STAT 306 Group Project Report

group: E2
group members: Manqin Cai, Alice Duan, Jack Fan

Introduction

1. Background Introduction:

Right now, countries worldwide are facing stagnant population growth and plummeting fertility rates. For any country in the world, the population is a manifestation of national conditions and national strength and one of a nation's most important strategic resources. Population affects every aspect of our lives. So the significant changes in fertility rates affect individuals, families, culture, the economy, the environment and politics.

Research Question:

In today's world, the fertility rates of the young generation have dropped significantly compared with the previous data, which has led to the aggravation of the aging of the population in some regions and is also a tremendous challenge for all countries in the world. Our research questions will look for the effects of GDP, population, province, marriage rate, average income and employment rate on fertility and the relationship between them.

Preliminary guesses about factors influencing fertility rates

① **GDP:** GDP may have a negative relationship with fertility rates.

② **Population:** Population may have a positive relationship with fertility rates.

③ **Province:** (We focus on the four provinces in Canada: British Columbia/ Quebec/ Alberta/ Ontario) Province(or regional difference) may impact fertility rates.

④ **Marriage rate:** Marriage rate may positively correlate with fertility rates.

⑤ **Income:** Income may have a positive relationship with fertility rates.

⑥ **Employment Rates:** The employment rate may positively affect fertility rates.

2. Source of the data:

Response variable:

Fertility rate(FR) from Statistics Canada
https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310041801&pickMembers%5B0%5D=1.1&cubeTimeFrame.startYear=2000&cubeTimeFrame.endYear=2020&referencePeriods=20000101%2C20200101

Categorical variable:

The region in Canada from Statistics Canada
BC: British Columbia
QC: Quebec
AB: Alberta
ON: Ontario

Explanatory variable:

Population from Statistics Canada
https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901&cubeTimeFrame.startMonth=01&cubeTimeFrame.startYear=1990&cubeTimeFrame.endMonth=10&cubeTimeFrame.endYear=2021&referencePeriods=19900101%2C20211001

Average income from Statistics Canada
https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110023901&pickMembers%5B0%5D=1.13&pickMembers%5B1%5D=2.1&pickMembers%5B2%5D=3.1&pickMembers%5B3%5D=4.1&cubeTimeFrame.startYear=2000&cubeTimeFrame.endYear=2019&referencePeriods=20000101%2C20190101

GDP from Statistics Canada
https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610022101&pickMembers%5B0%5D=1.11&cubeTimeFrame.startYear=2000&cubeTimeFrame.endYear=2020&referencePeriods=20000101%2C20200101

marriage rate from Statistics Canada

https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710006001&pickMembers%5B0%5D=1.11&pickMembers%5B1%5D=3.1&pickMembers%5B2%5D=4.1&cubeTimeFrame.startYear=2000&cubeTimeFrame.endYear=2020&referencePeriods=20000101%2C20200101

employment rate from Statistics Canada
https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410002001

3. Description of the variables:

A description of the variables measured (including when, where, how, in what units, plus any other important information).

The variables are collected about four major areas in Canada: BC, QC, AB, ON from 2000 to 2020.

Response variable:

Total fertility rate (N/A): an estimate of the average number of live births a female can be expected to have in her lifetime, based on the age-specific fertility rates (ASFR) of a given year. The total fertility rate (TFR) = (SUM of a single year of age-specific fertility rates) /1000

Explanatory variables:

GDP (million-dollar): Gross domestic product at market prices

Population (N/A): The number of people in the area

Province (N/A): BC, QC, AB, ON

Marriage rate (%): the number of people with legal marital status in the population

Income (dollar): Average income of the population 16 years of age and over

Employment rate (%): the number of persons employed expressed as a percentage of the population 15 years of age and over

Analysis

A summary of the data:

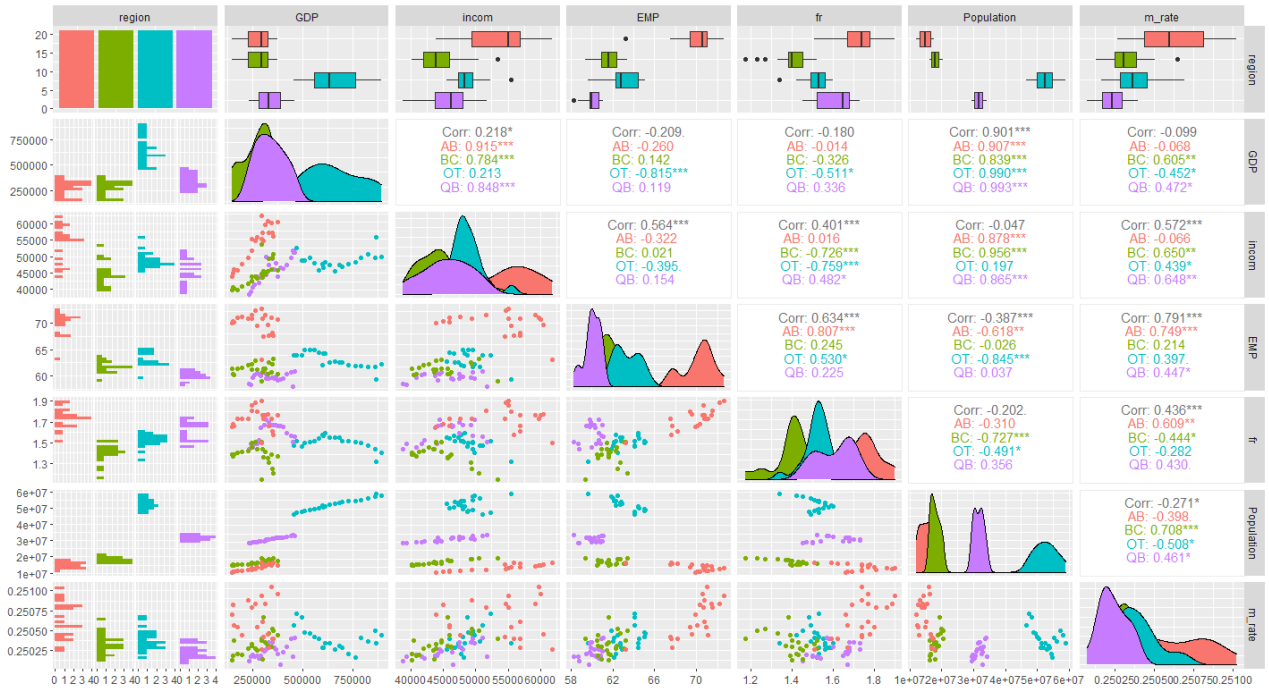

figure1: summary plot of all the variables

According to the graph, we can get that under the different regions, the fertility rate has strong patterns with other variables we choose. And next, we will try to fit models to find out the relationships between these variables.

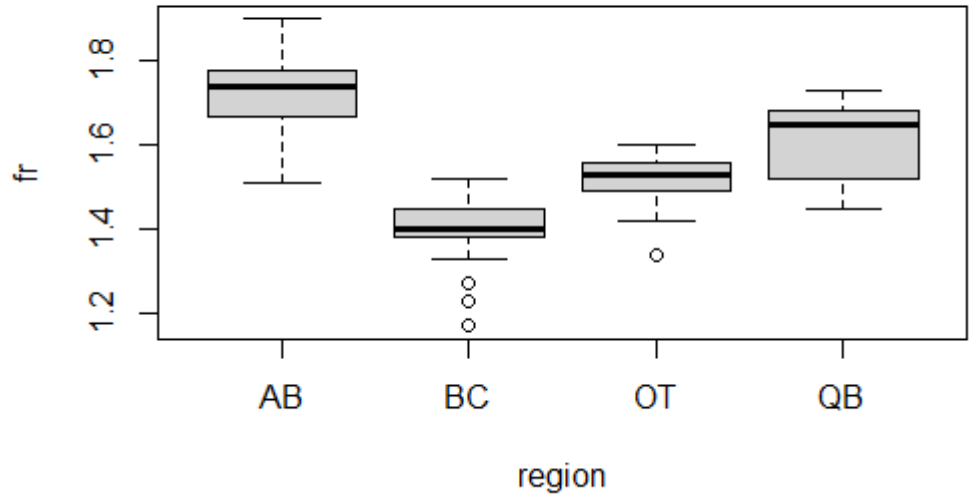

figure2: box plot of fertility rate in four regions

This box plot has been computed to visualize the difference in fertility rates between different regions. According to the graph, we can see that Alberta has an overall highest median, and British Columbia has the lowest fertility rate. Three outliers were found in BC's data, and one outlier was found in the OT data. Therefore, the different regions seem to be a factor that is connected with the fertility rate.

Methodology:

Model 1: the full linear model

A linear model containing all predictors was fitted first. Below is the summary of the model.

```
Call:
lm(formula = fr ~ ., data = p3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.190120 -0.048928  0.004699  0.044362  0.146499

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.115e+01  1.783e+01   1.747  0.08468 .
regionBC    -2.654e-02  6.753e-02  -0.393  0.69538
regionOT     8.990e-01  4.686e-01   1.918  0.05888 .
regionQB     6.153e-01  2.213e-01   2.781  0.00685 **
GDP          7.733e-07  3.716e-07   2.081  0.04083 *
incom        7.823e-07  2.769e-06   0.283  0.77829
EMP          2.877e-02  9.226e-03   3.118  0.00258 **
Population  -3.323e-08  1.684e-08  -1.973  0.05214 .
m_rate      -1.244e+02  7.232e+01  -1.720  0.08950 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07267 on 75 degrees of freedom
Multiple R-squared:  0.7875,    Adjusted R-squared:  0.7648
F-statistic: 34.74 on 8 and 75 DF,  p-value: < 2.2e-16
```

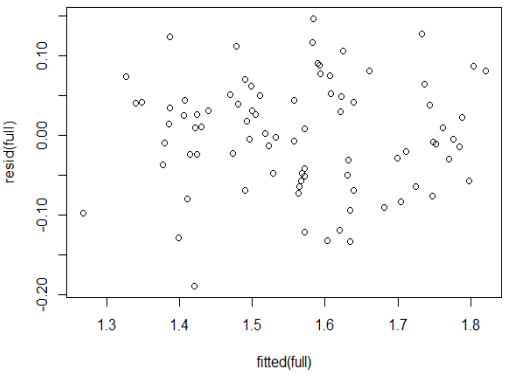

figure3: summary of model1 & corresponding residual plot

Although the residual plot looks randomly scattered around, we can see that some of the explanatory variables have non-linear relationships with the fertility rate from *figure 1*. Therefore, we can add some terms to improve the model according to the scatterplots.

Model 2: the full polynomial model

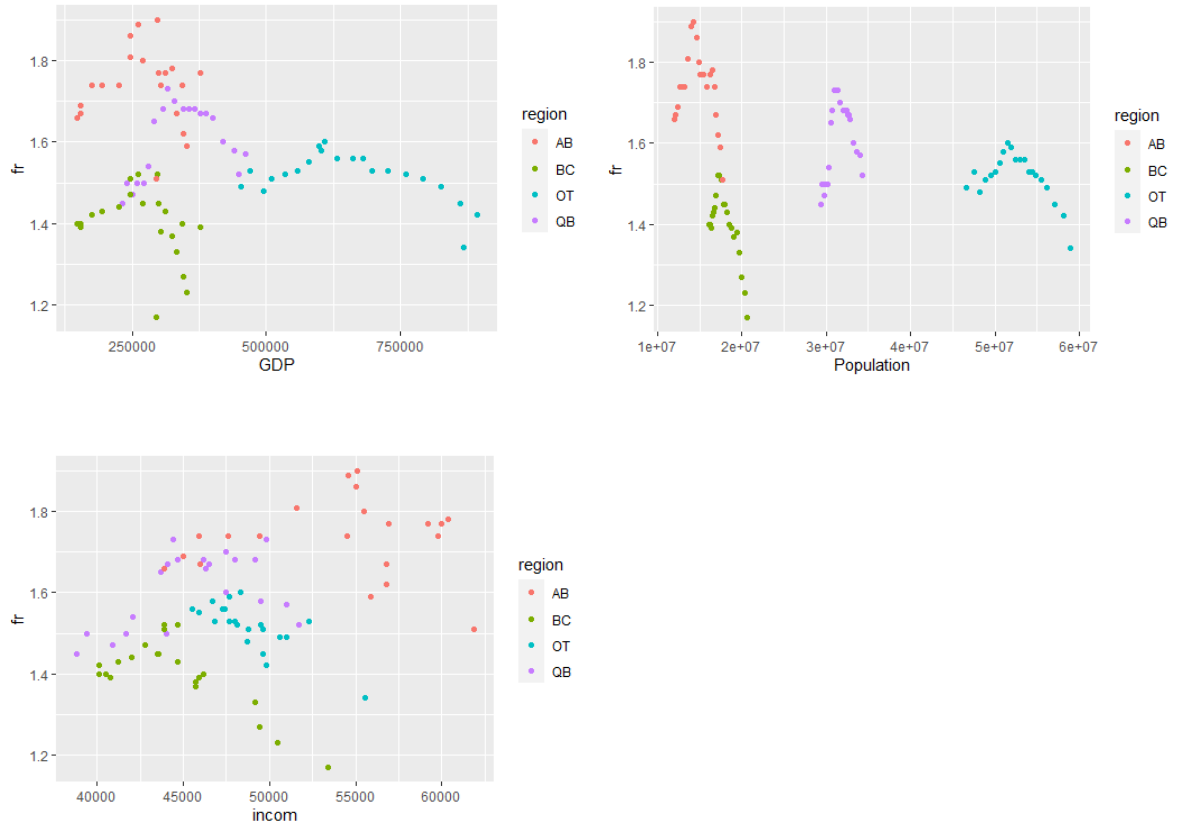

figure4: scatter plots of (1)GDP (2)Population (3)Income

As shown in figure4, GDP, population and income do not seem to have a linear relationship with fertility rate. The scattered plot is heavily curved. Thus, we decide to add a low order polynomial to each of the three predictors.

Below is the summary of the full model after adding three additional squared terms.

```
Call:
lm(formula = fr ~ . + I(GDP^2) + I(incom^2) + I(Population^2),
    data = p3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.149750 -0.035047  0.004819  0.043315  0.125917

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.821e+01  1.698e+01   1.661 0.101009
regionBC         8.351e-02  8.367e-02   0.998 0.321574
regionOT         8.949e-01  4.749e-01   1.884 0.063550 .
regionQB         9.799e-01  2.567e-01   3.817 0.000283 ***
GDP              2.673e-06  5.617e-07   4.758 9.80e-06 ***
incom            9.040e-07  2.806e-05   0.032 0.974385
EMP              1.648e-02  8.846e-03   1.863 0.066564 .
Population      -1.316e-07  2.757e-08  -4.776 9.16e-06 ***
m_rate          -1.067e+02  6.921e+01  -1.541 0.127601
I(GDP^2)        -2.719e-12  6.328e-13  -4.297 5.34e-05 ***
I(incom^2)       5.762e-11  2.738e-10   0.210 0.833911
I(Population^2)  1.547e-15  3.371e-16   4.587 1.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06491 on 72 degrees of freedom
Multiple R-squared:  0.8373,    Adjusted R-squared:  0.8124
F-statistic: 33.67 on 11 and 72 DF,  p-value: < 2.2e-16
```

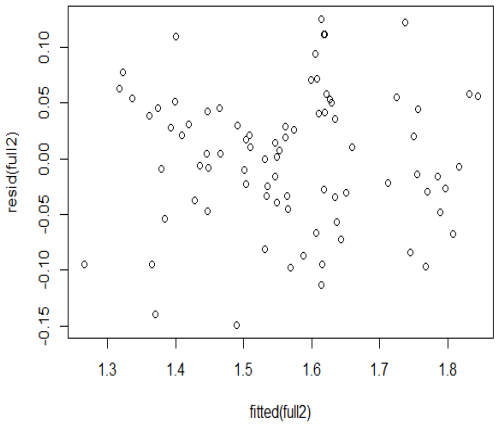

figure5: summary of model2 & corresponding residual plot

The adjusted R square had increased from 0.7648 to 0.8124, suggesting an improvement in the second model. Also, the residual plot looks patternless, which is excellent.
But since some of the coefficients are insignificant, we decided to make some further improvements to our model.

Model 3: removing the square of income

First, we can see that the I(income^2) in the previous model is insignificant, so we try to remove the I(income^2) term in model3.

Below is the summary of the model3.

```
> reg <- lm(fr~.+I(GDP^2)+I(Population^2), data = p3)
> summary(reg)

Call:
lm(formula = fr ~ . + I(GDP^2) + I(Population^2), data = p3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14873 -0.03536  0.00522  0.04238  0.12354

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.870e+01  1.671e+01   1.718 0.090061 .
regionBC         8.041e-02  8.182e-02   0.983 0.328976
regionOT         9.028e-01  4.703e-01   1.920 0.058793 .
regionQB         9.779e-01  2.548e-01   3.837 0.000262 ***
GDP              2.639e-06  5.353e-07   4.930 5.00e-06 ***
incom            6.750e-06  3.907e-06   1.728 0.088282 .
EMP              1.616e-02  8.659e-03   1.866 0.066005 .
Population      -1.311e-07  2.728e-08  -4.808 7.97e-06 ***
m_rate          -1.091e+02  6.777e+01  -1.610 0.111642
I(GDP^2)        -2.682e-12  6.036e-13  -4.443 3.10e-05 ***
I(Population^2)  1.534e-15  3.293e-16   4.657 1.40e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06448 on 73 degrees of freedom
Multiple R-squared:  0.8372,    Adjusted R-squared:  0.8149
F-statistic: 37.53 on 10 and 73 DF,  p-value: < 2.2e-16
```

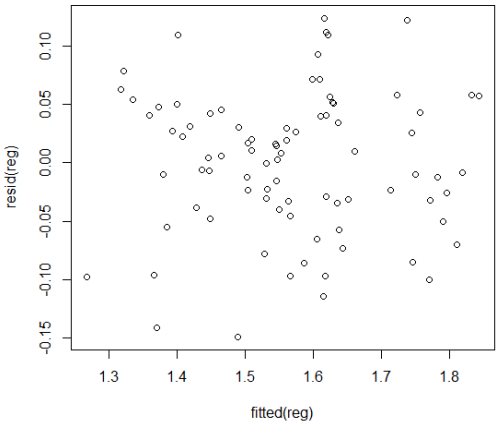

figure6: summary of model3 & corresponding residual plot

With a patternless residual plot, an increase in adjusted R square is observed from 0.8124 to 0.8149, suggesting an improvement in the model. Thus, the square term of income can be removed from the model.

Model 4: dropping out the marriage rate

Looking at the other insignificant variables, we first decided to check the relationship between marriage and fertility rates.

```
> lm <- lm(fr~region+m_rate, data=p3)
> summary(lm)

Call:
lm(formula = fr ~ region + m_rate, data = p3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23444 -0.05447  0.01410  0.06180  0.13623

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.05305   15.43921  -1.493   0.1394
regionBC     -0.30780    0.03196  -9.630 5.76e-15 ***
regionOT     -0.19552    0.03000  -6.517 6.17e-09 ***
regionQB     -0.09009    0.03537  -2.547   0.0128 *
m_rate       98.90828   61.60191   1.606   0.1124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08429 on 79 degrees of freedom
Multiple R-squared:  0.6989,    Adjusted R-squared:  0.6836
F-statistic: 45.84 on 4 and 79 DF,  p-value: < 2.2e-16
```

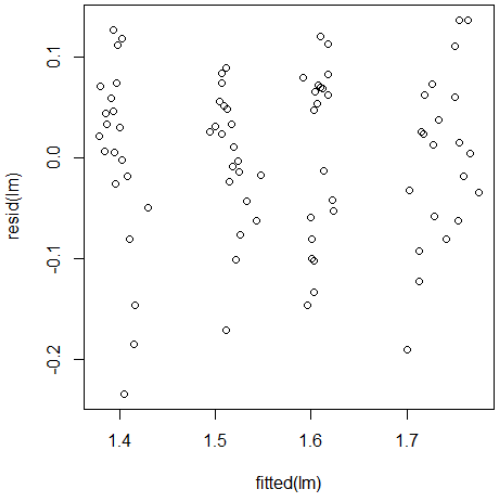

figure7: summary of the linear relationship between fertility rate and marriage rate

From the summary, we observed that the marriage rate is not significant, and the scatterplot is very strange, indicating that the effect of the marriage rate is very weak. So, we decided to ignore it and refit the model.

```
Call:
lm(formula = fr ~ . + I(GDP^2) + I(Population^2) - m_rate, data = p3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.150844 -0.040960  0.007874  0.047104  0.122491

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.819e+00  6.884e-01   2.642 0.010053 *
regionBC        1.544e-02  7.194e-02   0.215 0.830619
regionOT        6.557e-01  4.493e-01   1.459 0.148673
regionQB        8.242e-01  2.388e-01   3.451 0.000925 ***
GDP             2.667e-06  5.408e-07   4.931 4.88e-06 ***
incom           3.969e-06  3.542e-06   1.121 0.266085
EMP             1.009e-02  7.878e-03   1.280 0.204393
Population     -1.240e-07  2.720e-08  -4.559 1.99e-05 ***
I(GDP^2)       -2.830e-12  6.030e-13  -4.693 1.21e-05 ***
I(Population^2)  1.532e-15  3.328e-16   4.602 1.70e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06517 on 74 degrees of freedom
Multiple R-squared:  0.8314,    Adjusted R-squared:  0.8109
F-statistic: 40.54 on 9 and 74 DF,  p-value: < 2.2e-16
```

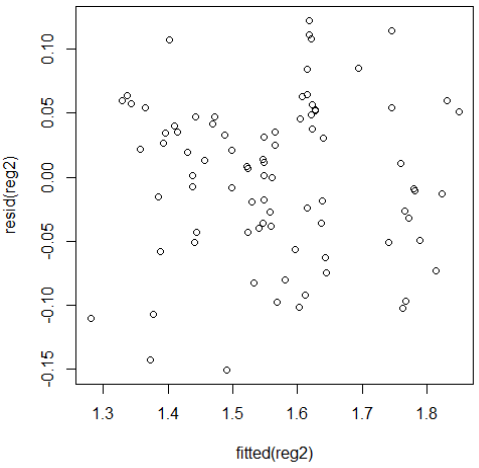

figure8: summary of model4 & corresponding residual plot

Model 5: dropping out of the income

Since the coefficient of income in the previous model is pretty small, the p-value is insignificant. We decided to check the linear relationship between income and fertility rate.

```
Call:
lm(formula = fr ~ region + incom, data = p3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.205044 -0.041796  0.006764  0.057872  0.167050

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.881e+00  1.309e-01  14.365  < 2e-16 ***
regionBC    -3.624e-01  3.437e-02 -10.545  < 2e-16 ***
regionOT    -2.332e-01  2.895e-02  -8.056 6.77e-12 ***
regionQB    -1.509e-01  3.299e-02  -4.575 1.74e-05 ***
incom       -2.686e-06  2.405e-06  -1.117    0.267
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08499 on 79 degrees of freedom
Multiple R-squared:  0.6939,    Adjusted R-squared:  0.6784
F-statistic: 44.77 on 4 and 79 DF,  p-value: < 2.2e-16
```

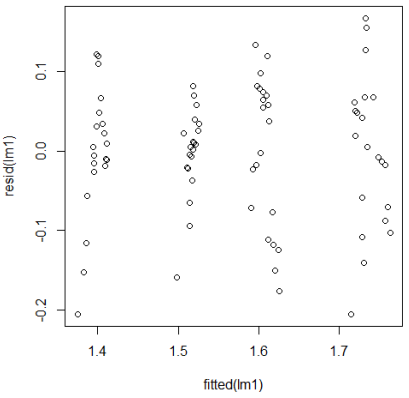

figure9: summary of the linear relationship between income and marriage rate

The same as the marriage rate, since the p-value of income is insignificant and the corresponding coefficient is minimal, we decided to remove income in our model.

Below is a summary of model 5.

```
Call:
lm(formula = fr ~ . + I(GDP^2) + I(Population^2) - m_rate - incom,
    data = p3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.146921 -0.042195  0.004296  0.046646  0.127640

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.718e+00  6.837e-01   2.513  0.01411 *
regionBC        -4.068e-02  5.173e-02  -0.786  0.43411
regionOT         4.651e-01  4.165e-01   1.116  0.26779
regionQB         6.718e-01  1.966e-01   3.417  0.00103 **
GDP              2.630e-06  5.407e-07   4.864 6.19e-06 ***
EMP              1.172e-02  7.755e-03   1.511  0.13493
Population      -1.077e-07  2.303e-08  -4.677 1.26e-05 ***
I(GDP^2)        -2.727e-12  5.970e-13  -4.568 1.90e-05 ***
I(Population^2)  1.352e-15  2.924e-16   4.625 1.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06528 on 75 degrees of freedom
Multiple R-squared:  0.8285,    Adjusted R-squared:  0.8102
F-statistic: 45.29 on 8 and 75 DF,  p-value: < 2.2e-16
```

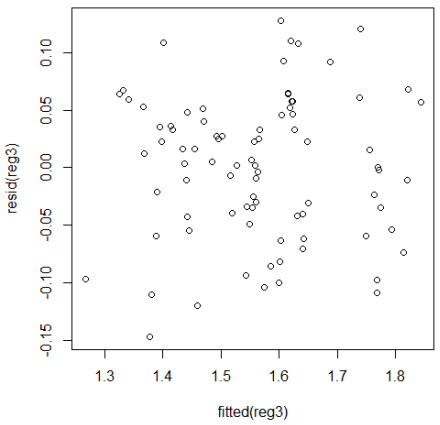

figure10: summary of model5 & corresponding residual plot

Also, we noticed that the EMP(employment rates) in our model is insignificant. A linear model is applied to the employment and fertility rates to check their relationship.

```
Call:
lm(formula = fr ~ region + EMP, data = p3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22298 -0.05527  0.01596  0.05777  0.13026

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.40374    0.40830  -0.989  0.32576
regionBC    -0.08528    0.05322  -1.602  0.11309
regionOT    -0.01459    0.04522  -0.323  0.74784
regionQB     0.17176    0.06162   2.787  0.00665 **
EMP          0.03063    0.00584   5.245 1.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07377 on 79 degrees of freedom
Multiple R-squared:  0.7694,    Adjusted R-squared:  0.7577
F-statistic: 65.88 on 4 and 79 DF,  p-value: < 2.2e-16
```

figure11: model summary of the linear regression between fertility rate and employment rate

The above model shows that the coefficient for the employment rate is significant, which does not suggest we simply drop it. But why does it become insignificant when adding other variables? This looks like there is collinearity. Next, we decided to compute the correlation matrix to find out the problem.

```
> cor(pdata)
                 fertility         gdp  gdp_square       income marrage_rate employment_rate  population population_square
fertility        1.0000000 -0.17991737 -0.20540614  0.401318896   0.43568438       0.6336404 -0.20212415      -0.192719071
gdp             -0.1799174  1.00000000  0.97861987  0.217902634  -0.09859621      -0.2090455  0.90094779       0.930625391
gdp_square      -0.2054061  0.97861987  1.00000000  0.160731363  -0.09703263      -0.1879266  0.86588907       0.914996285
income           0.4013189  0.21790263  0.16073136  1.000000000   0.57242334       0.5636203 -0.04729315      -0.004137495
marrage_rate     0.4356844 -0.09859621 -0.09703263  0.572423340   1.00000000       0.7907580 -0.27091810      -0.198743107
employment_rate  0.6336404 -0.20904553 -0.18792663  0.563620339   0.79075800       1.0000000 -0.38653388      -0.290170125
population      -0.2021242  0.90094779  0.86588907 -0.047293146  -0.27091810      -0.3865339  1.00000000       0.987527121
population_square -0.1927191  0.93062539  0.91499629 -0.004137495  -0.19874311      -0.2901701  0.98752712       1.000000000
```

figure12: summary of the correlation matrix

According to the correlation matrix, we found that the employment rate does not have any high correlation with other included variables. Thus there does not exist collinearity in the employment rate. To check if we should remove the employment rate in our model, we decided to run a revised model, model5-1, on that.

Below are the summary statistics for model5-1.

```
Call:
lm(formula = fr ~ . + I(GDP^2) + I(Population^2) - m_rate - incom -
    EMP, data = p3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.130952 -0.046060  0.003109  0.045797  0.133573

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.725e+00  1.547e-01  17.619  < 2e-16 ***
regionBC        -8.198e-02  4.430e-02  -1.851   0.0681 .
regionOT         8.050e-01  3.535e-01   2.277   0.0256 *
regionQB         8.025e-01  1.781e-01   4.507 2.35e-05 ***
GDP              3.088e-06  4.517e-07   6.836 1.79e-09 ***
Population      -1.292e-07  1.827e-08  -7.069 6.50e-10 ***
I(GDP^2)        -2.972e-12  5.794e-13  -5.129 2.16e-06 ***
I(Population^2)  1.472e-15  2.839e-16   5.185 1.74e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06583 on 76 degrees of freedom
Multiple R-squared:  0.8233,    Adjusted R-squared:  0.807
F-statistic: 50.58 on 7 and 76 DF,  p-value: < 2.2e-16
```

figure12: summary of model5-1 & corresponding residual plot

According to the summary, the adjusted r square had dropped from 0.8102 to 0.807, which shows that dropping the employment rate does not improve the model. Thus, we decided not to drop the EMP.

Model 6: dealing with collinearity between GDP and population

According to the figure11, the correlation between GDP and Population is 0.9009, which would cause collinearity in our model. To revise our model, we first divide the GDP data by the population to decrease the effect by changing its unit to GDP per person. Then we centred predictors on Population and modified GDP. After that, we refitted our model.

Below is a summary of the statistics of model 6.

```
Call:
lm(formula = fr ~ . + I(GDP/Population) + I((GDP/Population -
    mean(GDP/Population))^2) + I((Population - mean(Population))^2) -
    m_rate - incom - GDP, data = p3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.164020 -0.057284  0.003122  0.056555  0.131407

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 7.497e-01  6.064e-01   1.236 0.220201
regionBC                                   -5.569e-02  5.656e-02  -0.985 0.327993
regionOT                                    9.953e-01  3.444e-01   2.890 0.005034 **
regionQB                                    7.116e-01  1.859e-01   3.827 0.000267 ***
EMP                                         1.458e-02  8.269e-03   1.764 0.081873 .
Population                                 -2.980e-08  1.013e-08  -2.941 0.004353 **
I(GDP/Population)                           2.176e+01  6.886e+00   3.160 0.002275 **
I((GDP/Population - mean(GDP/Population))^2) -1.451e+03  6.726e+02  -2.158 0.034142 *
I((Population - mean(Population))^2)        2.790e-16  1.345e-16   2.074 0.041497 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07045 on 75 degrees of freedom
Multiple R-squared:  0.8003,    Adjusted R-squared:  0.779
F-statistic: 37.57 on 8 and 75 DF,  p-value: < 2.2e-16
```

figure14: summary of model 6

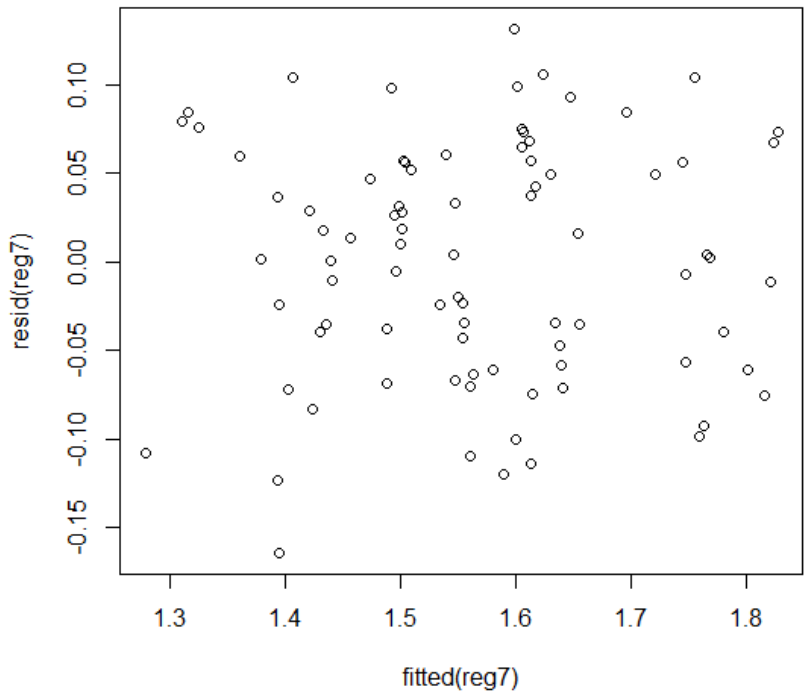

figure15: residual plot of model 6

We observed that the residual plot is randomly distributed, and the coefficients are almost significant, according to the above summary. Although the adjusted R square had dropped from 0.8102 to 0.779, seventy-five percent coverage of variance is quite enough in our case.

```
  (Intercept) regionBC regionOT regionQB incom   EMP Population m_rate I((Population - mean(Population))^2) I((GDP/Population - mean(GDP/Population))^2)
1        TRUE     TRUE    FALSE    FALSE FALSE FALSE      FALSE  FALSE                                FALSE                                          FALSE
2        TRUE    FALSE    FALSE     TRUE FALSE  TRUE      FALSE  FALSE                                FALSE                                          FALSE
3        TRUE     TRUE    FALSE     TRUE FALSE  TRUE      FALSE  FALSE                                FALSE                                          FALSE
4        TRUE     TRUE    FALSE     TRUE FALSE  TRUE      FALSE   TRUE                                FALSE                                          FALSE
5        TRUE     TRUE    FALSE     TRUE FALSE  TRUE      FALSE   TRUE                                FALSE                                          FALSE
6        TRUE    FALSE     TRUE     TRUE FALSE  TRUE       TRUE  FALSE                                 TRUE                                          FALSE
7        TRUE    FALSE     TRUE     TRUE FALSE  TRUE       TRUE  FALSE                                 TRUE                                           TRUE
8        TRUE    FALSE     TRUE     TRUE FALSE  TRUE       TRUE   TRUE                                 TRUE                                           TRUE
  I(GDP/Population)
1             FALSE
2             FALSE
3             FALSE
4             FALSE
5              TRUE
6              TRUE
7              TRUE
8              TRUE
> ss$cp
[1] 145.362870  20.210161  12.953797  12.739360  13.295484  12.273866   9.910372   8.786641
```

figure16: summary of cp

The leap package is used for checking the cp of the models. The cp summary shows that model 7 is the best, which includes employment rate, (Population - mean(Population))^2, ((GDP/Population) - mean(GDP/Population))^2 and GDP/Population. This also agreed with model 6. Consequently, model 6 is our final model.

Conclusion

Result Summary

Firstly, the box plot shows that the Province(regions) does impact the fertility rates. In other words, the fertility rates vary in different provinces in those four we selected from Canada. From the plot, we can conclude that Alberta has an overall highest median, and British Columbia has the lowest fertility rate, with three outliers found in BC and one outlier found in OT.

Then, for the research question: After model selection, the attributes that we selected are Employment Rate, Population and GDP per Person, which means the other two attributes(Income, Marriage Rate) have only a minor relationship with the fertility rates when holding others constant.

And for the preliminary guesses:

① GDP(per person) and Population: We can see from the model that GDP(per person) has a positive relationship with fertility rates, and Population has a moderate negative relationship with fertility rates, which are neither consistent with the preliminary guesses. After reviewing other academic studies about the impact of GDP(per person) and Population on the fertility rates(or birth rates), a large proportion of them is consistent with the preliminary guesses. So I think it is a limitation that the number of our regions selection is too small and cannot represent the world's condition.

② Province: As we demonstrated above, the Province(or regional difference) does impact fertility rates.

③ Marriage Rates and Income: After model selection, we know that Marriage Rates and Income have no or minor relationship with the fertility rates, which is not consistent but predictable from our preliminary guesses because we also could say they have a super small positive relationship with the fertility rates.

④ Employment Rates: We can see from the model that the Employment Rates have a positive relationship with fertility rates, which is consistent with the preliminary guess, which probably means the stability of life provided by the employer will make people more willing to have children.

Overall speaking, GDP(per person) and Employment Rates have a positive impact on fertility rates, Population has a negative effect on fertility rates, Province(regional difference) has an effect on fertility rates, Marriage Rates and Income have no effect on fertility rates.

Further Questions

“What could the government do to increase the fertility rates and restrain the aging tendency of the population?”

“What is behind millennials' growing resistance to childbearing?”

“What exactly will the reduction of the birth rates bring to society and the world?”