1. **Chọn hai "đề tài"tự do để tìm hiểu "nghiên cứu", vận dụng những kiến thức học được từ môn này.**

1. Dữ liệu 1: mô hình hồi quy đa biến.

- Tên "đề tài"

- Nguồn gốc của dữ liệu, giới thiệu các biến, ...

- Mô hình chọn được

- Phân tích kết quả

- Đưa ra những phương pháp/phân tích khác có thể giúp cho kết quả tốt hơn nếu có thể.

- Kết luận.

Sử dụng bộ dữ liệu về chất lượng rượu vang (nguồn: <https://archive.ics.uci.edu/ml/datasets/wine+quality> ) gồm các biến sau:

1 - fixed acidity: lượng acid cố định

2 - volatile acidity: lượng acid dễ bay hơi

3 - citric acid: lượng acid citric

4 - residual sugar: lượng đường

5 - chlorides: lượng clorua

6 - free sulfur dioxide: sulfur dioxide tự do

7 - total sulfur dioxide: sulfur dioxide tổng

8 - density: tỷ trọng

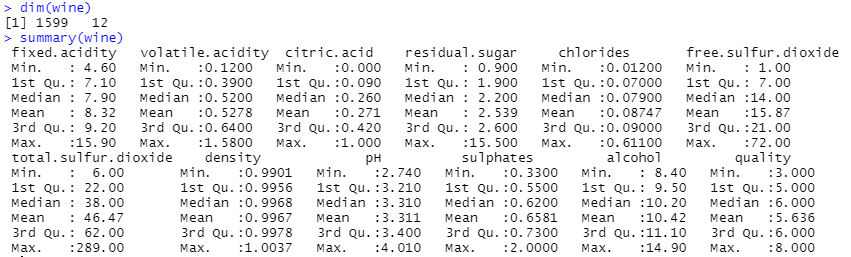
9 - pH: pH

10 - sulphates: lượng sulfate

11 - alcohol: độ cồn

12 - quality (score between 0 and 10): điểm từ 1 - 10

Tổng quan về bộ dữ liệu:

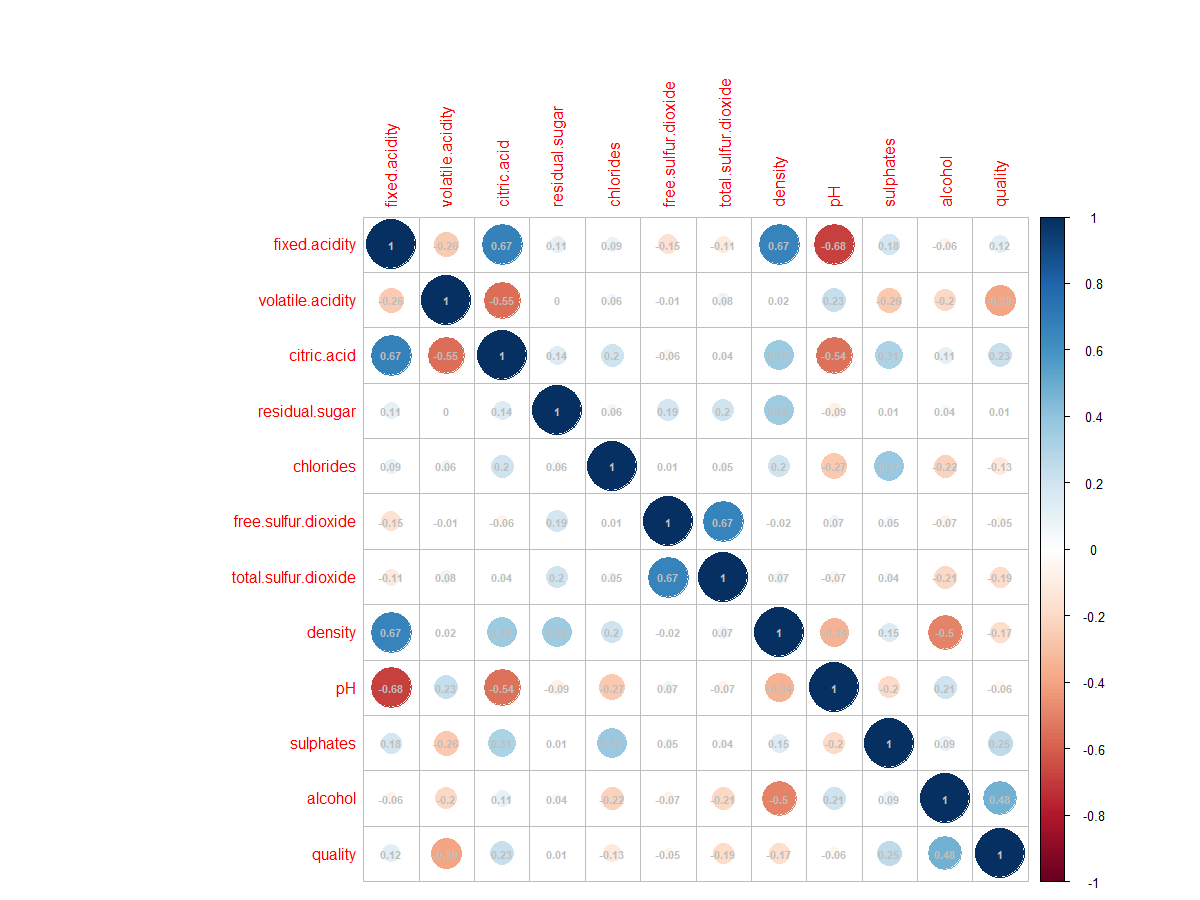


Dữ liệu gồm 1599 quan trắc và 12 biến.

Xét hệ số tương quan giữa các biến trong bộ dữ liệu, ta thấy có các biến về acid có liên quan đến nhau như:

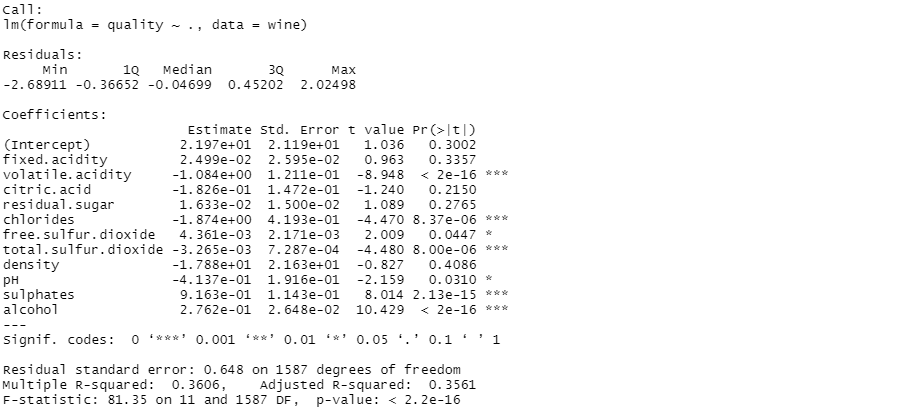
* fixed.acidity và citric.acid: lượng acid cố định và lượng acid citric
* fixed.acidity và density: lượng acid cố định và tỷ trọng
* fixed.acidity và pH: lượng acid cố định và pH
* Và một số biến khác

|  |
| --- |
| corrplot::corrplot(cor(wine), addCoef.col = "grey", number.cex = 0.7) |



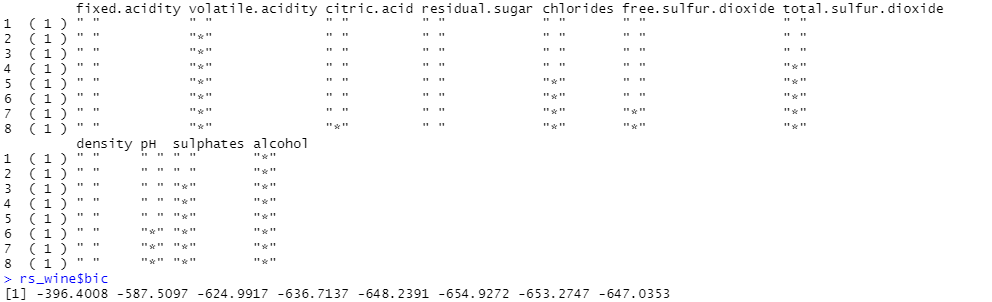
Ta sẽ đi tìm một mô hình hồi quy tuyến tính để giải thích biến chất lượng rượu quality dựa trên các biến định lượng còn lại.

Ta xét mô hình tuyến tính của quality với đầy đủ các biến của bộ dữ liệu.



Mô hình đầy đủ có nhiều biến có ảnh hưởng đến chất lượng rượu vang, các biến fixed.acidity, citric.acid, residual.sugar, density không có ý nghĩa thống kê. Để loại bỏ các biến không có ảnh hưởng đến mô hình, trước tiên ta thử tìm các tập mô hình tốt nhất:

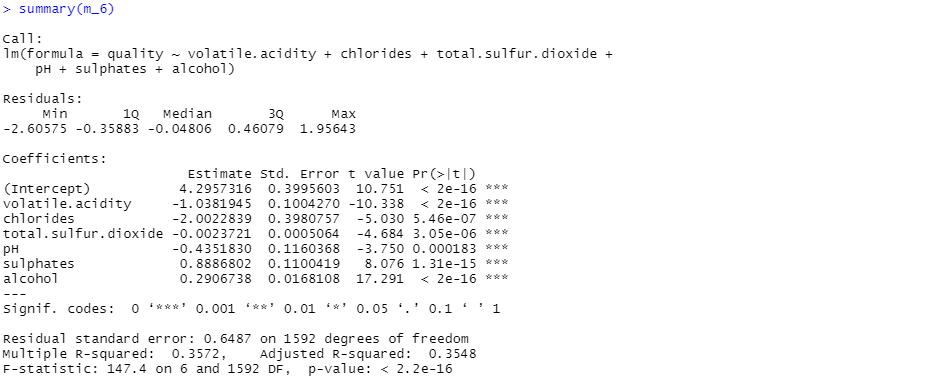
|  |
| --- |
| X\_wine <- cbind(fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol)  b\_wine <- regsubsets(as.matrix(X\_wine), quality)  rs\_wine <- summary(b\_wine) |



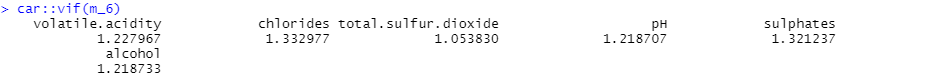
Dựa vào tiêu chuẩn BIC, ta chọn mô hình của quality gồm 6 biến: volatile.acidity, chlorides, total.sulfur.dioxide, pH, sulphates, alcohol.

Ta fit mô hình hồi quy tuyến tính giải thích cho quality theo 6 biến.

|  |
| --- |
| m\_6 <- lm(quality ~ volatile.acidity + chlorides + total.sulfur.dioxide + pH + sulphates + alcohol) #bic |



Kiểm tra hệ số phóng đại phương sai của mô hình được chọn.

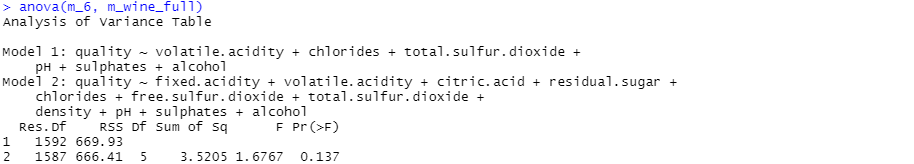


Các biến có hệ số phóng đại phương sai VIF < 5, nghĩa là các biến trong mô hình được chọn không có hiện tượng đa cộng tuyến.

Ta thực hiện kiểm định Fisher từng phần cho mô hình được chọn so với mô hình đầy đủ ban đầu.

Giả thuyết H0: Mô hình được chọn được giải thích bởi 6 biến

Giả thuyết đối H1: Mô hình đầy đủ các biến ban đầu



Dựa vào bảng phân tích phương sai của hai mô hình trên, giá trị p-value là 0.137 > ⍺ = 0.05, với mức ý nghĩa 5%, ta không đủ cơ sở bác bỏ giả thuyết H0, nên mô hình của quality được giải thích bởi 6 biến có ý nghĩa thống kê.

Chất lượng rượu vang được giải thích bởi các yếu tố như lượng acid bay hơi, nồng độ clorua, sulfur dioxide tổng, pH, lượng sulfate, và độ cồn theo công thức sau:

quality = 4.296 - 1.038✕volatile.acidity - 2.002✕chlorides - 0.002✕total.sulfur.dioxide- 0.435✕pH + 0.889✕sulphates + 0.291✕alcohol