **
## alcohol:sulphates                     0.05486    0.01939   2.829  0.00473 **
## sulphates:total.sulfur.dioxide       -0.07848    0.01340  -5.856 5.75e-09 ***
## volatile.acidity:total.sulfur.dioxide  0.07081    0.01789   3.958 7.90e-05 ***
## sulphates:citric.acid                -0.02908    0.01638  -1.776  0.07600 .
## total.sulfur.dioxide:citric.acid      0.03475    0.01935   1.796  0.07268 .
## chlorides:free.sulfur.dioxide        -0.02534    0.01677  -1.511  0.13103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6319 on 1584 degrees of freedom
## Multiple R-squared:  0.3932, Adjusted R-squared:  0.3878
## F-statistic: 73.31 on 14 and 1584 DF,  p-value: < 2.2e-16
```

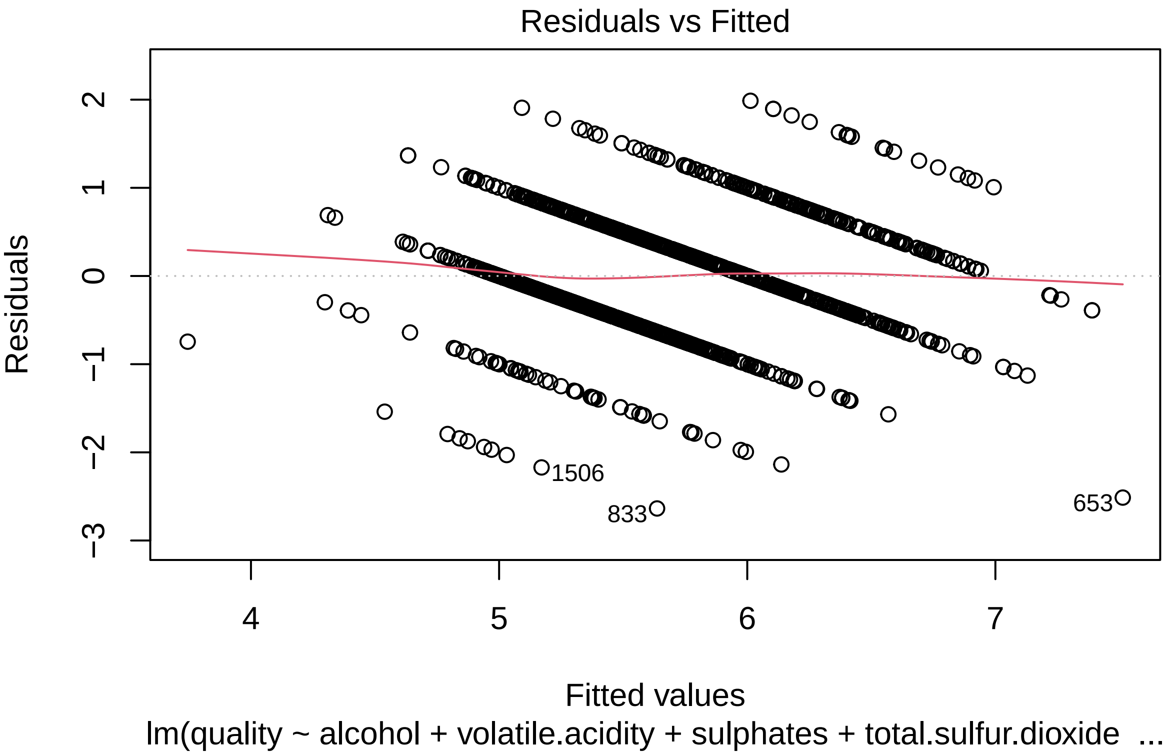
Almost every variable in this new model is significant (other than a few interaction variables that are near the boundary) with R^2_{adj} increasing by $0.3838 - 0.3561 = 0.0277$ over the original full model. This suggests an improvement over the original model which allows us to be slightly more confident with our inferences. We can also see the improvement in AIC.

```
AIC(fit0,fit1)
```

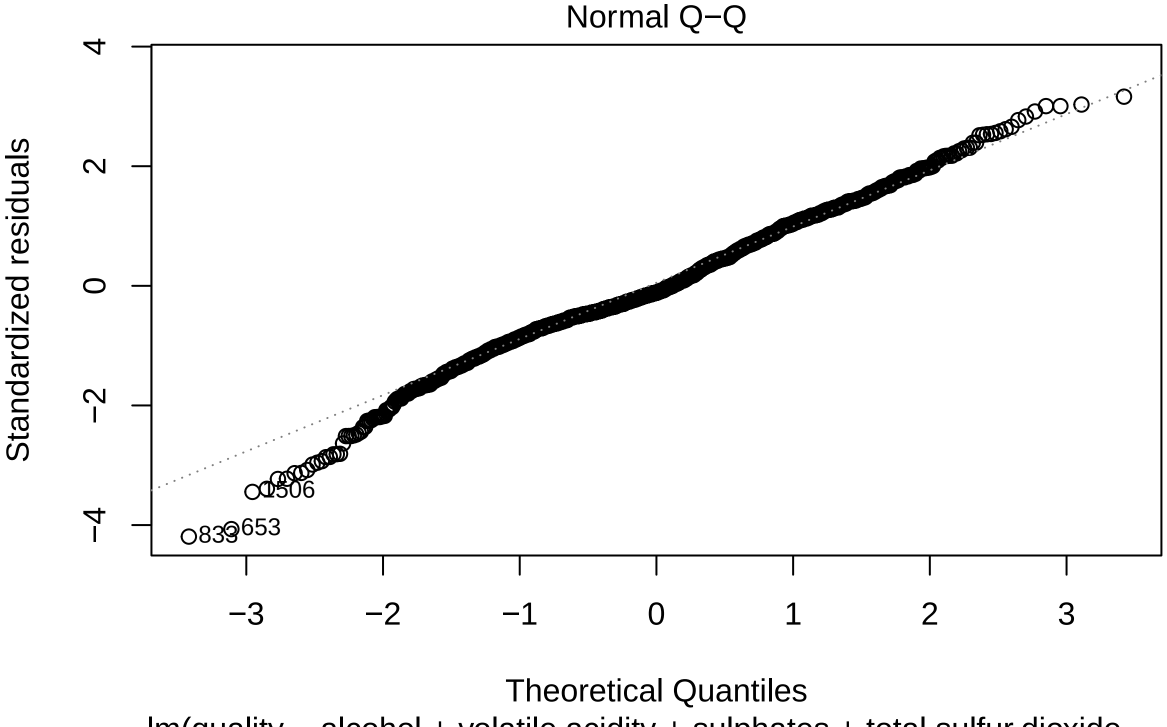
```
##      df      AIC
## fit0 13 3164.277
## fit1 16 3086.550
```

However, even with an improvement, R^2_{adj} is still moderately low which suggests that the variance in data isn't explained too well by the model. We can examine this by viewing the residual plots.

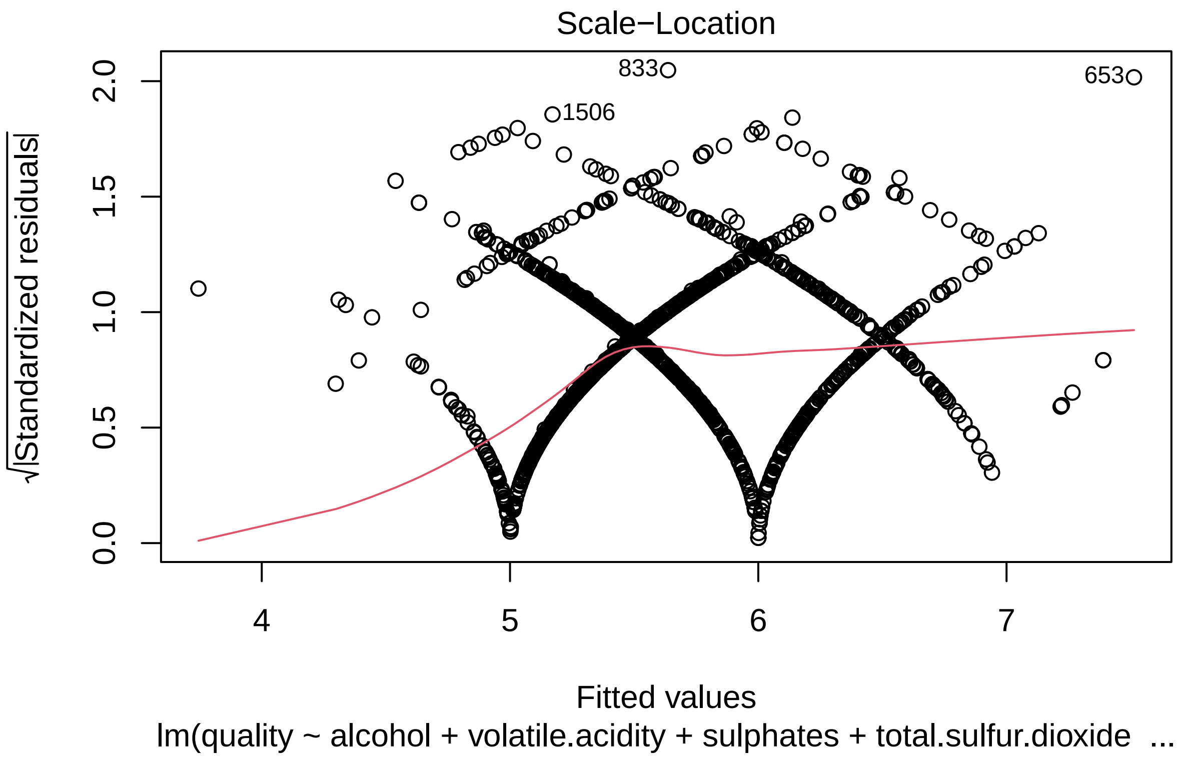
```
plot(fit1)
```

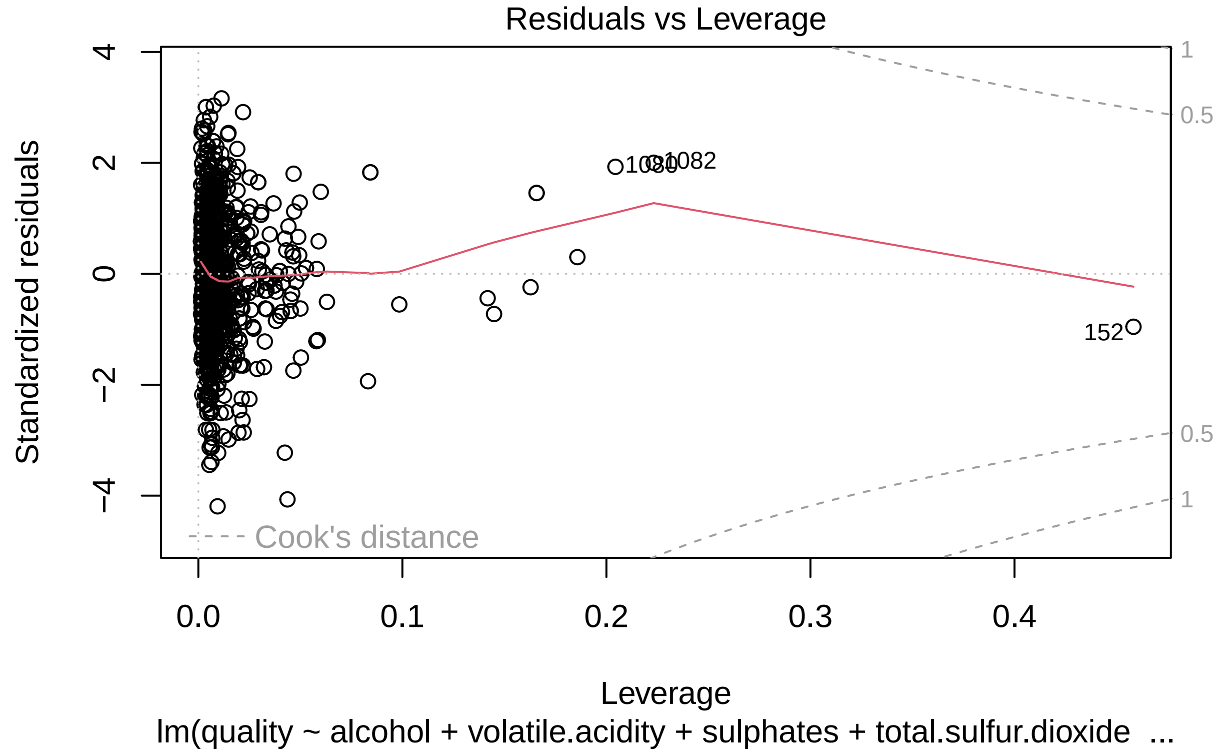


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
1506
833
653
4
5
6
7
Fitted values
lm(quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide ...



Normal Q−Q
Standardized residuals
4
2
0
−2
−4
1506
833
653
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide ...

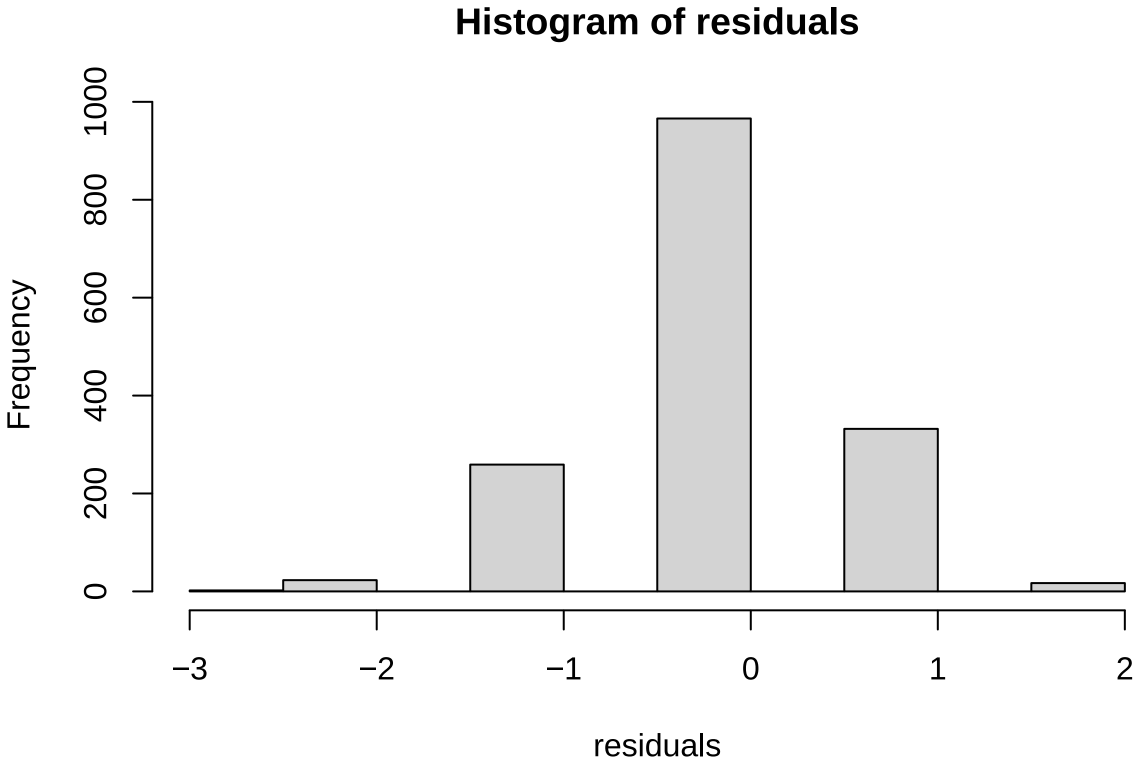




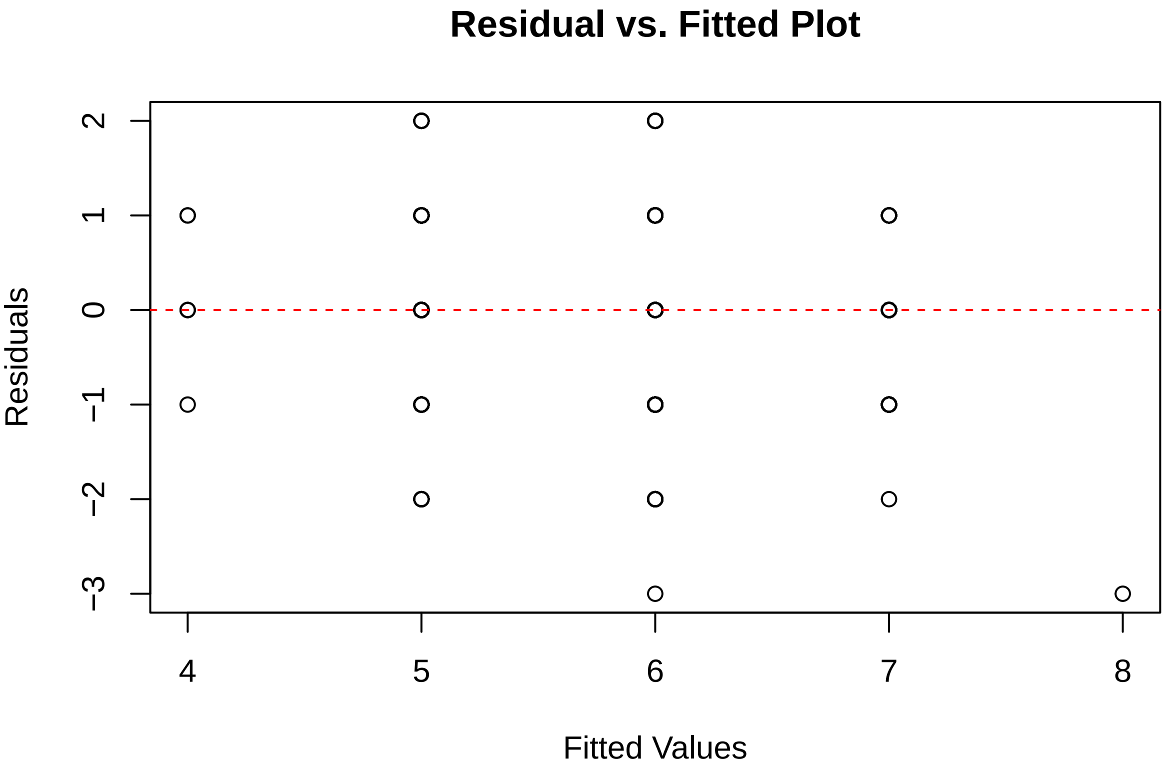
The Q-Q plot shows that the residuals are mostly normally distributed with a slight left skew and the leverage plot shows that no particular values have too much influence on the model (as none pass the Cook's distance levels of 0.5 or 1). However, both of the other plots look highly irregular with diagonal parallel lines in the fitted vs residual plot and strange wave patterns in the scale-location plot. This would usually suggest that the model assumptions have been completely violated and that all of our previous inferences are void. I will attempt to convince you that this is not the case.

This can all be explained by the response variable taking the form of an integer, a type of ordinal data. If we round the fitted values from the model to the nearest integer and then plot the new residuals on a histogram and on a residual vs fitted plot, we can see that the residuals have mean 0 and are distributed pretty evenly around 0, suggesting independence and (some analogue of) linearity. It shows that the model is correct with its predictions more often than not. In addition, the scale-location plot shows that variance remains relatively stable with respect to quality.

```r
rounded_fitted <- round(fitted(fit1))
# Calculate the residuals
residuals <- wine$quality - rounded_fitted
#Plot histogram
hist(residuals)
```



```r
# Create the residual vs. fitted plot
plot(rounded_fitted, residuals,
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residual vs. Fitted Plot")
abline(h=0, col="red", lty=2)
```



All of this along with the Q-Q plot showing normality suggests that the conclusions we draw from this model are somewhat grounded in reality.

Given the form of the response, a more suitable model would be the ordered logit model which is a type of generalised linear model for ordinal data. It is ordinal as the response is split into 11 categories with $i < i + 1$. If $\mathbf{x}$ is a set of n observations and Y n responses, with Y a non-decreasing vector, then this model is formulated as

$$\Pr(Y_j \leq i|\mathbf{x}_j) = g(\theta_i - \mathbf{x}_j^\intercal \beta)$$

where θ is a set of thresholds splitting the real line into 11 distinct points and g is the logisitic function

$$\sigma(\theta_i - \mathbf{x}_j^\intercal \beta) = \frac{1}{1 + e^{-(\theta_i - \mathbf{x}_j^\intercal \beta)}}.$$