Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.36726

Housing Price Modelling with Multiple Linear Regression and Regularisation

Xiaoyan Liu

Reading Academy, Nanjing University of Information Science and Technology, Nanjing, China
202383100049@nuist.edu

Abstract. Housing price models in a feature-price system can provide strong support for the housing-purchase decisions of households, real estate appraisal, bank credit risk control and urban planning. To determine the added value of various factors in the construction characteristics, location features and neighbourhood environments on house prices, several regressions can be employed. Real-world housing datasets generally have the following statistical problems: multicollinearity, outliers and model instability. This paper systematically reviews the theoretical basis and empirical methods for multiple linear regression and regularisation (Ridge regression, Lasso and Elastic Net) in housing price modelling and introduces fundamental concepts such as the response variable, explanatory variables and design matrix, as well as data standardisation and pre-processing techniques. Cutting-edge empirical research based on public datasets, such as the Ames Housing, California Housing, and King County House Sales datasets, is reviewed in this paper, together with model diagnostics, evaluation metrics, and the strengths and weaknesses of least squares and regularised regression. The traditional linear housing price model does not consider spatial differences or nonlinear relationships; therefore, in the future, researchers will integrate spatial regression, robust modelling and multiple sources of data to improve both the accuracy and interpretability of this model. Collectively, this paper shows how linear algebra, statistical inference and data science can support interpretable regression modelling for socioeconomic research.

Keywords: housing price modelling, hedonic price model, multiple linear regression, regularisation

1. Introduction

Housing price analysis offers a reference for social and economic policies, helps households decide where to live, buy or sell houses, take out mortgages, etc. The reasons for changes in house prices are varied, including floor area, number of rooms, age of the house, location in terms of transportation, schools, economy of the area, etc. Differences in housing prices for urban administrators are due to varying levels of capitalised value in public facilities, such as transportation and schools, and provide empirical support for public resource allocation and housing policies. These factors can be classified into structural, location-based and neighbourhood attributes. Under the traditional hedonic pricing system, a house price can be expressed as a function of several

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.36726

characteristics of the property, and regression coefficients indicate how much these attributes change the value of the house [1]. Data on public housing transactions have been collected in several studies for various multiple linear regression analyses of housing prices. Real-world housing data also have some well-known statistical problems, such as feature selection, multicollinearity, measurement error and model instability; these are still significant applied issues in statistics and data science [2, 3].

A large body of literature applies hedonic pricing theory in conjunction with multiple linear regression to predict housing prices using benchmark public datasets, such as Ames Housing, California Housing and King County House Sales. These studies build prediction models based on predictors such as floor area, house age, number of bedrooms, location, neighbourhood median income and community facilities. Ordinary Least Squares (OLS) can be sensitive to severe multicollinearity and outliers. To address multicollinearity and coefficient instability, regularisation methods such as Ridge, Lasso and Elastic-Net are frequently employed; Ridge regression stabilises parameter estimation through L_2-penalisation, Lasso performs embedded feature selection using an L_1-penalty, and Elastic-Net combines L_1 and L_2 penalties to handle groups of highly correlated predictors [4, 5]. In addition to penalised linear regression, several other nonlinear machine-learning algorithms have been proposed to address complex location-effect models for prediction [6]. Recently, some research has compared regularised linear estimators and non-linear approaches on public housing data and discussed the circumstances under which each method is more suitable [7]. Although many empirical results have been reported, fewer studies have systematically explained how the linear-algebraic construction of design matrices is applied in statistical inference for hedonic housing regression.

This paper reviews the application of multiple linear regression and regularisation to house price prediction. This paper shows the organisation of the raw housing features in design matrices for estimating feature effects and evaluating prediction errors. The following sections will present the basic ideas of housing data and linear regression, demonstrate a typical model construction and diagnostic process, compare some representative empirical studies, and discuss methodological deficiencies and future directions. Through this analysis, this paper demonstrates how linear algebra, statistical inference and data-science techniques jointly support interpretable housing-price analysis.

2. Basic concepts of housing data and multiple linear regression

In housing price modelling under the hedonic pricing method, the housing price is taken as the dependent variable and observable property attributes are used as independent variables. In this model, housing prices are the response variables to be predicted; the explanatory variables are various observable attributes of houses that are thought to affect housing prices [1]. Common predictors of house prices in research include floor area, number of rooms, location, neighbourhood income and school quality. The above buildings and their surrounding areas represent the features of a home and provide the foundation for hedonic pricing theory [8].

Observed housing-price data constitute the outcome response vector in the matrix-based framework for multiple linear regression. All attribute-related characteristics are contained in the predictor design matrix; each row corresponds to an individual house sample, and each column is one explanatory housing-related feature.

The regression model expresses house prices as a linear function of the predictors, weighted by regression coefficients, and includes a random error term. The set of regression coefficients includes the intercept term of the model. A residual is the difference between the actual observed house prices and the house prices predicted by the regression model.

DOI: 10.54254/2753-8818/2026.36726

Before regression fitting, the raw housing data need to be preprocessed; a training set and a test set are separated from the data for parameter estimation and evaluation of generalisation ability, respectively. Housing features are either continuous variables, such as floor area and neighbourhood income, or categorical variables, such as administrative district and building type. The first few processing steps generally include feature scaling, handling missing values, outlier detection and log transformation of skewed price attributes. Logarithmic transformation is commonly used in the study of housing prices to address right-skewness, and some coefficients can be interpreted as semi-elasticities [7]. The concepts and preprocessing methods mentioned above provided the foundation for building, diagnosing and evaluating the performance of the multiple regression model for housing prices.

3. Regression and regularisation methods for housing price modelling

Multiple linear regression for housing-price modelling uses ordinary least squares (OLS) to find the regression coefficients that minimise the sum of squared errors. Hedonic studies have been further extended in practice by adding dummy variables for categorical attributes, interaction terms to consider the joint effect of predictors, and polynomial terms to model partial non-linear relationships. OLS estimators are unstable when there is severe multicollinearity among the housing features. Variance Inflation Factor (VIF), residual analysis and cross-validation are typical diagnostic tools; VIF measures the degree of multicollinearity, residual analysis tests model misspecification, and cross-validation assesses out-of-sample generalisation performance.

Regularisation methods, such as Ridge, Lasso and Elastic-net, add a penalty term to the loss function to prevent coefficient instability and thus reduce model complexity. Ridge Regression adds a penalty to the size of the coefficients to keep all predictors. Lasso uses an L1 norm penalty and thus performs automatic feature selection by setting some coefficients to zero. Elastic Net is a combination of the L1 and L2 norms that performs well with datasets containing groups of highly correlated housing attributes [4, 5].

For model evaluation, in-sample goodness-of-fit measures and out-of-sample prediction metrics are used. The coefficient of determination and adjusted coefficient of determination indicate the proportion of price variation explained by the predictor(s). Prediction error is shown as the mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE). Explanatory power should not be confused with predictive ability, and a high coefficient of determination does not ensure good generalisation of the model to new housing data. A good housing-price model should have both an easily understandable coefficient output and high predictive accuracy.

4. State-of-the-art studies of regression-based housing price modelling

Recent research on housing price modelling has used the same public benchmark datasets, such as Ames Housing, California Housing and King County House Sales, but they differ significantly in terms of data sources, sample sizes, feature engineering and preprocessing. Researchers address locational heterogeneity in many urban case studies by introducing neighbourhood dummy variables or geospatial attributes; standard preprocessing steps, such as missing value imputation, outlier filtering and logarithmic transformation of housing prices, are performed before building the feature matrix. The split of the training and test sets, methods for variable transformation, and selection of evaluation metrics all affect the final reported model results; thus, when experimental designs are not the same, the raw prediction errors from different housing research reports cannot be directly

DOI: 10.54254/2753-8818/2026.36726

compared across studies. Multicollinearity is a typical problem in practice: high correlation among predictor variables related to building characteristics, location and neighbourhood can increase the variance of ordinary least squares (OLS) estimates obtained by matrix-based OLS. To address the instability caused by matrix properties, recently, researchers have systematically compared benchmark OLS with ridge regression, Lasso and elastic net regularised regression; Lasso performs embedded variable selection for high-dimensional housing attributes, ridge regression reduces coefficient magnitudes without eliminating variables, and elastic net handles groups of highly correlated housing features.

Polynomial and interaction terms have also been added to the linear model by some scholars to investigate local non-linear relationships between housing features and home prices. Although RMSE, MAE and MAPE may decrease, one must still consider whether the in-sample explanatory power shown by R^2 and adjusted R^2 corresponds to good out-of-sample performance. The well-known Ames Housing dataset is used as an example to demonstrate a general empirical process: selecting building, age and neighbourhood attributes; building a design matrix from processed observations; estimating regression coefficients by minimising least squares or regularised loss; and finally, comparing prediction errors under different regression models. At the base of this housing price model application are fundamental mathematical tools: matrix construction, least-squares optimisation, multicollinearity diagnostics derived from the properties of the design matrix, regularisation terms, and quantitative error assessment. Therefore, research on house prices can be considered an application in the field of real estate and a typical case for applying linear algebra and statistical inference, etc., to build an interpretable socioeconomic model.

5. Limitations and prospects

Linear Regression and Regularisation methods have achieved good interpretability and stable performance in housing price modelling; however, the traditional hedonic linear model still has some shortcomings for practical data analysis. First, standard multiple linear regression assumes a fixed linear relationship between housing features and property prices and fails to capture complex nonlinear patterns in the real urban housing market. Changes in house prices are often due to the indirect effects of urban development policies, traffic changes and shifts in population distribution, and pure structural or neighbourhood characteristics cannot be explained by linear models. Second, regularisation models such as Ridge, Lasso and Elastic Net can address multicollinearity and overfitting; however, they are still fundamentally linear estimators and lack the ability to model spatial heterogeneity and regional price differences [2, 9]. Most public datasets, such as Ames Housing and California Housing, are static cross-sectional records that do not consider the impact of time-varying factors on price fluctuations.

Traditional housing regression models also require manual feature engineering, such as variable transformation, outlier removal and missing value imputation. Subjective preprocessing choices can affect the model's robustness and reproducibility across research environments. In addition, linear regression models produce a single set of coefficients for the entire dataset and cannot describe different local effects of school quality, neighbourhood income and traffic accessibility in various urban areas.

Future work will improve the model's generality, spatial sensitivity and ability to capture dynamic effects. First, a model of spatial regression and Geographically Weighted Regression (GWR) will be employed to obtain location-varying feature coefficients and enhance the explanation of spatial differences in housing prices. Second, interpretable machine learning methods such as Shapley Additive Explanations and Local Interpretable Model-agnostic Explanations can enhance the

DOI: 10.54254/2753-8818/2026.36726

interpretability of nonlinear housing-price models [10]. Thirdly, in the future, housing price models will gradually move away from static cross-sectional regression to dynamic time-series analysis and add interest rates, inflation, urban policies, etc., to build an all-encompassing socioeconomic prediction system. At present, most research in data science for housing price data aims to balance mathematical interpretability, statistical stability and predictive accuracy.

6. Conclusion

This paper systematically reviews the applications of multiple linear regression and regularisation methods for interpretable housing price models. First, this paper introduces the general concepts of housing datasets, such as the response variable, explanatory attributes, design matrices and necessary data preprocessing steps, including standardisation, transformation and outlier detection. Second, it presents the foundations and distinctions of ordinary least squares, Ridge, Lasso and Elastic Net regression, as well as popular model diagnostic tools for multicollinearity, coefficient instability and prediction bias. Thirdly, recent state-of-the-art empirical studies based on public housing datasets are compared to identify existing methodological problems in feature engineering, nonlinear fitting and model generalisation. In the future, Spatial regression will be used to construct robust statistical models with multiple data sources for a comprehensive study of the complex price system in urban areas. In short, this paper introduces a way to combine linear algebra, statistical inference and data science techniques to build stable, interpretable and reliable continuous-variable models for socio-economic research. Based on the above studies, different models may be selected according to the characteristics of the housing data and different aims of inquiry. Ordinary least squares is still suitable if its assumptions are reasonably met and coefficient interpretation is required; otherwise, Ridge, Lasso and Elastic Net can be used to address problems such as many predictors or multicollinearity. Data preprocessing, diagnostic tests and out-of-sample evaluation need to be carried out, and a higher in-sample fit does not guarantee improved prediction performance on new data. Other problems in the standard linear model include spatial non-uniformity, time-varying effects, non-linear relationships and poor data quality. Therefore, the following work will combine spatial and robust regression methods with interpretable nonlinear models under consistent data division and transparent metrics. The above comparisons can reveal whether an increase in model complexity adds value, and when a simpler regression model is more suitable for interpretation, reproducibility and policy-oriented analysis.

References

[1] Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55. https://doi.org/10.1086/260169

[2] Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. *Land, 11*(3), Article 334. https://doi.org/10.3390/land11030334

[3] Zhan, W., Hu, Y., Zeng, W., Fang, X., Kang, X., & Li, D. (2024). Total least squares estimation in hedonic house price models. *ISPRS International Journal of Geo-Information, 13*(5), Article 159. https://doi.org/10.3390/ijgi13050159

[4] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[5] Simlai, P. E. (2021). Predicting owner-occupied housing values using machine learning: An empirical investigation of California census tracts data. *Journal of Property Research, 38*(4), 305–336. https://doi.org/10.1080/09599916.2021.1890187

DOI: 10.54254/2753-8818/2026.36726

[6] Rey-Blanco, D., Zofío, J. L., & González-Arias, J. (2024). Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. *Expert Systems with Applications, 235*, Article 121059. https://doi.org/10.1016/j.eswa.2023.121059

[7] Khoshnoud, M., Sirmans, G. S., & Zietz, E. N. (2023). The evolution of hedonic pricing models. *Journal of Real Estate Literature, 31*(1), 1–47. https://doi.org/10.1080/09277544.2023.2201020

[8] Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*(2), 132–157. https://doi.org/10.1086/259131

[9] Barde, S., Cherodian, R., & Tchuente, G. (2025). Moran's I lasso for models with spatially correlated data. *The Econometrics Journal, 28*(3), 423–441. https://doi.org/10.1093/ectj/utaf008

[10] Hussain, M. I., Munir, A., Mamun, M., Chowdhury, S. H., Uddin, N., & Hossain, M. M. (2025). A transparent house price prediction framework using ensemble learning, genetic algorithm-based tuning, and ANOVA-based feature analysis. *FinTech, 4*(3), Article 33. https://doi.org/10.3390/fintech4030033