Below the red and white wine data is analysed using linear regression. Initially

linear models are calculated for all predictors, and subsequently the linear models

are recalculated without using the predictors with p>0.05.

> reds=read.csv("winequality-red.csv")

> whites=read.csv("winequality-white.csv")

> reds.fit = lm(quality~., data=reds)

> whites.fit = lm(quality~., data=whites)

> summary(reds.fit)

Call:

lm(formula = quality ~ ., data = reds)

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) 2.197e+01 2.119e+01 1.036 0.3002

fixed.acidity 2.499e-02 2.595e-02 0.963 0.3357

volatile.acidity -1.084e+00 1.211e-01 -8.948 < 2e-16 \*\*\*

citric.acid -1.826e-01 1.472e-01 -1.240 0.2150

residual.sugar 1.633e-02 1.500e-02 1.089 0.2765

chlorides -1.874e+00 4.193e-01 -4.470 8.37e-06 \*\*\*

free.sulfur.dioxide 4.361e-03 2.171e-03 2.009 0.0447 \*

total.sulfur.dioxide -3.265e-03 7.287e-04 -4.480 8.00e-06 \*\*\*

density -1.788e+01 2.163e+01 -0.827 0.4086

pH -4.137e-01 1.916e-01 -2.159 0.0310 \*

sulphates 9.163e-01 1.143e-01 8.014 2.13e-15 \*\*\*

alcohol 2.762e-01 2.648e-02 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

Initially the predictors with the most influence on quality in the red wines

group are alcohol, sulphates, pH, total.sulfur.dioxide, free.sulfur.dioxide,

chlorides, and volatile.acidity.

> summary(whites.fit)

Call:

lm(formula = quality ~ ., data = whites)

Residuals:

Min 1Q Median 3Q Max

-3.8348 -0.4934 -0.0379 0.4637 3.1143

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.502e+02 1.880e+01 7.987 1.71e-15 \*\*\*

fixed.acidity 6.552e-02 2.087e-02 3.139 0.00171 \*\*

volatile.acidity -1.863e+00 1.138e-01 -16.373 < 2e-16 \*\*\*

citric.acid 2.209e-02 9.577e-02 0.231 0.81759

residual.sugar 8.148e-02 7.527e-03 10.825 < 2e-16 \*\*\*

chlorides -2.473e-01 5.465e-01 -0.452 0.65097

free.sulfur.dioxide 3.733e-03 8.441e-04 4.422 9.99e-06 \*\*\*

total.sulfur.dioxide -2.857e-04 3.781e-04 -0.756 0.44979

density -1.503e+02 1.907e+01 -7.879 4.04e-15 \*\*\*

pH 6.863e-01 1.054e-01 6.513 8.10e-11 \*\*\*

sulphates 6.315e-01 1.004e-01 6.291 3.44e-10 \*\*\*

alcohol 1.935e-01 2.422e-02 7.988 1.70e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7514 on 4886 degrees of freedom

Multiple R-squared: 0.2819, Adjusted R-squared: 0.2803

F-statistic: 174.3 on 11 and 4886 DF, p-value: < 2.2e-16

Initially the predictors with significant p-values are fixed.acidity, volatile.acidity,

residual.sugar, free.sulfur.dioxide, density, pH, sulphates, and alcohol.

Below, the tests are rerun, using only the significant predictors.

Call:

lm(formula = quality ~ volatile.acidity + chlorides + total.sulfur.dioxide +

sulphates + alcohol, data = reds)

Residuals:

Min 1Q Median 3Q Max

-2.67443 -0.38254 -0.06368 0.44893 2.07310

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.0048920 0.2037663 14.747 < 2e-16 \*\*\*

volatile.acidity -1.1419024 0.0969400 -11.779 < 2e-16 \*\*\*

chlorides -1.7047871 0.3916886 -4.352 1.43e-05 \*\*\*

total.sulfur.dioxide -0.0023096 0.0005082 -4.544 5.92e-06 \*\*\*

sulphates 0.9148320 0.1102702 8.296 2.26e-16 \*\*\*

alcohol 0.2770979 0.0164836 16.811 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6514 on 1593 degrees of freedom

Multiple R-squared: 0.3515, Adjusted R-squared: 0.3495

F-statistic: 172.7 on 5 and 1593 DF, p-value: < 2.2e-16

As is evident above, the F-statistic for this model is much higher than

it was initially. Also, the RSE and R-squared values are lower than either

was initially.

> summary(whites.fit2)

Call:

lm(formula = quality ~ volatile.acidity + residual.sugar + free.sulfur.dioxide +

sulphates + alcohol + density + pH + fixed.acidity, data = whites)

Residuals:

Min 1Q Median 3Q Max

-3.8246 -0.4938 -0.0396 0.4660 3.1208

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.541e+02 1.810e+01 8.514 < 2e-16 \*\*\*

volatile.acidity -1.888e+00 1.095e-01 -17.242 < 2e-16 \*\*\*

residual.sugar 8.285e-02 7.287e-03 11.370 < 2e-16 \*\*\*

free.sulfur.dioxide 3.349e-03 6.766e-04 4.950 7.67e-07 \*\*\*

sulphates 6.285e-01 9.997e-02 6.287 3.52e-10 \*\*\*

alcohol 1.932e-01 2.408e-02 8.021 1.31e-15 \*\*\*

density -1.543e+02 1.834e+01 -8.411 < 2e-16 \*\*\*

pH 6.942e-01 1.034e-01 6.717 2.07e-11 \*\*\*

fixed.acidity 6.810e-02 2.043e-02 3.333 0.000864 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7512 on 4889 degrees of freedom

Multiple R-squared: 0.2818, Adjusted R-squared: 0.2806

F-statistic: 239.7 on 8 and 4889 DF, p-value: < 2.2e-16

Again, the F-statistic is significantly better for this new test, however

the test did not significantly reduce RSE or R-squared values.

It appears, as evidenced by both linear models having p-values well below 0.01,

that there is a Linear Regression Model to explain Quality based on some or all

of the predictors provided by the data.