Understanding Alcohol Content in Wine

STAT 425: Statistical Modeling I

Matthew Klima & Gabe Price

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Background

Introduction to the Case Study:

- The objective of this analysis is to predict and understand the factors influencing alcohol content in wine based on its chemical properties.
- We aim to identify the best model through statistical analysis, model selection techniques, and diagnostic validation.

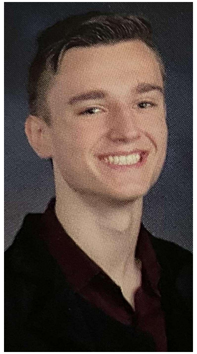
Team Members:

- Gabe Price and Matthew Klima

Overview of the Presentation:

- Introduction and Objectives
- Data Analysis Steps
- Results and Key Findings
- Conclusions and Takeaways

Gabe Price - B.S. in Statistics



Matthew Klima - B.S. in Statistics

Exploratory Data Analysis

Goal:

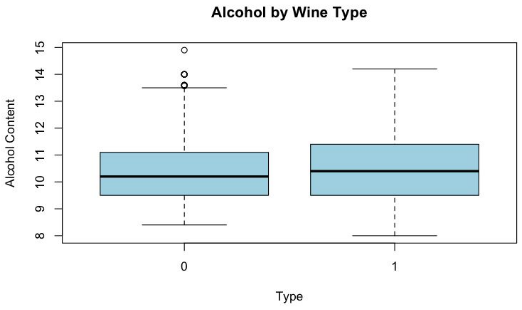
- Understand the data structure, check for missing values, and identify patterns.

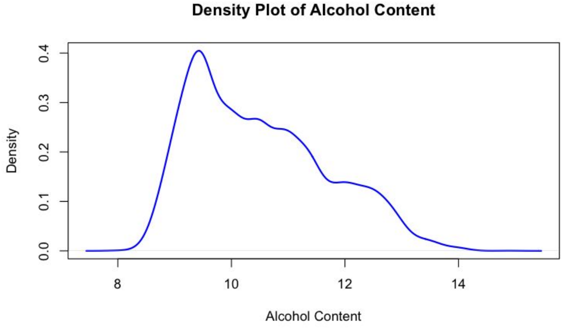
Steps Taken:

- Checked for missing data and ensured no NAs existed.
- Converted type into a binary variable (0 = redwine, 1 = whitewine).
- Used summary statistics to evaluate predictors.
- Created density plots of alcohol content to analyze the central tendency and variability of our response variable.
- Utilized a correlation matrix to surface level check for multicollinearity.

Insightful Conclusion:

- We identified potential multicollinearity issues and patterns in residual plots, guiding the next steps for modeling.





Model A - Subset Selection

Goal:

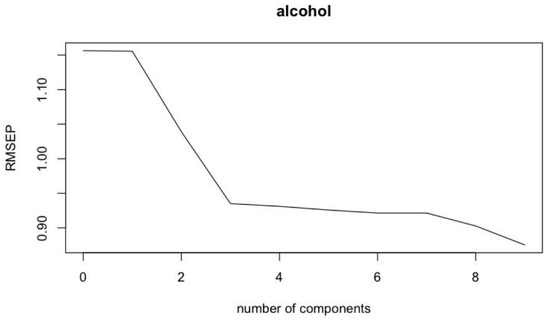
- Identify the most significant predictors of alcohol content while addressing multicollinearity concerns.

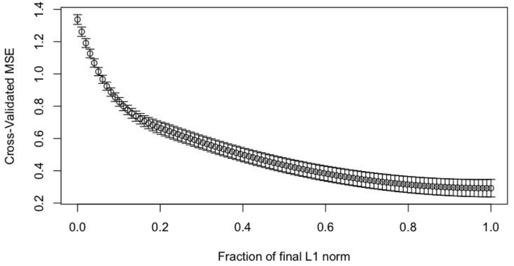
Steps Taken:

- Used stepwise selection to identify the best predictors for the model (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, density, pH, sulphates, and type).
- Selected the optimal number of components using cross-validation to minimize RMSE. (Training: 0.53, Testing: 0.38).
- Identified the best lambda, regularization parameter, using Generalized Cross Validation.

Insightful Conclusion:

- Model A was efficient in identifying key predictors; however, residual diagnostics showed concerns regarding heteroscedasticity.





Model B - Weighted Least Squares

Goal:

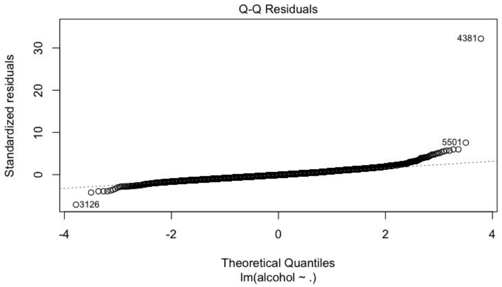
- Find the best model possible, utilizing model diagnostics and remedial measures.

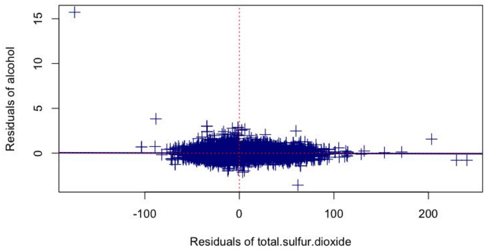
Steps Taken:

- Checked for high leverage points, outliers, and high influential points.
- Reran diagnostics for linearity, normality, and constant variance.
- Compared model fit and predictive performance using permutation tests.
- Applied Weighted Least Squares using weights into a lack-of-fit test to stabilize variance and check overall model fit.

Insightful Conclusion:

- WLS successfully addressed the shortcomings of Model A, improving model reliability while maintaining interpretability.
- Permutation test p-value was 0, suggesting statistically significant predictors.





Main Results

Model A - Subset Selection:

- Achieved an R-Squared of 82%, explaining a significant portion of the variability in alcohol content.
- Principal Component Regression and Ridge Regression were explored but yielded less favorable results.
- Diagnostics revealed issues with heteroscedasticity, leading us to seek further improvements.

Model B - Weighted Least Squares (WLS):

- Addressed heteroscedasticity by applying weights to the model.
- Resulted in a higher R-Squares 99.98%, improving overall model fit.
- Diagnostics confirmed the model effectively resolved linearity and variance issues.

AIC Comparisons:

- Subset Selection Model: AIC = 9615.
- Weighted Least Squares Model: AIC = 117.8.
- The WLS model demonstrated superior performance.

Food For Thought

Final Model:

- The Weighted Least Squares model outperformed the Subset Selection model.
- By addressing heteroscedasticity, WLS provided a better fit to the data, with significantly improved diagnostics.

Challenges and Steps Taken:

- Understanding how to analyze code that we generated, not someone else's.
- Diagnosed issues like heteroscedasticity and non-normality using diagnostic plots and tests.

Conclusion:

- The WLS model effectively resolves issues in the data and serves as the best model for predicting alcohol content.

ANOVA	AIC	Multiple-R^2	MSE	P-Value
Full Model	9540.27	82.2%	.2537	2.2e-16
Model A	9615.816	81.97%	.2568	2.2e-16
Model B	117.8112	99.98%	1.212	2.2e-16

Gratitude

STAT 425: Statistical Modeling I

Matthew Klima & Gabe Price

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN