fixed.acidity: As the quantity decreses, the quality seems to increse, except for quality 9 where it is higher than qualtiy 4

volatile.acidity: The less, seem to be better, but not too little

citric.acid: we weem to want a very accid wine

residual.sugar: Even thought "7" is not as right as I would expect, we seem to be looking for "less sugar"

chlorides: clearlly follows a descending aptter, where wed be lookig for less chlorides

free.sulfur.dioxide: Too much is correlated with a very bad wine (class 3) where too little is not good either (class). Any thing between 36.4 and 33.4 is goog, the later being ideal.

total.sulfur.dioxide: descending pattern

density: doesnt seem to matter that much, the difference between worst and best being 0.0034

pH: we see a tendency of having better quality whith higher PH (not very steep thoug)

alcohol: same behaviour as pH, only in this case the slope seems to be biger

sulphates: this is a tricky one, from quality 3 to 7 we seem to be getting better quality as the number of sulphates increases, whihc doesnt hold true for better qualities

Since “wine quality” is a factor (categorical) variable, it doesn't make sense to talk about correlation because it doesn’t have a numerical value that can go up and down. But there are measures of strength of association we can use that are somewhat analogous.

As an example, let’s look at the “alcohol” variable. By using “alcohol “as the reference level for "Quality", we can perform a multiple regression.

model.lm <- lm(alcohol ~ quality, data =df)  
summary(model.lm)

Coefficients:

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

(Intercept) 10.3450 0.2432 42.537 < 2e-16 \*\*\*

quality4 -0.1925 0.2577 -0.747 0.454974

quality5 -0.5362 0.2449 -2.190 0.028597 \*

quality6 0.2304 0.2443 0.943 0.345739

quality7 1.0229 0.2459 4.159 3.25e-05 \*\*\*

quality8 1.2910 0.2567 5.029 5.11e-07 \*\*\*

quality9 1.8350 0.5438 3.374 0.000746 \*\*\*

---

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

Residual standard error: 1.088 on 4891 degrees of freedom

Multiple R-squared: 0.2199, Adjusted R-squared: 0.2189

F-statistic: 229.7 on 6 and 4891 DF, p-value: < 2.2e-16

We can interpret the estimated intercept as giving the mean “alcohol” on “quality 3” Wines as 10.3450 (% by volume), and the estimated coefficients for the subsequent qualities as showing “quality 4” Wines having on average -0.1925 (% by volume) less than “quality 3”, “quality9” Wines having 1.8350 (% by volume) more than “quality 3” and so on.

The coefficient of determination R2=0.2199 is quite small (one interpretation it is that the model explains only 22% of variance).

Note that 0.2199 isn't the correlation between “alcohol” and “quality” - we can't correlate those two variables because Quality is categorical. What it actually represents is the correlation between the observed values for alcohol, and the ones predicted (fitted) by the model. Both of these variables are numerical so we are able to correlate them. In fact the fitted values are just the mean durations for each group:

> tail(model.lm$fitted)

4893 4894 4895 4896 4897 4898

9.80884 10.57537 9.80884 10.57537 11.36794 10.57537

> tail(select(df, alcohol, quality))

alcohol quality

4893 9.7 5

4894 11.2 6

4895 9.6 5

4896 9.4 6

4897 12.8 7

4898 11.8 6

We can easily see that the 10.57537 on the fitted model is actually 10.3450 (average) + 0.2304 (coefficient from wine quality 6 – which is the quality for observation 4898)

But that was only one variable. Bellow I run a similar code to output the coefficients of all measures on the same grid. On the last line I add the R2 for that particular measure.

Looking at the r-squares we can see that no measure by itself explains much of the model. Actually, alcohol and density are the two variables that individually explain more about the model (21% and 11% respectively) followed by acidity with 7% and chlorides with roughly 5%.

Correlation

citric.acid and fixed.acidity: 0.2891807

free.sulfur.dioxide and residual.sugar : 0.2990984

total.sulfur.dioxide and residual.sugar: 0.4014393

total.sulfur.dioxide and free.sulfur.dioxide: 0.6155010

density and residual.sugar: 0.8389665

density and free.sulfur.dioxide: 0.2942104

density and total.sulfur.dioxide: 0.5298813

pH and fixed.acidity: -0.4258583

alcohol and residual.sugar: -0.4506312

alcohol and chlorides: -0.3601887

alcohol and total.sulfur.dioxide: -0.4488921

alcohol and density: -0.7801376

alcohol: 4  
residual.sugar: 4  
total.sulfur.dioxide 4  
density: 4  
free.sulfur.dioxide: 3  
fixed.acidity: 2  
pH: 1   
chlorides: 1  
citric.acid: 1