MODELING THE PREFERENCE OF WINE QUALITY USING LOGISTIC REGRESSION TECHNIQUES BASED ON PHYSICOCHEMICAL PROPERTIES

by

Perpetual Opoku Agyemang

Submitted in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

in

Mathematics

YOUNGSTOWN STATE UNIVERSITY

December, 2010

Modeling the Preference of Wine Quality Using Logistic Regression Techniques Based on Physicochemical Properties

Perpetual Opoku Agyemang

I hereby release this thesis to the public. I understand that this thesis will be made available from the OhioLINK ETD Center and the Maag Library Circulation Desk for public access. I also authorize the University or other individuals to make copies of this thesis as needed for scholarly research.

Signature:

_____

Perpetual Opoku Agyemang, Student Date

Approvals:

_____

Dr Thomas P. Wakefield, Thesis Advisor Date

_____

Dr. G. Andy Chang, Committee Member Date

_____

Dr. G. Jay Kerns, Committee Member Date

_____

Peter J. Kasvinsky, Dean of School of Graduate Studies and Research Date

ABSTRACT

Wine quality is attributed to many different factors of the wine working collectively to bear a sensory experience that is not apparent from considering these components in isolation. The various chemical components in wine give the wine its distinct taste and aroma. Appreciation of wine quality involves moving beyond our innate preferences. Currently, about 1300 components relating to wine quality have been identified in wine and new components continue to be found. These physicochemical properties can be used to model wine quality.

This review presents an analysis to extend what P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis accomplished using support vector machine and neural network methods for modeling wine preferences by data mining from physicochemical properties. Two logistic regression approaches are used to predict human wine taste preferences with the goal of better predictions. The data were subject to the logistic regression analysis to develop suitable equations to predict which components were significant in the determination of quality of wine. Since ordering exists in the dependent variable, we first considered using ordinal logistic regression. Ordinal logistic regression is a statistical technique whose dependent variable is the order response category variable and the independent variables may be categorical, interval or ration scale. An order response variable is useful for subjective assessment of quality, importance or relevance. After applying this technique, we realized that sulphate, which improves the scent of wine, and citric acid were significant as an indication of quality in both red and white wine. As some of the assumptions of the ordinal logistic model were violated, we employed multinomial logistic regression as well. Multinomial logistic regression is used when the dependent (response) variable in question is nominal, i.e. a set of categories which cannot be ordered in any meaningful way (for example, societal class) and consists of more than two categories. It assumes that data are case specific (each independent variable has a single value for each case), independent of inappropriate options. Using this technique, alcohol was statistically significant and had a negative effect throughout the various quality levels of red wine. pH was statistically significant and had a negative effect throughout the various quality levels of white wine.

This research provides a useful basis for assessing the various chemical components in wine that give wine its quality, using two regression approaches. The model built for wine quality in this analysis is anticipated to be of great use because of its dependence on only seven components for red wine and eight components for white wine.

ACKNOWLEDGEMENTS

I owe my deepest gratitude to the Almighty God for his mercies and guidance towards the success of my thesis. I am grateful to my supervisor, Dr Wakefield, whose encouragement, supervision and support from the preliminary to the concluding level enabled me to develop an understanding of the subject. I also want to thank my committee members Dr. Kerns and Dr. Chang for their time and support. I offer my regards and blessings to faculty and staff of the Department of Mathematics and Statistics, the YSU community as a whole and members of the African Students Union for supporting me in any respect during the completion of my graduate study. Finally, I would like to show my profound gratitude to my daughter, Princess Jynelle Ampaben-Kyereme, my mum, Rosemary Opoku Agyemang, my dad, Isaac, my only sister, Tracy Opoku Agyemang, my auntie and her husband, Mr. and Mrs. Poku, Mr. and Mrs. Kum, my sister and best friend, Kacie Nevels, Kwabena Ampaben-Kyereme, Theresah Owusua, my friend, Sasha Annan, and Dr. and Mrs. Akpadock for their prayers, words of encouragement, support, and love.

Contents

1 Introduction

Wine is a beverage that is associated with relaxing, socializing with others, and is complementary to food consumption. Vinho verde wine accounts for 15% of the total Portuguese wine production [7] and around 10% is exported, most of this being white wine. The determination of wine quality deals with an artistic scheme or culture that is outside our mainstream preferences. Wine quality, therefore, is incredibly 'outside ourselves'. According to the immeasurable possible disparities in its making, wine varies greatly in scent and taste. Over time, wines made by a certain winery or from a certain region or vineyard can develop a reputation as being better, and be more desired and thus more costly than those from other sources.

Quality as a concept within the discipline of marketing is a vast subject, characterized by its complexity. A more relevant aspect of this topic is how quality is evaluated and how it is conceptualized. Different perceptions of the quality of wine exist. Thus, a production management method tends to view it as an objective concept, measurable alongside external norms, whereas economists may contend that it is relative to price [20]. Perceived quality tends to be the prevailing perspective within the discipline of marketing [6], [21]. Thus it is not a concrete characteristic of a product but it is 'abstracted' from those attributes.

Quality of wine is therefore a characteristic involving the combination of different components of the wine to give a sensory experience. All of these components have a strong influence on the quality and character of wine, and are therefore not only important for the characterization and differentiation of wines, but also for the detection of frauds [4]. The whole chemical composition of a wine reflects the stages of the wine producing process, including the grape variety, yeast strain, the containers used for fermentation and storage, and the enological practice [3]. To the user deciding whether to purchase a wine, fulfilling requirements is linked with the taste of wine. Such a status for quality for a particular wine product or origin, and the price a consumer is prepared to pay, can be improved or ruined over time.

The basic factors of wine quality include wine components, quality evaluation and wine certification. There is a convergence of objective and subjective characteristics which describe the relative 'greatness' of a particular wine. The simple action of marking down one's thoughts about a wine forces one to examine the quality of each wine and put those intuitions into words. Such a practice can be very helpful and concentrates one's analysis. Instead of receiving a general instinctual reaction for a wine, it allows one to research profoundly into its depths and really appreciate, or criticize, the wine's components and its features. Quality evaluation is a critical part of the certification process which can be used to enhance wine

making by identifying the major factors affecting wine quality. Wine certification is usually completed by physicochemical and sensory analysis.

Knowledge of wine sensory characteristics and wine composition is an extremely demanding task. From tasting grapes for estimation of the development and quality in the vineyard to the evaluation of completed wine post-bottling, the observations based on sensory assessment are made throughout the winemaking process. Sensory evaluation or taste is the least understood of the human senses making wine classification a complicated task [17]. For instance, in commercial wine treatment, wines are given a quality category label on the basis of standard physicochemical properties confirmed by a certified laboratory as based upon the sensory assessment of authorized expert wine tasters.

The concept of unbalanced and badly managed vines producing poor wines is commonly espoused in the wine industry, and it is believed that the assessment of aspects of a vineyard by experienced practitioners can allow judgment of the likely quality of wine produced from the vines. This sensory evaluation is completed by highly experienced winemakers based on their own sensory impressions and experiences.

Using physicochemical laboratory analysis, density, alcohol and pH values have been used to describe wine. Wine also consists of more specific chemical components which give it its characteristics. These include sulphates, total sulphur dioxide, alcohol, volatile acidity, free sulfur dioxide, fixed acidity, residual sugar, chloride and citric acid.

This research aims to predict whether sulphates, total sulphur dioxide, alcohol, volatile acidity, free sulfur dioxide, total sulfur dioxide, fixed acidity, residual sugar, chloride or citric acid can be used to predict wine quality in both red wine and white wine from a region in Portugal (vinho verde), using both ordinal and multinomial logistic regressions.

Linear/multiple regression (MR) is the typical approach used when modeling continuous data. Regression estimates are often biased, but the bias is small with large samples. The ordinal logistic regression model assumes that the response variable to be analyzed has more than two categories, which are ordered qualitatively. It also assumes ordinality of the outcomes. The proportional odds assumption under the ordinal logistic regression model plays an imperative role in this analysis.

Multinomial logistic regression makes no statistical assumptions concerning normality, linearity and homogeneity of variance for the independent variables. However it assumes that the different outcomes are classified nominally and they are mutually exclusive [1]. It is possible to model multinomial data in an ordinal way and vice versa, but if the wrong method is used this may introduce bias or loss of efficiency and information.

According to Paulo Cortez et al [9], the support vector machine achieved promising results, outperforming the multiple regression and neural network methods in modeling the

preference of wine using physicochemical properties. They proposed that their model be used for understanding how physicochemical tests affect the sensory preferences. Moreover, they concluded that their model can support the expert wine evaluations and ultimately improve the production.

In 2001 Block et al [5], categorized three sensory attributes of Californian wine using neural networks (NNs), on the basis of grape maturity levels and chemical analysis. Only 6% error was achieved after using 36 samples.

Most recently, partial least-squares (PLS) regression models have been used to suggest that defective and negative odorants exert a strong aroma suppression effect on fruity aroma in the work of Cullere et al [10]. Their result shows that the quality of red wine is primarily related to the presence of defective or negative odorants, and secondarily to the presence of a relatively large number of fruit-sweet odorants.

In 1993, Lawrence S. Lockshin and W. Timothy Rhodus [16] compared wine quality evaluations by wine consumers and wine wholesalers for the same Chardonnay wine at three price levels and four different oak levels using multiple regression. They concluded that consumers judged wines mainly by price, regardless of the oak level. Wholesalers ignored the prices and judged the wines by the oak level. They concluded that wholesalers predicted that consumers would respond based on the wholesalers' quality judgments, and were unable to accurately predict the consumers' responses.

In recent research [15], Tony Lima examined price and quality in the California wine industry using medals won in nine tasting events in 1995 as indices of quality using multiple regression. He then proposed that a wine that wins a medal in a particular tasting is valued more highly by consumers because of the medal.

Chapter 2 presents an overview. Chapter 3 explains the technique used for performing this research. Chapter 4 consists of the analysis of the results obtained using both the ordinal and the multinomial logistic regressions. Chapter 5 interprets the results in the previous chapter. Conclusions are presented in Chapter 6.

2 Overview

This research focuses on modeling the quality of wine based on its various components. This work employs multivariate data obtained from the University of California Irvine Machine Learning Repository and donated by Professor I-Cheng Yeh of Chung-Hua University on 2009-10-07 [2]. The University of California Repository contains large data sets composed of different kinds of data types, task types and submission areas. The main purpose of this repository is to motivate researchers to scale existing and upcoming data analysis algorithms to extremely large and complicated datasets.

Each wine obtained in the laboratory consists of the raw form of a mixture with respect to a specific quality. White wines are wines that contain little or no red pigmentation. These wines are almost always made from white grapes, but can be made from black grapes as well. Red wine is derived from a vast assortment of grape varieties ranging from grapes that are reddish, deep purple, or even blue. The main difference between red and white wines is the amount of tannins they possess. Tannins are compounds present in grapes and other plants. For the purpose of this study, two datasets related to 1599 red wine samples and 4898 white wine variants of the Portuguese "Vinho Verde" wine were evaluated. Each evaluation consisted of eleven input variables and one output variable. Only physicochemical (inputs) and sensory (the output) variables are available. The inputs involved unbiased tests made by laboratory technicians. The output is the median of at least three assessments of wine quality made by wine experts. The wine quality was measured on a scale of $0 - 10$, with 0 representing poor quality and 10 representing superior quality. The variables assessed in this analysis are presented in Table 1.

The choice of appropriate statistical models is important, as it can affect the outcome and thus the interpretation of results. To determine the quality of wine, ordinal and multinomial logistic regression approaches were used to fit the model. Outliers were present in the data set. During the selection of the best model, the stepwise regression technique was used to delete some of the red wine and white wine variables from the dataset due to departures from the linearity assumption and the normal quartile-quartile plot. In all, the proposed best model was fit to eight variables for red wine and seven variables for white wine.

Residual tests were carried out to determine whether the normality, constant variance and independence assumptions were satisfied. This was also examined through the use of diagrams as shown in the Analysis section. Scatterplots of the original datasets were drawn to determine if there exists a relationship between the various variables involved. The prediction and confidence intervals were also analyzed to see whether the model was a good fit. SPSS statistical software was used in the analysis of the data.

alcohol	pH
chloride	quality
citric acid	residual sugar
density	sulphate
fixed acidity	total sulfur dioxide
free sulfur dioxide	volatile acidity

Table 1: Wine Component Variables

We will now further define the components in the data set to measure the quality of wine.

Alcohol Alcohol is a series of hydroxyl compounds, the simplest of which is derived from saturated hydrocarbons, has the general formula $C_nH_{2n+1}OH$, and include ethanol and methanol. It is measured in percentage of volumes (vol.%).

Chloride Chloride is a highly irritating, greenish-yellow gaseous halogen, capable of combining with nearly all other elements. It is a component of salt produced principally by electrolysis of sodium chloride. It is normally found in combination with sodium ions or potassium ions and is often found in large amounts in processed foods. It is usually absorbed completely by the human digestive system. It is measured in milligrams of sodium chloride per cubic decimeter.

Citric Acid Citric acid is a colorless crystalline acid originating from the fermentation of carbohydrates or from lemon, lime, and pineapple. It is present in almost all plants (especially citrus fruits) and in many animal tissues and fluids. It has a sharp sour taste and is used in many foods, confections, and soft drinks to improve their stability in metal containers. Its standard unit is grams per cubic decimeter.

Density Density is defined in a qualitative manner as a measure of the relative 'heaviness' of an object with constant volume. It is a physical characteristic of a material, as each element and compound has density associated with it. It is measured in grams per cubic decimeter.

Fixed Acidity Fixed acids are fruit acids (nonvolatile) that are organic to grapes. The predominant fixed acids found in wines are tartaric, malic, citric, and succinic. All of these acids are derived from grapes with the exception of succinic acid, which is produced by yeast during fermentation process. It is measured in grams of tartaric acid per cubic decimeter.

Free Sulfur Dioxide Free sulfur dioxide is that which does not combine with wine. Excessive amounts of it can produce an undesirable trait indicated by a slight biting sensation at the back of the throat and in the upper part of the nose [18]. The standard unit for free sulfur dioxide is milligrams per cubic decimeter.

pH Potential of Hydrogen (pH) is an expression used in food processing which is the measure of acidity or basicity of a solution. It is determined by the hydrogen concentration in water and is presented on a scale from 0 to 14. A solution with a pH value of 7 is neutral; a solution with a pH value less than 7 is acidic; a solution with a pH value greater than 7 is basic. pH is measured in moles.

Quality Quality is the degree to which a specific product satisfies the wants of a specific consumer [13]. It is a measure of the level of excellence or standard of a product or service.

Residual Sugar Residual sugar is the sugar remaining in wine after its fermentation. It is the amount of sugar not converted to alcohol throughout fermentation and affects a wine's relative sweetness. Its standard unit of measurement is grams per cubic decimeter.

Sulphate Sulphate is a chemical compound containing the sulphate radical. Sulphates are salts or esters of sulfuric acid formed by replacing one or both of the hydrogens with a metal (e.g., sodium) or a radical (e.g., ammonium or ethyl) [12]. It is measured in grams of potassium sulphate per cubic decimeter.

Total Sulfur Dioxide Total sulfur dioxide is a measure of both the bound and free sulfur dioxide in wine. Sulfur Dioxide is used throughout all stages of the winemaking process to prevent oxidation and microbial growth. It can inhibit fermentation and cause undesirable sensory effects when used excessively [14]. It is measured in milligrams per cubic decimeter.

Volatile Acidity Volatile acidity is an unstable acid formed by dissolving carbon dioxide in water. It is the basis of carbonated beverages and is related to the carbonate group of compounds. It is an acetic acid (vinegar) created by spoilage organisms that are introduced by contact with fruit flies or other air-borne insects and contaminants. It is regarded as a fault in wine since it is quite disagreeable when excessive, although a tiny amount may enhance aromas in wine. Its standard unit of measurement is grams of acetic acid per cubic decimeter.

3 Methodology

3.1 Regression

Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables). It takes a group of random variables, thought to be useful in the prediction of Y, and attempts to find a mathematical relationship between them. This relationship is typically in the form of a straight line that best approximates all the individual data points. The striking idea about statistical properties of regression concerns the relationship between the probability distribution of the parameter estimates and the actual values of those parameters.

In situations where nonlinear relationships exist using scatterplots, we attempt to transform them into a linear path (for example, log-transformation might fix this issue). The reason for modeling nonlinear relationships in this manner is that the estimation of linear regressions is much easier and their statistical properties are very recognized. However, when this approach is impossible, techniques for the estimation of nonlinear regressions have been made available. The two basic types of regression are linear regression and multiple regression.

3.2 Linear Regression

Linear regression uses one independent variable to explain and/or predict the outcome of Y. Since the response variables are normally distributed, we use t-test or F-test statistics for testing significance of explanatory variables. It takes the form

$$Y = \alpha + \beta X + u$$

where Y is the variable that we are trying to predict, X is the independent variable, α is the intercept, β is the slope, and u is the regression residual.

The assumptions of linear regression are listed in the next section.

3.3 Multiple Regression

Multiple regression uses two or more independent variables to predict the outcome. Multiple regression analysis is in fact capable of dealing with a large number of explanatory variables.

The general form of this type of regression is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + u$$

where Y is the dependent variable, $X_1, X_2, \ldots, X_k$ are the independent variables, α is the intercept, $\beta_1, \beta_2, \ldots, \beta_k$ are the slopes and u is the regression residual.

Linear and multiple regression models have the following assumptions.

Homoscedasticity The variance of the error terms is constant for each value of X. We normally use scatter plots and plot(s) of the residuals versus the X value(s) to verify this.

Linearity The relationship between each independent variable and dependent variable is linear. To verify this, we usually look at the plot(s) of Y versus the X value(s).

Normally Distributed Error Terms The error terms follow the normal distribution (multivariate normality). This assumption can be checked with a histogram, goodness-of-fit test, a fitted normal curve, or a $P - P$ Plot.

Independence of Error Terms There must be no serial correlation in the error terms. The Durbin Watson statistic and scatter plots are used to check this assumption.

3.4 Logistic Regression

We considered logistic regression in our modeling of wine preference because it is a variation of ordinary regression which is used when the dependent (response) variable is a dichotomous variable (i.e., it bears only two values, which normally correspond to the occurrence or non-occurrence of some outcome event, usually coded as 0 or 1) and the independent (input) variables are continuous, categorical, or both. For instance, logistic regression could be used in analyzing the factors that influence whether a political candidate wins or loses an election. The outcome variable is binary (i.e., either a win or a loss). The independent variables of interest could include the amount of money spent on the campaign, the amount of time spent campaigning and whether the candidate is an incumbent.

Logistic regression can handle all sorts of relationships, because it applies a non-linear log transformation to the predicted odds ratio. The "odds" of an event is defined as the probability of the outcome event occurring divided by the probability of the event not occurring. Therefore, in general, the "odds ratio" is one set of odds divided by another.

In logistic regression, hypotheses on significance of explanatory variables cannot be tested in relatively the same manner as in linear regression. Since the response variables are

Bernoulli distributed with mean value (i.e. the probability of success) related to the explanatory variables through the logit transformation, we use different test statistics (for example, the likelihood ratio statistic and Wald statistic) whose distributions are fair approximations to the distributions of the test statistics.

Although logistic regression requires much more data (at least 50 data points per predictor is necessary) to achieve stability, it is a simple method both computationally and theoretically because the independent variables do not have to be normally distributed or have equal variance in each group. It assumes independent error terms. Moreover, it can handle ordinal and nominal data as independent variables and the explanatory variables do not need to be metric. As a contrast to ordinary linear regression, logistic regression does not assume that the relationship between the dependent variables and the independent variable is linear. Multicollinearity occurs when the independent variables are not independent from each other. The multicollinearity effect is fixed by centering the variables involved. This minimizes the mean of each variable. Applications of logistic regression have also been extended to cases where the dependent variable is of more than two cases, known as multinomial or polytomous logistic regression.

The logistic regression function uses the logit transformation of θ which takes the form:

$$\theta = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}$$

where θ is the logit, α is the intercept of the equation, the logistic regression parameters $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the predictor variables $X_1, X_2, \ldots, X_k$ and exp is the base of the natural logarithm (about 2.718).

There are numerous types of logistic regression that can be used for exposition and thesis analysis. They consist of direct, stepwise and sequential logistic regressions. The type of research determines which one to use.

Direct logistic regression Direct logistic regression involves entering all the predictor variables into the equation at the same time. This is used when there is no indication about the order of the predictor variables or the importance of them in relation to the constant (for example, multinomial logistic regression).

Stepwise logistic regression Stepwise logistic regression is viewed as a data screening tool because it is used to test the involvement of all the variables to determine if the variables are significant after new variables have been added. It is a regression model in which the selection of independent variables is carried out by a step-by-step automatic procedure that allows the most significant variable to enter at each step.

Sequential logistic regression Sequential logistic regression is used in situations where there is an indication of a certain order for the predictor variables (for example, ordinal logistic regression). Regrettably, there is no easy way to achieve this with most statistical software packages. Hence multiple "runs" are then used to complete the analysis.

3.5 Ordinal Logistic Regression

Ordinal logistic regression was used because there is ordering (from low to high) in the dependent variable (quality). It models the probability of an event in comparison to all other events. The ordinal logistic regression model is known as the proportional-odds model since the odds ratio of the outcome is independent of the category j. The odds ratio is assumed to be constant for all categories. It concurrently generates multiple equations (cumulative probability). The number of equations it estimates is one less than the number of categories in the dependent variable. Ordinal logistic regression gives only one set of coefficients for each independent variable. Thus, the coefficients for the variables in the equations do not differ significantly if they were estimated individually. The intercepts differ, but the slopes are fundamentally the same. An ordinal logistic regression model is given by

$$\ln\left(\frac{P(\text{Event})}{1-P(\text{Event})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

where the quantity to the left of the equal sign is called logit representing the dependent variable, P is the probability, $X_1, X_2, \ldots, X_k$ are the independent variables, β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_k$ are the slopes.

Logit is the log of the odds that an event occurs. The odds that an event occurs is the ratio of the probability that the event occurs to the probability that the event does not occur. The expected values for the ordinal logit model are replications of the predicted probabilities for each category and given by

$$E(Y=j) = \frac{\exp(\tau_j - X_i\beta)}{1+\exp(\tau_j - X_i\beta)} - \frac{\exp(\tau_{j-1} - X_i\beta)}{1+\exp(\tau_{j-1} - X_i\beta)}$$

where τ_j represents the j^{th} category.

The predicted value is drawn from the logit and observed as one of the J discrete outcomes.

The difference in each one of the predicted probabilities is given by

$$P(Y=j|X_i) - P(Y=j|X)$$

for $j = 1, 2, \ldots J$.

As a contrast to simple linear regression, we did not use the $R-$square in ordinal regression because $R-$square provides information about how much variance is explained by the independent variable.

3.6 Assumptions of Ordinal Logistic Regression

We will now itemize the assumptions of ordinal logistic regression.

One dependent variable There should be no multiple dependent variables in ordinal regression.

Parallel lines assumption There should be one regression equation for each category. Thus, the coefficients across these equations should not vary. This assumption is responsive to the number of cases. Samples with larger numbers of cases are more likely to show a statistically significant test, and indicate that the parallel regression assumption has been violated.

Adequate cell count It is required that 80% of cells must have more than 5 counts. There should not be a zero count for any of the cells.

3.7 Multinomial Logistic Regression

The multinomial logit model generates sets of parameter estimates, comparing different levels of the dependent variable to a base level (i.e., one category of the dependent variable is chosen as the comparison category). This makes the model considerably more complex. The model can be written as

$$\log\left(\frac{P(i)}{P(r)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

where the quantity to the left of the equal sign is called relative risk (odds) representing the dependent variable with i standing in for an event or a particular level of the dependent variable, r is the comparison category (reference group), $X_1, X_2, \ldots, X_k$ are the independent variables, β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_k$ are the slopes.

Our goal is to associate the quality with the predictor variables. We assumed a linear relationship between the outcome variable and our predictor variables. Since there are multiple categories, we chose a base category as the comparison group.

An important feature of the multinomial logit model is that it estimates $k - 1$ models, where k is the number of levels of the outcome variable. In this instance, SPSS, by default,

sets one quality level as the reference group, and therefore estimated a model for each of the quality levels. Since the parameter estimates are relative to the reference group, the standard interpretation of the multinomial logit is that for a unit change in the independent variable, the logit of outcome m relative to the reference group is expected to change by its respective parameter estimate (which is in log-odds units) given that the other variables in the model are held constant.

The multinomial logit model can also be interpreted using relative risk ratio (RRR). Relative risk is the ratio of the probability of choosing one outcome category over the probability of choosing the reference category. It is also sometimes referred to as odds. Separate relative risk ratios are determined for all predictor variables for each category of the independent variable with the exclusion of the comparison category of the dependent variable, which is excluded from the analysis. They can be attained by exponentiating the multinomial logit coefficients. The RRR of a coefficient shows how the risk of the outcome falling in the comparison group compared to the risk of the outcome falling in the reference group alters with the variable to be examined. This is a change in the odds of being in the dependent variable category versus the comparison category related with a one unit change on the predictor variable. An RRR > 1 implies that the risk of the outcome falling in the comparison group comparative to the risk of the outcome falling in the reference group increases as the variable increases. An RRR < 1 indicates that the risk of the outcome falling in the comparison group relative to the risk of the outcome falling in the reference group diminishes as the variable increases. Thus, generally, if the RRR < 1, the outcome is more likely to be in the reference category.

3.8 Assumptions of Multinomial Logistic Regression

We now outline the assumptions of multinomial logistic regression.

- Data are case specific. Each predictor variable has a single value for each case.
- Collinearity is assumed to be relatively low. It is very difficult to differentiate between the effect of several variables if they are highly correlated.
- The dependent variable cannot be perfectly predicted from the independent variables for any case.
- Independence of irrelevant alternatives (IIA). This assumption states that the odds do not rely on other alternatives that are not relevant.

To aid in model selection, Portuguese red and white wine data stored in Excel was analyzed with the SPSS statistical software to obtain the output for both ordinal and multinomial logistic regression models.

To eliminate insignificant variables from the model, we used stepwise regression which was proposed by Efroymson [11]. Stepwise regression is an automatic process for statistical model selection in situations where there are a large number of possible explanatory variables and no underlying theory on which to base the model selection. The procedure is used mainly in regression analysis, though the basic approach is applicable in many forms of model selection. The fit of the model was tested after the elimination of each variable to ensure that the model still adequately fit the data. The process terminates when the available improvement falls below some critical value or when the measure is maximized. The main methods are:

Forward Selection Forward Selection entails starting with no variables in the model, trying out the variables one by one and incorporating them if they are statistically significant.

Backward Elimination Backward Elimination involves beginning with all candidate variables and testing them one by one for statistical significance, deleting any that are not significant.

Methods that are a blend both Forward and Backwards Selections This involves testing at each stage for variables to be included or excluded. This procedure was used in this analysis.

The Wald test, described by Polit [19], was used to test whether the parameters associated with the explanatory variables were significantly different from zero. It has the null hypothesis that the parameters associated with these variables are zero (i.e., not significant). Thus, those explanatory variables could be omitted from the model. Given a single parameter the Wald statistic is just the square of the t-statistic. The Wald statistic is the square of the Z-value in the equation

$$Z = \frac{SS}{SE}$$

where SS represents the point estimate for each coefficient, β, in the model and SE is the standard error. It is asymptotically chi-square distributed with estimated degrees of freedom of $n - 1$, where n is the sample size. The reason for considering the Wald statistic was that it was computationally easy and was given automatically in the output of the SPSS package.

The likelihood ratio significance test was computed by executing a logistic regression with each parameter omitted from the model and evaluating the log likelihood ratio for the model with and without the parameter. This significance test was used because it was more

reliable (i.e., it maintained a nominal level with higher accuracy) than the Wald significance test. The likelihood ratio (log likelihood) statistic is denoted by $-2\log L$ and is defined as the difference between the model fit for the reduced model and the full model. It is an observation from a chi-square distribution with $n-1$ degrees of freedom, where n is the sample size. It has the null hypothesis that the explanatory variables are non-significant (i.e., the reduced model explains the data as well as the full model). This happens when the difference between the reduced model and the full model is close to zero.

The Durbin-Watson test was used to check independence of the residuals. It has the null hypothesis that the error terms are not linearly auto-correlated. The Durbin-Watson statistic is denoted by d. It is approximately equal to $2(1-R)$, where R equals the coefficient of correlation between successive residuals. The statistic takes on values between 0 and 4. It is more likely that the residuals are independent (no autocorrelation) of each other whenever the Durbin-Watson statistic approaches 2. However the Durbin-Watson test is restricted to linear autocorrelation and direct neighbors.

The Pearson's and Deviance goodness-of-fit tests were used to measure how well the given model explained the data (the lower the better). These two tests have the null hypothesis that the difference between observed and expected events is simultaneously zero for all the groups. Thus, the model does not fit. The Pearson statistic is given by

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - N_i\Pi_i)^2}{N_i\Pi_i(1-\Pi_i)}$$

where N_i is the total frequency of subjects in the i^{th} group, O_i is the observed number of cases in the i^{th} group, and Π_i is the average estimated probability of an event outcome for the i^{th} group. Large values of X^2_{HL} show a lack of fit of the model. It is a chi-square distribution with $(n-g)$ degrees of freedom, where g is the number of categories and n is the sample size.

The Deviance statistic is given by

$$D = 2\sum_{i=1}^{n}\sum_{j=1}^{n} O_{ij}\ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

where O_{ij} is the observed value for the various categories and E_{ij} is the expected value.

A scatterplot of the entire data was employed to determine whether there is a relationship between the dependent and independent variables. Prediction and confidence intervals were also evaluated to check the accuracy of the best model.

4 Analysis

This research work considers vinho verde wine from the Minho (far north) region of Portugal. It is a fizzy wine of less than 10% of alcohol. The major export markets are France, the United States, and Germany, followed by Angola, Canada, and the United Kingdom [8]. This analysis focuses on the two most common Portuguese red and white wines. Using only the protected designation of source samples that were tested at the official certification center (CVRVV), the data were collected from May 2004 to February 2007. Concerning the preferences, each sample was estimated by sensory assessors which graded the wine on a scale that varies from 0 (very bad) to 10 (excellent). Since there was no quality level as bad as zero, one, or two, the data had entries of quality levels starting from three. The analysis was performed individually, since the red and white wine samples are relatively diverse. We first looked at the dataset as being ordered by performing the ordinal logistic regression followed by the non-ordered (multinomial) logistic regression. Parameter estimate tables, goodness-of-fit tables, results of independence tests, model-fitting information, parallel line test results, normal $P - P$ plots and histograms for both red and white wines will be discussed.

4.1 Specifying the Analysis for Ordinal Regression for Red Wine

The SPSS Ordinal Regression process or Polytomous Universal Model (PLUM) is an extension of the general linear model to ordinal categorical data. We considered the probability of an event and all other events that were ordered before it. Using graphs and displays we proceeded to test if the red wine data met the assumptions of ordinal regression.

4.1.1 Non-Normal Standardized Residual for Red Wine

The normal P-P plot was used to check if the error terms were not normally distributed. The normal plot in Figure 1 has some departures which indicates non-normality in the error terms. Thus, the error terms are not normally distributed. Hence this assumption was met.

4.1.2 Non-Normal Error Terms for Red Wine

The normal P-P plot was also used to check the assumption of non-normally distributed error terms. The histogram in Figure 2 is slightly skewed which confirms from the normal P-P plot that the normality assumption has been violated. Hence the non-normality assumption has been met.

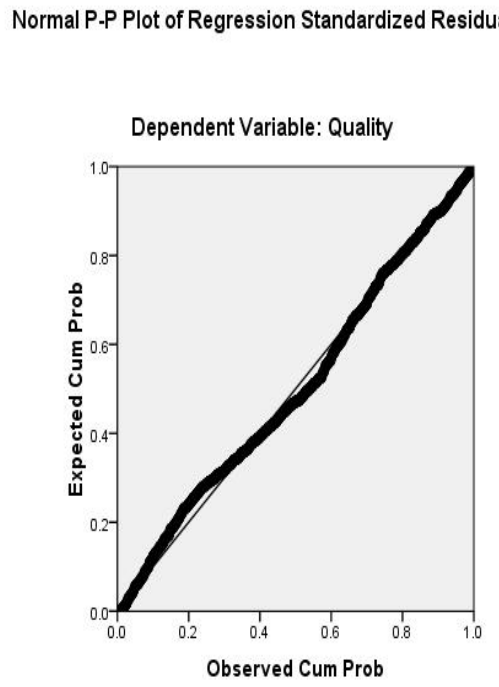

Figure 1: Normal P-P Plot of Expected Cumulative Probability vs. Observed Cumulative Probability for Red Wine

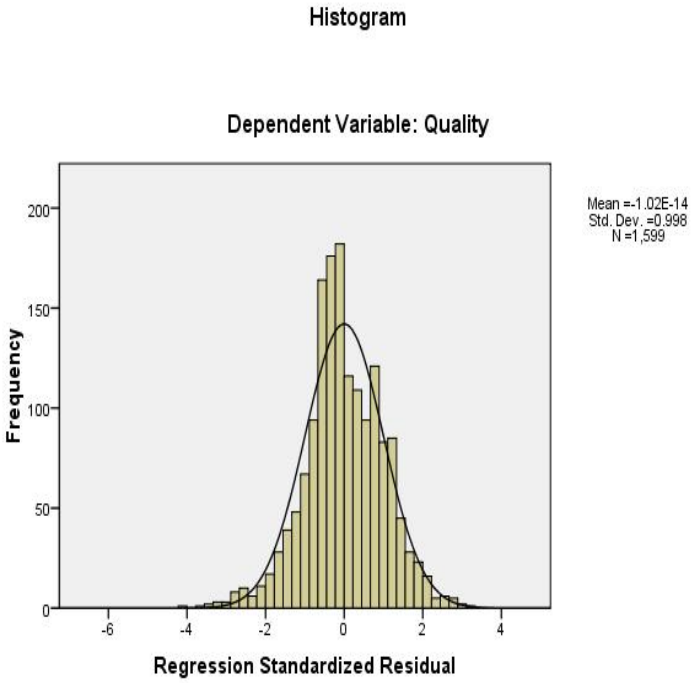

Figure 2: Histogram of Frequency vs. Regression Standardized Residual for Red Wine

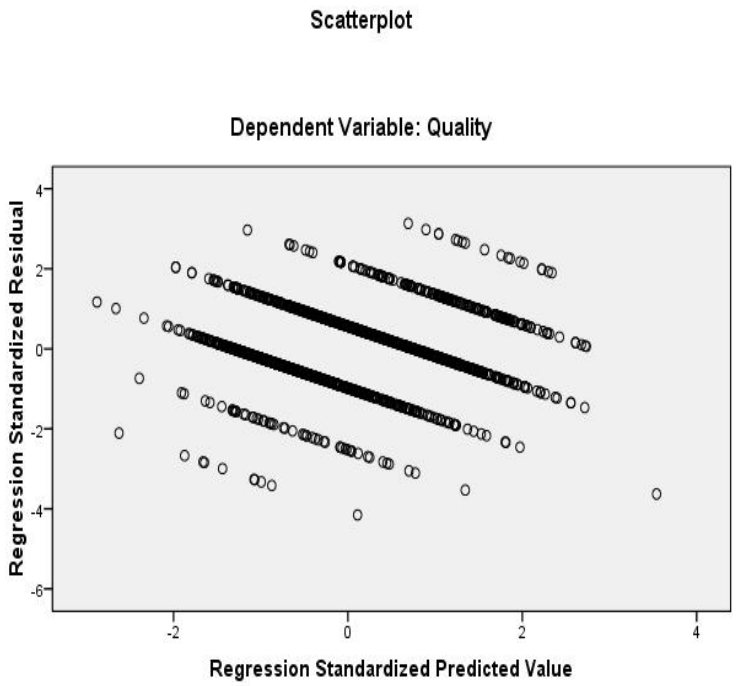

Figure 3: Scatterplot of Regression Standardized Residual vs. Regression Standardized Predicted Value for Red Wine

4.1.3 Non-Equal Variance of Regression Standardized Residual for Red Wine

The scatterplot of the regression standardized residual was used to check the assumption of heteroscedasticity. Figure 3 shows how the error terms are not scattered evenly. This indicates that the variance of the error terms is not homogeneous.

4.1.4 Stepwise Regression for Red Wine

Stepwise regression analysis was used to remove an already selected variable if that variable was not significant because of its association with the other variables. Table 2 displays the significant variables for red wine. Only seven of the predictor variables were significant so those seven variables were used in this analysis. Fixed acidity, citric acid, residual sugar and density were not significant. "VIF" represents the variance inflation factor. The cut-off point for the variance inflation factor is 10. A predictor variable is omitted from our model if its variance inflation factor is more than 10. Tolerance is the reciprocal of the variance inflation factor.

Model	Tolerance	VIF
Quality		
Alcohol	.820	1.220
Sulphate	.805	1.242
Residual Sugar	.024	27.543
Density	.064	13.567
Volatile Acidity	.756	1.332
Chloride	.514	1.944
Total Sulfur Dioxide	.750	1.333
Free Sulfur Dioxide	.797	1.255
pH	.531	1.883
Fixed Acidity	.084	11.987
Citric Acid	.056	14.654

Table 2: Stepwise Regression For Red Wine

4.1.5 Parameter Estimates for Red Wine

To fit the ordinal logit model, we found the parameter estimates for red wine. We estimated coefficients that capture differences between all possible pairs of groups. These coefficients tell how much the logit changes based on the values of the predictor variables. Multicollinearity in the model was checked by examining the standard errors for the estimated coefficients. A standard error bigger than 2.0 indicates numerical problems.

Table 3 contains the Wald statistic, the estimated coefficients, the standard errors for the coefficients and associated p-values for the model. The estimates labeled "Threshold" are α_j, the intercept equivalent terms. The estimates labeled "Location" are of most interest. They are the coefficients for the predictor variables. Considering all the independent variables at 95% confidence interval, we realized they are all statistically significant since all the p-values are less than $\alpha = .05$. Moreover, none of the standard errors exceeded 2.0, Hence, there was no multicollinearity problem.

Let A=Alcohol, V=Volatile Acidity, S=Sulphate, T=Total Sulfur Dioxide, F=Free Sulfur Dioxide, C=Chloride and P=pH.

Since qualities 3 and 4 are not statistically significant, the model for the fifth quality is then written as

$$P(Y = \text{Quality } 5|\text{X}) = \frac{\exp^{g(X)}}{1 + \exp^{g(X)}}$$

where

$$g(X) = 3.981 + .884A - 3.070V + 2.784S - .012T - 5.778C - 1.312P + .017F.$$

		Estimate	Standard Error	Wald	Sig
Threshold	quality= 3	−1.638		1.640	.200
	quality= 4	.279		.050	.823
	quality= 5	3.981		10.236	.001
	quality= 6	6.830		29.649	.000
	quality= 7	9.834		58.767	.000
Location	Alcohol	.884	.045	245.416	.000
	Volatile Acidity	−3.070	.456	89.816	.000
	Sulphate	2.784	1.75	66.194	.000
	Total Sulfur Dioxide	−.012	1.974	31.203	.000
	Chloride	−5.778	.097	21.126	.000
	pH	−1.312	1.256	12.952	.000
	Free Sulfur Dioxide	.017	.975	6.846	.009

Table 3: Parameter Estimates for Red Wine

Therefore,

$$P(Y = \text{Quality } 5|\text{X}) = \frac{\exp^{3.981+.884A-3.070V+2.784S-.012T-5.778C-1.312P+.017F}}{1+\exp^{3.981+.884A-3.070V+2.784S-.012T-5.778C-1.312P+.017F}}.$$

The sixth quality has the model

$$P(Y = \text{Quality } 6|\text{X}) = \frac{\exp^{g(X)}}{1+\exp^{g(X)}}$$

where

$$g(X) = 6.830 + .884A - 3.070V + 2.784S - .012T - 5.778C - 1.312P + .017F.$$

Hence,

$$P(Y = \text{Quality } 6|\text{X}) = \frac{\exp^{6.830+.884A-3.070V+2.784S-.012T-5.778C-1.312P+.017F}}{1+\exp^{6.830+.884A-3.070V+2.784S-.012T-5.778C-1.312P+.017F}}.$$

The seventh quality has the same model but with a different intercept.

As an interpretation of the coefficients in the model, for every increase of q grams of potassium sulphate per cubic decimeter of sulphate content of red wine, the log of odds (chance) of being in the fifth quality level increases by $2.784q$. The log odds change is $q\beta$ and the associated odds ratio is $\exp(q\beta)$, where β is the estimated coefficient for the predictor variable sulphate.

	Chi-square	Sig
Pearson	7287.238	.000
Deviance	3080.848	1.000

Table 4: Goodness-of-Fit for Red Wine

4.1.6 Goodness-of-Fit Statistics for Red Wine

Using the observed and expected frequencies, we computed the Pearson and Deviance goodness-of-fit measures. The Pearson goodness-of-fit statistic is

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - N_i\Pi_i)^2}{N_i\Pi_i(1-\Pi_i)}$$

where N_i is the total frequency of subjects in the i^{th} group, O_i is the observed number of cases in the i^{th} group, and Π_i is the average estimated probability of an event outcome for the i^{th} group. Large values of X^2_{HL} show a lack of fit of the model. It is a chi-square distribution with $(n-g)$ degrees of freedom, where n is the sample size and g is the number of categories.

The deviance goodness-of-fit statistic is

$$D = 2\sum_{i=1}^{n}\sum_{j=1}^{n} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

where O_{ij} is the observed value for the various categories and E_{ij} is the expected value.

These two goodness-of-fit statistics are used only for models that have relatively large expected values in each cell. Since we had many continuous independent variables, we had many cells with small expected values.

Table 4 shows that the observed significant level for Pearson is .000 and Deviance is 1.000 which is more than .05, hence the model fits the data well.

4.1.7 Independence Test for Red Wine

To check if there was little or no autocorrelation in the data, the independence test was used. Autocorrelation occurs when the residuals are not independent from each other. The Durbin-Watson test was employed to check this assumption.

From the SPSS output, the Durbin Watson statistic for red wine is 1.750. As the Durbin-Watson statistic approaches 2, it is more likely that the residuals are independent of each other, at least successively. Therefore, there is not sufficient evidence to reject the null hypothesis H_0 at $\alpha = 0.05$. Hence the error terms are independently distributed.

Model	2-log likelihood	Chi-square	Sig
Intercept Only	3788.451		
Final	3080.848	707.602	.000

Table 5: Model-Fitting Information for Red Wine

Model	2-log likelihood	Chi-square	Sig
Null	3080.848		
General	2961.298	119.550	.000

Table 6: Parallel Line Test for Red Wine

4.1.8 Model-Fitting Information for Red Wine

Model-fitting information was employed to check whether there is a relationship between the model without independent variables and the model with independent variables.

From Table 5, the entry labeled "Model" indicates the parameters of the model for which the model fit is evaluated. "Intercept Only" shows a model that does not control for any predictor variables and simply fits an intercept to predict the outcome variable. The entry labeled "Final" describes a model that involves the specified predictor variables. This was obtained through a process which maximized the log likelihood of the outcome variables. The final model shows an improvement over the "Intercept Only" model. The entry labeled "Chi-square" is the difference between the two -2 log-likelihood values. The observed significance level is .000 which is less than $\alpha = .05$. Hence we reject the null hypothesis that the model without predictors is as good as the model with the predictors.

4.1.9 Parallel Line Test for Red Wine

This test was carried out to check if the regression coefficients are the same for all the various categories. This is a very strong assumption for the ordinal logistic regression technique since the relationship between the independent variables and the logits must be the same. Thus, they must have the same slope.

From Table 6, the row labeled "Null" contains -2 log-likelihood value for the constrained model, the model that assumes the lines are parallel. The row labeled "General" is for the model with separate lines or planes. The entry labeled "Chi-square" is the difference between the two -2 log-likelihood values. The p-value is .000 which is less than α, so we reject the null hypothesis and conclude that there is significant difference in the coefficients between the models. This is a violation of the parallel line assumption since the relationship between the independent variables and the logits are not the same for all the logits. We then used the

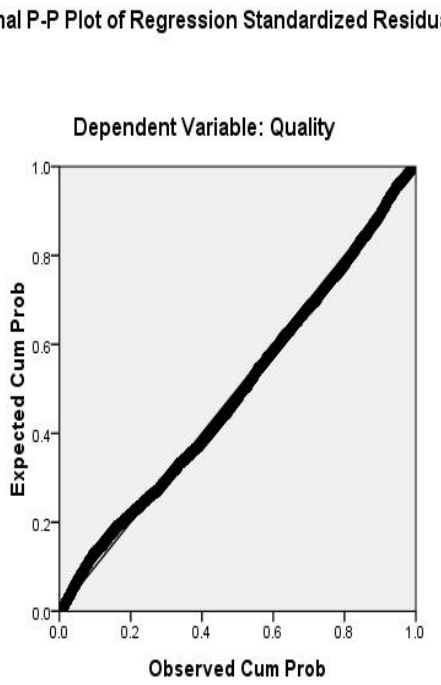

Figure 4: Normal P-P Plot of Expected Cumulative Probability vs. Observed Cumulative Probability for White Wine

multinomial logistic regression technique since the parallel line test assumption for ordinal logistic regression was violated.

4.2 Specifying the Analysis for Ordinal Regression for White Wine

We once again used graphs and displays to check the assumptions of ordinal regression for white wine.

4.2.1 Non-Normal Standardized Residual for White Wine

The normal $P - P$ plot was used to check how the error terms were distributed. Figure 4 shows no departures from the straight line which indicates normality in the error terms. Hence, the error terms are normally distributed and so this assumption was not met.

4.2.2 Non-Normal Error Terms for White Wine

A histogram was used to check if the residuals were not following a normal distribution. The histogram in Figure 5 shows a bell-shape in the various observations which confirms from

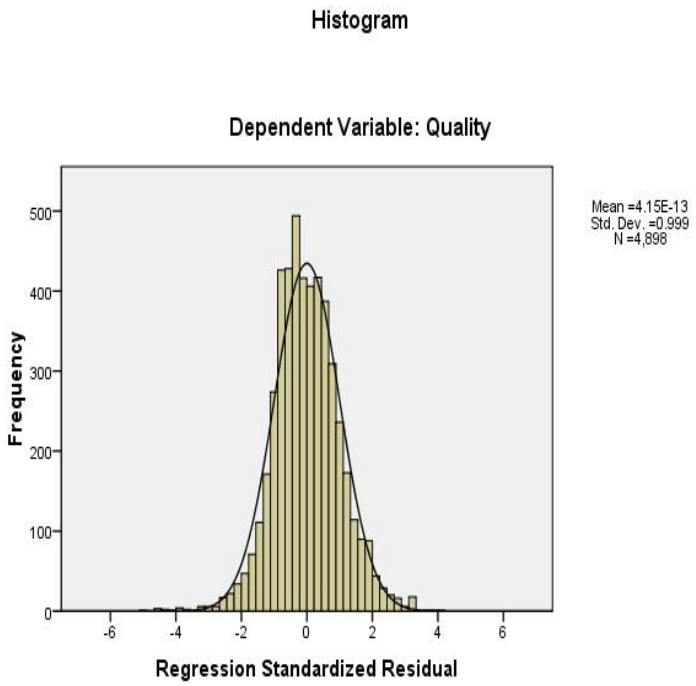

Figure 5: Histogram of Frequency vs. Regression Standardized Residual for White Wine

the normal P-P plot and the goodness-of-fit test that the non-normality assumption has not been met.

4.2.3 Non-Equal Variance of Regression Standardized Residual for White Wine

To check if the variance of the error terms is constant, we used the scatterplot of the regression standardized residual. The variance of the error terms is not constant in Figure 6 since they are not nicely scattered. Hence, the non-constant variance assumption is met.

4.2.4 Stepwise Regression for White Wine

To test the significance of variables after new variables have been added, the stepwise regression technique was used. We realized only eight of the independent variables were significant for the white wine sample. Citric acid, chloride and total sulfur dioxide were not significant. The results are displayed in Table 7. "VIF" represents the variance inflation factor. If the variance inflation factor for a particular predictor variable is more than 10, that predictor variable is omitted from our model since it is not significant. Tolerance is the reciprocal of the variance inflation factor.

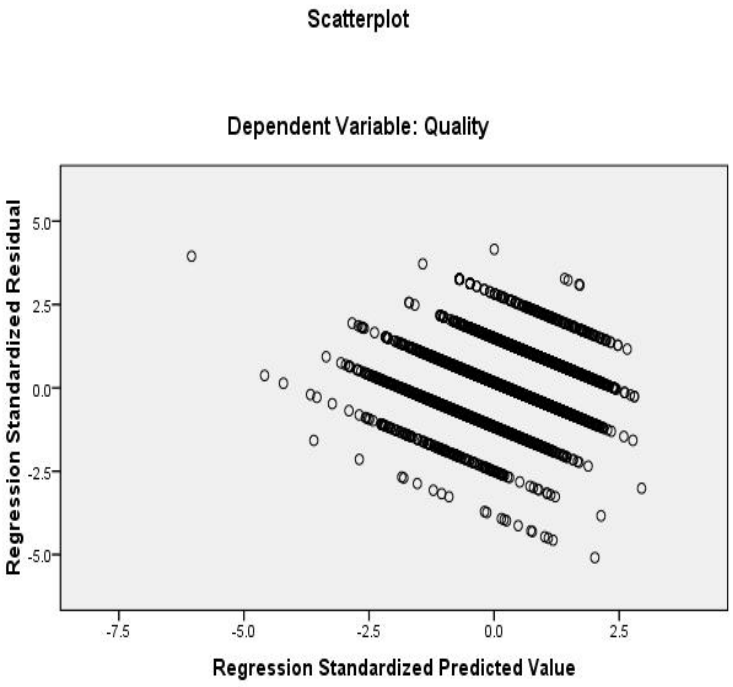

Figure 6: Scatterplot of Regression Standardized Residual for White Wine vs. Regression Standardized Predicted Value

Model	Tolerance	VIF
Quality		
Alcohol	.131	7.623
Volatile Acidity	.946	1.057
Citric Acid	.085	11.828
Residual Sugar	.129	7.612
Chloride	.079	12.579
Free Sulfur Dioxide	.870	1.149
Density	.238	8.123
pH	.473	2.114
Sulphates	.885	1.130
Fixed Acidity	.338	2.580
Total Sulfur Dioxide	.038	26.123

Table 7: Stepwise Regression for White Wine

		Estimate	Standard Error	Wald	Sig
Threshold	quality= 3	−467.769		73.625	.000
	quality= 4	−465.407		72.895	.000
	quality= 5	−462.367		71.959	.000
	quality= 6	−459.780		71.169	.000
	quality= 7	−457.527		70.477	.000
	quality= 8	−453.847		69.344	.000
Location	Alcohol	.422	1.100	34.699	.000
	Volatile Acidity	−5.073	.024	300.520	.000
	Residual Sugar	.236	.110	126.080	.000
	Free Sulfur Dioxide	.011	.007	38.298	.000
	Density	−478.478	.001	75.105	.000
	pH	2.095	1.344	56.253	.000
	Sulphates	1.816	.103	49.395	.000
	Fixed Acidity	.244	.100	18.537	.000

Table 8: Parameter Estimates for White Wine

4.2.5 Parameter Estimates for White Wine

We used the estimated coefficients to fit the various models for white wine. We checked multicollinearity in the model by examining the standard errors for the estimated coefficients. A standard error bigger than 2.0 indicates numerical problems.

Table 8 contains the estimated coefficients, the standard errors for the coefficients, the Wald test and associated p-values for the model. The estimates labeled "Threshold" are the α_j. The estimates labeled "Location" are the coefficients for the predictor variables. At 95% confidence interval, we realized all the independent variables are statistically significant since all the p-values are less than $\alpha = .05$. Furthermore, none of the standard errors exceeded 2.0, Hence, none of the independent variables depended on each other.

Again, let A=Alcohol, V=Volatile Acidity, R=Residual Sugar, F=Free Sulfur Dioxide, D=Density, P=pH, S=Sulphate and FA=Fixed Acidity.

All the quality levels are significant and so the model for the third quality is then written as

$$P(Y = \text{Quality } 3|\text{X}) = \frac{\exp^{g(X)}}{1 + \exp^{g(X)}}$$

where

$$g(X) = -467.769 + .422A - 5.073V + .236R + .011F - 478.478D + 2.095P + 1.8165S + .244FA.$$

	Chi-square	Sig
Pearson	56848.357	.000
Deviance	10902.106	1.000

Table 9: Goodness-of-Fit for White Wine

Hence,

$$P(Y = \text{Quality } 3|\text{X}) = \frac{\exp^{-467.769+.422A-5.073V+.236R+.011F-478.478D+2.095P+1.8165S+.244FA}}{1+\exp^{-467.769+.422A-5.073V+.236R+.011F-478.478D+2.095P+1.8165S+.244FA}}.$$

The model for the fourth quality is

$$P(Y = \text{Quality } 4|\text{X}) = \frac{\exp^{g(X)}}{1+\exp^{g(X)}}$$

where

$$g(X) = -465.407+.422A-5.073V+.236R+.011F-478.478D+2.095P+1.8165S+.244FA.$$

Therefore,

$$P(Y = \text{Quality } 4|\text{X}) = \frac{\exp^{-465.407+.422A-5.073V+.236R+.011F-478.478D+2.095P+1.8165S+.244FA}}{1+\exp^{-465.407+.422A-5.073V+.236R+.011F-478.478D+2.095P+1.8165S+.244FA}}.$$

The model for the other quality levels follow the same equation with different intercepts, although the coefficients across these equations do not vary.

The coefficients in the model are interpreted as, for each increase of c grams per cubic decimeter of residual sugar content of white wine, the log of odds (chance) of being in the third quality level increases by $.236c$. Then the log odds change is $c\beta$ and the associate odds ratio is $\exp(c\beta)$, where β is the estimated coefficient for residual sugar.

4.2.6 Goodness-of-Fit Statistics for White Wine

The goodness-of-fit test was used to check how well the model actually reflected the data. It verifies how close the observed values match the expected under the fitted model.

Table 9 shows the column labeled "Deviance" which is the difference between the observed values and the expected values. Thus, it can be thought of as a chi-square value. Its observed significance level is 1.000. The Pearson's observed significance level is .000 which is less than .05, hence the model fits the data well.

Model	2-log likelihood	Chi-square	Sig
Intercept Only	12641.174		
Final	10902.106	1739.069	.000

Table 10: Model-Fitting Information for White Wine

Model	2-log likelihood	Chi-square	Sig
Null	10902.106		
General	10614.963	287.42	.000

Table 11: Parallel Line Test for White Wine

4.2.7 Independence Test for White Wine

The independence test was used to verify if there was little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other. We employed the Durbin-Watson test to check this assumption.

The Durbin Watson statistic for white wine is 1.621. As this statistic approaches 2, it is more likely that the residuals are independent of each other. Hence, there is not sufficient evidence to reject the null hypothesis, H_0, at $\alpha = 0.05$. Thus, they are independently distributed.

4.2.8 Model-Fitting Information for White Wine

We then check if the current model for white wine fits better than a model with just an intercept.

Table 10 shows an observed significance level of .000 which is less than $\alpha = .05$. Hence we reject the null hypothesis and conclude that the model without predictors is not as good as the model with the predictors.

4.2.9 Parallel Line Test for White Wine

From Table 11, we rejected the null hypothesis since the p-value is .000 which was less than $\alpha = .05$ and conclude that there is significant difference in the coefficients between the models. This assumption was violated since the relationship between the independent variables and the logits are not the same for all the logits. Thus, the regression coefficients are not the same for all the various categories. This then called for multinomial logistic regression which estimates separate coefficients for each category.

Coefficients
Quality
Alcohol
Sulphate
Volatile Acidity
Chloride
Total Sulfur Dioxide
Free Sulfur Dioxide
pH

Table 12: Stepwise Regression For Red Wine

4.3 Specifying the Analysis for Multinomial Logistic Regression for Red Wine

We then proceeded to analyze the data by treating it as nominal. This was required as the parallel line test for the ordinal logistic regression technique failed. Multinomial logistic regression requires a minimum ratio of valid cases to independent variables to be at least 10 to 1 (preferably 20 to 1). This requirement was satisfied by our data.

4.3.1 Table for the Stepwise Regression for Red Wine

A forward selection algorithm was used to test the contribution of all the variables after new variables have been added. As shown in Table 12, seven of the independent variables were significant.

4.3.2 Parameter Estimates for Red Wine

SPSS was used to generate the various equations. The insignificant variables were omitted from the equations. The parameter estimates have five parts, labeled with the categories of the outcome variable of quality. We checked multicollinearity in the model by examining the standard errors for the estimated coefficients. A standard error greater than 2.0 indicates numerical problems. We compared the ratio of the probability of choosing a particular outcome category over the probability of choosing the reference category.

From Tables 13 and 14, there are implausible odds caused by too few cases in some categories or complete separation of two groups. SPSS treats the eighth quality as the reference group and therefore estimated a model for the third quality relative to the eighth quality, fourth quality relative to the eighth quality, and so on. The entry labeled "B" depicts the estimated multinomial logistic regression coefficients for the models. "Std. Error"

Quality		B	Std. Error	Sig	Exp(B)
3	Intercept	.220	1.984	.988	
	Alcohol	−3.233	.755	.000	.039
	Sulphate	−6.907	1.333	.038	.001
	Volatile Acidity	9.071	1.756	.001	$8.698E3$
	Chloride	50.505	17.469	.004	$8.593E21$
	Total Sulfur Dioxide	−.045	.045	.315	.956
	Free Sulfur Dioxide	.084	.084	.317	1.087
	pH	8.585	1.681	.020	$5.352E3$
4	Intercept	−2.877	1.441	.699	
	Alcohol	−1.848	.311	.000	.158
	Sulphate	−6.897	1.972	.000	.001
	Volatile Acidity	4.898	1.087	.019	133.999
	Chloride	41.941	16.845	.013	$1.640E18$
	Total Sulfur Dioxide	.026	.018	.145	1.026
	Free Sulfur Dioxide	−.051	.046	.264	.950
	pH	6.965	1.140	.001	$1.059E3$
5	Intercept	9.892	1.723	.141	
	Alcohol	−2.185	.262	.000	.112
	Sulphate	−6.683	1.520	.000	.001
	Volatile Acidity	2.125	1.947	.275	8.372
	Chloride	39.505	16.531	.017	$1.435E17$
	Total Sulfur Dioxide	.040	.016	.013	1.041
	Free Sulfur Dioxide	−.033	.039	.396	.967
	pH	5.200	1.896	.006	181.254

Table 13: Parameter Estimates for Red Wine

Quality		B	Std. Error	Sig	Exp(B)
6	Intercept	3.210	1.616	.628	
	Alcohol	−1.393	.254	.000	.248
	Sulphate	−4.417	1.478	.003	.012
	Volatile Acidity	.095	1.923	.961	1.100
	Chloride	35.788	16.492	.030	$3.488E15$
	Total Sulfur Dioxide	.022	.016	.166	1.023
	Free Sulfur Dioxide	−.008	.039	.839	.992
	pH	4.873	1.865	.009	130.687
7	Intercept	−1.168	1.649	.861	
	Alcohol	−.649	.249	.009	.522
	Sulphate	−1.929	1.474	.191	.145
	Volatile Acidity	−2.612	1.951	.180	.073
	Chloride	29.150	16.526	.078	$4.569E12$
	Total Sulfur Dioxide	.013	.016	.434	1.013
	Free Sulfur Dioxide	−.007	.039	.854	.993
	pH	3.458	1.866	.064	31.747

Table 14: Parameter Estimates for Red Wine (Continued)

represents the standard errors for the coefficients. We did not use chloride in any of our analysis since the standard error for its coefficient was greater than 2.0 throughout the various levels of quality of wine. "Intercept" is the multinomial logit estimate for the various categories of quality relative to the eighth quality when the predictor variables in the model are evaluated at zero.

Considering the third quality relative to the eighth quality, all the independent variables are significant except free sulfur dioxide, total sulfur dioxide and the intercept according to the $p-$value. Alcohol was significant in all the quality levels and has a negative effect throughout the various levels of quality. pH had a statistically significant positive influence in all the levels except the seventh.

Let A=Alcohol, V=Volatile Acidity, S=Sulphate, T=Total Sulfur Dioxide, F=Free Sulfur Dioxide, and P=pH. The model for this quality level is given by

$$\log\left(\frac{P(\text{Quality 3})}{P(\text{Quality 8})}\right) = .220 - 3.233A - 6.907S + 9.071V + 8.585P.$$

As interpretation of the coefficients in the model, for each increase in one percentage of volume in alcohol content of red wine, the odds of being in the third quality level decreased by 96.1% ($0.039 - 1 = -0.961$).

	Chi square	Sig
Pearson	7216.183	.000
Deviance	2961.626	1.000

Table 15: Goodness-of-Fit for Red Wine Test

Model	-2log likelihood	Chi-Sq	Sig
Intercept	3.788E3		
Final	2.962E3	826.824	.000

Table 16: Model-Fitting Information for Red Wine

4.3.3 Goodness-of-Fit Statistics for Red Wine

The goodness-of-fit test was used to check if the sample came from the population with the specified distribution.

From Table 15 we see that the observed significance level for Pearson is .000 and Deviance is 1.000 which is more than .05, hence the model fits the data well.

4.3.4 Model-Fitting Information for Red Wine

Checking if the model without predictors is as good as the model with the predictors, we used the model-fitting information.

As shown in Table 16, the entry labeled "Model" shows the parameters of the model for which the model fit is evaluated. "Intercept Only" indicates a model that does not control for any predictor variables and simply fits an intercept to predict the outcome variable. The entry labeled "Final" describes a model that involves the specified predictor variables. This was obtained through a process which maximized the log likelihood of the outcome variables. The final model shows an improvement of the intercept only model. The entry labeled "Chi-square" is the difference between the two -2 log-likelihood values. The observed significance level of .000 is less than $\alpha = .05$. Hence we reject the null hypothesis and conclude that the model without predictors is not as good as the model with the predictors.

4.4 Specifying the Analysis for Multinomial Logistic Regression for White Wine

4.4.1 Table for the Stepwise Regression for White Wine

We employed stepwise regression analysis to eliminate an already selected variable if that variable was not significant because of its relationship to the other variables. Table 17 gives

Coefficients
Quality
Alcohol
Volatile Acidity
Residual Sugar
Free Sulfur Dioxide
Density
pH
Sulphates
Fixed Acidity

Table 17: Stepwise Regression for White Wine

the eight significant predictor variables for white wine.

4.4.2 Parameter Estimates for White Wine

Since we saw ordering in the data, SPSS nominated the ninth quality of the response categories as a baseline or reference cell. Multicollinearity in the model was examined using the standard errors for the coefficients. A standard error greater than 2.0 indicates numerical problems. We calculated log-odds for all six categories relative to the baseline and then let the log-odds be a linear function of the predictors.

Considering Tables 18 and 19, there are implausible odds as a results of complete separation of two groups or too few cases in some categories. The standard errors for the coefficients of density, chloride and sulphates were bigger than 2.0. Hence, there was a multicollinearity problem so none of those independent variables were used in our analysis. The ninth quality is the reference group, we therefore estimated a model for the third quality relative to the ninth quality, fourth quality relative to the ninth quality, and so on.

Residual sugar, fixed acidity and pH were statistically significant according to the p-values of the fourth quality relative to the ninth quality. Moreover, pH was significant and had a very small effect on wine quality throughout the various quality levels of the white wine.

Let A=Alcohol, V=Volatile Acidity, R=Residual Sugar, F=Free Sulfur Dioxide, P=pH, and FA=Fixed Acidity.

The model for this quality level is given by

$$\log\left(\frac{P(\text{Quality4})}{P(\text{Quality 9})}\right) = -.610R - 13.080P - 1.533FA.$$

Interpreting the coefficients in this model, wine with one gram per cubic decimeter in-

Quality		B	Std. Error	Sig	Exp(B)
3	Intercept	−925.135	1.851	.193	
	Alcohol	−.938	1.659	.380	.392
	Volatile Acidity	8.288	10.315	.143	$3.975E3$
	Residual Sugar	−.493	.035	.118	.611
	Free Sulfur Dioxide	.019	7.692	.587	1.019
	Density	982.330	714.285	.169	
	pH	−10.373	1.235	.015	$3.127E-5$
	Sulphates	−.966	.235	.854	.381
	Fixed Acidity	−.637	.409	.119	.529
4	Intercept	$-1.049E3$	.410	.129	
	Alcohol	−1.336	1.026	.193	.263
	Volatile Acidity	9.750	5.443	.073	$1.716E4$
	Residual Sugar	−.610	.303	.044	.543
	Free Sulfur Dioxide	−.060	.035	.089	.942
	Density	1.130E3	694.195	.103	
	pH	−13.080	1.995	.001	$2.087E-6$
	Sulphates	.550	4.774	.908	1.733
	Fixed Acidity	−1.533	.395	.000	.216
5	Intercept	−881.509	1.563	.199	
	Alcohol	−1.577	1.021	.122	.207
	Volatile Acidity	6.123	5.417	.258	456.343
	Residual Sugar	−.475	.300	.114	.622
	Free Sulfur Dioxide	−0.19	.035	.591	.982
	Density	967.449	690.205	.161	
	pH	−13.315	.933	.001	$1.649E-6$
	Sulphates	.776	4.711	.869	2.173
	Fixed Acidity	−1.699	.380	.000	.188

Table 18: Parameter Estimates for White Wine

Quality		B	Std. Error	Sig	Exp(B)
6	Intercept	−765.441	.283	.264	
	Alcohol	−.820	1.019	.421	.441
	Volatile Acidity	.336	5.409	.951	1.399
	Residual Sugar	−.374	.300	.212	.688
	Free Sulfur Dioxide	−.013	.035	.698	.987
	Density	842.352	688.906	.221	
	pH	−12.861	.926	.001	$2.598E-6$
	Sulphates	2.081	7.702	.645	8.010
	Fixed Acidity	−1.711	.378	.000	.181
7	Intercept	−109.598	.412	.873	
	Alcohol	−.943	.299	.354	.389
	Volatile Acidity	−1.476	5.035	.785	.229
	Residual Sugar	−.100	.149	.738	.905
	Free Sulfur Dioxide	−.009	.920	.797	.991
	Density	165.854	834.702	.810	$1.070E72$
	pH	−9.632	.374	.014	$6.561E-5$
	Sulphates	4.064	10.374	.387	58.221
	Fixed Acidity	−1.137	.284	.002	.321
8	Intercept	1.197	.584	.999	
	Alcohol	−.786	1.038	.449	.456
	Volatile Acidity	−.903	12.407	.869	.405
	Residual Sugar	−.009	5.469	.976	.991
	Free Sulfur Dioxide	.004	.306	.906	1.004
	Density	47.056	241.035	.947	$2.730E20$
	pH	−8.841	.527	.027	.000
	Sulphates	3.346	3.991	.480	28.393
	Fixed Acidity	−1.082	.407	.008	.339

Table 19: Parameter Estimates for White Wine (Continued)

	Chi-square	Sig
Pearson	36405.487	.000
Deviance	10637.549	1.000

Table 20: Goodness-of-Fit for White Wine Test

Model	-2log likelihood	Chi-sq	Sig
Intercept	1.264E4		
Final	1.064E4	2.00E4	.000

Table 21: Model-Fitting Information for White Wine

crease in residual sugar content for instance, was 45.7% ($0.543 - 1 = -0.457$) less likely to be in the fourth quality level than the ninth.

4.4.3 Goodness-of-Fit Statistics for White Wine

To determine if the observed values were significantly different from the expected values, the goodness-of-fit test was used.

From Table 20 we see that the model fits well since the observed significance level for Pearson is .000 and Deviance is 1.000 which is more than .05.

4.4.4 Model-Fitting Information for White Wine

Model-fitting information was used to check if the current model fits better than a model with just an intercept.

By looking at the results presented in Table 21, we realized the entry labeled "Sig" has observed significant level of .000 which is less than $\alpha = .05$. Hence we reject the null hypothesis and conclude that the current model fits better than a model with just an intercept.

4.5 Comparing Accuracy Rates for Red Wine

To distinguish our red wine model as useful, we compared the overall percentage accuracy rate (using the case processing summary table) to the proportional by chance accuracy (using the classification accuracy table). A useful model generally has a 25% or higher classification accuracy rate than the proportional by chance accuracy rate.

		N	Marginal Percentage
Quality	3	10	.6%
	4	53	3.3%
	5	681	42.6%
	6	638	39.9%
	7	199	12.4%
	8	18	1.1%
Valid		1599	100%
Missing		0	
Total		1599	

Table 22: Case Processing Summary Table for Red Wine

4.5.1 Case Processing Summary Table for Red Wine

Table 22 summarizes the red wine data. All 1599 observations in our data set were used in the analysis. The table shows the number and percentage of cases in each level of our response variable. The column labeled "N" presents the number of observations fitting the description in the first column. For instance, the first eight values give the number of observations for which the subject's preferred quality of wine is the third quality through to eighth one. "Valid" shows the number of observations in the data set where the outcome variable and all predictor variables are non-missing. The marginal percentage gives the proportion of valid observations found in each of the outcome variable's groups. It was calculated by dividing the N for each group by the N for "Valid". Of the 1599 subjects with valid data, 10 subjects preferred the third quality to all the other quality levels. Hence, the marginal percentage for this group is $(10/1599) \times 100 = .6\%$.

In this regression, the outcome variable is quality which contains a numeric code for the subject's preferred quality of wine. The data includes nine levels of quality representing nine different preferred quality levels of wine. "Missing" represents the number of observations in the dataset where data are missing from the response variable or any of the predictor variables. "Total" shows the total number of observations in the dataset.

Based on the "Case Processing Summary" in Table 22, we computed the proportional by chance accuracy rate using the fraction of cases for every group found on the number of cases in every group. The result was achieved by squaring and summing the proportion of cases in each group as

$$(0.006^2 + 0.033^2 + 0.426^2 + 0.399^2 + 0.124^2 + 0.011^2) = 0.357299.$$

We then multiplied our results by 1.25 since the classification accuracy rate must be 25%

				Predicted			
Observed	3	4	5	6	7	8	Percentage Correct
3	1	1	7	1	0	0	10.0%
4	1	1	34	17	0	0	1.9%
5	0	2	518	157	4	0	76.1%
6	0	0	209	386	43	0	60.5%
7	0	0	12	129	58	0	29.1%
8	0	0	0	10	8	0	.0%
Overall Percentage	.1%	.3%	48.8%	43.8%	7.1%	.0%	60.3%

Table 23: Classification Accuracy Table for Red Wine

larger than the proportional by chance accuracy rate. Hence proportion of chance criteria is 44.7% ($0.357299 \times 1.25 = .447 = 44.7\%$). That is, we hoped to see a classification accuracy of 44.7% or higher.

4.5.2 Classification Accuracy Table for Red Wine

The classification accuracy table is a table of predicted group membership against actual group membership. Generally, the classification accuracy rate must be 25% higher than the proportional by chance accuracy rate. From Table 23, the classification accuracy rate was 60.3%.

The classification accuracy rate for red wine was 60.3% which was greater than the proportion of chance criteria of 44.7%. Therefore, the classification accuracy criteria is satisfied.

4.6 Comparing Accuracy Rates for White Wine

To characterize our white wine model as useful, we compared the overall percentage accuracy rate to the proportional by chance rate.

4.6.1 Case Processing Summary Table for White Wine

We see from Table 24 that all 4898 white wine observations in our data set were used in the analysis. By looking at N, the first seven values give the number of observations for which the subject's preferred quality of wine is the third quality through ninth one. Considering the marginal percentages, of the 4898 subjects with valid data, 20 samples were rated at a quality level of 3. Hence, the marginal percentage for this group is $(20/4898) \times 100 = .4\%$.

Using Table 24, we then calculated the proportional by chance accuracy rate using the

		N	Marginal Percentage
Quality	3	20	.4%
	4	163	3.3%
	5	1457	29.7%
	6	2198	44.9%
	7	880	18.0%
	8	175	3.6%
	9	5	.1%
Valid		4898	100%
Missing		0	
Total		4898	

Table 24: Case Processing Summary Table for White Wine

percentage of cases for each group found on the number of cases in each group. We arrived at the results by squaring and adding the percentage of cases in each group as

$$(0.004^2 + 0.033^2 + 0.297^2 + 0.449^2 + 0.18^2 + 0.036^2 + 0.001^2) = 0.324612.$$

Since the classification accuracy rate must be 25% higher than the proportional by chance accuracy rate we proceeded to multiply this result by 1.25. Therefore, the criteria for proportion of chance is 40.6% ($.324612 \times 1.25 = .4058 = 40.6\%$). Hence, a classification accuracy of 40.6% or higher is expected.

4.6.2 Classification Accuracy Table for White Wine

The classification accuracy table has each case predicted to be a member of the group to which it has the highest probability of belonging. Table 25 shows that the classification accuracy rate for white wine was 53.8%.

In conclusion, the classification accuracy rate for white wine was 53.8% which was greater than the proportion of chance criteria of 40.6%. Therefore, the classification accuracy criteria was satisfied.

4.7 Comparison with Previous Model

We then compared our model to the results of Cortez et al [9]. Their accuracy rates for support vector machines (SVM) were 89.0% for red wine and 86.8% for white wine which were relatively higher than ours. However, our model is much simpler.

				Predicted				
Observed	3	4	5	6	7	8	9	Percentage Correct
3	2	0	8	8	2	0	0	10.0%
4	0	8	94	60	1	0	0	4.9%
5	0	2	780	664	11	0	0	53.5%
6	0	2	410	1639	146	0	1	74.6%
7	0	0	39	636	205	0	0	23.3%
8	0	0	11	117	47	0	0	.0%
9	0	0	0	1	4	0	0	.0%
Overall Percentage	.0%	.2%	27.4%	63.8%	8.5%	.0%	.0%	53.8%

Table 25: Classification Accuracy Table for White Wine

5 Conclusion

The estimated parameters for the SPSS output contain six (from 3 to 8) classes of quality level for red wine and seven (from 3 to 9) classes for white wine. We used ordinal logistic regression and the multinomial logit model to estimate the preference of wine based on its physicochemical properties. Although the ordinal logistic regression model minimized the sum of squared errors, it failed the parallel line test so the multinomial logistic regression technique was used. The multinomial logistic regression model was less sensitive to outliers.

5.1 Results for Red Wine

The accuracy rate for red wine was 60.3% which satisfied the classification accuracy criteria. Alcohol was statistically significant and had a negative effect throughout the various levels of red wine. As we moved from a lower quality level to a higher quality level, sulphate and pH went from being statistically significant to not significant statistically.

5.2 Results for White Wine

Our model has an accuracy rate of 53.8% for white wine which met the classification accuracy criteria. The relative importance of the inputs brought interesting insights regarding the impact of the components on the quality of white wine. We realized from our model that pH was statistically significant and had a negative effect throughout the various levels of white wine. Moreover, fixed acidity and residual sugar also had negative impact on the predicted levels of white wine quality.

Since the result of this work is an important tool for the global wine market, such a model could be used to enhance the training of enology students and improve the quality of wine.

References

[1] A. Agresti, "An Introduction to Categorical Data Analysis", John Wiley and Sons, Inc., New Jersey, 1996.

[2] A. Asuncion and D. Newman, UCI Machine Learning Repository, University of California, Irvine, http://archive.ics.uci.edu/ml/, Accessed 10 July 2010.

[3] V. Barthet, H. M. Chan, H. V. Kuhnlein and D. Leggee, Macronutrient, mineral and fatty acid composition of Canadian arctic traditional food, *Journal of Food Composition and Analysis* **15** (2002), 545-566.

[4] L. Bisson, S. Ebeler, J. Lapsley, A. Waterhouse and M. Walker, The present and future of the international wine industry, *Insight Progress* **418** (2002), 1-4.

[5] D. Block, J. Ferrier and S. Vlassides, Using historical data for bioprocess optimization: Modeling wine characteristics using artificial neural networks and archived process information, *Biotechnology and Bioengineering* **73** (2001), 20-26.

[6] R.N. Bolton, and J.H. Drew, A multistage model of customers' assessments of service quality and value, *Journal of Consumer Research* **17** (1991), 375-84.

[7] Comissão de viticultura da região dos vinhos verdes (CVRVV), http://www.vinhoverde.pt, Accessed 5 July 2010.

[8] Comissao de viticultura da regiao dos vinhos verdes (CVRVV), Export Statistic, http://www.vinhoverde.pt/EN/estatistica/exportacao.htm. Retrieved 26 July 2010.

[9] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems, Elsevier **47** (2009), 820-854.

[10] L. Cullere, A. Escudero, V. Ferreira, F. S. Juan, Juan Cacho, P. N. F. Ndez-Zurbano and M. P. Saenz-Navajas, Modeling quality of premium Spanish red wines from gas chromatography-olfactometry data, *Journal on Agric. Food Chem* **57** (2009), 7490-7498.

[11] M. A. Efroymson, Mathematical methods for digital computers, in A. Ralston and H. S. Wilf, ed, "Multiple regression analysis" John Wiley and Sons, New York, 1960.

[12] N. L. Gilbert, The atom and the molecule, *Journal of the American Chemical Society* **38** (1986), 762-786.

[13] H. L. Gilmore, "Product Conformance Cost", Quality Progress, 1974.

[14] A. F. Holleman and E. Wiberg, "Inorganic Chemistry", Academic Press, San Diego, 2001.

[15] T. Lima, Price and quality in the California wine industry: An empirical investigation, *Journal of Wine Economics* **1** (2006), 176-190.

[16] L. S. Lockshin and W. T. Rhodus, The effect of price and oak flavor on perceived wine quality, *International Journal of Wine Marketing* **5** (1993), 13-25.

[17] R. Margolskee and D. Smith, Making sense of taste, *Scientific American* **3** (2006), 84-92.

[18] P. Patnaik, "Handbook of Inorganic Chemicals", McGraw-Hill, New York, 2002.

[19] D. Polit, "Data Analysis and Statistics for Nursing Research", Appleton and Lange, Stamford, Connecticut, 1996.

[20] J-B. E. M. Steenkemp, "Product Quality", Van Gorcum, Assen, 1989.

[21] V.A. Zeithaml, Consumer perceptions of price, quality and value: a means-end model and synthesis of evidence, *Journal of Marketing* **52** (1988), 2-22.