MATH 564 PROJECT

Regression Analysis of California Housing Prices

Authors:

Meghana Thimmaiah

Submission Date: November 15, 2024

Abstract:

This project investigates the factors influencing housing prices in California by analyzing the California Housing dataset using multiple linear regression techniques. Key predictor variables, including median income, housing median age, total rooms, and proximity to the ocean, were identified to assess their impact on median house value. After data cleaning and exploration, an initial regression model was fitted, which exhibited issues with autocorrelation. To address this, a lagged median house value variable was incorporated, significantly improving model reliability by reducing autocorrelation as verified by the Durbin-Watson test.

Comprehensive diagnostic checks identified additional model issues, such as heteroscedasticity and influential points. To mitigate these, a log transformation on the response variable was applied, further improving model assumptions. Variance Inflation Factor (VIF) analysis confirmed low multicollinearity among predictors, validating the model's robustness. Influential points were managed by Cook's Distance, resulting in a refined model with enhanced predictive accuracy.

The final model reveals that higher income levels and proximity to the ocean positively correlate with housing prices, while inland location negatively impacts value. These insights, along with practical recommendations for real estate investors and policy developers, demonstrate the model's utility in predicting housing prices and understanding market trends.

Introduction:

In recent years, housing markets have become an area of keen interest for stakeholders ranging from investors and developers to policymakers and individual homebuyers. California, with its diverse geography and booming economy, has seen rapid fluctuations in housing prices, making it an ideal candidate for a detailed regression analysis. This project aims to understand the key factors influencing median housing prices in California by analyzing various socioeconomic and geographical variables. The dataset used for this analysis, sourced from Kaggle, contains housing data from California, including features such as median income, housing age, total rooms, population, and proximity to the ocean.

Understanding the impact of these variables on housing prices is critical for making informed decisions in real estate investments, urban planning, and policy formation. However, the analysis presents several challenges, including addressing multicollinearity among predictors, handling potential autocorrelation in the data, and ensuring the regression model meets the necessary assumptions. These issues require careful examination through diagnostic tests and, when necessary, remedial steps such as variable transformations.

In this report, we explore these challenges and detail the strategies employed to address them, including using diagnostic tests like the Durbin-Watson test for autocorrelation and Variance Inflation Factor (VIF) for multicollinearity. Additionally, transformations were applied to some variables to stabilize variance and improve model interpretability.

The structure of this report proceeds as follows: first, we describe the data sources and methodology; next, we present the key findings and diagnostics; finally, we conclude with practical implications and recommendations for stakeholders in the California housing market.

Problem Statement and Data Sources

Dataset Description:

- **Name of the dataset: California Housing Prices (source: Kaggle).**
- **Link to the dataset: Kaggle California Housing Prices.**
- **Key Variables:**
    - **Response variable: median_house_value.**
    - **Predictor variables: median_income, housing_median_age, total_rooms, ocean_proximity (categorical).**
- **Data Type: Mixed data types (continuous and categorical variables)**

The primary objective of this analysis is to investigate the factors influencing housing prices in California and to build a regression model that accurately predicts median house values based on key demographic, economic, and geographic features. By identifying and understanding the drivers of housing prices, this study aims to provide actionable insights for real estate investors, policymakers, and urban planners.

Data Sources

The data used for this analysis is sourced from the California Housing Prices dataset on Kaggle. This dataset includes information collected across various neighborhoods in California, containing features such as:

- **Median Income**: The median income level of residents within a housing block.
- **Housing Median Age**: The median age of houses in the block.
- **Total Rooms**: The total number of rooms across all houses in the block.
- **Total Bedrooms**: The total number of bedrooms across all houses in the block.
- **Population**: The population count within the block.
- **Households**: The number of households in the block.
- **Median House Value**: The median value of houses, serving as the target variable for prediction.
- **Ocean Proximity**: A categorical variable indicating the block's proximity to the ocean (e.g., "NEAR BAY," "INLAND").

The dataset provides a comprehensive view of various socioeconomic and location-based factors that are hypothesized to influence housing prices. Given the range and diversity of data points across California, this dataset presents an ideal basis for developing a regression model that captures the nuances of the state's housing market. However, certain challenges, such as multicollinearity among predictors, potential outliers, and possible autocorrelation, must be carefully addressed to ensure model reliability and validity.

Proposed Methodology:

This section outlines the steps taken to build and refine the regression model for analyzing housing prices. The methodology includes data preprocessing, initial model fitting, diagnostic checks, and remediation techniques to ensure that the model adheres to the assumptions of linear regression.

1. Model Specification

- The response variable for the analysis is median_house_value, representing housing prices.
- The predictors include:
    - Continuous variables: median_income, housing_median_age, total_rooms.
    - Categorical variable: ocean_proximity, which includes levels like "NEAR BAY," "INLAND," and "NEAR OCEAN."
- An initial multiple linear regression model was specified as:

$$\text{median\_house\_value} = \beta_0 + \beta_1 \times \text{median\_income} + \beta_2 \times \text{housing\_median\_age} + \beta_3 \times \text{total\_rooms} + \beta_4 \times \text{ocean\_proximity} + \epsilon$$

2. Data Cleaning and Preprocessing

- Handling Missing Values:
    - The dataset contained missing values in the total_bedrooms variable. Rows with missing values were removed using the na.omit() function to ensure data consistency.
- Categorical Variable Conversion:
    - The ocean_proximity variable was converted into a factor to allow its inclusion in the regression model as dummy variables.

3. Initial Model Fitting

- A baseline model was fitted using all identified predictors to assess their individual and combined contributions to explaining variability in median_house_value.
- Key metrics such as Adjusted R^2 and p-values were used to evaluate the model's performance and predictor significance.

4. Diagnostic Checks

To ensure the model adheres to the assumptions of linear regression, the following diagnostic tests were conducted:

- Autocorrelation:
    - The Durbin-Watson test was used to detect autocorrelation in the residuals.
    - Initial results indicated significant positive autocorrelation, prompting the inclusion of a lagged variable.
- Heteroscedasticity:
    - Residual plots revealed heteroscedasticity, suggesting the need for a transformation of the response variable.
- Multicollinearity:
    - Variance Inflation Factor (VIF) analysis confirmed that multicollinearity among predictors was not a significant concern.
- Influential Points:
    - Cook's Distance identified influential points that disproportionately affected the regression coefficients.

5. Remediation Techniques

Lagged Variable:

A lagged version of the response variable (lag_median_house_value) was introduced to address autocorrelation. The modified model showed improved residual behavior, as confirmed by a Durbin-Watson statistic close to 2.

Log Transformation:

The response variable was log-transformed to stabilize variance and address heteroscedasticity.

Handling Influential Points:

Observations with high Cook's Distance values were removed, resulting in a refined dataset and improved model stability.

6. Model Refinement

- The final model incorporated the lagged variable and log-transformed response, with significant improvements in Adjusted R^2 (approximately 70%) and Akaike Information Criterion (AIC) scores.

Analysis and Results :

1. Initial Model Results

The initial regression model included the predictors median_income, housing_median_age, total_rooms, and ocean_proximity. Key findings:

- **Adjusted R^2**: The initial model explained approximately 60% of the variance in housing prices.
- **Significant Predictors**:
    - **Median Income**: A strong positive association with housing prices, with a unit increase in median_income leading to a significant rise in house value.
    - **Ocean Proximity**: Properties closer to the ocean were associated with higher values, while inland locations showed significantly lower prices.
    - **Housing Median Age**: Older houses were positively correlated with price, though the effect was less pronounced compared to median_income.
- **Model Issues**:
    - The Durbin-Watson test indicated significant positive autocorrelation.
    - Residual vs. fitted plots revealed heteroscedasticity.
    - Influential points were identified using Cook's Distance.

2. Diagnostic Insights

Key diagnostics from the initial model:

- **Autocorrelation**:
    - The Durbin-Watson statistic was 0.825, far below the threshold of 2, indicating strong positive autocorrelation.
- **Heteroscedasticity**:
    - Residuals showed a funnel shape in the residual vs. fitted plot, confirming non-constant variance.
- **Influential Points**:
    - Cook's Distance highlighted several points with disproportionate influence on the model.

3. Refined Model Results

To address the identified issues, a lagged variable and log transformation were incorporated, and influential points were removed. The refined model demonstrated significant improvements:

- **Adjusted R^2**: Increased to 70%, indicating a better fit.
- **AIC**: Decreased significantly, confirming improved model parsimony.
- **Predictor Significance**:
    - **Median Income**: Continued to be the strongest predictor, with a log-transformed effect indicating that a unit increase corresponds to an approximate 18% increase in house prices.
    - **Ocean Proximity**:
        - Homes near the ocean showed higher values than those inland.
        - The INLAND category had a large negative coefficient, emphasizing lower property values in these areas.
    - **Housing Median Age**: Positively correlated with house prices, though its effect was smaller compared to other predictors.

4. Diagnostic Improvements

- **Autocorrelation**:
    - The inclusion of the lagged variable improved the Durbin-Watson statistic to 2.179, indicating no significant residual autocorrelation.
- **Heteroscedasticity**:
    - The log transformation of the response variable reduced variance instability, as observed in residual plots.
- **Influential Points**:
    - Removing high-leverage points based on Cook's Distance improved model robustness and reliability.

5. Final Model Equation

The final regression model can be expressed as:

$$\log(\text{median\_house\_value}) = \beta_0 + \beta_1 \times \text{median\_income} + \beta_2 \times \text{housing\_median\_age} + \beta_3 \times \text{total\_rooms} + \beta_4 \times \text{ocean\_proximity} + \beta_5 \times \text{lag\_median\_house\_value} + \epsilon$$

Conclusions

This analysis provides a detailed examination of the factors influencing housing prices in California, with key insights derived from a multiple linear regression model. By addressing diagnostic issues and refining the model, we have arrived at reliable and interpretable results. Below are the main conclusions:

Key Findings

1. **Significant Predictors:**
    - **Median Income**: The most influential predictor, with higher income levels strongly associated with increased housing prices. This finding emphasizes the importance of economic factors in determining property values.
    - **Ocean Proximity**: A major geographic determinant, with houses near the ocean valued significantly higher compared to inland properties. This reflects the desirability of coastal locations.
    - **Housing Median Age**: Older houses showed a modest positive correlation with price, possibly reflecting their association with established neighborhoods or desirable locations.
    - **Total Rooms**: A small but significant positive impact on house prices, indicating that larger houses tend to have higher values.

2. **Model Improvements:**
    - The refined model explained approximately 70% of the variability in housing prices (Adjusted $R^2 = 70\%$), a substantial improvement over the initial model (Adjusted $R^2 = 60\%$).

- Diagnostic checks confirmed that issues such as autocorrelation and heteroscedasticity were mitigated through the inclusion of lagged variables, log transformation of the response variable, and removal of influential points.

Practical Implications

- **For Real Estate Developers**: Focus on high-income areas and coastal regions for premium housing projects, as these factors are significant drivers of housing prices.
- **For Policymakers**: Develop equitable housing strategies that address the disparity in property values between inland and coastal areas.
- **For Urban Planners**: Use the identified predictors to design and allocate resources to neighborhoods likely to experience growth or require support.

Challenges and Limitations

- **Residual Heteroscedasticity**: While reduced, slight heteroscedasticity remains, which may affect confidence intervals and p-values. Future models could explore weighted least squares (WLS) or alternative transformations to address this further.
- **Non-Normal Residuals**: Minor deviations from normality in the residuals suggest the need for robust regression techniques in future analyses.

Future Research Directions

- Incorporating neighborhood-specific factors (e.g., public transportation access, school ratings) could improve model accuracy and applicability.
- Expanding the dataset to include temporal data could allow the modeling of trends over time, providing insights into the dynamics of the housing market.

In summary, this analysis highlights the key factors influencing California housing prices, with median income and ocean proximity emerging as the most significant predictors. The refined regression model, which explained 70% of the variance in housing prices, addressed issues such as autocorrelation and heteroscedasticity, ensuring reliability and interpretability. These findings provide valuable insights for real estate investors, policymakers, and urban planners, offering a foundation for data-driven decision-making and future research to further enhance predictive accuracy and market understanding.

Bibliography and Credits

Bibliography

1. California Housing Prices Dataset. Kaggle. Link to Dataset.
2. ISLR Resources. *An Introduction to Statistical Learning* (2nd Edition). Available at: https://www.statlearning.com/resources-second-edition.
3. R Documentation. *lm Function* in R. Accessed from https://www.rdocumentation.org/.
4. Kutner, Nachtsheim, and Neter. *Applied Linear Statistical Models*. McGraw-Hill Education.

Credits

Authors:

- Meghana Thimmaiah (A20563769): Contribution includes data analysis, diagnostics, and report preparation.
- Jagadeeswar Reddy Jillella (A20542891): Contribution includes model refinement, results interpretation, and documentation.
- Deepakshi Indresh (A20541754): Contribution includes data cleaning, visualizations, and report finalization.

Acknowledgments:

- Course Instructor: Prof.Kia Ong for guidance and support throughout the project and the course
- Data Source: Kaggle for providing the California Housing Prices dataset.