Regression Analysis of House Price Of Unit Area

Prepared for:
Sudeep Bapat
Instructor of PSTAT 126

Prepared by:
Shirley Wang
Zhongtian Jiang

December 11, 2019

Abstract
This project focuses on the performance of house price of unit area, based on the transaction date, house age, distance to the nearest MRT station, number of convenience stores, latitude and longitude. The data can be found from the UCI Machine Learning Repository. We will figure out if the house price of unit area can be predicted by selected factors. Moreover, we want to explore which is the most influential predictor on house price among these factors.

Question of Interest
We consider the following questions:
Question 1: What factors affect the price of the house most?
Question 2: What is the best set of factors to predict the price of the house?

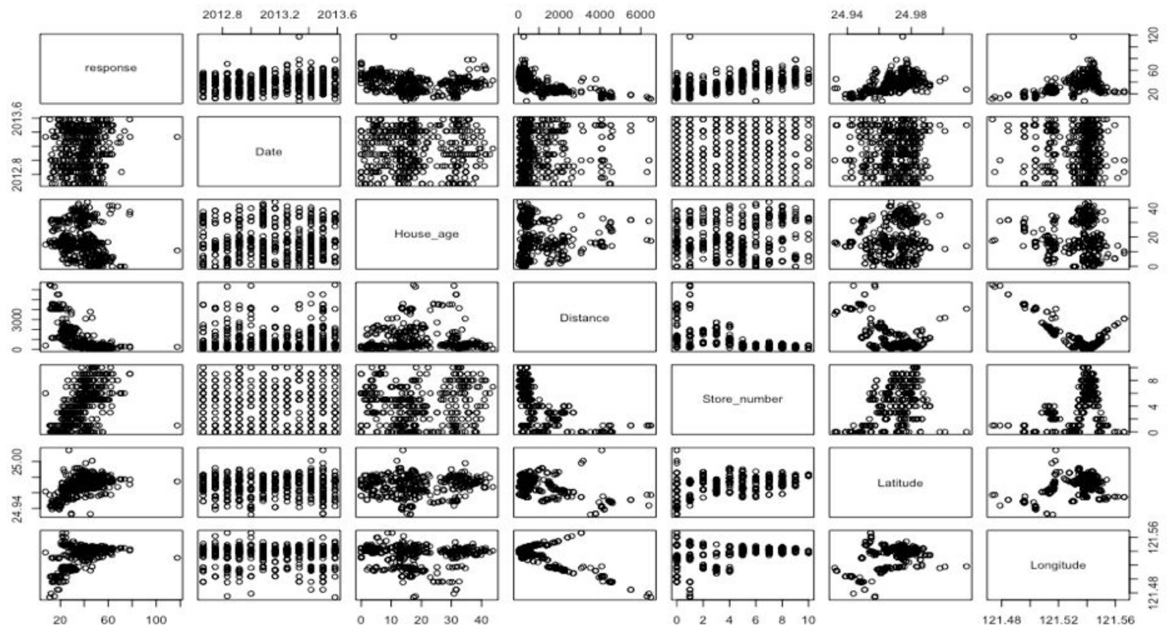
Regression Method
Firstly, we need to construct a model to meet all four LINE conditions to explore further relationship. We will first use means of residual analysis to figure out most important predictors in our models with their transformations if needed. Then we can judge which predictors are suitable for our model to explain the variation of house prices in unit area by using variable selection procedures.
After finding our best final model, we can answer some related research questions:
Question 1: Is there an association between House age and house price of unit area?
Question 2: What price do we expect a house with age 30 and with average values of the other predictors to have?

Regression Analysis, Results and Interpretation

Model Building Process

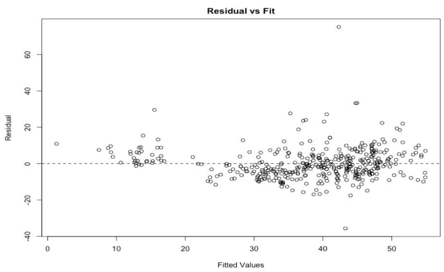
Define the variables:
To begin our analysis, we define our variables as follows:
response= house price of unit area
Date=transaction date
House_age=house age
Distance=distance to the nearest MRT station
Store_number=number of convenience store
Latitude=latitude
Longitude=longitude

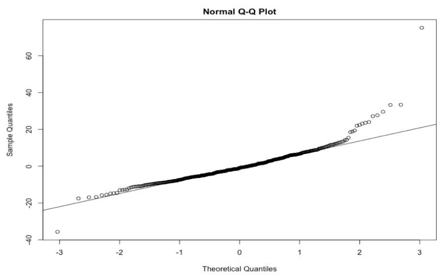
Check the Scatterplot Matrix of Data:
To begin, we plot each potential predictor (Date, House_age, Distance, Store_number, Latitude, Longitude) against the response variable (house price of unit area) using a pairs() function in R. Our results are as follows:

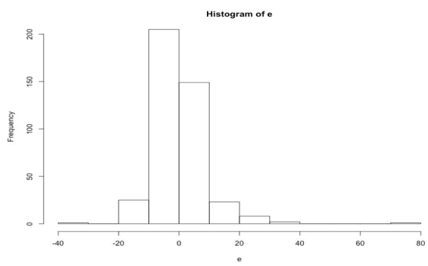


From the scatterplot matrix, we can see there is likely to exist positive linear relationships between house price of unit area and the Store_number, Latitude, Longitude as well as negative linear relationships between house price of unit area and house_age and distance.

Fit Model to check LINE conditions for the model:
We will get a first-order model on all the predictors:
$E(response) = \beta_0 + \beta_1 Date + \beta_2 House\ age + \beta_3 Distance + \beta_4 Store\ number + \beta_5 Latitude + \beta_6 Longitude$
We use Residuals vs. Fitted, Normal Q-Q and Histogram of the Residuals to check the "LINE" conditions for this model:
fit=lm(response~Date+House_age+Distance+Store_number+Latitude+Longitude)



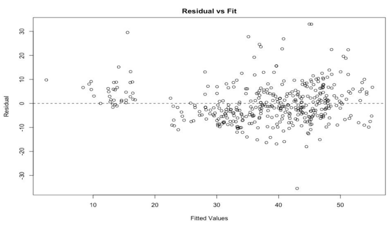


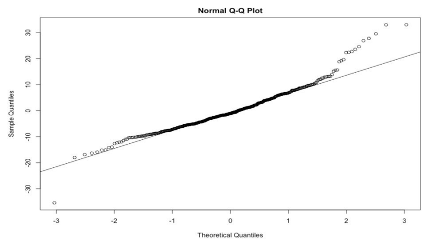


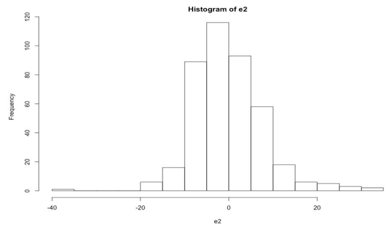
The Residuals vs. Fitted plot indicates that the linearity is violated since most of the points are focused on the right part of the graph which is not randomly distributed. Moreover, the Normal Q-Q plot and histogram of the residuals shows non-Normality of the error terms and outliers exists.
We can see from the Normal Q-Q plot, a huge outliers exists which will influence our final model a lot. We decide to use unstandardized deleted residuals to find it and delete it.

Redefine the variables and check LINE conditions again:
After deleting one huge outlier in our data, we update the data set and redefine the variables:
response2= house price of unit area

Date2=transaction date
House_age2=house age
Distance2=distance to the nearest MRT station
Store_number2=number of convenience store
Latitude2=latitude
Longitude2=longitude
Then we use lm() funciton to fit model fit2 and to recheck the LINE conditions by using Residuals vs. Fitted, Normal Q-Q and Histogram of the Residuals:
fit2=lm(response2~Date2+House_age2+Distance2+Store_number2+Latitude2+Longitude2)

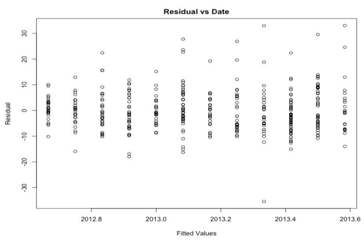


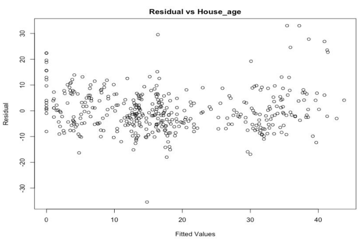


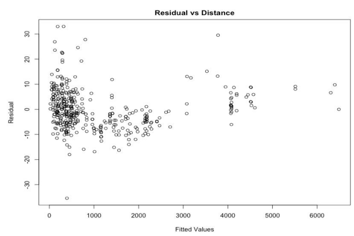


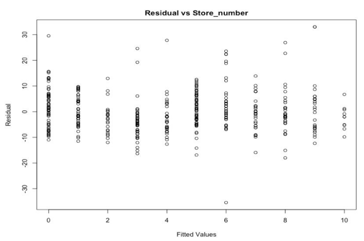
The Residuals vs. Fitted plot indicates that the linearity is still violated. The Normal Q-Q plot and histogram of the residuals shows non-Normality of the error terms and outliers exists. Although the huge outlier is deleted and the Normal Q-Q plot looks better than before, we still need to make some transformations to find the best model.

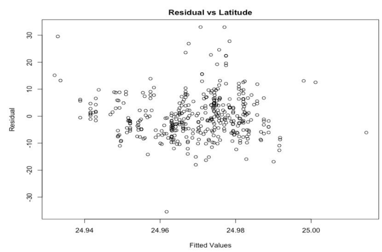
Data Transformations:
Since the fit2 model does not meet LINE conditions, we decide to use the natural logarithmic transformation to improve the regression model. It seems “everything” wrong here, we need to transform both the response Y and all/selected predictors(x) values. In order to find the predictor which need to be corrected, we use plot function to detect the most non-linearity predictor:

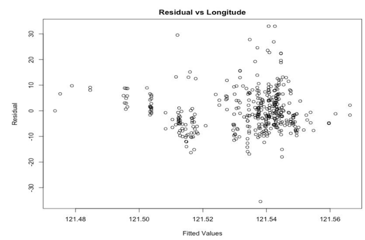




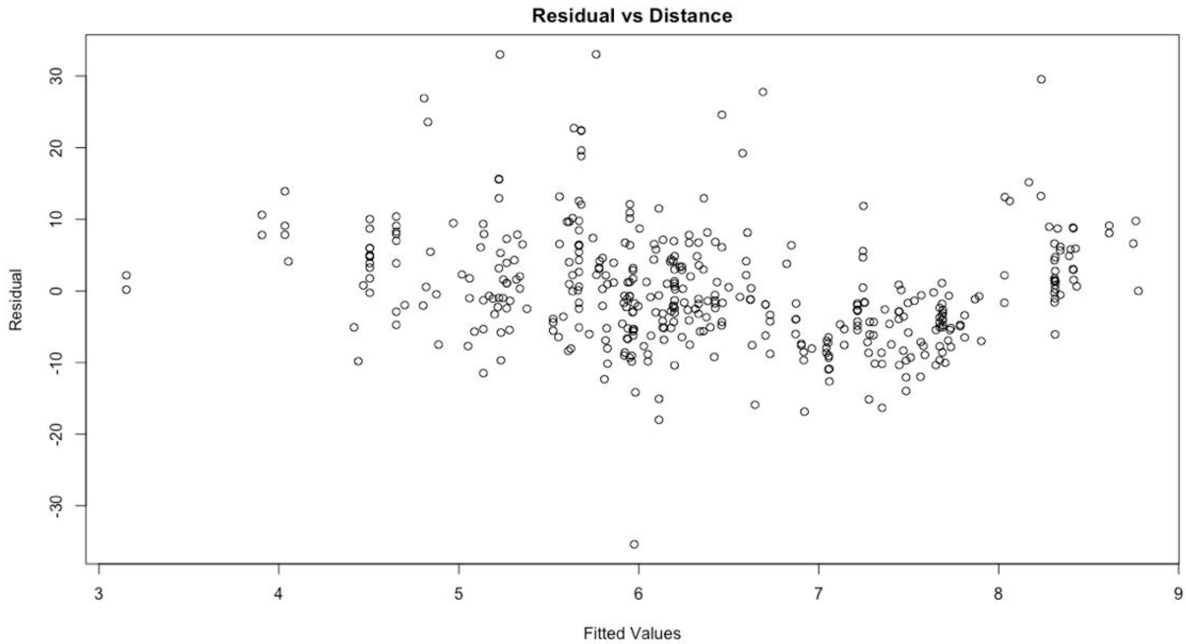




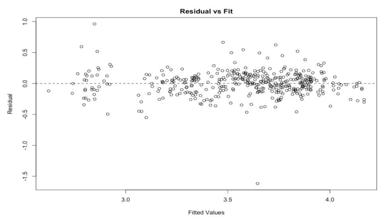


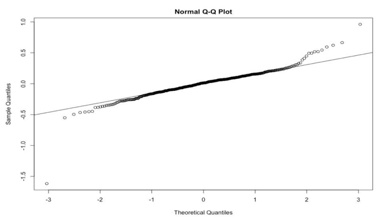


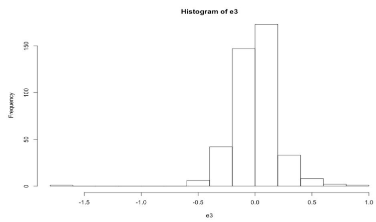
The above plots show that the nonlinearity seems to mostly come from the predictors Distance. We log-transform Distance and relook at the Residuals vs. Fitted plot:



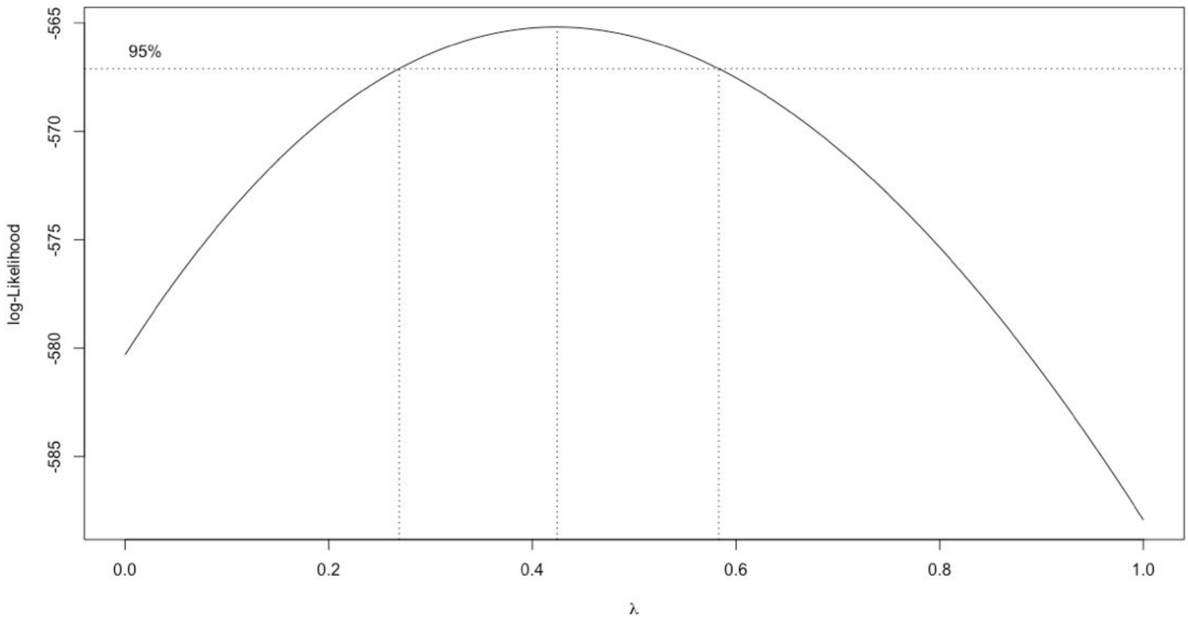
It seems better than before. Therefore, we play transformation on both response and predictor Distance. We fit a new first-order model fit3 to recheck the "LINE" condition. We still use Residuals vs. Fitted, Normal Q-Q and Histogram of the Residuals to check:
fit3=lm(log(response2)~Date2+House_age2+log(Distance2)+Store_number2+Latitude2+Longitude2)



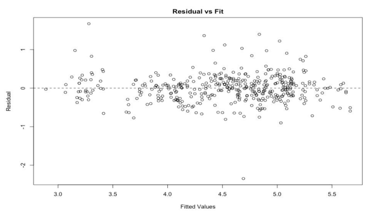


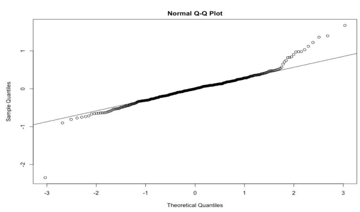


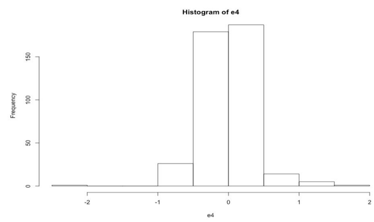
The Residuals vs. Fitted plot shows the linearity. However, the Normal Q-Q plot and histogram of the residuals still shows non-Normality of the error terms and outliers still exists. It implies that this transformation does not work well on Y. We use Box-Cox transformations to determine which transformation on Y to use:



According to the plot, we can find that lambda is around 0.42. We can get a new first-order model fit4 and use the same three plots to check the "LINE" condition:
fit4=lm((response2)^(0.42)~Date2+House_age2+log(Distance2)+Store_number2+Latitude2+Longitude2)



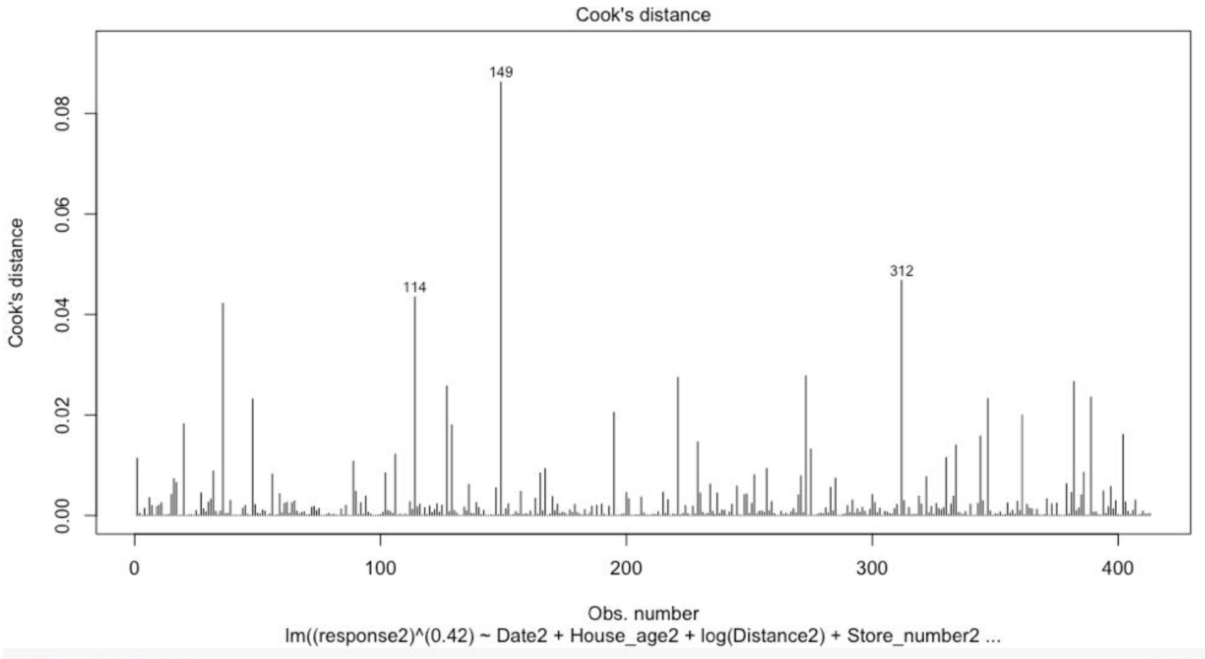




The fit4 model is better than the previous three. The model meets the four LINE conditions now.

Variable Selection
We use Stepwise Regression with F-tests and Akaike's Information Criterion (AIC) as criteria to determine the predictors in the final model. For both methods, our standing model includes log(Distance2) and House_age2 since these two predictors are important to our research questions.
We first check the p-value for the F-test for rest predictors to check their prediction availabilities and then we apply AIC to compare each model.
Finally we get:
$E(reponse2^{0.42}) = \beta_0 + \beta_1 Date2 + \beta_2 House\ age2 + \beta_3 log(Distance2) + \beta_4 Store\ number2 + \beta_5 \text{ Latitude2} + \beta_6 \text{ Lonitude2}$
AIC gives us the same model as Stepwise Regression did. The output of both methods will show in the appendix.

Checking the Influential Points
Before building the final model, we will use Cook's Distance to detect any influential observations. To do this, we look at a plot that visualizes Cook's Distance of all observations:



The plot shows that all observations are well below the threshold 1, and most of them are below the threshold 0.5. This implies that there are only 1 points might be influential. So we have finished the model building process.

The Final Model

Our final model is given by

$E(reponse2^{0.42}) = \beta_0 + \beta_1 Date2 + \beta_2 House\ age2 + \beta_3 log(Distance2) + \beta_4 Store\ number2 + \beta_5$ Latitude2 $+ \beta_6$ Lonitude2

The summary output for this model can be found in the appendix.

Research Questions

Question 1: Is there an association between House age and house price of unit area?

The submodel for house price is

$E(reponse2^{0.42}) = \beta_0 + \beta_1 Date2 + \beta_2 House\ age2 + \beta_3 log(Distance2) + \beta_4 Store\ number2 + \beta_5$ Latitude2 $+ \beta_6$ Lonitude2

To answer this question, we merely test the null hypothesis H0: $\beta_2 = 0$ using either the F-test or the equivalent t-test. We use summary(fit4)$coefficients to find the p-value:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.448154e+03	2.251565e+02	-6.431765	3.551510e-10
Date2	2.986865e-01	6.476030e-02	4.612186	5.348354e-06
House_age2	-1.133801e-02	1.602632e-03	-7.074618	6.609708e-12
log(Distance2)	-3.140218e-01	2.675346e-02	-11.737615	1.360459e-27
Store_number2	2.658325e-02	8.725063e-03	3.046768	2.464193e-03
Latitude2	1.658667e+01	1.700508e+00	9.753947	2.471521e-20
Longitude2	3.614851e+00	1.569492e+00	2.303198	2.177234e-02

Since the p-value is very close to 0, we conclude that there is a linear association between the House_age and the house price of unit area at any α level, say 0.05 or 0.01.

Question 2: What price do we expect a house with age 30 and with average values of the other predictors to have?

To answer this question, we calculate a 95 percent prediction interval for a new response with predictor values House_age=30, and average values of log(Distance), Latitude, Longitude, Date and Store_number. This interval is given by

```
      fit      lwr      upr
1 5.810626 5.04948 6.571771
```

We then convert to original units of the response using the exponential:

```
      fit      lwr      upr
1 66.00522 47.24946 88.48362
```

Thus, we predict that a 30-year house with average values of log(Distance), Latitude, Longitude, Date and Store_number will have price of 66.005 of unit area and we are 95 percent confident that its price of unit area is between 47.249 and 88.484.

Conclusion

In conclusion, we get the same prediction model from our regression analysis and AIC. We can say house prices in unit area is influenced by the natural logarithm of distance to the nearest mrt station, the house age, latitude, longitude, the transaction date and number of convenience stores around the house, and among these predictors, the natural logarithm of distance to the nearest mrt station and the house age have strongest influences on house prices in unit area. We also determined we can be 95 percent confident that the 30-year house price of unit are between 47.249 and 88.484.

Since the year of our data ranges from 2012 to 2013, the house price is likely to be influenced by other social factors, our model is more accurate for transactions took place in 2002 and 2003 and our house price prediction can only be used for house price of unit area in 2012 and 2013.

Appendix

Relevant code and code outputs used in our regression analysis can be found below:

Code

```
#load packages
library(readxl)
library(leaps)
library(MASS)

#import data
RE<-read_xlsx("Real estate valuation data set.xlsx")

#make column names meaningful
Date<-RE$`X1 transaction date`
House_age<-RE$`X2 house age`
Distance<-RE$`X3 distance to the nearest MRT station`
Store_number<-RE$`X4 number of convenience stores`
Latitude<-RE$`X5 latitude`
Longitude<-RE$`X6 longitude`
response<-RE$`Y house price of unit area`

#fit the linear model(fit)
fit=lm(response~Date+House_age+Distance+Store_number+Latitude+Longitude)

#get the estimated regression equation
coef(fit)

#scatterplot matrix
pairs(response~Date+House_age+Distance+Store_number+Latitude+Longitude)

#summary of fit
summary(fit)

#Residual vs. Fitted plot
yhat = fitted(fit)
e = response - yhat
plot(yhat, e, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
abline(h = 0, lty = 2)

#Normal Q-Q plot
qqnorm(e)
qqline(e)
```

```
#Histogram of Residuals
hist(e)

#find out outliers
n=length(Distance)
outlier=rstudent(fit)
for (i in 1:n){
  if (outlier[i]>3){
    print (outlier[i])
  }
}

#update data
RE2<-read_xlsx("Real estate valuation data set_2.xlsx")

#redefine the column names
Date2<-RE2$`X1 transaction date`
House_age2<-RE2$`X2 house age`
Distance2<-RE2$`X3 distance to the nearest MRT station`
Store_number2<-RE2$`X4 number of convenience stores`
Latitude2<-RE2$`X5 latitude`
Longitude2<-RE2$`X6 longitude`
response2<-RE2$`Y house price of unit area`

#fit the linear model(fit2)
fit2=lm(response2~Date2+House_age2+Distance2+Store_number2+Latitude2+Longitude2)

#summary fit2
summary(fit2)

#Residuals vs. Fitted, Normal Q-Q plot, Histogram of Residuals (fit2)
yhat2=fitted(fit2)
e2=response2 - yhat2
plot(yhat2, e2, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
abline(h = 0, lty = 2)
qqnorm(e2)
qqline(e2)
hist(e2)

#check linearity of each predictor
plot(Date2, resid(fit2), xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Date')
plot(House_age2, resid(fit2), xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs
House_age')
plot(Distance2, resid(fit2), xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs
Distance')
plot(Store_number2, resid(fit2), xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs
Store_number')
```

plot(Latitude2, resid(fit2), xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Latitude')
plot(Longitude2, resid(fit2), xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Longitude')

#check linearity of log(Distance)
plot(log(Distance2), resid(fit2), xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Distance')

#fit the linear model(fit3)
fit3=lm(log(response2)~Date2+House_age2+log(Distance2)+Store_number2+Latitude2+Longitude2)

#summary fit3
summary(fit3)

#Residuals vs. Fitted, Normal Q-Q plot, Histogram of Residuals (fit3)
yhat3=fitted(fit3)
e3=log(response2) - yhat3
plot(yhat3, e3, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
abline(h = 0, lty = 2)
qqnorm(e3)
qqline(e3)
hist(e3)

#boxcox to find best lamdba value
boxcox.trans=boxcox(response2~Date2+House_age2+log(Distance2)+Store_number2+Latitude2+Longitude2,lambda = seq(0,1,length=10))

#fit the linear model(fit4)
fit4=lm((response2)^(0.42)~Date2+House_age2+log(Distance2)+Store_number2+Latitude2+Longitude2)

#summary fit4
summary(fit4)

#Residuals vs. Fitted, Normal Q-Q plot, Histogram of Residuals (fit4)
yhat4=fitted(fit4)
e4=response2^(0.42) - yhat4
plot(yhat4, e4, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
abline(h = 0, lty = 2)
qqnorm(e4)
qqline(e4)
hist(e4)

#Stepwise Regression using F-tests
mod0=lm((response2)^(0.42)~log(Distance2)+House_age2)

#check the p-value for the F-test for rest predictors
add1(mod0,~.+Date2+Store_number2+Latitude2+Longitude2,test='F')
#It is clear to see that F-statistic (p-value) of Latitude is the largest (smallest) one.
mod1=update(mod0,~.+Latitude2)
add1(mod1,~.+Date2+Store_number2+Longitude2,test='F')
#It is clear to see that F-statistic (p-value) of Date is the largest (smallest) one.
mod2=update(mod1,~.+Date2)
summary(mod2)
#The p-value for log(Distance), House_age and Latitude shows that adding Date in the model doesn't affect the significance of log(Distance), House_age and Latitude.
add1(mod2,~.+Store_number2+Longitude2,test='F')
#It is clear to see that F-statistic (p-value) of Store_number is the largest (smallest) one.
mod3=update(mod2,~.+Store_number2)
summary(mod3)
#The p-value for log(Distance), House_age, Latitude and Date shows that adding Store_number in the model doesn't affect the significance of log(Distance), House_age, Latitude and Date.
add1(mod3,~.+Longitude2,test='F')
#It is clear to see that p-value of Longitude is less than 0.5.

<u>#Stepwise Regression using Akaike's Information Criterion (AIC)</u>
mod.upper=lm((response2)^(0.42)~Date2+Store_number2+Latitude2+Longitude2+log(Distance2)+House_age2)
step(mod0,scope=list(lower=mod0,upper=mod.upper))

<u>#plot Cook Disance to check influential observation</u>
plot(fit4,which=4)
abline(h=1,lty=2)

<u>#question 1</u>
fit4=lm((response2)^(0.42)~Date2+House_age2+log(Distance2)+Store_number2+Latitude2+Longitude2)
new =
data.frame(Date2=mean(Date2),House_age2=30,Distance2=mean(log(Distance2)),Store_number2=mean(Store_number2),Latitude2=mean(Latitude2),Longitude2= mean(Longitude2))
ams=predict(fit4,new, interval="predict", level=0.95,type = "response")
ams^(1/0.42)

<u>#question 2</u>
(summary(fit4))$coefficients
anova(fit4)[4]

Relevant Code Output

<u>#coef(fit)</u>

```
 (Intercept)          Date     House_age      Distance  Store_number      Latitude
-1.444198e+04  5.149017e+00 -2.696967e-01 -4.487508e-03  1.133325e+00  2.254701e+02
   Longitude
-1.242906e+01
```

#summary(fit)

```
Call:
lm(formula = response ~ Date + House_age + Distance + Store_number +
   Latitude + Longitude)

Residuals:
   Min      1Q  Median      3Q     Max
-35.667  -5.412  -0.967   4.217  75.190

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.444e+04  6.775e+03  -2.132  0.03364 *
Date          5.149e+00  1.557e+00   3.307  0.00103 **
House_age    -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
Distance     -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
Store_number  1.133e+00  1.882e-01   6.023 3.83e-09 ***
Latitude      2.255e+02  4.457e+01   5.059 6.38e-07 ***
Longitude    -1.243e+01  4.858e+01  -0.256  0.79820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.858 on 407 degrees of freedom
Multiple R-squared:  0.5824,	Adjusted R-squared:  0.5762
F-statistic:  94.6 on 6 and 407 DF,  p-value: < 2.2e-16
```

```
#n=length(Distance); outlier=rstudent(fit)
for (i in 1:n){
  if (outlier[i]>3){
    print (outlier[i])
  }
}
```

```
   127        149        221        271        313        390
3.176409   3.430879   3.861481   9.451489   3.861317   3.128453
```

#summary(fit4)

```
Call:
lm(formula = (response2)^(0.42) ~ Date2 + House_age2 + log(Distance2) +
   Store_number2 + Latitude2 + Longitude2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.34737 -0.20318  0.00148  0.18451  1.67080

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.448e+03  2.252e+02  -6.432 3.55e-10 ***
Date2           2.987e-01  6.476e-02   4.612 5.35e-06 ***
House_age2     -1.134e-02  1.603e-03  -7.075 6.61e-12 ***
log(Distance2) -3.140e-01  2.675e-02 -11.738  < 2e-16 ***
Store_number2   2.658e-02  8.725e-03   3.047  0.00246 **
Latitude2       1.659e+01  1.701e+00   9.754  < 2e-16 ***
Longitude2      3.615e+00  1.569e+00   2.303  0.02177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3659 on 406 degrees of freedom
Multiple R-squared:  0.7285,	Adjusted R-squared:  0.7245 
F-statistic: 181.6 on 6 and 406 DF,  p-value: < 2.2e-16
```

#Stepwise using F-test

```
>add1(mod0,~.+Date2+Store_number2+Latitude2+Longitude2,test='F')
Single term additions

Model:
(response2)^(0.42) ~ log(Distance2) + House_age2
              Df Sum of Sq    RSS     AIC F value    Pr(>F)    
<none>                     77.410 -685.50                      
Date2          1    4.7714 72.639 -709.77  26.866 3.434e-07 ***
Store_number2  1    4.2063 73.204 -706.57  23.501 1.776e-06 ***
Latitude2      1   17.8651 59.545 -791.86 122.711 < 2.2e-16 ***
Longitude2     1    2.4264 74.984 -696.65  13.235 0.0003099 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>add1(mod1,~.+Date2+Store_number2+Longitude2,test='F')
Single term additions

Model:
(response2)^(0.42) ~ log(Distance2) + House_age2 + Latitude2
              Df Sum of Sq    RSS     AIC F value    Pr(>F)    
<none>                     59.545 -791.86                      
Date2          1    3.2942 56.251 -813.37 23.8940 1.466e-06 ***
Store_number2  1    1.5723 57.973 -800.92 11.0656 0.0009591 ***
Longitude2     1    0.6978 58.847 -794.73  4.8383 0.0283951 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>summary(mod2)
Call:
lm(formula = (response2)^(0.42) ~ log(Distance2) + House_age2 + 
    Latitude2 + Date2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.30114 -0.19484  0.01285  0.18307  1.67075 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -1.091e+03  1.345e+02  -8.109 6.04e-15 ***
log(Distance2) -3.871e-01  1.860e-02 -20.810  < 2e-16 ***
House_age2     -1.088e-02  1.616e-03  -6.732 5.68e-11 ***
Latitude2       1.820e+01  1.669e+00  10.903  < 2e-16 ***
Date2           3.197e-01  6.541e-02   4.888 1.47e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3713 on 408 degrees of freedom
Multiple R-squared:  0.719,	Adjusted R-squared:  0.7163 
F-statistic:   261 on 4 and 408 DF,  p-value: < 2.2e-16
```

```
>add1(mod2,~.+Store_number2+Longitude2,test='F')
Single term additions

Model:
(response2)^(0.42) ~ log(Distance2) + House_age2 + Latitude2 +
   Date2
            Df Sum of Sq    RSS     AIC F value   Pr(>F)
<none>                   56.251 -813.37
Store_number2  1   1.18235 55.068 -820.14  8.7385 0.003297 **
Longitude2     1   0.64974 55.601 -816.17  4.7561 0.029767 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>summary(mod3)
Call:
lm(formula = (response2)^(0.42) ~ log(Distance2) + House_age2 +
   Latitude2 + Date2 + Store_number2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.33726 -0.20289  0.00205  0.18339  1.66987

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.032e+03  1.348e+02  -7.653 1.43e-13 ***
log(Distance2) -3.440e-01  2.349e-02 -14.646  < 2e-16 ***
House_age2     -1.141e-02  1.611e-03  -7.087 6.10e-12 ***
Latitude2       1.727e+01  1.683e+00  10.262  < 2e-16 ***
Date2           3.016e-01  6.509e-02   4.633 4.86e-06 ***
Store_number2   2.591e-02  8.766e-03   2.956   0.0033 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3678 on 407 degrees of freedom
Multiple R-squared:  0.7249,           Adjusted R-squared:  0.7216
F-statistic: 214.5 on 5 and 407 DF,  p-value: < 2.2e-16

>add1(mod3,~.+Longitude2,test='F')
Single term additions

Model:
(response2)^(0.42) ~ log(Distance2) + House_age2 + Latitude2 +
   Date2 + Store_number2
          Df Sum of Sq    RSS     AIC F value  Pr(>F)
<none>                 55.068 -820.14
Longitude2  1   0.71023 54.358 -823.50  5.3047 0.02177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Stepwise using AIC

```
>mod0=lm((response2)^(0.42)~log(Distance2)+House_age2)
>mod.upper=lm((response2)^(0.42)~Date2+Store_number2+Latitude2+Longitude2+log(Distance2)+House_age2)
```

```
>step(mod0,scope=list(lower=mod0,upper=mod.upper))


Start:  AIC=-685.5
(response2)^(0.42) ~ log(Distance2) + House_age2

               Df Sum of Sq    RSS     AIC
+ Latitude2     1   17.8651 59.545 -791.86
+ Date2         1    4.7714 72.639 -709.77
+ Store_number2  1    4.2063 73.204 -706.57
+ Longitude2    1    2.4264 74.984 -696.65
<none>                     77.410 -685.50

Step:  AIC=-791.86
(response2)^(0.42) ~ log(Distance2) + House_age2 + Latitude2

               Df Sum of Sq    RSS     AIC
+ Date2         1    3.2942 56.251 -813.37
+ Store_number2  1    1.5723 57.973 -800.92
+ Longitude2    1    0.6978 58.847 -794.73
<none>                     59.545 -791.86
- Latitude2     1   17.8651 77.410 -685.50

Step:  AIC=-813.37
(response2)^(0.42) ~ log(Distance2) + House_age2 + Latitude2 +
    Date2

               Df Sum of Sq    RSS     AIC
+ Store_number2  1    1.1823 55.068 -820.14
+ Longitude2    1    0.6497 55.601 -816.17
<none>                     56.251 -813.37
- Date2         1    3.2942 59.545 -791.86
- Latitude2     1   16.3879 72.639 -709.77

Step:  AIC=-820.14
(response2)^(0.42) ~ log(Distance2) + House_age2 + Latitude2 +
    Date2 + Store_number2

               Df Sum of Sq    RSS     AIC
+ Longitude2    1    0.7102 54.358 -823.50
<none>                     55.068 -820.14
- Store_number2  1    1.1823 56.251 -813.37
- Date2         1    2.9043 57.973 -800.92
- Latitude2     1   14.2486 69.317 -727.10

Step:  AIC=-823.5
(response2)^(0.42) ~ log(Distance2) + House_age2 + Latitude2 +
    Date2 + Store_number2 + Longitude2

               Df Sum of Sq    RSS     AIC
<none>                     54.358 -823.50
- Longitude2    1    0.7102 55.068 -820.14
- Store_number2  1    1.2428 55.601 -816.17
- Date2         1    2.8481 57.206 -804.41
- Latitude2     1   12.7379 67.096 -738.55

Call:
```

```
lm(formula = (response2)^(0.42) ~ log(Distance2) + House_age2 +
    Latitude2 + Date2 + Store_number2 + Longitude2)

Coefficients:
   (Intercept)  log(Distance2)      House_age2       Latitude2           Date2
    -1.448e+03      -3.140e-01      -1.134e-02       1.659e+01       2.987e-01
 Store_number2      Longitude2
     2.658e-02       3.615e+00
```

#question 1

```
>fit4=lm((response2)^(0.42)~Date2+House_age2+log(Distance2)+Store_number2+Latitude2
+Longitude2)
>new =
data.frame(Date2=mean(Date2),House_age2=30,Distance2=mean(log(Distance2)),Store_nu
mber2=mean(Store_number2),Latitude2=mean(Latitude2),Longitude2= mean(Longitude2))
>ams=predict(fit4,new, interval="predict", level=0.95,type = "response")
>ams
       fit     lwr      upr
1 5.810626 5.04948 6.571771
>ams^(1/0.42)
       fit      lwr      upr
1 66.00522 47.24946 88.48362
```

#question 2

```
>(summary(fit4))$coefficients
                   Estimate   Std. Error     t value     Pr(>|t|)
(Intercept)    -1.448154e+03 2.251565e+02  -6.431765 3.551510e-10
Date2           2.986865e-01 6.476030e-02   4.612186 5.348354e-06
House_age2     -1.133801e-02 1.602632e-03  -7.074618 6.609708e-12
log(Distance2) -3.140218e-01 2.675346e-02 -11.737615 1.360459e-27
Store_number2   2.658325e-02 8.725063e-03   3.046768 2.464193e-03
Latitude2       1.658667e+01 1.700508e+00   9.753947 2.471521e-20
Longitude2      3.614851e+00 1.569492e+00   2.303198 2.177234e-02

>anova(fit4)[4]
               F value
Date2           8.5532
House_age2     59.9737
log(Distance2) 884.2534
Store_number2   24.8091
Latitude2      106.4228
Longitude2       5.3047
Residuals
```