Multiple Linear Regression on housing prices

Ankit Gahlawat

2022-09-29

I aim to predict the house price of unit area while considering several factors.

I intend to use multiple regression model for the same . Considering house price as the dependent variable and transaction date , age of the house , distance from nearest station , convenience store in the locality, latitude , longitude of the house as the explanatory variables.

Importing libraries

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(ggpubr)
```

Importing Dataset

```
housing_data=read.csv("C:/Users/Ankit kumar gahlawat/Desktop/R
project/Realestate.csv")
head(housing_data)
```

```
##   No X1.transaction.date X2.house.age
X3.distance.to.the.nearest.MRT.station
## 1  1            2012.917         32.0
84.87882
## 2  2            2012.917         19.5
306.59470
## 3  3            2013.583         13.3
561.98450
## 4  4            2013.500         13.3
561.98450
## 5  5            2012.833          5.0
```

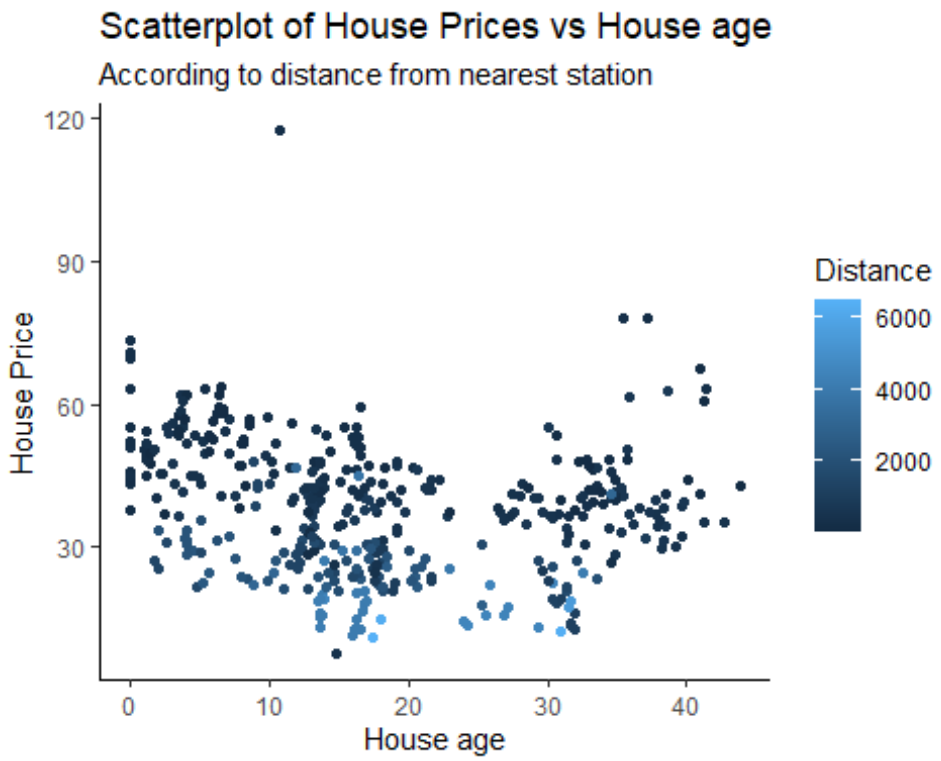
```
390.56840
## 6  6            2012.667          7.1
2175.03000
##   X4.number.of.convenience.stores X5.latitude X6.longitude
## 1                              10    24.98298     121.5402
## 2                               9    24.98034     121.5395
## 3                               5    24.98746     121.5439
## 4                               5    24.98746     121.5439
## 5                               5    24.97937     121.5425
## 6                               3    24.96305     121.5125
##   Y.house.price.of.unit.area
## 1                       37.9
## 2                       42.2
## 3                       47.3
## 4                       54.8
## 5                       43.1
## 6                       32.1

summary(housing_data)

##        No        X1.transaction.date  X2.house.age
##  Min.   :  1.0   Min.   :2013        Min.   : 0.000
##  1st Qu.:104.2   1st Qu.:2013        1st Qu.: 9.025
##  Median :207.5   Median :2013        Median :16.100
##  Mean   :207.5   Mean   :2013        Mean   :17.713
##  3rd Qu.:310.8   3rd Qu.:2013        3rd Qu.:28.150
##  Max.   :414.0   Max.   :2014        Max.   :43.800
##  X3.distance.to.the.nearest.MRT.station X4.number.of.convenience.stores
##  Min.   :  23.38                        Min.   : 0.000
##  1st Qu.: 289.32                        1st Qu.: 1.000
##  Median : 492.23                        Median : 4.000
##  Mean   :1083.89                        Mean   : 4.094
##  3rd Qu.:1454.28                        3rd Qu.: 6.000
##  Max.   :6488.02                        Max.   :10.000
##   X5.latitude     X6.longitude   Y.house.price.of.unit.area
##  Min.   :24.93   Min.   :121.5   Min.   :  7.60
##  1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 27.70
##  Median :24.97   Median :121.5   Median : 38.45
##  Mean   :24.97   Mean   :121.5   Mean   : 37.98
##  3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 46.60
##  Max.   :25.01   Max.   :121.6   Max.   :117.50

colnames(housing_data)[2:8]=c('t_date','house_age','dist','con_store','lattit
ude',
                              'longitude','house_price'  )

glimpse(housing_data)

## Rows: 414
## Columns: 8
## $ No          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
```

```
17,…
## $ t_date      <dbl> 2012.917, 2012.917, 2013.583, 2013.500, 2012.833,
2012.667…
## $ house_age   <dbl> 32.0, 19.5, 13.3, 13.3, 5.0, 7.1, 34.5, 20.3, 31.7,
17.9, …
## $ dist        <dbl> 84.87882, 306.59470, 561.98450, 561.98450, 390.56840,
2175…
## $ con_store   <int> 10, 9, 5, 5, 5, 3, 7, 6, 1, 3, 1, 9, 5, 4, 4, 2, 6, 1,
8, …
## $ lattitude   <dbl> 24.98298, 24.98034, 24.98746, 24.98746, 24.97937,
24.96305…
## $ longitude   <dbl> 121.5402, 121.5395, 121.5439, 121.5439, 121.5425,
121.5125…
## $ house_price <dbl> 37.9, 42.2, 47.3, 54.8, 43.1, 32.1, 40.3, 46.7, 18.8,
22.1…
```

All the variables are continuous , we will perform some EDA now.

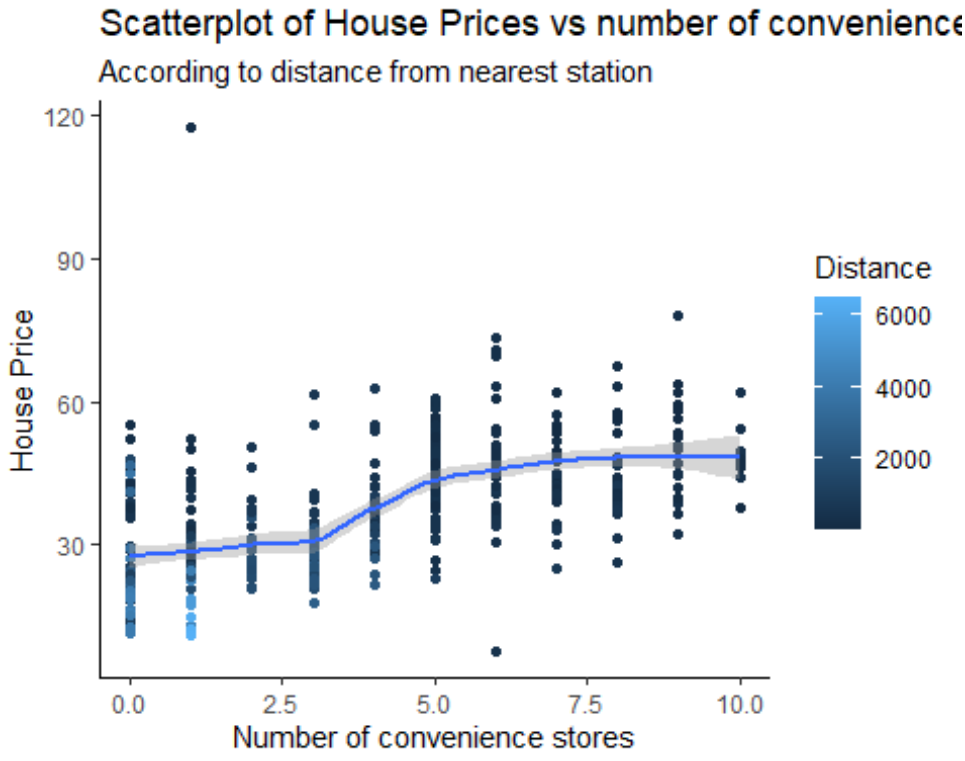
1.

```
g1 = housing_data %>%
  ggplot(aes(house_age, y = house_price)) +
    geom_point(aes(col = dist)) +
  labs(title = "Scatterplot of House Prices vs House age",
       subtitle = "According to distance from nearest station",
       col = "Distance") +
  xlab("House age") + ylab("House Price") +
  theme_classic()

g2 = housing_data %>%
  ggplot(aes(con_store, y = house_price)) +
    labs(title = "Scatterplot of House Prices vs number of convenience
stores",
         subtitle = "According to distance from nearest station",
       col = "Distance") +
  xlab("Number of convenience stores") + ylab("House Price") +
    geom_point(aes(col = dist)) +
  geom_smooth() +
  theme_classic()

g1
```

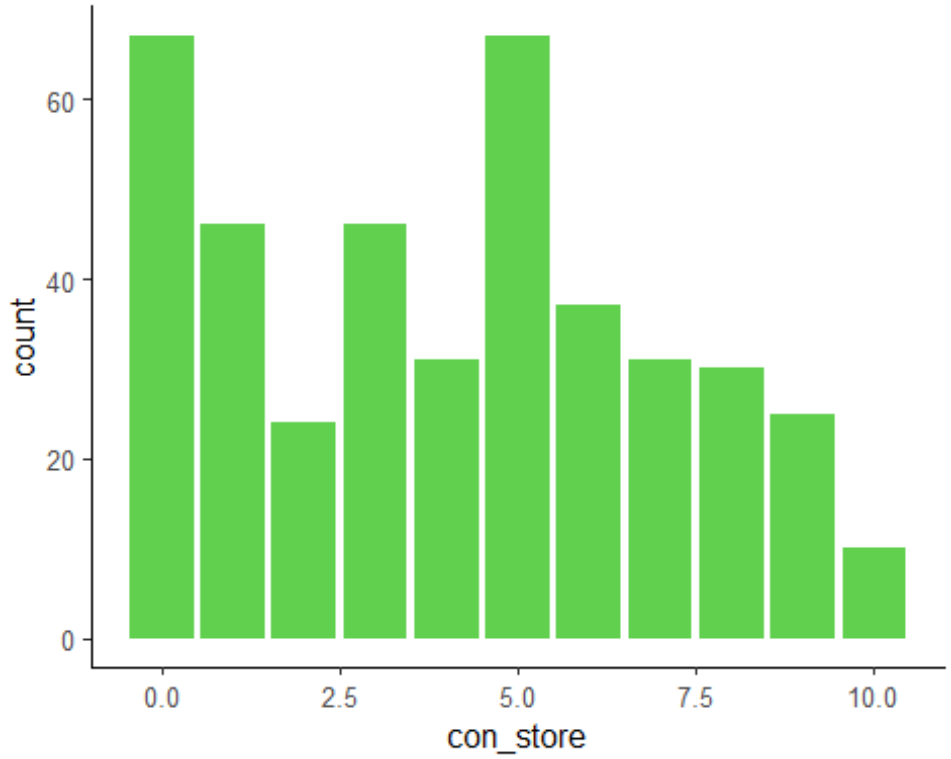



```
g2
```

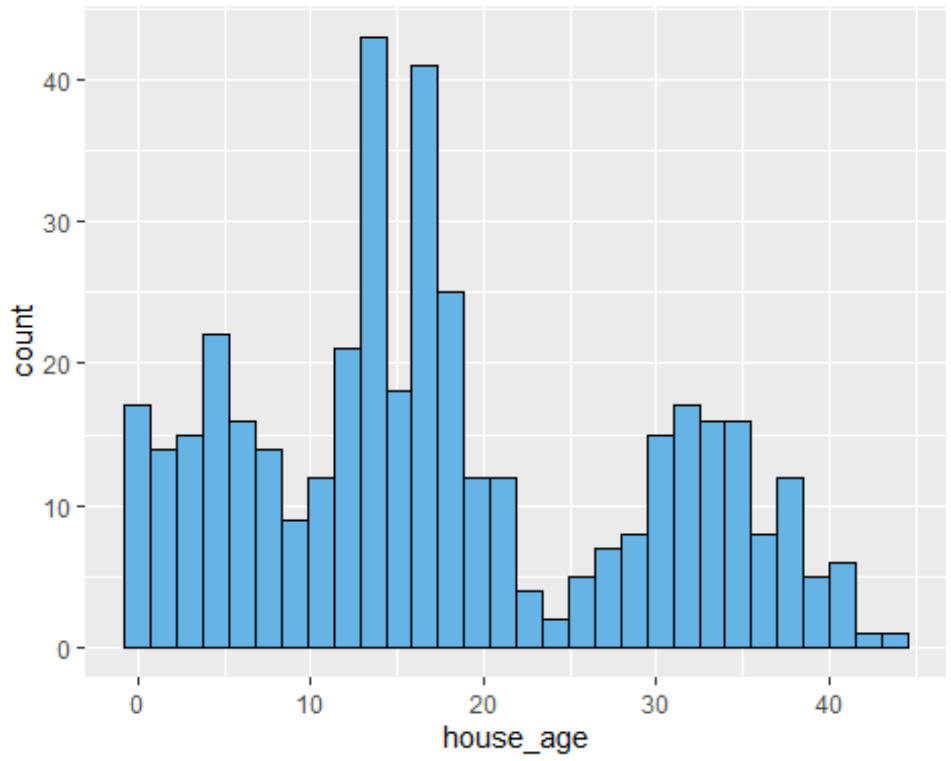
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
housing_data %>%
  ggplot() +
  geom_bar(aes(x = con_store), fill = 3) +
  theme_classic2()
```




```
housing_data %>%
  ggplot() +
  geom_histogram(aes(x = house_age), col = 1, fill = rgb(0.4, 0.7, 0.9))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

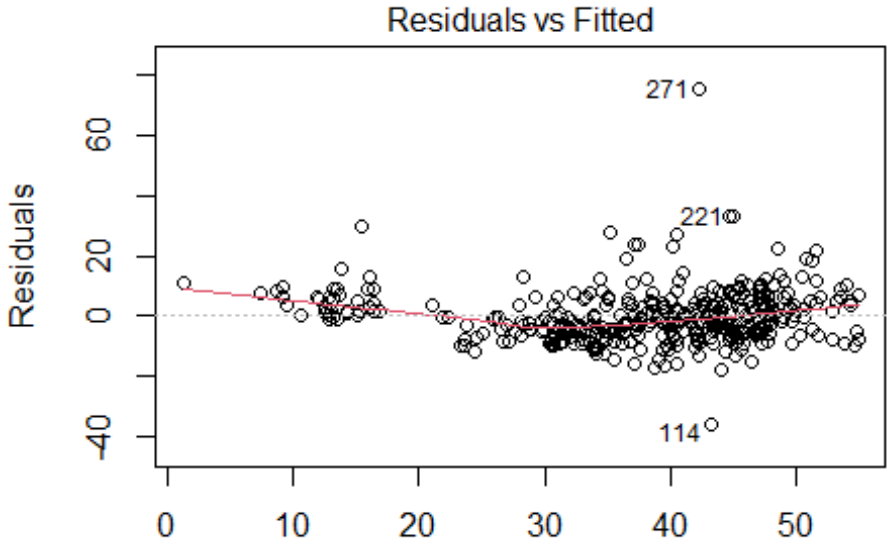


Checking the assumptions for multiple linear regression

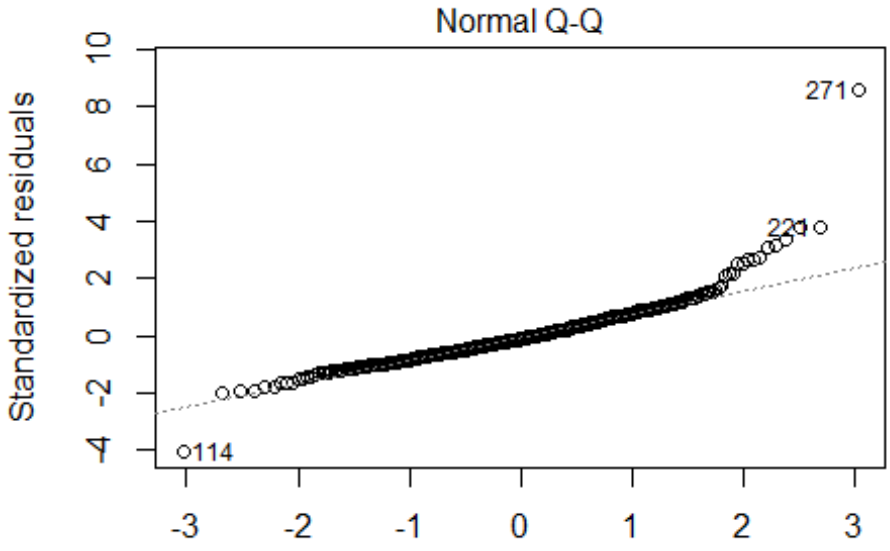
Applying the regression model now

```
mult_regression1= lm(house_price ~
t_date+house_age+dist+con_store+lattitude+longitude,
                     data=housing_data)
anova(mult_regression1)

## Analysis of Variance Table
##
## Response: house_price
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## t_date       1    585     585   7.4598  0.006584 **
## house_age    1   3441    3441  43.8559  1.12e-10 ***
## dist         1  34857   34857 444.2734 < 2.2e-16 ***
## con_store    1   3576    3576  45.5748  5.08e-11 ***
## lattitude    1   2065    2065  26.3187  4.49e-07 ***
## longitude    1      5       5   0.0654  0.798293
## Residuals  407  31933      78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(mult_regression1)

##
## Call:
## lm(formula = house_price ~ t_date + house_age + dist + con_store +
##     lattitude + longitude, data = housing_data)
```

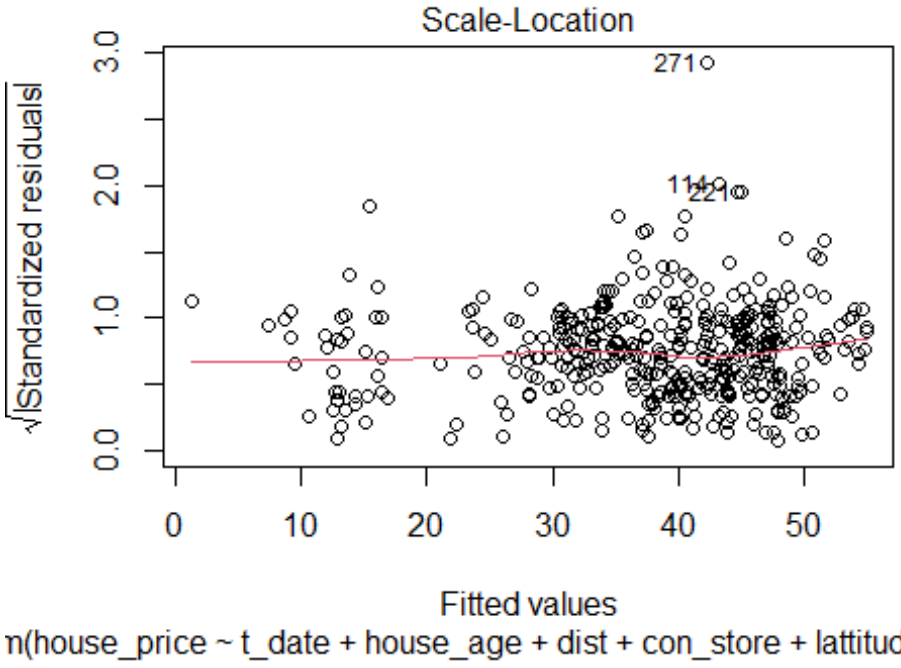
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.664  -5.410  -0.966   4.217  75.193
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.444e+04  6.776e+03  -2.131  0.03371 *
## t_date       5.146e+00  1.557e+00   3.305  0.00103 **
## house_age   -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
## dist        -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
## con_store    1.133e+00  1.882e-01   6.023 3.84e-09 ***
## lattitude    2.255e+02  4.457e+01   5.059 6.38e-07 ***
## longitude   -1.242e+01  4.858e+01  -0.256  0.79829
## ---
## 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: 94.59 on 6 and 407 DF,  p-value: < 2.2e-16

plot(mult_regression1)
```



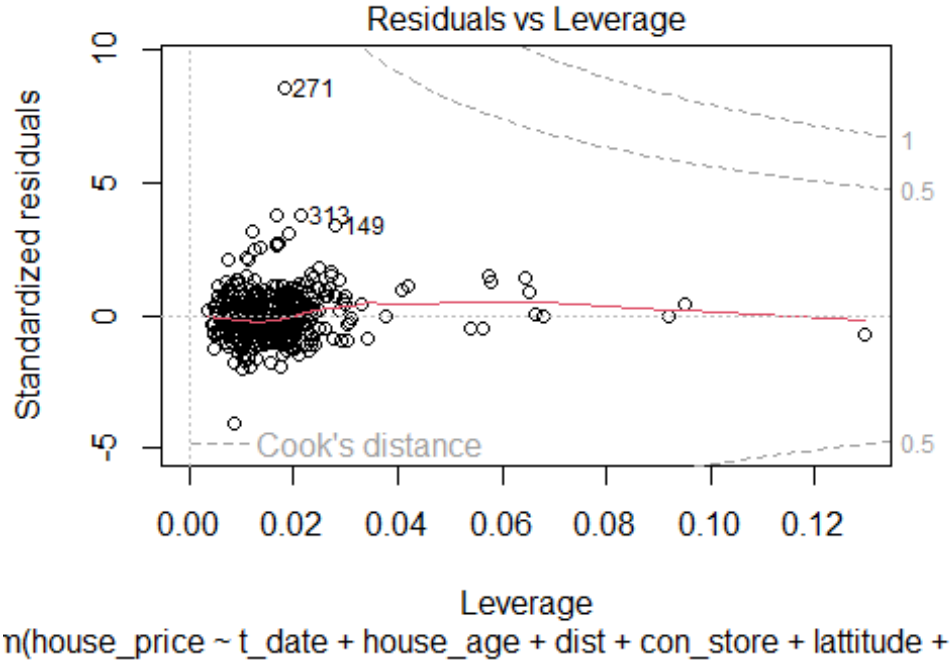
n(house_price ~ t_date + house_age + dist + con_store + lattitude + lo



n(house_price ~ t_date + house_age + dist + con_store + lattitude + lo



n(house_price ~ t_date + house_age + dist + con_store + lattitude + lo



n(house_price ~ t_date + house_age + dist + con_store + lattitude + lo

By inspection of the data, we can see that longitude should be dropped since the p-value for this variable is large.

Running the linear regression model without longitude , and plotting it.

By inspection of the linear regression models, the p-values and the R^2 are acceptable. Moreover, from the first graph , the homoscedasticity is respected since the lines are almost horizontal. However, from the second graph, the residuals do not seem normally distributed since there are many points that are far from the straight line. I will then try to improve the model by applying a log transformation to the model.

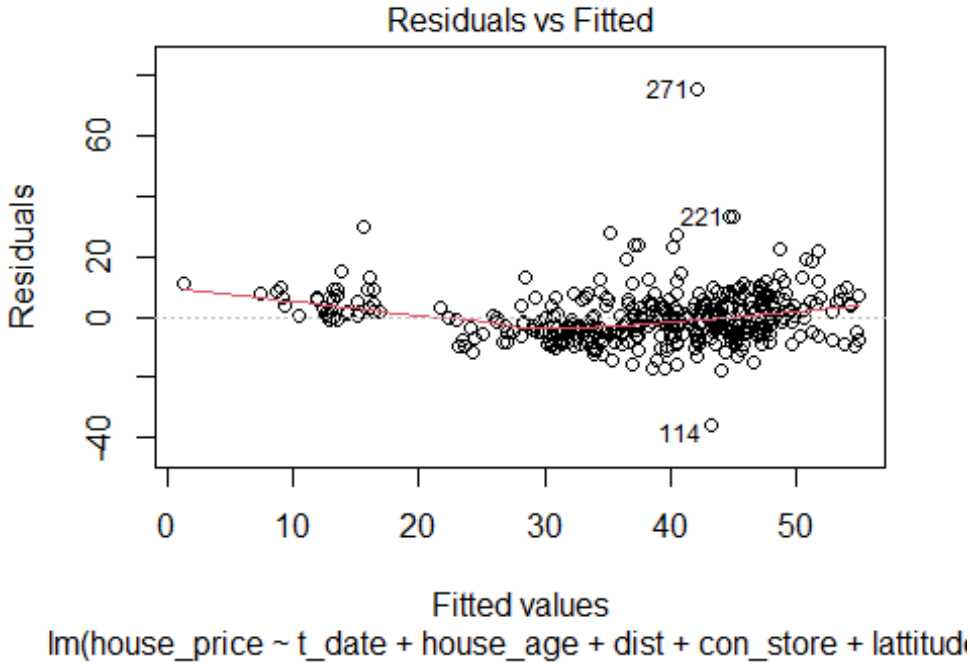
```
mult_regression2= lm(house_price ~ t_date+house_age+dist+con_store+lattitude,
                     data=housing_data)
summary(mult_regression2)
```

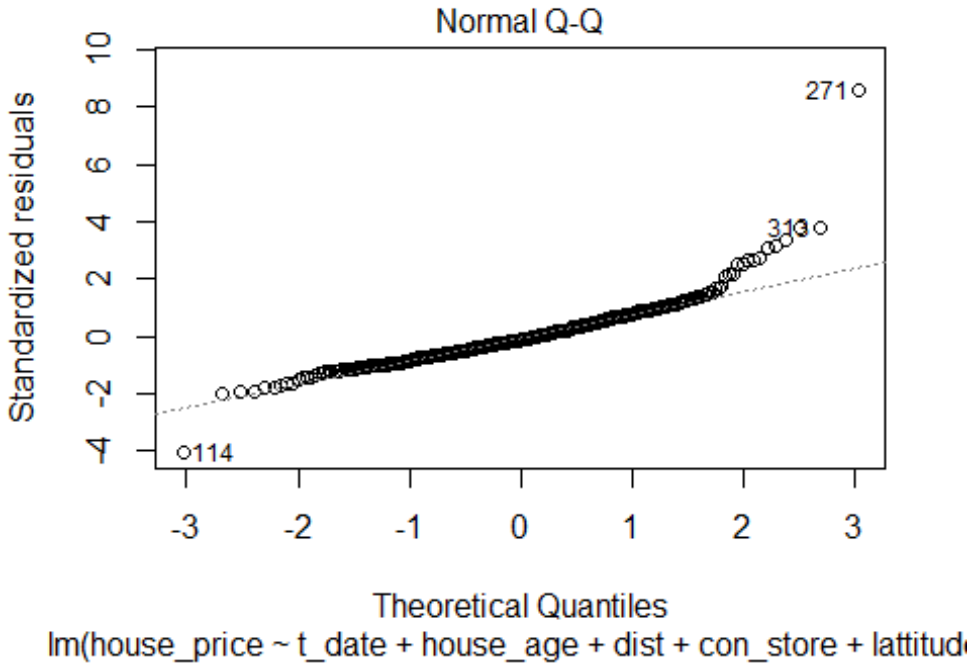
```
##
## Call:
## lm(formula = house_price ~ t_date + house_age + dist + con_store +
##     lattitude, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.623  -5.371  -1.020   4.244  75.346
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.596e+04  3.233e+03  -4.936 1.17e-06 ***
## t_date       5.135e+00  1.555e+00   3.303  0.00104 **
## house_age   -2.694e-01  3.847e-02  -7.003 1.04e-11 ***
## dist        -4.353e-03  4.899e-04  -8.887  < 2e-16 ***
## con_store    1.136e+00  1.876e-01   6.056 3.17e-09 ***
## lattitude    2.269e+02  4.417e+01   5.136 4.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.848 on 408 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5772
## F-statistic: 113.8 on 5 and 408 DF,  p-value: < 2.2e-16
```

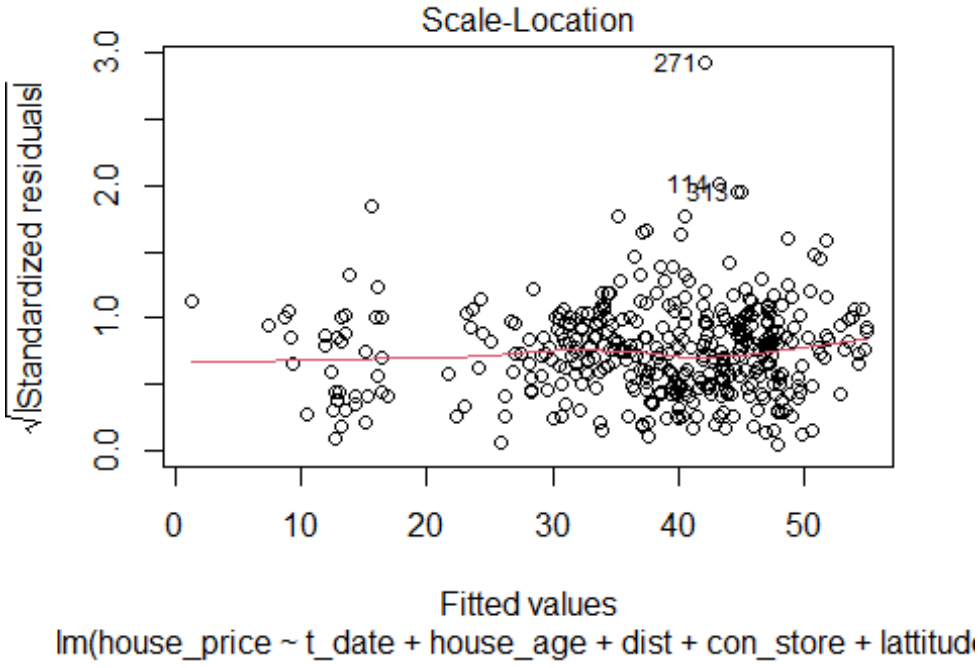
```
anova(mult_regression2)
```

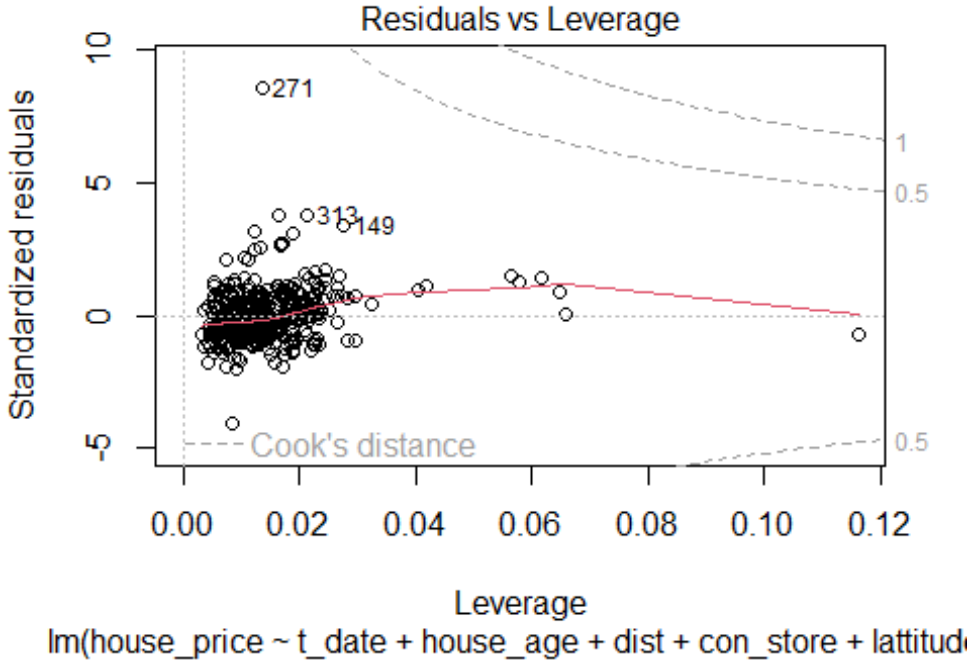
```
## Analysis of Variance Table
##
## Response: house_price
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## t_date      1    585     585   7.4769  0.006522 **
## house_age   1   3441    3441  43.9566 1.067e-10 ***
## dist        1  34857   34857 445.2934 < 2.2e-16 ***
## con_store   1   3576    3576  45.6794 4.828e-11 ***
## lattitude   1   2065    2065  26.3791 4.355e-07 ***
## Residuals 408  31938      78
```

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

plot(mult_regression2)
```






Scale-Location
271
149
313
√|Standardized residuals|
3.0
2.0
1.0
0.0
0
10
20
30
40
50
Fitted values
lm(house_price ~ t_date + house_age + dist + con_store + lattitud


Residuals vs Leverage
271
313
149
Standardized residuals
10
5
0
-5
Cook's distance
1
0.5
0.5
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Leverage
lm(house_price ~ t_date + house_age + dist + con_store + lattitud

Running the linear regression with log transformation

conclusion :

1. The p-values stays acceptable

2. The R^2 value improved from 0.5823 to 0.6857

3. The homoscedasticity has also improved since the line in the first graph are more horizontal than in the previous model

4 The points of the residuals are closer to the straight line compared to the last model, but there is still room for improvement. I will try to remove one variable from the model to get a better result.

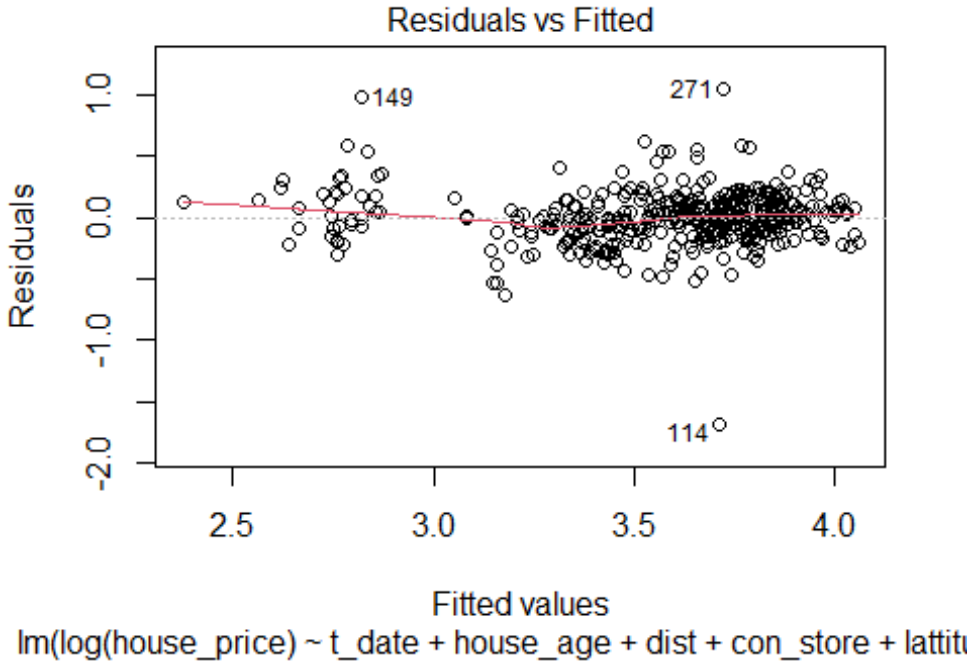
```
mult_regression3= lm(log(house_price) ~
t_date+house_age+dist+con_store+lattitude,
                     data=housing_data)
summary(mult_regression3)
```

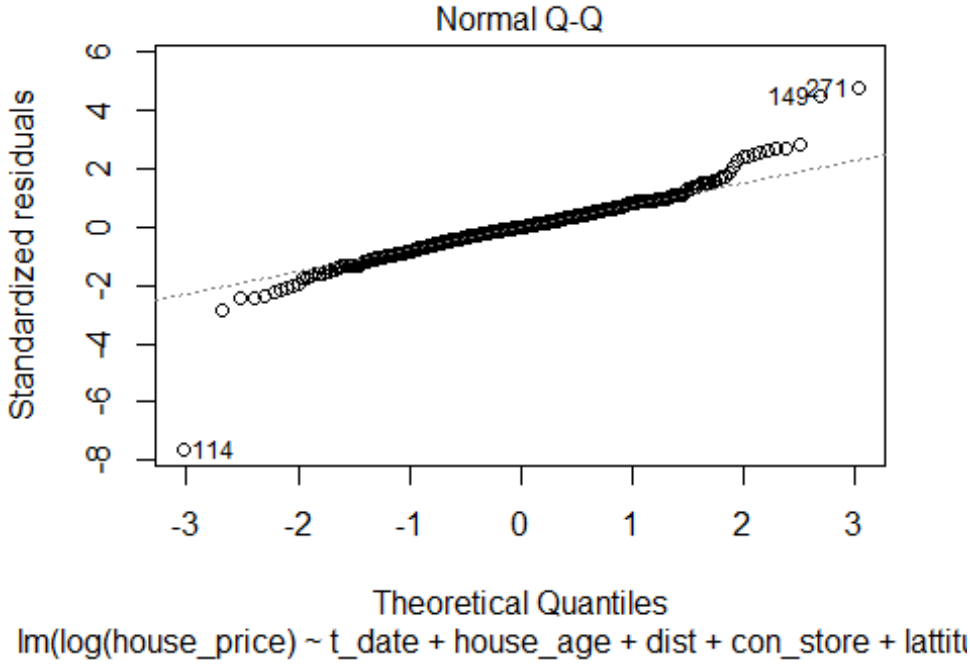
```
##
## Call:
## lm(formula = log(house_price) ~ t_date + house_age + dist + con_store +
##     lattitude, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68218 -0.11505  0.00055  0.11262  1.04395
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.665e+02  8.091e+01  -5.766 1.61e-08 ***
## t_date       1.358e-01  3.890e-02   3.491 0.000533 ***
## house_age   -6.977e-03  9.625e-04  -7.248 2.13e-12 ***
## dist        -1.495e-04  1.226e-05 -12.194  < 2e-16 ***
## con_store    2.766e-02  4.694e-03   5.892 7.97e-09 ***
## lattitude    7.883e+00  1.105e+00   7.132 4.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2214 on 408 degrees of freedom
## Multiple R-squared:  0.6857, Adjusted R-squared:  0.6818
## F-statistic:   178 on 5 and 408 DF,  p-value: < 2.2e-16
```

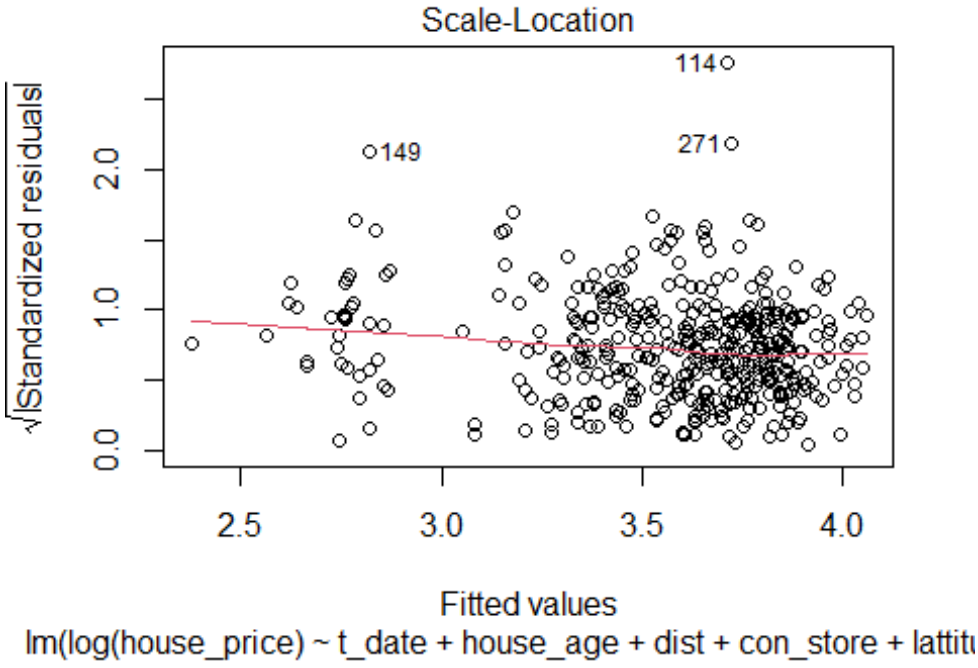
```
anova(mult_regression3)
```

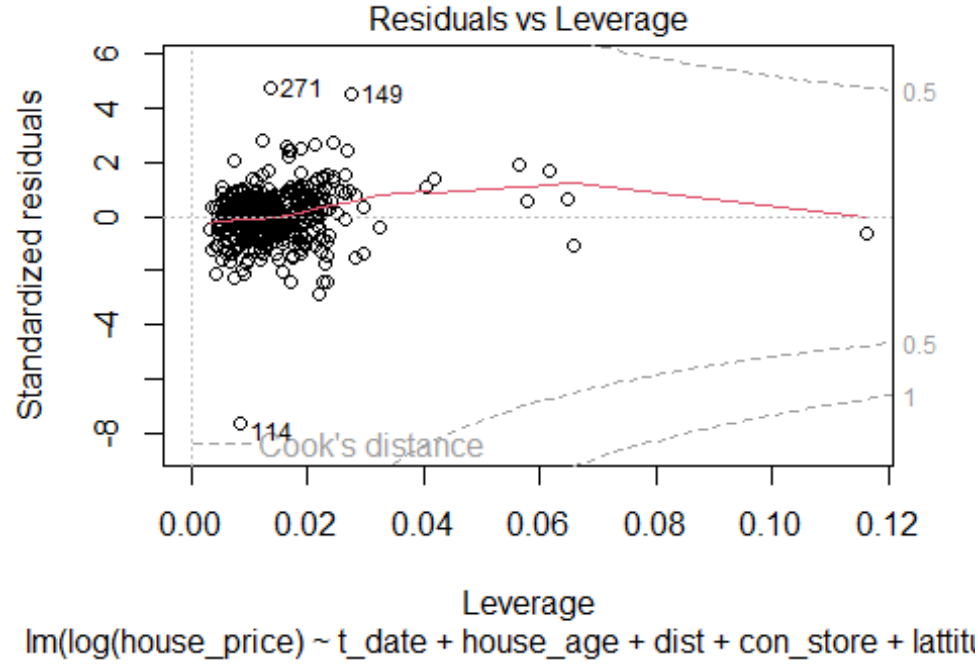
```
## Analysis of Variance Table
##
## Response: log(house_price)
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## t_date      1  0.363   0.363   7.3976   0.00681 **
```

```
## house_age   1  2.311   2.311  47.1596 2.450e-11 ***
## dist        1 36.155  36.155 737.7156 < 2.2e-16 ***
## con_store   1  2.298   2.298  46.8982 2.761e-11 ***
## lattitude   1  2.493   2.493  50.8630 4.541e-12 ***
## Residuals 408 19.996   0.049
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(mult_regression3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(house_price) ~ t_date + house_age + dist + con_store + lattit
149
271
114
1.0
0.0
-1.0
-2.0
2.5
3.0
3.5
4.0


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(house_price) ~ t_date + house_age + dist + con_store + lattit
149
271
114
6
4
2
0
-2
-4
-6
-8
-3
-2
-1
0
1
2
3






```
#ggplot(data=housing_data, aes(log(house_price))) +
  #geom_histogram(breaks=seq(0, 2, by=0.15),
   #              col="red",
    #             aes(fill=..count..)) +
```

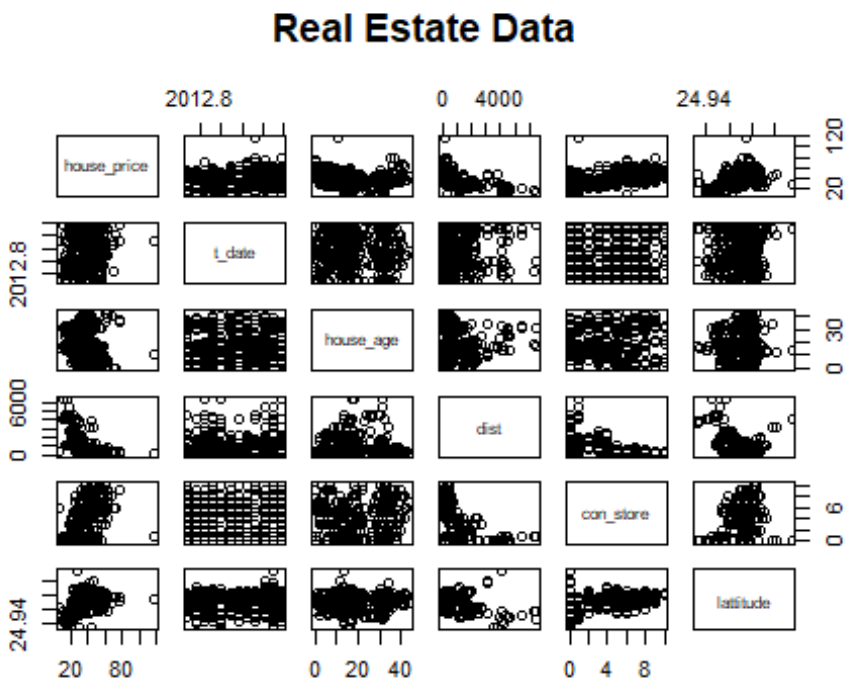
```
 # scale_fill_gradient("Count", low="green", high="red")+
  #labs(title="Histogram Log House Price Unit Area",x="Log Unit
Area",y="Count")
```

Analyzing which variable should be dropped

The variables that have the lowest correlation with house price is t_date and house_age.

I will try to run the regression without this variable in the model to see what is going to happen.

```
pairs(~ house_price + t_date + house_age + dist + con_store +
    lattitude, data = housing_data,  main = "Real Estate Data")
```



Applying Regression again without using t_date and house_age

Even if the R^2 has decreased a little bit comparatively the normality of residuals has gotten much better without t_date and house_age in the set. The results in general are similar in both sets. I will therefore accept this model. Hence , my final adjusted R^2 comes out to be 0.6334

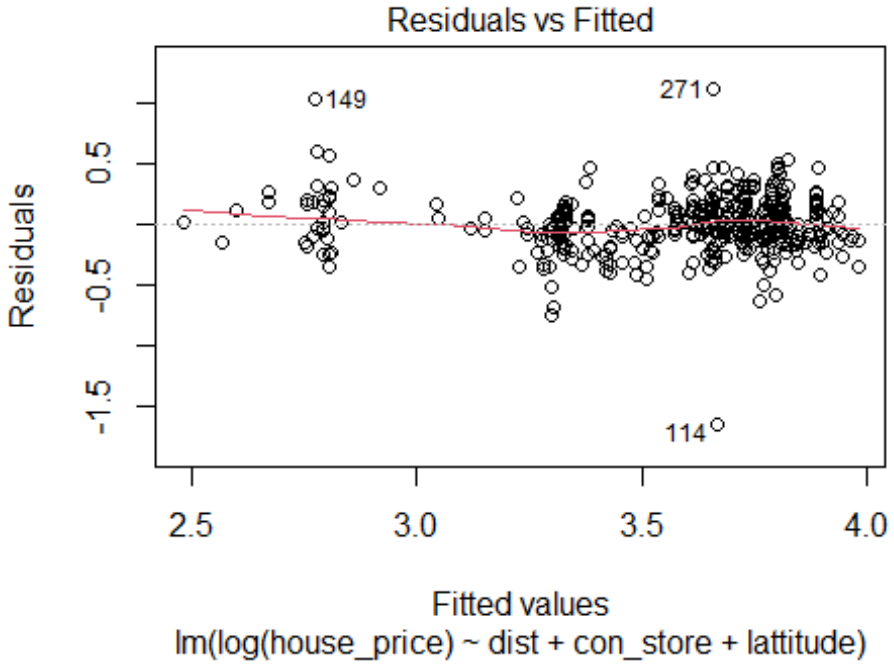
```
mult_regression4= lm(log(house_price) ~ dist+con_store+lattitude,
                     data=housing_data)
summary(mult_regression4)

##
## Call:
## lm(formula = log(house_price) ~ dist + con_store + lattitude,
##     data = housing_data)
##
## Residuals:
```

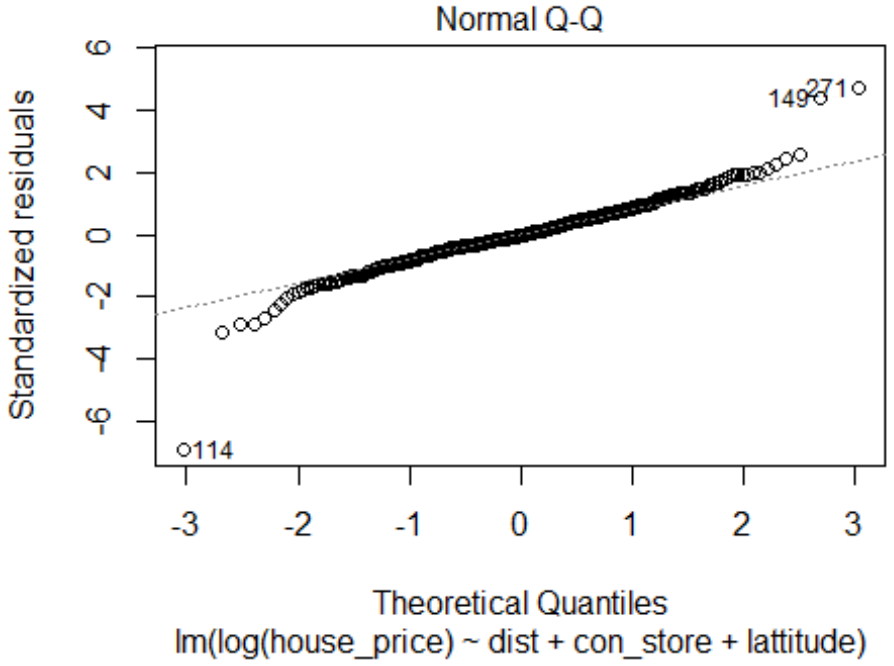
```
##      Min       1Q   Median       3Q      Max
## -1.63852 -0.12207 -0.00658  0.13008  1.11060
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.858e+02  2.945e+01  -6.310 7.23e-10 ***
## dist        -1.533e-04  1.302e-05 -11.773  < 2e-16 ***
## con_store    2.602e-02  5.021e-03   5.183 3.44e-07 ***
## lattitude    7.588e+00  1.179e+00   6.434 3.46e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2376 on 410 degrees of freedom
## Multiple R-squared:  0.636,  Adjusted R-squared:  0.6334
## F-statistic: 238.8 on 3 and 410 DF,  p-value: < 2.2e-16
```

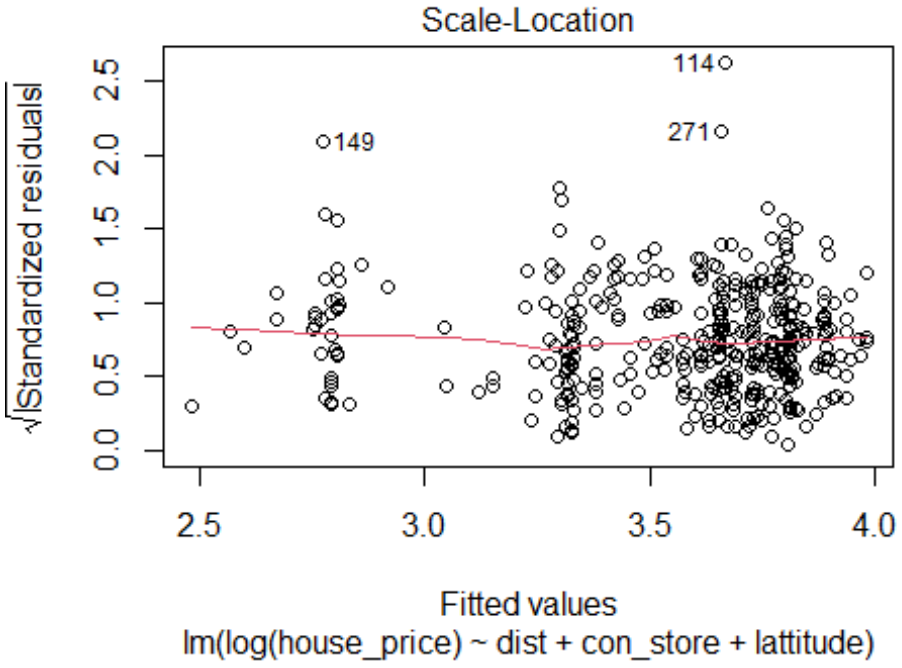
```
anova(mult_regression4)
```

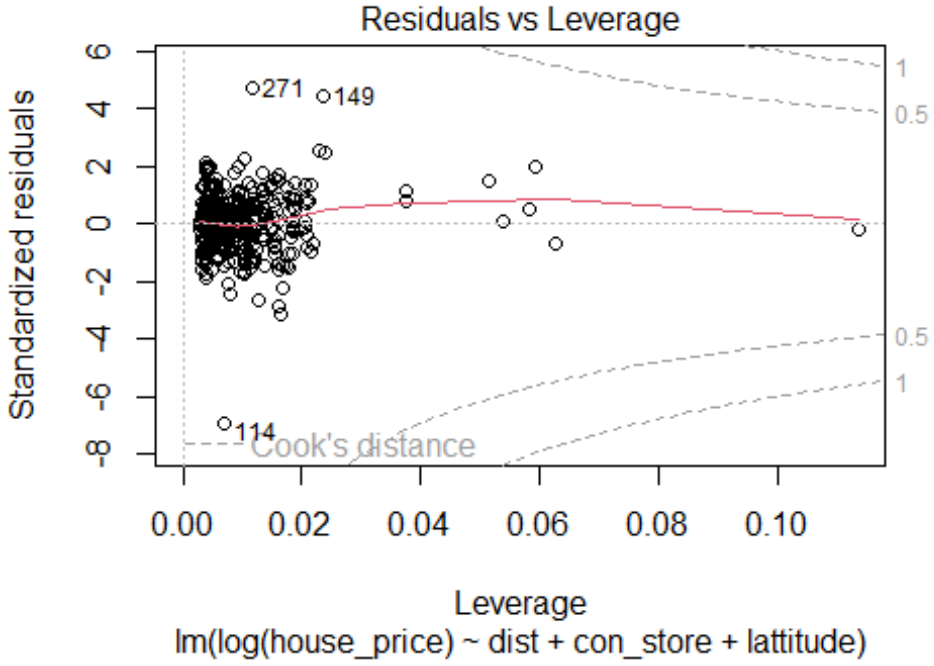
```
## Analysis of Variance Table
##
## Response: log(house_price)
##            Df Sum Sq Mean Sq F value    Pr(>F)
## dist        1 36.007  36.007 637.566 < 2.2e-16 ***
## con_store   1  2.115   2.115  37.452 2.191e-09 ***
## lattitude   1  2.338   2.338  41.401 3.464e-10 ***
## Residuals 410 23.155   0.056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(mult_regression4)
```


Residuals vs Fitted
149
271
114
Residuals
0.5
-0.5
-1.5
2.5
3.0
3.5
4.0
Fitted values
lm(log(house_price) ~ dist + con_store + lattitude)


Normal Q-Q
149
271
114
Standardized residuals
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(log(house_price) ~ dist + con_store + lattitude)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
114
271
149
2.5
3.0
3.5
4.0
Fitted values
lm(log(house_price) ~ dist + con_store + lattitude)


Residuals vs Leverage
Standardized residuals
6
4
2
0
-2
-4
-6
-8
271
149
114
Cook's distance
1
0.5
0.5
1
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(log(house_price) ~ dist + con_store + lattitude)