Housing Price Prediction using Linear Regression with One Feature

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

Abstract

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

Keywords

Linear Regression, Real Estate, Price Prediction

ACM Reference Format:
Kyla Ronquillo. 2024. Housing Price Prediction using Linear Regression with One Feature. 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

The rapid urbanization and increasing population in metropolitan areas have led to fluctuating housing prices, making it essential to develop predictive models that can estimate housing costs accurately. In this study, we explore the use of linear regression, a fundamental machine learning algorithm, to predict housing prices based on various input features.

2 METHODOLOGY

This section outlines the steps taken to apply 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 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.

*

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

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

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 rule:

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

4 RESULTS AND DISCUSSION

Table 1 presents the results of univariate linear regression applied to individual features of the housing dataset. For each feature, the learned parameters θ_0 (bias) and θ_1 (weight) were optimized using gradient descent. The final cost represents the mean squared error (MSE) after training the model solely on that feature.

Feature	Theta_0	Theta_1	Final Cost
Transaction Date	-0.0000	0.0875	0.4962
House Age	0.0000	-0.2106	0.4778
Distance to MRT Station	0.0000	-0.6736	0.2731
Number of Convenience Stores	0.0000	0.5710	0.3370
Latitude	-0.0000	0.5463	0.3508
Longitude	0.0000	0.5233	0.3631

Table 1: Learned Parameters and Final Cost for Each Feature

Among the features, *Distance to MRT Station* achieved the lowest final cost of 0.2731, identifying it as the most predictive individual variable for housing price. Conversely, *Transaction Date* resulted in the highest cost (0.4962), indicating minimal contribution to price prediction when used alone.

Other features such as *Number of Convenience Stores*, *Latitude*, and *Longitude* demonstrated moderate predictive power, with positive θ_1 values suggesting a direct correlation with house price. These findings serve as a foundation for feature selection in subsequent multivariate regression models.

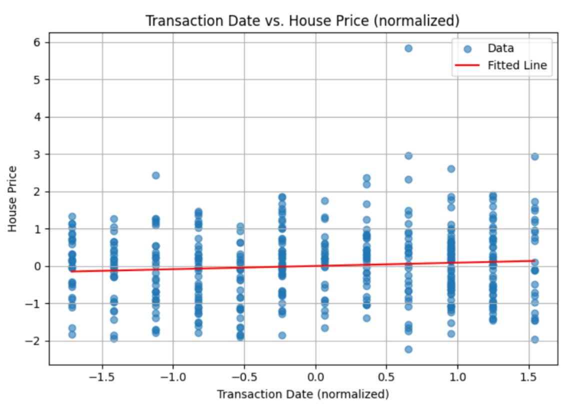

Figure 1: Transaction Date vs. House Price

Figure 1 shows a nearly flat regression line, consistent with the minimal θ_1 value. This confirms that transaction date has negligible predictive value for housing prices in this dataset.

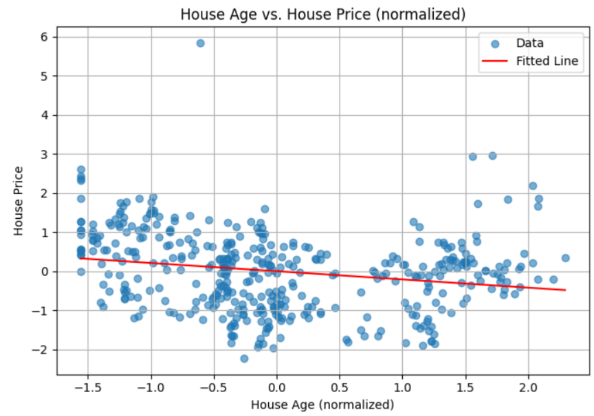

Figure 2: House Age vs. House Price

As shown in Figure 2, there is a slight negative slope, indicating an inverse relationship: older houses tend to be priced slightly lower. However, the scatter suggests a weak correlation.

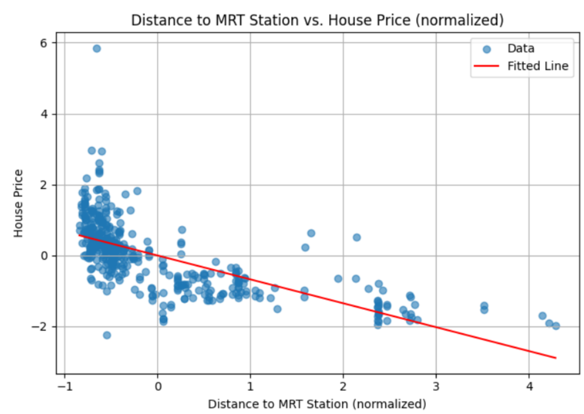

Figure 3: Distance to MRT Station vs. House Price

Figure 3 illustrates a clear negative correlation. The steeper slope indicates that house prices decrease as the distance from the MRT station increases—likely due to the premium on accessibility and convenience associated with proximity to public transport.

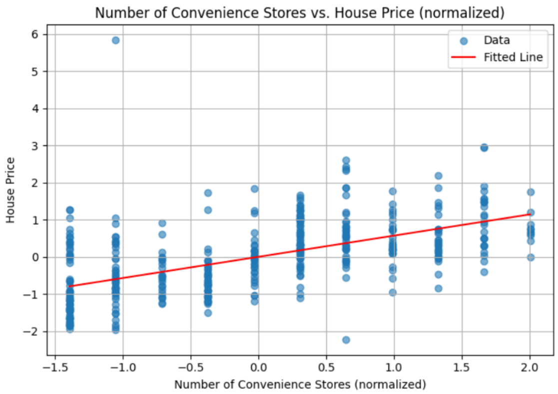

Figure 4: Number of Convenience Stores vs. House Price

A positive slope in Figure 4 suggests that houses located near more convenience stores tend to be priced higher. This reflects the desirability of easy access to daily necessities.

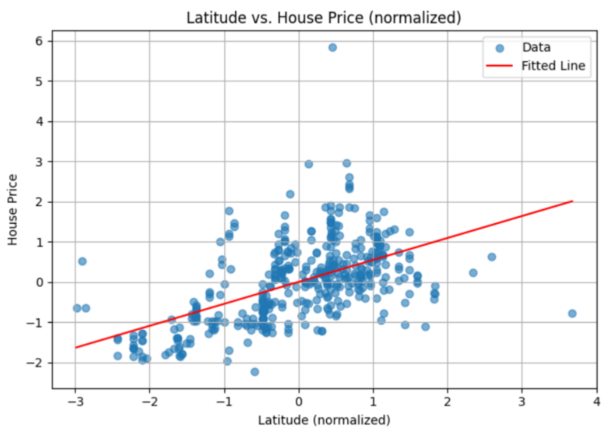

Figure 5: Latitude vs. House Price

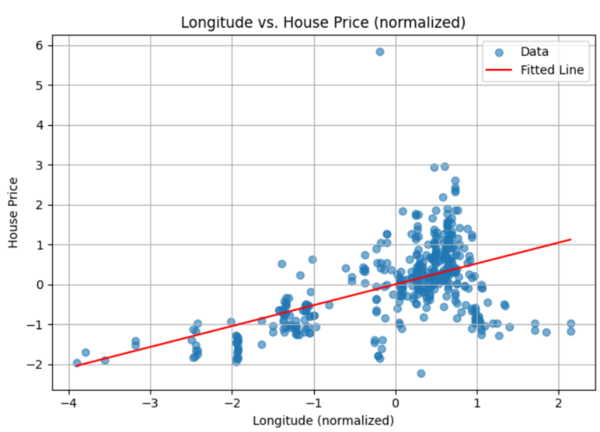

Figure 6: Longitude vs. House Price

Figures 5 and 6 both reveal strong positive relationships between geographic location and house price. The relatively steep upward slopes suggest that location, as defined by latitude and longitude, is a key factor influencing property values in the dataset.

Overall, the univariate analysis highlights the varying predictive power of each feature. These insights not only guide initial feature selection but also emphasize the importance of combining multiple variables in a multivariate regression framework for improved accuracy.

5 CONCLUSION

This study applied univariate linear regression to evaluate the individual predictive power of six features in relation to housing prices. The analysis revealed that *Distance to MRT Station* is the most influential single predictor, exhibiting the lowest final cost and a clear negative correlation with house price. In contrast, *Transaction Date* showed the weakest predictive performance, with the highest final cost and a nearly flat regression line.

Features such as *Number of Convenience Stores*, *Latitude*, and *Longitude* demonstrated moderate predictive capabilities, with positive associations to house price. While *House Age* showed a slight negative trend, its influence was minimal.

Overall, the results highlight the varying degrees of importance among individual features and underscore the value of geographic and accessibility-related variables in predicting housing prices. These findings provide a solid foundation for future work involving multivariate regression, where combining features is expected to yield more accurate and robust predictive models.

References

GeeksforGeeks. (2023). *Univariate Linear Regression in Python*. Retrieved May 13, 2025, from https://www.geeksforgeeks.org/univariate-linear-regression-in-python/