Discovering The Relationship Between When Canadians Get Married And When They Have Their First Child

Kexin Qin

Oct.19.2020

Code and data supporting this analysis is available at: https://github.com/qinkexinnn/STA304

Abstract

My approach uses linear regression for finite population to model the age when people get married, and when they have their first child. A positive association was found between the two variables. This is important because we get to see whether couples choose to wait longer, or have kids right away; giving us some information about how the future demographics might be like.

Introduction

Getting married and becoming a parent seems like a milestone more many people in life. In this report, I will try to figure out while these two milestones are correlated. A simple linear regression model will be used to see if there are any association between the age of marriage of a person, and the age when they have their first baby. I will start by discussing how the data was collected, justify why I used the simple linear regression model, and end with my findings and conclusions.

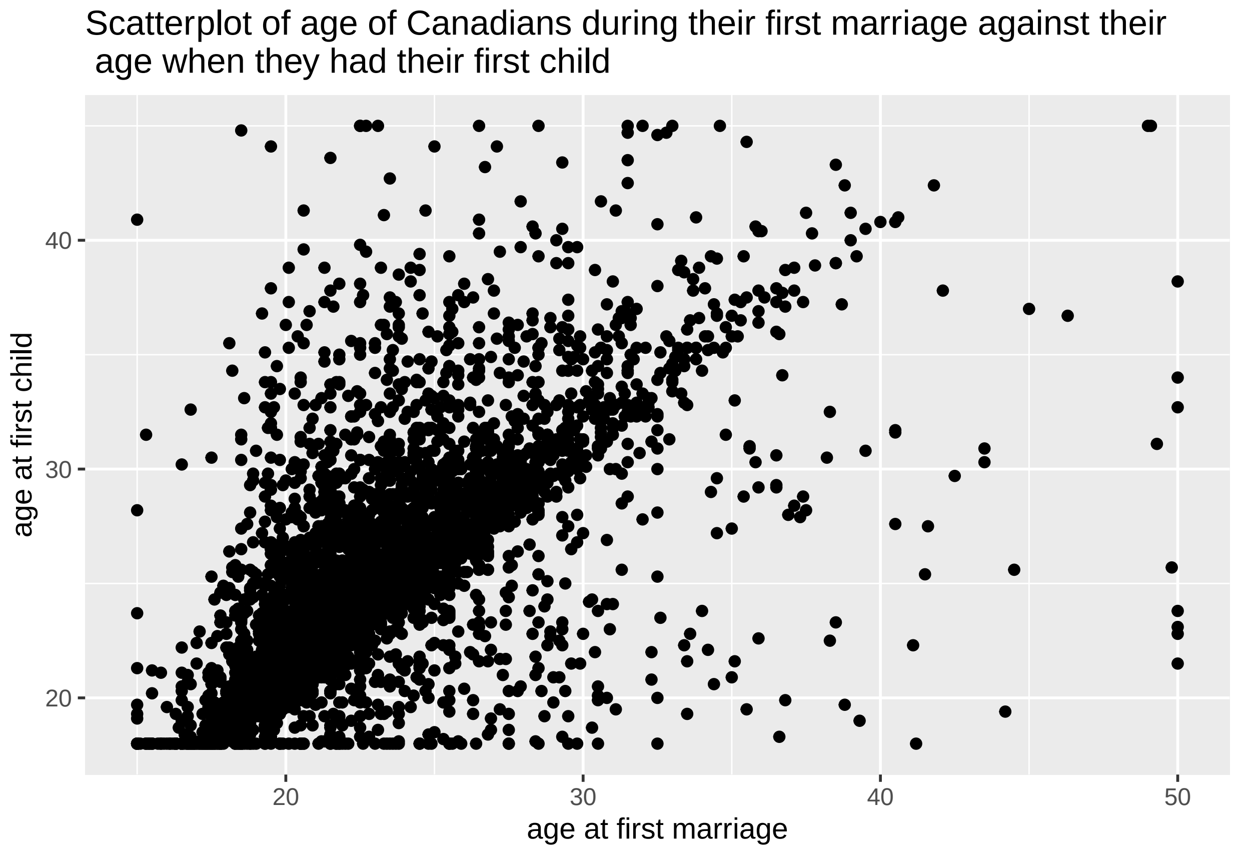
Data

The data that I have chosen comes from the General Social Survey of 2017. The purpose of the survey is to gather data that will help us better understand Canadian families. The data was collected through stratified sampling, where each of the ten provinces were divided into different stratas. Then, a simple random sample without replacement was performed within each of the stratum.

The target population of the survey is persons 15 years of age and older that live in the 10 provinces of Canada. The frame was created using a list of telephone numbers from Statistics Canada, and the Address Register. The target sample size for this data was 20,000. The respondants were contacted via telephone interviews, with no mention of what happened to the non-response.

This data has its own strengths and weaknesses. Since the target population of this survey is people 15 years of age and older that live in the 10 provinces of Canada, it has very good potential for generalizability. We would be able to generalize our findings to a large population. The strengths also include having a large number of topics, and that it comes from a very reliable source. However, the data also has limitations. For example, its response rate. The response rate of this survey was 52.4%. This can lead to potential sampling bias. There might also be communication failings. People can interpret the question in different ways that might affect their answers.

The data consists of many variables, but I will mainly be focused on how old the respondant is when they first got married, and the age when they had their first child. Below is a scatterplot showing us how the raw data looks like. We will be exploring and discussing this data in the next sections.



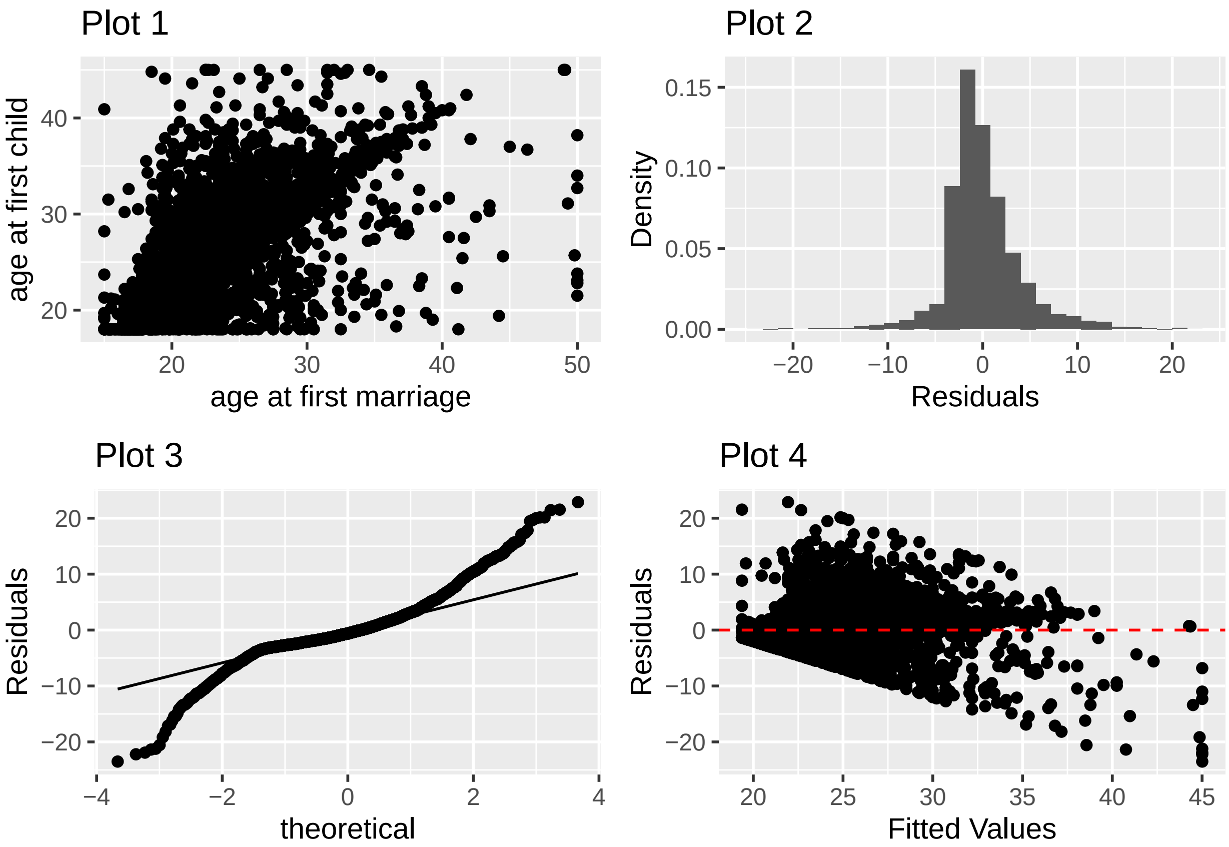
Model

From the scatterplot in the data section, I observed weak positive trend between the two variables. I want to determine whether one's age at first marriage and one's age when they have their first child are associated and the strength of the association if it exists. Specifically, I would like to find out if we can use one's age when they get married to predict their age when they have their first child. Therefore, I would like to use R to explore this further with the following simple linear regression model:

$Y = \beta_0 + \beta_1 X + e_i$

X is our predictor variable, which is one's age when they get married; Y is the response variable, which is one's age when they have their first child. β_0 represents the average value of Y when X is equal to zero, which has no interpretation in this case because being 0 years old is outside the range of the data. β_1 is our slope parameter, which tells us the corresponding average change in Y when X changes by 1 unit; that is, the average change in one's age when they have their first child when their age at marriage changes by 1 year. Lastly, e_i is the random noise, the variation in measures that we cannot account for.

A simple linear regression model require that we verify some assumptions. I will be using the four following plots to help me verify model assuptions for the model.



Plot 1 is the plot of our raw data, which shows a weak positive trend. From Plot 2 we can see that the histogram residuals is approximately normally distributed, we can also kind of see this Plot 3, except we see that the tails deviate in the QQ plot, a sign that that the residuals might not be exactly normal. Using Plot 4, we can also check that the expected value of the residuals is approximatly zero. Based on the plots of residuals, it seems like the assumptions more or less hold, and the SLR would be appropriate for this data. We will look at the results in the next section.

Since the data was collected through stratified sampling, I will account for this method and use finite population correction. The population I will use is $N = 10020035$, the number of people with children according to the 2017 census in Canada.

Results

Table 1: Summary of SLR Results

Coefficient	*Estimate*	*Error*	*t-value*	*p-value*
Intercept	8.38	0.33	25.15	2e-16
x	0.73	0.01	52.06	2e-16

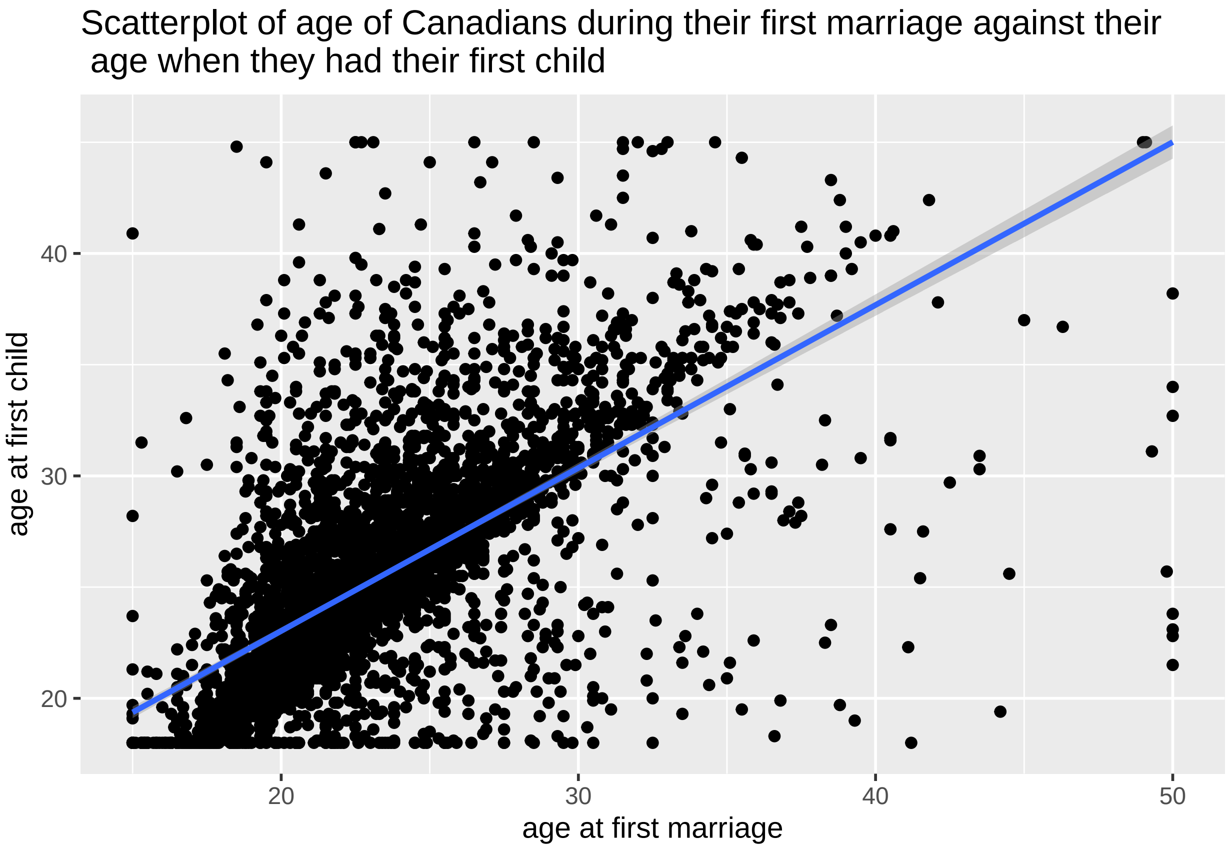
Table 2: Summary of SLR Results, accounted for survey method and population

Coefficient	*Estimate*	*Error*	*t-value*	*p-value*
Intercept	8.38	0.69	11.83	2e-16
x	0.73	0.03	23.84	2e-16

Based on the fitted model, it looks like the age when one gets married, and the age when they have their first child is associated. The small p value suggests that we have evidence against the hypothesis that there is no association between the predictor and the response variable. The estimated model shows how age when one has their first child (Y) as a function of their age when they are married (X):

$Y = 8.38 + 0.73X$

We can visualize this line on our scatterplot:



Based on the data, it appears that for 1 year increase in age when they are married, their age when they have their first child increases by 0.73 years. The intercept parameter has no interpretation since it is outside the range of our data. In table 2, we can see that we still get the same results when we account for the survey method and the population.

Discussion

From the results section, we can see that parents who chose to have kids tend to not wait too long to have one. This is interesting because we can see that parents across the ages make this similar choice.

There are weaknesses to the data as well. These numbers just show when the respondant had their first child, without indicating whether or not it was planned. This means that we may have misleading results, since some of the births might not be planned in the first place.

Although we see a weak correlation between age at first marriage and age at first child for people who got married before 30, that might not be the case for people who are older than 30. Since the QQ plot also showed that the residuals do not follow the normal distribution at the tails, maybe it would be better to fit the model for people who got married between 20-30 years old rather than everyone who took the survey.

References

R. Alexander (2020) “gss_cleaning.R”, tellingstorieswithdata.com

B. Auguie (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

“Canadian general social surveys (GSS)”, By Statistics Canada under the terms of the Data Liberation Initiative (DLI), https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

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.

Guide book of 2017 General Social Survey (GSS): Families Cycle 31 https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

S. Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

T. Lumley (2020) “survey: analysis of complex survey samples”. R package version 4.0.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Y. Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963