**BỘ GIÁO DỤC VÀ ĐÀO TẠO**



**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**

****

BÁO CÁO BÀI TẬP LỚN 2

**XÁC XUẤT THỐNG KÊ**

**Giảng viên hướng dẫn : Hoàng Văn Hà**

# Lớp : L13

Sinh viên thực hiện

Võ Mạnh Hào 1913238

Nguyễn Huy Phúc 1914697

Trần Thế Quang 1914806

Võ Đình Thanh 1912041

Lý Vĩnh Khang 1852443

Hồ Trung Nghĩa 1911681

# I.Phần chung:

Bài tập 1:

Hồi quy tuyến tính bội

Mô tả dữ liệu

Tập tin "gia\_nha.csv" chứa thông tin về giá bán ra thị trường (đơn vị đô la) của 21613 ngôi nhà ở quân King nước Mỹ trong khoảng thời gian từ tháng 5/2014 đến 5/2015. Bên cạnh giá nhà, dữ liệu còn bao gồm các thuộc tính mô tả chất lượng ngôi nhà. Dữ liệu gốc được cung cấp tại: https://www.kaggle.com/harlfoxem/housesalesprediction.

Các biến chính trong bộ dữ liệu:

• price: Giá nhà được bán ra.

• sqft\_living15: Diện tích trung bình của 15 ngôi nhà gần nhất trong khu dân cư.

• floors: Số tầng của ngôi nhà được phân loại từ 1-3.5.

• condition: Điều kiện kiến trúc của ngôi nhà từ 1 − 5, 1: rất tệ và 5: rất tốt.

• sqft\_above: Diện tích ngôi nhà.

• sqft\_living: Diện tích khuôn viên nhà.

**Yêu cầu:** xây dựng mô hình hồi quy bội (multiple regression) để phân tích tác động của các điều kiện lên giá nhà và diễn giải kết quả.

1.Đọc dữ liệu (Import data):

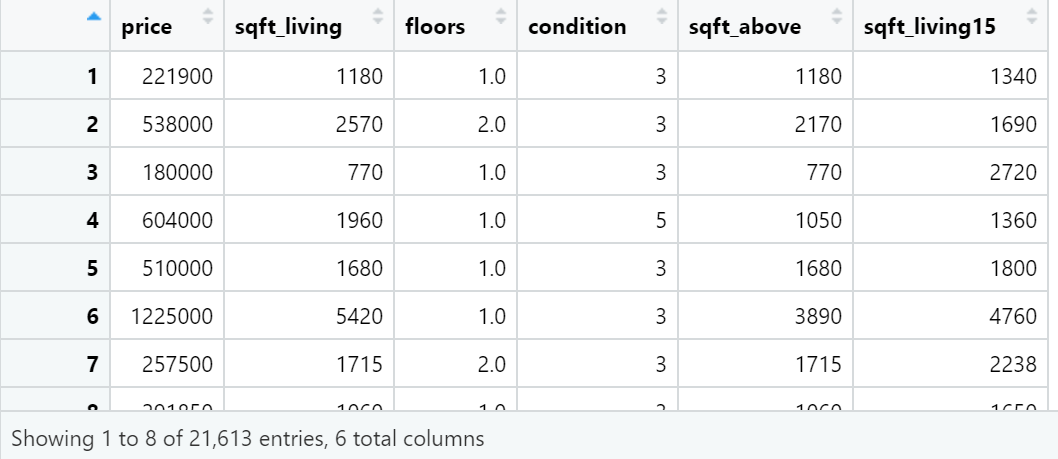
Hãy dùng lênh read.csv() để đọc tệp tin:



2. Làm sạch dữ liệu (Data cleaning):

(a) Hãy trích ra một dữ liệu con đặt tên là new\_DF chỉ bao gồm các biến chính mà ta quan tâm như đã trình bày trong phần giới thiệu dữ liệu. Từ câu hỏi này về sau, mọi yêu cầu xử lý đều dựa trên tập dữ liệu con new\_DF này:



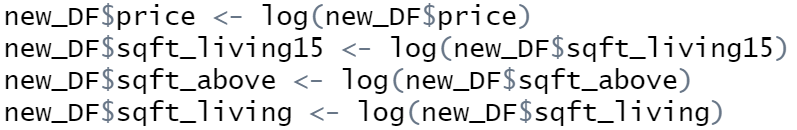


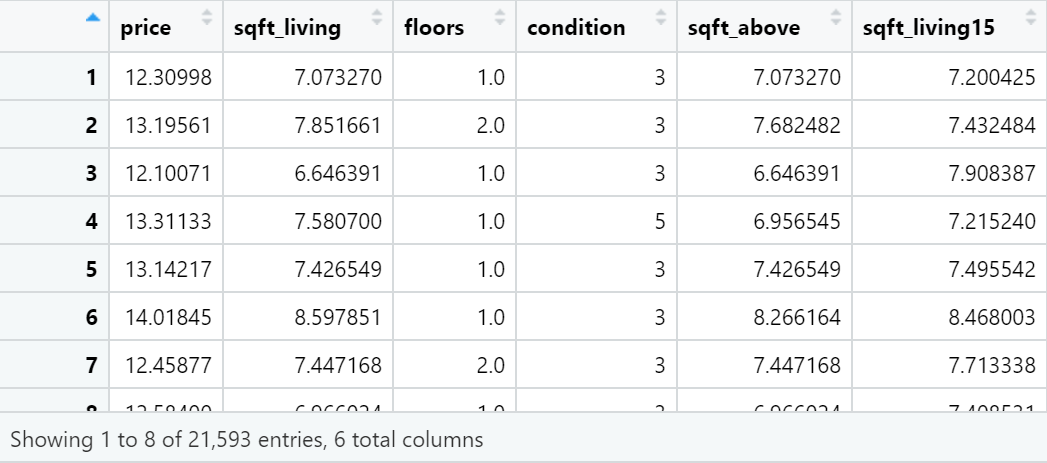
(b) Kiểm tra các dữ liệu bị khuyết trong tập tin. (Các câu lênh tham khảo: is.na(), which(), apply()). Nếu có dữ liệu bị khuyết, hãy đề xuất phương pháp thay thế cho những dữ liệu bị khuyết này:



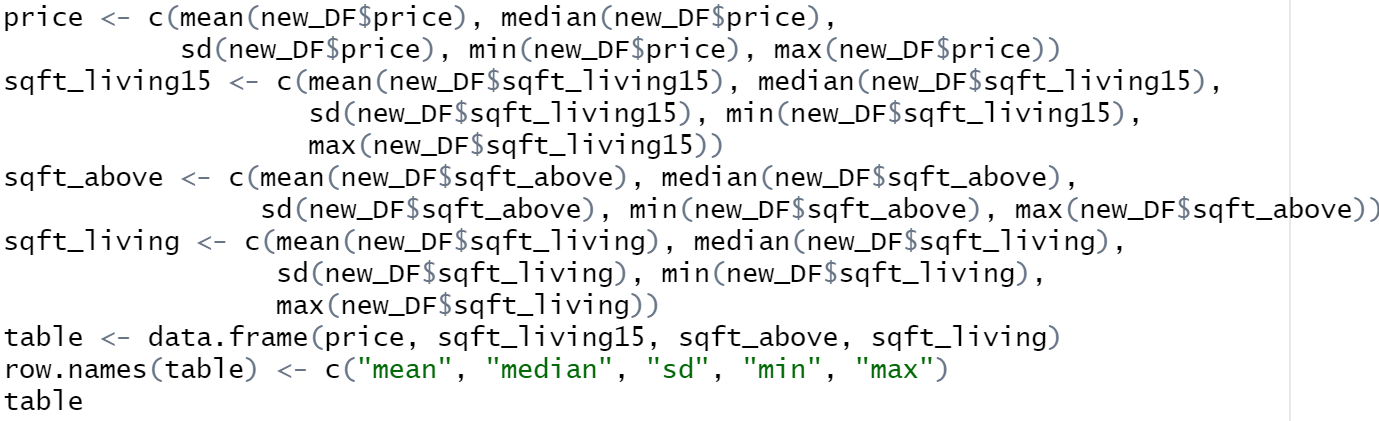
3. Làm rõ dữ liệu (Data visualization):

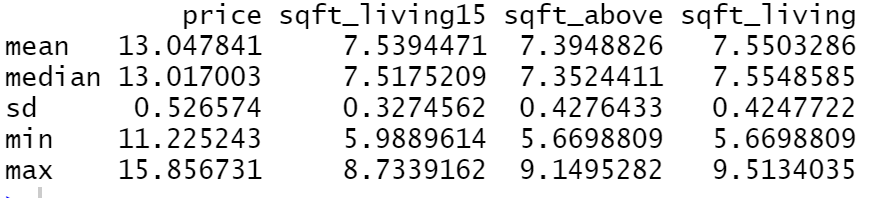
(a) Chuyển đổi các biến price, sqft\_living15, sqft\_above, sqft\_living lần lượt thành log(price), log(sqf t\_living15), log(sqf t\_above), và log(sqf t\_living). Từ đây mọi sự tính toán với các biến trên được hiểu là đã qua đổi biến dạng log.



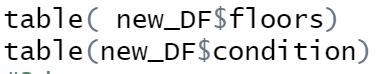


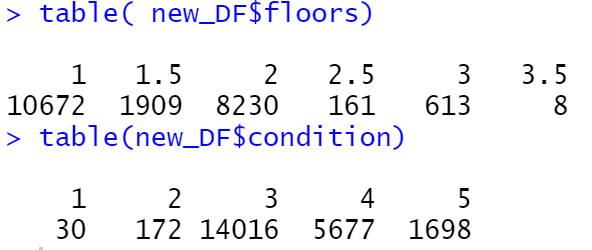
(b) Đối với các biến liên tục, hãy tính các giá trị thống kê mô tả bao gồm: trung bình, trung vị, độ lệch chuẩn, giá trị lớn nhất và giá trị nhỏ nhất. Xuất kết quả dưới dạng bảng. (Hàm gợi ý: mean(), median(), sd(), min(), max() , apply(), as.data.frame(), rownames())





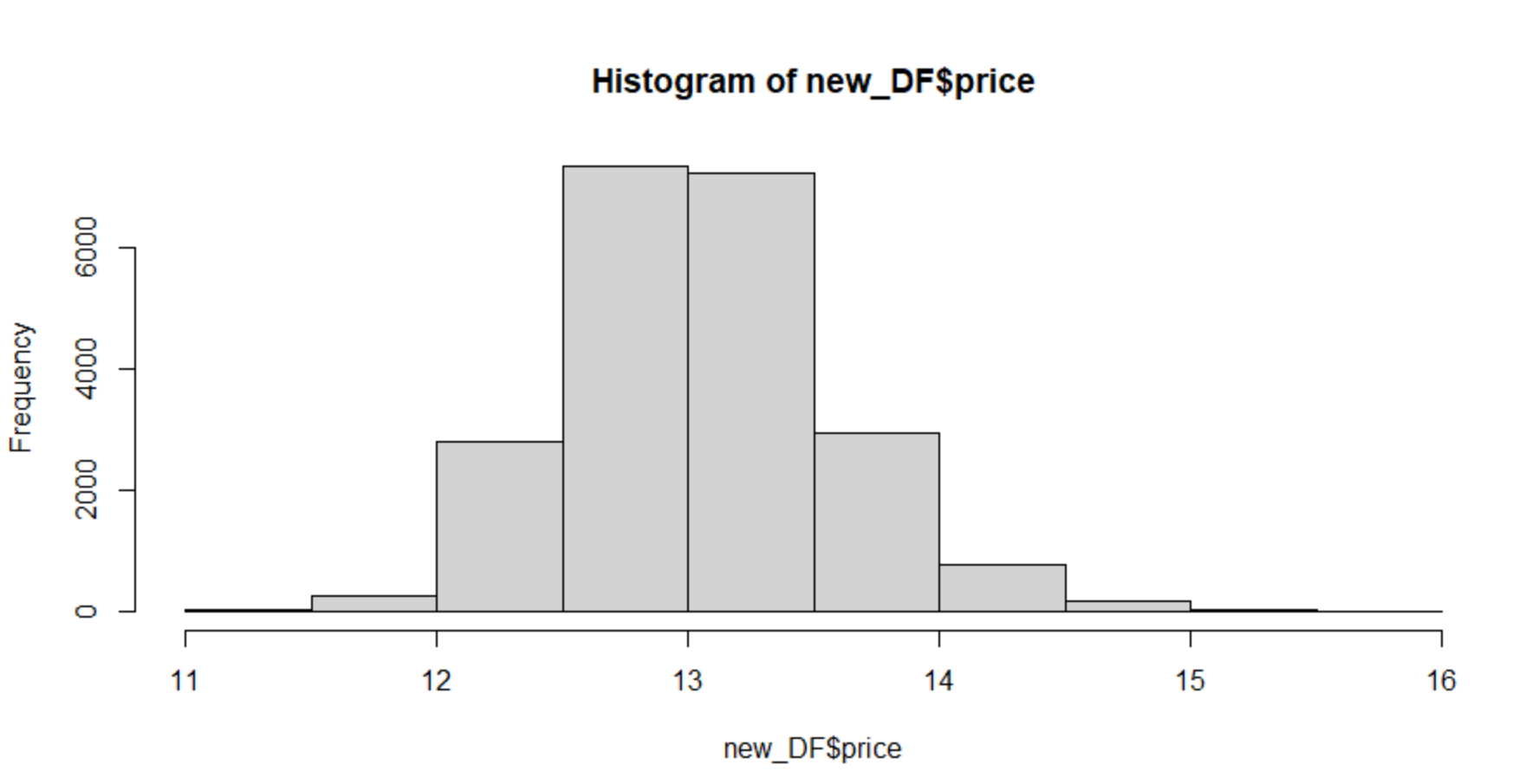
(c) Đối với các biến phân loại, hãy lập một bảng thống kê số lượng cho từng chủng loại (Hàm gợi ý: table()).



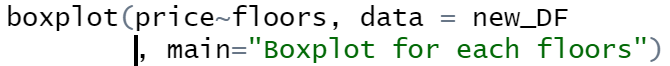


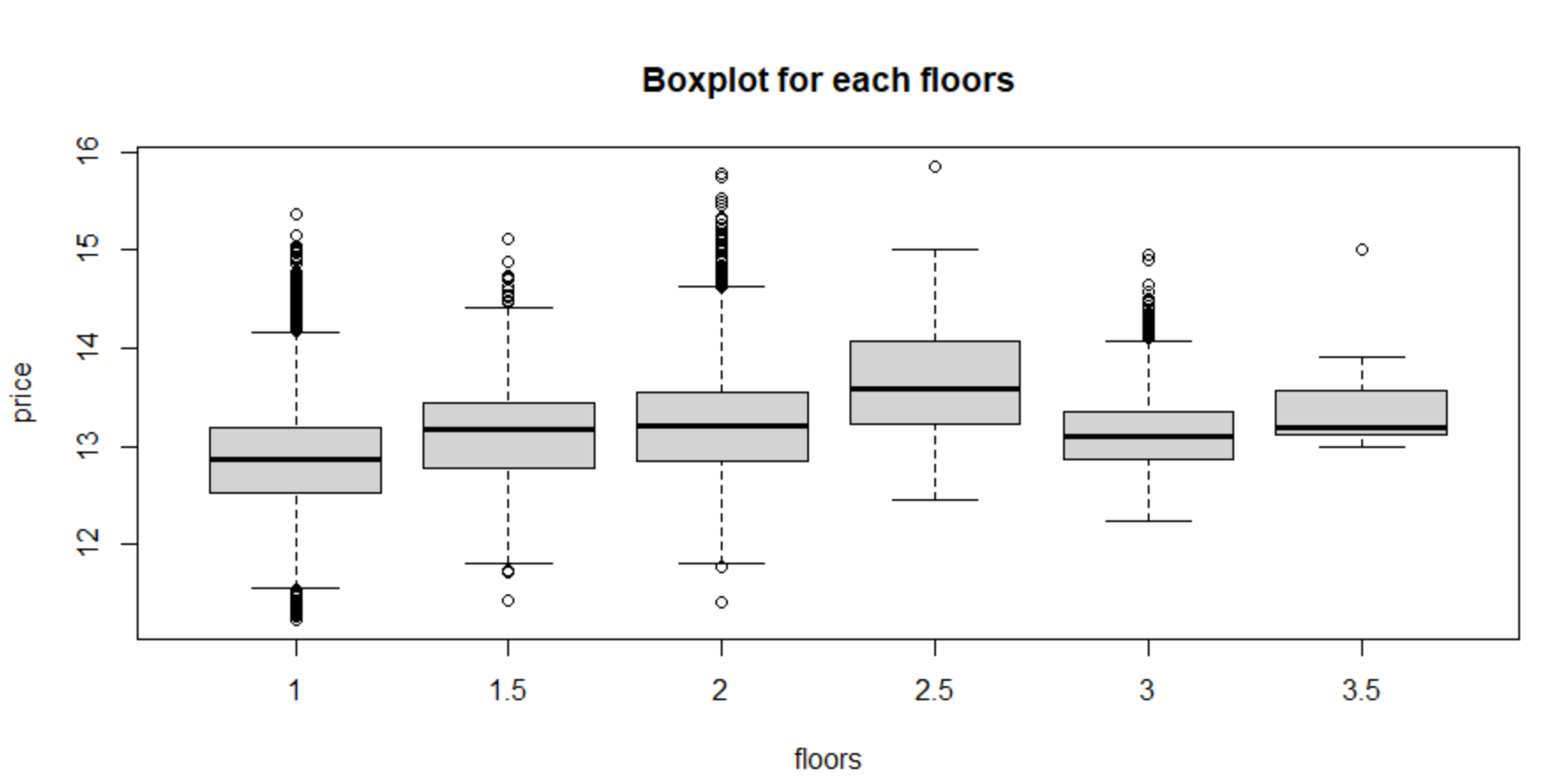
(d) Hãy dùng hàm hist() để vẽ đồ thị phân phối của biến price.

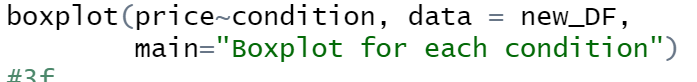


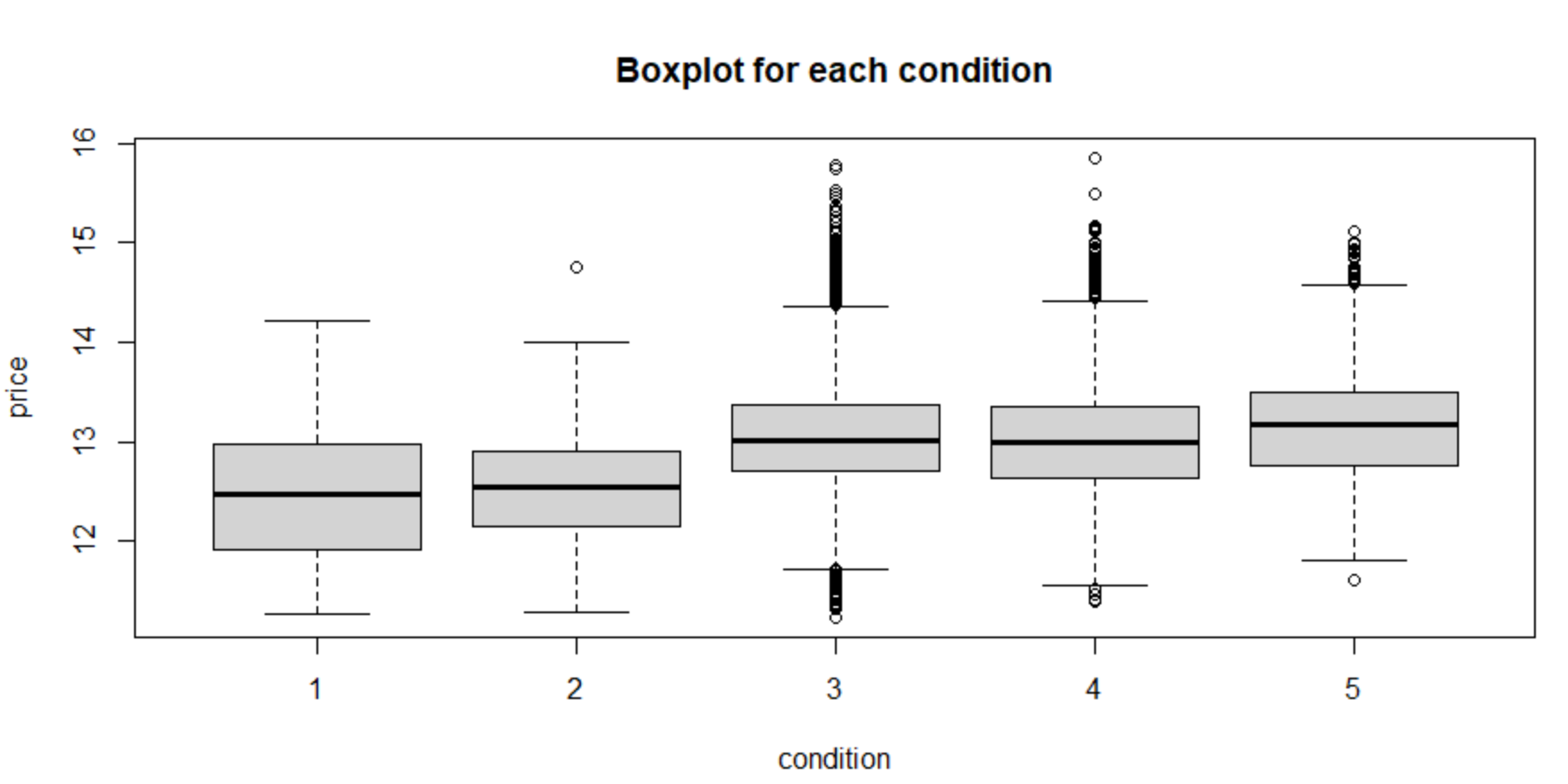


(e) Hãy dùng hàm boxplot() vẽ phân phối của biến price cho từng nhóm phân loại của biến floors và biến condition.

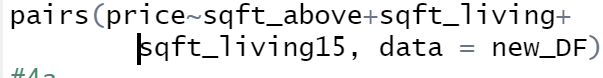


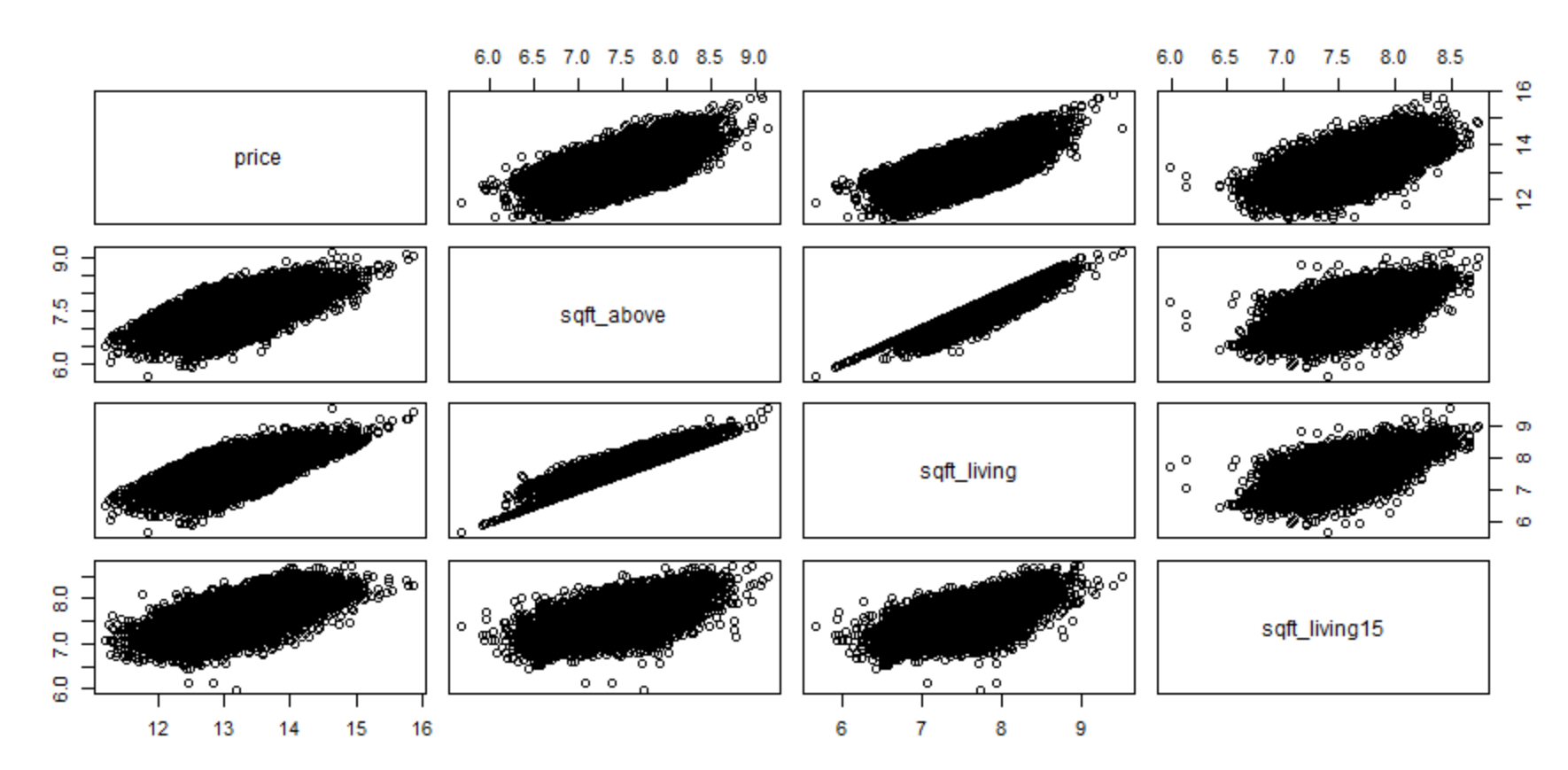






(f) Dùng lệnh pairs() vẽ các phân phối của biến price lần lượt theo các biến sqft\_living15, sqft\_above, và sqft\_living



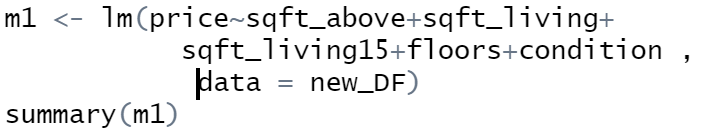


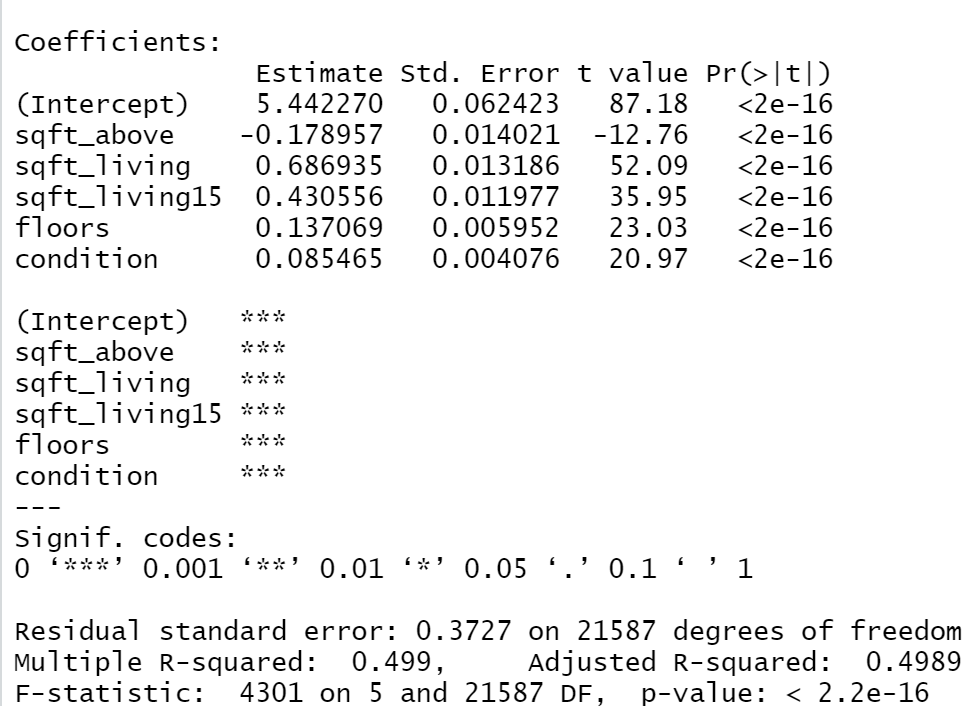
4. Xây dựng các mô hình hồi quy tuyến tính (Fitting linear regression models):

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

**Mô hình được biểu diễn như sau:**

price=β0+β1×sqft\_above+β2×sqft\_living+β3×sqft\_living15+ β4×floors+ β5×condition +ϵ.





Từ kết quả phân tích, ta thu được đường thẳng hồi quy ước lượng cho bởi phương trình sau:

price=5.442270-0.178957×sqft\_above+0.688935×sqft\_living+0.430556×sqft\_living15+ 0.137069×floors+ 0.085465×condition +ϵ.

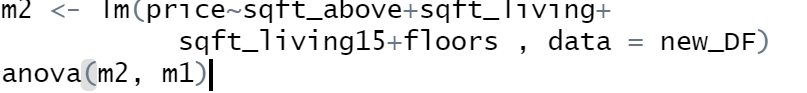
(b) Dựa vào kết quả của mô hình hồi quy tuyến tính trên, những biến nào bạn sẽ loại khỏi mô hình tương ứng với mức tin cậy 5%?

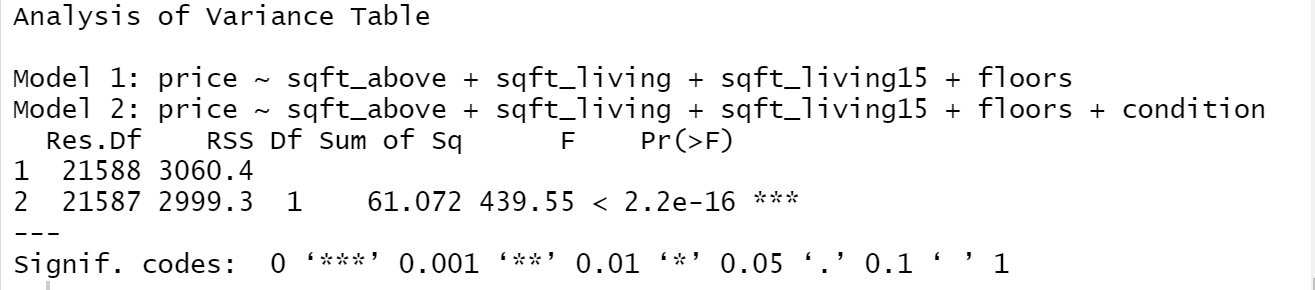
Ta thấy rằng p-value rất bé nên các biến đều có ý nghĩa rất cao. Điều này chỉ ra rằng, không có biến dự báo trong mô hình ta có loại bỏ.

(c) Xét 2 mô hình tuyến tính cùng bao gồm biến price là biến phụ thuộc nhưng:

• Mô hình M1 chứa tất cả các biến còn lại là biến độc lập

• Mô hình M2 là loại bỏ biến condition từ mô hình M1. Hãy dùng lệnhh anova() để đề xuất mô hình hồi quy hợp lý hơn.





Kết quả của Df là 1 (chỉ ra rằng mô hình phức tạp hơn nhiều hơn một biến dự báo), và kết quả p value rất bé (< .001). Có nghĩa là việc thêm biến condition vào mô hình dẫn đến sự cải thiện đáng kể cho mô hình. Do đó ta chọn mô hình với biến condition là mô hình hợp lý hơn.

(d) Chọn mô hình hợp lý hơn từ câu (c) hãy suy luận sự tác động của các biến lên giá nhà.

Để xét ảnh hưởng cụ thể của từng biến độc lập, ta xét trọng số (hệ số βi) và p-value tương ứng. Ta thấy rằng p-value tương ứng của tác cả các biến đều bé hơn 2.2×10-16, điều này nói lên rằng ảnh hưởng của tất cả các biến có ý nghĩa rất cao lên biến giá nhà price. Mặt khác, hệ số hồi quy βi của một biến dự báo cũng có thể được xem như ảnh hưởng trung bình lên biến phụ thuộc doanh thu khi tăng một đơn vị của biến dự báo đó, giả sử rằng các biến dự báo khác không đổi. Cụ thể :

Với sqft\_above giảm 0.178957 thì price tăng 1

Với sqft\_living tăng 0.688935 thì price tăng 1

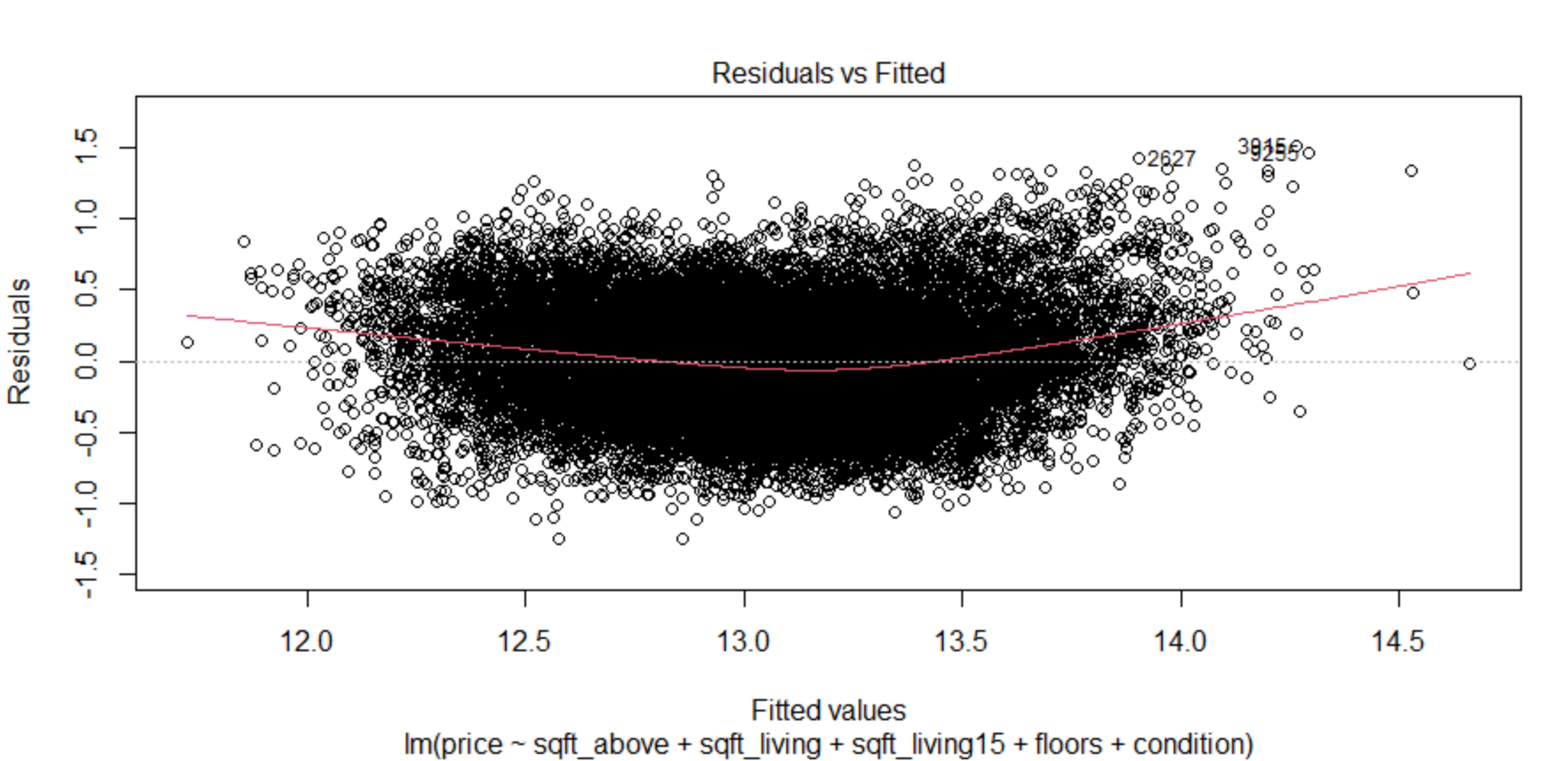
Với sqft\_living15 tăng 0.430556 thì price tăng 1

Với floors tăng 0.137069 thì price tăng 1

Với conditioin tăng 0.085465 thì price tăng 1

(e) Từ mô hình hồi quy mà bạn chọn ở câu (c) hãy dùng lệnh plot() để vẽ đồ thị biểu thị sai số hồi quy (residuals) và giá trị dự báo (fitted values). Nêu ý nghĩa và nhận xét đồ thị.

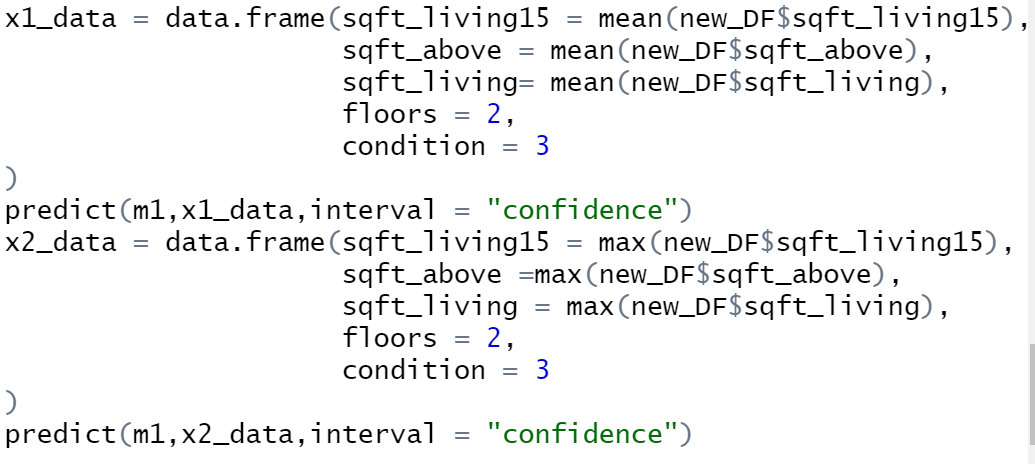


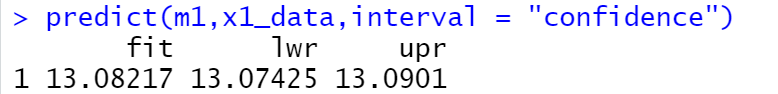


**Ý nghĩa :** Đồ thị (Residuals vs Fitted) vẽ các giá trị dự báo với các giá trị thặng dư (sai số) tương ứng, dùng để kiểm tra tính tuyến tính của dữ liệu và tính đồng nhất của các phương sai sai số. Nếu như giả định về tính tuyến tính của dữ liệu **KHÔNG** thỏa, ta sẽ quan sát thấy rằng các điểm thặng dư (residuals) trên đồ thị sẽ phân bố theo một hình mẫu (pattern) đặc trưng nào đó (ví dụ parabol). Nếu đường màu đỏ trên đồ thị phân tán là đường thẳng nằm ngang mà không phải là đường cong, thì giả định tính tuyến tính của dữ liệu được thỏa mãn. Để kiểm tra giả định thứ 3 (phương sai đồng nhất) thì các điểm thặng dự phải phân tán đều nhau xung quanh đường thẳng y=0.

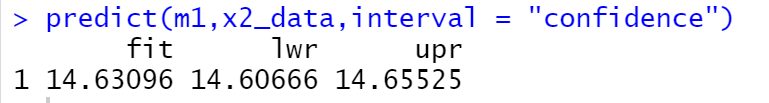
**Nhận xét:** Đồ thị thứ 1 (*Residuals vs Fitted*) cho thấy giả định về tính tuyến tính của dữ liệu hơi bị vi phạm, ta có thể thấy rằng sự vi phạm này bởi vì mối quan hệ giữa price và các biến dự báo là phi tuyến tính.

5. Dự báo (Predictions:) (a) Từ mô hình bạn chọn trong câu (c), hãy dùng lệnh predict() để dự báo giá nhà tại 2 thuộc tính như sau: x1: sqft\_living15 = mean(sqft\_living15), sqft\_above = mean(sqft\_above), sqft\_living = mean(sqft\_living), floor = 2, condition = 3 x2: sqft\_living15 = max(sqft\_living15), sqft\_above = max(sqft\_above), sqft\_living = max(sqft\_living), floor = 2, condition = 3. So sánh khoảng tin cậy cho 2 giá trị dự báo này.





Khoảng tin cậy : 13.07425 – 13.0901

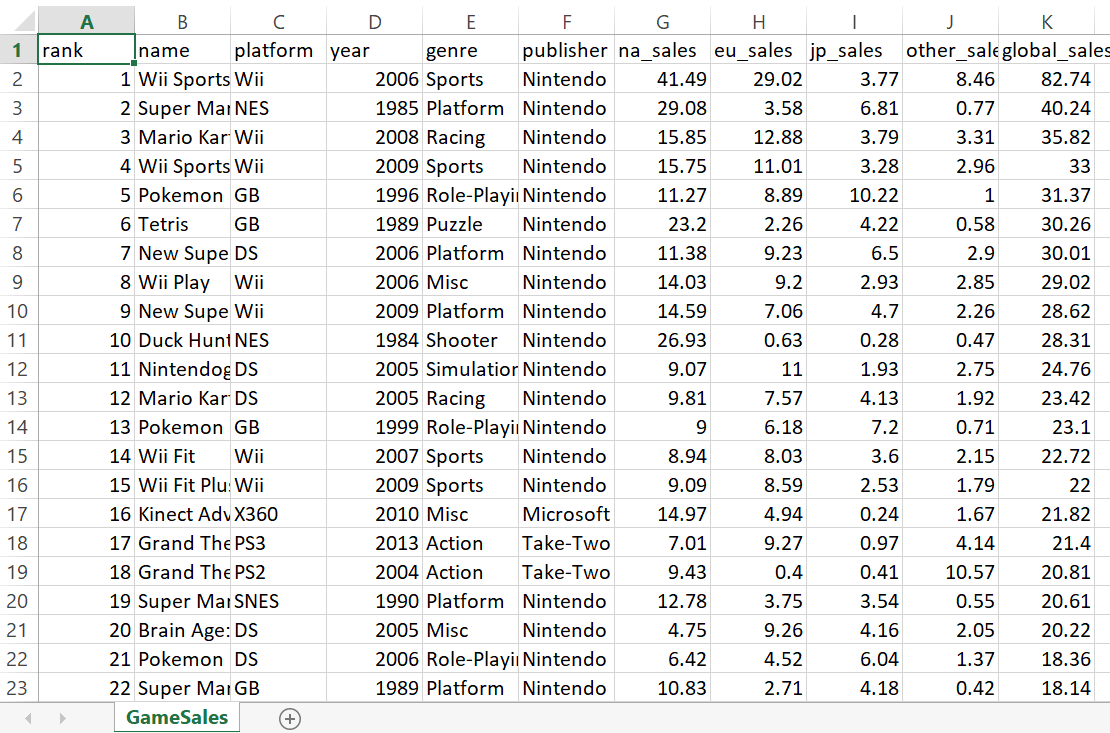


Khoảng tin cậy : 14.60666 – 14.65525

# II.Phần riêng:

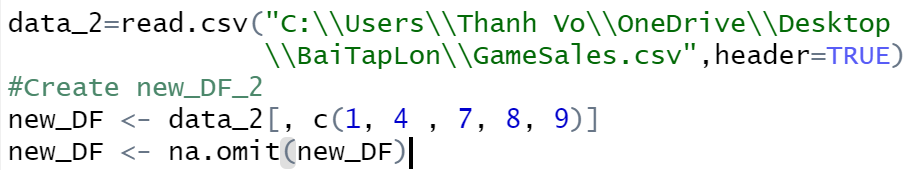
**Mô tả dữ liệu:**

Tập tin GameSales.csv là tập số liệu về doanh thu bán game toàn cầu trong giai đoạn (1983-2017)

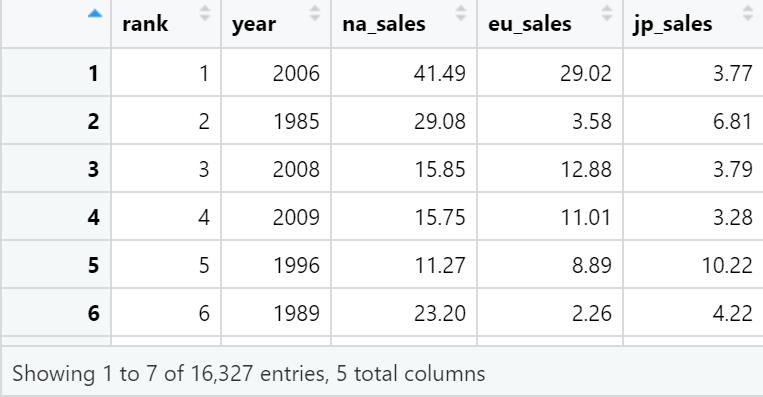


**Mục tiêu:** xây dựng mô hình hồi quy bội (multiple regression) để phân tích tác động của biến do ta lựa chọn tác động lên na\_sales và diễn giải kết quả.

**Data cleaning:**

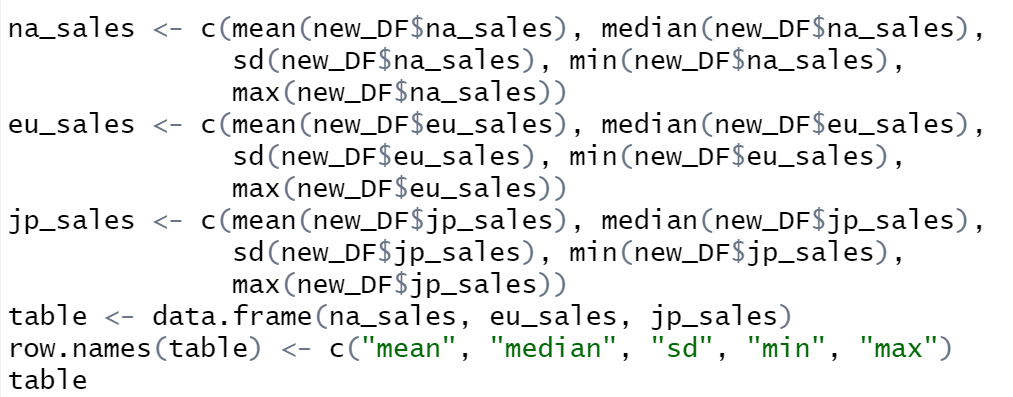


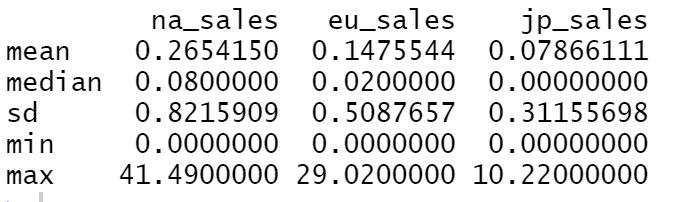
Đọc file dữ liệu và chọn các biến chính mà ta quan tâm. Sau đó loại bỏ hàng có giá trị NA.



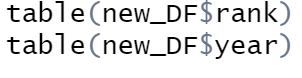
**Data visualization:**

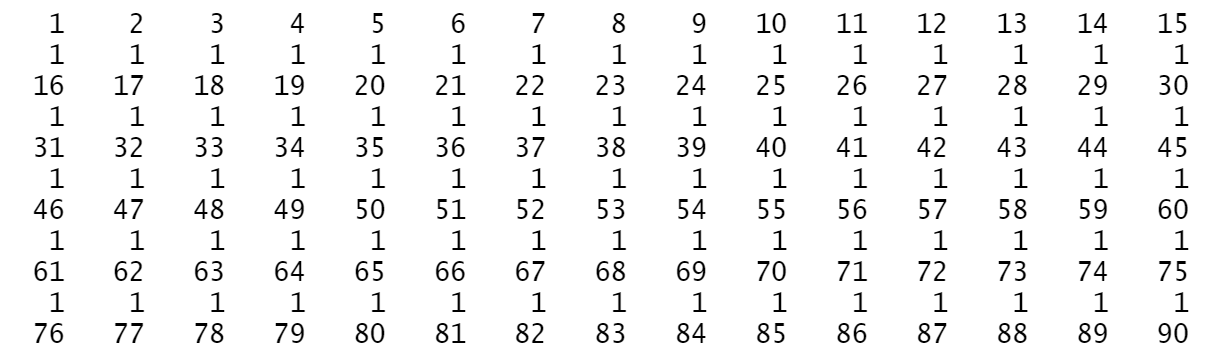
Đối với các biến liên tục, ta tính các giá trị thống kê mô tả bao gồm: trung bình, trung vị, độ lệch chuẩn, giá trị lớn nhất và giá trị nhỏ nhất. Xuất kết quả dưới dạng bảng.

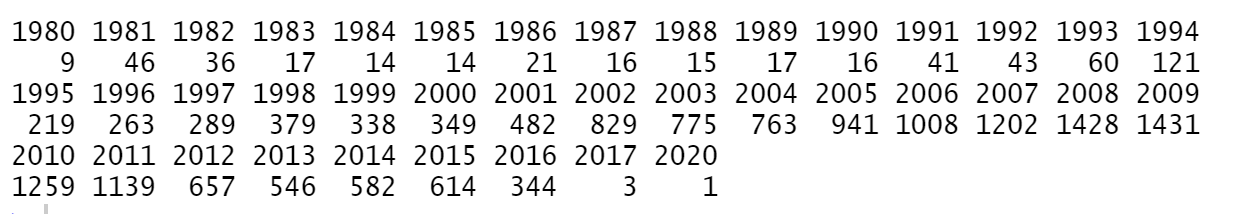
****

****

Đối với các biến phân loại, lập một bảng thống kê số lượng cho từng chủng loại:

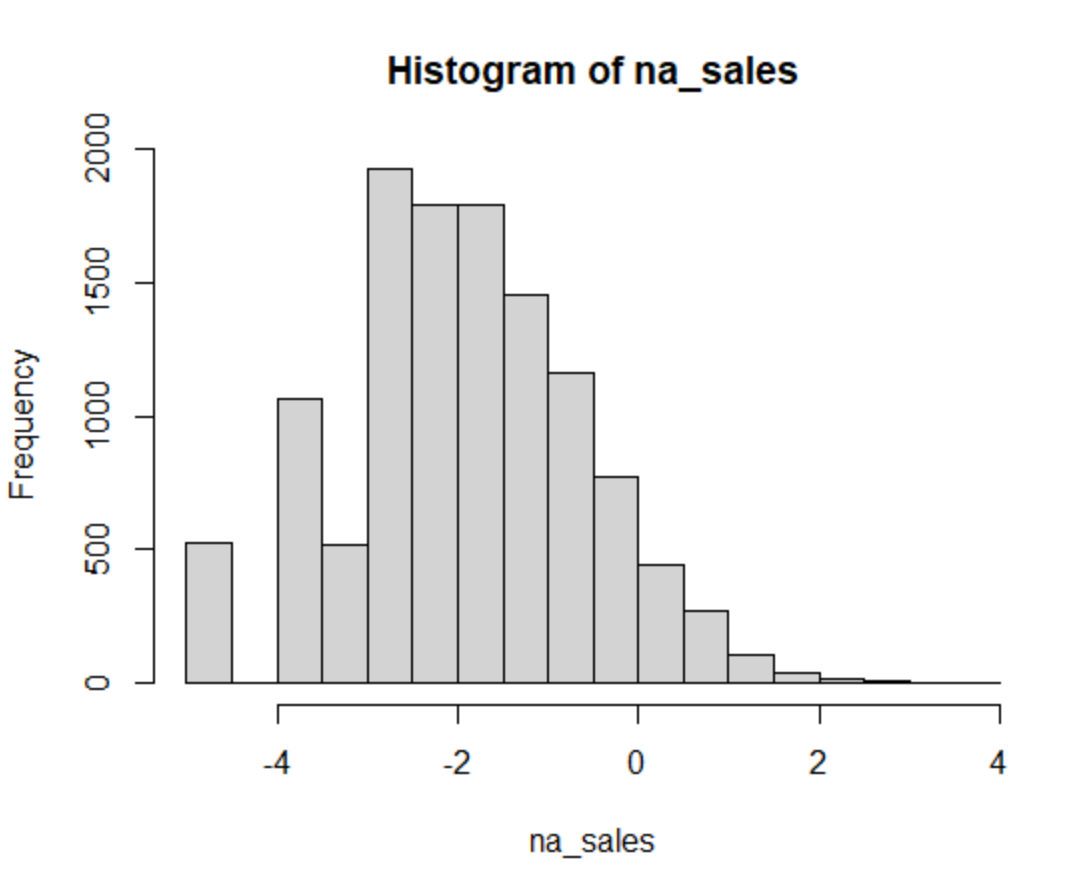






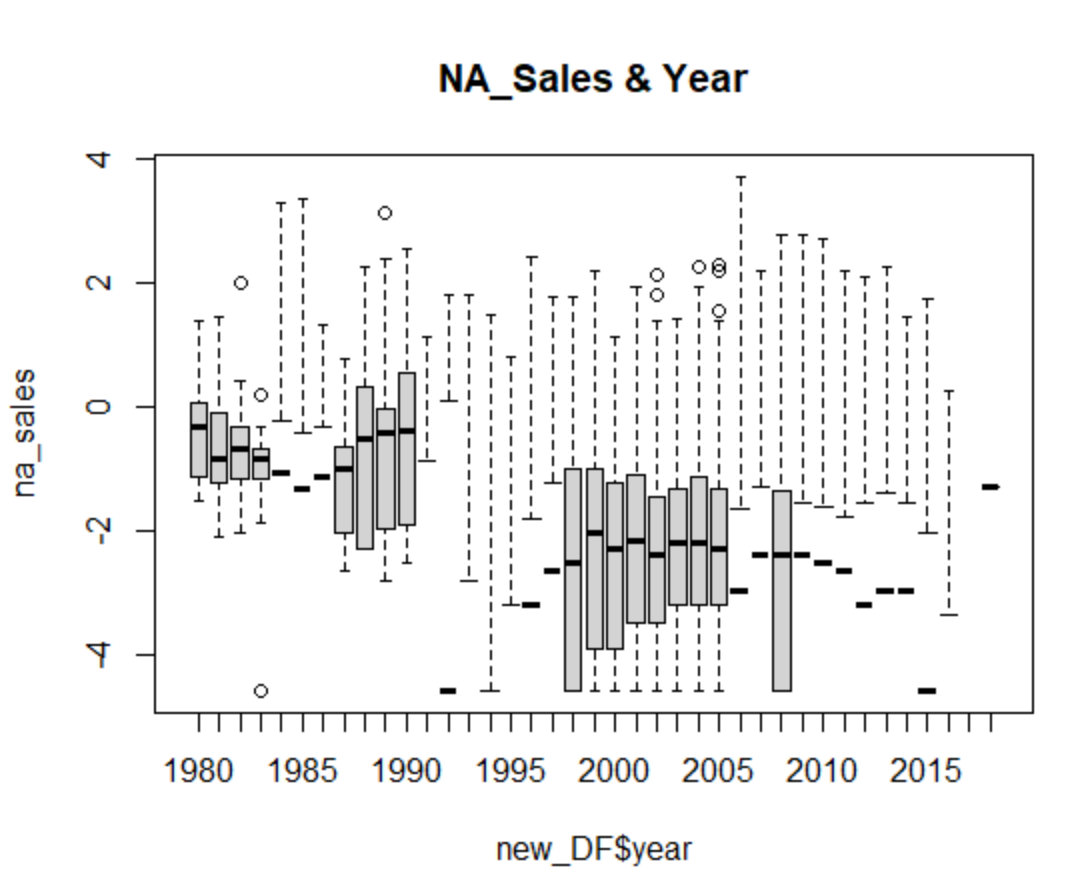
Dùng hàm hist() để vẽ đồ thị phân phối của biến na\_sales dưới dạng log.



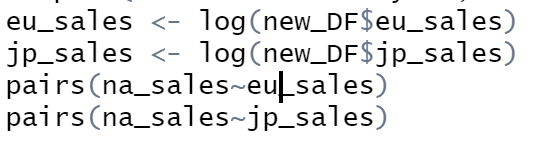


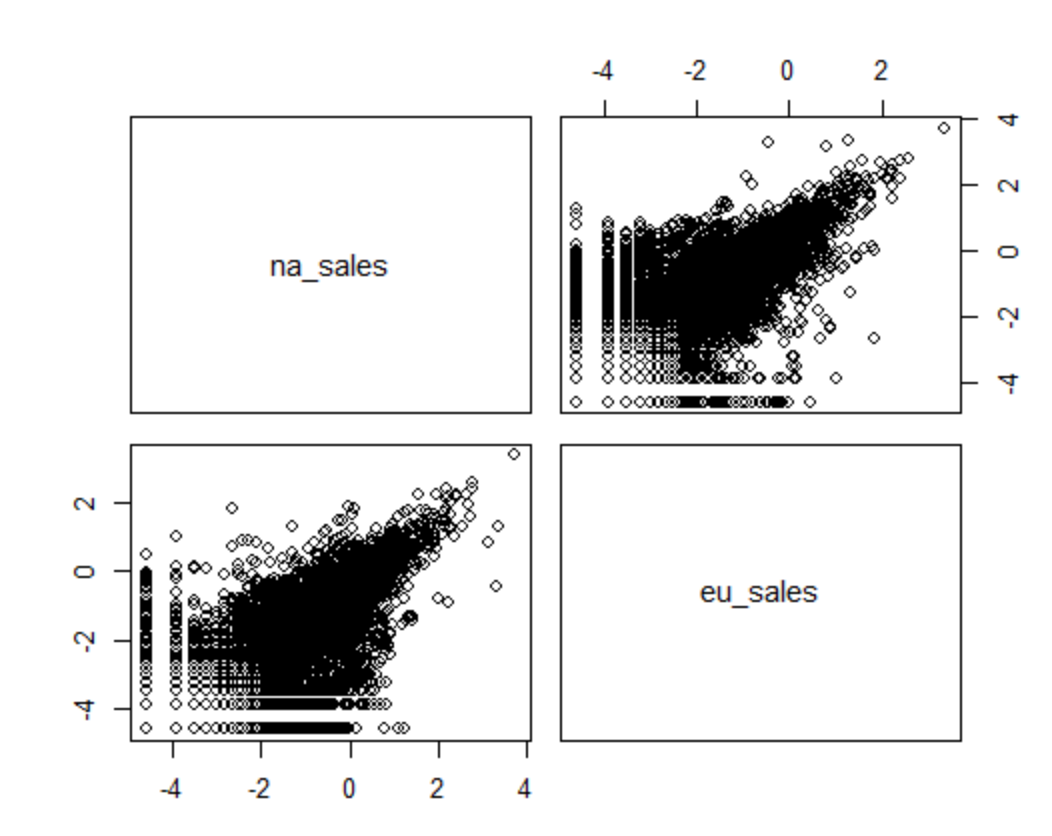
Hãy dùng hàm boxplot() vẽ phân phối của biến na\_sales cho từng nhóm phân loại của biến rank và biến year.

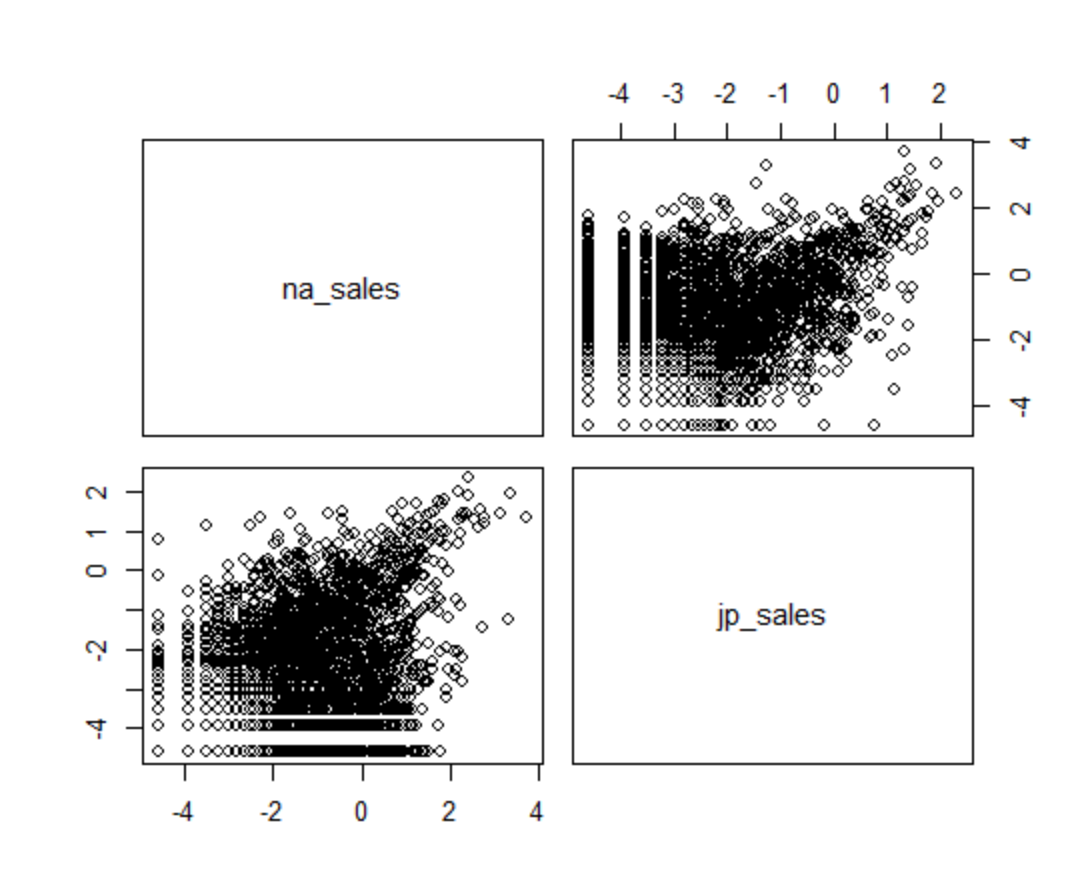




Dùng lệnh pairs() vẽ các phân phối của biến na\_sales lần lượt theo các biến eu\_sales và jp\_sales







**Modeling:**

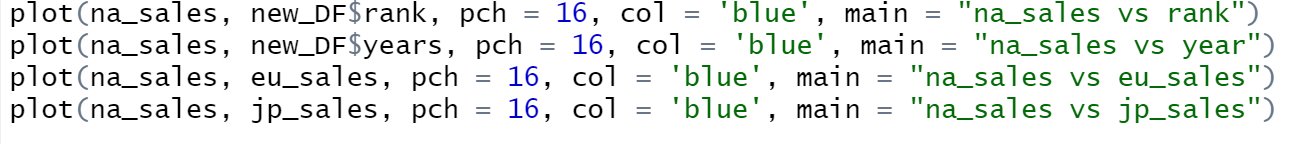
Mô hình hồi quy bội bao gồm:

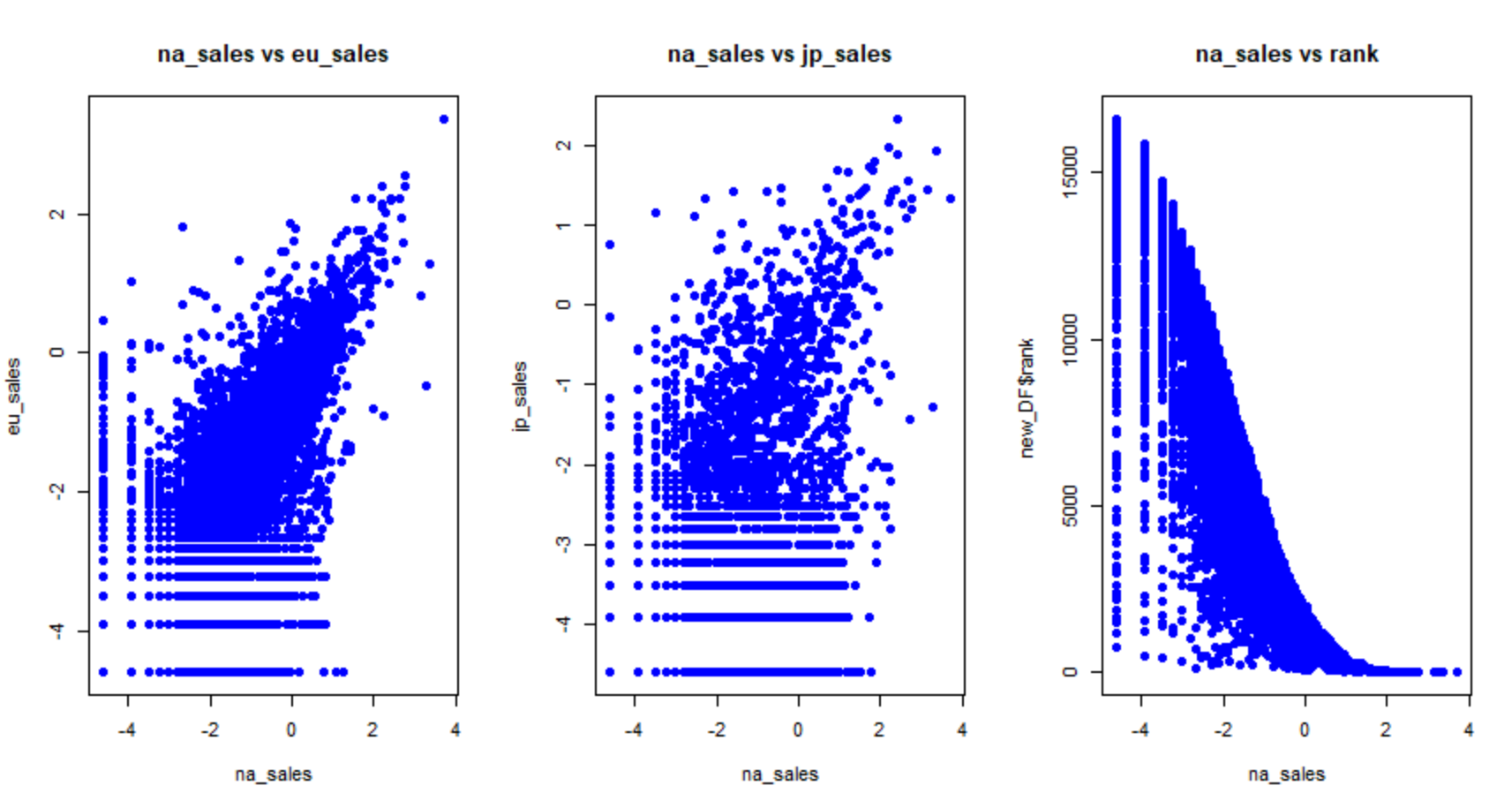
Biến phụ thuộc: doanh thu na\_sales

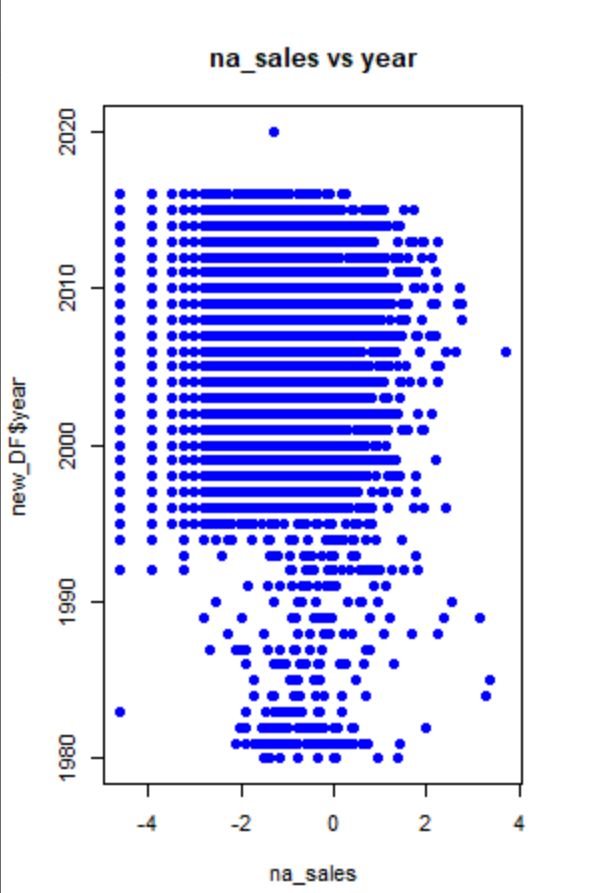
Biến dư báo (biến độc lập): rank, year, eu\_sales, jp\_sales lần lượt là hạng của game, năm xuất bản, doanh thu của Châu Âu , doanh thu ở Nhật Bản.

**Mô hình được biểu diễn như sau:**

na\_sales=β0+β1×rank+β2×year+β3×eu\_sales+ β4×jp\_sales +ϵ.



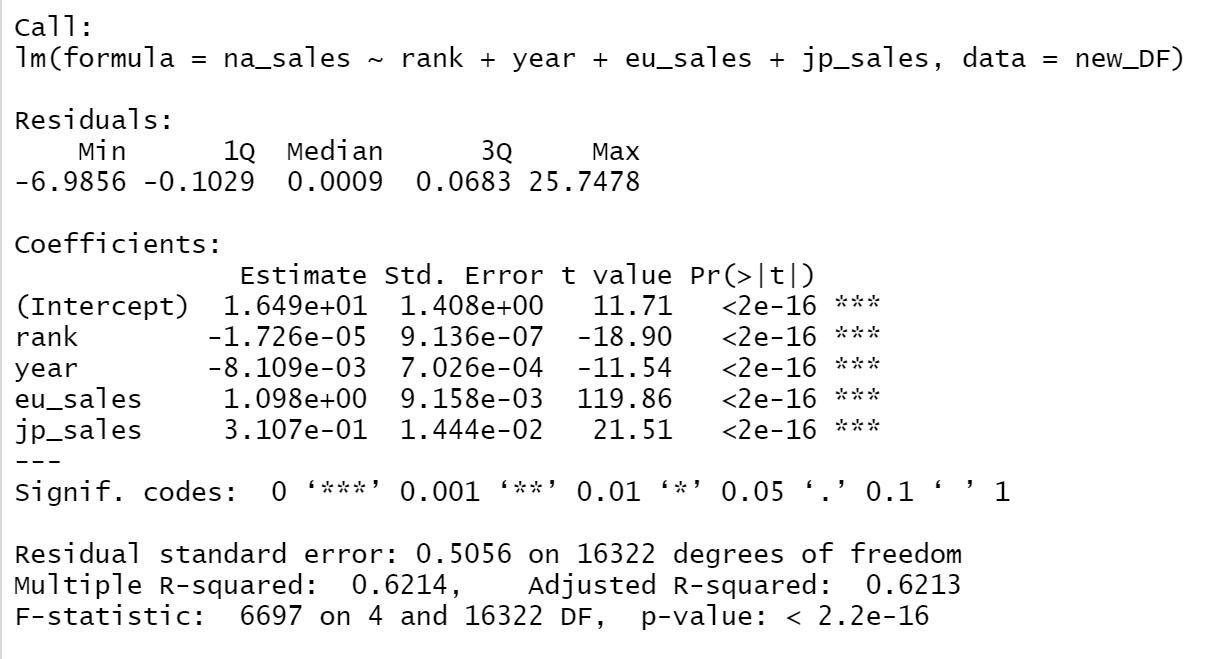


****

Từ các đồ thị phân tán, ta có thể nhận xét rằng các biến jp\_sales và eu\_sales có mối quan hệ tuyến tính với biến na\_sales trong khi biến years không có. Mặt khác, mối quan hệ giữa eu\_sales, jp\_sales, rank, nói một cách chính xác hơn là phi tuyến tính hơn là tuyến tính.

Ta ước lượng các hệ số βi, i=0,…,3sử dụng lệnh **lm()**:

****

****

2) Phân tích kết quả

Trước hết, ta thấy rằng p-value tương ứng với thống kê F bé hơn 2.2×10-16, có ý nghĩa rất cao. Điều này chỉ ra rằng, ít nhất một biến dự báo trong mô hình có ý nghĩa giải thích rất cao cho biến doanh thu na\_sales.

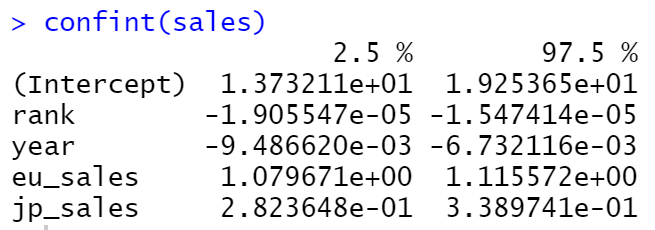
Để xét ảnh hưởng cụ thể của từng biến độc lập, ta xét trọng số (hệ số βi) và p-value tương ứng. Ta thấy rằng p-value tương ứng của tất cả các biến đều bé hơn 2×10-16, điều này nói lên rằng ảnh hưởng của tất cả biến có ý nghĩa rất cao lên biến doanh thu na\_sales.

Mô hình hồi quy bội về ảnh hưởng của rank, year, eu\_sales, jp\_sales lên na\_sales cho bởi:

na\_sales=16.49-1.726×10-5×rank – 8.109×10-3×year+1.098×eu\_sales+3.107×10-1 ×jp\_sales

Hệ số R2 hiệu chỉnh bằng 0.6214 nghĩa là 62.14% sự biến thiên trong doanh thu na\_sales được giải thích bởi các biến rank, year, eu\_sales, jp\_sales

Đề tìm khoảng tin cậy cho các hệ số hồi quy, ta sử dụng hàm **confint()**:



Khoảng tin cậy 95% cho các hệ số hồi quy cho bởi:

13.73≤β0≤19.25,

-1.9×10-5≤β1≤-1.5×10-5,

-9.4×10-3≤β2≤-6.7×10-3,

1.07≤β3≤1.16,

2.8×10-1≤β4≤3.4×10-1.

3) Kiểm tra các giả định của mô hình

Nhắc lại các giả định của mô hình hồi quy:

Yi=β0+β1X1+⋯+βpXp+ϵi,i=1,…,n

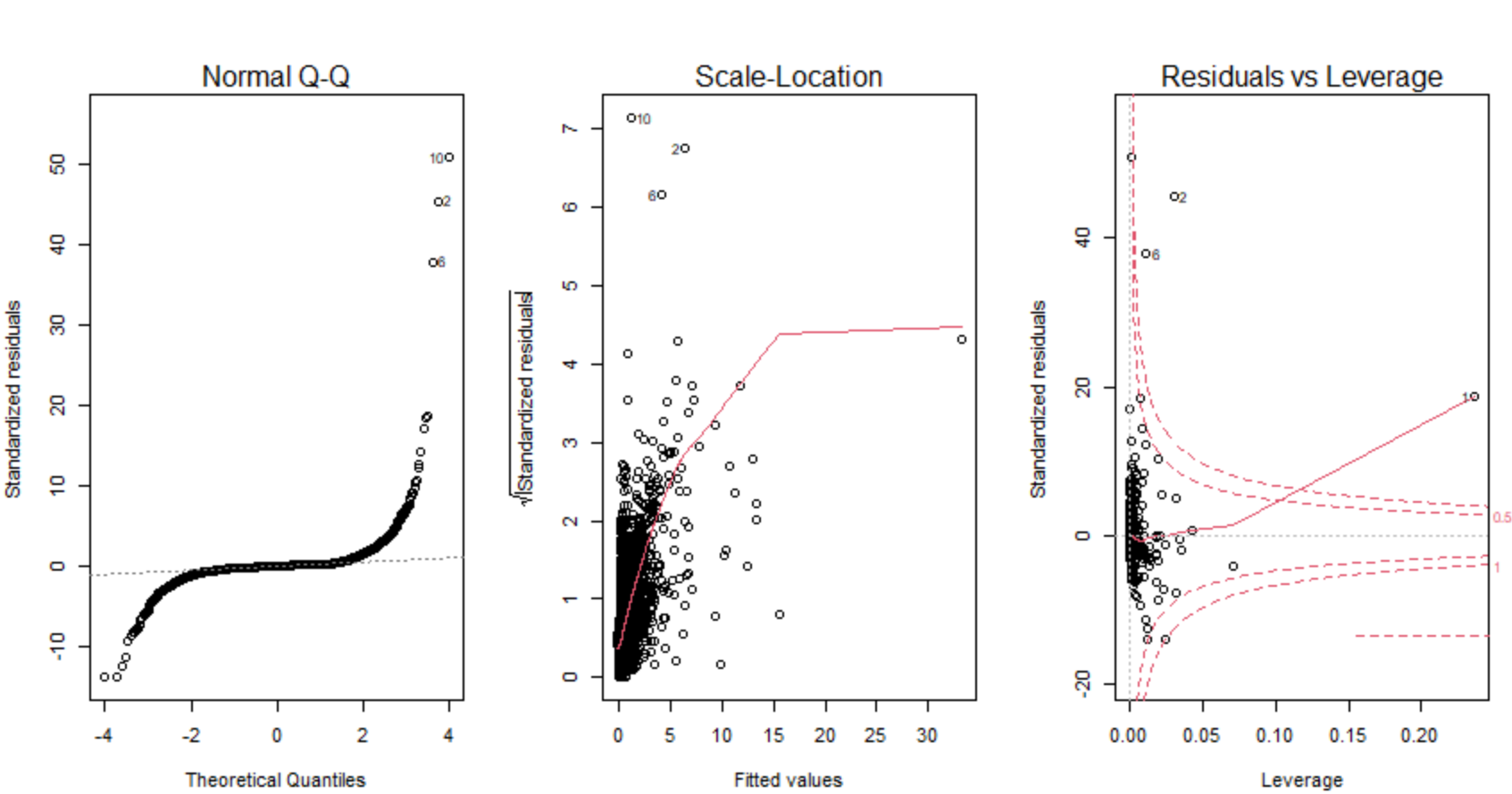
 1.**Tính tuyến tính của dữ liệu:** mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính

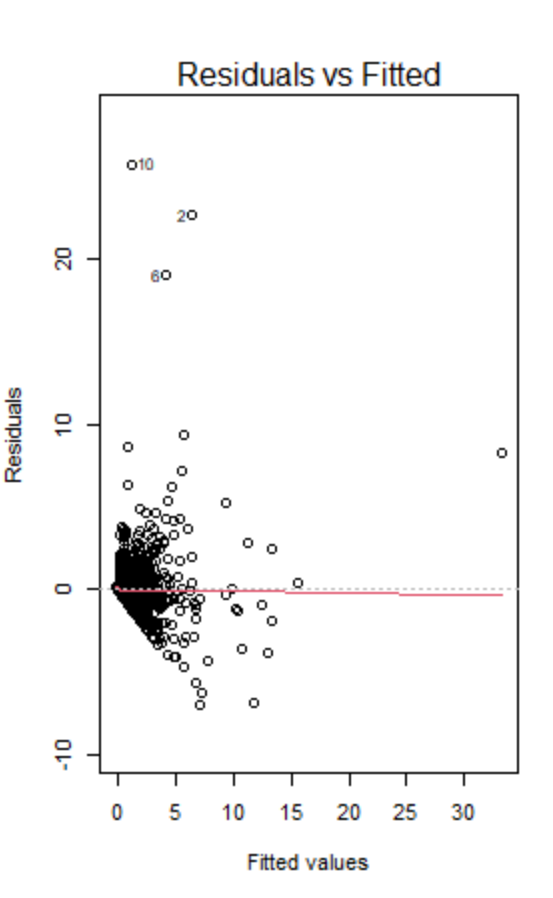
**2.Sai số có phân phối chuẩn**

**3.Phương sai của các sai số là hằng số:** ϵi∼N(0,σ2)

**4.Các sai số**ϵ1,…,ϵn **thì độc lập với nhau**

Ta thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình:





Đồ thị (Residuals vs Fitted) vẽ các giá trị dự báo (y^iy^i) với các giá trị thặng dư (sai số) tương ứng, dùng để kiểm tra tính tuyến tính của dữ liệu (giả định 1) và tính đồng nhất của các phương sai sai số (giả định 3). Nếu như giả định về tính tuyến tính của dữ liệu **KHÔNG** thỏa, ta sẽ quan sát thấy rằng các điểm thặng dư (residuals) trên đồ thị sẽ phân bố theo một hình mẫu (pattern) đặc trưng nào đó (ví dụ parabol). Nếu đường màu đỏ trên đồ thị phân tán là đường thẳng nằm ngang mà không phải là đường cong, thì giả định tính tuyến tính của dữ liệu được thỏa mãn. Để kiểm tra giả định thứ 3 (phương sai đồng nhất) thì các điểm thặng dự phải phân tán đều nhau xung quanh đường thẳng y=0y=0.

Đồ thị (Normal Q-Q) cho phép kiểm tra giả định về phân phối chuẩn của các sai số. Nếu các điểm thặng dư nằm trên cùng 1 đường thẳng thì điều kiện về phân phối chuẩn được thỏa.

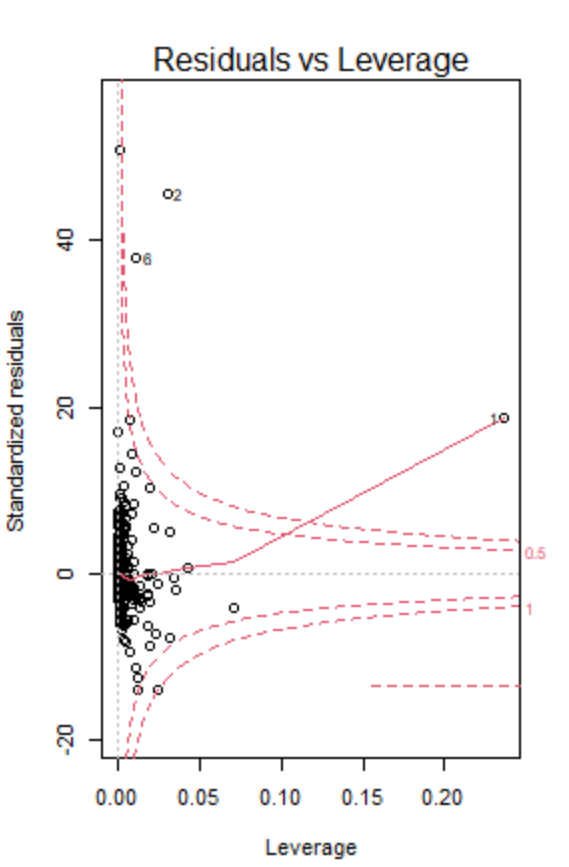
Đồ thị (Scale - Location) vẽ căn bậc hai của các giá trị thặng dư được chuẩn hóa với các giá trị dự báo, được dùng để kiểm tra giả định thứ 3 (phương sai của các sai số là hằng số). Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm thặng dư phân tán đều xung quanh đường thẳng này thì giả định thứ 3 được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm thặng dư phân tán không đều xung quanh đường thẳng này, thì giả định thứ 3 bị vi phạm.

Đồ thị (Residuals vs Leverage) cho phép xác định những điểm có ảnh hưởng cao (influential observations), nếu chúng có hiện diện trong bộ dữ liệu. Những điểm có ảnh hưởng cao này có thể là các điểm outliers, là những điểm có thể gây nhiều ảnh hưởng nhất khi phân tích dữ liệu. Nếu như ta quan sát thấy một đường thẳng màu đỏ đứt nét ([Cook’s distance](https://en.wikipedia.org/wiki/Cook%27s_distance)), và có một số điểm vượt qua đường thẳng khoảng cách này, nghĩa là các điểm đó là các điểm có ảnh hưởng cao. Nếu như ta chỉ quan sát thấy đường thẳng khoảng cách Cook ở góc của đồ thị và không có điểm nào vượt qua nó, nghĩa không có điểm nào thực sự có ảnh hưởng cao.

Nhận xét:

* Đồ thị (*Normal Q-Q)* cho thấy giả định sai số có phân phối chuẩn được chưa được thỏa mãn.
* Đồ thị (Residuals vs Fitted) cho thấy giả định về tính tuyến tính được thõa mãn.
* Đồ thị (Scale - Location) cho ta thấy rằng giả định về tính đồng nhất của phương sai đã bị vi phạm
* Đồ thị (Residuals vs Leverage) chỉ ra có các quan trắc thứ 1, 2 và 6 có thể là các điểm có ảnh hưởng cao trong bộ dữ liệu.





Tuy nhiên ta cũng quan sát thấy rằng các điểm này chưa vượt qua đường thẳng khoảng cách Cook (đường thẳng đứt nét màu đỏ Cook’s distance). Do vậy, các điểm này chưa thực sự là các điểm có ảnh hưởng cao trong bộ dữ liệu. Do đó ta không cần phải loại bỏ chúng khi phân tích.

***Các website dùng để tham khảo:***

[YaRrr! The Pirate’s Guide to R (bookdown.org)](https://bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-anova.html)

[MT2013\_010100\_DH\_HK201: Ví dụ về Hồi quy bội trên R (hcmut.edu.vn)](http://e-learning.hcmut.edu.vn/mod/resource/view.php?id=515616)