Project Report - Predicting House prices from different factor

Group 3

Introduction

This report aims to develop a regression model to estimate real estate prices per unit area in Taiwan, expressed in units of 10,000 New Taiwan Dollars per Ping (where 1 Ping equals 3.3 square meters). Specifically, the research question is: How can house price per unit area be accurately predicted using factors including transaction date, house age, proximity to metro stations, number of nearby convenience stores, and geographic coordinates?

The dataset using in this project originates from the UCI Machine Learning Repository, titled "Real Estate Valuation Data Set," and consists of data collected on real estate transactions within a specific urban area in Taiwan. This publicly available and open-source dataset contains 414 observations, 8 variable. This is the link for this dataset: https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set.

To analyze this data, we first draw a scatter plot to estimate the relationship between variables. We then use a multiple linear regression model to construct the full model. We use summary information, diagnostic plots, and other tools to determine if we need to perform a transformation or model selection, or make any other changes. After finding the best final model, we check the model performance using summary, diagnostic plots, ANOVA, and so on.

The remainder of this report is structured as follows: Section 2 reports the summary information about the data and different variables, Section 3 reports how we construct the model and determine the final model, and Section 4 summarizes the project and discusses the model limitations and the real-world significance of this regression model.

Data describtion

Pre-adjustment

Before describe the data, we need do some pre-adjustment. First we need to remove the meaningless variable "NO", It does not affect the regression model. And then for X1_transaction_date, it indicate the transaction date, but it is hard to interprets. For example, 2013.250=2013 March, 2013.500=2013 June, etc. The minimum of X1 is 2012.667, so we could use this date as baseline date. we define a new variable as X1_transaction_month=(X1_transaction_date - 2012.667) * 12). It represents how many months pass since 2012.677, that is August, 2012. This change would make data easier to interpret. We would use X1_transaction_month as the predictor variable.

Overview of the Data

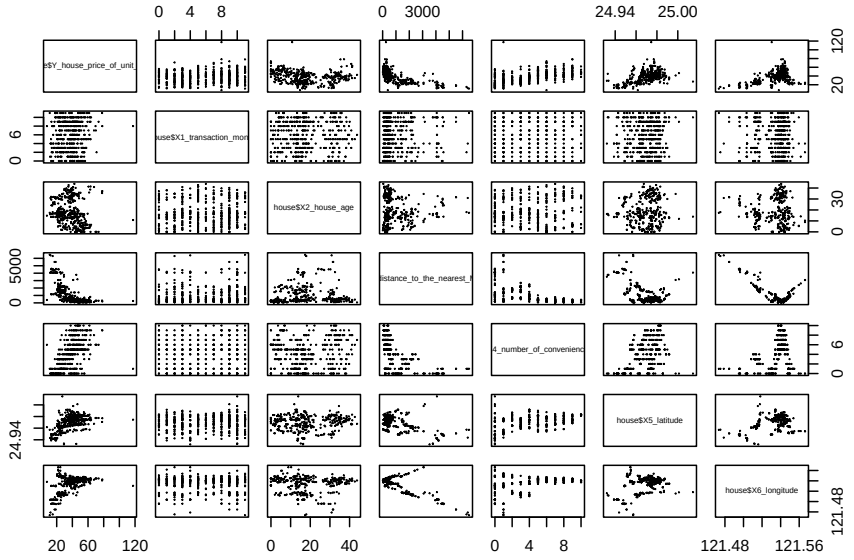
Table 1 gives the descriptive statistics of the key variables in our study, including their means, standard deviations, medians, and correlation coefficients with house price. The variables are defined as follows:
Transaction Month: The number of months relative to a baseline date. averages 5.79 months after baseline month (SD = 3.38) and shows a correlation of 0.09 with house price, suggesting later house transaction with higher price.
House Age: The age of the house in years. It has averages 17.71 years (SD = 11.39) and shows a negative correlation of -0.21 with house price, suggesting older houses tend to be slightly cheaper.
Distance to MRT: Distance to the nearest MRT station (in meters). It has a mean of 1083.89 meters (SD = 1262.11), with a -0.67 correlation to house price, indicating that farther distances strongly decrease property value.
Convenience Stores: The count of convenience stores within walking distance. It averages 4.09 (SD = 2.95) and correlates positively (0.57) with house price, implying that better convenience or amenities can raise housing values.
Latitude: The latitude coordinate of the house. It (mean = 24.97) correlates moderately (0.55) with house price, suggesting higher latitudes (further north in this region) are associated with higher prices.
Longitude: The longitude coordinate of the house. It (mean = 121.53) also shows a moderate correlation (0.52) with price, but the scatterplot reveals a potential curved relationship, indicating possible nonlinearity.
House Price: The unit-area house price (in local currency per "ping").It has a mean of 37.98 and standard deviation of 13.61, with a median of 38.45, reflecting a somewhat right-skewed distribution and possible outliers at the high end.

Distribution and Pairwise Relationships

Analyzes scatter matrix plot below and Histogram in the Appendix. We could know:
Y_house_price_of_unit, it show somewhat right skewed.
X1_transaction_date, it is a discrete variable. We could see a slight increasing trending to Y when X1 increases.
X2_house_age, it is the age of house, it show a slight down trending when age increases.
X3_distance_to_the_nearest_MRT_station, when distance increase, house price down. And we could see a left skewed, so consider transformation later.
X4_number_of_convenience_stores, the relationship between number of store and house price is positive. When number of store increase, house price increases.
X5_latitude, it's the latitude of house. When latitude increases, house price increase. It might imply the house price of Northern City is higher.
X6_longitude, it shows a curve shape, it might be not linear.

Table 1: Descriptive Statistics of Housing Variables

Variable	Mean	Std. Dev.	Median	Corr.
Transaction Month	5.79	3.38	6.00	0.09
House Age	17.71	11.39	16.10	-0.21
Distance to MRT	1083.89	1262.11	492.23	-0.67
Convenience Stores	4.09	2.95	4.00	0.57
Latitude	24.97	0.01	24.97	0.55
Longitude	121.53	0.02	121.54	0.52
House Price	37.98	13.61	38.45	NA



Results and interpretation

Construct a full model

Firstly, we use construct a full model m1.
The model is: Y = -4.079e+03 + 4.291e-01 * X1 - 2.697e-01 * X2 - 4.488e-03 * X3 + 1.133e+00 * X4 + 2.255e+02 * X5 - 1.243e+01 * X6. The Adjusted R-squared is 0.5762, F-statitic is 94.6 with p-value $<$ 2.2e-16. And for each predictor, except X6, all predictor are significant. This summary shows a good performance. We then use diagnostic plots to check the model assumption.
From Residual vs Fitted value and Scale-location, the residuals are constant and homoscedastic. The residuals are almost randomly scatter around 0 and no trending down or up.
From Q-Q plot, there is a slight deviation from normality in the tail part.
From Residual vs Leverage, we could see there are some outlier points, leverage points, and potential influential points.
The diagnostic output shows a residual is constant, almost normally distributed. We need to test the multicolinearity and consider do transformation to improve that.

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -4.079e+03  6.142e+03  -0.664  0.50705
X1_transaction_month                     4.291e-01  1.297e-01   3.307  0.00103 **
X2_house_age                            -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
X3_distance_to_the_nearest_MRT_station -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
X4_number_of_convenience_stores          1.133e+00  1.882e-01   6.023 3.83e-09 ***
X5_latitude                              2.255e+02  4.457e+01   5.059 6.38e-07 ***
X6_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
```

Figure 1: summary of full model

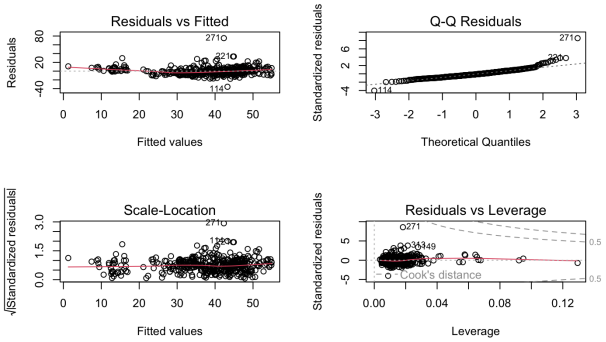


Figure 2: diagnostic plot

Transformation

Firstly, we consider to do log transformation to Y. Because Y is right-skewed, log transformation can make Y being more normality. And log transformation is easier to interprets, so we can intuitively understand the affect in Y for each predictor. Finally, log transformation can reduce the influence of outlier. And the box-cox also result a lambda 0.18, which supporting log transformation.

And then for X3, it is also right skewed. And X3 represents the Distance to MRT. When the distance is very close, each additional 100 meters will have a large impact on the house price; when the distance is already very far, an additional 100 meters will have a small impact. We consider do log transformation to it, which also make interpretation easier. The powerTransform function result a rounded value 0, which also support our Decision.

Now we get the Model m3, which is log(Y) = -5.666e+02 + 1.380e-02 * X1 -6.014e-03 * X2 -1.728e-01 * log(X3) + 1.067e-02 * X4 + 1.013e+01 * X5 + 2.620e+00 * X6.
From summary of model, we could see all predictor are significant right now, and R-squared increase from *0.5824 to 0.724*. The f-statistic 177.9 with a p-value $<$ 2.2e-16 also show the model significant.

Diagnostic analysis

And check model assumption. From Residual vs Fitted value and Scale-location, the residuals points are nearly scattered around zero, and nearly homoscedastic.
From Q-Q plot, the residuals are more normal than before, also show a good normality.
From Residual vs Leverage, we could see there are still some outliers points, leverage points. But the amount decrease, and reduced the impact of leverage points.
Overall, the model after transformation perform good and does not violate the assumption.

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                -5.666e+02  1.067e+02  -5.311 1.79e-07 ***
X1_transaction_month                        1.380e-02  3.060e-03   4.509 8.53e-06 ***
X2_house_age                               -6.014e-03  9.097e-04  -6.612 1.20e-10 ***
log(X3_distance_to_the_nearest_MRT_station) -1.728e-01  1.510e-02 -11.444  < 2e-16 ***
X4_number_of_convenience_stores             1.067e-02  4.921e-03   2.169  0.03066 *
X5_latitude                                 1.013e+01  9.648e-01  10.499  < 2e-16 ***
X6_longitude                                2.620e+00  8.898e-01   2.944  0.00342 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2077 on 407 degrees of freedom
Multiple R-squared:  0.724,     Adjusted R-squared:  0.7199
F-statistic: 177.9 on 6 and 407 DF,  p-value: < 2.2e-16
```

Figure 3: summary of full model

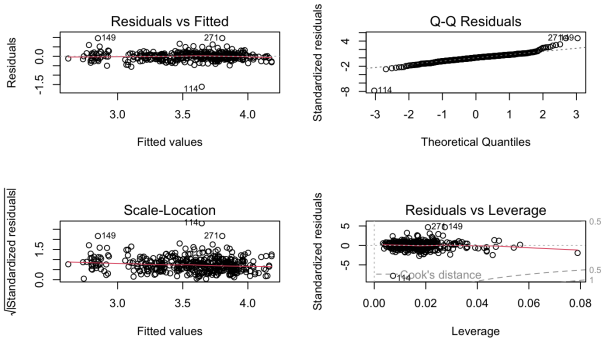


Figure 4: diagnostic plot

Check multicolinearity and model selection

all VIFs of each predictor less than 5. There is severe multicollinearity. Furthermore, we use backward step selection to select model(output shown in appendix). And result also suggest we should use all predictor in transformation model.

```
                X1_transaction_month                     X2_house_age
                            1.026494                         1.028020
log(X3_distance_to_the_nearest_MRT_station) X4_number_of_convenience_stores
                            2.736325                         2.011456
                        X5_latitude                      X6_longitude
                            1.372220                         1.785090
```

Figure 5: VIFs

Final model and interpretation.

Our final model is:

$$log(Y) = -566.6e{+}02{+}0.0138{*}X1{-}0.0060{*}X2{-}0.1728{*}log(X3){+}0.01067{*}X4{+}10.13{*}X5{+}2.62{*}X6$$

All predictor in this model are significant. The f-statistic 177.9 with a p-value $<$ 2.2e-16 also show the model significant. The R-Squared is 0.724, which mean approximately 72.4% variation can be explain by model.

Transaction month (X1): An increase of 1 month from the baseline date leads to a approximately 1.4% increase in house prices. This imply the house price is increasing over time.

House Age (X2): For each additional year of house age, the house price decreases by roughly 0.6%, demonstrating depreciation effects.

Distance to the Nearest MRT Station (X3): Each 1% increases in X3 with a 0.1728% decrease in house prices. This imply the farther away from MRT, the less convenient the transportation, the lower the price of housing.

Number of Convenience Stores (X4): Each additional nearby convenience store corresponds to a 1.1% increase in house prices.

Latitude (X5): An increase of 0.01-degree increase in latitude (~1.11 km) corresponds to about an 10.6% increase in house prices. This may imply the northern area has higher house price. We use 0.01 units here, because the are we study is small.

Longitude (X6): A 0.01-degree shift eastward (about 0.85–1 km) yields around a 2.7% increase in house prices. This may imply eastern area has higher house price.

Discussion

Summary

We analyzed a New Taipei city real estate dataset by removing an irrelevant ID variable and adjust transaction date. Log-transforming both the target price (per ping) and the distance to MRT station improved normality and model fit. The final model includes months since baseline date, house age, log-distance to MRT, number of convenience stores, latitude, and longitude. It explains about 72% of the log-price variance. House age and MRT distance have negative impacts on price and transaction month, convenience stores, and higher latitude/longitude have positive impacts prices.

Consistent With Real World

In the model, X1 imply the house price is increasing over time. This finding aligns with the report from Taipei Times, which also highlights a rising trend in housing prices (Taipei Times, 2023).
For X2, X3, X4, these imply a property with newer age, convenient transportation and close to more store has higher house unit price. This is consistent with the paper "PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction" (arXiv:2209.05471), which emphasizes the impact of property age, transportation convenience, and accessibility to amenities on real estate pricing.
For X5 and X6, these imply northern and eastern property with higher price. For, analysis, we draw the points indicating each observation in Google Map by the latitude and longitude. From the map, we could see the distribution of each observations. And we checke the price for different district in Housefeel (https://www.housefeel.com.tw/price-all/taipei-city/). The house in our dataset all located in Xindian District, with average unit price 628.3 thousand NTD per ping. Yonghe District is located north of Xindian District, which has average unit price 684.3 thousand NTD per ping. Wenshan District is located east of Xindian District, which has average unit price 986 thousand NTD per ping (price data is for 2025). This shows that the closer to the east, i.e. the closer to Wenshan District, the higher the house prices, and the closer to the north, i.e. the closer to Yonghe District, the relatively higher the house prices. This is consistent with our model result and interpretation.

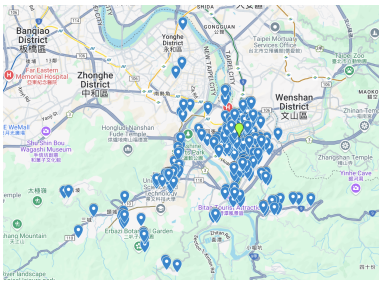


Figure 6: Map

Appendix

Reference

Taipei Times Article: Taipei Times. "Rising Trend in Housing Prices." 2023, https://www.taipeitimes.com/News/biz/archives/2023/03/07/2003795603.

arXiv Paper:
Zhang, Z., Pan, S., and Liu, Y. "PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction." arXiv preprint, 2022, https://arxiv.org/abs/2209.05471.

Housefeel Data Source:
Housefeel. "Housing Price Data for Taipei City." Housefeel, 2025. (Website content available primarily in Chinese.) Retrieved from https://www.housefeel.com.tw/price-all/taipei-city/.

Google Maps Data:
Google Maps. "Housing Price Distribution in Taipei." Google Maps, 2025, https://www.google.com/maps/d/u/0/edit?mid=1E_NBxM7kDgWVhTHdKTtgmdnHjF-BZYw&usp=sharing.

Addtional Plots and R Result

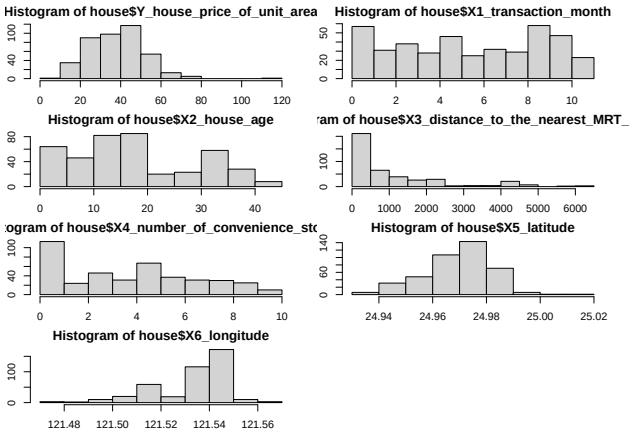
VIFs And AIC selection.

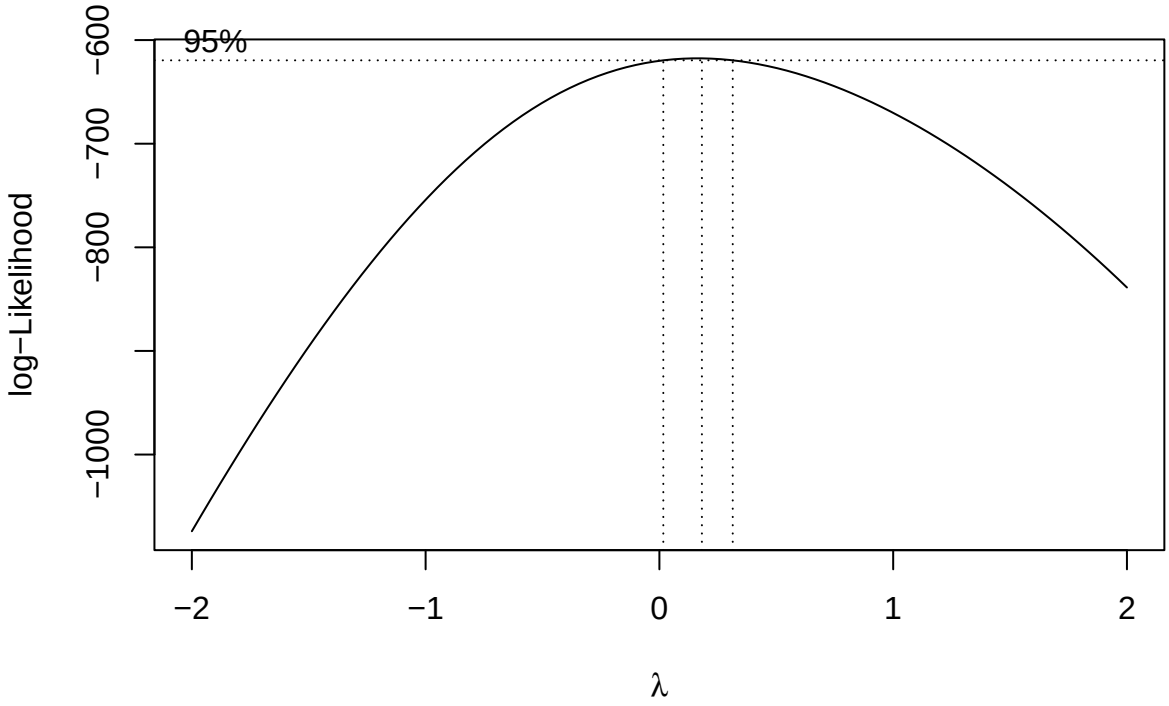
```
##                         X1_transaction_month
##                                     1.026494
##                                X2_house_age
##                                     1.028020
## log(X3_distance_to_the_nearest_MRT_station)
##                                     2.736325
##             X4_number_of_convenience_stores
##                                     2.011456
##                                 X5_latitude
##                                     1.372220
##                                X6_longitude
##                                     1.785090
```

```
## Start:  AIC=-1294.31
## log(Y_house_price_of_unit_area) ~ X1_transaction_month + X2_house_age +
##     log(X3_distance_to_the_nearest_MRT_station) + X4_number_of_convenience_stores +
##     X5_latitude + X6_longitude
##
##                                              Df Sum of Sq    RSS     AIC
## <none>                                                     17.561 -1294.3
## - X4_number_of_convenience_stores             1    0.2030 17.764 -1291.5
## - X6_longitude                                1    0.3740 17.935 -1287.6
## - X1_transaction_month                        1    0.8773 18.439 -1276.1
## - X2_house_age                                1    1.8861 19.447 -1254.1
## - X5_latitude                                 1    4.7557 22.317 -1197.1
## - log(X3_distance_to_the_nearest_MRT_station)  1    5.6511 23.212 -1180.8
```

```
##
## Call:
## lm(formula = log(Y_house_price_of_unit_area) ~ X1_transaction_month +
##     X2_house_age + log(X3_distance_to_the_nearest_MRT_station) +
##     X4_number_of_convenience_stores + X5_latitude + X6_longitude,
##     data = house)
##
## Coefficients:
##                                   (Intercept)
##                                     -5.666e+02
##                          X1_transaction_month
##                                      1.380e-02
##                                  X2_house_age
##                                     -6.014e-03
## log(X3_distance_to_the_nearest_MRT_station)
##                                     -1.728e-01
##              X4_number_of_convenience_stores
##                                      1.067e-02
##                                 X5_latitude
##                                      1.013e+01
##                                X6_longitude
##                                      2.620e+00
```

```
par(mfrow = c(4,2), mar = c(2,2,2,2))
hist(house$Y_house_price_of_unit_area)
hist(house$X1_transaction_month)
hist(house$X2_house_age)
hist(house$X3_distance_to_the_nearest_MRT_station)
hist(house$X4_number_of_convenience_stores)
hist(house$X5_latitude)
hist(house$X6_longitude)
```




95%
log−Likelihood
−600
−700
−800
−1000
−2
−1
0
1
2
λ