Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2025.22572

Hybrid House Price Prediction Model by Integration of Simple Linear Regression and Cubic Spline Interpolation

Yiwen Tang

Department of Mathematics, The University of Hong Kong, Hong Kong, China
u3609469@connect.hku.hk

Abstract: In this day and age, house-purchase has become a crucial consideration for almost everyone, whether seeking a residence or making an investment. Therefore, analyzing the relationship between these factors and house prices is vitally important for both buyers and sellers to make informed decisions. Some researchers have used a linear regression model that can predict the house price for a company or individual. This paper focuses on a target sample data set of "Houses in London" from Kaggle. The author firstly provides an analysis of two methods to model the relationship between the dependent variable, House Price, and an independent variable, Square Meters. These methods are the Simple Linear Regression Model (SLR) and Cubic Spline Interpolation polynomial (CSI), respectively. Then, a Hybrid House Price Prediction Model is established to predict the house price with specific Square Meters by integrating SLR and CSI. Finally, the author uses Multiple Linear Regression to model the effects of various independent variables from the target sample to the House Price. The research significance of this paper mainly includes increasing the accuracy and comprehensiveness of the House Prediction Model by constructing a Hybrid Model and using the MLR model, respectively.

Keywords: House Price Prediction, Simple Linear Regression, Cubic Spline Interpolation, Multiple Linear Regression.

1. Introduction

Housing price is an important factor affecting people's life happiness index [1]. For numerous individuals, acquiring real estate represents a significant and pivotal decision and investment in their lifetime [2]. Housing prices fluctuate constantly and are occasionally driven by hype rather than being grounded in proper valuation [3]. Numerous factors can affect the house price, which includes economic conditions, location, market speculation, government policies, square meters of the house, and built age, etc. The underlying relationship between these factors and house prices has been investigated for a long time by experts from different professional fields [4]. Although the variation of house price and complexity of considering numerous factors makes the analysis and prediction of house price extremely difficult, it is still important in an individual's daily life, investment, and the entire society.

House price prediction in the real estate sector is a complex challenge that has attracted considerable attention from researchers in recent years as they seek to develop an effective model for forecasting property prices [5]. Some advanced techniques like artificial intelligence and machine learning have been broadly embraced in various aspects of contemporary real estate industry research

DOI: 10.54254/2753-8818/2025.22572

[6]. It is known that the Random Forest forecasting model can be used to predict house prices [7], There are many other forecasting house price prediction models, both statistical and machine learning models have been studied and are relatively mature and the advantages and problems of each model are also discussed, and many combination models have been proposed [8].

It is known that the Simple Linear Regression (SLR) model can be used to predict future house prices by the linear equation constructed among the given sample data. A linear regression model can be developed to build a house prediction model [9]. However, for the case to predict house price with an independent variable within the sample domain, since the SLR model requires the assumption of linearity, the errors may be significant because the fitted model line does not need to pass through the given sample data points. Hence, polynomial interpolation would give a more accurate prediction. This article intends to construct a Hybrid House Prediction Model integrating the advantages of the SLR and Cubic Spline Interpolation polynomial (CSI).

Note that London's real estate market is one of the most dynamic and diverse in the world. Hence, this article chooses a dataset from Kaggle that contains 1000 house prices in London and conducts research to construct a valid house prediction model. The upcoming sections of this article will be arranged as follows: In section 2, SLR will be introduced, which includes a brief introduction of some preliminary concepts and then illustrates an analysis of how to use the SLR to fit the target sample data. Then, some advantages and drawbacks of the SLR are explained in section 2.3. The CSI is introduced in section 3, which is arranged similarly as the SLR, includes basic concept, benefits and problems. After conducting a comparison between the SLR and CSI, a Hybrid Prediction Model is developed in section 4, and some tested prediction data is presented in the same section. In section 5, the Multiple Linear Regression model is conducted. A conclusion of all the previous sections is shown in section 6.

2. Simple Linear Regression Model

2.1. Preliminary Concept

When the given sample has the data points $(x_1, y_1), \cdots, (x_n, y_n)$, where x_i denotes sample value of the independent variable while y_i denotes the observed value of the dependent variable, the simplest model that can be constructed to fit those data points is the simple linear regression model (SLR) given by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1,2, \cdots, n\,, \tag{1}$$

where $\beta_j, j = 1,2$ which are called the intercept and slope, respectively, are the regression parameters, ϵ_i's are random errors which cannot be observed. Note that in the SLR model, $X = [x_1, \cdots, x_n]^T$ is regarded as a constant vector; while $Y = [y_1, \cdots, y_n]^T$ a random vector.

To construct the best-fit line of the observed data is equivalent to find the estimation of β_0 and β_1. This article uses the common criterion to estimate the regression parameters, that is, minimize the Sum of the Squared Errors (SSE), where SSE stands for the sum of the square of the vertical deviations from the fitted linear line:

$$\mathrm{SSE} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \epsilon_{\mathrm{i}}^2 = \sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{y_i} - (\beta_0 + \beta_1 \mathrm{x_i})]^2 \tag{2}$$

This method to choose β_0 and β_1 is called the Method of Least Squares (LS), and the LS estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ are unbiased estimators for β_0, β_1. Figure 1 shows an example of how the SLR model works to fit a random data set generated by R. Here, $\boldsymbol{\beta} = \left[\widehat{\beta_0}, \widehat{\beta_1}\right]^T = [0.02837598, 1.97376419]^T$ and value of SSE is 92.34413.

DOI: 10.54254/2753-8818/2025.22572

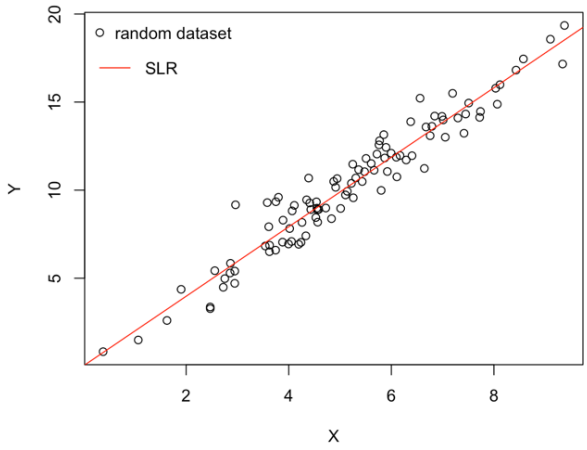

Figure 1: Simple Linear Regression Model fitting a random dataset.

2.2. Fit Sample Data by SLR

The article uses a sample data set from Kaggle to evaluate the linear relationship between the independent variable, Square Meters of the house, and the dependent variable, House Price. The data set consists of 1000 entries, with its scatter plot of Price against Square Meters shown in Figure 2. A hypothesis test is conducted to prove the existence of the linear relationship between Square Meters and House Price. Test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ at 5% level of significance. The test statistic is approximately 40.795, while the t-value is about 1.96. Therefore, the null hypothesis H_0 should be rejected since the test statistic is greater than the t-value, implying that using SLR to model their relationship is appropriate. Hence, the SLR model can be constructed to fit those 1000 data points from the sample, as shown in Figure 2, where the fitting model has the linear equation:

$$Y = 48933.05 + 11975.61\,X \tag{3}$$

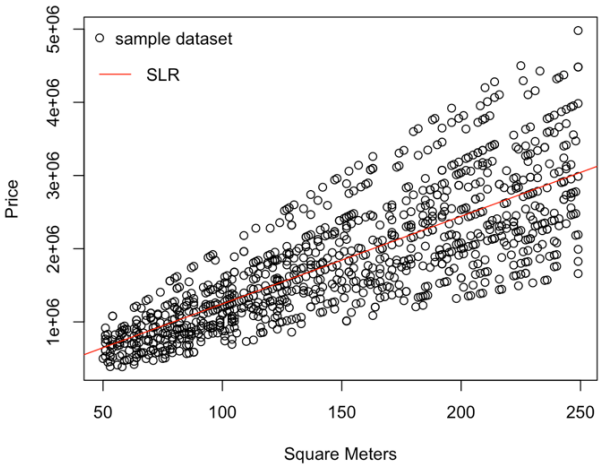

Figure 2: Scatter plot of Price against Square Meters of the sample dataset and the corresponding SLR model to fit the data.

2.3. Discussion of Advantages vs. Drawbacks of SLR

Some of the advantages and problems of the SLR model are discussed in this section.

The benefits of the SLR Model include its easiness of interpretation and implementation. Moreover, it can be used to predict future responses beyond those already observed in the sample data; for

DOI: 10.54254/2753-8818/2025.22572

instance, predicting the house price for properties larger than 250 square meters. This prediction encompasses the following two cases: the mean response of house prices whose areas are specified, as all equal to a specific value, say $x_0(x_0 > 250)$ well as the individual response y for properties of a particular size. It is important to note that the predicted individual responses have exactly the same value as the mean response, although they possess different prediction intervals. The mean response has the 100(1-α)onse$'$ s prediction interval is:

$$\widehat{\beta_0} + \widehat{\beta_1}x_0 \pm t_{\alpha/2,n-2}\hat{\sigma}\left[\frac{1}{n} + \frac{\left(x_0 - \overline{(x)}\right)^2}{\sum_i\left(x_i - \overline{(x)}\right)^2}\right]^{\frac{1}{2}} \tag{4}$$

where $\hat{\sigma} = \sqrt{SSE/(n-2)}$, i.e., the square root of the mean square error. The individual response has the 100(1-α)% prediction interval:

$$\widehat{\beta_0} + \widehat{\beta_1}x_0 \pm t_{\alpha/2,n-2}\hat{\sigma}\left[\frac{1}{n} + \frac{\left(x_0 - \overline{(x)}\right)^2}{\sum_i\left(x_i - \overline{(x)}\right)^2} + 1\right]^{\frac{1}{2}} \tag{5}$$

For example, the 95% prediction interval for the individual response at the Square Meters equals 300: $[2580499.33, 4702731.28]$, while that for the mean response: $[3548766.71, 3734463.90]$. Hence, the uncertainty associated with predicting individual response is greater compared to predicting the mean response, which can be explained as the following: predicting the future individual response means estimating the specific outcome of the dependent variable for a particular value of the independent variable outside the domain of the sample dataset, whereas mean response is predicted by estimating the average or expected value of all individual response for a given value of the independent variable, which leads to less uncertainty. Figure 3 shows the graphical illustration of ten predicted mean/individual responses corresponding to ten specific Square Meters (x_i).

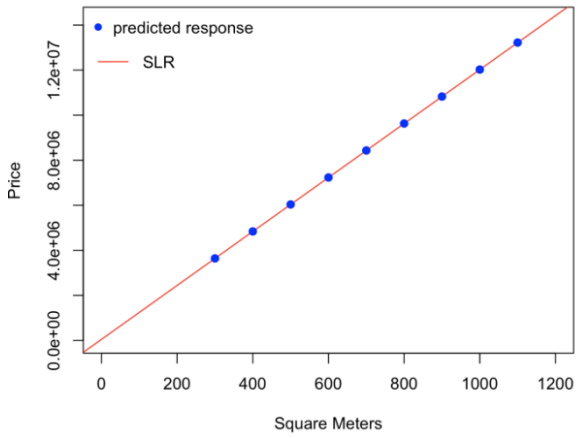

Figure 3: 10 example predicted mean/individual response of Price at 10 Square Meters shown on the SLR Model.

A disadvantage of the SLR Model is its assumption of linearity. Although the LS estimators of the regression coefficients $\widehat{\beta_0}, \widehat{\beta_1}$ are chosen so that the SSE is minimized, due to the fluctuation of the sample data points, the SSE is about 2.9×10^{14}, which is significant. The large SSE may lead to difficulty when predicting the square meters of the house within the domain of the sample data. If the same square meters coincide with one of the data points from the sample, the predicted individual house price on the constructed straight line would have a significant difference from the actual

DOI: 10.54254/2753-8818/2025.22572

observed house price. To address this issue and in pursuit of a better prediction within the Square Meters domain of the sample, the integration with the cubic spline interpolating polynomial would lead to substantial improvement.

3. Cubic Spline Interpolation

3.1. Fit Sample Data by CSI

A cubic spline interpolating polynomial (CPI) is defined piece-wisely as follows:

$$p_k(t) = a_k(t - t_{k-1})^3 + b_k(t - t_{k-1})^2 + c_k(t - t_{k-1}) + d_k \quad (6)$$

for $t \in [t_{k-1}, t_k], k = 1,2,\cdots,N$. One can fit data points $(t_0, y_0), (t_1, y_1), \cdots, (t_N, y_N)$: $p(t_i) = y_i, i = 0,1,\cdots,N$, and $p(t), p'(t), p''(t)$ are continuous. Figure 4 illustrates an example of fitting ten data points using the cubic spline interpolating polynomial.

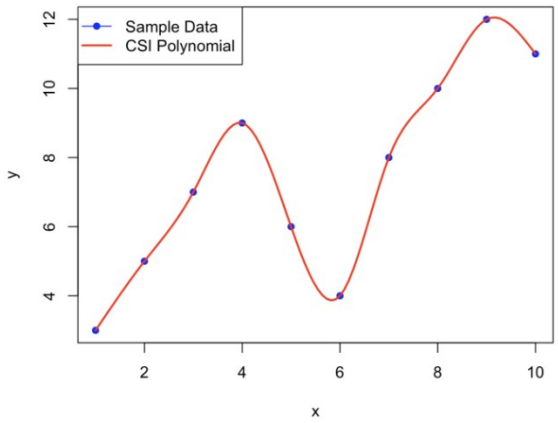

Figure 4: CSI of a sample with 10 data points.

Different from the CSI in Figure 4, the target sample set has a tremendous number of data points, so it will be computationally intensive and cause long processing times with large computation costs. Moreover, the noise of the sample data set can be seen in Figure 2. Hence, using CSI directly like in Figure 4 can cause overfitting of the model to the noise in the data due to large variance, leading to a less generalizable model that may not accurately capture the underlying trend.

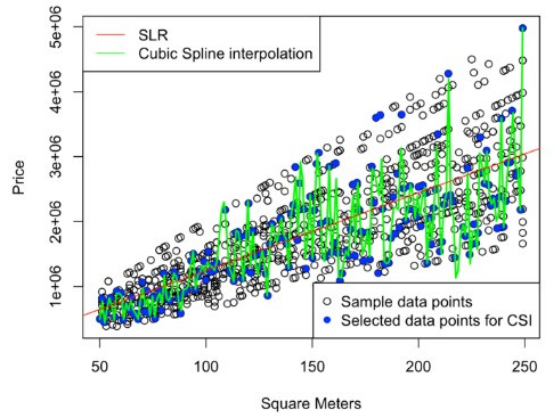

Figure 5: Comparison between SLR and CSI.

Therefore, it is necessary to select a subset of the sample nodes which are meaningful, far from noise, and then conduct CSI only on the selected data points. After selecting meaningful data points

DOI: 10.54254/2753-8818/2025.22572

from the large sample set, the CSI can be used to fit those selected points. Once the CSI is constructed, a comparison between SLR and CSI can then be conducted, which is shown in Figure 5.

3.2. Discussion of Advantages vs. Drawbacks of CSI

The CSI polynomial also has benefits and problems.

Clearly, by using CSI, the interpolating polynomials, which are smooth functions, passes through a subset of data points from the original sample, which makes it possible to have no difference between the predicted house price of a given square meters from 50 to 250 and the actual observed house price from the sample data. Additionally, by analyzing Figure 5, if the given x value, that is, the square meters, is within the domain from 50 to 250, the predicted house price on the cubic spline interpolating polynomial would be close to the observed data points, which makes the prediction within the sample domain mode precise.

The LI is when given the distinct points: $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, and the elementary Lagrange polynomial of degree n, $l_i(x)$, such that, $l_i(x) = \begin{cases} 1 \; if \; i = j \\ 0 \; if \; i \neq j \end{cases}$. Hence, the Lagrange interpolation polynomial which has degree ≤ n is given by:

$$p_n(x) = \sum_{i=0}^{n} y_i l_i(x) = y_0 l_0(x) + y_1 l_1(x) + \cdots + y_n l_n(x) \tag{7}$$

Note that, $p(x_i) = \sum_{i=0}^{n} y_i l_i(x) = y_0 l_0(x_i) + y_1 l_1(x_i) + \cdots + y_n l_n(x_i) = y_i \;\; \forall i \in \{0,1, \dots, n\}$. To fulfill the requirements of the construction $l_i(x)$, eventually,

$$l_i(x) = \frac{(x - x_0) \cdots \widehat{(x - x_i)} \cdots (x - x_n)}{(x_i - x_0) \cdots \widehat{(x_i - x_i)} \cdots (x_i - x_n)} \tag{8}$$

The LI can only be used to interpolate small and well-behaved datasets and is not suitable for the relatively huge house price versus Square Meters sample dataset since noise and the tremendous number of data points that can be seen from Figure 2 causes the impossible implementation of finding a smooth polynomial that can passes through all those data points simultaneously.

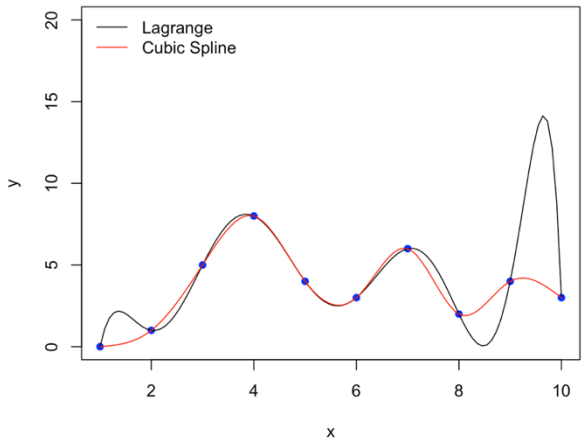

Figure 6: Comparison of CSI and LI for a sample with sample size equals 10.

In comparison, the CSI is piecewise continuous polynomial with each $p_k(t)$ and has degree 3, a much lower degree than the LI when the sample size is huge. Hence, the CSI is implementable. Moreover, even though a sample with a small sample size is given, the piecewise continuity of the

DOI: 10.54254/2753-8818/2025.22572

CSI leads to less variation of the interpolation polynomial from the fitted data points. Figure 6 shows the comparison between the CSI and LI to fit a sample set with a sample size equals 10, with the independent variable $x = [1,2,3,4,5,6,7,8,9,10]^T$, and the dependent variable $y = [0,1,5,8,4,3,6,2,4,3]^T$. There are big jumps on the intervals $[0,2], [8,10]$ of the LI, whereas the pointwise continuity of the CSI makes its interpolation polynomial close to the fitted sample data points over the entire domain of x, i.e., [0,10]. The huge variation of the LI over specific intervals may lead to inappropriate prediction over those intervals when predicting individual y values, given its x values lie between two adjacent sample nodes. Therefore, the implementation difficulties and significant errors of the prediction caused by huge variation from the fitted sample nodes of the LI make this article use CSI as a representative of the polynomial interpolation.

The SLR model is considered an approximation, while CSI is a polynomial interpolation method. The key distinction between approximation and interpolation lies in the error values: approximation involves non-zero errors at each data point, whereas interpolation requires the polynomial to pass through all selected data points, resulting in zero error at each node x_i. Consequently, the SLR model provides a global prediction validity, whereas the CSI's appropriateness is primarily guaranteed locally.

This distinction is rooted in the fact that the interpolation polynomial must pass through the sample data points, ensuring accurate predictions around these nodes. However, when considering future trends of the dependent variable, predictions made using the constructed CSI may incur significant errors. This is because there are no sample data points available in the future for the CSI to adapt to, leading to potential inaccuracies in predicting future trends.

4. House Price Prediction Hybrid Model Formulation

Drawing from the analysis of the strengths and weaknesses of SLR and CSI, this section presents a holistic prediction model that combines the benefits of both SLR and CSI while mitigating their individual drawbacks. The aim is to leverage the advantages of SLR and CSI effectively, resulting in more accurate predictions and minimizing the potential for significant prediction errors arising from their limitations.

SLR can give an initial individual prediction of the House Price due to its validity among the entire domain of the Square Meters, that is, $[0, \infty)$. For any prediction with Square Meters within the sample nodes, the CSI can give a better predicted House Price than the SLR because the linearity assumption of the SLR model leads to greater error. For example, if the input Square Meters that requires the prediction of the House Price coincides with one of the nodes in the sample, then the CSI would give zero error while the SLR model would give a predicted House Price on the fitted line, which is most likely different from the exact observed House Price since it if not necessary for the SLR model to passes though the observed sample data points. For any input Square Meters that lie between any two adjacent sample nodes, the prediction based on the CSI should also have less error than the SLR. On the other hand, for the future prediction, that is, Square Meters greater than 250, not inside the sample domain, the SLR model can give a better prediction with less error due to the local appropriation of the CSI.

Therefore, the HPM would like to achieve the following goals: let the user to input the Square Meters that they want to get a predicted House Price, if the input value coincides with any observed sample nodes or lies between any two adjacent observed sample nodes, then the HPM would output the predicted House Price using the CSI. If the input value is greater than 250 square meters, then the SLR should be used to give the predicted House Price and print (exists error).

The decision tree used for constructing the HPM should condition on the value of the Square Meters (SM) that want to predict. If SM is greater than or equal to 250, then HPM uses the SLR model

DOI: 10.54254/2753-8818/2025.22572

for prediction, whereas if SM is less than 250, then HPM uses the CSI polynomial to predict the house price.

The article chooses ten Square Meters to demonstrate how the HPM works, where five Square Meters lie between 50 to 250 so that the predicted individual House Price should be made by CSI while the remaining five Square Meters are greater than 250 so that the predicted individual House Price should be made by SLR. The graphical demonstration of the ten selected data points is shown in Figure 7, where the black points represent Square Meters lie between some adjacent observed sample data points so that it should lie on the green CSI polynomial.

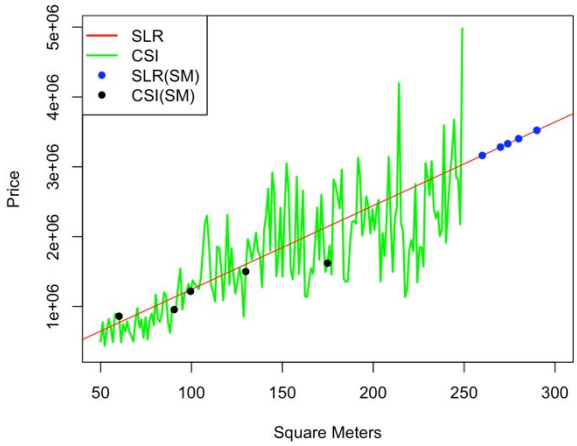

Figure 7: Extract the Model of SLR and CSI, representation of five predicted data points using SLR and five predicted data points using CSI on the model plot.

Whereas the blue data points have Square Meters greater than 250, which exceed the sample domain so that it should lie on the SLR line. The numerical value of the predicted house price of the previous ten square meters is illustrated in Table 1.

Table 1: Numerical House Price prediction of the ten Square Meters, where the black ones are predicted by CSI while the blue ones are predicted by SLR.

Square Meters	Predicted House Price	Square Meters	Predicted House Price
60.12	860998.8	260.00	3162591.0
90.47	955719.7	270.00	3282347.1
99.47	1217167.6	274.00	3330249.5
129.82	1500045.3	280.00	3402103.2
174.80	1619837.1	290.00	3521859.2

5. Multiple Linear Regression Model

House Price is influenced by various factors and characteristics of the particular property. Based on prior research, certain researchers have suggested several variables that have a notable impact on the total housing cost. As outlined by Kusan et al. [1], these variables can be categorized into three groups: housing-related factors, environmental factors, and transportation factors. Therefore, in this section, a Multiple Linear Regression Model (MLR) is introduced to predict the House Price according to various factors other than the Square Meters of the house.

Given a sample that contains n observations $(x_{i1}, x_{i2}, \cdots, x_{im}, y_i), i = 1,2, \cdots, n$, where m is the number of independent variables, x_{ij} the observed values of one of the independent variables, and y_i

DOI: 10.54254/2753-8818/2025.22572

denotes the observed values of the dependent variable corresponding to x_{ij}. Then, the MLR model has the matrix form [10]:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{9}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ & & \vdots & \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{10}$$

Similar as the SLR case, $\beta_0, \beta_1, \cdots, \beta_m$ denote the $(m+1)$ regression coefficients, which are unknown.

The data points from the sample can be fitted by the MLR model. The Sample that are used for the previous SLR and CSI also has other independent variables that would affect the House Price, which include the number of bedrooms, number of bathrooms, square meters, building ages, and floors. Set those five independent variables as $x_i, i = 1,2,3,4,5$.

This section tests the existence of a regression relation between the House Price and the set of independent variables in the sample. Test at 5% level of significance: $H_0: \beta_i = 0 \; \forall \; i = 0,1,\cdots,5$ $versus$ $H_1: \exists \beta_j \neq 0$ $where$ $j \in \{0,1,\cdots,5\}$. The test statistic has distribution $F = \frac{SSR/m}{SSE/(n-m-1)} = \frac{MSR}{MSE}$ where SSR is the regression sum of squares and SSE the sum of squared errors. By testing the target sample, $F = 332.4439 > 2.223107 = F_{0.95,m,n-m-1}$. Hence, the null hypothesis H_0 should be rejected, which implies that the set of independent variables in this sample $\{X_i: i \in \{1,2,\cdots,5\}\}$ collectively is effective in explaining the dependent variable House Price variation. Therefore, the MLR model can then be constructed.

By the same settings and let m equals five while n equals 1000 since the sample contains 1000 entries, the regression coefficient vector of the MLR model can be calculated:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 29854.54288 \\ 30.91769 \\ -6567.52007 \\ 11970.14565 \\ 741.65039 \\ -2058.87587 \end{pmatrix} \tag{11}$$

The MLR Model can then be:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\ \widehat{\boldsymbol{\beta}} \tag{12}$$

where the matrix X is defined the same as before.

Hence, the MLR model can help to predict the House price for any independent variables x_i from the sample outside the observed sample domain. For example, to predict the House Prices for two specific houses with first one has 7 bedrooms, 4 bathrooms, 300 square meters, building age equals 2, and at floor 8, while the second one has 10 bedrooms, 5 bathrooms, 330 square meters, building age equals 1, and at floor 10. The MLR model predicts as follows by using the Eq. (10):

$$\widehat{\boldsymbol{Y}} = X\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1\ 7\ \ 4\ 300\ 2\ 8 \\ 1\ 10\ 5\ 330\ 1\ 10 \end{pmatrix}, \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 3579856.875 \\ 3927627.076 \end{pmatrix}. \tag{13}$$

DOI: 10.54254/2753-8818/2025.22572

6. Conclusion

This work focuses on the issue of house-purchase. The real estate market is a complex and dynamic environment affected by numerous factors such as location, square meters of the house, amenities, and economic conditions. Taking several factors that can affect the house price into consideration can make the prediction more appropriate and precise. The author constructs a Hybrid House Price Prediction Model (HPM) based on the integration of SLR and CSI to make the prediction of House Price implementable for any particular Square Meters of the House as well as minimizing the potential errors. In order to let the prediction more appropriate, the author also conducts the MLR model to consider various factors affecting the House Price. Admittedly, HPM may not work well as a consequence of the limitations of the single independent variable. Although MLR model can alleviate this problem, the necessity of assuming the linearity in MLR model may cause significant errors in reality. Therefore, further research can be conducted to evaluate how to integrate the ideas of the CSI polynomial with the MLR model to construct a more comprehensive and accurate house price prediction model.

References

[1] *Zhang, L.L., Ma, X. M., Wang, X. Y. & Sun, J.J. (2024). Application of housing price forecasting based on PCA-BPNN algorithm. Journal of Changchun Institute of Technology (Natural Science Edition), 25 (02), 114-118.*

[2] *Ng, A., & Deisenroth, M. (2015). Machine learning for a London housing price prediction mobile application. Imperial College London.*

[3] *Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018, April). House price prediction using machine learning and neural networks. In 2018 second international conference on inventive communication and computational technologies (ICICCT) (pp. 1936-1939). IEEE.*

[4] *H.Kusan, O.Aytekin, and I.Ozdemir, (2010). "The use of fuzzy logic in predicting house selling price," Expert Systems with Applications, 37(3), 1808-1813.*

[5] *Mohd, T., Jamil, N. S., Johari, N., Abdullah, L., & Masrom, S. (2020). An overview of real estate modelling techniques for house price prediction. In Charting a Sustainable Future of ASEAN in Business and Social Sciences: Proceedings of the 3rd International Conference on the Future of ASEAN (ICoFA) 2019—Volume 1 (pp. 321-338). Springer Singapore.*

[6] *Zhang, Q. (2021). Housing price prediction based on multiple linear regression. Scientific Programming, 2021(1), 7678931.*

[7] *Qin Y J. (2024). A comparative study of housing price forecasting models based on multiple linear regression and Random forest algorithm. Modern Information Technology, 8 (22), 127-131.*

[8] *Huo, Y J. (2024). Research on the Prediction of Second-hand Housing Price Based on Single and Combination Model (Master's Thesis, China University of Geosciences for Master of Professional Degree).*

[9] *Sagala, N. T., & Cendriawan, L. H. (2022, July). House Price Prediction Using Linear Regression. In 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED) (pp. 1-5). IEEE.*

[10] *Soffritti, G., & Galimberti, G. (2010). Multivariate linear regression with non-normal errors: A solution based on mixture models. Statistics and Computing, 21(4), 523–536.*