**1. Đề tài**

Đánh giá các yếu tố ảnh hưởng chất lượng rượu vang đỏ. Tập tin **winequality-red.csv** chứa thông tin về các yếu tố ảnh hưởng chất lượng rượu vang đỏ. Các thuộc tính dữ liệu dựa trên các phép thử hóa lý bao gồm pH, axit citric, SO2 tự do, độ cồn (phần trăm rượu trong dung dịch) và một số yếu tố khác. Dữ liệu gốc được cung cấp tại:

<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

***Tổng quan***

Dựa vào lý thuyết và đồ thị tương quan các yếu tố ảnh hưởng chất lượng rượu, các biến có ảnh hưởng cao tới chất lượng rượu là volatile acidity (Nồng độ axit dễ bay hơi (/ dm3)); density: tỷ trọng của rượu (g/cm3); alcohol: Độ cồn; chlorides: Nồng độ muối chloride(g/dm3). Có thể tổng kết *density* ảnh hưởng đến chất lượng rượu lớn nhất và *Nồng độ muối chloride* ảnh hưởng đến chất lượng rượu thấp nhất. (sự thay đổi tính trên trên một đơn vị đo lường)

**2. Giải quyết bài toán bằng phần mềm RStudio**

***2.1. Đọc dữ liệu:***

Thực hiện đọc dữ liệu CSDL đã Downloads từ web: **winequality-red.csv**

#Cài đặt các gói

install.packages("dplyr")

install.packages("ggplot2")

install.packages("corrplot")

install.packages("psych")

# Mở tệp dữ liệu dưới dạng bảng

rm(list=ls())

options(warn=-1)

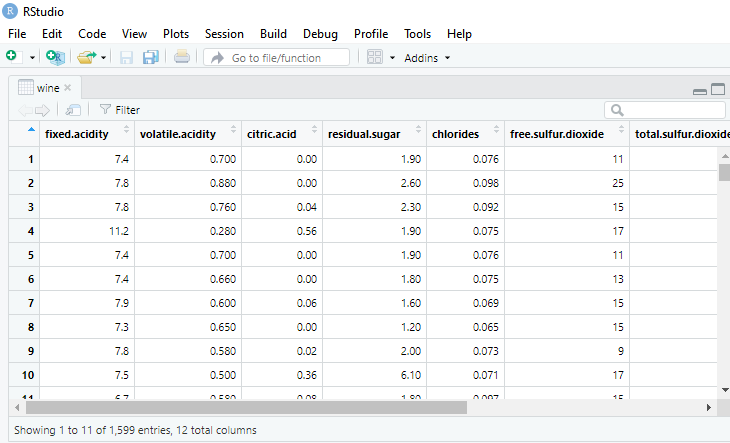
library(dplyr)

library(ggplot2)

library(corrplot)

library(readr)

Read.csv2(“D:/winequality-red.csv”)

Ta được bảng dữ liệu sau:

**2.2. Làm sạch dữ liệu**

Bài toán này gồm các biến như sau: fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol. Với dữ liệu đầu ra là quality (điểm từ 0 đến 10)

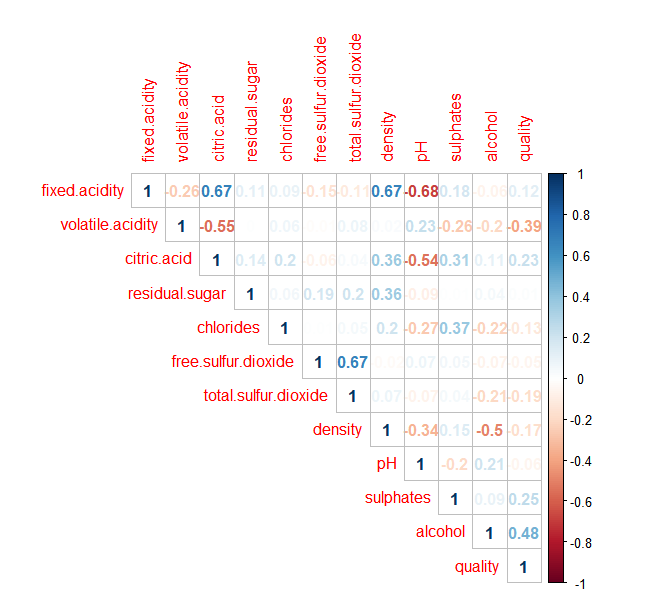
Ta tiến hành vẽ biểu đồ thể hiện hệ số tương quan giữa từng biến trên đối với biến quality:

attach(wine)

winecor= cor(wine)

corrplot(winecor, type = "upper", method= "number")

Kết quả cho ra biểu đồ hệ số tương quan giữa các biến như sau:



Qua biểu đồ trên, ta thấy rằng các biến **volatitle acidity, chlorides, density và alcohol** có độ tương quan khá cao với quality lần lượt với hệ số tương quan r lần lượt là: -0.39, 0,13, -0.17, 0.48

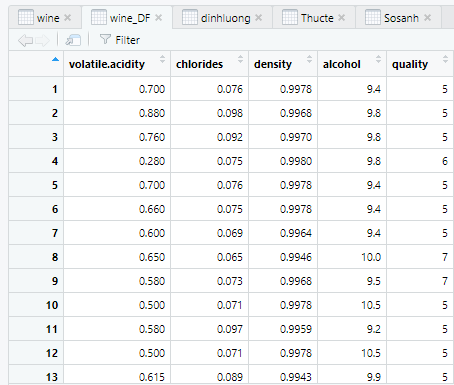
Sau khi chọn biến xong, ta tiến hành tạo một tập dữ liệu con có tên là wine\_DF, với các cột là các dữ liệu đã chọn và tiến hành kiểm tra dữ liệu khuyết:

wine\_DF=wine[c(2,5,8,11,12)]

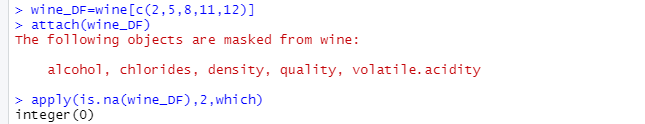
attach(wine\_DF)

apply(is.na(wine\_DF),2,which)

Ta được bảng dữ liệu con wine\_DF sau:



Vì không có dữ liệu khác nên kết quả trả lại Integer(0)



**Giả sử dữ liệu bị khuyết ta sửa bằng cách sau:** ta cần loại bỏ các dòng chứa những dữ liệu khuyết này bằng cách dùng hàm **na.omit()** và sau đó gán lại dữ liệu mới không bị khuyết cho **new\_DF**

* Dùng cú pháp **new\_DF=na.omit(new\_DF)** để tiến hành xóa các dòng có dữ liệu khuyết
* Kế tiếp dùng hàm **anyNA(new\_DF)** để kiểm tra xem trong **new\_DF** còn chứa dữ liệu khuyết hay không. Nết kết quả trả về **FALSE** chứng tỏ không còn dữ liệu khuyết

**2.3\*. Làm rõ dữ liệu (Data visualization):**

**a. Tính các giá trị thống kê mô tả bao gồm:**

*Đối với các biến liên tục, hãy tính giá trị thống kê mô tả bao gồm: trung bình, trung vị, độ lệch chuẩn, giá trị lớn nhất và giá trị nhỏ nhất. Xuất kết quả dưới dạng bảng. (Hàm gợi ý:* ***mean()****,* ***median()****,* ***sd()****,* ***min()****,* ***max()****,* ***apply()****,* ***as.data.frame()****,* ***rownames()****)*

Các biến liên tục trong bộ dữ liệu wine\_DF là các biến sau:biến **volatitle acidity, chlorides, density và alcohol.**

Mean= apply(wine\_DF, 2 , mean)

Median = apply(wine\_DF, 2 , median)

Sd = apply(wine\_DF, 2 , sd)

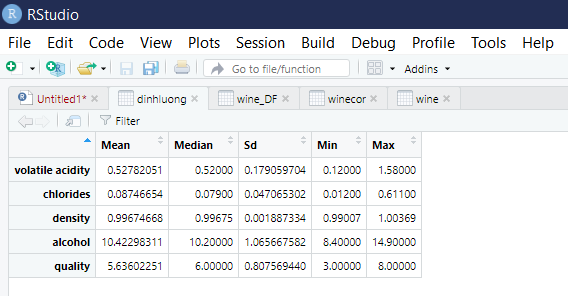
Min= apply(wine\_DF, 2 , min)

Max= apply(wine\_DF, 2 , max)

dinhluong = data.frame(Mean, Median, Sd, Min, Max )

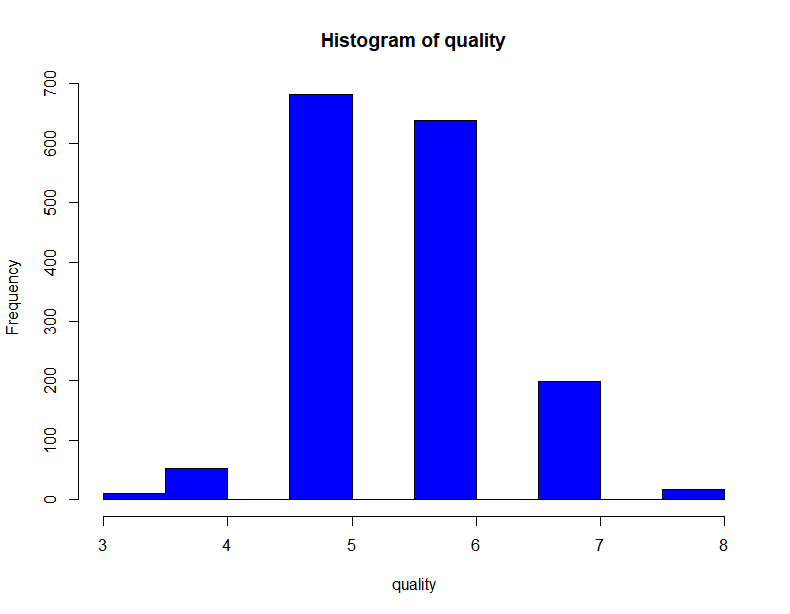
View(dinhluong)

Ta được bảng định lượng như sau:

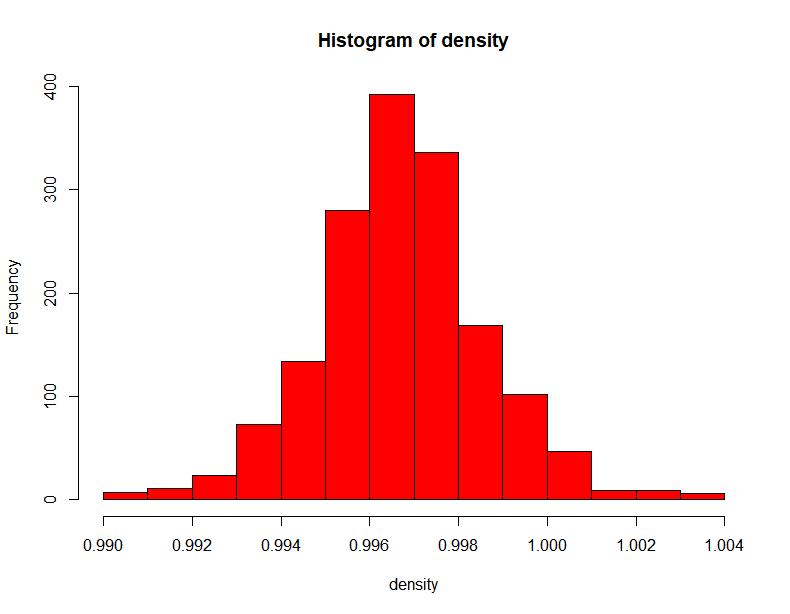


**b. Vẽ** **biểu đồ phân phối của các biến:**

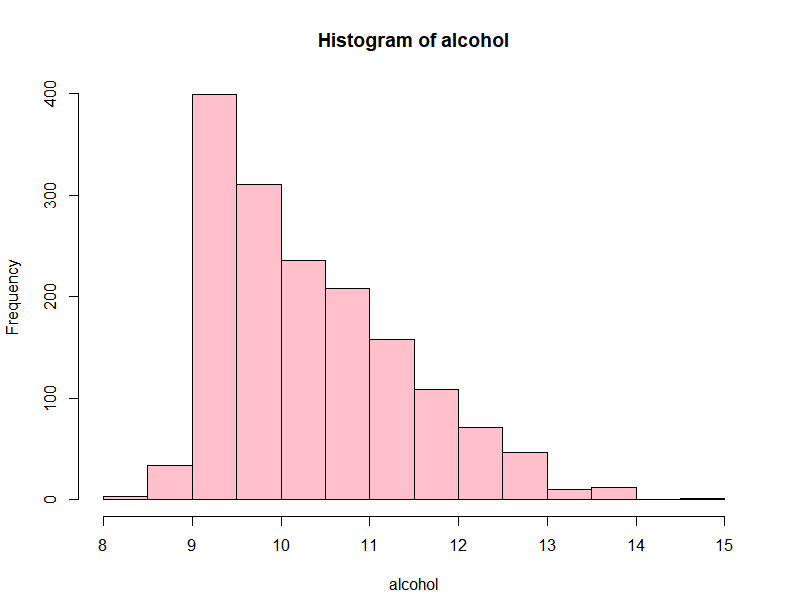
hist(quality , col = "blue")



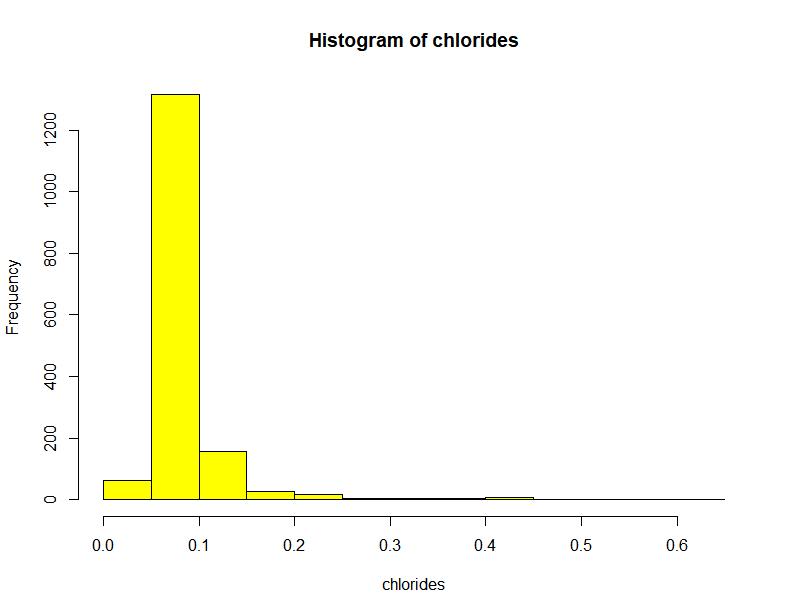
hist(density, col = “red”)



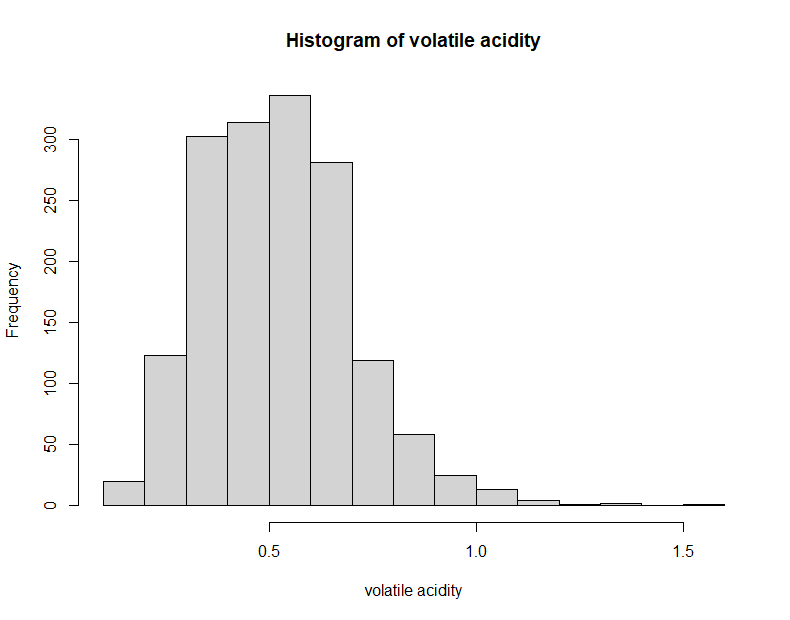
hist(alcohol, col = "pink" )



hist(chlorides, col = "yellow")



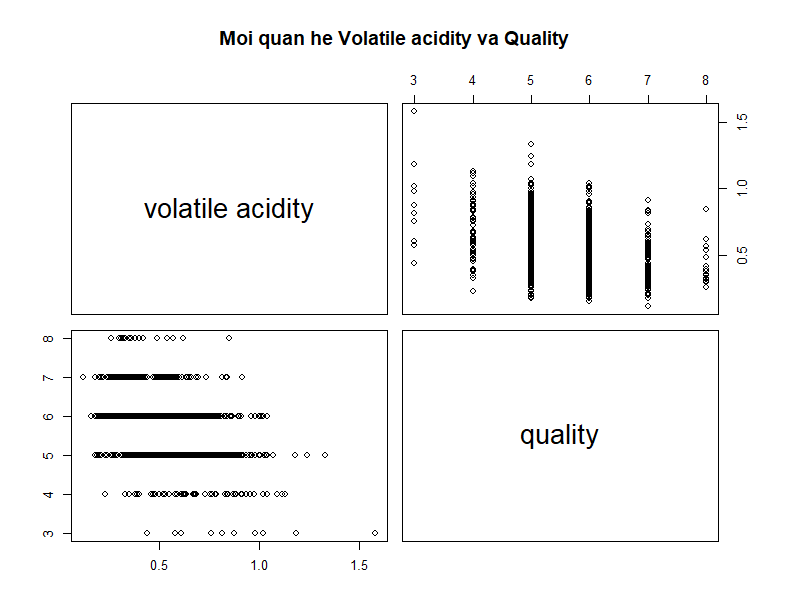
hist(volatile.acidity)



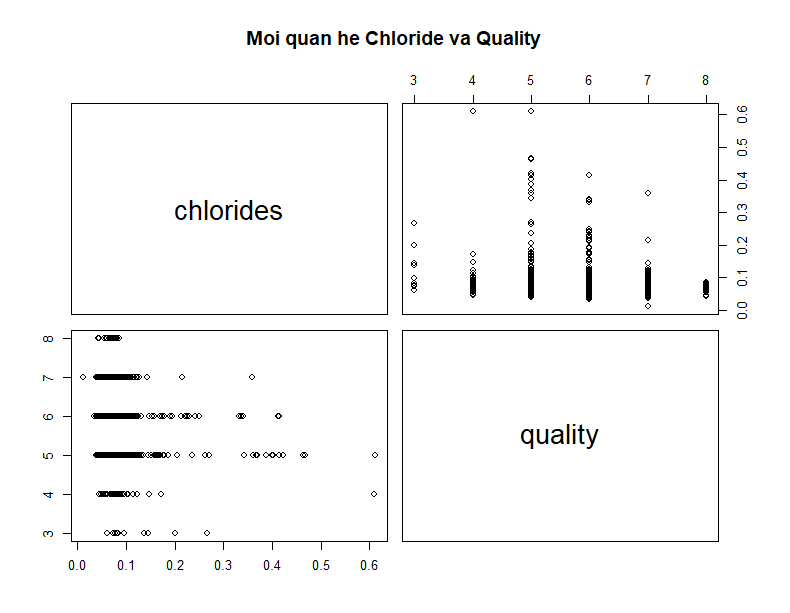
Do các biến đang khảo sát đều là biến liên tục, Mô hình có dạng hình chuông Gauss nên ta dự đoán mô hình theo phân phối chuẩn.

Ta sử dụng pairs() để vẽ phân phối của quality theo các biến khảo sát. Đánh giá quan hệ của quality với các biến còn lại.

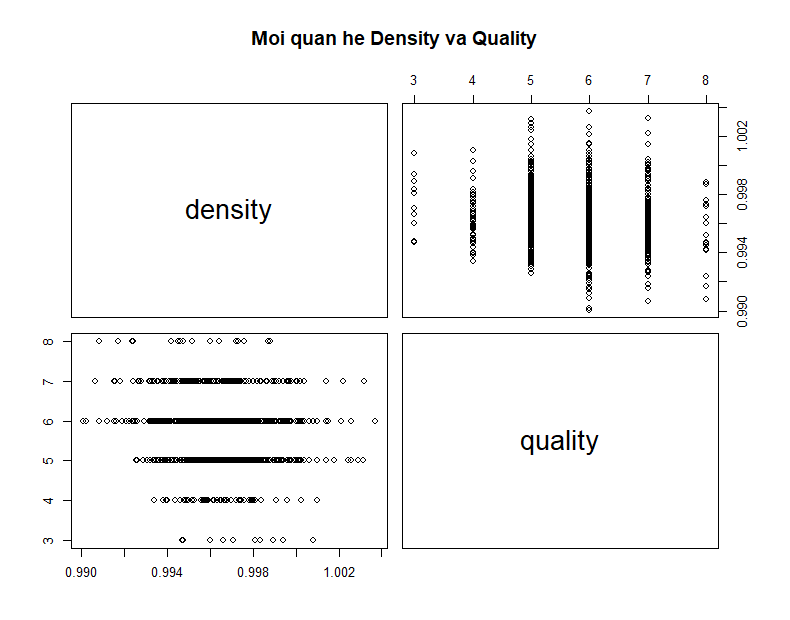
pairs(`volatile acidity`~quality, main= "Moi quan he Volatile acidity va Quality")



pairs(chlorides~quality, main= "Moi quan he Chloride va Quality")



pairs(density~quality, main= "Moi quan he Density va Quality")



pairs(alcohol~quality, main= "Moi quan he Alcohol va Quality")



**2.4.\* Xây dựng các mô hình hồi quy tuyến tính (Fitting linear regression models):**

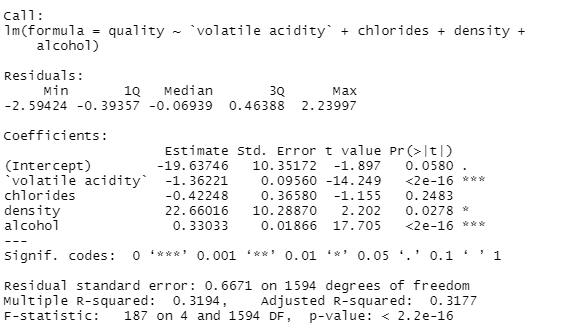
a) Xét mô hình hồi quy tuyến tính bao gồm biến quality là một biến phụ thuộc, và tất cả các biến còn lại đều là biến độc lập. Dùng lệnh lm() để thực thi mô hình hồi quy tuyến tính bội:

Input:

m1=lm(quality~`volatile acidity`+chlorides+density+alcohol)

summary(m1)

Output:



b) Dựa vào kết quả của mô hình hồi quy tuyến tính, loại những biến với các mức tin cậy 5% khỏi mô hình tương ứng.

**Giả thiết:** H0: Các hệ số hồi quy đều không có ý nghĩa thống kê.

H1: Các hệ số hồi quy đều có ý nghĩa thống kê.

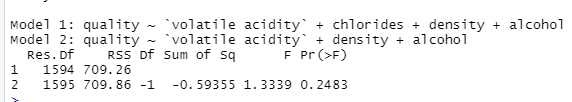
Dựa vào kết quả mô hình tuyến tính, vì biến "chlorides" có Pr(>|t|)= 0.2483 > 0.05 nên các hệ số hồi quy tương ứng với biến này không có ý nghĩa thống kê. Do đó ta loại biến "chlorides" ra khỏi mô hình.

c) Dùng Anova để đề xuất mô hình hợp lý

Input:

Giải thích: mô hình m2 là mô hình đã loại bỏ biến “chlorides”

Output:



Đặt H0: hai mô hình tuyến tính giống nhau

H1: hai mô hình tuyến tính khác nhau

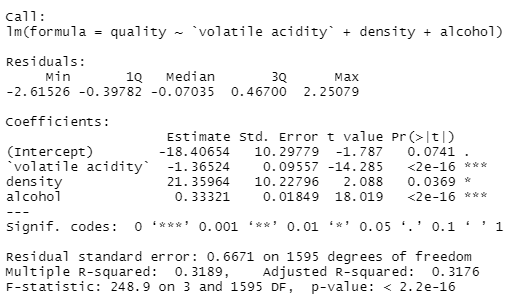
Ta có Pr >(F) là P -value của kiểm định này. Vì P -value = 0,2843 > 0.05

Nên chưa bác bỏ H0, nghĩa là hai mô hình giống nhau do đó có thể chọn mô hình m2 để đánh giá vì nó có sai lệch không lớn với mô hình ban đầu.

1. Từ mô hình hồi quy hợp lý nhất ở câu (c) hãy suy luận tác động của các biến đến chất lượng rượu.

m2=lm(quality~`volatile acidity`+density+alcohol)

summary(m2)



\* Từ mô hình 2, ta xây dựng được phương trình quy hồi tuyến tính sau:

Y (quality) = -18.40654 - 1.36524 (volatile acidity) + 21.35964 (density) + 0.33321 (alcohol)

Nghĩa là: Khi axit bay hơi tăng thì chất lượng giảm 1.36524, mật độ tăng thì chất lượng tăng 21.35964, cồn tăng thì chất lượng tăng 0.33321 (sự thay đổi tăng hay giảm trên một đơn vị đo lường)

\* Từ mô hình 1, ta xây dựng được phương trình quy hồi tuyến tính sau:

Y ( quality ) = -19.63746 - 1.36221 ( volatile.acidity ) - 0.42248 (chlorides) + 22.66016 (density) + 0.33033 ( alcohol )

Nghĩa là: Khi axit bay hơi tăng thì chất lượng giảm 1.36221, clorua tăng thì chất lượng giảm 0.42248, mật độ tăng thì chất lượng tăng 22.66016, cồn tăng thì chất lượng tăng 0.33033 ( sự thay đổi tăng hay giảm trên một đơn vị đo lường)

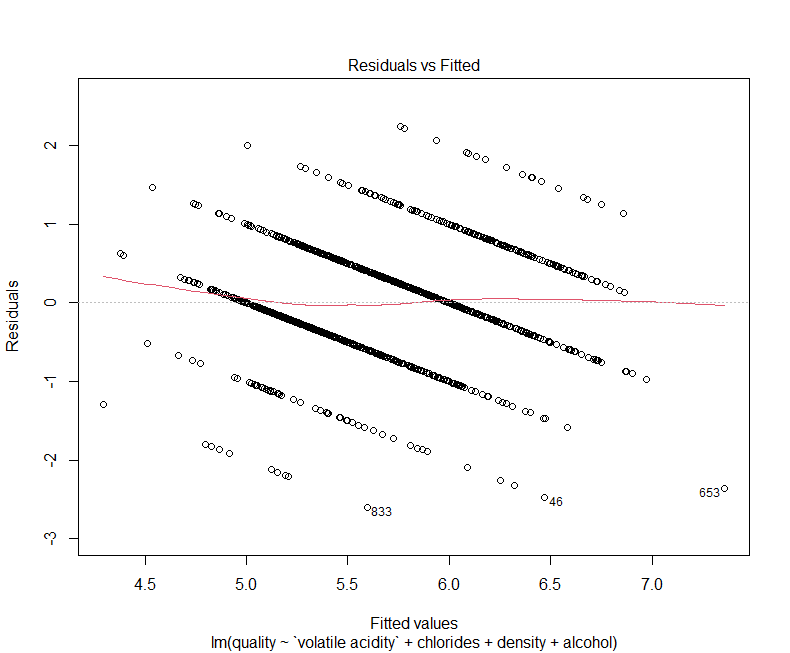
1. Từ mô hình hợp lý nhất ở câu c) hãy dùng lệnh **plot()** để vẽ đồ thị biểu thị sai số hồi quy và dự báo. Nêu ý nghĩa và nhận xét.

Vẽ biểu đồ biểu thị quan hệ giữa giá trị dự báo và sai số hồi quy của mô hình m1:

m1=lm(quality~`volatile acidity`+chlorides+density+alcohol)

plot(fitted(m1),resid(m1))

plot(m1,which=1)

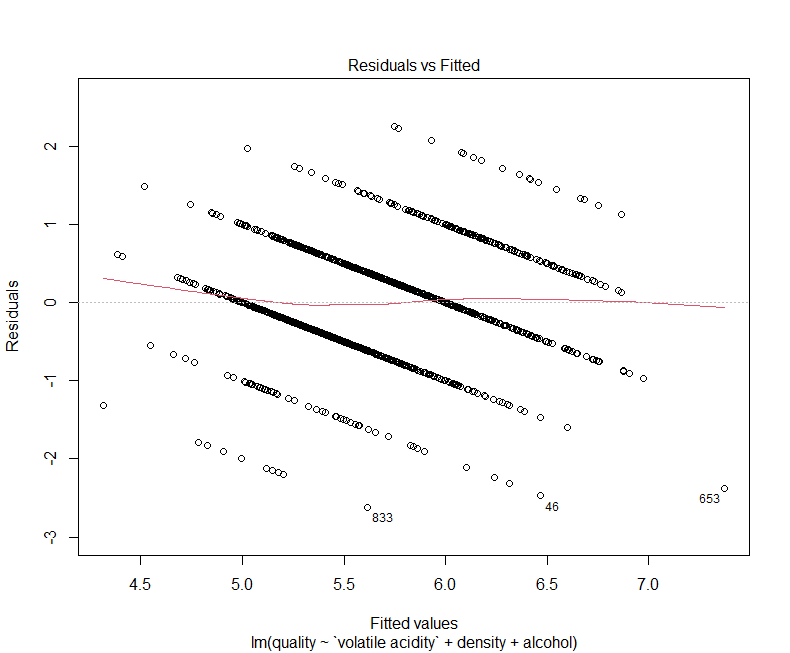


Vẽ biểu đồ biểu thị quan hệ giữa giá trị dự báo và sai số hồi quy của mô hình m2:

m2=lm(quality~`volatile acidity`+density+alcohol)

plot(fitted(m2),resid(m2))

plot(m2,which=1)



**Nhận xét:** đường hồi quy(đường màu đỏ) hai mô hình gần trùng với đường Residuals = 0 (đường nét đứt)

**Ý nghĩa:** Sự sai số hai mô hình m1, m2 là không đáng kể, khẳng định ở câu c là phù hợp. Về mặt đơn giản ta sẽ chọn mô hình 2 để khảo sát chất lượng rượu.

Code tổng của bài:

#Cài đặt các gói

install.packages("dplyr")

install.packages("ggplot2")

install.packages("corrplot")

install.packages("psych")

# Mở tệp dữ liệu dưới dạng bảng

rm(list=ls())

options(warn=-1)

library(dplyr)

library(ggplot2)

library(corrplot)

library(readr)

Read.csv2(“D:/winequality-red.csv”)

View(wine)

# Lọc những dữ lệu có mức ảnh hưởng lớn từ tệp gốc

wine\_DF=wine[c(2,5,8,11,12)]

attach(wine\_DF)

apply(is.na(wine\_DF),2,which)

# Làm rõ dữ liệu (Data visualization)

# Tính toán dữ liệu thống kê

Mean= apply(wine\_DF, 2 , mean)

Median = apply(wine\_DF, 2 , median)

Sd = apply(wine\_DF, 2 , sd)

Min= apply(wine\_DF, 2 , min)

Max= apply(wine\_DF, 2 , max)

dinhluong = data.frame(Mean, Median, Sd, Min, Max )

# biểu đồ phân phối của các biến

hist(quality , col = "blue")

hist(density, col = “red”)

hist(alcohol, col = "pink" )

hist(chlorides, col = "yellow")

hist(volatile.acidity)

# Mối quan hệ giữa các bán đến chất lượng

pairs(`volatile acidity`~quality, main= "Moi quan he Volatile acidity va Quality")

pairs(chlorides~quality, main= "Moi quan he Chloride va Quality")

pairs(density~quality, main= "Moi quan he Density va Quality")

pairs(alcohol~quality, main= "Moi quan he Alcohol va Quality")

#Xây dựng mô hình hồi quy tuyến tính

#Kiểm tra mức ảnh hưởng

m1=lm(quality~`volatile acidity`+chlorides+density+alcohol)

summary(m1)

#mô hình m2

m2=lm(quality~`volatile acidity`+density+alcohol)

#Dùng Anova để kiểm tra mô hình phù hợp

anova(m1,m2)

# Vẽ biểu đồ biểu thị quan hệ giữa giá trị dự báo và sai số hồi quy của mô hình m1

plot(fitted(m1),resid(m1))

plot(m1,which=1)