**WINE QUALITY PREDICTION THROUGH GENERALISED LINEAR MODEL AND HYPOTHESIS**

**SELVA VIGNESH.M**

M.Sc. Data Science,

Department of Data Science,

Coimbatore Institute of Technology,

Coimbatore.

[saiselva2302@gmail.com](mailto:saiselva2302@gmail.com)

[nikinprasad1845@gmail.com](mailto:nikinprasad1845@gmail.com)

**ABSTRACT:**

This project will explore wine dataset for red wine quality. The objective is to explore which chemical properties influence the quality of red wines. As interesting relationships in the data are discovered, produce and refine plots to illustrate them. First try to get a feel of the variables on their own and then try to find out the correlation between them and the Wine Quality with other factors thrown. Finally a linear model is created to predict the outcome of a test set data. Various methods are performed to find the actual quality of wine. It also proves that the better prediction can be made if selected features (variables) are being considered rather than considering all the features.

**KEYWORDS:**

ANOVA, Linear regression, Tests.

**INTRODUCTION:**

Today, all type of industries is improving by adopting the new technologies and applying these in all areas. These technologies are also helpful to enhance the production and making the whole process smooth. But, still there are different areas, which demands human expertise such as product quality assurance. Nowadays, it becomes an expensive process as the demand of product is growing over the time. Therefore, this paper explores different statistical techniques such as linear regression, ANOVA and tests for product quality assurance. These techniques performs quality assurance process with the help of available characteristics of product and automate the process by minimizing human interfere. Wine quality can be assessed by two types of tests, first is physicochemical test and second is sensory test [1]. Physicochemical test can be determined by lab tests and no human expert is required but for sensory test, a human expert is required. The work also identifies the important features to predict the values of dependent variables. Moreover, Wine quality assessment is very difficult as the relationships between the physicochemical and sensory analysis are complex and still not fully understood [2]

**DATASET:**

In this work, Wine dataset is used for all the experiments. The dataset is taken from the website [www.kaggle.com/dataset/winedata](http://www.kaggle.com/dataset/winedata). The wine data set consists of 75 wines and 12 physiochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol, and quality [3]. The quality rating is based on a sensory test and the quality classes from 0 - very bad to 10 - very excellent. Moreover, some variables have values between 0 and 1. Such type of inconsistency may affect predictions due to more influence making by some variables than others.

**STATISTICAL TOOLS:**

In this project, we used some statistical concepts to interpret the quality of wine. The tools are Generalized Linear Regression Model, ANOVA, Sign test, Kruskal Wallis test, Man-Whitney Wilcoxon test, Spearman’s rank Correlation Coefficient and Wilcoxon signed rank test [4]. The analysis of variance is done by one way and two way ANOVA. They give us the significant difference between the chemicals that influence the taste and quality of wine. The other tests gives whether the variables are identical or not. The answers from the tests are used for the correction in the quantity of chemicals of wine.

**ANOVA:** ANOVA or Analysis of Variance is a group of statistical models to test if there exists a significant difference between means. It tests whether the means of various groups are equal or not. In ANOVA, the variance observed in a particular variable is partitioned into different components based on the sources of variation. An important fact to note is that while we use ANOVA to find out whether the means differ significantly. Here we have used one-way and two-way ANOVA. The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. The weight and primary qualities are taken into consideration for ANOVA. They show a significance in quality. Specifically, it tests the null hypothesis

One-way ANOVA Null Hypothesis

Where µ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (HA), which is that there are at least two group means that are statistically significantly different from each other. The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable. The interaction term in a two-way ANOVA informs you whether the effect of one of your independent variables on the dependent variable is the same for all values of your other independent variable (and vice versa).The R functions used for one way and two way ANOVA are aov and oneway.test

**SIMPLE LINEAR REGRESSION:** Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. The weight and quality variables produce intercepts which is used to refine plots for regression and observe their change. The R function used here is lm and glm.

*Y (Pred) = a + b\*x*

**MAN-WHITNEY WILCOXON TEST:** A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test. The Man Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations. Recall that the parametric test compares the means (H0: μ1=μ2) between independent groups. Where H0 states that the two populations are equal and H1, the two populations are not equal. The weight and quality variables doesn’t satisfy this test.

U1=n1n2+ (n1 (n1+1))/2-R1

U2=n1n2+ (n2 (n2+1))/2-R2

Here U1, U2 is the Man Whitney calculation for sample 1 and sample 2.n1, n2is the weight and quality variables. R1 and R2 are the sum of ranks

**SIGN TEST:** The Sign test is a non-parametric test that is used to test whether or not two groups are equally sized. The sign test is used when dependent samples are ordered in pairs, where the bivariate random variables are mutually independent. The sign test is considered a weaker test, because it tests the pair value below or above the median and it does not measure the pair difference. Here the factors taken are quality and weight of primary substances. SIGN.test is the function used.

**KRUSKAL WALLIS TEST:** The Kruskal-Wallis test is a nonparametric (distribution free) test, and is used when the assumptions of one-way ANOVA are not met.  Both the Kruskal-Wallis test and one-way ANOVA assess for significant differences on a continuous dependent variable by a categorical independent variable (with two or more groups). Kruskal-Wallis test can be used for both continuous and ordinal-level dependent variables. Kruskal.test function uses the same factors for consideration.

**WILCOXON SIGNED RANK TEST:** The Wilcoxon Sign test is a repeated measures test of dependency. This test is mathematically similar to conducting a Mann-Whitney U-test. It is a non-parametric analysis that statistically compared of the average of two dependent samples (weight, Qual) and assess for significant differences and also a non-parametric alternative of the dependent samples t-test. The R function used here is Wilcox.test

**SPEARMANS RANK CORRELATION COEFFICIENT:** The Spearman’s Rank Correlation Coefficient is the non-parametric statistical measure used to study the strength of association between the two ranked variables. This method is applied to the ordinal set of numbers, which can be arranged in order, i.e. one after the other so that ranks can be given to each. In the rank correlation coefficient method, the ranks are given to each individual on the basis of its quality, such as ranking starts from position 1st and goes till Nth position for the one ranked last in the group. D is the difference between primary chemicals and weight. The R function used is simply correlation factor.

**PLOTS:**

These are the plots for linear regreesion. They infer the realtion between the residuals and leverage. The dots are the measures of chemicals used and the line connects them to have a linear trend.

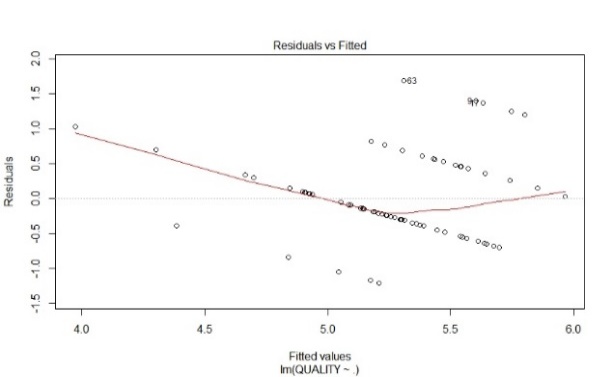
****

Figure 1: Linear plot

The figure 1 represents the linear model with x axis fitted values lm(weight ~ Qual) and y axis as residuals

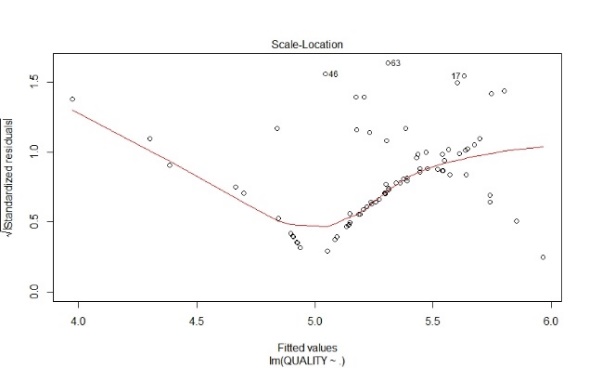
****

Figure2: Quality Plot

The figure 2 represents the quality plot with x axis and square root of residuals as y axis.

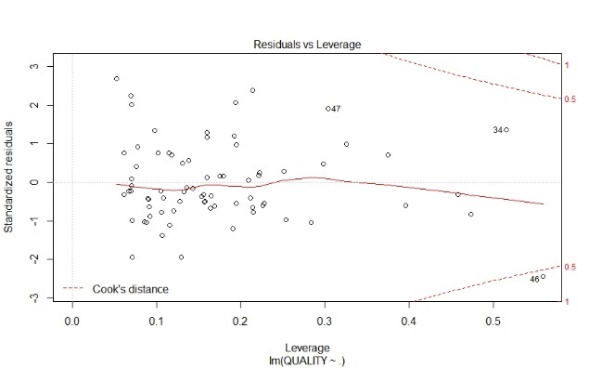
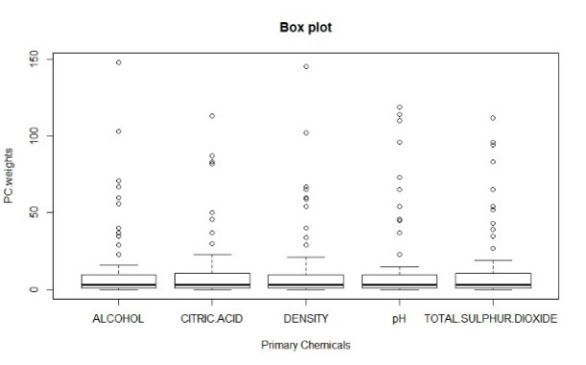
****

Figure 3 : Residual plot

.

The figure 3 represents the x axis as theoretical quantities lm(weight\*Qual) and y axis as standardized residuals.

****

*Figure 4: Box plot*

The figure 4 states that there is no significant change in the middle line, which shows that the primary chemicals (citric acid, pH, alcohol, density and total sulphur dioxide) are in their right proportions. The middle line helps us to identify that there are any variations in their values in graph. The chemicals are on the x-axis and their respective weights in the y-axis. This helps in better understanding of the model.

**INFERENCE:**

The interest has been increased in wine industry in recent years which demands growth in this industry. Therefore, companies are investing in new technologies to improve wine production and selling. This paper explores the usage of R and statistical techniques in two ways. Firstly, how linear regression determines important features for prediction. Secondly, the usage of tests in predicting the values. One (quality) is dependent variable and other 11 are predictors.

The experiments shows that the value of dependent variable can be predicted more accurately if only important features are considered in prediction rather than considering all features. The wine quality certification plays a very important role for both processes and it requires wine testing by human experts. From the tests conducted above we infer that the hike in any of the primary substances may affect the quality of wine. We conclude that the taste of wine depends on the primary chemicals like density, alcohol, pH, total sulphur dioxide, citric acid. Absence of variation in the chemicals infers that they are in right proportions and makes the wine tasty. Rise in their values make it a bitter one and thus reduces quality. From the tests we can know the limits and can prepare the wine accordingly for the satisfaction of taste and quality of the customers.

**RESULTS AND DISCUSSIONS:**

In this project the quality of wine is predicted .The dataset is obtained from rstudio website [6]. Consider the input variables as the chemicals used for the production of wine. It is separated to two types of variables. The primary variables that determine the good quality of wine such as Density, pH, Alcohol, Citric acid, Total sulphur dioxide. The secondary variables present in the wine are fixed acidity, Volatile acidity, free sulphur dioxide, Residual sugar and Chlorides. ANOVA concepts are used to show the difference in the variance of variables. The difference in the variance shows the improper mixture of chemicals. To know this, use the boxplot to find the midline where the mean of variances are equal or not. One way and two way ANOVA are done. For the dataset, ANOVA is rejected as the null hypothesis value is greater than 0.05. As it got rejected we are proceeding with the tests to find whether it is getting accepted or not. The tests used are Sign test, Wilcoxon signed rank test, Man Whitney Wilcoxon, Kruskal Wallis, Spearman’s correlation coefficient test. They show their significances between the chemicals used for the production of wine.

**REFERENCES:**

**[1]** Ebeler S. (1999) “Flavour Chemistry — Thirty Years of Progress: chapter Linking flavour chemistry to sensory analysis of wine”. Kluwer Academic Publishers, 409–422.

**[2]** Legin, Rudnitskaya, Luvova, Vlasov, Natale and D'Amico. (2003) “Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception”. Analytica Chimica Acta.

**[3]** Quantitative analysis and correlation with human sensory perception”. Analytica Chimica Acta 484 (1): 33–34. [3] Sun, Danzer and Thiel. (1997) “Classification of wine samples by means of artificial neural networks and discrimination analytical methods”. Fresenius Journal of Analytical Chemistry

**[4]** Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani, ISLR: Data for An Introduction to Statistical Learning with Applications in R, <http://www.StatLearning.com>

**[5]** <http://www.datasciencemade> simple.com/r/

**[6]** <https://www.kaggle.com/> winedata