# Family Total Children Number Analysis - by linear regression model

Ruotong Wang & Shengnan Mao

October 17th, 2020

**Family Total Children Number Analysis - by linear regression model**

## Ruotong Wang & Shengnan Mao

## October 17th, 2020

### Abstract

It is not easy to raise a child as a family either from the financial aspect or personal aspect. The whole family needs to take responsibility from the childbirth to their adulthood or even longer. The number of children of each family can be affected by many factors, and we focused on investigating into respondents' personal aspect. We obtained the dataset of the number of total children in each family in Canada in 2017 from the GSS database, and variables such as the respondent's age, age at the first child, and their evaluation of feelings to life were also collected from the same database. Therefore, we built a linear regression model to study the relationship between the total children of each family with the other three variables. The result shows that people's life feelings have a weak correlation with their numbers of children, and people's age and their ages at first child's birth have strong relationships with their numbers of children.

### Introduction

The data we used is from the 2017 General Social Survey (GSS) on the Family. In this analysis, we focused on the total number of children in each family. Also, we chose three variables that we think affect the total number of children. These variables include the age of the respondent, the age of this respondent while his/her first child was born and the evaluation of feelings to life from this respondent. Instead of some financial factors, these three variables all belong to the respondent's personal factors, and we wonder which variable is strongest related to the total number of children in each family. To prove our opinion, we would make a linear regression model based on the independent variable and the dependent variables. Due to this model, we would do a comparison between our original opinion and the final result provided by the dataset. Finally, the result we have conducted overall can help us know some potential reasons that each family decides to have a number of children differently.

### Data

This survey was taken from February 2nd to November 30th in 2017 as the 2017 General Social Survey (GSS) on the Family. The goal of this survey is to gather information on social trends in order to know well about changes in the living conditions and well-being of Canadians. The target population is all non-institutionalized persons aged from 15 years old and older and lived in the 10 provinces of Canada. This survey used a frame that created in 2013, and it combines people's telephone number and their address that they leave at Statistics Canada's Address Register. The data was collected via a telephone survey. In this survey, each respondent was randomly selected from each household and it's stratified random sampling. The target sampling size for this survey was 20,000, but the actual number of respondents is 20,602.

The sampling method is not very good, because it is more like a volunteer sampling. The data was taken by answering the phone which means the respondents are only the people who want to answer questions for this survey. People who are not willing to answer questions would not be collected in this dataset. Then this data could not explain the living conditions and well-being of Canadians well. The questions in this survey are great because these questions could define people living conditions and well-being as well as possible.

## Model

```
##
## Call:
## svyglm(formula = total_children ~ age + feelings_life + age_at_first_birth,
##     design = children.design, family = "gaussian")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3921 -0.7102 -0.1208  0.4640  5.5853
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.7937835  0.0799401  34.948  < 2e-16 ***
## age                 0.0186037  0.0006274  29.652  < 2e-16 ***
## feelings_life       0.0242251  0.0058010   4.176 2.99e-05 ***
## age_at_first_birth -0.0629030  0.0017573 -35.795  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.05 on 12561 degrees of freedom
## Multiple R-squared:  0.1751, Adjusted R-squared:  0.1749
## F-statistic: 888.9 on 3 and 12561 DF,  p-value: < 2.2e-16
```

The linear regression model between the total children of each family with the other three variables is $\hat{y} = \beta_0 + \beta_a * age + \beta_f * feelings + \beta_b * firstbirth$. And the linear regression model with actual number in is $\hat{y} = 2.7938 + 0.0186 * age + 0.0242 * feelings - 0.0629 * firstbirth$

In this linear regression model, we want to predict the total children of each respondent's family. The variables we used to predict it are the age of the respondent, the age of this respondent while his/her first child was born and the evaluation of feelings to life from this respondent.

From the linear regression model above: $\hat{y}$ represents the predicted total children of each respondent's family; $\beta_0 = 2.7938$ represents there are 2.7938 children in a family while age, feelings of life, and age at first child born are zero. $age$ represents the age of the respondent. $\beta_a = 0.0186$ represents when other variable doen not change, if the age of the respondent increase 1 unit, the total children in each family would increase by 0.0186 unit. $feelings$ represents the evaluation of feelings to life. $\beta_f = 0.0242$ represents when other variable doen not change, if the evaluation of feelings to life increase 1 unit, the total children in each family would increase by 0.0242 unit. $firstbirth$ represents the age of this respondent while his/her first child was born. $\beta_b = -0.0629$ represents when other variable doen not change, if the age of this respondent while his/her first child was born increase 1 unit, the total children in each family would decrease by 0.0629 unit.

Based on the result of the linear regression model above, the R-squared equals to 0.1751. This means there are 17.51% variation in total number of children can be explined by the model. Each variable has a p-value, and p-value is a valuation to see if each variable is significant. Every variables' p-value are smaller than 0.05 which means all of the variables are significant to this model.

## Results

**Figure1: Scatterplot between total children and age**

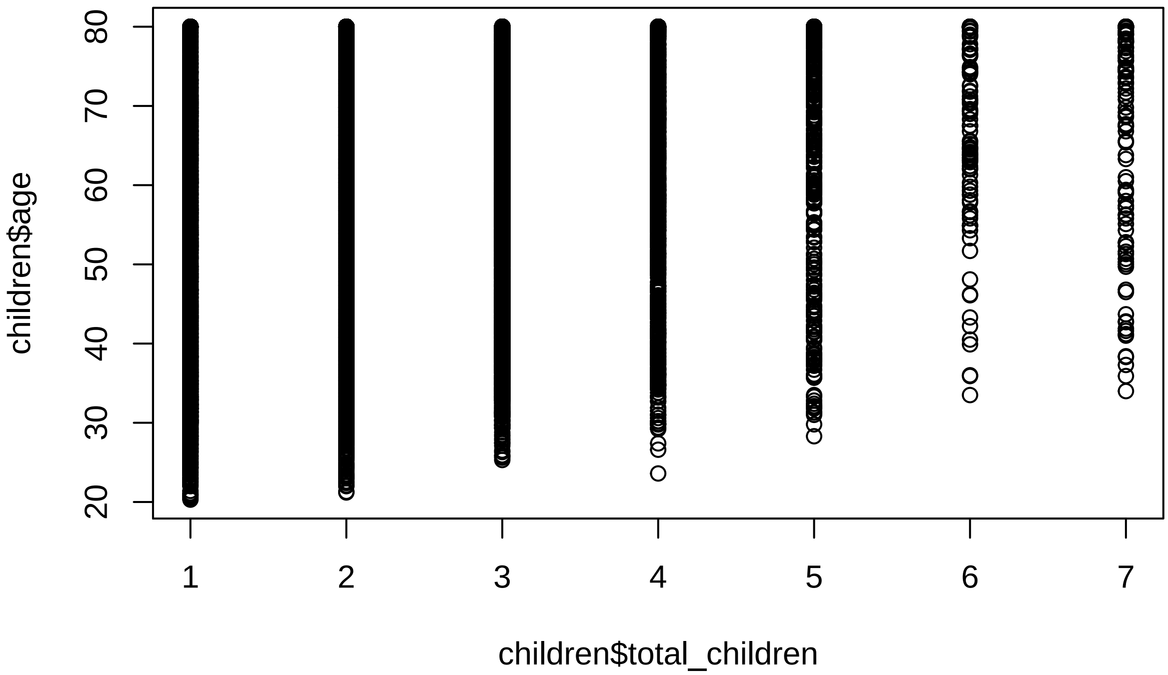


**Figure2: Scatterplot between total children and feelings**

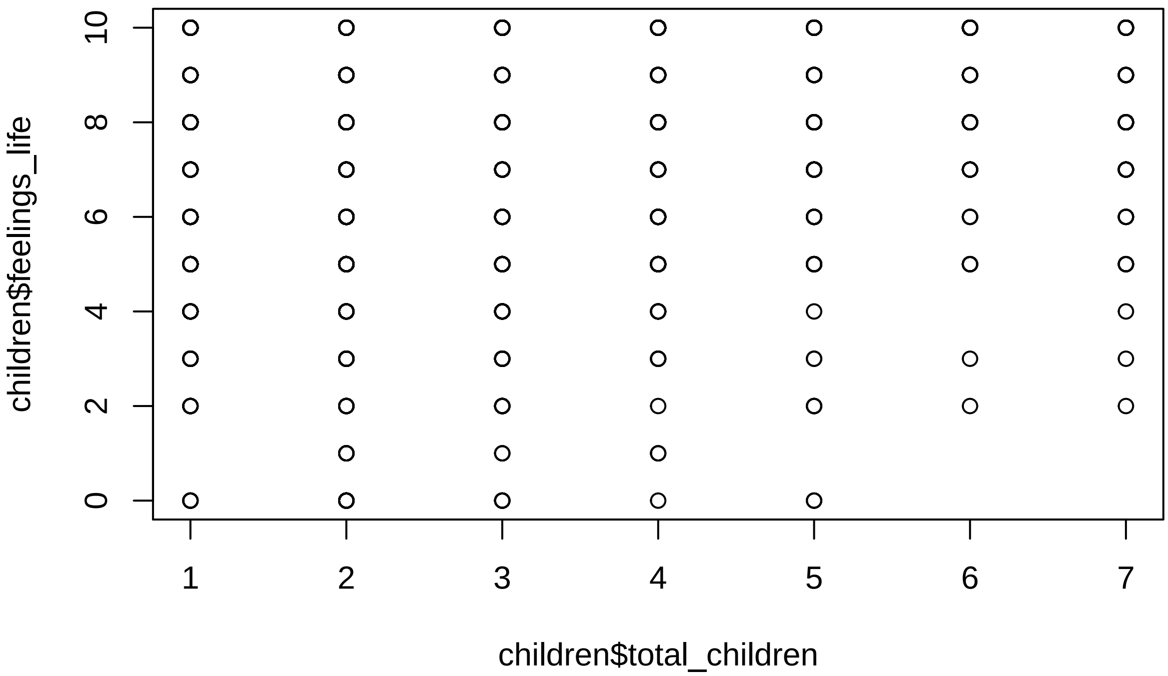
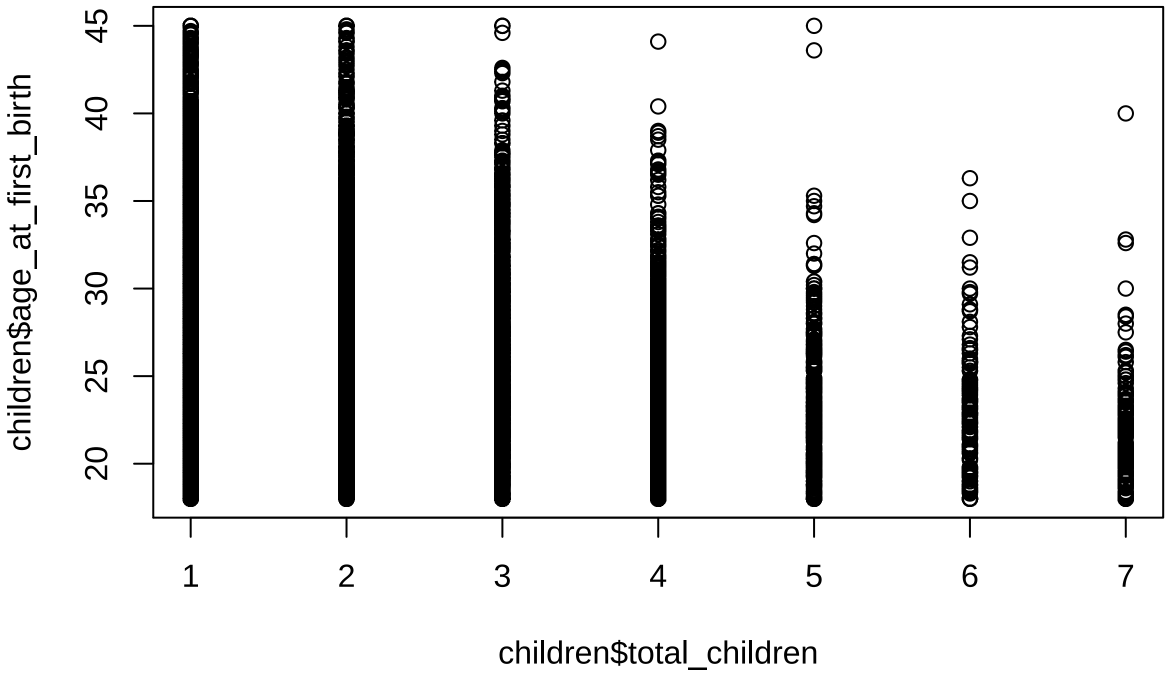


**Figure3: Scatterplot between total children and age when having first c**



By looking at FIGURE 1 which shows the scatter plot of total children and age in each family, and it is obvious to see the group of respondents that have 6 or 7 children are much less than other groups. The age range for respondents with 6 or 7 children is mostly above 40 years old, but it is noticeable that there exists one outlier which reveals a 25-year-old respondent already has 6 children. Most respondents have 1 to 3 children, and the age range for these respondents is between 20 years old and 80 years old. The group that respondents have zero children takes the biggest share of the whole population, and the group contains each age stage, also the majority of young people have no children. The overall information can summarize that most middle and old age respondents already have at least one child, but some respondents have more than five children in their families, and no respondent under 20 has children.

By looking at FIGURE 2 that shows the relationship between the respondent's feeling of life and the number of children he/she has, it is almost an evenly distributed scatter plot. All respondents rated their feelings of life on a scale of 0 to 10 which 0 is the lowest along with 10 is the highest. For respondents with 2 to 4 children, the scale was fully rated from 0 to 10. For respondents with 5 to 7 children, the rate is mostly above level 2. It is still hard to conclude that respondents' feelings of life are related to the number of children they have. Due to each point on the scatter plot is distributed uniformly, no trend or difference among groups is showed from the plot. Thus, respondents' feelings of life may have no or very weak relationship with the number of children they have.

By looking at FIGURE 3, the scatter plot of total children each respondent has and the age of each respondent when the first child was born shows a downward trend. Firstly, for respondents with 1 to 2 children, the range of their ages at the birth of the first child is from under 20 to 45 and It is the most normal phenomenon nowadays. Secondly, for respondents with 3 to 4 children, their ages at the birth of the first child have some changes. Fewer respondents have their first child at 35. Furthermore, for respondents with 5 or more children, most of them have their first child under 30. These respondents have already had many children that they must have their first child when they were young, due to the chance of having a child decreases with age. Therefore, the information collected through the scatter plot can summarize that the total number of children of each respondent has a relation with the age of each respondent when the first child was born, but this relation is not as strong as the relation with the respondent's age.

### Discussion

Figures from the above result part show three different relations among the total children respondents have, respondents' age, and their rates of life feelings. Through creating the scatter plot and the data modeling, we can demonstrate that the total number of respondents have has a relation with their ages and their ages at the birth of their first child. Compared these two relations, the total number of children of each respondent and the respondent's age have a stronger relationship. However, the total number of children of each respondent does not seem to have much to do with the respondent's feeling of life.

We make this study to find out how respondents themselves affect the number of children they have. We initially thought that the respondent's age, the respondent's age at birth of a first child, and their feelings of life would all have an impact on the number of children in each family. However, the feelings of life are failed to be a factor in this situation, and the other two variables can be seen as significant factors.

## Weaknesses

First of all, the weakness of the analysis is our model is too simple, and it's a very basic model in statistics. There are many other more accurate models to help to do the analysis, but we have no ability or no enough knowledge to use them yet.

Secondly, the dataset is not large enough to represent the people in the whole of Canada. From research, we can know that there are 37.89 million people live in Canada in 2020. In this data set, there are only 20,602 people took this survey.

Thirdly, the actual number of total children is not very accurate, which means it's not efficient for making the analysis. The total number of children in each family would include the adopted children. The adopted children would affect the relationship between the total children and the other three variables, which would also make us think the predict variable has weak correlation with other three variables.

## Next Steps

First, we could use a more accurate model to do this analysis when we have enough knowledge, for example, we could use the Bayesian Inference to do it.

Secondly, when collecting the data, the people who are collecting data should increase the sampling size.

Thirdly, maybe it should add one question in the survey: how many children in your family are biologically related to you? After that, we would use the data collected by this question to know well about the correlation between the total children and the other three variables.

### References

Gagné, C., Roberts, G. and Keown, L.-A. (2014) "Weighted estimation and bootstrap variance estimation for analyzing survey data: How to implement in selected software". The Research Data Centres Information and Technical Bulletin. (Winter) 6(1):5-70. Statistics Canada Catalogue no. 12-002-X. http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002- X20040027032&lang=eng

https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf