Predictive Modeling of Taiwanese Real Estate

Elaine Prathiksha
and
Eric Antillon

April 9, 2025

Abstract

This study investigates the factors shaping residential property prices in southern Taipei, focusing on the roles of MRT accessibility, geographic location, and local amenities. Using multiple regression with LASSO variable selection, we find that proximity to transit and convenience stores, as well as advantageous northeastern locations, are associated with higher housing values. Notably, the relationship between house age and price is non-linear: both modern and vintage homes tend to command higher prices, while middle-aged properties are less valued, reflecting a U-shaped effect. Although the model explains a substantial portion of price variation, some residual non-normality and heteroscedasticity suggest caution when interpreting uncertainty estimates. These findings provide valuable insights into how transit access, neighborhood amenities, and architectural era interact to influence urban housing markets, offering guidance for developers, investors, and city planners in optimizing property value and urban development strategies.

1 Introduction

Urbanization is accelerating at an unprecedented pace in Taiwan, and with it, the need for efficient city planning has become more critical than ever. In rapidly growing metropolitan areas, public transportation networks and the availability of nearby amenities play a central role in shaping both the desirability and the value of residential properties. Home buyers, investors, and real estate developers alike are continuously seeking insights that can help them make more informed decisions about property investments. One important factor influencing these decisions is the proximity to mass rapid transit (MRT) stations, which significantly enhances the convenience and accessibility of daily commutes.

This study investigates how multiple factors can be used to predict house unit prices in the southern part of Taiwan's capitol, Taipei. While distance to MRT stations is a key variable, other factors such as house age, transaction date, geographic location (latitude and longitude), and the number of nearby convenience stores are also considered. By analyzing these diverse attributes, we aim to uncover patterns that explain how various urban dynamics influence property values.

The central research question guiding this study is:

How do factors such as proximity to MRT stations, house age, transaction timing, and nearby amenities collectively influence residential property prices in southern Taipei, Taiwan?

By addressing this question, we aim to provide actionable insights for stakeholders in the Taiwanese real estate market. These insights can assist in evaluating property prices, identifying prime locations for new developments, and fostering economically vibrant and accessible urban environments.

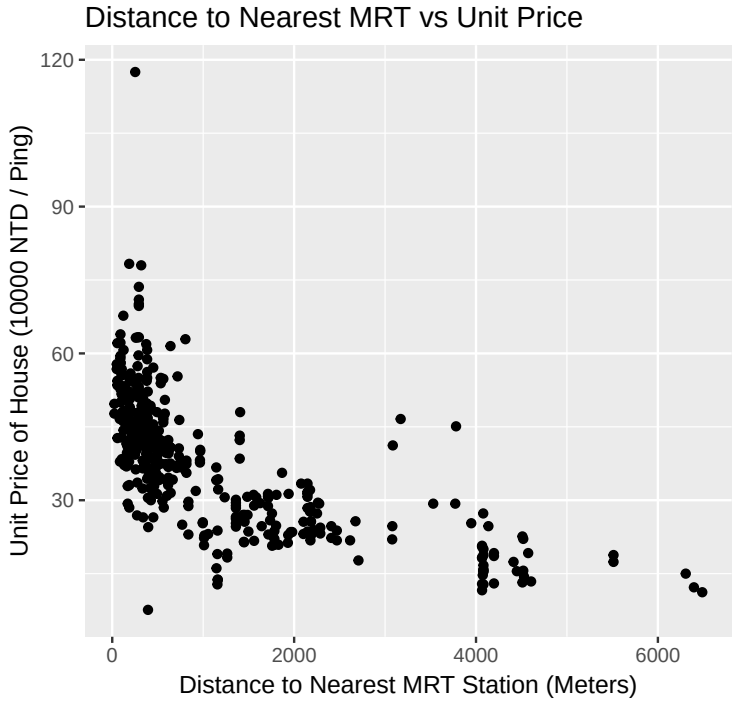

Figure 1: Scatterplot of Unit Price of House (10000 NTD / Ping) by Distance to Nearest MRT Station (Meters).

Data Description

This dataset contains observations for 414 houses in Taiwan. Each row in the dataset represents an individual house, with several variables recorded to describe its characteristics and surroundings. Below, we provide definitions for key terms and details about the variables.

Term Definitions

- **MRT Station:** A Mass Rapid Transit Station, which is a high-capacity transport system designed to move large numbers of people efficiently.
- **New Taiwan Dollar (NTD):** The local currency used in Taiwan.
- **Ping:** A local unit of measurement, where 1 Ping = 3.3 m^2.

Variables The dataset includes the following variables:

- **trans_date:** The transaction date, recorded as a numerical value in the format *year.proportion_of_year_completed*. For example, 2013.250 corresponds to March 2013, while 2013.500 corresponds to June 2013.
- **house_age:** The age of the house, measured in years.
- **dist_to_mrt:** The distance to the nearest MRT station, measured in meters.
- **num_stores:** The number of convenience stores within walking distance of the house (recorded as an integer).
- **lat:** The latitude coordinate of the house, measured in degrees.
- **lon:** The longitude coordinate of the house, measured in degrees.
- **unit_price:** The price of the house per unit area, expressed in units of 10,000 NTD per Ping.

	trans_date	house_age	dist_to_mrt	num_stores	lat	lon	unit_price
1	2012.92	32.00	84.88	10.00	24.98	121.54	37.90
2	2012.92	19.50	306.59	9.00	24.98	121.54	42.20
3	2013.58	13.30	561.98	5.00	24.99	121.54	47.30
4	2013.50	13.30	561.98	5.00	24.99	121.54	54.80
5	2012.83	5.00	390.57	5.00	24.98	121.54	43.10
6	2012.67	7.10	2175.03	3.00	24.96	121.51	32.10

Table 1: Sample of observations from the dataset

Key Observations The dataset provides valuable insights into real estate trends in Taiwan by capturing spatial and temporal aspects of housing prices. Some notable features of the data include:

- Using the mean `lat` and `lon`, we find that the "average" house is located in Taiwan, New Taipei City, Xindian District (Northeastern part of Taiwan, but southern part of the capitol). See Figure 5 in the appendix to see the distribution of all houses.

- This data comes from a one-year timespan (Mid 2012 - Mid 2013).
- Exactly $\frac{2}{3}$ of the houses in the dataset are within 1km of an MRT station.

2 Model Selection and Validation

Model Selection After variable selection via LASSO regression and ANOVA comparisons of nested models with interaction terms, the final model is specified as:

$$\begin{aligned} \text{UnitPrice}_i = \beta_0 &+ \beta_1 \log(\text{DistanceToMRT}_i) \\ &+ \beta_2 \text{NumStores}_i \\ &+ \beta_3 \text{TransactionDate}_i \\ &+ \beta_4 \text{ScaledLatitude}_i \\ &+ \beta_5 \text{ScaledLongitude}_i \\ &+ \beta_6 \text{ScaledHouseAge}_i \\ &+ \beta_7 (\text{ScaledLatitude}_i \times \text{ScaledLongitude}_i) \\ &+ \beta_8 \text{I}(\text{ScaledHouseAge}_i^2) \\ &+ \epsilon_i, \end{aligned} \tag{1}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

Variable Selection Cross-validation identified the optimal regularization parameter (λ) for the LASSO regression model, which retained all predictor variables—a result underscoring their collective significance in explaining price variation. ANOVA and partial F-tests further validated the statistical necessity of interaction terms (all $p < 0.05$). Notably, the inclusion of a longitude $\times$ latitude interaction term accounts for geographic spatial heterogeneity, enabling the model to discern how property values shift across distinct regional patterns. A quadratic effect was also added for house age as it has a quadratic relationship with unit price (see Figure 4).

Key Transformations

- `DistanceToMRT` log-transformed to address its non-linear relationship with price as shown in Figure 1 (see Figure 3 in the appendix for transformed scatter plot).
- Latitude/Longitude as well as house_age were standardized (converted to Z-scores) to mitigate multicollinearity in interaction and quadratic terms. This will change our model interpretation of these variables.
- An outlier with a Cook's distance of approximately 0.22, which was also the sole observation exhibiting an unusually high unit price, was removed from the analysis.

Diagnostic Validation Added-variable plots (Figure 2) confirm approximately linear relationships between predictors and response after adjusting for other variables, satisfying the linearity assumption. However, distinct trends emerge for houses in the high unit-price range, suggesting potential confounding factors such as local school quality that warrant further investigation. The multicollinearity assumption remains tenable with all variance inflation factors (VIFs) below 3.

Residual analysis reveals two primary concerns:

- A Shapiro-Wilk test ($p < 0.01$) indicates non-normal error distribution, and a histogram of these residuals seems to be slightly skewed (see Figure 6)
- Heteroscedasticity manifests through funnel-shaped patterns in residual-vs-predicted plots

Despite these violations, we proceed with cautious interpretation of the model. The normality violation may marginally affect prediction intervals but leaves point estimates unbiased. Heteroscedasticity primarily impacts the efficiency of standard errors, suggesting hypothesis tests and confidence intervals should be interpreted with reduced confidence.

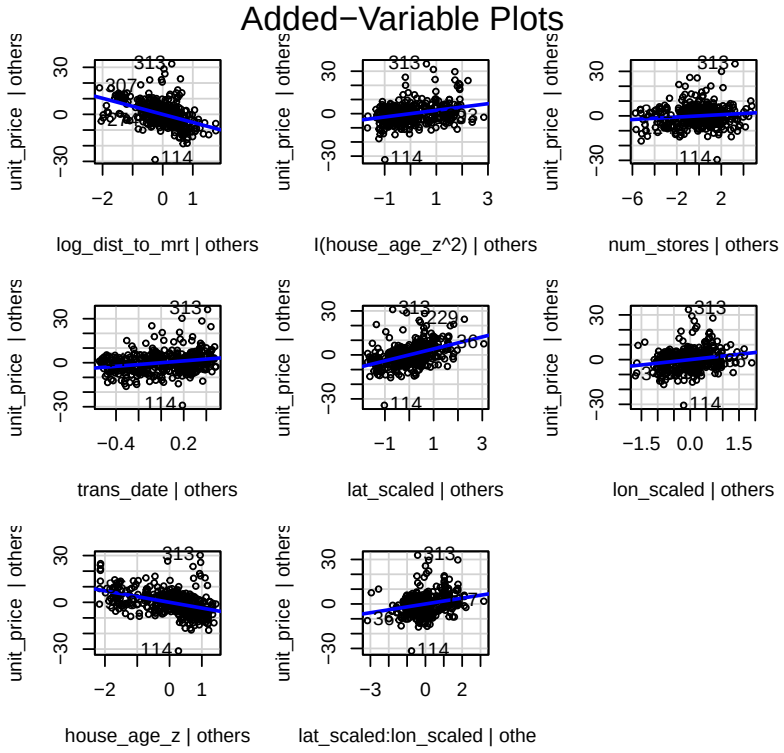

Figure 2: Added variable plot of the variables in final model.

3 Analyses, Results, and Interpretation

The fitted model summary is shown by the table below. The model explains 73% of variation in the unit price of the Taiwanese houses (adjusted $R^2 = 0.7304$). The fit of our model can be assessed by examining the added variable plot shown in Figure 2. With an RMSE of approximately 6.7, we can conclude that, on average, our model's predicted unit prices deviate from the actual observed prices by around 67,000 NTD per Ping. This error represents roughly 18% of the average home value.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12286.3444	2413.7909	-5.09	0.0000
log_dist_to_mrt	-5.1215	0.5166	-9.91	0.0000
I(house_age_z^2)	2.3456	0.3697	6.34	0.0000
num_stores	0.4084	0.1653	2.47	0.0139
trans_date	6.1357	1.1993	5.12	0.0000
lat_scaled	4.0909	0.4128	9.91	0.0000
lon_scaled	2.3867	0.5615	4.25	0.0000
house_age_z	-3.6794	0.3723	-9.88	0.0000
lat_scaled:lon_scaled	1.9677	0.3725	5.28	0.0000

Table 2: Regression summary output for the model shown in Equation 1.

Geographical Interaction Effects

The model includes a significant interaction between latitude and longitude (1.97, $p < 0.001$), displaying a spatial dependence pattern.

- **Latitude effect:** Holding longitude constant, a 1 SD northward move changes unit price by $(40,910 + 19,680 \times \text{lon\_scaled})$ NTD/Ping on average.
- **Longitude effect:** Holding latitude constant, a 1 SD eastward move changes unit price by $(23,870 + 19,680 \times \text{lat\_scaled})$ NTD/Ping on average.

For example:

- At mean longitude (lon_scaled = 0), a 1 SD northward move increases price by **40,910 NTD/Ping** on average
- At +1 SD longitude, a 1 SD northward move increases price by **60,590 NTD/Ping** on average
- At mean latitude (lat_scaled = 0), a 1 SD eastward move increases price by **23,870 NTD/Ping** on average
- At +1 SD latitude, a 1 SD eastward move increases price by **43,550 NTD/Ping** on average

Transit and Market Trends

- **MRT Proximity**: 10% increase in distance reduces price by **4,880 NTD/Ping** on average (95% CI: \$-5,843–\$-3,914)[1]
- **Annual Appreciation**: Prices rise by **61,360 NTD/Ping/year** on average (95% CI: \$37,780–84,930)

Property Characteristics

- **House Age Effect**: Shows a U-shaped relationship:
  - At mean age: 1 SD older decreases price by **36,790 NTD/Ping** on average (95% CI: \$-44,110–\$-29,470)

[1] Calculated as $-5.122 \times \ln(1.10) \times 10,000$

  - Quadratic term (2,346 NTD/Ping/SD^2) indicates diminishing depreciation
  - Depreciation turns to appreciation at **0.78 SD** above mean age

- **Retail Access**: Each nearby store adds **4,084 NTD/Ping** on average (95% CI: $835–$7,334)

Practical Considerations

- The intercept (-122,863,444 NTD/Ping) reflects prices when all metrics are 0. This has no practical application as no predictions will be made for houses sold in the year 0 AD.

- **Strategic Priorities**:
  - Target northeast regions for maximum geographical premium (properties closer to central Taipei)
  - Target properties with close proximity to MRT stations and surrounding stores
  - The market is currently bullish, making it an excellent time to invest, as property appreciation rates are significantly outpacing depreciation
  - Modern houses or vintage houses have the most value, with average houses (about 20 years old) being worth less.

4 Conclusions

Our findings suggest that housing prices in southern Taipei are influenced by a combination of geographical, temporal, and structural factors. We found that homes located in the northeastern regions, those with higher latitude and longitude, tend to be worth significantly more, especially when both values are high. This refers to the houses closer to the center of Taipei and reinforces the idea of a location-based premium, likely due to higher demand in those areas. Additionally, proximity to MRT stations and access to convenience stores were both positively associated with property value, emphasizing the role of accessibility and amenities in urban real estate pricing. Time effects aligned with expectations, as houses sold later in the study period had higher prices. Notably, both modern and vintage homes tended to command higher values, while middle-aged properties were less valued, highlighting changing buyer preferences across different eras of construction. Although our model explains over 73% of the variability in unit price, some assumptions were violated, particularly the normality of residuals and constant variance. This does not invalidate our findings, but it does mean that prediction intervals and p-values should be interpreted with caution, as our standard errors may be somewhat inflated or inefficient. However, the observed patterns align with practical expectations and offer valuable insights for real estate investors, planners, and policy makers interested in Taiwan's urban housing dynamics.

APPENDIX

Here, a description of all help received, as well as additional plots are shared.

Gen AI

Here we include how we received assistance from Gen AI (which prompts were used).

- Prompt: "How to format _____ in LaTeX" Solution: We followed the formatting that was given to us.
- Prompt: "How to reword _____ to make it more understandable for the reader" Solution: Taking into account the recommendations, we used Gen AI to help reword certain words or phrases in our analysis.
- Prompt: "Why is the interaction effect between lat:lon NA" Solution: Standardization of the lat and lon fixes this issue and provides added interpretability.
- Note that we turned to Gen AI for the interpretation of the quadratic coefficient.

Additional Figures

Here we include any other figures that could prove useful to the reader.

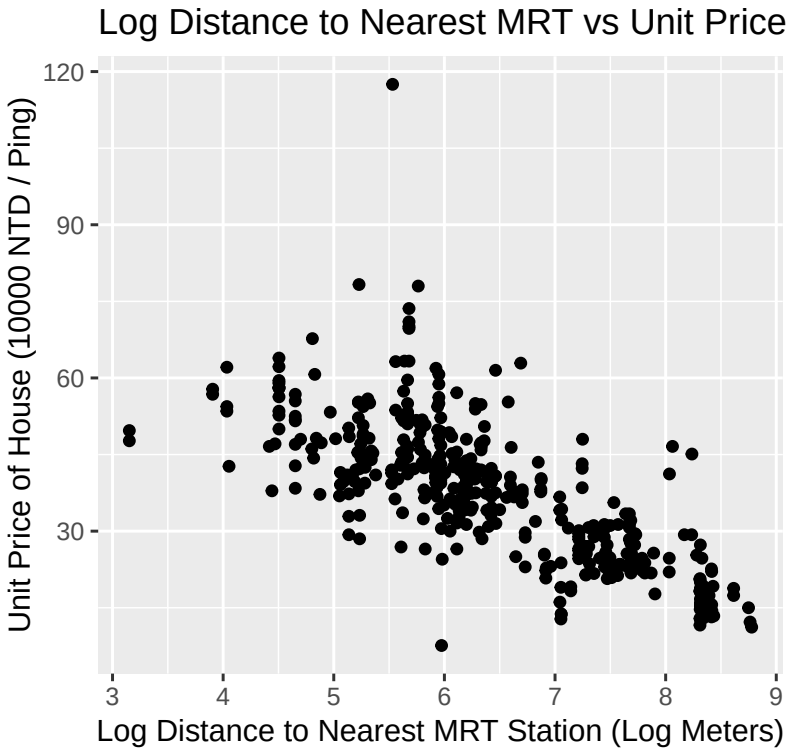

Figure 3: Distance to MRT vs unit price after log transformation

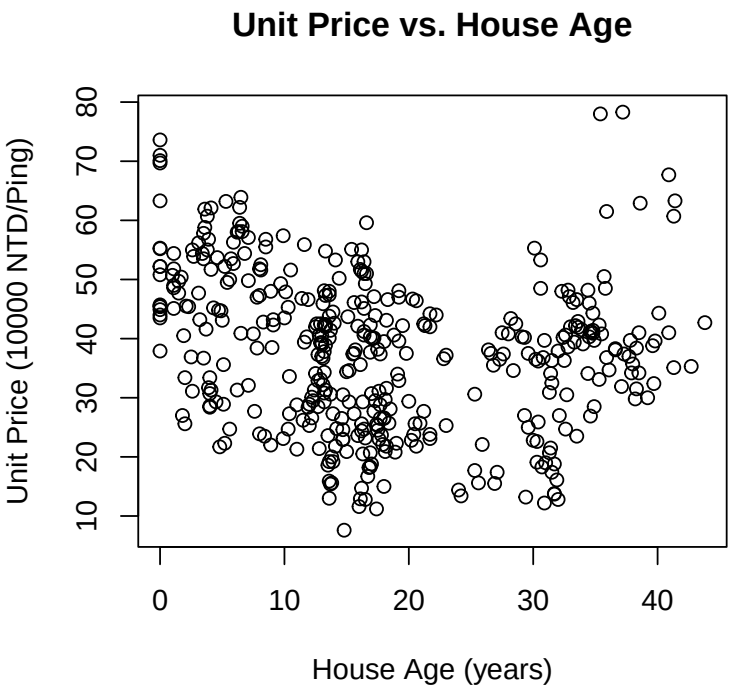

Figure 4: Unit Price vs House Age

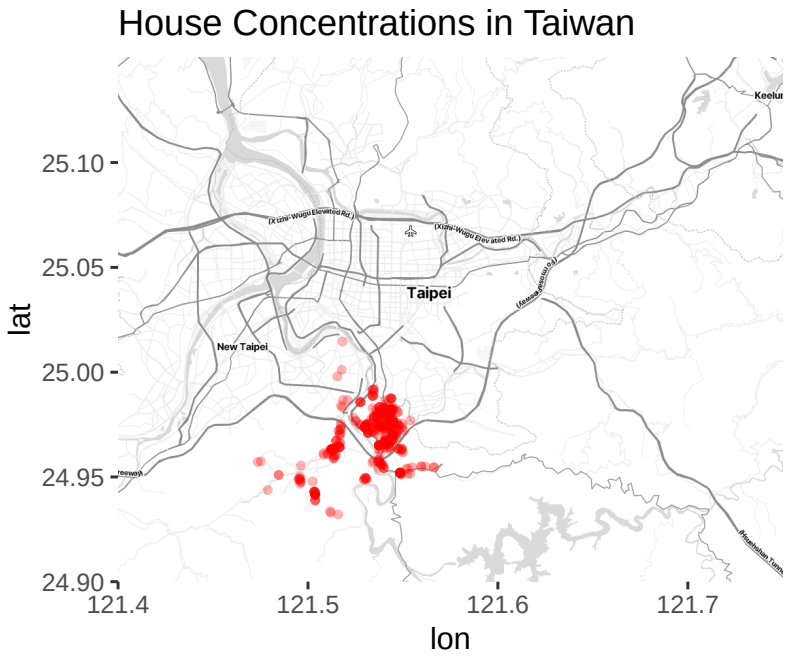

Figure 5: Distribution of houses in the dataset (located in Taipei, Taiwan)

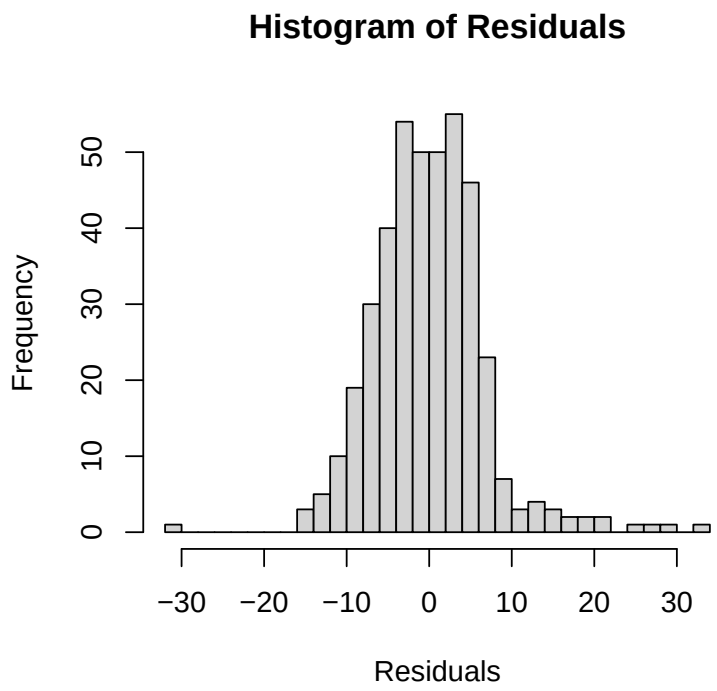

Figure 6: Histogram of residual errors