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)

5.1 Model Fitting

6 Load Data

- $Quarterly_{A}verage$: the body mass index of an individual as the response
- $Detached_{A}bsorption_{Q}uarterly_{A}vg$: age of the individual (rounded to an integer)
- $Starting_{D}etached_{C}onstruction$: Male or Female designation for gender (this survey predates when other options for this category were provided)
- $Completed_{C}onstruction_{D}etached$: a self-identified race with options White, Black, Hispanic, Mexican, Other
- $rates$: The midpoint of a household income bracket
- $Detached_{U}nabsorbed_{Q}uarterly_{A}vg$: Hours of sleep per night on weekdays, self-reported

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(ggplot2)
library(readr)
# loads the dataset into R
data <- read_csv(here::here("data/analysis_data/analysis_data.csv"))
```

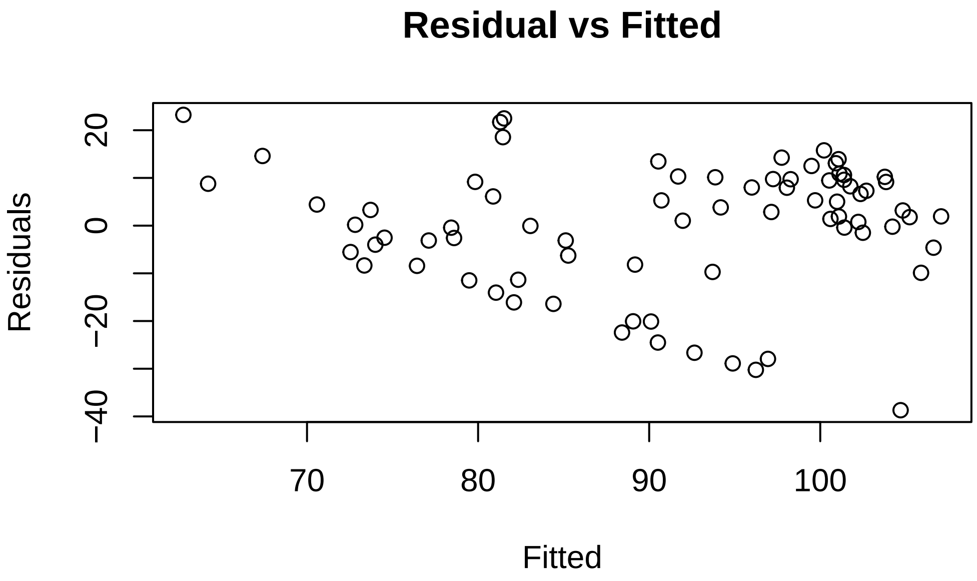
```
Rows: 76 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl  (14): Quarterly_Average, Detached_Absorption_Quarterly_Avg, Semi_Absorp...
date  (1): Quarter

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

7 Multiple Linear Model

```
model <- lm(Quarterly_Average ~ Detached_Absorption_Quarterly_Avg +
              Detached_Unabsorbed_Quarterly_Avg +
              Starting_Detached_Construction +
              Completed_Construction_Detached +
              rates,
            data = data)
intercept <- round(as.numeric(coef(model)[1]))
```

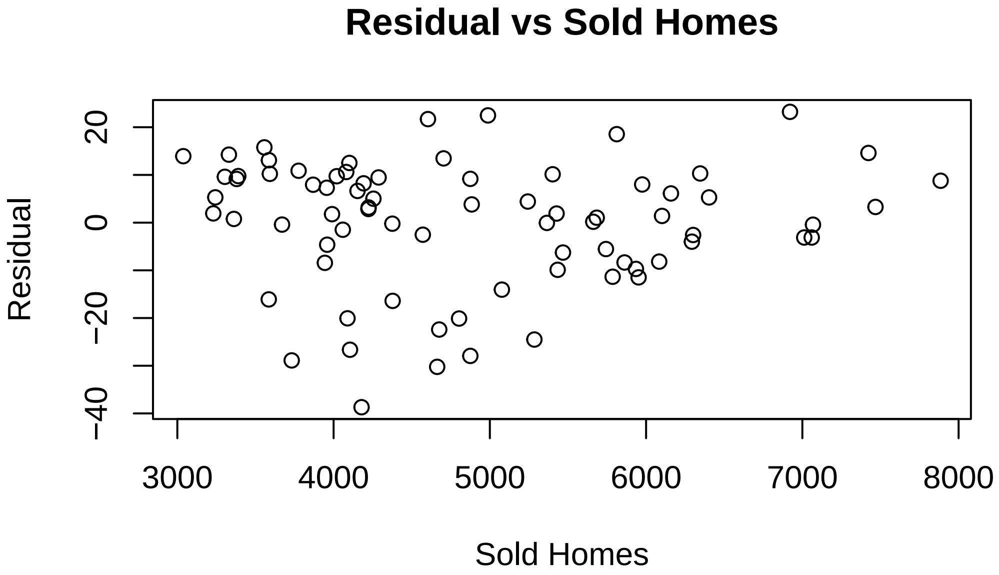
8 Assumptions of this model

```
# replace NULL with appropriate values to be used in y and x axes of plot respectively.
y_value <- resid(model)
x_value <- fitted(model)

# plots the residual vs fitted plot
plot(x = x_value, y = y_value, main="Residual vs Fitted", xlab="Fitted",
     ylab="Residuals")
```

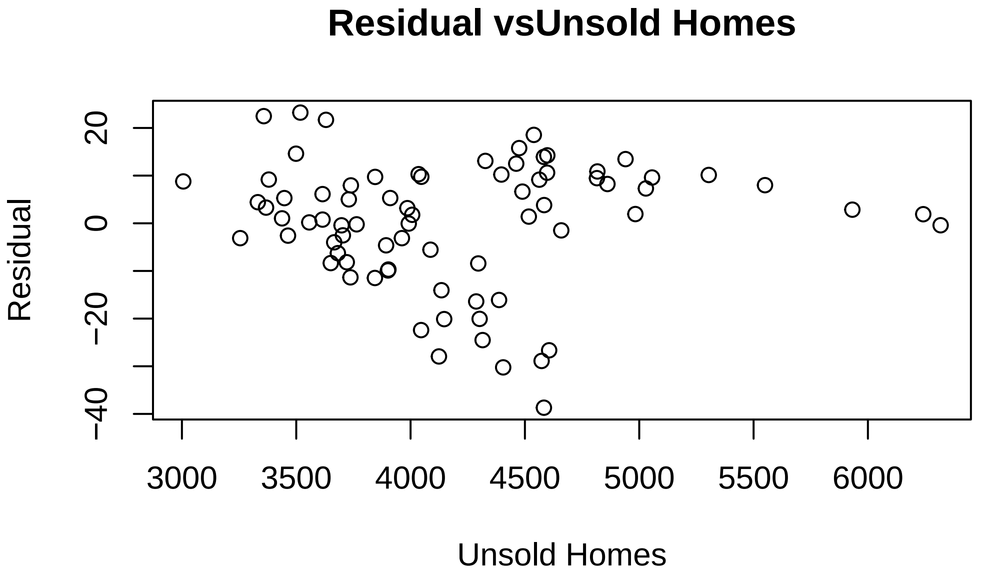


9 Residuals vs each predictor

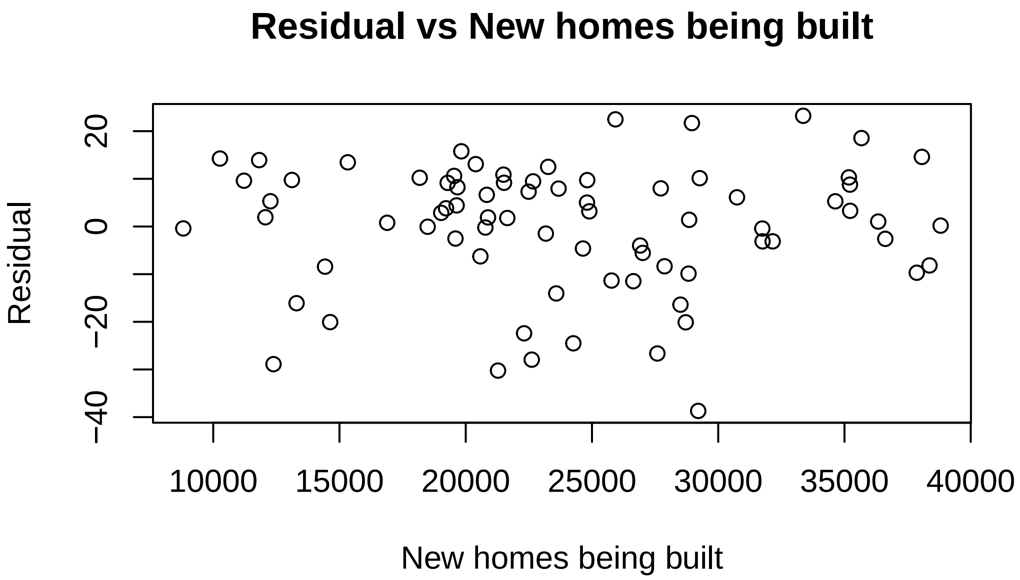
```
# Plot multiple graphs in a single grid (1 row, 3 columns)
par(mfrow=c(1,1))

# Plot Residual vs Sold Homes
plot(x = data$Detached_Absorption_Quarterly_Avg, y = y_value, main="Residual vs Sold Homes",
```



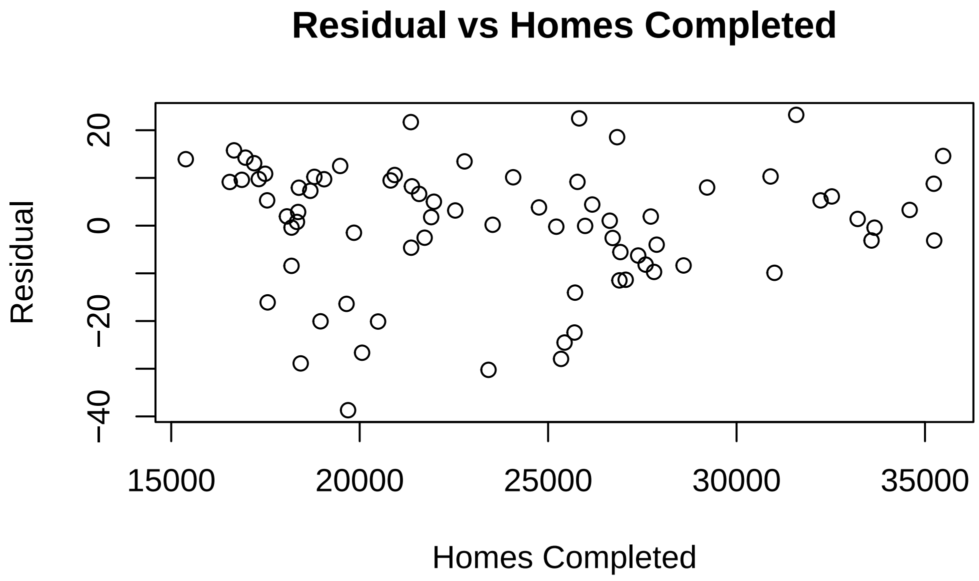
```
# Plot Residual vs Unsold Homes
plot(x = data$Detached_Unabsorbed_Quarterly_Avg, y = y_value, main="Residual vsUnsold Homes"
     xlab="Unsold Homes", ylab="Residual")
```



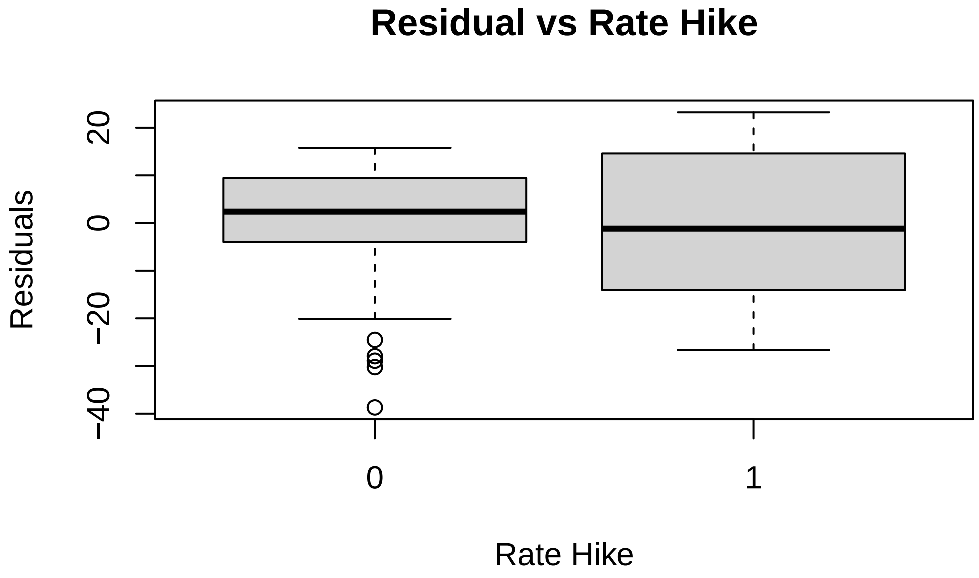
```
# Plot Residual vs New homes being built
plot(x = data$Starting_Detached_Construction, y = y_value, main="Residual vs New homes being
     xlab="New homes being built", ylab="Residual")
```



```
# Plot Residual vs Homes Completed
plot(x = data$Completed_Construction_Detached, y = y_value, main="Residual vs Homes Complete
     xlab="Homes Completed", ylab="Residual")
```

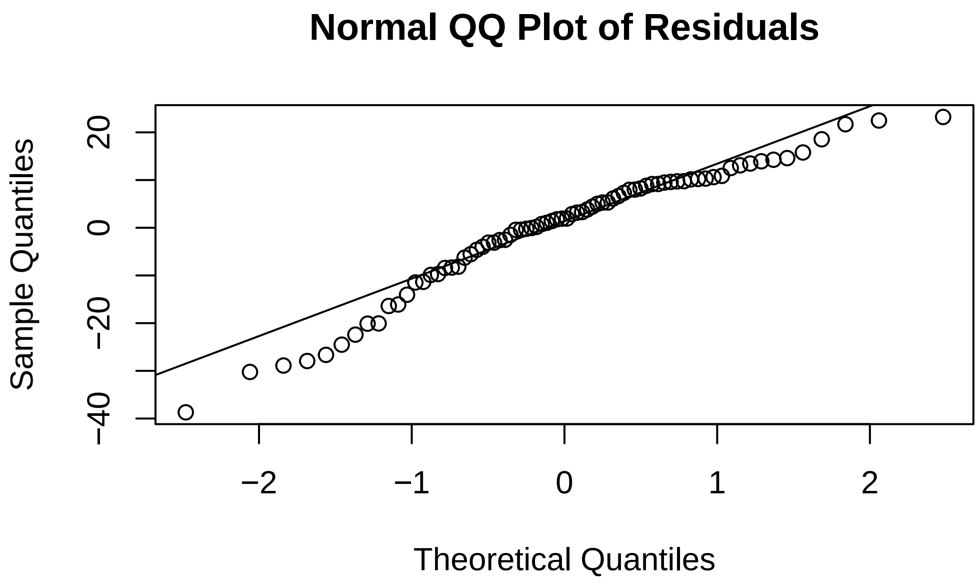


```
# Boxplot Residual vs Rate Hike
boxplot(y_value ~ data$rates, main="Residual vs Rate Hike",
        xlab="Rate Hike", ylab="Residuals")
```



10 QQ Plot

```
# Create QQ plot for residuals
qqnorm(resid(model), main = "Normal QQ Plot of Residuals")
qqline(resid(model))
```



10.1 Residual Analysis

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.