Analysis for women's age of first marriage and age at first date

Ziyi Shen

10.19.2020

1 Abstract

The present study aims to help sociologist find an association between women's age at first marriage and women's age at first relationship. We employ simple linear regression model to justify our hypothesis. The result shows that our hypothesis is valid. Therefore, we might conclude there is a positive linear relationship between woman's age at first marriage and woman's age at first relationship.

2 Introduction

The purpose of this data analysis report is to help sociologist study the relationship between women's age of their first relationship and the age of their first married. Typically, people will think if women start their first relationship early, they will get married early. However, nowadays, an increasing number of generations start to be married very late compared to the last century. Therefore, the current study will use simple linear regression model to see whether the association exists. First, we want to build this model based on the general idea.

3 Data

3.1 dataset

I use dataset from the CHASS data centre of University of Toronto. This data has initially been from general social survey. General social survey is a five-year cycle survey and designed by Statistics Canada. The dataset is from 2017. The primary purpose of the survey is to monitor changes in Canadian family.This survey helps policymakers change policy, such as parental benefits, child care policy and income inequality. In this report, programming language R is used to complete the data cleaning and wrangling.

3.2 Survey population

The survey's target population is all non-institutionalized persons at 15 years of age or older, living in the ten provinces of Canada. Excluding the residents of the Yukon, Northwest Territories, and Nunavut and full-time residents of institutions, the sampling population is participants who voluntarily filled the question. "Note that GSS only selects one eligible person per household to be interviewed."

3.3 Questionnaire

"The questionnaire was designed based on the research and extensive consultations with key partners and data users. Qualitative testing was conducted by Statistics Canada's Questionnaire Design Resource Center (QDRC)". The questionnaire enjoys an edge in easy use, low cost, and acquisition of a lot of information in a short time. However, the accuracy rate may depend on the response rate. The overall response rate is 52.4%. The level is not very high. But, as 13,000 samples have been gathered in the dataset, the response rate is acceptable. The most common disadvantage is response bias and non-response bias (more details in the weakness section). The last weakness of the questionnaire is that the question cannot be accessed anymore.

On the website, it states, “For the 2017 GSS, significant efforts were made to minimize bias by using a well-tested questionnaire, a proven methodology, specialized interviewers and strict quality control.” However, as we can not access a copy of the questionnaire, we do not know the potential bias in the questionnaire. For example, the question is very misleading. Then participants might give a different answer in this part.

3.4 Sampling method

This survey’s primary sampling method is stratified sampling, which is sampling from a population which can be partitioned into sub populations.To complete the survey, Statistics Canada divides ten provinces of the target population into strata. Many of the CMAs (Census metropolitan areas) are considered separate strata, such as Toronto, Hamilton, Montreal and Quebec City. Next, the non-CMA areas of each of the ten provinces are grouped to form 10 more strata. After the completion of sampling, there are 27 strata in the survey. Additionally, Statistics Canada also uses the bootstrap method to estimate the sampling variability for 2017 GSS survey.

3.5 Variables

The independent variable is age of first date. The unit of measurement is age (years). The dependent variable is age of first marriage. The unit of measurement of age is (years). Since age is a numerical and continuous variable, we can directly input our model’s data(GSS.cvs) without modifying data type.

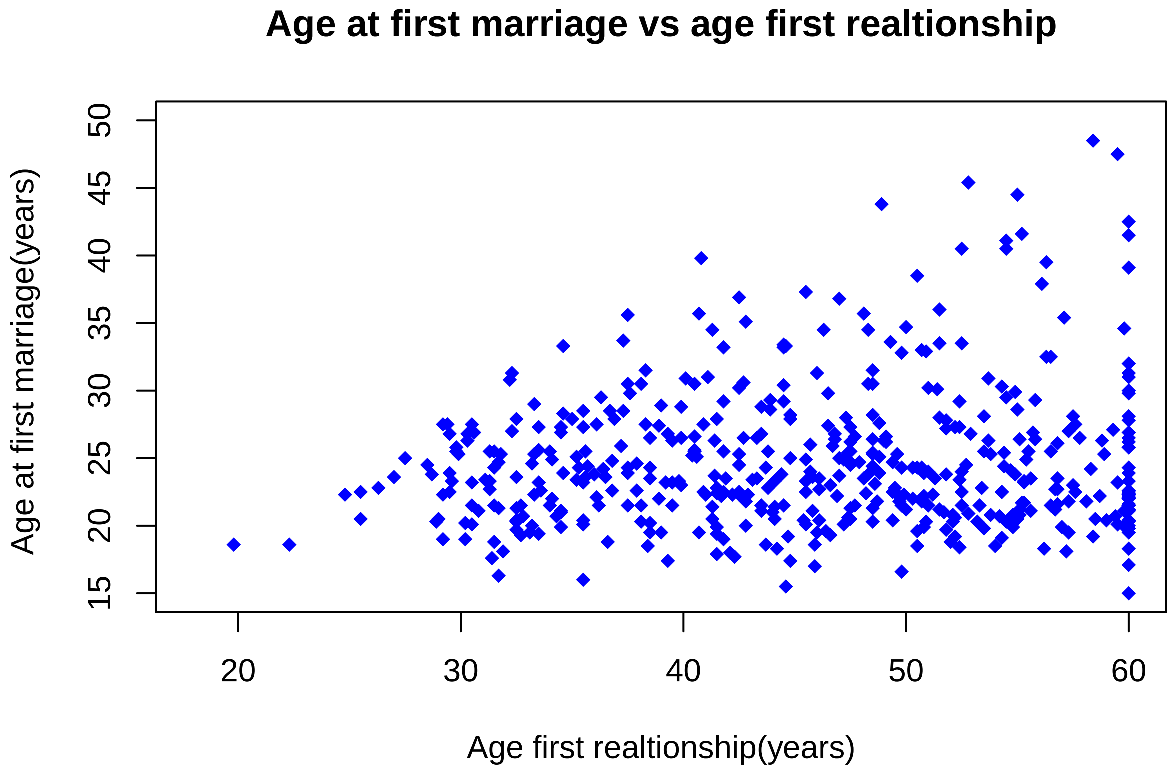


Figure1

4 Model

To examine our initial hypothesis, we should first run a hypothesis test. The null hypothesis is that there exists no correlation between the age of first date and the first marriage age. The alternative hypothesis is that there is a correlation between age of the first date and first marriage age. Second, we use the simple linear regression model to examine the hypothesis. The reason is that this model predicts the value of Y for a new value of X, which is our goal.Also,two variable are numerical and continuous. So the linear regression model easily be fitted. The theoretical model of linear regression is

$$Yi = E(Y|X = x) + ei = \beta_0 + \beta_1 x + ei.$$

beta0represents the intercept of the model. beta1 represents the slope of the model, and e defines the error term. The practical model in our report is

$$\widehat{Age\ first\ marrige} = \beta_0 + \beta_1 * Age\ first\ marrige$$

In our study, we will use lm() function in R to run our data.In lm function, y represent the dependent variable, and x stands for independent variable. The sample, will impact the result accurately (more will discuss in limitation section). However, it does not affect the running lm() function. Since lm() and gglplot() automatically remove the NA samples from data frame. Finally, we need to check four assumptions of simple linear regression model. These assumptions are necessary for us to make inferences about the unknown model parameters. The alternative model is use svydesign() function to correct our original model. However, we do not know the exact population size. When we summarize the function,there are a lot information are missing. Additionally, the alternative model only minimize standard error, t- value and p-value.These three value in our original model is very small. In conclusion, we want to use the lm() function to run our model.

4.1 Diagnosed test

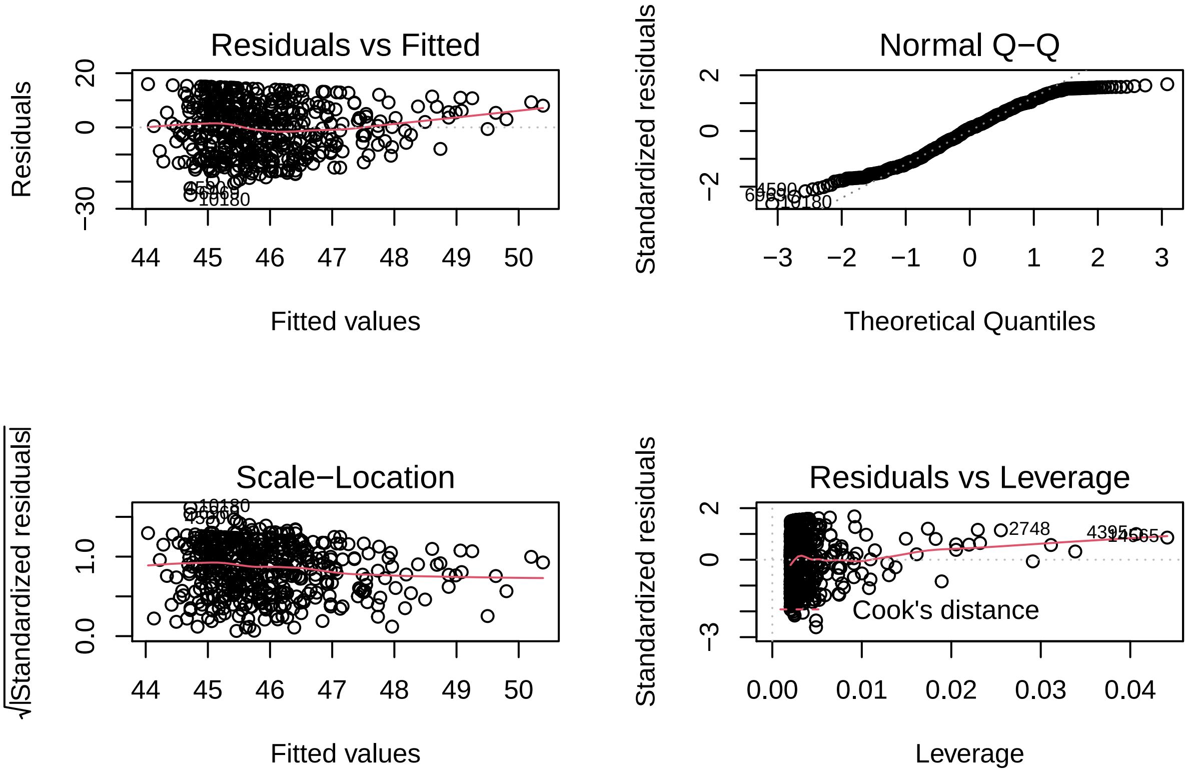


Figure 2

Residuals v.s. Fitted

Even though some points seem to be denser as the fitted value gets larger, we can tell from the graph that the residuals around the horizontal line spread without distinct patterns, and thus we can conclude that the linearity assumption holds.

Scale Location

We can tell that there is a straight line with randomly spread points on the graph. Hence, there exists no indication of heteroscedasticity.

Normal Q-Q

The Normal Q-Q plot points lift off the line at the ends wiggle around the line, but they are not crazy. It can be concluded that it follows the assumption of normality.

Residuals vs Leverage

We can tell from the graph that there are only a few observations outside Cook's distance.

5 Results

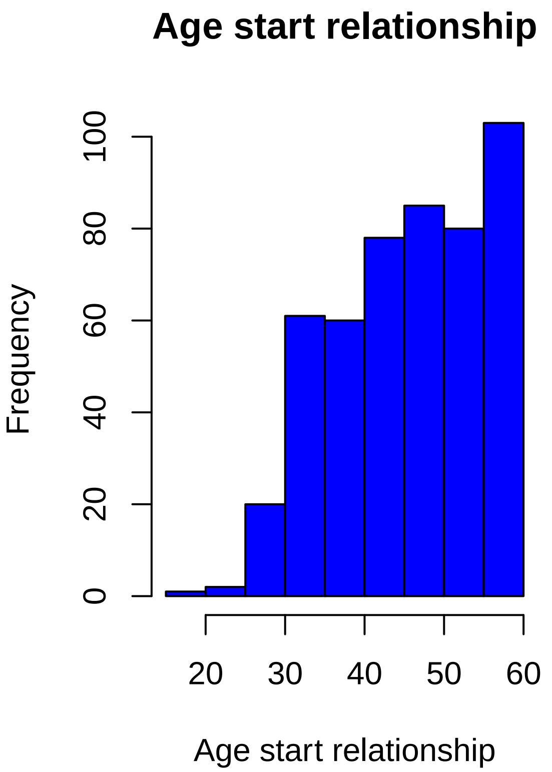


Figure2

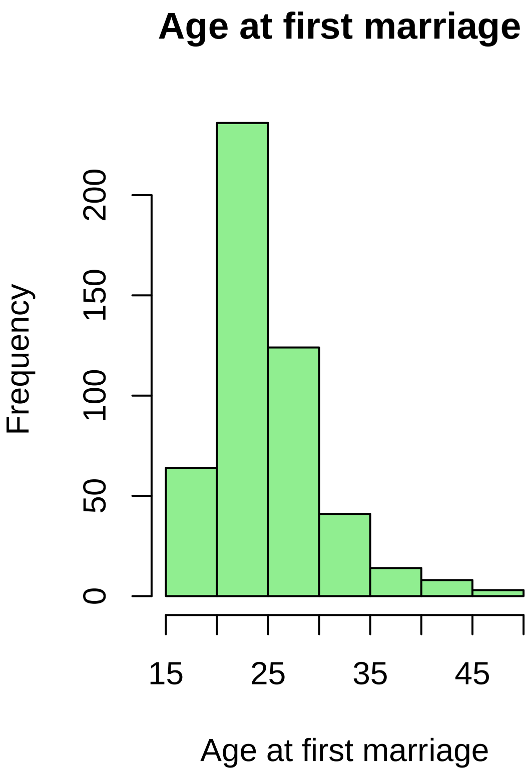


Figure3

Age of first marriage

This variable has an asymmetrical distribution with positive skewness. The majority of the data is located within the interval of 20-25, and the frequency displays a general pattern of descending as age increases.

Age of fist relationship

The majority of the data is distributed between the interval 55-60. The distribution of age of the first relationship is approximately skewed to the left and single-peaked. However, starting at the age of 45, the frequency of each bar becomes very close.

5.2 Best measure of central tendency

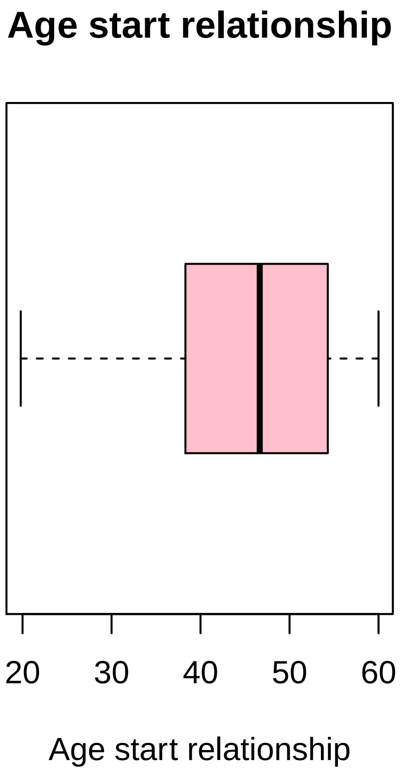


Figure4

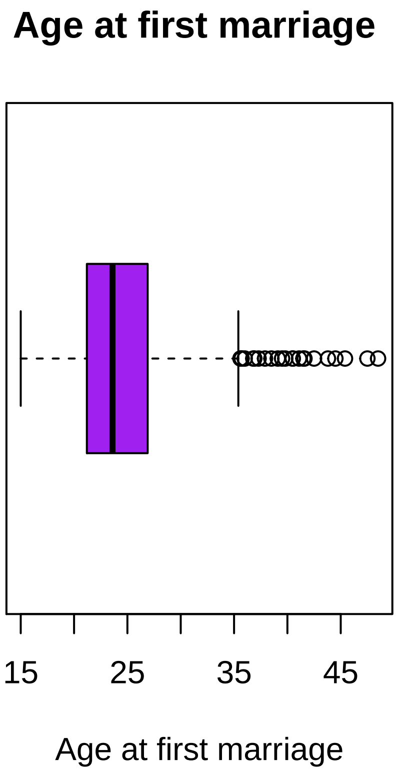


Figure5

Table 1: Summary of age start relationship

min	max	mean	median	sd
19.8	60	45.89041	46.65	9.589506

Table 2: Summary of age at first marriage

min	max	mean	median	sd
15	48.5	24.77	23.6	5.230368

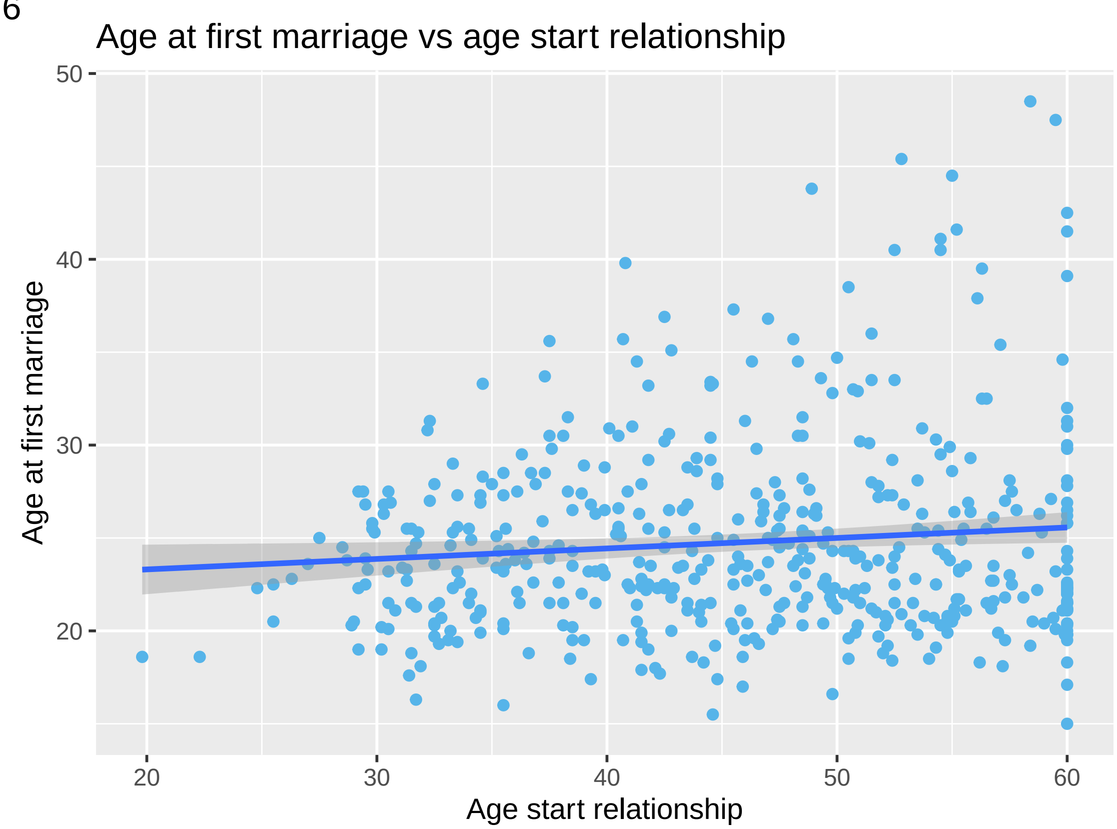
Age of first marriage

The best measure of central tendency for this variable is median. The box plot distribution is approximately symmetric, indicating that the mean, mode, and median are extremely close. Even though table2 shows that median and mean are very closed, the presence of outliers in this data set has stronger influence on mean than median. The skewed values cause the mean to raise and drag it away from the most accurate central position. However, the median receives fewer impacts from the skewed values than mean does. Consequently, median is the most representative and provides the best measure of central tendency.

Age of first date

The distribution of box plot is approximately skewed to indicate the same result as a histogram. From table1, the mean, median is very close. However, the skewed value still impacts the mean, so the best measured central is median.

Figure 6



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.18	1.151	19.27	6.236e-62
age_start_relationship	0.05643	0.02456	2.298	0.022

Table 4: Summary of linear regression

Observations	Residual Std. Error	R^2	Adjusted R^2
490	5.208	0.0107	0.008675

These plotted points constitute the scatter diagram. The general trend seems to be reasonably well represented by an upward-sloping line segment by inspecting the figure, as shown in the diagram. From table3 our linear regression model table, the equation is

$$\widehat{y} = 0.05643x + 22.18.$$

The interpretation of slope is when age of first date change 1 unit, the corresponding average difference of 0.05643 change in age of first marriage. The interpretation of intercept is the average value of age first marriage 22.18 when the age start relationship is 0. Note, there is no possible practical interpretation.In table 4 R ˆ2 is 0.01 means 1% of the data can be explained by the model, Rˆ 2= rˆ2.So r =0.103, which indicates a weak positive relationship between age of first marriage and age of first date. We use 0.05 as a benchmark. The p-value is less than benchmark. Therefore, we rejected our hypothesis.

6 Discussion

In the result section, we reject our null hypothesis, indicating that our alternative hypothesis is valid. Therefore, our assumption of age of first date and age of first marriage is correct. Hence, my model can perform an analysis between age of first marriage and age of first date.However, my model can not draw on all available evidence, as it is deliberately parsimonious. From Figure 6 we can conclude the model is not fitted very well .More detail will be discussed in the weakness section. The data that we are using is after a lot cleaning,wrangling and adjusting. However, there still exist some bias in the data which we discuss in next subsection.

6.1 Weaknesses

6.11 Limitation of dataset

Non-response bias

Non-response bias occurs when some individuals are under-represented, since they refuse to participate. In this study, non-response bias still exists because there are many NA samples in the dataset. This factor may favour specific outcomes as the missing data may contain outliers, which could influence the current study's results on the website, Statistics Canada attempts to reduce non-response bias involved a series of adjustments to the survey weights to account for non-response as much as possible. Although this might decrease the error rate, we cannot eliminate non-response bias.

Response bias

Response bias occurs when participants in the survey deliberately give a false or misleading answer. The respondents may want to influence the results unduly or be afraid or embarrassed to answer sensitive questions honestly. In this case, the particle might say wrong age of their first date.

Coverage error

Coverage error means there is no one-to-one correspondence between the target population and the sampling frame from which a sample can be drawn. The frame for GSS was created using several linked sources, such as the Census, administrative data and billing files. Coverage was improved (over coverage and under coverage may still exist). Statistics Canada uses adjusted (weighted) to represent all persons in the target population, including those not covered by the survey frame. However, the error can not be eliminated.

6.12 Limitation of the model

Firstly after we clean the data, the sample size is 490, indicating that the actual response rate for these two questions is much lower than 56% from the GSS website. Secondly, the R^2 is very small which indicate the model can not explain the data very well. Thirdly, the problem refers to the confounding variable, such as a new understanding of marriage. Nowadays, the Yonge generation may not want to get married to improve their relationship, and the law also protects the common-law partner, which could also mislead the result. Another limitation is that we lack multiple trials with data from different years to prove the correlation between the independent and dependent variables. The data varies from year to year.We only cover data obtained in 2017 from GSS, and this variation may cause the overall correlation to change so that the correlation may become more accurate if we run multiple trials with data from different years.

6.2 Next Steps

As we mention in the last section about modelling problem, we can try to use linear transformation to increase the accuracy of the model. Additionally, we can also attempt to use another model, such as logistic and multivariable, to eliminate the existing bias. To solve small sample size problem, we re-run the survey to increase the accuracy rate. Moreover, we can also use the old data set and other data to compare our data to improve validity.Finally,We also can use another similar data set to re-run the model to improve the reliability.

References

Gergely Daróczi and Roman Tsegelskyi (2018). pander: An R ‘Pandoc’ Writer. R package version 0.6.3.

General Social Survey - Family (GSS), Statistics Canada, 6 Feb. 2019, www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurv

GSS. (2019. Public Use Microdata File Documentation and User’s Guide.

GSS. (2019). Data. Retrived on Oct. 15th, 2020 from

https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm

Kahle, D. and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1):144–161

Kay, M. (2019). tidybayes: Tidy Data and Geoms for Bayesian Models. R package version 1.1.0.

Wu, Changbao, and Mary E. Thompson. “Basic Concepts in Survey Sampling.” Sampling Theory and Practice. Springer, Cham, 2020. 3-15.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report,Generation in R. R package version 1.30.