Feeding the Children: Investigating the Likelihood of Individuals Intending to Start Families

Karrie Chou

October 19, 2020

Abstract

The following statistical analysis attempts to answer the question of what identifiable characteristics of a person's background significantly impact the likelihood that they would want to start a family by having children in the future. While there is a significant body of research devoted to how families, once started, function as a unit, but the area of study devoted to how these units form in the first place deserves more attention. By analyzing the 2017 Canadian General Social Survey (GSS) on families compiled by Statistics Canada and constructing a potential statistical model to analyze the effects of certain given answers on a respondent's stated likelihood that they intend to have children and essentially start their own family in the future.

For this statistical analysis, a logistic regression model based on the frquentist approach to statistical inference was created using some of the numeric and categorical variables with factor levels that could be translated into numeric scales as predictor variables and a binary "dummy" variable that captured the stated intent to have children in the future in a numeric form as the response variable. By running a regression of this response variable on each of these predictor variables, estimates for model coefficients were computed and evaluated based on the significance of their impact on the response variable. The conclusion of the statistical analysis is that there exist factors such as age, personal income, and number of marriages which significantly decrease an individual's likelihood that they want to have children in the future.

Code and data supporting this statistical analysis can be found at https://github.com/karriechou/sta304ps2.

Introduction

Family, defined by the Marriam-Webster Dictionary as "the basic unit in society traditionally consisting of two parents rearing their children", is an integral aspect of many cultures around the world. In Canada, whose reputation as a stable and multicultural nation has made it a location of interest for many people looking to settle in communities that they are familiar with, families manifest in many different conditions and have different characteristics across the board. From a family's parents' marriage histories, ethnicity, socioeconomic background, and even priorities when it comes to raising children, we can identify trends among how families stay united or even how they separate across different cultures.

This report will focus specifically on identifying factors related to an individual's cultural, socioeconomic, and biological background which affect their intent to have children and essentially start a family. By buildng a statistical model to track the significance of the impact of each of these chosen factors on an individual's probability that they want to have children in the future, the understanding of how families are formed will be deepened. The findings of this research can be then used to predict potential demographic trends such as birth rate and population growth using historical population data, which is highly relevant to any country which aims to have sufficient resources to support its citizens.

Data

The data used within this report comes from the 2017 GSS on families compiled by Statistics Canada and distributed through the Computing in the Humanities and Social Sciences (CHASS) Data Service, created by the Faculty of Arts and Sciences at the University of Toronto. Each data entry comes from a survey respondent's answers to a questionnaire on topics related to characteristics of themselves and their families. Within this dataset, there are 20,602 entries which form a collectively representative sample of the associated survey's target population, which is all Canadian residents.

Important features of the data and potential drawbacks

According to the user guide that accompanies this dataset, every respondent was verified to be at least 15 years of age or older. This check was limited at 80 years; anyone older than 80 years of age had their age simply listed as 80.0 in the age variable column. A respondent's age was also used as a way for respondents to avoid answering irrelevant questions, marked by the possibility of 'Valid skip' as an answer to several of the questions represented by the different variables in the dataset. Because the 'Valid skip' was implemented in various questions according to the age of the respondent, the cleaned dataset contains a lot of missing data, making it difficult to conduct any statistical analysis involving a given variable without introducing some degree of non-response bias.

Another aspect of the data that is important to highlight as a potential drawback to its usability is that to obtain categories for family income and respondent income, the questionnaire was linked to respondents' tax information. This was presumably done in place of directly asking respondents to disclose their income. Tax information is likely not a perfectly accurate source to determine actual income from, as tax return amounts are calculated based on placing the individual in different tax brackets which encompass ranges of income. Additionally, tax deductions listed in tax information can be subjectively applied to individuals; even if two individuals have the same income and expenditures, their underlying financial circumstances may differ in such a way as to allow each of them to have different exemptions. Based on these fact, having incomes in this dataset listed as ranges is a logical decision. However, any statistical analysis involving family and respondent income will inherently have introduced bias by attempting to place a numerical value on the specified ranges, as conventional methods such as taking the midpoint of the income range may not reflect the true mean of incomes of respondents who fall within that same range.

Initial data cleaning

Within the initial data file downloaded from CHASS, there are both numeric and categorical variables. The values of categorical variables for each observation in this data file take the form of numeric codes which correspond to certain factor levels, or English text categories. For humans, these codes are not readable. In order to make the data file usable, a couple of tasks need to be done:

- All numeric codes representing categorical variable levels must be replaced with their corresponding English text.
- All variable names, which are currently alphanumeric codes, must be replaced with English readable names.

These processes are completed using gss_cleaning.R.

Preparing cleaned data for statistical analysis

Once the data has been cleaned, several new variables need to be constructed in order to create the framework for the model. Because the response variable of interest is the stated intent to have children in the future, a dummy variable based on the existing future_children_intention variable needs to be created. In the cleaned

data, future_children_intention is a categorical variable with 9 different factor levels, 5 of which result in an empty data entry. The other 4 factor levels are:

- No, definitely not
- Probably not
- Probably yes
- Definitely yes

The dummy variable that will be constructed to represent a respondent's stated intent to have children in the future, intent_children, takes a value of 1 if the value of future_children_intention for that respondent is "Definitely yes" or "Probably yes". Otherwise, it takes a value of 0. intent_children serves as the response variable in the model proposed later on in this report.

Two other variables that were created for statistical analysis are num_income_family and num_income_respondent. These variables take the midpoint of each of the factor levels within the categorical variables income_family and income_respondent in order to better assist with quantitative statistical analysis and data plotting. For the factor level "$125,000 and more", a numerical value of $125,000 was used because it is impossible to determine a midpoint between $125,000 and what could potentially be an infiniely large number.

Model

Selecting a model

The model that will be used in this analysis is a *logistic model based on frequentist inference*. A logistic model is used when fitting a stragiht-line or a linear regression model is not approrpiate for the data, which is typically the case for categorical response variables that are binary (i.e. they only have two possible values). A logistic model is approrpiate for this analysis because the response variable intent_children is a binary categorical variable which equals 1 if the respondent indicated they were definitely or probably interested in having children in the future, and 0 otherwise, provided that the respondent did not skip the relevant question in the original GSS questionnaire. Using a *frequentist* approach to inference means that the statistical analysis in this report assumes that the response variable's probability of success (i.e. the probability that a resopndent will say that they intend on having children in the future) is fixed in the population. Thus, the data in this statistical analysis is treated as historical data that can be used to predict the long-run trend in how likely someone is to say that they intend to have children based on certain predictor variables.

For a general logistic model, the corresponding mathematical expression is:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$

where the predictor variables $X_1, X_2, ...X_k$ can either be numeric or categorical and the response variable (i.e. variable of interest) is a binary categorical variable with a probability p of occurring. The predictor variables of interest for this particular statistical analysis are:

- X_1, age of the respondent (age)
- X_2, the respondent's self-rated health, if rated 'Excellent' (self_rated_health)
- X_3, the respondent's self-rated health, if rated 'Fair' (self_rated_health)
- X_4, the respondent's self-rated health, if rated 'Good' (self_rated_health)
- X_5, the respondent's self-rated health, if rated 'Poor' (self_rated_health)
- X_6, the respondent's self-rated health, if rated 'Very good' (self_rated_health)
- X_7, the respondent's self-rated mental health, quantified by a numerical rating of their "feelings about life as a whole" (feelings_life)
- X_8, income of the respondent's family (num_income_family)
- X_9, the resopndent's own income (num_income_respondent)

- X_{10}, number of marriages the respondent experienced (number_marriages)
- X_{11}, sex of the respondent (is_male)

Note that because self_rated_health is a categorical variable with multiple factor levels, there are 5 different dummy variables in the model associated with self_rated_health. For a given data observation, each dummy variable will have a value of 0 unless the respondent associated with that observation rated their health in the way that corresponds to the factor level associated with a certain dummy variable (in which case, the value of the dummy variable will be 1).

Setting up the model

The model will be set up using the *surveys* package in R. To further refine the model, a population correction will be implemented which accounts for the difference in size of the sample and the population within the study setting. Although the sample size is easy to find — it is the number of observations in the dataset — the size of the population is more difficult to pinpoint.

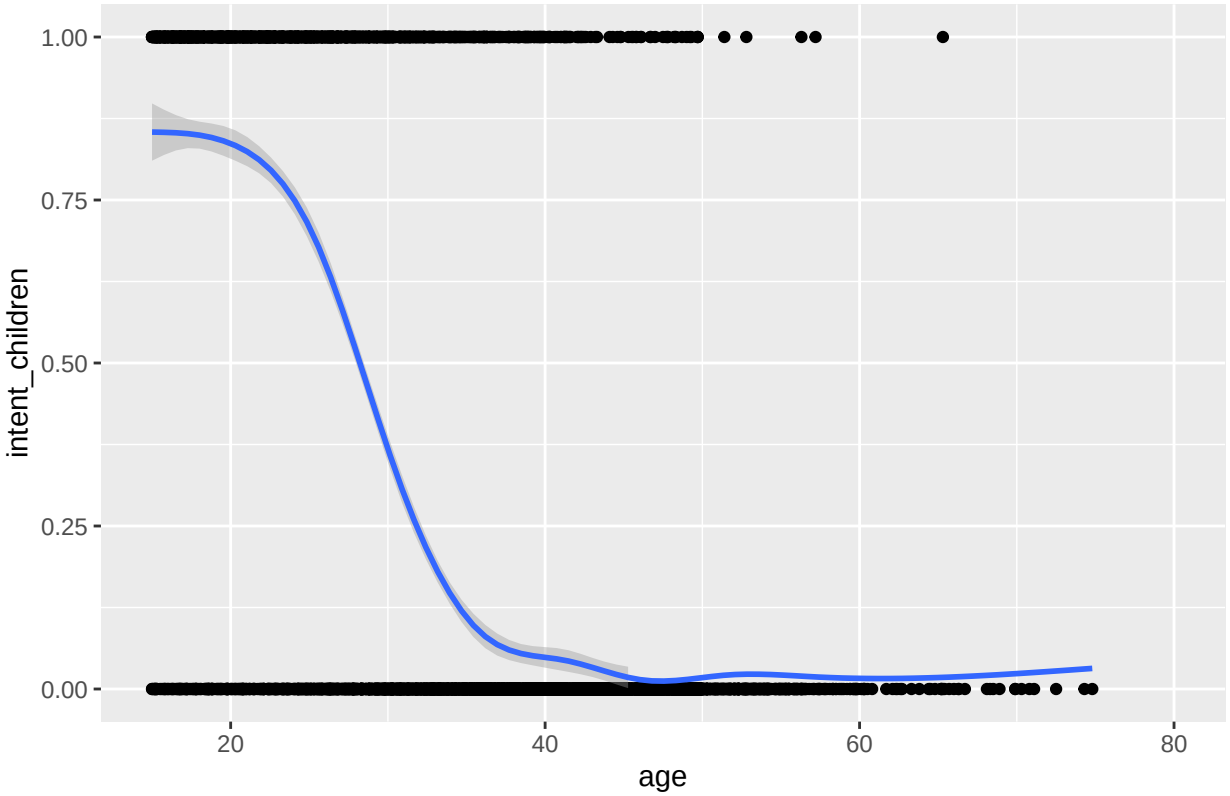
From Table 1 of Statistics Canada's *Canada at a Glance 2017 — Population*, in 2016 Canada's population was estimated to be 35,151,728. From Table 3 of this same source, the percentage of the population between 18 and 90 years of age is $9.2 + 27.1 + 27.8 + 16.5 - 0.8 = 79.8\%$. Based on this information, the population size used in this statistical analysis will be $35,151,728 \times 0.798 \approx 28,051,103$. One drawback to using this particular value of a population, or any population value for that matter, is that it is impossible to determine whether or not it is a good estimate of the true population of Canada. This is because such data is obtained using a census, which attempts to account for everyone in the population by having the entire population be the size of the respondent pool; however, it is difficult to know if every individual in the population actually responded to the census. Thus, the assumed population size can be a source of bias and in the event that it is not actually a good estimate of the true population size, it can skew the proposed model and make it less representative of population trends.

Results

Examining the significance of predictor variables to impacting the stated intent to have children

To check the model, whether the selected predictor variables in the model seem to have any relation with a resopndent's stated intent to have children needs to be evaluated. To do so, this report has compiled various bar and scatter plots, all with the intent to have children on the y-axis and each predictor variable on the x-axis. The predictor variable with the most visible relationship to a stated intent to have children, age, is displayed below; all other plots can be found in the Appendix.

Figure 1: Intent on having children based on respondents' ages



By fitting of a logistic model with the intent to have children as the response variable and age as the only predictor variable to the above scatter plot, an easily observable trend of a respondent being less likely to want to have children the older they are can be seen This aligns with popular thought; as people grow older, they are more likely to already have established families if they ever wanted to do so. Additionally, older people are unlikely to be able to have children for biological reasons.

Running the proposed model and and examining its significance

Figure 2: Summary statistics for the proposed model

predictor_variables	coef_estimate	standard_error	t_value	p_value
(Intercept)	2.9063167	1.7521522	1.6587124	0.0972
age	-0.1792433	0.0067563	-26.5297037	0.0000
as.factor(self_rated_health)Excellent	1.9833366	1.7527820	1.1315363	0.2579
as.factor(self_rated_health)Fair	1.4544170	1.7602304	0.8262651	0.4087
as.factor(self_rated_health)Good	1.5302064	1.7518440	0.8734833	0.3824
as.factor(self_rated_health)Poor	1.1558690	1.7768064	0.6505318	0.5154
as.factor(self_rated_health)Very good	1.9091175	1.7517831	1.0898138	0.2758
feelings_life	0.0444370	0.0288025	1.5428151	0.1229
num_income_family	-0.0000011	0.0000011	-0.9862227	0.3241
num_income_respondent	-0.0000060	0.0000018	-3.3963722	0.0007
number_marriages	-1.3570310	0.1125165	-12.0607342	0.0000
is_male	0.8638764	0.0822282	10.5058408	0.0000

From the estimates of the coefficients associated with each predictor variable given in the above table, we

see that the most significant predictor variables in determining how likely an individual is to intend to have children in the future are:

- age
- the individual's personal income
- the number of marriages an individual has experienced
- the respondent's identified gender

For each of these variables, the following conclusions can be drawn from the values of their associated coefficients:

- The older an individual is, the less likely they are to intend to have children in the future. For every additional year of age, an individual is on average 17.92% less likely to want to have children in the future.
- The higher an individual's personal income is, the less likely they are to intend to have children in the future. For every additional dollar of personal income an individual makes, they are on average 0.0059% less likely to want to have children in the future.
- The more marriages an individual has participated in, the less likely they are to intend to have children in the future. For every additional marriage an individual particpates in, that individual is 135.7% less likely to want to have children in the future.
- If a respondent identifies as male, they are 86.39% more likely to want to have children in the future.

Discussion

From the summary statistics, this analysis concludes that the predictor variables with the most significant impact on an individual's likelihood that they will want to have children in the future are their age, their personal income, the number of marriages they have had, and whether or not they identify as male. These findings support the qualification of the proposed model as a strong one which identifies relevant predictor variables that influence the response variable, otherwise known as the variable of interest.

The conclusion drawn from the model supports the initial goal of this statistical analysis to "[identify] factors related to an individual's cultural, socioeconomic, and biological background which affect their intent to have children and essentially start a family". The fact that the majority of proposed predictor variables in the model significantly influenced the probability that an individual would intend to have children indicates that the statistical analysis outlined in this report is a good starting point for continuing research on family formation. As stated previously, knowing the exact characteristics of an individual's background affect their desire for children can help inform the forecasting of population trends due to changes in birth rate or general population growth attributed to the creation of family units. Demographic study is especially important to various disciplines, not limited to but including public policy, economics, anthropology, and urban development.

Weaknesses

Many of the weaknesses in the above outlined statistical analysis can be attributed to the dataset that was used. As previously stated in the "Data" section of this report, several parts of the methodology used in collecting data for some variables such as income greatly affected the data's accuracy and ability to be representative of the target population of all Canadian residents. As a result, the statistical analysis conducted has multiple sources of bias, including:

- Non-response bias attributable to the fact that a respondent's age limited the questions they were eligible to answer in the survey that produced this GSS dataset.

- The use of the midpoint approach to assign concrete numerical values to each income range for the income_family and income_respondent variables.

The quantitative magnitude of the effects of these various sources of bias are not easily measurable, as we do not have sufficient data to calculate an unbiased estimator of any of the regression model coefficient estimates. In order to calculate unbiased estimates, we would have needed to run prior tests on this dataset to ensure that the following conditions are met:

- There are visibly significant relationships between the response variable and each of the proposed predictor variables in the model.
- Residual (error) values in the estimation of each coefficient in the proposed model are independent of each toher.
- Residual values in the estimation of each coefficient in the proposed model have constant variance.
- Residual values are normally distributed.

For this statistical analysis, because so much of the data was categorical and not quantitative, and even among categorical variables there was not an easy way to translate the factor levels of each one into a numeric scale, it would have been difficult to run the traditional residual plot tests to confirm that the above four assumptions about the data are true.

Next Steps

Mitigating existing weaknesses with this research

As previously stated, the primary weaknesses of the statistical analysis undertaken for this report lie in the composition and collection methods of the initial dataset. If this statistical analysis were to be repeated to further refine the proposed model, one way to mitigate the effects that these weaknesses have on the certainty with which conclusions can be drawn from quantitative evaluation of the model's strength would be to re-issue surveys with questions that all participants can answer. This can be done in place of implementing a skipping mechanism for particular questions if participants were not eligible to answer them due to their previous answers (e.g. due to their calculated age). Doing so would likely result in a cleaner dataset with a set of answers that is representative of the study's entire sampling frame rather than just a small subsection of it.

Another way to mitigate the effects of unquantifiable categorical data is to re-issue surveys to the same participants whose responses were recorded in the 2017 GSS on families with a signficant portion of the questionnaire rephrased to allow answers to be given on a numeric scale, similar to how the variable feelings_life in the original survey was coded. Alternatively, more Likert scales could be used with qualitative answers that can easily be translated into numeric scales in data cleaning and manipulation. Both of these methods would help with generating quantifiable responses that can be easily translated into numeric scales and used for data analysis.

Use of this research in prompting additional questions

This report specifically targeted the identification of factors that would increase the likelihood that an individual would want to start a family by having children. Now that the proposed model has provided examples of significant personal factors that affect the likelihood of an individual wanting to have children, this research can be expanded by looking at the opposite side of the proverbial coin: what personal factors would prompt someone to want to separate from their family?

Right from the start, there are a few distinct groups of people who would be likely to want to leave the family group(s) they are a part of, including:

- Younger people who are just starting to live "adult" lives by pursuing post-secondary education, searching for jobs, or leaving their country of birth.
- People who are unsatisfied wtih their current marriages.
- People who experience domestic conflict within their current family units.

There may be more as additional research on this question is conducted; however, any research project should start with using these initially described groups and later examining each group to find trends that suggest the existence of different possible classifications of respondents. The same dataset used in the above statistical analysis, the 2017 Canadian GSS on families, can be used to propose a statistical model that can potentially answer this new research question.

References

Government of Canada, S. C. (2017, March 31). Canada at a Glance 2017 — Population. https://www150.statcan.gc.ca/n1/pub/12-581-x/2017000/pop-eng.htm.

Merriam-Webster. (n.d.). Family. In Merriam-Webster.com dictionary. Retrieved from https://www.merriam-webster.com/dictionary/family.

Statistics Canada. (2017). 2017 General Social Survey (GSS): Families Cycle 31 [Data file]. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31.

Appendix

A. Bar graphs of respondents' intent to have children based on their identified sex

The figures below were constructed using the intent_children dummy variable on the y-axis. The is_male dummy variable was used to sort the dataset into two groups: one with all male-identifying responses and another with all female-identifying responses.

Figure 3: Male−identifying respondents' intent to have children in the future

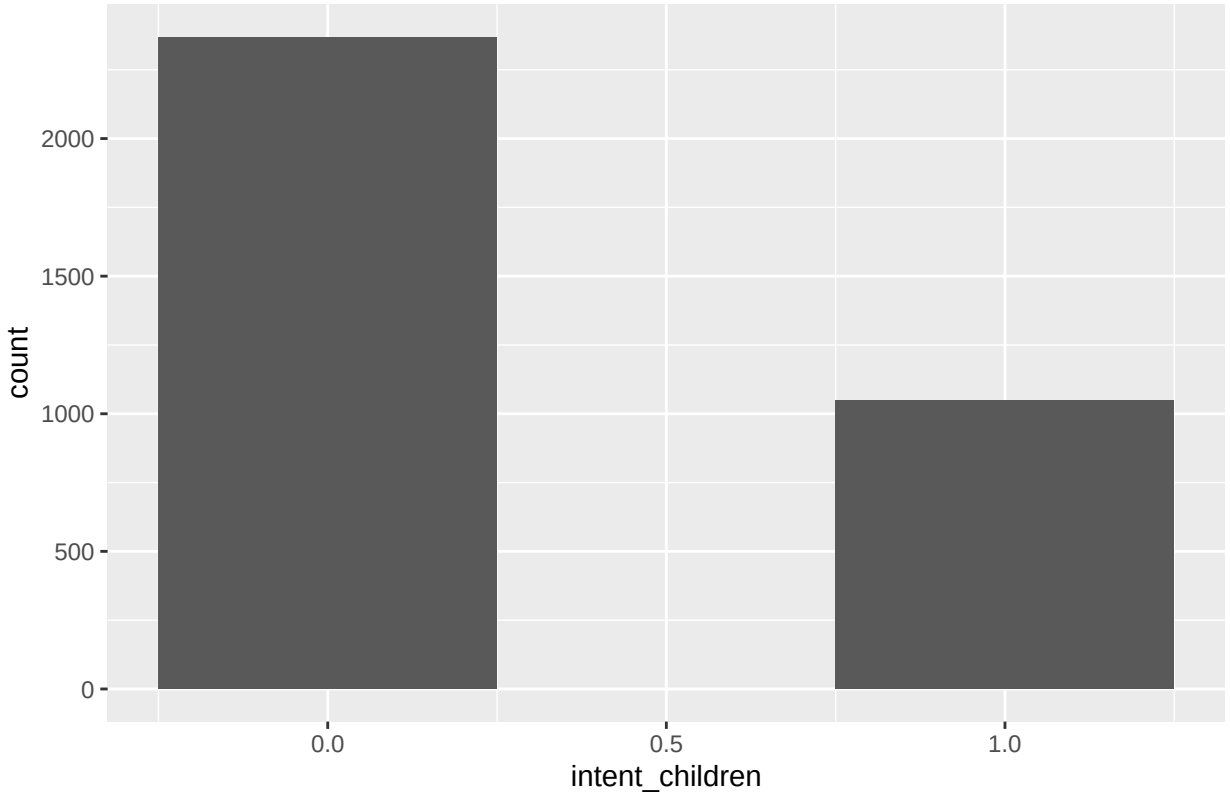
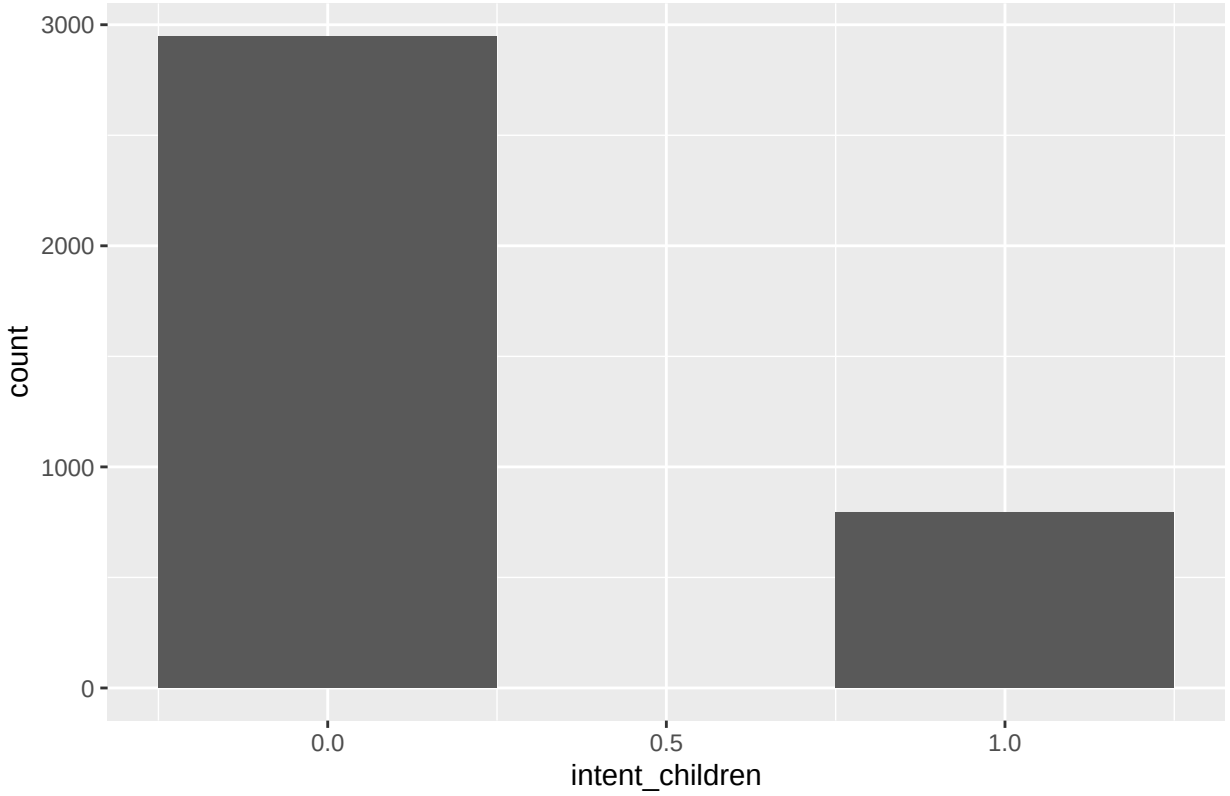


Figure 4: Female−identifying respondents' intent to have children in the futur



Something interesting to note is that according to the above two bar graphs, a larger proportion of male-

identifying respondents in the sample stated that they intended to have children in the future than female-identifying respondents. This finding goes against traditional schools of thought that female-identifying individuals are more likely to want to start families.

B. Bar graphs of respondents' intent to have children based on indivdual and family income levels

Figure 5: Intent on having children among respondents with $125,000 or more in income

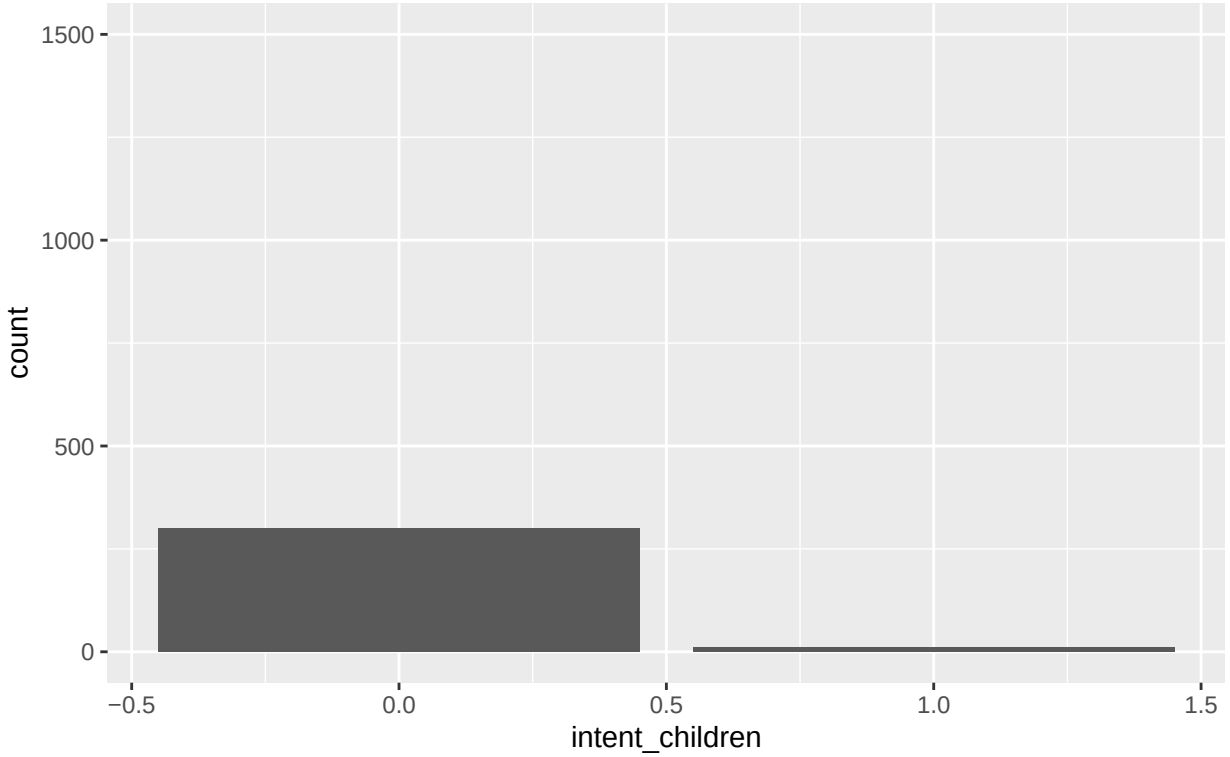
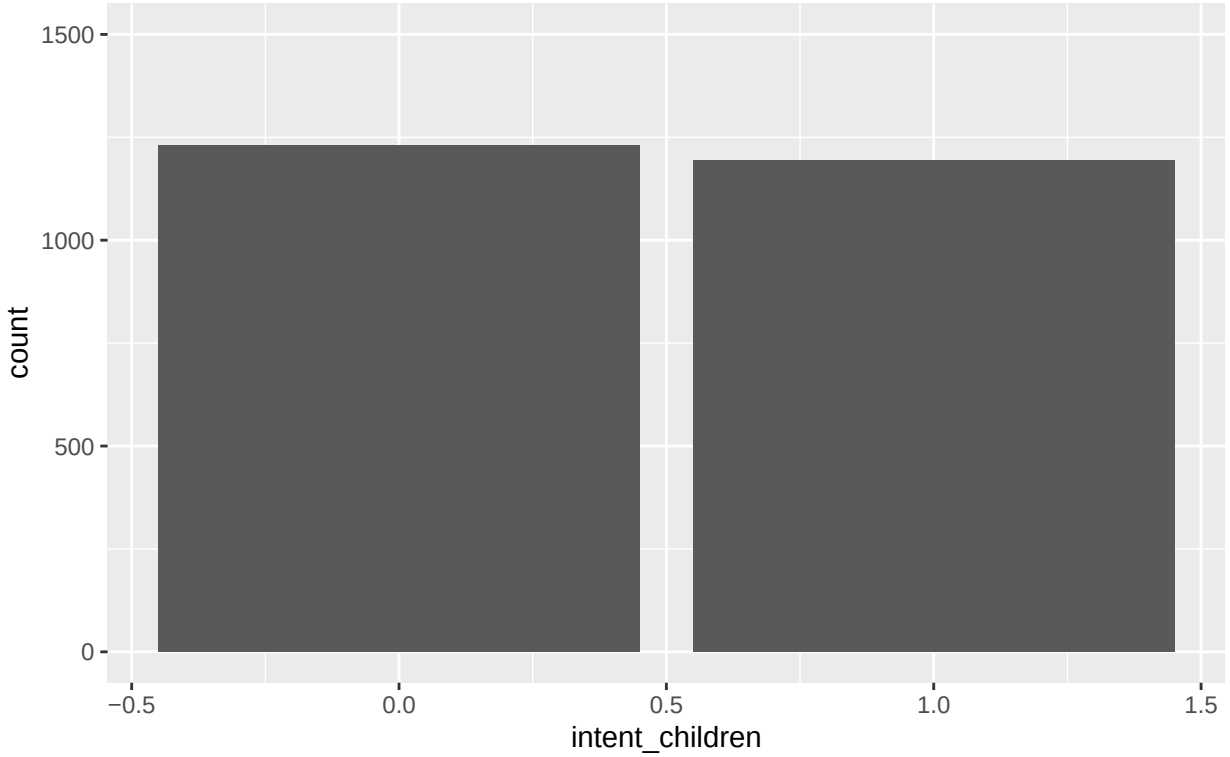


Figure 6: Intent on having children among respondents with less that $25,000 in income



From these graphs, we see that in the group of respondents with less than $25,000 in personal income, there is a much larger proportion of respondents who intended to have children than in the group of respondents with $125,000 or more in personal income.

The same phenomenon can be seen in similarly-made bar graphs for respondents with family income less than $25,000 and for respondents with family income totalling $125,000 or more. Observe below:

Figure 7: Intent on having children among respondents with $125,000+ family income

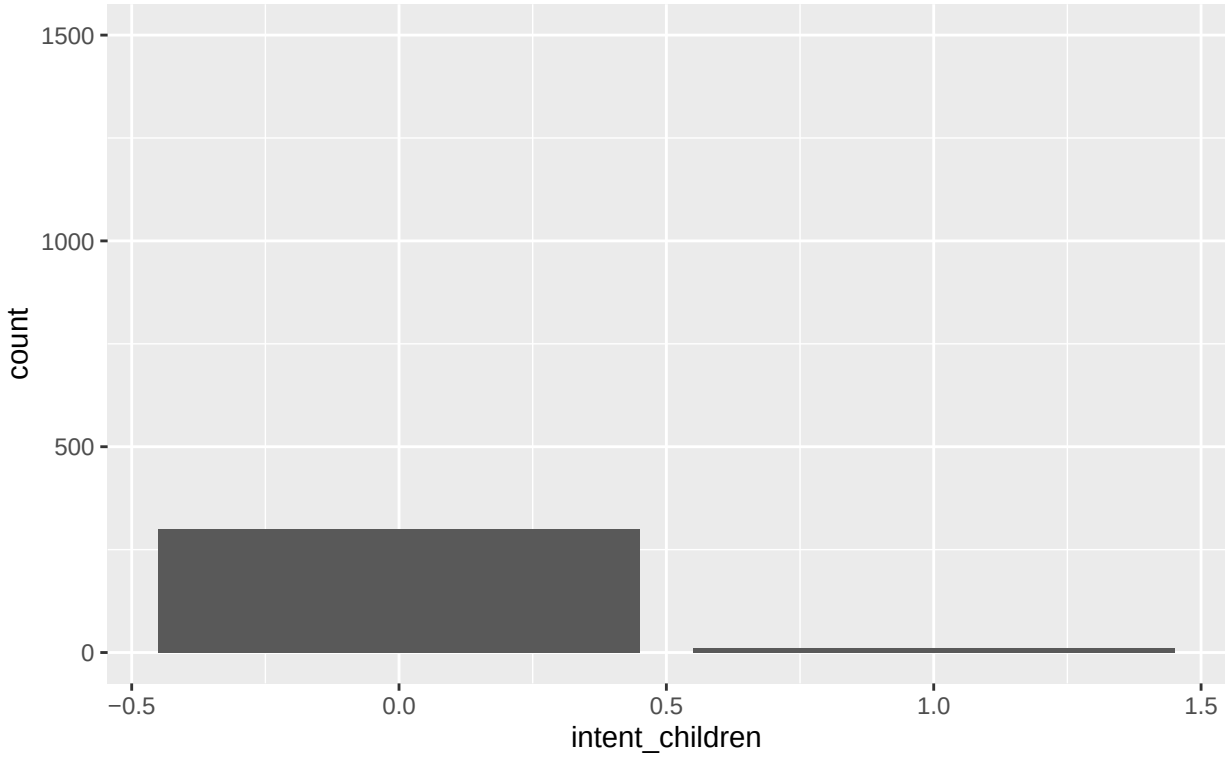
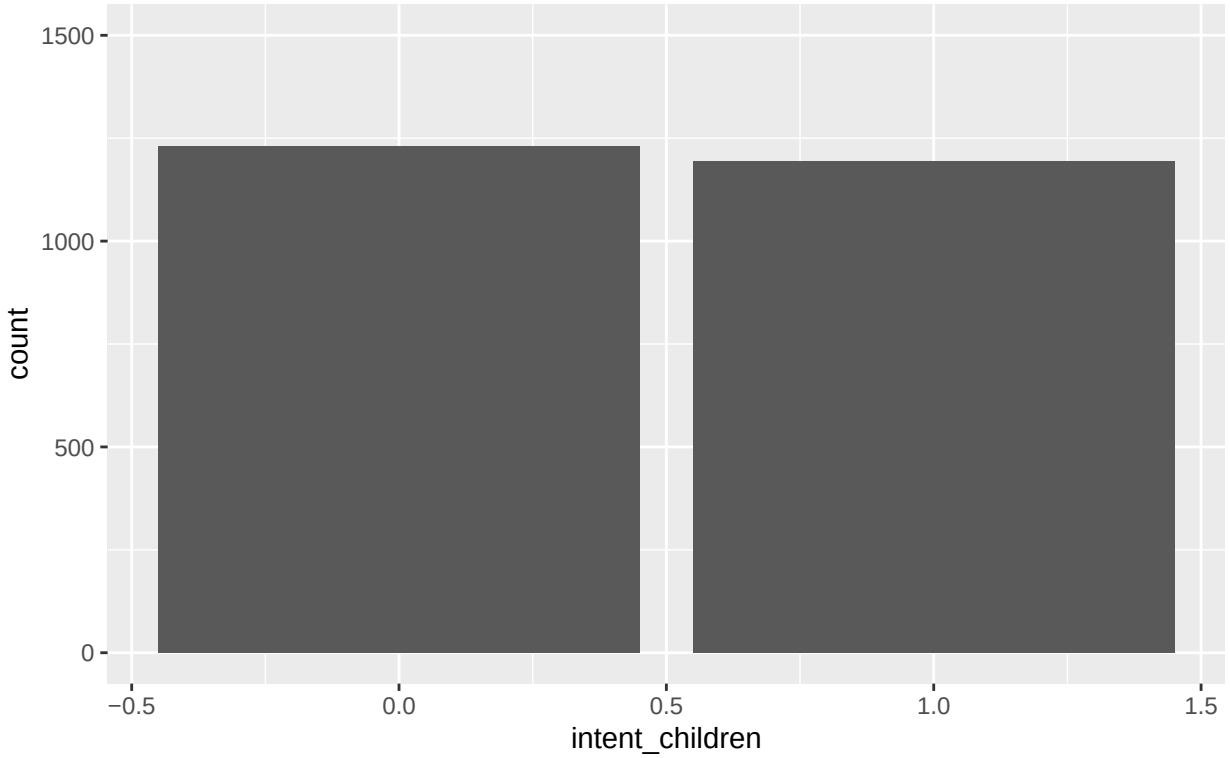
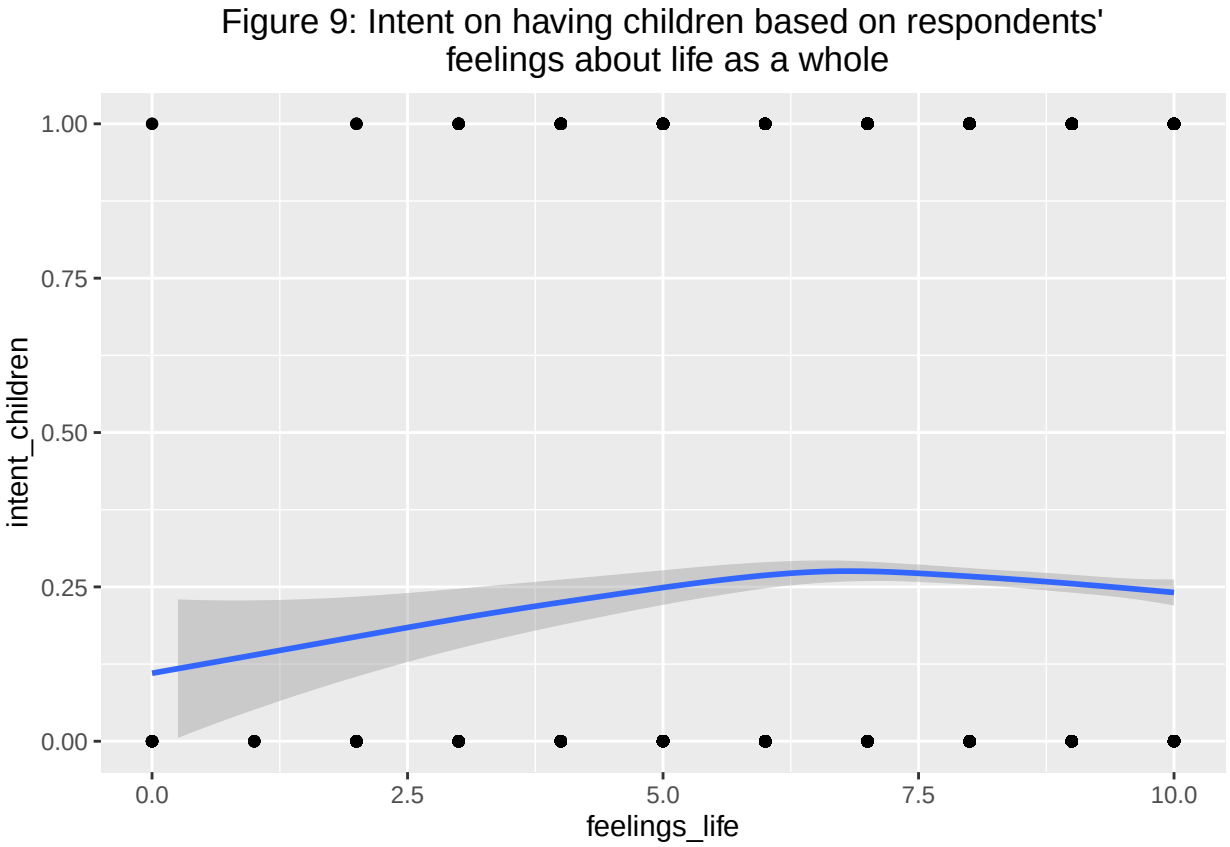


Figure 8: Intent on having children among respondents with less that $25,000 in family income

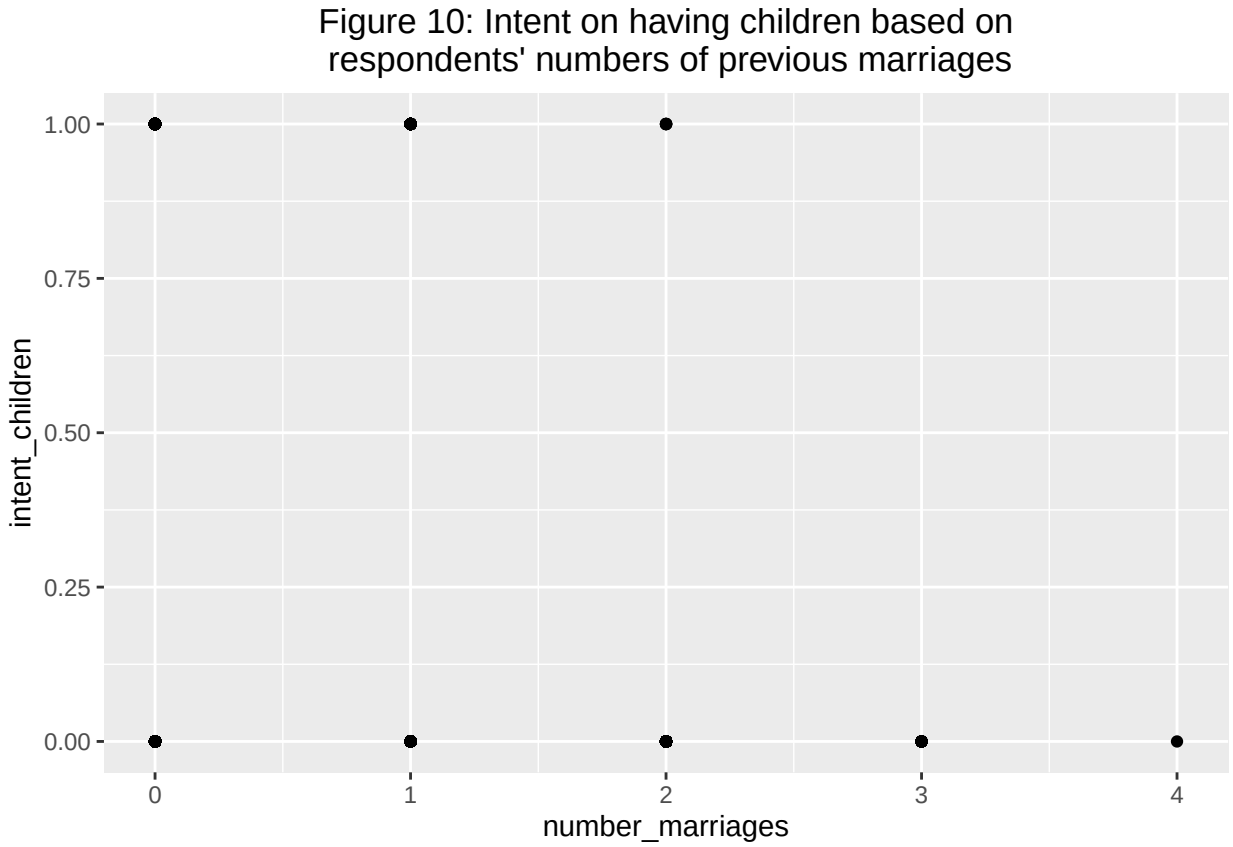


C. Scatter plot of respondents' intent to have children based on self-assessed mental health



The logistic regression model of the stated intent to have children on "feelings on life as a whole", a quantitative approximation of a respondent's self-rated mental health, suggests that as a respondent feels "better" and is more satisfied with their life as a whole they are more likely to want to have children in the future.

D. Scatter plot of respondents' intent to have children based on number of marriages including their current one, if it exists



From this scatter plot, it appears that respondents who have gone through marriages are less likely to intend to have children. This is supported by the fact that while having had no marriages, 1 marriage, or 2 marriages produced scatter plot points representing both the intent to have and to not have children, there are no scatter plot points representing respondents who have gone through 3 or 4+ marriages and intended to have children. It is possible that this is the case because people who have gone through multiple marriages tend to be older, making age a confounding variable in this situation. It is also possible that people who have gone through multiple marriages are also more likely to already have children from previous marriages, therefore minimizing the desire for more children. It is uncertain whether the apparent relationship between number of previous marriages and the intent to have children is due to a genuine direct correlation between the two, or to a confounding variable producing a correlation where none exists.