House Price Analysis

Quanliang Chen
Weiling Luo
Jingxian Shu
Jingjing Yu

Introduction

- Datasets : House prices in Sindian District, New Taipei City, Taiwan based on 6 factors
- Variables
  - Y.house.price.of.unit.area (Response Variable) – House price per unit area.
  - X1.transaction.date – Date of property transaction.
  - X2.house.age – Age of the house (in years).
  - X3.distance.to.the.nearest.MRT.station – Distance to the closest MRT station (in meters).
  - X4.number.of.convenience.stores – Number of convenience stores nearby.
  - X5.latitude – Geographic latitude of the property.
  - X6.longitude – Geographic longitude of the property.

Data Description

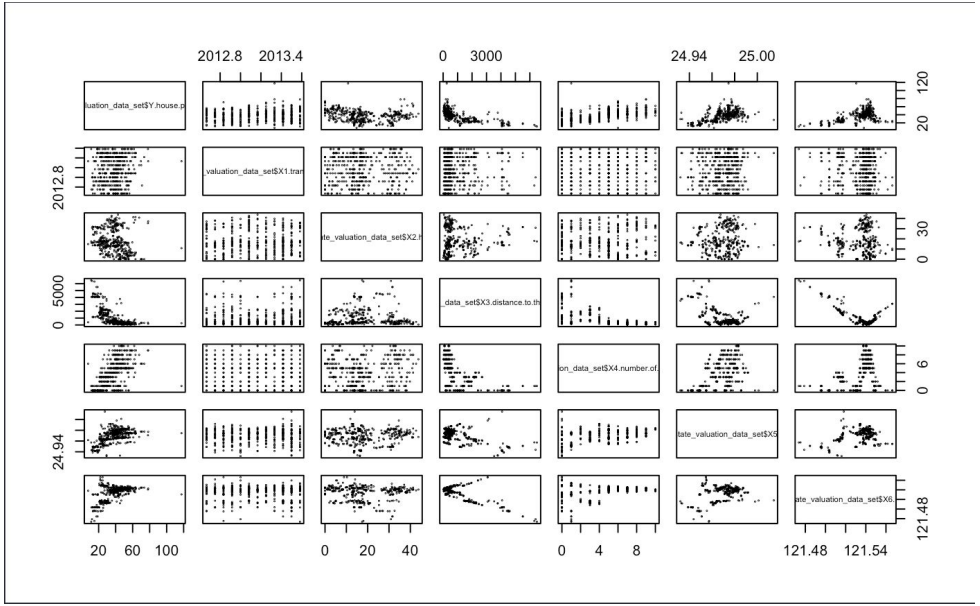
- Transaction Date: Normally distributed
- House Age: Normally distributed
- Distance to Nearest MRT Station: Right Skewed
- Number of Convenience Stores: Right Skewed
- Latitude: Normally distributed
- Longitude:Left Skewed
- Transaction Date: Normally distributed
- House Age: Normally distributed
- Distance to Nearest MRT Station: Right Skewed
- Number of Convenience Stores: Right Skewed
- Latitude: Normally distributed
- Longitude:Left Skewed

Description: df [8 × 3]

	Variable <chr>	Mean <dbl>	SD <dbl>
No	No	207.500000	1.196558e+02
X1.transaction.date	X1.transaction.date	2013.148971	2.819672e-01
X2.house.age	X2.house.age	17.712560	1.139248e+01
X3.distance.to.the.nearest.MRT.station	X3.distance.to.the.nearest.MRT.station	1083.885689	1.262110e+03
X4.number.of.convenience.stores	X4.number.of.convenience.stores	4.094203	2.945562e+00
X5.latitude	X5.latitude	24.969030	1.241020e-02
X6.longitude	X6.longitude	121.533361	1.534718e-02
Y.house.price.of.unit.area	Y.house.price.of.unit.area	37.980193	1.360649e+01

Scatter Plot Matrix of Original Variables

- Transaction Date (X1) vs. House Price (Y): No correlation
- House Age (X2) vs. House Price (Y): Negative correlation
- Distance to Nearest MRT Station (X3) vs. House Price (Y): Negative correlation
- Number of Convenience Stores (X4) vs. House Price (Y): Positive correlation
- Latitude (X5) and Longitude (X6) vs. House Price (Y): No correlation



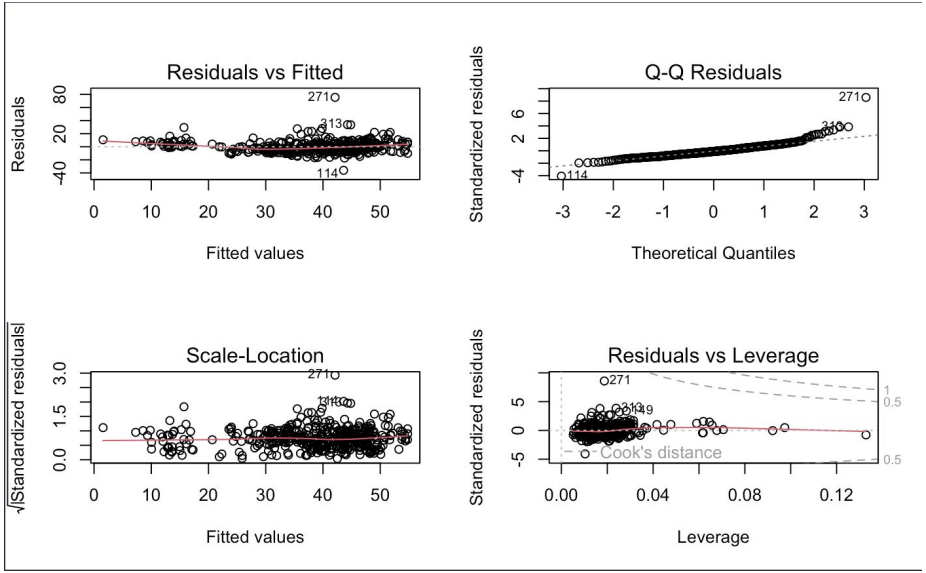
Full Model Fitting and R Output

- R-squared (R^2): The Multiple R-squared value is 0.5834.
- p-value of ANOVA: p-value: < 2.2e-16

```
Call:
lm(formula = Y.house.price.of.unit.area ~ ., data = Real_estate_valuation_data_set)

Residuals:
    Min      1Q  Median      3Q     Max
-36.003  -5.196  -0.990   4.181  75.384

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -1.404e+04  6.788e+03  -2.068  0.03927 *
No                                     -3.593e-03  3.653e-03  -0.984  0.32590
X1.transaction.date                     5.079e+00  1.559e+00   3.259  0.00121 **
X2.house.age                           -2.708e-01  3.855e-02  -7.026 9.04e-12 ***
X3.distance.to.the.nearest.MRT.station -4.521e-03  7.189e-04  -6.289 8.28e-10 ***
X4.number.of.convenience.stores         1.129e+00  1.882e-01   6.000 4.37e-09 ***
X5.latitude                             2.247e+02  4.458e+01   5.040 7.02e-07 ***
X6.longitude                           -1.442e+01  4.863e+01  -0.297  0.76691
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.858 on 406 degrees of freedom
Multiple R-squared:  0.5834,    Adjusted R-squared:  0.5762
F-statistic: 81.21 on 7 and 406 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: Y.house.price.of.unit.area
                                       Df Sum Sq Mean Sq  F value    Pr(>F)
No                                      1     62      62   0.7964   0.37271
X1.transaction.date                     1    568     568   7.2410   0.00742 **
X2.house.age                            1   3470    3470  44.2205 9.495e-11 ***
X3.distance.to.the.nearest.MRT.station  1  34892   34892 444.6857 < 2.2e-16 ***
X4.number.of.convenience.stores         1   3551    3551  45.2575 5.893e-11 ***
X5.latitude                             1   2054    2054  26.1809 4.806e-07 ***
X6.longitude                            1      7       7   0.0880   0.76691
Residuals                             406  31857      78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

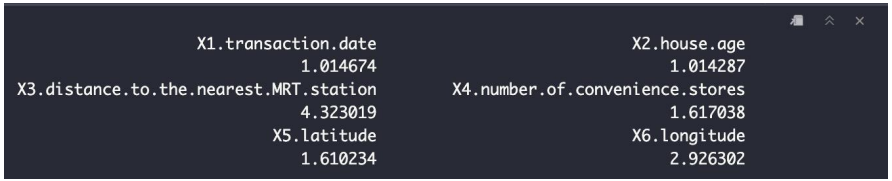
Diagnostic Plots and Model Assumptions

- Residuals vs. Fitted: Linearity assumption is satisfied
- Normal Q-Q Plot: Normality assumption is somewhat satisfied
- Scale-Location Plot: Homoscedasticity assumption is satisfied
- Residuals vs. Leverage: Model fit



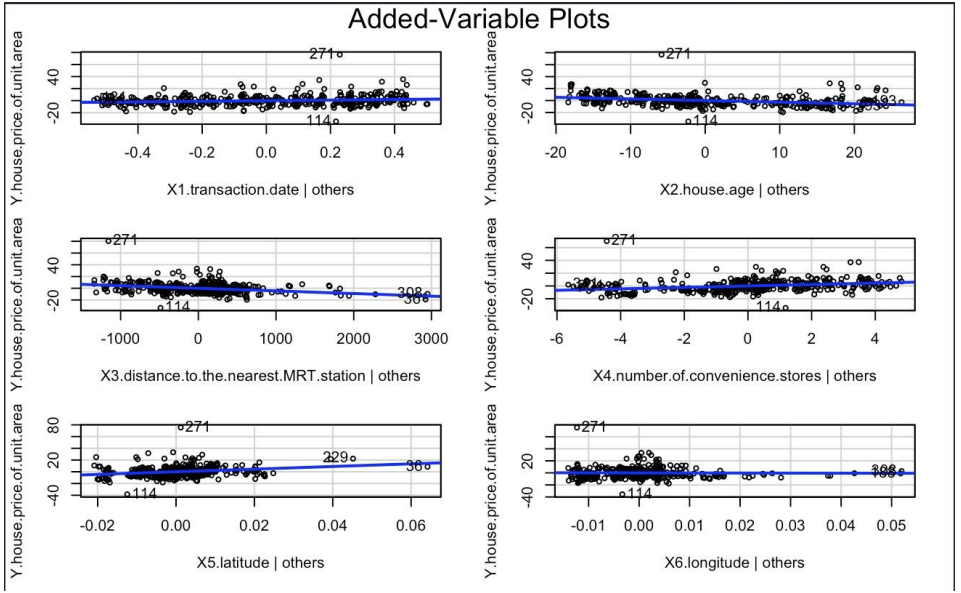
VIF

- X1.transaction.date: 1.014674
- X2.house.age: 1.014287
- X3.distance.to.the.nearest.MRT.station: 4.323019
- X4.number.of.convenience.stores : 1.617038
- X5.latitude: 1.610234
- X6.longitude: 2.926302



Added Variable Plots

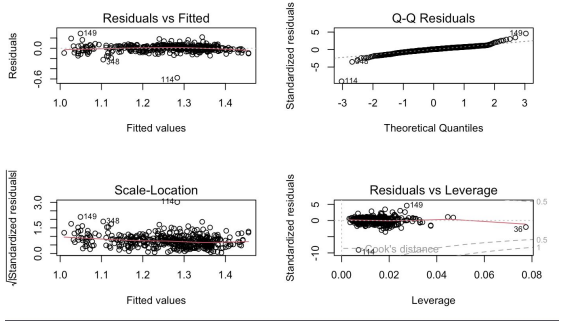
- X1.transaction.date: weak relationship
- X2.house.age: slight negative trend
- X3.distance.to.the.nearest.MRT.station: clear negative trend: clear negative trend
- X4.number.of.convenience.stores : positive trend
- X5.latitude: strong positive trend
- X6.longitude: weak relationship



Consideration of Transformations

- λ = 0.1818182
- Variable need to be transform:
  - Y.house.price.of.unit.area (Response Variable) – House price per unit area.
  - X3.distance.to.the.nearest.MRT.station – Distance to the closest MRT station (in meters).

```
[1] 0.1818182

Call:
lm(formula = log(Y_log) ~ X1.transaction.date + X2.house.age +
    log(X3.distance.to.the.nearest.MRT.station) + X4.number.of.convenience.stores +
    X5.latitude, data = Real_estate_valuation_data_set)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57631 -0.02983  0.00577  0.03524  0.28816

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.767e+02  2.332e+01  -7.576 2.41e-13 ***
X1.transaction.date                      4.747e-02  1.127e-02   4.213 3.10e-05 ***
X2.house.age                            -1.718e-03  2.790e-04  -6.157 1.77e-09 ***
log(X3.distance.to.the.nearest.MRT.station) -5.616e-02  4.051e-03 -13.862  < 2e-16 ***
X4.number.of.convenience.stores          2.817e-03  1.509e-03   1.866   0.0627 .
X5.latitude                              3.315e+00  2.915e-01  11.372  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06373 on 408 degrees of freedom
Multiple R-squared:  0.705,     Adjusted R-squared:  0.7014
F-statistic:   195 on 5 and 408 DF,  p-value: < 2.2e-16
```



Variable Selection Methodology

- Adjusted R^2: The highest Adjusted R^2 is 0.681826527 (Model 5)
- Model 5 has the lowest AIC (-1242.5596), AICc (-1242.2851), and BIC (-1218.4044).
- Final Regression Model:
  - ln(Real_estate_valuation_data_set$Y.house.price.of.unit.area) = B0 + B1 * X1.transaction.date + B2 * X2.house.age + B3 * X3.distance.to.the.nearest.MRT.station + B4 * X4.number.of.convenience.stores + B5 * X5.latitude + e

```
6 Variables  (and intercept)
                                Forced in Forced out
X1.transaction.date                 FALSE      FALSE
X2.house.age                        FALSE      FALSE
X3.distance.to.the.nearest.MRT.station FALSE   FALSE
X4.number.of.convenience.stores     FALSE      FALSE
X5.latitude                         FALSE      FALSE
X6.longitude                        FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         X1.transaction.date X2.house.age X3.distance.to.the.nearest.MRT.station
X4.number.of.convenience.stores
1  ( 1 ) " "                 " "          "*"                                    " "
2  ( 1 ) " "                 " "          "*"                                    " "
3  ( 1 ) " "                 "*"          "*"                                    " "
4  ( 1 ) " "                 "*"          "*"                                    "*"
5  ( 1 ) "*"                 "*"          "*"                                    "*"
6  ( 1 ) "*"                 "*"          "*"                                    "*"
         X5.latitude X6.longitude
1  ( 1 ) " "         " "
2  ( 1 ) "*"         " "
3  ( 1 ) "*"         " "
4  ( 1 ) "*"         " "
5  ( 1 ) "*"         " "
6  ( 1 ) "*"         "*"
```

Description: df [6 × 4]

Adjusted_R2 <dbl>	AIC <dbl>	AICc <dbl>	BIC <dbl>
0.003285732	-773.7874	-773.7292	-765.7357
0.037369029	-787.1982	-787.1009	-775.1206
0.607512690	-1157.6334	-1157.4871	-1141.5300
0.643036579	-1195.9210	-1195.7156	-1175.7917
0.681826527	-1242.5596	-1242.2851	-1218.4044
0.681116880	-1240.6532	-1240.2994	-1212.4721

Final Model R Output and Interpretation

- R^2 (Multiple R-squared): R^2: 0.6857
- p-value of ANOVA: p-value: < 2.2e-16
- X1.transaction.date: 1.358e-01
- X2.house.age: -6.977e-03
- X3.distance.to.the.nearest.MRT.station: -1.495e-04
- X4.number.of.convenience.stores: 2.766e-02
- X5.latitude: 7.883e+00

```
Call:
lm(formula = Y_log ~ X1.transaction.date + X2.house.age + X3.distance.to.the.nearest.MRT.station +
    X4.number.of.convenience.stores + X5.latitude, data = Real_estate_valuation_data_set)

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 ***
X1.transaction.date                     1.358e-01  3.890e-02   3.491 0.000533 ***
X2.house.age                           -6.977e-03  9.625e-04  -7.248 2.13e-12 ***
X3.distance.to.the.nearest.MRT.station -1.495e-04  1.226e-05 -12.194  < 2e-16 ***
X4.number.of.convenience.stores         2.766e-02  4.694e-03   5.892 7.97e-09 ***
X5.latitude                             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

Analysis of Variance Table

Response: Y_log
                                       Df Sum Sq Mean Sq  F value    Pr(>F)
X1.transaction.date                     1  0.363   0.363   7.3976   0.00681 **
X2.house.age                            1  2.311   2.311  47.1596 2.450e-11 ***
X3.distance.to.the.nearest.MRT.station  1 36.155  36.155 737.7156 < 2.2e-16 ***
X4.number.of.convenience.stores         1  2.298   2.298  46.8982 2.761e-11 ***
X5.latitude                             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
```

The End