Comparative Analysis of Linear and Polynomial Regression Models in Predicting Shanghai Housing Prices

Zhengnan Wang *

Ulink college of Shanghai, Shanghai, 201600, China

* Corresponding Author Email: wangeden643@gmail. com

Abstract. This study presents a comparative analysis of linear and polynomial regression models for predicting housing prices in Shanghai's dynamic real estate market. Utilizing a dataset of 100 second-hand housing listings sourced from Lianjia between 2020 and 2023, with living area as the primary predictive feature, we rigorously train and evaluate both models. Our findings demonstrate that a second-degree polynomial regression model (MSE = 0. 58, R^2 = 0. 89)significantly outperforms a simple linear regression model (MSE = 0. 97, R^2 = 0. 78)in predictive accuracy and explanatory power. Residual analysis further reveals that the linear model suffers from heteroscedasticity, indicating its inadequacy in capturing the market's complexity, particularly for high-value properties. The results underscore the importance of selecting appropriately complex models for economic forecasting. We conclude that polynomial regression offers a superior fit for the non- linear trends prevalent in Shanghai's housing market. This study provides practical insights for stakeholders in real estate investment, policy formulation, and urban economics, while also acknowledging limitations related to dataset size and feature selection, suggesting directions for future research involving multivariate approaches and regularization techniques.

Keywords: Housing Price Prediction, Linear Regression, Polynomial Regression, Shanghai Real Estate, Model Comparison.

1. Introduction

The Shanghai real estate market has experienced rapid growth and transformation, becoming one of the most significant and volatile markets globally. Accurate prediction of housing prices is not merely an academic exercise but a critical necessity for a multitude of stakeholders. Policymakers rely on accurate valuations to design effective property tax systems and urban development plans. Investors and financial institutions require precise forecasts to make informed investment decisions and assess mortgage risks. Traditional modeling approaches, particularly ordinary linear regression, are often favored for their simplicity and interpretability. However, the assumption of a linear relationship between housing characteristics and price is frequently violated in complex, dynamic markets like Shanghai's, where factors such as location prestige, property type, and market sentiment introduce non-linearities.

This study aims to rigorously compare the predictive performance of two fundamental regression models: linear regression and second-degree polynomial regression. The primary objective is to empirically determine which model more accurately captures the relationship between living area and housing price in the Shanghai market, especially within the higher-value segment where non-linear trends are more pronounced. The significance of this research lies in its practical contribution to the field of real estate economics. By providing a clear, evidence-based comparison, this study offers valuable guidance for model selection, helping stakeholders avoid the pitfalls of underfitting models and improve the reliability of their price forecasts.

The research employs a quantitative methodology based on real-world market data. We utilize a dataset of 100 observations from Lianjia, featuring living area(sqm)as the independent variable and price (million RMB) as the dependent variable [1]. After standard preprocessing steps including outlier removal, we fit both a linear model and a second-degree polynomial model. Model performance is evaluated using standard metrics: Mean Squared Error (MSE)and R-squared(R^2), supplemented by visual residual analysis to diagnose model adequacy. The remainder of this paper is structured as follows: Section 2 reviews relevant literature; Section 3 details the methodology; Section

4 presents the results; Section 5 discusses the implications and limitations; and Section 6 concludes with findings and future research directions. Model performance is evaluated using standard metrics: Mean Squared Error (MSE)and R-squared(R^2), supplemented by visual residual analysis to diagnose model adequacy. The remainder of this paper is structured as follows: Section 2 reviews relevant literature; Section 3 details the methodology; Section 4 presents the results; Section 5 discusses the implications and limitations; and Section 6 concludes with findings and future research directions.

2. Literature Review

2.1. Linear Regression Applications and Limitations in Housing Economics

Linear regression has been a cornerstone of hedonic pricing mode ls in real estate due to its straightforward implementation and ease of interpretation. Studies like Li and Zhang have demonstrated its utility in housing valuation across Chinese cities, highlighting its effectiveness as a baseline model [2]. Its coefficients offer direct insights into the marginal contribution of a unit change in a feature (e. g. square meters) on the house price. However, a significant body of literature also exposes its limitations. The primary criticism is its inability to capture non-linear relationships and interaction effects that are inherent in real estate markets. It often fails to adequately model the accelerating price growth for luxury properties or the diminishing returns for extremely large areas, leading to systematic prediction errors.

2.2. Nonlinear Models and the Bias-Variance Tradeoff

To address the limitations of linear models, researchers have turned to more flexible nonlinear approaches. Kumar&Singh advocated for the use of polynomial regression in urban housing markets, arguing that it can better capture the curvature in the relationship between features and price [3]. Polynomial regression extends the linear model by adding higher-order terms (e. g., x^2, x^3), allowing it to fit a wider range of trends. This introduction of model complexity, however, directly engages with the fundamental bias-variance tradeoff discussed by Ng [4]. A model that is too simple (high bias) like linear regression may underfit the data, while a model that is too complex (high variance) may overfit the training data and perform poorly on new, unseen data. Techniques like regularization, using methods like Ridge or Lasso regression, have been developed to manage this tradeoff by penalizing excessive complexity. More advanced studies, such as Wu and He have applied neural networks to Shanghai data, showing even greater performance but at the cost of interpretability [5]. Furthermore, research by Tanaka and Xie et al. emphasizes the importance of geographic segmentation and sophisticated nonlinear techniques like kernel regression in achieving accurate urban price models [6,7]. This study positions itself within this discourse by providing a foundational comparison between the simple linear model and the next step in complexity—the polynomial model—using Shanghai-specific data.

3. Methodology

3.1. Data Description

The dataset used in this analysis was sourced from the publicly available listings on Lianjia, a leading real estate platform in China. It comprises 100 observations of second-hand housing units in Shanghai, transacted between 2020 and 2023. The two key variables are:

(1) Independent Variable(x): Living area, measured in square meters(sqm).

(2) Dependent Variable(y): Transaction price, measured in million RMB.

Data preprocessing was conducted to ensure analysis quality. This involved the removal of extreme outliers, properties with anomalously high or low prices per square meter that could disproportionately influence the model fit. The remaining data was standardized to facilitate model convergence and interpretation.

3.2. Model Specifications

This study defined and compared two regression models:

Linear Regression Model: This model assumes a straight-line relationship between the area and the price. Its equation is:

$$y = \beta_0 + \beta_1 x \quad (1)$$

Where y is the predicted price, x is the living area, β_0 is the y-intercept, and β_1 is the coefficient representing the marginal price per square meter.

Polynomial Regression Model (Degree 2): This model extends the linear model by introducing a quadratic term(x^2) to capture curvature in the data. Its equation is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \quad (2)$$

Where β_2 is the coefficient for the quadratic term. A positive β_2 indicates an accelerating relationship, while a negative β_2 indicates a decelerating one.

Both models were implemented and trained using the scikit-learn library in Python, which employs the ordinary least squares (OLS)method to minimize the sum of squared differences between the observed and predicted values.

3.3. Evaluation Metrics

To objectively compare the performance of the two models, we employed the following metrics and techniques:

(1) Train-Test Split: The dataset was split into a training set (80%of the data) to train the models and a test set (the remaining 20%) to evaluate their performance on unseen data, providing a measure of generalizability.

(2) Mean Squared Error (MSE): This calculates the average of the squared differences between predicted and actual values. A lower MSE indicates a better fit.

(3) R-squared(R^2): This statistic represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1, with a higher value indicating a better fit.

(4) Residual Analysis: Plots of the residuals (the differences between actual and predicted values) were examined. A well-fitted model will have residuals randomly scattered around zero. Patterns in the residuals, such as a funnel shape, indicate model misspecification (e. g., heteroscedasticity).

4. Results

4.1. Performance Comparison

The quantitative results from the model evaluation on the test set clearly demonstrate the superiority of the polynomial regression model for this dataset.

Model Mean Squared Error (MSE)R-squared(R^2) Linear Regression0. 970. 78Polynomial Regression0. 580. 89Performance Comparison Between Linear and Polynomial Regression Models

The polynomial model achieved a significantly lower MSE (0. 58 vs. 0. 97), meaning its predictions were, on average, much closer to the actual observed prices. Furthermore, its R^2 value of 0. 89 indicates that it explains 89%of the variance in housing prices based on area, a substantial improvement over the 78%explained by the linear model.

4.2. Residual Analysis

Visual inspection of the residual plots provides further evidence for the better fit of the polynomial model.

The plot of residuals versus predicted values revealed a distinct funnel pattern for the linear model, indicating the presence of heteroscedasticity, where the spread of the residuals increased significantly

as the predicted price increased. This systematic error demonstrates that the linear model's predictions become less reliable and more volatile for higher-priced, larger properties. In contrast, the residuals for the polynomial model were scattered randomly around zero with a constant variance, exhibiting homoscedasticity. This suggests the polynomial model does not suffer from the same systematic bias across different price levels and is consequently a more appropriate and robust specification for the data.

5. Discussion

5.1. Interpretation of Results

The results strongly support the hypothesis that the relationship between living area and price in the Shanghai housing market is not linear but contains significant curvature. The superior performance of the second-degree polynomial regression model, both in terms of lower error and higher explanatory power, confirms that the marginal price per additional square meter is not constant. For higher-end properties, the price appears to increase at an accelerating rate. Th could be driven by factors such as luxury amenities, superior building quality, prestigious locations, and scarcity value, which are often associated with larger units and are not captured by a simple linear relationship. The linear model, while intuitive, oversimplifies this dynamic, leading to significant under-prediction for high-value homes and contributing to the observed heteroscedasticity.

5.2. Policy and Practical Implications

The choice of model has direct real-world consequences:

For Government and Policy, accurate valuation is crucial for fair property tax assessment. Using an underfitting linear model could lead to systematic undervaluation of luxury properties, resulting in lost tax revenue and inequitable tax burdens across different market segments. Urban planners also rely on accurate market models to understand housing trends.

For Real Estate and Investment, Developers and real estate agents can use more accurate nonlinear models to refine pricing strategies for new launches and existing inventory. Investors can make better-informed decisions about asset valuation and potential returns.

For Banking and Finance, Financial institutions base mortgage lending and risk assessments on property valuations. A more accurate model reduces credit risk by providing a more reliable estimate of a property's market value, which is critical for loan-to-value calculations.

6. Conclusion

6.1. Summary of Findings

This study conducted a rigorous comparison of linear and polynomial regression models for predicting housing prices in Shanghai. Based on empirical evidence from recent market data, we conclusively find that a second-degree polynomial regression model provides a significantly better fit than a simple linear model. It achieves a lower prediction error (MSE of 0. 58 vs. 0. 97)and explains a greater proportion of the price variance (R^2 of 0. 89 vs. 0. 78). The analysis of residuals confirms that the polynomial model adequately captures the underlying non-linear trend, whereas the linear model exhibits systematic flaws.

6.2. Limitations Analysis

While the findings are compelling, this study has several limitations, primarily arising from the simplicity of the model and the constrained scale of the dataset. The model uses only living area as a predictor, omitting other key determinants of housing prices such as location, age, number of bedrooms, school district quality, and proximity to amenities. Furthermore, the dataset of 100 observations is relatively small and covers only a specific time frame, limiting the robustness and

generalizability of the findings to the entire Shanghai housing market. While the polynomial regression outperformed the linear model, exploring more complex techniques like regularized regression or tree-based algorithms could yield further improvements and help prevent potential overfitting. These factors collectively constrain the model's overall predictive power and its broader applicability.

6.3. Future Research Directions

Future work should focus on addressing these l imitations through several key avenues. First, developing multivariate models that incorporate other critical features—such as location, property age, and number of rooms—would more accurately reflect the multifaceted nature of housing prices. Second, sourcing larger and more granular datasets would substantially improve the model' s robustness and generalizability. Third, exploring more sophisticated machine learning techniques—including regularized regression methods like Lasso and Ridge, support vector machines with nonlinear kernels, and ensemble method s such as Random Forests and XGBoost—would help capture complex, nonlinear relationships and allow for performance comparisons against the baseline models established in this study. Finally, investigating temporal aspects to understand how these predictive relationships evolve over time could further enhance the model' s relevance and accuracy.

By building upon this foundational comparison, future research can develop even more powerful and accurate tools for understanding and predicting real estate dynamics in complex markets like Shanghai.

References

[1] Lianjia. Shanghai Housing Dataset. 2025, sh.lianjia.com/ershoufang/.

[2] Li, Jian, and Yiming Zhang. "Linear Models for Housing Valuation in Chinese Cities." Journal of Real Estate Research, vol. 40, no. 3, 2018, pp. 345–362.

[3] Kumar, Sanjay, and Rajesh Singh. "Polynomial Regression in Housing Price Forecasting." International Journal of Data Science, vol. 15, no. 2, 2020, pp. 112–125.

[4] Ng, Andrew. Bias-Variance Tradeoff Explained. Stanford Machine Learning Lecture Notes, 2012.

[5] Tanaka, Kenji. "Geographic Segmentation in Urban Price Modeling." Urban Economics Review, vol. 55, no. 1, 2021, pp. 88–104.

[6] Wu, Lei, and Xuan He. "Deep Learning Applications in Chinese Real Estate." Applied AI, vol. 7, no. 4, 2023, pp. 201–220.

[7] Xie, Dong, et al. "Kernel Regression in Real Estate." China Economic Journal, vol. 13, no. 2, 2020, pp. 234–250.