Investigation of Family related Factors that Have Impact on Life Satisfaction in Canada

Jianzhong You 1003042628, Ziyue Xu 1002924026

2020/10/19

Abstract

The Bayesian logistic regression approach is first utilized to draw conclusion about the association between the predictors (number of children, family income, and living arrangement) of interest and the dichotomized version of the response variable, family life satisfaction. However, from the diagnostic plots, we see that the logistic regression might not be the best fit for this particular data set. Thus, we eventually decides to use the Chi-Square test of Independence approach to perform a more detail analysis and draw the final conclusion the association. Indeed, the each predictors mentioned above has a positive association with one's life satisfaction.

Keywords: Family related factors, Life satisfaction, Stratified Random Sampling, Bayesian inference, Frequentist inference, Chi-Square test of Independence, Credible Intervals, Logistic Regression.

Introduction

In this paper, the data collected by 2017 General Social Survey (GSS) on the Family is used to analyze the family related factors that have impact on one's whole life satisfaction in Canada. The 2017 GSS survey focuses on family since the family is of significant importance in one's life and the data about family could reflect Canadians' living conditions and well-being. To be specific, the family-related factors, such as number of children, family income, living arrangement considerably influence one's life satisfaction. How and to what extent does each factor impact one's whole life satisfaction will be investigated and discussed in this paper based on statistical data analysis of 2017 GSS survey data.

Since survey is a form of observational study, we could not draw any casual-and-effect conclusion from our analysis, we could only analyze the association between the variables[8]. We first uses the Bayesian approach to perform a logistic regression model on each of the interested predictors, we realize from the diagnostic plots that logistic regression might not be the most ideal approach to analyze this particular data set as it comprise mostly categorical variables. Thus, we use the Chi-Square test of Independence to draw final conclusion about the association between number of children vs. family life satisfaction, living arrangement vs. family life satisfaction, and finally, family income vs. family life satisfaction. And we will soon see that the positive association for each pair does exist.

Data

The information of data and survey questionnaire is in the User's Guide and Codebook in the documentation of the 2017 General Social Survey(GSS): Family Cycle 31[1]. All description of data and survey questionnaires below uses the information in the documentation and is paraphrased.

Data source

The data set is the General Social Survey(GSS) 2017 downloaded as a csv file from CHASS[2]. Then the data is cleaned and processed by using the code provided by Rohan Alexander and Sam Caetano[3].

Methodology: Sampling strategy

1. First do stratification and the population is stratified as following for carrying out the survey, and the stratification is based on the geographic areas:
    - 14 Census Metropolitan Areas (CMAs) were each considered separate strata, including Montreal, Quebec City, Toronto, Ottawa, etc.
    - 3 more strata were formed by grouping the remaining CMAs (except Moncton, which is included in the non-CMA stratum for New Brunswick) in each of Quebec, Ontario and British Columbia.
    - Finally, the non-CMA areas of each of the ten provinces were also grouped to form 10 more strata

    So there are 14 + 3 + 10 = 27 strata in total.
2. Then each of the record in the sampling frame was assigned to a stratum respectively, and a simple random sampling without replacement of records is performed within each stratum.
3. Then a respondent was then select from each sampled household (record), by using random sampling without replacement, to participate in a telephone interview.

Data collection and processing approach

Data for the 2017 GSS was collected via computer assisted telephone interviews (CATI). Centralized telephone facilities in five of Statistics Canada's regional offices are used to perform the interview. Then the responses were entered directly into computers as the interview goes. For the questions allow for write-in responses, the responses were coded into existing categories when a match was possible. If a match was not possible, new categories will be created or left in "other-specify" if the responses' frequencies were too small. When the data was entered into CATI, any "out-of-range" values will be identified by CATI, some of the problems can be solved by the interviewer instantly with the respondent, and the other problems will be forward to Head Office for determination. Then the final output data was sent to Ottawa electrically. By reasoning, the cost mainly consists of payment to the interviewers (including the training), the fee for telephone calls and network, and the fee for devices maintenance.

The population, the frame, and the sample

The Target Population: All people of age 15 and above (15 years of age or older) who lived in Canada excluding the full-time residents of institutions and the residents of the Yukon, Northwest Territories, and the Nunavut.

Sampling frame: The survey sampling frame was created using two different components: 1. Lists of telephone numbers in use available to Statistics Canada from various sources; 2. The Address Register (AR), which is the list of all residences within the ten provinces. Then the AR is used to group together all telephone numbers associated with the same valid address. For the telephone numbers that cannot be linked to the AR are either grouped using address information from administrative sources or treated as single "record". A "record" is a grouping of telephone numbers that consists of sampling unit.

Sample: After mapping each record to a stratum, a simple random sampling without replacement of records is performed within each stratum. Then one respondent from each record was selected by using simple random sampling without replacement method as well. The target sample size (number of respondents) was 20000 while the actual sample size was 20602. The sample size in each stratum is determined by the population within each stratum to ensure that bias and sampling variability are bearable.

Find respondents and handle non-response

First use the telephone numbers in the sampled records to reach the households. Then a respondent is randomly selected from each household to take this interview. For those who at first refused to participate, the interviewer will contact them up to two more times to describe the importance of this survey and encourage them to take it. For the respondents who are inconvenient to do an interview when contacted, the interviewer will arranged an appointment to do this interview at a convenient time. For cases that no one at home, the interviewer will make phone calls numerous times. As a result, the overall response rate for the 2017 GSS was 52.4%.

Survey strengths and weaknesses

The strengths of this survey are:

1. The conjugal history questions of this questionnaire not only collected the information of respondents' histories of relationships and marital status, but also whether or not children were born during each relationship. "These data allow for rich historical analyses which are not possible using other sources." as claimed in the User's Guide in the documentation of the 2017 General Social Survey(GSS): Family Cycle 31[1].
2. The flow of this questionnaire's questions is well-designed. For example, a question asking for the respondent's age is asked in the beginning of survey. The benefits are the interviewer can confirm the respondent is a valid (15 years old or above) and this age information could be used to validate other responses involving age information such as the conjugal history.
3. There is an improvement in this 2017 survey questionnaire that the income information is collected through a linkage to tax data instead of asking a question to the respondents as previous GSS questionnaires. This improvement made the income data more accurate and saved respondents' time for recalling or looking up.
4. All questions include "Valid Skip; Don't know; Refused; Not stated" options, which cover the whole spectrum of possible answers and the respondents can have a response to each question in any case.

The weaknesses of this survey are:

1. The range of the rating scale questions is too broad: "Using a scale of 0 to 10 where 0 means"Very dissatisfied" and 10 means "Very satisfied", how do you feel about your life as a whole right now?". This scale consists of 11 values and the respondents may don't have a clear perception of the difference between two consecutive scales. For example, a respondent understand that 6 to 10 represents positive feelings but he or she may not be able to tell the difference between 7 and 8. A suggestion is that the scale can be collapsed into 5 scales of 1 to 5, where 1 means"Very dissatisfied", 2 means"Dissatisfied", 3 means"Neutral", 4 means"Satisfied" and 5 means "Very satisfied".
2. Although the data regarding to family collected by this questionnaire is very comprehensive and detailed, the survey is relatively long and takes a lot of time and patience for a respondent to complete. As a result, the the responses' accuracy of the later questions might be lower than the former question.
3. Some of the questions such as "What is the main reason why you intend to continue living in a common-law/Blank relationship rather than getting married to your current partner?" are too private and the respondents are apt to conceal their genuine reason and make some responses that sounds better.

Dataset visualization

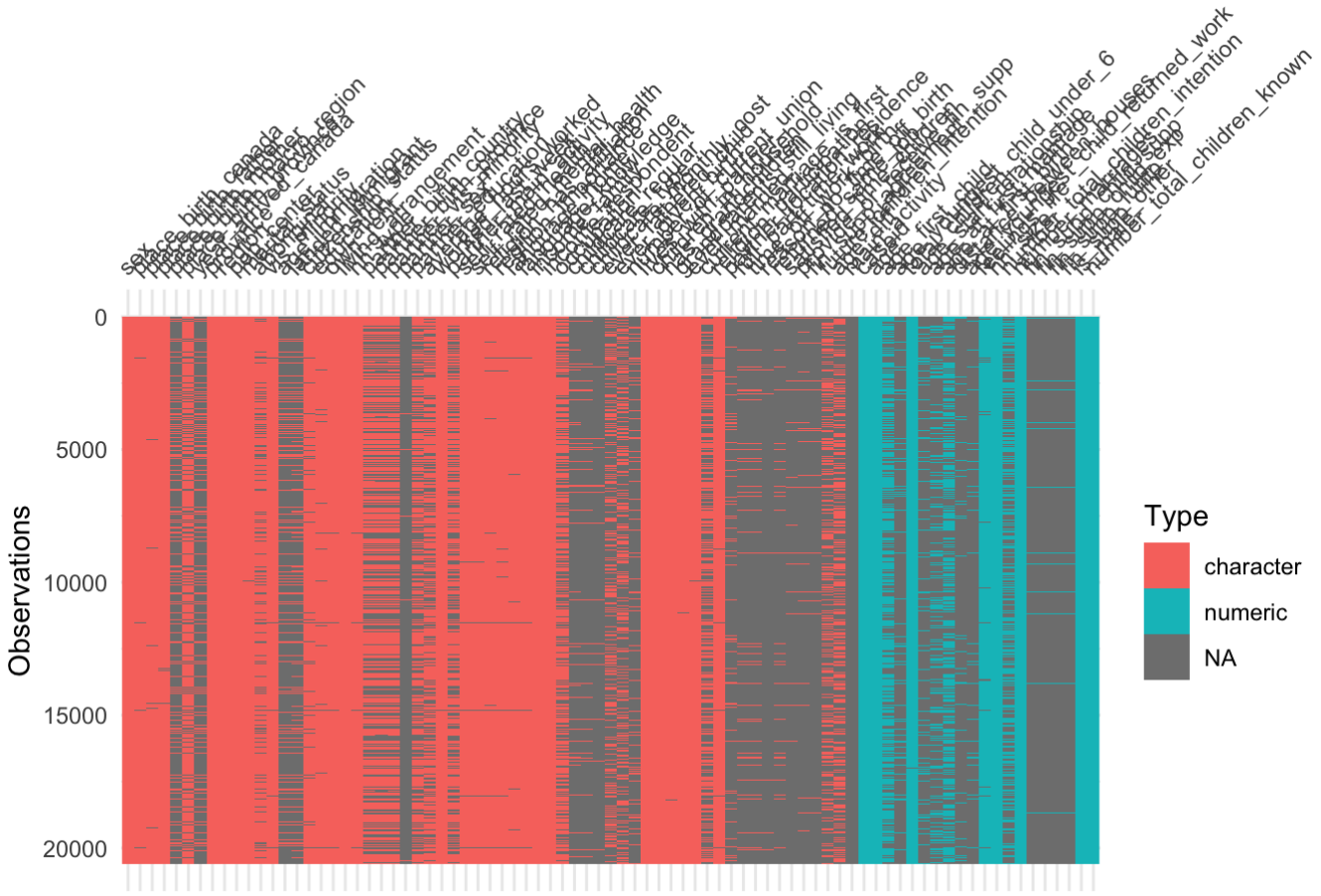

Figure1: Visualization of the raw GSS data

Variables discussion

From Figure 1, the key features of the data are:

1. There is a large amount of variables (81) in the dataset and the data consists of some numeric variables but mainly categorical variables.
2. There are many missing values in the dataset. Note that most of the missing should exist since many questions are valid skip for some respondents.

Since some of the predictors have large amount of missing values and when we want to model the response variable, family satisfaction, with multiple predictors, it required us to find out the rows that does not have missing values in each of the desired predictor we want to model. Thus, in some cases, we have a smaller set of observations than the original one (which has more than 20,000 observations). So, the small drawbacks of this dataset are the complexity of this dataset, the data cleaning is needed for NA values and as a result we have a smaller set of observations.

Model

We employ two approaches to analyze this data set, In the first approach, we use Bayesian logistic regression to fit various predictor that we are interested in; to do that, we need to abstract away the details of the response variable *family satisfaction about life* by collapsing the number of categories from 11 (scale for the degree satisfactions of life) to 2 (either not satisfy about life or satisfy about life) and see if there are any association

between the desired predictor(s). We will subsequently describe the fallback of this approach and employ a more appropriate statistical method, known as Chi-Square test of Independence, to draw conclusion for this particular data set.

Before we articulate in details about the models we are using for the Bayesian approach, we first need to discuss the difference between the two statistical inferences, the Frequentist approach and the Bayesian approach. In Frequentist inference, any conclusion drawn from it rely on the assumption that the same experiment is repeated infinite number of times and thus any probabilities estimated from sample data only hold true in a long run. In contrast, from a Bayesian inference perspective, repeated experiment under the exact same setting is not required in order to define a probability, the definition of probability is merely a number between 0 and 1 inclusively to express our uncertainty toward the variable of interest, which match our intuition of probability nicely. Thus we use the Bayesian inference to do the analysis for the following reasons:

1. The Bayesian approach allows us to go from the effect (the data we have) back to the causes (the parameters of interest) with a degree of uncertainty[5]
2. Since the survey of this focus most likely only perform once instead of doing it repeatedly for infinite many times, Bayesian approach makes more logical sense in this setting.
3. if we have our prior belief of the possible underlying distribution, Bayesian is the approach that can incorporate our belief into the model and the parameters would subsequently adjust our beliefs based on the provided data according to the following property.

$$P(\theta|data) \propto P(data|\theta)P(\theta)$$

The term $P(\theta|data)$ is the posterior distribution, whereas the term $P(data|\theta)$ is the likelihood (given the parameters, what is the probability that we see all these observations) and lastly the term $P(\theta)$ is our prior.

4. the statistical results such as the estimate and the credible interval (defined below) are easier to interpret, we do not need to repeatedly perform the exact same experiment to validly interpret the statistics[6]

We are using the 95% credible interval and its interpretation has a subtle difference between that of confidence interval. if the credible interval has lower bound of x and upper bound of y, we interpret it merely as there is 95% of chance that the estimate fall within this range

For all the Bayesian modeling, we use the Rstudio along with the brms R package to help us fit the model using the data, we subsequently using the package functions mcmc_plot to check convergence and credible interval of each parameters in each model.

For the first model we are fitting, we use Bayesian approach to fit the logistic model $\log \frac{p}{1-p} = \beta_{0c} + \beta_{1c}x_c$, where x_c is the predictor (total number of children), β_{0c} is the intercept term, β_{1c} is the coefficient of predictor, and p is the probability that the family feel satisfied for the given number of children. For this model, based on the observations of the data set as well as the intuition behind having more than one child, we use a informative prior normal distribution $N(1, 1)$ since we believe that there is a positive association between number of kids and the family satisfaction about life (if a family hate having child, they should never having more than one at the first place). and let posterior distribution extrapolate information from the data and correct our belief, according to the above property. In this Bayesian model, the prior is $P(\beta_{0c}, \beta_{1c})$, where β_{0c} and β_{1c} are parameters to be estimated.

The second model we are fitting is the Bayesian approach to fit the logistic model $\log \frac{p}{1-p} = \beta_{0l} + \beta_{1l}x_l$, where x_l is the predictor (living arrangement), β_{0l} is the intercept term, β_{1l} is the coefficient of predictor, and p is the probability that the family feel satisfied given the living arrangement. For this model, we use a uninformative prior

$N(0, 1)$ since we do not notice any obvious pattern in the data. If our believe is incorrect, the evidence should adjust itself properly. In this Bayesian model, the prior is $P(\beta_{0l}, \beta_{1l})$, where β_{0l} and β_{1l} are parameters to be estimated.

The third model we are fitting is the Bayesian approach to fit the logistic model $\log \frac{p}{1-p} = \beta_{0i} + \beta_{1i} x_i$, where x_i is the predictor (family income), β_{0i} is the intercept term, β_{1i} is the coefficient of predictor, and p is the probability that the family feel satisfied given its income level. For this model, we also use a uninformative prior $N(0, 1)$ since we do not notice any obvious pattern in the data. Again, if our believe is incorrect, the evidence should be able to adjust itself. In this Bayesian model, the prior is $P(\beta_{0i}, \beta_{1i})$, where β_{0i} and β_{1i} are parameters to be estimated.

There are at least two drawbacks of using this approach to analyze this particular data set. We made some informal diagnostics for model checking and from Figure A, Figure B, and Figure C in the Appendix section, it is obvious that Figure B (the second model above) and C (the third model above) violate the linear relation between the logit term and the fitted linear model[7]. Thus, logistic regression might not be the best statistical model to analyze this particular data for some predictors. Second, for the above approach to work, we are force to abstract away the detail of the response variable into only two levels, and we want to analyze more than two levels. For those reasons, it suggested that we should use another approach to further analyze the data, called Chi-Square Test of Independence, to test the independence between categorical variables of interest.

The following equation is used to check the linear relationship assumption of logistic regression:

$$\log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1} \beta_i x_i$$

where the term $\log(\frac{p}{1-p})$ is known as the logit term and the term $\beta_0 + \sum_{i=1} \beta_i x_i$ is the fitted model of the predictors of interest.

Recall from the previous approach, we collapse the response variable to only 2 levels (satisfied or dissatisfied) due to the constraints of the logistics regression model. In the following approach, since we are using the Chi-Square Test of Independence to examining the possible association between two variables, to make analysis a bit more interesting, we relax the abstraction of the response variable to include three levels, that is dissatisfy (scale from 0-3), neutral (scale from 4-6), and satisfy (scale from 7-10). There is a logical reason behind this adjustment. We consider that the survey takers might not have a rigid scale in their mind when they takes the survey, that is, when they answer the satisfaction with a 7, it might just be 8 or 9, we consider the difference is negligible, thus we group them together.

Chi-Square test of Independence is a non-parametric tool of analysis, that is, it does not assume the distribution of the underlying data, but one of the critical assumption it has is that the expected frequency of each cell in the table should be> 5[4] at least 80% of the time

In particular, the Chi-Square statistics is calculated as

$$\sum \chi^2_{ij} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where χ^2_{ij} is the Chi-Square statistics of each cell ij, O_{ij} is the observed frequency value of cell ij in the table, and E_{ij} is the expected value (calculated as follow) of the cell ij in the table, and i, j range from 1 to number of levels of its respective categorical variable.

One thing to note is that the term $r = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$ is known as the residual term and will be used and discussed in the Discussion section.

In the Chi-Square test of Independence, the expected value represent the estimate of cases be distributed if there is no association between the two categorical variables being tested, also known as the Null hypothesis of the Chi-Square test of Independence.

$$E_{ij} = \frac{R_i \times C_j}{n}$$

Where R_i is the sum of row i in the table, C_j is the column sum of column j, and n is the grand total number of observations

For this analysis, we use the same predictors as the previous approach, that is, association between number of children and Family Life Satisfaction, association between living arrangement and Family Life Satisfaction, and association between income family and Family Life Satisfaction.

From Table C, Table D, and Table E in Appendix section, we see that the expected count of each cell in each table is > 5, thus the assumption is satisfy and we can perform the Chi-Square Test of Independence. Note that for each categorical variable, we group its levels in a sensible way so that the expected count assumption is satisfied.

Results

Descriptive Statistics for Bayesian Logistic Regression Model

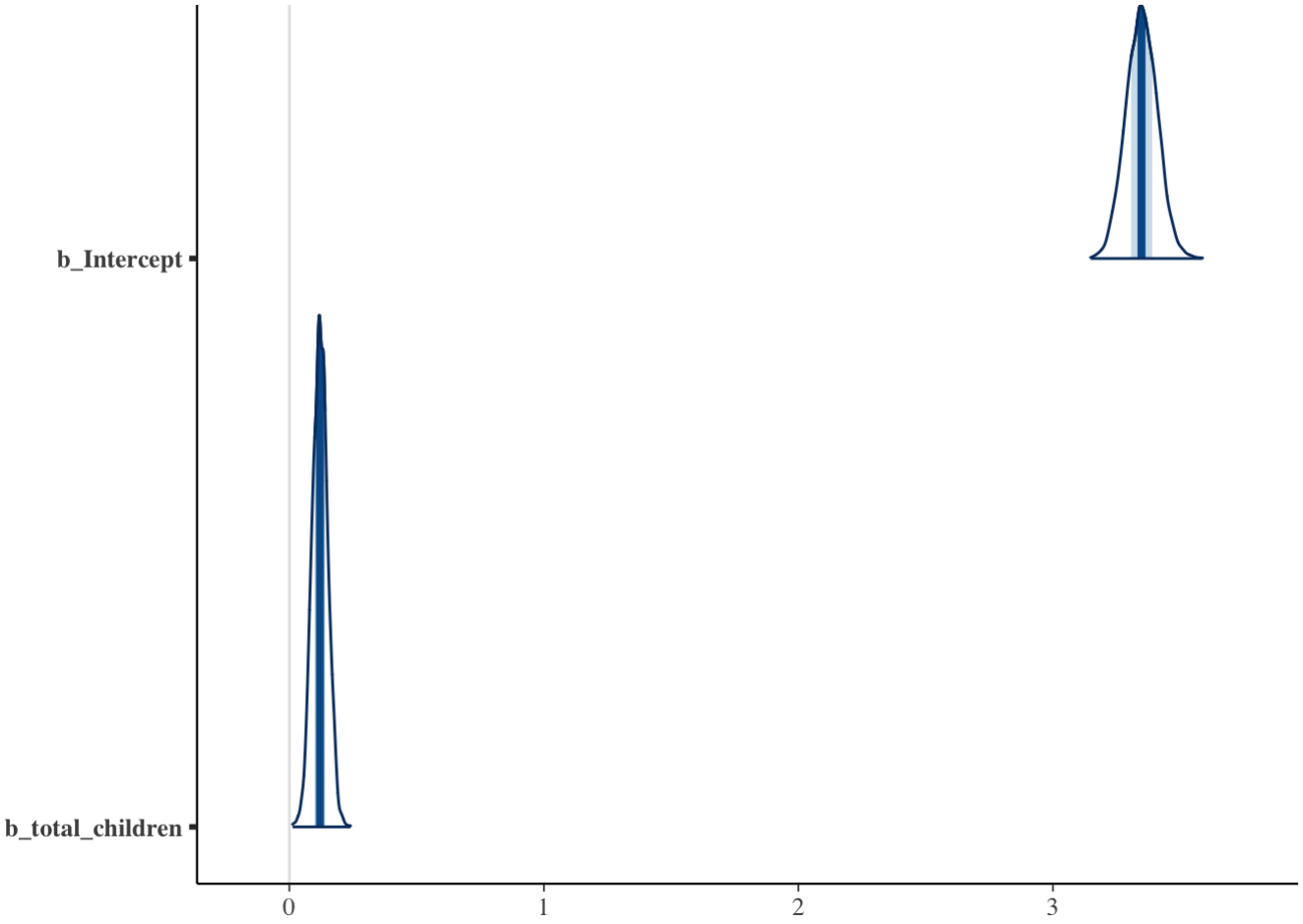

Figure 2: Distributions of Parameters - Family Life Satisfaction vs. Number of Children

This plot shows the distribution of the predictor *total_children* and the intercept.

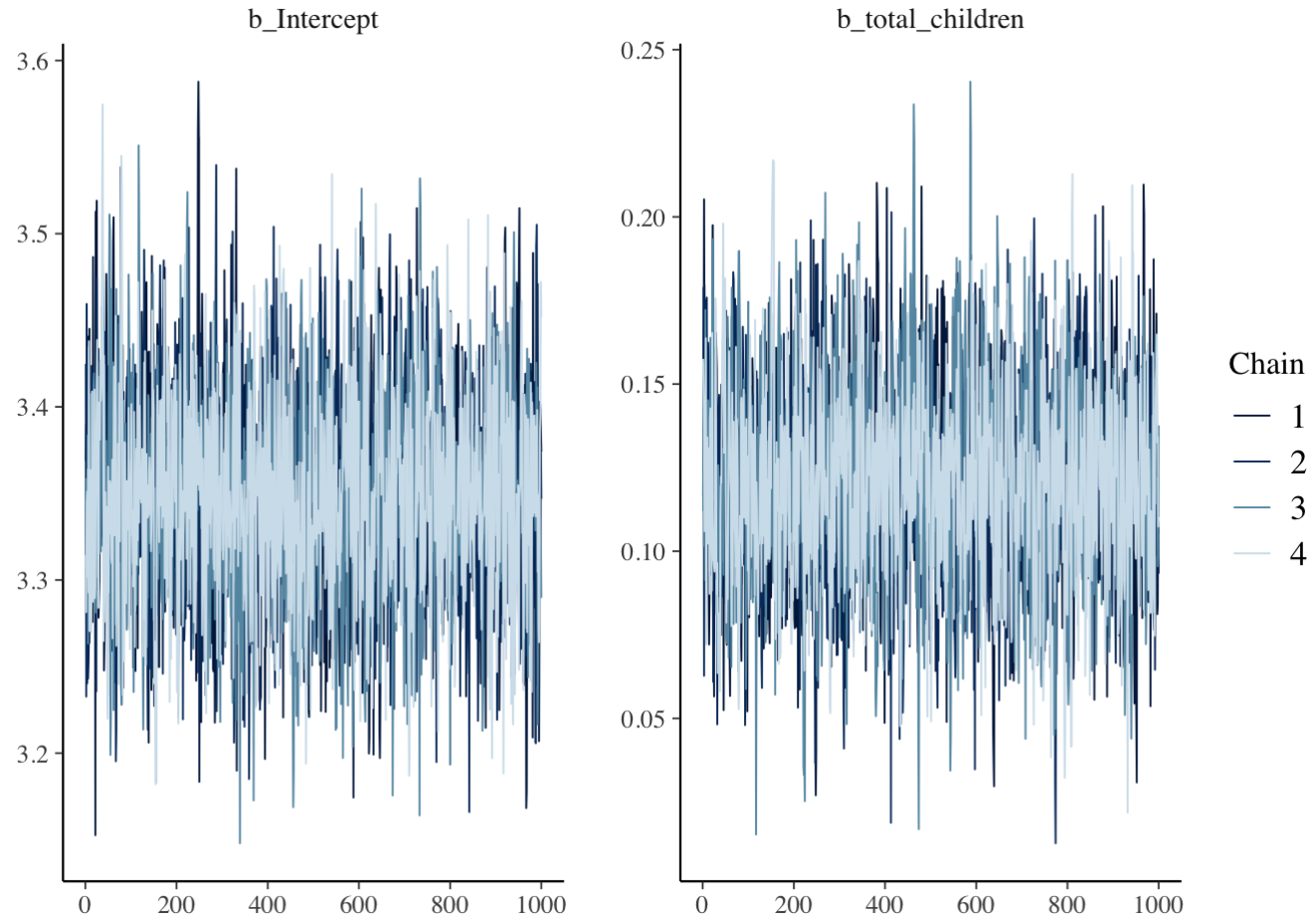

Figure 3: Convergence of Parameters - Family Life Satisfaction vs. Number of Children

This plot shows the convergence of the predictor *total_children* and the intercept.

Table 1: The statistic table for Family Satisfaction vs. Number of Children

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	3.3484795	0.0612448	3.228804	3.4695206
total_children	0.1204836	0.0305314	0.062656	0.1801069

This table shows that the association between number of children and the Family Life Satisfaction is 0.121 and affirm above statement that the 95% credible interval range lie above 0 with lower bound of 0.0611357 and upper bound of 0.1802879

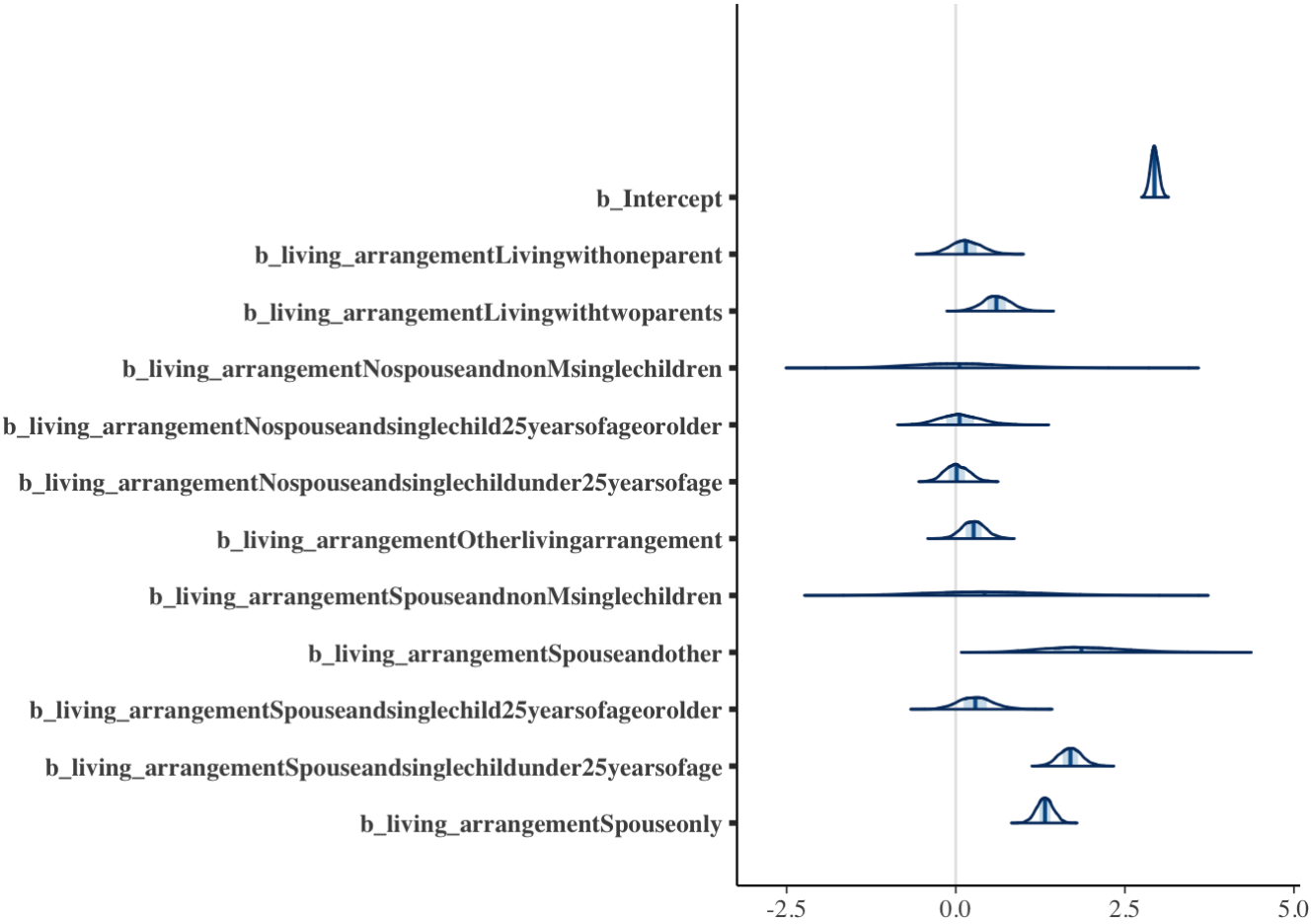

Figure 4: Distributions of Parameters - Family Life Satisfaction vs. Living Arrangement

We see some of the categorical level with 95% credible interval that contains 0, which implies that this level is not as interesting because it implies that this level could either negative association, no association, or positive association.

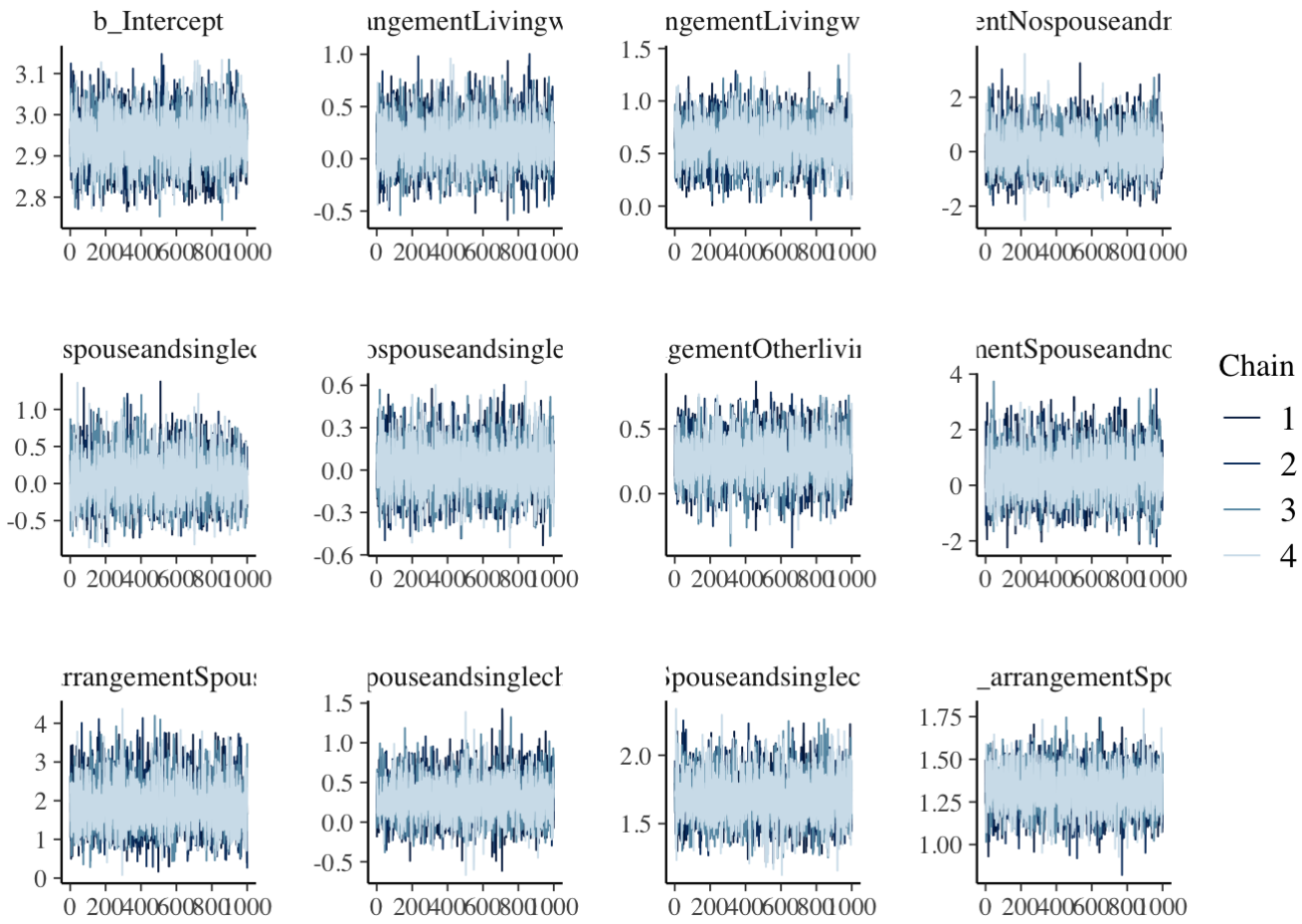

Figure 5: Convergence of Parameters - Family Life Satisfaction vs. Living Arrangement

The trace plot of each categorical level implies that they all converged since there is no trend of divergence

Table 2: The statistic table for Family Satisfaction vs. Living Arrangement

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	2.9385153	0.0605762	2.8197292	3.0610383
Living with one parent	0.1575063	0.2241667	-0.2538169	0.6075129
Living with two parents	0.6056439	0.2032271	0.2136043	1.0047168
No spouse and nonsingle children	0.0881786	0.7361745	-1.2384554	1.6319376
No spouse and single child 25 years of age or older	0.0702572	0.3117087	-0.5015591	0.7329970
No spouse and single child under 25 years of age	0.0180823	0.1757162	-0.3115767	0.3731651
Other living arrangement	0.2655835	0.1729159	-0.0637076	0.6138041
Spouse and nonsingle children	0.4533896	0.8630937	-1.1516099	2.2186941
Spouse and other	1.9021981	0.6217572	0.8211444	3.2379078
Spouse and single child 25 years of age or older	0.2950445	0.2555952	-0.1837594	0.8248686
Spouse and single child under 25 years of age	1.6962581	0.1693068	1.3722847	2.0436923
Spouse only	1.3201663	0.1210981	1.0800583	1.5532159

This is the numerical representation of Figure 4 above, as you can see, some of the levels with 95% credible interval that contains negative and positive values.

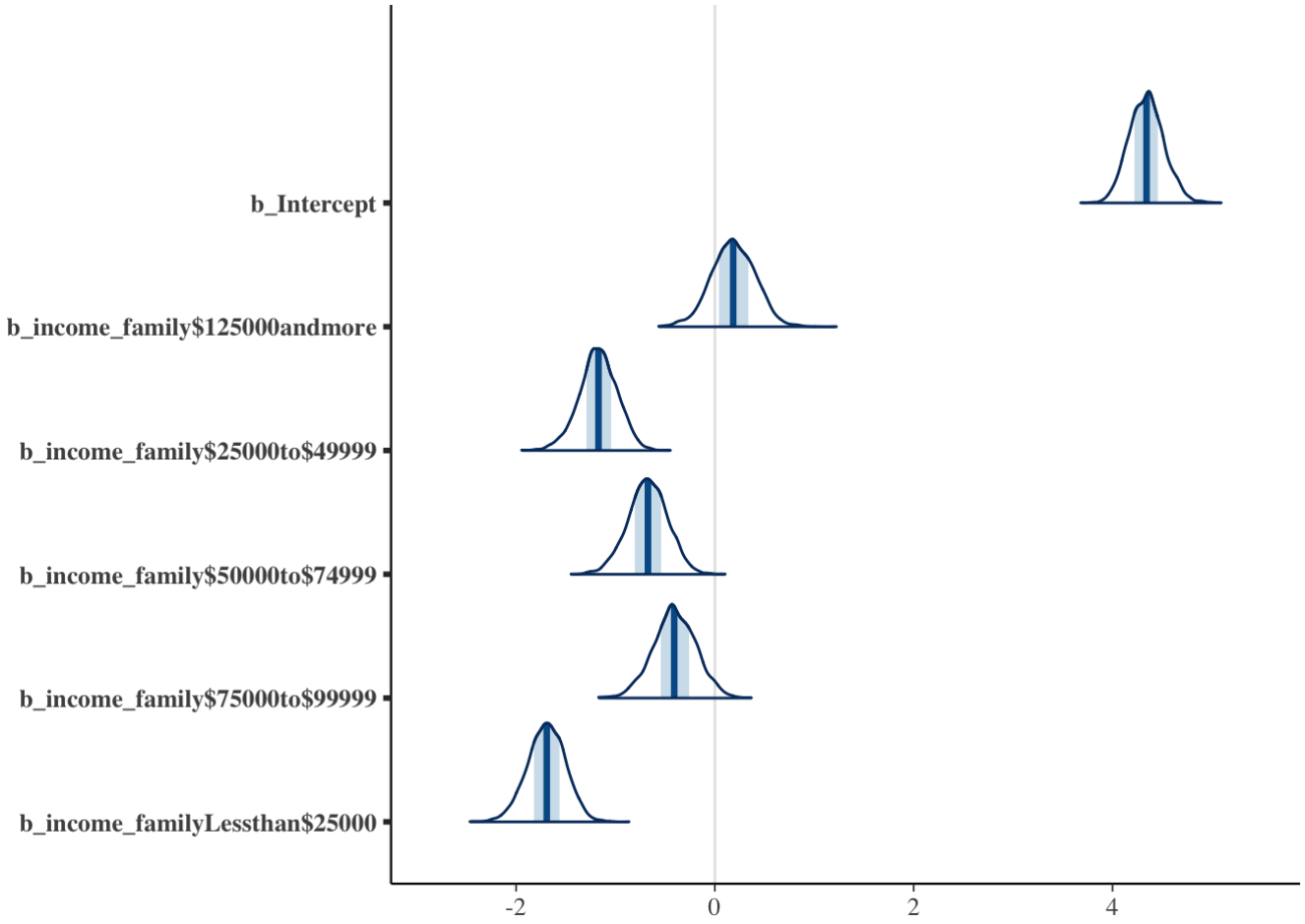

Figure 6: Distributions of Parameters - Family Life Satisfaction vs. Income Family

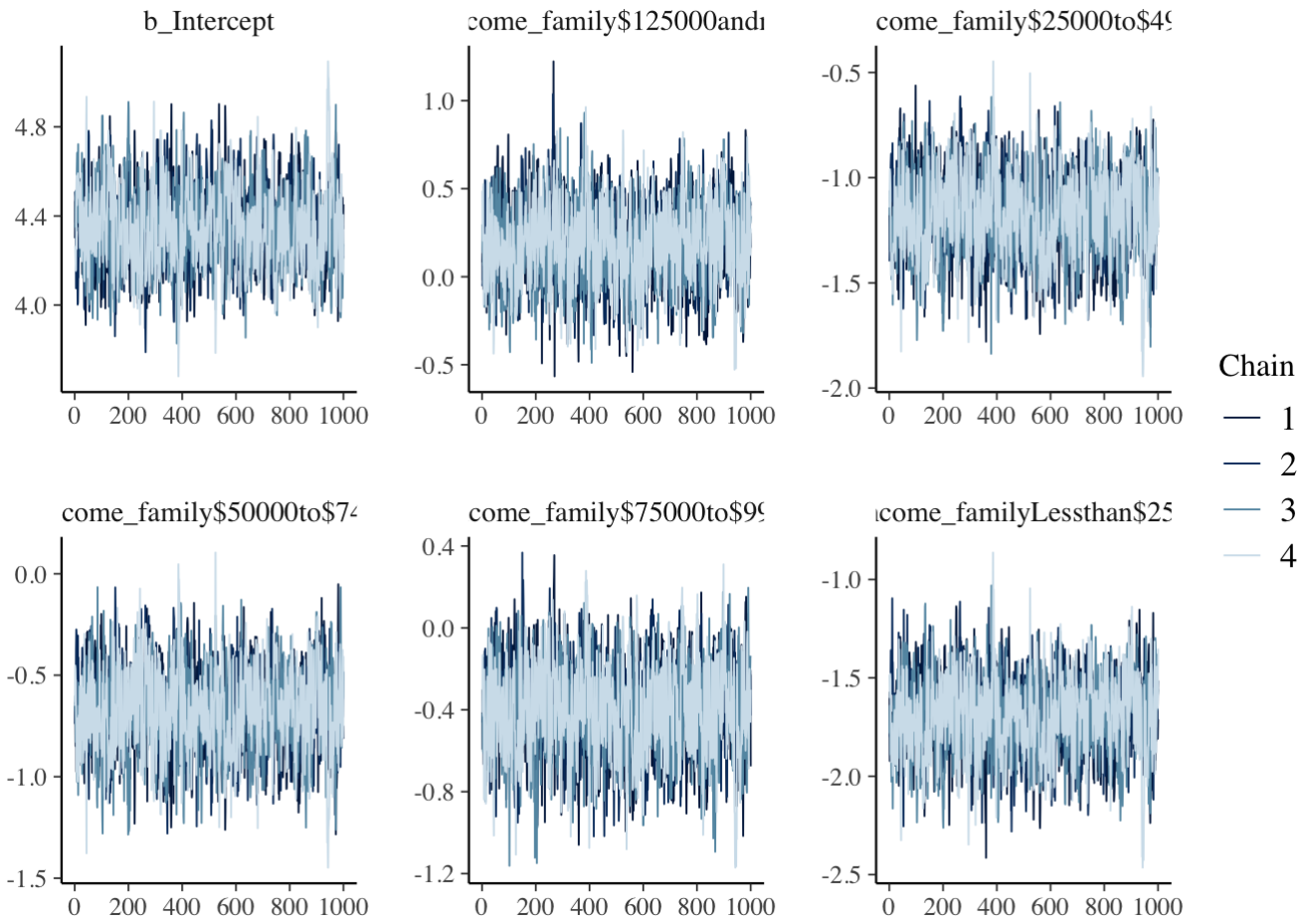

Figure 7: Convergence of Parameters - Family Life Satisfaction vs. Income Family

Table 3: The statistic table for Family Satisfaction vs. Income Family

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	4.3435628	0.1750100	4.0178822	4.6940995
Family Income $125000 and more	0.1866580	0.2201178	-0.2594290	0.6072299
Family Income $25000 to $49999	-1.1731421	0.1911527	-1.5695238	-0.8117577
Family Income $50000 to $74999	-0.6745281	0.2005980	-1.0777945	-0.2896857
Family Income $75000 to $99999	-0.4048893	0.2165724	-0.8380218	0.0119429
Family Income Less than $25000	-1.6951224	0.1924738	-2.0779786	-1.3335206

Descriptive Statistics for Chi-Square test of Independence

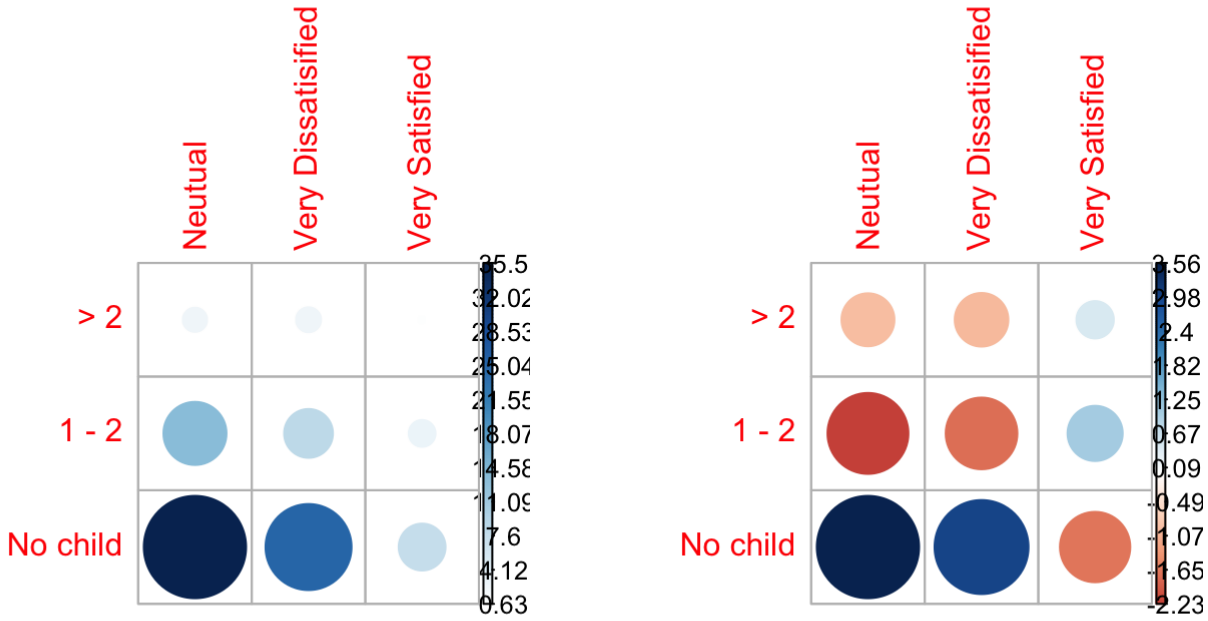

Figure 8: Contribution and Correlation Plot - Family Life Satisfaction vs. Number of Children

Refer to the Discuss section for detail interpretation

Table 4: Residual Table for Category Variables Family Life Satisfaction vs. Number of Children

	Neutual	Very Dissatisfied	Very Satisfied
> 2	-0.9533327	-0.9772801	0.4739335
1 - 2	-2.2258307	-1.7474327	1.0349230
No child	3.5604539	3.0046041	-1.6835349

Table 4.1: Chi-Square Statistics for Family Life Satisfaction vs. Number of Children

X statistics	P-value
35.70621	3e-07

The statistics for the Chi-Square test of Independence

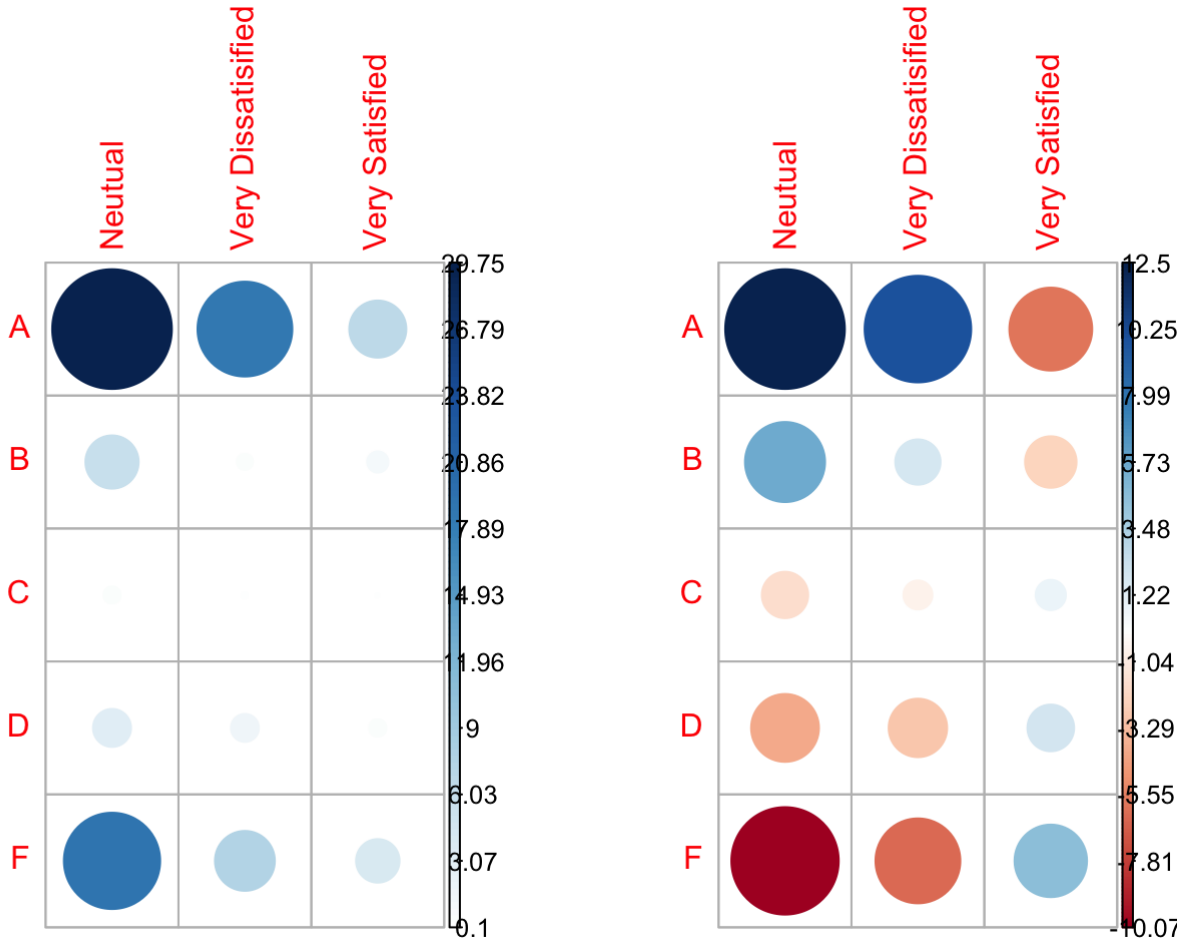

Figure 9: Contribution plot and Correlation plot - Family Life Satisfaction vs. Family Income

Refer to the Discuss section for detail interpretation and table.A in the Appendix section for the label mappings

Table 5: Residual Table for Category Variables Family Life Satisfaction vs. Family Income

	Neutual	Very Dissatisified	Very Satisfied
A	12.503977	9.9153976	-5.9934783
B	5.593399	1.7832826	-2.3107102
C	-1.871494	-0.7240695	0.7909304
D	-4.016901	-2.9785764	1.8965343
F	-10.065608	-6.2835399	4.5875563

Table 5.1: Chi-Square Statistics for Family Life Satisfaction vs. Family Income

X statistics	P-value
525.4935	0

The statistics for the Chi-Square test of Independence

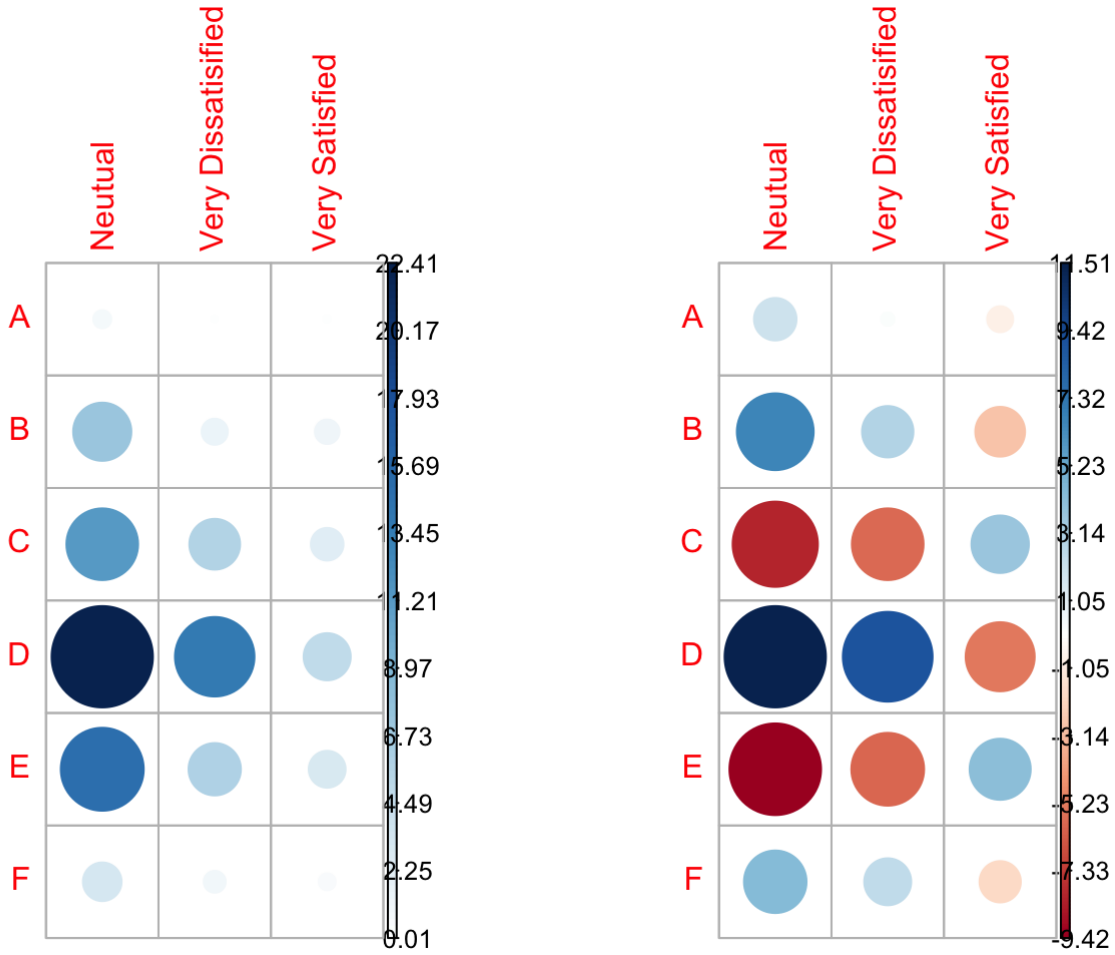

Figure 10: Contribution plot and Correlation plot - Family Life Satisfaction vs. Living Arrangement

Refer to the Discuss section for detail interpretation and table.B in the Appendix section for the label mappings

Table 6: Residual Table for Category Variables Family Life Satisfaction vs. Living Arrangement

	Neutual	Very Dissatisfied	Very Satisfied
A	2.019239	0.1847436	-0.7508194
B	6.611572	2.9556353	-2.7737944
C	-8.146852	-5.7647906	3.7027169
D	11.510825	9.0306533	-5.3504351
E	-9.418812	-5.9434169	4.1840259
F	4.384301	2.4679218	-1.9075264

Table 6.1: Chi-Square Statistics for Family Life Satisfaction vs. Living Arrangement

X statistics	P-value
591.3065	0

The statistics for the Chi-Square test of Independence

Discussion

Bayesian Logistic Regression Model Interpretation

Figure 2, Figure 3, and Table 1 are for model $\log \frac{p}{1-p} = \beta_{0c} + \beta_{1c} x_c$ (refer Model section for variable interpretation), we see that the Bayesian logistic model does converged (From Figure 3) and from the estimate and the 95% credible intervals of the predictor, total children, we see that there are some positive association between number of children and family satisfaction (Table 1). That is, the more children in the family, the family tends to be more satisfy about life.

Figure 4, Figure 5, and Table 2 are for model $\log \frac{p}{1-p} = \beta_{0l} + \beta_{1l} x_l$ (refer Model section for variable interpretation), we see that the Bayesian logistic model does converged (From Figure 4) and from the estimates and the 95% credible intervals, we see that some predictors tend to have some positive association with the family satisfaction (Table 2), for example, the predictor living with two parent has estimate of 0.6049157 and the credit interval range all lie above 0, that is, if the family live with two parent, the family satisfaction tends to be around 0.6 higher compare to the reference level. There are also some predictors that could not draw conclusion about such as living with one parent because the credible interval contains 0

Figure 6, Figure 7, and Table 3 are for model $\log \frac{p}{1-p} = \beta_{0i} + \beta_{1i} x_i$ (refer Model section for variable interpretation), we see that the Bayesian logistic model does converged (Figure 7), the reference level is family with income between $10,000 and $125000 and we see some negative associations(Table 3) for income family of lower income and family satisfaction relative to the reference level. For example, for family with income of $25,000 to $49,000, the family satisfaction is 1.1743058 lower compare to the families with income $10,000 and $125000

Chi-Square test of Independence Interpretation

Since this data set comprise of mostly categorical variables and our goal is to find the possible association between them and the family life satisfaction, Bayesian logistics model might not be the ideal model for fitting this data. The following are the interpretation for the Chi-Square test of Independence and it responsible for our final conclusion of this analysis.

Note for Figure 8, Figure 9, and Figure 10, the blue color represent positive association, the red color represent negative association, and the size of the circle and darkness of color represents the contribution to the total Chi-Square statistics. On the left of each figure is the contribution plot, each cell in the plot is calculated as $100\frac{r^2}{\chi^2}$ and we call it as the contribution to the final Chi-Square statistics because it indicates the nature of dependency between the row and column of the table. We can think of it as which residual contribute the most to make the final χ^2 statistics large enough to reject the Null hypothesis[9]. On the right side of each figure is the correlation plot, it is the similar the contribution plot except that it takes the sign into account, that is, each cell is calculated using the residual r(check the equation in the Model section)[9] in table directly below, negative residual is in red and positive residual is in blue. After knowing each plot is about, we can start to interpret it as follows for each category variable of interest. One interest observation of those plots is that, the the circles in the last column of each correlation plots (right side of the Figure 8, 9, 10) is the opposite color of the first two columns. That is, if the first two cell of a row in the correlation plot is blue, it is equivalent of saying that the category level is negatively correlated with the satisfaction; if the first two cell of a row is red, it is equivalent of saying that the category level is positively correlated with the satisfaction.

Figure 8 and Table 4 are the descriptive statistics to examine the association between the number of children in the family and the family life satisfaction. In particular, the two plots in Figure 8 are drawn according to Table 4. In the correlation plot (right of Figure 8), we see that having child in family tends to have a positive association with the family's life satisfaction as the first two circles of the first two rows are red circles. Furthermore, family have no child tends to have a negative association with the family satisfaction. Overall, more children in the family

associate with higher life satisfaction and Table 4.1 shows that the χ^2 statistics is large enough that the P value is way less than 5%, which implies the Null hypothesis of no association between number of children and Family life satisfaction is rejected.

Figure 9 and Table 5 are the descriptive statistics to examine the association between the family income level and the family life satisfaction. The family income level is in ascending order from top to bottom (see the mapping in Table A in appendix). We see that lower incomes tend to have negative association with family satisfaction (the first two rows in the correlation plot). Start from family income higher than $25,000, there is a positive association between the life satisfaction. Overall, higher income associate with higher life satisfaction and Table 5.1 shows that the χ^2 statistics is large enough that the P value is way less than 5%, which implies the Null hypothesis of no association between family income and Family life satisfaction is rejected.

Figure 10 and Table 6 are the descriptive statistics to examine the association between the living arrangement and the family life satisfaction. The most obvious observation from the correlation plot is that we see living with spouse along or with other family members have a positive association with the family satisfaction (row C and E) and living alone (row C) have a negative association with the family satisfaction. Another subtle finding is that living with parent seems to have a slight negative association with response variable. Overall, we can conclude at least living with spouse will have a positive association with the life satisfaction and Table 6.1 shows that the χ^2 statistics is large enough that the P value is way less than 5%, which implies the Null hypothesis of no association between living arrangement and Family life satisfaction is rejected.

Weaknesses

The weaknesses of this general survey is listed and discussed in the **Data** section. Here, let's focus on the sampling approach and analysis. One weakness of the sampling approach is when choosing the respondent in each household, the male and female respondents' ratio is not very well balanced. The 2017 GSS data includes 9399 male respondents and 11203 female respondents, which is acceptable but the ratio should be closer to 1. Males and females may likely have different feelings of life, even in the same living condition, and the unbalance of male and female respondents might bring some bias in our analysis.

In terms of the overall analysis, one of the major flaw is that we does not quantity the association. In the Bayesian approach, we realize that some predictors violate the linear relationship assumption between the logic of odds (the term $\frac{p}{1-p}$) and the fitted linear model. Thus interpret the estimated parameters might not be as meaningful. However, although we subsequently use another approach , the Chi-Square test of Independence, to conclude the association between categorical variables of interest, but we do not have the quantitative number measure of the strength of the association. But this will be the good next step to keep on the analysis.

Next Steps

The analysis in this report is just the beginning, we only touch the surface of the data, for the next step of our analysis, we can perform a three-way contingency table to analyze three categorical variables; also we can perform Cramer's V strength test[4] to give the quantitative measure of the strength of the association, known as the correlation. Last but certainly not least, this analysis does not screen all the possible predictors, so analyze on other categorical variables or even any combination of them are desired.

Appendix

The Github repository of the code: https://github.com/xuziyue/GSS2017-Analysis (https://github.com/xuziyue/GSS2017-Analysis)

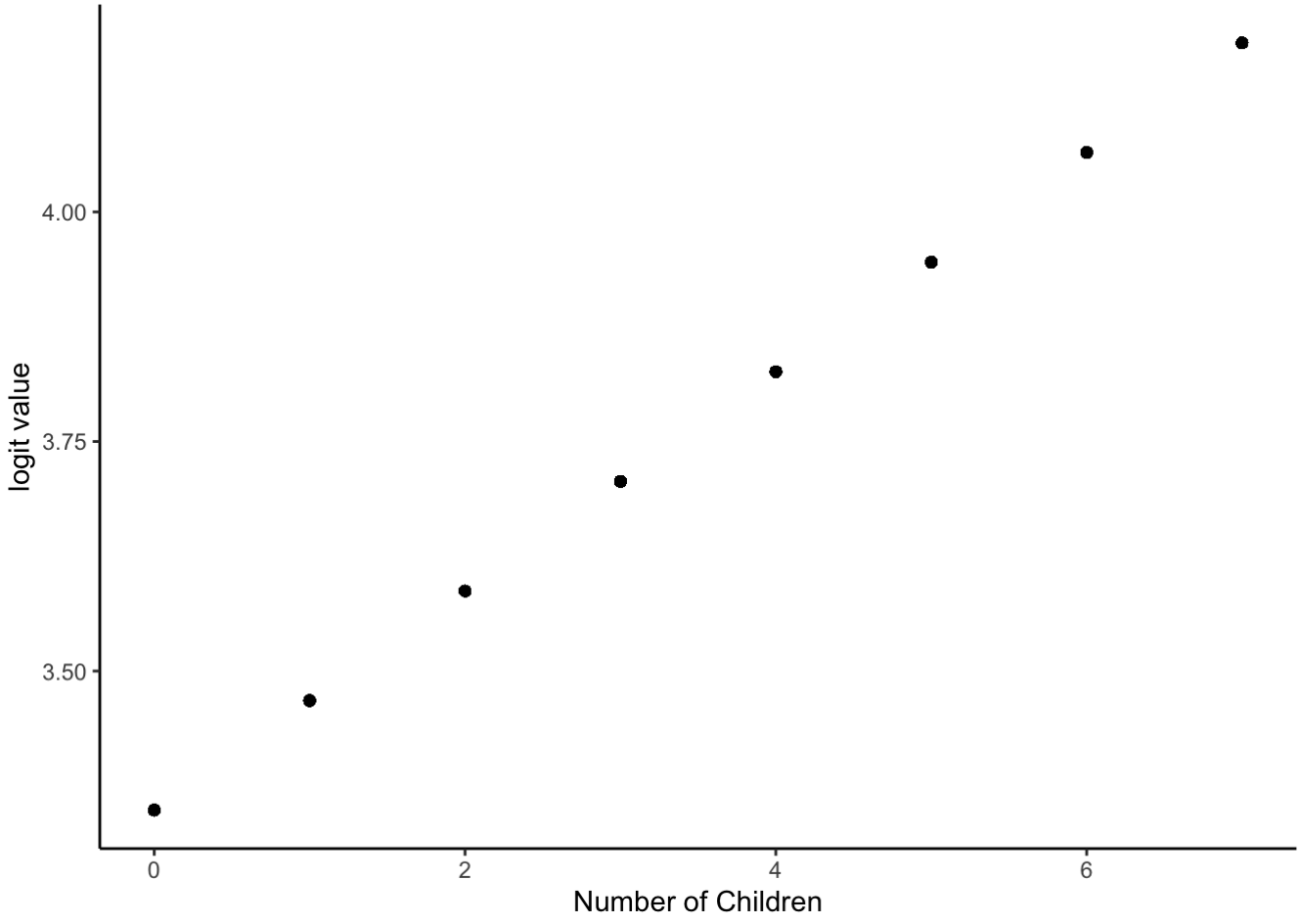

Figure A

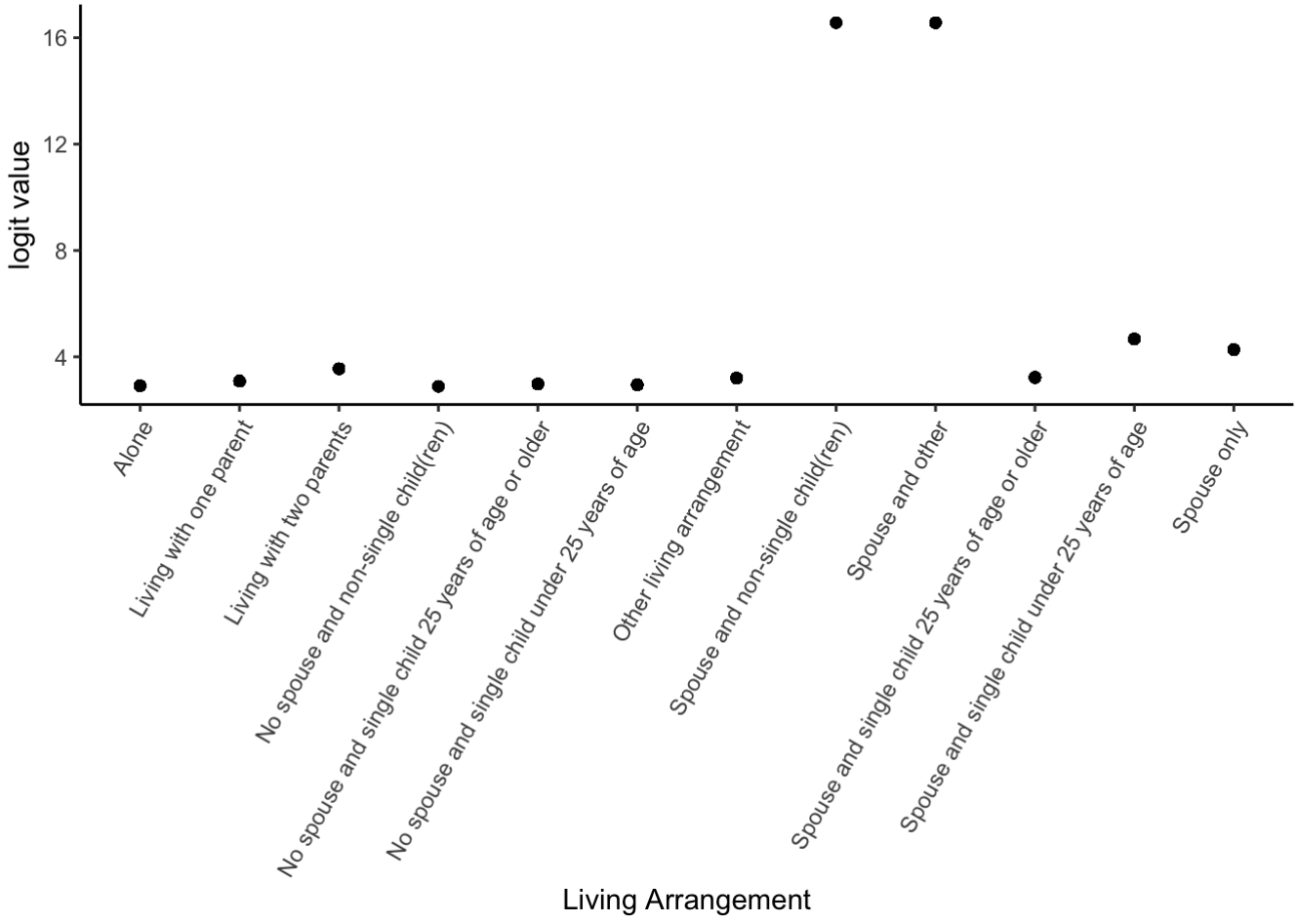
16
12
8
4
logit value
Alone
Living with one parent
Living with two parents
No spouse and non-single child(ren)
No spouse and single child 25 years of age or older
No spouse and single child under 25 years of age
Other living arrangement
Spouse and non-single child(ren)
Spouse and other
Spouse and single child 25 years of age or older
Spouse and single child under 25 years of age
Spouse only
Living Arrangement

Figure B

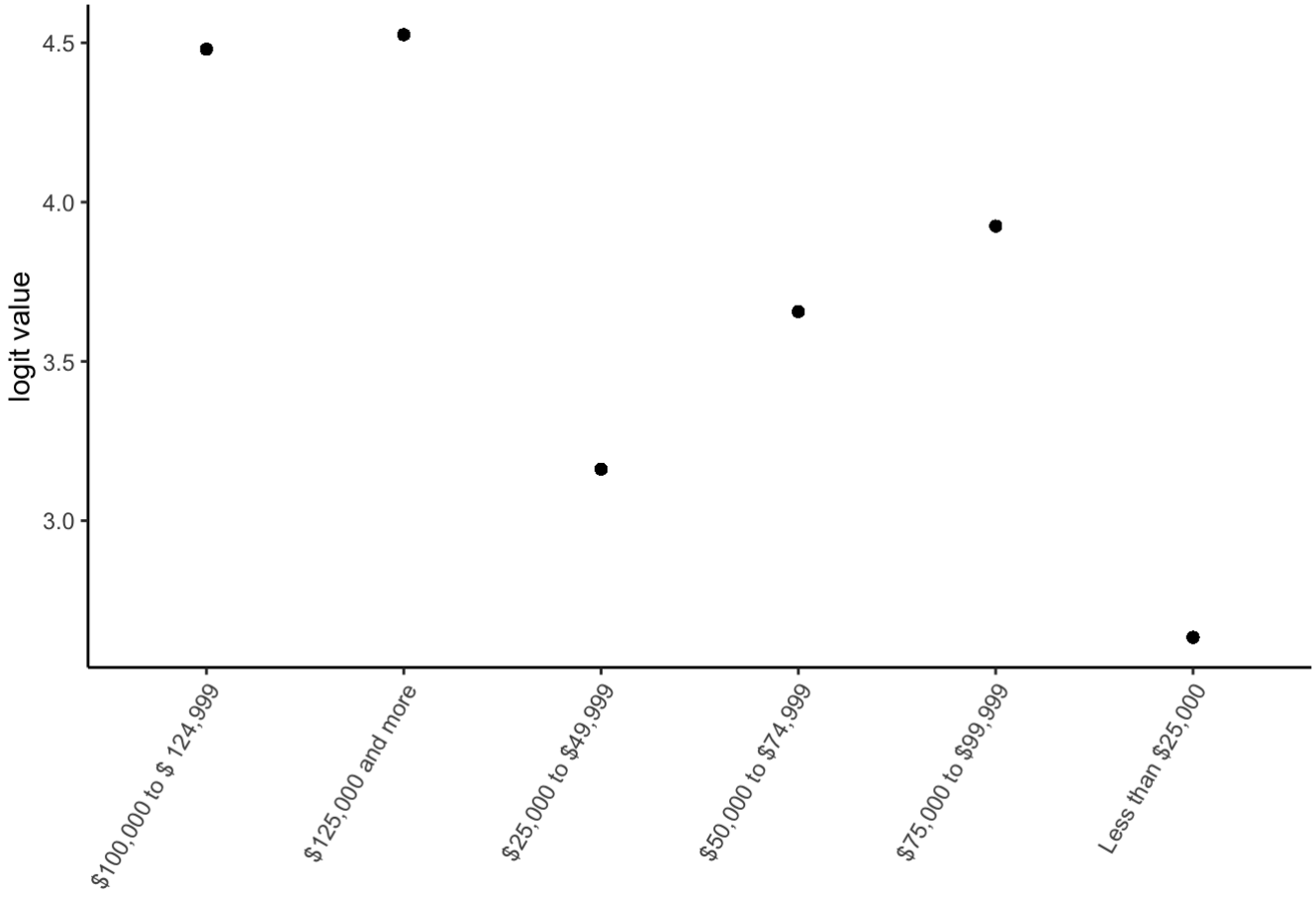

Figure C

Table A

Labels	Mappings
E	$100,000 to $124,999
F	125,000 and more
B	$25,000 to $49,999
C	$50,000 to $74,999
D	$75,000 to $99,999
A	Less than $25,000

Table B

Labels	Mappings
A	Living with one parent
A	Living with two parents
B	No spouse and non-single child(ren)
B	No spouse and single child 25 years of age or older

Labels	Mappings
B	No spouse and single child under 25 years of age
C	Spouse and non-single child(ren)
C	Spouse and other
C	Spouse and single child 25 years of age or older
C	Spouse and single child under 25 years of age
D	Alone
E	Spouse only
F	Other living arrangement

Table C Expected value of each cell - Number of Children vs. Family Life Satisfaction

	Neutual	Very Dissatisified	Very Satisfied
> 2	581.9988	80.78377	4497.217
1 - 2	1013.8735	140.72972	7834.397
No child	695.1277	96.48651	5371.386

Table D Expected value of each cell - Living Arrangement vs. Family Life Satisfaction

	Neutual	Very Dissatisified	Very Satisfied
A	165.9851	23.11185	1283.9031
B	115.8402	16.12965	896.0301
C	527.0280	73.38365	4076.5883
D	643.0937	89.54469	4974.3617
E	712.3950	99.19424	5510.4108
F	126.6580	17.63593	979.7061

Table E Expected value of each cell - Income Family vs. Family Life Satisfaction

	Neutual	Very Dissatisified	Very Satisfied
A	319.4977	45.85591	2351.646
B	500.8247	71.88087	3686.294
C	429.7991	61.68692	3163.514
D	340.0763	48.80946	2503.114
F	548.8023	78.76685	4039.431

References

1: “More Documentation 2017 General Social Survey (GSS): Families Cycle 31.” My.access - University of Toronto Libraries Portal, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/index.htm.

2: “Canadian General Social Surveys (GSS).” My.access - University of Toronto Libraries Portal, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm.

3: Alexander, Rohan. Telling Stories With Data, 17 May 2020, www.tellingstorieswithdata.com/.

4: McHugh, Mary L. “The Chi-Square Test of Independence.” Biochemia Medica, Croatian Society of Medical Biochemistry and Laboratory Medicine, 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/.

5: Lambert, Ben. A Student’s Guide to Bayesian Statistics. SAGE, 2018.

6: Makowski, Dominique, et al. “BayestestR: Describing Effects and Their Uncertainty, Existence and Significance within the Bayesian Framework.” Journal of Open Source Software, 13 Aug. 2019, joss.theoj.org/papers/10.21105/joss.01541.

7: Kassambara, and Michael U. “Logistic Regression Assumptions and Diagnostics in R.” STHDA, 11 Mar. 2018, www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/.

8: “3.4 - Experimental and Observational Studies: STAT 800.” PennState: Statistics Online Courses, online.stat.psu.edu/stat800/lesson/3/3.4.

9: “Chi-Square Test of Independence in R.” STHDA, www.sthda.com/english/wiki/chi-square-test-of-independence-in-r.

Package references:

Firke, Sam. “Janitor v2.0.1.” Janitor Package | R Documentation, www.rdocumentation.org/packages/janitor/versions/2.0.1.

Wickham, Hadley. “Easily Install and Load the ‘Tidyverse’ [R Package Tidyverse Version 1.3.0].” The Comprehensive R Archive Network, Comprehensive R Archive Network (CRAN), 21 Nov. 2019, cran.r-project.org/web/packages/tidyverse/index.html.

Bürkner, Paul-Christian. “Bayesian Regression Models Using ‘Stan’ [R Package Brms Version 2.14.0].” The Comprehensive R Archive Network, Comprehensive R Archive Network (CRAN), 8 Oct. 2020, cran.r-project.org/web/packages/brms/index.html.

“Visualization of a Correlation Matrix [R Package Corrplot Version 0.84].” The Comprehensive R Archive Network, Comprehensive R Archive Network (CRAN), 16 Oct. 2017, cran.r-project.org/web/packages/corrplot/index.html.

Tierney, Nicholas. “Preliminary Visualisation of Data [R Package Visdat Version 0.5.3].” The Comprehensive R Archive Network, Comprehensive R Archive Network (CRAN), 15 Feb. 2019, cran.r-project.org/web/packages/visdat/.