DATA2:

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.

Mô tả các biến trong bảng dữ liệu như sau:

|  |  |  |
| --- | --- | --- |
| 1 | id | id của một ngôi nhà |
| 2 | date | ngày nhà được bán |
| 3 | price | giá nhà |
| 4 | bedrooms | số phòng ngủ |
| 5 | bathrooms | số phòng tắm |
| 6 | sqft\_living | diện tích ngôi nhà |
| 7 | sqft\_lot | diện tích lô đất |
| 8 | floors | tổng số tầng |
| 9 | waterfront | hướng nhà bờ sông |
| 10 | view | số lượt xem công ty bất động sản thu được |
| 11 | condition | đánh giá tình trạng của nhà (1-cũ, ..., 6-tuyệt vời) |
| 12 | grade | điểm tổng thể cho nhà ở (theo phân loại quận King: 1-kém, ..., 13-xuất sắc |
| 13 | sqft\_above | diện tích ngôi nhà ngoài tầng hầm |
| 14 | sqft\_basement | diện tích tầng hầm |
| 15 | yr\_built | năm xây dựng |
| 16 | yr\_renovated | năm ngôi nhà được cải tạo |
| 17 | zipcode | mã vùng |
| 18 | lat | vĩ độ |
| 19 | long | kinh độ |
| 20 | sqft\_living15 | diện tích ngôi nhà năm 2015 |
| 21 | sqft\_lot15 | diện tích lô đất năm 2015 |

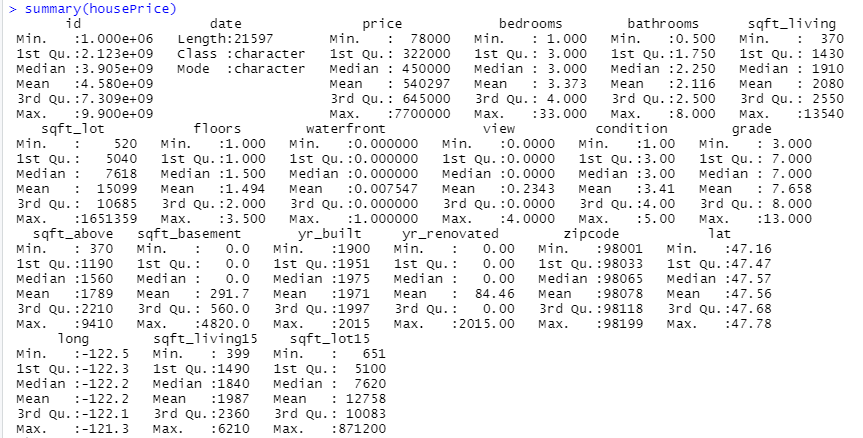
Nhập bộ dữ liệu vào R để và xem xét tổng quan về bộ dữ liệu

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

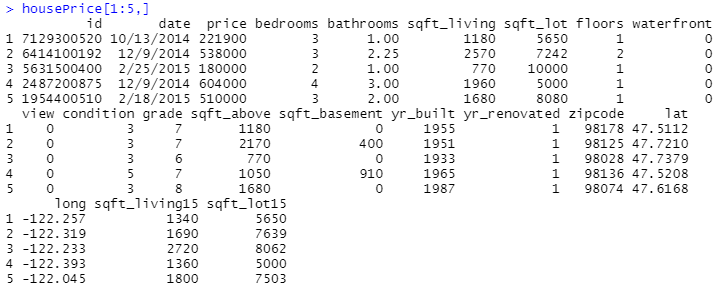
Kích thước bộ dữ liệu:



Số lượng quan trắc có được là lớn, thiếu dữ liệu phân tích không còn là vấn đề trong xây dựng mô hình cho bô dữ liệu. Cụ thể thì số quan sát lớn gấp hơn 1000 lần số lượng biến. Với kích thước như vậy nhiều vấn đề khác như quá khớp cũng sẽ được giải quyết, lựa chọn phương pháp tiếp cận trong quá trình xây dựng mô hình cũng sẽ trở nên đa dạng.

Tóm tắt cơ bản các biến:

Quan sát cụ thể cho 5 mẫu đầu tiên:



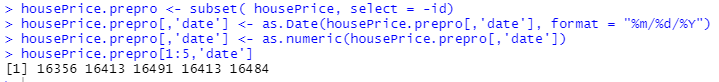
Dựa trên thông số thống kê mô tả kết hợp các hiểu biết của cá nhân người phân tích, ta có một số điểm cần lưu ý như sau:

1. ID là 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) |

1. Biến date, thời gian hoàn toàn có khả năng có có tương quan đến giá nhà, nhưng đang ở dạng biến chuỗi kí tự, ta sẽ xử lý 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'] |

Kết quả dữ liệu ngày bán ở dạng số: 

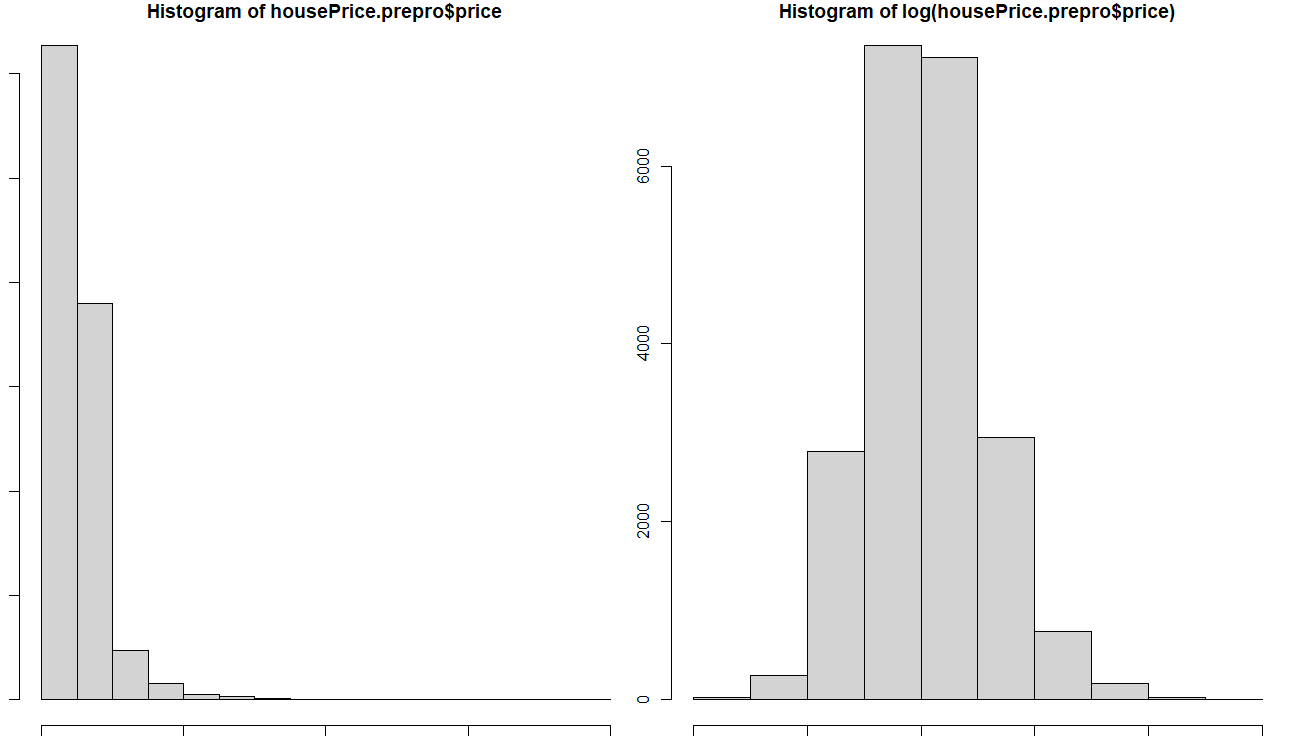
1. Price chính là biến phụ thuộc trong mô hình hồi quy. Biểu đồ phân tán giữa price và các biến phụ thuộc, có thể được biểu diễn trên R với lệnh:

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

Tuy nhiên số lượng quan trắc là rất lớn, việc biểu diễn scatterplot dường như không hiệu quả, cần máy tính có khả tính toán lớn để trực quan hóa cũng như sẽ mất quá nhiều thời gian.

Thống kê mô tả cho thấy giá nhà - price có khoảng giá trị khá lớn và có dấu hiệu của bị lệch ở đây. Xử lý log transform cho price trước khi xây dựng mô hình sẽ là giải pháp hợp lý để xử lý vấn đề này. Có thể kiểm tra bằng histogram:

|  |
| --- |
| par(mfrow=c(1,2))  hist(housePrice.prepro$price)  hist(log(housePrice.prepro$price)) |



Kết quả cho thấy log transform cho biến phụ thuộc – giá nhà là hợp lý. Log có dạng phân phối chuẩn khá tốt.

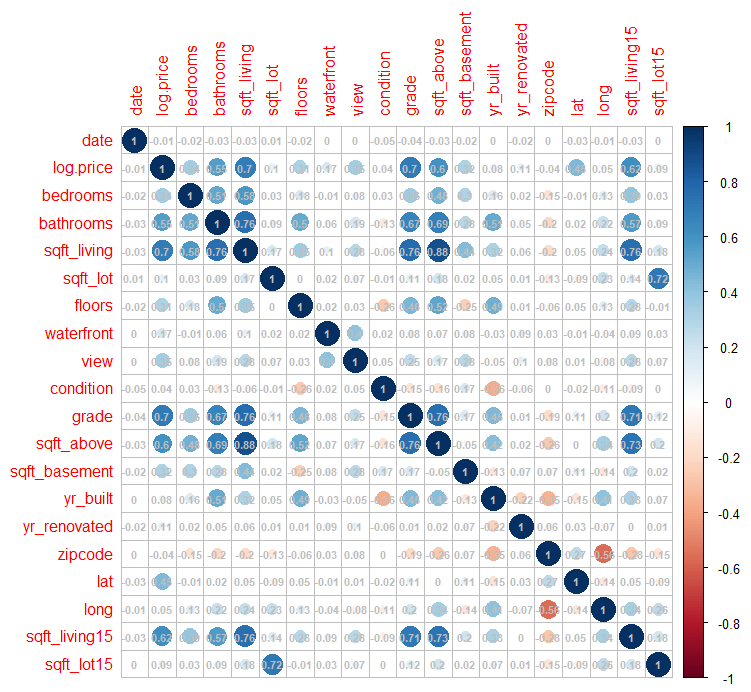
|  |
| --- |
| attach(housePrice.prepro)  housePrice.prepro[["price"]] <- log(price)  colnames(housePrice.prepro)[colnames(housePrice.prepro) == "price"] <- "log.price" |

1. Có rất nhiều biến liên quan đến diện tích nhà được sử dụng, đặc biệt là các biến:

* sqft\_living, sqft\_living15: diện tích nhà và diện tích nhà năm 2015
* sqft\_lot, sqft\_lot15: diện tích lô đất và diện tích lô đất năm 2015
* sqft\_above, sqft\_living: diện tích ngôi nhà và diện tích trừ tầng hầm
* một số biến khác...

Những tương quan trên các cặp biến này chắc chắn sẽ rất cao, nguy cơ của đa cộng tuyến sẽ ảnh hưởng rất lớn đến mô hình hồi quy đa biến, đó là giảm khả năng dự đoán, phương sai cao, thậm chí là tính tồn tại của mô hình.

|  |
| --- |
| round(cor(housePrice.prepro), 2)  corrplot::corrplot(cor(housePrice.prepro), addCoef.col = "grey", number.cex=0.7) |



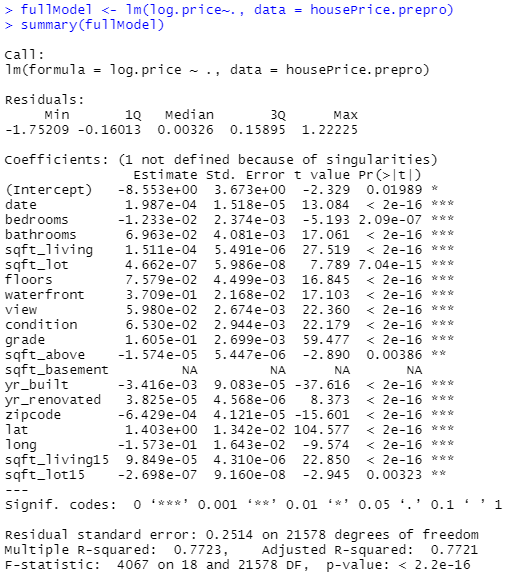
1. Mô hình sẽ có một số các biến Dummy, có thể cần được chú ý trong quá trình xây dựng mô hình.

Hiện tại ta đã có được bộ dữ liệu với một số tiền xử lý đơn giản được lưu trong data frame housePrice.prepro

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:



Mô hình thu được:

|  |
| --- |
| log.price ~ date + 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 |

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 sqft\_basement ta có 2 phương án xử lý như sau:**

I/ Bỏ qua biến này trong mô hình hồi quy tuyến tính. Tiếp tục sử dụng mô hình fullModel này để phân tích.

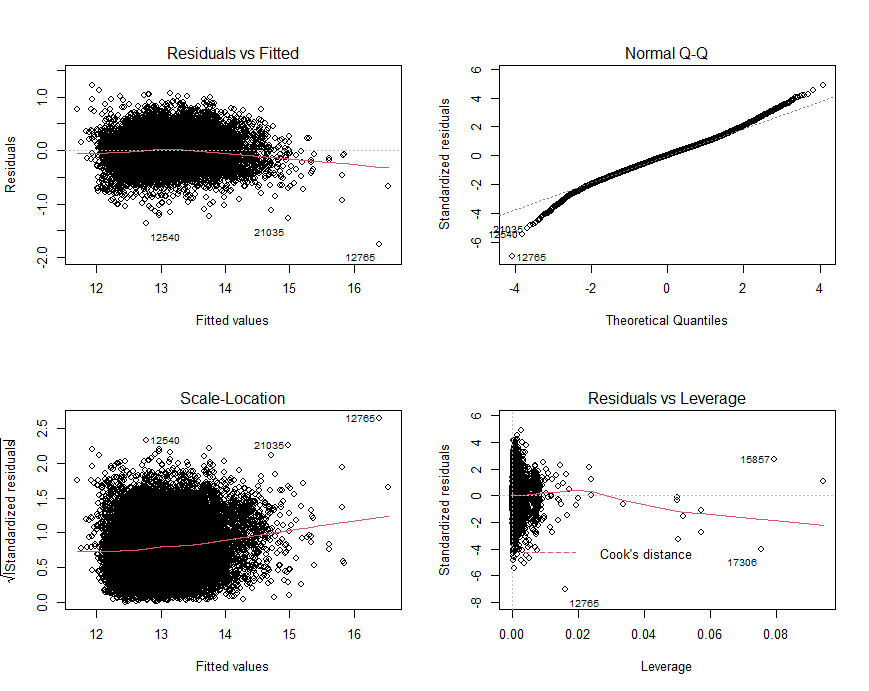
Toàn hệ số hồi quy đều có ý nghĩa thống kê cao, mô hình giải thích được khoảng 77.2% biến phụ thuộc.

Kiểm tra tính tồn tại của mô hình này:

|  |  |
| --- | --- |
| par(mfrow=c(1,1))  plot(fullModel$fitted.values,  log.price,xlab="Fitted Values")  abline(lsfit(fullModel$fitted.values,  log.price), col = "red") |  |

Mối quan hệ tuyến tuyến trên biểu đồ giá trị fit và giá trị thực tế cho thấy sự hợp lý của mô hình.

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

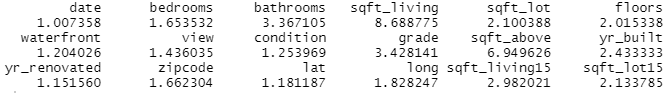


Biểu đồ diagnostic cho thấy thêm những bằng chứng cho sự tồn tại của mô hình này. Đó là phân phối chuẩn của phần dư, trung bình phần dư bằng 0, phương sai không thay đổi với giá trị dao động ở khoảng 0.8.

Tuy nhiên khi xem xét hệ số VIF của mô hình:

|  |
| --- |
| car::vif(lm(log.price ~ date + bedrooms + bathrooms + sqft\_living + sqft\_lot + floors + waterfront + view + condition + grade + sqft\_above + yr\_built + yr\_renovated + zipcode + lat + long + sqft\_living15 + sqft\_lot15)) |

Hiện tượng đa cộng tuyến xảy ra như dự đoán ban đầu:



Nếu chấp nhận sử dụng 18 biến này để dự đoán thì vấn đề đa cộng tuyến sẽ làm giảm khả năng dự đoán của mô hình. Ngoài ra thì kết hợp với số lượng biến quá lớn, hiện tượng đa cộng tuyến càng làm cho ý nghĩa từng biến đến giá trị dự đoán trở nên không chính xác do các biến đa cộng tuyến tương tác làm dấu đi mối quan hệ này. Bỏ đi biến sqft\_living diện tích căn nhà sẽ giúp các nhân tố lạm phát phương sai VIF giảm về dưới 5, tuy nhiên rõ ràng diện tích ngôi nhà sẽ có mốt quan hệ tương quan cao với giá nhà, ta không thể tùy tiện bỏ biến như vậy.

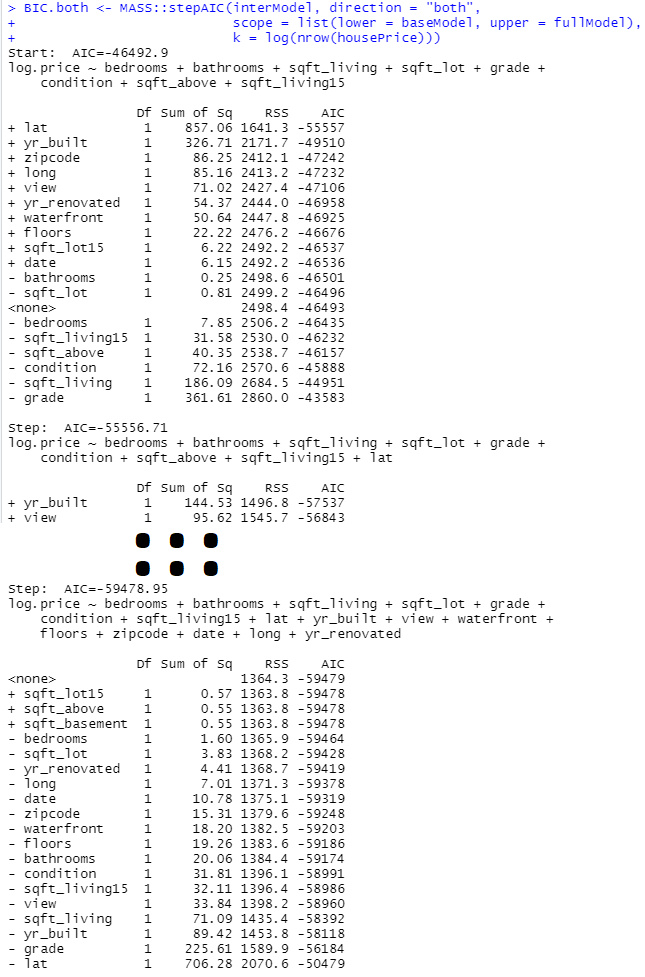
Do đó, bước tiếp theo để xác định mô hình hồi quy tuyến tính cho bộ dữ liệu sẽ là: Chọn mô hình hay chọn biến.

Như nhận xét ban đầu, nhiều biến độc lập đang tồn tại hiện tượng đa cộng tuyến. Hơn nữa ta mong muốn xây dựng mô hình đơn giản nhưng đủ tốt để giải thích giá nhà trên bộ dữ liệu khá nhiều biến, có nhiều quan trắc. Trong trường hợp này BIC sẽ là tiêu chuẩn chọn hợp lý hơn bởi 2 lý do: 1) BIC sẽ có thành phần phạt cao hơn AIC đối với mô hình phức tạp hơn và 2) số lượng quan trắc là lớn hơn nhiều so với số biến.

Phương pháp tiếp cận được lựa chọn là stepwise bởi với tổng số biến là 19 biến độc lập có thể có thì việc chọn bestsubset là không khả thi. Sử dụng both – cả 2 hướng thêm bớt biến từ một mô hình trung bình ở giữa. Nguyên nhân là từ cảm quan cá nhân của người làm thống kê mô hình, sẽ có một số biến như diện tích nhà, diện tích lô đất, điểm tình trạng căn nhà, điểm đánh giá tổng quát căn nhà, số phòng ngủ sẽ là những biến có thể có ảnh hưởng lớn nhất đến giá bán nhà. Kết hợp với đồ thị tương quan ở trên ta chọn mô hình trung bình và thực hiện chọn biến như sau:

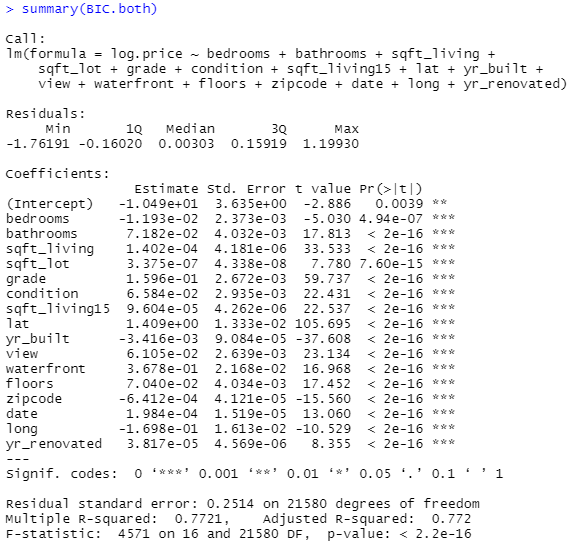
|  |
| --- |
| fullModel <- lm(log.price~., data = housePrice.prepro)  interModel <-lm(log.price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot + grade + condition + sqft\_above + sqft\_living15)  BIC.both <- MASS::stepAIC(interModel, direction = "both",  scope = list(lower = baseModel, upper = fullModel),  k = log(nrow(housePrice))) |

Kết quả chọn mô hình:



Mô hình cuối cùng chọn được bằng stepwise sử dụng tiêu chuẩn BIC theo 2 hướng như sau:

|  |
| --- |
| summary(BIC.both) |



Toàn bộ hệ số hồi quy đều rất có ý nghĩa thống kê, mô hình vẫn giải thích được 77.2% giá nhà, tương đương với mô hình full ban đầu.

Tiếp tục đánh giá tính tồn tại của mô hình được chọn:

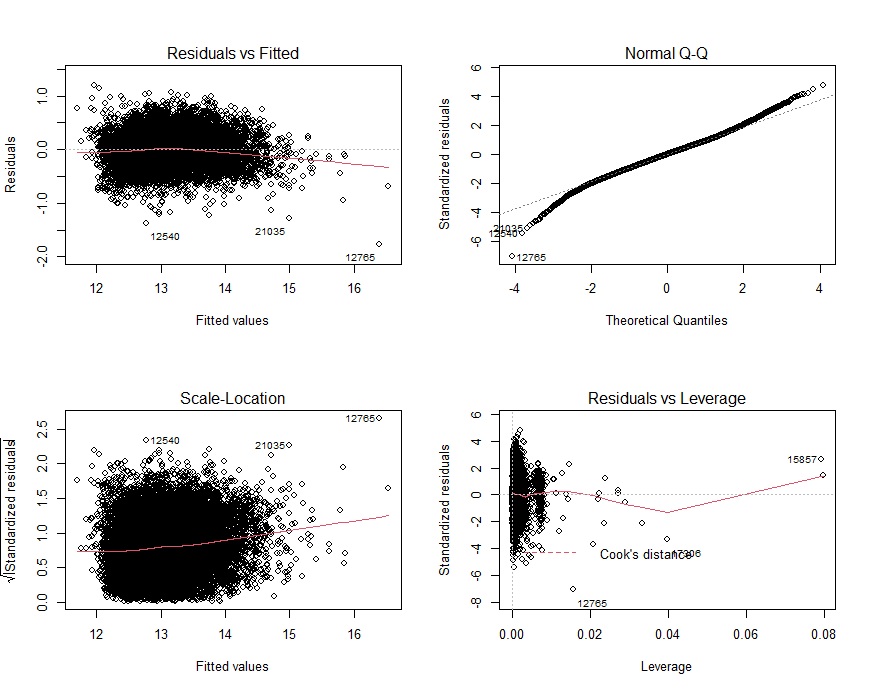
|  |
| --- |
| par(mfrow=c(1,1))  plot(BIC.both$fitted.values,log.price,xlab="Fitted Values")  abline(lsfit(BIC.both$fitted.values,log.price), col = "red")  par(mfrow=c(2,2))  plot(BIC.both)  vif(BIC.both) |

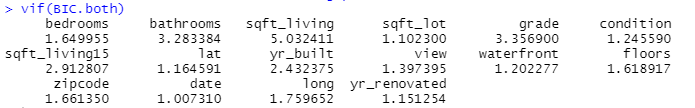
Các biểu đồ đánh giá vẫn tương tự mô hình full đã đánh giá ở trước:

Bên trái là biểu đồ, mối quan hệ tuyến tuyến trên biểu đồ giá trị fit và giá trị thực tế cho thấy sự hợp lý của mô hình chọn. Bên phải là Histogram của phần dư ở dạng phân phối chuẩn.

|  |  |
| --- | --- |
|  |  |

Biểu đồ diagnostic cho thấy thêm những bằng chứng khẳng định cho sự tồn tại của mô hình. Đó là phân phối chuẩn của phần dư, với trung bình là 0, phương sai không thay đổi với giá trị dao động ở khoảng từ 0.8 đến 1, số lượng điểm đòn bẩy không đáng kể so với số quan trắc lên đến hơn 21000.





Một điều cần quan tâm hơn đó là hệ số VIF, và kết quả thu được rất tốt, chỉ có sqft\_living có VIF lớn nhất khoảng 5, giá trị hoàn toàn chấp nhận được. Ta có được mô hình đầu tiên giải thích cho bộ dữ liệu như sau:

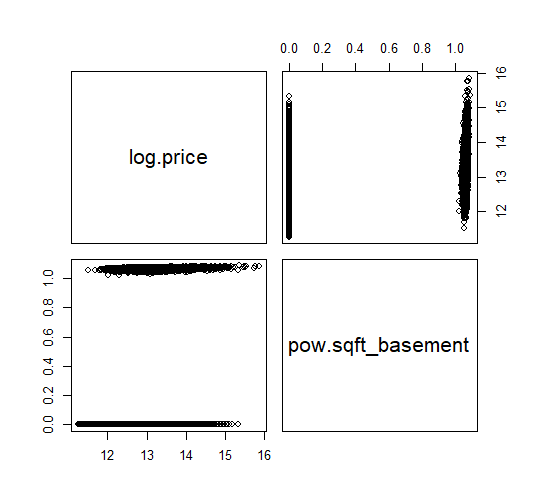
|  |
| --- |
|  |

Mô hình này là một hình sử dụng để dự đoán giá nhà tốt bởi vì toàn bộ hệ số hồi quy đều có ý nghĩa thống kê. Kiểm tra tính tồn tại mô hình cho kết quả tốt ở toàn các phương diện: tính tuyến tính, tính đồng nhất của phương sai. Đặc biệt là phân phối chuẩn của sai số như biểu đồ ở trên, và cuối cùng mô hình cũng giải quyết được vấn đề đa cộng tuyến.

II/ 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.

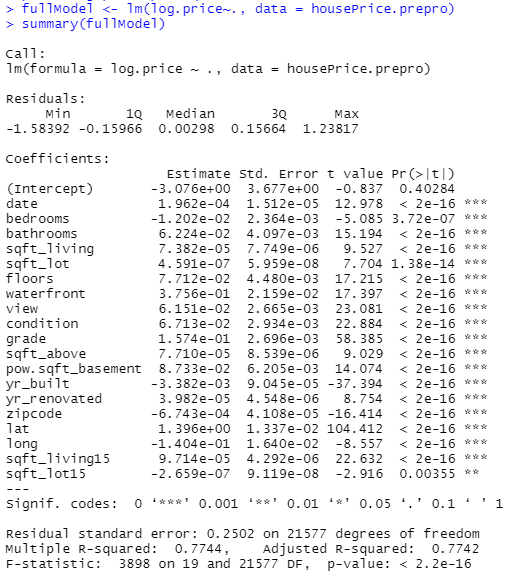
|  |
| --- |
| housePrice.prepro[["sqft\_basement"]] <- sqft\_basement\*\*0.01  colnames(housePrice.prepro)[colnames(housePrice.prepro) == "sqft\_basement"] <- "pow.sqft\_basement"  hist(housePrice.prepro$pow.sqft\_basement)  pairs(log.price~pow.sqft\_basement) |



Xây dựng mô hình trên toàn bộ biến sau transform:

|  |
| --- |
|  |

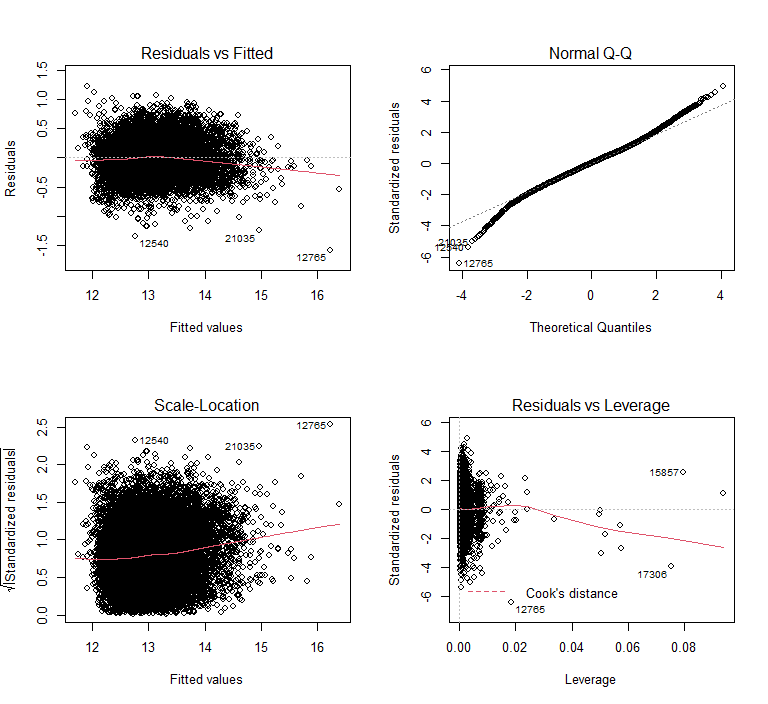
Kết quả:



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:

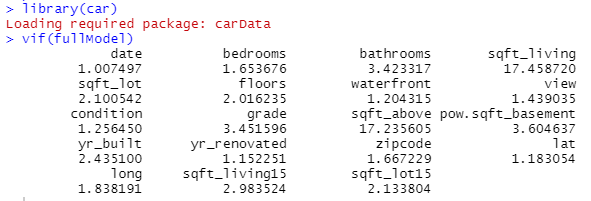
|  |
| --- |
| par(mfrow=c(2,2))  plot(fullModel2) |



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) |



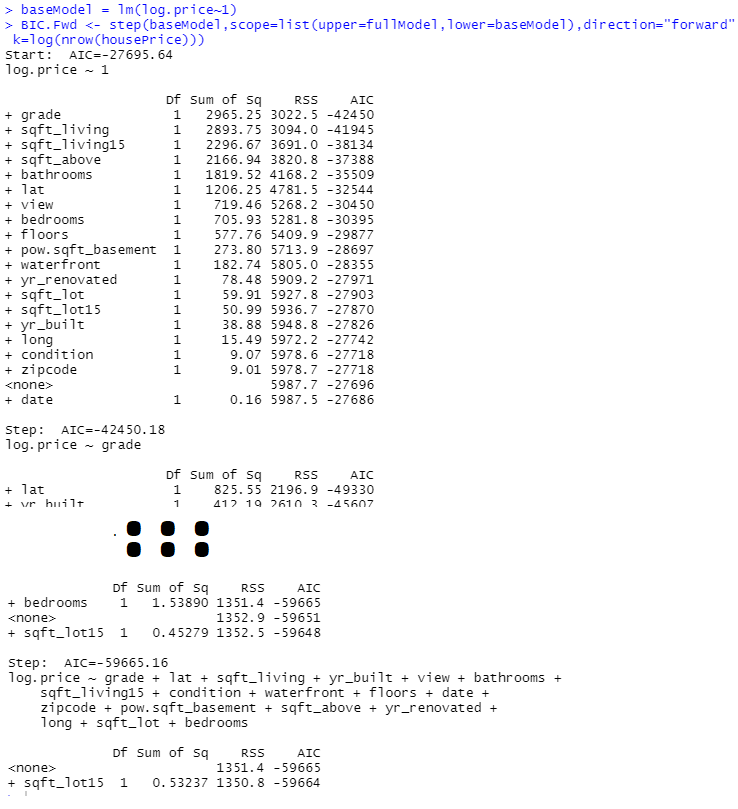
Có sqft\_living và sqft\_above có hiện tương đa cộng tuyến rất nặng, lớn hơn giá trị giới hạn của VIF là 5 rất nhiều. Giả thiết về sự tồn tại của mô hình bị phá vỡ. Có thể dễ thấy là giữa có thông số về diện tích như diện tích phòng khách, diện tích tầng hầm, diện tích các phần phía trên trừ hầm, ... 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:

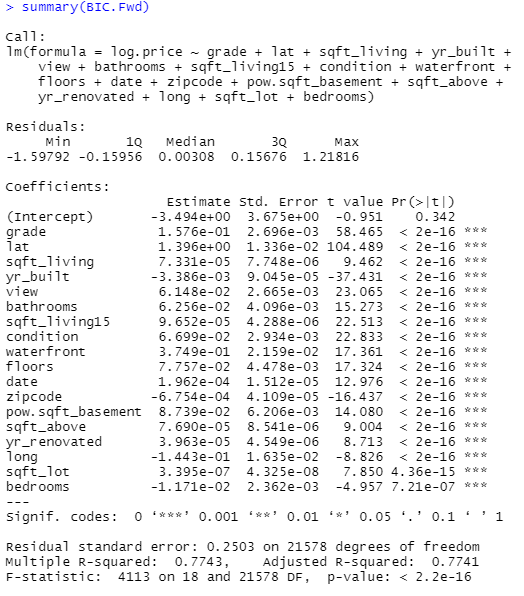


Bây giờ ta xem xét mô hình được chọn bởi thuật toán stepwise:

|  |
| --- |
| 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) |

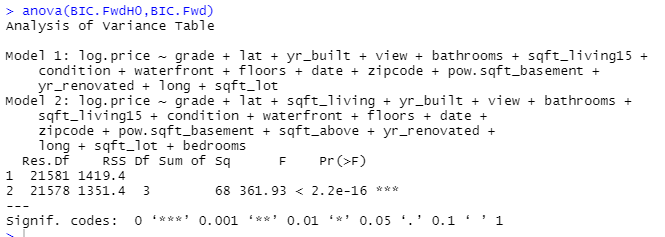
H1: Mô hình đã được chọn với bằng step wise

|  |
| --- |
| 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.

…..