STAT 306 Project Report

Date: 2022-04-04

Group: E5

Group Members: Conan Liu, Dan Liu, Rhi Ann Ng, Sandy Wu

Contents

1. Introduction

Motivation

Maintaining an ideal population plays an indispensable and positive role in a country's labor force, economic growth, and human rights. We need to know what we can do to keep the population at an excellent level. This study aims to explain which factors affect the population and how these variables interact with the population. This study involves five explanatory variables in nine Canadian provinces, including GDP, income, life expectancy, employment rate, and province. The response variable is the population estimates in the 4th quarter of each year of each region.

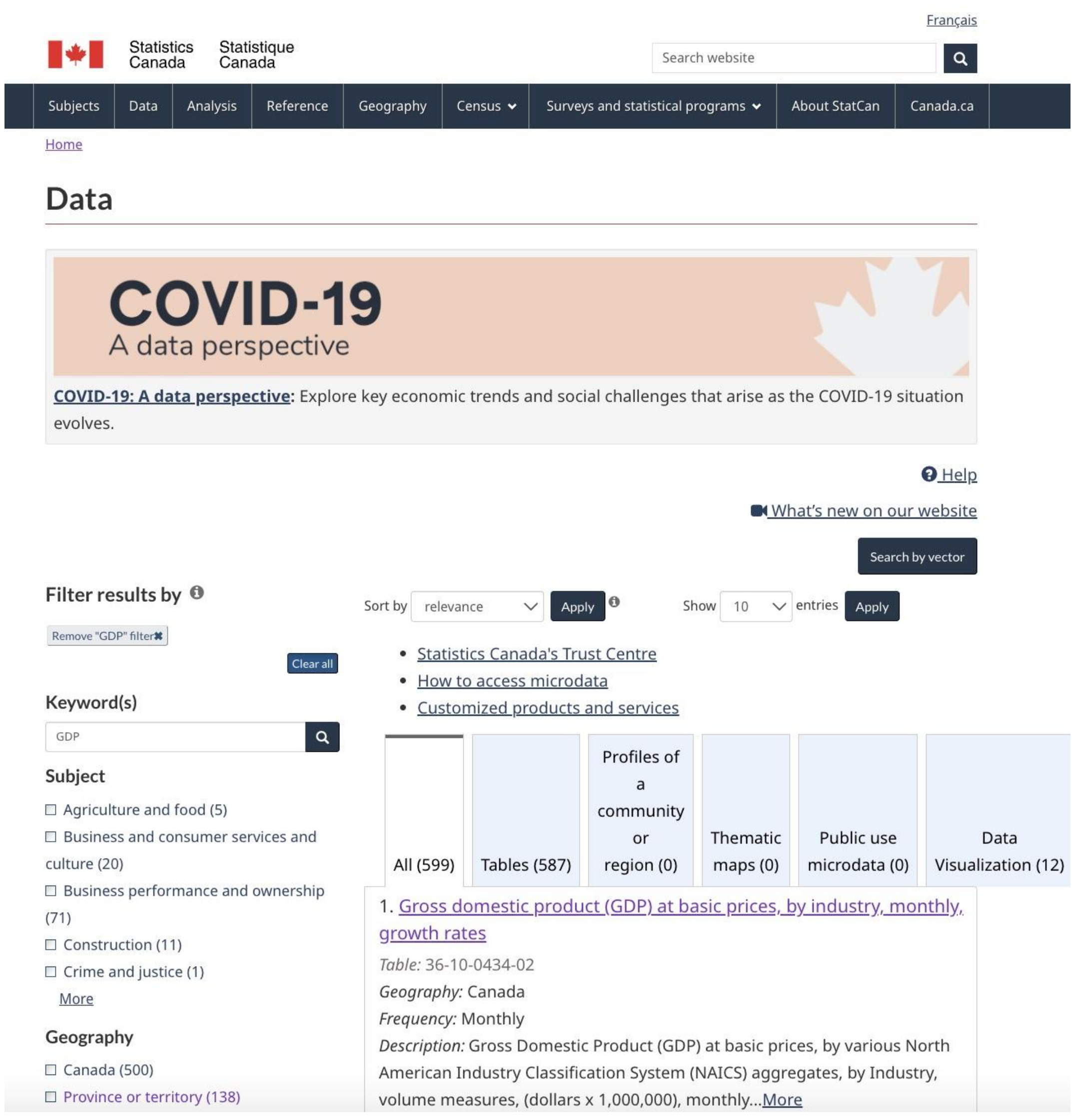
Data collected

The data set used in the study was historical data collected by Statistics Canada, which is official. The variables were measured from 2005 to 2010. The study focused on nine provinces in Canada: British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick, Nova Scotia, Newfoundland and Labrador. Below is a brief description of the variables and its notation in our model:

- *Population* ~ Population Estimates (in number of people)
  - Taken at quarter 4 of each year
- *GDP_PC* ~ Gross domestic product per capita (in million dollars)
- *Income* ~ Average Income (in dollars)
  - For economic families and persons not in an economic family
- *Life_Exp* ~ Life Expectancy (in years)
  - For both sexes for people aged 30
- *Empl_Rate* ~ Employment Rate (in percentage)
  - For both sexes, 15 years and older, all education levels
- *Province* ~ Province (categorical)
  - British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick, Nova Scotia, Newfoundland and Labrador

After looking through the initial dataset that was proposed in the project proposal, we made a few adjustments to our dataset.
The Statistics Canada website had more information about provinces than it did for cities, so we decided to change our cities variable to provinces. Unfortunately,

not all provinces and territories had data on the Statistics Canada website, so not all of them are included in our dataset. Given that we now had less observations, we also decided to include more years into our dataset. The years 2005-2010 were chosen arbitrarily. Whilst time does affect most, if not all, of the explanatory variables, we have chosen to ignore the time effect given that 5 years is not too long of a time frame. Since we have chosen to consider provinces, we had to remove the housing benchmark price variable. The prices in rural areas and urban areas vary significantly, even within the same province, thus it did not seem appropriate to include housing prices when considering provinces. On a similar note, we removed the education rate variable as there was not enough information. After further consideration, we also removed the language variable since only one province speaks French completely, whilst the others mainly speak English or a combination of both.



2. Analysis

a. Visualizations of Data and key Features

Since the GDP variable is actually dependent on the population, we took our observations for the GDP and divided it by the population number in order to get GDP per capita. Furthermore, min-max scaling was used on the data in order to normalize it. Mean centering the data was initially considered; however, it would not have been appropriate since it would give negative values, making it harder to interpret.

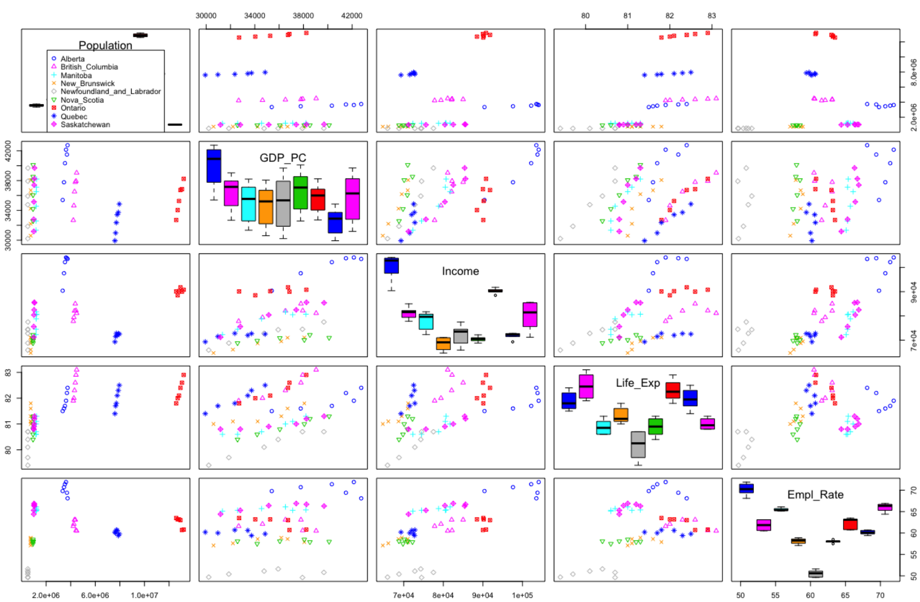


Figure 1. Scatterplot for Each Pair of Response and Explanatory Variable

After looking through this scatterplot, we can summarize these features:

When paying attention to population and other variables, we found that Ontario is the most populous province in Canada, with far more people than any other province. And then it comes to Quebec. Coming in third and fourth are British Columbia and Alberta. The two are relatively close in population size. Lastly, the remaining provinces are exactly close to each other and the amount is really low compared to four populous provinces. It is because of this large population gap

that it is difficult to see a linear relationship between population and other variables together with years for each province in the scatterplots.
Regarding GDP per capita, Alberta's GDP per capita is much higher than the average of other provinces. Quebec, on the other hand, is well below the national average. The differences in GDP per capita in the other provinces are not so significant. Only by seeing the scatterplots, seems GDP_PC has some relatively apparent linear relationship with life expectancy. As for other variables, there are no clear linear relationships in these scatterplots.
What can be determined is that life expectancy and employment rate have nothing to do with each other.This may be an important point for us to determine the best model at a later step.

b. Methodology and Analysis

To get a better sense of the relationship between the response variable and all the explanatory variables, the collected data was first visualized with a matrix of scatterplots for each pair of response/explanatory variables, as shown in Figure.1. It appeared that the data was highly clustered in terms of province, and a linear relationship was observed between response variable and each continuous explanatory variable.

Based on findings from data visualization, it would make sense to include the categorical variable (*Province*) in the model due to the clustering of data. To further explore the explanatory power for the rest of continuous variables, a “Full” model including all possible variables was fitted in R, and the result is shown in Figure.2.

```
Call:
lm(formula = Population ~ ., data = population)

Residuals:
    Min      1Q  Median      3Q     Max 
-182535  -36958    5344   38075  155316 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -4.217e+06  8.469e+06  -0.498 0.621203    
GDP_PC                           -3.824e+00  1.688e+01  -0.227 0.821915    
Income                            9.841e+00  7.129e+00   1.380 0.174951    
Life_Exp                          1.304e+05  9.756e+04   1.337 0.188710    
Empl_Rate                        -5.310e+04  1.797e+04  -2.955 0.005164 ** 
ProvinceBritish_Columbia          4.313e+05  1.171e+05   3.684 0.000665 ***
ProvinceManitoba                 -2.289e+06  1.647e+05 -13.896  < 2e-16 ***
ProvinceNew_Brunswick            -3.094e+06  2.512e+05 -12.314 2.33e-15 ***
ProvinceNewfoundland_and_Labrador -3.616e+06  4.471e+05  -8.087 5.04e-10 ***
ProvinceNova_Scotia              -2.864e+06  3.048e+05  -9.398 8.80e-12 ***
ProvinceOntario                   8.922e+06  1.006e+05  88.683  < 2e-16 ***
ProvinceQuebec                    3.902e+06  1.535e+05  25.418  < 2e-16 ***
ProvinceSaskatchewan             -2.474e+06  1.515e+05 -16.331  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69980 on 41 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9997 
F-statistic: 1.44e+04 on 12 and 41 DF,  p-value: < 2.2e-16
```

Figure 2. Summary Statistics of the “Full” Model

According to the p-value in hypothesis testing in the model fitted, only the employment rate seemed to have enough explanatory power with a p-value of 0.005164 that fell into the upper 5% tail of the t- distribution. However, a clear linear relationship was also observed for the rest of continuous variables which was not reflected in this “Full” model. Such findings were suspected to be a result from the correlation and difference in magnitude between the explanatory variables. Therefore, four separate linear models were added to the analysis to further examine the explanatory power of each continuous explanatory variable, where the adjusted R^2 and p-values of these models would be compared and analyzed. These four models included the categorical variable (*Province*), plus only one continuous variable in each of them: *GDP_PC*, *Income*, *Life_Exp*, and *Empl_Rate.* The fitted models, *gdp_model, income_model, life_exp_model, and empl_model*, were shown from Figure.3 to Figure.6.

```
Call:
lm(formula = Population ~ GDP_PC + Province, data = population)

Residuals:
    Min      1Q  Median      3Q     Max 
-232594  -46819  -11810   53004  257690 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       2.697e+06  1.974e+05  13.664  < 2e-16 ***
GDP_PC                            2.172e+01  4.851e+00   4.478 5.29e-05 ***
ProvinceBritish_Columbia          8.562e+05  5.464e+04  15.670  < 2e-16 ***
ProvinceManitoba                 -2.260e+06  5.711e+04 -39.578  < 2e-16 ***
ProvinceNew_Brunswick            -2.702e+06  5.792e+04 -46.640  < 2e-16 ***
ProvinceNewfoundland_and_Labrador -2.944e+06  5.709e+04 -51.565  < 2e-16 ***
ProvinceNova_Scotia              -2.553e+06  5.441e+04 -46.929  < 2e-16 ***
ProvinceOntario                   9.404e+06  5.598e+04 167.992  < 2e-16 ***
ProvinceQuebec                    4.359e+06  6.319e+04  68.985  < 2e-16 ***
ProvinceSaskatchewan             -2.455e+06  5.578e+04 -44.011  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89820 on 44 degrees of freedom
Multiple R-squared:  0.9996,	Adjusted R-squared:  0.9995 
F-statistic: 1.166e+04 on 9 and 44 DF,  p-value: < 2.2e-16
```

Figure 3. Summary Statistic of gdp_model

```
Call:
lm(formula = Population ~ Income + Province, data = population)

Residuals:
    Min      1Q  Median      3Q     Max 
-294673  -42530  -11567   32095  307440 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       2.378e+06  4.256e+05   5.586 1.36e-06 ***
Income                            1.185e+01  4.225e+00   2.805  0.00747 ** 
ProvinceBritish_Columbia          1.004e+06  9.882e+04  10.162 4.06e-13 ***
ProvinceManitoba                 -2.104e+06  1.100e+05 -19.126  < 2e-16 ***
ProvinceNew_Brunswick            -2.441e+06  1.460e+05 -16.718  < 2e-16 ***
ProvinceNewfoundland_and_Labrador -2.719e+06  1.315e+05 -20.685  < 2e-16 ***
ProvinceNova_Scotia              -2.274e+06  1.385e+05 -16.421  < 2e-16 ***
ProvinceOntario                   9.428e+06  7.154e+04 131.786  < 2e-16 ***
ProvinceQuebec                    4.535e+06  1.334e+05  33.998  < 2e-16 ***
ProvinceSaskatchewan             -2.308e+06  1.028e+05 -22.452  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99820 on 44 degrees of freedom
Multiple R-squared:  0.9995,	Adjusted R-squared:  0.9994 
F-statistic:  9438 on 9 and 44 DF,  p-value: < 2.2e-16
```

Figure 4. Summary Statistic of income_model

```
Call:
lm(formula = Population ~ Life_Exp + Province, data = population)

Residuals:
    Min      1Q  Median      3Q     Max 
-198293  -42673    2946   39645  196603 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -12476786    2439554  -5.114 6.60e-06 ***
Life_Exp                            195920      29791   6.577 4.79e-08 ***
ProvinceBritish_Columbia            664836      47709  13.935  < 2e-16 ***
ProvinceManitoba                  -2171428      53494 -40.592  < 2e-16 ***
ProvinceNew_Brunswick             -2706149      47530 -56.935  < 2e-16 ***
ProvinceNewfoundland_and_Labrador -2714556      67746 -40.070  < 2e-16 ***
ProvinceNova_Scotia               -2431084      53494 -45.446  < 2e-16 ***
ProvinceOntario                    9227729      46133 200.026  < 2e-16 ***
ProvinceQuebec                     4180850      44501  93.950  < 2e-16 ***
ProvinceSaskatchewan              -2374060      51639 -45.974  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76960 on 44 degrees of freedom
Multiple R-squared:  0.9997,	Adjusted R-squared:  0.9996 
F-statistic: 1.588e+04 on 9 and 44 DF,  p-value: < 2.2e-16
```

Figure 5. Summary Statistic of life_exp_model

```
Call:
lm(formula = Population ~ Empl_Rate + Province, data = population)

Residuals:
    Min      1Q  Median      3Q     Max 
-241995  -38387    7886   52550  226116 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        7429444    1103078   6.735  2.8e-08 ***
Empl_Rate                           -55024      15700  -3.505  0.00106 ** 
ProvinceBritish_Columbia            317836     142770   2.226  0.03117 *  
ProvinceManitoba                  -2627797      92642 -28.365  < 2e-16 ***
ProvinceNew_Brunswick             -3482965     197858 -17.603  < 2e-16 ***
ProvinceNewfoundland_and_Labrador -4133947     313936 -13.168  < 2e-16 ***
ProvinceNova_Scotia               -3298301     199366 -16.544  < 2e-16 ***
ProvinceOntario                    8879256     134612  65.962  < 2e-16 ***
ProvinceQuebec                     3641432     167940  21.683  < 2e-16 ***
ProvinceSaskatchewan              -2779142      86272 -32.214  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95820 on 44 degrees of freedom
Multiple R-squared:  0.9995,	Adjusted R-squared:  0.9994 
F-statistic: 1.024e+04 on 9 and 44 DF,  p-value: < 2.2e-16
```

Figure 6. Summary Statistic of empl_model

Between these four fitted models, the adjusted R^2 had no apparent difference and they all achieved at least 0.999. Additionally, each explanatory variable showed enough explanatory power according to their t-test, which all fell into the extreme upper tail of the t-distribution. However, some of the variables had extremely large coefficients due to their much smaller magnitude compared to the response variable. Lastly, all the signs on coefficients in individual models agree with the signs on all coefficients in the “Full” model.

Based on the above findings, it appeared that either one of the four models would be an acceptable model for this study. It was reasonable to include all the explanatory variables in the “Full” model from the point moving forward. Also, since the adjusted R^2 in these models were extremely close to 1, it is decided there is no need to add any interaction term. However, it would be appropriate to normalize all the data values between 0 and 1, so that the potential influence of the magnitude difference could be eliminated.

To select the explanatory variables for the final model, Mallows' C_p and adjusted R^2 were used as selection criterion, based on the “Full” model with all continuous data columns normalized. The best model for each variable size p is shown in Figure.7.

	(Intercept)	GDP_PC	Income	Life_Exp	Empl_Rate	ProvinceBritish_Columbia	ProvinceManitoba	ProvinceNew_Brunswick	ProvinceNewfoundland_and_Labrador	ProvinceNova_Scotia	ProvinceOntario	ProvinceQuebec	ProvinceSaskatchewan
1	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
2	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
3	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
4	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
5	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
6	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
7	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
8	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
9	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
12	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Figure 7. Best Model Selections of Various Model Size P

Mallows' C_p was plotted against size p as shown in Figure.8, and adjusted R^2 were listed in Table.1. Based on these statistics, the best model could be *Population~Income+Life_Exp+Empl_Rate+Province (p =11),* since it has the highest adjusted R^2 value and smallest difference between C_p and size p. However, the model *Population~Life_Exp+Empl_Rate+Province (p = 10)* would also be acceptable based on these two statistics and with fewer explanatory variables.

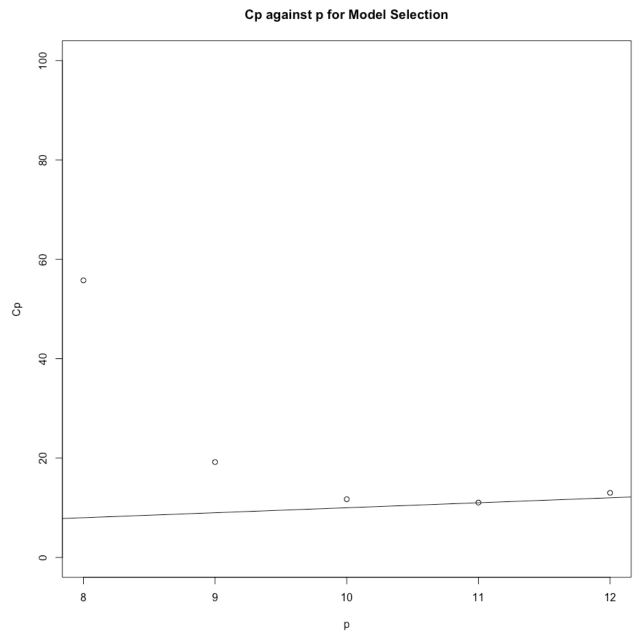


Figure 8. Mallows' Cp against Size p for Model Size Selection

Table 1. Adjusted R2 for the Best Models of Model Size from 8 to 12

p	8	9	10	11	12
Adjusted R^2	0.9993748	0.9996293	0.9996883	0.9997004	0.9996934

An additional train/test based method was conducted to make the final decision and to eliminate the scenario of overfitting. A 2-fold-test was applied to the data set, and the RMSE was calculated for the two models described above. According to the RMSE value, there was no apparent difference between RMSE of these two models. Thus, we chose the model with fewer parameters, which is *Population~Life_Exp+Empl_Rate+Province*.

Table 2. RMSE for Model Size from 8 to 12 in 2-Fold Test

p	8	9	10	11	12
RMSE	0.003846629	0.002450979	0.002399223	0.002297815	0.002250364

The summary statistics of the final model is shown in Figure.9, note that all data was normalized between 0 and 1 before fitting the linear model.

```
Call:
lm(formula = Population ~ Life_Exp + Empl_Rate + Province, data = norm_pop)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.013329 -0.003155  0.001012  0.002822  0.011928 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        0.266186   0.021611  12.317 1.08e-15 ***
Life_Exp                           0.050850   0.008236   6.174 2.04e-07 ***
Empl_Rate                         -0.064167   0.021011  -3.054  0.00387 ** 
ProvinceBritish_Columbia           0.029310   0.008322   3.522  0.00103 ** 
ProvinceManitoba                  -0.186590   0.006337 -29.443  < 2e-16 ***
ProvinceNew_Brunswick             -0.249220   0.012212 -20.408  < 2e-16 ***
ProvinceNewfoundland_and_Labrador -0.273669   0.020112 -13.607  < 2e-16 ***
ProvinceNova_Scotia               -0.228554   0.012660 -18.053  < 2e-16 ***
ProvinceOntario                    0.706012   0.007874  89.667  < 2e-16 ***
ProvinceQuebec                     0.300824   0.010002  30.077  < 2e-16 ***
ProvinceSaskatchewan              -0.200885   0.005823 -34.500  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.005566 on 43 degrees of freedom
Multiple R-squared:  0.9997,	Adjusted R-squared:  0.9997 
F-statistic: 1.7e+04 on 10 and 43 DF,  p-value: < 2.2e-16
```

Figure 9. Summary Statistics of the Final Model Based on Normalized Data

C. Further Analysis (Discussion)

Given our initial analysis of the effect of the explanatory variables on the population, further discussion was done in order to continue our analysis. As mentioned previously, we get the best model to be an additive model where Population is with respect to the Life expectancy variable, employment rate variable, and the province categorical variable. According to the model summary, it appears that life expectancy has the smallest p-value, suggesting that it has the most explanatory power of all the variables. This is validated by the summaries of the models exploring each of the explanatory variables, where *Population~Life_Exp+Province* has the highest adjusted R^2 among those models. This observation makes sense since the death rate will directly affect the population number, whilst the other factors affect the population in other indirect ways, perhaps by causing migration or affecting birth rates. Also notice that our final model has an extremely high adjusted R^2 of 0.997, suggesting that 99.7% of the variation in the population variable is explained by life expectancy, the employment rate, and provinces. This high value may be due to the fact that our data came from Statistics Canada. The quality of the data from Statistics Canada is quite high since they can get data from the entire population of Canada and their data may be more robust to outliers and errors.

In our analysis of the data, we chose to ignore the year effect on the variables. The Year variable directly affects all the variables chosen, and hence not an interesting explanatory variable to consider. Taking different years gave us more observations to consider, however, the time effect was not of interest in our study as we wanted to consider other driving forces in the changes of the population.

3. Conclusion

According to our analysis, the best fitted model that explains the relationship between population and the suitable explanatory variables is shown as below:

Y = 0.266 + 0.051*X1 - 0.064*X2 + 0.029*X3 - 0.187*X4 - 0.249*X5 - 0.273*X6 - 0.229*X7 + 0.706*X8 + 0.301*X9 - 0.201*X10

Y: Population
X1: Life expectancy
X2: Employment Rate
X3: Province: British Columbia
X4: Province: Manitoba
X5: Province: New Brunswick
X6: Province: Newfoundland and Labrador
X7: Province: Nova Soctia
X8: Province: Ontario
X9: Province: Quebec
X10: Province: Saskatchewan

There are positive linear relationships between population and Life expectancy / British Columbia province / Ontario province / Quebec province, while there are negative relationships between population and Employment Rate / Manitoba province / New Brunswick province / Newfoundland and Labrador province / Nova Scotia province / Saskatchewan province.

For provinces, only three provinces have positive relationships with the population. The three provinces are Ontario, Quebec, and British Columbia. The three provinces not only have the largest population, but also have three of the most developed cities in Canada (Toronto, Montreal, Vancouver), which is also a reason why the three provinces are popular for attracting people and increasing the population.

Besides, it is determined that the population has a linear relationship with life expectancy, employment rate, and provinces. The categorical variable (province) plays a leading role. The p-value of the variable in each province is lower than that of other variables, which means that this predictive variable is significant to the model. One of the most popular provinces is Ontario. From the model, we can also know that life expectancy positively impacts the population. The main reason is that the higher the life expectancy of the local population, the more the elderly will be and the higher the total population will be. Finally, the coefficient of the employment rate is negative in the model, which means that the population will decrease with the increase in the employment rate. On the contrary, it should be that the number of people will affect the employment rate. Population growth will lead to increased competition for employment.

The intensification of job competition makes it more difficult to find a job, thus reducing the employment rate. We have noticed the limitations of this variable, so we can further explore and analyze this problem in the future.

Given our conclusion, there are many other ways that can improve or add to our insights on the factors that affect the population. For instance, further consideration of larger datasets and more explanatory variables will provide a deeper understanding of the factors that affect the population. Exploring explanatory variables like the housing benchmark prices, education rate, immigration rates, and birth rates would have provided better insight into the rise and fall of the population, and perhaps their effect on other explanatory variables as well. Another interesting consideration would have been to look for variables that would decrease the population. And the adjusted R^2 we got in part b are all above 0.999, which means that the covariates can jointly explain the variation in the outcome Y. This means Y can be accurately predicted using the covariates. Maybe the reason is that the time span is very small. If more years were added as observations, a time effect would have been important to consider, and perhaps major world events can also be taken into account in the analysis.