Exploration of Factors that Influence One's Intention of Having Children Using Logistic Regression Model

Mengxin Zhao 1004762296

December 21, 2020

Code and data supporting this analysis is available at: https://github.com/jomczyhao/Exploration-of-Factors-that-Influence-One-s-Intention-of-Having-Children-Using-Logistic-Regression-M

I. Abstract

Facing the situation of the globally aging of population, it is important to understand the reasons behind the declining fertility rates. This study aims to find the main factors that affect one's intention of having children, based on the General Social Survey on Family (2017) conducted by Statistics Canada. The analysis shows that one's choice of having children or not is related to his/her personal features and conditions.

II. Keywords

Aging of population; Fertility rate; Intention of having children; Logistic regression model

III. Introduction

With the progress in society and economics, many new issues have emerged. The aging of population is one of the consequences of development, and has become a serious problem for the government agencies world wise. The aging of population has both advantages and disadvantages. According to U.S Department of Health and Human Services, on the good side, it "expanded social benefits, increased home ownership, elder-friendly housing, and an emphasis in many nations on community care". However, the aging of population also reduces the labour force participation rate, increases the dependency rates and increases the government's expense on health care and pensions. The aging of population is a global issue. Based on the data from World Population Prospects: the 2019 Revision, 16% of the population in the world will be over age 65 by 2050, compared to the 9% in 2019. Measures need to be taken to alleviate the situation before this global issue becomes even more serious.

According to WHO (World Health Organization), the two main causes of the aging of population are the increased life expectancy and the decline in fertility rate. In this study, the latter cause will be focused on. People in many countries nowadays intend to have less kids, or even no kid at all. It is concerning that whether Canada is in the same situation, and more importantly, the factors that affect the Canadian residents' choice between whether to have children or not should be investigated, in order to make the future plans for relieving the pressure.

In this report, the population who has either kids or the intention to have kids will be compared with the population of people who don't have kids, nor the intention. After that, a logistic regression model will be fitted onto the data to see the potential factors that affect people's choice of having a kid or not, so that the interaction between the variables and the outcome can be studied.

The data from General Social Survey on Family (2017) will be used to investigate the proportion of people who plan to not have kids and the potential reasons behind both choices. In the Methodology section, the modified data will be explained, and the model that is used to perform the analysis will be introduced. The result of the analysis will show under the Results section. The inferences of the data and the conclusions based on this analysis will be provided in the Conclusion section. All the references that are used in this study will be listed at the end of the report.

IV. Methodology

Data:

In order to investigate the topic of whether people have children or the intention to have children, the data in General Social Survey on Family (2017) is selected among all the Canadian general surveys, as it monitors changes in Canadian families and provides information and sights on social issues, and is the newest general survey on family. The target population is all Canadian residents. As it is a voluntary survey, there is no strict frame population. The sampled population is the people who participated in the survey. Since it is a voluntary survey, the respondents participated in the survey and provided their answer voluntarily. For the questions which the respondents didn't answer, it is recorded as NA in the dataset.

The data are divided into two groups based on their responses on these two variables -

total_children - "Total number of children reported by respondent"

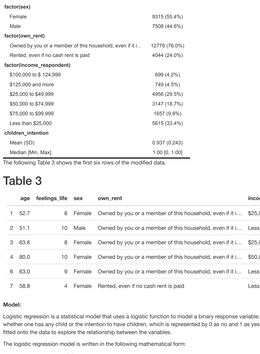
future_children_intention - "Total number of children intending to have"

If one doesn't have any child and denied the intention of having children in the future by answering "No, definitely not" or "Probably not", he/she will be marked as 0 for children_intention, meaning the one has no child, nor the intention to have children. In contrast, if one has had children, or affirmed the intention to have children in the future by answering "Definitely yes" or "Probably yes", he/she will be marked as 1, meaning the one has children, or the intention to have children, or both.

In this way, children_intention becomes a binary variable, which would result in answers of 0 or 1.

Figure 1 is a bar plot that shows the counts of people who doesn't have any child, nor the intention to have children (represented by 0), and who do have children or the intention to have children (represented by 1). According to the bar plot, it is obvious that the difference in population counts is significant. The majority of the population do have at least one child, or are planning to have children in the future.

Figure 1



This raises the question of what factors would potentially have impact on one's choice of having children or not. This will be the main focus of this report.

General Social Survey on Family (2017) contains a variety of variables coming from different aspects of life collected from the respondents. As this study focuses on the intention of having children and the potential influential factors, the variables which are related to the topic are selected from the original dataset. The selected variables include -

age - "Age of respondent with decimal at time of the survey interview"

feelings_life - "Feelings about life as a whole"

sex - "Sex of respondent"

own_rent - "Dwelling type of the respondent"

income_respondent - "Income of respondent - Total (before tax)"

The missing and unsure responses are removed from the data.

Among the five selected variables, age and sex are personal features. These two variables represents the potential internal causes that might influence one's intention of having children. One might have different opinions towards having children, or have different life goals due to their age and gender.

The other three variables - feelings_life, own_rent and income_respondent - reflects one's personal conditions. These potential external causes can also affect one's choice, as the surroundings can influence one's perspectives on children. One also needs to think about his/her situation before he/she makes such an important decision of having children or not.

Table 2 contains the baseline characteristics of the data that were mentioned. From the table, the respondents have an average age of 52.6, which means that the survey tends to reflect the features of the population that are relatively more mature. Most of the respondents own their fixed asdoe instead of renting a place, and have a relatively lower amount or a moderate amount of income.

Table 2

	Overall (N=16823)
age	
Mean (SD)	52.6 (17.8)
Median [Min, Max]	54.3 [15.0, 80.0]
feelings_life	
Mean (SD)	8.13 (1.62)
Median [Min, Max]	8.00 [0, 10.0]
factor(sex)	
Female	9315 (55.4%)
Male	7508 (44.6%)
factor(own_rent)	
Owned by you or a member of this household, even if it i...	12779 (76.0%)
Rented, even if no cash rent is paid	4044 (24.0%)
factor(income_respondent)	
$100,000 to $ 124,999	699 (4.2%)
$125,000 and more	749 (4.5%)
$25,000 to $49,999	4956 (29.5%)
$50,000 to $74,999	3147 (18.7%)
$75,000 to $99,999	1657 (9.8%)
Less than $25,000	5615 (33.4%)
children_intention	
Mean (SD)	0.937 (0.243)
Median [Min, Max]	1.00 [0, 1.00]

The following Table 3 shows the first six rows of the modified data.

Table 3

	age	feelings_life	sex	own_rent	income_respondent	children_intention
1	52.7	8	Female	Owned by you or a member of this household, even if it i...	$25,000 to $49,999	1
2	51.1	10	Male	Owned by you or a member of this household, even if it i...	Less than $25,000	1
3	63.6	8	Female	Owned by you or a member of this household, even if it i...	$25,000 to $49,999	1
4	80.0	10	Female	Owned by you or a member of this household, even if it i...	$50,000 to $74,999	1
6	63.0	9	Female	Owned by you or a member of this household, even if it i...	Less than $25,000	1
7	58.8	4	Female	Rented, even if no cash rent is paid	Less than $25,000	1

Model:

Logistic regression is a statistical model that uses a logistic function to model a binary response variable. In this study, the dependent variable is whether one has any child or the intention to have children, which is represented by 0 as no and 1 as yes. Thus a logistic regression model will be fitted onto the data to explore the relationship between the variables.

The logistic regression model is written in the following mathematical form:

$$\begin{aligned} log(\frac{(p)}{1-(p)}) = \beta_0 + \beta_1 X_{age} + \beta_2 X_{feelings\_life} + \\ \beta_3 X_{sex:female} + \beta_4 X_{own:male} + \\ \beta_5 X_{own\_rent:rented\ by\ you\ or\ a\ member\ of\ this\ household} + \beta_6 X_{own\_rent:rented,\ even\ if\ no\ cash\ rent\ is\ paid} + \\ \beta_7 X_{income\_respondent:\$125000\ and\ more} + \beta_8 X_{income\_respondent:\$100000\ and\ \$124999} + \\ \beta_9 X_{income\_respondent:\$75000\ to\ \$99999} + \beta_{10} X_{income\_respondent:\$50000\ to\ \$74999} + \\ \beta_{11} X_{income\_respondent:\$25000\ to\ \$49999} + \beta_{12} X_{income\_respondent:less\ than\ \$25000} \end{aligned}$$

In this formula, β_i's are the parameters of the logistic regression model. β_0 represents the intercept of the model, which is the average value of the log odds of the response variable when all the predictor variables are 0. β_1 to β_{12} are the estimates for the slope of the corresponding predictor variable. The name of the predictor variables are marked as the subscripts under X's. For the categorical variables, such as sex, own_rent and income_respondent, they are treated as dummy variables. Rstudio is used to run the logistic regression model.

The population that will be used as the finite population correction is 36,708,083, which is the population of Canada in 2017 according to Statistics Canada. The reason why no other number is subtracted from the population (such as the age groups) is that, the intention of having children is not restricted by age. People at any age can have their future plan on children.

V. Results

Table 4

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.5091557	0.2328861	-10.774176	0.0000000
age	0.0620669	0.0017804	34.860919	0.0000000
feelings_life	0.2124258	0.0184703	11.500942	0.0000000
factor(sex)Male	0.2002339	0.0699812	2.861252	0.0042249
factor(own_rent)Rented, even if no cash rent is paid	-0.3990187	0.0748654	-5.329815	0.0000001
factor(income_respondent)$125,000 and more	0.3709279	0.2102249	1.807245	0.0707419
factor(income_respondent)$25,000 to $49,999	0.7515255	0.1508510	4.981905	0.0000006
factor(income_respondent)$50,000 to $74,999	0.5117005	0.1524434	3.356658	0.0007907
factor(income_respondent)$75,000 to $99,999	0.4548893	0.1658493	2.742787	0.0060985
factor(income_respondent)Less than $25,000	1.3329881	0.1533002	8.693948	0.0000000

Table 4 shows the summary of the model. According to the output, most of the variables have small p-value which are smaller than 0.05, except the variable for the income of respondent of $125,000 and more.

As the individual t-test for this variable is not significant, we fail to reject the null hypothesis of having children or the intention to have children is independent from people with income level higher or equal to $125,000. In other words, this variable is not useful for predicting the response variable.

Among the other useful explanatory variables, most of them have positive effect on the response variable. For every additional unit increase in age, we expect the log odds of having children or the intention to have children to increase by 0.0620669, and the increase is 0.2124258 for the rating of life satisfaction. Males have an extra 0.2002339 in the log odds of the response variable compared to females. All variables related to income also have positive influence. However, as the income level goes up, the amount of positive effect decreases steadily.

The only predictor variable that causes negative effect on the log odds of the response variable is the variable for people who rent for their dwellings. If one's dwelling is rented instead of owning, the log odds of having children or the intention to have children is expected to decrease by 0.3990187.

The final model is shown as:

$$\begin{aligned} log(\frac{(p)}{1-(p)}) = -2.5091557 + 0.0620669 X_{age} + 0.2124258 X_{feelings\_life} + \\ 0.2002339 X_{sex:male} - 0.3990187 X_{rented,\ even\ if\ no\ cash\ rent\ is\ paid} + \\ 0.3709279 X_{income\_respondent:\$125000\ and\ more} + 0.4548893 X_{income\_respondent:\$75000\ to\ \$99999} + \\ 0.5117005 X_{income\_respondent:\$50000\ to\ \$74999} + 0.7515255 X_{income\_respondent:\$25000\ to\ \$49999} + \\ 1.3329881 X_{income\_respondent:less\ than\ \$25000} \end{aligned}$$

VI. Discussion

Summary:

Base on the data from General Social Survey on Family (2017) conducted by Statistics Canada, the interaction between age, rating of life satisfaction, gender, dwelling type, income level and the chance of having children or the intention to have children was investigated. The result shows that most of the variables are significant and have a positive influence on the log odds of the response variable.

Conclusions:

First, the positive coefficient of the age variable indicates that the elder people usually have higher chance of having children or the intention to have children. In other words, younger people tend to not have children. It matches with the fact that the fertility rates have been steadily declining as younger generations are relatively less willing to have children. As the elder people usually have more children, the population of each generation is imbalanced, and would eventually reach to a point where there are way more old people than younger ones in the society.

Secondly, people would have a higher chance to choose to have children if their rating for life satisfaction is higher. This indicates that one's choice of having children or not is related to their overall life quality. It might seem contradictory as the aging of population is exactly the issue of the development in society and economics, where people's life quality has been improving. This is because feelings_life is based on individuals' feelings, but not the society as a whole. One would rate their life satisfaction based on their own personal conditions and the compare with other individuals, instead of comparing their life or the society they are in with the past ones. This can also explains the lower chance of people having children when they do not have a stable abode. Compared to people who own their dwelling places, the ones who have to rent the abode would tend to not have kids.

Next, the difference in gender is also noticeable, as males have a higher chance of choosing to have children compared to females. It might be because giving birth brings more disadvantages to females. Physically, around 15% of all pregnant women will encounter a potentially life-threatening complication, according to WHO. Other than that, mentioned in Becker S., Fernandes A. and Weichselbaumer D.'s paper on discrimination in hiring based on potential and realized fertility, females are more likely to be in charge of childcares, and it is more difficult for females who might have kids in the future to get a job. Females who decide to give birth to child would be under an unfair disadvantage in the society.

As for the income levels, the respondents are less willing to have children according. According to the article written by Balbo N., Billari F. and Mills M., the three theories behind the phenomenon are that, first, people warning more have a higher opportunity cost if they focus on childbirth and parenting rather than their continued career. Second, women who can economically sustain themselves have less incentive to become married. Thirdly, higher-income parents value quality over quantity and so spend their resources on fewer children.

The declining fertility rates is becoming more and more serious global wise. According to the logistic regression model in this study, in order to alleviate the situation, more career opportunities should be provided for the younger people to improve their overall life quality. The child benefit for the ones who have their abode rented should be increased to ease their burden on childbirth. It is also necessary to offer more physical, psychological and social supports to the females, so that they can be less worried about the potential disadvantages that childbirth would bring.

Weakness & Next Steps:

In General Social Survey on Family (2017), the completion rate for the variable total_children is 99.9%, whereas the completion rate for future_children_intention is only 34.8%. Due to the large amount of missing values in the data, the model fitted and the outcomes are less accurate. Also, the data we chose is from 2017. The data might be outdated, especially that many global events (such as the global pandemic of COVID-19) that have happened would change people's thoughts and behavior towards having children.

As the declining fertility rates is just one of the causes of the aging of population, the next steps of the study is to focus on other factors, such as the increased life expectancy and immigration in Canada.

VII. References

1. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Statistics Canada (2017). General Social Survey on Family. Retrieved from SDA @ CHASS
3. Alexander Rohan and Caetano S. (2020, Oct 7). GSS.cleaning. Retrieved from U of T Quercus
4. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
5. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
6. Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor
7. T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.
8. Benjamin Rich (2020). table1: Tables of Descriptive Statistics in HTML. R package version 1.2.1. https://CRAN.R-project.org/package=table1
9. Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra
10. U.S. Department of State (2007, March). Why Population Aging Matters: A Global Perspective. Retrieved from https://www.nia.nih.gov/sites/default/files/2017-06/WPAM.pdf
11. World Health Organization (2011, October). Global Health and Aging. Retrieved from https://www.who.int/ageing/publications/global_health.pdf
12. United Nations (2019). World Population Prospects 2019. Retrieved from https://population.un.org/wpp/
13. Statistics Canada (2018, March 27). Canada at a Glance 2018 Population. Retrieved from https://www150.statcan.gc.ca/n1/pub/12-581-x/2018000/pop-eng.htm
14. Becker S., Fernandes A. and Weichselbaumer D. (2019 August). Discrimination in hiring based on potential and realized fertility: Evidence from a large-scale field experiment. Retrieved from https://doi.org/10.1016/j.labeco.2019.04.009
15. Balbo N., Billari F. and Mills M. (2012, September 12). Fertility in Advanced Societies: A Review of Research. doi: 10.1007/s10680-012-9277-y