# Analysis of number of total children in Canadian family

Jiayang Wang & Yiyan Zhao & Xuefen Luo & Zhanhe Zhang

October 19

## Abstract

As the fertility declines, population aging will likely lead to a decline in sustainable development capacity and increased financial pressures in a country. This leads us to investigate into the total number of children in a family for studying the fertility trends and natural population growth in Canada. Using the dataset from the 2017 Canadian General Social Survey (GSS) on families, we build a linear regression model to predict the total number of children of a respondent, and analyze how it is affected by other factors, such as age, sex, education, etc. According to the results of the model, the factors,which significantly impact children's number in a Canadian family, should be taken seriously when concerning about the natural growth rate of population and the aging of population.

## Introduction

Our main objective is to find out what factors have significant impacts on Canadian's children number, and analyze how these variables affect people's willingness to have children. We obtain a dataset from the 2017 GSS on families and further narrow down the scope to the variables we are interested in. Developing a demographic analysis on the data to see which variables are correlated with the total number of children in a family. To further analyze how they affect, we fit a stratified survey linear regression model to predict the children's number based on these independent variables (age, sex, education, income, number of marriage, household size). By the model performance, we can know whether the it is significant. Through the analysis of model summary statistics, tables and graphs, we can conclude the influence of each variable on the children's number in a Canadian family. Based on these conclusions, the factors that are negatively related to the total number of children can be considered to impact Canadians to have fewer children. Therefore, these factors are also of great significance to the study of natural fertility and population aging in Canada. Besides, we will analyze the weakness of the data and our model. Then, discussing what can be improved for the next step.

## Data

2017 GSS is designed as an independent, annual and cross-sectional sample survey, containing several survey contents for each respondent, such as entry component, conjugal history, fertility intentions, etc.

A stratified random sampling design(STRS) is carried out for sampling, dividing the population into 27 groups known as strata by geographic areas, and then randomly sampled by telephone interviews within each strata. The target population is all non-institutionalized people above 15 years old living in Canada, and the sampling frame combines both telephone numbers and the Address Register[1](AR) to ensure good coverage of all households. Then the data is collected by telephone interviews, and for those non-response or refused, more calls will be made to contact or explain the importance of the survey as well as encourage participation. The STRS is an appropriate method in this case because the strata are formed based on respondents' shared

attributes, geographic areas, including not only 10 provinces but 17 Metropolitan areas, and also it involves the random sampling from the target population, so it is random enough that each sample is equal likely to be selected, and the sample population can be the best representative of the target population. However, this method can be time-consuming and high-cost because of the large sample size and survey contents.

2017 GSS contains 20602 observations and 461 variables, which is good because the dataset is large and has a lot of attributes so that we have multiple options for research. However, it is missing the variable of the strata name that divides the population by geographic areas, which may affect model setting. Besides, the dataset is not clean enough, resulting in difficult code reading. Therefore, our data is obtained by cleaning and summary analyzing from the 2017 GSS data, including 20244 observations and 7 variables:

Total children($num^2$): Total number of children of the respondent, which is our variable of interest.

Age(num): Age of the respondent at time of the survey interview. We use it as one of our predictor because we believe a person's age may be positively correlated with the number of children they have.

Sex(chr): Sex of the respondent. We use it as an explanatory variable because we want to study which sex group has more children.

Education(chr): Education background of the respondent with highest certificate, diploma or degree. We will use it as an explanatory variable because we believe people with different education backgroup would have different number of children.

Number marriages(num): Number of marriages the respondent has ever had. There is also a variable of marital status which is similar to number marriages. We use number marriages instead of marital status because we believe that generally, people who marry more are likely to have more children, while marital status may have little to do with the number of children, so that number of marriages can be more precise for predicting total number of children compared to marital status.

Household size(num): Household size of the respondent. We use it as an predictor because generally if members are more in a family, this family may have more children.

Income respondent(chr): Total income of the respondent. We use it as an explanatory variable because we want to study the influence of income on the number of children for the respondent.

Province(chr): Province of residence of the respondent, which we will use as the strata in modeling.

N(num): We create a variable records total number of observations in each province, and we decide group size based on this value in our STRS model.

# Model

We continue our study by building a linear regression model that is a linear approach to modeling the relationship between variable of interest and explanatory variables, predicting the number of total children using age, sex, education, number of marriages, household size and income as predictors. Since 2017 GSS is sampled by STRS method that divides population into 27 geographic groups, we choose to build a stratified linear regression model to make our regression more consistent and accurate. Our sample is first grouped by 10 provinces of residence of the respondent, where the reason we use 10 provinces as strata in modeling is that GSS missing the variables of 27 geographic groups. Then randomly sampled from each group and build a linear regression model in R with the following formula:

$\mathbf{\hat{TotalChildren}}$ = -1.9623423 + 0.0401249$\mathbf{age}$ + 0.5828348$\mathbf{mariages}$ + 0.4211213$\mathbf{HouseholdSize}$ - 0.1094921$\mathbf{SexMale}$ + 0.1772940$\mathbf{EduCollege}$ + … - 0.3758132$\mathbf{EduTrade}$ +

0.0083069**IncomeOver125k**+ … - 0.0855531**IncomeUnder25k**

Where the variables are:

**$\widehat{\text{TotalChildren}}$** is the estimate of total number of children for each respondent, and it is a discrete random variable and does not make any sense that the total number of children is decimals, but it can be decimals when fitting model and calculation, so we only use it to study the relationship between TotalChildren and other variables or round it to the nearest integer when needed.

**Intercept** is -1.9623423, which means when all other predictors are 0, the expected total number of children is -1.9623423. Again note that 0 is outside the range of age and household size for numeric predictors, and for indicators, it has to be whether male or female, one of the 7 education backgrounds, and one of the 6 income types, so it is not safe to extrapolate beyond the range of our data.

**age** means when other variables are unchanged, for each additional unit increase of age, the expected total number of children increase by 0.0401249 on average, which is a positive relationship.

**mariages** means when the number of marriages increase by 1, we expect the total number of children to increase by 0.5828348, which is a positive relationship.

**HouseholdSize** means when household size increases by 1, the expected total number of children increase by 0.4211213 on average, and it is a positive relationship.

**SexMale** is a dummy variable and sex of female is the baseline for variables of sex, which means when SexMale is 0, the equation represents the total number of children of a female with other predictors(i.e. age, marriages, etc.). When other predictors are unchanged, if SexMale changes from 0 to 1, it means the sex changes to a male from a female, and we expect the total number of children will decrease by 0.1094921, which is a negative relationship.

**EduCollege** is also a dummy variable and education of Bachelor's degree is the baseline for variables of education background, which means when all variables of education(i.e. EduCollege, EduTrade, etc.) is 0, the equation represents the total number of children of a person that has a Bachelor's degree with other predictors. When other predictors are unchanged, if EduCollege changes from 0 to 1, it means the education changes from Bachelor's degree to College certificate or diploma, and the expected total number of children increase by 0.1772940 on average, which is a positive relationship. Same interpretation for other variables of education. When other predictors are unchanged, if the education level changes from 0 to 1, it means it changes from Bachelor's degree to the corresponding education level with corresponding coefficients, and we expect the total number of children changes by that coefficient.

**IncomeOver125k** is also a dummy variable of income, and income from 100k to 149.999k is the baseline for income variables, meaning when all variables of income is 0, the equation represents the total number of children of a person that has income over 125k with other predictors. When other predictors are unchanged, if IncomeOver125k changes from 0 to 1, it means the income level changes from 100k-149.999k to over 125k, and the expected total number of children increase by 0.0083069 and it is a positive relationship. Same interpretation for other variables of income. When other predictors are unchanged, if the income level changes from 0 to 1, it means it changes from income over 125k to the corresponding income level with corresponding coefficients, and we expect the total number of children changes by that coefficient.

We use income as a variable for predicting total number of children instead of occupation because we believe income may have more influence on the number of children. However, when testing whether the coefficients of variables of income equals to 0(Null hypothesis), all of the p-values are from 0.0526 to 0.8831 which are greater than the significance level of 5%, so we do not reject our null hypothesis, and we consider coefficients

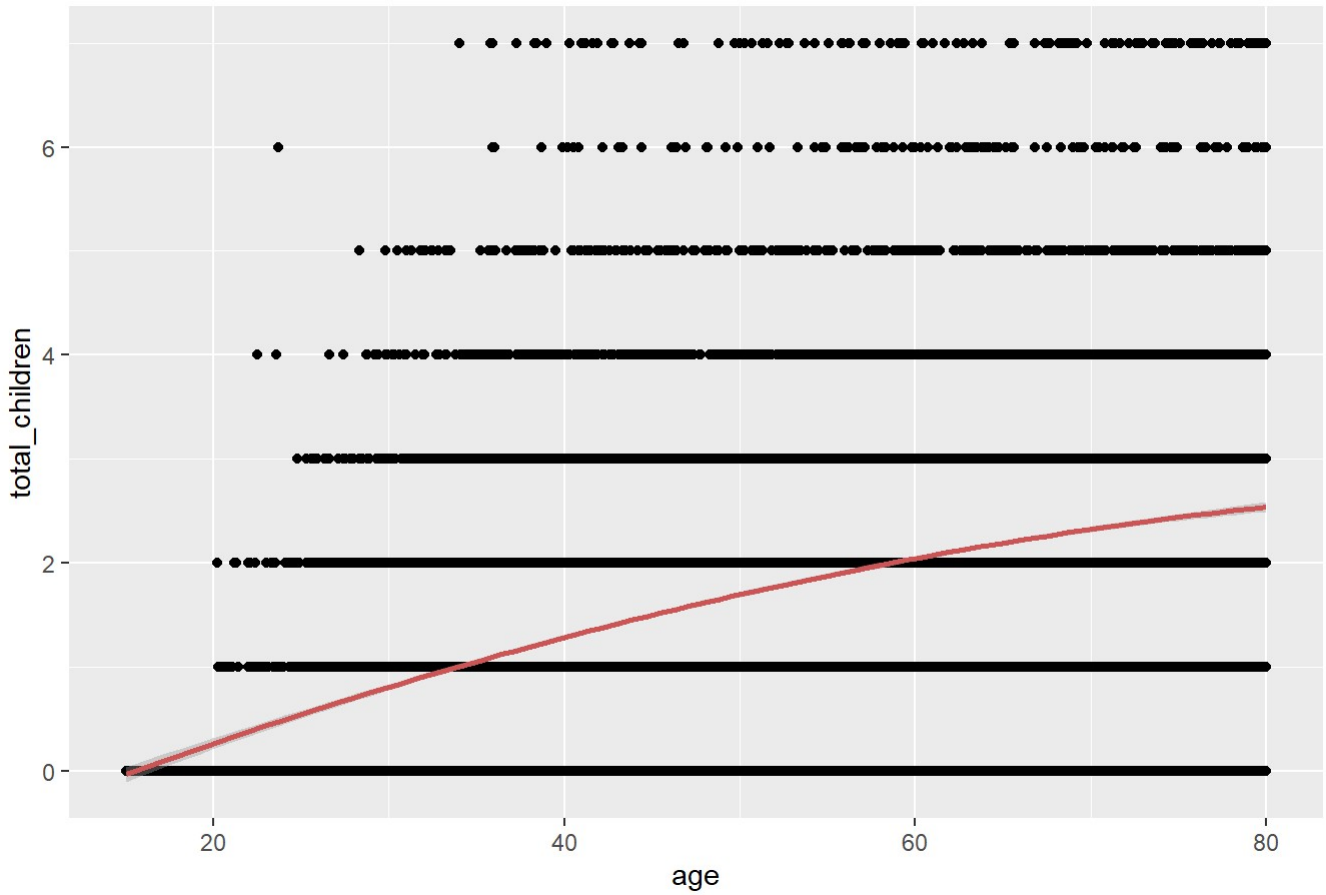
of income variables are not statistically different from 0, which means variables of income are bad predictors for predicting total number of children. When testing whether the coefficients of variables of age, sex, number of marriages, household size and education equal to 0(Null hypothesis), we see that all of the p-values are less than significance level of 5%, so we reject our null hypothesis, say that these variables are not 0, so variables of age, sex, number of marriages, household size and education are good predictors for our fitted model.

Overall the performance of our model is good with 0.3889 R-squared which is a measure that explains the strength of the relationship between our variable of interest and predictors. 0.3889 means our model has a positive and relatively strong relationship. Besides, our p-value for testing whether response variable is related to explanatory variables is significantly less than significance level of 5%, which means there is a relationship between total number of children, and the 6 predictors. However, there might be some caveats; firstly when grouping, we use 10 provinces as strata in modeling instead of 27 geographic groups in GSS sampling because of missing variables, which may lead to inconsistent from 2017 GSS, and inaccurate predictions. And there is relatively less data for total children number over 6, so it may not be very accurate to predict high number of children.

# Results

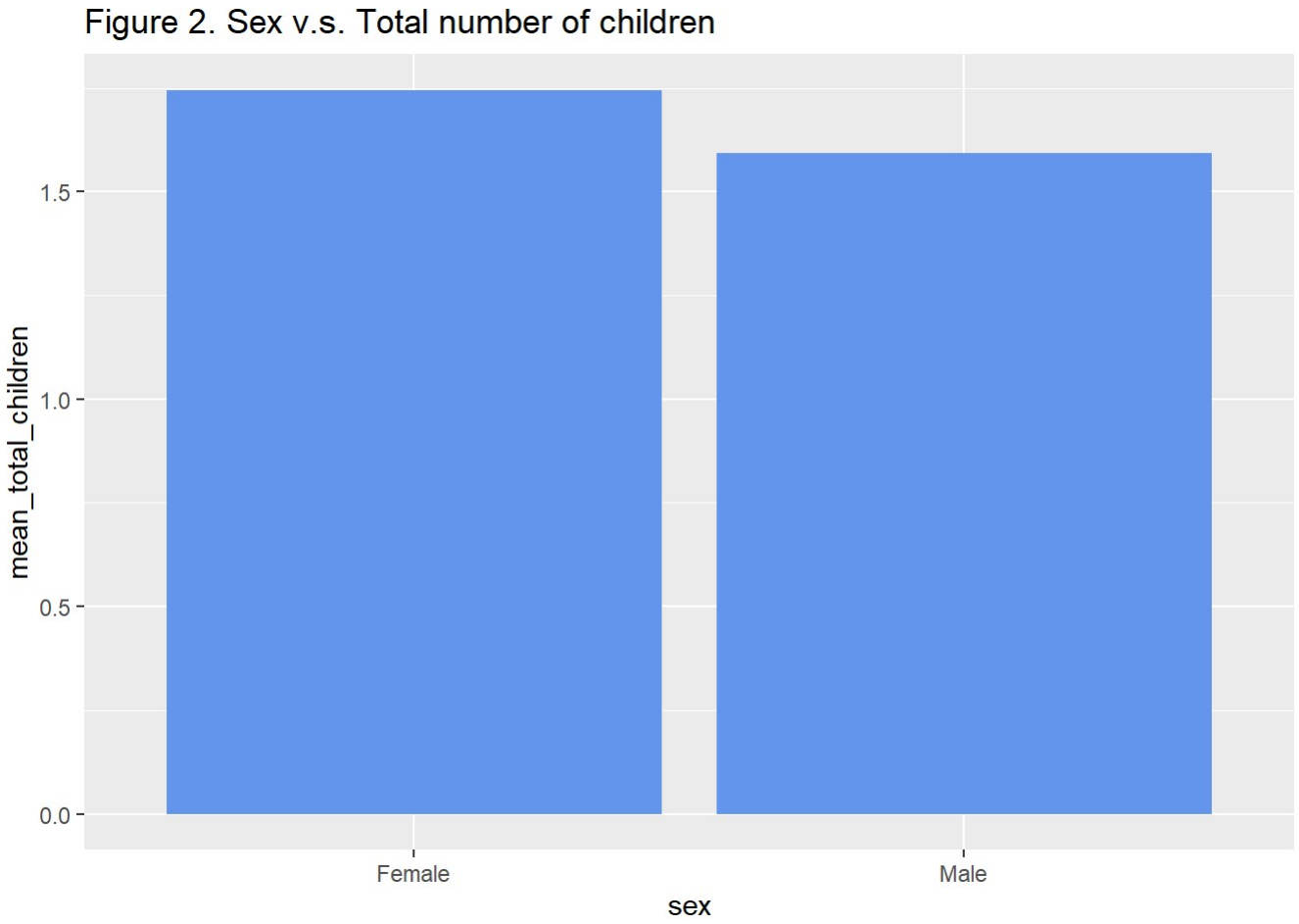
#1.

Figure 1. Age v.s. Total number of children




By looking at the Figure 1, we can know that there is a positive relationship between age and the total number of children. It is true since as we are getting older, having marriages, the number of children will increase. #2.

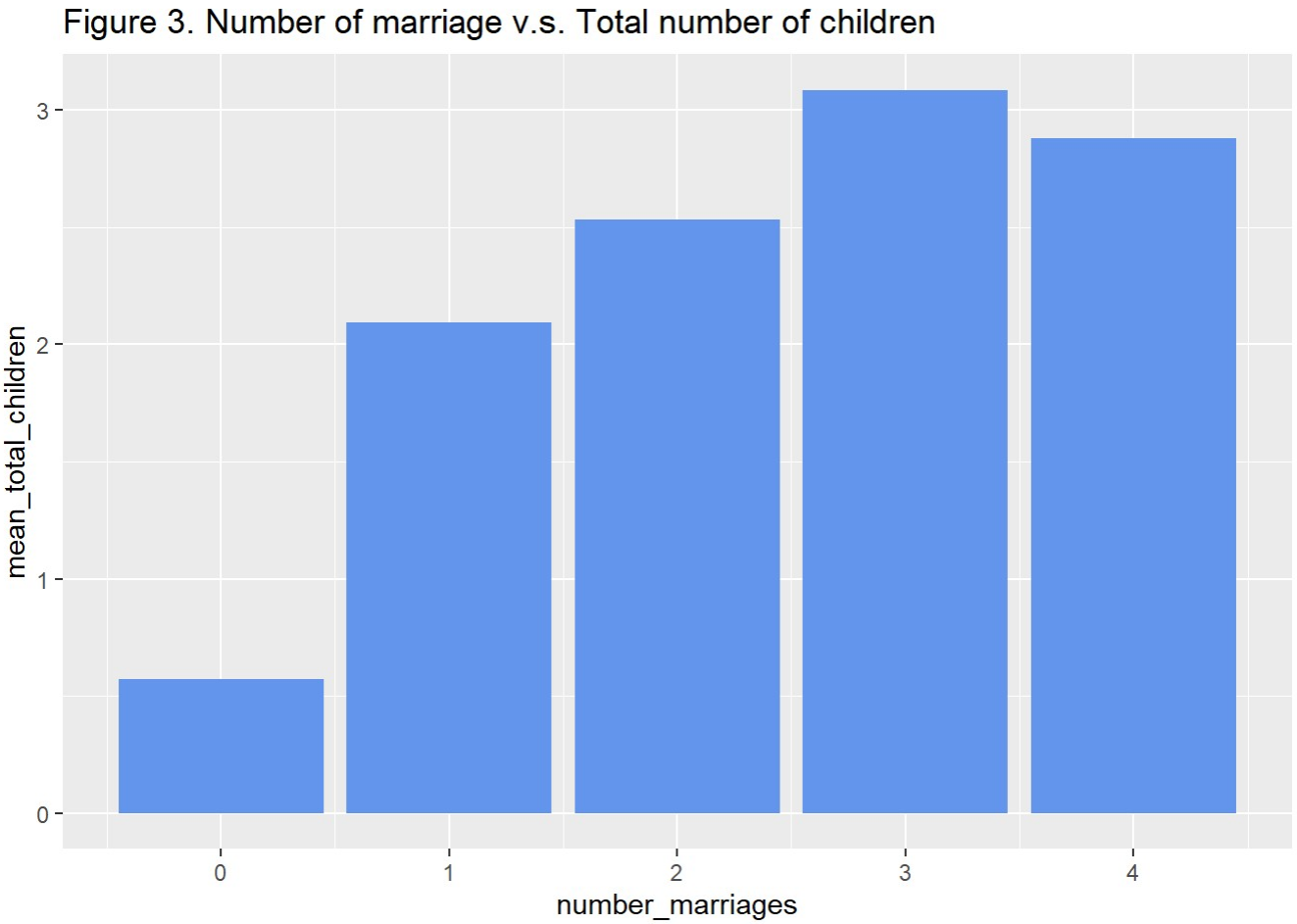
```
## `summarise()` ungrouping output (override with `.groups` argument)
```




By looking at the Figure 2, we can know that there is not too much difference in female and male about having total number of children. It may because most of the respondents are the representation of their family.

#3.

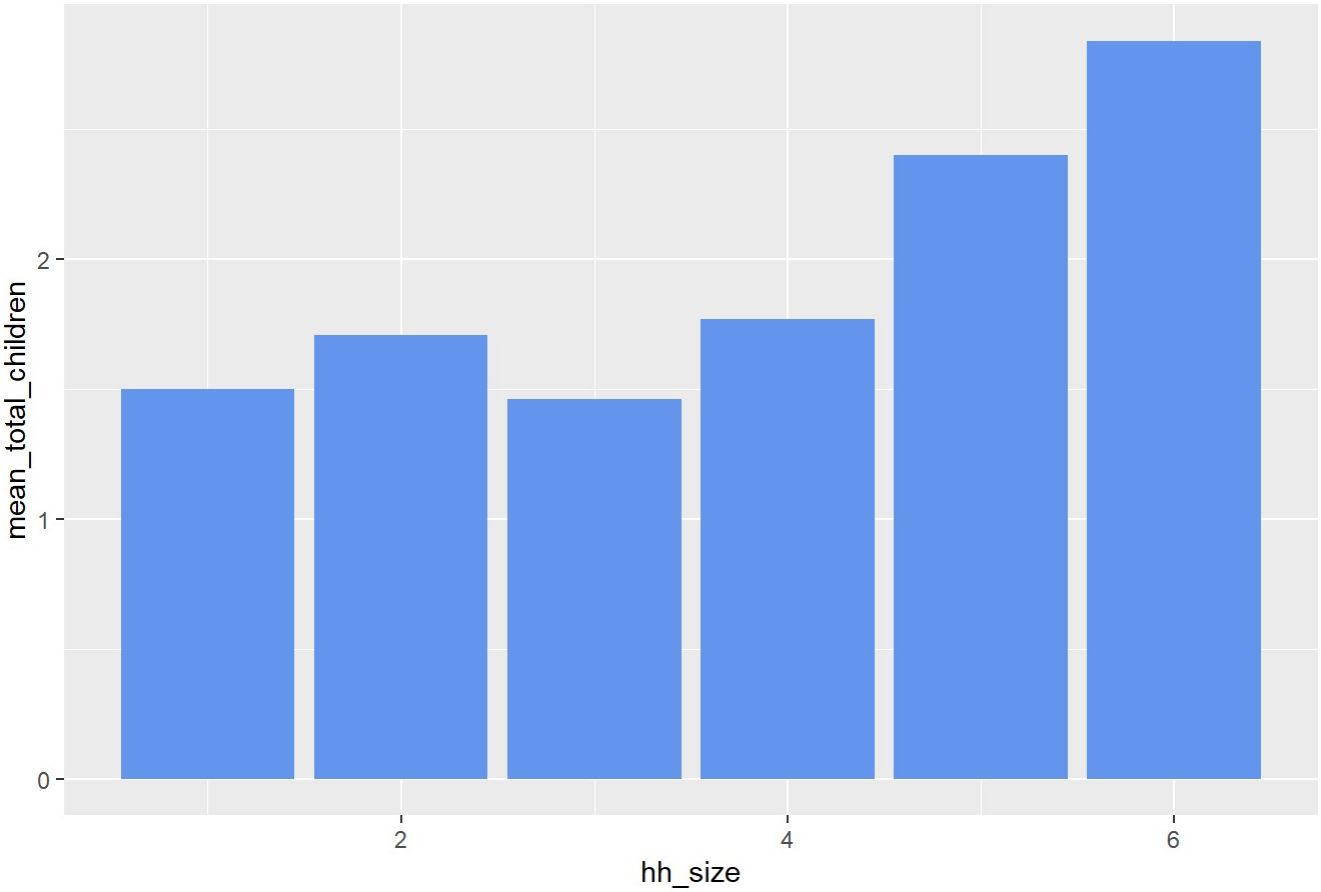
```
## `summarise()` ungrouping output (override with `.groups` argument)
```




By looking at the Figure 3, we can know that there is a positive relationship between the number of marriage and the total number of children. It is mostly likely because as one person have more marriages, she or he will have new children in new families. #4.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

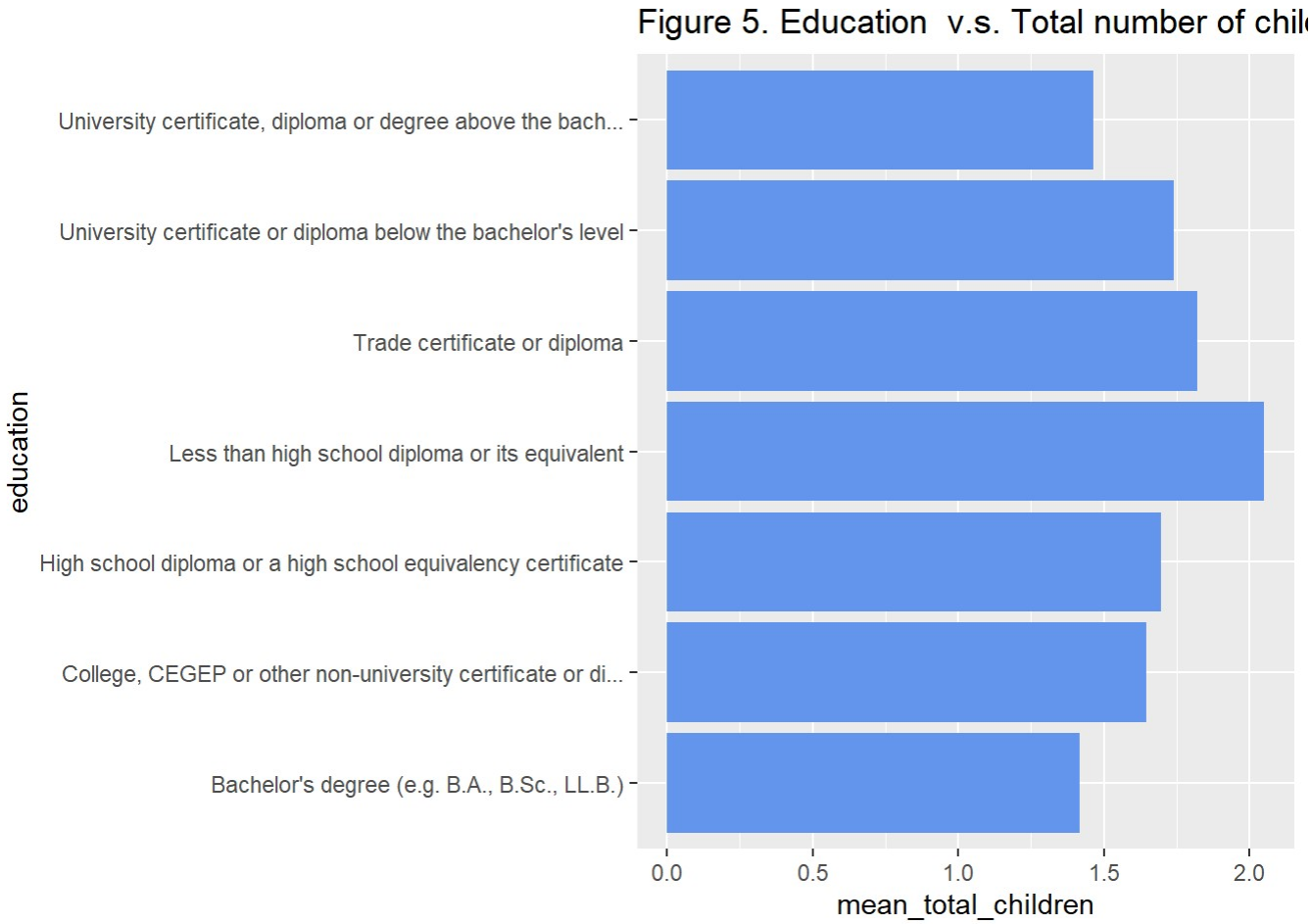
Figure 4. Household size v.s. Total number of children



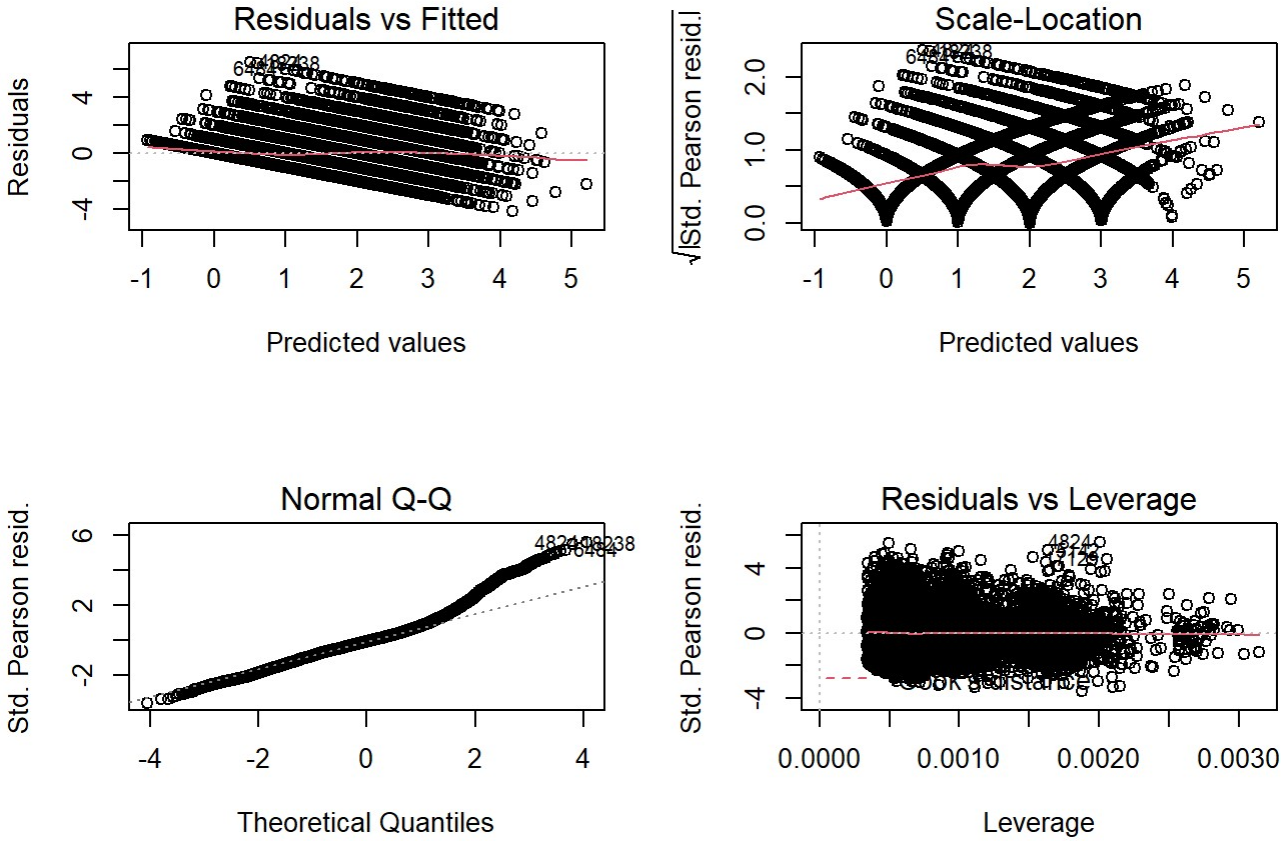

By looking at the Figure 4, we can know that there is a positive relationship between the household size and the total number of children. It is mainly due to the fact larger size families may have more children than other families.

#5.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```




By looking at the Figure 5, it is obvious that people who have less than high school diploma or its equivalent tend to have more children in their families than others with different educational level. And it can be explained that people with high educational level may pay more attention on the workplace, pursuing higher life quality than that of relative low educational levels.




By the Residual vs Fitted plot, we can know that the residuals have non-linear patters since there is a flat straight line. And we can say that our model is good. From the Normal Q-Q plot, we can conclude that our model is not perfect since the residuals are not lined well on the straight dashed line. Based on the Scale-Location plot, we can say that our residual are spread equally and randomly since the red smooth line is almost horizontal. Based on the residual vs Leverage plot, we can conclude that there are no any influential cases since all cases are within in the dashed Cook's distance line.

Therefore, we can say that our model is good in general, but not perfect enough. And it may due to the missing of other useful variables that we are not considered or missing interaction terms in our model.

# Discussion

The model aims to provide a reference to government institutions which are care about fertility rate of Canada. Fertility decline is a serious problem nowadays. Much more surveys show that more and more young people do not want to get marriage or babies. In addition, as people's open thinking, more types of families are accepted. Therefore, this will cause new born become less and less. For now, this case will not be a problem but in a long-range sight like 50 years later, a serious problem, aging population will come out. There will be not enough young people to work which due to economic receding. Therefore, people can find the numeric relationship between total children and other elements in linear regression model and we can find the most important element which effect the number of children in a family. Then we can focus on improving this part in order to rerise fertility rate. From the graphs in result , we can see that age, number of marriages and household size have positive relationships with the number of total children and sex is not a significant reason which can affect the number of total children. And for other elements, education and income, we cannot find any significant

relationship from their graphs. However, combined with line regression model, it shows they are the significant elements because of some strange p-value. In terms of education, education University certificate or diploma below the bachelor's level (p-value0.0370) shows weaker relationship than other education levels. So we can definethat people who are in education University certificate or diploma below the bachelor's level have less contribution on increasing the number of total children. In addition, for income, we can find that people whose income less than $25,000 per year have strong relationship with increasing total children number because income respondent Less than $25,000 has the least p-value (0.0526) among other levels.

Based on these two findings, if governments want to solve the problem, fertility decline, university education and improving income are two very important points which cannot be ignored. At first, for university education , we have found that people who are in university education below bachelor shows weak relationship with total children number. The reason why cause this is new university students do not put their all energy on studying because most of them need to earn their tuition by themselves. It's not like people who are in high school or below high school education level. After high school, they get into the society early and through their effort they may live better than people who graduate from university. Besides, new graduate students are hard to get a good job so it is very easy to bring them the idea, do not want to get married and babies because of financial issues. Therefore, government may keep more eyes on health of university students, especially mental health and try to give them more financial support in order to let them put them all on studying. Finally, people in university education above bachelor increase, the number of children will increase as well.

In terms of income, we can easily find that people who have more money, the less contribution to total children number they do according to their p-value. The reason why causes this situation is plenty. For example, a person who are rich is strongly connected with his hard work, but his energy is limited. With work pressure, he will not want to have children because they will affect their work and become their burden. In addition, people who are in high income have plenty of interesting things to do on their daily life. They prefer to use their money to enjoy the life, not for children. To solve the fertility problem in this point of view, government may still focus on education. Schools should teach them not only how to live better, but also how to be a man with responsibility.

# Weaknesses

For data and its analysis, we choose linear regression model, but we found that R^2 (0.4) is not very big. So, it means our prediction is not accurate enough. There must be another significant element like people's health which we do not conclude in our study. In addition, there is also a bad predictor, income in our analysis. In histogram of income and total children number, we can not clearly define their relationship, but in our model, we find income is a significant element.

# Next Steps

Obviously, our data is not enough to find what we want. So, we plan to collect more data about health, first birth age and marriage age in order to help us improve our model. And we will also try another model like logistic or Bayesian because not all data only fits linear regression model.

# References

Alexander, Rohan, and Caetano, Sam."gss_cleaning". 7 Oct 2020. https://q.utoronto.ca/courses/184060/files/9422740?module_item_id=1867317 (https://q.utoronto.ca/courses/184060/files/9422740?module_item_id=1867317)

“Welcome to My.access – Please Choose How You Will Connect.” My.access - University of Toronto Libraries Portal，sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm.

---

1. List of all dwellings, used to group together telephone numbers associated with the same address↩
2. Data type: num: numerical; chr: categorical↩