

# Econometrics Project

Denise Berthold (01626568), Katrin Deisl (01625727)

28.06.2020

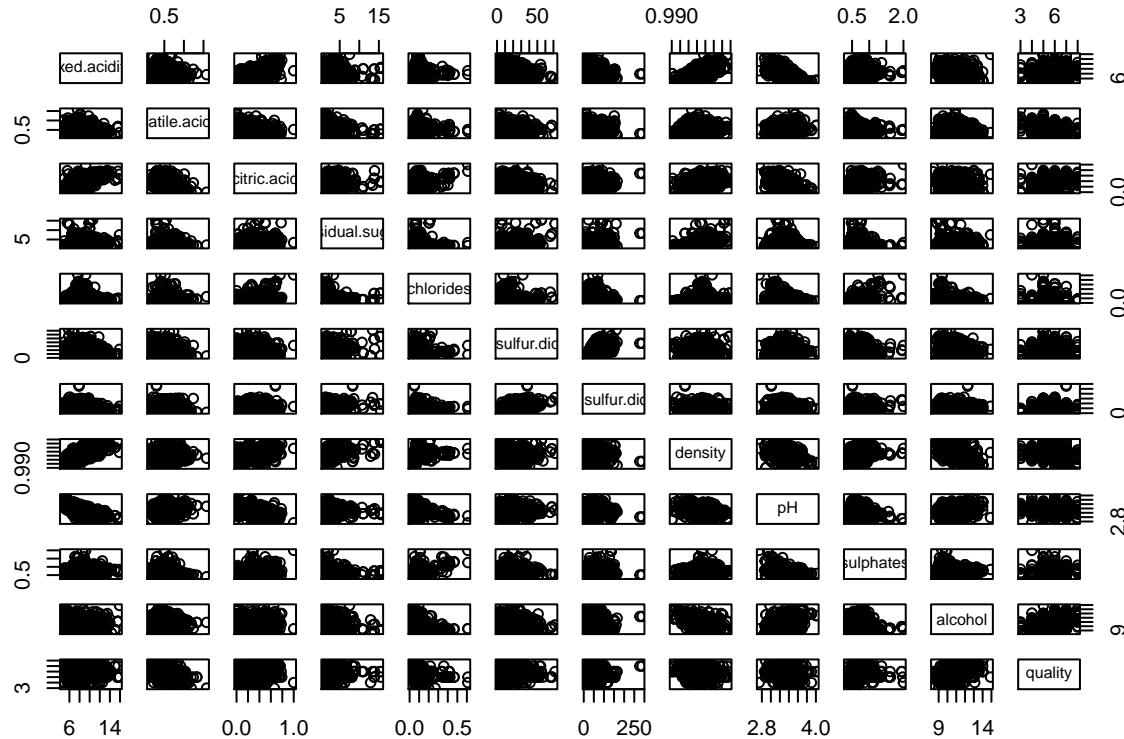
## Einleitung

In der folgenden Ausarbeitung werden verschiedene Regressions-Modelle anhand des “winequality-red” Datensatzes erstellt. Die Ausarbeitung widmet sich der Frage,

## Datensatz

Laden des Datensatzes und Plotten für einen ersten Überblick

```
data <- read.csv("winequality-red.csv")
plot(data)
```



## Beschreibung des Datensatzes

Der Datensatz stammt von Kaggle (Quelle: <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>) und enthält die folgenden beschreibenden Variablen:

- **fixed.acidity**

Diese Variable beschreibt den Anteil der nichtflüchtigen Säuren im Wein. Diese bleiben im Wein und verdampfen nicht. Der Großteil der im Wein enthaltenen Säuren sind nichtflüchtige Säuren.

- **volatile.acidity**

Der Anteil an flüchtigen Säuren (Essigsäuren) im Wein. Bei zu hohem Anteil kann der Geschmack beeinträchtigt werden.

- **citric.acid**

Gehalt an Zitronensäure. Bei kleinen Anteilen kann diese Frische und Geschmack verleihen.

- **residual.sugar**

Diese Variable enthält die Menge an Zucker, die nach der Gärung im Wein verbleibt. Die meisten Weine enthalten mehr als ein Gramm Zucker pro Liter. Weine mit bis zu 4 Gramm Zucker gelten als trockene Weine. Bei 4 - 9 Gramm Zucker handelt es sich je nach Säuregehalt um trockene oder halbtrockene Weine (je höher der Säuregehalt desto trockener). Bis zu 12 - 18 Gramm handelt es sich um halbtrockene Weine. Liebliche Weine haben einen Zuckergehalt von bis zu 45 Gramm. Weine mit mehr als 45 Gramm pro Liter werden als süß bezeichnet. (<https://www.oesterreichwein.at/unser-wein/weintypen/10-edelsuesse-weine/geschmacksrichtungen>)

- **chlorides**

Der Anteil an Salz im Wein.

- **total.sulfur.dioxide**

Schwefeldioxide werden dem Wein hinzugefügt, um das Wachstum von Bakterien und Wildhefen zu hemmen und die Oxidation des Weines zu verhindern. Ein Teil des Schwefeldioxides bindet sich im Wein, der übrige Anteil bleibt ungebunden.

- **free.sulfur.dioxide**

Der nicht gebundene Anteil wird freies Schwefeldioxid genannt. Ab 50 ppm (parts per million) kann man die freien Schwefeldioxide riechen und im Geschmack erkennen.

- **density**

Die Dichte des Weines ist je nach Alkohol- und Zuckeranteil nahe der Dichte von Wasser.

- **pH**

Der pH-Wert beschreibt, wie sauer oder basisch eine Lösung ist. Die meisten Weine liegen zwischen drei und vier auf der Skala von 0 bis 14 (0 – sauer; 14 – basisch).

- **sulphates**

Sulfat ist ein Weinzusatz, der zum Schwefeldioxidgas beitragen kann und somit ebenfalls das Wachstum von Bakterien und Wildhefen hemmt.

- **alcohol**

Der Prozentsatz des Alkoholgehalts liegt bei den meisten Weinen zwischen acht und 15 %.

Diese Variablen haben unterschiedliche Einflüsse auf die Qualität des Weins, welche in einer eigenen abhängigen Variable gespeichert wird. Die Werte liegen zwischen null und zehn.

## Vermutungen über den Einfluss der einzelnen Variablen

- **fixed.acidity**

Vorraussichtlich ist ein geringer Säureanteil wünschenswert. Höhere Werte würden in einem zu sauren Wein resultieren und die Qualität negativ beeinflussen.

- **volatile.acidity**

Da es bei einem zu hohen Anteil an flüchtigen Säuren zu einer Geschmacksbeeinträchtigung kommen kann, vermuten wir einen negativen Einfluss dieser Variable auf die Qualität.

- **citric.acid**

Bei dieser Variable wir einen zunehmenden negativen Einfluss an, da kleine Mengen als angenehm empfunden werden, bei großen Mengen jedoch der Geschmack beeinträchtigt werden könnte.

- **residual.sugar**

Da es sich bei verschiedenen Zuckergehältern um unterschiedliche Weinarten (trockener Rotwein, halbtrockener Rotwein, lieblicher Rotwein, süßer Rotwein) handelt, vermuten wir hier keinen direkten Einfluss auf die Qualität. Stattdessen vermuten wir pro Kategorie Schwankungen in der Qualität.

- **chlorides**

Hier vermuten wir einen ähnlichen Zusammenhang wie bei der Variable citric acid. Allgemein tendieren wir bei höheren Mengen zu einem negativen Einfluss.

- **sulfur.dioxide**

Unserer Vermutung nach wirkt sich der gesamte Anteil an Schwefel und vor allem der Anteil an freien Schwefeldioxiden, negativ auf die Qualität des Weines aus.

- **density**

Aufgrund unserer geringen Weinkenntniss, können wir uns hier keine Meinung bilden.

- **pH**

Je niedriger der pH-Wert, desto saurer der Wein. Je höher desto näher liegt der Wein bei der Skala bei neutral. Wir vermuten, dass ein bestimmter Säuregehalt für eine gute Qualität notwendig ist. Deshalb nehmen wir an, dass der Säuregehalt in einer schmalen Bandbreite liegen muss und Ausreißer in beide Richtungen einen negativen Einfluss haben.

- **sulphates**

Hier sehen wir den gleichen Zusammenhang wie bei den Schwefeldioxiden, da die Sulfate sehr eng mit den Schwefeldioxiden in Beziehung stehen. Der Einfluss könnte daher negativ sein.

- **alcohol**

Ähnlich wie beim pH-Wert sollte auch beim Alkoholgehalt, für eine gute Qualität, der Wert innerhalb einer gewissen Spanne liegen.

Signifikanz der einzelnen Variablen inkl. Kombinationen + t-test überprüfung

## Regressions-Modelle

### 1. Fit

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
summary(fit)
```

```

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

```

### fixed.acidity

Vorraussichtlich ist ein geringer Säureanteil wünschenswert. Höhere Werte würden in einem zu sauren Wein resultieren und die Qualität negativ beeinflussen.

```

fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
           chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
           sulphates + alcohol, data = data)
#summary(fit)

```

### volatile.acidity

Da es bei einem zu hohen Anteil an flüchtigen Säuren zu einer Geschmacksbeeinträchtigung kommen kann, vermuten wir einen negativen Einfluss dieser Variable auf die Qualität.

```

fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
           chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
           sulphates + alcohol, data = data)
#summary(fit)

```

## citric.acid

Bei dieser Variable wir einen zunehmenden negativen Einfluss an, da kleine Mengen als angenehm empfunden werden, bei großen Mengen jedoch der Geschmack beeinträchtigt werden könnte.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## residual.sugar

Da es sich bei verschiedenen Zuckergehältern um unterschiedliche Weinarten (trockener Rotwein, halbtrockener Rotwein, lieblicher Rotwein, süßer Rotwein) handelt, vermuten wir hier keinen direkten Einfluss auf die Qualität. Stattdessen vermuten wir pro Kategorie Schwankungen in der Qualität.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## chlorides

Hier vermuten wir einen ähnlichen Zusammenhang wie bei der Variable citric acid. Allgemein tendieren wir bei höheren Mengen zu einem negativen Einfluss.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## sulfur.dioxide

Unserer Vermutung nach wirkt sich der gesamte Anteil an Schwefel und vor allem der Anteil an freien Schwefeldioxiden, negativ auf die Qualität des Weines aus.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## density}

Aufgrund unserer geringen Weinkenntniss, können wir uns hier keine Meinung bilden.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## pH

Je niedriger der pH-Wert, desto saurer der Wein. Je höher desto näher liegt der Wein bei der Skala bei neutral. Wir vermuten, dass ein bestimmter Säuregehalt für eine gute Qualität notwendig ist. Deshalb nehmen wir an, dass der Säuregehalt in einer schmalen Bandbreite liegen muss und Ausreißer in beide Richtungen einen negativen Einfluss haben.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## sulphates

Hier sehen wir den gleichen Zusammenhang wie bei den Schwefeldioxiden, da die Sulfate sehr eng mit den Schwefeldioxiden in Beziehung stehen. Der Einfluss könnte daher negativ sein.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## alcohol

Ähnlich wie beim pH-Wert sollte auch beim Alkoholgehalt, für eine gute Qualität, der Wert innerhalb einer gewissen Spanne liegen.

```
fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
          chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
          sulphates + alcohol, data = data)
#summary(fit)
```

## Statistische Tests