STA302 Fall 2024 Final Project Part 1: Research Proposal and Data Introduction

Research Proposal and Data Introduction

Wendy Huang Gadiel David Flores

2024-04-10

1 Work Division

1.1 Group Member 1:

Main Focus: Research and Data Analysis

- **Contributions**:
  - Locate and clean the dataset.
  - Write the **Data Description** section (300 words).
  - Fit the **Multiple Linear Regression Model** and write the **Preliminary Results** section (300 words).
  - Conduct **Residual Analysis** and create the required residual plots.

1.1.1 Specific Tasks:

1. **Locate the dataset** that fits the project requirements (open-source, contains a response variable and at least 9 predictors).
2. **Clean the data**: Handle missing values, prepare numerical and categorical variables, and ensure it's ready for regression.
3. **Write the Data Description**:
   - State where the data was found.
   - Describe how the data was originally collected by the curator.
   - Summarize the response variable and justify why it's suitable for linear regression.
   - Summarize the predictor variables, provide numerical/graphical summaries, and interpret the descriptive statistics.

4. **Fit the multiple linear regression model** using R, incorporating at least 5 predictors, including one categorical predictor.
5. **Conduct residual analysis**:
    - Check for violations of linear regression assumptions.
    - Create residual plots and interpret their results.
6. **Write the Preliminary Results**:
    - Report on the fitted model.
    - Discuss model assumptions (whether violated or not).
    - Compare the preliminary model results to findings from the literature (summarized by Group Member 2).

1.2 Group Member 2:

Main Focus: Research Question, Literature Review, and Report Writing

- **Contributions**:
    - Write the **Introduction** section (350 words).
    - Conduct the **Literature Review**: Summarize three peer-reviewed academic papers related to the research question.
    - Write the **Ethics Discussion** section (100-200 words).
    - Compile the **Bibliography** and format citations.
    - Final editing and formatting of the proposal.

1.2.1 Specific Tasks:

1. **Formulate the research question**: Define a clear, focused research question based on the chosen dataset and explain why it's suitable for linear regression.
2. **Conduct the literature review**:
    - Find and summarize three peer-reviewed academic papers related to the topic.
    - Describe how each paper relates to the proposed research question and supports the use of linear regression.
3. **Write the Introduction**:
    - Introduce the relevance/importance of the research topic.
    - Clearly state the research question.
    - Summarize the results of the three academic papers and explain how they connect to your research.
    - Justify why linear regression is an appropriate tool for this analysis.

4. **Write the Ethics Discussion**:
    - Assess whether the dataset is trustworthy and ethically collected
5. **Compile the bibliography**:
    - Ensure proper formatting and inclusion of references in the proposal using APA format.
6. **Final editing and formatting**:
    - Review the entire document

2 Introduction (350 words)

3 Data Description (300 words)

3.1 Data Source

Using data from Statistics Canada, we extracted five datasets. Statistics Canada (2024e) served as the foundation for this research paper, as it contained the main statistic we wanted to analyze. Statistics Canada (2024e) represents the new housing price index, which calculates the average cost of new housing each month by multiplying by 100 and dividing by the base year. This provides an index reflecting housing costs for any given year and month. We then utilized Statistics Canada (2024c), Statistics Canada (2024d), Statistics Canada (2024b), Statistics Canada (2024f), and Statistics Canada (2024a) to extract our 13 predictors, including Absorption, GDP, CPI, interest rates, and Construction. Specifically, Absorption and Construction are further divided into two parts. Absorbtion is divided into semi-detached homes and single homes as well as empty and sold houses for any given month. Construction is categorized into three parts: new construction, under construction, and finished construction. These datasets were cleaned by removing missing values, converting data into whole integers, and merging them into one dataset. Additionally, the data spans from 1997 to 2016 and is divided quarterly, providing 76 entries. Lastly, we used interest rates to create our categorical predictor. This was done by comparing the entry i with entry i+1 in order to determine if there was an interest hike.

3.2 Response and Predictor Variables

Given our research question [TODO: what specific factors have the greatest effect on home prices in Canada]. The response variable chosen to help answer our research question is 'new housing price index'. This variable is suitable for linear regression because it is continuous, enabling us to model its relationships with multiple predictors. Additionally, since the new

housing price index is influenced by economic variables, it allows us to identify trends and patterns through our linear regression model. For these reasons, it is a strong candidate for our model. Our five predictor variables are the number of available homes, homes sold, homes starting construction, new homes completed, and rate hikes

4 Ethics Discussion (100-200 words)

5 Preliminary Results (300 words)

6 Residual Analysis

In this section, we will explore the assumptions made for our linear regression models and check for any violations. To begin, using our model in Section 6.1, we created residual plots to verify these assumptions. The first graph we will examine is Figure 1, which is our Residuals vs. Fitted Values plot. This graph shows some clustering around a fitted value of 100 and above the 0 residual line. It also displays a decreasing fanning pattern. These two characteristics suggest potential non-linearity and some correlated errors.

Next, we will review the residuals vs. predictors. In Figure 4 and Figure 5 The plots for homes being built and completed homes show the least amount of assumption violations, as the graphs exhibit good spread and randomness. In contrast, the plots for homes sold Figure 2 and unsold homes Figure 3 display some fanning and clustering, which may indicate correlated errors and non-constant variance. Our categorical predictor, rate hike Figure 6 , reveals some non-constant variance, as the spread of residuals for "1" is wider than for "0."

Lastly, while most residuals align well with the line in the QQ plot Figure 7 , there are deviations at the lower and upper extremes, suggesting some skewness. This indicates that the residuals are not perfectly normally distributed, which may signal a violation of the normality assumption.

6.1 Model Fitting

Data
- **Quarterly Average**: This is our response variable, representing the average cost of a home in Canada.
- **Detached Absorption Quarterly Avg**: This is our first predictor variable, describing the number of houses sold in a quarter.
- **Starting Detached Construction**: This predictor variable indicates the number of houses that began construction in a quarter.
- **Completed Construction Detached**: This represents the number of homes completed in

a given quarter.
- **Rates**: This is our categorical predictor, with a value of 1 if there was a rate hike in that quarter or 0 if there was not.
- **Detached Unabsorbed Quarterly Avg**: This is our final predictor, indicating the number of available houses that were not sold during that quarter.

```
Call:
lm(formula = Quarterly_Average ~ Detached_Absorption_Quarterly_Avg +
    Detached_Unabsorbed_Quarterly_Avg + Starting_Detached_Construction +
    Completed_Construction_Detached + rates, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-38.698  -6.746   1.921   9.503  23.227

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       8.852e+01  1.728e+01   5.124 2.54e-06 ***
Detached_Absorption_Quarterly_Avg -2.270e-02  5.407e-03  -4.199 7.76e-05 ***
Detached_Unabsorbed_Quarterly_Avg  5.315e-03  2.848e-03   1.867 0.066157 .
Starting_Detached_Construction     9.611e-04  3.713e-04   2.588 0.011725 *
Completed_Construction_Detached    2.977e-03  1.028e-03   2.896 0.005032 **
rates                             -1.341e+01  3.841e+00  -3.490 0.000839 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.66 on 70 degrees of freedom
Multiple R-squared:  0.438, Adjusted R-squared:  0.3979
F-statistic: 10.91 on 5 and 70 DF,  p-value: 8.736e-08
```

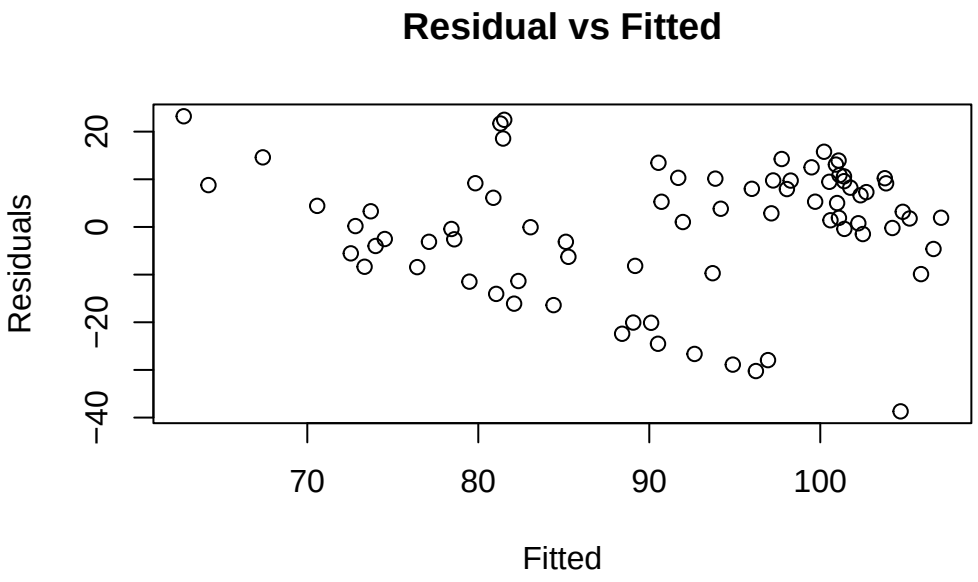

Figure 1: Residual vs Fitted Values

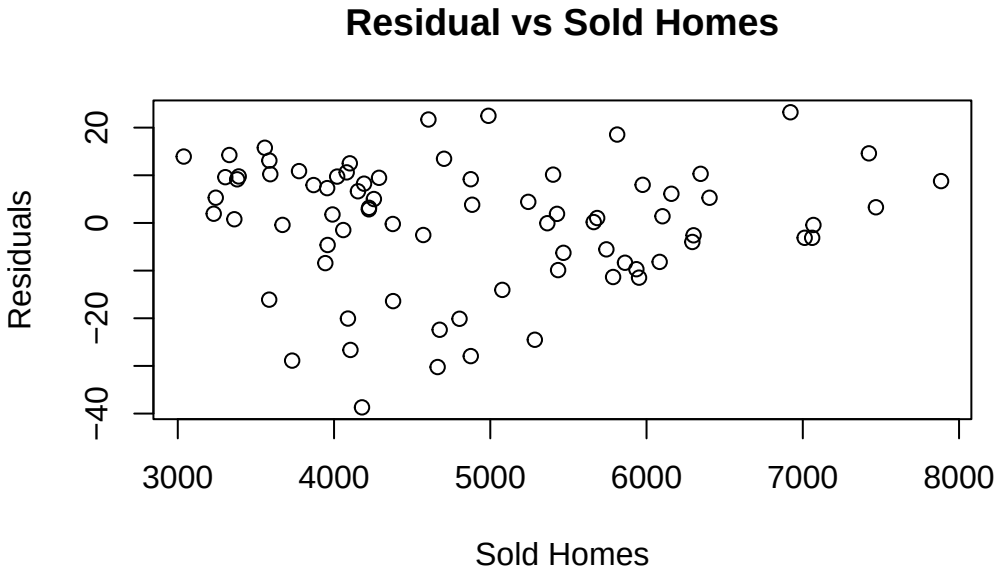

Figure 2: Residuals vs Each Predictor

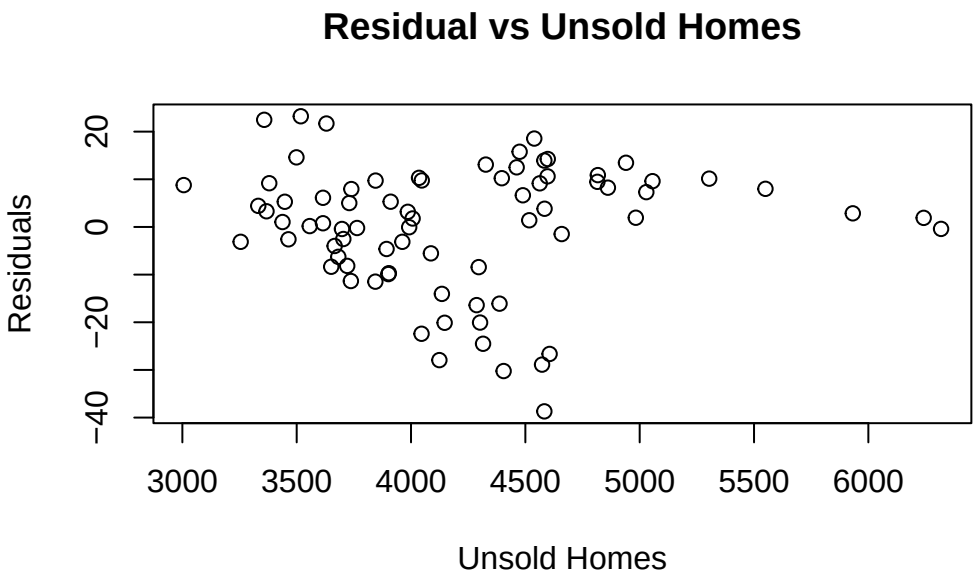

Figure 3: Residuals vs Each Predictor

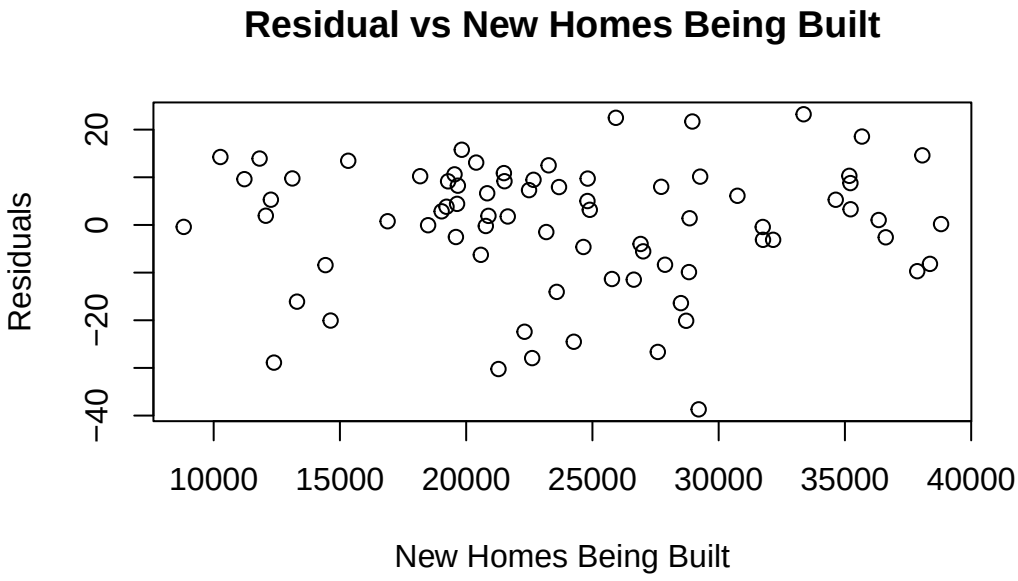

Figure 4: Residuals vs Each Predictor

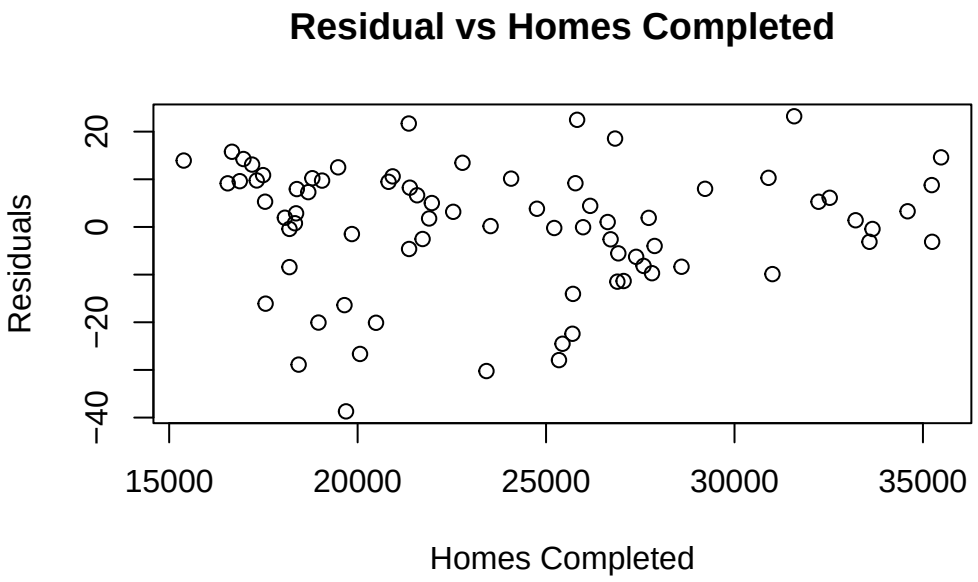

Figure 5: Residuals vs Each Predictor

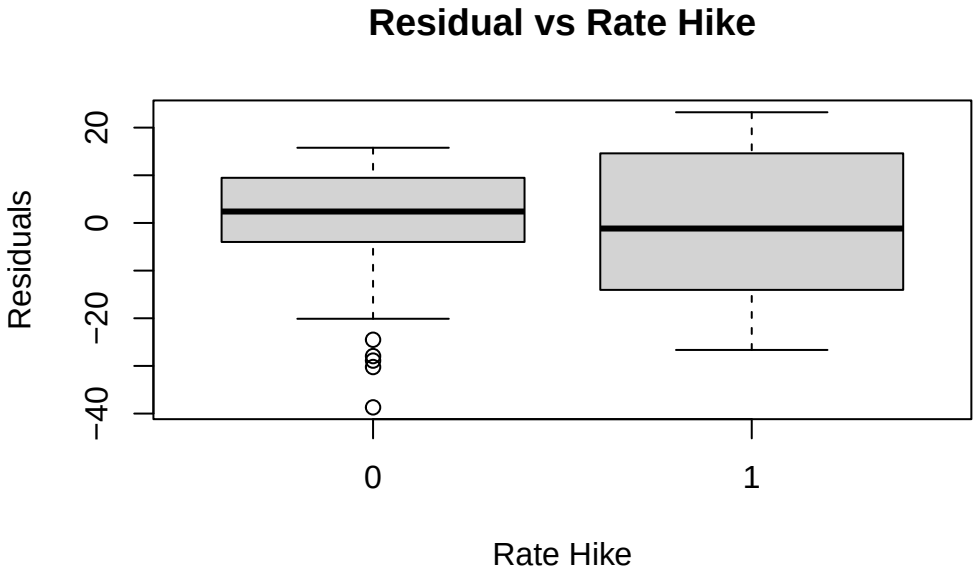

Figure 6: Residuals vs Each Predictor

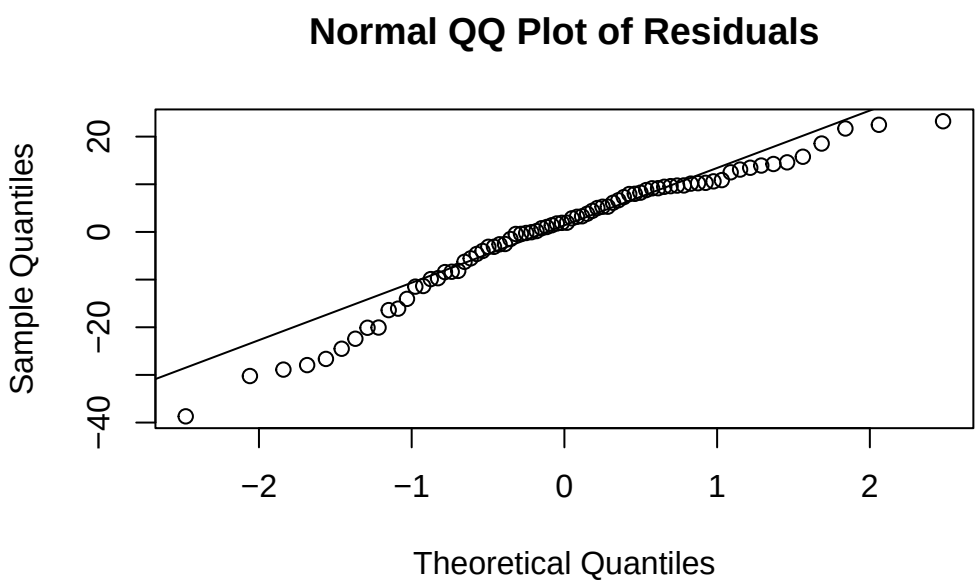

Figure 7: QQ Plot

Bibliography

Statistics Canada. 2024a. *Canada Mortgage and Housing Corporation, Absorptions and Unabsorbed Inventory, Newly Completed Dwellings, by Type of Dwelling Unit in Census Metropolitan Areas*. Statistics Canada. https://doi.org/https://doi.org/10.25318/3410014901-eng.

———. 2024b. *Canada Mortgage and Housing Corporation, Housing Starts, Under Construction and Completions, All Areas, Quarterly*. Statistics Canada. https://doi.org/https://doi.org/10.25318/3410013501-eng.

———. 2024c. *Consumer Price Index, Monthly, Not Seasonally Adjusted*. Statistics Canada. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000401.

———. 2024d. *Gross Domestic Product (GDP) at Basic Prices, by Industry, Monthly (x 1,000,000)*. Statistics Canada. https://doi.org/https://doi.org/10.25318/3610043401-eng.

———. 2024e. *New Housing Price Index (2007=100)*. Statistics Canada. https://doi.org/https://doi.org/10.25318/1810005201-eng.

———. 2024f. *Table 10-10-0122-01 Financial Market Statistics, Last Wednesday Unless Otherwise Stated, Bank of Canada*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1010012201.