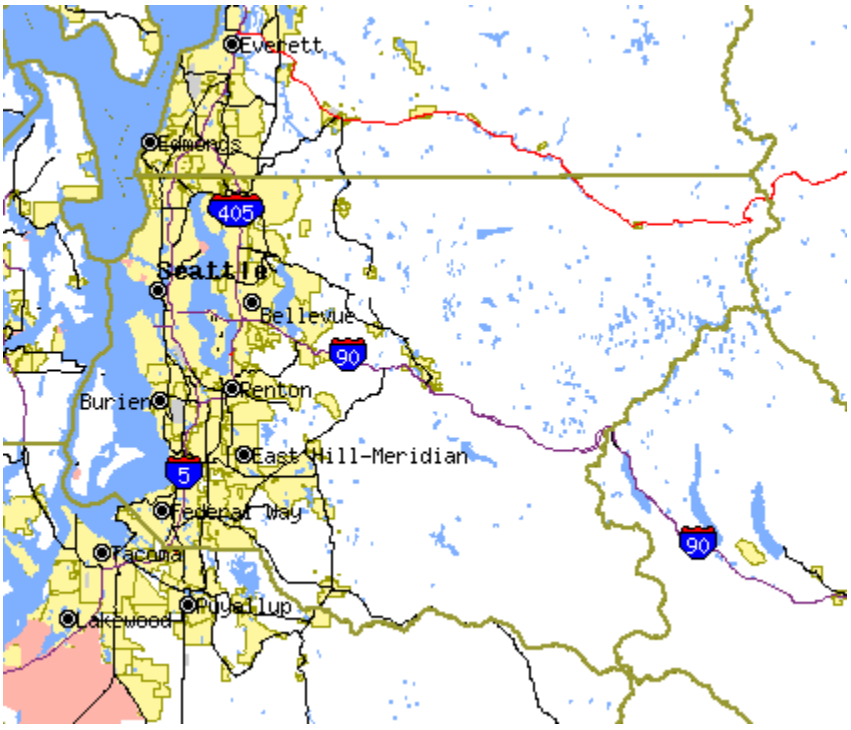


Predicting House Price in King County, Washington

Li Wiles
Department of Mathematics
Illinois State University
March 22nd, 2017

Contents

I. BACKGROUND

Housing price has always been playing an important role in the economy. The National Bureau of Economic Research states that "housing markets are becoming increasingly significant in shaping the economic and social well-being of many Americans". The research also points out housing price assumes considerable importance given housing expenditures are a large component of household budget. In this paper, the main purpose is building a regression model to predict house price in King County, Washington. Relations between house price and house features will be discussed. Potential business use for the model is to help individuals to evaluate whether a house is priced fairly and be used as reference in flipping houses.

II. DATA

Data for this project comes from Kaggle.com updated by a user named *harlfoxem*. The dataset includes 21,613 houses sold in King County, Washington between May 2014 and May 2015. I split the data into training data and testing data. 60% of data is used for training and 40% of data is used for testing.

The response variable is sales price of a house. There are 20 predictor variables including 19 house features and 1 dummy variable. Examples of house features are purchase date of a house, number of bedrooms, number of bathrooms, square feet of living space, the year a house is built, number of floors etc. Four house features are categorical including grade and condition of house. Data dictionary is attached in the Appendix.

III. MODEL BUILDING PROCESS

1) Preliminary Check

    a. Check for multicollinearity

From the scatter plot matrix and correlation matrix, house price has fairly strong positive correlation with square feet of living, grade and square feet above the ground. Surprisingly house price is slightly negatively correlated with zip code. House price has very weak correlation with the date a house's bought, longitude and latitude. That can be explained by dataset that location is narrowed and time frame is only for a year. Looking at the predictor variables only, most variables don't have strong correlation with one another. However, several house features do have positive correlations. For example, square feet of

living have a strong positive correlation with square feet above the ground with r = 0.87. The square feet of living is also more highly correlated with the number of bathrooms than with the number of bedrooms. All these imply there could be serious multicollinearity issue. To further investigate the linear dependencies of the predictor variables, I checked VIF values. Error showed up indicating there are aliased coefficients in the model. In the preliminary model, X13 showed "NA". Since X13 (Square feet above the ground) can be explained by square feet of living and grade, I decided to drop X13. Checking VIF values again, X5(Square feet of living) and X12(square feet above the ground) have the highest VIF values. X12 is removed because X5 has higher correlation with house price. After dropping X12, VIF values remain values between 1 and 5, which is acceptable in this project.

b. Check normality assumption & error variance constancy

The residual plot displays megaphone shape which indicates the need for a curvilinear regression function. It also tells the error variance is not constant. Some outliers are found in the residual plot. Shapiro Wilk normality test failed because of data size. But looking at the normal Probability plot, it shows the error term distribution is symmetrical with heavy tails. BP test also shows the error terms variance is not constant. It seems reasonable to conduct Y transformation. According to Box-Cox approach, logarithmic transformation Y' = log (Y) is the best to use.

c. Check Outliers

outlierTest() function is used to find outliers. 5 houses are found to be outliers. Digging into these houses, I find one house's square feet of lot is 172 times of its living space which seems rare. But considered the small number of outliers, I decided to keep them for now.

2) Model Selection

After logarithmic transformation of Y, the preliminary model is much better model. Data now seems to be fairly normally distributed. But error term variance is still not constant. Could we find a better model with fewer yet important predictor variables? To select the best subset model, I used adjusted R^2, Mallow's Cp, AIC, BIC and stepwise procedure methods and following are the top 4 models I select:

Tentative Model 1:

$$\log(\hat{Y}) = \beta_0 + \beta_4X_4+\beta_5X_5+ \beta_9X_9+ \beta_{10}X_{10}+ \beta_{11}X_{11}+ \beta_{14}X_{14}+ \beta_{17}X_{17}+ \beta_{19}X_{19}$$

Tentative Model 2:

$$\log(\hat{Y}) = \beta_0 + \beta_4X_4+\beta_5X_5+ \beta_9X_9+ \beta_{11}X_{11}+ \beta_{14}X_{14}+ \beta_{17}X_{17}+ \beta_{19}X_{19}$$

Tentative Model 3:

$\log(\hat{Y}) = \beta_0 + \beta_5X_5 + \beta_9X_9 + \beta_{11}X_{11} + \beta_{14}X_{14} + \beta_{17}X_{17} + \beta_{19}X_{19}$

Using stepwise procedure, all the predictor variables are retained, therefore, it is not the best subset model. Whether using adjusted R^2, AIC or BIC method, the best models are exactly the same.

3) Model Validation

	R_a^2	AIC	BIC	MSPR	MSE	MSPR-MSE
Model 1	0.7651	-23614.09	-23546.86	0.06500461	0.07	-0.004995
Model 2	0.7619	-23497.28	-23437.52	0.065904	0.07	-0.004096
Model 3	0.7563	-23297.61	-23245.32	0.06745934	0.07	-0.002541

Comparing R_a^2, AIC,BIC, MSPR, it can be concluded that Model 1 is the best model because it has the smallest values. The final model is

$\log(\hat{Y}) = \beta_0 + \beta_4X_4 + \beta_5X_5 + \beta_9X_9 + \beta_{10}X_{10} + \beta_{11}X_{11} + \beta_{14}X_{14} + \beta_{17}X_{17} + \beta_{19}X_{19}$

Or, more specifically,

log(Ŷ) = -12.92 + 1.17 (# of bathrooms) + 0.025(Square footage of Living Space)+ 1.057(View)+ 0.758(Condition)+0.628(Grade)-0.162(Year Built)+1.364(Latitude)+0.059(Average house square footage of the 15 closest neighbors)

Price = e ^ (-12.92 + 1.17 (# of bathrooms) + 0.025(Square footage of Living Space)+ 1.057(View)+ 0.758(Condition)+0.628(Grade)-0.162(Year Built)+1.364(Latitude)+0.059(Average house square footage of the 15 closest neighbors))

IV. DISCUSSION

1) Model results

For the final model, normality assumption is met but constant variance assumption is not met. Final model suggests house price could be fairly explained by latitude, average house square footage of the 15 closest neighbors, condition, Grade and the number of bathrooms, year built, view and square footage of living space. In King County, Washington, house closer to the north (with higher latitude) is more expensive. The

better condition a house holds, more expensive it will be. More views a house gets, more pricy it will be.

2) Thoughts and future Ideas

House price could be greatly affected by location and time. The model could be improved by collecting more data in different years. In this paper, the training data and testing data are chosen by the order of house purchasing time, not in a random order. To ensure randomness of data points, it can be a good idea to work around "sample" command in R software to divide data (Agrawal). Outlier issue isn't fully addressed and resolved in this paper. To alleviate the multicollinearity, ridge regression could be used. More variables can be considered such as number of garages, distance to school, style of house, etc.

V. REFERENCES

Ritesh Agrawal, "DIVIDING DATA INTO TRAINING AND TESTING IN R" https://ragrawal.wordpress.com/2012/01/14/dividing-data-into-training-and-testing-dataset-in-r/

Data Source Link: https://www.kaggle.com/harlfoxem/housesalesprediction

VI. APPENDIX

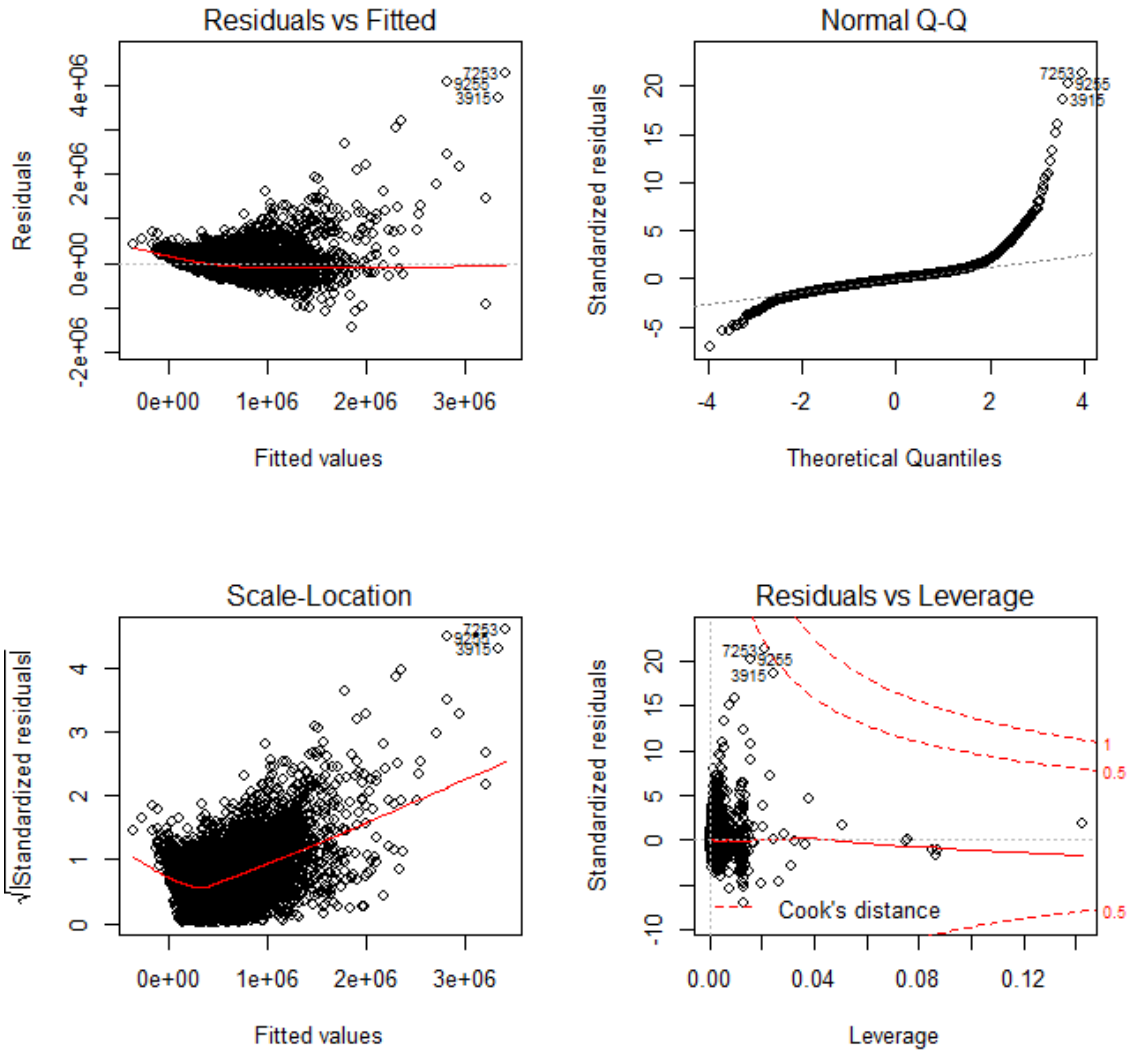
Appendix 1: Data dictionary

Variables	code	Description
price	y0	Sales price of a house
id	x1	Identity column
date	x2	Date the house is sold
bedrooms	x3	Number of bedrooms
bathrooms	x4	Number of bathrooms
sqft_living	x5	Square feet of living space
sqft_lot	x6	Square feet of lot
floors	x7	Number of floors
waterfront	x8	Waterfront house? 1 means yes,0 means no
view	x9	House with # of views? (Views:City,Lake,Mountain,…)
condition	x10	Condition of the house
grade	x11	Classification by construction quality which refers to the types of materials used and the quality of workmanship. Buildings of better quality (higher

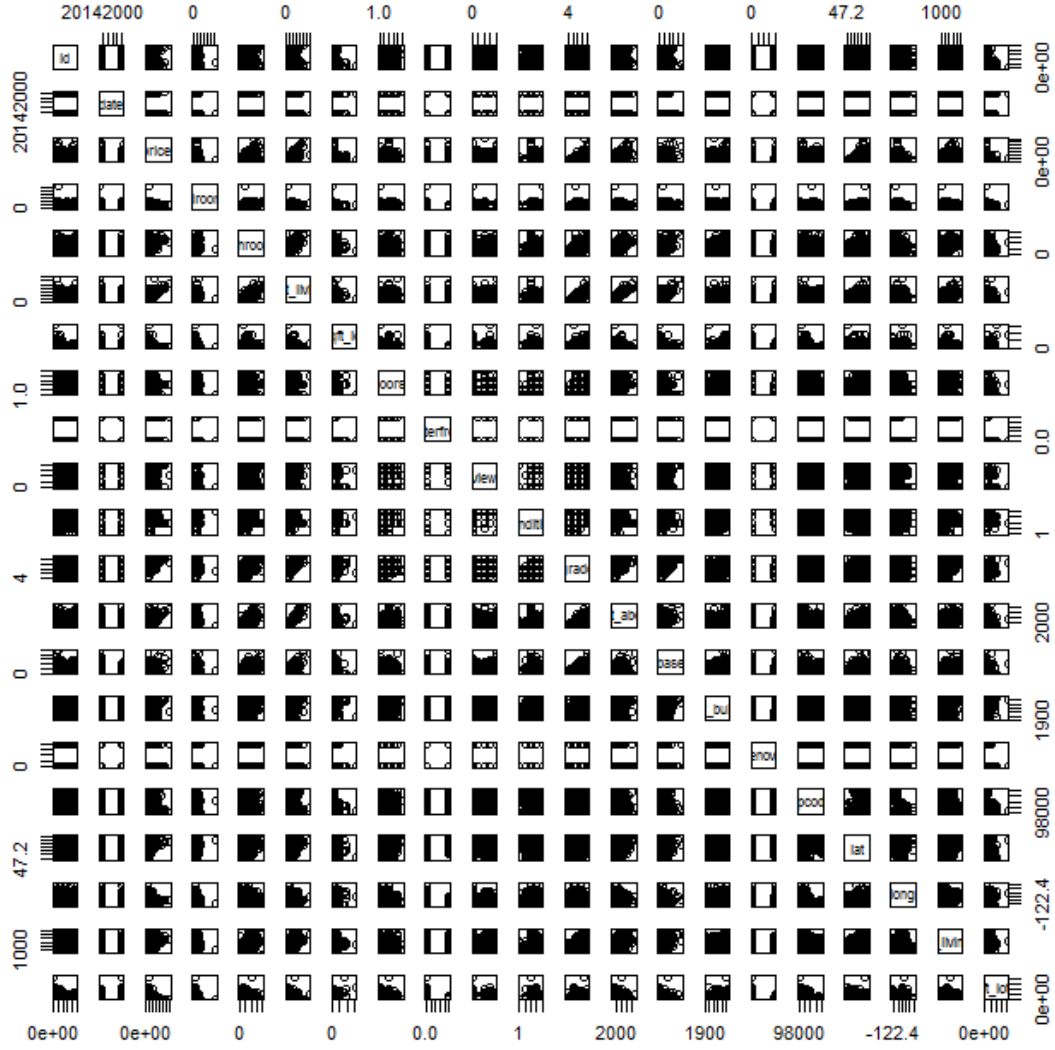
		grade) cost more to build per unit of measure and command higher value.
sqft_above	x12	Square feet above the ground=sqft of Living – sqft of basement
sqft_basement	x13	Square feet of basement
yr_built	x14	Year the house is built in
yr_renovated	x15	Year the house is renovated
zipcode	x16	Zip code
lat	x17	Latitude
long	x18	Longitude
sqft_living15	x19	Average house square footage of the 15 closest neighbors
sqft_lot15	x20	Average lot square footage of the 15 closest neighbours

Appendix 2: Preliminary checks

Residual Plot, Normality plot of preliminary model:



Scatter Plot of Variables



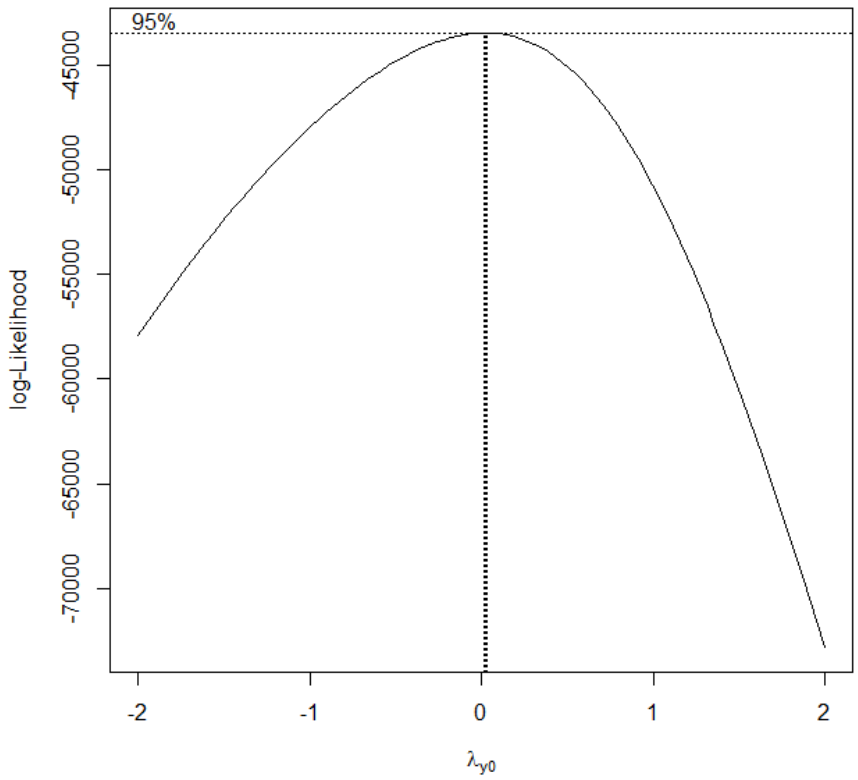
Correlation matrix: (shows partial variables)

	row.names	id	date	price	bedrooms
1	id	1	0.01109435	-0.007851806	0.001532534
2	date	0.01109435	1	0.006405634	-0.007123419
3	price	-0.007851806	0.006405634	1	0.3081297
4	bedrooms	0.001532534	-0.007123419	0.3081297	1
5	bathrooms	0.0004772125	-0.02686452	0.5211293	0.5390104
6	sqft_living	-0.008832309	-0.02816719	0.7007091	0.5801447
7	sqft_lot	-0.1339972	-0.0009212252	0.09484964	0.03886822
8	floors	0.005439891	-0.0274789	0.2786668	0.212037
9	waterfront	-0.008019784	0.00389632	0.2911057	0.000245488
10	view	0.02327302	0.005331071	0.3979616	0.09547846
11	condition	-0.02116349	-0.04649222	0.04341442	0.02873726
12	grade	0.005828838	-0.02823056	0.6563093	0.3670903
13	sqft_above	-0.01513679	-0.02261619	0.6059811	0.4700441
14	sqft_basement	0.009327408	-0.01627017	0.3269761	0.3275206
15	yr_built	0.01224224	0.004046732	0.0357529	0.1674234
16	yr_renovated	-0.0205612	-0.02523879	0.1389503	0.02223116
17	zipcode	-0.0006368121	0.001294096	-0.04973848	-0.1478424
18	lat	0.0005591173	-0.03008406	0.3014539	-0.001635754
19	long	0.008425518	0.004770038	0.01767432	0.1159023
20	sqft_living15	-0.01414907	-0.01194969	0.5926339	0.3845021
21	sqft_lot15	-0.1450588	-0.007730032	0.08609152	0.03310552

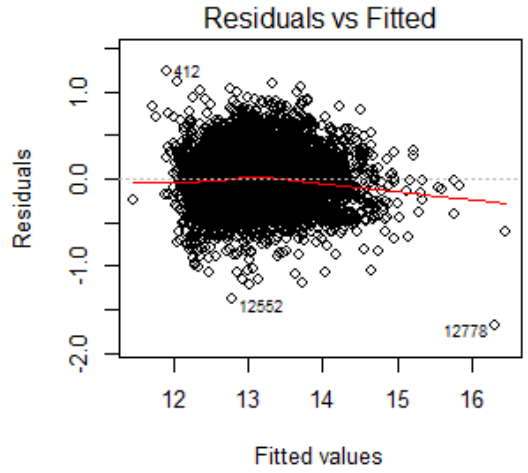
Analysis of preliminary model:

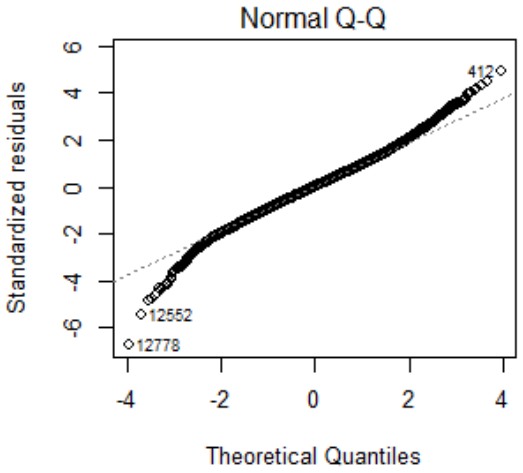
```
> anova(m1)
Analysis of Variance Table

Response: y0
              Df     Sum Sq    Mean Sq    F value    Pr(>F)
x1             1 1.0904e+11 1.0904e+11     2.6518 0.1034551
x2             1 7.4567e+10 7.4567e+10     1.8135 0.1781130
x3             1 1.6799e+14 1.6799e+14  4085.6527 < 2.2e-16 ***
x4             1 3.1483e+14 3.1483e+14  7656.7709 < 2.2e-16 ***
x5             1 4.1256e+14 4.1256e+14 10033.6347 < 2.2e-16 ***
x6             1 3.4818e+12 3.4818e+12    84.6782 < 2.2e-16 ***
x7             1 3.0966e+10 3.0966e+10     0.7531 0.3855139
x8             1 7.0386e+13 7.0386e+13  1711.8182 < 2.2e-16 ***
x9             1 2.7114e+13 2.7114e+13   659.4221 < 2.2e-16 ***
x10            1 1.0801e+13 1.0801e+13   262.6782 < 2.2e-16 ***
x11            1 5.5771e+13 5.5771e+13  1356.3794 < 2.2e-16 ***
x12            1 4.5846e+11 4.5846e+11    11.1500 0.0008427 ***
x14            1 9.3228e+13 9.3228e+13  2267.3497 < 2.2e-16 ***
x15            1 2.0488e+11 2.0488e+11     4.9827 0.0256189 *
x16            1 2.8699e+11 2.8699e+11     6.9798 0.0082534 **
x17            1 7.1931e+13 7.1931e+13  1749.3839 < 2.2e-16 ***
x18            1 4.5900e+12 4.5900e+12   111.6311 < 2.2e-16 ***
x19            1 1.3160e+12 1.3160e+12    32.0057 1.570e-08 ***
x20            1 1.1076e+12 1.1076e+12    26.9380 2.133e-07 ***
Residuals 12947 5.3235e+14 4.1118e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

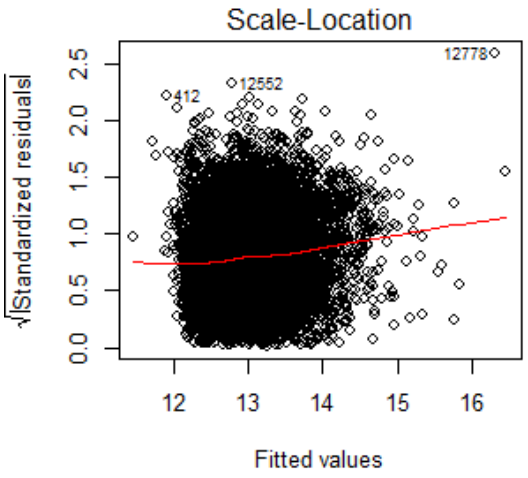
Box-Cox procedure for transformation:

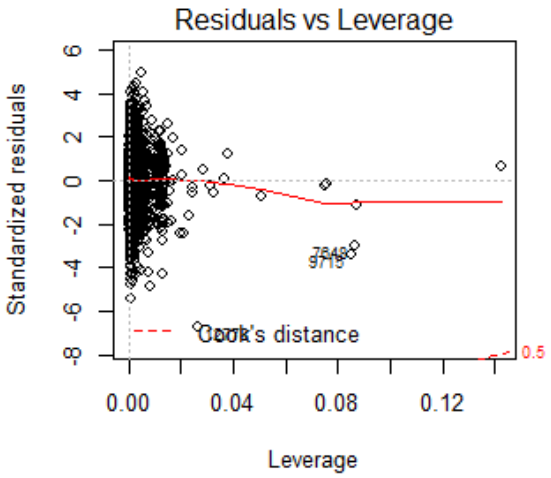


Appendix 3: Plot of Model after logarithmic transformation


Residuals vs Fitted
Residuals
Fitted values
412
12552
12778


Normal Q-Q
Standardized residuals
Theoretical Quantiles
412
12552
12778


Scale-Location
√|Standardized residuals|
Fitted values
12778
12552
412


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5

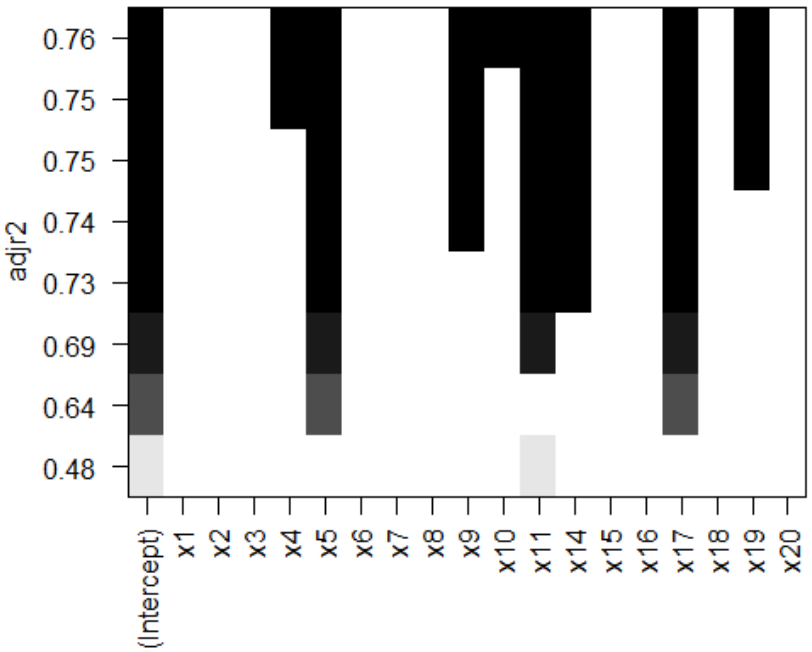
```
lm(formula = log(y0) ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 +
    x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16 + x17 + x18 +
    x19 + x20)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67125 -0.15971  0.00118  0.15847  1.23522

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.116e+02  1.110e+01 -10.050  < 2e-16 ***
x1           1.320e-12  7.835e-13   1.685   0.0920 .
x2           5.788e-06  4.990e-07  11.599  < 2e-16 ***
x3          -1.262e-02  3.141e-03  -4.017 5.92e-05 ***
x4           6.955e-02  5.279e-03  13.175  < 2e-16 ***
x5           1.395e-04  6.965e-06  20.022  < 2e-16 ***
x6           5.027e-07  7.343e-08   6.846 7.94e-12 ***
x7           7.498e-02  6.063e-03  12.368  < 2e-16 ***
x8           3.963e-01  2.821e-02  14.048  < 2e-16 ***
x9           5.881e-02  3.444e-03  17.077  < 2e-16 ***
x10          6.625e-02  3.643e-03  18.187  < 2e-16 ***
x11          1.570e-01  3.524e-03  44.548  < 2e-16 ***
x12         -1.278e-05  7.178e-06  -1.781   0.0749 .
x13                 NA         NA      NA       NA
x14         -3.912e-03  1.197e-04 -32.697  < 2e-16 ***
x15          3.755e-05  5.739e-06   6.543 6.27e-11 ***
x16         -7.276e-04  5.264e-05 -13.824  < 2e-16 ***
x17          1.401e+00  1.712e-02  81.786  < 2e-16 ***
x18         -1.498e-01  2.148e-02  -6.974 3.23e-12 ***
x19          1.164e-04  5.724e-06  20.328  < 2e-16 ***
x20         -2.683e-07  1.172e-07  -2.289   0.0221 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2515 on 12947 degrees of freedom
Multiple R-squared:  0.7722,    Adjusted R-squared:  0.7719
F-statistic:  2310 on 19 and 12947 DF,  p-value: < 2.2e-16
```

outlierTest result:

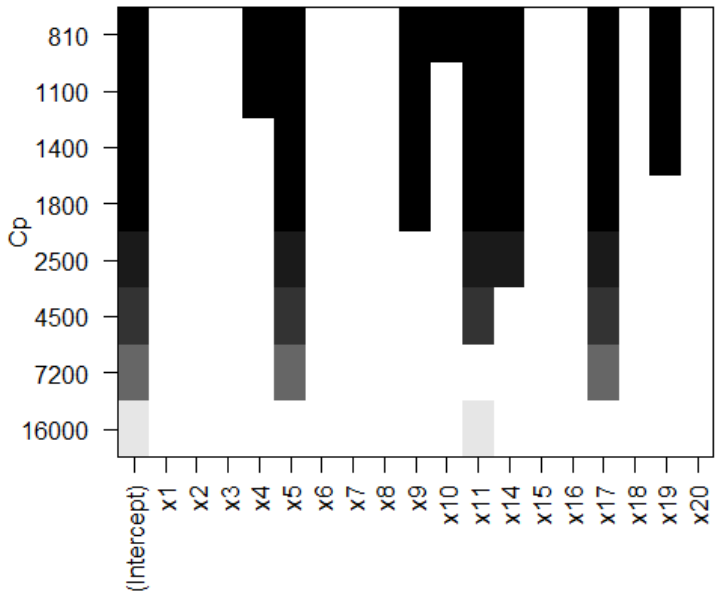
```
> outlierTest(m2)
       rstudent unadjusted p-value Bonferonni p
12778 -6.747094         1.5725e-11   2.0391e-07
12552 -5.466595         4.6726e-08   6.0590e-04
412    4.921578         8.6906e-07   1.1269e-02
2590  -4.806399         1.5540e-06   2.0151e-02
327   -4.779316         1.7783e-06   2.3059e-02
```

Appendix 4: Subset selection

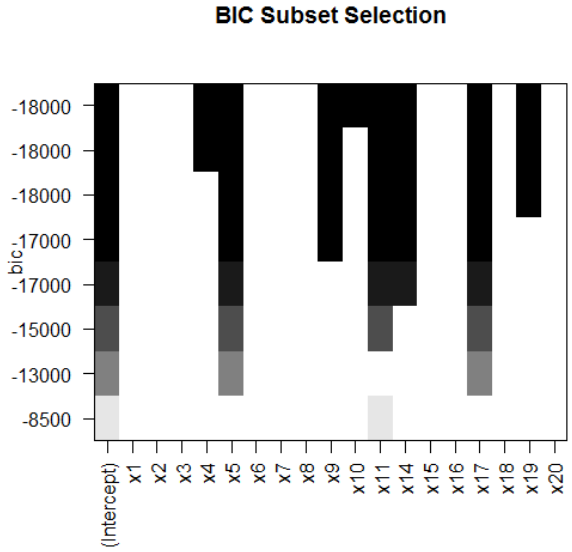
Best Subset Model based on R_a^2 :



Best subset Model based on Mallow's C_p :



Best subset models based on BIC:



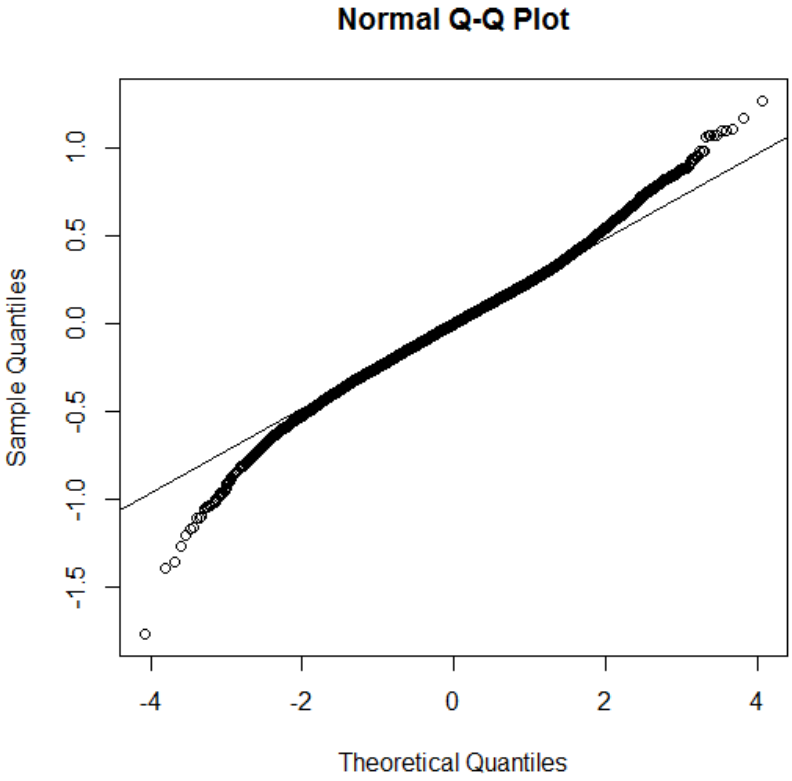
Appendix 3: Model Validation

Variance analysis of tentative model 1:

```
> summary(bestm1)

Call:
lm(formula = log(t_y0) ~ t_x4 + t_x5 + t_x9 + t_x10 + t_x11 +
    t_x14 + t_x17 + t_x19)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41793 -0.15975  0.00412  0.15729  1.11050

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.913e+01  1.072e+00  -45.83   <2e-16 ***
t_x4         8.411e-02  6.087e-03   13.82   <2e-16 ***
t_x5         1.465e-04  6.210e-06   23.59   <2e-16 ***
t_x9         8.001e-02  3.832e-03   20.88   <2e-16 ***
t_x10        5.355e-02  4.898e-03   10.93   <2e-16 ***
t_x11        1.731e-01  4.037e-03   42.88   <2e-16 ***
t_x14       -2.749e-03  1.249e-04  -22.01   <2e-16 ***
t_x17        1.377e+00  2.090e-02   65.87   <2e-16 ***
t_x19        6.312e-05  6.256e-06   10.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2551 on 8637 degrees of freedom
Multiple R-squared:  0.7653,    Adjusted R-squared:  0.7651
F-statistic:  3521 on 8 and 8637 DF,  p-value: < 2.2e-16
```

```
> summary(m_final)

Call:
lm(formula = log(Fy0) ~ Fx4 + Fx5 + Fx9 + Fx10 + Fx14 + Fx17 +
    Fx19)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51564 -0.17686 -0.00279  0.17484  1.29860

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.651e+01  7.289e-01  -77.54   <2e-16 ***
Fx4          1.104e-01  4.232e-03   26.09   <2e-16 ***
Fx5          2.218e-04  4.096e-06   54.14   <2e-16 ***
Fx9          9.203e-02  2.676e-03   34.39   <2e-16 ***
Fx10         5.204e-02  3.188e-03   16.33   <2e-16 ***
Fx14        -1.899e-03  8.504e-05  -22.34   <2e-16 ***
Fx17         1.516e+00  1.420e-02  106.71   <2e-16 ***
Fx19         1.680e-04  4.383e-06   38.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2825 on 21605 degrees of freedom
Multiple R-squared:  0.7126,    Adjusted R-squared:  0.7125
F-statistic:  7652 on 7 and 21605 DF,  p-value: < 2.2e-16
```

Final Model :




```
> vif(m_final)
     Fx4      Fx5      Fx9     Fx10     Fx11     Fx14     Fx17     Fx19
2.903441 4.346769 1.147962 1.165908 3.103844 1.811020 1.083826 2.665786
```

```
> bptest(m_final)

        studentized Breusch-Pagan test

data:  m_final
BP = 831.87, df = 8, p-value < 2.2e-16
```

```
Call:
lm(formula = log(Fy0) ~ Fx4 + Fx5 + Fx9 + Fx10 + Fx11 + Fx14 +
    Fx17 + Fx19)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76265 -0.16465  0.00029  0.16068  1.26810

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.745e+01  6.811e-01  -69.66   <2e-16 ***
Fx4          8.656e-02  3.888e-03   22.26   <2e-16 ***
Fx5          1.326e-04  3.990e-06   33.24   <2e-16 ***
Fx9          7.786e-02  2.457e-03   31.69   <2e-16 ***
Fx10         5.637e-02  2.916e-03   19.33   <2e-16 ***
Fx11         1.712e-01  2.634e-03   64.99   <2e-16 ***
Fx14        -3.253e-03  8.052e-05  -40.40   <2e-16 ***
Fx17         1.361e+00  1.320e-02  103.10   <2e-16 ***
Fx19         8.951e-05  4.187e-06   21.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2584 on 21604 degrees of freedom
Multiple R-squared:  0.7596,    Adjusted R-squared:  0.7595
F-statistic:  8532 on 8 and 21604 DF,  p-value: < 2.2e-16


9 x 1 sparse Matrix of class "dgCMatrix"
                           1
(Intercept) -4.730425e+01
Fx4          8.324405e-02
Fx5          1.343865e-04
Fx9          7.710795e-02
Fx10         5.490505e-02
Fx11         1.701338e-01
Fx14        -3.165878e-03
Fx17         1.355207e+00
Fx19         8.804117e-05
```