Final Project Part 3 Report

Gadiel David Flores Wendy Huang

2024-12-05

Contents

Contributions

- **Gadiel David Flores:** Description of contributions.
- **Wendy Huang:** Description of contributions.

Introduction

Housing affordability has become a pressing issue in many urban cities, including Toronto, where rising house prices have placed both homeowners and tenants out of reach for many years. This paper is conducted to investigating several key house indicators to the influence of house price. This is essential for policymakers , real estate professionals and potential homeowners seeking to navigate this challenging market.

Based on prior experience, we hypothesize that housing supply is a primary determinant of house prices. Therefore, we selected key indicators of housing supply as predictors: the number of available homes, homes sold, homes under construction, new homes completed, and the rate hike, which serves as a proxy for investment conditions. This analysis aims to use the multiple linear regression model to investigate the relationship between house prices and these factors, with the rate hike modeled as a categorical variable capture the impact of interest rate changes on housing affordability.

We conducted literature reviews to validate the selection of our predictors based on prior studies. Housing supply is consistently shown to be a critical factor influencing house prices. Research indicates that higher sales activity (houses sold) is often associated with increased demand, which drives prices upward. Conversely, an oversupply of unsold houses tends to push prices downward. These relationships are commonly examined using hedonic and linear regression models, where transactional data serve as key predictors for estimating price changes effectively (@housepriceprediction).

The number of newly constructed and completed houses influences house prices by affecting market supply. An increase in construction tends to stabilize prices by meeting demand, while lower construction rates can contribute to supply constraints and higher prices. "Global House Prices: Trends and Cycles" highlights the influence of new constructions and completions on house prices. Regression analyses reveal that increases in supply through construction stabilize prices, while supply shortages contribute to price hikes (@housingmarkets)

In the context of rate hikes, the CEPR report discusses their inclusion in econometric models to study the effects of borrowing costs on housing affordability and prices. It finds that higher rates reduce demand, reflected in lower prices, a relationship frequently captured through regression techniques (@whatdriveshouseprices)

Our literature review confirms that the selected predictors—houses sold, houses constructed, houses under construction, unsold houses, and rate hikes—have been previously studied and exhibit a linear relationship with house prices. Linear regression is recognized as a valid method for this analysis, offering a clear way to quantify the relationships between house prices (the response) and these predictors. The validity of the linear relationship will be tested and thoroughly discussed later, as this is a key objective of this paper.

For this analysis, the focus will be on identifying and validating the linear relationship between the predictors and the response variable. We will employ exploratory data analysis (EDA) and evaluate model performance using metrics such as AIC, BIC, and adjusted R^2. Through these statistical methods, we aim to establish that a linear regression relationship exists between house prices and the predictors: the number of available homes, homes sold, homes under construction, homes completed, and rate hikes.

The structure of this paper is as follows: In the @sex-methods section, we describe the dataset acquisition process and the data cleaning steps undertaken for subsequent analysis. @sex-results introduce the result of the linear regression and the validation of the model. @sex-limitation discuss the limitation of the model and dataset selection and processing. Ethics are discussed in @sex-ethics.

Methods

1. Data Preparation

- **Tools Used**: Analysis used **R** with `tidyverse` for data handling and `car` for diagnostics.
- **Steps Taken**:
  - Defined **New Housing Price Index** as the response variable and identified predictors like absorption, construction, GDP, and CPI.

- Removed missing values and duplicates, merged data from 1997-2016, and divided it quarterly.
- Used Cook's Distance to remove influential outliers.
  These steps ensured a clean dataset for accurate regression analysis.

2. Exploratory Data Analysis (EDA)

- **Tools Used**: `ggplot2` for visualization and `corrplot` for visualizing correlations.
  Scatterplots were used to check for linearity between predictors and the response variable. Some predictors showed potential relationships and multicollinearity issues. Variance inflation factors (VIFs) flagged these predictors with high multicollinearity for further review. This exploratory data analysis provided essential insights into variable relationships, guiding decisions on transformations and feature selection for improved model accuracy.

3. Model Building

- **Tools Used**: `MASS` for stepwise regression and `caret` for cross-validation.
  Both AIC- and BIC-based stepwise regressions tested multiple models to identify the most efficient predictor combinations.Log transformations were used for skewed predictors like `Detached_Unabsorbed_Quarterly_Avg`, improving linearity and stabilizing variance. A Box-Cox transformation ($\lambda = -0.3434$) was applied to the response variable, addressing residual non-normality and heteroscedasticity. These steps balanced model simplicity with predictive accuracy for optimal performance.

4. Assessment of Assumptions

- **Tools Used**: Residual plots, Q-Q plots, and statistical tests like Cook's Distance.
  The model's assumptions were thoroughly checked: linearity and homoscedasticity were verified using residuals vs. fitted value and residual vs. predictors plots. Normality of residuals was assessed with Q-Q plots and histograms, revealing minor tail deviations. These steps ensured the regression model's validity, enhancing confidence in its results and inferences.

5. Model Diagnostics

- **Tools Used**: Performance metrics such as Adjusted R^2, AIC, BIC, RMSE, and MAE.
  Model performance was compared using AIC and BIC criteria. The simpler 8-predictor BIC model was chosen for its nearly equivalent performance to the more complex AIC model. These diagnostics ensured the selected model effectively balanced simplicity and performance, adhering to best practices in evaluating regression models.

6. Mitigating Issues

- **Tools Used**: Log transformations and feature selection techniques.
  To enhance model reliability, predictors with high VIFs, such as `GDP_Quarterly_Avg`, were excluded to reduce multicollinearity. Log transformations addressed non-linearity in variables like `Completed_Construction_Semi`. Outliers identified through Cook's Distance were removed to minimize their impact on coefficients. These actions ensured the model adhered to regression assumptions and improved its interpretability and accuracy.

7. Conclusion of Methods

The stepwise, structured approach to developing the MLR model ensured robustness and validity while aligning with theoretical principles from the course material. Preprocessing and feature selection techniques addressed potential biases and assumption violations. Through careful validation and diagnostics, the BIC-selected model was identified as the final model, balancing simplicity with high predictive performance. This methodical process provides confidence in the model's applicability and reliability for predicting **Quarterly_Average**.

Results

1. Introduction to Results This section presents the comprehensive results of the multiple linear regression (MLR) analysis. It details the evaluation, refinement, and selection of the final model, emphasizing the predictive performance, key decisions, and diagnostic tests conducted to validate the model's robustness and reliability.

2. Model Comparison The analysis involved testing various models to balance predictive power and interpretability. Key model selection methods included: - **AIC-based Stepwise Regression**: Focused on optimizing model fit while allowing for complexity. - **BIC-based Stepwise Regression**: Prioritized simpler models to reduce overfitting risks.

Table 1: Comparison of AIC- and BIC-Selected Models

Metric	AIC-Selected Model	BIC-Selected Model
Adjusted R^2	0.99370	0.99370
Residual Standard Error (RSE)	0.01029	0.01029
RMSE (Cross-Validation)	0.01088	0.01110
MAE (Cross-Validation)	0.00918	0.00953
Predictors Selected	10.00000	10.00000

Both models performed similarly, with the AIC-selected model slightly outperforming on cross-validation metrics.

3. Key Decisions To build a strong and reliable model, important features like `Detached_Unabsorbed_Quarterly_Avg` and `CPI_Quarterly_Avg` were kept because they were both statistically significant and theoretically important. Less useful features, like `Completed_Construction_Detached`, were removed to avoid overfitting the model. Transformations were used to improve accuracy: log transformations helped manage non-linearity and stabilize data for variables like `Detached_Unabsorbed_Quarterly_Avg`, while a Box-Cox transformation adjusted the response variable `Quarterly_Average` for better fit. For model selection, AIC was used to focus on minimizing errors, and BIC helped ensure the model remained simple and avoided overfitting.

4. Model Diagnostics and Assumptions To ensure the model's accuracy and reliability, several checks were performed. Residual analysis showed that the errors were mostly normal, as confirmed by the Q-Q plot, and there were no noticeable patterns in residual variance, confirming consistent variance (homoscedasticity). Multicollinearity was flagged between predictors like GDP_Quarterly_Avg and CPI_Quarterly_Avg using Variance Inflation Factors (VIFs), but these variables were kept because of their importance in the model. Influential data points, identified using Cook's Distance, were removed to avoid skewed results. Overall, the model satisfied all key assumptions of linear regression, including linearity, normality, independence, and homoscedasticity, ensuring it is both valid and robust.

5. Interpretation of Final Model The final BIC-selected model included 10 significant predictors. Key metrics:
- **Multiple** R^2: 0.9946 (99.46% of the variability explained).
- **Adjusted** R^2: 0.9937 (robust against overfitting).
- **Residual Standard Error**: 0.01029.
- **Cross-Validation Results**:
- **RMSE**: 0.01110
- **MAE**: 0.00953
- R^2: 0.9933

Significant Predictors:
- **Detached__Absorption__Quarterly__Avg** ($\beta = -1.552 \times 10^{-5}, p < 0.001$): Negative impact.
- **Detached__Unabsorbed__Quarterly__Avg** ($\beta = 2.236 \times 10^{-5}, p < 0.001$): Positive impact.
- **Semi__Unabsorbed__Quarterly__Avg** ($\beta = -2.864 \times 10^{-5}, p < 0.05$): Negative impact.
- **GDP__Quarterly__Avg** ($\beta = -3.004 \times 10^{-7}, p < 0.001$): Negative impact.
- **CPI__Quarterly__Avg** ($\beta = 0.01315, p < 0.001$): Strong positive effect.
- **Starting__Detached__Construction** ($\beta = -4.393 \times 10^{-6}, p < 0.001$): Negative impact.
- **Under__Construction__Detached** ($\beta = 6.240 \times 10^{-6}, p < 0.001$): Positive impact.
- **Under__Construction__Semi** ($\beta = 2.180 \times 10^{-5}, p < 0.001$): Positive impact.
- **Completed__Construction__Semi** ($\beta = -1.250 \times 10^{-5}, p < 0.01$): Negative impact.

6. Visual Representation

- **Residuals vs. Fitted Values**: No patterns detected, confirming linearity.
- **Q-Q Plot**: Residuals followed a normal distribution.
- **Predicted vs. Actual**: Strong alignment along the diagonal, indicating high predictive accuracy.

7. Conclusion of Results The analysis focused on understanding the factors influencing new home prices, and evaluating a robust model to predict them. Using multiple linear regression, the final model highlighted several key predictors with significant effects on new home prices. For instance, unabsorbed homes had a strong positive impact. Conversely, absorption and completed construction negatively affected price. Macroeconomic indicators like GDP showed a negative effect, while CPI had a substantial positive impact. Construction activity variables, such as starting construction and under construction, further revealed nuanced effects on price trends. The model, explaining over 99% of the variability in new home prices, provides a comprehensive view of how inventory, construction activity, and macroeconomic factors collectively shape the housing market, offering valuable insights for policymakers and market stakeholders. Our model failed validation with our test and train datasets, indicating that it struggles to generalize accurately across different data subsets. This highlights potential limitations in its ability to make reliable predictions. However, it is important to note that the primary objective of this model is not solely predictive accuracy, but rather to provide insights into underlying behaviors and relationships within the data. Despite its shortcomings in predictive performance, the model's ability to identify significant patterns and key drivers of variability remains valuable for understanding the dynamics at play.

Conclusion and Limitations

Ethics Discussion

References