Multiple Linear Regression Model: Using Age, Sex, Income and Number of Marriages to Predict Total Number of Children

Yingren Luo(1004194873), LiSi Xuan(1004941325), Youwen Xu(1006675493), Alvin Gong(1005121556)

2020/10/16

Abstract

Population is a quite important thing that need to be monitored and forecast, since it has large impacts on many aspects. This study aims to investigate the relationship between respondents' age, sex, family income, number of marriages and total number of children in respondents family using 2017 General Social Survey (GSS) data on families. The stratified random sampling is used and data is collected via computer assisted telephone interviews (CATI). Using the cleaned data, we build a multiple linear regression model for these 5 variables. Our results show that increasing number of marriage and age has positive impact on the total number of children per family, while for sex and family income, male and low income would give lower number of total number of children comparing with female and high income. In conclusion, there are many factors influencing the number of children in each family such as age, sex, family income and number of marriage, but may not be limited to those factors. Therefore, further research need to be done to improve this model.

Introduction

Population, the total number of inhabitants in a particular area, has significant impacts, both positive and negative on society, economy and environment. The negative effects are usually caused by rapid population growth. For instance, for the aspect of social and economic effect, the rapid population growth affects the availability of education and health services, and increases the unemployment rate (13). For the environmental effect, the rapid population growth will generate more waste (6). It seems that there are many negative effects for rapid population growth. However, it does not mean that a lower population growth rate would be better. The reason is that lower population growth rate indicates that there will be fewer young people and more older people in the total population, which means that there might not be enough workers in the future. Therefore, population growth rate should be a meaningful object to study with.

In this case, we would like to focus on the population in Canada in 2017 and investigate the association between age, sex, income, marriage and number of children for each family using multiple linear regression model. In 2017, there were around 36.7 million people in Canada with a 0.96% growth rate (4), which could be considered as a low population growth rate. In addition, the population growth rate is affected by birth rate, death rate and also immigration rate. In 2017, the birth rate was 10.549 births per 1000 people with a 0.91% decline from 2016 (5), which could also be considered as a low birth rate. And, the fertility rate in Canada is 1.53 births per woman, which is below the population replacement rate (4), which shows the immigration plays an important role in population growth in Canada. However, nowadays immigration is heavily influenced by the COVID-19 pandemic (14), which means that fertility rate would affect population growth more than before. That is the reason why we want to demonstrate what might be related to the total number of the children per family.

Data

The data that we use to do the analysis comes from the 2017 General Social Survey (GSS), which is a sample survey with cross section design (11). For the sampling, the population is divided into 27 stratus based on the geographical areas and within each stratum, a simple random sample with replacement is performed to collect enough number of data reaching the minimum sample size for each province (11). For the collection, it is performed via computer assisted telephone interviews (CATI), and if there is no person within the household that meets eligibility criteria, it will terminate after initial questions (11). For people who refuse to do the survey, they contact those people two more times to explain the significance of the survey (11). They make appointments for people who are not convenient at that time (11). For non-responsive phone calls, they call back numerous times (11).

For the 2017 General Social Survey (GSS), the target population is all 15 years of age and older persons who are not institutionalized and living in the 10 provinces of Canada, Quebec, Ontario, British Columbia, Alberta, Saskatchewan, Manitoba, Newfoundland and Labrador, New Brunswick, Nova Scotia, and Prince Edward Island (11). The list of telephone numbers from both Statistic Canada and Address Register (AR) are used to create the survey frame (11). The target sample size is 20000, while the actual sample size is 20602 (11).

There are 5 variables in the cleaned dataset, **gss.csv**, which are **age**, **total_children**, **sex**, **number_marriages**, and **income_family** containing 20602 observations. The variable, **age**, gives the exact age of the respondent whose age is between 15 years old and 80 years old with decimal at time of the survey interview and gives 80.0 for all the respondents who are 80 years old or over(12). For the variable, **sex**, it gives the sex of respondents (12). The variable, **number_marriages**, gives the exact number of marriages the respondent has ever had within 3 times and gives 4 for all respondents who have 4 times or more marriages (12). For the variable, **income_family**, it gives the total incomes of all members in the census family received in 2016 from all kinds of sources before income taxes (12). It was separated into 6 groups, "Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "$100,000 to $ 124,999" and "$125,000 and more",in the original dataset (12). During the cleaning of data, we divide into two groups, "low income" and "high income" that "low income" contains the first three values and "high income" contains the last three values in the original dataset. The variable, **total_children**, gives the exact total number of children reported by the respondents within 6 children and gives 7 for all respondents who report 7 or more children (12). There is no "Valid skip", "Don't know", "Refusal", and "Not stated" for independent variables, while there are 18 for dependent variable which should show "NA" in the cleaned dataset. **age**, **sex**, **number_marriages**, and **income_family** are the independent variables, and **total_children** is the dependent variable.

For the drawbacks, as we mentioned above, the answer for the variable, **age**, **number_marriages**, and **total_children**, does not give all the exact numbers, which means that the data is not so accurate. In addition, we divide respondents generally into "high income" and "low income". Even within the same group, the difference between the actual income might be large. For the advantages, the questionnaire has a similar structure for all the questions and is easy to answer.

Model

We got our data from the website of **CHASS Data center**, and we have picked a **general social survey** from **Canada**, which is about **family**, and it was collected in **2017**. The **raw data** has **20602** observations of **461** variables.

Then we **cleaned** the data and just left **five columns**, which are **age**, **total_children**, **sex**, **number_marriages** and **income_family**. For income_family, it is a new column **mutated** by us; to get the income_family, we split different intervals of income into two groups, which are **low income** and **high income**. Thus, our cleaned data has **20602** observations of **5** variables.

We analyze the cleaned data by building a **multiple linear regression model**. A multiple linear regression model "is a statistical technique that uses several explanatory variables to predict the outcome of a response variable"(Kenton,2020).

We choose this model because we want to see how **ages**, **number of marriages**, **sex** and **incomes** can affect **total number of children**. Also, before seeing the results, we believe that people have higher ages might have more children, people have more marriages might have more children, males might have less children than females, and people have high incomes might have less children than people have low incomes.

Moreover, we want to predict total number of children using **ages**, **number of marriages**, **sex** and **incomes**. Thus, the **explanatory variables** are **total number of children**, and the **response variable** are **ages**, **number of marriages**, **sex** and **incomes**, where total number of children, ages, number of marriages are numerical, and sex, incomes are categorical.

We want to find **the linear relationship** between explanatory variables and response variable, and the formula is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,age} + \hat{\beta}_2 x_{i,numbermarriages} + \hat{\beta}_3 x_{i,male} + \hat{\beta}_4 x_{i,lowincome}$. The coefficients, p-values of coefficients, and R^2 are important, so we will analyze and interpret them in details.

We cleaned the data and analyze the data by **RStudio**.

The P-value is small, which means the performance of model is significant. Therefore, it can explain the relationship between the **total number of children** and **age**, **sex**, **income** and **the number of marriages** in this model. Overall the performance is good.

The caveats are :

(1) The data is from part of Canada only, so this result does not apply to other regions and is not widespread.

(2) The survey is conducted by telephone, so there may have dishonest answers that make the results inaccurate.

(3) For those who refuse to be interviewed by telephone for the first time, they may be impatient when conducting the investigation again, which leads to insufficiently detailed answers or misunderstanding of the question. Then this affects the authenticity of the data.

(4) When we analyzed income, we divided it into high income (more than $75,000) and low income ($0-$74,999). This may be different from the real life.

Result

Coefficients	Estimate	Std. error	t value	P-value
(intercept)	-0.159939	0.029174	-5.482	4.25e-08
age	0.025225	0.000583	43.271	$<$ 2e-16
number_marriages	0.770303	0.016414	46.928	$<$ 2e-16
sexMale	-0.066686	0.017690	-3.770	0.000164
income_familylow income	-0.119110	0.018287	-6.513	7.52e-11

R^2	0.2849
$\sqrt{MSE}$	1.26

An **estimated multiple linear regression model** between response variable and explanatory variables is:
$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i,age} + \hat{\beta}_2 x_{i,numbermarriages} + \hat{\beta}_3 x_{i,male} + \hat{\beta}_4 x_{i,lowincome}$
$= -0.159939 + 0.025225 x_{i,age} + 0.770303 x_{i,numbermarriages} - 0.066686 x_{i,male} - 0.119110 x_{i,lowincome}$.

The **standard error** of $\hat{\beta}_0$ is **-0.159939**.

The **standard error** of $\hat{\beta}_1$ is **0.000583**.

The **standard error** of $\hat{\beta}_2$ is **0.016414**.

The **standard error** of $\hat{\beta}_3$ is **0.017690**.

The **standard error** of $\hat{\beta}_4$ is **0.018287**.

For β_0:

Assume $H_0 : \beta_0 = 0$, $H_a : \beta_0 \neq 0$. The **p-value**= 4.25e-08, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_0 \neq 0$.

For β_1:

Assume $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$. The **p-value**= $<$ 2e-16, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_1 \neq 0$, which means there is a relation between x_1 and y.

For β_2:

Assume $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$. The **p-value**= $<$ 2e-16, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_2 \neq 0$, which means there is a relation between x_2 and y.

For β_3:

Assume $H_0 : \beta_3 = 0$, $H_a : \beta_3 \neq 0$. The **p-value**= 0.000164, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_3 \neq 0$, which supports that there is a difference in average value of y_i between male and female.

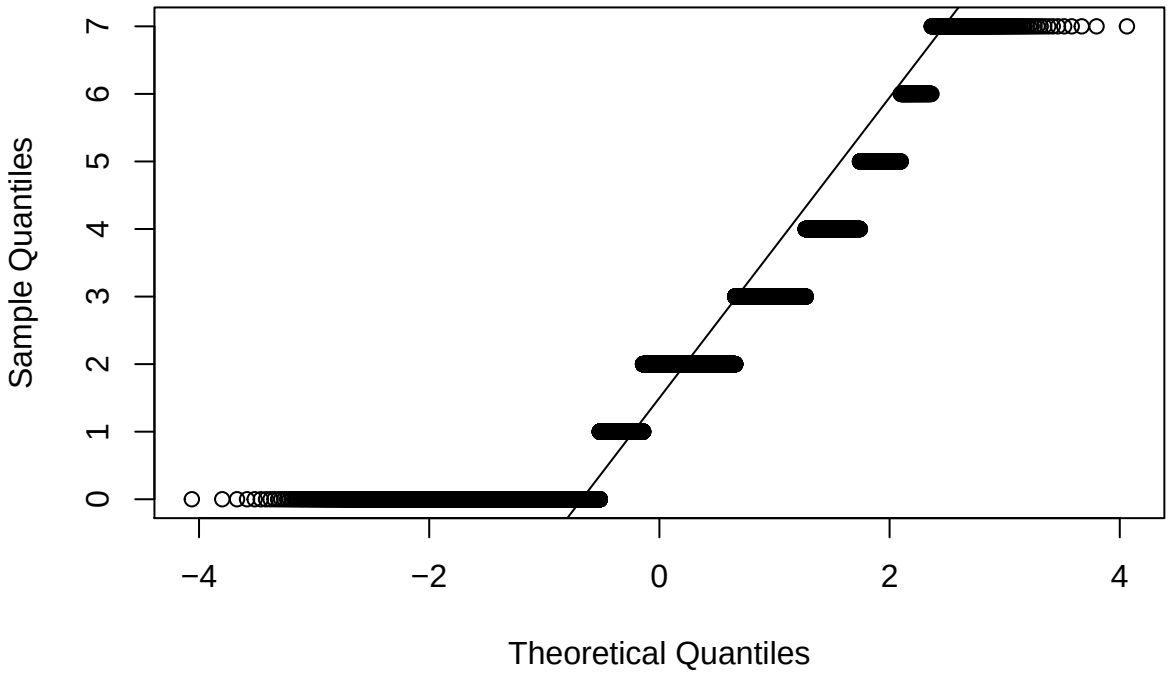
For β_4:

Assume $H_0 : \beta_4 = 0$, $H_a : \beta_4 \neq 0$. The **p-value**= 7.52e-11, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_4 \neq 0$, which supports that there is a difference in average value of y_i between low income and high income.

The interpretation of the estimates of the regression parameters will be explained in discussion section.

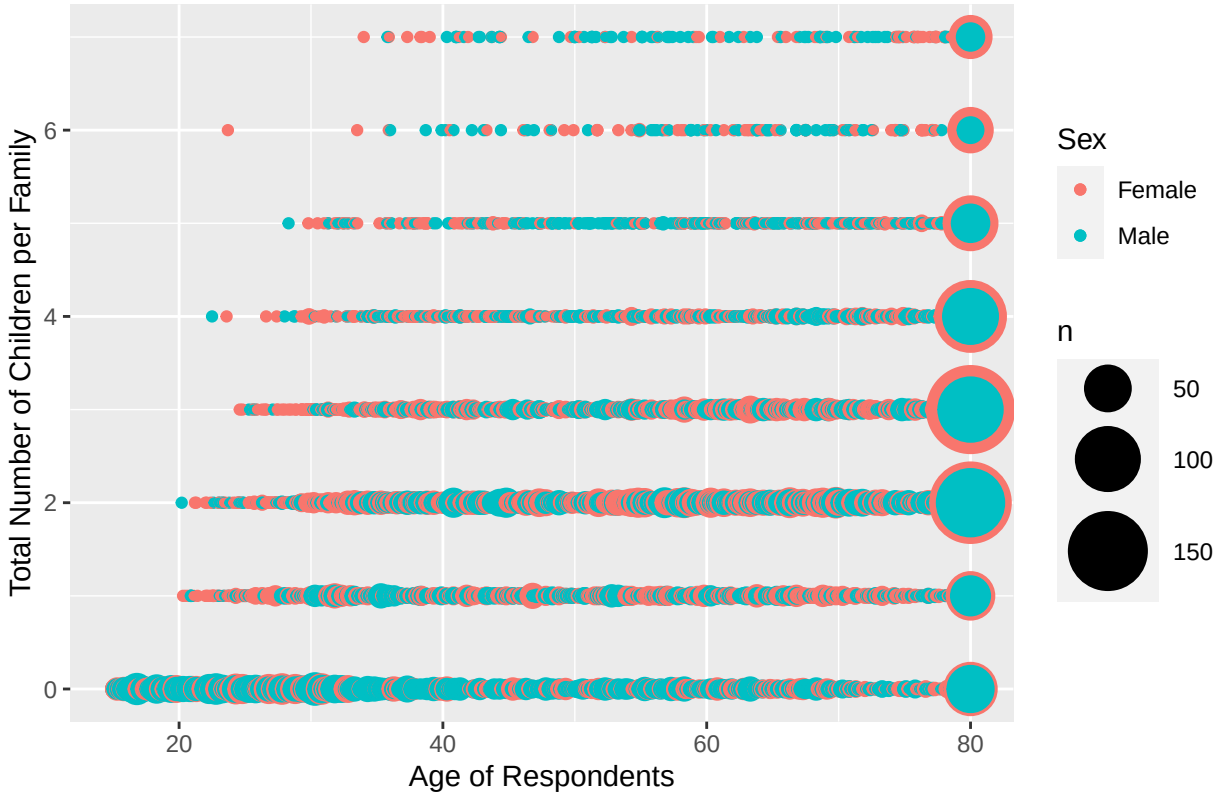
The $\sqrt{MSE}$ of this estimated model is **1.26**. And the R^2 of this estimated model is **0.2849**, which means 28.49% of the total variation in y is explained by the regression line.

Figure 1: Normal QQ−Plot and QQ−line of Total Number of Childrer



The first graph is **normal QQ-plot and QQ-line** of the response variable total number of children. From this graph, it can be seen that most dots are **not fall on** the QQ-line, so the response variable is **not Normal**.

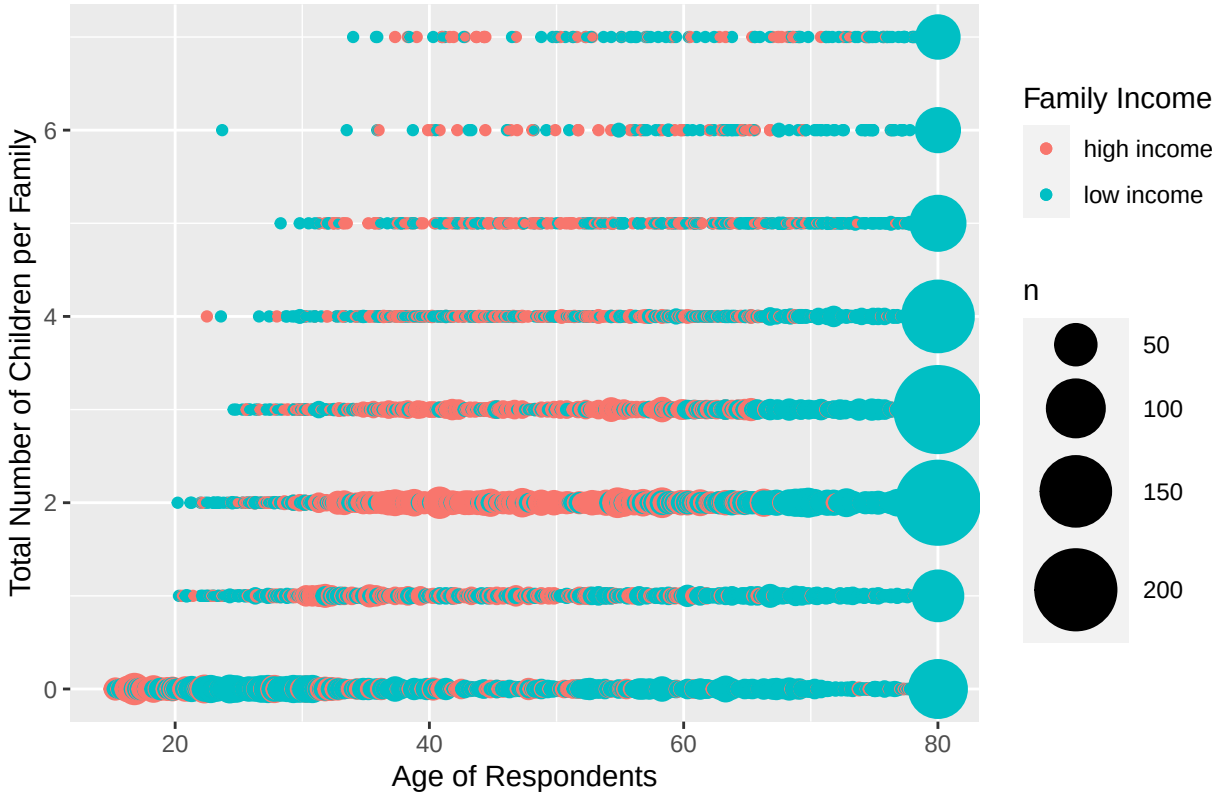
Figure 2: Scatterplot of Total Number of Children and Age (By Sex)



The second graph is **The Scatter Plot of Total Number of Children and Age (By Sex)**. From the overall graph, the older people, the more children they have. This is true for both sexes. Among them, female who with 3 children are the most, and male who with 2 children are the most. Besides, for people without children, relatively speaking, the proportion of men is relatively larger than females, ranging from about 20 to 80. However, there are also a few females who have 6 children when they are about 22 or 23 years old.

Comparing the two sexes, the proportion of females with children is higher than that of males. It can be predicted that between the ages of 20 and 40, some unmarried females are likely to **adopt children** or choose to **give birth after pregnancy**. Between the ages of 40 and 80, females may be divorced and have more children in subsequent marriages.

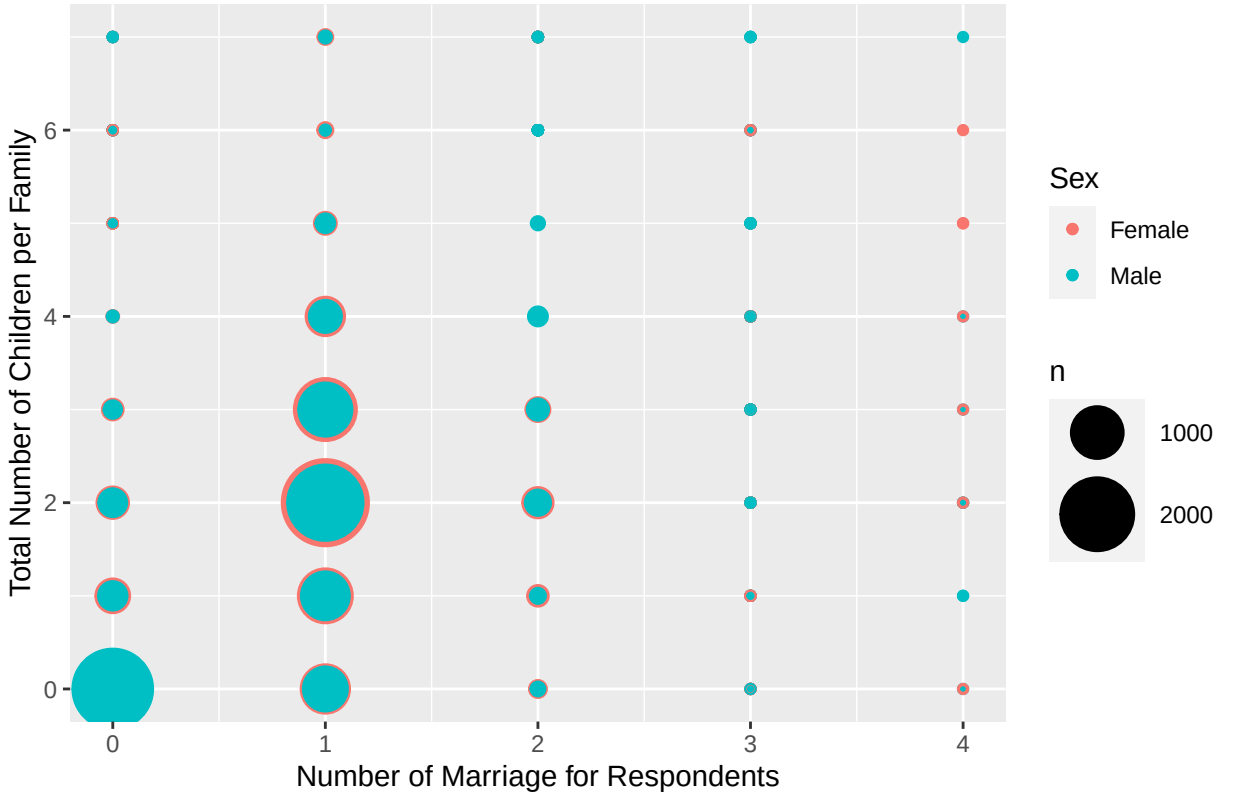
Figure 3: Scatterplot of Total Number of Children and Age (By Income)



The third graph is **The Scatter Plot of Total Number of Children and Age (By Income)**. It can be seen from the graph that when the age is about 30 to 55, **high-income** people usually have 2 or 3 children, and **low-income** people usually have no children or 6 to 7 children.

From this point of view, it can be concluded that people will relatively have good job opportunities during the rising period of their careers. Also, they may have the strength to work hard for their children. Moreover, since they are a family, the income is from both females and males. For people without children, They may not have much demand for money. On the other hand, too many children may prevent them from going to work or pursue better job opportunities.

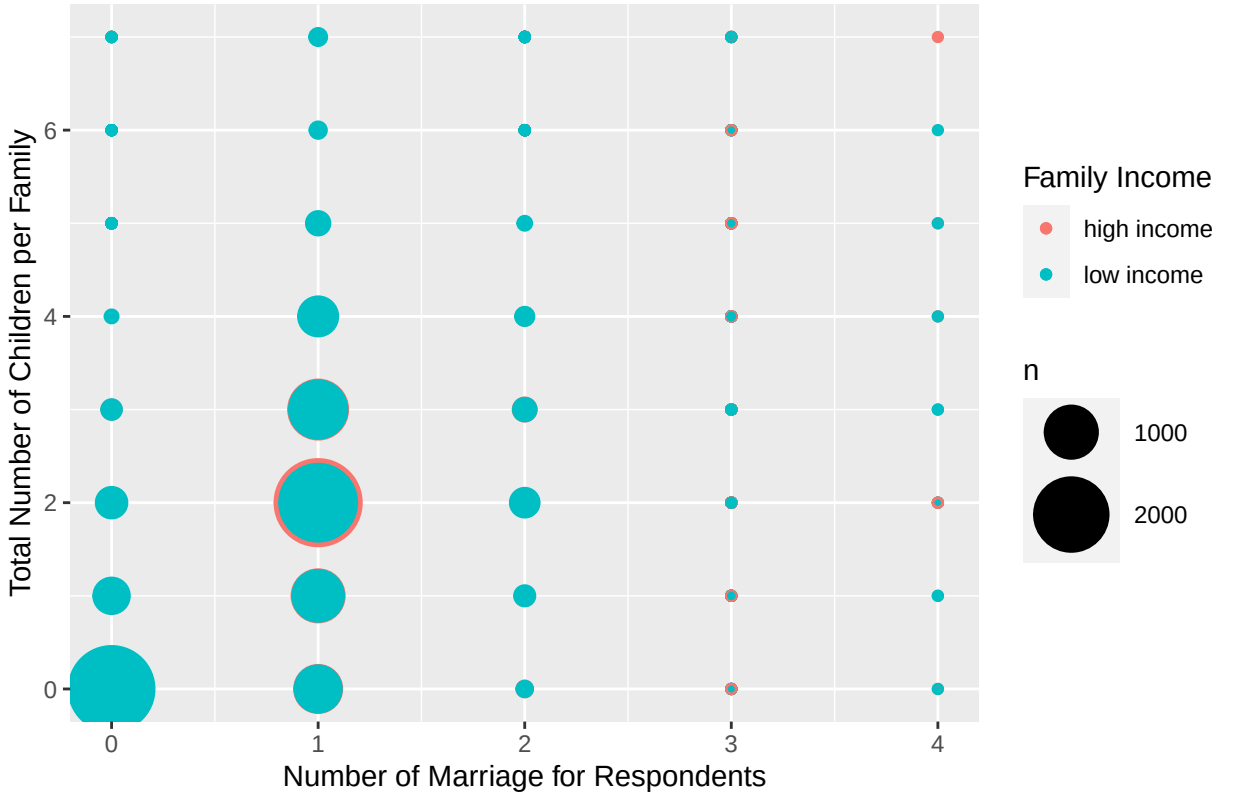
Figure 4: Scatterplot of Total Number of Children and Number of Marriages (



The fourth graph is **The Scatter Plot of Total Number of Children and Total Number of Marriages (By Gender)**. First of all, this graph clearly shows that **the number of marriages** for most people with 2 children is 1. But it can also be seen that for unmarried male or female, they are having 1 to 7 children also account for a part of the proportion.

This means that some of them may be **single-parent families** or **adopted children**. Also, the female who has married 4 times have more children than men. This may represent that females usually continue to raise children after divorce. Another possibility is that most male may not know that they have children due to the lack of contact after a breakup or divorce. However, pregnancy represents that females know they are having children. In this case, most females choose to **give birth** because females are caring for children, and **abortion** is prohibited in Canada.

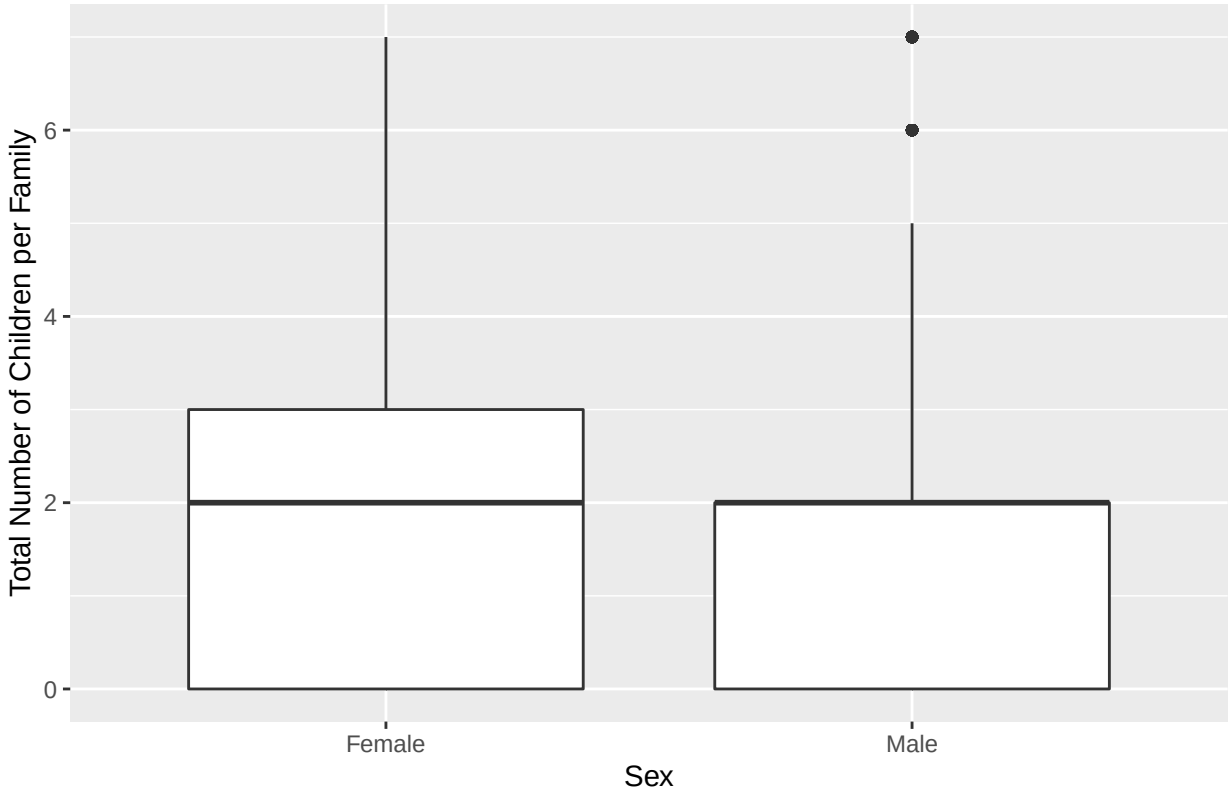
Figure 5: Scatterplot of Total Number of Children and Number of Marriages (I



The fifth graph is **The Scatter Plot of Total Number of Children and Total Number of Marriages (By Income)**. This graph shows that the average of **high and low income** is for people with two children in a marriage. The income of unmarried people is generally not high, because the possibility of raising children alone is relatively high, the income is not as high as the family income.

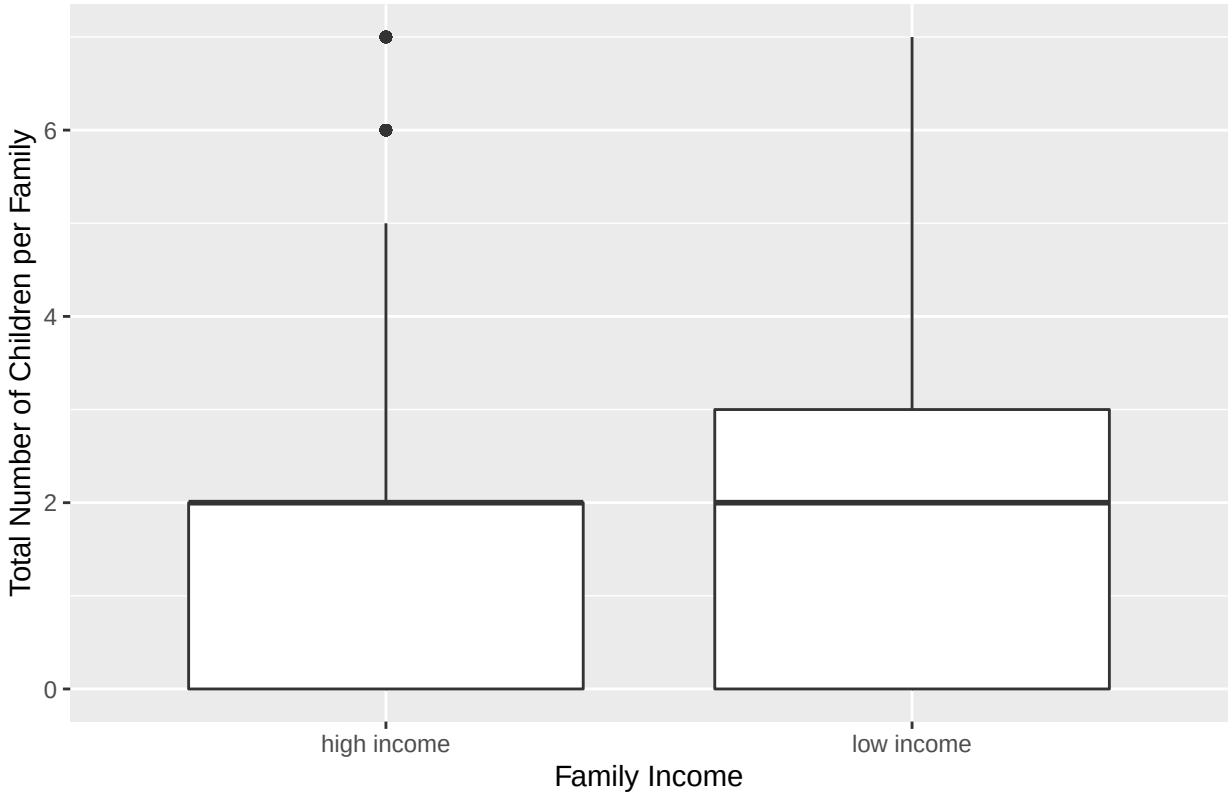
Besides, People with **high income** have relatively fewer children than people with **low incomes**. In this situation, it can be predicted that families with fewer children will have more time and energy on work. If a family has less than 2 children, the female can resume work after 2-3 years and keep up with the pace of society faster. However, if there are more than 3 children in a family. The chance of one birth is very low. This shows that females will not be able to devote themselves to work for 5-6 years, while men will need to support the entire family expenditure.

Figure 6: Box Plot of Total Number of Children and Sex



The sixth graph is **The Box plot of Total Number of Children and Sex**. This graph demonstrates that the median of the **total number of children** is 2 for both sexes. Most females have children between 0 to 3, most males have children between 0 to 2. But some people have as many as 7 children. Also, the median is more close to the top of the box for both sexes, so the distribution is negatively skewed.The **IQR (Interquartile Range)** of the female is 3 which is more than the **IQR** of the male is 2.

Figure 7: Box Plot of Total Number of Children and Income



The seventh graph is **The Box plot of Total Number of Children and Income**. This graph demonstrates that the median of the **total number of children** is 2 for both levels of income. Most high-income people have children between 0 to 2, and most **low-income** people have children between 0 to 3. But some people have as many as 7 children. Also, the median is more close to the top of the box for both sexes, so the distribution is negatively skewed. The **IQR (Interquartile Range)** of the high income is 2 which is lower than the **IQR** of the low income is 3.

```
## # A tibble: 9 x 2
##   total_children mean_age
##            <dbl>    <dbl>
## 1              0     40.5
## 2              1     51.3
## 3              2     56.4
## 4              3     59.8
## 5              4     63.1
## 6              5     64.2
## 7              6     68.4
## 8              7     68.2
## 9             NA     53.8
```

From the above table of **total number of children** and **mean of age**, we can see that **ages generally increase as total number of children increases**. This is a logical conclusion because produce a baby will costs lots of time.

```
## # A tibble: 5 x 2
##   number_marriages mean_total_children
##              <dbl>               <dbl>
## 1                0               0.572
## 2                1               2.10
## 3                2               2.53
## 4                3               3.13
## 5                4               2.88
```

From the above table of **number of marriages** and **mean of total number of children**, we can see that **total number of children generally increase as number of marriages increases**. This is a logical conclusion because when people are married more than once, they might have children with their exes.

Discussion

Thus, this is the final multiple linear regression model we get: $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i,age} + \hat{\beta}_2 x_{i,numbermarriages} + \hat{\beta}_3 x_{i,male} + \hat{\beta}_4 x_{i,lowincome}$
$= -0.159939 + 0.025225 x_{i,age} + 0.770303 x_{i,numbermarriages} - 0.066686 x_{i,male} - 0.119110 x_{i,lowincome}$. The $\hat{\beta}_0$ intercept here is negative, which doesn't have an actual meaning when interpreting, $\hat{\beta}_1$ is 0.025225 which means increasing age by 1 will increase the number of children by 0.025225 assuming other variables are fixed. Similarly, $\hat{\beta}_2$ is 0.770303 which means increasing number of marriage by 1 will increase the number of children by 0.770303 assuming other variables are fixed. $\hat{\beta}_3$ and $\hat{\beta}_4$ means the differences of mean of two groups, because x_3 and x_4 are indicator variables which have values of 0 or 1. If the target is male,then the mean number of children will decrease by 0.066686, and if the target has low income, then the number of children will decrease by 0.119110 in average assuming other variables are fixed.

According to the coefficients, the coefficient of sex indicator and income indicator are both negative, meaning they have a negative impact on the mean. Male tend to have less children and people with low income tend to have less children, which is consistent with our guess at the beginning, because some male may never married thus they will not have children and people suffered from poverty may not have the resources to raise a child. Overall, the result can be used to predict potential population growth rate in the future, as we can approximately calculate the mean of number of children next year based on the model, then multiply it by the population in Canada to get the total number of children for next year, and we minus the number this year to get predicted new birth number, eventually using it to calculate potential population growth rate. This allows the government to make useful predictions about the future population and make policies ahead of time. If the population growth rate is way too high, they may want to tax on families with multiple babies ; if the rate is too low, they can subsidize them instead. This way the government of Canada can limit the population growth to a certain range which is beneficial to the society.

Weaknesses

First,the target population for the 2017 GSS included all persons 15 years of age and older in Canada, but excluded all residents of the Yukon, Northwest Territories, and Nunavut. People between 15 to 18 will most likely don't have children, but when operating the data, we didn't separate them. This can be improved next time, we could just filter out the observation with age greater than 18. In the analysis, we first picked out 5 variables and investigate the relationship, but we don't know if there are more potential variables which may affect mean number of children as well. We could improve this by using AIC/BIC algorithm to find optimal number of variables for the model.

Next Steps

The future study of this topic can include a follow-up survey to collect the data for number of new born babies this year. To approximate the population growth rate, there's another way. We can just use another data set which contains new birth numbers and death numbers. Then the population growth rate is (#birth-#death)/#people in Canada, which is easier to calculate and more accurate.The result for the subsequent study can be used to verify our results from previous study. We can also conduct a confidence interval test after or a hypothesis test, the hypothesis would be the growth rate in previous study and growth rate in subsequent study is the same.

References

1. Caetano S, Alexander R. (2020). gss_cleaning.csv.
2. Caetano S, Alexander R. (2020). gss_dict.text.
3. Caetano S, Alexander R. (2020). gss_label.text.
4. Canada Population 2020 (Live). World population review. https://worldpopulationreview.com/countries/canada-population (2020)
5. Canada Birth Rate 1950-2020. Macrotrends. https://www.macrotrends.net/countries/CAN/canada/birth-rate (2020)
6. Effects of Population Growth on our Environment. Young article library. https://www.yourarticlelibrary.com/essay/effects-of-population-growth-on-our-environment/39624 (2020)
7. Firke S. (2014). Rdocumentation. https://www.rdocumentation.org/packages/janitor/versions/2.0.1
8. GitHub link: https://github.com/Yingren-Luo/STA304-PS2
9. General Social Survey, Cycle 31: 2017: Family. AAYNFDZK.csv. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31
10. Kenton, W. (2020, September 21). How Multiple Linear Regression Works. Retrieved October 16, 2020, from https://www.investopedia.com/terms/m/mlr.asp
11. Public Use Microdata File Documentation and User Guide. CHASS. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf (2017)
12. Public Use Microdata File. CHASS. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_Codebook.pdf (2017)
13. Social and Economic Effects of Population Growth. Kullabs. https://kullabs.com/class-9/enviroment-population-and-health-9/causes-and-effects-of-population-change/social-and-economic-effects-of-population-growth (2020)
14. What COVID-19 means for immigration. UConn Today. https://today.uconn.edu/2020/03/qa-covid-19-means-immigration/# (2020)
15. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686