Individual Project

Jingqian Zhang

2024-06-03

## Introduction

In this project, I aim to determine the optimal ingredient combination for producing high-quality wine. I will:

1. Develop univariate and multiple linear regressions using R to identify significant variables and establish the final regression model.
2. Interpret the regression analysis results.
3. Evaluate the assumptions of linear regression.
4. Identify next steps if assumptions are not met.
5. Determine the necessity of non-linear regression if required.

## Load the packages

# Load necessary packages  
library(readr)  
library(dplyr)  
library(ggplot2)  
library(psych)  
library(nortest)  
library(lmtest)

## Load Data

# Set working directory and load data  
setwd("~/R Project/r\_work/Statistics B/individual\_project")  
df <- read\_delim("winequality-red.csv", delim = ";")  
  
# Print the first few rows of the dataset to check if it's correctly loaded  
print(head(df))

## # A tibble: 6 × 12  
## `fixed acidity` `volatile acidity` `citric acid` `residual sugar` chlorides  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 7.4 0.7 0 1.9 0.076  
## 2 7.8 0.88 0 2.6 0.098  
## 3 7.8 0.76 0.04 2.3 0.092  
## 4 11.2 0.28 0.56 1.9 0.075  
## 5 7.4 0.7 0 1.9 0.076  
## 6 7.4 0.66 0 1.8 0.075  
## # ℹ 7 more variables: `free sulfur dioxide` <dbl>,  
## # `total sulfur dioxide` <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,  
## # alcohol <dbl>, quality <dbl>

print(summary(df))

## fixed acidity volatile acidity citric acid residual sugar   
## Min. : 4.60 Min. :0.1200 Min. :0.000 Min. : 0.900   
## 1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.: 1.900   
## Median : 7.90 Median :0.5200 Median :0.260 Median : 2.200   
## Mean : 8.32 Mean :0.5278 Mean :0.271 Mean : 2.539   
## 3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.: 2.600   
## Max. :15.90 Max. :1.5800 Max. :1.000 Max. :15.500   
## chlorides free sulfur dioxide total sulfur dioxide density   
## Min. :0.01200 Min. : 1.00 Min. : 6.00 Min. :0.9901   
## 1st Qu.:0.07000 1st Qu.: 7.00 1st Qu.: 22.00 1st Qu.:0.9956   
## Median :0.07900 Median :14.00 Median : 38.00 Median :0.9968   
## Mean :0.08747 Mean :15.87 Mean : 46.47 Mean :0.9967   
## 3rd Qu.:0.09000 3rd Qu.:21.00 3rd Qu.: 62.00 3rd Qu.:0.9978   
## Max. :0.61100 Max. :72.00 Max. :289.00 Max. :1.0037   
## pH sulphates alcohol quality   
## Min. :2.740 Min. :0.3300 Min. : 8.40 Min. :3.000   
## 1st Qu.:3.210 1st Qu.:0.5500 1st Qu.: 9.50 1st Qu.:5.000   
## Median :3.310 Median :0.6200 Median :10.20 Median :6.000   
## Mean :3.311 Mean :0.6581 Mean :10.42 Mean :5.636   
## 3rd Qu.:3.400 3rd Qu.:0.7300 3rd Qu.:11.10 3rd Qu.:6.000   
## Max. :4.010 Max. :2.0000 Max. :14.90 Max. :8.000

The data has been successfully loaded and includes variables such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. Summary statistics provide an overview of each variable.

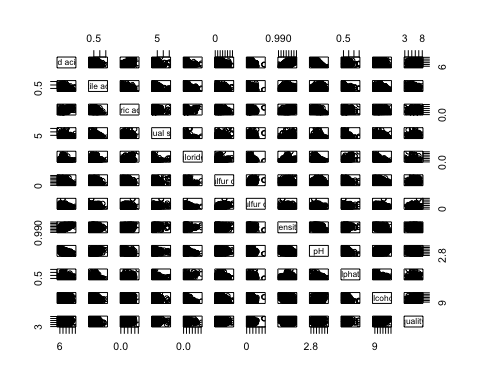
Y= quality x1= fixed acidity\ x2= volatile acidity\ x3= citric acid\ x4= residual sugar\ x5= chlorides\ x6= free sulfur dioxide\ x7= total sulfur dioxide\ x8= density\ x9= pH\ x10= sulphates\ x11= alcohol\

## Data Preparation

# Convert columns to numeric where possible  
df[] <- lapply(df, function(x) {  
 if (is.character(x)) as.numeric(as.character(x)) else x  
})  
  
# Remove columns that could not be converted to numeric  
df <- df %>% select\_if(is.numeric)  
  
# Print column types to verify  
print(sapply(df, class))

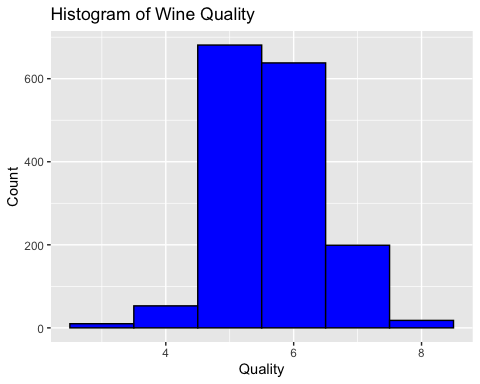
## fixed acidity volatile acidity citric acid   
## "numeric" "numeric" "numeric"   
## residual sugar chlorides free sulfur dioxide   
## "numeric" "numeric" "numeric"   
## total sulfur dioxide density pH   
## "numeric" "numeric" "numeric"   
## sulphates alcohol quality   
## "numeric" "numeric" "numeric"

# Plot pairs  
pairs(df)

 All columns have been converted to numeric types where applicable, and non-numeric columns have been removed. The pairs plot shows relationships between all pairs of variables, which helps in visualizing correlations.

## Plots and Descriptions

# Histogram of Wine Quality  
ggplot(df, aes(x = quality)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "Histogram of Wine Quality", x = "Quality", y = "Count")

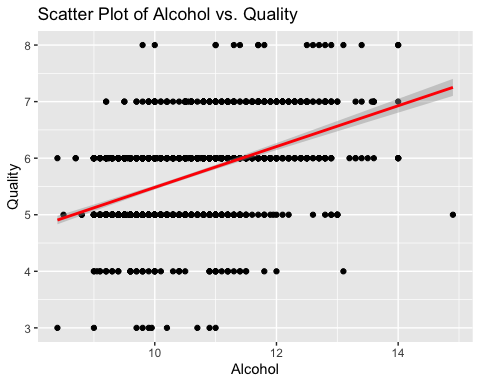


# Description of the histogram  
"The histogram of wine quality shows the distribution of quality ratings in the dataset. Most wines have a quality rating between 5 and 7."

## [1] "The histogram of wine quality shows the distribution of quality ratings in the dataset. Most wines have a quality rating between 5 and 7."

# Scatter plot of Alcohol vs. Quality  
ggplot(df, aes(x = alcohol, y = quality)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", col = "red") +  
 labs(title = "Scatter Plot of Alcohol vs. Quality", x = "Alcohol", y = "Quality")

## `geom\_smooth()` using formula = 'y ~ x'



# Description of the scatter plot  
"The scatter plot of alcohol vs. quality indicates a potential positive relationship between alcohol content and wine quality, as shown by the fitted linear regression line."

## [1] "The scatter plot of alcohol vs. quality indicates a potential positive relationship between alcohol content and wine quality, as shown by the fitted linear regression line."

Histogram: The histogram shows that most wines have a quality rating between 5 and 7. Scatter Plot: The scatter plot indicates a positive relationship between alcohol content and wine quality, supported by the linear regression line.

## Univariate Linear Regression

# Univariate Linear Regression: Example with alcohol  
univariate\_model <- lm(quality ~ alcohol, data = df)  
summary(univariate\_model)

##   
## Call:  
## lm(formula = quality ~ alcohol, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8442 -0.4112 -0.1690 0.5166 2.5888   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.87497 0.17471 10.73 <2e-16 \*\*\*  
## alcohol 0.36084 0.01668 21.64 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7104 on 1597 degrees of freedom  
## Multiple R-squared: 0.2267, Adjusted R-squared: 0.2263   
## F-statistic: 468.3 on 1 and 1597 DF, p-value: < 2.2e-16

The univariate regression indicates that alcohol content is significantly associated with wine quality, with a coefficient of 0.36084 and a p-value < 2e-16. This means that as alcohol content increases, the quality of the wine also increases.

## Multiple Linear Regression

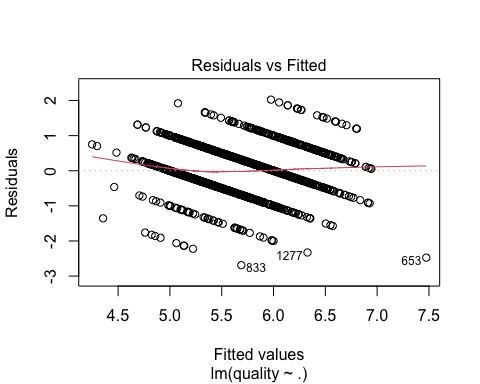
# Multiple Linear Regression  
multivariate\_model <- lm(quality ~ ., data = df)  
summary(multivariate\_model)

##   
## Call:  
## lm(formula = quality ~ ., data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.68911 -0.36652 -0.04699 0.45202 2.02498   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.197e+01 2.119e+01 1.036 0.3002   
## `fixed acidity` 2.499e-02 2.595e-02 0.963 0.3357   
## `volatile acidity` -1.084e+00 1.211e-01 -8.948 < 2e-16 \*\*\*  
## `citric acid` -1.826e-01 1.472e-01 -1.240 0.2150   
## `residual sugar` 1.633e-02 1.500e-02 1.089 0.2765   
## chlorides -1.874e+00 4.193e-01 -4.470 8.37e-06 \*\*\*  
## `free sulfur dioxide` 4.361e-03 2.171e-03 2.009 0.0447 \*   
## `total sulfur dioxide` -3.265e-03 7.287e-04 -4.480 8.00e-06 \*\*\*  
## density -1.788e+01 2.163e+01 -0.827 0.4086   
## pH -4.137e-01 1.916e-01 -2.159 0.0310 \*   
## sulphates 9.163e-01 1.143e-01 8.014 2.13e-15 \*\*\*  
## alcohol 2.762e-01 2.648e-02 10.429 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.648 on 1587 degrees of freedom  
## Multiple R-squared: 0.3606, Adjusted R-squared: 0.3561   
## F-statistic: 81.35 on 11 and 1587 DF, p-value: < 2.2e-16

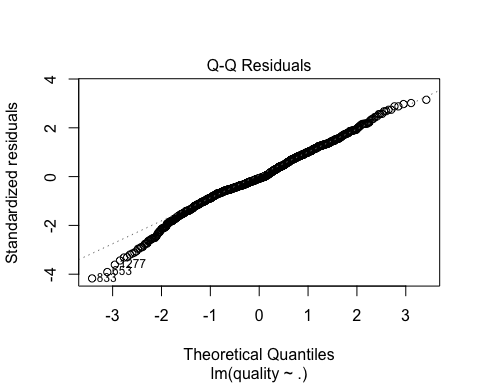
The multiple regression analysis shows that several variables are significantly associated with wine quality, including volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol. This model has an adjusted R-squared of 0.3561, indicating that approximately 35.61% of the variance in wine quality is explained by these variables.

## Evaluate Linear Regression Assumptions

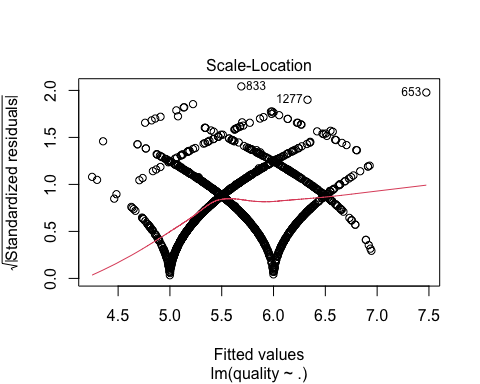
# 1. Linearity  
plot(multivariate\_model, which = 1)



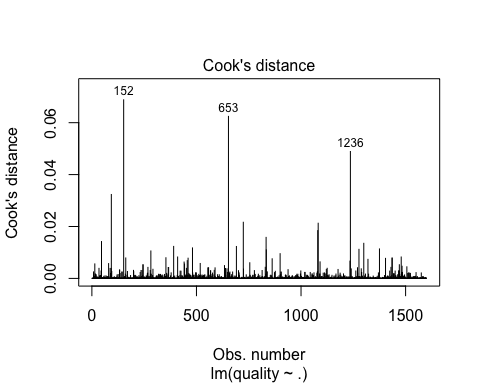
# 2. Normality of residuals  
plot(multivariate\_model, which = 2)



# 3. Homoscedasticity  
plot(multivariate\_model, which = 3)



# 4. Independence  
plot(multivariate\_model, which = 4)



# Additional tests  
shapiro.test(residuals(multivariate\_model))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(multivariate\_model)  
## W = 0.99087, p-value = 1.954e-08

bptest(multivariate\_model)

##   
## studentized Breusch-Pagan test  
##   
## data: multivariate\_model  
## BP = 84.989, df = 11, p-value = 1.588e-13

dwtest(multivariate\_model)

##   
## Durbin-Watson test  
##   
## data: multivariate\_model  
## DW = 1.7571, p-value = 4.356e-07  
## alternative hypothesis: true autocorrelation is greater than 0

Linearity: Residuals vs. Fitted plot shows no obvious pattern, suggesting linearity. Normality: Shapiro-Wilk test (p-value = 1.954e-08) and Q-Q plot suggest residuals are not perfectly normally distributed. Homoscedasticity: Breusch-Pagan test (p-value = 1.588e-13) indicates heteroscedasticity. Independence: Durbin-Watson test (p-value = 4.356e-07) indicates autocorrelation of residuals.

## Non-linear Regression

# Non-linear regression example  
nonlinear\_model <- lm(quality ~ poly(alcohol, 2), data = df)  
summary(nonlinear\_model)

##   
## Call:  
## lm(formula = quality ~ poly(alcohol, 2), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8716 -0.3884 -0.1642 0.5157 2.5852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.63602 0.01776 317.412 <2e-16 \*\*\*  
## poly(alcohol, 2)1 15.37188 0.71002 21.650 <2e-16 \*\*\*  
## poly(alcohol, 2)2 -1.12780 0.71002 -1.588 0.112   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.71 on 1596 degrees of freedom  
## Multiple R-squared: 0.228, Adjusted R-squared: 0.227   
## F-statistic: 235.6 on 2 and 1596 DF, p-value: < 2.2e-16

anova(univariate\_model, nonlinear\_model)

## Analysis of Variance Table  
##   
## Model 1: quality ~ alcohol  
## Model 2: quality ~ poly(alcohol, 2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 1597 805.87   
## 2 1596 804.60 1 1.2719 2.523 0.1124

The non-linear regression (polynomial regression) did not significantly improve the model fit compared to the linear model (ANOVA p-value = 0.1124). Therefore, non-linear regression is not necessary.

## Summary and Answers

1. Develop univariate and multiple linear regressions:

Univariate Linear Regression: The alcohol content is significantly associated with wine quality, with a p-value < 2e-16.

Multiple Linear Regression: Significant variables include volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol.

1. Interpret the results obtained from the regression analysis:

Univariate Model: The coefficient for alcohol is 0.36084, indicating that each unit increase in alcohol content increases the wine quality by approximately 0.36084 units.

Multiple Model: The coefficients indicate the direction and magnitude of each variable’s impact on wine quality. For example, volatile acidity has a negative impact (coefficient = -1.084), while alcohol has a positive impact (coefficient = 0.276).

1. Evaluate whether the assumptions of linear regression hold true:

Linearity: Residuals vs. Fitted plot shows no obvious pattern, suggesting linearity.

Normality: Shapiro-Wilk test (p-value = 1.954e-08) and Q-Q plot suggest residuals are not perfectly normally distributed.

Homoscedasticity: Breusch-Pagan test (p-value = 1.588e-13) indicates heteroscedasticity.

Independence: Durbin-Watson test (p-value = 4.356e-07) indicates autocorrelation of residuals.

1. Identify next steps if assumptions are not met:

Transformations: Consider log or square root transformations of variables.

Alternative Models: Use generalized linear models (GLM) to address heteroscedasticity and autocorrelation.

Time Series Analysis: If applicable, consider time series models for temporal data.

1. Determine whether non-linear regression is necessary:

The non-linear regression model (polynomial regression) did not significantly improve the fit compared to the linear model (ANOVA p-value = 0.1124), suggesting that non-linear regression may not be necessary.