Final Project Part 3 Report

Gadiel David Flores Wendy Huang

2024-12-02

Contents

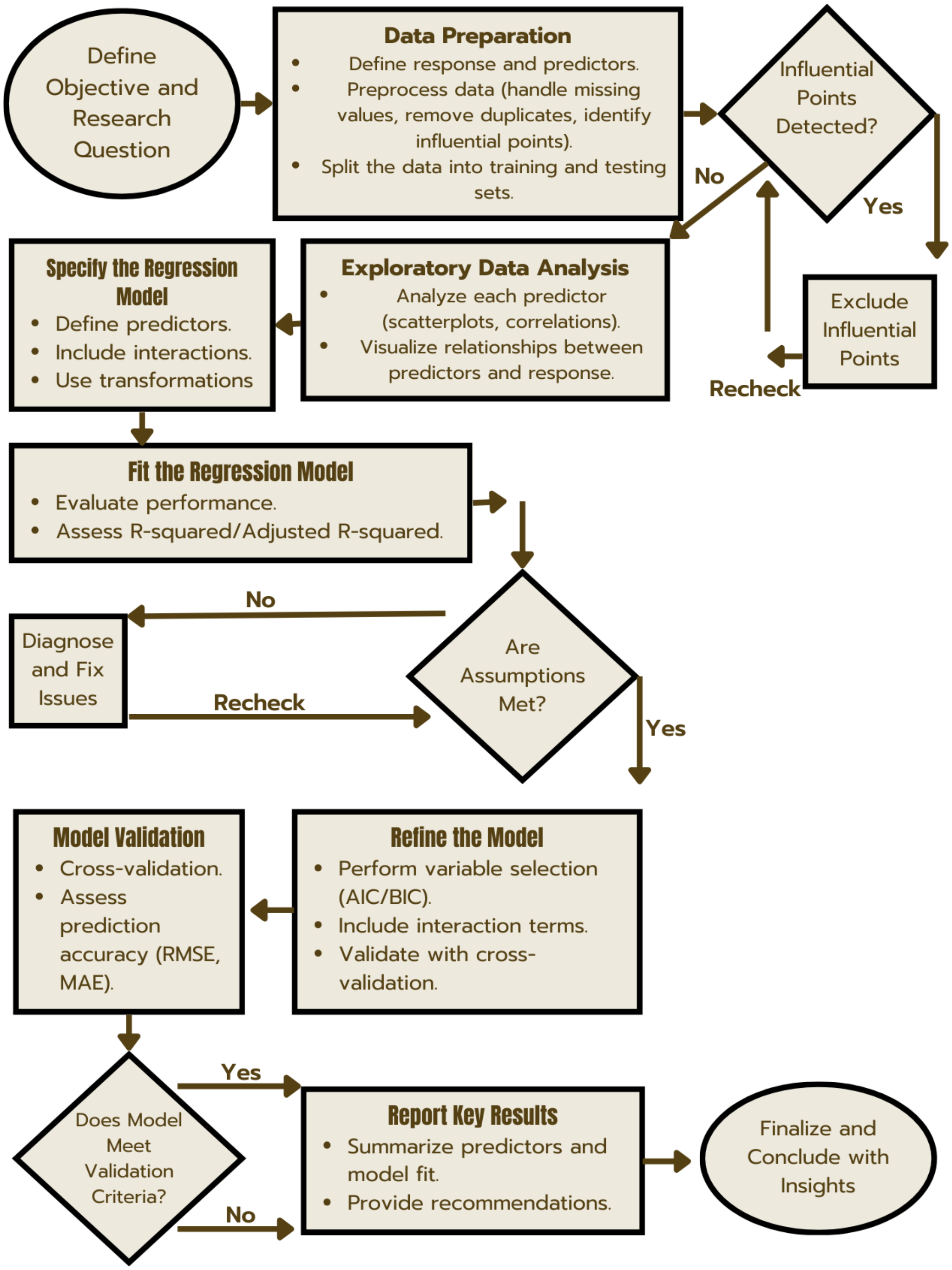
Contributions

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

Introduction

Methods

```r
knitr::include_graphics(here::here("final_paper", "Flow Chart.png"))
```



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?

Exclude Influential Points

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?

Diagnose and Fix Issues

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?

Report Key Results

- Summarize predictors and model fit.
- Provide recommendations.

Finalize and Conclude with Insights

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 mini-

mize 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.

```r
library(knitr)

comparison_table <- data.frame(
  Metric = c(
    "Adjusted R^2",
    "Residual Standard Error (RSE)",
    "RMSE (Cross-Validation)",
    "MAE (Cross-Validation)",
    "Predictors Selected"
  ),
  `AIC-Selected Model` = c(
    0.9937,
    0.01029,
    0.01088,
    0.00918,
    10
  ),
  `BIC-Selected Model` = c(
    0.9937,
    0.01029,
    0.01110,
    0.00953,
    10
  )
)

kable(
  comparison_table,
  col.names = c("Metric", "AIC-Selected Model", "BIC-Selected Model"),
  caption = "Comparison of AIC- and BIC-Selected Models",
  align = "lcc"
)
```

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

Metric	AIC-Selected Model	BIC-Selected Model
Adjusted Rˆ2	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.
- Residuals vs. Predictors:
- Q-Q Plot: Residuals followed a normal distribution.

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