DATA 2:

Chúng ta quan tâm đến bộ dữ ghi lại lịch sử những ngôi nhà được bán từ 5/2014 đến 5/2015 ở quận King bang Washington.

Price

Bedrooms

Bathrooms

sqft\_living

sqft\_lot

floors

waterfront

view

condition

grade

sqft\_above

sqft\_basement

yr\_built

yr\_renovated

zipcode

lat+long

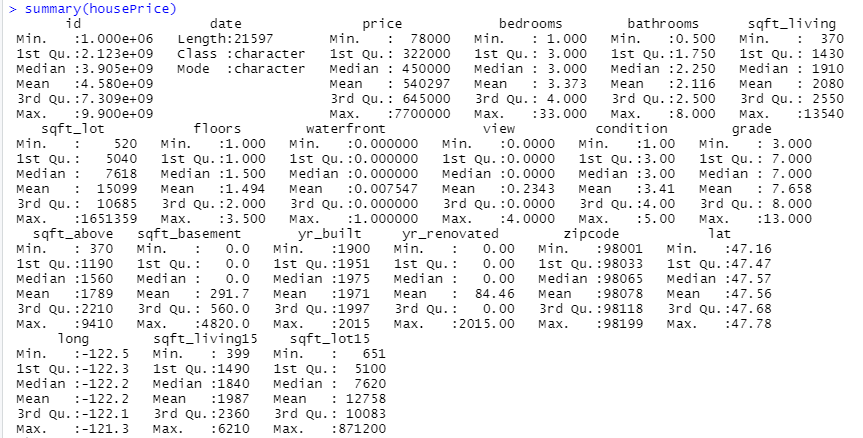
sqft\_living15

sqft\_lot15

Các biến như sau:

Nhập dữ liệu vào R, xem xét tổng quan về bộ dữ liệu

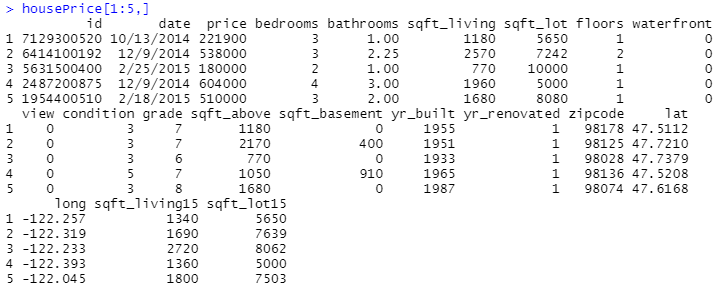
|  |
| --- |
| housePrice <- read.csv('data2.csv', header = TRUE)  summary(housePrice)  dim(housePrice)  housePrice[1:5,] |

Tóm tắt cơ bản các biến:

Kích thước bộ dữ liệu:



Số lượng quan trắc lớn gấp hơn 1000 lần số lượng biến. Vấn đề quá khớp sẽ là vấn đề cần chú ý hơn khi xây dựng mô hình cho bộ dữ liệu giá nhà này.



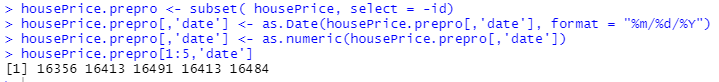
Nhận xét tổng quan:

ID và ngày là các biến sử dụng để theo lưu trữ, truy xuất dữ liệu là chính. Ta có thể loại biến này khỏi mô hình ngay từ đầu.

|  |
| --- |
| housePrice.prepro <- subset( housePrice, select = -id) |

Biến date có thể cũng sẽ có tương quan đến giá nhà, nhưng đang ở dạng biến chuỗi kí tự, cần biến đổi về dạng số để có thể sử dụng trong mô hình.

|  |
| --- |
| housePrice.prepro[,'date'] <- as.Date(housePrice.prepro[,'date'], format = "%m/%d/%Y")  housePrice.prepro[,'date'] <- as.numeric(housePrice.prepro[,'date'])  housePrice.prepro[1:5,'date'] |

Dữ liệu ngày bán ở dạng số: 

Price chính là biến phụ thuộc trong mô hình hồi quy,. Biểu đồ phân tán giữa price và các biến còn lại như sau:

|  |
| --- |
| pairs(price~., data = housePrice.prepro) |

price có giá trị quá lớn so với các biến độc lập, và có dấu hiệu của data bị lệch ở đây. Thực hiện LOG transform cho price trước khi xây dựng mô hình

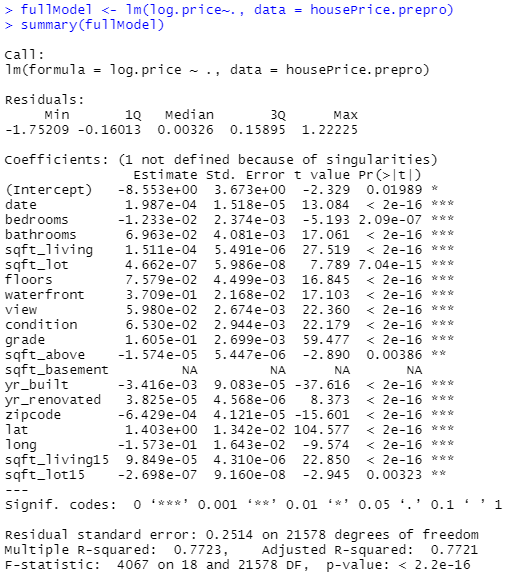
|  |
| --- |
| attach(housePrice.prepro)  housePrice.prepro[["price"]] <- log(price)  colnames(housePrice.prepro)[colnames(housePrice.prepro) == "price"] <- "log.price" |

|  |
| --- |
|  |

Thực hiện phép hồi quy tuyến tính trên toàn bộ data đã được tiền xử lý này:

|  |
| --- |
| fullModel <- lm(log.price~., data = housePrice.prepro)  summary(fullModel) |

Ta có kết quả trên R:



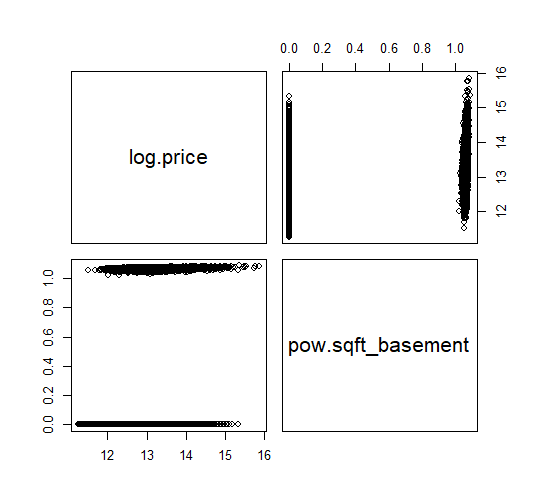
Đa số các biến đều có ý nghĩa thống kê ngoại trừ biến sqft\_basement diện tích tầng hầm, R bỏ qua luôn ảnh hưởng của biến này trong mô hình, vấn đề có thể là sự phụ thuộc tuyến tính với các biến khác. Thực tế thì có rất nhiều biến diện tích khác được thu thập trong bộ dữ liệu cho nên việc phụ thuộc tuyến tính giữa chúng là hoàn toàn có thể xảy ra.

Với biến ta có 2 phương án xử lý như sau:

1. Bỏ qua biến này trong mô hình hồi quy tuyến tính.
2. Xem xét sử dụng biến một phép biến đổi để loại bỏ sự phụ thuộc tuyến tính.

Ta có một nhận xét đối với biến sqft\_basement đó là Min = Median = Q1 = 0, một số lượng lớn quan trắc nhận giá trị 0, đối với data này ta có thể dùng phép lũy thừa với để có được dạng biến tương tự gần tương tự dummy.

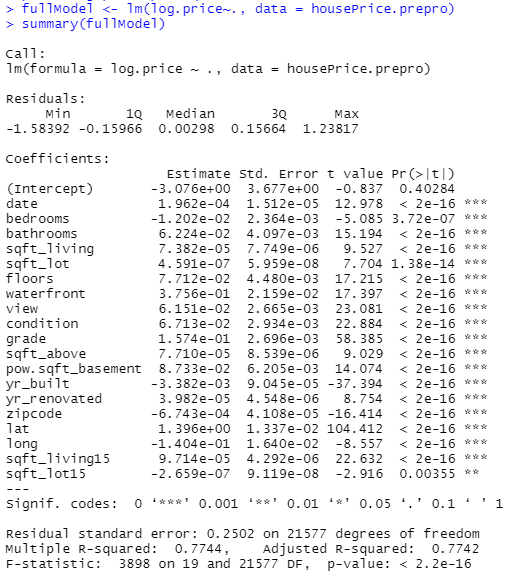
|  |
| --- |
| housePrice.prepro[["sqft\_basement"]] <- sqft\_basement\*\*0.01  colnames(housePrice.prepro)[colnames(housePrice.prepro) == "sqft\_basement"] <- "pow.sqft\_basement"  hist(housePrice.prepro$pow.sqft\_basement)  pairs(log.price~pow.sqft\_basement) |



Xây dựng mô hình trên toàn bộ biến sau transform:

|  |
| --- |
|  |

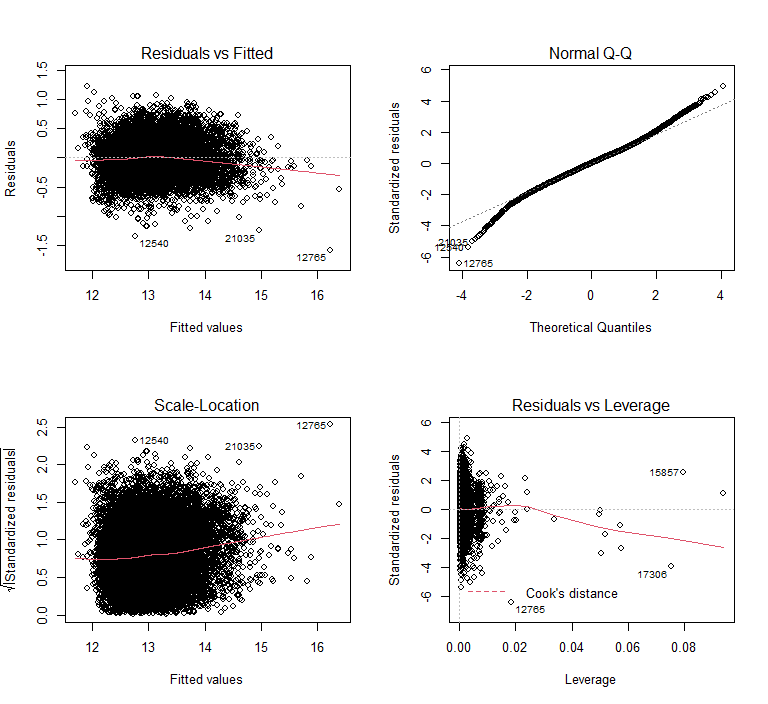
Kết quả:



Mô hình hiện tại có 19 biến đều có ý nghĩa thống kê rất cao, giải thích được 77,4% biến thiên của giá nhà ở quận King.

Kiểm tra tính tồn tại của mô hình bằng các biểu đồ phần dư như sau:

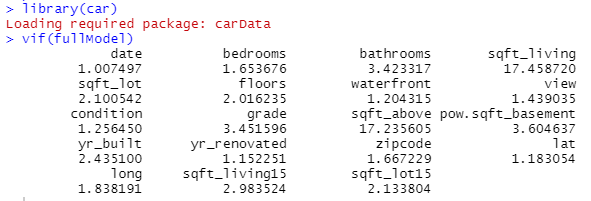
|  |
| --- |
| par(mfrow=c(2,2))  plot(fullModel2) |



Từ biểu đồ ta thấy được phân phối chuẩn của phần dư, giá trị trung bình cũng bám sát giá trị 0 cho ta bằng chứng để có thể chấp nhận sự tồn tại của mô hình.

Tiếp theo, kiểm tra sự đa cộng tuyến trong mô hình ta dùng hệ số VIF như sau:

|  |
| --- |
| library(car)  vif(fullModel) |



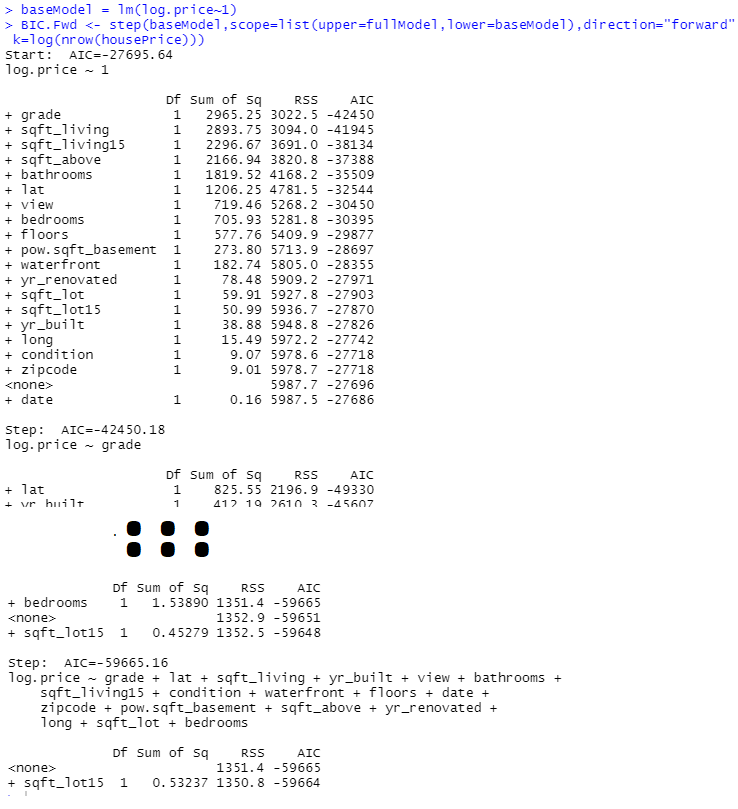
Có sqft\_living và sqft\_above có hiện tương đa cộng tuyến rất nặng, lớn hơn giá trị giới hạn của VIF là 5 rất nhiều. Giả thiết về sự tồn tại của mô hình bị phá vỡ. Có thể dễ thấy là giữa có thông số về diện tích như diện tích phòng khách, diện tích tầng hầm, diện tích các phần phía trên trừ hầm, ... có tương quan với nhau rất cao và hiện tương đa cộng tuyến thực sự rất dễ xảy ra. Thực tế diện tích tầng hậm bị R loại bỏ giúp ta nhận ra sớm và xử lý nó

Tuy nhiên, trong tình huống này, vừa mong muốn loại bỏ hiện tượng đa cộng tuyến giúp giảm phương sai của sai số dự đoán, ta cũng mong muốn xây dựng một mô hình đơn giản hơn mà vẫn đủ khả năng giải thích giá nhà, mô hình hiện tại sử dụng tới 21 biến cộng thêm số quan trắc lớn sẽ rất dễ bị quá khớp bộ dự liệu này , vì vậy bước tiếp theo ta sẽ thực hiện là chọn mô hình với số biến phù hợp hơn.

Với số lượng biến tương đối lớn và mục tiêu xây dựng mô hình hồi quy đơn giản, ít biến hơn. Trong bước chọn mô hình này ta lựa chọn phương pháp stepwise với tiêu chuẩn chọn là BIC.

|  |
| --- |
| baseModel = lm(log.price~1)  BIC.Fwd <- step(baseModel,scope=list(upper=fullModel,lower=baseModel),direction="forward", k=log(nrow(housePrice))) |

Quá trình chọn mô hình:

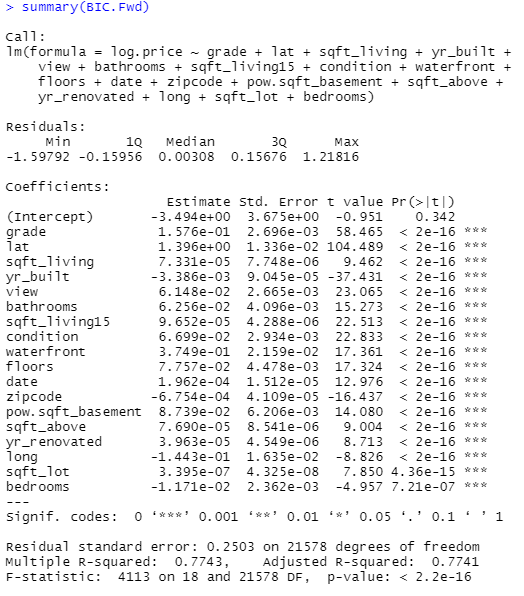


Bây giờ ta xem xét mô hình được chọn bởi thuật toán stepwise:

|  |
| --- |
| log.price ~ grade + lat + sqft\_living + yr\_built + view + bathrooms + sqft\_living15 + condition + waterfront + floors + date + zipcode + pow.sqft\_basement + sqft\_above + yr\_renovated + long + sqft\_lot + bedrooms |

|  |
| --- |
| summary(BIC.Fwd) |

Kết quả cho thấy với 18 biến được chọn đều có ý nghĩa thống kê và mô hìnhvẫn giải thích được tới 77.4%. Đây tiếp tục là dấu hiệu của đa cộng tuyến.



Ta thấy những biến được chọn vẫn còn những biến về diện tích bị đa cộng tuyến đã kiểm tra ở trên đó là sqft\_living và sqft\_above

Đến đây ta thực hiện kiểm định Fisher từng phần cho việc loại bỏ luôn 2 biến này ra khỏi mô hình:

Giả thuyết:

H0: Mô hình bỏ 2 biến sqft\_living và sqft\_above

|  |
| --- |
| BIC.FwdH0 <- lm(log.price ~ grade + lat + yr\_built + view + bathrooms +  sqft\_living15 + condition + waterfront + floors + date +  zipcode + pow.sqft\_basement + yr\_renovated +  long + sqft\_lot) |

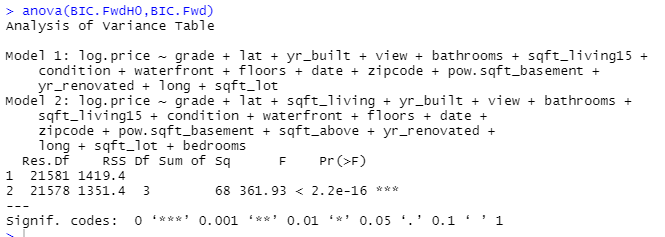
H1: Mô hình đã được chọn với bằng step wise

|  |
| --- |
| BIC.Fwd <- log.price ~ grade + lat + sqft\_living + yr\_built + view + bathrooms + sqft\_living15 + condition + waterfront + floors + date + zipcode + pow.sqft\_basement + sqft\_above + yr\_renovated + long + sqft\_lot + bedrooms |

Sử dụng R để kiểm định:

|  |
| --- |
| anova(BIC.FwdH0,BIC.Fwd) |

Giá trị thống kê của kiểm định: Fobs = 361.93, p-value <2.2e-16



Như vậy ta bác bỏ giả thuyết H0, điều này có nghĩa ta không đủ cơ sở để loại biến đa cộng tuyến trong mô hình.

…..