Logistic Regression: Analysis of factors affecting whether a person has a child

Zijun Ye, Yuhan Zhao, Zuoyu Wang

2020.10.16

Abstract

The decreasing birth rate of Canada is an interesting topic. Fortunately, the 2017 General Social Survey(GSS) on families has offered a wide range of data regarding individuals' social, financial, educational background. This survey provided an opportunity to study the factors affecting whether a person had a child. This study has conducted a logistic model with six predictors to analyse the probability for a person to have a child. The model can be considered as a guidance for constructing further studies on the topic of birth rate.

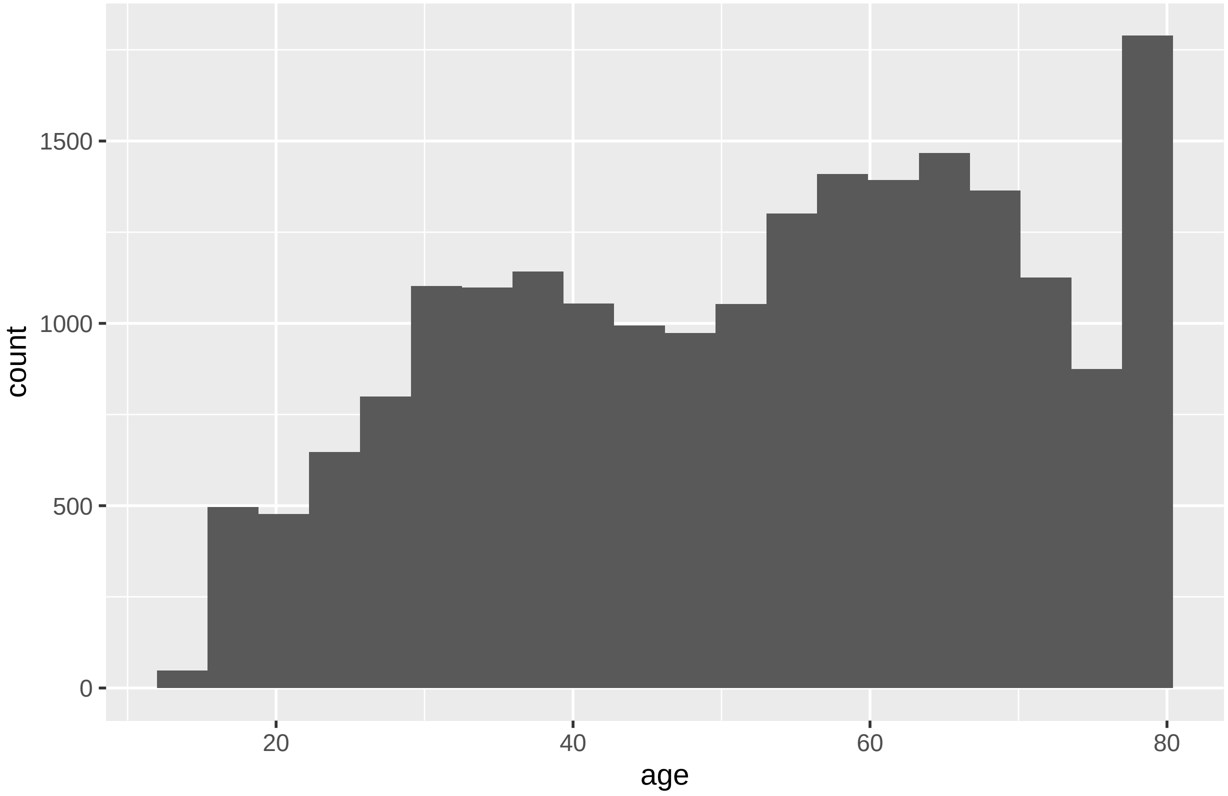
Introduction

Canada has had a low birth rate for a long time and the birth rate has reduced over years. There are many reasons that cause this phenomenon such as the individuals' satisfaction of life, their income level, education experience, religiousness level etc.. It's interesting to identify those individuals who are likely to have a child and what demographic characteristics they might have. This report summarizes the statistical model and analysis results associated with the main factors which would have affected individuals' tendency of having a child. The purpose of this report is to investigate which of those Canadian individuals with some particular features are more likely to have a child. The aim of this report is to build a logistic regression model that is able to make a good prediction on the probability of Canadian individuals ever having a child, and it will be helpful for the future study on solving the problem of the aging population.

Data

The original data is based on the 2017 General Social Survey (GSS) on the family.[1] It has been well cleaned in and formatted in a readable way. The provided dataset contains 81 factors consisting of both numeric and categorical measurement. It has a huge size of 20,602. Due to sufficient variables, this dataset has covered information of individuals from different perspectives such as family, economy, education, etc. As a result, this dataset is detailed enough for building a valid model. The following histogram and summary provided a brief overview of variable "age". The age for citizens involved in this age has mean of 53 with the minimum of 15 and the maximum of 80.

Figure 1. Histogram for Age



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   37.30   54.20   52.19   66.78   80.00
```

Among the 81 factors, the variables "ever_fathered_child" and "ever_given_birth" are primary concerns since it tells whether a male or female has ever had a child. Combining these two variables into a new variable "have_child" which will yield whether a person has ever had a child, which is the response variable of this study. Then the other 79 factors are considered as potential predictor variables. However, lots of them will be neglected since most variables contain huge amounts of missing values, which is a drawback of this dataset. Missing values will shrink the data size, so the data will be less informative.

Model

Since the response variable contains binary outcome, it is reasonable to build a logistic regression model with following equation: $log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$

In this equation, each x corresponds to a predictor which affects the probability of having a child, and each corresponding β is the parameter that will be estimated to indicate the scale of such effect, while β_0 is the bias term. p is the probability of whether a corresponding observation has a child or not. Once a valid model has been built, the probability for an individual of having a child can be estimated by calculating p in this equation.

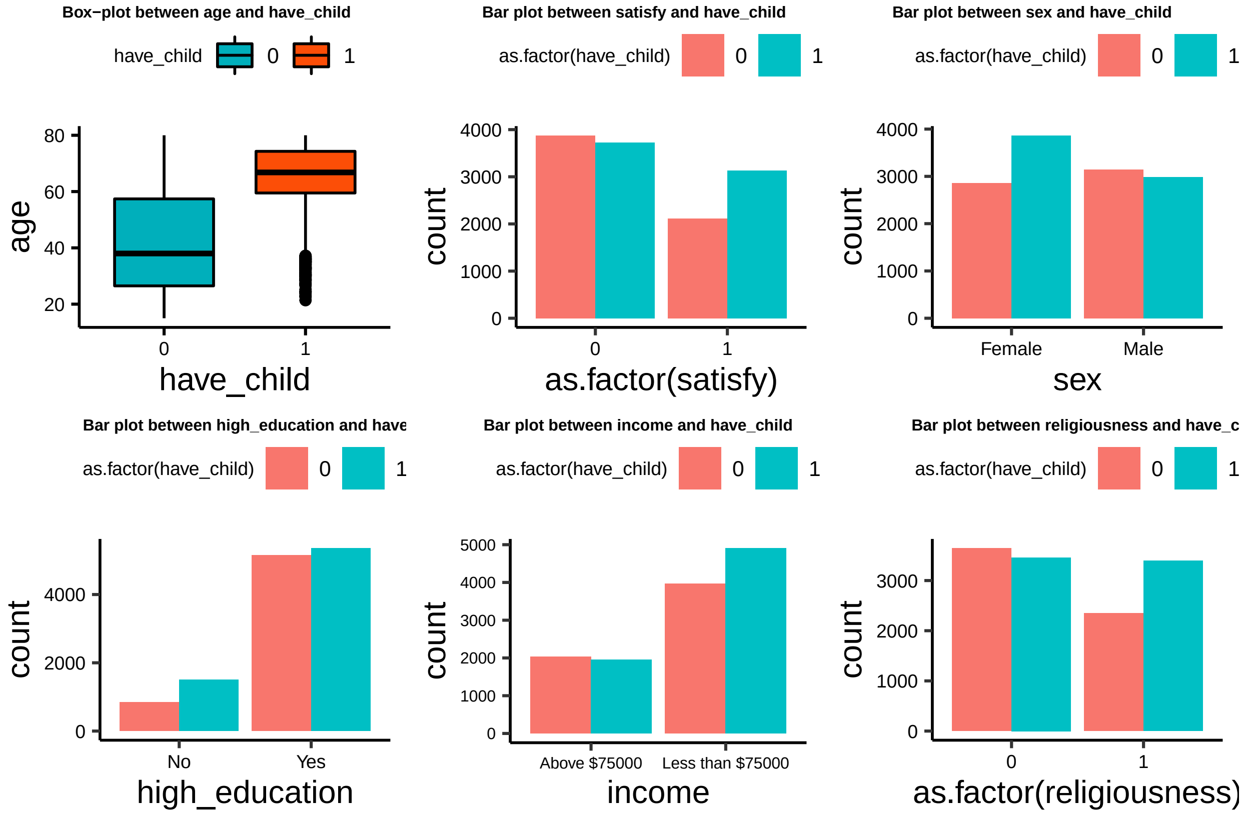
The model's coefficients β are usually estimated by using maximum likelihood estimation. In order to calculate the β that best fits our data, an iterative process associated with maximizing the log-likelihood of a Bernoulli distribution using Newton's method called IRLS is used. This process starts from setting tentative values for β, and repeating a slight revision to adjust the values until they converge to a point which no more improvement has been done. Fail to converge might be due to various reasons such as having a large ratio of predictors, multicollinearity, sparseness etc.. The AUC-ROC curve is used to check the model's goodness of fit. This curve tells how much a model is capable of distinguishing between classes. AUC indicates the area under the ROC curve. The higher the AUC, the better the model is at predicting for the true value. The model's estimated coefficients are required to be exponential which could be interpreted as the odds ratio of the interest, controlling other covariates constant and their confidence intervals are used to assess the significance of the estimates. [2]

Results

There are 79 variables considered to be the potential predictor variables. Since some of them contain huge amounts of missing value and some of them intuitively do not have a relationship with our response variable "have_child", most of the factors are eliminated. After taking into consideration all the factors which have strong potential relationship with the response variable, the candidate model contains six predictor variables which are shown in the table X below. These variables are being transformed to the new type in which make the relationship between each predictors and the response variable shown in the statistic plots and the further interpretation of the model coefficients easier to understand. The "satisfy" variable is transformed from the "feelings_life" variable in the original dataset, where value 1 indicates the rating score of an observation is greater than the average value of the sample, otherwise value 0 is assigned. If the education level of an observation has value of "Less than high school diploma or its equivalent", the new variable "high_education" has the value of "No", otherwise "Yes" indicates the education level of an observation is higher or equal to high school diploma. Moreover, the new rescaled variable "income" contains binary data which indicates whether their family annual income is "Less than $75000" or "Above $75000". The last rescaled variable "religiousness" is transformed from "religion_participation", which value 1 means that this observation has participated in religious activities, services or meeting at least once a year and value 0 means he/she has never participated. Figure 1 demonstrates the relationship between each of these 6 predictors and the response variable. The box-plot shows that observation with higher age tends to have a child. The rest of the bar plot illustrates that it seems like the observation who have the characteristics of satisfying with their life, is female, less education, less family income and having religiousness tend to have a child.

Figure 2

Combination plot of Six predictors and have_child



The mathematical notation of the candidate model can be express as follow:
$log(\frac{p}{1-p}) = \beta_0 + \beta_1$ age $+ \beta_2$ satisfy$+$ β_3 sex $+$ β_4 high_education $+$ β_5 income $+$ β_6 religiousness

where p indicates the probability that an observation has ever had a child. Table 2 illustrates the estimated coefficients β of each covariates and Table 3 shows the odd ratio and their confidence interval of each covariates which the coefficients β in the model could be interpreted better. In order to check the goodness of the model, the AUC-ROC curve of the model is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

Discussion

For our model, the candidate Model in the result part contains 5 categorical binary variables and 1 continuous variable and it has become the final model. In table 3, it shows the lower and upper bound of the confidence interval of odds ratio of each of the covariates. If the confidence interval does not contain 1, it indicates this covariate is significant. The odd ratio of age which can also express as exponential of $beta1$ can be interpreted as that when age increases by 1, the odds ratio of people have a child increases by 1.1044065, controlling other categorical variable at the reference group value, which indicates people have the greater the age would have the higher chance of having a child; The odds ratio of "incomeLess than $75000" can be interpreted as that comparing to people who have family income greater than $75000, the people whose family income less than $75000 has 0.7291754 times odds ratios, controlling other continuous covariates constant and categorical variable at the reference group value, which means that people who have less family income have less chance of having a child. The ROC curve shows the AUC is 86% which means that this logistic regression is able to distinguish whether people have a child 86 percent of the time.

In the 2017 General Social Survey On The Family, the target population is all non-institutionalized persons 15 years of age or older, living in the 10 provinces of Canada. The frame population is non-institutionalized 15 years old or order people who live in 10 provinces of Canada and has landline and cellular telephone numbers from the Census and various administrative sources with Statistics Canada's dwelling frame. And the sampling population are people who were invited to this survey. The sampling unit is person. Besides, these qualitative tests conducted by Statistics Canada's Questionnaire Design Resource Centre (QDRC) were conducted in two cities, with respondents selected according to representativeness criteria, so we can know this survey was a Two Stage Cluster sampling. And the main approach to reducing the unresponsive bias involves a series of adjustments to the survey weight to explain as much as possible of the unresponsive bias. For this survey, the most important good points is that the dataset is huge enough- 2% of the whole Canada population, so the result we get would be significant. And the survey was sponsored by the government, so that people who got interviewed would answer the question reliably. What's more, this survey contains 81 factors, which are convenient for scholars to do a large number of kinds of research. On the other hand, the 2017 General Social Survey On The Family still has some shortages. Firstly, it has lots of missing value, which is useless. Secondly, there are some non-sampling errors during the survey, like sampling bias, we will discuss the details in the Weakness section.

(1) Weaknesses

As mentioned in the Discussion Part, the biggest weakness of our research is that some factors of the dataset are missing value. In statistics, we call this non-response problem, it is a kind of non-sampling error. [3] Therefore, it can not be sure that the information is completely valid. Besides, although this survey uses Two Stage Cluster sampling, the sampling units are the groups of telephone numbers, which means people who do not use telephone can not be invited in this survey. Moreover, in this survey, respondents could be interviewed in French or English, in this way, people who can not use these two languages fluently(for example, indigenous) can not join this survey either. These also caused others non-sampling error, imperfect sampling population and selection bias. And since these data are based on a person's sample, sampling errors are easy to occur. In other words, sample based estimates vary from sample to sample, and the survey we are using right now was launched three years ago(2017), so our result may not work with the next survey. Last but not least, two variables we are mainly using are "ever_fathered_child" and "ever_given_birth", but there might be some special situation, for example, an aunt lives with her nephew, she is not the child's mother or adoptive mother, but she raising the kid. In this example,this family also has a child, but it cannot be counted in our model. In the model, we assume all the predictors are independent. However, in reality, some of them may have associations These potential correlations could impact on the results and reduce the accuracy of the model.

(2) Next Steps

In the next step, we can respond to our results with the related government to let them follow up. For example, the government may launch the policy to let families which have less income and low education but have kids get some economic subsidies. In fact, after this report, we already know the factors for a family or person to have a child, so we can move further in this aspect by using our model. For instance, we might research what is the correlation between the number of children and other factors, such as family income, education level of parents, religiousness of family, etc. to determine what kind of family has more probability of having one child, or more than one child. In addition, we can also make a model to research what kind of family or person tends to live with their grandparents or grandchildren.

References

[1] General social survey on Family (cycle 31), 2017 Retrieved 2020.10.15 from: http://dc.chass.utoronto.ca/myaccess.html

[2] Alan Agresti "An Introduction to Categorical Data Analysis" Second Edition. John Wiley & Sons, Inc. 2007. 99-163.

[3] Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-52.

Appendix

Table 1: Table 1. Variables Summary Table

Names	Types	Levels
have_child	Binary	[0,1]
age	Numeric	/
satisfy	Binary	[0,1]
sex	Binary	[Female, Male]
high_education	Binary	[Yes, No]
income	Binary	[Less than $75000, Above $75000]
religiousness	Binary	[0,1]

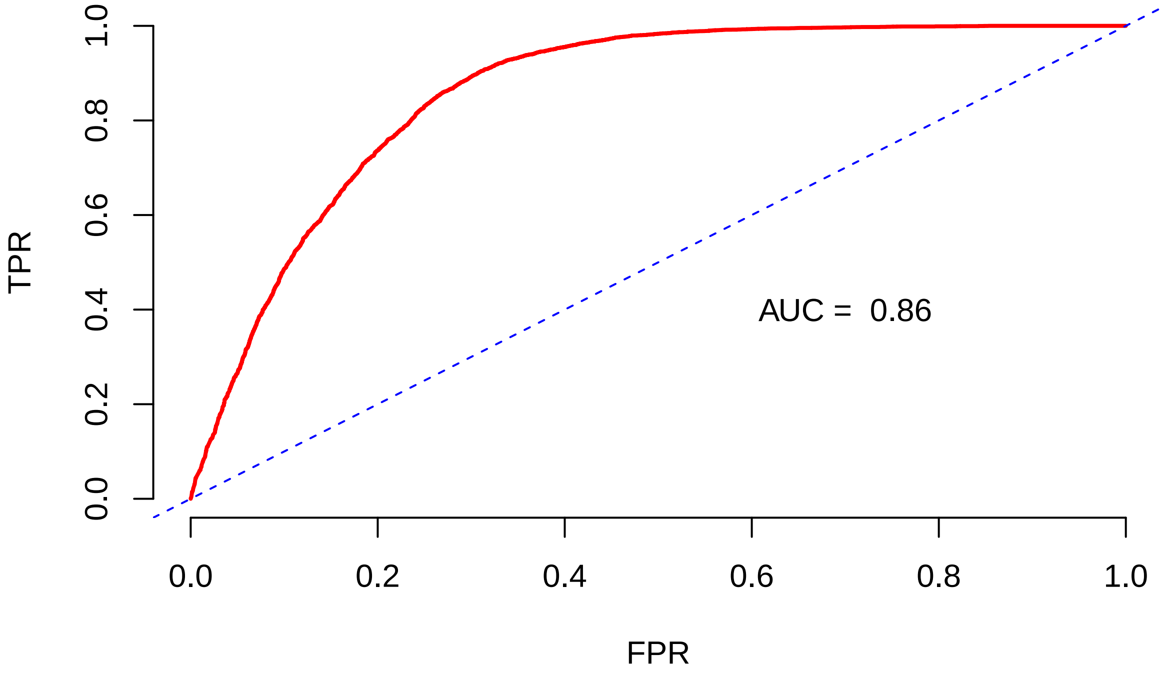
Table 2: Table 2. Model Coefficients Table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.9958704	0.1316605	-37.945086	0.0000000
age	0.0993081	0.0017803	55.782195	0.0000000
as.factor(satisfy)1	0.2251104	0.0478758	4.701970	0.0000026
sexMale	-0.2068588	0.0467626	-4.423592	0.0000097
high_educationYes	-0.3057569	0.0673355	-4.540798	0.0000056
incomeLess than $75000	-0.3158410	0.0515065	-6.132062	0.0000000
as.factor(religiousness)1	0.1604903	0.0473926	3.386396	0.0007082

Table 3: Table 3. Odds Ratio and 95% Confidence Interval of the Coefficients

	est	2.5	97.5
(Intercept)	0.0067658	0.0052270	0.0087577
age	1.1044065	1.1005597	1.1082669
as.factor(satisfy)1	1.2524609	1.1402820	1.3756758
sexMale	0.8131344	0.7419214	0.8911828
high_educationYes	0.7365657	0.6454987	0.8404803
incomeLess than $75000	0.7291754	0.6591581	0.8066300
as.factor(religiousness)1	1.1740864	1.0699398	1.2883704

Figure 3. ROC curve



github repo link: https://github.com/zuoyuwang/STA304-ProblemSet2.git