Final project : Predicting the real estate price by using Multiple Linear Regression

Uid : 606107407

Kim Tae Hee

**Introduction**

The real estate price is always a controversial subject to any country since it is known to be highly correlated with an overall economic condition of the country. The price of real estate is also intertwined with many aspects of life from the perspective of households. For instance, since the value of the house accounts for most of the households' equity, predicting the house price helps us to estimate the purchasing power of the household.

A lot of factors are associated with the house price so a sophisticated approach using quantitative methods is needed to predict the appropriate value of the house.

This paper mainly focus on Taiwan's real estate price and the price of the real estate refers to the price of the house that people live in. Therefore, to predict the house price using given datasets of diverse variables, we can set those factors that affect the house price as explanatory variables while setting the house price as a response variable. Then, by applying multiple linear regression model, we can find the best fitted model to properly predict the house price.

Under these circumstances, this paper aims to propose a quantitative real estate valuation approach through the multiple linear regression model by calculating correlation coefficient, t-value, F-value, and coefficient of explanatory variables.

This dataset includes 6 explanatory variables but it would be reasonable to exclude several explanatory variables that seem not to have any impact on the house price in our common knowledge. For instance, it would be reasonable to rule out the transaction date and geospatial data like longitude and latitude. This paper will focus on the three variables, 'the house age', 'the distance to the nearest MRT station', and 'the number of convenience stores in the living circle on foot' since those three criteria major concerns when predicting the house price.

The factors can be shortened to following x variables.

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
X2=the house age (unit: year)
X3=the distance to the nearest MRT station (unit: meter)
X4=the number of convenience stores in the living circle on foot (integer)
X5=the geographic coordinate, latitude. (unit: degree)
X6=the geographic coordinate, longitude. (unit : degree)

My research interest is to figure out the best fitted model that explains the house price from three predictors, 'the house age, the distance to MRT station, and the number of convenience stores nearby'. To specify, in case of latitude and longitude, the standard deviation values of those variables are too less, proving that the values are concentrated close to the mean. Therefore, it is less meaningful to analyze the relationship between X5,X6 and the response variable.

Therefore, the X-variables should be reduced to below.
X2=the house age (unit: year)

X3=the distance to the nearest MRT station (unit: meter)
X4=the number of convenience stores in the living circle on foot (integer)

Accordingly, we set the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared ) as a response variable, referring to Y.

The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan. Multiple linear regression model would be the best approach since there are more than one predictor variables. To best fit the model that best explains the relationship between those six x variables and the house price, this paper will find out best multiple linear regression model under the following order.

1. Draw the Scatter plot for each variable.
2. Draw the ordinary least squares and test the ordinary multiple linear model.
3. Do the diagnostic test of the ordinary model. (See if there is any violation of 'linearity, equal variance, and equal residuals'.)
4. Test if there is an issue of multicollinearity. (add variable test)
5. Use the transformed model (Box-cox transform, log transform) and perform the diagnostic test.
6. Eliminating x-variables through BIC, AIC
7. Derive the model that best explains the relationship between x-variables and the y-variables
8. Try ANOVA test for the reduced model and the full model.
9. Supplement the fitted model with any external data that exist.
10.

**Data description**

Before we run the R code to discover a linear association between explanatory variables and the house price, we can draw scatter plots to intuitively see if there is a linear relationship between each x-variable and the response variable. After checking the scatter plots, we can start to find out more accurate model by using the summary table, transformation, and ANOVA test.

1) Summary statistics for mean

To understand the traits of each variable, we can create summary tables for each variable.

```
>summary(X2.house.age)
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
  0.000   9.025   16.100    17.713  28.150    43.800
> summary(X3.distance.to.the.nearest.MRT.station)
   Min.   1st Qu.  Median    Mean     3rd Qu.    Max.
  23.38  289.32  492.23     1083.89  1454.28   6488.02
> summary(X4.number.of.convenience.stores)
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
```

```
   0.000   1.000   4.000   4.094   6.000  10.000
> summary(X5.latitude)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   24.93   24.96   24.97   24.97   24.98   25.01

> summary(X6.longitude)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   121.5   121.5   121.5   121.5   121.5   121.6
> summary(Y.house.price.of.unit.area)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   7.60   27.70   38.45   37.98   46.60  117.50
```

Judging from the summary table, it is reasonable to exclude X5 and X6 from our analysis since the values are concentrated on the average.

2) Summary statistics for standard deviation

```
> sd(X5.latitude)
[1] 0.0124102
> sd(X6.longitude)
[1] 0.01534718
```

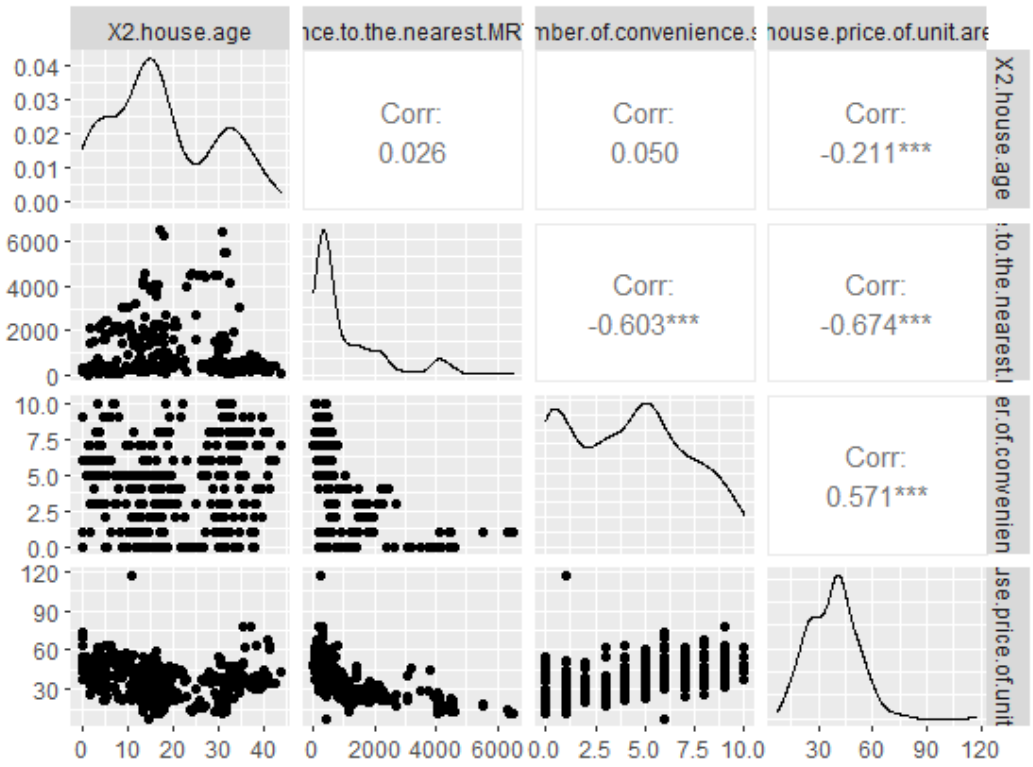
As shown in above, the standard deviation of the two excluded values are too small to be used in finding the best model that explain the relationship between explanatory variables and the response variable.

3) We can look overall scatter plots that show a rough relationship between each explanatory variable and the response variable.

The R code for the scatter plots is below. With GGally package.

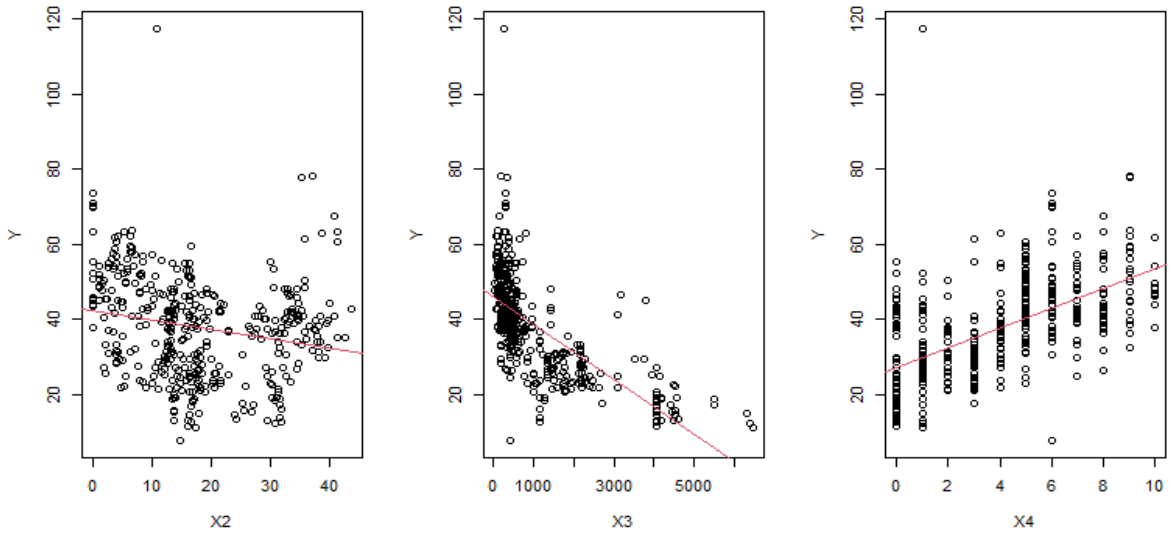
```
>pairs(realestate[,c(3,4,5,8)])
```

We can see the overall distribution of each variable.




4) Before delving into the correlation between the variables, this paper simplified the corresponding variable names to 'X2, X3, X4, and Y,' using 'rename' command.

5) We can first look at the simple linear regression model in each separate plot.




The plots show that each variable seems to be linearly associated to the house price.

6) Next, we can check the correlation between the variables.

X2 X3 X4 Y

X2 1.00000000 0.02562205 0.04959251 -0.2105670

X3 0.02562205 1.00000000 -0.60251914 -0.6736129

X4 0.04959251 -0.60251914 1.00000000 0.5710049

Y -0.21056705 -0.67361286 0.57100491 1.0000000

This correlation chart shows that none of the x variables are highly correlated each other. However, X3 and X4 is somewhat correlated(about 0.6), but the chance of multicollinearity is quite low.

To conclude, it seems that all the variables seem to have a linear association with the response variable, ignoring the effect of other predictors.

**Results and interpretation**

1) Let's begin by fitting an ordinary linear regression model. (R codes are attached to a separate file)

>m1=lm(Y~X2,X3,X4, data=data_new)

>summary(m1)

Call:

lm(formula = Y ~ X2 + X3 + X4, data = data_new)

Residuals:

Min 1Q Median 3Q Max

-37.304 -5.430 -1.738 4.325 77.315

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.977286 1.384542 31.041 < 2e-16 ***

X2 -0.252856 0.040105 -6.305 7.47e-10 ***

X3 -0.005379 0.000453 -11.874 < 2e-16 ***

X4 1.297443 0.194290 6.678 7.91e-11 ***

---

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.251 on 410 degrees of freedom

Multiple R-squared: 0.5411, Adjusted R-squared: 0.5377

F-statistic: 161.1 on 3 and 410 DF, p-value: < 2.2e-16

The result of the summary table shows that the overall f value is significantly low, so we reject the null hypothesis that none of the x variables has linear association with the response variable.

Furthermore, in this table, all of the explanatory variables show significantly low p-value so all the x variables are significantly showing linear association with the y variable.

To illustrate, If X3 and X4 are the same, an increase in house age by a year decrease the house price by 0.252856. If X2 and X4 are the same, an increase in the distance to the nearest MRT station decreases the house price by 0.005379. If X2 and X3 are the same, an increase in the number of convenience stores increase the house price by 1.297443 in average.

For the intercept, the average house price is 42.977286 is all values of the explanatory variables are 0. However, when we look at the Adjusted R-squared, it is '0.5377' so only 54percent of the response variable is explained by the explanatory variables.

Under these circumstances, the ordinary linear model should be set like this.

Predicted house price =

42.977286 - 0.252856*X2(House age) - 0.005379*X3(distance.to.the.nearest.MRT.station) +1.297443*X4(number.of.convenience.stores)

2) Now let's compare the previous summary table with the ANOVA table.

>anova(m1)

>Analysis of Variance Table

Response: Y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| X2 | 1 | 3390 | 3390 | 39.611 | 7.974e-10 *** |
| X3 | 1 | 34164 | 34164 | 399.165 | < 2.2e-16 *** |
| X4 | 1 | 3817 | 3817 | 44.594 | 7.908e-11 *** |
| Residuals | 410 | 35091 | 86 | | |

---

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

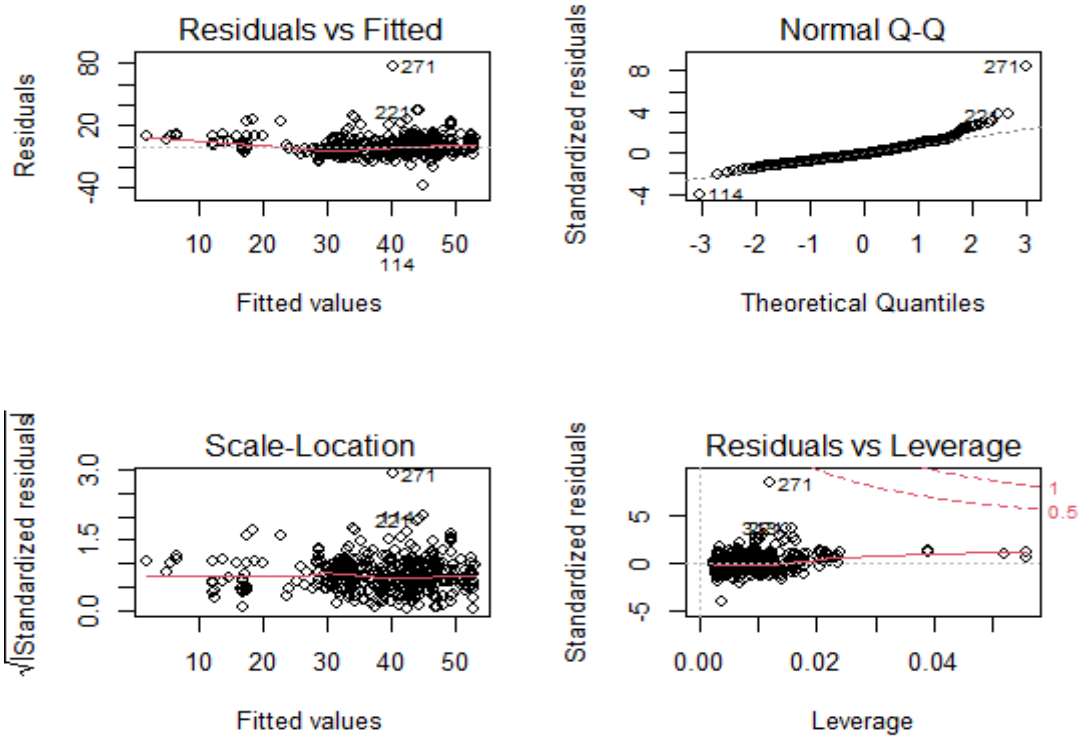
The ANOVA table also shows that all those x variables are significant and they are linearly associated with the house price like the summary table. In the ANOVA, the order of the X variables in the table matters so only the p-value of X4 is the same as the value shown in the summary table.

3) Next, we should perform diagnostic test for the model to determine whether our ordinary model satisfies the basic assumptions as follows.

1. Independent residuals
2. No relation between residuals and fitted values

3. Normally-distributed residuals
4. Homoscedastic (equal-variance) residuals, at each fitted value
5. No high-leverage points with large residuals
6. Linearity in all predictors, independently (perhaps also with interaction)

#diagnostic test for the multiple linear regression model




4) Let's test if there is any violation in following standards.

-linearity : The horizontal line in 'Residuals vs Fitted' proves that there is a linear association between explanatory variables and the response variable.

-a random scatter of points around the horizontal axis : It shows random scatter points around the horizontal axis.

-constant variability: Homoscedasticity is also satisfied since standardized residual plots show horizontal axis.

-normality of errors : It seems that errors show the normal distribution with a slightly longer right tail.

#Possible leverage points could be as follows.

Since the total observation number is 414, n=414

hii>2*4/414 = 0.01932367

Therefore, observations that show the leverage value more than 0.01932367 could be some leverage points.

#We can make a new datasets, 'eraseleverage', by erasing the high leverage points, using the command

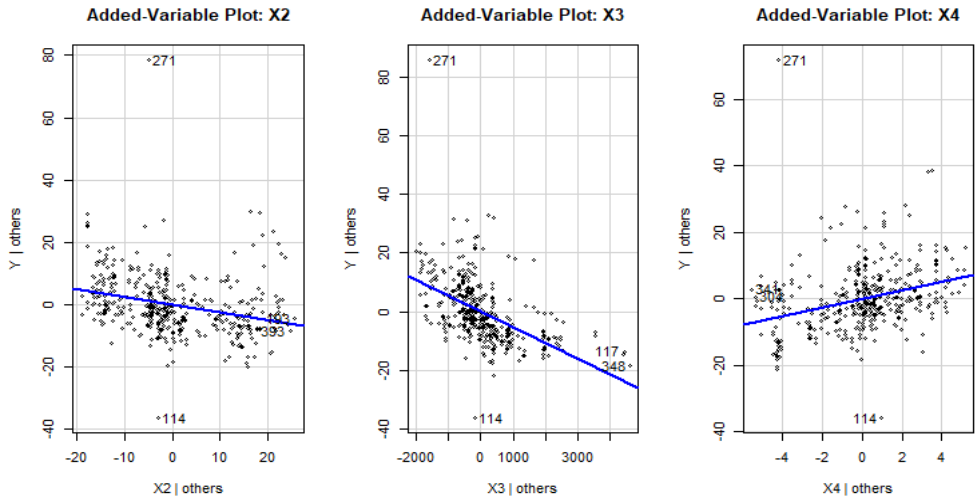
‘hatvalues.’ ( R codes are attached to a separate file )

Added-variable plots help to determine the effect of each explanatory variable to the response variable while considering effects of other predictors.

avPlot(m1,variable="X2",ask=FALSE)

avPlot(m1,variable="X3",ask=FALSE)

avPlot(m1,variable="X4",ask=FALSE)




Having adjusted for the effects of other predictors, all three explanatory variables seem to have linear relationship to the house price.

vif(m1)

| X2 | X3 | X4 |
|---|---|---|
| 1.007349 | 1.577579 | 1.580431 |

Since the value is less than 5, there is no issue of multicollinearity.

5) Now, we can Analyze the linear regression model m2 based on the data ‘eraseleverage’ to see if the adjusted model without the leverage points show the improvement.

m2=lm(Y~X2+X3+X4, data=eraseleverage)

summary(m2)

Call:

lm(formula = Y ~ X2 + X3 + X4, data = eraseleverage)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -37.462 | -4.969 | -1.205 | 4.089 | 76.816 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 44.0076330 | 1.6097428 | 27.338 | < 2e-16 | *** |
| X2 | -0.2772727 | 0.0419797 | -6.605 | 1.32e-10 | *** |
| X3 | -0.0063571 | 0.0006027 | -10.547 | < 2e-16 | *** |
| X4 | 1.2763175 | 0.2139414 | 5.966 | 5.51e-09 | *** |

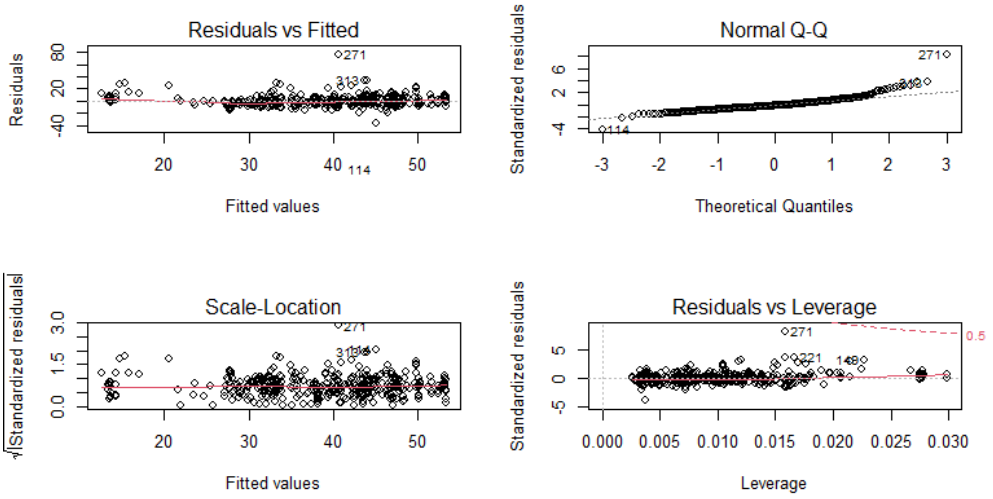
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.257 on 386 degrees of freedom

Multiple R-squared: 0.517, Adjusted R-squared: 0.5133

F-statistic: 137.7 on 3 and 386 DF, p-value: < 2.2e-16

The summary table also shows that all three explanatory variables are significant. We can now look at the plot of m2 to do the diagnostic test.

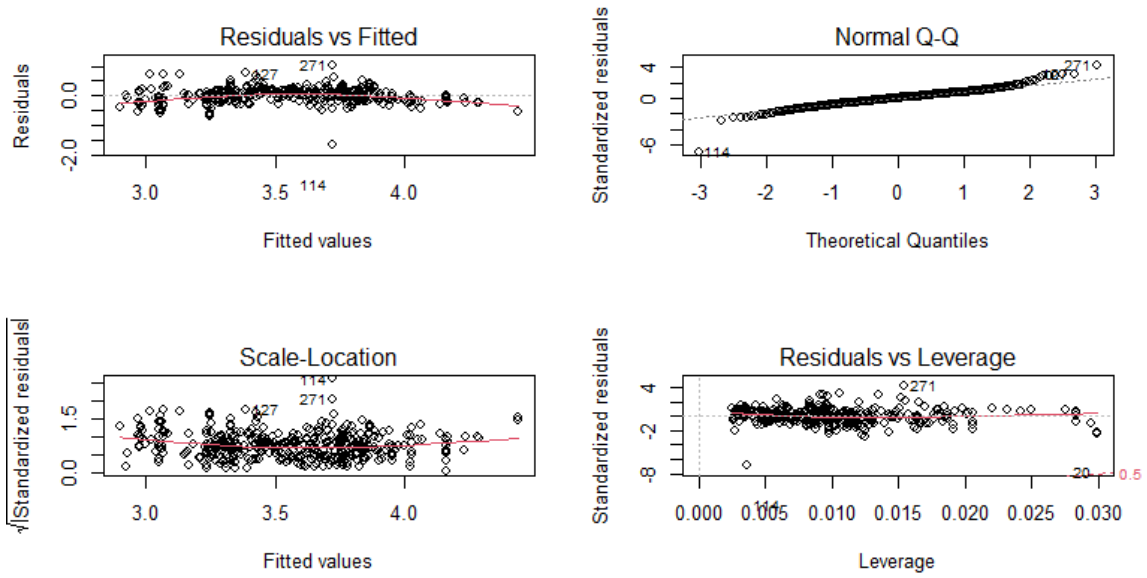



This m2 is also quite good under the conditions of basic assumptions that are mentioned above.

6) Now let's do the log transformation.

Since there are several '0' values in X2 and X4 which makes the log transformation impossible, we can first try transforming the response variable and X3 first. The formula will be as follows.

```
#log transformation for Y and X3

log_realestate = transform(realestate, lY=log(Y), lX3 = log(X3))

m3 = lm(lY~X2+lX3+X4, data=log_realestate)

summary(m3)

par(mfrow=c(2,2))

plot(m3)
```



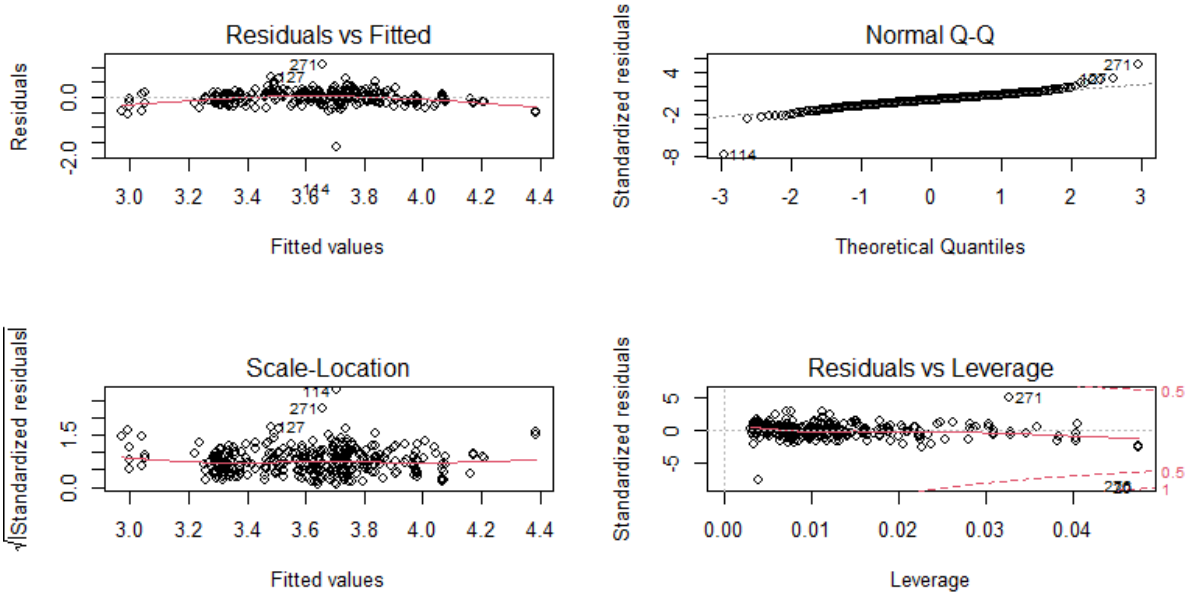

As you can see above, the 'Residuals vs Fitted' scatter plot portrays a slightly curved horizontal line, instead of a straight line. Other plots seem quite good and does not violate the basic assumptions but this model shows less linear relationship between the explanatory variables and the response variable. Therefore, m3 turns out to be less fitting to the dataset than the previous models, m2 and m1.

7) Alternative model with log transformation that has removed values with 0

To remove the values with '0', we can use subset command.

```
>subset(realestate, X4>0)

>realestate[apply(realestate,1, function(x) all(x!= 0)),]

>without0 <- realestate[apply(realestate,1, function(x) all(x!= 0)),]

>without0

>log_without0 = transform(without0, lY=log(Y),lX2=log(X2),lX4=log(X4), lX3 = log(X3))

>m4 = lm(lY~lX2+lX3+lX4, data=log_without0)
```

Now, let's apply diagnostic test for the model m4.




The plots quite resemble the plots of previous model m3, the horizontal line in 'Residuals vs Fitted' is also a little curved. Therefore, m4 does not fit better to the dataset than m1 and m2.

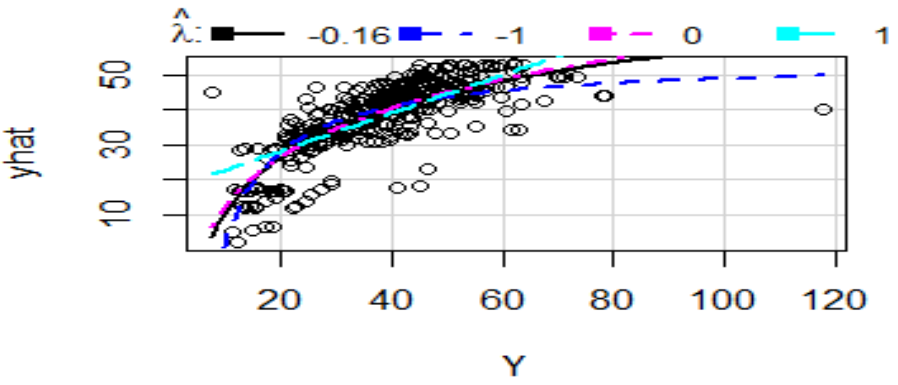
8) Let's figure out which variable needs transformation.

To target exact variables that need transformation, we can use 'inverseResponsePlot' command in package 'car.'

#inverseResponsePlot transformation (transform the response variable only)

inverseResponsePlot(m1, key=TRUE)

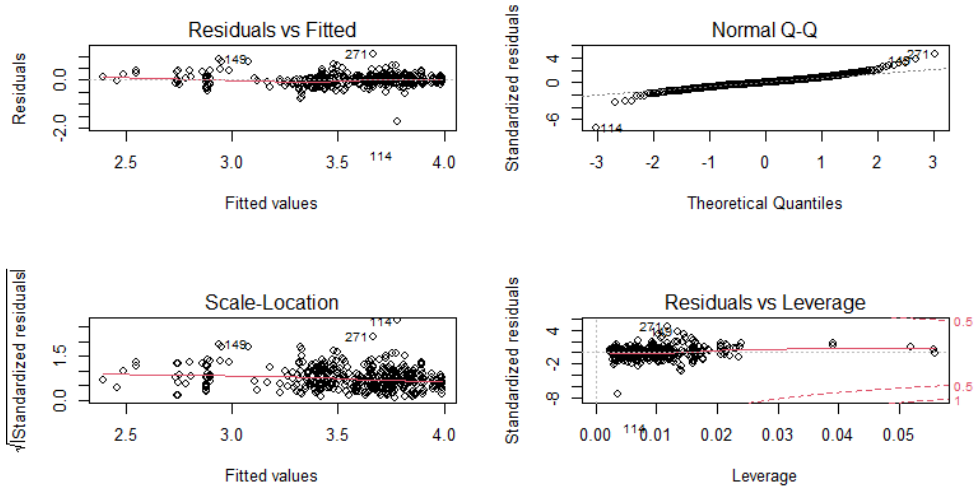
par(mfrow=c(1,1))




# Among the curves, it is evident that lamda of 0.16 provides the closest fit to the data. Since we can round 0.16 to 0, we can try log transformations to the response variable.

9) Log transformation only to the response variable. (Codes are attached separately)

#log transformation to the y. (The model below is named m5)




The plots satisfy the assumptions for the multiple linear regression in diagnostic test.

Also, when we look at the summary table, m5 shows the same with m1 and m2. Therefore, m5 is also quite a good model.

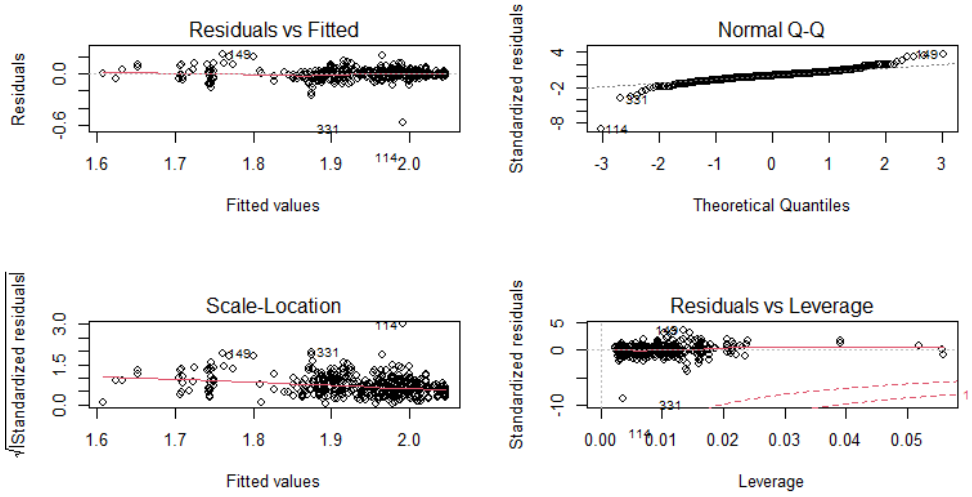
10) In addition, we can also use BC transform formula to find the best lamda and figure out a new model. The codes will be attached in the separate file.

>Y.bc <- BCTransform(Y, bc.power)

#Re-run the linear model with this transformation:

summary(m1.bc <- lm(Y.bc ~X2+X3+X4, data=realestate_new))

We can name the new model 'm1.bc' and the diagnostic test shows that this model is also good. plot(m1.bc)




#This is also quite good.

11) another transformation function using car package

>(bc.realestate <- powerTransform(m1)) : An estimated transformation parameter turns out to be 0.2302808.

>str(bc.realestate, max.level=1)

>print(bc.realestate$lambda)

>summary(bc.realestate)

12) histogram before log transformation vs after the transformation

#histogram before log transformation

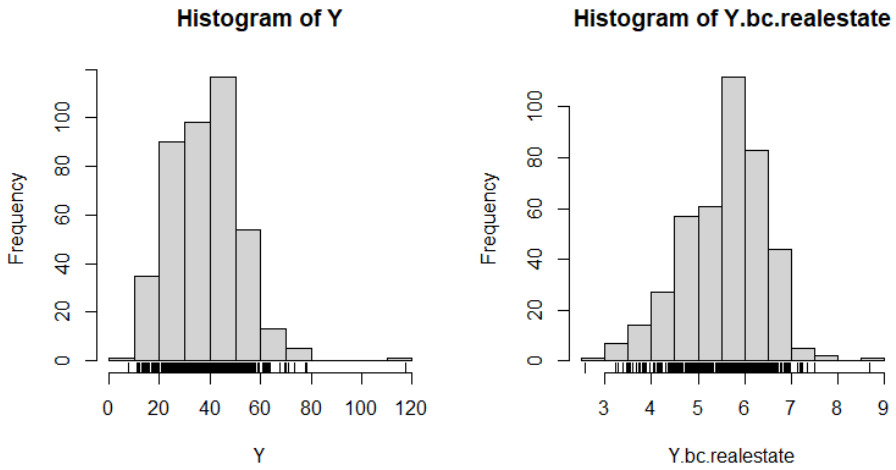
summary(Y)

hist(Y, breaks=12); rug(Y)

Y.bc <- bcPower(Y, lambda=bc.realestate$lambda)

#histogram after log transformation

hist(Y.bc, breaks=12); rug(Y.bc)



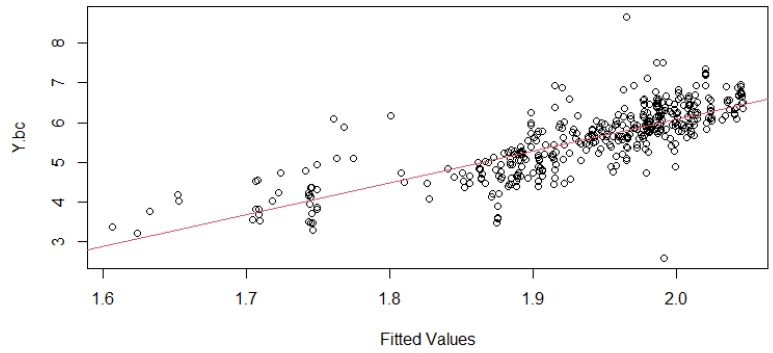

#Both distributions are quite similar, so does not matter whether we transform the reponse variable or not.

13) fitted values and house price.

#fitted values and house price

plot(m1.bc$fitted.values,Y.bc.realestate,xlab="Fitted Values",ylab=expression(Y.bc))

abline(lsfit(m1.bc$fitted.values,Y.bc.realestate),col=2)




There are some outliers but the model m1.bc seems good.

14) Now let's use variable selection to obtain a final model and remove redundancy in the full model. The results will be attached to the separate R script file.

Remember that, m1=lm(Y~X2+X3+X4, data=realestate)

#backward elimination using AIC

backAIC<-step(m1, direction="backward", data=realestate)

#backward elimination usic BIC

backBIC <-step(m1, direction="backward", data=realestate, k=log(414))

nrow(realestate)

#Forward selection using AIC

mint<-lm(Y~1, data=realestate)

summary(mint)

forwardAIC <- step(mint, scope=list(lower=~1, upper=~Y ~ X2 + X3 + X4),direction="forward", data=realestate)

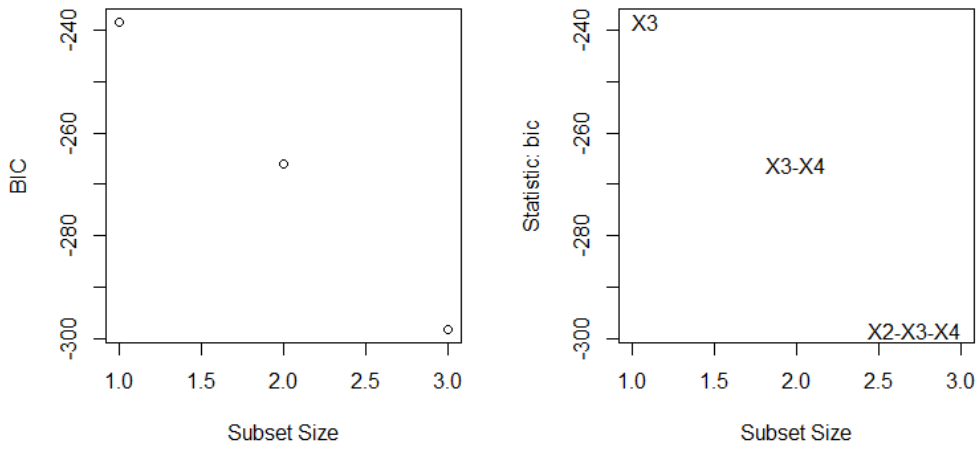
#Forward selection using BIC

forwardBIC <- step(mint, scope=list(lower=~1, upper=~ Y ~ X2 + X3 + X4),direction="forward", data=realestate, k=log(414))

Consequently, including all three variables seems to be the most significant since all three selection methods show the same results.

Since we have to figure out the optimal subset size with the minimal AIC or BIC value with the highest R-adjusted square value, another option is to draw a plot using leap package. (Codes are attached separately)

>subsets(b1, statistic=c("bic"))




The plots above shows that the optimal subset size of the model is 3, and the model should include all the three explanatory variables(house age, distance to MRT, number of convenience stores).

15) Even though including all three variables turns out to be valid, we can confirm our full model by doing hypothesis test with the reduced model.

m.ols.reduced <- lm(Y~ X3+X4, data=realestate)

anova(m.ols.reduced, m1 )

The p-value of the reduced model turns out to be 7.47e-10 ***. Since it is significantly small, we reject the null hypothesis, so the full model better fits the data.

**Further interpretation**

1) R results and interpretation

So far, we have tried several models to predict the house price from three predictors X2, X3, and X4. We can conclude that four models, **'m1, m2, m5, m1.bc'** fit well to our data set.

All those models verify that X2(house age) and X3(distance to the MRT station) are significant factors that are negatively related to the house price. In contrast, X4(number of convenience stores) is positively related to the house price. Among all those valid models, the model m1 deserves the optimality that explains our dataset. To can verify our prediction model, the average house price could be estimated as follows.

**Predicted house price =**

**42.977286 - 0.252856*X2(House age) - 0.005379*X3(Distance to the nearest MRT station) +1.297443*X4(Number of convenience stores)**

Also, the added-variable plots enhanced our model by confirming that with the consideration of other variable's impact on the explanator variable, all three explanatory variables have linear association with the response variable.

2) The diagnostic tests also proved that those models satisfy the basic assumptions of multiple linear regression, including linearity, constant variance, and normality of errors.

**Discussion**

1) Summary of the model

To summarize my project, the house price in Taiwan could be predicted with the multiple linear regression model, using three predictors, 'house age, distance to the nearest MRT station, and the number of convenience stores nearby. Our model does not have an issue of multicollinearity and it satisfies all the basic assumptions of linear regression, in terms of the constant variance.

2) External data strengthen our conclusions.

Our model not only holds in Taiwan. According to the dissertation 'Prediction of real estate prices with data mining algorithms,' MRT stations, and the house age play as key factors in determining the price of real estate. However, this article shows quite a different result in the sense that the number of convenience store did not influence the house price significantly. However, still, the house price can be estimated by the remaining two variables.[1]

3) Several limitations of the model.

The model can be improved in the future if it contains more factors that reflect the macroeconomic condition, such as index of exchange rate, GDP, interest rate, and the mortgage interest rate. These variables are known for affecting the housing market significantly but are excluded in this paper. Since omitting important variables can lead to bias in the model and estimated parameters, predicting the house price only through previous three variables could result in inaccurate p-values.

Another limitation in these models is that the dataset does not contain numerical variables that explains the feature of the house itself as a commodity, such as house size, land space, and floor area ratio. Moreover, several categorical variables such as 'construction companies, existence of competitive schools that are nearby, and proficiency of overall education' also remain to be accounted for.

Finally, this model does not consider difference of geological traits of existing countries in the world, making it difficult to generalize the model to a broader respect. For example, Countries with bigger land size to build additional real estates might have different formula that fits the dataset that better reflect their countries' situation.

Including every variable that affects the house price might be useful in deriving the best model that fits our dataset. However, overfitting data is also problematic since it reduces the predictive ability of the model. Therefore, even though the future model might emerge with the largest adjusted r square value, we might beware of the overfitting when assessing the prediction ability.

[1] Uzut, O. G., and S. Buyrukoglu. "Prediction of real estate prices with data mining algorithms." *Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences* 8.9 (2020): 77-84.