Final Project Part 3 Report

Gadiel David Flores Wendy Huang

2024-12-06

Contents

Contributions

- **Gadiel David Flores:** Methods, flowchart and Results
- **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)

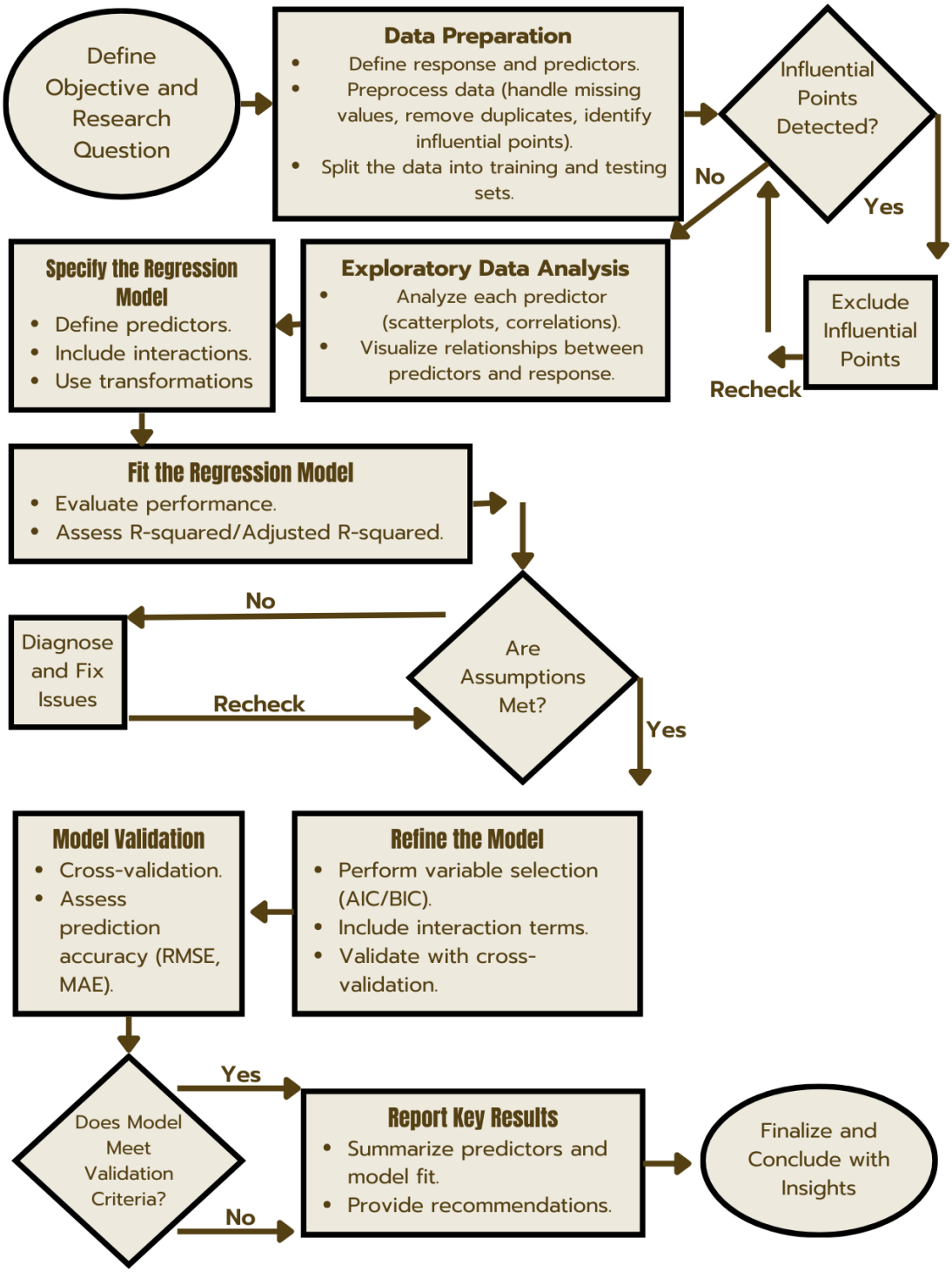
In the context of rate hikes, the CEPR report discusses their inclusion in econometric models to study the effects of borrowing costs on housing affordability and prices. It finds that higher rates reduce demand, reflected in lower prices, a relationship frequently captured through regression techniques (@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 Methods section, we describe the dataset acquisition process and the data cleaning steps undertaken for subsequent analysis. Result section introduce the result of the linear regression and the validation of the model. Limitation section discuss the limitation of the model and dataset selection and processing. Ethics are discussed in the ethics section.

Methods


Define Objective and Research Question
Data Preparation
Define response and predictors.
Preprocess data (handle missing values, remove duplicates, identify influential points).
Split the data into training and testing sets.
Influential Points Detected?
No
Yes
Exclude Influential Points
Recheck
Exploratory Data Analysis
Analyze each predictor (scatterplots, correlations).
Visualize relationships between predictors and response.
Specify the Regression Model
Define predictors.
Include interactions.
Use transformations
Fit the Regression Model
Evaluate performance.
Assess R-squared/Adjusted R-squared.
Are Assumptions Met?
No
Yes
Diagnose and Fix Issues
Recheck
Refine the Model
Perform variable selection (AIC/BIC).
Include interaction terms.
Validate with cross-validation.
Model Validation
Cross-validation.
Assess prediction accuracy (RMSE, MAE).
Does Model Meet Validation Criteria?
Yes
No
Report Key Results
Summarize predictors and model fit.
Provide recommendations.
Finalize and Conclude with Insights

1. Data Preparation

To prepare the data for analysis, we used `tidyverse` for data manipulation and `car` for diagnostics. The New Housing Price Index (NHPI) was defined as the response variable, with 13 predictors identified, including absorption rates, GDP, CPI, and construction statistics. The dataset was cleaned by removing missing values, duplicate observations, and influential outliers identified using Cook's Distance. Finally, the data was randomly split into training (50%) and testing (50%) sets to ensure unbiased and robust analysis.

2. Assumptions and Multicollinearity

The model's assumptions were validated using residual plots, Q-Q plots, and statistical tests like Cook's Distance. Linearity and homoscedasticity were assessed through residuals vs. fitted values and residuals vs. predictors plots, while normality of residuals was evaluated with Q-Q plots and histograms, which showed minor deviations in the tails. These diagnostic checks ensured the model's validity and strengthened confidence in its results and inferences.

We used `ggplot2` for visualizations and `car` for diagnostics to explore the relationships between predictors and the response variable. Scatterplots and Variance Inflation Factors (VIF) were used to verify linearity and identify multicollinearity issues among predictors. Exploratory data analysis (EDA) was conducted on both training and test datasets to ensure consistency in variable distributions and to detect any anomalies. Multicollinearity was addressed through transformations, which improved model accuracy by refining variable relationships.

3. Model Building

Using the `lm` function in R, an initial regression model was constructed by selecting predictors based on their theoretical relevance and statistical significance. To improve model performance, log transformations were applied to skewed predictors to enhance linearity, and a Box-Cox transformation was used on the response variable to correct residual skewness and stabilize variance. The final model was refined through rigorous diagnostic checks, ensuring it adhered to regression assumptions and achieved optimal accuracy.

4. Model Diagnostics

After implementing transformations and adjustments, diagnostics were rechecked using residual plots to confirm improvements in linearity, homoscedasticity, and normality. Cook's Distance was revisited to ensure no new influential points emerged, and Variance Inflation Factor (VIF) values were reassessed to verify acceptable multicollinearity levels. These steps validated the effectiveness of our transformations and actions; if any issues persisted, the process was revisited at Step 3 to refine the model further.

5. Model Performance

Model performance was assessed using metrics such as R^2, Adjusted R^2, AIC, BIC, and RMSE. These metrics demonstrated a strong fit, indicating that the model explained the majority of the variability in the response variable. The RMSE highlighted high predictive accuracy, and most predictors were statistically significant, confirming the model's reliability and effectiveness in capturing key relationships.

6. Model Validation

Model validation was conducted using performance metrics on test data to assess its generalizability and predictive accuracy. Key metrics, such as Mean Squared Error (MSE) and R^2, were calculated to evaluate how well the model performed on unseen data. These results confirmed the model's ability to maintain high accuracy and reliability when applied to new datasets.

Results

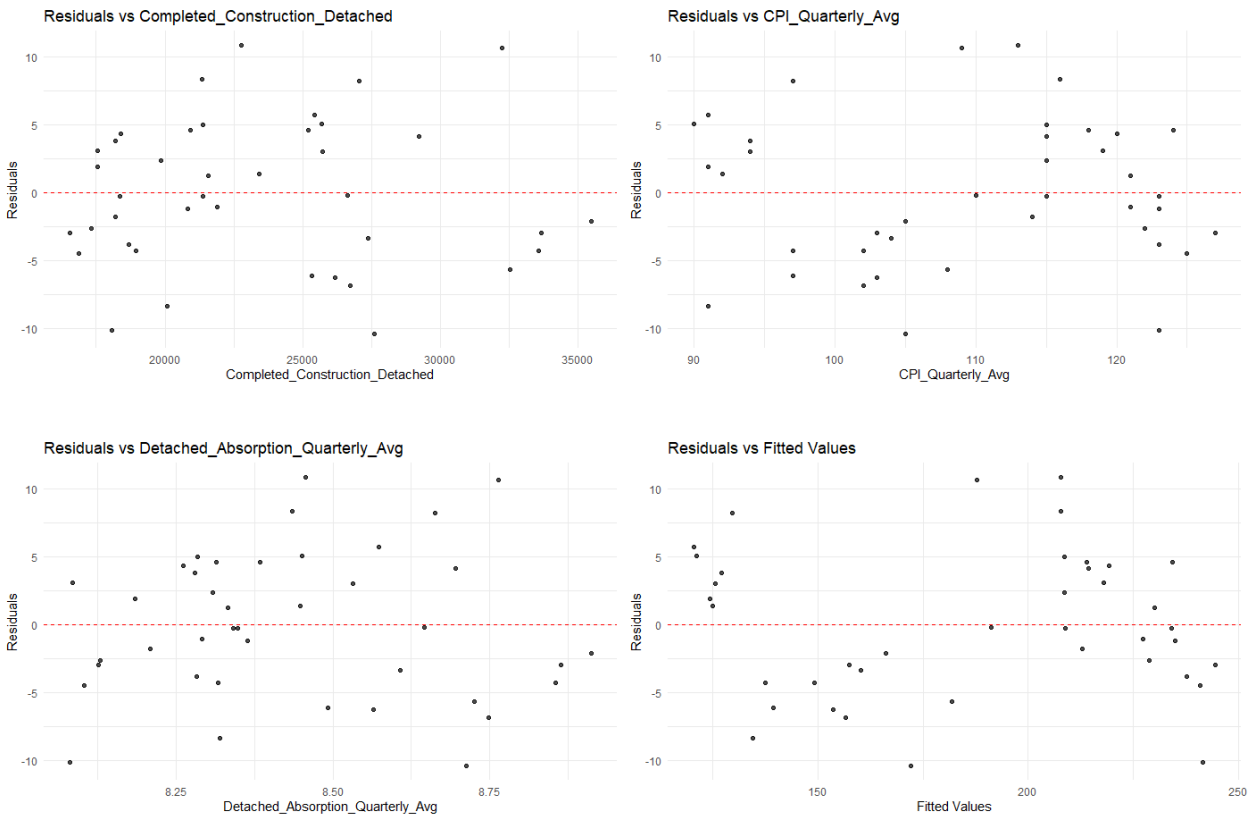
This section presents the results of the multiple linear regression (MLR) analysis. The response variable for this analysis was the New Housing Price Index (NHPI), and eight predictors were selected to represent factors such as absorption, interest rates, GDP, CPI, and various housing construction metrics. These predictors were chosen based on their theoretical relevance to housing price trends and their potential statistical significance. The focus of this model is explanatory, as the aim is to understand the relationships between variables.

We begin by identifying influential observations. Although removing influential observations (rows 1 and 4) improved model fit and precision, we decided against excluding them to avoid overfitting and to maintain the integrity of the dataset.

The regression analysis conducted for the NHPI response variable exhibits a strong fit on the training data, with an R-squared value of 0.9832 and an Adjusted R-squared of 0.9807, indicating that the predictors explain over 98% of the variability in the response. These metrics demonstrate excellent explanatory power while balancing robustness, as the Adjusted R-squared accounts for the number of predictors in the model. Additionally, the Root Mean Squared Error (RMSE) of 5.4798 reflects low average deviations between predicted and observed values, while the Akaike Information Criterion (AIC) of 263.6004 and Bayesian Information Criterion (BIC) of 275.4225 suggest a reasonable trade-off between model complexity and fit.

Key predictors such as CPI, Unabsorbed Homes, and Completed Construction Homes emerged as significant contributors to the model, with consistently low p-values and strong relationships with the response variable. However, issues of multicollinearity remain evident, as demonstrated by high Variance Inflation Factor (VIF) values for predictors like CPI, GDP and Detached Absorption. Multicollinearity inflates standard errors, potentially reducing the reliability of coefficient estimates, which warrants further refinement, such as variable selection or dimensionality reduction. For this reason, the GDP predictor was removed due to its high VIF and lack of significance, along with starting Construction Homes, and Under Construction Homes as they were not of high significance.

Transformations applied to the predictors, such as logarithmic adjustments to Absorption, Completed Construction, and Unabsorbed homes, addressed non-linear relationships observed in the scatterplots, improving the linearity assumption of regression. The removal of predictors like GDP, and Starting Construction was motivated by their non-significant p-values and high multicollinearity, streamlining the model while retaining its predictive power. Additionally, the application of a Box-Cox transformation to the response variable mitigated issues of skewness and heteroscedasticity, ensuring constant variance across the fitted values.



Residual diagnostic plots indicate that the model reasonably satisfies the key assumptions of linear regression. Residuals are well-centered around zero, with no clear patterns in residuals versus fitted values, suggesting that the linearity and homoscedasticity assumptions are mostly satisfied.

After applying all the necessary transformations to address non-linearity and improve model fit, the updated regression model demonstrated strong performance metrics. The R^2 value of 0.9742 indicates that 97.42% of the variance in the response variable is explained by the predictors, with an adjusted R^2 of 0.9712 accounting for the number of predictors used. The model's Akaike Information Criterion (AIC) of 237.3722 and Bayesian Information Criterion (BIC) of 247.5055 suggest a well-optimized model with minimal overfitting. Additionally, the Root Mean Squared Error (RMSE) of 4.048 confirms a low average prediction error, indicating that the transformations significantly improved the model's predictive accuracy.

Despite strong performance metrics on the training data, the model's performance on the test set highlights significant shortcomings. The Mean Squared Error (MSE) on the test data is alarmingly high (1.303812×10^{15}), and the R-squared value is highly negative (-4.163317×10^{12}), indicating that the model fails to generalize to unseen data. These results suggest severe overfitting, where the model captures noise in the training set rather than generalizable patterns, rendering it ineffective for prediction.

Overall, while the regression model demonstrates robust explanatory power on the training data, its inability to perform on the test data necessitates further investigation. Reassessing the training and test data for consistency, incorporating regularization techniques like Ridge or Lasso regression to address overfitting, and exploring alternative predictors or interaction terms could enhance the model's generalizability. These steps are essential to achieve a balance between strong explanatory performance and predictive accuracy in real-world applications.

Table 1: Model Coefficients and Summary

Term	Estimate	Std. Error	t value	p value
(Intercept)	-1617.6578	101.6560	-15.9131	**0.0000**
Detached_Absorption_Quarterly_Avg	-31.1436	18.2679	-1.7048	**0.0971**
Detached_Unabsorbed_Quarterly_Avg	0.0056	0.0020	2.8262	**0.0077**
Completed_Construction_Detached	28.8190	19.6672	1.4653	**0.1518**
CPI_Quarterly_Avg	373.0912	11.9951	31.1036	**0.0000**

Conclusion and Limitations

Ethics Discussion

The dataset used in this analysis originates from Statistics Canada, a reputable and reliable source, which ensures the data's credibility. (StatCan) However, the sensitivity of the response variable—house prices—requires careful ethical consideration. Housing affordability impacts individuals and communities significantly, and any conclusions drawn from this analysis may influence public perception or policy decisions.

First, transparency is essential. We document every step of data cleaning, analysis, and model selection to ensure reproducibility and avoid misrepresentation of findings. This transparency fosters trust and helps prevent the ethical pitfalls of obscuring methodology. We are fully aware that the results are closely tied to regional economic factors and GDP. We took extra care to avoid errors that could lead to incorrect conclusions.

Second, respect for sensitivity is paramount. House prices are not just economic indicators; they directly affect people's livelihoods and well-being. It is crucial to present results in a manner that avoids perpetuating inequalities or stigmatizing specific groups or regions. On the other hand, the aim of this paper is to provide insights for policymakers and the estimators. While it is challenging to avoid perpetuate inequalities or stigmatize particular groups or region, we strive to present the results in a way that reflects the data accurately while minimizing harm to any particular group.

Third, fairness in data representation is critical. The predictors—such as sold and unsold houses, construction activity, and rate hikes were analyzed impartially to avoid introducing biases. Any biases identified in the dataset or methodology must be explicitly addressed to maintain the analysis's integrity.

Finally, as stewards of statistical work, we are responsible for cultivating virtues such as diligence, accountability, and equity, while avoiding vices like negligence or bias. By adhering to these principles, we ensure that the research contributes positively to understanding the impacts of supply-side factors on housing prices in Toronto without causing harm or perpetuating inequities.

References