Red Wine Project Step 2

Abel Tekle and Ethan Mayoss

11/5/2023

The red wine data set contains a collection of red wine samples with factors that make up the wine such as alcohol content, pH level, and residual sugar quantity, along with a quality score as rated by critics. The goal of this step of the project was to determine if there is a linear relationship between two continuous quantitative variables. Looking at the heat map created for step one of the project, the two variables that were chosen for analysis were "alcohol" and the "density" of the wine, as they have a relatively high correlation on the map. The alcohol content variable, or "alcohol" in the data set, is the percentage of alcohol that makes up each wine.

The null hypothesis addressed is that the alcohol content and the density of wines are not linearly related, while the alternative hypothesis is that the two variables are linearly related. In order to reject the null hypothesis, β_1 needs to be significantly different than zero:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

For ease, since alcohol is listed as a percentage rather than a decimal, a new variable for the data set was created, called alcoholUpdate, where the values are equal to the values of alcohol divided by 100. This does not change the model other than making the alcohol content values and the resulting statistics from the model more readable.

With density as the response variable, and alcoholUpdate as the predictor, the following fitted linear model is produced(rounded to 3 decimal places):

$$\text{density} = 1.004 - 0.071 * \text{alcoholUpdate}$$

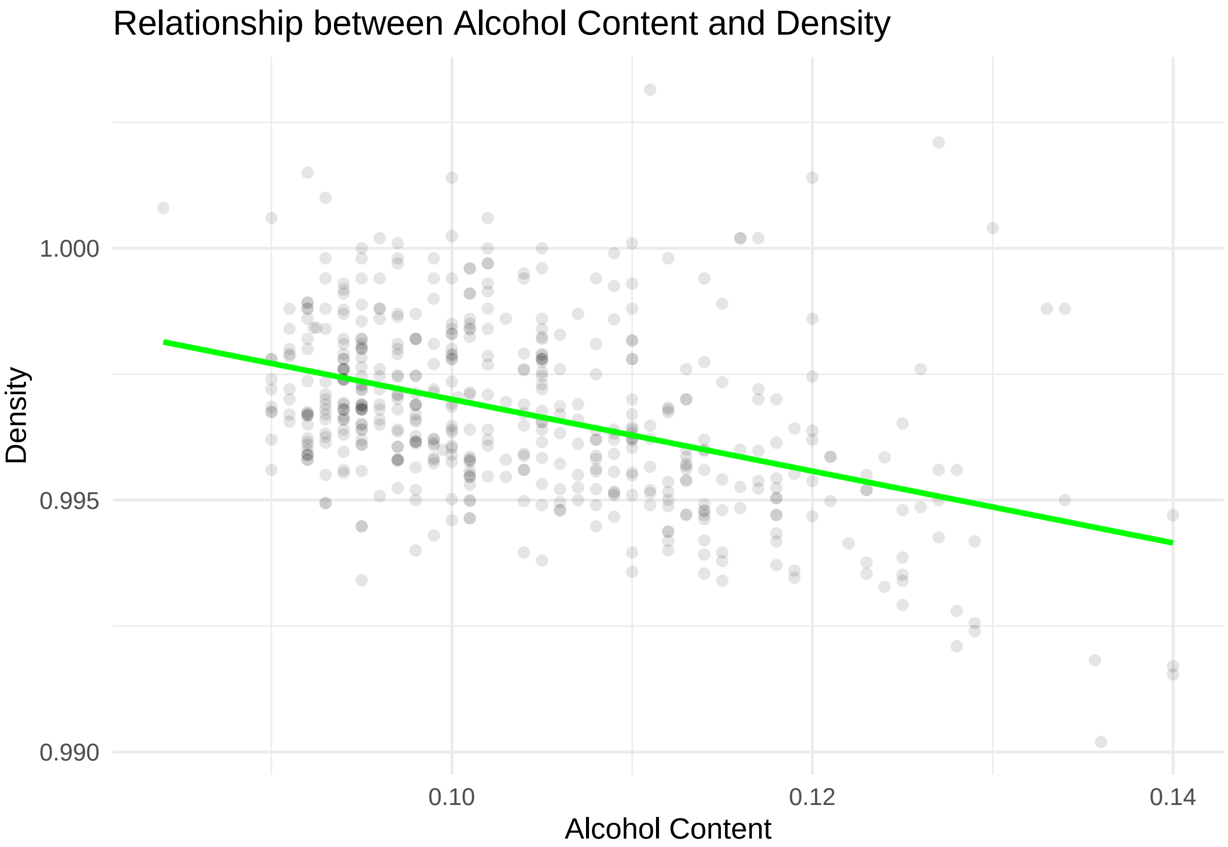
This suggests that for every one unit increase in alcohol content, the density decreases by 0.071 g/dm^3. This means that every percentage increase in alcohol of the wine, the density decreases by 0.00071 g/dm^3. Considering that all of the wines in the sampled data set have a density less than 1 g/dm^3, this is not a very small change.

To conduct the hypothesis test stated above, testing if β_1 is significantly different than zero, which suggests that alcohol and density have a linear relationship, a 95% confidence interval of the true population value of β_1 is made:

The confidence interval has a lower bound of -0.08488468 and an upper bound of -0.05762622, suggesting at the 95% confidence level, β_1 is significant, as the interval does not contain 0. This shows us that the two variables have a linear relationship, and the null hypothesis can be rejected.

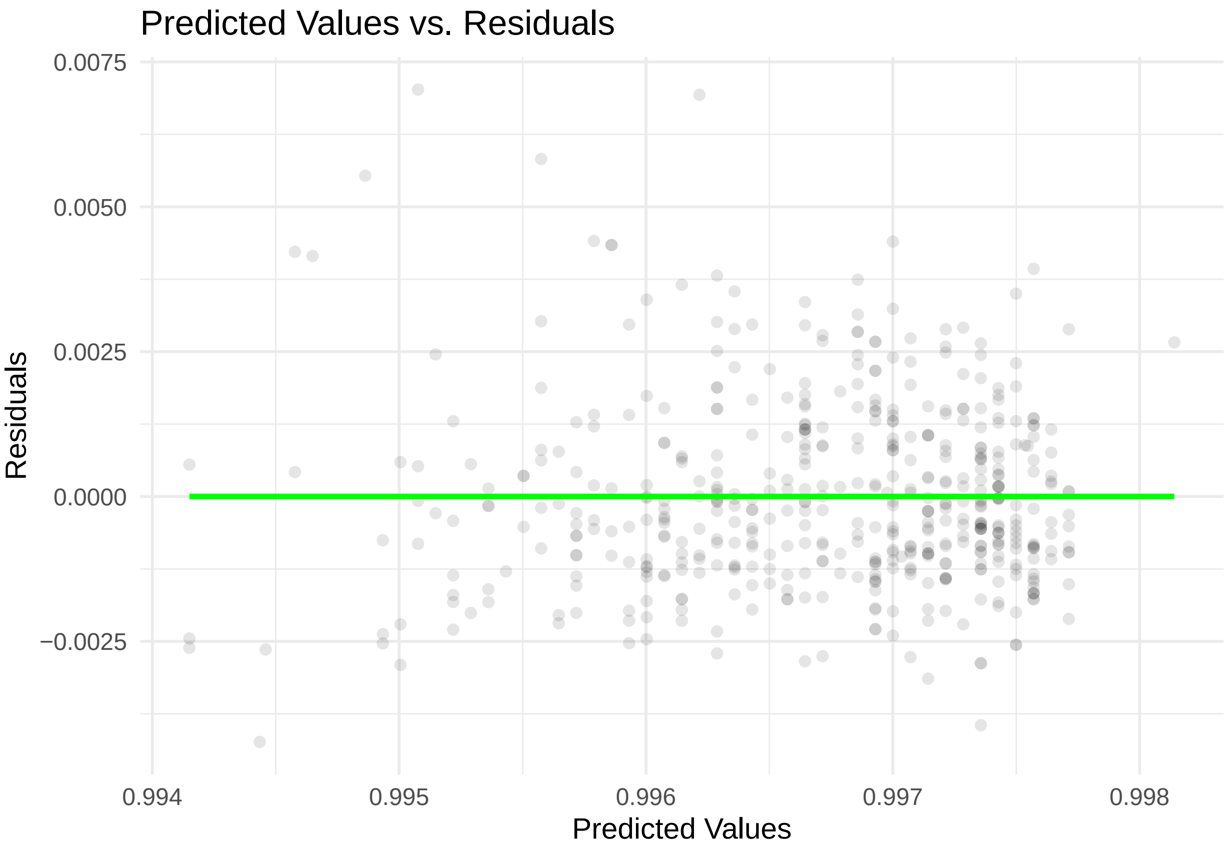
Before the analysis of a linear model between the alcohol content and density can take place, the assumptions for linear regression must be checked. This can be done through a series of plots.

To check the linearity assumption, $\mathbb{E}[\varepsilon_i] = 0$, a scatter plot was created with the values of alcoholUpdate on the x-axis, and the values of density on the y-axis:



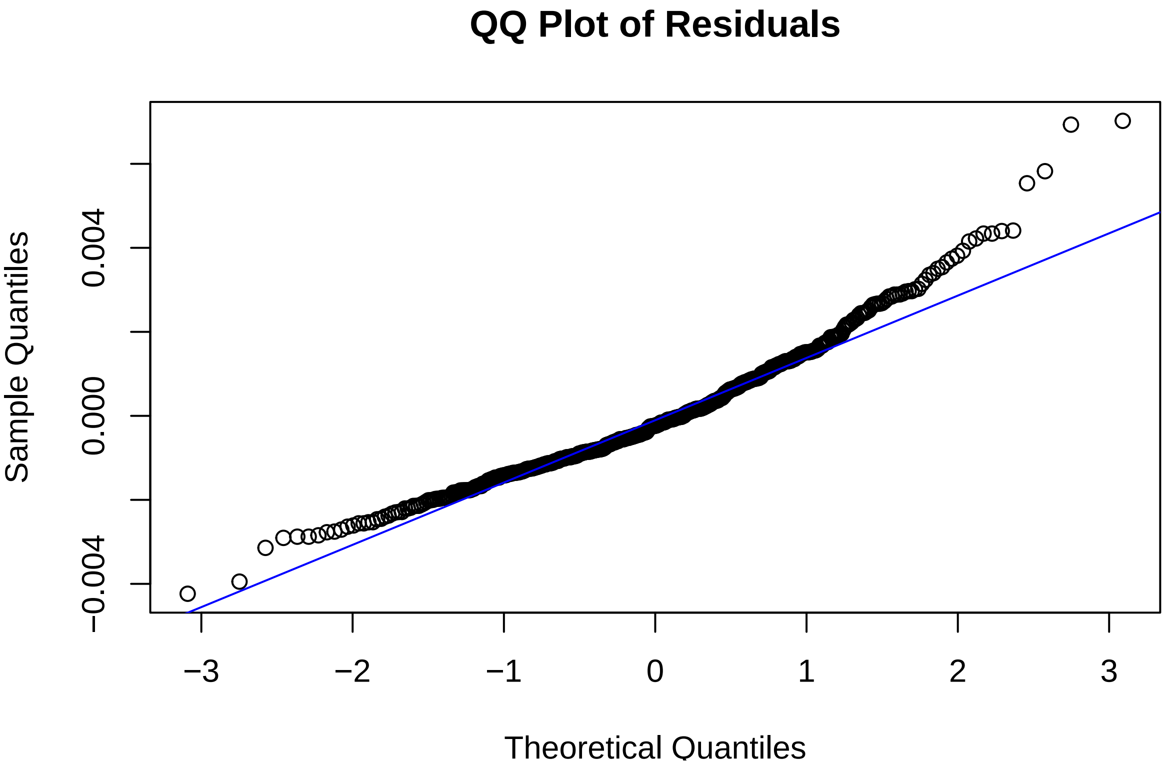
Based on how the distance between the plot points and the fitted line, or the error terms, are evenly spread across both sides of the line, the alcohol content and density of the wines do appear to fulfill the linearity assumption.

A plot is formed with the residuals of the model on the y-axis, and predicted values on the x-axis to show the assumptions of homoscedasticity, or equal variance, and independence:



The plot shows a fairly random scatter, with no distinguishable pattern, suggesting that the variances of the error terms are constant, and the homoscedasticity assumption is fulfilled. Additionally, since the correlation of the points is approximately zero, suggesting that there is no correlation between the predicted values of density and the residuals of the model, the independence assumption is satisfied by the model. While the residuals vs. predicted values plot suggests that the assumptions are met, the R^2 value of the linear model is 0.1748, which suggests there are other factors contributing to density. This makes sense, as wine is made up of many different components that could affect the density, rather than just alcohol content, as seen in the model. Additionally, we attempted to apply a log transformation to see if it would help the model become more linear, but we actually found that the transformation slightly worsened our results. R^2 decreased slightly while the range of residuals increased insignificantly.

Additionally, the distribution of the residuals of the model were checked to see if they resembled a normal distribution, using a Q-Q plot:



In the Q-Q plot analysis, the residuals were largely consistent with the expected pattern for a normal distribution, except for deviations at the tails. These deviations suggest the presence of outliers or extreme values in the data set. These outliers could potentially have a disproportionate effect on the regression analysis, exerting leverage and possibly skewing the results. While such deviations from normality do not necessarily invalidate the regression model, they do warrant further investigation.

Analyzing the relationship between alcohol content and wine density, we centered on the mean alcohol content to compute the mean density prediction. The 95% confidence interval was extremely precise (0.99657, 0.9968543), suggesting a strong estimation of the mean density at this average alcohol content. Conversely, the prediction interval was broader, (0.9935308, 0.9998935), accounting for individual sample variability and indicating where a random wine's density might lie, given the mean alcohol content. The tight confidence interval conveys robustness in our model's ability to predict average density, while the wider prediction interval acknowledges the expected fluctuations in individual observations.

In conclusion, the analysis showed a significant inverse relationship between alcohol content and density of the wine samples. The linear model explains approximately 17.48% of the variability in wine density, which suggests that other factors also contribute to density. The residual plots did not reveal any obvious patterns, indicating that the model's assumptions are reasonable. The data largely confirmed expectations, but the relatively low R^2 value suggests that further investigation could reveal more about the factors influencing wine density. Future analysis could explore the effects of other variables, such as acidity or sugar content on the density of wine.