Assignment: Wines

Image: Horst J. Meuter, via Wikimedia Commons, under CC BY-SA 4.0

Learning goals

In this assignment, you:

1. learn to conduct linear regression analysis.
2. improve your data manipulation skills in Python.

Assignment

In this assignment, you analyse numerical data on wine properties.

The data sets are available in the Documents/Methods/Data/Wine folder in the course's Oma workspace. Alternatively, the data sets can be downloaded from UCI repository at http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

First, choose either red or white wines as the target of the study.

Then choose a trait from two options: 1) wine quality or 2) wine alcohol content, whichever you find more interesting.

Now, your task is to build a regression model that predicts the values of your chosen response variable as well as possible.

You should provide evidence-based answers to the following questions:

1. What is the regression equation for estimating your chosen trait values?
2. What are the five most useful variables for estimating the trait values?
3. Provide a validation-based error estimate for your model. As the data set is large, use split validation that divides the data set into separate training and testing sets.

Deliverables

Your deliverable should include both the Python codes and the results needed to verify the conclusions.

Submit your work preferably in pdf format. The deliverable should contain the information specified in the points 1 to 3 above.