Debapratim Ghosh, Lavanya Upadhyaya, Shantanu Solanki, Tanishq Khurana
STAT 425 UIUC
Oct 25, 2021

# Case Study 1: Executive Summary

## Problem Description

Vinho Verde wine comes from a small region in Northern Portugal known for its high quality white, red and rose`s wine. The wines from this region are truly magnificent in taste and are known for its lower alcohol content. The red variant is known for its red and tannic color, mostly due to the use of Vinhão, Borraçal and Amaral grapes. Since wine is becoming increasingly popular across the world, sufficient certification and quality assessment is needed to standardize the making and selling of wines. Wine certification is usually assessed through physicochemical tests.

To make sure that the alcohol level of the red variant of Vinho Verde is up to the market standards, we need to understand the relationship of the alcohol level with the various physicochemical attributes obtained through the tests. Data on the physicochemical tests were collected and donated to the UCI repository for analysis.

**The purpose of this case study is to take these physicochemical attributes and use them to model their relationship with the alcohol level of the red variant.**

## *Dataset*

The redwines dataset from the UCI repository has the following features [X1, X2……X10 are the independent variables and Y is our dependent variable]:

X1: fixed acidity
X2: volatile acidity
X3: citric acid
X4: residual sugar
X5: chlorides
X6: free sulfur dioxide

X7: total sulfur dioxide

X8: density

X9: pH

X10: sulphates

Y : alcohol

## Objective and Methods

The objective covered in this case study is to identify the linear relationship between our physicochemical attributes and the alcohol level in the redwines dataset. In order to achieve that, we fit a linear regression model to our dataset using the independent variables as our predictors and Y as our response.

There are many possible combinations of predictors, but it is necessary to keep those that best explain the variation in our response. Additionally, we need to make sure that the assumptions of a linear regression model are satisfied. If any of the assumptions are violated, we would be inaccurately estimating the regression coefficients of our predictors, therefore leading to higher errors.

Once we examine the full model, the relationships between the variances, we can determine an effective reduced model, for which we will check Model Diagnostics. Sometimes , unusual observations are indeed present in the dataset which might change the regression structure and we might misinterpret our results. Therefore we need to perform additional diagnostics to detect and remove harmful unusual observations.

Finally, there are situations when the existing set of responses and predictors don't meet the initial assumptions of linear regression. Therefore as a remedial measure, we need to transform either the response or the predictors until the assumptions are met. One such transformation is the Box Cox transformation of the response ( in our case the alcohol level) which ideally reduces the degree of non-linearity and normalises the errors. Another method to make sure that linearity assumption is met is to transform the predictors to a higher order polynomial so that the response is linearly associated with the higher order predictor.

Lastly, if we find that there is evidence of non-constant variance in the error terms, we can utilize generalised least squares/weighted least squares to remedy that.

## Summary of Findings

Overall, we can make some key assumptions about this dataset. Firstly, there is no clear evidence of collinearity, which indicates that there are no "problems" with the dataset (as independent variables should be independent of each other). We also ran through a couple of different versions of the model to see if it would give better results than the full model. For example, we first ran a reduced model to check its validity, and since it had a greater p-value, we verified model diagnostics to see if we needed to improve the model further.

We concluded that there are no influential points or outliers in any of the models that we fit. Therefore, we did not need to remove any data points from our dataset or account for any data that may skew our results.

After exploring a variety of methods and re-fitting our data model a couple of times after exploring potential issues with it, we decided to select the model option that had the least variability in terms of outliers. This was the last model that we fit -- the Transformed Model -- which has the following variables as predictors: fixed acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, and sulphates. This model shows better normality plots, fewer points deviating from the straight line in the last QQ plot, and a less askew histogram. We ran as many iterations as possible to ensure that the model we fit minimized the violation of model assumptions to the best extent possible!

## Conclusion

In conclusion, after many iterations of attempting to fit the data "well," we came up with our final model: the Transformed WLS model. This model ensures that the data follows a linear pattern without considerable deviation, indicated by a couple of diagnostics including a high R^2 value (close to 1). Moreover, there are no outlandish points that are affecting the slope or general behavior of the data.