Have you ever married? A cross-sectional study on factors impacting conjugal history

Jingying Liu, Zhijian Zhu, Mingze Xu

October 19, 2020

Contents

Github link:

https://github.com/liujin59/sta304.git

0.1 Abstract

We were interested that whether an individual is married or not is significantly related to his/her age, income and education degree in Canada. We used a logistic regression model to address this question. We found out that age, income and education degree are all associated with the outcome of interest. Specifically, being older, earning over $50,000 and having a University degree all lead to a higher odds of getting married. The next steps would be fixing some study flaws and analysis shortcomings and account for other factors to improve the study and analysis results.

0.2 Introduction

Every day there are many couples register marriage in Canada. They are of different ages, having different social backgrounds and coming from all kinds of sectors and professions. Motivated by this phenomenon, we would like to predict whether an individual is married or not. Possible factors suggested by current evidence in the society includes age, income and education level are pivotal on a person's marriage.

We established a statistical model to assess the prediction power of variables like age, income, and education level on the binary outcome - Have the individual ever married or not. The data was extracted from the 2017 General Social Survey(GSS) on the Family, which is a nation-wide survey that taking in Canadians' social-demographic variables, as well as seizing the socio-economic trend over the country.

The factors that are predictive to individual's conjugal behavior would help social researcher learn the intentions and reason to union better, provide reference and evidence for governors to make policy and social decisions, and popularize the social knowledge to the public through various media.

0.3 Data

We got the data from the General Social Survey, which was introduced in the class.

The target population for the 2017 GSS is all people except all age smaller than 15 and all people from Yukon, Northwest Territories, and Nunavut. They use stratified random sampling. The stratum of sampling is the different province, and the sampling frame is all available telephone numbers and addresses registered in the same province. In each stratum, the sampling strategy is a simple random sample without replacement.

As a result of sampling, there are 20,602 data in the dataset, and the response rate was 52.4%. It would induce some non-sampling errors to appear because of the non-response on the survey results, but they used adjustments to the weights to minimize the problem. For the sampling errors, it is an unavoidable bias because the cost of sampling is finite, and we usually use a statistical method to avoid it such as bootstrap weights or cross-validation.

The use of stratified random sampling benefited the costs of the survey, but it hard to find the subgroup to represent the whole province.

Since the data set has plenty of non-response data, then after checked the data, I decided to use some parameters of low non-response(missing) and related data from the dataset, which are age income and education. The actual data plots are in the result.

0.4 Model

We use logistic regression model here to address the research question, assuming the outcome is Bernoulli distributed. It's a type of generalized linear models with a logit link function of the mean response. The outcome ever married, has a person ever married or not, is denoted as Y_i.

$$Y_i = \begin{cases} 1 & \text{if had ever married} \\ 0 & \text{if had never married} \end{cases}$$

Such that $Y_i \sim \text{Bernoulli}(p_i)$, where $p_i = P(Y_i = 1)$.

The three main predictors are age, income (over \$50,000 annual income or not), and education (with a University degree or not) The regression models log odds of the expected value of the response variable $\frac{p_i}{1-p_i}$. The coefficient β_i would be interpreted as, one unit increase in the value of predictor variable X_i leads to a β_i increase in log $\frac{p}{1-p}$, controlling other covariates to be held constant. The model equation could be written as follows:

$$log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_{age,i} + \beta_2 X_{income-over50K,i} + \beta_3 X_{education-university,i}$$

where $i = 1...n$.

The model selects age instead of age-group since the continuous data in age have more information than the Categorical age-group data, which makes us have less bias in the model. The income we divided to over \$50,000 annual income and less than the \$50,000 , since \$50,000 is like a mean of income in the dataset. The education level is can not be valuable, so we can only fit it in the model as a Categorical parameter.

0.5 Results

We used plots to visually check the outcome of interest the exposures, as well as their relationships.

We see that in Figure 1 our sample approximately 65% had married before.

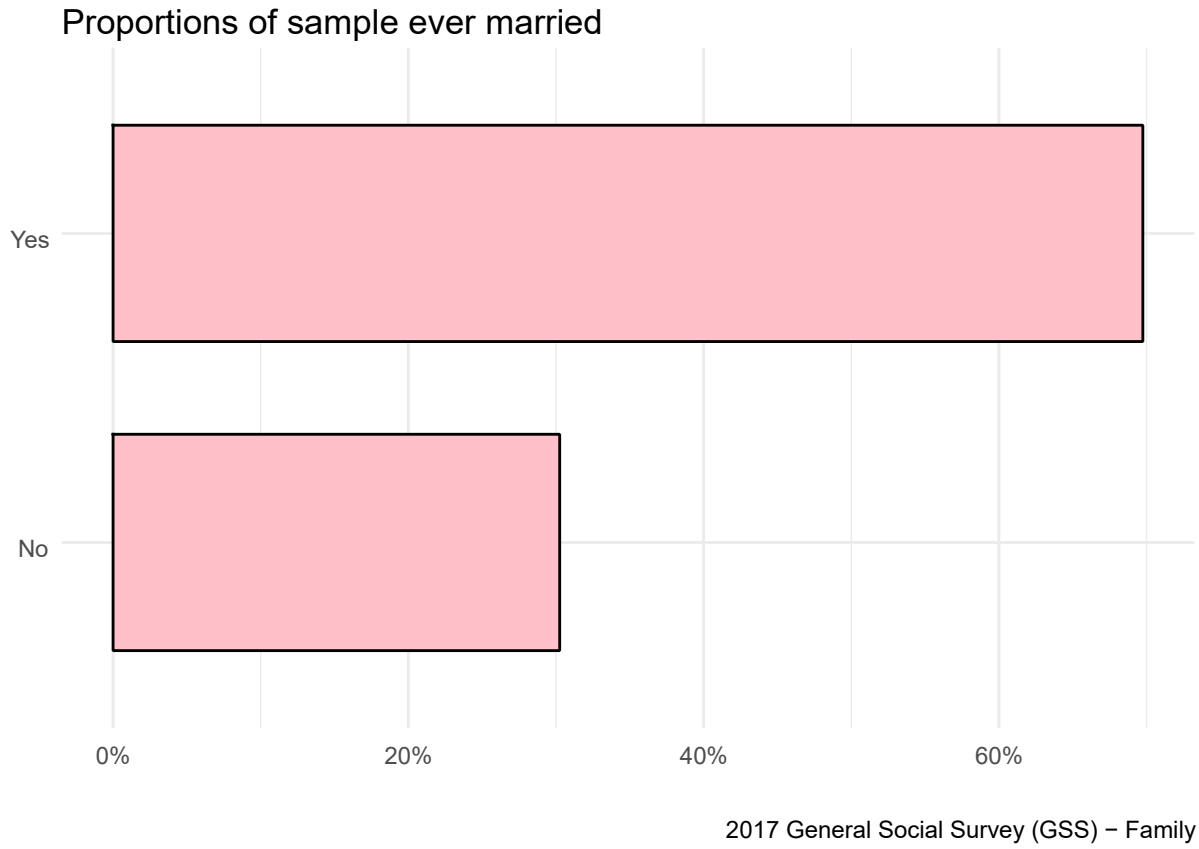


Figure 1: Percentage of selected individuals who had ever married in the 2017 CSS survey

The distribution of age is slightly left skewed and the median is about 55 (Figure 2).

We see that in figure 3, approximately 65% of our sample earn greater than $100,000 a year.

We see that in Figure 4 close to 20% of our sample have a University degree.

We see that in Figure @ref(fig:figure5) education does not really have an association with whether an individual had ever married or not. Also the age distribution for those who had ever married is generally larger in median, almost equally spread, which makes sense socially.

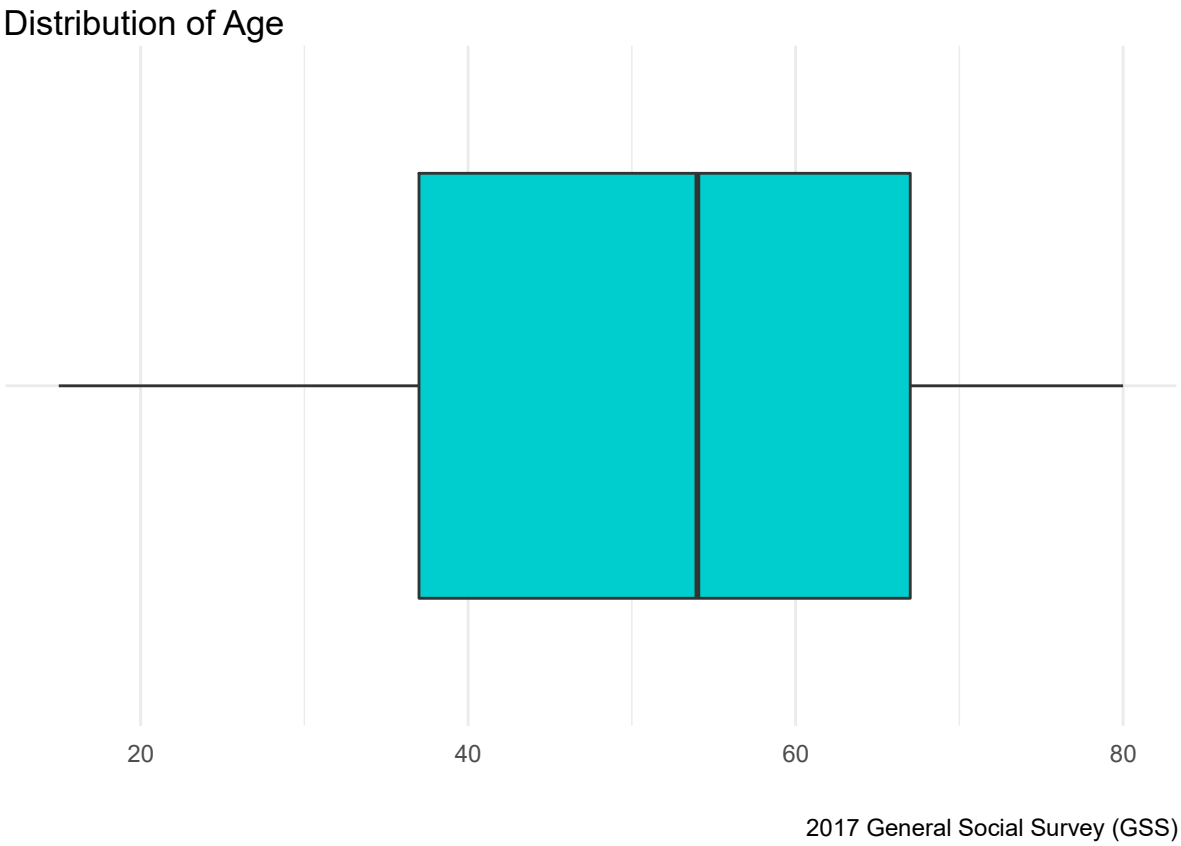


Figure 2: Distribution of age of selected individuals in the 2017 CSS survey

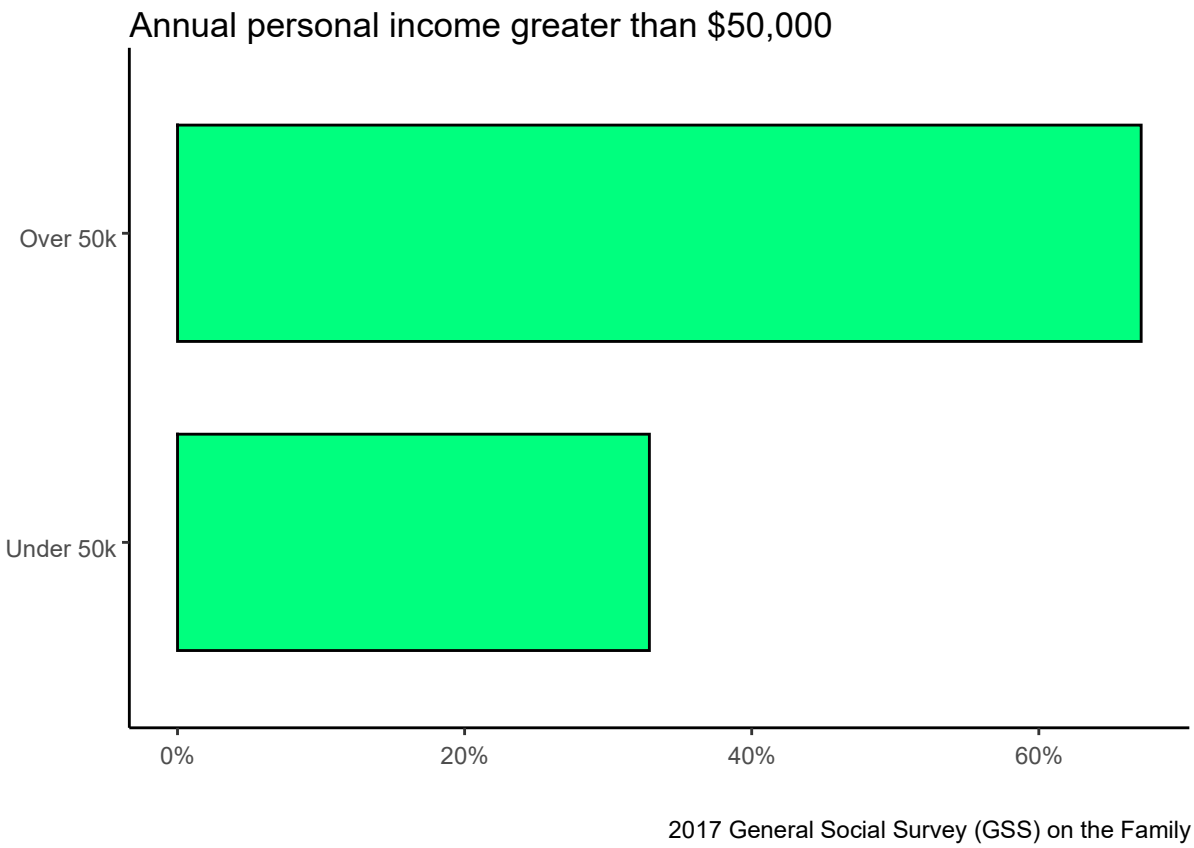
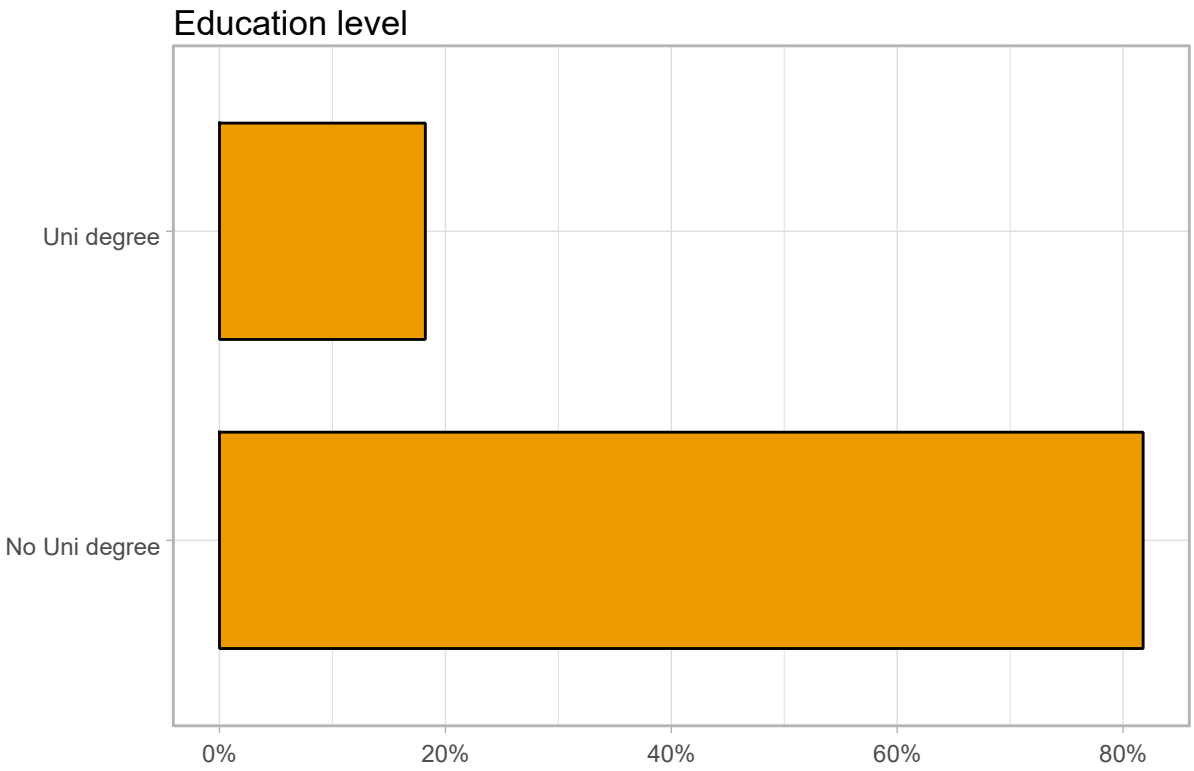
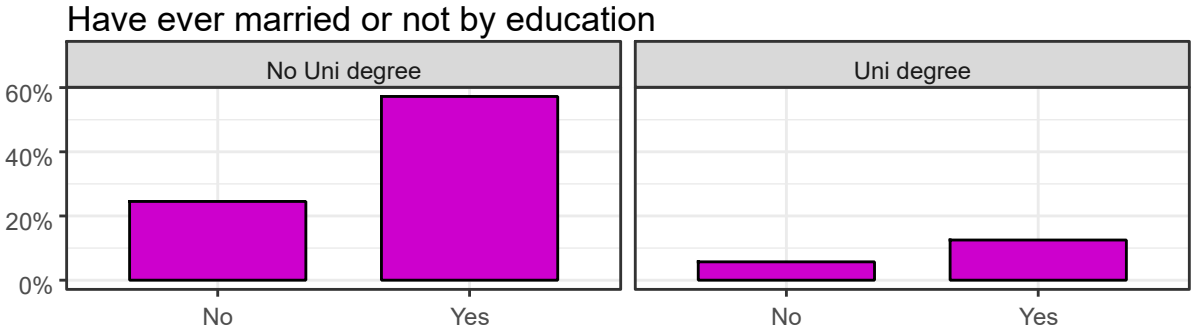


Figure 3: Percentage of selected individuals who earn greater than $50,000 a year in the 2017 CSS survey



2017 General Social Survey (GSS) on the Family

Figure 4: Breakdown of education degree of selected individuals in the 2017 CSS survey



2017 General Social Survey (GSS)

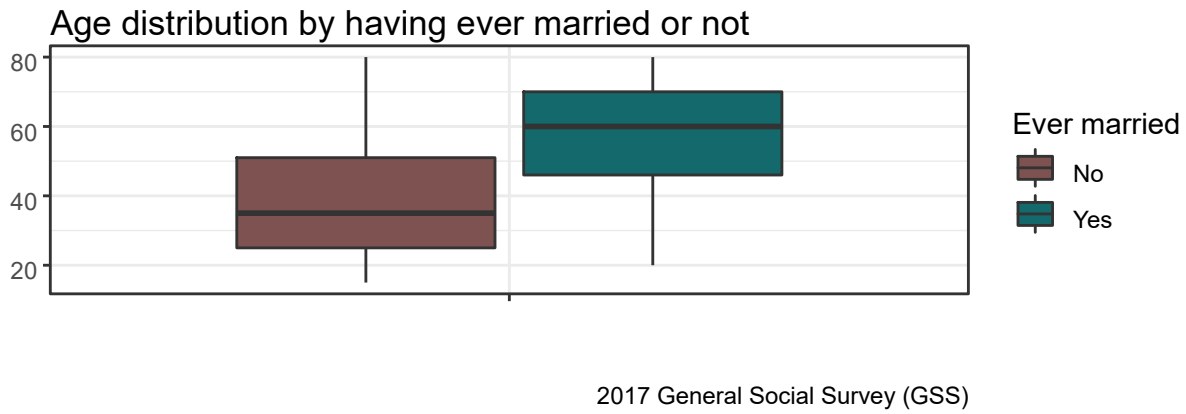


Figure 5: Percentage of selected individuals who had ever married and age distribution by education.

Model results

Table 1 contains the logistic regression model output. We see that the P-values for age, income and education degree are all far less than 0.05 (significance level), which means highly significant. So age, income and education degree are statistically significant in predicting whether an individual has ever married.

Table 1: Logit Model Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.15	0.06	-49.06	0
age	0.08	0.00	63.73	0
income_newOver 50k	0.41	0.04	10.61	0
education_newUni degree	0.17	0.04	3.69	0

Table 2 reports the estimated odds ratio with the 95% confidence interval from the logit model. We see that individuals of older age, with higher income, and with a university degree are more likely to get married than those younger, less-educated people with lower income.

Table 2: Estimated Odds Ratio and 95% Confidence Interval

	OR	2.5 %	97.5 %
(Intercept)	0.04	0.04	0.05
age	1.08	1.08	1.08
income_newOver 50k	1.51	1.40	1.63
education_newUni degree	1.18	1.08	1.29

0.6 Discussion

The odds of having ever married increases by 0.08 with an additional year increase in age, keeping other predictors fixed. The odds of having ever married for the richer people (who earn more than $50,000 annually) is 1.51 times that of people with less income. For those with a university degree, the odds is 1.18 times that of people without one.

Weaknesses

One major weakness in our study is the lack of ability to generalize the results. We did not consider all relative parameters in the dataset and did not consider to minimize the sampling error. Furthermore, we did not create a new model for the model comparison. Also, the bias seem becomes higher since we divide the income in to two unbalanced subsets and the people who have a University degree are much less than the people without one in our sample. The dataset is overpowered for our logisitic regression model due to such a huge sample size, which resulted in extremely small P-values but not so large odds ratios (not large in relative magnitude). Additionally, model fit was not assessed and the causality pathway is still unclear, we cannot establish conclusions implying strong causality.

Next Steps

Beside the quantification, we are not surprised that age, income and education degree are highly associated with a person's conjugal history, as they are pivotal factors that would affect people's intentions and reasons to unite. But it's not a simple outcome that is caused by only these 3 factors and there are still other factors that influences the incidence, such as race, religious belief, communication ability, times have spent in Canada, etc, could be confounding factors. The next steps would be building another logistic model that assess the predictiveness and influence of these additional factors. What's more, we will look into the causality pathway and see whether they actually predict the outcome or simply relates to it and remove the bias in interpretation. If data not be available from the 2017 CSS survey, we could try to find other national surveys that capture the information we would require to conduct further analysis.

References

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Francois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.Rproject.org/package=scales.