

## Hausarbeit 4

Der *winequality.csv* Datensatz enthält Daten zu Portugiesischen 'Vinho Verde' Weinen.<sup>1</sup>. Sie werden versuchen, die Weinqualität mittels physikalisch-chemischen Eigenschaften der Weine zu schätzen. Die folgende Tabelle beschreibt die Variablen des Datensatzes:

Variable	Beschreibung
acidity	Säuregehalt [g/l]
sulfur_dioxide	Sulfitgehalt [mg/l]
density	Dichte [g/l]
pH	PH-Wert
alcohol	Alkoholgehalt [% vol]
color	Weintyp [rot/weiss]
quality	Weinqualität [1,...,10]

Für die Beantwortung der folgenden Fragen können Sie als Vorlage das R-Script von Moodle verwenden.

### Fragen

1. Laden sie den Datensatz *winequality.csv* und verschaffen Sie sich einen Überblick.  
Kommentieren Sie die vorliegenden Variablentypen (numerisch, kategorisch, ...).  
Wieviele Rot-, bzw. Weissweine gibt es im Datensatz.

---

<sup>1</sup>Die Daten sind aus dem Artikel 'Wine Quality' von Paulo Cortez et al. (2009), siehe <http://www3.dsi.uminho.pt/pcortez/wine/> für mehr Details.

2. Regressieren sie *quality* auf das ganze Set an unabhängigen Variablen (ohne Interaktionsterme und Transformationen). Beurteilen Sie die Signifikanz der Variablen im Modell.
3. Wir möchten nun überprüfen, ob es ein unterschiedliches Modell für Rotweine und Weissweine gibt. Schätzen Sie ein uneingeschränktes Modell, indem sie die Dummy Variable zu Rotwein mit den restlichen unabhängigen Variablen interagieren. Beurteilen Sie die Signifikanz der Interaktionsterme im erweiterten Modell.
4. Führen Sie einen Heterogenitätstest durch um zu entscheiden, ob Rotweine und Weissweine einem unterschiedlichen Modell folgen. Wie lautet Ihre Schlussfolgerung?