An Exposition on Multiple Regression: A Case Study for Applications

Gregory A. Bauer, Divya Kamath

Department of Engineering, San Diego University

Prof. Dallin Munger

**An Exposition on Multiple Regression: A Case Study for Applications**

A recent study involving the categorization of wines based on its quality involved both objective and subjective approaches. In the later case, wine quality is generally assessed by physiochemical and sensory tests (Cortez, Paulo et. Al, July 2008). These sensory tests are performed by human ‘expert’ taste testers which are assumed to be the final arbiter of “quality.” As part of the study’s literature review, (Cortez, Paulo et al. Al, July 2008) reviewed several Data Mining (DM) techniques that were used to map the physiochemical parameters (e.g. alcohol and density) with a sensory taste panel which often proved difficult – frequently mitigated by using an “electronic tongue.” These above referenced DM techniques involved the us of Neural Networks (NN) and Support Vector Machines (SVM) to discriminate wines into separate classifications; and not necessarily into the subjective categories of “Quality.”

Cortez’s (July 2008) study on Decision Support Systems used both NN and SVM techniques to build models that support wine evaluation or “quality.” This in practical terms, involves taking the physiochemical properties from the wine data set and performing variable and model selection, using sensitivity analysis and kernel selection methods, to select the best performing model in terms of the how accurate the model is at predicting a wine’s “Quality.”

Using a classical approach to modeling continuous data, Multiple Regression (MR), Cortez, Paulo et. Al (July 2008) sought to gauge the performance of NN and SVM models they had developed in classifying the quality of wine based on physiochemical properties. Our approach, as an academic exercise, seeks to develop an MR model based on the same data set to illustrate how that process looks from beginning to end; something the authors left out completely and due in whole or in part in the interest of brevity.

# Methods

## Data

Cortez’s et al. (2008) study on Decision Support Systems sought to compare the performance of novel methods of classification of data, to include Neural Nets (NN) and Support Vector Machines (SVM), against the classical approach – Multiple Regression. The data used in their study can be found online at UC Irvine Machine Learning Repository and consists of two separate data sets: one consisting of all red wines and the other consisting of all white wines. The “Red Wine” data set consists of 1599 observations and 12 predictors as shown in Figure 1.

Red Wine Data Set 


Figure 1 – Red Wine Data Set (First 3 Observations)

The “White Wine” data set consists of 4899 observations and 12 predictors as shown in Figure 2.

A white background with black text

AI-generated content may be incorrect.

Figure 2 – White Wine Data Set (First 3 Observations)

Both data sets include continuous predictors and an ordinal response variable, quality, to indicate the wine category. Since our primary goal here is to highlight an MR model from start to finish, the most appropriate model at this point, considering all continuous predictors and a single response, is a Generalized Linear Model (GLM) using Ordinal Logistic Regression (OLR). Before jumping right into an OLR, we first must take a closer look at the data in the following section and, as an academic exercise, answer the question as to why both wine data sets were treated separately as opposed to a single data set in Cortez’s et al. (2008) study. Are they significantly different? One thing that was clear from the study was that during the modeling phase, three predictors stood out as being the most influential as inputs to both models. The three predictors common to both wine types were Alcohol, Sulphates, and Total Sulfur Dioxide. We will compare the two means of both groups of wine to determine if there is a statistically significant difference between them. We follow with the results of the regression and conclude with a discussion of topics for further study.

## Exploratory Analysis

## Significance Testing

## Multiple Regression Analysis

As the section on Data pointed out, both the red and white wine data sets contain 12 continuous predictors and 1 ordinal response variable. Multiple Regression using the Ordinary Least Squares (OLS) method assumes the dependent variable is continuous and normally distributed. This is not the case with both of our data sets. A more appropriate method for modeling an ordinal response variable using continuous predictors is the Ordinal Logistic Regression (OLR) method. The (OLR) method models the cumulative odds of an observation falling within or below a certain category, while assuming proportional odds across various levels of predictor variables. OLR regression model relies on several assumptions that are enumerated below.

### Modeling Assumptions

* *Ordinality*. The dependent variable must be ordinal. Within the white wine data set, quality rankings or scores consist of ordinal values beginning at 3 and ending at 9. In practical terms, there we no wines with a quality ranking lower than 3 and no higher 9 in sensory tests. Similarly, red wines consist of ordinal values beginning at 3 and ending at 8.
* *Proportional Odds Assumption.* The relationship between each pair of outcome groups is the same.At all levels of the predictor variables the odds are constant or consistent across outcome groups. “In other words, ordinal logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc.” (UCLA, 2024). This can be tested using statistical tests such as the Brant test. We assume the optimistic case of proprietary for taste testers on the sensory panel.
* *Linearity of the Predictors and Log Odds.* The relationship between the predictors and the response variable should be linear.
* *Independence of Observations.* Each observation (row of data) should be from a different wine and there should be no correlation between these observations.
* *No Perfect Collinearity.* The independent variables should not be perfectly correlated with each other.
* *Large Sample Size.* This is preferred to ensure accuracy of the results. Both our data sets are sufficiently large.

### Ordinal Logistic Regression (OLR) Model

We used the Python statsmodels library, miscmodels package, and ordinal\_model module to construct our OLR model. The results for both the red wine and white wine models are shown below as Figure 3 and Figure 4 respectively.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3 – OLR Model, Red Wine

A screenshot of a computer

AI-generated content may be incorrect.

Figure 4 - OLR Model, White Wine

Discussion and interpretation for both models follow in the Results section.

# Results

## Discussion

# References

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis, Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems, Volume 47, Issue 4, 2009, Pages 547-553, <https://doi.org/10.1016/j.dss.2009.05.016>

UCLA Advance Research Computing – Statistical Methods and Data Analytics. (2024). Ordinal Logistic Regression | R Data Analysis Examples. <https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/>