1.giaNha <- read.csv("C:\\Users\\Thanh Vo\\Desktop\\BaiTapLon\\gia\_nha.csv", header = TRUE)

2a. new\_DF <- giaNha[, c(6, 9, 11, 14, 16, 23)]

b. new\_DF <- na.omit(new\_DF)

3

a.

new\_DF$price <- log(new\_DF$price)

new\_DF$sqft\_living15 <- log(new\_DF$sqft\_living15)

new\_DF$sqft\_above <- log(new\_DF$sqft\_above)

new\_DF$sqft\_living <- log(new\_DF$sqft\_living)

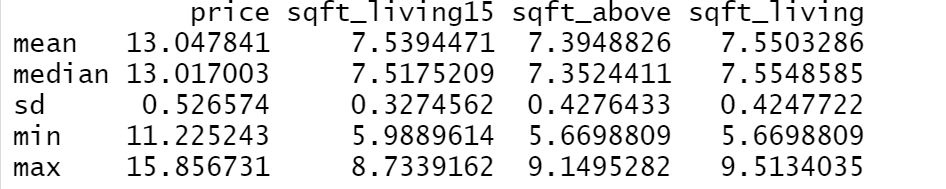
b.

new\_DF$price <- log(new\_DF$price)

new\_DF$sqft\_living15 <- log(new\_DF$sqft\_living15)

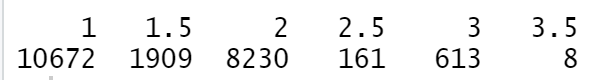
new\_DF$sqft\_above <- log(new\_DF$sqft\_above)

new\_DF$sqft\_living <- log(new\_DF$sqft\_living)

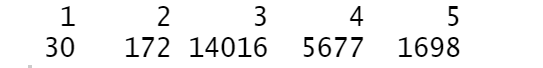


c.

table(new\_DF$floors)

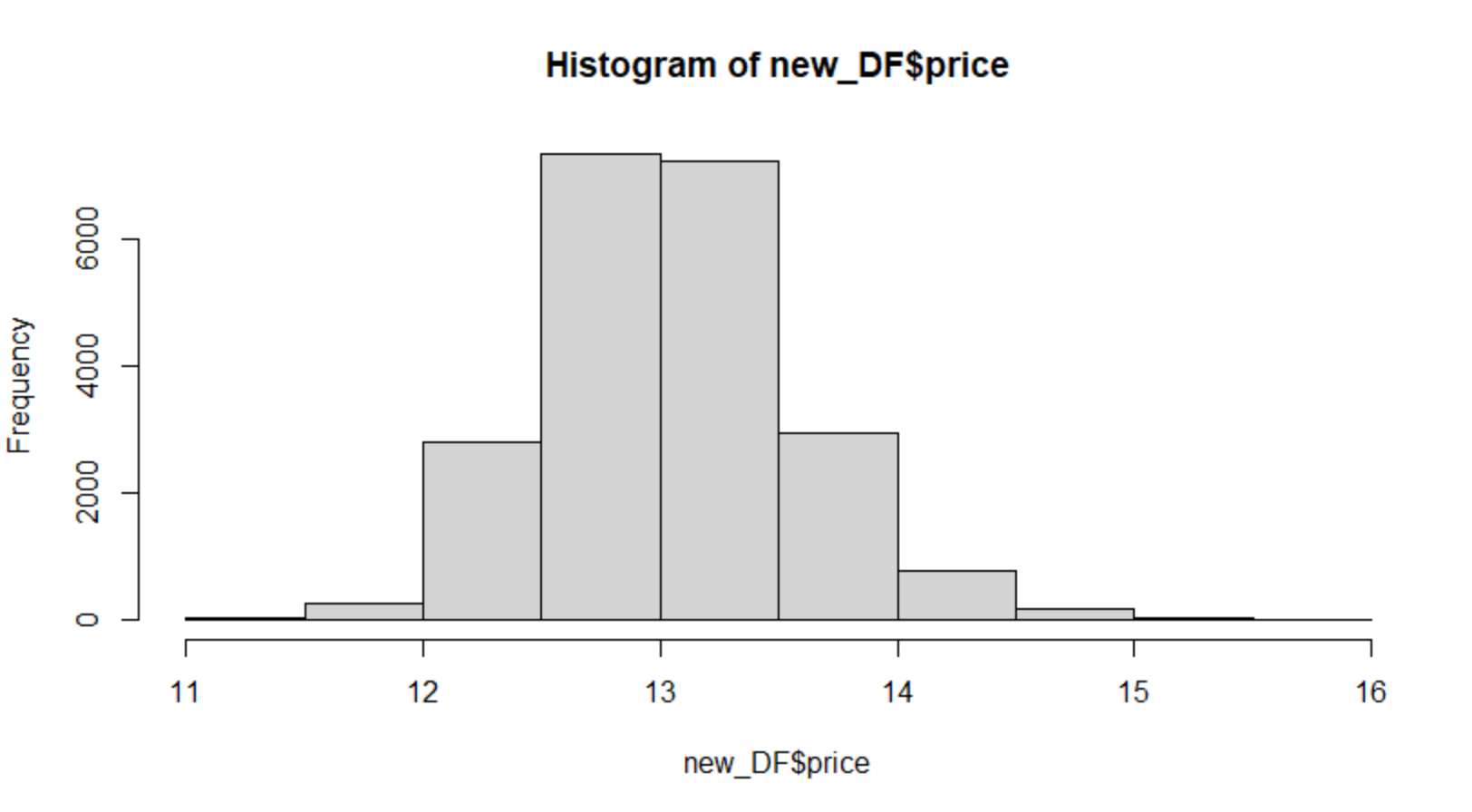


table(new\_DF$condition)



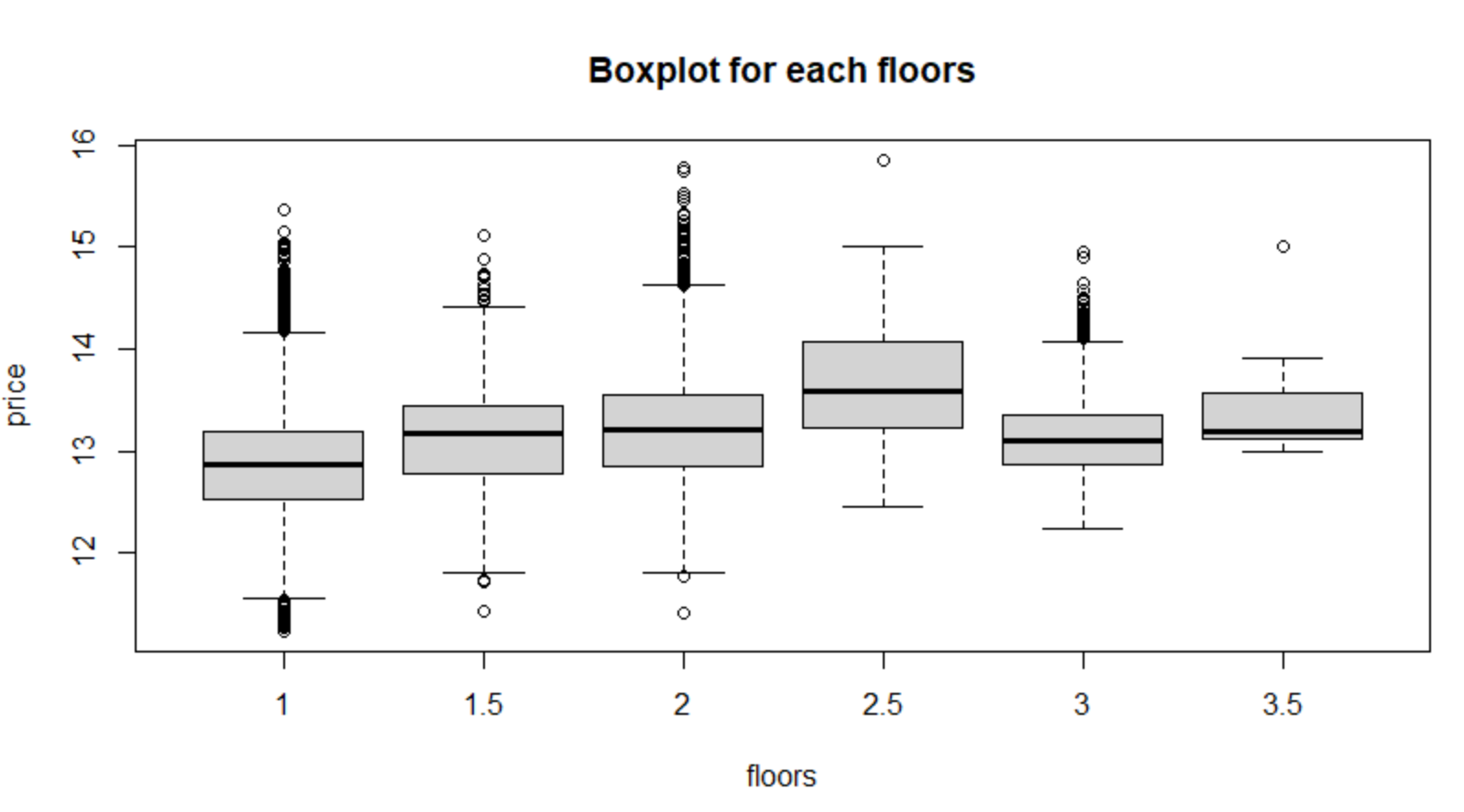
d.

hist(new\_DF$price)

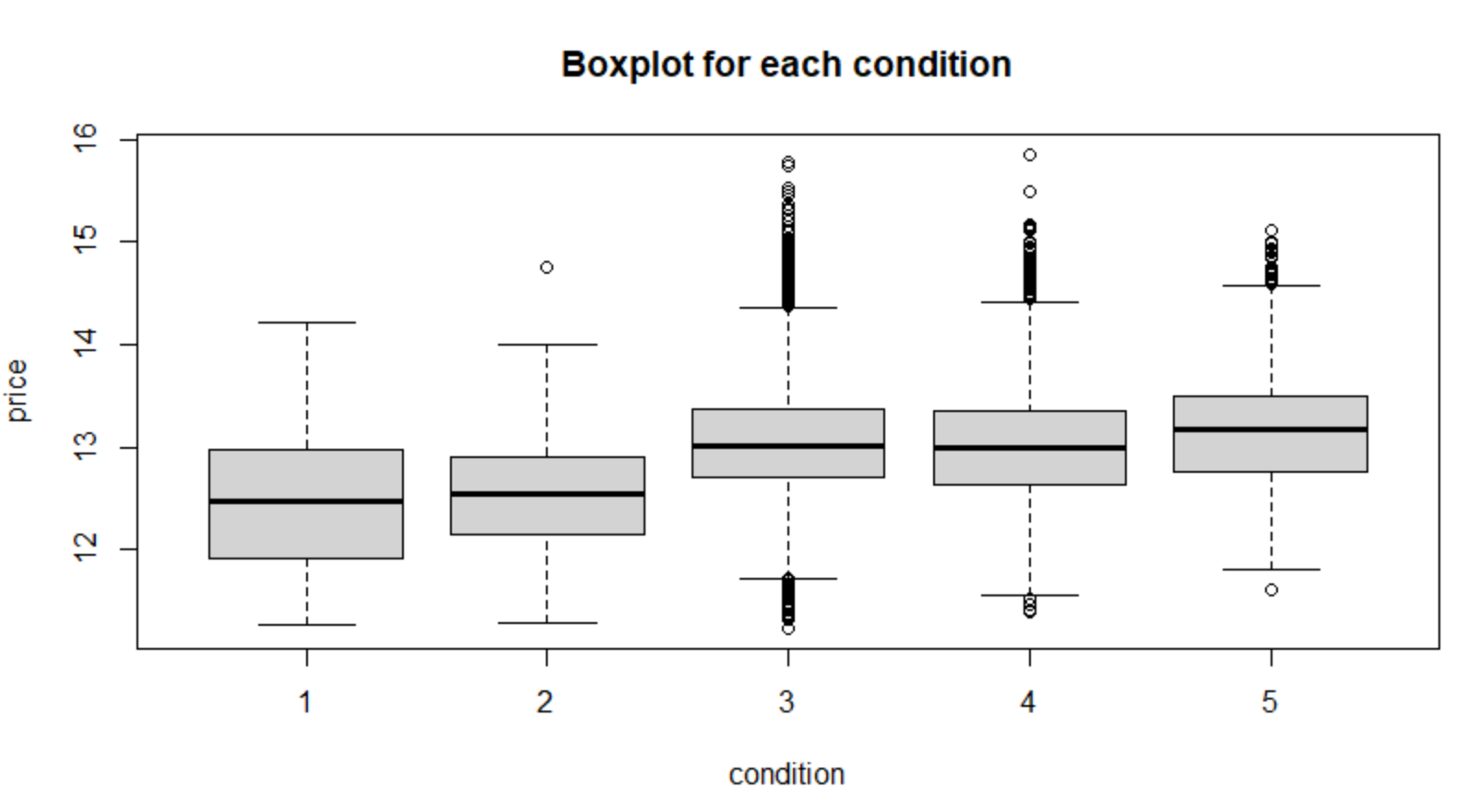


e.

boxplot(price~floors, data = new\_DF, main="Boxplot for each floors")

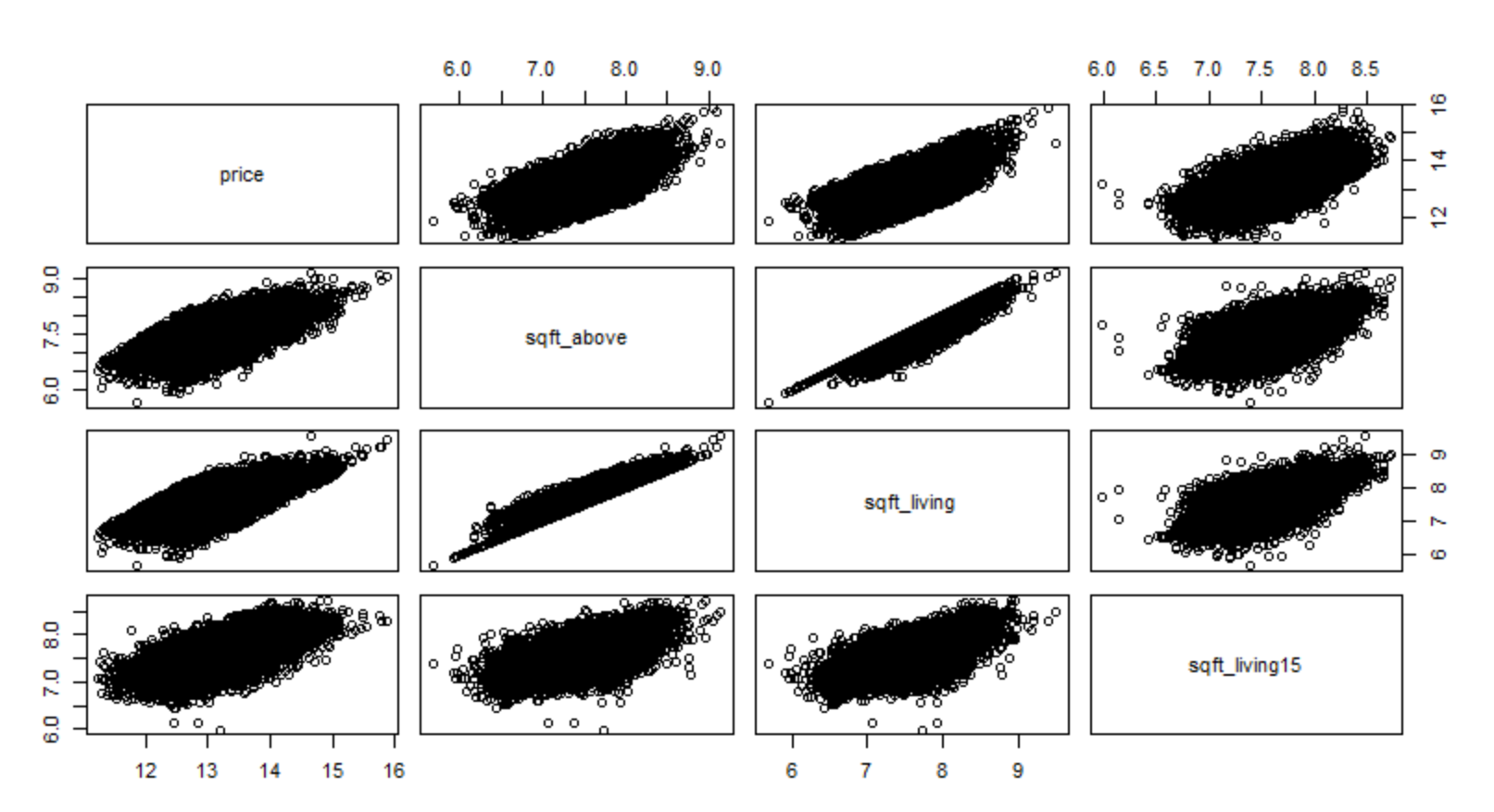


boxplot(price~condition, data = new\_DF, main="Boxplot for each condtion")



f.

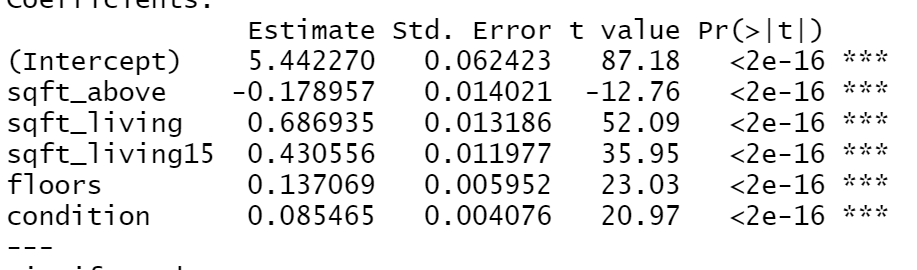
pairs(price~sqft\_above+sqft\_living+sqft\_living15, data = new\_DF)



4a.

m1 <- lm(price~sqft\_above+sqft\_living+sqft\_living15+floors+condition , data = new\_DF)

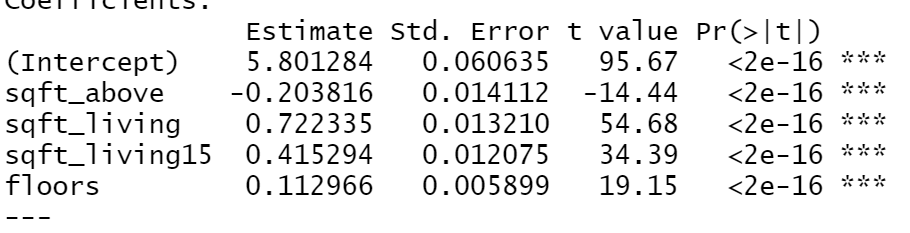
summary(m1)



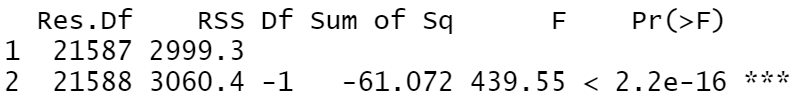
b) Không loại bỏ được biến nào

c)m2 <- lm(price~sqft\_above+sqft\_living+sqft\_living15+floors , data = new\_DF)

summary(m2)



anova(m1, m2)



Mô hình m1 hợp lý hơn

d) Với sqft\_above giảm 0.178957 thì price tăng 1

Với sqft\_living tăng 0.686935 thì price tăng 1

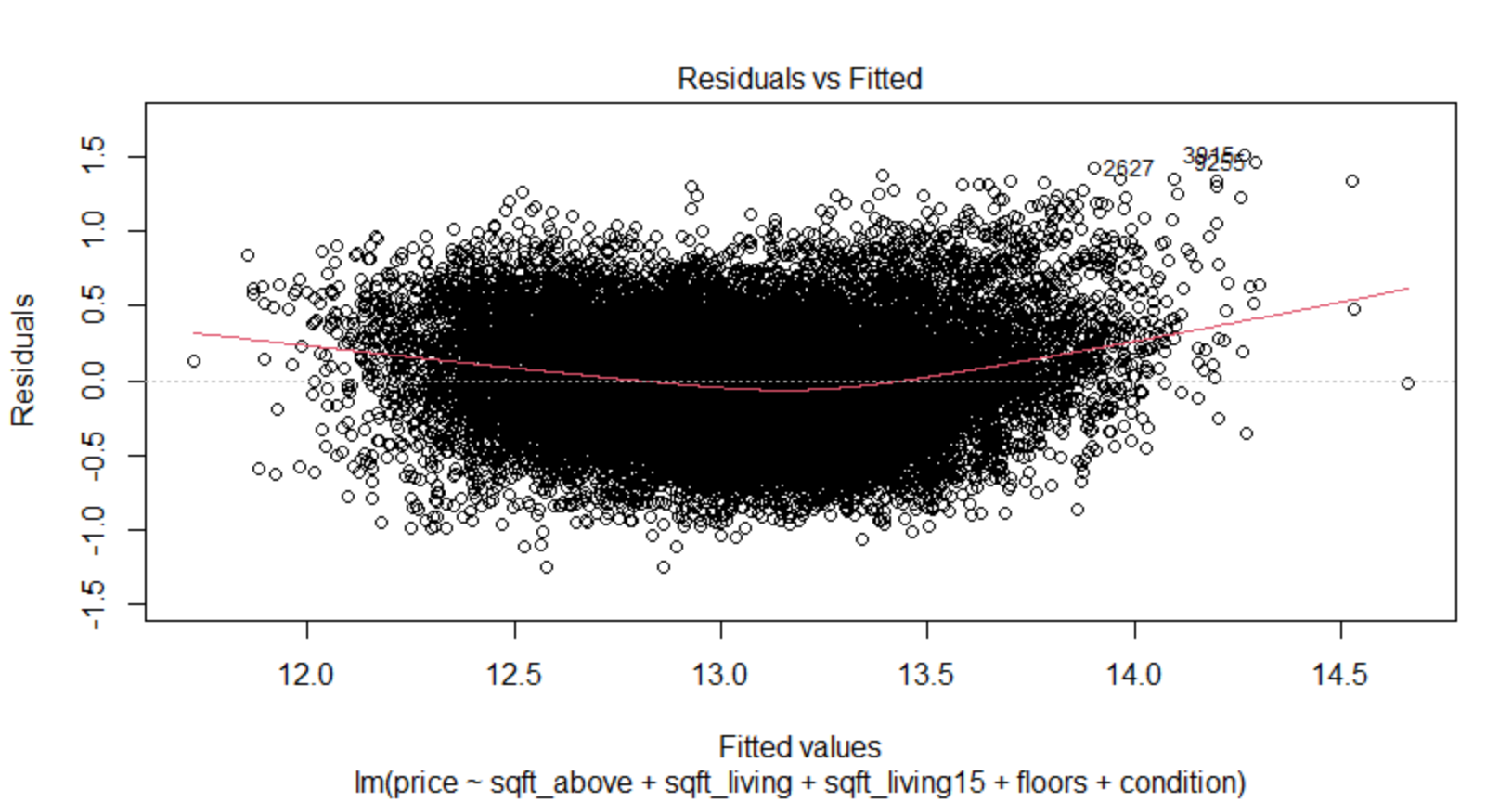
Với sqft\_living15 tăng 0.430556 thì price tăng 1

Với floors tăng 0.137069 thì price tăng 1

Với conditioin tăng 0.085465 thì price tăng 1

e)

plot(m1, which = 1)

Ý nghĩa:

Dùng để xét tính tuyến tính của mô hình. Trong đồ thì nó được chỉ ra bởi đường màu đỏ có gần đường gạch ngang không.

Dùng để kiểm tra phương sai không đổi. Sự chênh lệch của residual phải xấp xỉ như nhau trên trục x.

Kiểm tra các điểm outlier. Với các điểm outlier có residual nằm xa các điểm còn lại.

Nhận xét:

Ở đây ta thấy tính tuyến tính dường như bị vi phạm. Có phương sai không đổi và vài điểm outlier ở residual xấp sỉ 1,5.

5.

x1\_data = data.frame(sqft\_living15 = mean(new\_DF$sqft\_living15),

sqft\_above = mean(new\_DF$sqft\_above),

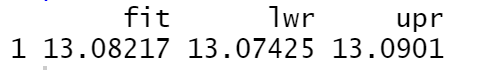
sqft\_living= mean(new\_DF$sqft\_living),

floors = 2,

condition = 3

)

predict(m1,x1\_data,interval = "confidence")



X1 có khoảng tin cậy 13.07425-13.0901 = 0.01585

x2\_data = data.frame(sqft\_living15 = max(new\_DF$sqft\_living15),

sqft\_above =max(new\_DF$sqft\_above),

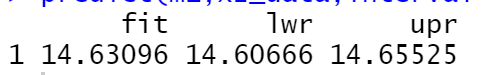
sqft\_living = max(new\_DF$sqft\_living),

floors = 2,

condition = 3

)

predict(m1,x2\_data,interval = "confidence")



X2 có khoảng tin cậy 14.60666-14.65525=0.04859

X2 có khoảng tin cậy lớn hơn.