Studying of Canadians Using Regression Model

Haona Yang 1004949531, Minchen Cai 1000523800, Mingkai Zhang 1004903063

2020-10-19

Abstract

- The 2017 GSS is conducted as a cross-sectional telephone survey across the ten provinces of Canada. This survey is conducted from Feb. 2nd to Nov. 30th 2017, targeting all non-institutionalized people at the age of 15 or older, residents of the ten provinces of Canada. The main objectives of this survey measure the living condition of Canadians based on social trends and deriving the social policy issues behind that. The data was collected through a telephone interview, for those that refused to complete the survey, they were contacted up to two times to make sure they understand the importance of the survey. The participant will inconvenient time were rescheduled at a later time. The overall response rate of this survey was 52.4%. This data set consists of two components; one is core content which measures the relationship between social trend, living condition and social policies. The other component is classification variable such as age, sex, education, etc., this helps the analysis to be more specific and meaningful when interpreting the results. Based on the dataset, we are able to build models to study the satisfaction levels among different age groups and significant social trends, seeing as an influential factor on Canadians' living conditions.

Introduction

- The goal of our report is to find a relationship based on the factors of age, age at first birth, feelings life and sex to illustrate the impact how they affect Canadians born their first kids and feelings of life according to living conditions. It is aimed to get a better understanding of the diversity of families in the future. We have our dataset from the 2017 GSS, which contains 81 variables at all. At the beginning of our work, we do a data cleaning to choose 4 variables: age, age at first birth, feelings, life and sex. This step is easy for us to see which variables are correlated with the age at first birth and feelings of life conditions. We build both a linear regression model and logistics regression to predict age at first birth and happiness to see the significant impacts involved in this dataset. This analysis will perform a better understanding of how families in Canada at different stages look like and the general satisfaction of social life from the General Social Survey.

Data

- The 2017 GSS dataset was downloaded from CHASS website and utilized professor's code to clean it. And,the dataset contains 81 variables, we will focus on four of them that are age, age at first birth, feelings life and sex because we want to use these variables to known when Canadians born first kids and their feelings of life. The target population for 2017 GSS is all persons 15 years of age and older in Canada.The survey frame was created using lists of telephone numbers and register addresses. The sampling method is Stratified Random Sampling, and this method provides better coverage of the population and enables control of the subgroups.The dataset is good because it has larger sample size and contains lots of attributes. However, the drawbacks are some people may not provide full information and it causes missing data. Part of the information that people fill in could be inaccurate since they can not remember clearly. Also, some residents might not respond to the questionnaire.

Model

For further analysis, we use R to build a linear regression model to predict age at first birth using factors of age, feelings, life and sex. A linear regression model is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. The reason why we choose this model is that we have generated two histograms above with some related distributions involved. Moreover, we would like to get a better understanding of the relationship between each other to find out whether the analysis is significant or not. In our linear regression model, variables of age, feelings, life and age at first birth are numeric. For variable sex, it is categorical.

	age at first birth		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	29.22	28.64 – 29.80	**<0.001**
age	-0.07	-0.07 – -0.06	**<0.001**
feelings_life	0.03	-0.02 – 0.09	0.269
sex [Male]	2.82	2.64 – 3.00	**<0.001**
Observations		12566	
R^2 / R^2 adjusted		0.100 / 0.100	

We build a linear regression model with the following formula by using data we selected before:

$$Age.at.first.birth = \beta_0 + \beta_1 \times age + \beta_2 \times feelings.life + \beta_3 \times sex + \epsilon_i$$

When age increases one unit, the age at first birth will decrease by 0.065944 on average, given other predictors hold constant.When feelings life raises one unit, the age at first birth will increase by 0.031444 on average, given other predictors hold constant.There is no practical interpretation here for the intercept estimate since the age can not be 0.The variable, sex, is dummy variable. It equals to 1 if sex is male and equal to 0 if it is female. The value of estimation standard deviation means the year that male has first child later than female has; the age of male who has first child is 2.820747 years later than female has.

R-squared is 0.1001 and adjusted R-squared is 0.1004. It performs that approximately 10.0% variation in age at first birth can be explained by this linear regression model.

Assume that the null hypothesis is $\beta_0 = 0$, the alternative hypothesis is $\beta_0 \neq 0$.Since the p-value of β_0 is obviously less than 0.05 level of significance, we can reject the null hypothesis.Thus, we get $\beta_0 \neq 0$. Similarly, the p-value of β_1 is also less than 0.05, so we can reject the null hypothesis and $\beta_1 \neq 0$. For β_2, its p=value is 0.269 which is larger than 0.05, since we fail to reject the null hypothesis, and $\beta_2 = 0$.The p-value for the whole model is 2.2e-16, which is less than 0.05 and significant to the whole model we predict.

Thus, there is a negative correlation between age and age at first birth, feelings life has no correlation with age at first birth since its p-value is larger than 0.05. In conclusion, the performance of a linear regression model is good with a smaller p-value of the whole model though R-squared is not large enough. However, there are still some caveats we have to pay attention to such as the variable of feelings life has few impacts on the age at first birth in Canada. It presents that happiness has no correlation with the age of having a first child, and more surveys regarding social psychology should be taken in the future to collect appropriate dataset.

	happiness		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.67	0.53 – 0.86	**0.002**
age	1.00	1.00 – 1.01	**0.001**
age_at_first_birth	1.00	0.99 – 1.00	0.422
Observations	12566		
R^2 Tjur	0.001		

In addition we model a logistic regression to predict the happiness level to be 1(happy) and 0(unhappy) based on age and age at first birth. Logistic regression is used to predict a categorical response variable of two levels usually 1 and 0, and fit an “s” shape function, by using the predictor variables and they could be numerical or categorical. Based on our summary we get β_0 = -0.394, β_1 = 0.004, β_2 = -0.003. The intercept is our β_0, age estimate and age_at_first_birth estimate are our slope for each parameter.

Based on these we get our logistic function as the following:

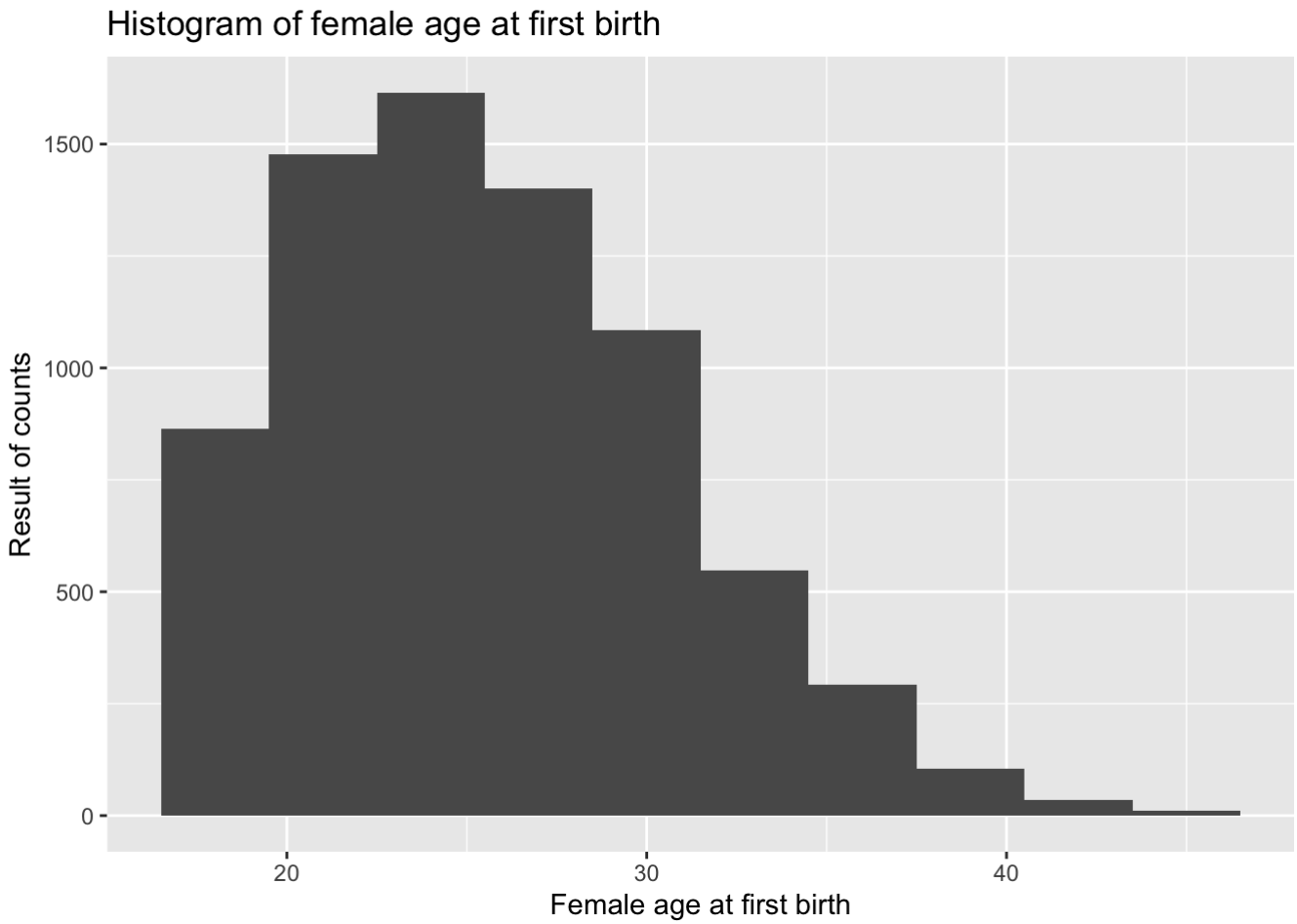
$$log(p/(1-p)) = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2$$

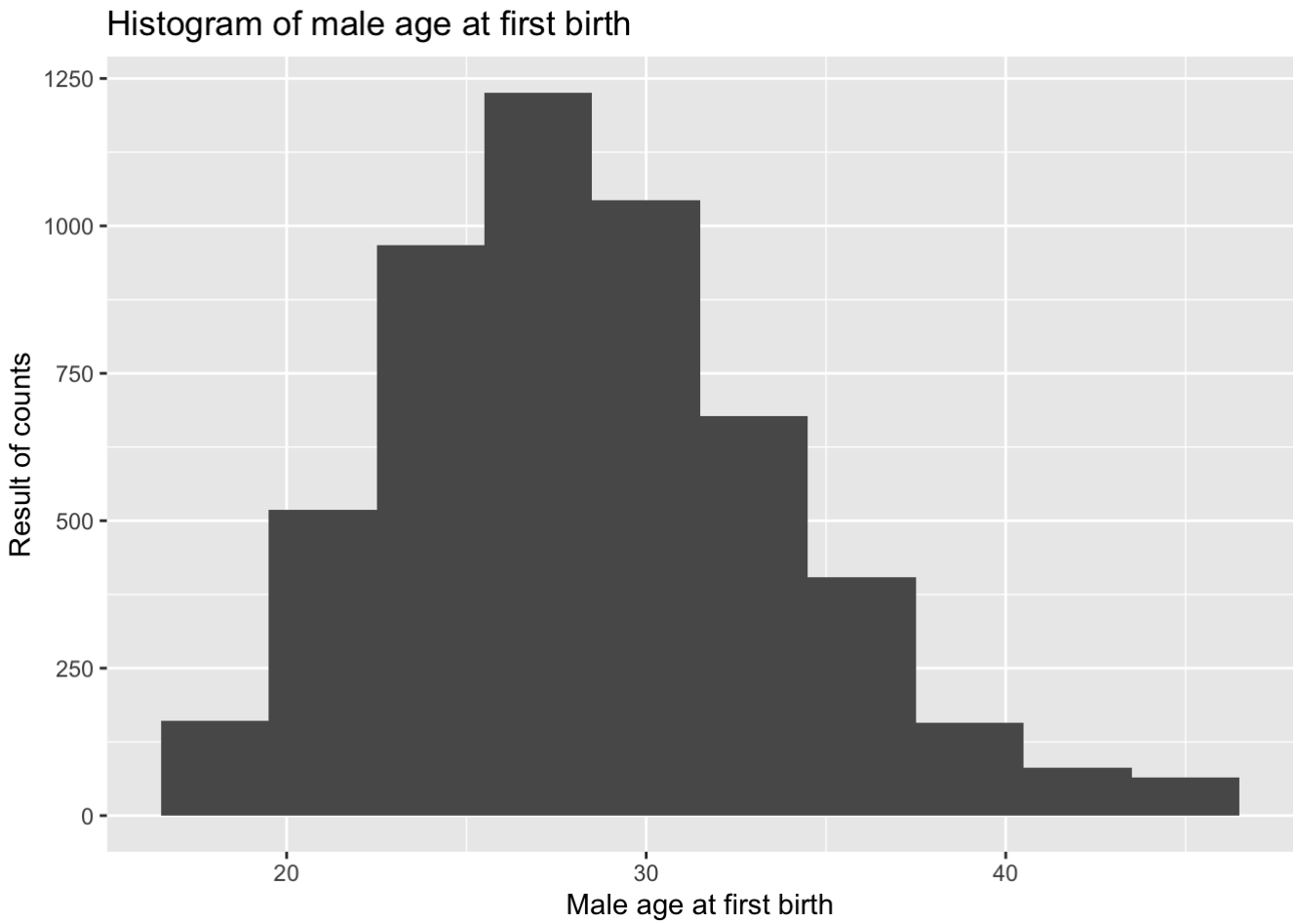
where x1 represents age and x2 represents age at first birth. This function means when age and age at first birth our interceptor or log odds of a person is happy is -0.394. As age(x1) increases by 1 the log odds of happiness is increased by 0.004, this means that there is a positive relationship between age and happiness. Then we look at age at first birth(x2) increases by 1, log odds of happiness decreases by 0.003, showing a negative relationship.

Moreover at a significant level of < 0.05, the p value for age is fairly small, therefore we have the reason to believe that the parameter age is significant in relation to the happiness level of an individual. However, the parameter age_at_first_birth has a relatively large p value of 0.422, this means that this parameter does not have a significant meaning to our model.

The overall performance of our model is fair since we have one parameter - age, that is meaningful in predicting the happiness level of an individual, and the other parameter is not very useful in predicting the same parameter. This model allows us to study there is a positive relationship between happiness and age. However, it could also include other parameters such as income level, education level, etc., to predict how happiness changes relevant to different income levels as age varies. Moreover, we should also be cautious about ‘bad’ leverage points because they could have a large effect on the estimated coefficients of our model.

Results





- From the two histograms which perform the amount of people who have first birth based on factors of sex, we could clearly witness there is a slight difference between the gender of female and male. The graph of male illustrates approximately a normal distribution though male whose age at first birth is

likely to be a bit right skewed. And we could tell from the graph that most male have their first born kids at the age with an interval 23-31. Turning to the graph of females who have their first kids, it seems to be a right skewed distribution with most data falling to the positive side. Females who aged at a range of 23-31 are most willing to have first born kids in Canada.

- Moving on we built a linear regression model to predict how age at first birth relate to age, feelings of life and sex. Based on the p value of each parameter we concluded that feelings of life have no influence on when people have their first birth, but age and sex does have a meaning to the the parameter age_at_first_birth. Therefore we could conclude that people's preference on when to have their first kid is dependent on their age and gender.
- Lastly we built a logistic regression model to indict the level of happiness with the parameter age and age at first birth. The parameter age_at_first_birth has a large p-value relevant to our significant level of 0.05, therefore it is not meaningful in our interpretation. However age with a small p value at 0.0005 can be meaningful in our interpretation, also the estimate of the parameter age is positive. Holding all other parameter constants as age grows by 1 happiness level increases by 0.004. In conclusion, the happiness levels of each individual are positively related to their age and not relevant to when they have their first kid.

Discussion

- The goal of this analysis is to generate a better understanding of social trends and the well-being of Canadian over decades. As a result, it is more helpful to monitor changes in living conditions. Additional information such as social policy issues of current or emerging interest will be presented through this data analysis by using software RStudio. All factors related to life satisfaction are involved in building two regression models in assessing the well-being of Canadians. According to our graphs and tables above, we get the number of females and males who have first birth at different ages. Females prefer to have their first kids younger than males. We also obtain this conclusion from the summary of a linear regression model. It indicates that the age of males who have their first child is approximately 2.820747 years later than females on average. From a logistic regression model, it presents that the happiness levels of each individual are positively related to their age and not relevant to when they have their first kid. From this data analysis, we are getting close to knowing how to build different models depending on different kinds of research specific interests.

Weakness

- For our linear regression model,the residuals were high which affect the accuracy of the prediction.
- For our logistic regression model, we only focus on one variable. We could include some other parameters such as income level, education level, etc., to have a stronger model in predicting our parameter of happiness.
- To determine people's happiness, we could introduce more variables rather than age and age at first birth.

Next Steps

- This report researches all Canadians in 2017, to get further conclusions, we may analyze the GSS data for future years. Thereby, we could obviously change the data in various years, and provide more reliable results. Also, we could find a survey of different countries in 2017, then contrast it with Canada GSS data. Moreover, the overall response rate for the 2017 GSS was 52.4% which caused several non-sampling errors and non-responses on the survey results. For further steps, the collection method of the survey may be improved by using new techniques like using Big Data to generate more accurate raw data variables through all information from Statistic Canada and other federal departments.

References

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)
2. General Social Survey,Cycle 31:Families,Public Use Microdata File Documentation and User's Guide
3. General Social Survey (GSS), Cycle 30, 2017: Canadians at Work and Home

Appendix

Appendix A

- Linear Regression Model

```
##
## Call:
## lm(formula = age_at_first_birth ~ age + feelings_life + sex,
##     data = df_gss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7143  -3.7091  -0.4942   3.1461  20.7395
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.221557   0.294215  99.321   <2e-16 ***
## age           -0.065944   0.003021 -21.825   <2e-16 ***
## feelings_life  0.031444   0.028435   1.106    0.269
## sexMale        2.820747   0.093144  30.284   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.146 on 12562 degrees of freedom
##   (171 observations deleted due to missingness)
## Multiple R-squared:  0.1004, Adjusted R-squared:  0.1001
## F-statistic: 467.1 on 3 and 12562 DF,  p-value: < 2.2e-16
```

- Logistic Regression Model

```
##
## Call:
## glm(formula = happiness ~ age + age_at_first_birth, family = "binomial",
##     data = df_log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.130  -1.087  -1.043   1.266   1.334
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.394445   0.125106  -3.153 0.001617 **
## age                 0.004148   0.001205   3.442 0.000578 ***
## age_at_first_birth -0.002710   0.003374  -0.803 0.421752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 17256  on 12565  degrees of freedom
## Residual deviance: 17242  on 12563  degrees of freedom
## AIC: 17248
##
## Number of Fisher Scoring iterations: 3
```

Appendix B

- Github Link: https://github.com/minchencai/STA304PS2.git