Regression analysis and correlation

In the prior chapter, I have shown summaries and basic, simple graphics of each variable. The main goal of this dataset analysis is that we determine which variables make a wine good, or which variables give higher quality of red wines in the sample, but also in the population as well. For that, I am going to calculate correlations, to see what the relationship between variables is. After that, I am going to calculate linear regression models where the y-variable (dependent variable) is quality variable, and all other variables will be independent variables or x-variables. So, let us start with correlation first. Since there are 66 combination of correlation, I am making a correlation matrix, and then proceed from there to extract only significant correlations (above 0.5) and make a regression model. Since the output correlation matrix is very strange-looking in Word output, I will skip it here.

Correlation is a statistical measure of a relationship between two variables. It can go between -1 and +1. If it is positive, that means that if a value of one variable goes up, the value of the other variable goes up too. The same logic goes for negative values (it goes down). The best values for correlation is above 0.4, as all below that aren't that significant.

From the correlation matrix, I see 6 correlation coefficients higher than 0.50. Those are:

1) Citric acid and fixed acidity = 0.672
2) Density and fixed acidity = 0.668
3) pH and fixed acidity = -0.683
4) Citric acid and volatile acidity = -0.55
5) pH and citric acid = -0.54
6) total sulfur dioxide and free sulfur dioxide = 0.66

Earlier, during the summary part, I saw what could be a higher correlation coefficient, between density, residual sugar and alcohol. Unfortunately, coefficient between density and residual sugar variables, and coefficient between residual sugar and alcohol variables aren't high enough. Only the coefficient between Alcohol and density is -0.496, which is pretty close to our limit (0.5), so I'll take that one in consideration also. Now, at last, we have 7 significant correlations to check via regression model. Most of them aren't really significant, when looking at *adjusted R-squared which explains how much of the variability is being explained by the model. To be significant, that R squared should be around 60-70%.*

When talking about regression model, think about it as linear equations you learn in school. Mostly, they start with y = ax + b, just like the regression equation which is linear. But the equation can be not-linear, which doesn't have a linear look on the graph, but as a polynomial. Those are harder to explain and not used in normal examples as much as a linear equation.

First regression model to check is citric acid= -0.354270 + 0.075153*fixed.acidity.

As we can see, the p-values are all less than 0,05, so the means of the model and the variables in are significant, but the adjusted R squared is less than 0.50 (it is actually 0.4508). Still, that level of R-squared is still semi-significant and useful.

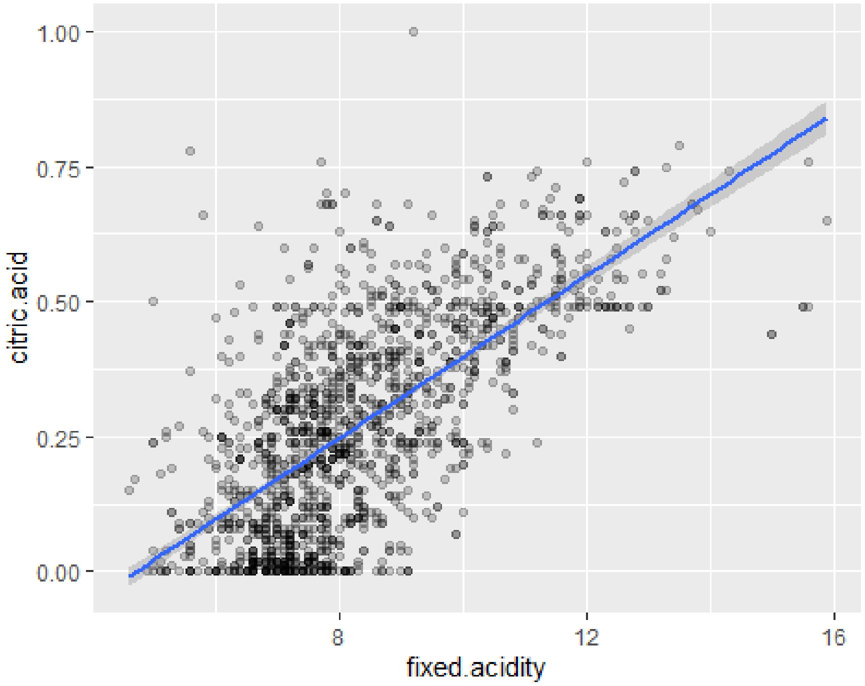
```
reg1<-lm (citric.acid~fixed.acidity)
print (summary (reg1))

##
## Call:
## lm(formula = citric.acid ~ fixed.acidity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33302 -0.11017 -0.00938  0.09317  0.71341
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.354270   0.017629  -20.09   <2e-16 ***
## fixed.acidity  0.075153   0.002074   36.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1444 on 1597 degrees of freedom
## Multiple R-squared:  0.4512, Adjusted R-squared:  0.4508
## F-statistic:  1313 on 1 and 1597 DF,  p-value: < 2.2e-16
```

```
ggplot (dfrm, aes(x=fixed.acidity, y=citric.acid))+ geom_point (alpha=.2) + stat_smooth (method=lm, level=0.95)

## `geom_smooth()` using formula 'y ~ x'
```



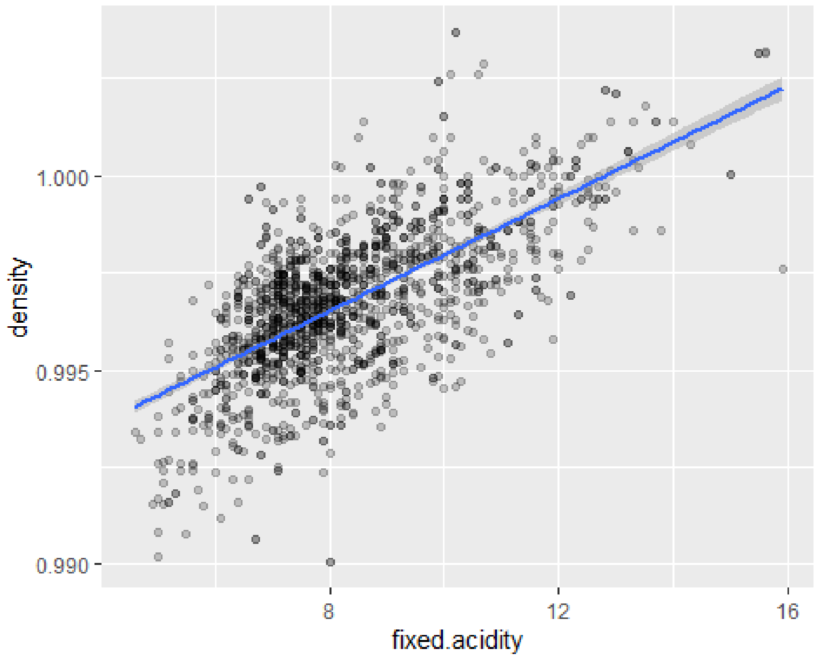
The next regression model is density = 9.907e-01 + 7.242e-04*fixed acidity. Here we are really talking about big decimal numbers after a point, and both p-values are lower than 0.05, which makes the model and the variables in it significant. The R squared value is 0.4459, which isn't particularly high.

```
reg2<-lm (density~fixed.acidity)
print (summary(reg2))

##
## Call:
## lm(formula = density ~ fixed.acidity)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0064452 -0.0007700  0.0000738  0.0009434  0.0055816
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.907e-01  1.716e-04 5774.70   <2e-16 ***
## fixed.acidity 7.242e-04  2.018e-05   35.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001405 on 1597 degrees of freedom
## Multiple R-squared:  0.4463, Adjusted R-squared:  0.4459
## F-statistic:  1287 on 1 and 1597 DF,  p-value: < 2.2e-16

ggplot (dfrm, aes(x=fixed.acidity, y=density))+ geom_point (alpha=.2) + stat_smooth(method=lm, l
evel=0.95)

## `geom_smooth()` using formula 'y ~ x'
```



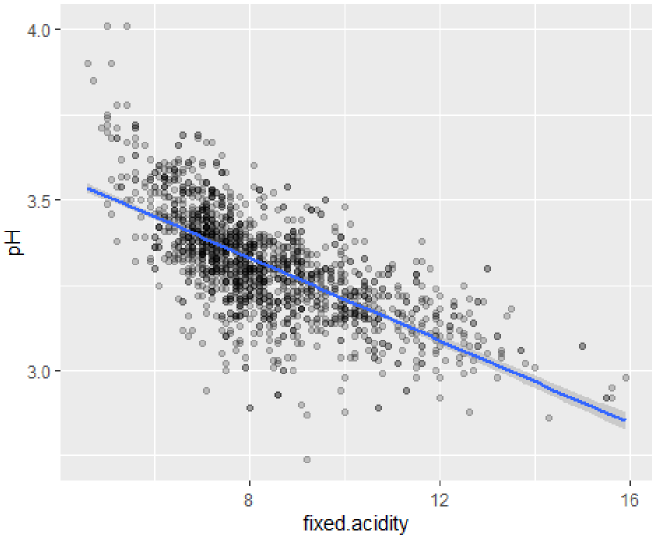
The next regression model is pH = 3.81 - 0,06*fixed.acidity. Again, p-values are lower than 0.05, which makes the model and the variables in it significant. The R squared value is 0.4661, which again isn't that particularly high to be seen significant.

```
reg3<-lm (pH~fixed.acidity)
print (summary(reg3))

##
## Call:
## lm(formula = pH ~ fixed.acidity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51780 -0.06547  0.00164  0.06488  0.52207
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.814959   0.013776  276.93   <2e-16 ***
## fixed.acidity -0.060561   0.001621  -37.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1128 on 1597 degrees of freedom
## Multiple R-squared:  0.4665, Adjusted R-squared:  0.4661
## F-statistic:  1396 on 1 and 1597 DF,  p-value: < 2.2e-16

ggplot (dfrm, aes(x=fixed.acidity, y=pH))+ geom_point (alpha=.2) + stat_smooth (method=lm, level
=0.95)

## `geom_smooth()` using formula 'y ~ x'
```



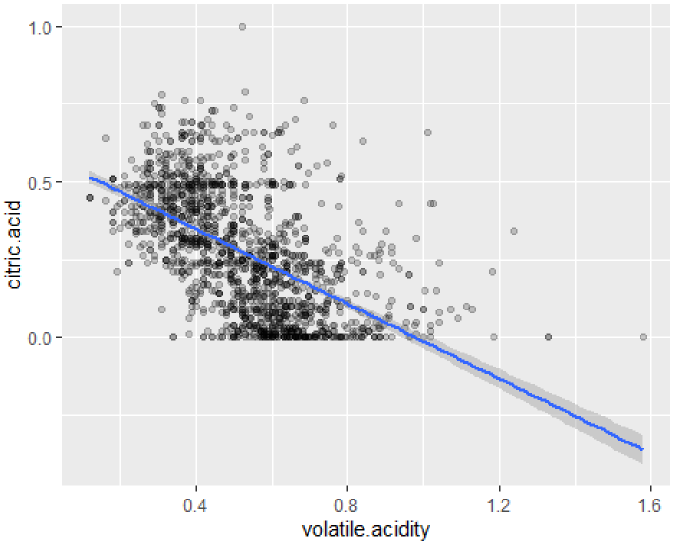
The next regression model is citric.acid = 0.58 – 0.6*volatile.acidity. Again, p-values are lower than 0.05, which makes the model and the variables in it significant. The R squared value is 0.3048, which again isn't that particularly high to be seen significant.

```
reg4<-lm (citric.acid~volatile.acidity)
print (summary(reg4))
```

```
##
## Call:
## lm(formula = citric.acid ~ volatile.acidity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38387 -0.12073 -0.01748  0.09632  0.72432
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.58823    0.01265   46.51   <2e-16 ***
## volatile.acidity -0.60107    0.02269  -26.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1624 on 1597 degrees of freedom
## Multiple R-squared:  0.3053, Adjusted R-squared:  0.3048
## F-statistic: 701.7 on 1 and 1597 DF,  p-value: < 2.2e-16
```

```
ggplot (dfrm, aes(x=volatile.acidity,y=citric.acid))+ geom_point (alpha=.2) + stat_smooth(method=lm, level=0.95)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



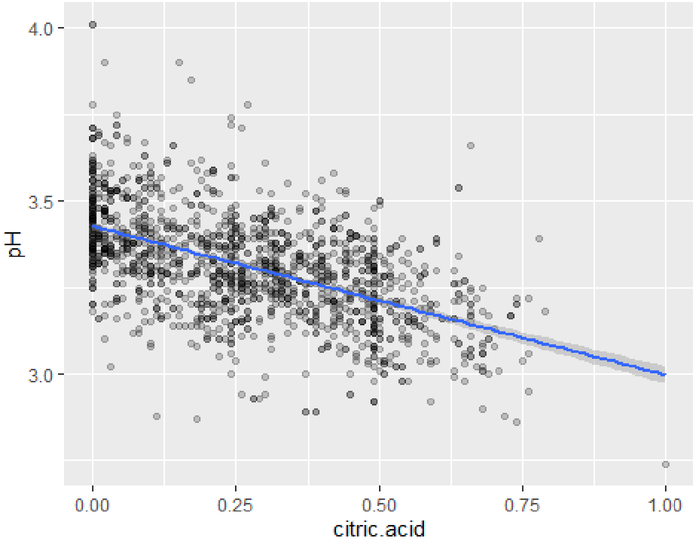
The next regression model is pH = 3.427 – 0.429*citric acid. Even though the scatter plot looks like the R squared could be higher, it is only 0.2932. The R squared is higher when the dots are close to the line.

```
reg5<-lm (pH~citric.acid)
print (summary(reg5))

##
## Call:
## lm(formula = pH ~ citric.acid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50025 -0.07733 -0.00570  0.08251  0.58251
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.427491   0.005562  616.25   <2e-16 ***
## citric.acid -0.429477   0.016668  -25.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1298 on 1597 degrees of freedom
## Multiple R-squared:  0.2937, Adjusted R-squared:  0.2932
## F-statistic:   664 on 1 and 1597 DF,  p-value: < 2.2e-16

ggplot (dfrm, aes(x=citric.acid, y=pH))+ geom_point (alpha=.2) + stat_smooth(method=lm, level=0.95)

## `geom_smooth()` using formula 'y ~ x'
```



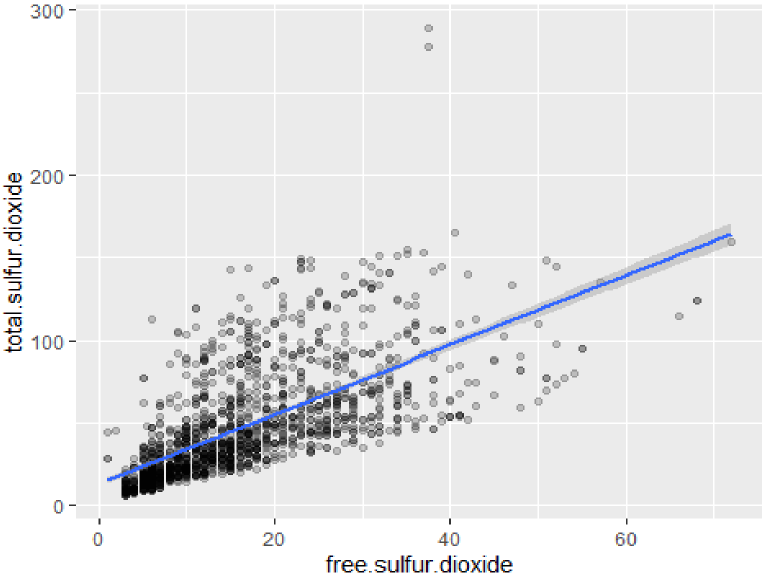
The sixth regression model is total.sulfur.dioxide = 13.13 + 2.09*free.sulfur.dioxide. P-values are lower than 0.05, and the R squared is 0.4454.

```
reg6<-lm (total.sulfur.dioxide~free.sulfur.dioxide)
print (summary(reg6))

##
## Call:
## lm(formula = total.sulfur.dioxide ~ free.sulfur.dioxide)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.120 -13.534  -7.325   7.570 197.126
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         13.13535    1.11367   11.79   <2e-16 ***
## free.sulfur.dioxide  2.09969    0.05858   35.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.5 on 1597 degrees of freedom
## Multiple R-squared:  0.4458, Adjusted R-squared:  0.4454
## F-statistic:  1285 on 1 and 1597 DF,  p-value: < 2.2e-16

ggplot (dfrm, aes(x=free.sulfur.dioxide, y=total.sulfur.dioxide))+ geom_point(alpha=.2) + stat_smooth (method=lm, level=0.95)

## `geom_smooth()` using formula 'y ~ x'
```



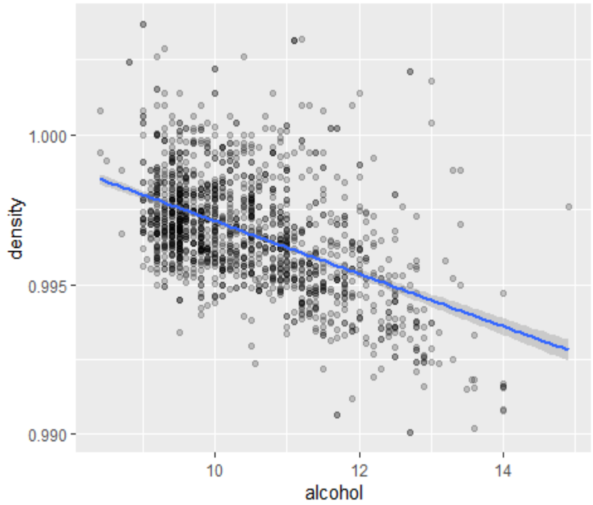
The last regression model is density = 1.006e+00 – 8.787e-04*alcohol. So again - big decimal numbers after a point. I investigated this regression model because it is said that density of the wine depends on the percent of alcohol in it, among other things. It turns out that R squared is very low here, only 0.2457 and we can see the dots are being scattered around the line away.

```
reg7<-lm (density~alcohol)
print (summary(reg7))
```

```
##
## Call:
## lm(formula = density ~ alcohol)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0049845 -0.0010951 -0.0002456  0.0008467  0.0073543
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.006e+00  4.031e-04 2495.19   <2e-16 ***
## alcohol     -8.787e-04  3.848e-05  -22.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001639 on 1597 degrees of freedom
## Multiple R-squared:  0.2462, Adjusted R-squared:  0.2457
## F-statistic: 521.6 on 1 and 1597 DF,  p-value: < 2.2e-16
```

```
ggplot (dfrm, aes(x=alcohol, y=density))+ geom_point (alpha=.2) + stat_smooth (method=lm, level
=0.95)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



With this, I am done with individual regression models. The last thing I want to see and check is how much variability is accounted by all of the variables together, versus Quality variable. In order to do that, I am going to create a multiple regression model.

Now, in statistics you have a linear regression model and a multiple regression model. Linear model takes in account just one x-variable, whereas the multiple regression model has more x-variables.

```
reg8<-lm (quality~fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dio
xide+total.sulfur.dioxide+density+pH+sulphates+alcohol)
print (summary(reg8))

## 
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid + 
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + 
##     density + pH + sulphates + alcohol)
## 
## 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
```

As you can see, at first, R squared isn’t particularly high - only 0.3561, which means that there are other factors and variables in real life that are having an effect on quality of red wines, which aren’t a part of this sample. Also, not all p-values in the models are lower than 0.05. For example, we would say that variables like Fixed acidity, citric acid, residual sugar and density aren’t significant in the model itself. The proposition to this sample is that we find other variables that are found in the red wine production, in order to better the whole model and raise the R squared value.

If I were to delete those variables from the model and only take those which are significant, the R squared value doesn’t change almost anything. The only thing we can do, with the data in this