ANALYSIS AND CONSTRUCT A REGRESSION MODEL FOR HOUSE PRICE PREDICTION IN SEATLE

Author: Viet Le
Date: 12/18/2022

ABSTRACT

This projects is to analyze the effects of some properties on the house price of the Seattle metropolitan area from the dataset on https://www.kaggle.com/datasets/shree1992/housedata?select=data.csv. Since the price, living area, and parking lot area are "very" large values to be compared with other predictor variables, a transformation with log function is applied into these variables to initialize the model (M1). Also, since the raw data of year of construction, renovation, and location are intuitively not linearly related to the price, the addition of binary values is applied to transform them into dummy variables. Hence, we come up with the model (M1) as below

$$\text{Log}(Y) = \text{intercept} + \beta_1 X1 + \beta_2 \log(X2) + \beta_3 \log(X3) + \beta_4 X4 + \beta_5 X5 + \beta_6 X6 + \beta_7 X7 + \beta_8 X8 + e \qquad \text{(M1)}$$

, in which
Y = the price of apartment (Price)
X1 = the number of bedrooms (Bedrooms)
X2 = the area of living space in square feet (sqft_living)
X3 = the area of parking lot in square feet (sqft_lot)
X4 = the number of floors of the apartment (floors)
X5 = the condition of the apartment out of 5 (condition)
X6 = yr_built = if the house was built after the year 2000 (dummy variable)
X7 = yr_renovated = if the house was renovated after the year 2000 (dummy variable)
X8 = city = if the house is in or near the central Seattle-Bellevue area (Bellevue, Sammamish, Kirkland, Seattle, Redmond, Shoreline, Issaquah, Bothell, Edmonds, Renton) (dummy variable)

Regression Ouput from R

```
Call:
lm(formula = log(price) ~ bedrooms + log(sqft_living) + log(sqft_lot) +
    floors + condition + yr_built + yr_renovated + city, data = maindf)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8240 -0.2189 -0.0058  0.1975  4.7188

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.536415   0.117409  47.155  < 2e-16 ***
bedrooms         -0.071331   0.007697  -9.267  < 2e-16 ***
log(sqft_living)  0.929025   0.018966  48.985  < 2e-16 ***
```

```
log(sqft_lot)      0.001116   0.007536   0.148  0.88234
floors             0.101681   0.012659   8.032 1.21e-15 ***
condition          0.092977   0.009000  10.331  < 2e-16 ***
yr_built1         -0.037605   0.016614  -2.263  0.02365 *
yr_renovated1      0.040940   0.014069   2.910  0.00363 **
city1              0.369811   0.012063  30.655  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3563 on 4542 degrees of freedom
Multiple R-squared:  0.5698,   Adjusted R-squared:  0.5691
F-statistic: 752.1 on 8 and 4542 DF,  p-value: < 2.2e-16
```

DIAGNOSIS AND CONSTRUCT POSSIBLE MODELS

The Figure 1 shows a scatter plot matrix of the observation variable and the predictor variables. The observation variable and the predictor variables appear linearly at least approximately. To access the extent of collinearity among the predictors, we might consider the Variance Inflation Factors. The Variance Inflation Factors for the predictor variables are founded as follows:

```
        bedrooms log(sqft_living)    log(sqft_lot)           floors
condition
        1.737458         2.368026         1.692643         1.665600
1.323138
        yr_built     yr_renovated             city
        1.639316         1.175637         1.145977
```

Since all VIFs are less than 5, the multicollinearity is not a problem to our model. That means our predictor variables are moderately independent. In other words, our model (M1) basically holds no multi-collinearity.

We now consider to the plots of standardized residuals against each other predictors variables (see Figure 2 and Figure 3). The random nature of these plot within values from -4 to 4 indicative that model (M1) is a valid model for data since there are no significant number of outliers shown.

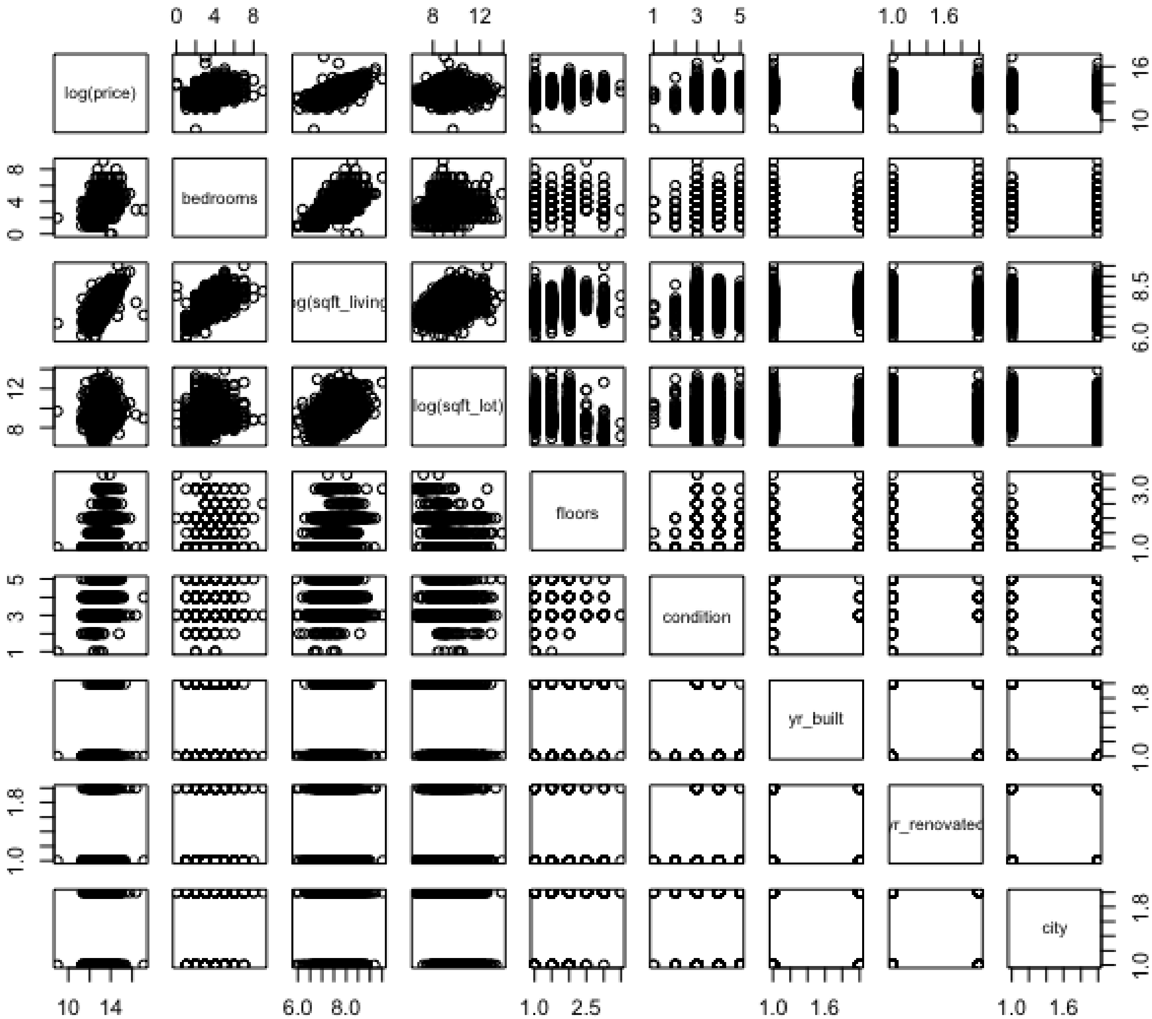

FIGURE 1.SCATTER MATRIX

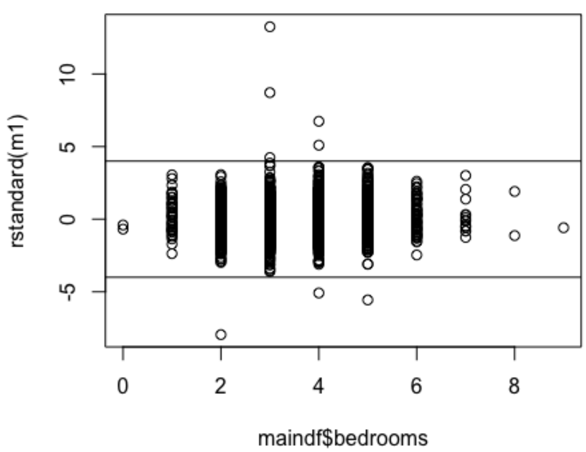

FIGURE 2. NUMBER OF BEDROOMS VS STANDARD RESIDUAL

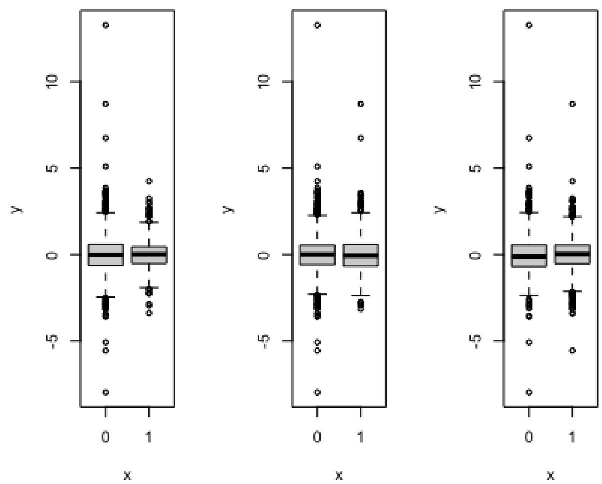

FIGURE 3. RESIDUAL STANDARD VS DUMMY VARIABLES

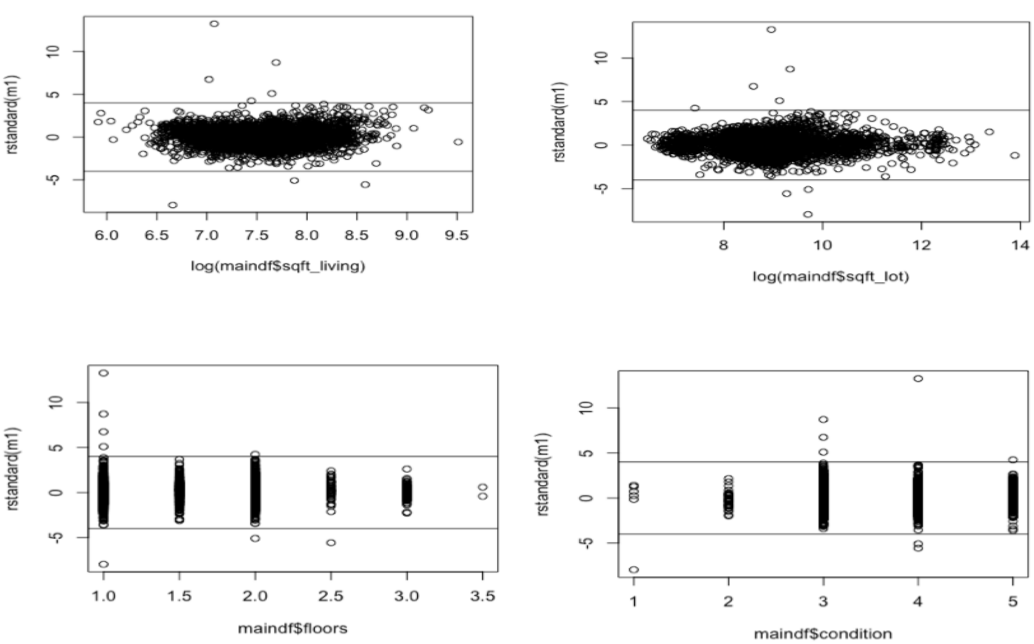

FIGURE 4. STANDARD RESIDUALS PLOT OF MODEL M1

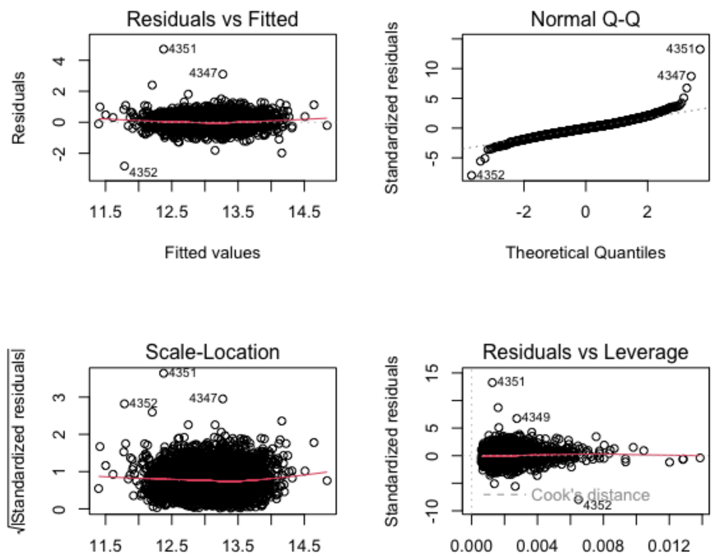

FIGURE 5. DIAGNOSTIC PLOTS

Figure 5 contains plots of the Standardized Residuals against Fitted Values, normal Q-Q, Scale-Location, and Cook's Distance of model (M1).

As we can see on the Residuals vs Fitted plot, the datapoints are scattered randomly about zero in no particular pattern. That means the constant variance properties of our model (M1) is hold.

The normal Q-Q plot states that the model holds normality since the scatter appears on the straight line. In other words, our data is normally distributed.

The Cook's distance plot is given without significant number of bad leverage points shown. That means the model (M1) is valid. Generally, there are still some points with influence to the model, such as 4347, 4349, 4351, and 4352 that should be investigated.

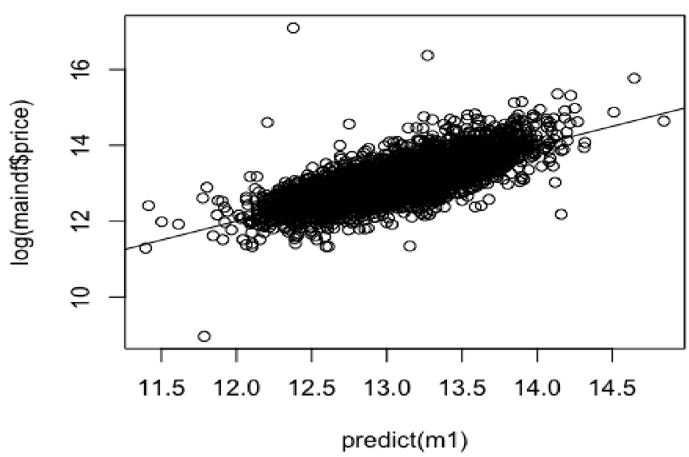

FIGURE 6. FITTED VALUES VS OBSERVES

Figure 6 contains a plot of observations against the fitted values. The straight-line fit to this plot provides a reasonable fit. This provides further evidence that model (M1) is a valid model for the data.

Now we consider the correlation of the interested predictor variables vs observations given the remaining predictors. There is statistically significance of bedrooms, living area (see Figure 7) and condition, and city (see Figure 8) to the model (M1). The lack of statistical significance of the regression coefficient associated with the variable log(sqft_lot), floors, year_built, and year_renovated are clear in the Figure 8. Thus, these predictors variables add little to the prediction of *Y*, Price.

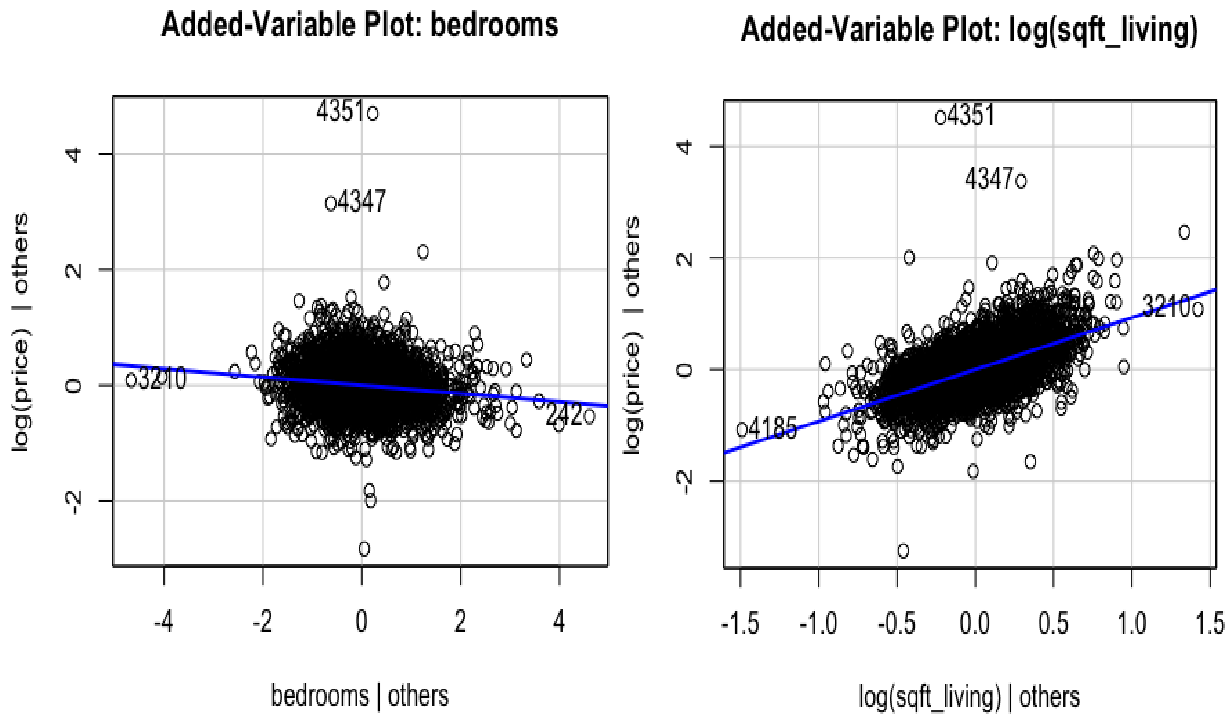

FIGURE 7. ADDED-VARIABLES PLOT

Some points are identified in the Added Variables plot of Bedrooms vs Price as having a large influence on the least squares estimate of the regression coefficient for Bedrooms. These points correspond to cases 3210, 4185, 4347, 4351 and 242 and should be investigated. Similarly, the right-hand side plot of Figure 7 shows the case 3210, 4185, 4347, and 4351 should be investigated. In Figure 8, the two cases 4347 and 4351 should be investigated.

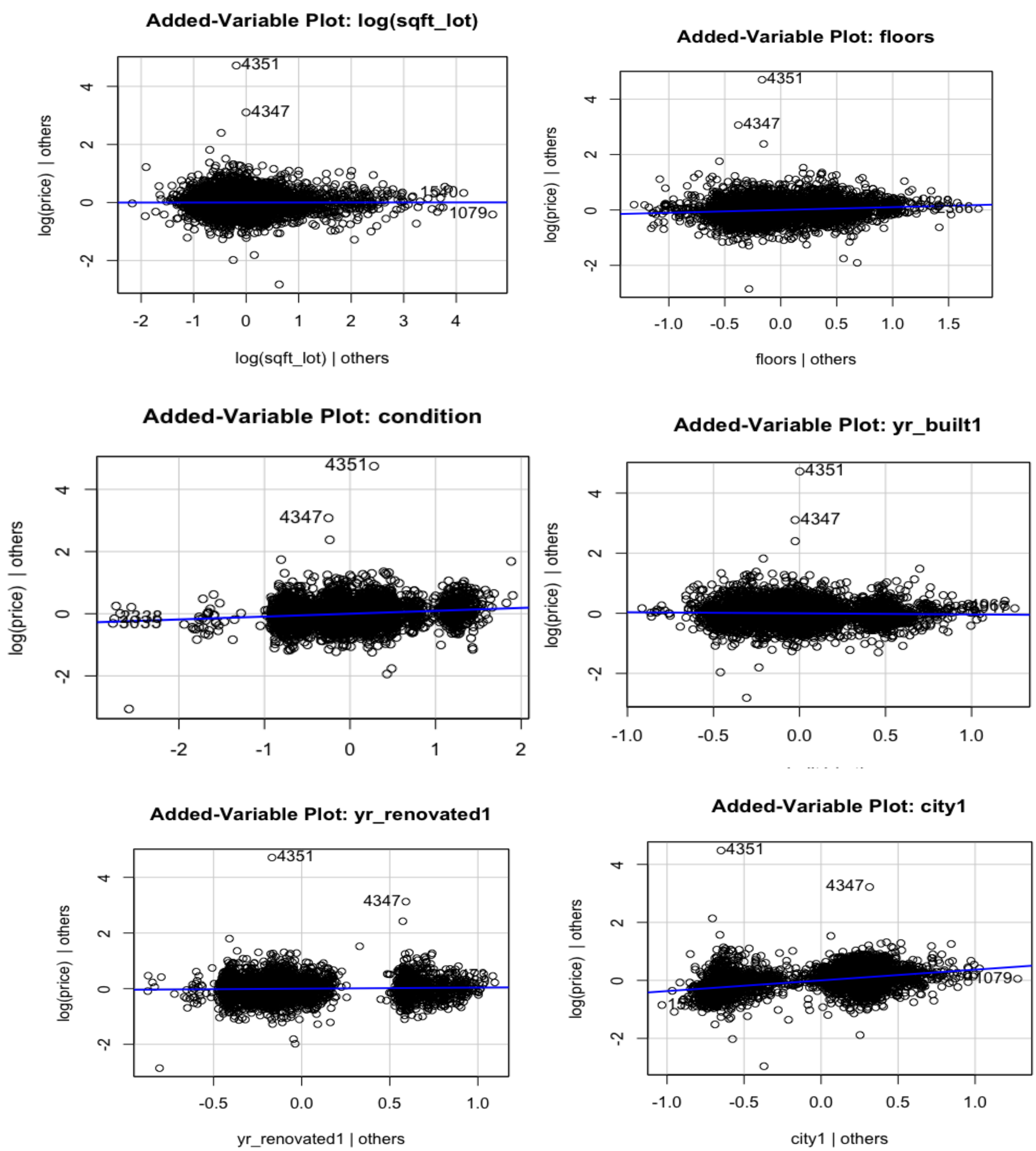

FIGURE 8. THE REMAINING ADDED VARIABLES PLOTS

COMPARE MODELS

From the analysis of added-variables plot, we can come up with a new model (M2), which is:

$Log(Y) = intercept + \beta_1 X1 + \beta_2 log(X2) + \beta_5 X5 + \beta_8 X8 + e$ (M2)

, in which
Y = the price of apartment (Price)
X1 = the number of bedrooms (Bedrooms)
X2 = the area of living space in square feet (sqft_living)
X5 = the condition of the apartment out of 5 (condition)
X8 = city = if the house is in or near the central Seattle-Bellevue area (Bellevue, Sammamish, Kirkland, Seattle, Redmond, Shoreline, Issaquah, Bothell, Edmonds, Renton) (dummy variable)

Regression Output from R of Model M2

```
Call:
lm(formula = log(price) ~ bedrooms + log(sqft_living) + condition +
    city, data = maindf)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8867 -0.2165 -0.0026  0.2014  4.6999

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.483552   0.114356  47.951   <2e-16 ***
bedrooms         -0.072000   0.007701  -9.350   <2e-16 ***
log(sqft_living)  0.966442   0.016272  59.393   <2e-16 ***
condition         0.073164   0.007917   9.241   <2e-16 ***
city1             0.376620   0.011379  33.098   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3589 on 4546 degrees of freedom
Multiple R-squared:  0.5631,     Adjusted R-squared:  0.5627
F-statistic:  1465 on 4 and 4546 DF,  p-value: < 2.2e-16
```

Since the VIFs are less than 5 and there is a large number of observations, the backward elimination by BIC is preferable for variables selection. The output of progress is shown as follows:

Regression Output from R

```
Start:  AIC=-9326.01
log(price) ~ bedrooms + log(sqft_living) + log(sqft_lot) + floors +
    condition + yr_built + yr_renovated + city
```

```
                    Df Sum of Sq    RSS     AIC
- log(sqft_lot)      1     0.003 576.65 -9334.4
- yr_built           1     0.650 577.29 -9329.3
<none>                           576.64 -9326.0
- yr_renovated       1     1.075 577.72 -9326.0
- floors             1     8.191 584.83 -9270.2
- bedrooms           1    10.903 587.55 -9249.2
- condition          1    13.551 590.19 -9228.7
- city               1   119.309 695.95 -8478.6
- log(sqft_living)   1   304.637 881.28 -7404.1

Step:  AIC=-9334.41
log(price) ~ bedrooms + log(sqft_living) + floors + condition + 
    yr_built + yr_renovated + city

                    Df Sum of Sq    RSS     AIC
- yr_built           1      0.75 577.40 -9336.9
<none>                           576.65 -9334.4
- yr_renovated       1      1.07 577.72 -9334.4
- floors             1      8.63 585.28 -9275.2
- bedrooms           1     10.98 587.62 -9257.0
- condition          1     13.62 590.27 -9236.6
- city               1    134.36 711.00 -8389.6
- log(sqft_living)   1    378.33 954.98 -7047.0

Step:  AIC=-9336.89
log(price) ~ bedrooms + log(sqft_living) + floors + condition + 
    yr_renovated + city

                    Df Sum of Sq    RSS     AIC
<none>                           577.40 -9336.9
- yr_renovated       1      1.42 578.82 -9334.1
- floors             1      7.99 585.38 -9282.8
- bedrooms           1     10.72 588.12 -9261.6
- condition          1     16.67 594.07 -9215.8
- city               1    133.83 711.23 -8396.6
- log(sqft_living)   1    377.60 954.99 -7055.4
```

With the view of Backward BIC, we can come up the new model (M3) as below:

$$\text{Log}(Y) = \text{intercept} + \beta_1 X1 + \beta_2 \log(X2) + \beta_4 X4 + \beta_5 X5 + \beta_7 X7 + \beta_8 X8 + e \qquad \text{(M3)}$$

, in which
Y = the price of apartment (Price)
X1 = the number of bedrooms (Bedrooms)
X2 = the area of living space in square feet (sqft_living)
X4 = the number of floors of the apartment (floors)
X5 = the condition of the apartment out of 5 (condition)

X7 = yr_renovated = if the house was renovated after the year 2000 (dummy variable)
X8 = city = if the house is in or near the central Seattle-Bellevue area (Bellevue, Sammamish, Kirkland, Seattle, Redmond, Shoreline, Issaquah, Bothell, Edmonds, Renton) (dummy variable)

Regression Output from R of Model M3

```
Call:
lm(formula = log(price) ~ bedrooms + log(sqft_living) + floors +
    condition + yr_renovated + city, data = maindf)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8064 -0.2184 -0.0074  0.1975  4.7174

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.536694   0.116090  47.693  < 2e-16 ***
bedrooms         -0.070488   0.007673  -9.186  < 2e-16 ***
log(sqft_living)  0.928680   0.017036  54.512  < 2e-16 ***
floors            0.089340   0.011269   7.928 2.79e-15 ***
condition         0.098827   0.008628  11.454  < 2e-16 ***
yr_renovated1     0.046468   0.013882   3.347 0.000822 ***
city1             0.368264   0.011348  32.453  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3565 on 4544 degrees of freedom
Multiple R-squared:  0.5693,     Adjusted R-squared:  0.5687
F-statistic:  1001 on 6 and 4544 DF,  p-value: < 2.2e-16
```

We now consider an ANOVA analysis to compare the full models (M1) to (M2), and to (M3) respectively. From the output below, we can see that the removing predictor variables out of full model (M1) to come up with model (M2) are valid since the F-value is statistically significant. The removing predictor variables out of full model (M1) to come up with model (M2) are not statistically significant as we can see from the output as follows.

Regression Output from R of ANOVA between M1 and M2

```
Analysis of Variance Table

Model 1: log(price) ~ bedrooms + log(sqft_living) + log(sqft_lot) +
floors +
    condition + yr_built + yr_renovated + city
Model 2: log(price) ~ bedrooms + log(sqft_living) + condition + city
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   4542 576.64
2   4546 585.71 -4   -9.0636 17.848 1.501e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Output from R of ANOVA between M1 and M3

```
Analysis of Variance Table

Model 1: log(price) ~ bedrooms + log(sqft_living) + log(sqft_lot) +
floors +
    condition + yr_built + yr_renovated + city
Model 2: log(price) ~ bedrooms + log(sqft_living) + floors + condition
+
    yr_renovated + city
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1   4542 576.64
2   4544 577.40 -2  -0.75541 2.975 0.05115 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence, the finalized model should be model (M2). Now, we might consider some diagnosis to guarantee that the model (M2) works well. The diagnosis plots (see Figure 10) agree the validation of model (M2).

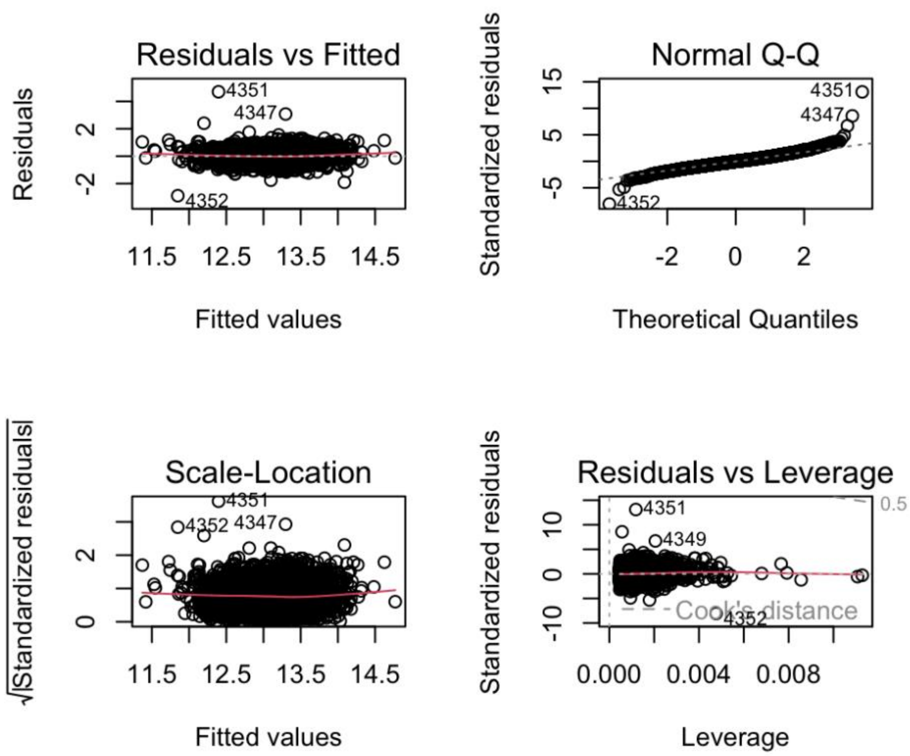

FIGURE 10. DIAGNOSIS PLOT OF FINALIZED MODEL