Individual Assignment 5, Dion Chang - 20812576

1) Problem statement and data used: The red and white wine data sets were combined in order to build a logistic regression model that predicts whether a wine is red or white. The data set has 1599 observations for red wine and 4898 observations for white wine, making a total of 6497 observations. The features used in the model are pH, sulphates, chlorides, volatile acidity, total sulfur dioxide, and density.

2) Planning: For the data wrangling and cleaning, a new column called "winetype" was added for both red and white wine data sets. It was a binary variable where 1 was for red wine while 0 was for white. The data sets were then combined to make one large data frame to work with. Since the assumption of linearity of the logit was initially not met, a square root transformation on the selected predictor variables was performed.

After the logistic regression model is built, the model will be interpreted by looking for significant predictors as well as deviance statistics. The odd ratios for the predictors in the model will also be determined. Furthermore, the assumptions of logistic regression will also be tested. The following three are to be diagnosed: linearity of the logit, independence of errors using the Durbin-Watson test, and multicollinearity using VIF statistics.

3) Analysis:

Model: From the summary of the logistic regression model performed in R, the value of the deviance for the "null" model (only the constant in the model) was determined to be 7250.98. However, when the square-root transformed features: pH, sulphates, chlorides, volatile acidity, total sulfur dioxide, and density was included, the deviance value was reduced to 610.45. The model chi-square statistic was determined to be **6640.53** by calculating the difference between the residual and null deviance. This reduction meant that the model with the features added was much better at predicting whether a wine is red or white than the "null" model. The degrees of freedom is 6, reflecting the number of variables in the model. This information was used to determine the probability associated with the chi-square statistic ($p < .001$); since the probability was less than .05, this further shows that the model fits the data significantly.

Table 1 shows that all the variables are significant predictors for the type of wine at the 5% level of significance. Since the data was transformed, the it would be difficult to interpret the coefficients and therefore they were omitted form this report. Since logistic regression has similar assumptions as linear regression, the following from Pek et al.(2017) applies: "While the square root transformation can be said to stabilize the variance of the residuals and remove nonlinearity in effects, applying the linear regression to the transformed data results in virtually uninterpretable regression coefficients in the square root scale" (p. 8). Although the transformation could be reversed, it would give large coefficient values and therefore not recommended. As Pek et al. (2017) stated:

> Because such transformed data are on a different scale compared to the original data (e.g., natural log of reaction time instead of reaction time), the nature of the effect as operationalized by the original variable and its interpretation changes due to the transformation. To address this complication in interpretation, several authors and textbooks have recommended reverse transformations after conducting inference on the transformed variables. However, inferential results (i.e., NHSTs and CIs) associated with the simple reverse transformation does not necessarily map back onto the original effect of interest, and we strongly discourage the use of reverse transformations. (p. 4)

Table 1: p-value of the variables after square-root transformation.

Variable	pvalue
Intercept	$< .001$
squareroot.pH	$< .001$
squareroot.sulphates	$< .001$
squareroot.chlorides	$< .001$
squareroot.volatile.acidity	$< .001$
squareroot.total.sulfur.dioxide	$< .001$

Variable	pvalue
squareroot.density	$< .001$

Note: $R^2 = .92$ (Hosmer-Lemeshow)

Linearity of of the logit testing: In order to test this assumption, the logistic regression must include the interaction between each predictor and the log of the same predictor. Therefore, interaction terms of each feature was created by multiplying the log of a variable by the same original variable. For example in Table 2, the interaction of pH is equal to the log value of pH multiplied by the original value of pH:

Table 2: Output of the linearity of the logit assumption test.

Variable	Coefficient	Pvalue
log.pH.int	-177.38	.22
log.sulphates.int	-11.68	.36
log.chlorides.int	-76.04	$< .001$
log.volatile.acidity.int	-9.03	.30
log.total.sulfur.dioxide.int	11.78	.75
log.density.int	-175042.27	.24

Table 2 shows that since all interactions except for chlorides have significant values, this indicates that the assumption of linearity of the logit has been met for pH, sulphates, volatile acidity, total sulfur dioxide, and density.

Multicollinearity: Table 3 gives the VIF and tolerance (1/VIF) values for the predictors in the logistic regression model. As shown, there is no concern for collinearity as the VIFs are way below 10 and the tolerances are much greater than 0.2. The average VIF was determined to be 1.26, close to 1 which is good.

Table 3: VIF and tolerance for multicollinearity testing.

	VIF	Reciprocal.VIF
square.pH	1.167824	0.8562932
square.sulphates	1.190816	0.8397604
square.chlorides	1.248153	0.8011836
square.volatile.acidity	1.152564	0.8676310
square.total.sulfur.dioxide	1.468560	0.6809393
square.density	1.320105	0.7575154

Independence of errors: The Durbin-Watson test determined that the d value was close to 2 (1.50), which means that there was no autocorrelation detected.

Conclusion: A logistic regression model predicting if the wine is red or white was built from pH, sulphates, chlorides, volatile acidity, total sulfur dioxide, and density. Assumptions of the model were all met, except for chloride with the linearity of the logit test. This means that we would be able to generalize our findings beyond just this data set sample with pH, sulphates, volatile acidity, total sulfur dioxide, and density. Since the data underwent a square-root transformation, the coefficients cannot be interpreted directly as its values have been inflated dramatically. In other words, the odds ratio are very large. Overall, the data fits the model very well as proven by the Hosmer-Lemeshow test which is a test for the goodness of fit for logistic regression. This proves that the logistic regression model built predicts the outcome variable (red or white wine) really well.

References

Field, A., Miles, J., &; Field Zoe. (2014). Discovering statistics using R. SAGE Publ.

Pek, J.; Wong, O.; and Wong, A. C. (2017) "Data Transformations for Inference with Linear Regression: Clarifications and Recommendations," Practical Assessment, Research, and Evaluation: Vol. 22 , Article 9. DOI: https://doi.org/10.7275/p86s-zc41. https://scholarworks.umass.edu/pare/vol22/iss1/9