What causes the Canadian citizens' to have their first child late

Ziqi Gao 1003051092, Shidong Gui 1003592506, Cheng Qian 1004484569, Kaixi Zhang 1005059268

Oct.19th,2020

What causes the Canadian citizens' to have their first child late

Ziqi Gao, Shidong Gui, Cheng Qian, Kaixi Zhang

Oct.19th,2020

Abstract

This paper builds a multi-linear regression model investigates how age, sex, age people first get married and income of the family, influence age of having their first child, using a 2017 Canadian survey from CHASS. From this model, we would see that the older generation, the family group that has the lowest income, people who get married earlier and women tend to have their first child earlier. With these implications from the model, we can now roughly infer a Canadian's age at first birth if given these information. These results are mostly aligned with our common sense, however, more investigation can be done from sociological aspects by utilizing this model. Code and data supporting this analysis is available at: https://github.com/KaixiZhang99/STA304-PS2-Group164

Introduction

As society develops, it is believed that the younger generation is having their children later and later, or even choose to be DINK. There are definitely a lot of reasons cause the late birth of the first child in Canada, no matter the global extrinsic reasons like increasing cost of raising a child, or personal intrinsic reasons like more freedom desired by young couples. It is questionable if personal extrinsic factors will affect people's age of having their first child. Thus, our goal of analysis is to find the relationship between Canadian citizen's age at first birth and their age, gender, income level and age at first marriage. We want to see how they together predict the age of having their first child, and how important and predictive each individual factor is.

In this report, we first obtain this data set from the General Social Survey 2017, which is a survey focused on family. This data set contains a wide variety of variables about a family like their background, work-related information etc. But here we narrow down our focus on age, gender, income level and age at first marriage as we believe these are the variables that are more closely linked to the age at first birth from common sense. A multi-linear regression model will then be built using these variables as independent variables. We would analyze the model's validity to see how reliable the model is, and use the model to predict someone's age at first birth given these contributors. Multiple graphics would be used as well in order to provide a more straight forward view on how the variables impact the response variable age.

From this model, we would get the result that people having the following characteristics are more likely to have their first child earlier: older generation, people who get married earlier, female, and family with lower income. These are quite the same as our common knowledge, however, it does to be meaningful because with this model, we now have more statistically supportive knowledge instead of pure common sense. Sociologist

might be able to use this model to do more research from sociological aspect, for example, investigate why under the pressure of increasing cost of raising up a child, the lower income group actually has their first child earlier.

Data

'age_at_first_birth'	'age'	'sex'	'income_family'	'age_at_first_marriage'	'gender'	'income_level'
Min. :18.00	Min. :25.70	Length:4049	Length:4049	Min.:15.00	Min.:0.0000	Min.:0.00
1st Qu.:21.50	1st Qu.:58.20	Class :character	Class :character	1st Qu.:20.30	1st Qu.:0.0000	1st Qu.:1.000
Median :24.60	Median :67.60	Mode :character	Mode :character	Median :22.30	Median :0.0000	Median :1.000
Mean :25.41	Mean :65.97			Mean :23.25	Mean :0.3258	Mean :1.831
3rd Qu.:28.30	3rd Qu.:75.90			3rd Qu.:25.40	3rd Qu.:1.0000	3rd Qu.:3.000
ax. :45.00	Max. :80.00			Max. :50.00	Max. :1.0000	Max. :5.000

The dataset we chose is the 2017 GSS obtained from the CHASS website published by Statistics Canada. This dataset is a survey sample with a cross-section method. The target population for the 2017 GSS has counts of 30,154,347; which included all persons (excluding residents of Yukon, Northwest Territories, and Nunavut; and full-time residents of institutions) 15 years of age and older in Canada. This number is computed from the 2017's census provided by Statistics Canada. Full-time residents of institutions are not given, so the number is estimated using 2016's census.

The frame population is all people approachable through telephone, including landline and cellular, from all available sources that can provide telephone numbers. Another frame list is the address, but it is largely overlapped with telephone numbers. The target sample size is 20,000 and the actual sample size 20,602. The response rate is 52%. Respondents are mostly found by telephone interviews, and non-responded people would be re-contacted or schedule a time to call back. The interviewer would also explain the importance of the interviewee and encourage citizens to participate.

Among the total of 81 variables, we selected "age_of_first_birth" as our y variable and "age", "sex", "income_family", "age_at_first_marriage" as our x variables, intend to investigate the relations between a person's age, sex, age of first marriage and income level with their age when they have their first child born. It is worth mentioning that we purposely selected the "income_family" variable rather than other similar variables such as "income_respondant". We believe that accounting for the effect of the income of a family would make more sense than just individuals since having the first child is more likely a family decision instead of an individual one. These selected variables are not combined with other variables.

In the final data frame, there are two more variables that didn't get mentioned previously. Since both "sex" and "income_family" are categorical variables, we require indicators to represent these variables in order to build a solid multiple linear regression model. Thus, we added columns of "gender" and "income_level" to our data frame. For the "gender" column, we record males as "1" and females as "0" from the original "sex" column. There are six categories in the "income_family" column, which are "Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "100,000 to $124,999" and "125,000 and more". We record these categories as "0" to "6" into the new "income_level" column accordingly. ("0" being "Less than $25,000" and "5" being "$125,000 and more")

After filtering out all null observations, we were given 4049 valid observations for our analysis. One of the good things about this dataset is that it is large enough to be trustworthy. As we can see, after the filtration, we still have over 4,000 observations for us to investigate. Also, the source of data is creditworthy since it is officially conducted by a Canadian agency. However, to construct such a large dataset, it requires a huge

amount of time and resources.
Also, as this is a voluntary sample, the sample size is actually a lot smaller than the target population, which might cause biases.

Model

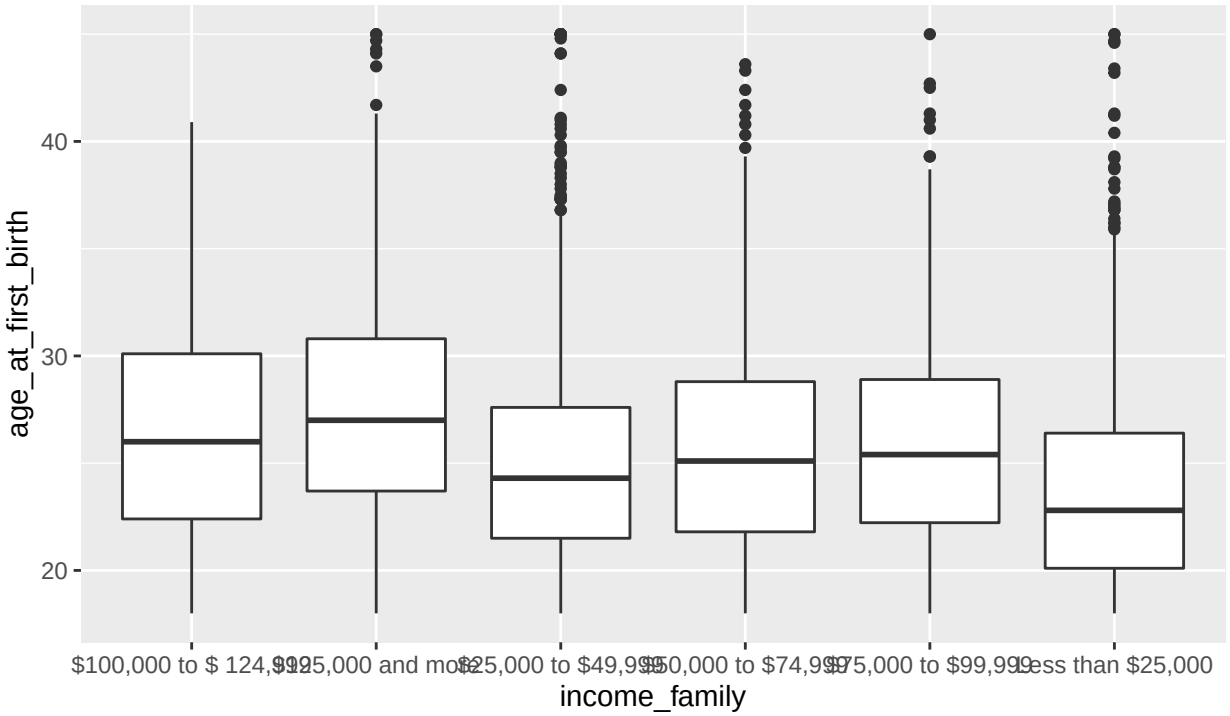
```
##
## Call:
## svyglm(formula = age_at_first_birth ~ age + gender + income_level +
##     age_at_first_marriage, design = data.design, data = selected_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.5243  -2.0757  -0.5081   1.6766  23.3153
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              8.665938   0.565885  15.314  < 2e-16 ***
## age                     -0.003764   0.005828  -0.646    0.518
## gender                   1.106547   0.139504   7.932 2.77e-15 ***
## income_level             0.378839   0.041892   9.043  < 2e-16 ***
## age_at_first_marriage    0.685612   0.014414  47.566  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.967 on 4044 degrees of freedom
## Multiple R-squared:  0.4282, Adjusted R-squared:  0.4277
## F-statistic: 757.2 on 4 and 4044 DF,  p-value: < 2.2e-16
```

We started out our analysis by constructing a multiple linear regression model. The reason we chose this model is that we have several variables and we expected a linear relationship between the dependent variable(Y) and independent variables(X). We follow the formula of (). We have a total of 4 independent variables, "age","gender","income level" and "age_at_first_marriage", and were labeled X_1, X_2, X_3 and X_4 respectively.
Among these explanatory variables, gender and income level are categorical, because for gender, we only have two categories of male and female, and for income level, the family falls to one of the sections depending on their income. On the other hand, age and age at first marriage are numerical. The type of variables is crucial to how we choose our graph type in the later section.
Before running the model, we construct a formula reference from the variables we have selected and the equation of multiple linear regression: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$ The beta values are interpreted as regression weights , which measure the association between the predictor variable and the outcome.
Before we decided to construct a multiple linear regression model, we have thought of using several single linear regression models to examine how the change of every single variable influences the dependent variable. However, the two methods give the exact results and the multiple linear regression has higher efficiency than the simple linear regression.

Results

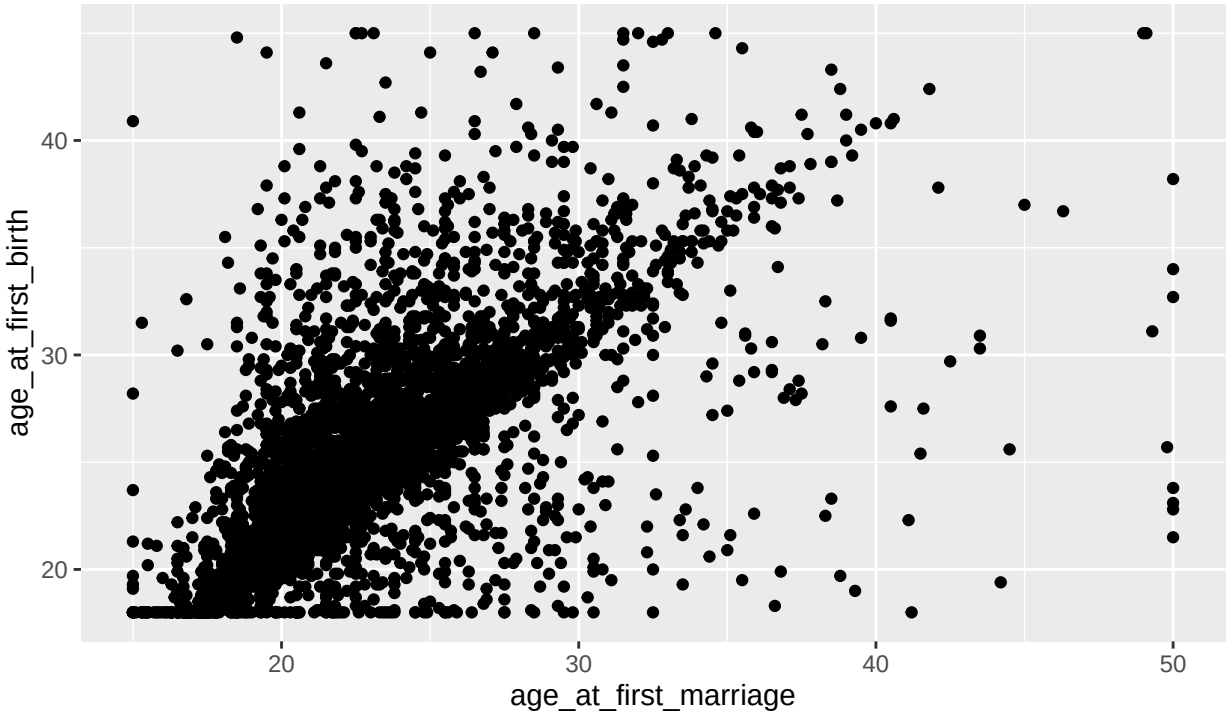
Boxplots of
Incomes of Families vs. Age of Individuals at Their First Child's Birth

Figure 1



Scatterplot of
of Individuals at Their First Marriage vs. Age of Individuals at Their First Chi

Figure 2



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

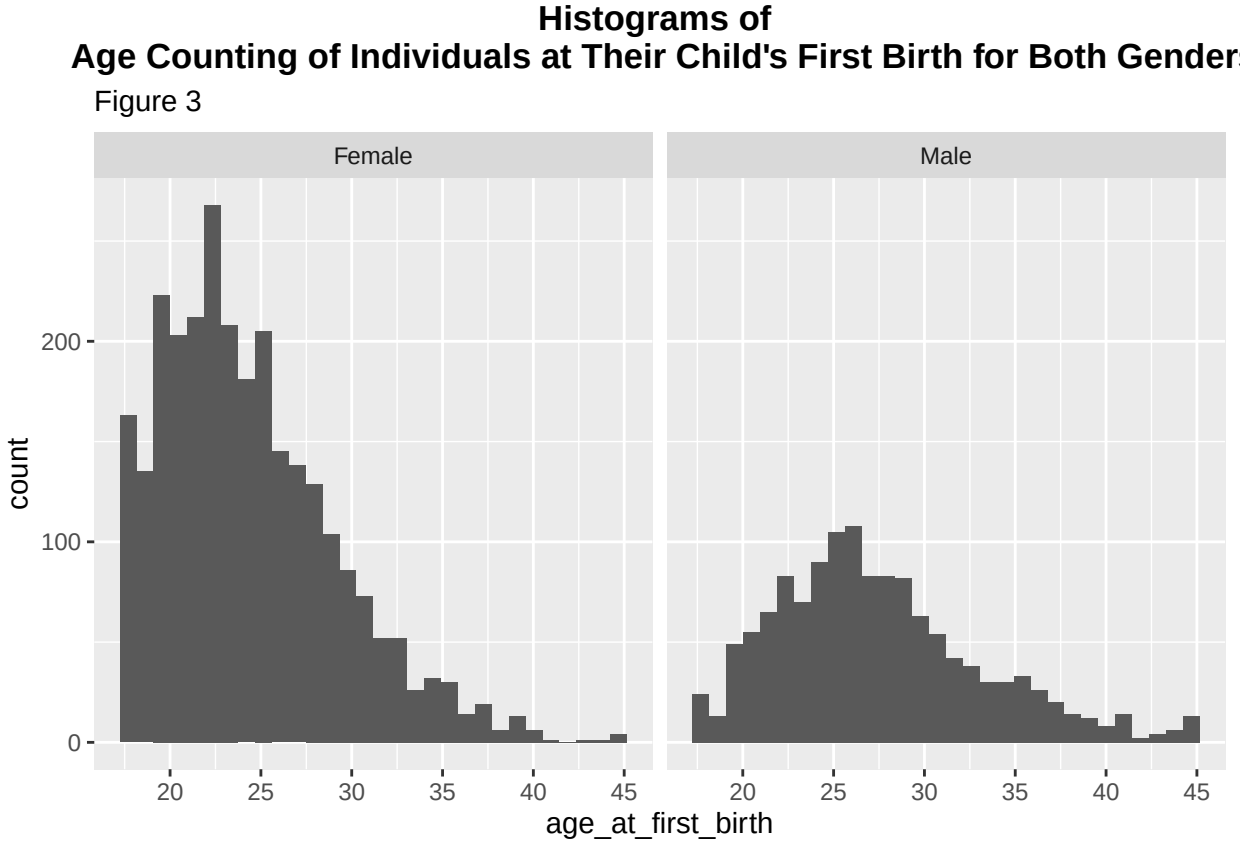


Figure 3

We use R to run the model, which gives out the following equation: $\hat{Y}_i = 8.666 - 0.00376X_1 + 1.107X_2 + 0.3789X_3 + 0.6856X_4$

From the equation, we see that $\hat{\beta}_0$ equals 8.666, this is a constant term and the y-intercept. In other words, it is the value of the independent variable Y if the values of Xs are all equal to zero. It has no meaningful value in our case since the X values have no tendency approaching zero. $\hat{\beta}_1$ has a value of - 0.00376, which means one unit increase in age leads to a 0.00376 decrease in age of first birth holding everything else constant. $\hat{\beta}_4$ gives us the value of 0.6856, which indicates that one unit increase in age of first marriage leads to a 0.6856 increase in the age of first birth. Since income level and gender are categorical variables, we would get more detailed information from graphs.

Figure 1 is the boxplot of income_family vs. Age_of_first_birth. We can see the median level of each category of income level. The family who has income less than 25,000 a year has the lowest median of age_of_first_birth among other categories. The highest median of age_of _first_birth falls into the category with the highest income level. There are multiple outliers in 4 of the 5 categories, however, due to the large size of the dataset, it should not have much influence on our final result.

Figure 2 is the scatterplot of age_at_first_marriage vs. age_at_first_graph. Since these two are numeric variables, we use the scatterplot to see their correlations. From the graph, we can tell that these two variables have a strong, positive, linear correlation with noticeable outliers. There is also a cluster in the bottom left corner, which indicates most individuals had their first marriage and first child born during their 20s to 30s.

Figure 3 is the histograms of age counting of Individuals at their first child born for both female and male. We can see from the graph that females have a lower median compared to males, which indicates that females, on average, would have their first child earlier than a male would have.

The model result shows the p value is statistically significant ($P<0.05$). It basically tells us the data are unlikely to be due to random chance, in other words, the model has strong evidence of being valid. The R-square has the value of 0.4282, it is a measure of strength of relationship between our model and dependent variables. And this value is an acceptable value for our model.

Discussion

Our dataset comes from the General Social Survey (GSS) in 2017. It discusses the trend of living conditions and well-being of Canadians during this period. The sampling approach of this dataset is stratified sampling, which is to select a simple random sample from each strata (candidates' province) without replacement. In addition, the data is collected by random select telephone numbers from households and complete the telephone interviews. According to the User_GuideBook, it illustrates this survey has 52.4% in overall response rate. Although this sample has many constraints to minimize the error, it still has some sampling error and non-sampling errors in its result, especially the effect from non-response from partial questions to total questions due to not understanding or misinterpreted questions. These errors would lead to some biases in the sampling result.

According to the limitations of linear regression model and non-sampling errors from the origin dataset, our group selected the "age_of_first_birth" as the dependent variable, and "age", "gender", "income_level" and "age_at_first_marriage" as the independent variables to provide the correlation between them. In our model, "gender', "income_level" and "age_ at_first_marriage" have a strong positive correlation with the response variable "age_of_first_birth" due to the extremely small p_value. Smalle p_values means that under the null hypothesis, the probability of independent variables not linearly correlated with the dependent variable is very small. Thus, it is supportive that the 4 independent variables are strongly correlated with the response variable. Moreover, the R square in our model is 0.4282, it represents there are 42.82 percentage of variation in dependent variable ("age_of_first_birth") can be explained by independent variables ("sex", "gender", "income_level" and "age_at_first_marraige") in our linear regression model.

According to Figure 1 in our project, it shows the distribution of age of first birth in different income level groups, this boxplot graph has a clear result that the group who has a higher income level would have dropped in an older range when they have a first child. The reason for this finding may relate to high income level groups paying more attention in their work rather than their life, compared with the lower income level group. This distribution may exacerbate the gap between the rich and the poor, and intensify the inequity of education quality between groups.

On the other hand, our graph has discussed the relationship between age at first marriage and age at first birth, and plot the points into a scatter plot (Figure 2). The graph clarifies there is a positive relationship between the two numerical variables, which means people who are younger in first marriage will have the first baby early that fits the common sense. This condition is due to the quality of people's bodies that is easier to make pregnant in a younger stage. Nevertheless, there are still some points outside the linear, they can be separated by two groups which is get a child before marriage and get a child after a long marriage period.This trend may push people to have a child at an early age that would improve the decline of birth rate in Canada.

In this model, it can provide reference to the male group, for example, the average age of male is older than female when they have their first birth. Hence, by this model, the male can predict their age of first birth depending on their income level and age when first marriage. However, from the "Histogram of Age Counting of Individuals at Their First Child's First Birth for Both Gender", it illustrates that there is a significant difference of number between male and female, it may be caused by many males not knowing they have a child before getting a divorce. This result also represents the child's birth in a single parent family usually raised by their mother. This trend would have an obvious negative effect on children and teenagers' well-being during this period.

Weaknesses

There are several weaknesses of this survey and analysis. First, the variables examined by the survey are mostly categorical variables. Using categorical variables in a linear regression model works, but might limit the result compared to using continuous variables.

Secondly, the original survey is conducted using stratified sampling technique as stated in the user guide. A fpc correction should be made but due to inability of coding, this analysis uses simple random sampling. This may give misleading results.

Lastly, a bayesian linear regression model might give better prediction than simply a multi linear regression. In bayesian models, we have some prior knowledge so that the prediction can be improved.

Next Steps

To improve the analysis, the first thing we could do is to do more research on coding and use fpc correction for stratified sampling. Also, we can build a bayesian linear regression model. This can not only give another set of predictions, but also can be compared with our current model so that we know how bayesian actually improves prediction.
Lastly, if we want to investigate more on the contributors to the age of having the first child, another survey needs to be conducted and collect other variables. For example, the new survey could try to collect information on the family's health conditions, which is another aspect that may cause early or late first birth. They should also try to collect more continuous variables, for example, respondents' BMI. In this way, the data set can be more diversified and incorporates more variables.

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

2. Government of Canada, Statistics Canada. Population Estimates on July 1st, by Age and Sex, Government of Canada, Statistics Canada, 29 Sept. 2020, www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501.

3. Government of Canada, Statistics Canada. "Census Profile, 2016 Census Canada [Country] and Canada [Country]." Census Profile, 2016 Census - Canada [Country] and Canada [Country], 18 June 2019, www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E. http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

4. Technology, Advancing Knowledge through. Computing in the Humanities and Social Sciences, www.chass.utoronto.ca/index.html. Alexander,Rohan, and Sam Caetano. "gss_cleaning.R", 7 Oct,2020 Christopher JohnChristopher John 3, et al. "Center Plot Title in ggplot2." Stack Overflow, Nov. 2016, stackoverflow.com/questions/40675778/center-plot-title-in-ggplot2.

5. DQdlMDQdlM 7, et al. "Replace a Value in a Data Frame Based on a Conditional (`If`) Statement." Stack Overflow, May 2011, stackoverflow.com/questions/5824173/replace-a-value-in-a-data-frame-based-on-a-conditional-if-statement.

6. Grolemund, Garrett. "Hands-On Programming with R." 7 Modifying Values, rstudio-education.github.io/hopr/modify.html.

7. rdatasculptorrdatasculptor 6, et al. "Dplyr Mutate with Conditional Values." Stack Overflow, Mar. 2014,stackoverflow.com/questions/22337394/dplyr-mutate-with-conditional-values.