The relationship between ‘total number of children each Canadian family’ and ‘age of the participants at first birth, place of their birth outside or in Canada and their sexes’

Wen Huang

2020-10-19

The relationship between ‘total number of children each Canadian family’ and ‘age of the participants at first birth, place of their birth outside or in Canada and their sexes’ Wen Huang 2020-10-19

Part 1: Abstract

The following report is going to focus on the study of the correlation between the variables total_children, age_at_first_birth, sex and place_birth_canada. We are going to see if age_at_first_birth can correlate to total number od children the Canadian families have. Stratified sampling will be used when collecting the data and multiple linear regression model will be used when estimating the response variable using the predictors.

Part 2: Introduction

Our report is going to focus on the study of the correlation between the variables total_children, age_at_first_birth, sex and place_birth_canada. We are going to see if age_at_first_birth can correlate to total number of children the Canadian families have. Stratified sampling will be used when collecting the data and multiple linear regression model will be used when estimating the response variable using the predictors.

The data we are going to use is selected from GSS 2017 survey data. We are going to focus on the topic described above to gain insight about Canadians’ idea of fertility and how important they think to have children in their life.

In the following report, we are going to discuss about the data we are going to use in our report and the weaknesses and advantages of these data. Then we discuss about the methodology we use when collecting the data. Afterwards, we explain the reason why we choose multiple linear regression model and use it to create a fitted model for the variables we are interested in. We will then analyze the result and discuss about the weaknesses of the study. Then conclusion will be included at last to end this report.

Part 3: Data

The data was selected from surveys related to Canadian families, conducted by GSS in 2017. These data were selected because it provides us with insights into Canadian families and other issues related to them. As mentioned in the user guide about this survey,(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017) the Canadian families are becoming more and more diverse now. There might be more single-parent families now or there are more families with gay or lesbian families. I am very interested about these changes and would like to explore, for instance, how would features of Canadian families affect future Canadian policies, or how are the features of Canadian families from the data recently collected help me gain insight into the social change, custom changes and social conditions, etc.

There some important features about the data I selected.

Firstly, the data includes two parts: “core content” and “classification variables”(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017) . “Core content” helps us learn about social changes about people’s health and living conditions and also helps us to decide what policies to change or create. “classification variables” classify participants of the survey into different groups according to their features.(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017)

I think this is a very precise, clear and comprehensive way of collecting data. By collecting “core content”, we can get the direct answer to the questions we would like to know; afterwards, we collect “classification variables” data to let us categorize the “core content” data from their sources and able to analyze them more easily to find specific patterns about these data. The participants, or the target population of this set of surveys are 15 years old or order persons in Canada, excluding “residents of Yukon, northwest territories and Nunavut and also full-time residents of institutions”(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017).The target population includes people of different age, from different region and different family conditions, etc. But when we would like to see for instance, whether the older the participants, more children they will have, not all the participants’ answer can help us gain better insight as those participants who are under 20 years old commonly may not have any children.

Another advantage of including these two components is that this can help us avoid reaching unreasonable conclusions or get strange patterns of the data. Using the example we talked about above, for participants who are under 20 years old, they commonly don’t have children. If we include them when summarizing the data and get a conclusion about the question “do persons who are above 20 years or older have more children?”, we may get a wrong conclusion as those people who are under 20 years old are included in this analysis.

Another feature of these data is that stratified random sampling is used. This method can provide us with greater precision. It can more accurately reflects population studied.It can also help us eliminate any unrelated or unnecessary samples. Also, the targeted sample size is 20,000 and the actual number of respondent is 20,062(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017). That is, the sample collected is 20,062 persons who are 15 years old or order persons in Canada, excluding “residents of Yukon, northwest territories and Nunavut and also full-time residents of institutions”(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017). Using stratified sampling method enables us to get data better organized and makes the data easier to analyze. Thus the data I selected is very easy and convenient for us to use during analysis. However, there is a potential drawback of stratified sampling method, which is it can be tedious and time consuming as what the strata(the groups we classify these samples we collected into) to use and how to stratify are difficult to use and may require some pre-survey research.

The size of the sample of this survey is 20,062. The main goal of this survey is to study the families in Canada. Compared to the whole Canadian population, according to Statistics Canada at the end of 2017(the year when this survey was conducted), 36,721,242. Compared to the total population size, the sample size is relatively small so one weakness of these data could be that they are not representative enough. The conclusion we get from these data may not be persuasive enough, either.

The survey frame was created by including two components: lists of telephone numbers in use available to Statistics Canada from various resources and the Address Registrar. (General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017) Most data(84%) was from the Address Registrar and the others were from Statistics Canada. (General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017) There could be more than one phone numbers related to an address and we use the first one we can get and we prefer landline to cell phone numbers. (General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017)

Data for this 2017 GSS survey was collected via computer assisted telephone interviews.(CATI) The telephone number are Interviewers were given two language options which were English or French. To be eligible for the survey, the household called should at least include one 15 years old or older person. For those who refuse to participate, they will be called several more times and will be encouraged to participate; for those calls which are not convenient to responded when called, the calls will be rescheduled; for those which are not responded, the calls backs will be made(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017). CATI may bring us weaknesses related to the data. Firstly, the languages the participants may choose are only French or English. For immigrants or other people from different nations, they may not be able to use these two languages very well and may misunderstand the interview questions and provide reliable information. So the reliability of the data could be not very good.

Another thing to be noticed about CATI is that the overall response rate was 52.4% for 2017 GSS. This further reduces the sample size and broadness of the data. It may, as a result, limit the reliability of the persuasiveness of the data collected.

The questionnaire of the survey includes the following parts(General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide,2017): Entry component (respondent’s date of birth) Family origins Leaving the parental home Conjugal history Intentions and reasons to form a union Respondent’s children Fertility intentions Maternity/parental leave Organization and decision making within the household Arrangements and financial support after a separation/divorce Labour market new and education Health and subjective well-being Characteristics of respondent’s dwelling Characteristics of respondent of spouse/partner

The advantages of the questionnaire is that it asked the date of birth of the respondents. This helps us avoid the inclusion of samples which do not meet our requirement and improve the accuracy of the following analysis.The broadness of the survey content is also very good so that when making decisions on related policies, the government are able to more comprehensively understands current situations of Canadian population and families and come up with more applicable and helpful policies.

There is also a downside of the questionnaire which is that it may cover too many questions. This is the trade-off caused by the broadness of the survey. As we mentioned before, this survey was conducted on phone. The language could be a factor that affects the accuracy of transferring the questions to the participants and including so many questions may make this survey even more difficult for people who can speak neither English or French very well. Also, the survey is conducted through a call, without any text or images, so the even for those participants can are native speakers, the questions could sometimes be unclear and they may sometimes misinterpret the questions. Also, considering people’s life pace nowadays, some people may not have enough patience to carefully participate in this survey and thus may provide unreliable information.

Part4: Model

From the GSS 2017 data, I selected four columns. They are total_children, age_at_first_birth, sex and birth_place in Canada.I choose these data because I think we can get some very interesting insights and conclusion about Canadian society and Canadian’s idea about topics related to family, marriage and children.

I choose multiple linear regression model and the software I use is R studio. Before we creating the model, let’s first look at different parts of the data we will use in the model.

I draw the following three figures which are Figure 1, Figure 2 and Figure 3. Figure 1 shows us the number of families which have different number of children. We may notice that most families have 0-3 children. The amount of families which have 0 and the amount of families which have 2 children are the highest, exceeding 6000 respectively. A relatively small proportion of Canadian families have 4 or more than 4 children. One thing very interesting here is that more than 6000 Canadian families don’t have any children. This could be explained by the increasing acceptance of the idea of “DINK family” in the west, or more specifically, in North American region(Thomas Baudin et al, 2013). The decreasing fertility rate is a phenomenon that is worth of attention in Canada as the fertility rate has hit the record low in 2019 and may even be lower in 2020 due to Covid-19(Miriam Halpenny,2020).

Figure 2 shows the relationship between total_children and age_at_first_birth. We may see that the earlier a person have his or her first child, the more children in total he or she will have. This may imply that the younger a person has her or his first children, the more willing he or she is to have more children. The reason for those people who have their first child at a larger age could probably be the difficulty for them to get a child at an older age or there could exist danger during pregnancy or simply because their prefer the idea of “DINK family”.

Figure 3 shows that for those people who were born outside Canada, they will have their first child born slightly earlier than those who were born in Canada. The reason why such relationship exists can be cultural difference the participants grew up in.

Now, let’s get into the multiple linear regression model.

We choose to use multiple linear regression model because we would like to understand the value of one variable, which is total_children in this case, based on the values of more than one variable, which are age_at_first_birth, sex and place_birth_canada in this case. Using multiple linear regression model also enables us to determine the overall variation explained of the model and how each explanatory variable contribute to the total variation explained.(Laerd Statistics, 2018)

In order to see if we can apply multiple linear regression model to this case, we are going to see if the data we investigate on meet some basic assumptions of multiple linear regression model(Laerd Statistics, 2018). Although, it can be noticed that it is not necessary for the data to meet all the requirements as in reality the data could be complex. There can always be a solution(Laerd Statistics, 2018).

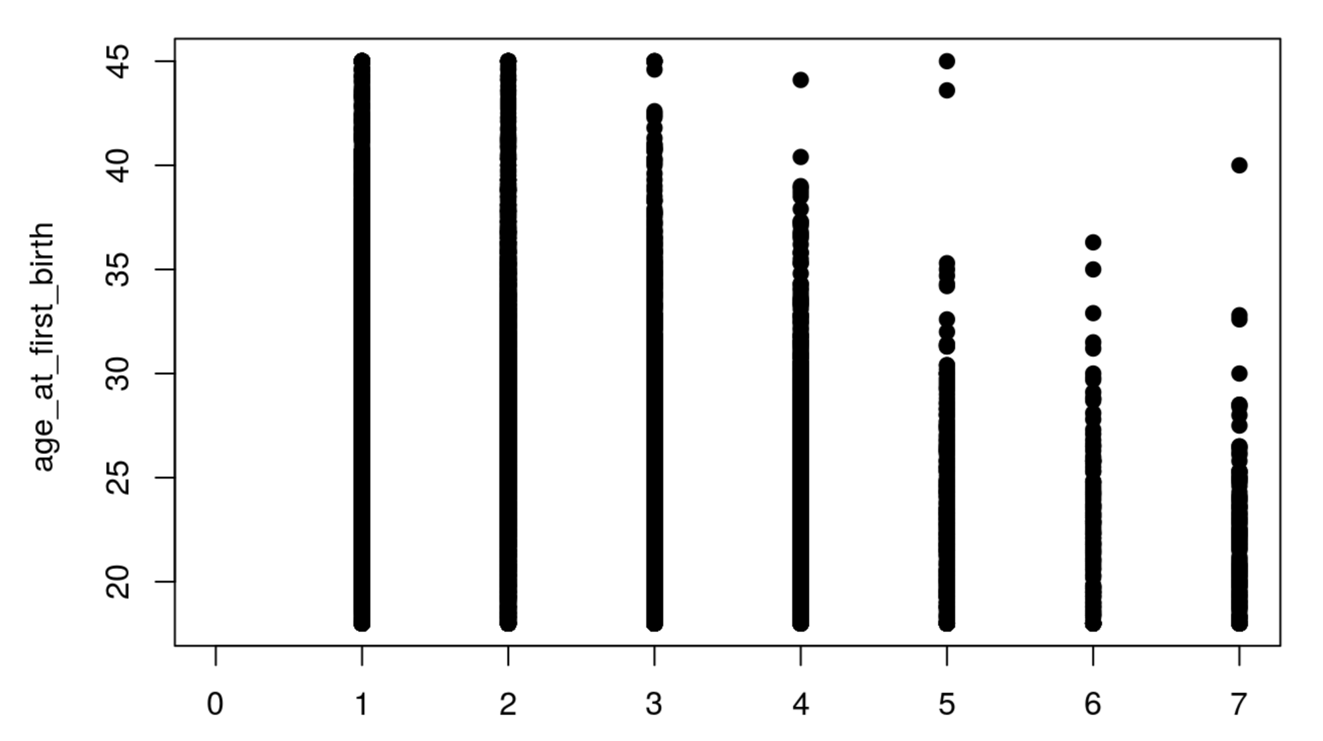
The first basic assumption that the data we use is that our response variables should be measured on a continuous base(Laerd Statistics, 2018). Back to our data, the response variable is total number of children in a family and it is an interval variable, which is also a continuous variable.

The second assumption is that the independent variables, or the explanatory variables should be either continuous or categorical(Laerd Statistics, 2018). We have three explanatory variables here, which are age_at_first_birth, sex and place_birth_canada. Age_at_first_birth is a continuous variable; sex and place_birth_canada are two categorical variables so this assumption is satisfied as well.

The third assumption is that we should have independence of observations(Laerd Statistics, 2018). We could use Durbin-Watson statistic to test this but there is a easier way to analyze. The way these data that GSS collected in 2017 was mainly from Statistics Canada. The survey method is stratified sampling so the samples were randomly selected from the target population. Thus it can be almost certain that the independence of observations assumption is satisfied.

The fourth assumption is that there should exist a linear relationship between that response variable and each explanatory variable. There should also exist a linear relationship between the response and explanatory variables collectively(Laerd Statistics, 2018). This can be proved by drawing scatterplot graph. Thus the following scatterplot graphs are created.

Figure 4: relationship between total_children and age_at_first_birth



From the figure above, we may approximately see an overall linear relationship between total_children and age_at_first_birth. For other two independent variables, sex and place_birth_canada, the scatterplots can be drawn but we first assume there exist a linear relationship between them and total_children. The scatterplot for the response and explanatory variables collectively will be included after we create the multiple linear regression model.

There are several other assumptions that needed to be satisfied, including homoscedasticity (which is where the variances along the line of best fit remain similar as you move along the line), multicollinearity (which occurs when you have two or more independent variables that are highly correlated with each other), no significant outliers, high leverage points or highly influential points, residuals (errors) are approximately normally distributed(Laerd Statistics, 2018). We will test these assumptions after we create the multiple linear regression model as it is not yet able to be discussed before the model is created.

As it is already confirmed that we can apply multiple linear regression model to this case, now we will go into more details of the model we are going to create. The response variable for this multiple linear regression model is total_children, which is the total number of children each family have. The explanatory variables for this model are age_at_first_birth(meaning the age at which the participants have their first children), sex(the sex of the participants) and place_birth_canada(whether the participant was born in Canada or not). This model helps understand if age_at_first_birth, sex and place_birth_canada are the factors that affect the total number of children a family has. It can provide us with more accurate and precise understanding about the association between each explanatory variable and the response variable(Keith A. Marill, MD, 2003).

The general form of multiple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i$$

$$\beta_0 \; is \; the \; intercept \; value \; and \; \beta_1 \ldots to \; \beta_p \; are \; the \; coefficients \; of \; the \; explanatory \; variables. \epsilon_i \; is \; the \; random \; error.$$
$$y_i \; is \; the \; response \; variable \; value \; and \; x_{i,1} \ldots \; x_i, p \; are \; the \; explanatory \; variables.$$

It is expected that the multiple linear regression model we create will be:

$$total \widehat{\; children} = \beta_0 + \beta_1(age \; at \; first \; birth) + \beta_2(sex) + \beta_3(place \; birth \; canada)$$

Notice that the independent variables sex and place birth canada here are categorical variables as the sex variable only include male and female and the answer to place birth canada is only yes or no. Thus only 0 and 1 are used when indicating these two variables. We choose age rather than age_groups is because this will help us more clearly and precisely understand the data and results will be more specific instead of being too general.

The response variable total_children here is an estimated value so we add a hat here to indicate this.

If the model is lack of convergence, then this indicates that the data do not fit the model very well because there are too many poorly fitted observations(rasch.org, 2020)

To check the model, we need to see linearity, variance homogeneity and deviations from normality(distance to the line) and also graphical checks.

Part 5: Results

```
##
## Call:
## lm(formula = total_children ~ age_at_first_birth + sex + place_birth_canada,
##     data = data_HW_9248)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1861 -0.6858 -0.1792  0.4894  5.5241
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            4.371429   0.048906  89.385   <2e-16 ***
## age_at_first_birth                    -0.077739   0.001855 -41.913   <2e-16 ***
## sexMale                                0.214014   0.020214  10.587   <2e-16 ***
## place_birth_canadaBorn outside Canada -0.015027   0.024000  -0.626    0.531
## place_birth_canadaDon't know           0.224195   0.198285   1.131    0.258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.084 on 12678 degrees of freedom
##   (7900 observations deleted due to missingness)
## Multiple R-squared:  0.1242, Adjusted R-squared:  0.124
## F-statistic: 449.6 on 4 and 12678 DF,  p-value: < 2.2e-16
```

From the summary above, we may get the value of

$$\beta_0, \; \beta_1, \; \beta_2, \; and \; \beta_3$$

(two decimals):

$$\beta_0 = 4.37$$
$$\beta_1 = -0.08$$
$$\beta_2 = 0.21$$
$$\beta_3 = -0.02$$
$$\beta_4 = 0.22$$

By getting these values, we now can summarize the multiple linear regression model we get. Notice that for the sex variable, we assign 1 to male and 0 to female. For place_birth_canada variable, we assign 1 to outside canada and 0 to in canada.

For males who were born outside Canada, the estimated multiple linear regression model is:

$$total \widehat{\; children} = 4.37 - 0.08(age \; at \; first \; birth) + 0.21 - 0.02$$

In this model, when age at first birth increases by one year, on average, the number of total children decreases by 0.08, given other predictors held constant.

For females who were born outside Canada, the estimated multiple linear regression model is:

$$total \widehat{\; children} = 4.37 - 0.08(age \; at \; first \; birth) - 0.02$$

In this model, when age at first birth increases by one year, on average, the number of total children decreases by 0.08, given other predictors held constant.

For males who were born in Canada, the estimated multiple linear regression model is:

$$total \widehat{\; children} = 4.37 - 0.08(age \; at \; first \; birth) + 0.21$$

In this model, when age at first birth increases by one year, on average, the number of total children decreases by 0.08, given other predictors held constant.

For females who were born in Canada, the estimated multiple linear regression model is:

$$total \widehat{\; children} = 4.37 - 0.08(age \; at \; first \; birth)$$

In this model, when age at first birth increases by one year, on average, the number of total children decreases by 0.08, given other predictors held constant.

From the summary of the data, we may also see that the standard error of age_at_first_birth is 0.001855, which is relatively small. The p-value for age_at_first_birth is smaller than 2×10^{-16}, which is much smaller than 0.05(we assume the significance level is 0.05). Thus for the hypothesis test in which $H_0 : \beta_1 = 0 \; and \; H_a \neq 0$, the null hypothesis is rejected and there does exist a correlation between total_children and age_at_first_birth.

Part 6： Discussion

From the analysis of the model above, we may see that for different sexes and places where the participants were born, there does exist a linear correlation between independent variable age_at_first_birth and the dependent variable total_children. They are negatively correlated and the correlation is moderately weak.

We may see that the later the Canadian families have their first child, the less children they will have in total. So there does exist a correlation between total number of children and the age at which they have their first children. There will be only very slight difference for different sexes and birth place(in or outside Canada).

As the sample size of this survey is relatively small, we will not be very confident to say that our result will be very representative. For future work, we may further investigate into why the earlier people in Canada have children, more children they will have in total? Is there any implications about Canadian society or people’s opinion about family and children?

The weakness for future work is that our conclusion from this study is not very representative and may lead us to the completely wrong direction in further research. So the representativeness of this study shoud be further investigated on in the future.

Part7: Weaknesses

The weakness of this study could be that the result is not representative enough. As we only collected about 20,000 samples from the survey. Compared to the whole population in Canada, although we collect data from all different regions in Canada, the sample size may still be too small.

Another weakness is that there may not exist a direct relationship between total_children and age_at_first_birth. For those families who have their child very early, they may not have many children. The could exist systematic bias for our conclusion.

Part 8： References 1. Rohan Alexander and Sam Caetano，license:MIT, file: gss_cleaning.R, 7 October 2020

2. Government of Canada, Statistics Canada. Population estimates, quarterly. Government of Canada, Statistics Canada, September 29, 2020. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901.
3. Pascale Beaupré, Statistics Canada- Diversity and Sociocultural Statistique. General Social Survey-Cycle 31: Families-Public Use Microdata File Documentation and User’s Guideuser guide, April 2020.
4. Murphy, Chris B. “Pros and Cons of Stratified Random Sampling.” Investopedia. Investopedia, August 28, 2020. https://www.investopedia.com/ask/answers/041615/what-are-advantages-and-disadvantages-stratified-random-sampling.asp.
5. Baudin, Thomas, et al. “Fertility and Childlessness in the United States.” American Economic Review, vol. 105, no. 6, 1 June 2015, pp. 1852–1882, 10.1257/aer.20120926. Accessed 19 Oct. 2020.
6. Halpenny, Miriam. “Canada’s Fertility Rate Has Hit a Record Low - Canada News.” Castanet. Accessed October 19, 2020.
7. Marill, Keith A. “Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression.” Academic Emergency Medicine, vol. 11, no. 1, Jan. 2004, pp. 94–102, 10.1111/j.1553-2712.2004.tb01379.x. Accessed 17 Dec. 2019.
8. “How to Perform a Multiple Regression Analysis in SPSS Statistics | Laerd Statistics.” Laerd.Com, 2018, statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php.
9. “Estimation: Iteration and Convergence.” Www.Rasch.Org, www.rasch.org/rmt/rmt11b.htm. Accessed 19 Oct. 2020.
10. Multiple Regression, https://publicifsv.sund.ku.dk/~lts/engelsk_basal/overheads/multiple_regression4.pdf