Multivariate Linear Regression Analysis on the Prediction of Real Estate Prices in Singapore

2022-07-15

Introduction

Linear regression is a statistical method in which the analyst attempts to find a relationship between two or more independent regressor variables to a dependent response variable. We assume that there is a relationship between the variables, such that the regressor variables can be used to predict the values of the response variables. In this analysis, we build a multivariate model using the dataset of real estate prices using six different regressor variables to predict real estate prices.

I. Data

The data we chosen is the real estate data (n = 414) of real estate prices in Singapore, with six different regressor variables. They are as follows:

- X1 transaction date
- X2 house age
- X3 distance to the nearest major public transportation station
- X4 number of convenience stores near the area of the house
- X5 latitude
- X6 longitude
- Y house price of unit area

II. Data Description

- Real Estate Price Prediction Data Set is used to predict the price of unit area for houses given their features.
- Before the analysis, we predict that X5 and X6 are not very significant regressor variables in predicting Y, because they are geographical coordinates, which do not contain any important information in predicting real estate prices.
- X1 may be a significant regressor variable, but it may contain seasonal factors on price, which may call for a Durbin-Watson test to deal with postive autocorrelation.

We will show proof of the insignificance of the mentioned variables in section III.

Source of the data:

Real Estate Price Prediction, by Sohaila Diab. (2022). https://www.kaggle.com/code/sohailadiab/real-estate-price-prediction

III. Linear Regression

As stated in the introduction, we think that variables x1, x5 and x6 are not significant due to the context of the data. However, we run a complete model below just to show the insignificance.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \epsilon$$

- As stated in the introduction, we suspect that X5 and X6 are not significant due the context of te data. However, we run a complete model of all regressor variables to get the total variation.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df <- read.csv("~/Desktop/Real estate data.csv")

head(df)
```

```
##   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 X3_dummy_low..0.3. X3_dummy_med..3.6.
## 1                       37.9                  0                  0
## 2                       42.2                  0                  0
## 3                       47.3                  0                  1
## 4                       54.8                  0                  1
## 5                       43.1                  0                  1
## 6                       32.1                  1                  0
##   X3_dummy_high..7.10.
## 1                    1
## 2                    1
## 3                    0
## 4                    0
## 5                    0
## 6                    0
```

```
y <- df$Y.house.price.of.unit.area #set house price of unit area as the response variable y
x1 <- df$X1.transaction.date
x2 <- df$X2.house.age
x3 <- df$X3.distance.to.the.nearest.MRT.station
x4 <- df$X4.number.of.convenience.stores
x5 <- df$X5.latitude
x6 <- df$X6.longitude
```

Now let's see how effective this complete model is.

```
fit <-lm(y~x1+x2+x3+x4+x5+x6,data=df)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = df)
##
## 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 *
## x1           5.146e+00  1.557e+00   3.305  0.00103 **
## x2          -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
## x3          -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
## x4           1.133e+00  1.882e-01   6.023 3.84e-09 ***
## x5           2.255e+02  4.457e+01   5.059 6.38e-07 ***
## x6          -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
```

We see from the R-Squared value of 0.5824, that this isn't the best model to predict y, but from the f-statistic of 94.59, the model isn't insignificant, which leads us to believe that some of the regressor variables are indeed significant but requires us to identify and drop some of the insignificant variables.

From the t-statistic of X6, it seems that X6 may be a non-significant regressor variable. It does seem strange that X6 (logitude) would be significant while X5(latitude) isn't, let's try dropping both X5 and X6.

```
fit2 <-lm(y~x1+x2+x3+x4,data=df)
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -38.389  -5.630  -0.987   4.306  76.006
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.159e+04  3.215e+03  -3.605 0.000351 ***
## x1           5.778e+00  1.597e+00   3.618 0.000334 ***
## x2          -2.545e-01  3.953e-02  -6.438 3.40e-10 ***
## x3          -5.513e-03  4.480e-04 -12.305  < 2e-16 ***
## x4           1.258e+00  1.918e-01   6.558 1.65e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.118 on 409 degrees of freedom
## Multiple R-squared:  0.5553, Adjusted R-squared:  0.5509
## F-statistic: 127.7 on 4 and 409 DF,  p-value: < 2.2e-16
```

From the summary above, we find that the R^2 value only drops about 0.03 while the f-statistic jumps up significantly from 94.59 to 161.1. The regresssor variables in our model becomes more significant with a small trade in R^2.

Least Square Estimators

With the model being $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \epsilon$, we fit the data to the multiple linear regression model.

$$\hat{y} = -11590 + 5.778x_1 - 0.2545x_2 - 0.005513x_3 + 1.258x_4$$

Interpretation of the regressor variables:

- $\hat{\beta_0}$: The mean of the house price of unit area when all other regressor variables are zero, but here it is meaningless since it is negative.
- $\hat{\beta_1}$,$\hat{\beta_2}$,$\hat{\beta_3}$,$\hat{\beta_4}$: The change in the mean of house price of a unit area associated with one unit increase in the jth regressor variables while all other regressor variables remained affixed.

T-test of the significant of the j-th regressor variables (j=1,2,3,4):

```
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.389  -5.630  -0.987   4.306  76.006
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.159e+04  3.215e+03  -3.605 0.000351 ***
## x1           5.778e+00  1.597e+00   3.618 0.000334 ***
## x2          -2.545e-01  3.953e-02  -6.438 3.40e-10 ***
## x3          -5.513e-03  4.480e-04 -12.305  < 2e-16 ***
## x4           1.258e+00  1.918e-01   6.558 1.65e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.118 on 409 degrees of freedom
## Multiple R-squared:  0.5553, Adjusted R-squared:  0.5509
## F-statistic: 127.7 on 4 and 409 DF,  p-value: < 2.2e-16
```

From the summary table, we see that the t values and their respective p-values at the significance level of $\alpha = 0.05$, are very small, so we are sure that each regresssor variables are individually significant in explaining the total variation.

Confidence Intervals

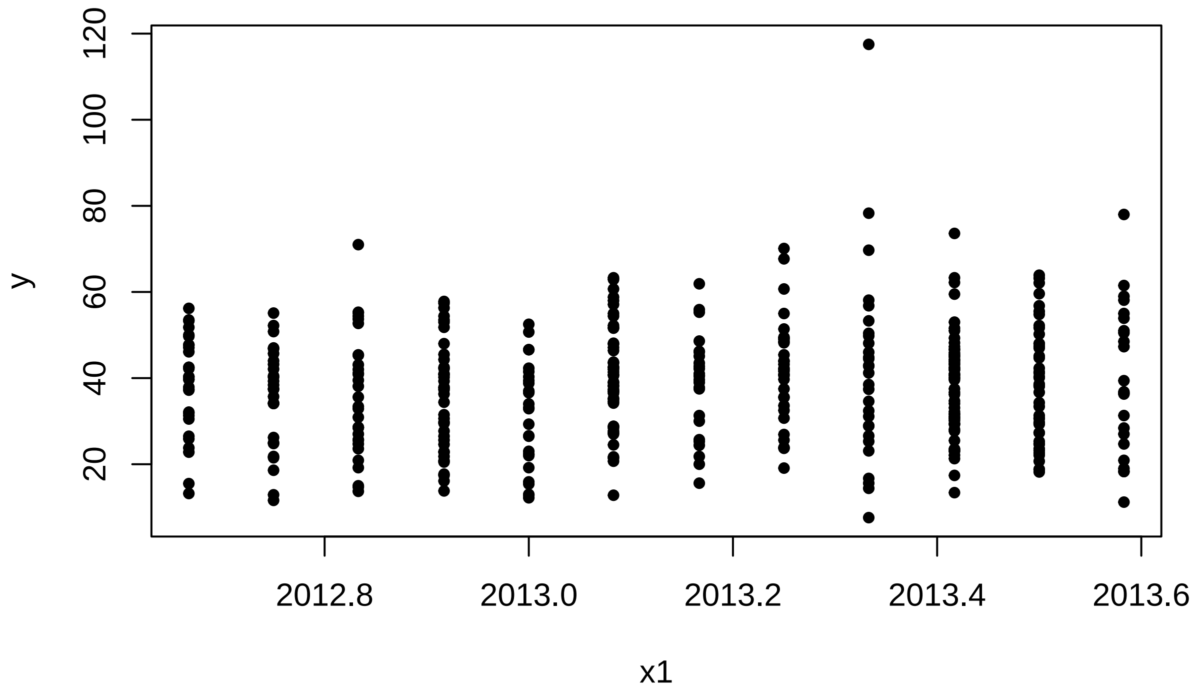
```
confint(fit2)
```

```
##                     2.5 %        97.5 %
## (Intercept) -1.790833e+04 -5.269161e+03
## x1           2.638797e+00  8.917288e+00
## x2          -3.321839e-01 -1.767721e-01
## x3          -6.393719e-03 -4.632275e-03
## x4           8.808293e-01  1.634899e+00
```

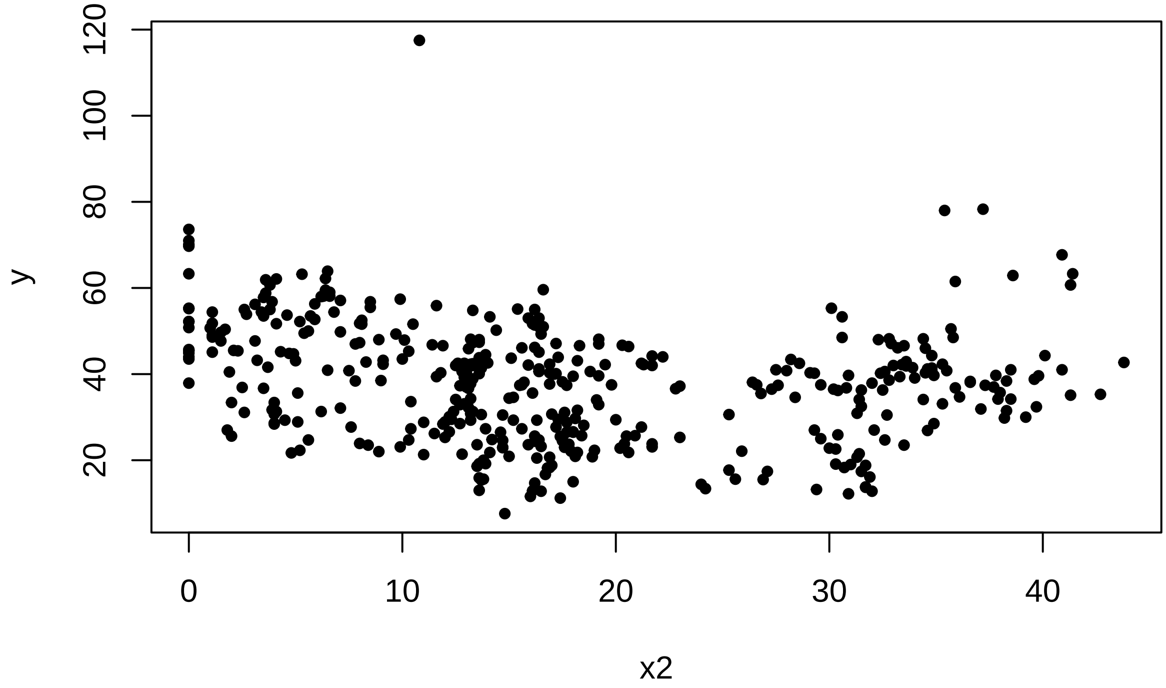
Above are the confidence intervals for each of the regressor variables at the 95% confidence interval. At 1 unit of change for any of the regressor variable, the mean response of y changes.

If the regressor variable β_j is greater than 1, it signifies that the mean of y changes more slowly than x_j, and if $\beta_j = 1$, it signifies that the mean of y changes roughly at the same rate as x_j. ### Plotting y vs. regressors

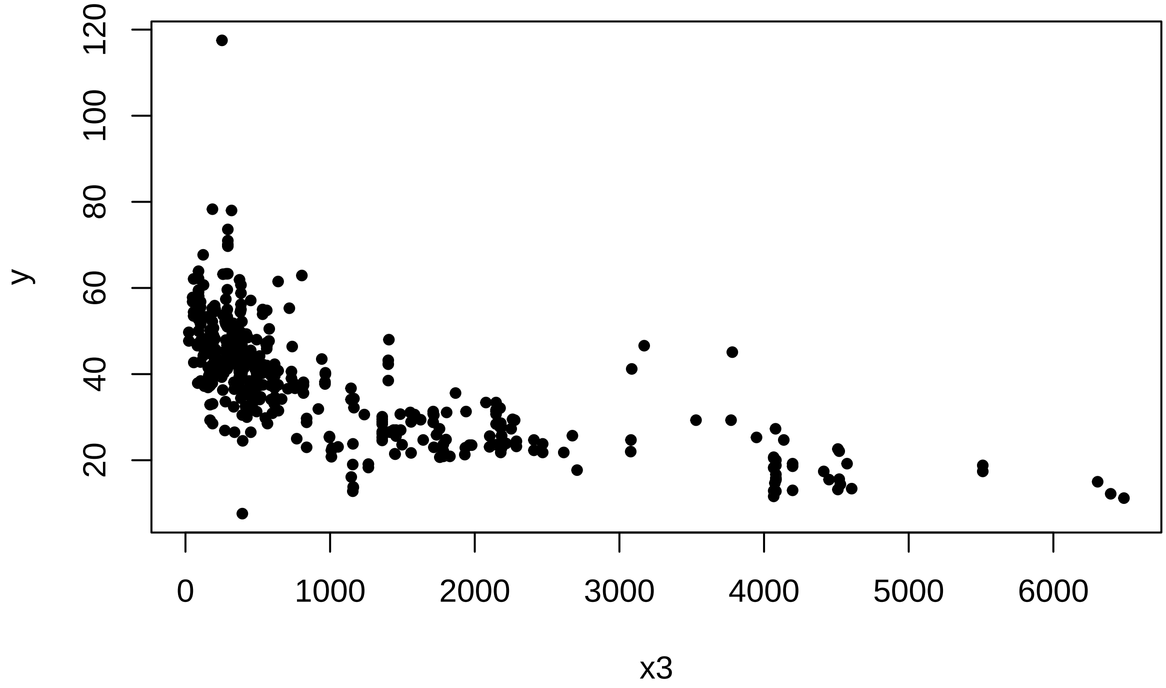
```
plot(x1,y,pch=20)
```



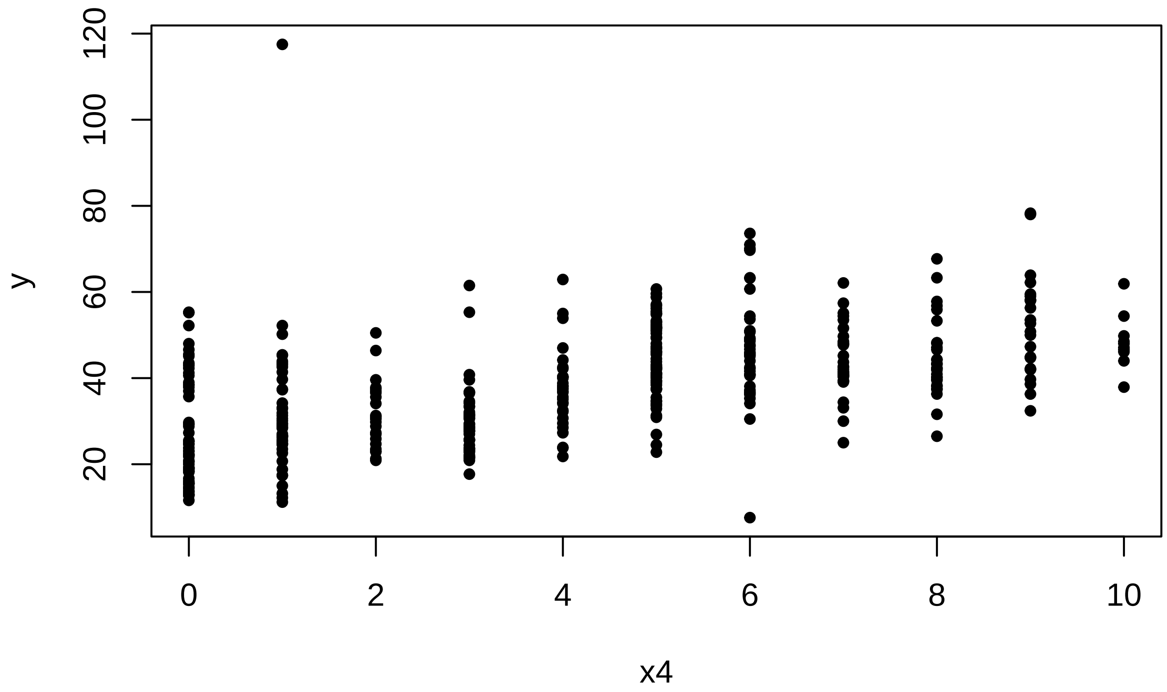
```
plot(x2,y,pch=20)
```



```
plot(x3,y,pch=20)
```



```
plot(x4,y,pch=20)
```



From the 4 plots, we can assume the following:

- **y vs x1**: There does not seem to be a consistent pattern in when plotting price vs. Time of the year.
- **y vs. x2**: We expected a negative linear relationship for older houses, but it appears that the spread of prices appear to follow a polynomial pattern that would require higher order polynomial regression.
- **y vs. x3**: Here we see a very clear negative linear relationship between price and distance to a metro station. This is most likely because houses near large cities would tend to have more metro stations than rural areas.
- **y vs. x4**: Here we see a very weak linear positive relationship between price and number of convenience stores. This could be explained by greater presence of convenience stores in large urban areas, as opposed to rural areas.

ANOVA, F-test, Lack of Fit Test

```
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.389  -5.630  -0.987   4.306  76.006
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.159e+04  3.215e+03  -3.605 0.000351 ***
## x1           5.778e+00  1.597e+00   3.618 0.000334 ***
## x2          -2.545e-01  3.953e-02  -6.438 3.40e-10 ***
## x3          -5.513e-03  4.480e-04 -12.305  < 2e-16 ***
## x4           1.258e+00  1.918e-01   6.558 1.65e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.118 on 409 degrees of freedom
## Multiple R-squared:  0.5553, Adjusted R-squared:  0.5509
## F-statistic: 127.7 on 4 and 409 DF,  p-value: < 2.2e-16
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## x1          1    585     585   7.0401  0.008281 **
## x2          1   3441    3441  41.3884 3.492e-10 ***
## x3          1  34857   34857 419.2766 < 2.2e-16 ***
## x4          1   3576    3576  43.0105 1.647e-10 ***
## Residuals 409  34003      83
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Sum of Squares, Mean Squares

- $SS_R = 585 + 3441 + 34857 + 2576 = 41459$
- $SS_{Res} = 34003$
- $SS_{Total} = 41459 + 34004 = 75463$
- $MS_{Res} = 83$
- $\sigma = \sqrt{MS_{Res}} = 9.1104$

Multiple Coefficients of Determination

- $R^2 = 0.5553$ –> R^2 isn’t very high, but not low either.
- Adjusted $R^2 = 0.5509$

Multiple Correlation Coefficient

- $R = \sqrt{R^2} = 0.74518$ –> If $\hat{\beta}_j$ is greater than 1, since R → 1 ,x_j and y are highly positively related.
- $R = -\sqrt{R^2} = -0.74518$ –> If $\hat{\beta}_j$ is less than 1, since R → -1, x_j and y are highly negatively related.

F-test for the complete model

- **Implement hypothesis**:

$H_0 : \hat{\beta}_2 = \hat{\beta}_3 = \hat{\beta}_4 = 0,$

H_a: At least one of $\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ are not zero, meaning that the model is significant in predicting y.

- Test Statistic: F(model) = **127.7**
- p value: $2.2 * 10^{-16} < 0$, reject H_0, model is significant.

Lack of Fit test

Here we compare the full model (including x5, x6) to our reduced model (x1,x2,x3,x4) to determine the lack of fit,

```
anova(fit,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6
## Model 2: y ~ x1 + x2 + x3 + x4
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    407 31933
## 2    409 34003 -2   -2070.1 13.192 2.811e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full model is not adequate.

Confidence Intervals and Prediction Intervals

```
head(predict(fit2,df,interval="confidence"))
```

```
##        fit      lwr      upr
## 1 45.94023 43.54707 48.33339
## 2 46.64102 44.71102 48.57102
## 3 45.62754 43.91119 47.34389
## 4 45.14796 43.63401 46.66192
## 5 44.35119 42.65987 46.04251
## 6 30.50417 28.37054 32.63781
```

Here we see the the first 6 intervals for the prediction intervals for the predicted y, and the confidence interval for the mean response y at 95% confidence.

Variance Inflation Factor

```
#install.packages(car)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

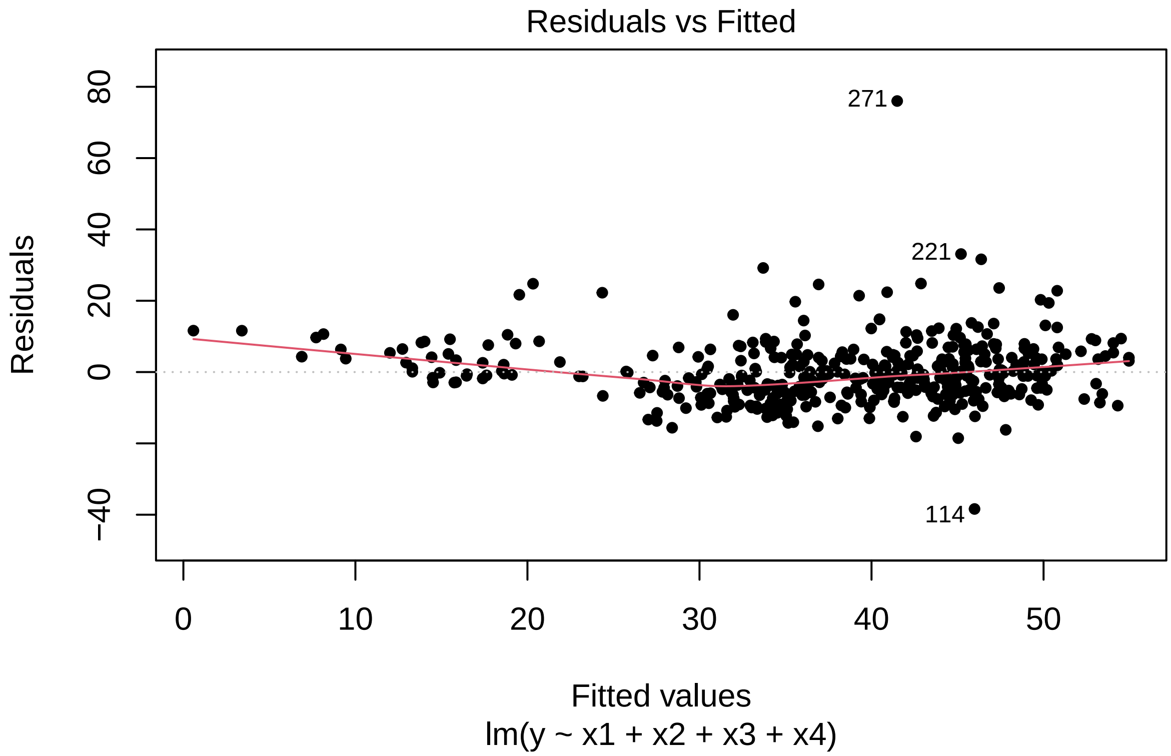
```
vif(fit2)
```

```
##       x1       x2       x3       x4
## 1.007254 1.007479 1.588412 1.585588
```

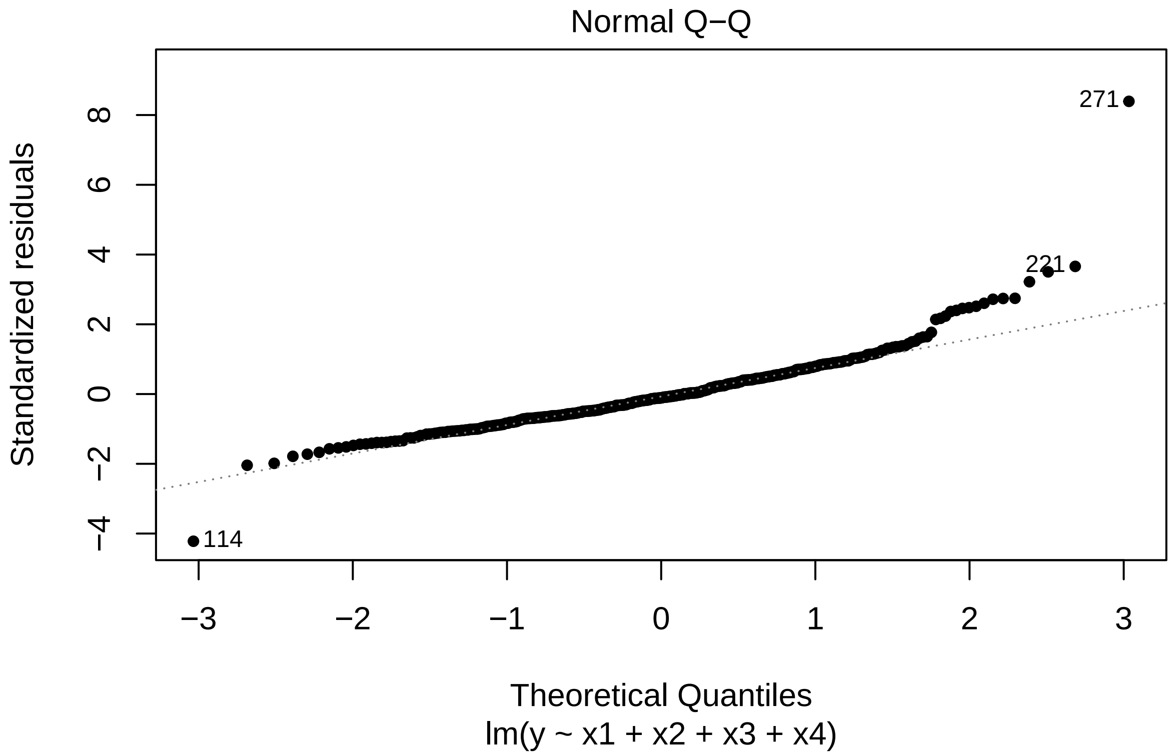
From the VIF values for x1 x2 x3 and x4, seeing as none of them are over 5, it is safe to say non of the regressors contribute too much to the standard error of the regression.

Residual Plots

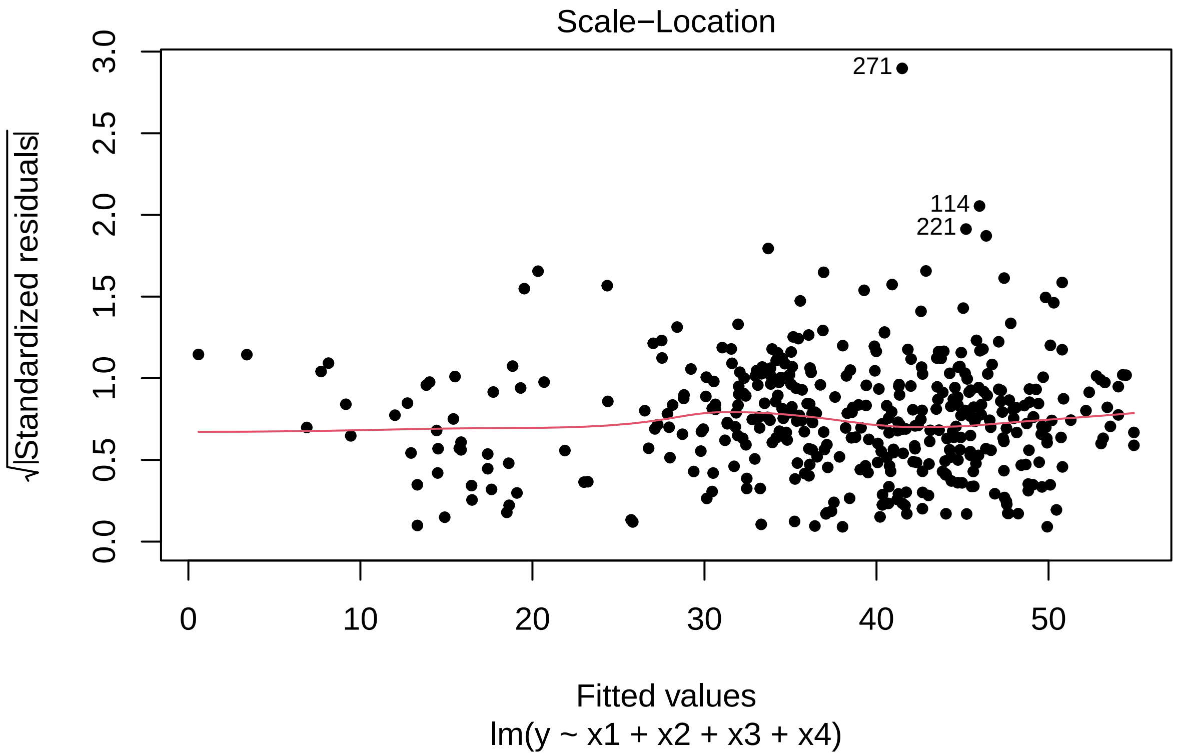
```
plot(fit2,which=1,pch=20)
```



```
plot(fit2,which=2,pch=20)
```



```
plot(fit2,which=3,pch=20)
```



- Residual vs Fitted Graph: it appears that the residuals aren't very spread out, with a heavy concentration of smaller residuals near the right side of the graph. This is likely to violate constant variance assumption.

- QQ plot: The population seems to be normally distributed.
- Scale Location: The variance of the residuals seem to fall within a certain range, matching our observation from the QQ plot.

Conclusion

As seen in section 3 of the analysis, the SSR of 41459, SSRes of 34004 and SStotal of 75463, gives us an R2 score of 0.5553. It is evident that where our model is somewhat proficient in prediction of real estate prices, using the regressor variables X1: Sale date, X2: House Age, X3: Distance to a major transit system, and X4: the number of convenience stores near the unit.

Prior to our analysis, we predicted that given the three regressor variables, we would yield a much higher R2 score than our result. We estimate that this is due to the nature of real life data, where not everything can be perfectly explained through a model.

Given the fact that the overall model's f-statistic and the t-statistics of the regressor variables are very significant at the 95% confidence interval, we believe it is safe to say that our model would be a well functioning model in predicting future data sets of real estate prices.