ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

TIỂU LUẬN CUỐI KÌ

MÔ HÌNH HÓA THỐNG KÊ

Giảng viên hướng dẫn: **TS. Nguyễn Thị Mộng Ngọc**

Nhóm thực hiện: **Nhóm 4**

Học viên:	**Phan Thị Thùy An**	MSHV: 20C29002
	Đinh Thị Nữ	MSHV: 20C29013
	Lý Phi Long	MSHV: 20C29028
	Đặng Khánh Thi	MSHV: 20C29038

TP. Hồ Chí Minh – Tháng 04, 2021

Mục lục

Chương 1

Dữ liệu tự chọn

- 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.
- Kết luận.

1.1 Dữ liệu 1: Mô hình hồi quy đa biến

1.2 Dữ liệu 2: Hồi quy thành phần chính

Chương 2

Dữ liệu có sẵn

- Chọn mô hình phù hợp nhất giải thích biến phụ thuộc với từng bộ dữ liệu.
- Nêu rõ phương pháp chọn mô hình và lý do chọn phương pháp đó.
- Nói rõ ý nghĩa của mô hình đã chọn.

2.1 Dữ liệu 1

2.2 Dữ liệu 2

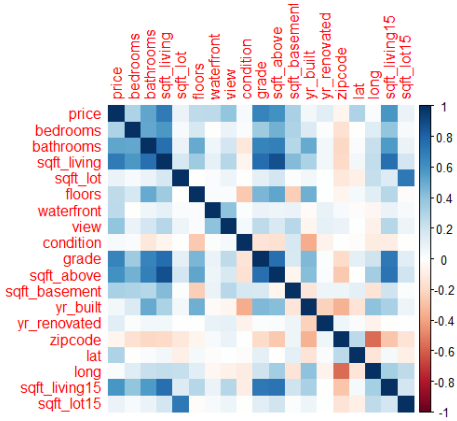
Bộ dữ liệu ghi lại lịch sử về những ngôi nhà được bán từ 5/2014 đến 5/2015 ở quận King, bang Washington, Hoa Kỳ. Bộ dữ liệu bao gồm 21613 quan trắc, gồm 21 biến.

∗ **Phương pháp chọn: Stepwise - lùi; tiêu chuẩn chọn: BIC**.

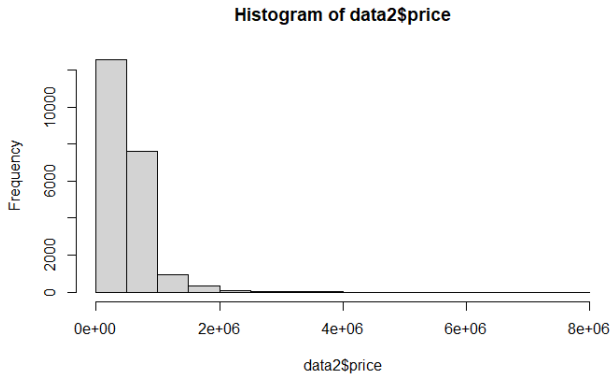
Tìm hiểu dữ liệu

```
> mydata <-read.csv("data2.csv")
> head(mydata)
           id       date   price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
1 7129300520 10/13/2014  221900        3      1.00        1180     5650      1          0    0
2 6414100192  12/9/2014  538000        3      2.25        2570     7242      2          0    0
3 5631500400  2/25/2015  180000        2      1.00         770    10000      1          0    0
4 2487200875  12/9/2014  604000        4      3.00        1960     5000      1          0    0
5 1954400510  2/18/2015  510000        3      2.00        1680     8080      1          0    0
6 7237550310  5/12/2014 1230000        4      4.50        5420   101930      1          0    0
  condition grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat     long
1         3     7       1180             0     1955            0   98178 47.5112 -122.257
2         3     7       2170           400     1951         1991   98125 47.7210 -122.319
3         3     6        770             0     1933            0   98028 47.7379 -122.233
4         5     7       1050           910     1965            0   98136 47.5208 -122.393
5         3     8       1680             0     1987            0   98074 47.6168 -122.045
6         3    11       3890          1530     2001            0   98053 47.6561 -122.005
  sqft_living15 sqft_lot15
1          1340       5650
2          1690       7639
3          2720       8062
4          1360       5000
5          1800       7503
6          4760     101930
```

(a) Một số quan trắc đầu tiên



(b) Hệ số tương quan giữa các biến



(c) Phân bố của biến phụ thuộc

Hình 2.1: Một số quan sát ban đầu của bộ dữ liệu

Bộ dữ liệu cung cấp gồm 21 biến, trong đó biến **id** và **date** được loại bỏ khỏi dữ liệu trước khi tiến hành phân tích, vì nhóm em nghĩ các biến này chỉ để ghi lại chỉ số và thời gian mua bán, không mang nhiều ý nghĩa thống kê.

Quan sát ban đầu cho thấy: các biến độc lập **sqft_living, grade, sqft_above,**

sqft_living15 có mối tương quan cao với biến phụ thuộc **Price**; biến phụ thuộc **Price** phân bố không đều, bị lệch hẳn về một phía và giá trị chủ yếu từ 0 đến 2 000 000.

Phân tích, chọn mô hình

```
> # Create full model
> mod_full_1 = lm(price ~ ., data2) #full model
> summary(mod_full_1)

Call:
lm(formula = price ~ ., data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1291631   -99089    -9569    77778  4330096

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.564e+06  2.933e+06   2.238  0.02523 *
bedrooms      -3.556e+04  1.901e+03 -18.707  < 2e-16 ***
bathrooms      4.128e+04  3.268e+03  12.632  < 2e-16 ***
sqft_living    1.496e+02  4.397e+00  34.033  < 2e-16 ***
sqft_lot       1.289e-01  4.792e-02   2.690  0.00714 **
floors         6.474e+03  3.602e+03   1.797  0.07229 .
waterfront     5.833e+05  1.736e+04  33.593  < 2e-16 ***
view           5.278e+04  2.141e+03  24.652  < 2e-16 ***
condition      2.679e+04  2.353e+03  11.387  < 2e-16 ***
grade          9.701e+04  2.161e+03  44.894  < 2e-16 ***
sqft_above     3.129e+01  4.361e+00   7.174 7.53e-13 ***
sqft_basement         NA         NA      NA       NA
yr_built      -2.628e+03  7.272e+01 -36.135  < 2e-16 ***
yr_renovated   1.983e+01  3.656e+00   5.425 5.87e-08 ***
zipcode       -5.819e+02  3.299e+01 -17.635  < 2e-16 ***
lat            6.022e+05  1.074e+04  56.071  < 2e-16 ***
long          -2.156e+05  1.316e+04 -16.385  < 2e-16 ***
sqft_living15  2.116e+01  3.451e+00   6.131 8.88e-10 ***
sqft_lot15    -3.907e-01  7.334e-02  -5.327 1.01e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 201300 on 21579 degrees of freedom
Multiple R-squared:  0.7001,    Adjusted R-squared:  0.6999
F-statistic:  2964 on 17 and 21579 DF,  p-value: < 2.2e-16
```

Hình 2.2: Mô hình hồi quy đầy đủ ban đầu

Bộ dữ liệu (sau khi loại bỏ id và date) có 18 biến giải thích, do đó nhóm em chọn phương pháp lùi (**stepwise - backward**) cho bộ dữ liệu này. Trong mô hình hồi quy đầy đủ (Hình 2.2), đa số các biến giải thích đều có ý nghĩa thống kê, do đó tiến hành phương pháp lùi (loại biến dần dần) sẽ tiết kiệm thời gian hơn so với các phương pháp còn lại. Tiêu chuẩn BIC có xu hướng chọn các mô hình ít phức tạp hơn so với tiêu chuẩn AIC, đặc biệt khi số lượng quan trắc lớn.

Bằng phương pháp lùi và tiêu chuẩn BIC (Hình 2.3), các biến **sqft_basement, floors, sqft_lot** đã bị loại bỏ khỏi mô hình. Mô hình được chọn có $R^2 = 0.7, R^2_{adj} = 0.6998$, các tham số ước lượng của mô hình đều có ý nghĩa thống kê.

```
> mod_BIC_1$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
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

Final Model:
price ~ bedrooms + bathrooms + sqft_living + waterfront + view +
    condition + grade + sqft_above + yr_built + yr_renovated +
    zipcode + lat + long + sqft_living15 + sqft_lot15


            Step Df     Deviance Resid. Df   Resid. Dev      AIC
1                                    21579 8.739836e+14 527659.9
2 - sqft_basement  0            0     21579 8.739836e+14 527659.9
3         - floors  1 130842954998     21580 8.741144e+14 527653.1
4       - sqft_lot  1 284802633671     21581 8.743992e+14 527650.2
```

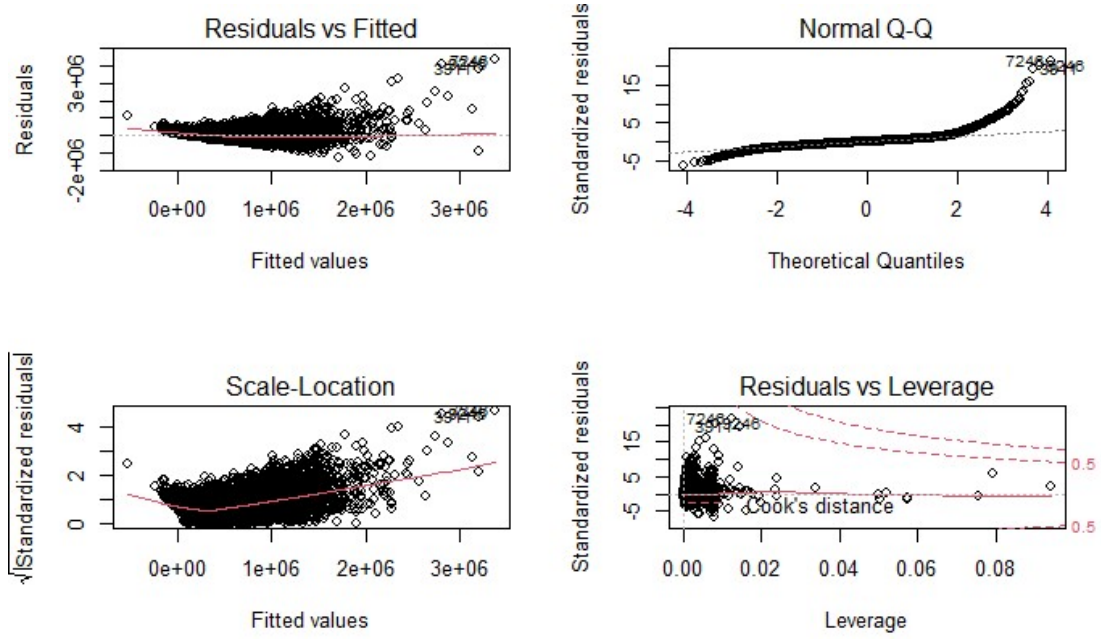
(a) Chọn biến

```
> summary(mod_BIC_1)

Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + waterfront +
    view + condition + grade + sqft_above + yr_built + yr_renovated +
    zipcode + lat + long + sqft_living15 + sqft_lot15, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1284799   -99272    -9674    77773  4326048

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.017e+06  2.885e+06   2.085   0.0371 *
bedrooms      -3.578e+04  1.900e+03 -18.826  < 2e-16 ***
bathrooms      4.285e+04  3.154e+03  13.587  < 2e-16 ***
sqft_living    1.475e+02  4.181e+00  35.280  < 2e-16 ***
waterfront     5.826e+05  1.736e+04  33.554  < 2e-16 ***
view           5.306e+04  2.140e+03  24.797  < 2e-16 ***
condition      2.645e+04  2.349e+03  11.256  < 2e-16 ***
grade          9.746e+04  2.152e+03  45.284  < 2e-16 ***
sqft_above     3.501e+01  3.910e+00   8.952  < 2e-16 ***
yr_built      -2.609e+03  7.094e+01 -36.779  < 2e-16 ***
yr_renovated   2.003e+01  3.651e+00   5.487 4.14e-08 ***
zipcode       -5.764e+02  3.286e+01 -17.542  < 2e-16 ***
lat            6.027e+05  1.070e+04  56.341  < 2e-16 ***
long          -2.152e+05  1.308e+04 -16.451  < 2e-16 ***
sqft_living15  1.989e+01  3.423e+00   5.811 6.30e-09 ***
sqft_lot15    -2.613e-01  5.311e-02  -4.919 8.74e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 201300 on 21581 degrees of freedom
Multiple R-squared:    0.7,     Adjusted R-squared:  0.6998
F-statistic:  3357 on 15 and 21581 DF,  p-value: < 2.2e-16
```

(b) Kết quả mô hình

Hình 2.3: Mô hình khi chọn bằng tiêu chuẩn BIC

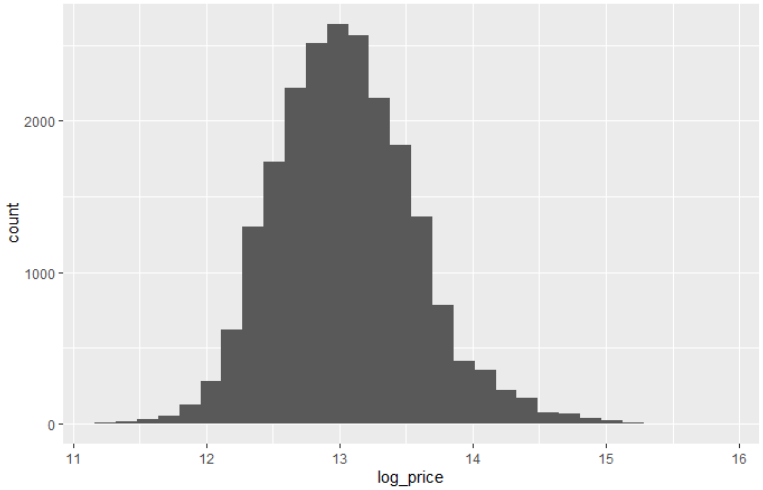
Ta tiến hành kiểm tra xem mô hình này có thỏa mãn các giả thiết của mô hình hồi quy hay không.



Hình 2.4: Các biểu đồ kiểm định mô hình

Dựa vào hình 2.4, phương sai của sai số không phải là hằng số, kì vọng của sai số bằng 0; sai số có vẻ tuân theo phân phối chuẩn nhưng phần đuôi trên bị lệch khá nhiều.

Kết hợp với nhận xét ban đầu, về việc biến **Price** phân bố không đều, nhóm em tiến hành biến đổi biến này thành **log(Price)**.



Hình 2.5: Phân bố của biến **Price** sau khi biến đổi

Sau khi biến đổi, tiến hành hồi quy cho mô hình có 15 biến đã chọn bằng tiêu chuẩn BIC trước đó (**Mô hình 1**) và hồi quy cho mô hình đầy đủ rồi áp dụng tiêu chuẩn BIC (**Mô hình 2**), ta có:

```
> summary(mod_2)

Call:
lm(formula = log(price) ~ bedrooms + bathrooms + sqft_living +
    waterfront + view + condition + grade + sqft_above + yr_built +
    yr_renovated + zipcode + lat + long + sqft_living15 + sqft_lot15,
    data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72685 -0.16385  0.00299  0.16386  1.18219

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.436e+01  3.645e+00  -3.940 8.18e-05 ***
bedrooms      -1.351e-02  2.400e-03  -5.629 1.83e-08 ***
bathrooms      8.720e-02  3.984e-03  21.891  < 2e-16 ***
sqft_living    1.238e-04  5.282e-06  23.444  < 2e-16 ***
waterfront     3.702e-01  2.193e-02  16.881  < 2e-16 ***
view           6.195e-02  2.703e-03  22.919  < 2e-16 ***
condition      5.984e-02  2.968e-03  20.163  < 2e-16 ***
grade          1.643e-01  2.719e-03  60.449  < 2e-16 ***
sqft_above     2.582e-05  4.939e-06   5.228 1.73e-07 ***
yr_built      -3.126e-03  8.960e-05 -34.882  < 2e-16 ***
yr_renovated   4.008e-05  4.612e-06   8.690  < 2e-16 ***
zipcode       -5.816e-04  4.150e-05 -14.014  < 2e-16 ***
lat            1.414e+00  1.351e-02 104.612  < 2e-16 ***
long          -1.741e-01  1.652e-02 -10.537  < 2e-16 ***
sqft_living15  8.802e-05  4.324e-06  20.355  < 2e-16 ***
sqft_lot15     1.512e-07  6.709e-08   2.254   0.0242 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2543 on 21581 degrees of freedom
Multiple R-squared:  0.767,     Adjusted R-squared:  0.7668
F-statistic:  4736 on 15 and 21581 DF,  p-value: < 2.2e-16
```

(a) Mô hình 1

```
> summary(mod_BIC_2)

Call:
lm(formula = log(price) ~ bedrooms + bathrooms + sqft_living +
    sqft_lot + floors + waterfront + view + condition + grade +
    yr_built + yr_renovated + zipcode + lat + long + sqft_living15,
    data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7953 -0.1615  0.0037  0.1590  1.1735

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.932e+00  3.639e+00  -1.905   0.0568 .
bedrooms      -1.174e-02  2.382e-03  -4.930 8.27e-07 ***
bathrooms      7.137e-02  4.047e-03  17.634  < 2e-16 ***
sqft_living    1.403e-04  4.197e-06  33.431  < 2e-16 ***
sqft_lot       3.426e-07  4.355e-08   7.868 3.78e-15 ***
floors         6.979e-02  4.049e-03  17.234  < 2e-16 ***
waterfront     3.686e-01  2.176e-02  16.937  < 2e-16 ***
view           6.148e-02  2.649e-03  23.205  < 2e-16 ***
condition      6.352e-02  2.941e-03  21.594  < 2e-16 ***
grade          1.591e-01  2.682e-03  59.299  < 2e-16 ***
yr_built      -3.419e-03  9.120e-05 -37.494  < 2e-16 ***
yr_renovated   3.650e-05  4.585e-06   7.962 1.78e-15 ***
zipcode       -6.441e-04  4.137e-05 -15.569  < 2e-16 ***
lat            1.404e+00  1.337e-02 104.988  < 2e-16 ***
long          -1.715e-01  1.619e-02 -10.590  < 2e-16 ***
sqft_living15  9.566e-05  4.278e-06  22.359  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2524 on 21581 degrees of freedom
Multiple R-squared:  0.7703,    Adjusted R-squared:  0.7702
F-statistic:  4826 on 15 and 21581 DF,  p-value: < 2.2e-16
```

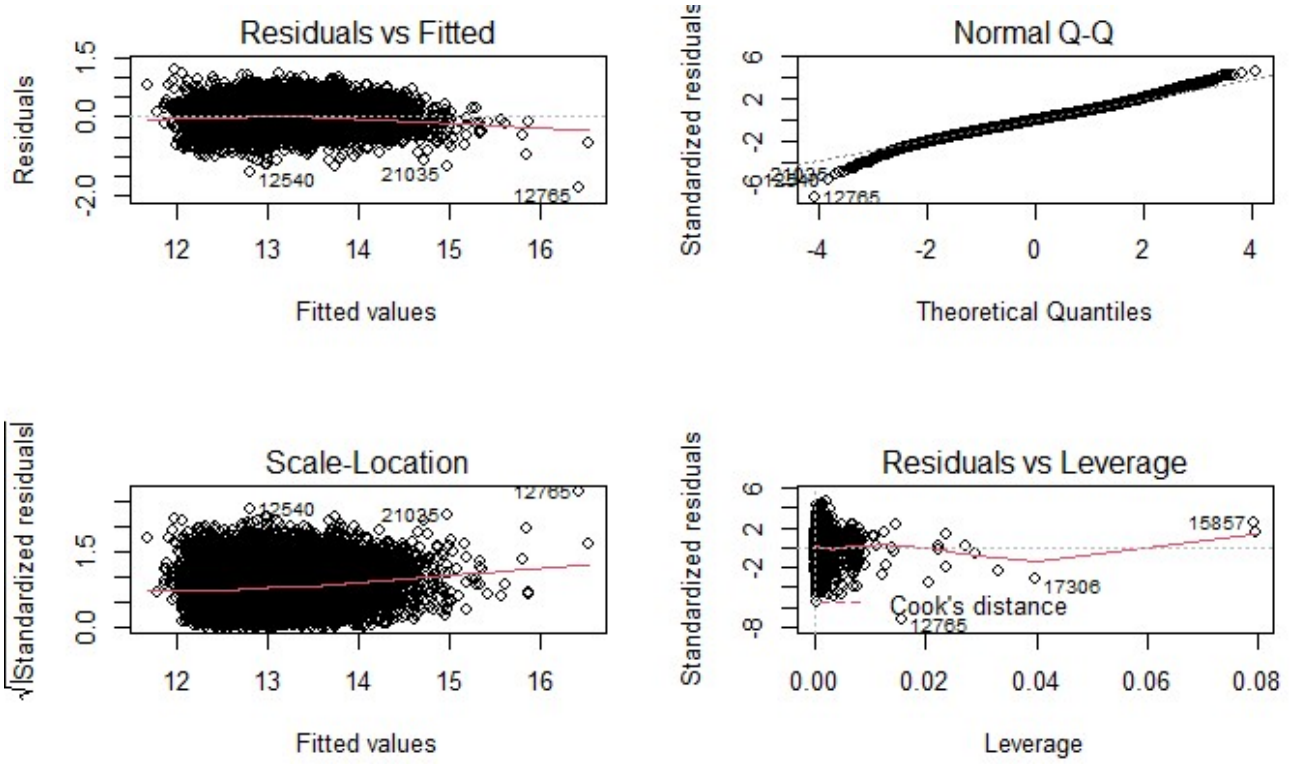
(b) Mô hình 2

Hình 2.6: Kết quả khi biến đổi **Price** thành **log(Price)**

Cả hai mô hình đều gồm 15 biến giải thích, mô hình 2 đã loại bỏ các biến **sqft_basement, sqft_above, sqft_lot15** khác với 3 biến đã loại trước khi biến đổi **Price**.

Nhóm em chọn **mô hình 2** là mô hình cuối cùng, vì: mô hình 2 có hệ số xác định lớn hơn ($R^2 = 77.03\%$), các biến liên quan đến diện tích tầng hầm (**sqft_basement**, **sqft_above**) đã được bao gồm trong **sqft_living**, diện tích khu đất vào năm 2015 cũng không mang nhiều ý nghĩa thống kê trong mô hình 1 nên có thể loại bỏ.

Kiểm tra giả thiết mô hình 2: phương sai của sai số không thay đổi, kì vọng bằng 0 và đã tuân theo phân phối chuẩn, chưa phát hiện hiện tượng đa cộng tuyến trong mô hình (các chỉ số VIF < 5) (Hình 2.7).



(a) Các biểu đồ kiểm định

```
> vif(mod_BIC_2)
    bedrooms    bathrooms  sqft_living     sqft_lot       floors   waterfront
    1.649892     3.283149     5.032386     1.102209     1.618698     1.202269
        view    condition        grade     yr_built yr_renovated      zipcode
    1.397185     1.241008     3.356002     2.432360     1.150352     1.661301
         lat         long sqft_living15
    1.163852     1.759544     2.912665
```

(b) Kiểm tra đa cộng tuyến

Hình 2.7: Kết quả khi biến đổi thành **log(Price)**

Kết luận

Vậy **mô hình cuối cùng được chọn** có các hệ số ước lượng như hình 2.8.

```
> coef(mod_BIC_2)
  (Intercept)      bedrooms     bathrooms   sqft_living      sqft_lot        floors
-6.932157e+00 -1.174353e-02  7.137346e-02  1.403104e-04  3.426024e-07  6.978707e-02
   waterfront          view     condition         grade      yr_built  yr_renovated
 3.685686e-01  6.147550e-02  6.351646e-02  1.590506e-01 -3.419313e-03  3.650388e-05
       zipcode           lat          long sqft_living15
-6.441469e-04  1.404181e+00 -1.714684e-01  9.565513e-05
```

Hình 2.8: Hệ số mô hình được chọn

Có 77.06% sự biến thiên của giá nhà ở quận King được giải thích bởi 15 biến độc lập, trong đó các yếu tố ảnh hưởng nhiều nhất gồm số phòng ngủ, số phòng tắm, diện tích nhà, số tầng, hướng nhà ra bờ sông, tình trạng của ngôi nhà, điểm tổng thể của ngôi nhà theo phân loại của quận, kinh độ - vĩ độ (vị trí), năm xây dựng căn nhà.

Giá trị của một căn nhà **không bị ảnh hưởng** nhiều bởi các yếu tố liên quan đến diện tích: diện tích tầng hầm, diện tích khu đất, diện tích ngoài tầng hầm, năm sửa chữa căn nhà, zipcode (mã vùng) của ngôi nhà.

2.3 Dữ liệu 3

2.4 Dữ liệu 4