Data 603 Report: Predicting Housing Affordability

2024-11-30

1. David Griffin - 30094819
2. Golin Chen - 30088451
3. Haneen Al-Ramahi - 30085815

Introduction

Housing affordability is at the forefront of Canadians' minds, significantly younger individuals like ourselves. Our project aims to develop a predictive model for the housing affordability index using data from the Bank of Canada's "Indicators of Capacity and Inflation Pressures for Canada" (Bank of Canada, 2024). The motivation for this arises from the growing concern about housing costs and their effects on economic stability. The goals intended for this model are to aid in helping to address the housing affordability crisis in Canada effectively, which may be useful for policymakers about where to focus efforts. Additionally, our analysis also seeks to enhance understanding of housing affordability and various economic factors, shedding light on the dynamics influencing the Canadian housing market.

The applied domains of our project include:

-Economics: Multiple economic indicators to understand the housing affordability index surround inflation, wages, labour, etc.

-Real Estate & Housing: Our project is directly attempting to predict and understand the real estate market (the Housing affordability index), a ratio measuring household costs to household income.

-Public Policy: This model may aid in supporting policymakers concerning housing, as these indicators are derived from Statistics Canada and understand the relationships between these economic variables

The overall problem our project is going to address is housing affordability within Canada. With rising housing costs influenced by various complex economic factors, many individuals struggle to understand what factors drive this issue. Policymakers also encounter difficulties in identifying which factors have the most significant impact on housing affordability and how to effectively address these issues. While this project doesn't capture every variable and every complex factor, the overall intent of this project is to help tackle this by creating a data and visual analytics solution in which we develop a predictive model for the Housing Affordability Index.

This project will involve Exploratory Data Analysis (EDA) by using visualizations to examine relationships between the variables. We also hope to achieve a multiple linear regression model for predictive modelling with a high adjusted R^2, of around 90%. Our project surrounds the following research questions:

What are the key factors driving the Housing Affordability Index? Can we build a predictive model to effectively predict the Housing Affordability Index based on our selected economic indicators, and can this model provide a minimum 90% accuracy?

Tackling this problem is challenging due to the complex nature of housing prices. Housing generally involves numerous interdependent economic and policy factors that can be difficult to predict and quantify through a statistical lens. For instance, immigration and migration are believed to be significant driving forces behind housing availability and affordability, but immigration and population variables are not captured within this model (CBC News, 2024). Additionally, the issue of housing affordability varies across locations within

Canada. In dense urban cities like Toronto and Vancouver, we see high levels of investment and population growth that may contribute to the large decrease in housing affordability, In contrast, smaller cities may experience this due to a lack of housing supply or unemployment. This problem is hard to solve as it goes beyond the simple concept of supply and demand. To help address this complexity for the nature of this project, a form of simplification is necessary.

Methodology, Workflow, and Contribution

For this project we will be creating a multiple linear regression model using a variety of techniques to determine the best model for predicting the affordability index. To begin we determined several variables of interest from 3 sources and created a combined dataset for all 33 variables. We then use the stepwise selection procedure to determine our variables of interest for our model using the olsrrr package. Since we have plenty of variables we will be using fairly strict selection criteria of p_enter of 0.01 and p_remove of 0.1. Afterwards we will then perform EDA on the selected variables to better understand why they were selected.

Moving into improving and selecting the model we will start by performing tests for heteroskedasticity and multicollinearity with Breusch- Pagan and VIF tests and adjust the model accordingly. To begin determining the model we will test for interaction effects and keep any significant interactions. Next we will use the GGally package and ggpairs command to identify any higher order effects and show the correlations between the selected variables. We will then test for higher order variables in the model and keep any that are significant. Finally we will assess the model one last time and remove any variables that are unnecessary through partial F tests before selecting a final version we are satisfied with. The final model will be interpreted and explanations of some of the economic theory behind the selections for the model will be broken down. For our project we will be using a 5% level of significance as is the general standard in regression modeling.

Our workflow follows a structured sequence of tasks designed to build a robust regression model for predicting INDINF_AFFORD, with a focus on statistical rigor and interpretability. Below are the key steps in our process:

1. Data Exploration and Cleaning: Import and inspect the dataset to understand its structure and quality. Handle missing values, remove duplicates, and check for consistency in variable definitions and units.
2. Exploratory Data Analysis (EDA): Visualize the data using correlation matrices, scatterplots, and time-series graphs to identify trends and relationships between variables. Summarize key descriptive statistics for all predictors and the response variable. Challenge: Interpreting non-linear patterns and multicollinearity among predictors. We decided to use advanced visualization tools and conduct a Variance Inflation Factor (VIF) analysis to pinpoint highly collinear variables.
3. Variable Selection: Perform stepwise regression to identify a set of significant predictors based on p-value thresholds. Retain variables with high predictive power while eliminating redundant or insignificant predictors.
4. Model Building and Diagnostics: Develop multiple linear regression models, incorporating interaction and polynomial terms to address potential non-linear relationships. Validate model assumptions through residual analysis and diagnostic tests (e.g., the Breusch-Pagan test for heteroscedasticity). Challenge: We had a problem addressing heteroscedasticity or violations of normality in residuals. So we decided to consider data transformation or weighted regression if needed.
5. Model Refinement and Validation: Compare different models using ANOVA, adjusted R^2, and predictive performance metrics. Conduct cross-validation to ensure the model generalizes well to unseen data. Challenge: Overfitting due to high-order interactions or complex terms: solve things by Simplifying the model by removing overly complex terms and using regularization techniques if necessary.

Golin Chen: Golin is responsible for data management, overseeing tasks such as data cleaning and verification to ensure the dataset is accurate and ready for analysis. Golin manages the multiple regression analysis as

the analysis lead and ensures statistical methods are applied correctly. He is also the KPI manager, tracking key performance indicators and milestones to maintain project progress. Golins's analytical strengths and organizational skills make him well-suited for these roles. Additionally, as the project visualization lead, Golin is responsible for designing clear and impactful visual representations of findings.

David Griffin: David takes the lead in statistical modelling, focusing on the regression analysis and ensuring proper methodology and interpretation of results. Additionally, David acts as the project scheduler, coordinating deadlines and ensuring tasks are completed on time. His expertise in statistical methods and creative visualizations supports the team's analytical and communication goals.

Haneen Al-Ramahi: Haneen focuses on data interpretation, breaking down and explaining analysis results to ensure the findings are meaningful and accessible. As the project planner, Haneen oversees the scope and objectives, ensuring the team aligns with project goals. She also serves as the meeting scheduler, organizing team meetings and maintaining clear communication on progress. Haneen's skills in project management and effective communication make her essential for ensuring cohesion and clarity within the team.

Data

Our primary sources of data are collected from the Bank of Canada: "Indicators of Capacity and Inflation Pressures for Canada", Statistics Canada: "Housing and labour market data" And lastly, Federal Reserve Bank of St. Louis (FRED): Interest rates data. The data was systematically collected, offering reliable quarterly measurements for macroeconomic, housing, and inflation metrics.

The data represents quarterly observations, covering macroeconomic variables for Canada over several years. Potential biases may exist due to shifts in data collection methodologies or definitions over time. The dataset contains 93 observations and 34 attributes. The data spans quarterly entries from 2001Q2 and includes a mix of quantitative and qualitative variables. The attributes cover various economic indicators, housing-related metrics, and expectations data, all of which are critical to analyzing housing affordability.

We have included a document containing all 34 variables of our dataset and the individual variable explanations. We're going to use stepwise regression to remove predictors based on the p-value stepwise until there is no more variable left to enter or remove.

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers

##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.029416 -0.011571  0.000098  0.010557  0.038868
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.3341964  0.0668704  -4.998 3.02e-06 ***
## INDINF_GRACE_Q           0.0041915  0.0002331  17.981  < 2e-16 ***
## INDINF_CPI_MEDIAN_Q      0.0236400  0.0036394   6.496 5.13e-09 ***
## X10.Year.Bond            0.0233224  0.0023383   9.974 5.14e-16 ***
```

```
## INDINF_OUTGAPMPR_Q      -0.0072644  0.0018234  -3.984 0.000141 ***
## INDINF_SLSALLSEC_Q      -0.0008204  0.0002471  -3.320 0.001322 **
## INDINF_EXPECTTWOTHREE_Q  0.0854940  0.0272886   3.133 0.002366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01491 on 86 degrees of freedom
## Multiple R-squared:  0.9449, Adjusted R-squared:  0.9411
## F-statistic: 245.9 on 6 and 86 DF,  p-value: < 2.2e-16
```

This leaves us with 6 variables and our responding variable:

The independent variable: -INDINF_GRACE_Q: Foreign demand for Canadian non-commodity exports (GRACE) (2007=100) -INDINF_CPI_MEDIAN_Q: CPI-median -X10.Year.Bond: 10-year bond yield -INDINF_OUTGAPMPR_Q: Current MPR output gap (%) -INDINF_SLSALLSEC_Q: Labour shortage (% firms, Business Outlook Survey) -INDINF_EXPECTTWOTHREE_Q: Percentage of firms expecting price increases over the next six months of: 3% or less

Responding variable -INDINF_AFFORD: Housing affordability index

Summary Stats:

```
##  pro.INDINF_AFFORD pro.INDINF_GRACE_Q pro.INDINF_CPI_MEDIAN_Q pro.X10.Year.Bond
##  Min.   :0.2730    Min.   : 74.00     Min.   :1.500           Min.   :0.5465
##  1st Qu.:0.3130    1st Qu.: 87.90     1st Qu.:1.800           1st Qu.:1.8578
##  Median :0.3300    Median : 98.90     Median :2.000           Median :3.0054
##  Mean   :0.3515    Mean   : 99.09     Mean   :2.196           Mean   :2.9677
##  3rd Qu.:0.3660    3rd Qu.:111.20     3rd Qu.:2.300           3rd Qu.:4.0650
##  Max.   :0.5410    Max.   :122.70     Max.   :5.300           Max.   :5.7203
##  pro.INDINF_OUTGAPMPR_Q pro.INDINF_SLSALLSEC_Q pro.INDINF_EXPECTTWOTHREE_Q
##  Min.   :-5.90000       Min.   : 7.00          Min.   :1.900
##  1st Qu.:-0.70000       1st Qu.:22.00          1st Qu.:2.000
##  Median : 0.30000       Median :30.00          Median :2.000
##  Mean   : 0.02796       Mean   :29.11          Mean   :2.028
##  3rd Qu.: 1.00000       3rd Qu.:36.00          3rd Qu.:2.100
##  Max.   : 2.20000       Max.   :50.00          Max.   :2.200
```

Investigating the Housing Affordability Index is particularly relevant for understanding how economic pressures, like rising mortgage rates or stagnant wage growth, impact housing markets and financial security. This makes it a key variable for your project's focus on addressing Canada's housing affordability crisis.

$$C = \frac{r}{(1-(1+r)^{-N})} * M_0 + U$$

M0 is the total mortgage value, U is the utility fees, r is a weighted average mortgage rate, and N is the number of mortgage payments. For a full in depth explanation the reference to the bank of Canada will lead to a full in depth explanation of the calculation and assumptions made.

The model above is the formula for the Housing Affordability Index. It is important to understand that Affordability Index is composed of different elements, and while making the regression model and breaking down the analysis different variables might have different interactions with elements within the model, something that the group will take into when doing the project.

For the basic EDA, we graphed all 7 variables over time to look at the general trends and to see if they generally correlate with the Housing Affordability Index. Trying to justify variables based on statistical significance and their theoretical relevance to housing affordability. The graphs and their code will be

found in the appendix at figures 1-7. From the graphs, we see some variables that have similar trends with INDINF_AFFORD, however there we don't see a direct correlation with the responding variables

The models choices of variables are interesting, while most of the variables are leading inflation indicators or in the case of CPI a record of inflation itself the stepwise selection model has chosen the output gap, median cpi, demand for exports, bond yield, and expectations for inflation. While housing affordability is very complicated with a variety of factors affecting both the supply and demand sides of the equation, often even local factors, these variables appear to have a decently good likelihood of predicting next period's affordability index value. The economic rationale of this is likely through the effect of how mortgage rates affect affordability and how inflation is one of the direct predictors of central bank interest rate changes, these changes which in turn influences how mortgage rates rise and fall alongside long term bond yields which was also selected.

Results

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.029416 -0.011571  0.000098  0.010557  0.038868
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.3341964  0.0668704  -4.998 3.02e-06 ***
## INDINF_GRACE_Q           0.0041915  0.0002331  17.981  < 2e-16 ***
## INDINF_CPI_MEDIAN_Q      0.0236400  0.0036394   6.496 5.13e-09 ***
## X10.Year.Bond            0.0233224  0.0023383   9.974 5.14e-16 ***
## INDINF_OUTGAPMPR_Q      -0.0072644  0.0018234  -3.984 0.000141 ***
## INDINF_SLSALLSEC_Q      -0.0008204  0.0002471  -3.320 0.001322 **
## INDINF_EXPECTTWOTHREE_Q  0.0854940  0.0272886   3.133 0.002366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01491 on 86 degrees of freedom
## Multiple R-squared:  0.9449, Adjusted R-squared:  0.9411
## F-statistic: 245.9 on 6 and 86 DF,  p-value: < 2.2e-16
```

To start, after our selection of important variables through stepwise selection we have a base model to begin our analysis. To begin, the intercept in this model has a negative value but is significant. For the simplicity of this analysis when I state that a coefficient is significant I mean that we can reject the null hypothesis stated above and accept the alternative hypothesis. For our base model we can see that it has an adjusted R squared of 0.9411 with an RSE of 0.01491.

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

```
## Warning: package 'lmtest' was built under R version 4.3.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.3.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
##  studentized Breusch-Pagan test
##
## data:  step$model
## BP = 5.0392, df = 6, p-value = 0.5388
```

```
##
## Call:
## imcdiag(mod = step$model, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                             VIF detection
## INDINF_GRACE_Q          4.5462         0
## INDINF_CPI_MEDIAN_Q     3.4711         0
## X10.Year.Bond           4.0677         0
## INDINF_OUTGAPMPR_Q      2.7377         0
## INDINF_SLSALLSEC_Q      2.2770         0
## INDINF_EXPECTTWOTHREE_Q 1.8334         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

H_0:heteroscedasticity is not present

H_1:heteroscedasticity is present

For this base model we will perform tests for heteroskedasticity and multicollinearity. From our breusch-pagan test we can see that since the p value is greater than the significance level of 0.05 we can not reject the null hypothesis and conclude that there is no heteroskedasticity in this initial model. Additionally, we can see from our test for multicollinearity that there is no multicollinearity in this initial model.

```
##
## Call:
## lm(formula = INDINF_AFFORD ~ (INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q +
##     X10.Year.Bond + INDINF_OUTGAPMPR_Q + INDINF_SLSALLSEC_Q +
##     INDINF_EXPECTTWOTHREE_Q)^2, data = pro)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0253150 -0.0063028 -0.0005258  0.0056815  0.0279351
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                               -1.160e-01  7.133e-01  -0.163
## INDINF_GRACE_Q                             7.566e-04  5.735e-03   0.132
## INDINF_CPI_MEDIAN_Q                        1.186e-01  7.009e-02   1.692
## X10.Year.Bond                             -3.699e-02  6.722e-02  -0.550
## INDINF_OUTGAPMPR_Q                        -1.538e-02  6.967e-02  -0.221
## INDINF_SLSALLSEC_Q                        -4.094e-03  1.055e-02  -0.388
## INDINF_EXPECTTWOTHREE_Q                    1.273e-01  3.441e-01   0.370
## INDINF_GRACE_Q:INDINF_CPI_MEDIAN_Q        -5.896e-05  2.633e-04  -0.224
## INDINF_GRACE_Q:X10.Year.Bond               6.502e-04  1.331e-04   4.886
## INDINF_GRACE_Q:INDINF_OUTGAPMPR_Q         -2.308e-04  1.996e-04  -1.157
## INDINF_GRACE_Q:INDINF_SLSALLSEC_Q          2.972e-05  3.445e-05   0.863
## INDINF_GRACE_Q:INDINF_EXPECTTWOTHREE_Q     1.337e-04  2.841e-03   0.047
## INDINF_CPI_MEDIAN_Q:X10.Year.Bond          2.949e-03  4.161e-03   0.709
## INDINF_CPI_MEDIAN_Q:INDINF_OUTGAPMPR_Q    -1.239e-03  4.726e-03  -0.262
## INDINF_CPI_MEDIAN_Q:INDINF_SLSALLSEC_Q    -2.066e-04  6.333e-04  -0.326
## INDINF_CPI_MEDIAN_Q:INDINF_EXPECTTWOTHREE_Q -4.237e-02  3.168e-02  -1.337
## X10.Year.Bond:INDINF_OUTGAPMPR_Q          -3.062e-03  1.598e-03  -1.916
## X10.Year.Bond:INDINF_SLSALLSEC_Q           1.146e-04  3.345e-04   0.343
## X10.Year.Bond:INDINF_EXPECTTWOTHREE_Q     -7.392e-03  3.304e-02  -0.224
## INDINF_OUTGAPMPR_Q:INDINF_SLSALLSEC_Q      1.440e-04  1.799e-04   0.801
## INDINF_OUTGAPMPR_Q:INDINF_EXPECTTWOTHREE_Q  1.686e-02  2.964e-02   0.569
## INDINF_SLSALLSEC_Q:INDINF_EXPECTTWOTHREE_Q  6.796e-04  4.233e-03   0.161
##                                           Pr(>|t|)
## (Intercept)                                 0.8712
## INDINF_GRACE_Q                              0.8954
## INDINF_CPI_MEDIAN_Q                         0.0951 .
## X10.Year.Bond                               0.5839
## INDINF_OUTGAPMPR_Q                          0.8259
## INDINF_SLSALLSEC_Q                          0.6991
## INDINF_EXPECTTWOTHREE_Q                     0.7126
## INDINF_GRACE_Q:INDINF_CPI_MEDIAN_Q          0.8235
## INDINF_GRACE_Q:X10.Year.Bond               6.16e-06 ***
## INDINF_GRACE_Q:INDINF_OUTGAPMPR_Q           0.2513
## INDINF_GRACE_Q:INDINF_SLSALLSEC_Q           0.3912
## INDINF_GRACE_Q:INDINF_EXPECTTWOTHREE_Q      0.9626
## INDINF_CPI_MEDIAN_Q:X10.Year.Bond           0.4808
## INDINF_CPI_MEDIAN_Q:INDINF_OUTGAPMPR_Q      0.7940
## INDINF_CPI_MEDIAN_Q:INDINF_SLSALLSEC_Q      0.7452
## INDINF_CPI_MEDIAN_Q:INDINF_EXPECTTWOTHREE_Q 0.1854
## X10.Year.Bond:INDINF_OUTGAPMPR_Q            0.0594 .
## X10.Year.Bond:INDINF_SLSALLSEC_Q            0.7329
## X10.Year.Bond:INDINF_EXPECTTWOTHREE_Q       0.8236
## INDINF_OUTGAPMPR_Q:INDINF_SLSALLSEC_Q       0.4261
## INDINF_OUTGAPMPR_Q:INDINF_EXPECTTWOTHREE_Q  0.5712
## INDINF_SLSALLSEC_Q:INDINF_EXPECTTWOTHREE_Q  0.8729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0111 on 71 degrees of freedom
```

```
## Multiple R-squared:  0.9748, Adjusted R-squared:  0.9673
## F-statistic: 130.7 on 21 and 71 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q +
##     X10.Year.Bond + INDINF_OUTGAPMPR_Q + INDINF_SLSALLSEC_Q +
##     INDINF_EXPECTTWOTHREE_Q + INDINF_GRACE_Q:X10.Year.Bond, data = pro)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032232 -0.007046 -0.000042  0.006584  0.027570
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -6.602e-02  6.188e-02  -1.067  0.28905
## INDINF_GRACE_Q                1.933e-03  3.411e-04   5.666 1.95e-07 ***
## INDINF_CPI_MEDIAN_Q           2.388e-02  2.797e-03   8.535 4.71e-13 ***
## X10.Year.Bond                -4.084e-02  8.437e-03  -4.840 5.74e-06 ***
## INDINF_OUTGAPMPR_Q           -8.846e-03  1.416e-03  -6.246 1.60e-08 ***
## INDINF_SLSALLSEC_Q           -2.954e-04  2.015e-04  -1.466  0.14644
## INDINF_EXPECTTWOTHREE_Q       5.891e-02  2.125e-02   2.772  0.00684 **
## INDINF_GRACE_Q:X10.Year.Bond  6.510e-04  8.365e-05   7.783 1.54e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01146 on 85 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9652
## F-statistic: 365.4 on 7 and 85 DF,  p-value: < 2.2e-16
```

In this section we tested for the presence of interaction effects between variables and we can see that the vast majority of the interactions are not significant and we will discard them. The one interaction that we will keep for future models will be the grace and bond yield interaction due to its significance. This interaction having a positive coefficient is surprising, long term bond yields and key interest rates show a positive relationship with the responding variable and given that economic theory states that higher interest rates raise the value of a countries currency and thus reduce exports. This theory is supported by their strong negative correlation in the ggpairs plot in the appendix at figure 8. Trying the model with just that interaction and all but one variable are significant.

```
##
## Call:
## lm(formula = INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q +
##     I(INDINF_CPI_MEDIAN_Q^2) + X10.Year.Bond + INDINF_OUTGAPMPR_Q +
##     I(INDINF_OUTGAPMPR_Q^2) + INDINF_SLSALLSEC_Q + INDINF_EXPECTTWOTHREE_Q +
##     INDINF_GRACE_Q:X10.Year.Bond, data = pro)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.025376 -0.006124 -0.000561  0.007098  0.032290
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.788e-02  5.637e-02  -0.317 0.751868
```

```
## INDINF_GRACE_Q                1.261e-03  3.356e-04   3.757 0.000319 ***
## INDINF_CPI_MEDIAN_Q           5.380e-02  9.233e-03   5.828 1.04e-07 ***
## I(INDINF_CPI_MEDIAN_Q^2)     -4.514e-03  1.336e-03  -3.379 0.001109 **
## X10.Year.Bond                -5.652e-02  8.244e-03  -6.855 1.17e-09 ***
## INDINF_OUTGAPMPR_Q           -1.268e-02  1.831e-03  -6.926 8.52e-10 ***
## I(INDINF_OUTGAPMPR_Q^2)      -1.414e-03  4.022e-04  -3.516 0.000712 ***
## INDINF_SLSALLSEC_Q            9.671e-05  2.156e-04   0.449 0.654845
## INDINF_EXPECTTWOTHREE_Q       4.626e-02  1.924e-02   2.405 0.018416 *
## INDINF_GRACE_Q:X10.Year.Bond  7.958e-04  8.143e-05   9.773 1.86e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01024 on 83 degrees of freedom
## Multiple R-squared:  0.9749, Adjusted R-squared:  0.9722
## F-statistic: 358.6 on 9 and 83 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q + I(INDINF_CPI_MEDIAN_Q^2) +
##     X10.Year.Bond + INDINF_OUTGAPMPR_Q + I(INDINF_OUTGAPMPR_Q^2) +
##     INDINF_EXPECTTWOTHREE_Q + INDINF_GRACE_Q:X10.Year.Bond
## Model 2: INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q + I(INDINF_CPI_MEDIAN_Q^2) +
##     X10.Year.Bond + INDINF_OUTGAPMPR_Q + I(INDINF_OUTGAPMPR_Q^2) +
##     INDINF_SLSALLSEC_Q + INDINF_EXPECTTWOTHREE_Q + INDINF_GRACE_Q:X10.Year.Bond
##   Res.Df       RSS Df  Sum of Sq      F Pr(>F)
## 1     84 0.0087237
## 2     83 0.0087026  1 2.1106e-05 0.2013 0.6548
```

$$H_0 : \beta_7 = 0$$
$$H_1 : \beta_7 \neq 0$$

We next tested for higher order effects, in the appendix we can see that many of the variables showed potential higher order effects. As such we tested all the variables for higher order effects and found that the median CPI and the MPR output gap variables both have a higher level term. In this model we can see that SLSALLSEC is still not significant. As such we perform a partial f test to see if we would be improving the model by dropping the variable and with a p value of 0.65 we fail to reject the null hypothesis and can conclude that the reduced model is better.

```
##
## Call:
## lm(formula = INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q +
##     I(INDINF_CPI_MEDIAN_Q^2) + X10.Year.Bond + INDINF_OUTGAPMPR_Q +
##     I(INDINF_OUTGAPMPR_Q^2) + INDINF_EXPECTTWOTHREE_Q + INDINF_GRACE_Q:X10.Year.Bond,
##     data = pro)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.025014 -0.006274 -0.000667  0.006924  0.032022
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.186e-02  5.540e-02  -0.395 0.694170
```

```
## INDINF_GRACE_Q                1.319e-03  3.085e-04   4.274 5.04e-05 ***
## INDINF_CPI_MEDIAN_Q           5.409e-02  9.166e-03   5.901 7.37e-08 ***
## I(INDINF_CPI_MEDIAN_Q^2)     -4.529e-03  1.329e-03  -3.408 0.001007 **
## X10.Year.Bond                -5.487e-02  7.344e-03  -7.471 6.87e-11 ***
## INDINF_OUTGAPMPR_Q           -1.208e-02  1.250e-03  -9.666 2.71e-15 ***
## I(INDINF_OUTGAPMPR_Q^2)      -1.315e-03  3.346e-04  -3.930 0.000174 ***
## INDINF_EXPECTTWOTHREE_Q       4.633e-02  1.915e-02   2.419 0.017708 *
## INDINF_GRACE_Q:X10.Year.Bond  7.792e-04  7.216e-05  10.797  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01019 on 84 degrees of freedom
## Multiple R-squared:  0.9749, Adjusted R-squared:  0.9725
## F-statistic: 407.2 on 8 and 84 DF,  p-value: < 2.2e-16
```

The model stated above is likely the best model for predicting the housing affordability index given the dataset we put together. Interpreting the model is complicated due to its multiple higher order terms and interaction variable. What can be said is that the GRACE and CPI expectations variable have an overall positive relationship with the affordability index holding the bond yield constant. The output gap has a negative relationship with the affordability index for both of its coefficients. The bond yield and the median CPI variables have a complicated relationship with their interaction and higher order relationship variables respectively have opposite relationships with the affordability index to the normal coefficients. The selected model does explain 97.25% of the total variation and the standard deviation of the unexplained variance of the selected model is 0.01019

$$Y_A = -0.02186 + 0.001319GRACE + 0.05409CPIMEDIAN - 0.004529CPIMEDIAN^2 - 0.05487BOND - 0.01208OUTGAP$$

$$-0.001315OUTGAPMPR^2 + 0.04633EXPECTTWOTHREE + 0.0007792GRACE * BOND$$

In figure 9 in the appendix we can see the residual plots, using these we can see how our assumptions for linear regressions hold up. From the graphs we can see that the assumption of normality holds up as the Q-Q residual plot shows a relatively straight line. The assumption of independence also holds as the residuals seem to be randomly distributed.

Discussion

In this project we used stepwise selection to see what variables would be chosen to predict the housing affordability index calculated by the bank of canada. When run, the selection chose variables for export demand, CPI, long term bond yields, output gap, labor shortage expectations, and expected CPI from businesses. We then found no evidence of heteroskedasticity or multicollinearity in the selected model. We found an additional interaction effect between long term bond yields and export demand and 2 higher order terms. After dropping the variable for labor shortage expectations we arrived at our final model with an adjusted R squared of 97.25% and and RSE of 0.01019. The final model showed an overall positive relationship of export demand and cpi expectations with the affordability index, the output gap and showed a negative relationship with the affordability index. The CPI variable and long term bond yield had conflicting coefficients with their higher order and interaction coefficients respectively. Overall the economic reasoning of which variables were selected is logical but some of the coefficients and interactions do not. Regardless, the model due to its high adjusted R squared is likely a good predictor. Due to the brief nature and focus of the project we could not go into further detail and explore the models selections and coefficients further. Housing is a complex topic in economics with a simple explanation of "supply and demand" being a gross simplification. Much work has been done on the topic but this brief project could maybe inform a more in depth exploration with the hope of better understanding what is causing the housing affordability crisis we are currently experiencing.

Refrences

-CBC News. (2021, January 21). Immigration, housing crisis, and costs: Why Canada's housing crisis is deepening. CBC News. https://www.cbc.ca/news/politics/immigration-housing-crisis-costs-1.7088878

-Bank of Canada. (2024, October 23). Capacity and inflation pressures. Bank of Canada https://www.bankofcanada.ca/rates/indicators/capacity-and-inflation-pressures/

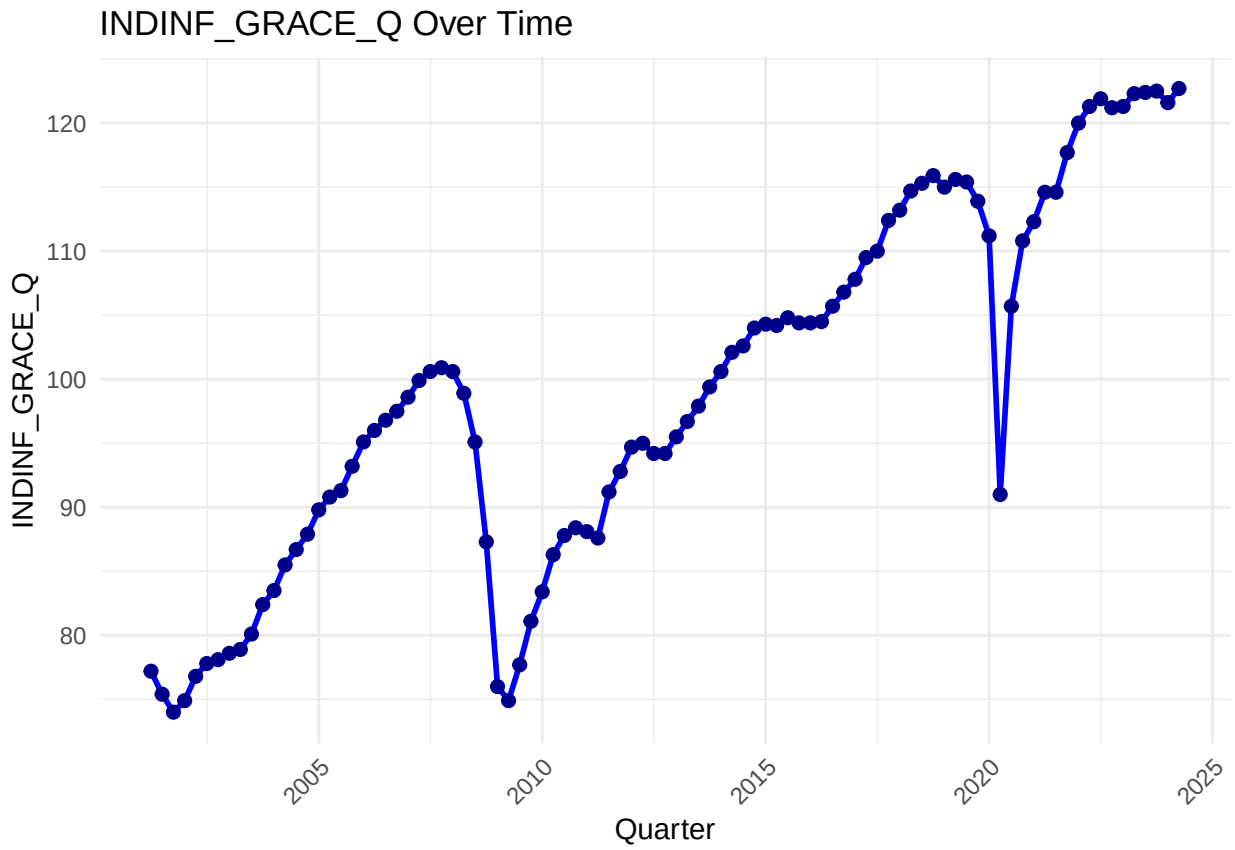
-Federal Reserve Bank of St. Louis. (n.d.). FRED economic data. Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/

-Government of Canada, Statistics Canada. (2024, October 16). Canada mortgage and Housing Corporation, housing starts, all areas, Canada and provinces, seasonally adjusted at annual rates, monthly. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410015801

Appendix

Figure 1:

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



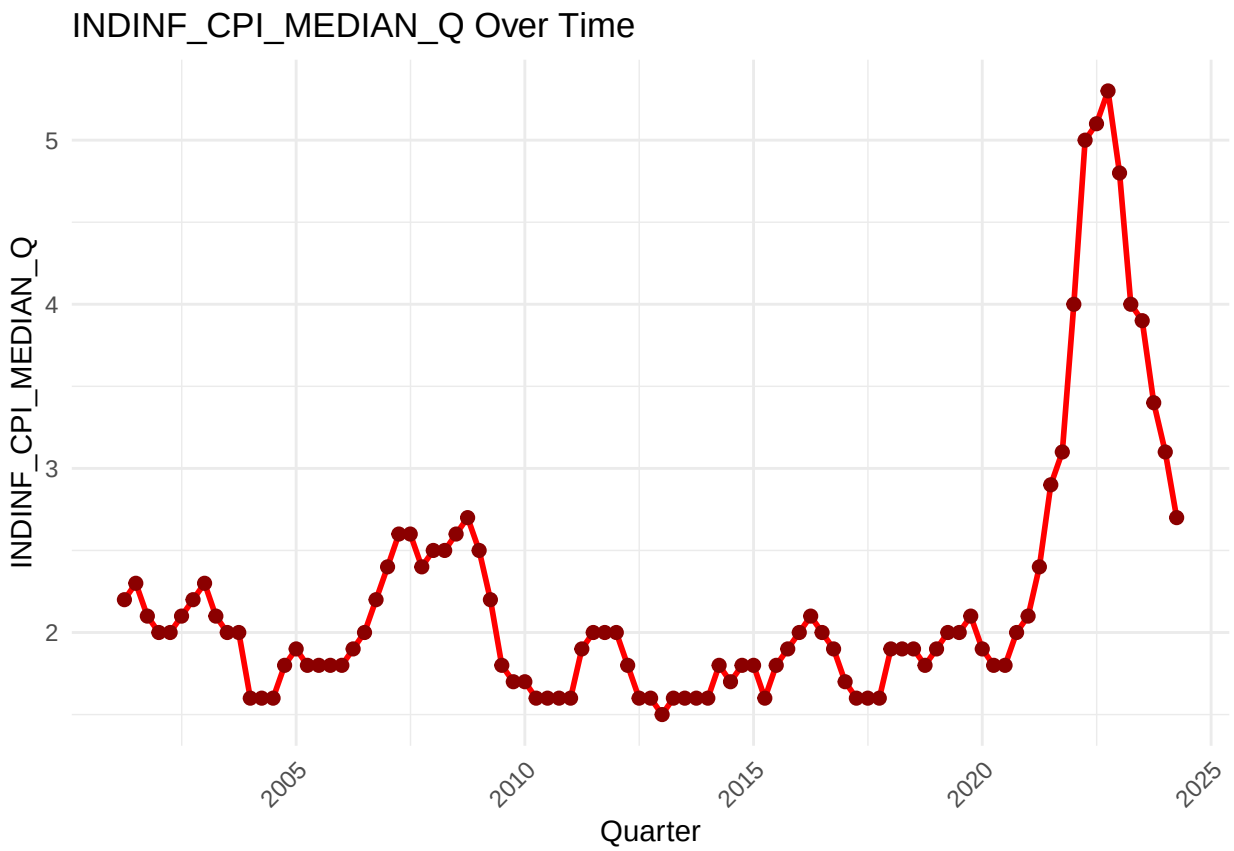

INDINF_CPI_MEDIAN_Q Over Time
INDINF_CPI_MEDIAN_Q
5
4
3
2
2005
2010
2015
2020
2025
Quarter

Figure 2:

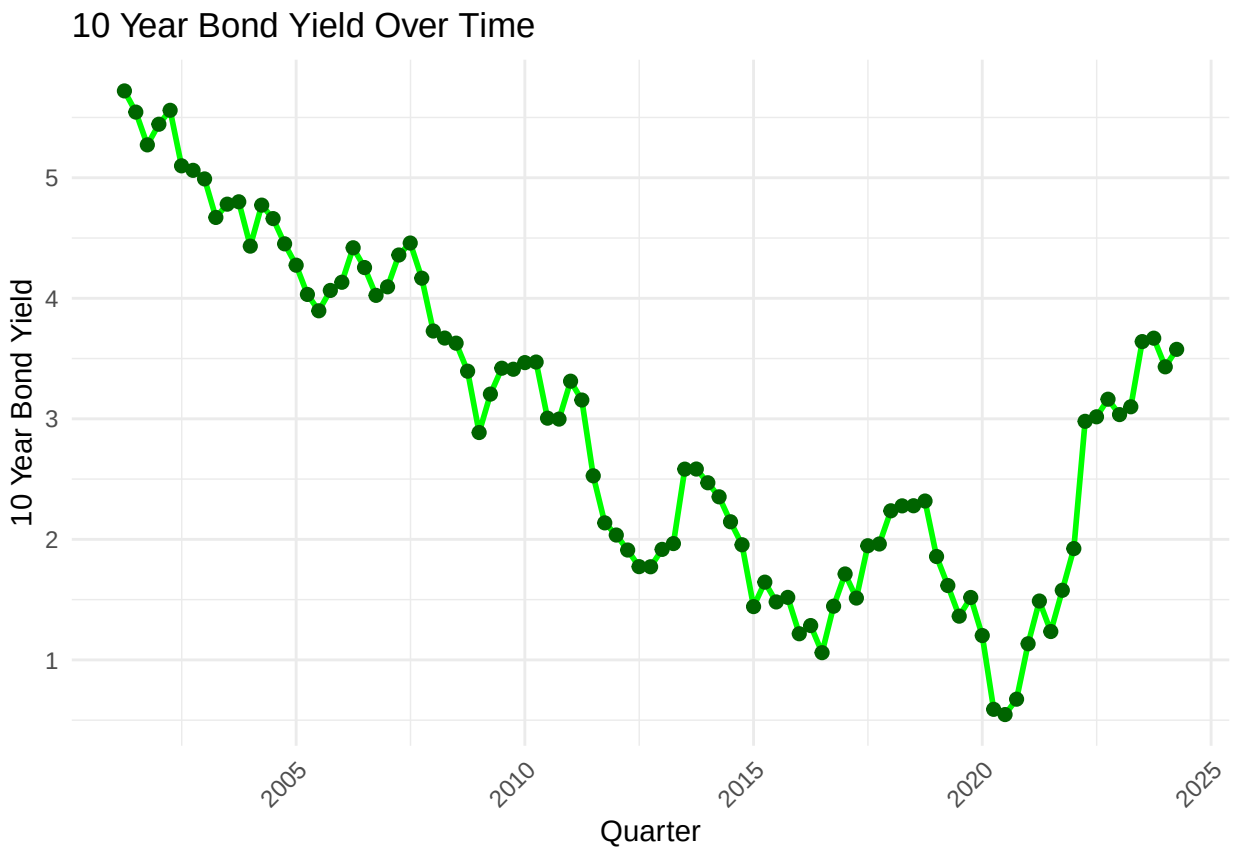

10 Year Bond Yield Over Time
10 Year Bond Yield
5
4
3
2
1
2005
2010
2015
2020
2025
Quarter

Figure 3:

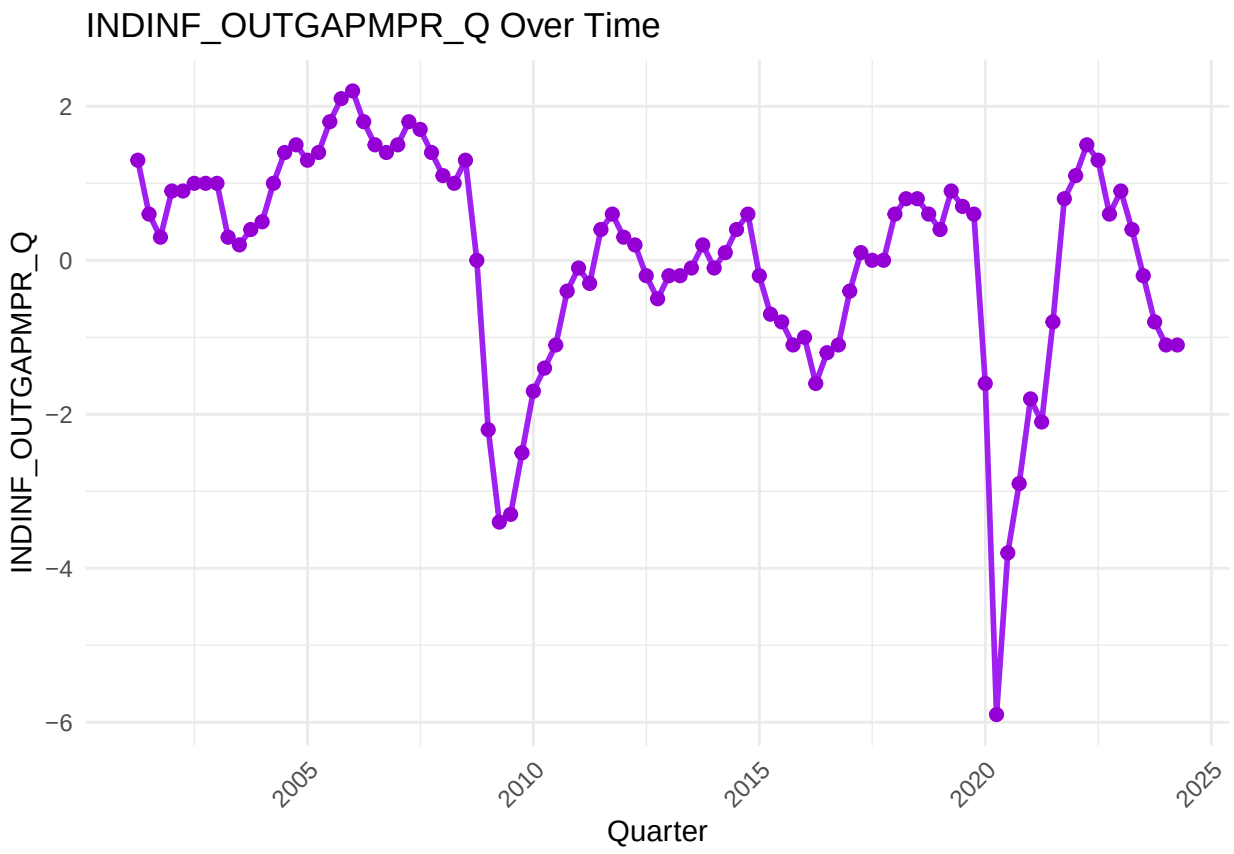

INDINF_OUTGAPMPR_Q Over Time
INDINF_OUTGAPMPR_Q
2
0
−2
−4
−6
2005
2010
2015
2020
2025
Quarter

Figure 4:

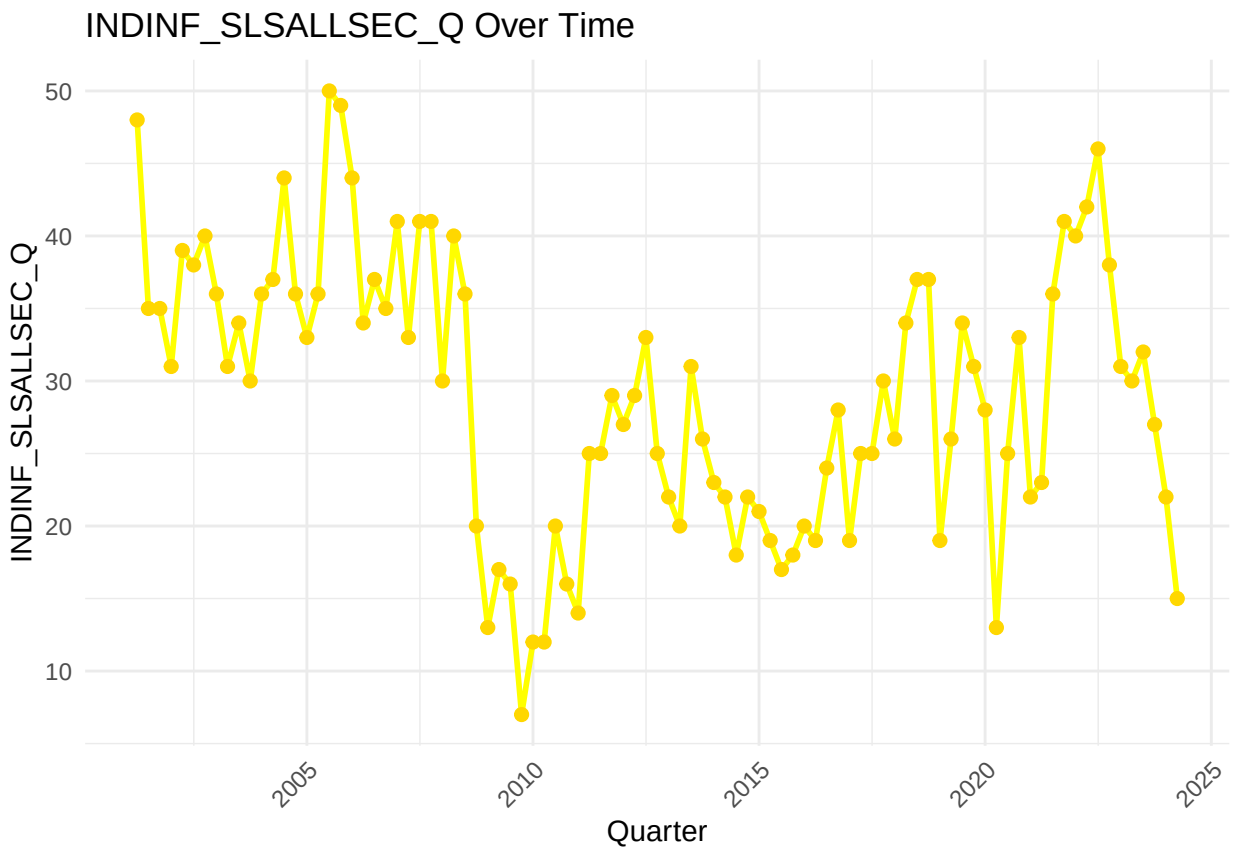

INDINF_SLSALLSEC_Q Over Time
INDINF_SLSALLSEC_Q
50
40
30
20
10
2005
2010
2015
2020
2025
Quarter

Figure 5:

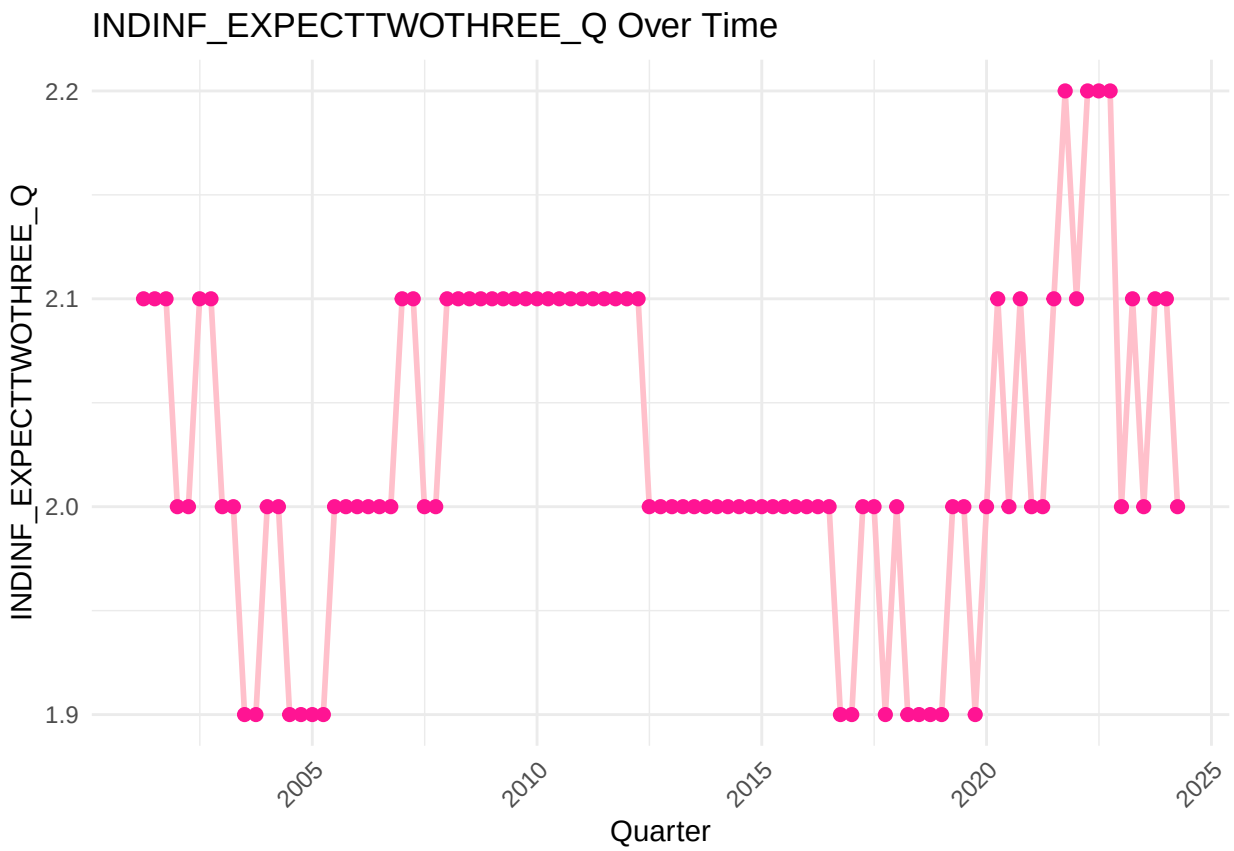

INDINF_EXPECTTWOTHREE_Q Over Time
INDINF_EXPECTTWOTHREE_Q
2.2
2.1
2.0
1.9
2005
2010
2015
2020
2025
Quarter

Figure 6:

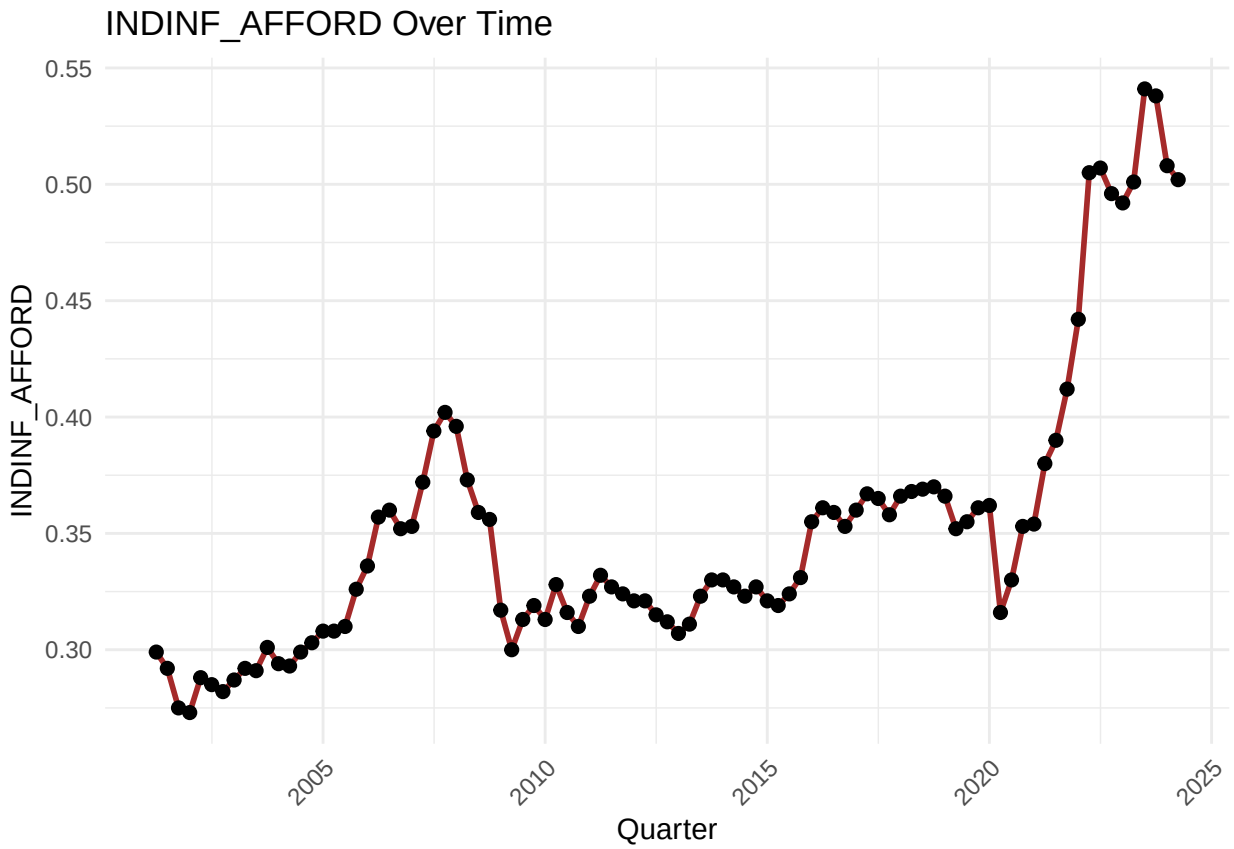


Figure 7:

Figure 8:

```
## Warning: package 'GGally' was built under R version 4.3.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

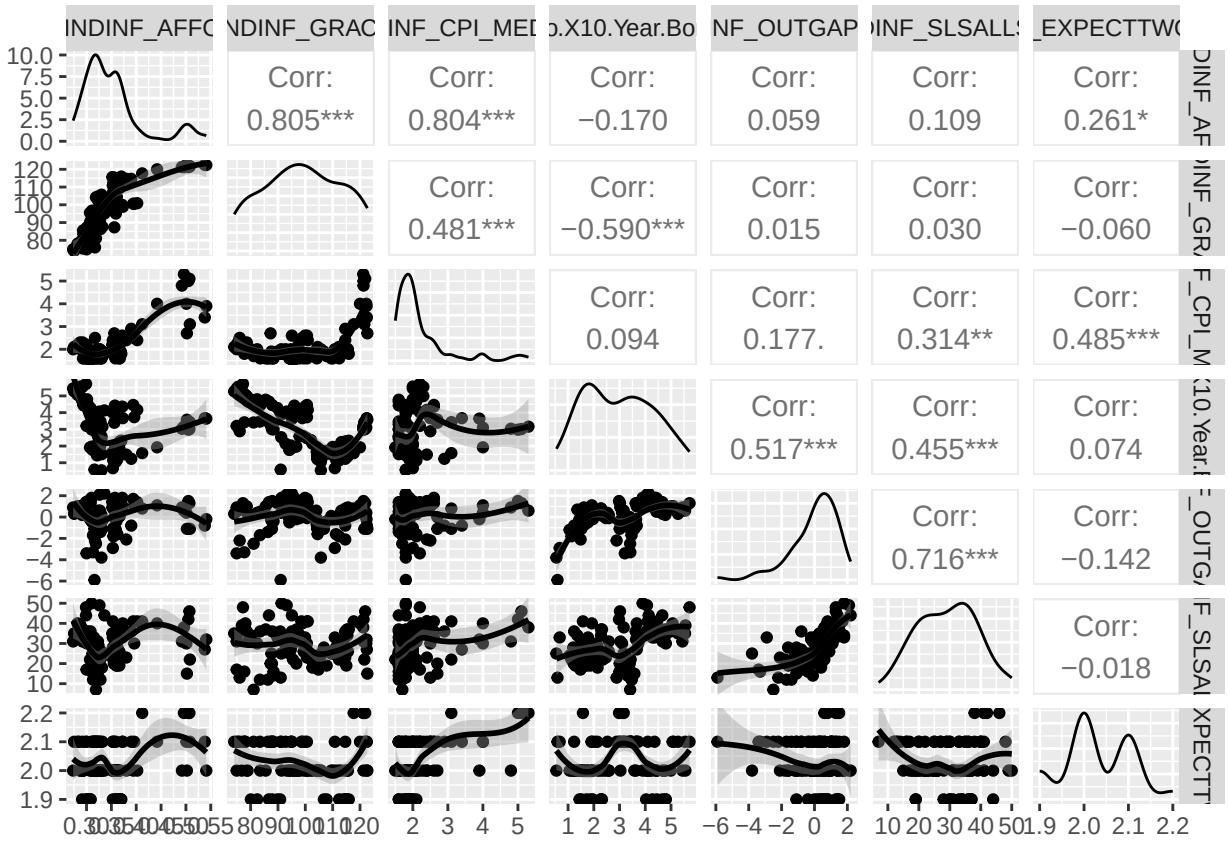
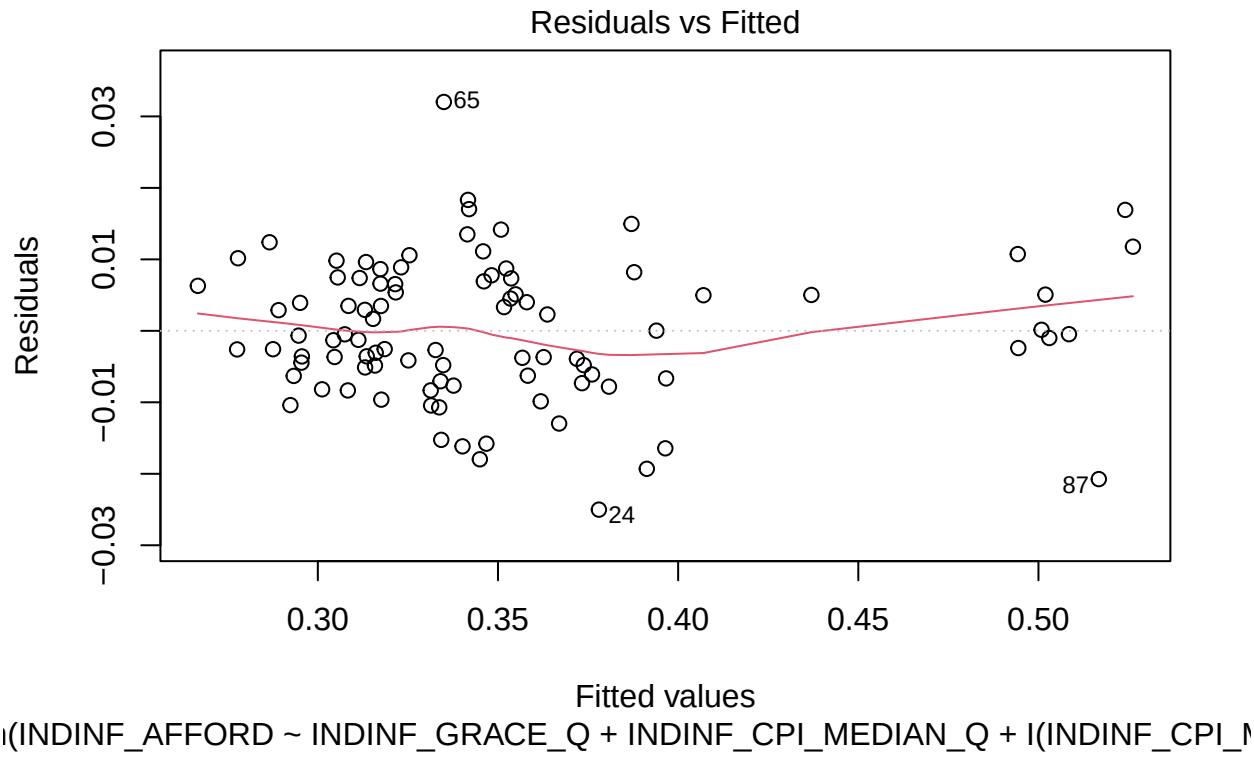
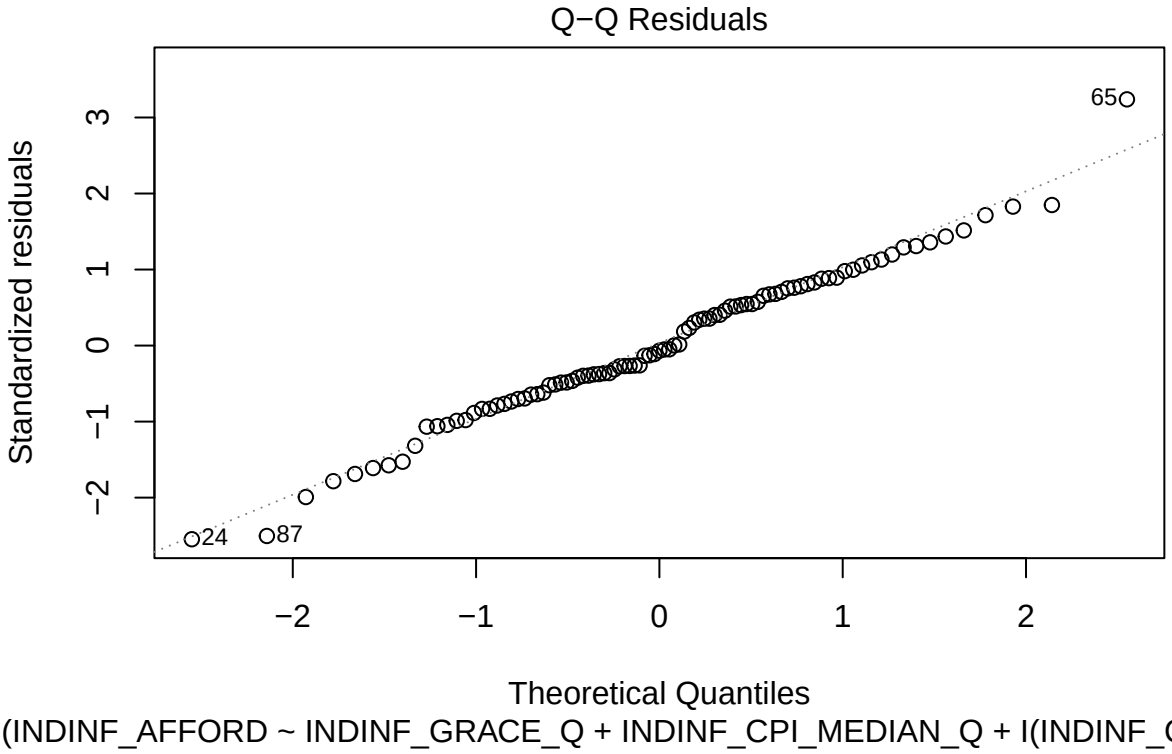


Figure 9:

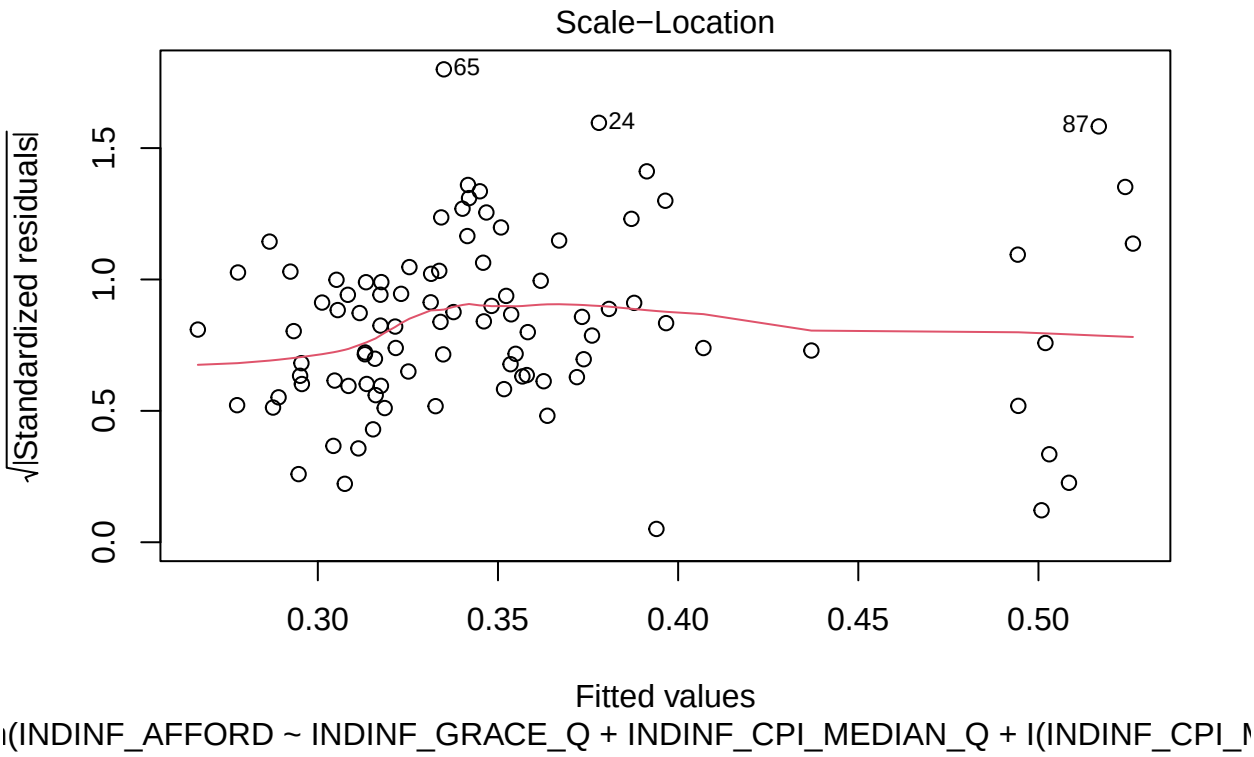
```
plot(reg5)
```

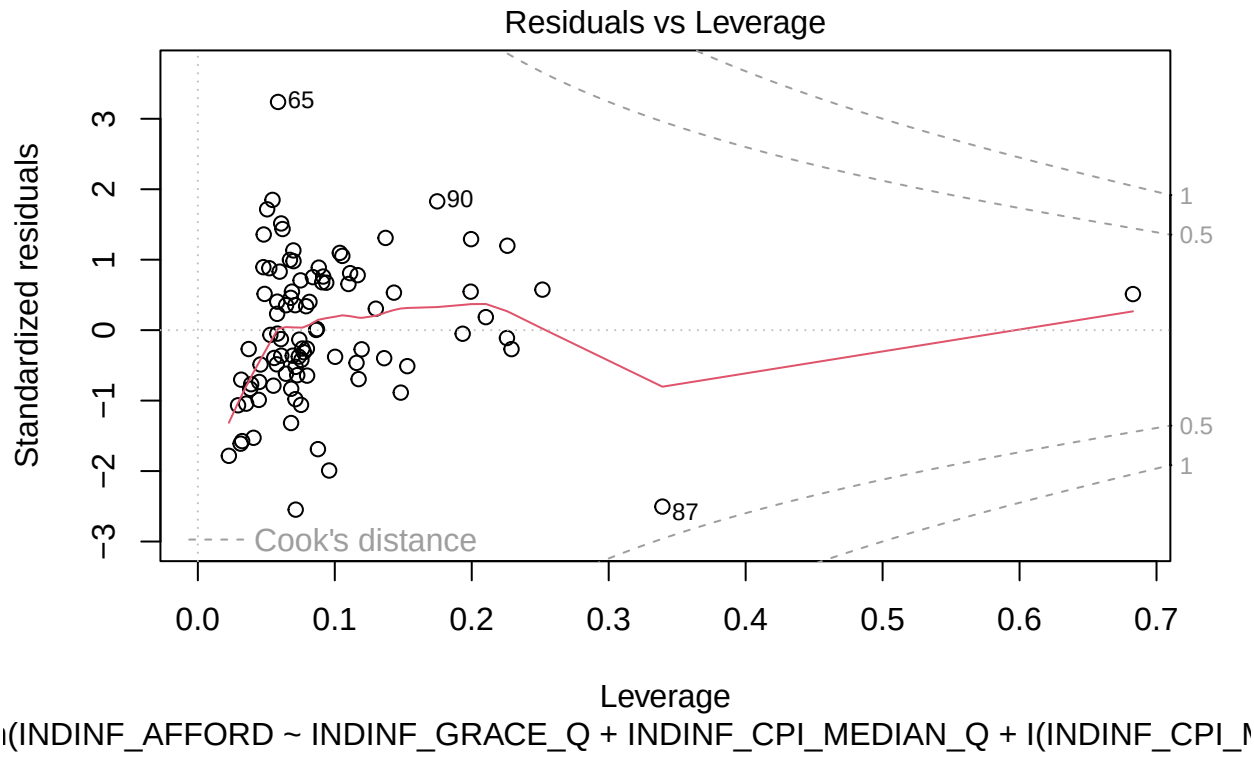


ı(INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q + I(INDINF_CPI_M



ı(INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q + I(INDINF_CPI_M


Scale−Location
65
24
87
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.30
0.35
0.40
0.45
0.50
Fitted values
ı(INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q + I(INDINF_CPI_M



I(INDINF_AFFORD ~ INDINF_GRACE_Q + INDINF_CPI_MEDIAN_Q + I(INDINF_CPI_M