An Analysis of Predictors for a Committed Relationship

Jasmine Carlos, Matthew Caringi, Haeun Choi, Mahmoud Elsheikh

10/19/2020

Abstract

The purpose of this study is an investigation into the optimal predictors for a committed relationship using data from the 2017 general social survey. Analysis was carried out through the creation of a standard logistic regression. It was found that the variables of level of education, total number of children, age of respondent's first child and income of the family are appropriate predictors of a committed relationship. The information found is important for the discovery of factors that affect whether or not Canadians stay in a committed relationship.

Note that all code and data used in this analysis can be found at the following link: https://github.com/jasminekcarlos/predictors-of-committed-relationships

Introduction

We examined the 2017 Canadian General Social Survey (GSS) conducted on the topic of "Families" in order to investigate the marital status of Canadians. The survey consists of numbers of unbiased questionnaires that asks the income, education and marital status etc of the participants. We primarily focused on what factors would affect the length of the committed relationships and chose 4 corresponding variables which are education group, total number of children, age of respondent's first child and income of family.

In order to reduce the data set to include only the variables of interest, we cleaned the data that we derived from the GSS. We cleaned the data which removed all the "N/A" values for the variables and created new groupings for the variable of education and marital status. Then we created a standard logistic regression model assigning the dependent variable as "Marital status" and the predictor variables as "education group, total number of children, age of respondent's first child and income of the family."

There are several main findings we discovered from this survey. First of all, the many Canadians who participated in the GSS survey seemed to have highschool diplomas or less and the next majority of the group has postsecondary education followed by trade school education. In all three categories of education, most of them were in committed relationships. Another aspect we found was that respondents without any kid seemed to be in non-committed relationships compared to respondents with kids. We also investigated that families with higher income have a higher chance of being in committed relationships. Some of the weaknesses that we have in this paper is that certain variables had "N/A" datas so we had to filter out the N/A responses. Due to this, 345 observations were lost which can lead to more skewed data. Also, since it was a telephone survey that was conducted, the response rate is not as high as other methods of surveys which can lead to limited data. In the future, we can consider other variables that are more related to the marital status.

Data

The 2017 General Social Survey (GSS) conducted by Statistics Canada was a questionnaire that focused on the topic of "Families" to analyze certain trends within the population of that year. The questionnaire

consisted of personal questions such as income, age, marital status, etc, which supported the survey's intention by covering a large area of subtopics under the general theme. These questions displayed no bias as they were not leading or self-incriminating questions that may influence the respondents' answers. Although there is an aspect of response bias due to the respondents self-reporting their answers, this relies heavily on the respondent's memory and ability to answer the questions. Cross-validations from other sources to support the respondent's answers would assist to avoid this type of bias.

The survey's target population consists of people 15 years old or older that reside in Canada. Exclusion criteria filters out residents living in the territories (Nunavut, Northwest Territories, and Yukon), as well as in institutions (hospitals, jails, and nursing homes). The survey's sample population consists of 20 602 households with telephone numbers in the 10 provinces. Statistics Canada used stratification sampling on geographic areas by dividing the 10 provinces into 27 stratas. Stratification sampling allows the survey to include any underrepresented groups that are considered the minority in the population that would be more likely to be unselected by clustered or random sampling. Therefore, stratification sampling allows for a more precise illustration when looking at a large diverse population such as the 10 provinces within Canada.

To establish the sample population, Statistics Canada cross referenced telephone numbers to the address register. The frame of the sample population included telephone numbers that were linked and not linked to an address, thus allowing better coverage. In the case where more than one telephone number was linked to an address, the landline number was used. Once contacted, random sampling of the respondent within the household who answered was selected and the questionnaire was conducted via computer assisted telephone interview (CATI). One drawback of telephone interviews is that they exclude households without telephones which tend to be residents with lower income and education.

The response rate of the 2017 GSS was 52.4%. Statistics Canada prepared procedures to counter non-responses. In cases where respondents initially refused to participate, they were re-contacted two more times. If calls were not answered, multiple attempts were made to reach the household. The survey did not ask the name of the respondent, so privacy concerns were kept minimal. In regard to the cost of the survey, Statistics Canada did not state any cost, only that the CATI method cost significantly less than in-person interviews. Some strengths of the entire survey include the cross-sectional design of using telephone numbers and home addresses, a large sample size with clear inclusion and exclusion criteria, and appropriate sampling methods based on geographic area. The weaknesses of the entire survey include the long interview time of 45 minutes which can prompt non-responses, the factor of response bias from the validity of the answers, and the exclusion of households without telephones that can omit data from lower income and less educated households.

The variable of interest in scope of this investigation is the respondent's marital status. The predictor variables include respondent's education, total number of children, age of their first child, and family income. Some of the variables were not applicable when observing influence on marital status. This includes predictor variables such as age of first marriage which would not impact our respondents who are single and have never been married. Therefore, the chosen predictors and variable of interest support the investigation's aim on the factors that influence an individual's marital status.

Model

$$P(y = 1) = \text{logit}^{-1}\left(\alpha^{margrp} + \alpha^{educ} + \alpha^{totalchild} + \alpha^{age1stchild} + \alpha^{incomefam}\right)$$

For our model we made use of a generalized linear model which is a flexible generalization of a regular linear regression. This allows for response variables with alternative distributions to the normal distribution. The dependent variable used was the marital group as this is our main variable of interest and the predictor variables used are education group, total number of children, age of respondent's first child and income of the family. We chose quantitative variables so that the GLM was consistent with our data properties in order to run the variables through the model. The categorical variables used are marital group and education group and the as.factor function was utilised to process them. Our reference variable for this model in the education group was the high school diploma or less and for the income group was the $100,000 – $124,999 group. All

Table 1: Summary of Model Statistics

term	estimate	std.error	statistic	p.value
(Intercept)	-1.7397	0.1445	-12.0405	0.0000
as.factor(education_group)Post-Secondary school certifcate or diploma	0.2018	0.0739	2.7289	0.0064
as.factor(education_group)Trade certificate or diploma	0.0987	0.0930	1.0618	0.2883
total_children	-0.1377	0.0258	-5.3468	0.0000
age_first_child	-0.0274	0.0019	-14.3214	0.0000
as.factor(income_family)$125,000 and more	-0.7380	0.1713	-4.3074	0.0000
as.factor(income_family)$25,000 to $49,999	2.1296	0.1386	15.3606	0.0000
as.factor(income_family)$50,000 to $74,999	1.4388	0.1404	10.2452	0.0000
as.factor(income_family)$75,000 to $99,999	0.7247	0.1514	4.7879	0.0000
as.factor(income_family)Less than $25,000	3.2845	0.1453	22.6000	0.0000

of our results from the model are at a p-value of under 0.05 making all the variable coefficients significant values.

The model was run using R through the use of the 'glm' function allowing us to interpret the intercept coefficient estimates. The generalised linear model allows for a clear understanding of how the predictors relate to the outcome. It allows us to interpret the data through obtaining coefficients for different factors such as family income at different income levels. The dependent variable from the data is set to "binomial" as the results of marital status have been divided into two sets of committed or non-committed relationships.

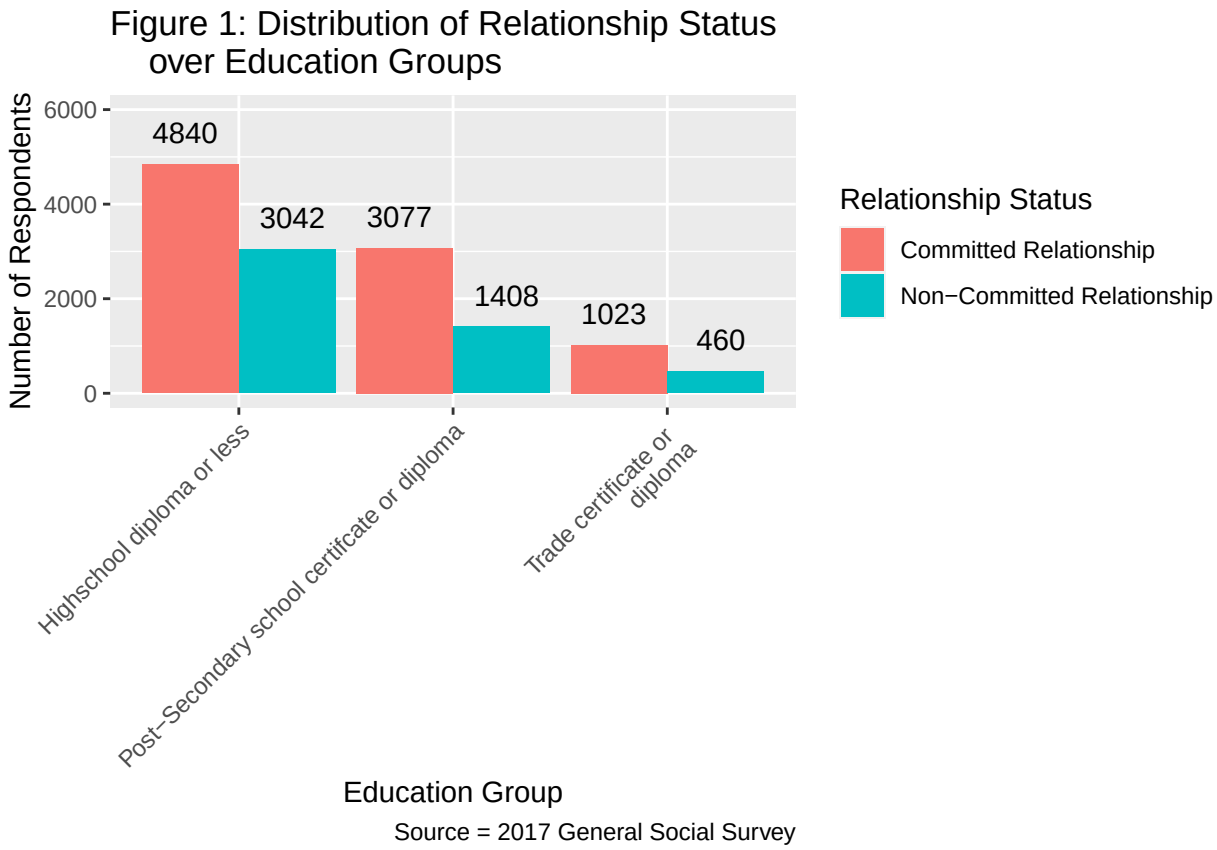
As you can see the negative value for the intercept for the age of the first child is negative; however, this is a very small value of -0.0274. This can be explained as the relationship approximates a normal distribution where the correlation is initially positive and then becomes negative as the age of the first child increases and evens out around -0.0274. The total number of children has a negative coefficient of -0.138 supporting our assumption that there is a weak yet significant correlation between an increase in the number of children and an increased likelihood of being in a non-committed relationship.

The positive coefficients for income brackets support the assumption that the model predicts the correlation between the likelihood of being in a committed relationship and each of our income brackets. The coefficient for income becomes more positive as income decreases. A reason for this may be that those in difficult financial situations are more stressed and unable to stay in a committed relationship.

The model is limited as it does not allow us to account for confounding factors such as age. An improvement would be to find a model that would allow us to find out how age was correlated to other factors in order to better understand if age is a confounding variable in our assumptions. While the Bayesian model takes a fuller account of all the uncertainties of the model, the use of prior distribution can carry significant weight on the distribution leading to misleading results. The GLM model runs with no issues and we came across no diagnostic issues in our model checks.

Results

In this study, the variable of interest is the marital group of the respondents. The chosen influential factors include education, number of children, age of first child, and income of family. Within committed relationship status, this covers respondents who are married, common-law, or widowed. While non-committed relationship status covers respondents who are divorced, separated, or single.



In Figure 1, the influence of education groups on marital status was investigated. Within all three categories, more respondents were in committed relationships rather than non-committed relationships. The calculation below describes the percentage of respondents in committed relationships within their education group.

Committed Relationship (High school Diploma or less)

$$= (\frac{4840}{4840 + 3042}) * 100\%$$

$$= (\frac{4840}{7882}) * 100\%$$
$$= 61.4\%$$

Committed Relationship (Trade Certificate or Diploma)

$$= (\frac{1023}{1023 + 460}) * 100\%$$

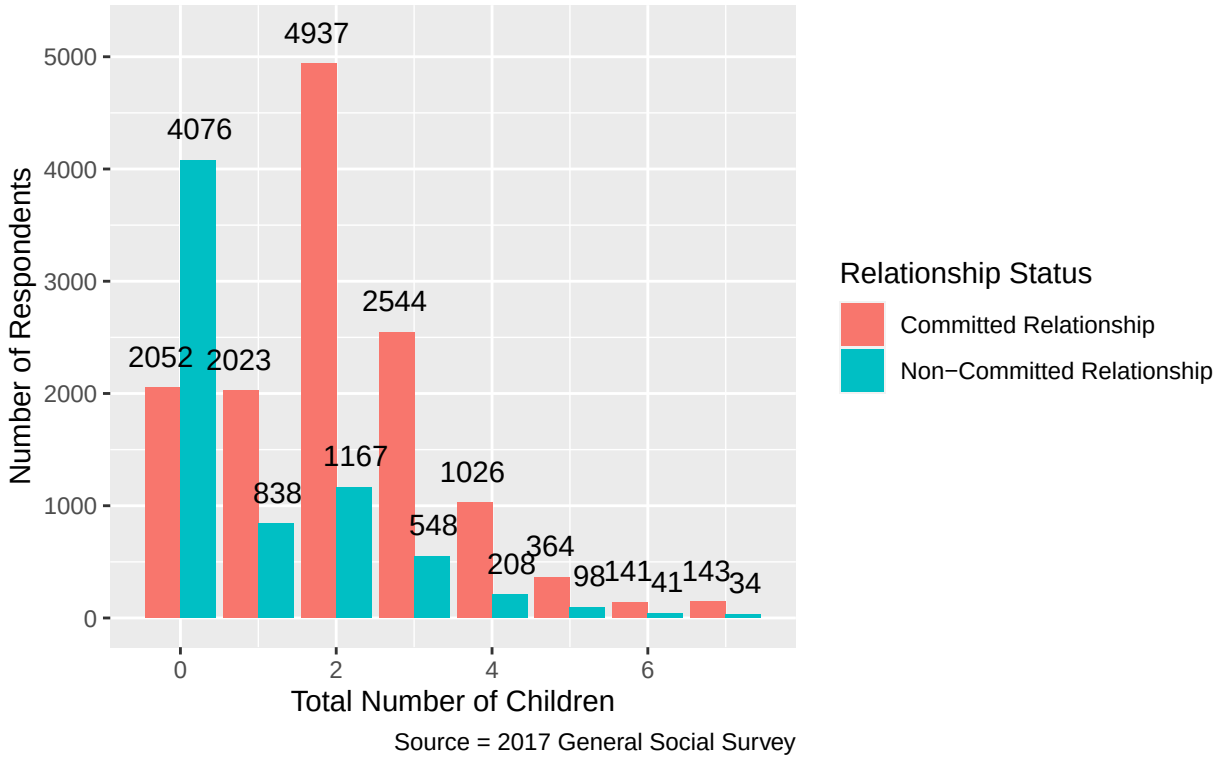
$$= (\frac{1023}{1483}) * 100\%$$
$$= 69.0\%$$

Committed Relationship (Post-Secondary School Certificate or Diploma)

$$= (\frac{3077}{3077 + 1408}) * 100\%$$

$$= (\frac{3077}{4485}) * 100\%$$
$$= 68.6\%$$

Figure 2: Distribution of
Relationship Status over Total Number of Children



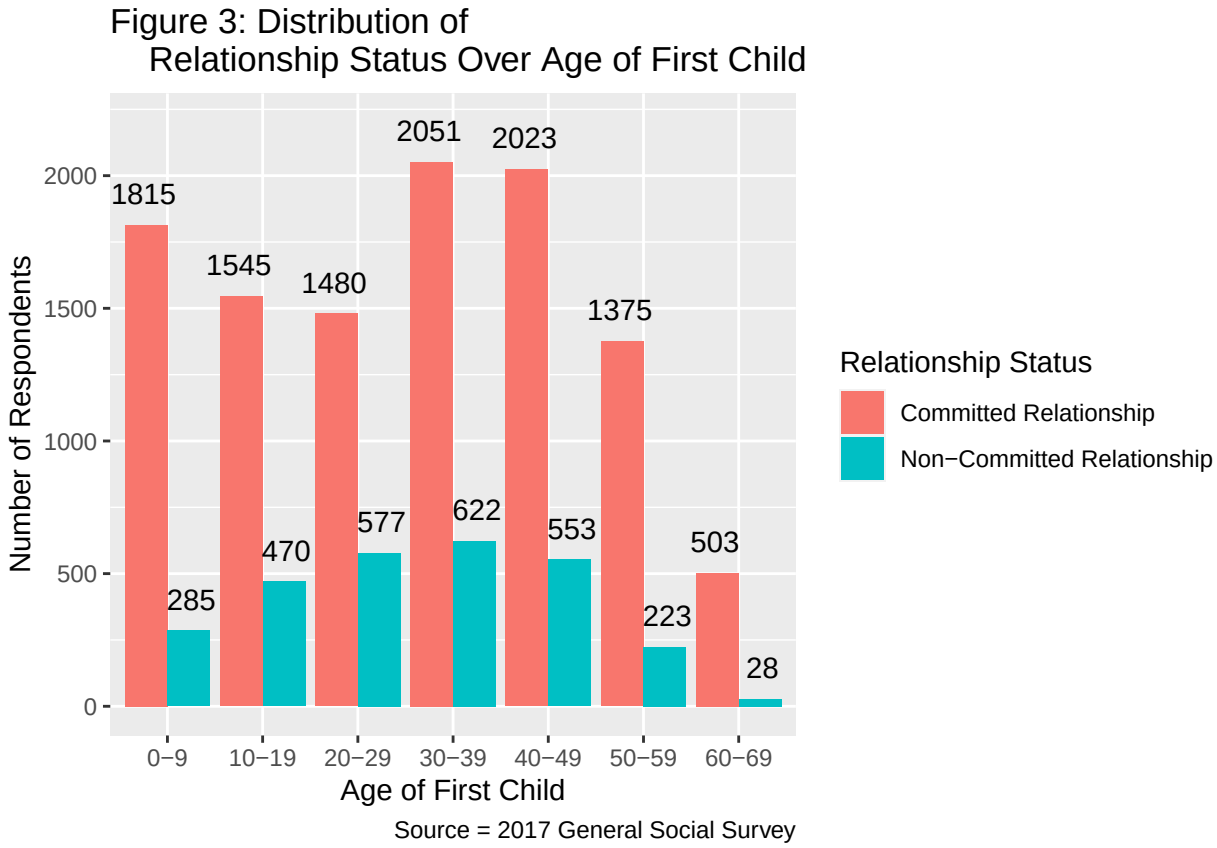
In Figure 2, the distribution of marital group over the total number of children is shown. One observation is that there is a peak of respondents in non-committed relationships when they have 0 children. Meanwhile, the majority of respondents with 1-7 children are in committed relationships, with a peak number of respondents with 2 children. The calculation below presents the relative ratio of respondents with children who are in committed relationships. Therefore, respondents with children are 2.4 times more likely to be in a committed relationship versus respondents without children.

Committed Relationship with Children = 11178 Committed Relationship with No Children = 2052 Non-Committed Relationship with Children = 2934 Non-Committed Relationship with No Children = 4076

Committed Relationship (CR) with Children Relative Ratio

$$= \frac{\frac{CR\&Child}{All\&Child}}{\frac{CR\&NoChild}{All\&NoChild}}$$

$$= \frac{\frac{11178}{14112}}{\frac{2052}{6128}}$$
$$= 2.4$$



In Figure 3, the age of the respondent's first child is compared against the respondent's marital status. In this graph, only respondents with children were included. As stated previously, respondents with children are more likely to be in a committed relationship. Listed below are calculations of respondents with children (age < 40) in committed relationships based upon their first child's age. Committed Relationship (0-9)

$$= (\frac{1815}{1815 + 285}) * 100\%$$

$$= 86.4\%$$

Committed Relationship (10-19)

$$= (\frac{1545}{1545 + 470}) * 100\%$$

$$= 76.7\%$$

Committed Relationship (20-29)

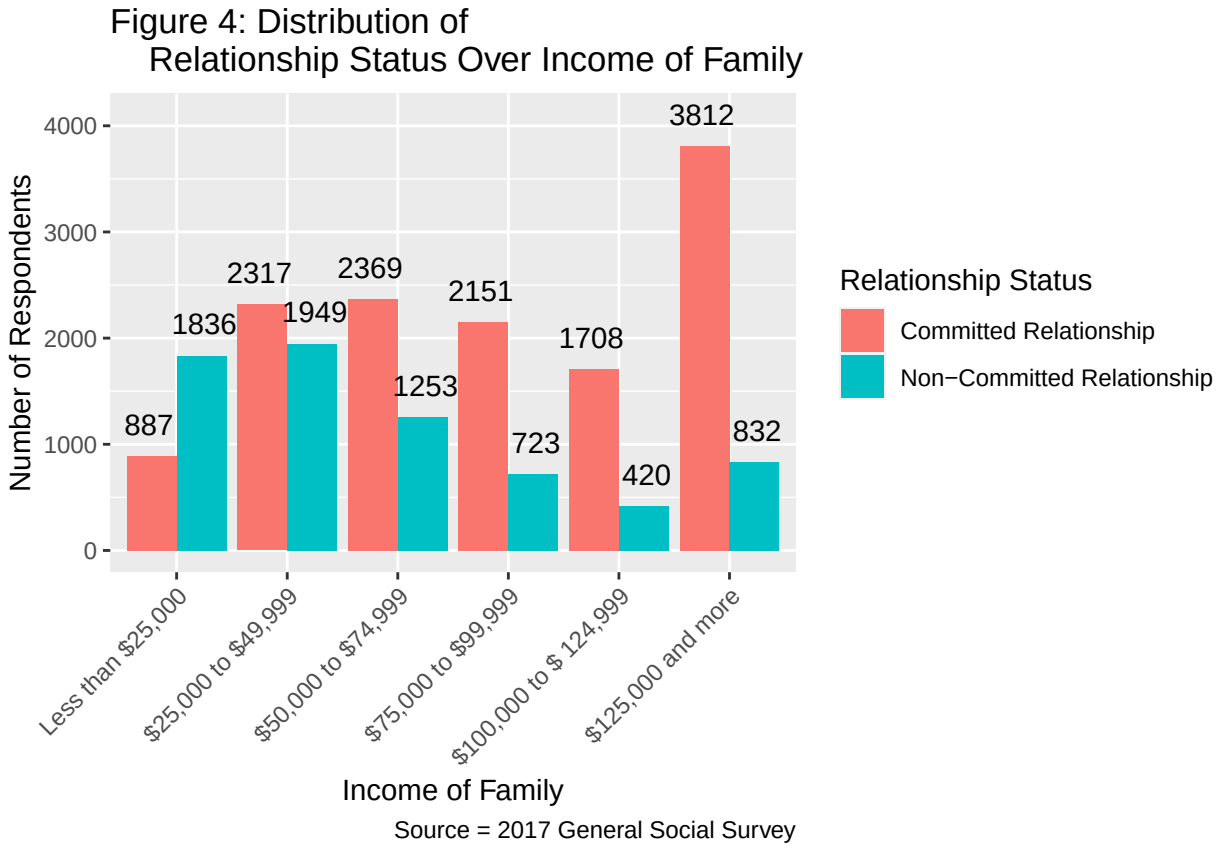
$$= (\frac{1480}{1480 + 577}) * 100\%$$

$$= 71.9\%$$

Committed Relationship (30-39)

$$= (\frac{2051}{2051 + 622}) * 100\%$$

$$= 76.7\%$$



In Figure 4, the distribution of marital status over family income is presented. Shown in the graph, respondents whose household income is less than $25,000 tend to be in non-committed relationships while respondents whose household income is $25,000 or more tend to be in committed relationships. Additionally, there is a spike of respondents in committed relationships when the family income is $125,000 or more.

In Table 1, the linear regression model is shown to observe the influence of education, total number of children, age of first child, and income of family on non-committed relationships. In the table, the "Intercept" is used as the initial value and the "Estimate" is used for the variables' coefficient. Negative coefficients represent a negative association with non-committed relationships, shown with the following variables: total number of children, age of first child, and family income $125,000. Positive coefficients represent a positive association with non-committed relationships, shown with the following variables: post-secondary school education, trade school education, and family income $99,999. High school education is used as a reference value for the rest of the education group, as well as family income of $100,000 to $124,999 is used as a reference value for the rest of the family income group, and thus not present in the table. A p-value of less than 0.05 was used to show statistical significance. Below is the logistic calculation of the model.

Post-Secondary School Certificate/Diploma = edu2

Trade Certificate/Diploma = edu1

Total Children = children

Age of First Child = age

Family Income $125,000 or more = income5

Family Income $75,000 - $99,999 = income4

Family Income $50,000 - $74,999 = income3

Family Income $25,000 - $49,999 = income2

Family Income less than $25,000 = income1

$$P(p/(1-p)) = -1.74 + 0.0987 * edu1 + 0.202 * edu2 - 0.138 * children - 0.0274 * age$$
$$+3.28 * income1 + 2.13 * income2 + 1.44 * income3 + 0.725 * income4 - 0.738 * income5$$

Discussion

This dataset was obtained from the Canadian General Social Survey (GSS) with 20 602 respondents and was conducted using stratified sampling with each of the 10 Canadian provinces being divided into individual strata. The results of the dataset were obtained through a random telephone survey which is a relatively effective and cost-effective method of obtaining data, allowing for a large sample size. Telephone survey response rates have been declining to 6-7% in recent years which can lead to a potentially significant non-response bias. This could remove certain demographics from the survey such as busier individuals with longer working hours who don't have time to respond to these surveys. The GSS obtained a response rate of over 52% due to recontacting respondents who chose not to participate allowing for data that is more representative of the population.

In our evaluation of what factors could potentially influence marital status, we have made some interesting findings. Figure 1 suggests that an increased level of education increased the likelihood of a subject being in a committed relationship. This interesting finding could be a result of a better education enabling better income and therefore incentivising subjects to enter committed relationships due to financial stability. This point is supported in figure 4 as there is a positive correlation between increased income and increased proportion of subjects being in a committed relationship. It is important to consider the possibility that age may be a confounding variable in this analysis as age is positively correlated with both income and the probability of being in a committed relationship. This could mean that age, rather than increased income, is positively correlated with the likelihood of being in a committed relationship.

Another interesting finding is that the proportion of subjects in a committed relationship is positively correlated with the number of children they have. This is visible in figure 2 where the proportion changes from 70% to 82% when increasing from one child to two children with the model also showing this correlation with a coefficient of -0.138. This could be explained as being in a committed relationship can lessen the burden of childcare on the parents or because people in committed relationships tend to have more children.

Figure 3 shows the relationship between marital status and the age of the first child. This figure suggests that as the age of the first child increases, the subjects are less likely to be in a committed relationship. This is an interesting finding that contradicts the notion that commitment and age are positively correlated as age and age of first child are positively correlated. This could be explained by assuming that as the age of the first child increases, the probability of divorce will also increase explaining the correlation.

A weakness that limits our analysis was the non-response in many of the fields in variables we are investigating such as first child age that could potentially manipulate the data if the subjects prefer not to answer share common characteristics. Another weakness could be in the consensus on what is classified as a committed relationship. There are levels to commitment in relationships and the definition may vary from subject to subject. This could alter the subject's responses and manipulate the data and consequently our findings.

References

- "General Social Survey: An Overview, 2019." February 20, 2019. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm.
- Government of Canada, Statistics Canada. 2016. "General Social Survey - Family (GSS)." December 20, 2016. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501.

- "GSS31_User_Guide.Pdf." n.d. Accessed October 19, 2020. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf.
- Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra
- "Institutional Resident - 2011 Census Dictionary." n.d. Accessed October 19, 2020. https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/pop053-eng.cfm.
- JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com.
- NW, 1615 L. St, Suite 800Washington, and DC 20036USA202-419-4300 | Main202-857-8562 | Fax202-419-4372 | Media Inquiries. n.d. "Response Rates in Telephone Surveys Have Resumed Their Decline." Pew Research Center (blog). Accessed October 19, 2020. https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.