Group Project

A3

2022-08-03

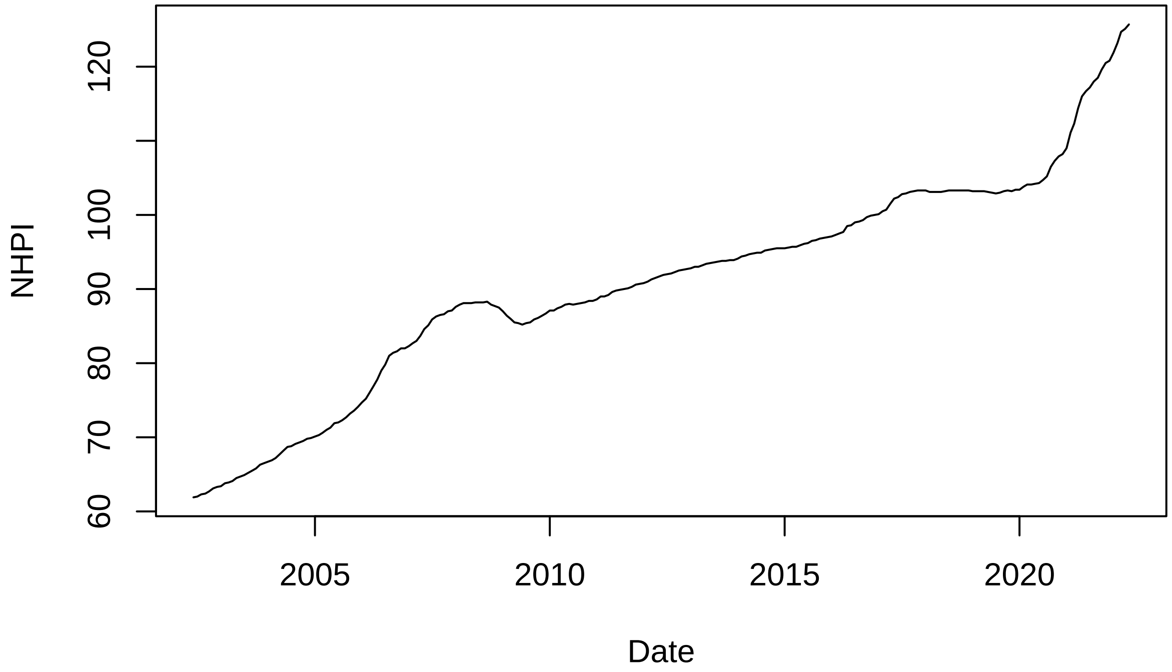
Introduction

With increasingly unaffordable housing having become a chronic social issue in Canada over the last two decades, the factors that contribute to rising housing cost have been in focus, and hotly debated among policy makers and voters alike. Here, we would like to measure the relative importance of the commonly referred contributors of rising cost of housing: interest rate, immigration, earnings increases, and general increase in consumer prices [3]. We will explore and quantify the significance of the predictors or drivers of rising housing prices. Furthermore, by better understanding possible causes and predictors of the present day problem, we aim to better inform policymakers and their electorate on the most important issues underlying the unaffordable housing markets of the metropolitan Canada.

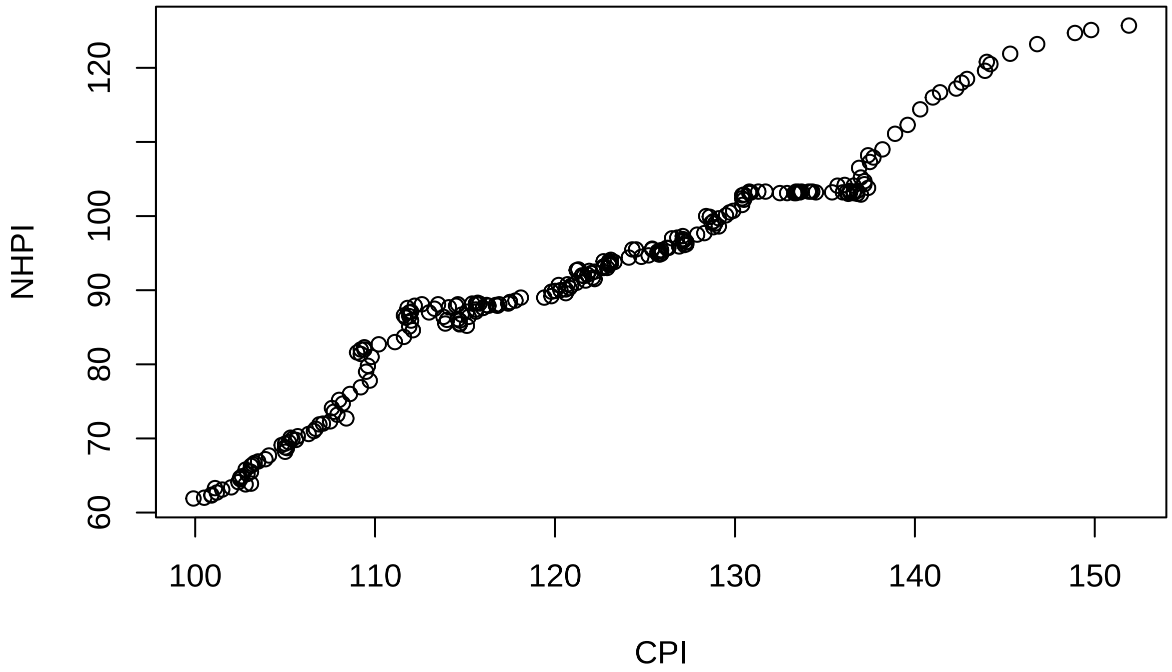
Variables

Name	Description	Unit
New Housing Price Index (y)	Monthly series that measure changes over time of the selling prices of new residential houses sold by builders in the Canadian metropolitan areas. The reference period is December 2006, for which the index value is set to 100.	-
Mortgage Rate (x_1)	Average annual mortgage lending rate for 5-year term.	%
Immigrants (x_2)	Population growth due to the total number of immigrants to Canada between the preceding two calendar years. Immigration, Refugees and Citizenship Canada does not make immigration data with higher frequency than yearly.	-
Average Weekly Earnings (x_3)	Average weekly earnings for all employees in Canada in Canadian dollar per week.	CAD/week
Consumer Price Index (x_4)	Indicator for changes in consumer prices of all goods and services experienced by Canadians. The time base is the period for which the CPI equals 100; currently this is the year 2002.	-

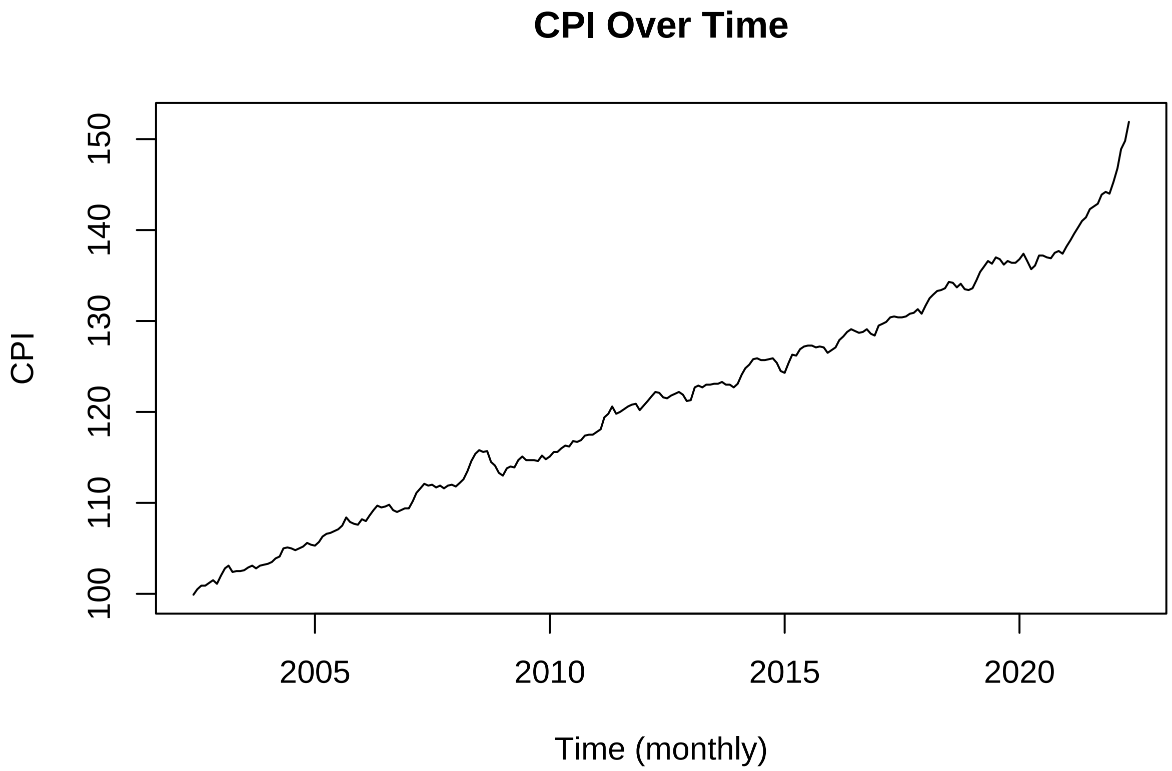
Analysis

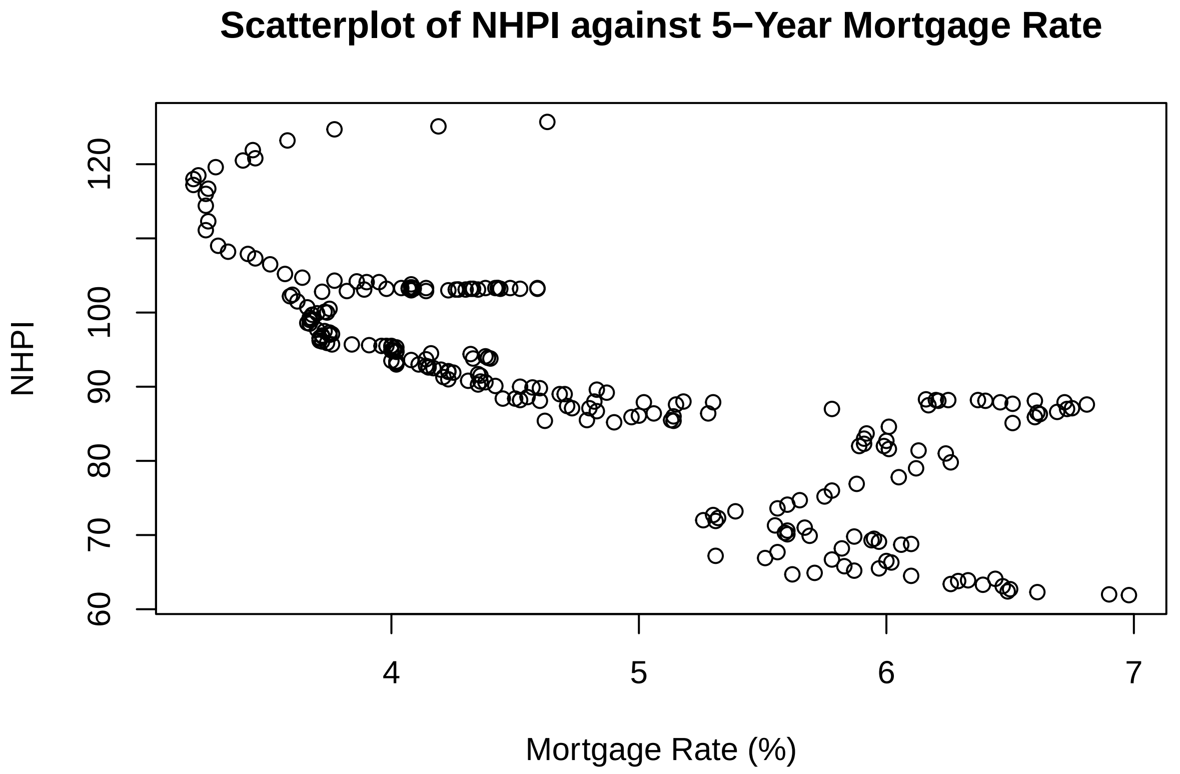


Line graph of Monthly NHPI



Scatterplot of NHPI against CPI

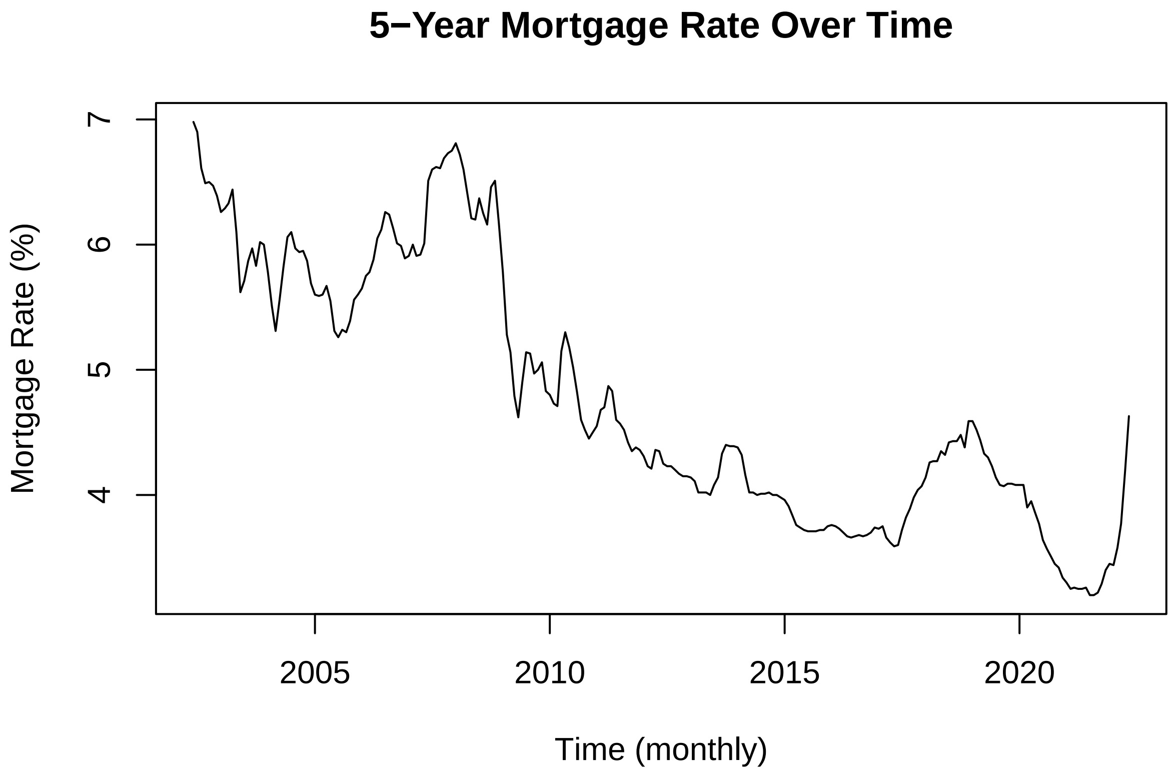

CPI Over Time
CPI
150
140
130
120
110
100
2005
2010
2015
2020
Time (monthly)



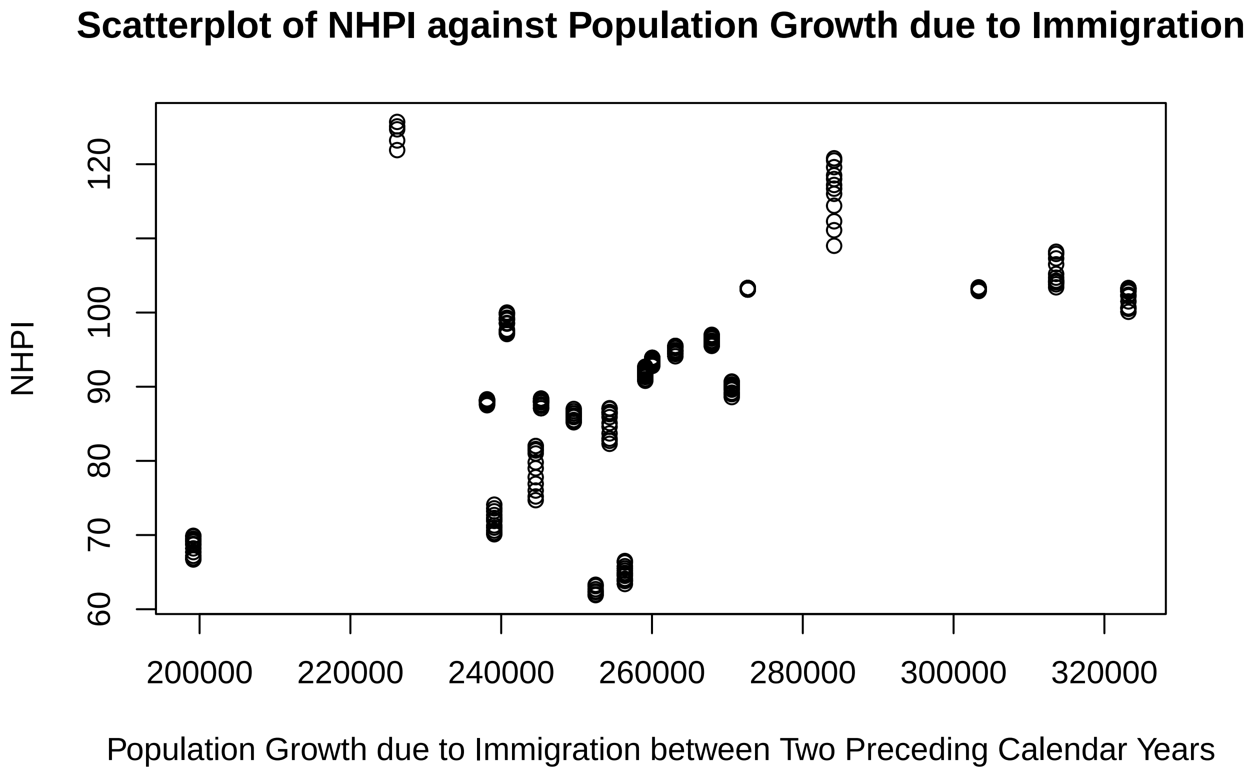
It seems like a quadratic curve centered around 7. Defend it based on the correlations against NHPI as well.

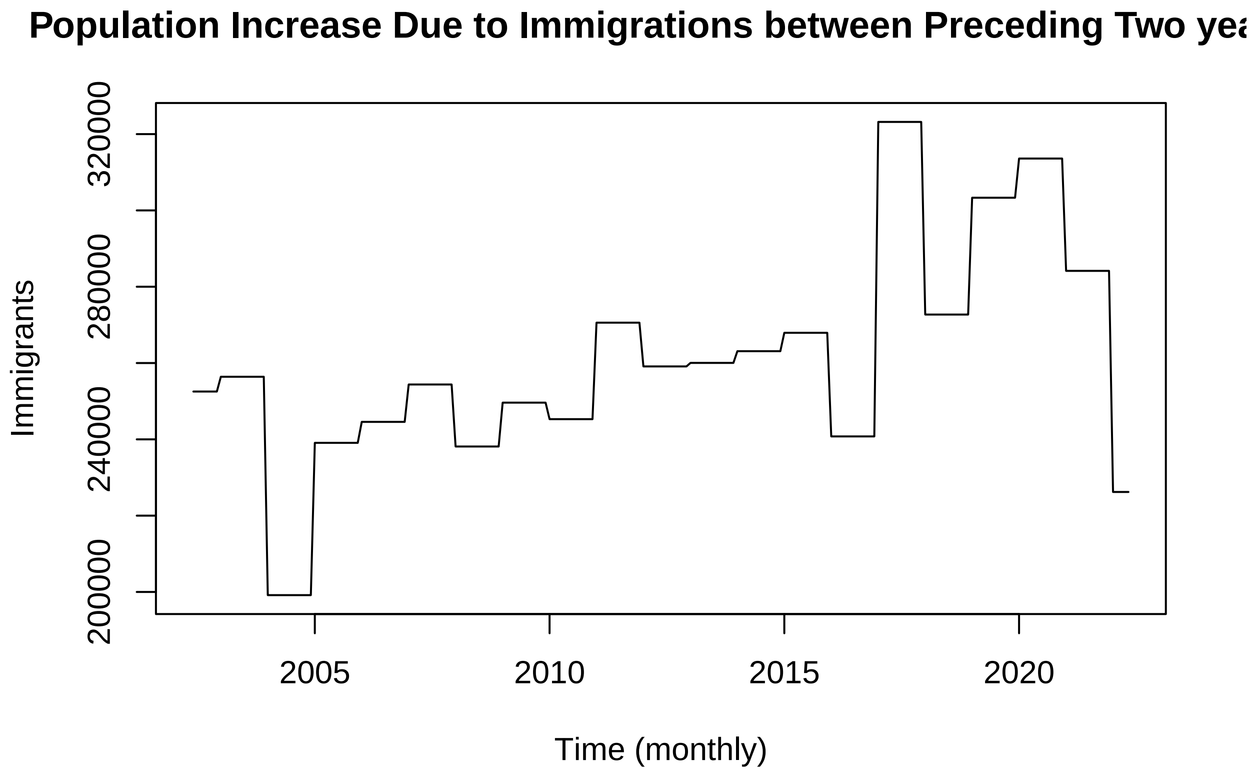
Transformation	Sample correlation (r)
Interest	-0.7978552
log(Interest)	-0.8128942
(Interest)$^-1$	0.8199765
Interest^2	-0.7762884
(Interest $-$ 7)^2	0.8253844

Table ?: Correlation between NHPI and Different Transformations of Interest

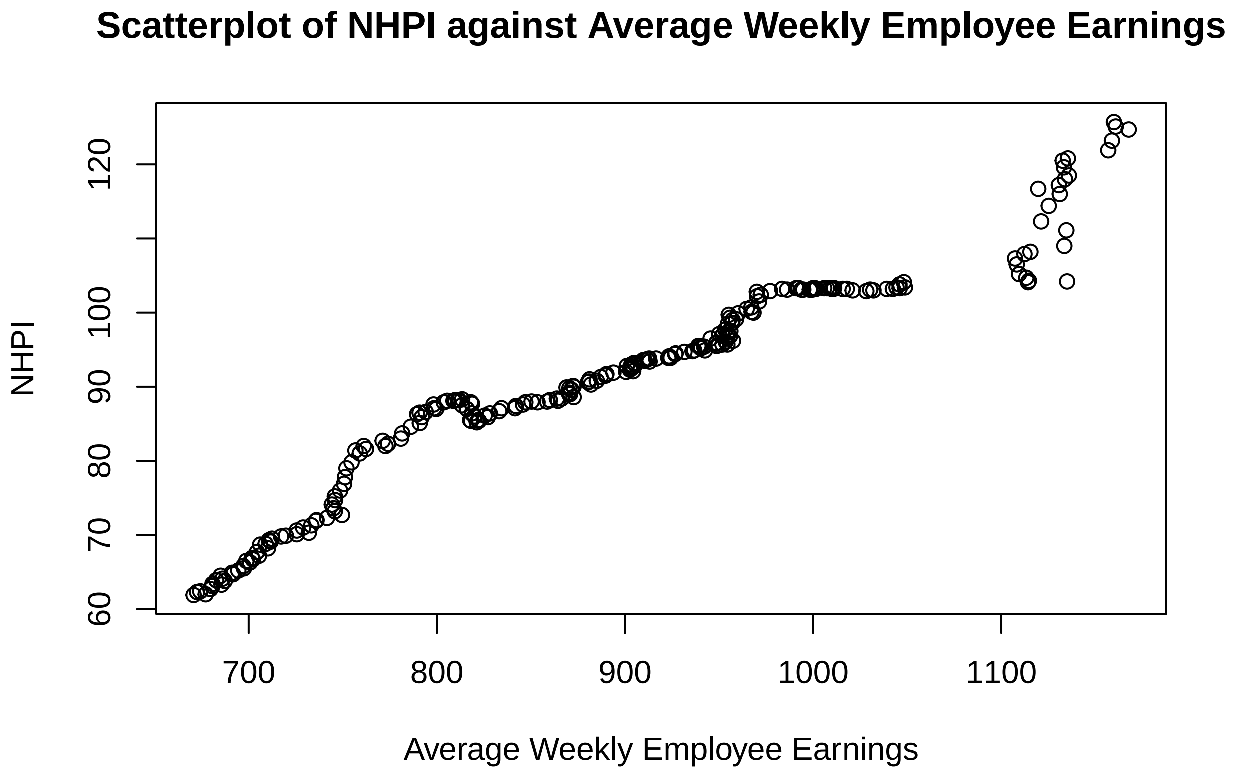


Because interest rates are mostly set by the Bank of Canada, the changes appear somewhat erratic.

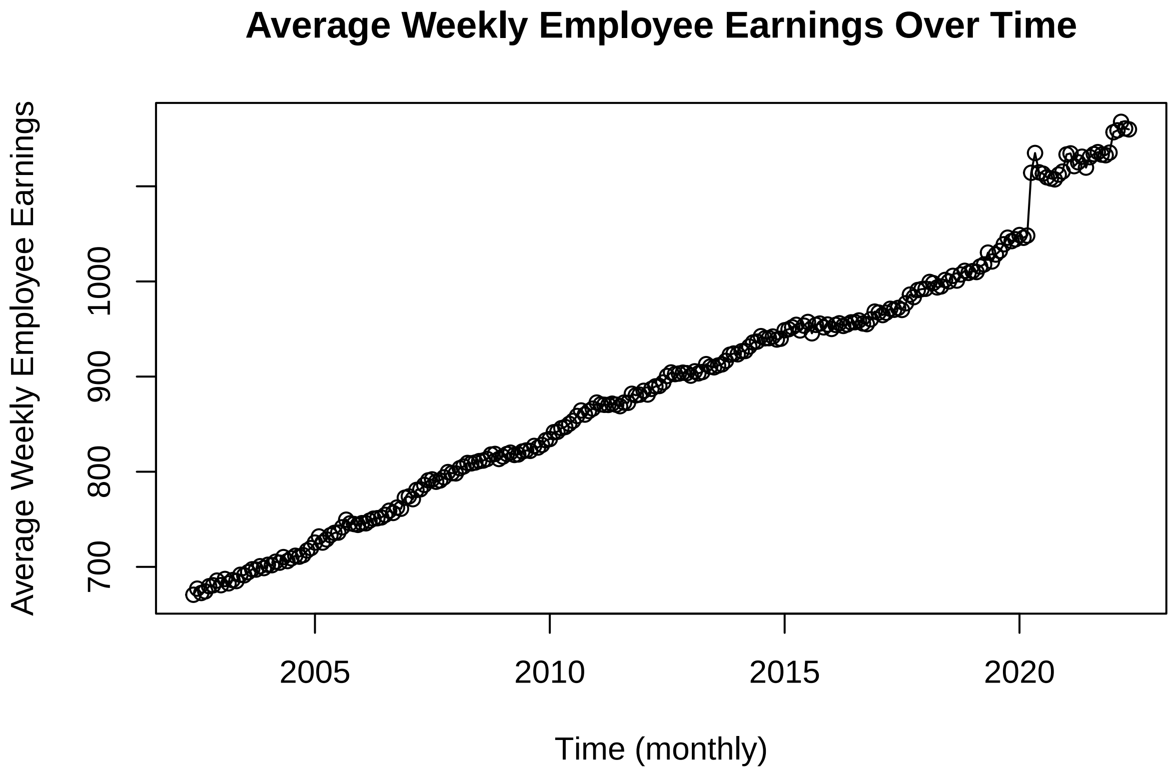

Scatterplot of NHPI against Population Growth due to Immigration
NHPI
60
70
80
90
100
120
200000
220000
240000
260000
280000
300000
320000
Population Growth due to Immigration between Two Preceding Calendar Years



Due to the lack of monthly data on immigration provided by Statistics Canada, the monthly data was imputed by applying the yearly data flatly across calendar years. The predictor variable itself doesn't have a strong linear pattern, but seems to have been generally increasing over the last two decades.

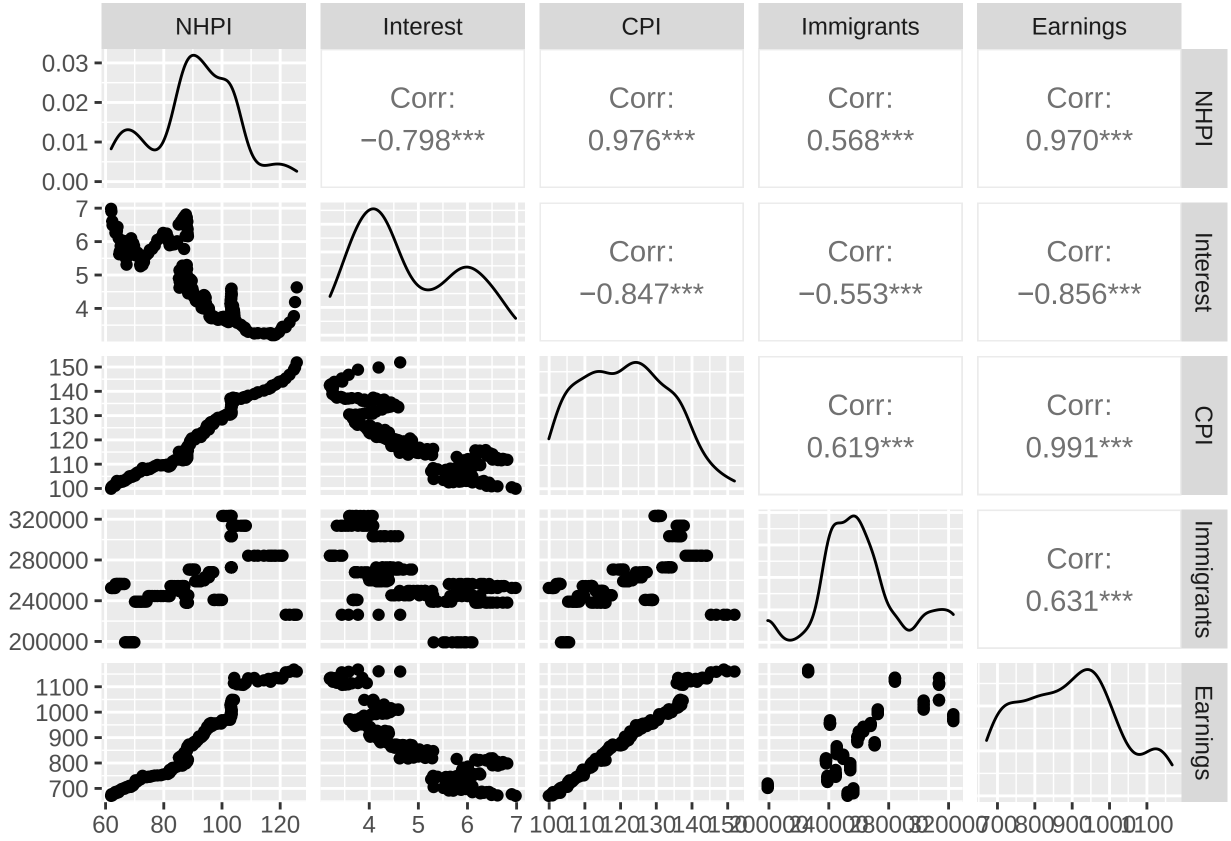


We observe a sudden jump in the average weekly employee earnings. The jump happened in March 2020, and is likely due to the pandemic.


Average Weekly Employee Earnings Over Time
Average Weekly Employee Earnings
700
800
900
1000
2005
2010
2015
2020
Time (monthly)

	VIF
Interest	3.752
CPI	58.437
Immigrants	1.672
Earnings	63.187

Collinearity



We measured the collinearity in the dataset with VIF (Variance Inflation Factor). The VIFs of CPI and Earnings is larger than 10, and therefore, the collinearity between them may be a problem.

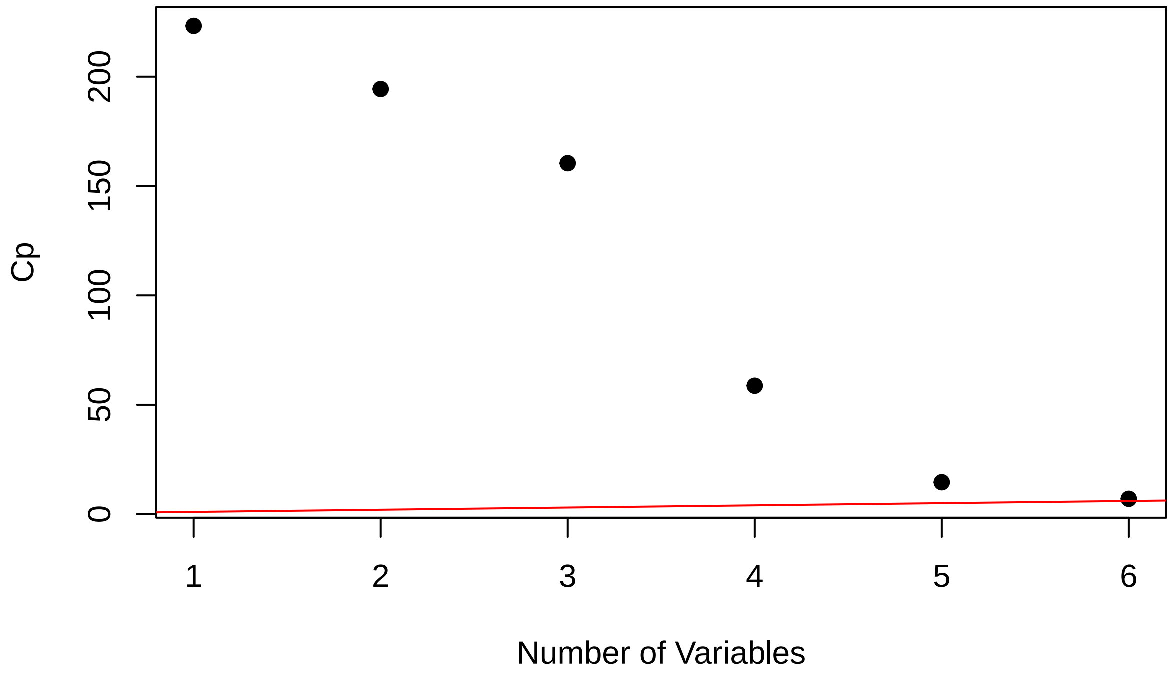
Transformation

An interaction terms is added to our model to address the high collinearity of CPI and Earnings. The quadratic term of Interest was introduced to improve linearizeability and sample correlation as shown in Figures 5 and 6.

p	r^2	r^2_{adj}	C_p
0	0.9522666	0.9520661	223.20787
1	0.9554736	0.9550978	194.35584
2	0.9592062	0.9586876	160.44748
3	0.9699909	0.9694801	58.69533
4	0.9747847	0.9742459	14.57800
5	0.9757803	0.9751566	7.00000

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-52.1039140	6.3503810	-8.204849	0.0000000
Interest	1.6062384	0.3581727	4.484536	0.0000114
CPI	0.9134636	0.1191467	7.666712	0.0000000
Immigrants	-0.0000324	0.0000087	-3.709042	0.0002597
Earnings	0.0370122	0.0114562	3.230756	0.0014113

Model Selection with Exhaustive Search



We chose the model with 6 predictor variables including the interaction term and the quadratic term because C_5 is very close to the corresponding value of p, and the model therefore appears the least unbiased.

We created 3 candidate models. full linear Model 1, Model 2 with an interaction term between CPI and average weekly earnings, and Model 3 with the same interaction term and a quadratic term for the mortgage rate. The quadratic term was added based on Figure 2, and the residual plots of the fully linear model, and the model with an interaction term.

The models appear excellent in terms of the significance of the predictor variables. However, they were based

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-183.1623129	12.5383819	-14.608130	0
Interest	2.9667882	0.3110743	9.537234	0
CPI	1.7966471	0.1228930	14.619600	0
Immigrants	-0.0000453	0.0000071	-6.374004	0
Earnings	0.2010940	0.0170363	11.803844	0
CPI:Earnings	-0.0011433	0.0000999	-11.440228	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-173.4710111	12.2219078	-14.193448	0.0000000
Interest	-6.4194332	2.0742415	-3.094834	0.0022101
CPI	2.0165623	0.1274021	15.828327	0.0000000
Immigrants	-0.0000482	0.0000069	-7.033372	0.0000000
Earnings	0.2025099	0.0163577	12.380118	0.0000000
I(Interest^2)	0.8988153	0.1965580	4.572773	0.0000078
CPI:Earnings	-0.0012664	0.0000996	-12.709211	0.0000000

on a time series data, which is most likely serially correlated. Hence, we need to examine, and possibly correct for the possible autocorrelation in the data.

Serial Correlation

The standard assumptions of linear regression includes serial independence of data. The Durbin-Watson statistic was used to measure the serial correlation of our model [1].

The Durbin-Watson statistic for our model is 0.1225723, and which is close to 0 with a very small p value, indicating high positive serial correlation. The standard errors of coefficients are underestimated if the data is positively serially correlated.

Newey-West Standard Errors

The standard errors of coefficients are underestimated due to the high serial correlation. Instead the Newey-West standard errors, both Heteorscedasticity and Autocorrelation (HAC), should be used instead [2]. The chosen delay truncation value is 12 intervals (months).

The significance of some of the variables appear much lower when corrected for autocorrelation, and most of them appear insignificant when corrected for serial correlation. This model is mostly based on the autocorrelation in the data.

When the interaction term is included, the significance of the predictor variables are improved overall. The increased complexity of the model better captures the significance of the predictor variables and their interactions.

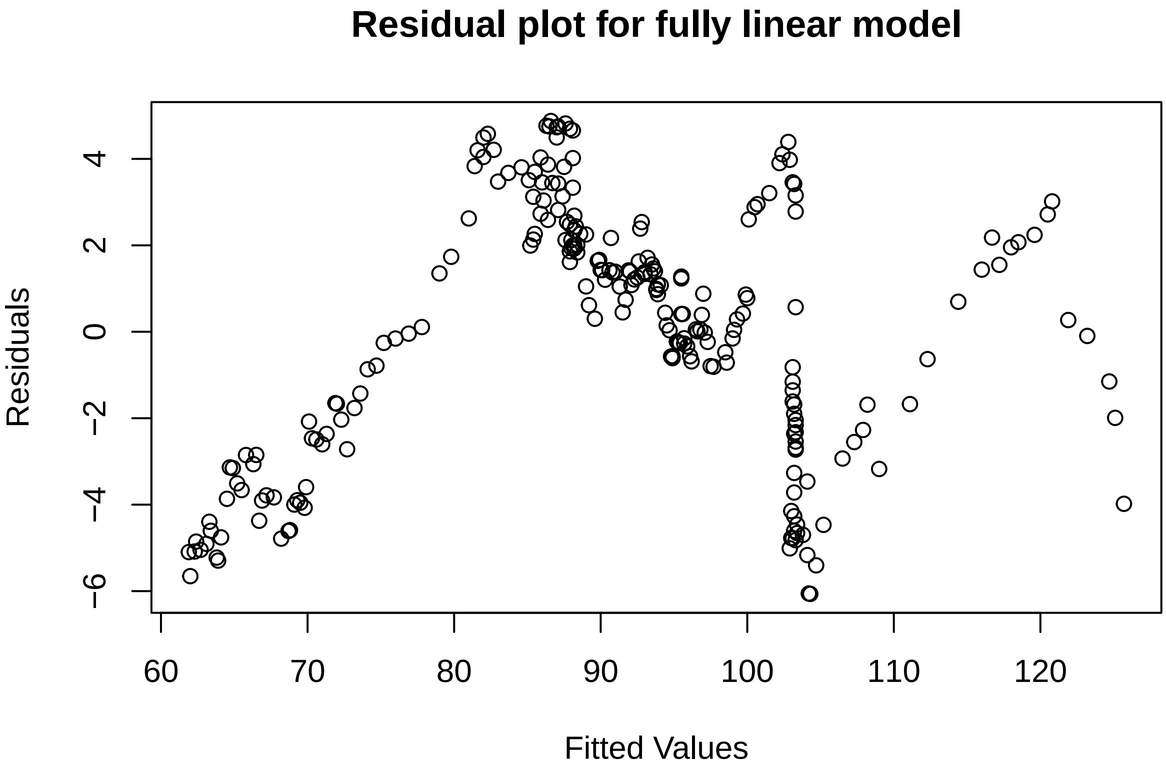
The predictor variables remain significant even after the correction. The lag is set at 12 months. The significance of Interest and Interest^2 is individually quite low, but cannot be discounted since the linear and quadratic terms may still be significant when combined.

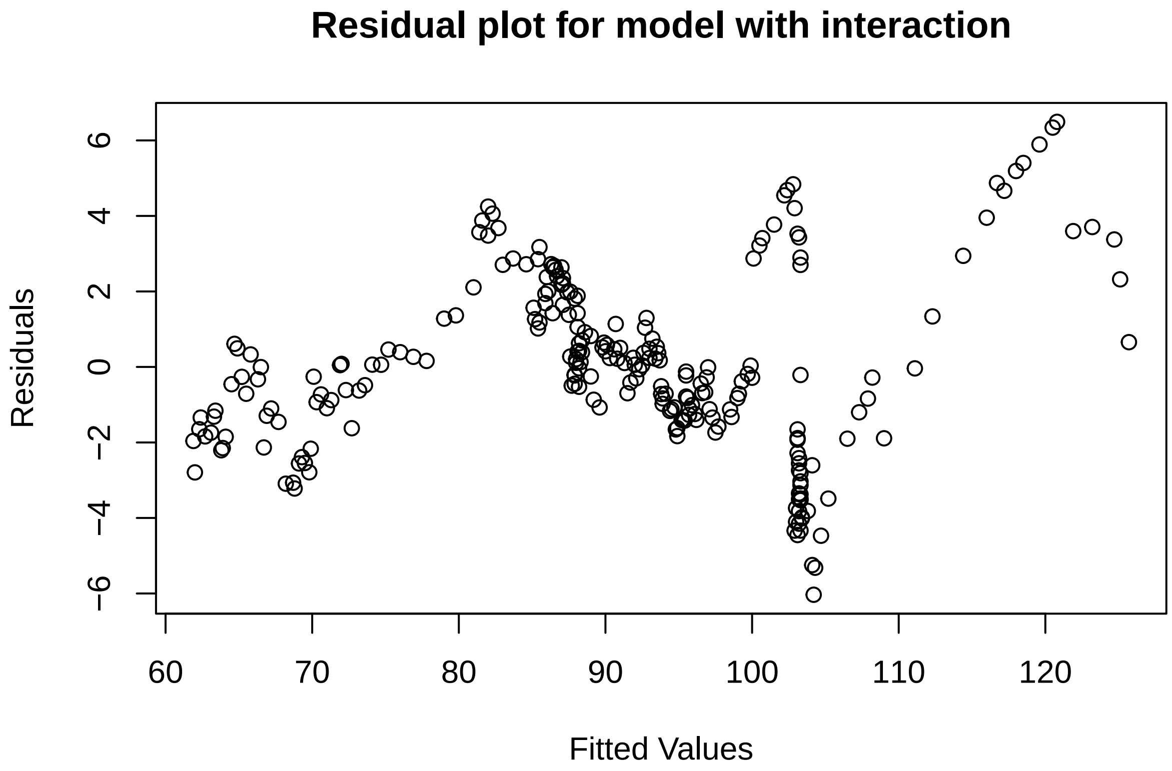
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-52.1039140	18.8291084	-2.767200	0.0061041
Interest	1.6062384	1.2335760	1.302099	0.1941572
CPI	0.9134636	0.3422132	2.669282	0.0081314
Immigrants	-0.0000324	0.0000276	-1.175946	0.2408065
Earnings	0.0370122	0.0343642	1.077057	0.2825592

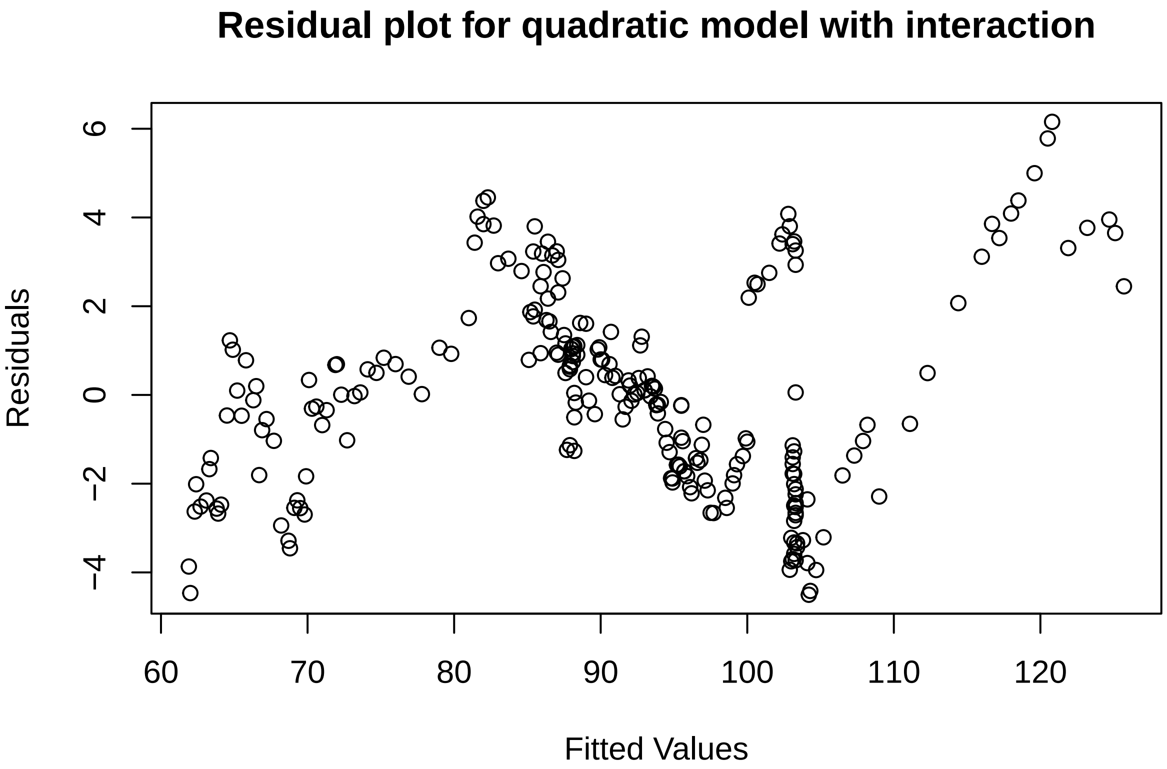
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-183.1623129	37.5845399	-4.873342	0.0000020
Interest	2.9667882	0.8128600	3.649815	0.0003237
CPI	1.7966471	0.3846929	4.670341	0.0000051
Immigrants	-0.0000453	0.0000252	-1.799818	0.0731775
Earnings	0.2010940	0.0490489	4.099865	0.0000570
CPI:Earnings	-0.0011433	0.0003127	-3.656415	0.0003159

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-173.4710111	35.6092516	-4.871515	0.0000020
Interest	-6.4194332	6.1070554	-1.051150	0.2942785
CPI	2.0165623	0.3380983	5.964426	0.0000000
Immigrants	-0.0000482	0.0000244	-1.972527	0.0497320
Earnings	0.2025099	0.0465326	4.352004	0.0000202
I(Interest^2)	0.8988153	0.5527657	1.626033	0.1052943
CPI:Earnings	-0.0012664	0.0002858	-4.430721	0.0000145

Residuals




Residual plot for model with interaction
Residuals
6
4
2
0
−2
−4
−6
60
70
80
90
100
110
120
Fitted Values


Residual plot for quadratic model with interaction
Residuals
6
4
2
0
−2
−4
60
70
80
90
100
110
120
Fitted Values

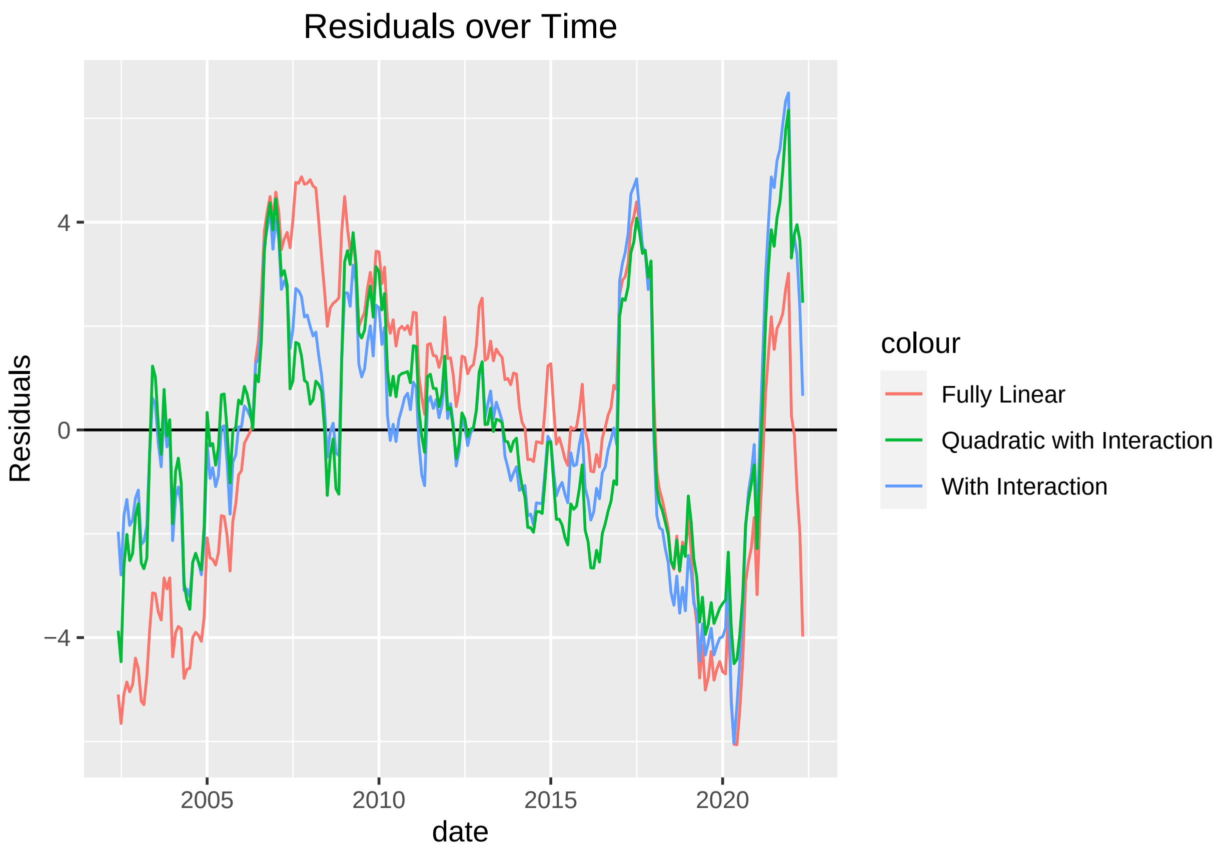
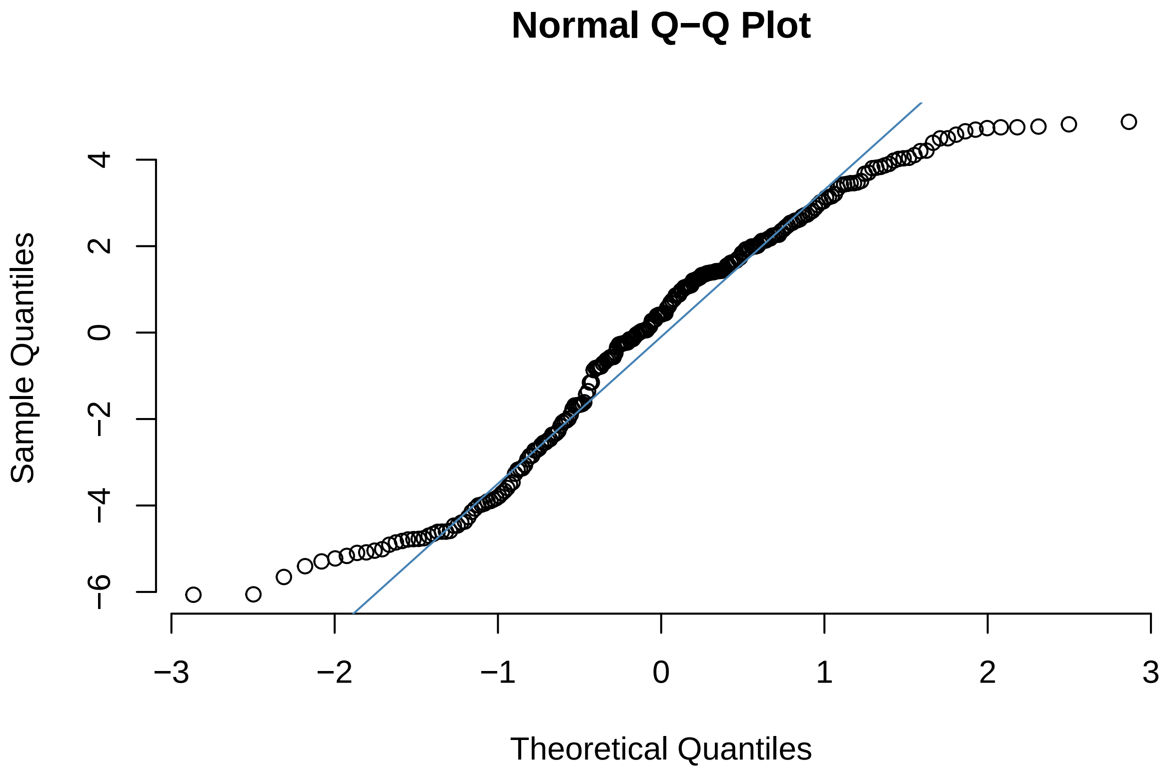
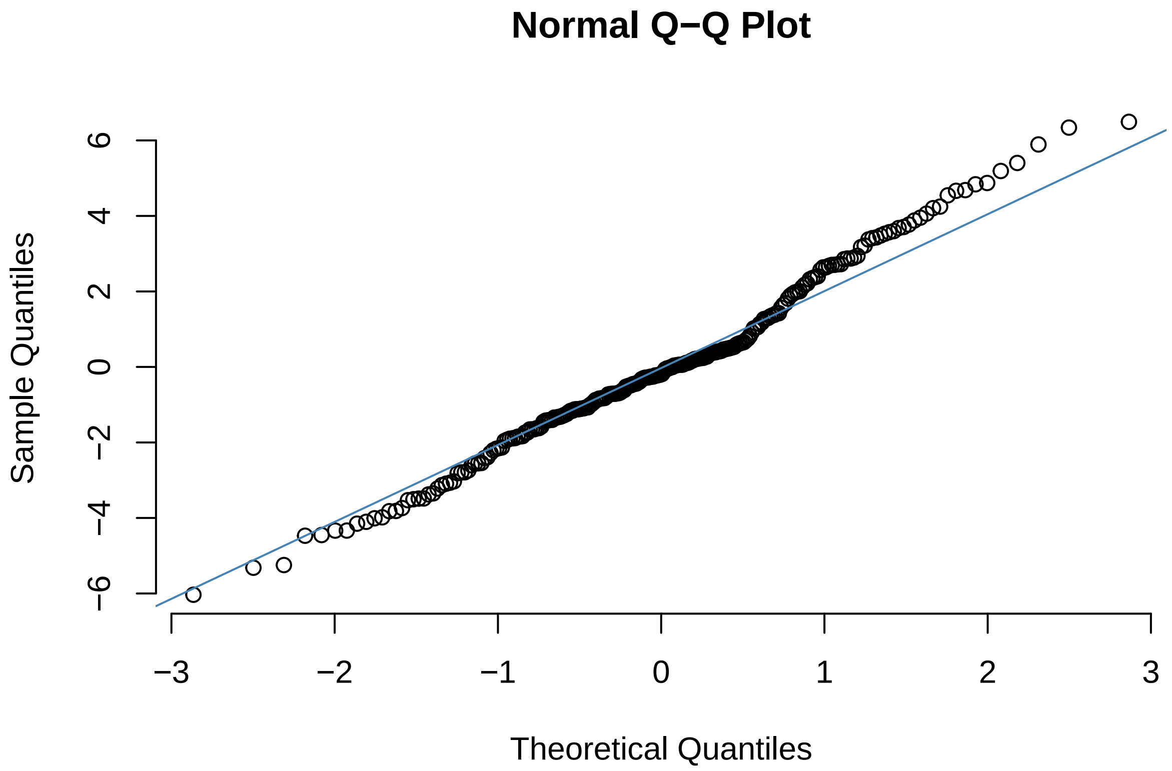


Figure ?: Residuals over Time of All Three Models. The lines show similar patterns to their respective residual plots.

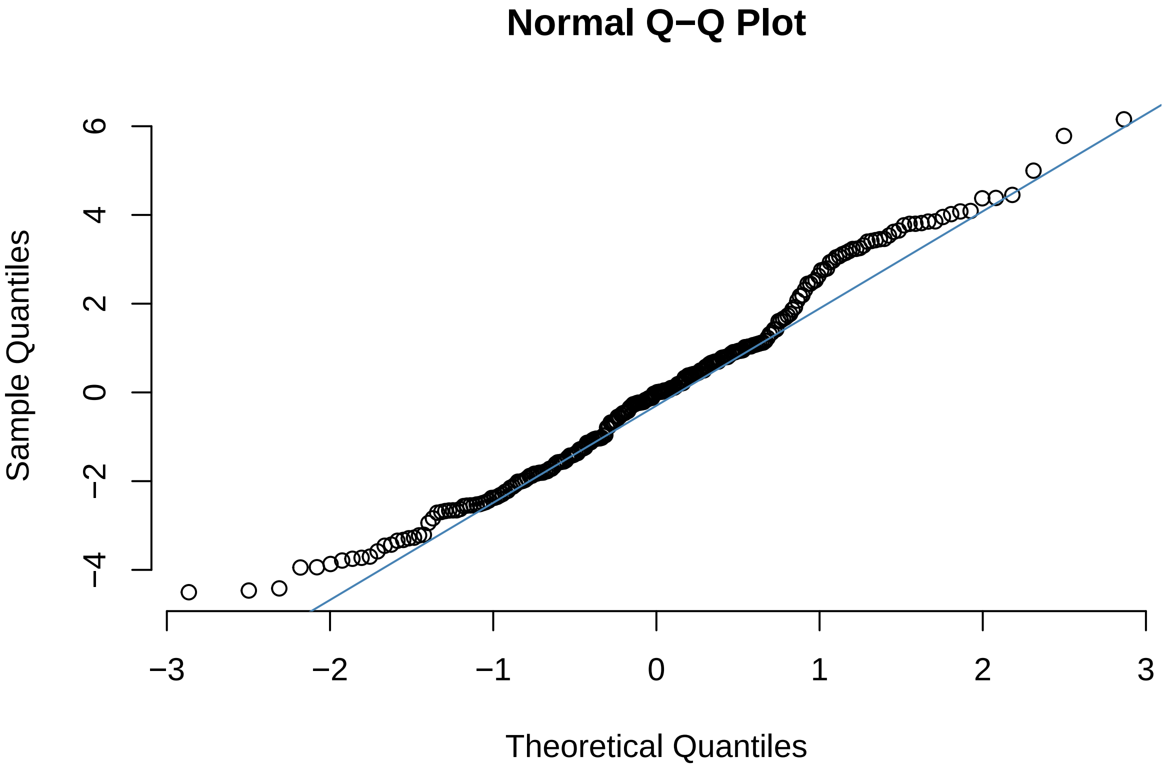
There seems to be patterns in residuals to the autorrelation in the form of a serial correlation in the data. The continuous patterns seem to arise from the fact that our data is a time series, and do not necessarily imply lack of linearizability. Still, we cannot explain with all variance in NHPI with the predictors here due to rapid, artificial shifts in variables such interest rates, which can change drastically based on the whims of the central bank. We may need more variables and average-over-time transformations for a more complete model that accounts for sudden shifts.



The error distribution of the fully linear model is light-tailed, and the normality-of-errors assumption may not hold. While The simplicity of the model is a positive attribute, the model is too light-tailed on the right side, and undermines the model's value for making forecasts.


Normal Q−Q Plot
Sample Quantiles
−6
−4
−2
0
2
4
6
Theoretical Quantiles
−3
−2
−1
0
1
2
3

Model	RMSE
Fully linear	1.426268
With an Interaction	1.130076
Quadratic	1.159623



The distribution of the residual is heavy-tailed, but not extremely so. The normal error assumption seems to mostly hold.

Forecast Tests with a Holdout Set

We cannot do a cross-validation on a time series data. We instead train our model on the first 80% of the data, and test it on the last 20% of the data. The root mean square prediction errors are used to evaluate the three models.

The model with an interaction term seems to perform the best. However, our data is a time series, and the holdout sets necessarily include the most recent data, which tend to be the most important data points in time series models. Furthermore, because the most recent data involves the highly unusual period of the COVID-19 pandemic, it is unlikely that we can build a strong model while excluding the most recent 20% .

Akaike Information Criterion (AIC)

Because of the reasons discussed in the previous sections, tests using holdout sets may be misleading in evaluating the models. As a result, AIC is particularly valuable for choosing the best model among our three models.

The quadratic model (Model 3) has the lowest AIC among the three models. It implies that the quadratic model has the smallest prediction error in general.

Model	AIC
Fully linear	1205.517
With an Interaction	1100.898
Quadratic	1082.272

Model	Month	lower	fit	upper	se
Fully Linear	June	114.3821	125.9675	137.5530	4.506933
Fully Linear	July	118.2203	130.3142	142.4081	4.704728
Interaction	June	104.6463	116.9696	129.2930	5.036294
Interaction	July	108.1633	121.3481	134.5329	5.388347
Quadratic	June	106.3749	125.7726	145.1703	8.203281
Quadratic	July	110.2866	125.0410	139.7953	6.239603

Forecasts Using Models

One major use of time series models is to forecast response variables in the future. In this section, we attempt to forecast NHPI for June and July 2022 with imputed values, and updated interest rates. All predictor variables were assumed to have remained the same since May except for the interest rate. The Newey-West standard errors were used to calculate the 95% prediction intervals. The real NHPI for June 2022 was 125.9 [4]. The NHPI for July 2022 has not been published yet.

```
kable(pred.df)  %>%
  kable_styling(bootstrap_options = c("striped", "condensed"), full_width = F)
```

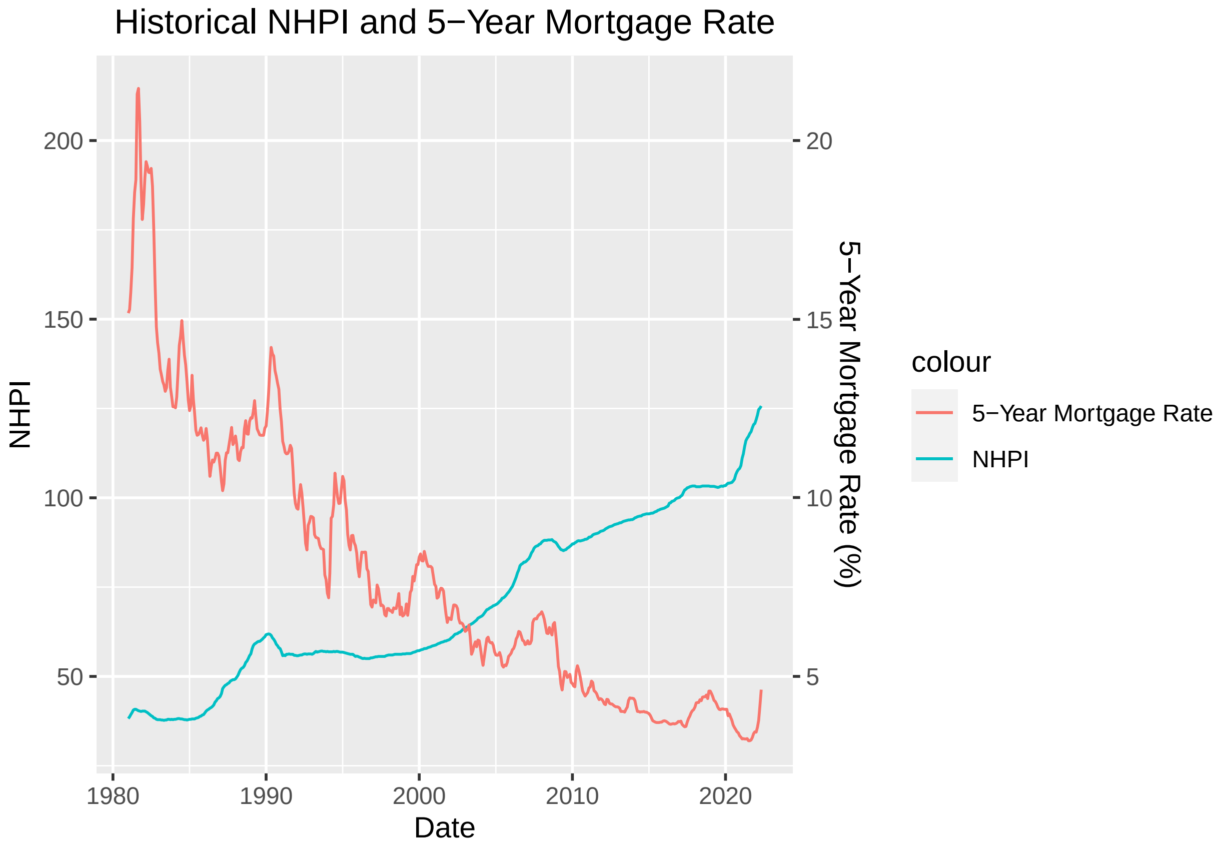
Figure ?: Prediction Intervals and Errors for June and July 2022 for All Three Models.

The model 2 is very far off with its predictions. The more complicated models are more conservative, and have wide prediction intervals.

Discussion

The minimum in the quadratic model is at $Interest = \frac{-b_{Interest}}{2 \cdot b_{Interest^2}}$. Note that $b_{Interest^2}$ positive.

The model suggests the 5-year mortgage rate to achieve the minimum NHPI is 3.571 % if all other predictor variables are held constant. The positive correlation is unintuitive as home price is expected to continue to decline with higher interest rates. This does not necessarily mean the model is biased. There have been periods of rapidly rising NHPI through interest hikes in the 80's and 2000's. **TODO: refer to the figure below.** However, the mortgage rate at the partial minimum is close to the historic low of 3.2%, and may imply a bias due to a limited range of interest rates in the data. We should also consider the fact that interest rates are largely controlled by the Bank of Canada, and are raised in response rising costs including housing costs. **TODO: Please find a reference for this.** Furthermore, the effects of raising interest rates to housing cost may come with significant delays. Hence, high interest rates past a threshold may be associated with high home price, and our quadratic model may reflect such tendencies. In future studies, we could explore the possible causal effect of interest rates on house prices by introducing delays to interest rates.



Model 1 is a simple model, which failed to capture the significance of most predictor variables, and relied mostly on autocorrelation for its predictions. Model 2 is a good candidate in terms of RMSE, but appears to be too sensitive to changes in the interest rate, and makes poor forecasts. Model 3 is most favorable in terms of improved significance of the predictor variables over the other models, the lowest AIC. It makes sensible forecast results for June and July 2022. Thus, we conclude that Model 3 best models the market behavior in regard to NHPI.

Conclusion

The best model is Model 3:

$$y = -173.47101 - 6.41943x_1 + 2.01656x_2 - 0.00005x_3 + 0.20251x_4 + 0.89882x_5 - 0.00127x_6$$

, where x_5 is x_1^2, and x_6 is the interaction between x_3 and x_4.

The most significant predictors of NHPI are CPI and employee earnings by far. Their significance implies that general income and affordability of goods form the baseline for the housing market. The number of recent immigrants is apparently the least significant factor, but it may be due to the poor quality of this data. Provided more accurate monthly data on immigration, the significance of immigration might increase.

The analysis yielded some unexpected results regarding the effects of interest rates and immigration. When adjusted for CPI and earnings, immigration and the interest rate seem to affect the NHPI in the opposite directions that they are typically associated with in the literature. High immigration lowers NHPI, and high interest rate is associated with high NHPI. Simply limiting immigration may not result in lower housing prices.

The significance of the mortgage rate is difficult to interpret. The bank rate set by the Bank of Canada essentially have a bidirectionally causal relationship with housing prices, and indirectly dictates mortgage rates. Thus, the effects of interest rates on the Canadian housing market is much more subtle, and hard to capture. Our quadratic model (Model 3) shows that the 5-year mortgage rate affects NHPI most negatively at near its historic low of 3.571 %, further suggesting the subtlety of the effects of interest rates on the housing market. Therefore, it may be dangerous to assume that the rising housing costs can be simply addressed by hiking interest rates to a very high value. If other factors such as inflation stays out of control, high interest rates could further exacerbate the housing crisis by driving cash flow towards the housing market to protect assets against inflation.

The coefficient for the interaction between earnings and CPI is negative, which means the positive impact of earnings and CPI on NHPI is attenuated when these values change in the same direction. In other words, if changes in average wage stay in parity with CPI, housing prices tend to be a little lower. Thus, addressing the longstanding gap between wage increase and inflation may be a key step to lower the current rampant housing prices.

References

Durbin, J.; Watson, G. S. (1971). "Testing for serial correlation in least squares regression.III". Biometrika. 58 (1): 1–19. doi:10.2307/2334313

Newey, Whitney K; West, Kenneth D (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix" (PDF). Econometrica. 55 (3): 703–708. doi:10.2307/1913610. JSTOR 1913610.

RBC-Pembina Location Matter series(2013). "Understanding the factors affecting home prices in the GTA.", https://www.pembina.org/

Statistics Canada. 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

Statistics Canada. Table 18-10-0205-01 New housing price index, monthly https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810020501

Statistics Canada. Table 34-10-0145-01 Canada Mortgage and Housing Corporation, conventional mortgage lending rate, 5-year term https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410014501

Statistics Canada. Table 17-10-0008-01 Estimates of the components of demographic growth, annual https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000801

Statistics Canada. Table 18-10-0004-01 Consumer Price Index, monthly, not seasonally adjusted https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000401

Statistics Canada. Table 14-10-0223-01 Employment and average weekly earnings (including overtime) for all employees by province and territory, monthly, seasonally adjusted https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410022301