Multivariate Regression on Real Estate Prices

Kyla Ronquillo*
kyronquillo@gbox.adnu.edu.ph
Ateneo de Naga University
Naga City, PH

Abstract

This paper shows how Multivariate Linear Regression is used to predict real estate prices.

Keywords

Multivariate Regression, Real Estate, Price Prediction

ACM Reference Format:
Kyla Ronquillo. 2024. Multivariate Regression on Real Estate Prices. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Multivariate regression is a method used to understand how multiple factors affect a single outcome. It extends simple linear regression by using more than one input (independent variable) to predict a result (dependent variable). This approach is useful when trying to understand complex relationships in data.

One of the main strengths of multivariate regression is its ability to analyze the combined effect of several features at once. This helps provide a more complete and accurate prediction compared to looking at one variable at a time. It also allows us to see the individual impact of each variable while controlling for the others, which is useful in decision-making and understanding real-world situations.

In this paper, we apply multivariate regression to a dataset on housing prices. The goal is to find out which features are most helpful in predicting house prices. By using this method, we can better understand which factors have the strongest influence on the value of a property.

2 METHODOLOGY

This section outlines the steps taken to apply multivariate linear regression to the housing price dataset, from data preprocessing to obtaining results and interpretation.

3 Methodology

This section outlines the steps taken to apply multivariate linear regression to the housing price dataset, from data preprocessing to obtaining results and interpretation.

3.1 Data Selection

The dataset used contains information on real estate transactions, including the house price per unit area and several features such as house age, distance to the nearest MRT station, number of convenience stores, transaction date, and geographic coordinates.

3.2 Feature Selection

We selected six features as independent variables:

- X1: Transaction date
- X2: House age
- X3: Distance to the nearest MRT station
- X4: Number of convenience stores
- X5: Latitude
- X6: Longitude

The target variable (dependent variable) is:

- Y: House price of unit area

3.3 Data Preprocessing

- **Normalization:** All features and the target variable were normalized using z-score normalization to ensure that variables with larger scales did not dominate the model.
- **Conversion to Arrays:** The features and target were extracted and converted into NumPy arrays for processing.

3.4 Model Training

A manual implementation of multivariate linear regression was developed, and this model was trained on the normalized dataset using the gradient descent algorithm for 10,000 iterations with a learning rate of 0.01. The theta values (coefficients) were updated at each iteration until the cost function reached a minimum.

- The hypothesis function was defined as a linear combination of the input features and their corresponding weights (theta parameters):

$$h_\theta(X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n = \theta^T X$$

- A cost function (mean squared error) was defined to evaluate the accuracy of the predictions:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

- Gradient descent was used to minimize the cost function by iteratively updating the weights (theta values) using the

*

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

rule:

$$\theta_j := \theta_j - \alpha \cdot \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)\cdot x_j^{(i)}$$

- The learning rate (α) and number of iterations were chosen experimentally to ensure convergence.

3.5 Model Evaluation

- After training, predictions were generated using the final theta values.
- The model’s performance was evaluated using the R^2 score, which measures how well the regression predictions fit the actual data.
- A scatter plot was generated to compare actual vs. predicted values visually.

4 RESULTS AND DISCUSSION

After training the model using gradient descent with a learning rate of 0.01, the following results were obtained:

- **Final Cost (MSE):** 0.2088
- R^2 **Score:** 0.5824

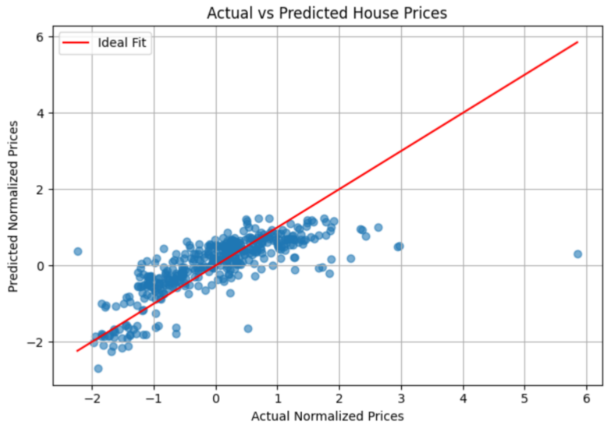

Figure 1: Comparison of actual vs. predicted house prices per unit area.

The R^2 score of 0.5824 indicates that approximately 58.24% of the variance in house prices can be explained by the selected features in the model. While this reflects a moderate level of predictive accuracy, there remains room for improvement, potentially through the inclusion of more relevant variables or non-linear modeling approaches.

4.2 R^2 Scores by Individual Feature

To assess each feature’s independent explanatory power, we also calculated the R^2 score for each feature using univariate regression. The results are summarized in Table 1.

As shown in Table 1, the most predictive single feature is the distance to the nearest MRT station, with an R^2 score of approximately 0.50. This aligns with the multivariate results, which also identified it as the most influential variable. Other features like latitude and the number of convenience stores provide moderate explanatory power. Meanwhile, transaction date and longitude have negligible R^2 values on their own. This confirms that some features contribute meaningful insights only when considered in combination with others, reinforcing the value of multivariate regression over univariate approaches in this context.

Table 1: Univariate R^2 Scores Per Feature

Feature	R^2 Score
X1 Transaction Date	0.0117
X2 House Age	0.0723
X3 Distance to MRT Station	0.5019
X4 Number of Convenience Stores	0.1665
X5 Latitude	0.2697
X6 Longitude	0.0002

4.1 Feature Coefficients

The final learned coefficients (theta values) from the model are as follows:

- Intercept (θ_0): **-0.0000**
- X1 Transaction Date (θ_1): **0.1066**
- X2 House Age (θ_2): **-0.2258**
- X3 Distance to the Nearest MRT Station (θ_3): **-0.4162**
- X4 Number of Convenience Stores (θ_4): **0.2453**
- X5 Latitude (θ_5): **0.2056**
- X6 Longitude (θ_6): **-0.0140**

The coefficients provide insights into the influence of each feature on housing prices per unit area:

- The most significant negative influence is from **distance to the nearest MRT station** ($\theta_3 = -0.4162$), which aligns with expectations: properties farther from transportation hubs tend to be less valuable.
- **House age** ($\theta_2 = -0.2258$) also negatively impacts price, suggesting older houses are generally less expensive.
- **Number of convenience stores** ($\theta_4 = 0.2453$) and **latitude** ($\theta_5 = 0.2056$) are positively associated with higher housing prices.
- **Transaction date** ($\theta_1 = 0.1066$) has a modest positive impact, possibly indicating appreciation of property values over time.
- **Longitude** ($\theta_6 = -0.0140$) has the least influence among all features.

These results highlight the relative importance of location and accessibility in determining real estate values in the dataset. Further analysis with additional features or advanced models may yield improved predictive performance and deeper insights.

5 CONCLUSION

This study applied multivariate linear regression to analyze factors influencing real estate prices. The model achieved an R^2 score of 0.5824, indicating that approximately 58.24

Analysis of the learned coefficients revealed that distance to the nearest MRT station ($\theta_3 = -0.4162$) had the most significant negative impact on price, followed by house age ($\theta_2 = -0.2258$), while number of convenience stores ($\theta_4 = 0.2453$) and latitude ($\theta_5 = 0.2056$) were positively correlated with price. These insights align with expectations in real-world real estate valuation, where accessibility and nearby amenities tend to drive up property value.

To further interpret the influence of individual variables, we computed R^2 scores for each feature using univariate regression. The results showed that while distance to MRT station independently explains a substantial portion of variance ($R^2 = 0.50$), other features such as longitude and transaction date have very low explanatory power when used alone. This reinforces the idea that multivariate regression is essential for capturing the combined influence of features.

Overall, this work demonstrates the effectiveness of multivariate regression for understanding price dynamics in real estate.

References

UCLA Statistical Consulting Group. n.d. *Multivariate Regression Analysis*. Retrieved August 7, 2023 from https://stats.oarc.ucla.edu/stata/dae/multivariate-regression-analysis/