12/11/2018

Analysis of House Prices in Taiwan

Prof. Sudeep Bapat

Niko Kalpakian & Riley Mault
SECTION TIME: W 11 A.M

Introduction

Our data set is the collection of historical data and statistics of real estate in Sindian Dist., New Taipei City, Taiwan from September 2012 to July 2013. The goal of the project is to discover if we can predict real estate value based on several predictors such as house age and number of convenience stores nearby. The real estate value is calculated by house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared). We have decided to split the data into two sets: In Sample and Out of Sample. In Sample will contain roughly ¾ of the data and we will use the In Sample to create a model to predict house price of unit area. Then we will test our final model with the Out of Sample data which it wasn't built on, to see how accurate the model is.

Questions of Interest

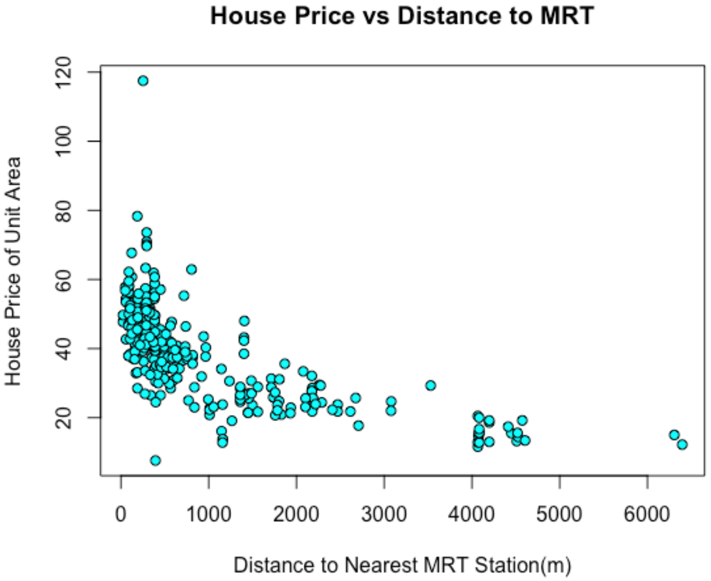
We consider the following research questions: Does the distance between a house and its nearest Metro Rapid Transit(MRT) station significantly affect the house's value? Is a model containing more than one predictor from our data set useful in predicting a house's value?

Regression Method

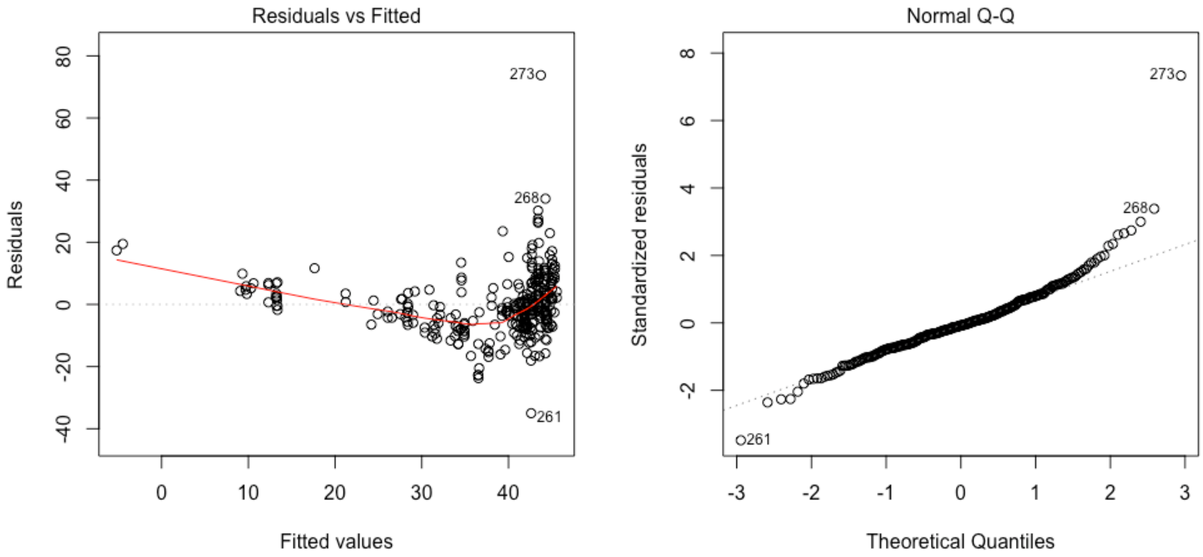
To answer our first question, we start by plotting house value against distance to nearest MRT. After analyzing this relationship, we develop a model to test all predictors.

Regression Analysis: House Value vs. Distance to Nearest MRT

The distance to nearest MRT is measured in meters. This scatterplot shows house value on distance.



Based on the scatterplot, we can identify that our assumptions of linearity, normality, and equal variances may be violated.



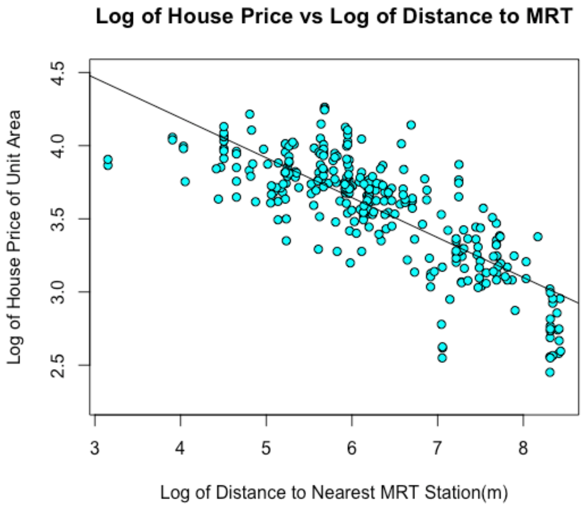
LINE Diagnostics

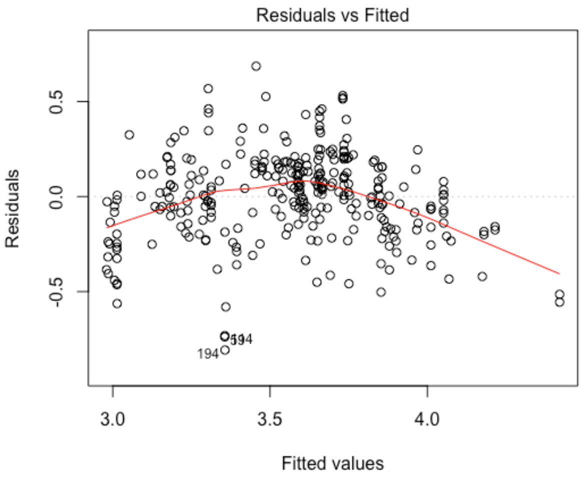
While conducting diagnostics, we conclude the LINE conditions are not all met. There is a large cluster on the Residuals vs Fitted graph which violates linearity and equal variances. The Normal Q-Q plot appears somewhat linear, not violating normality. To resolve the issues of linearity and equal variances, we must perform a transformation on our data. We perform a transformation and discover that the best regression is $\log(x)$ on $\log(y)$.

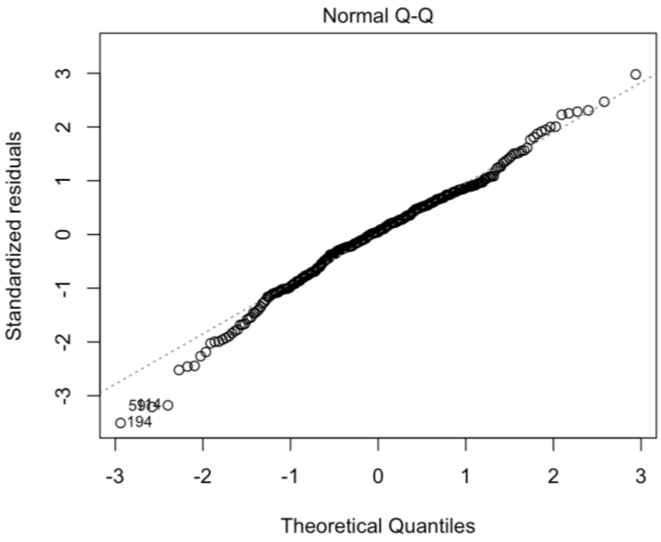
Outliers and Influential Points

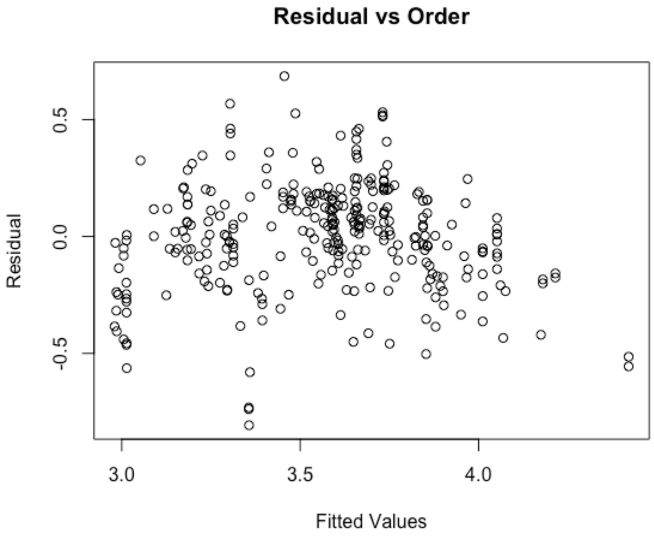
After conducting LINE diagnostics, we can notice there are potential outliers such as observation 261 and 273. Therefore, we conduct the Difference in Fits measurement to find and identify any influential points. After conducting DFFITS diagnostics, we choose to remove 6 observations that were signaled as influential.

Transformed Data with No Influential Points









From the scatter plot above, we see an existing linear relationship between transformed house price of unit area and transformed distance to nearest MRT. Its supporting Residuals vs Fitted plot shows a random scatter of points, resolving our violated assumptions from the non-transformed model. Also, our Normal Q-Q plot looks like it resembles normal data and appears more normal than our non-transformed Normal Q-Q Plot. Finally, our Residual vs. Order plot ensures there is no positive serial correlation among the error terms, therefore concluding that our errors are independent.

Research Question 1: *Does the distance between a house and its nearest Metro Rapid Transit(MRT) station significantly affect the house's value?*

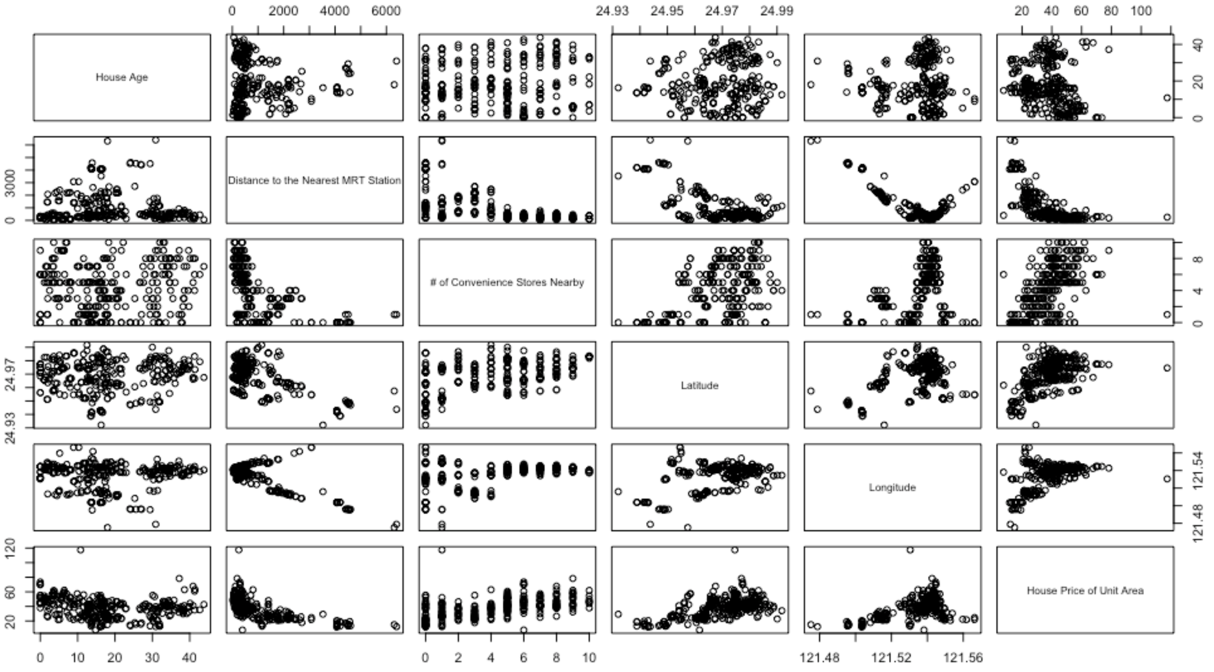
To answer our question, we conduct the following hypothesis test:

$$H_0: \beta_{\log(Distance)} = 0 \ vs \ H_1: \beta_{\log(Distance)} \neq 0$$

After constructing an ANOVA table of the transformed model, we are given a large F value of 502.78 as well as a small p-value of 2e-16. Since our p-value is less than $\alpha = 0.05$ significant value, we have evidence to reject the null hypothesis. **Thus, we can conclude that distance between a house and MRT is statistically significant in determining the house price per unit area.** In our model, the slope estimate for $\beta_{\log(Distance)}$ is -0.27263. For every increase in log of Distance to the Nearest MRT Station, the log of house price decreases by -.027263.

Regression Analysis: Constructing a Multi-Linear Model for House Price per Unit Area

We conduct a scatterplot matrix to observe for any correlation between House Price and all predictors in the dataset. These predictors include House Age, Distance to the Nearest MRT Station, # of Convenience Stores Nearby, Latitude, and Longitude.



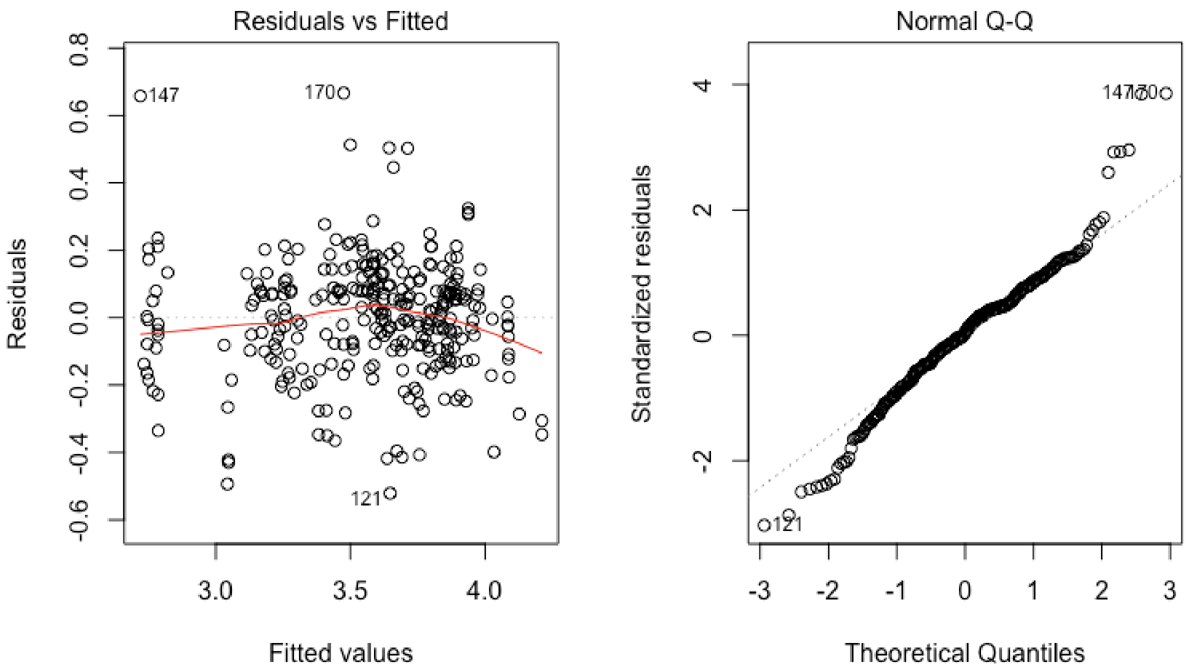
Observations: At a quick glance, some of these predictors could be useful for predicting house value. In order to figure out which predictors would make the best model, we must run stepwise regression.

Research Question: *Is a model containing more than one predictor from our data set useful in predicting a house's value?*

Regression Method: Stepwise Regression

We use Akaike's Information Criterion to perform stepwise regression. We are left with a model of all five of our predictors after performing stepwise computations. Another method we can use to develop a model is by comparing the Mallows' C_p value of different equations. However, we are not going to compare C_p values because C_p always equals p for the largest model containing all the predictors. Making sure there are no issues of multi-collinearity, we calculate the Variance Inflation Factor for each predictor. After calculating the VIF, none of them are above the value of 5 signaling that there are no instances of multi-collinearity.

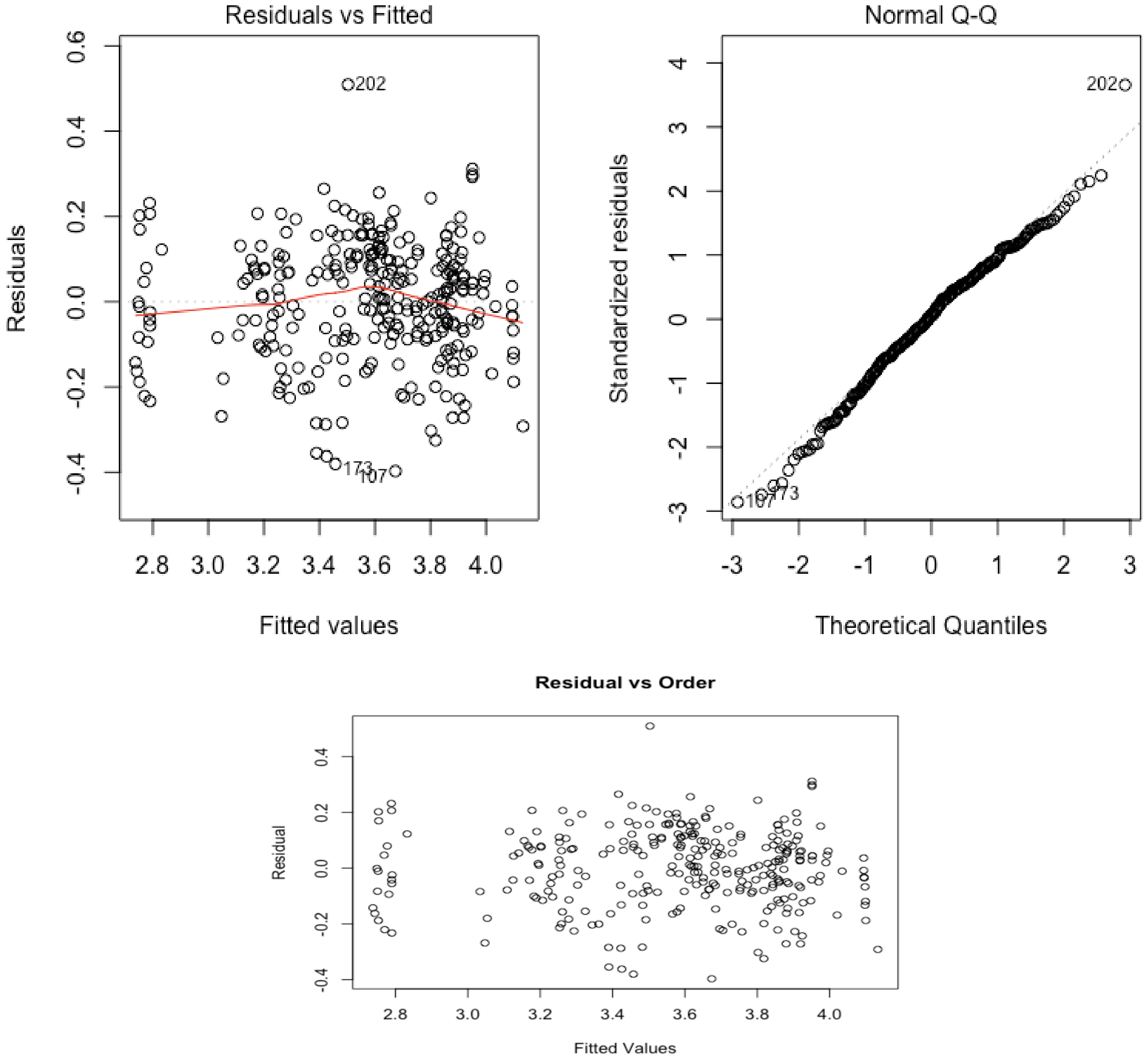
Residuals vs Fitted and Normal Q-Q with Influential Points



After constructing the models, we looked at our LINE Conditions (Graph Above). The Residuals vs. Fitted Values plot has a small cluster in the middle, but it still looks to bounce around 0 randomly. Our Normal Q-Q plot looks linear as well, however it seems like there are too many positive and negative residuals which means the distribution is heavy-tailed. To see if influential points are causing these problems, we ran a DFFITS Test. Based on the DFFITS Test, there are 16 points that are flagged as influential. We decided to remove these points and re-fit the model with data not containing the influential points.

We ran diagnostics again after refitting the line. Based on the graphs below, the Residuals vs. Fitted Plot is less clustered and looks to bounce around 0 more randomly. Also for the Normal Q-Q plot, we do not have the same heavy-tails that were present in the diagnostic with the influential points. Again, our Residual vs Order plot shows the errors are independent. Our diagnostics look relatively strong enough to hold the assumptions of linearity and equal variance of errors. **Thus, based on stepwise regression, all five predictors are significant in predicting the House's Price.**

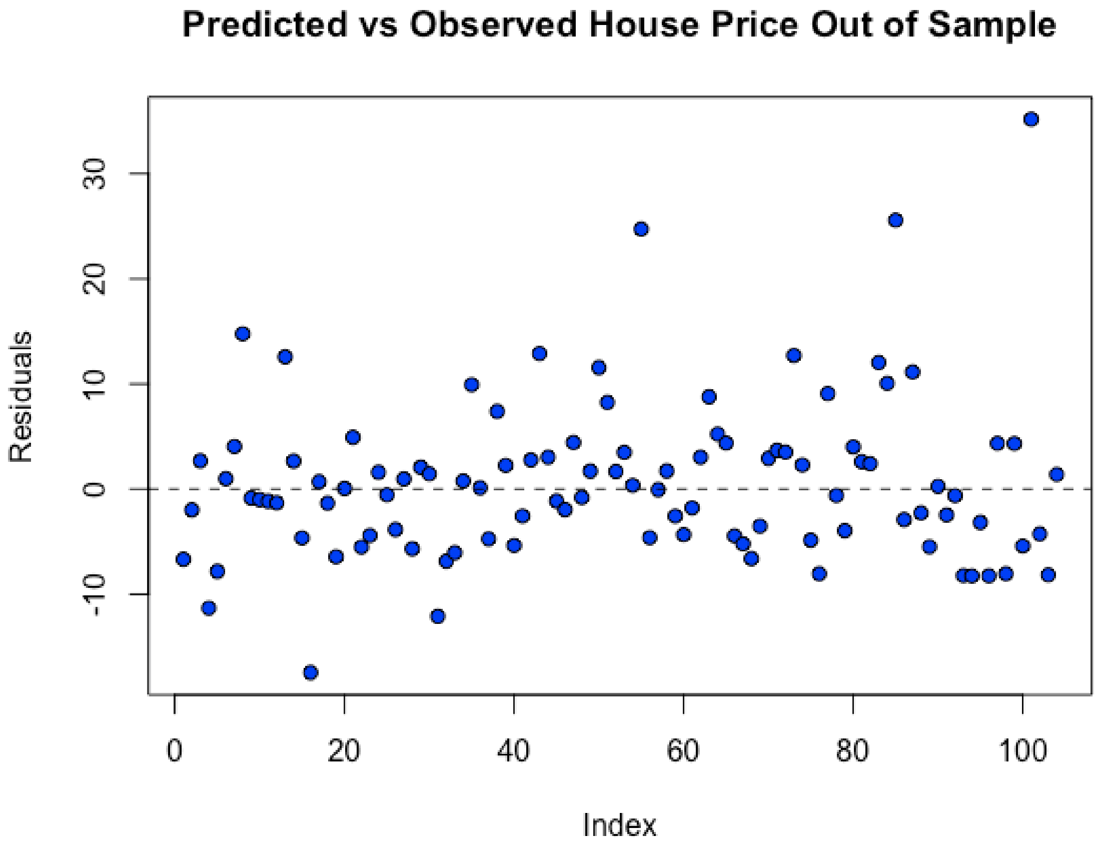
Residuals vs Fitted, Normal Q-Q, and Residuals vs Order without Influential Points



Final Model:

$$\log(Y_i) = \beta_0 + \beta_{House\ Age} x_{i,House\ Age} + \beta_{\log(Distance)} x_{i,\log(Distance)} + \beta_{Convenience\ Stores} x_{iConvenience\ Stores} + \beta_{Latitude} x_{i,Latitude} + \beta_{Longitude} x_{i,Longitude} + \varepsilon_i$$

We will now use the final model to predict our Out of Sample data, which are Houses' Prices between May 2013 and July 2013 in New Taipei City.



Based on the residual plot of Predicted vs Observed of Out of Sample data, our final model does a good job of predicting the House Price per unit area. The errors are small, mostly less than ± 10 and the errors bounce around 0 randomly. We do not know any other specifics of the house, therefore we cannot form a conclusion to why and how the outliers on residual plot came about.

Conclusion

In summary, we have found a model that predicts the House Price per unit area for houses in New Taipei City, Taiwan. Based on our first research question, the distance to the nearest MRT station is negatively correlated with the House Price. One reason this is true is because the farther away the house is from the MRT, the more likely the house is in a rural area, where prices are cheaper per unit area. For example, the actual price of a large farm in a rural county may be more expensive than an apartment in the downtown area. However, since we are looking at the price per unit area, the apartment may be more expensive per unit area.

In regards of the full final model, all predictors were deemed as significant in predicting the House's Price per unit area. This makes sense as well and was in line with our expectations. The House's Age is significant as newer houses are generally more expensive. The Longitude and Latitude are significant as that determines were the house is, the house price per unit area will always be based on where the house actually is. The Number of Convenience Stores Nearby is significant as many home seekers look at accessibility to markets and stores as an important factor, especially if they do not have a car.

Other predictors such as crime statistics, school district rankings, and unemployment level could have been significant for the model. In conclusion, our model is sufficient in predicting the House Price per unit area for houses in in Sindian Dist., New Taipei City, Taiwan.

Appendix

```
> library(car)
> library(leaps)
>
>
> ##To load data and create In vs. Out of Sample Data
> house_data = read.csv("Real estate valuation data set.csv")
> house_data = subset(house_data, select  = -c(No))
> house_data_original = house_data
> house_data = na.omit(house_data)
>
> colnames(house_data) <- c("Transaction", "House Age", "Distance to the Nearest MRT
Station", "# of Convenience Stores Nearby", "Latitude","Longitude", "House Price of U
nit Area")
>
> house_data <- house_data[order(house_data$Transaction),]
> rownames(house_data) <- 1:nrow(house_data)
> house_data_out_of_sample <- house_data[311:414,]
> house_data <- house_data[0:310,]
>
>
> house_data= subset(house_data, select = -c(Transaction))
>
> house_data_with_influential <- house_data
>
>
>
>
> ##Test with Distance as Predictor
> plot(((house_data$`Distance to the Nearest MRT Station`)), (house_data$`House Price
of Unit Area`),
+      main = "House Price vs Distance to MRT",xlab = "Distance to Nearest MRT Statio
n(m)",
+      ylab = "House Price of Unit Area", col = "black", pch = 21, bg = "cyan")
> mod= lm((house_data$`House Price of Unit Area`) ~ house_data$`Distance to the Neare
st MRT Station`)
> par(mfrow = c(1,2))
> plot(aov(mod))
Hit <Return> to see next plot:
Hit <Return> to see next plot:
>
> par(mfrow = c(1,1))
> mod.dffits = dffits(mod)
> dffits.influence.mod = 2 * sqrt((2+1)/(length(house_data$`House Age`) -2 -1))
> dffits_influential_obs = which(abs(mod.dffits) > dffits.influence.mod)
>
> #Take out influential points
> house_data <- house_data_with_influential[-dffits_influential_obs,]
>
> ##New Transformed Model with no Influential Points
>
```

```
> new.mod = lm(log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance
to the Nearest MRT Station`))
> par(mfrow = c(1,2))
> plot(log(house_data$`Distance to the Nearest MRT Station`),log(house_data$`House Pr
ice of Unit Area`),
+      main = "Log of House Price vs Log of Distance to MRT",xlab = "Log of Distance
to Nearest MRT Station(m)",ylim = c(2.25,4.5),
+      ylab = "Log of House Price of Unit Area", col = "black", pch = 21, bg = "cyan"
)
> abline(new.mod)
>
> plot(aov(new.mod))
Hit <Return> to see next plot: summary(new.mod)
Hit <Return> to see next plot: anova(new.mod)
>
> yhat.mod = fitted(new.mod)
> e.mod = log(house_data$`House Price of Unit Area`) - yhat.mod
> plot(yhat.mod, e.mod,  xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual
vs Order')
>
> ##Scatterplot and Correlation Matrix
> pairs(house_data)
> cor(house_data)
                                      House Age Distance to the Nearest MRT Station #
of Convenience Stores Nearby    Latitude
House Age                            1.00000000                         -0.02609305
0.07918783  0.07452151
Distance to the Nearest MRT Station -0.02609305                          1.00000000
-0.59330286 -0.65833265
# of Convenience Stores Nearby       0.07918783                         -0.59330286
1.00000000  0.45350495
Latitude                             0.07452151                         -0.65833265
0.45350495  1.00000000
Longitude                            0.01644948                         -0.75144530
0.42431910  0.41005801
House Price of Unit Area            -0.22773323                         -0.73361859
0.60992393  0.60490316
                                      Longitude House Price of Unit Area
House Age                            0.01644948               -0.2277332
Distance to the Nearest MRT Station -0.75144530               -0.7336186
# of Convenience Stores Nearby       0.42431910                0.6099239
Latitude                             0.41005801                0.6049032
Longitude                            1.00000000                0.5322850
House Price of Unit Area             0.53228497                1.0000000
>
> mod.full.log <- lm(log(house_data$`House Price of Unit Area`) ~ (house_data$`House
Age`) + log(house_data$`Distance to the Nearest MRT Station`)+
+                      (house_data$`# of Convenience Stores Nearby`) + (house_data$La
titude) +(house_data$Longitude))
>
> vif(mod.full.log)
                               house_data$`House Age` log(house_data$`Distance to the
Nearest MRT Station`)
                                             1.036392
2.421160
```

```
          house_data$`# of Convenience Stores Nearby`
house_data$Latitude
                                             1.913123
1.399200
                                 house_data$Longitude
                                             1.614318
>
> mod.reduced <- lm(log(house_data$`House Price of Unit Area`) ~ 1)
> step(mod.reduced, scope = list(lower = mod.reduced, upper = mod.full.log))
Start:  AIC=-593.77
log(house_data$`House Price of Unit Area`) ~ 1

                                                        Df Sum of Sq    RSS     AIC
+ log(house_data$`Distance to the Nearest MRT Station`)  1   26.7581 16.073 -889.73
+ house_data$Latitude                                    1   19.1913 23.639 -772.45
+ house_data$`# of Convenience Stores Nearby`            1   16.6405 26.190 -741.30
+ house_data$Longitude                                   1   14.9710 27.860 -722.51
+ house_data$`House Age`                                 1    1.5455 41.285 -602.94
<none>                                                               42.831 -593.77

Step:  AIC=-889.73
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest
MRT Station`)

                                                        Df Sum of Sq    RSS     AIC
+ house_data$Latitude                                    1    4.8037 11.269 -995.67
+ house_data$`House Age`                                 1    0.9348 15.138 -905.95
+ house_data$Longitude                                   1    0.9252 15.147 -905.76
+ house_data$`# of Convenience Stores Nearby`            1    0.7262 15.346 -901.79
<none>                                                               16.073 -889.73
- log(house_data$`Distance to the Nearest MRT Station`)  1   26.7581 42.831 -593.77

Step:  AIC=-995.67
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest
MRT Station`) + 
    house_data$Latitude

                                                        Df Sum of Sq    RSS      AIC
+ house_data$`House Age`                                 1    1.4979  9.771 -1037.03
+ house_data$Longitude                                   1    0.3351 10.934 -1002.85
+ house_data$`# of Convenience Stores Nearby`            1    0.1603 11.109  -998.03
<none>                                                               11.269  -995.67
- house_data$Latitude                                    1    4.8037 16.073  -889.73
- log(house_data$`Distance to the Nearest MRT Station`)  1   12.3705 23.640  -772.45

Step:  AIC=-1037.03
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest
MRT Station`) + 
    house_data$Latitude + house_data$`House Age`

                                                        Df Sum of Sq     RSS      AIC
+ house_data$Longitude                                   1    0.3970  9.3740 -1047.64
+ house_data$`# of Convenience Stores Nearby`            1    0.3246  9.4463 -1045.31
<none>                                                                9.7710 -1037.03
- house_data$`House Age`                                 1    1.4979 11.2689  -995.67
- house_data$Latitude                                   1    5.3669 15.1379  -905.95
- log(house_data$`Distance to the Nearest MRT Station`)  1   11.3910 21.1620  -804.11

Step:  AIC=-1047.64
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest 
MRT Station`) + 
    house_data$Latitude + house_data$`House Age` + house_data$Longitude

                                                       Df Sum of Sq     RSS      AIC
+ house_data$`# of Convenience Stores Nearby`           1    0.3248  9.0492 -1056.36
<none>                                                               9.3740 -1047.64
- house_data$Longitude                                  1    0.3970  9.7710 -1037.03
- house_data$`House Age`                                1    1.5598 10.9338 -1002.85
- house_data$Latitude                                   1    4.7182 14.0922  -925.71
- log(house_data$`Distance to the Nearest MRT Station`)  1    6.7318 16.1058  -885.11

Step:  AIC=-1056.36
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest 
MRT Station`) + 
    house_data$Latitude + house_data$`House Age` + house_data$Longitude + 
    house_data$`# of Convenience Stores Nearby`

                                                       Df Sum of Sq     RSS      AIC
<none>                                                               9.0492 -1056.36
- house_data$`# of Convenience Stores Nearby`           1    0.3248  9.3740 -1047.64
- house_data$Longitude                                  1    0.3972  9.4463 -1045.31
- house_data$`House Age`                                1    1.7268 10.7760 -1005.27
- log(house_data$`Distance to the Nearest MRT Station`)  1    3.6287 12.6779  -955.86
- house_data$Latitude                                   1    4.0838 13.1330  -945.14

Call:
lm(formula = log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to 
the Nearest MRT Station`) + 
    house_data$Latitude + house_data$`House Age` + house_data$Longitude + 
    house_data$`# of Convenience Stores Nearby`)

Coefficients:
                                          (Intercept)  log(house_data$`Distance to th
e Nearest MRT Station`)  
                                           -6.958e+02  
-1.562e-01  
                                  house_data$Latitude  
house_data$`House Age`  
                                            1.181e+01  
-6.562e-03  
                                 house_data$Longitude            house_data$`# of Con
venience Stores Nearby`  
                                            3.337e+00  
1.526e-02  

> 
> 
> ##Stepwise, Best Subsets
> 
> 
```

```
> ##Stepwise
> mod0 = lm(log(house_data$`House Price of Unit Area`) ~ 1)
> mod.upper = lm(log(house_data$`House Price of Unit Area`) ~ (house_data$`House Age`
) + log(house_data$`Distance to the Nearest MRT Station`)+
+                      (house_data$`# of Convenience Stores Nearby`) + (house_data$Lat
itude) + house_data$Longitude)
> step(mod0, scope = list(lower = mod0, upper = mod.upper))
Start:  AIC=-593.77
log(house_data$`House Price of Unit Area`) ~ 1

                                                        Df Sum of Sq    RSS     AIC
+ log(house_data$`Distance to the Nearest MRT Station`)  1   26.7581 16.073 -889.73
+ house_data$Latitude                                    1   19.1913 23.639 -772.45
+ house_data$`# of Convenience Stores Nearby`            1   16.6405 26.190 -741.30
+ house_data$Longitude                                   1   14.9710 27.860 -722.51
+ house_data$`House Age`                                 1    1.5455 41.285 -602.94
<none>                                                               42.831 -593.77

Step:  AIC=-889.73
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest 
MRT Station`)

                                                        Df Sum of Sq    RSS     AIC
+ house_data$Latitude                                    1    4.8037 11.269 -995.67
+ house_data$`House Age`                                 1    0.9348 15.138 -905.95
+ house_data$Longitude                                   1    0.9252 15.147 -905.76
+ house_data$`# of Convenience Stores Nearby`            1    0.7262 15.346 -901.79
<none>                                                               16.073 -889.73
- log(house_data$`Distance to the Nearest MRT Station`)  1   26.7581 42.831 -593.77

Step:  AIC=-995.67
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest 
MRT Station`) + 
    house_data$Latitude

                                                        Df Sum of Sq    RSS      AIC
+ house_data$`House Age`                                 1    1.4979  9.771 -1037.03
+ house_data$Longitude                                   1    0.3351 10.934 -1002.85
+ house_data$`# of Convenience Stores Nearby`            1    0.1603 11.109  -998.03
<none>                                                               11.269  -995.67
- house_data$Latitude                                    1    4.8037 16.073  -889.73
- log(house_data$`Distance to the Nearest MRT Station`)  1   12.3705 23.640  -772.45

Step:  AIC=-1037.03
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest 
MRT Station`) + 
    house_data$Latitude + house_data$`House Age`

                                                        Df Sum of Sq     RSS      AIC
+ house_data$Longitude                                   1    0.3970  9.3740 -1047.64
+ house_data$`# of Convenience Stores Nearby`            1    0.3246  9.4463 -1045.31
<none>                                                                9.7710 -1037.03
- house_data$`House Age`                                 1    1.4979 11.2689  -995.67
- house_data$Latitude                                    1    5.3669 15.1379  -905.95
- log(house_data$`Distance to the Nearest MRT Station`)  1   11.3910 21.1620  -804.11
```

```
Step:  AIC=-1047.64
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest
MRT Station`) +
    house_data$Latitude + house_data$`House Age` + house_data$Longitude

                                                        Df Sum of Sq     RSS      AIC
+ house_data$`# of Convenience Stores Nearby`            1    0.3248  9.0492 -1056.36
<none>                                                                9.3740 -1047.64
- house_data$Longitude                                   1    0.3970  9.7710 -1037.03
- house_data$`House Age`                                 1    1.5598 10.9338 -1002.85
- house_data$Latitude                                    1    4.7182 14.0922  -925.71
- log(house_data$`Distance to the Nearest MRT Station`)  1    6.7318 16.1058  -885.11

Step:  AIC=-1056.36
log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to the Nearest
MRT Station`) +
    house_data$Latitude + house_data$`House Age` + house_data$Longitude +
    house_data$`# of Convenience Stores Nearby`

                                                        Df Sum of Sq     RSS      AIC
<none>                                                                9.0492 -1056.36
- house_data$`# of Convenience Stores Nearby`            1    0.3248  9.3740 -1047.64
- house_data$Longitude                                   1    0.3972  9.4463 -1045.31
- house_data$`House Age`                                 1    1.7268 10.7760 -1005.27
- log(house_data$`Distance to the Nearest MRT Station`)  1    3.6287 12.6779  -955.86
- house_data$Latitude                                    1    4.0838 13.1330  -945.14

Call:
lm(formula = log(house_data$`House Price of Unit Area`) ~ log(house_data$`Distance to
the Nearest MRT Station`) +
    house_data$Latitude + house_data$`House Age` + house_data$Longitude +
    house_data$`# of Convenience Stores Nearby`)

Coefficients:
                                          (Intercept)  log(house_data$`Distance to th
e Nearest MRT Station`)
                                           -6.958e+02
-1.562e-01
                                  house_data$Latitude
house_data$`House Age`
                                            1.181e+01
-6.562e-03
                                 house_data$Longitude            house_data$`# of Con
venience Stores Nearby`
                                            3.337e+00
1.526e-02

>
>
> ##Best Subsets
> mod = regsubsets(cbind(house_data$`House Age`,log(house_data$`Distance to the Neare
st MRT Station`),
+                  house_data$`# of Convenience Stores Nearby`, house_data$Latitude,h
ouse_data$Longitude),log(house_data$`House Price of Unit Area`))
```

```
> summary.mod = summary(mod)
> summary.mod$which
  (Intercept)     a    b     c     d     e
1        TRUE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE TRUE FALSE  TRUE FALSE
3        TRUE  TRUE TRUE FALSE  TRUE FALSE
4        TRUE  TRUE TRUE FALSE  TRUE  TRUE
5        TRUE  TRUE TRUE  TRUE  TRUE  TRUE
> summary.mod$adjr2
[1] 0.6234979 0.7351485 0.7695890 0.7782107 0.7851778
> 
> summary.mod$cp
[1] 229.29183  73.09888  25.76973  14.69710   6.00000
> summary.mod$which
  (Intercept)     a    b     c     d     e
1        TRUE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE TRUE FALSE  TRUE FALSE
3        TRUE  TRUE TRUE FALSE  TRUE FALSE
4        TRUE  TRUE TRUE FALSE  TRUE  TRUE
5        TRUE  TRUE TRUE  TRUE  TRUE  TRUE
> 
> 
> 
> 
> 
> ##Testing Model with all predictors, with stepwise if one or more predictors is sig
nificant
> mod.full = lm(log(house_data$`House Price of Unit Area`) ~ (house_data$`House Age`)
+
+                  log(house_data$`Distance to the Nearest MRT Station`)+
+                  (house_data$`# of Convenience Stores Nearby`) + (house_data$Latitud
e) + house_data$Longitude)
> mod.reduced <- lm(log(house_data$`House Price of Unit Area`) ~ 1)
> 
> vif(mod.full)
                              house_data$`House Age` log(house_data$`Distance to the
Nearest MRT Station`)
                                            1.036392
2.421160
          house_data$`# of Convenience Stores Nearby`
house_data$Latitude
                                            1.913123
1.399200
                                house_data$Longitude
                                            1.614318
> 
> anova(mod.reduced,mod.full)
Analysis of Variance Table

Model 1: log(house_data$`House Price of Unit Area`) ~ 1
Model 2: log(house_data$`House Price of Unit Area`) ~ (house_data$`House Age`) +
    log(house_data$`Distance to the Nearest MRT Station`) + (house_data$`# of Conveni
ence Stores Nearby`) +
    (house_data$Latitude) + house_data$Longitude
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
1    303 42.831
2    298  9.049  5    33.782 222.49 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> par(mfrow = c(1,2))
> plot(aov(mod.full))
Hit <Return> to see next plot:
Hit <Return> to see next plot: ## we decide to remove influential points again
> mod.dffits.full = dffits(mod.full)
> dffits.influence.mod.full = 2 * sqrt((6+1)/(length(house_data$`House Age`) -6 -1))
> dffits_influential_obs.full = which(abs(mod.dffits.full) > dffits.influence.mod.ful
l)
> house_data_full_mod_no_inf <- house_data[-dffits_influential_obs.full,]
>
> mod.full.no.inf = lm(log(house_data_full_mod_no_inf$`House Price of Unit Area`) ~ (
house_data_full_mod_no_inf$`House Age`) +
+                         log(house_data_full_mod_no_inf$`Distance to the Nearest MRT
Station`)+
+                         (house_data_full_mod_no_inf$`# of Convenience Stores Nearby`
)
+                       + (house_data_full_mod_no_inf$Latitude) + house_data_full_mod_
no_inf$Longitude)
> mod.reduced.no.inf <- lm(log(house_data_full_mod_no_inf$`House Price of Unit Area`)
~ 1)
>
> vif(mod.full.no.inf)
                               house_data_full_mod_no_inf$`House Age`
                                                             1.033498
log(house_data_full_mod_no_inf$`Distance to the Nearest MRT Station`)
                                                             2.490143
          house_data_full_mod_no_inf$`# of Convenience Stores Nearby`
                                                             1.886543
                                  house_data_full_mod_no_inf$Latitude
                                                             1.385959
                                 house_data_full_mod_no_inf$Longitude
                                                             1.612192
>
> anova(mod.reduced.no.inf,mod.full.no.inf)
Analysis of Variance Table

Model 1: log(house_data_full_mod_no_inf$`House Price of Unit Area`) ~
    1
Model 2: log(house_data_full_mod_no_inf$`House Price of Unit Area`) ~
    (house_data_full_mod_no_inf$`House Age`) + log(house_data_full_mod_no_inf$`Distan
ce to the Nearest MRT Station`) +
        (house_data_full_mod_no_inf$`# of Convenience Stores Nearby`) +
        (house_data_full_mod_no_inf$Latitude) + house_data_full_mod_no_inf$Longitude
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    287 36.673
2    282  5.527  5    31.146 317.84 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> par(mfrow = c(1,2))
> plot(aov(mod.full.no.inf))
Hit <Return> to see next plot:
```

```
Hit <Return> to see next plot: yhat.mod.full.no.inf = fitted(mod.full.no.inf)
> e.mod.full.no.inf = log(house_data_full_mod_no_inf$`House Price of Unit Area`) - yh
at.mod.full.no.inf
Error: object 'yhat.mod.full.no.inf' not found
> plot(yhat.mod.full.no.inf, e.mod.full.no.inf,  xlab = 'Fitted Values', ylab = 'Resi
dual',
+      main = 'Residual vs Order')
Error in plot(yhat.mod.full.no.inf, e.mod.full.no.inf, xlab = "Fitted Values",  :
  object 'yhat.mod.full.no.inf' not found
>
>
>
>
>
>
>
> ##Testing observed vs predicted in in sample data
> estimate_response.in = fitted(mod.full.no.inf)
> estimate_response.in = exp(estimate_response.in)
>
> residual.in = house_data_full_mod_no_inf$`House Price of Unit Area` - estimate_resp
onse.in
> plot( (1:nrow(house_data_full_mod_no_inf)), residual.in,
+       main = "Predicted vs Observed House Price In Sample", ylab = 'Residuals',
+       xlab= "Index", pch = 21, bg = 'red')
>
> #Testing observed vs predicted in Out of Sample Data
> mod.full.out =  lm(log(house_data_out_of_sample$`House Price of Unit Area`) ~ (hous
e_data_out_of_sample$`House Age`) +
+                                log(house_data_out_of_sample$`Distance to the Neares
t MRT Station`)+
+                                (house_data_out_of_sample$`# of Convenience Stores N
earby`) + (house_data_out_of_sample$Latitude) + house_data_out_of_sample$Longitude)
>
> estimate_response.out = fitted(mod.full.out)
> estimate_response.out = exp(estimate_response.out)
> residual.out = house_data_out_of_sample$`House Price of Unit Area`  - estimate_respo
nse.out
> plot( (1:nrow(house_data_out_of_sample)), residual.out,
+       main = "Predicted vs Observed House Price Out of Sample", ylab = 'Residuals',
+       xlab= "Index", pch = 21, bg = 'blue')
> abline(h = 0, lty = 2)
```