STA302 Report

Research Question

What is the relationship between Canadian unemployment rate and the following economic indicators: GDP growth, number of new immigrants, investments growth, prime rate, and inflation rates?

Literature Review

Unemployment rate is a key Economic metric that concerns the entire society, especially as upper-year college students that will soon enter the job market. There are a number of researches that uses multiple linear regression to study the underlying causes of unemployment rates. For this project, 3 different studies were referenced below[1][2][3].

All referenced paper aims to model the relationship between macroeconomic indicators and the country's unemployment rate. Several common predictors include GDP growth, inflation, urbanization, and direct investments. Generally, predictors that are more directly related to economic performance (like GDP) have a statistically significant relationship with unemployment rate, whereas those that are more tangential (like urbanization) are insignificant.

The subjects from the referenced studies are all developing economies that heavily relies foreign investments to support their economies. In contrast, Canada is a relatively developed and wealthy country with a slower rate of economic growth. As a result, in addition to the established economic indicators, the following predictors are added to the model:

1. Immigration: Canada is an immigration country that admits a substantial amount of immigrants each year, and a majority of them are educated and of working age. Does the amount of immigration influence the unemployment in any way?
2. Prime interest rate: The interest rate plays a crucial role in regulating the economic cycle. The addition of this predictor seeks to investigate the effectiveness of fiscal policies on unemployment rates.

Methodology:

Data Collection

This study collected quarterly economic data from 2000-Q1 to 2022-Q4. The data for GDP, Investment, Prime interest rate, and projected inflation rate are collected from OECD database at the corresponding collection period. The number of quarterly new immigrant is collected from the Statistic Canada website. As a result, the resultant dataset contains 6 variables of 92 observations each.

Exploratory data analysis

A preliminary EDA is performed on the original dataset to identify any missing data and outliers. Since the amount of predictor in this model is limited, a scatter matrix is plotted to visually identify any outliers and verify the linearity of the model. A histogram and boxplot of each variable is the plotted to verify the normally of each variable. If any predictor has a skewed distribution, the predictor may be a candidate to a transformation of variables.

Selection and validation of models

The dataset is then divided into training and testing sets with a ratio of 0.5, and a preliminary linear model is built with all predictors included in this study. If any of the predictor is deemed statistically insignificant, a new

[1] Chowdhury, M., & Hossain, T. (2014). Determinants of Unemployment in Bangladesh: A Case Study. *Developing Country Studies*, 4(3).

[2] Ni, T. V., Yusof, Z. M., Misiran, M., &amp; Supadi, S. S. (2021). Assessing Youth Unemployment Rate in Malaysia using Multiple Linear Regression. *Journal of Mathematics and Computing Science*, 7(1), 23–34.

[3] Puspadjuita, E. (2017). Factors that Influence the Rate of Unemployment in Indonesia. *International Journal of Economics and Finance*, 10(1). https://doi.org/10.5539/ijef.v10n1p140

model would be constructed without said predictor, then a partial F-test will be conducted to see whether the predictor should be removed.

After the linear model is built, the assumptions of multiple linear regression were examined using the produced diagnostic plots. Specifically, the assumption of homoscedasticity is checked using the fitted value/variable vs. residual plots, the normality assumption is checked using the qq-plot, the multicollinearity of variables will be determined through the VIF values of the model, and the whether the predictor/response variables will be determined via the shape of the residual plot.

After the 'best' model for the training set has been selected, a linear regression will be performed on the testing set. If the values of the parameters are within the confidence interval of the training model, then the model will be validated.

Results

Exploratory Data Analysis (EDA)

During this step, 2 observations (#82 and #83) deviates from the main trend for all predictors. The time period corresponding to those observations are 2020-Q2 and 2020-Q3 respectively, when the COVID crisis was at its peak. due to the anomalous deviation from the rest of the points, as well as its underlying context, those 2 observations were removed from the dataset as outliers. The position of the outliers can be clearly observed on the scatterplot matrix below:

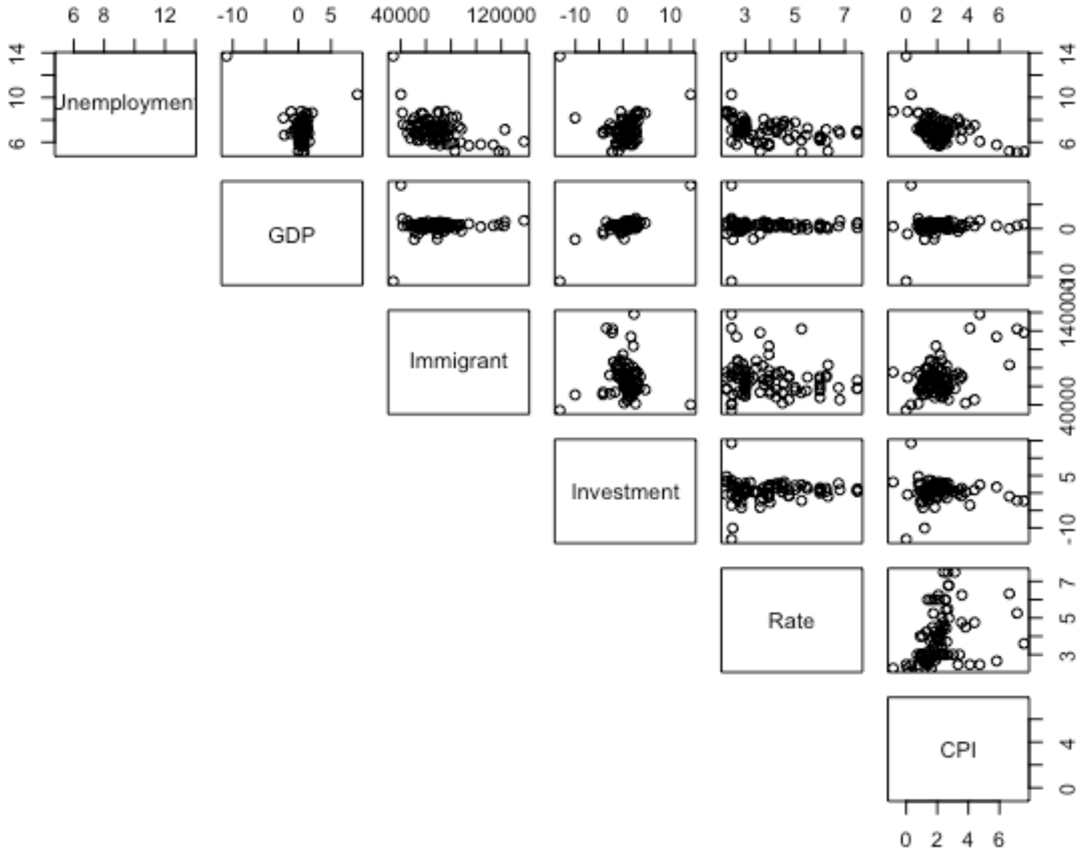

Figure 1. Scatterplot matrix of the combined data. Outliers can be clearly seen.

In addition, the distribution of prime rates and immigration is to be skewed to the right, which may warrant a transformation of variables. The distribution of the rest of the variables were not exactly normal, but by removing the outliers at the 'tail' of the distribution, the distribution of other variables is reasonably normal.

A preliminary model involving all predictors was built using the following formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

X_1 = GDP growth, X_2 = Number of new immigrants, X_3 = percentage change in investments, X_4 = Prime interest rates, X_5 = Inflation rates.

The multicollinearity of the variables is also tested with the following result: (1.491139, 2.370263, 1.554328, 1.432319, 2.436207), which is not problematic. With only the number of new immigrants and prime interest rates being significantly related to the response variable (p-value < 0.01), the AIC value (97.7219) and R^2_{adj} value (0.4656) of this linear model was taken for future comparisons, and a new model is built without the insignificant predictors are removed from the model.

All statistically insignificant predictors are removed, which yields:

$$Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4$$

with a AIC value of 95.42 (R^2_{adj} = 0.4612). A partial F-test was conducted to see whether the predictors X_1, X_3, X_5 were necessary in the model. The resulting p-value is 0.35, which suggests that they were not. Since the new model has 3 less predictors and a very minor decrease of R^2_{adj} value (a decrease of 0.0044), this model is chosen for further analysis. The diagnostic plots of the current model are plotted below:

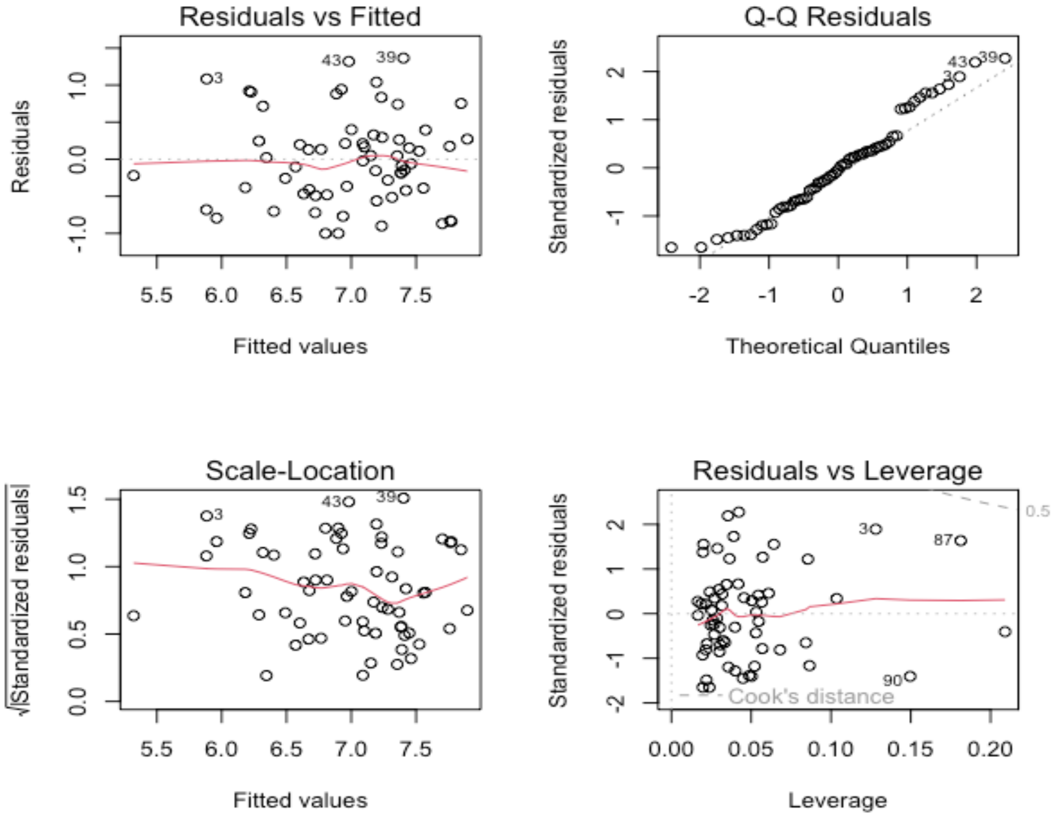

Figure 2. Diagnostic plots for initial linear fit

As observed from the residual plots, the residuals appear to be evenly distributed with no discernible patterns, which signifies that the assumption of equal variances is not violated. However, the QQ-plot has a notable disjoint from normality at the tails, which may suggest the assumption of normality may be violated and warrants a transformation of some kind. The predictor vs. residual plot for X2 indicates a non-random pattern, and the same plot for X4 indicates a slight fanning pattern, which supports the former hypothesis.

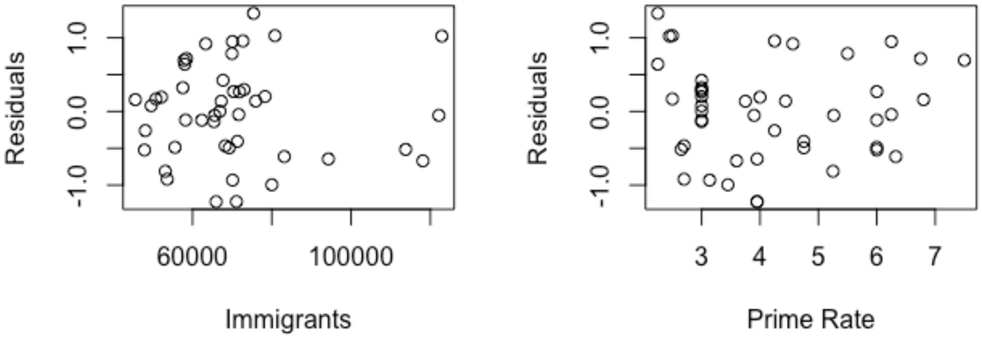

Figure 2. Predictor for linear fit. Note the fanning of the Rate vs. Residuals

A power transform function is used to provide the power coefficients of variables. For X_2, $\lambda_{X_2} = -0.6470$, which is rounded to the nearest integer of -1 and suggests a reciprocal function of $\frac{1}{X}$. $\lambda_{X_4} = -0.6682$, which is also rounded to -1, which suggests the same relationship. $\lambda_Y = 1.3704 \approx 1$, which corresponds to the identity function. A model that is transformed to the power coefficients of -0.5 (a reciprocal square root transformation of $\frac{1}{\sqrt{X}}$) is also considered and rejected during the validation process, as it exhibits similar behaviours to the less complex transformation.

The diagnostic plots of the transformed model are shown below:

$$Y = \beta_0 + \frac{\beta_1}{X_2} + \frac{\beta_2}{X_4}$$

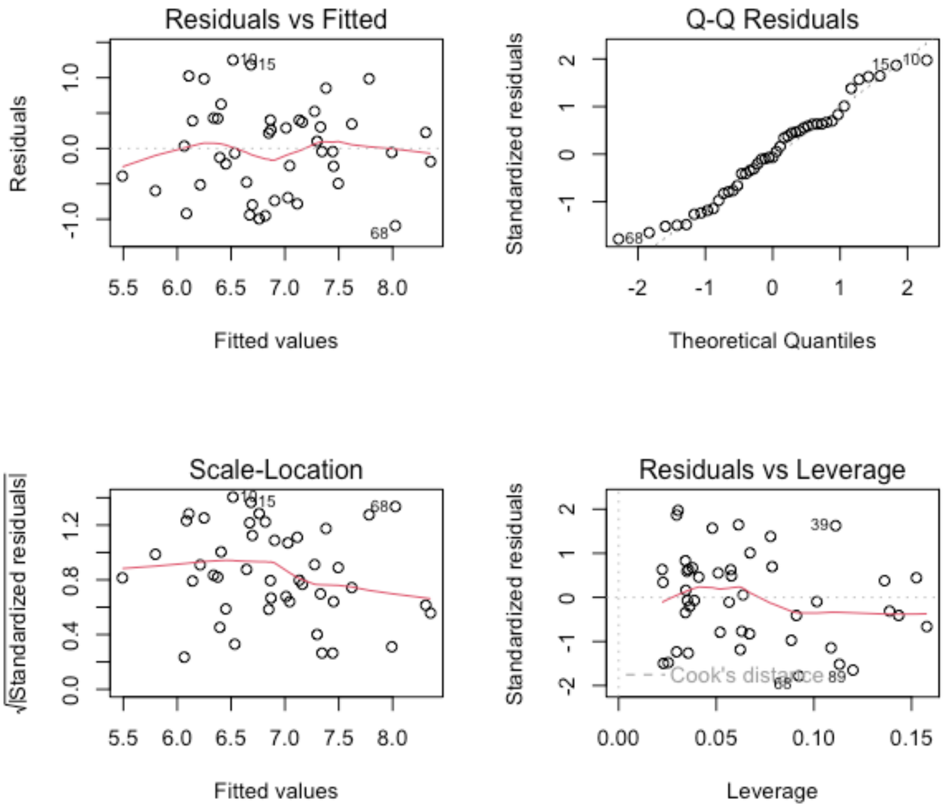

Figure 3. Diagnostics of reciprocal transformation

It can be seen that the normality of the residuals has improved significantly, although the QQ-plot indicates the presence of a heavier tail. There appears to be no visible patterns for the residual plots, which means that the assumption of equal variance and normality is not violated. The pattern of the leverage plot indicates that the cook's distance for each point is shorter than the model pre-transformation. The linearity condition is also satisfied, as there is no visible pattern on the predictor vs. residual plots:

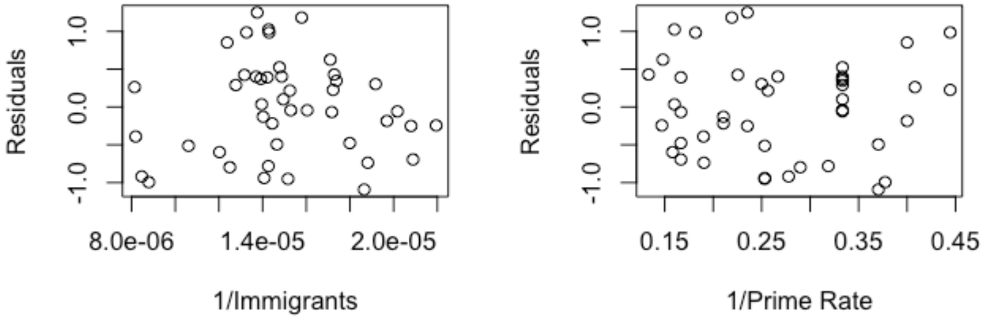

Figure 4. Predictor vs. Residual plot of reciprocal transformation

Table 1. Summary of final model (Training set)

Model	Estimated	Standard Error	p-value	95%CI
Intercept	3.197	0.5914	2.22e-06*	(2.003938, 4.390999)
1/Immigration	132700	29780	6.11e-05*	(72580.929061, 192779.6)
1/Rate	6.354	1.110	9.84e-07*	(4.2624022, 7.829822)
R^2	**0.5144**	**R^2_{adj}**	**0.4913**	

*Significance level set at p<0.01

Table 2: Variance Decomposition Table (ANOVA)

Model	Sum of Squares	dof	Mean Square	F-value
Regression	18.3894	2	9.1947	22.247
Residual	17.3583	42	0.4133	
Total	35.7477	44		

The overall F-value is 22.25, which signifies the overall model is significant. ($F_{0.99,\,42}$ = 5.18)

Table 3. Summary of the final model (testing set)

Model	Estimated	Standard Error	p-value	Validated
Intercept	3.778	0.6042	1.72e-07*	Yes
1/Immigration	118800	26540	5.73e-05*	Yes
1/Rate	5.050	1.103	4.11e-05*	Yes
R^2	**0.4334**	**R^2_{adj}**	**0.4064**	

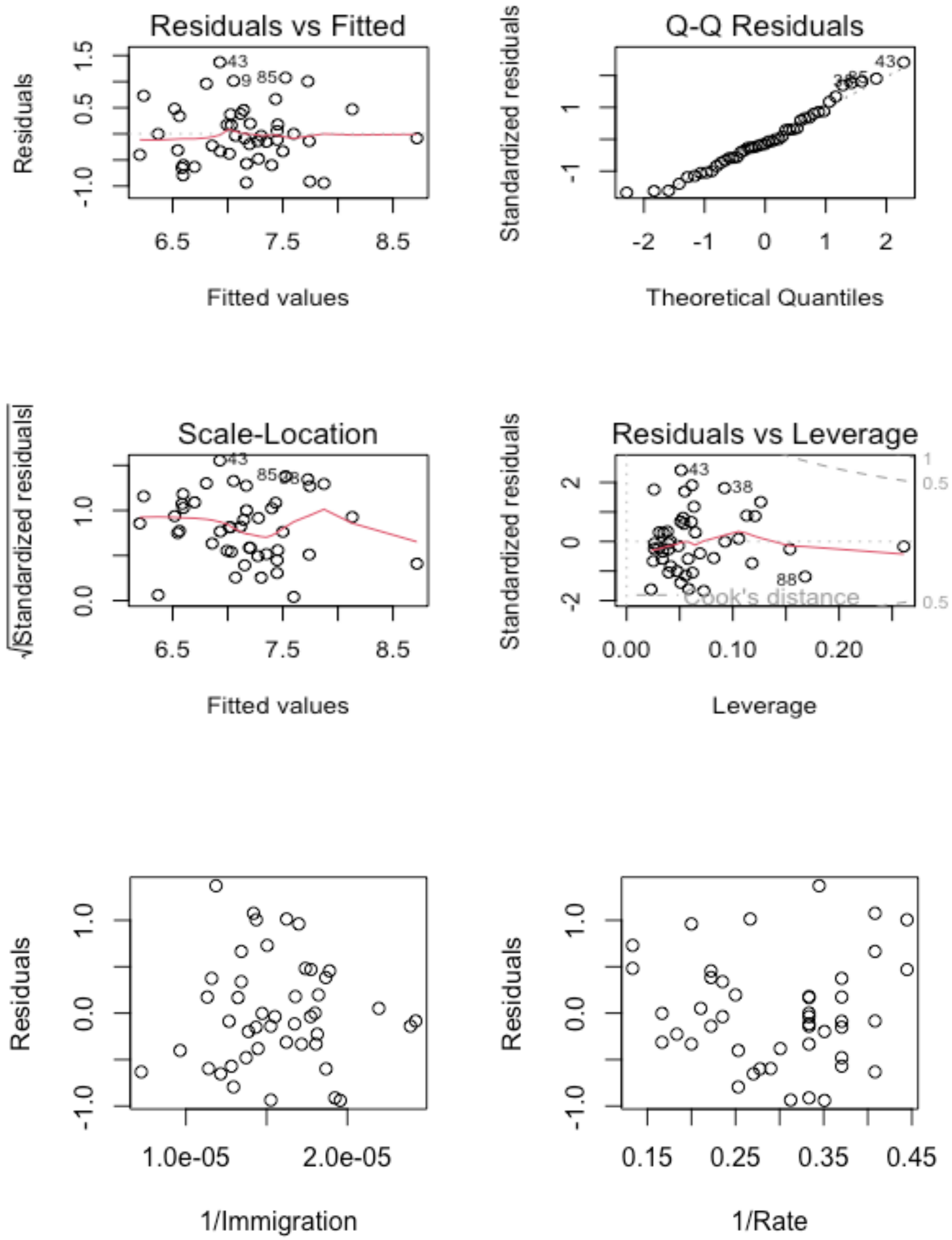

Figure 5, 6. Diagnostic plots of final model, predictor vs. residual plots (testing set)

According to the diagnostic plots, the requirements for normality, constant variance and linearity is reasonably satisfied, since the residuals and QQ-plots are similarly behaved, as the behaviour is similar to the training model. However, it is notable that the average Cook's distance for the testing set are longer than the training set. This means that there are more leverage points in the testing set, which may be the reason behind the notable drop in R^2_{adj} value (0.4913 vs. 0.4064).

The coefficients appear to be similar to that of the training model, and all testing coefficients fells in 95%CI of the training model with similar levels of standard error. As mentioned before, the R^2_{adj} value are notably different, but that can be explained by the variability of random data. Overall, the final model:

$$Y = 3.197 + \frac{1.327 \times 10^5}{X_2} + \frac{6.354}{\mathrm{X}_4}$$

is validated.

Discussion:

To summarize, among all predictors selected in this model (GDP growth, Immigration, Investment growth, Prime interest rate, and Inflation rates), only immigration (X_2) and prime interest rate (X_4) has a statistically significant relationship with unemployment rate, the response variable. The final model:

$$Y = 3.197 + \frac{1.327 \times 10^5}{X_2} + \frac{6.354}{\mathrm{X}_4}$$

contains two reciprocal relationships, which means that as one of the predictors increases (and the other predictor remains constant), the unemployment rate would decrease. Notably, as the value of one predictor grows, the reduction in the unemployment rate would get smaller.

The first relationship between immigration and the unemployment rate appears logical since an influx of people (with a majority of educated, working-age adults) would stimulate the economy and spawn new employment opportunities. In this case, the introduction of the reciprocal can be explained by the law of diminishing return, where adding more people to the economy would reduce efficiency and benefits.

The second relationship is more perplexing. This is because an increase in interest rate represents the beginning of contractionary policies, which aims to cool down the economy and increase unemployment. One possible explanation is that it takes time for the increasing interest rate to take effect, and a high-interest rate is an indicator of an overheating economy with plenty of job positions.

The intercept represents the ‘natural’ unemployment rate where the influence of both predictors are zero. In the context of this model, it means that in an environment with plenty of incoming talents and a well-performing economy with a high-interest rate, one would expect the unemployment rate to be around 3.197%. This value is close to the widely considered optimum of 3-4% under a healthy economy.

The analysis confirms that GDP has a statistically insignificant relationship with the unemployment rate. This result is surprising since it violates Okun’s law, which predicts a negative linear relationship between GDP growth and the unemployment rate. Additionally, multiple linear regression may not be the optimal way to analyze economic metrics. Since one value may be influenced by the previous value. Instead, a time series analysis may be more appropriate. As a result, future research should re-examine the relationship between GDP and unemployment rate, preferably integrating time series analysis in the study.

References:

1. Chowdhury, M., & Hossain, T. (2014). Determinants of Unemployment in Bangladesh: A Case Study. *Developing Country Studies*, 4(3).
2. Ni, T. V., Yusof, Z. M., Misiran, M., &amp; Supadi, S. S. (2021). Assessing Youth Unemployment Rate in Malaysia using Multiple Linear Regression. *Journal of Mathematics and Computing Science*, 7(1), 23–34.
3. Puspadjuita, E. (2017). Factors that Influence the Rate of Unemployment in Indonesia. *International Journal of Economics and Finance*, 10(1). https://doi.org/10.5539/ijef.v10n1p140
4. OECD (2023), Investment (GFCF) (indicator). doi: 10.1787/b6793677-en (Accessed on 18 June 2023)
5. OECD (2023), Inflation forecast (indicator). doi: 10.1787/598f4aa4-en (Accessed on 18 June 2023)
6. OECD (2023), Quarterly GDP (indicator). doi: 10.1787/b86d1fc8-en (Accessed on 18 June 2023)
7. OECD (2023), Unemployment rate (indicator). doi: 10.1787/52570002-en (Accessed on 18 June 2023)
8. OECD (2023), Interest Rates: Immediate Rates (< 24 Hrs): Prime Rates (Accessed on 18 June 2023).
9. Statistics Canada (2023). *Table 17-10-0040-01 Estimates of the components of international migration, quarterly*. DOI: https://doi.org/10.25318/1710004001-eng.