Towards a logistic analysis of satisfication of lifes

Lingyue Kong, Zhixing Hong, Jiayuan Pan

10/10/2020

Abstract

The purpose of this report is to evaluate whether a person is satisfied with life based on a collection of factors: age, total children, household size and family income. The data session summarizes the dataset used in this analysis, and key features are included. In the modeling session, the logistic regression modeling is used in analysing the probability of being satisfied with life, and there are compelling findings coming out of the result. Interestingly, as for all four factors, they are all negatively correlated with satisfaction of life, which helps the society to better predict one's feeling of life.

Key words: logistic modeling, logistic regression estimation, negative correlation

code and data supporting this analysis is available at: https://github.com/neverknowhen/STA304

Introduction

The feeling of life is in every moment of excitement, in any challenge and success. People are always pursuing happiness, because happy people can not only enjoy life, but also live longer. A new study has found that those in better moods were 35% less likely to die in the next 5 years. However, people still have not been able to get a final standard for measuring their own happiness.

The 2017 GSS is a sample survey with cross-sectional design. The target population are all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. The main content is to measure social changes related to living conditions and provide data to inform specific policy issues.

This study paid particular attention to the `feelings of lives`with respect to the numeric variables `age`, `total_children` and `hh_size`, and one categorical variable `income_family`, and will highlight major strengths and weaknesses while offering some discussion for raw data. The resulting model would be assessed to see how the response incorporated and had been influenced by five predictors. Finally, a recommended regression formula will be made about how `feelings of lives` could be measured to further meet the needs of people.

The aim of this study was primarily to determine whether there were some factors that could contribute significantly to the feelings of life. If so, it would be of interest to be able to define which factor can best confirm such feeling. The goal of the study was to use these factors as associated with 2017 GSS to develop a standardized tool for the evaluation of probability of the feelings of life.

Data

Key features of the data:

This data is from the 2017 General Social Survey (GSS): Families Cycle 31, provided by Statistics Canada under the terms of the Data Liberation. It was collected via computer assisted telephone interviews, and all the data output was transmitted electronically to Ottawa. There are 81 variables in the database. Some variables are straightforward, and some of them were created by combining two or three variables.

As feelings of life identified a large number of factors and characteristics, variables that were of particular intuitionistic interest were selected by the prior to analysis. The `feelings_life` was chosen to be the response variable and then created a 'new feel' which was dummied so as to only express the probability of being satisfied with life. As variable `new feel` is a categorical one, it has two levels: `good`(i.e., `feelings_life` with a value greater than 8) and `not so good`.

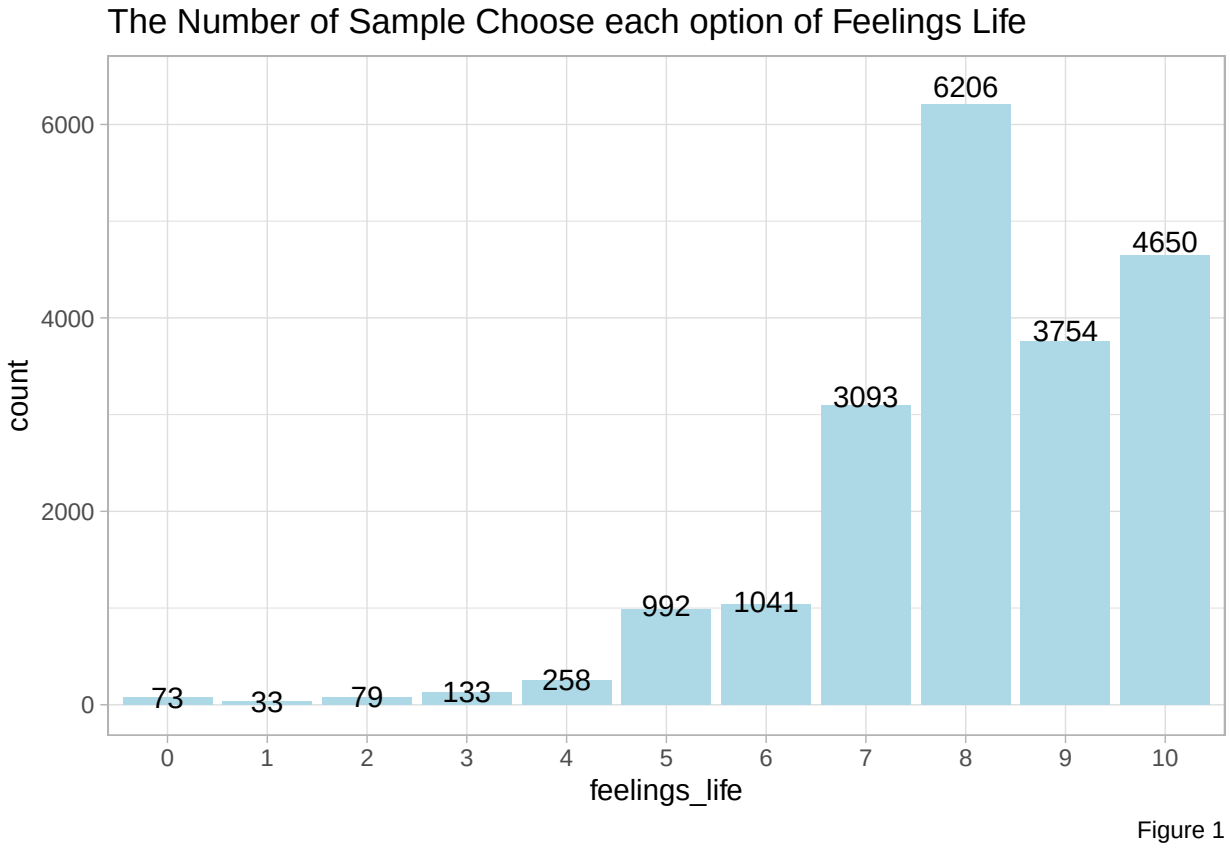

Figure 1

Figure1: This is a bar chart of the number of samples which were chosen for each level of the feelings of life for the GSS 2017 survey. The total number of samples was 20312, and over 70% of samples had high degrees of satisfaction which were over 8.

Figure 2: Number of Observations in Each Level of New_Feel

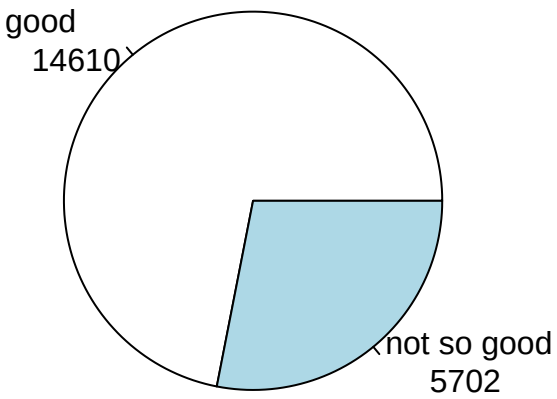

Figure2: This pie chart shows the number of samples at each level after we define the new variable— `New feel`, as it divided samples into satisfied and not so satisfied groups. From the chart, there were far more satisfied people than dissatisfied ones.

The numeric variables `age`, `total_children` and `hh_size`, and one categorical variable `income_family` were predictors. The variable `income_respondent` is very similar to variable `income_family`, the reason for choosing `income_family` over `income_respondent` is that we are more interested in the family level measure, as an individual is living in a family during his or her life. The variables categorized various aspects of study objects, including identifying locational features of the living conditions, degree of the economy of objects, as well as classifying the type of objects.(detailed data summry information in Appendix A)

It is also worth mentioning that since each of the variables in the data were used to identify the probability of high feelings of life, there was the possibility of collinearity between the variables (i.e. a number of variables measured common characteristics of feelings of life).

Limitations of the data:

Any missing values in the data were imputed as "Not Stated" or "Don't Know". While the assumption under which these imputations were conducted was probably reasonable, it may have been useful to perform a sensitivity analysis to determine whether such an imputation had a drastic effect on the results of the analyses.

Strengths of the data:

The method for sampling was the stratification, each of the ten provinces were divided into strata, and this will give us an unbiased sample. A simple random sample without replacement of records was next performed in each stratum which ensures each unit has equal probability of being selected. The dataset can be considered as a large data, and it is very representative of the population due to stratification sampling method.

Discussion of the questionnaire:

The questionnaire for 2017 GSS contained 460 questions, meeting two key objectives: monitoring the change of living standard and well-being of Canadians over time and gathering the data of specified social policy

issues. Details of the questions in the questionnaire can be found in (Appendix B). What is also worth mentioning is the underlying problems of the questionnaire. For telephone surveys, non-response is already increasingly prevalent. Unfortunately, the length of the GSS 2017 questionnaire raised the probability of non-response. Since the GSS program was started from 1985, the questionnaire contained two types of questions: existing questions and new questions. Both of these two kinds of questions were tested by Statistic Canada's Questionnaire Design Resource Centre. Though this increased the reliability of the questionnaire, and the pilot test could be included before finalizing the questions. The pilot test is able to reveal the potential problem when conducting the survey in the real world with low cost. And this can be considered as a good side of the questionnaire.

Discussion of the methodology:

The target population for the 2017 GSS included all non-institutionalized persons 15 years of age and older in 10 provinces of Canada. As for the stratification for sampling, each of the ten provinces were divided into strata. Details of stratified sampling can be found in (Appendix C). The survey frame was created using two different components: lists of telephone numbers in use available to Statistics Canada from various sources; the Address Register: list of all dwellings within the ten provinces. The word "record" will refer to the grouping of telephone numbers that consists of our sampling unit on the survey frame. The sampling approach is that each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records was next performed in each stratum. The sample formed by the above approach has the actual number of respondents to be 20,602. From this approach, some trade-offs may occur as it is very time consuming for gathering survey results. The extent of non-response varies from partial non-response (failure to answer one or a few questions) to total non-response. Total non-response was handled by adjusting the weight of households who responded to the survey to compensate for those who did not respond. And partial non-response was included in the sample, this is one downside of the data.

Model

In this section, we will discuss the logistic model we used to analysis Canadians' feeling of life, Design-Based inference in the modeling, and model assumptions and diagnostics.

The response variable we want to analysis is the `new feel` categorical variable. Since this variable has two levels, `good` and `not so good`, the logistic model was the best option to apply on the dataset. The logistic regression model is a statistical model that use the logistic function to model a binary response variable. In the content of `new feel`, the logistic model is predicting whether people are satisfied with their current life or not.

R software of version 4.0.2 was used to build the model. Considering how GSS data was collected, directly employing the logistic model on the data is not enough. Therefore, we need to add the design-based feature into the model to complete the logistics model. The function svydesign() in the `survey` package is used to including design method used by General Social Study. In the svydesign() function, we specified that the population is finite. After the design method was established and stored in the R, two logistic models were built in order to choose better one. The first one is with three variables: `age`, `total_children` and `hh_size`, while the second had one more variable `income_family`.

First Model:

$$log(\frac{p}{1-p}) = b_0 + b_1 age + b_2 totalchildren + b_3 hh\_size$$

Second Model:

$$log(\frac{p}{1-p}) = b_0 + b_1 age + b_2 totalchildren + b_3 hh\_size + b_4 income\_group_1 +$$

$$b_5 income\_group_2 + b_6 income\_group_3 + b_7 income\_group_4 + b_8 income\_group_5$$

Label: The b_n here are respenting the coefficients before each predictors. Age, total children and hh_size are same from the dataset. $income\_group_1$ is the income family level: 125,000 and more $income\_group_2$ is the income family level: 25,000 to 49,999 $income\_group_3$ is the income family level: 50,000 to 74,999 $income\_group_4$ is the income family level: 75,000 to 99,999 $income\_group_5$ is the income family level: Less than 25,000.

As `age` in the dataset was recorded as double, a subcategory belonging to number, `age` was keep as how it was in the logistic model. The variable `total_children` is the number of children reported by the respondent. Not only the number of children reported by the respondent is a numerical variable, but since the variable is not grouped, `total_children` used as numerical variable in the logistic regression model. Same for the `hh_size` (household size of the respondent). However, different from other three numerical variable, `income_family` is a grouped variable with 6 different levels. Recall that the income information recorded in the GSS dataset was obtained from the tax information each respondent reported. Therefore, it would be unreasonable to convert `income_family` into numerical variable by simply picking the mean from each income level. To concluded, there are three numerical predictors in the first model, whereas the second logistic model have three numerical variable and one categorical variable `income_family`.

For both two model, the design method, `gss.design` was included to adding survey sampling information. After using `svyglm()` function to get the result of each two model, we compared the accuracy of two models. For the first model, the accuracy is 0.719; the second model has the accuracy of 0.718. As the accuracy for the two model is very similar, the sensitivity and specificity rate for two models are compared. It is surprisingly to noticed that for the first model, the predicting results for all observations are `good`. This indicates that the first model is not good at predicting people who might feel `not so good` with their current living conditions. Due to the fact that it is more important to study why some Canadians feeling not so good about their lives, the second model is chosen as the final model, instead the first model.

After finalizing variables included in the logistic regression model, it is crucial to conduct the model checks and diagnostics. The model-checks include testing the linear relationship between continuous predictor variables and the logit of the outcome, and whether multicollinearity lies between predictors. To testify the relationship

between the numerical predictors and `new_feel`, we used the scatter plot to visualize the distribution of the data. Three scatter plots suggested that `hh_size`, `age`, `total_children` all have a linear relationship with the response variable. (Appendix D) The function `vif()` in the Car package is used to compute the variance inflation factor. Variance inflation factor represents how much larger the variance is due to the multicollinearity. The results of the four predictors are around 1 to 2 (Appendix F), suggesting that there is no multicollinearity between the predictors. As the logistic regression model has been tested that it satisfies the model assumptions, the influential observations are found to check their influence on the logistic model. Using `augment()` built in the `broom` package, observations with ten largest Cook's distance were listed and plotted. As the Cook's distance of these observations are much smaller than the 50th percentile of the F distribution with 4 and (20312 – 4 – 1) degrees of freedom, the observation should be included in the logistic model (details in Appendix E).

Result

The result of the logistic regression model is shown as below, in the table below.

Table 1: Logistic Regression Model Result

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.3646789	0.0984389	-3.704624	0.0002123
age	-0.0132052	0.0012147	-10.870796	0.0000000
hh_size	-0.0955609	0.0174726	-5.469176	0.0000000
total_children	-0.0316270	0.0132840	-2.380840	0.0172824
as.factor(income_family)$125,000 and more	-0.0862364	0.0646246	-1.334420	0.1820811
as.factor(income_family)$25,000 to $49,999	0.6371701	0.0633541	10.057278	0.0000000
as.factor(income_family)$50,000 to $74,999	0.3822981	0.0649425	5.886712	0.0000000
as.factor(income_family)$75,000 to $99,999	0.2529632	0.0681037	3.714381	0.0002042
as.factor(income_family)Less than $25,000	0.9843295	0.0678890	14.499104	0.0000000

By combining the coefficients listed in the table, the following logitic function was summarized to present the model result.

$$log(\frac{p}{1-p}) = -0.365 - 0.013age - 0.096hh\_size - 0.032total\_children -$$
$$0.086Income\_group\_1 - 0.637Income\_group\_2 + 0.382Income\_group\_3 +$$
$$0.253Income\_group\_4 + 0.984Income\_group\_5$$

The following figures showed how the probability of the Canadian felt good about their current life varies as the predictors(`age`, `hh_size`, `total_children` and `income_family`).

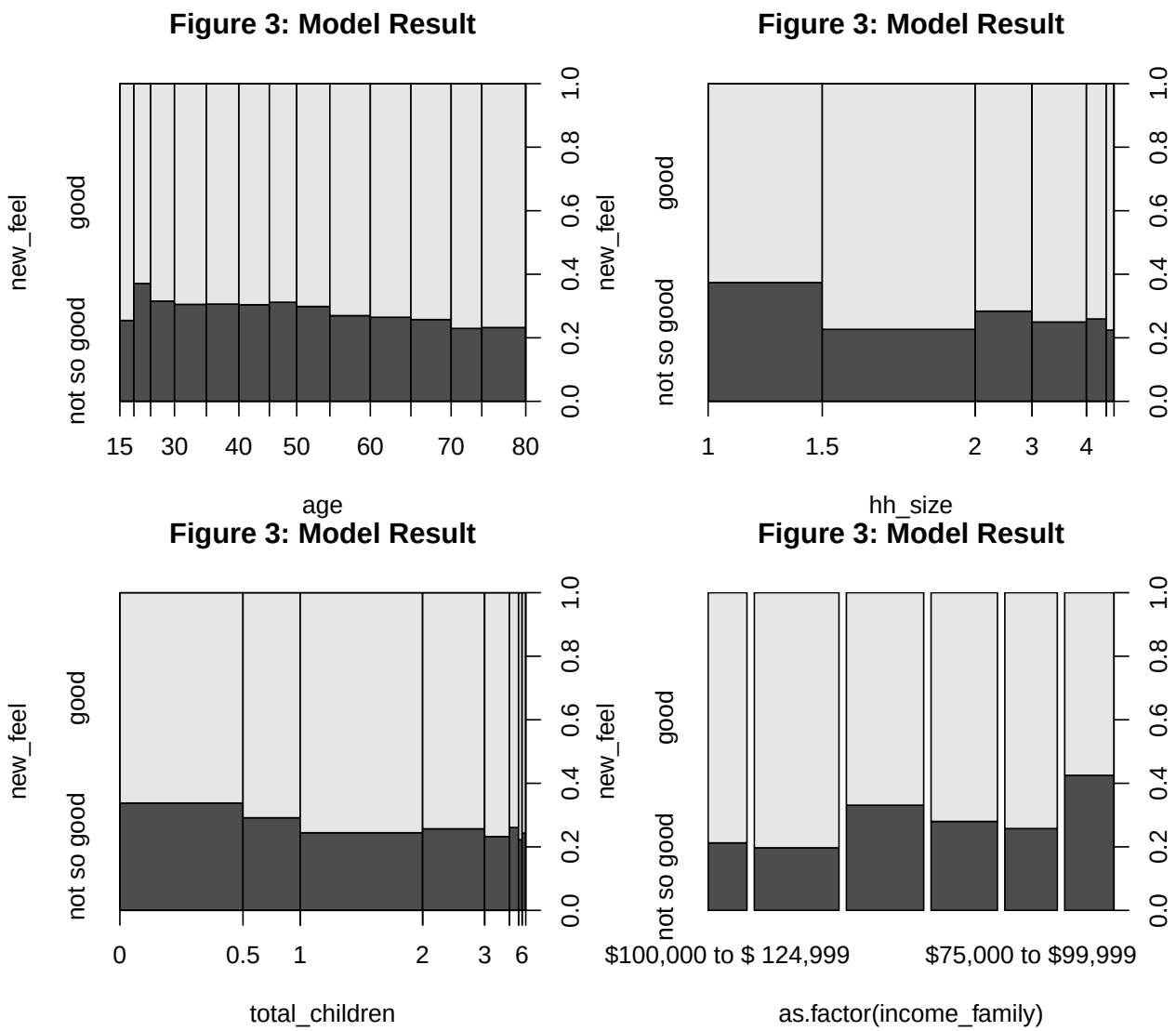

Figure 3: These four plots are demonstrating how the probability obtained from the logistic model varies as the predictors value changes. The X axis is representing the values for numerical predictors(`age`, `hh_size`, `totol_children`). As for the categorical predictor, `income_family`, the x-axis is divided into 6 bar, indicating 6 different income groups. The y-axis shows the probability of Canadians feeling good about their living conditions. The dark-gray area represents the proportion of people feeling not so good about their living conditions within that group, while the light area is for people feeling good about their current live. By viewing the change of the dark-grey area, the general relationship between `new feel` and the predictors.

Discussion

In this session, we will focus on discussing the main results of the model and interpretation of it. The following plots help us visualizing the dataset used in the model. First of all, Figure 1 shows the dataset from one variable `feelings_life`'s perspective, which is our main focus variable from the dataset. The variable `feelings_life` has 11 levels and it ranges from 0 to 10, for 0 means the person is not satisfied with life, and 10 means very satisfied with life. We are interested in determining the probability of a person with high degree of satisfaction(i.e., value of `feelings_life` is over 8) given predictor variables(i.e., `age`, `total_children`, `hh_size`, and `incom_family`). Moreover, Figure 2 is a pie chart showing that 70% of people in the data have a value greater than 8, and the rest 30% of them have a value less than or equal to 8. Regarding possible bias of the dataset, one can find that the distribution of feelings_life is left skewed, with few values far to the left and closer to zero. Skewness of the data may imply the dataset is not representative enough if assuming the distribution of feelings_life is uniform from 0 to 10.

The logistic regression model results are shown in the previous result session, and the following are interpretations of the result. In the following analysis, "p" will represent the probability the person is satisfied with his or her life(i.e., his or her feelings_like variable has value larger than 8). Interpreting the p_value from the model, one can see that the p_value is relatively small compared with a predetermined significance level of 5% except for the one associated with income group 2. With such low p_values, a moderate to strong relationship between the log odds of being satisfied with life and all predictor variables(age, hh_size, total_children, and income_family) can be established.

It is stated by James, G. (2013) that the estimated intercept is typically not of interest; its main purpose is to adjust the average fitted probabilities to the proportion of ones in the data. So in this model for the probability of being satisfied with one's life, the intercept with value -0.365 is not of interest. More meaningful estimates in the following are of interest. In terms of age, for every additional year increase in age, it is expected that the log adds of being satisfied with life to decrease by 0.013(i.e., probability of being satisfied is negatively related with age).

Similarly for estimates associated with household size`hh_size` and total children, the probability of being satisfied is also negatively related with those two predictor variables. For every additional unit increase in `hh_size`, it is expected that the log adds of being satisfied with life to decrease by 0.096. And for every children increase, it is expected that the log adds of being satisfied to decrease by 0.032. "Dummy" variable is used for categorical variable `income_family`, and if the person is in income group 2, then it is expected that the log adds of being satisfied to decrease by 0.086. However, for income group 6, it is expected that the log adds of being satisfied to increase by 0.984. It is shown that income_family is negatively related with being satisfied, it suggested that if a person has more family income, then he or she is less likely to be satisfied with life.

To sum up all relationships between predictor variables and the probability of being satisfied with life, it is interesting that all four variables are negatively correlated with the log odds of being satisfied with life. From the model, one can learn that if the person has more children, bigger household size, more family income and is older, he or she tends to be less satisfied with life. In reality, having more children and bigger household size is associated with more housework and trivial matters, which will lead to lower satisfaction of life. When people get older, they will miss their life when they are young, and will also be more satisfied with their previous life when they are younger. It is very interesting when talking family income into consideration, it has a negative impact on being satisfied with life. Perhaps, high income families will tend to think about earning more money and not be satisfied with their current life.

This model will represent the small world to an extent of people in Canada. In terms of the whole world, this model may not perform well, as different countries have different living conditions and measurements. It is limited to the society of Canada, and may not be able to use it globally. As for future work, opportunity is there to gather more information and data from different countries across the world. Therefore, future opportunities can focus on extending the model globally.

Summary

In this report, we demonstrate that the probability of being satisfied with one's life is negatively correlated with age, household size, total children, and family income. Under the logistic regression model, we built a tool for evaluating the probability of being satisfied with life based on above four factors. This potential can be used in Canada to assist the social changes related to living conditions and provide data to inform specific policy issues, which is the primary objective of the General Social Survey. It is hoped that in the future work, one can extend the model globally based on more global data.

Reference

1. Data source: 2017 General Social Survey (GSS): Families Cycle 31, provided by Statistics Canada under the terms of the Data Liberation

2. Gss31_user_Guide of 2017 General Social Survey (GSS): Families Cycle 31 https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

3. Wu, C., & Thompson, M. E. (2020). Sampling Theory and Practice (ICSA Book Series in Statistics) (1st ed. 2020 ed.). Springer.

4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) (1st ed. 2013, Corr. 7th printing 2017 ed.). Springer.

5. (2018, March 11). Logistic Regression Assumptions and Diagnostics in R. Articles - STHDA. http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

6. Population Data Source: Statistics Canada. Table 17-10-0005-01 Population estimates on July 1st, by age and sex DOI: https://doi.org/10.25318/1710000501-eng.

Appendix

Appendix A

Data Summary for Variables Used In the Model:

Table 2: Summary Data Information for age, total_Children, hh_size

age	total_children	hh_size
Min. :15.00	Min. :0.000	Min. :1.00
1st Qu.:37.30	1st Qu.:0.000	1st Qu.:1.00
Median :54.10	Median :2.000	Median :2.00
Mean :52.11	Mean :1.675	Mean :2.35
3rd Qu.:66.62	3rd Qu.:3.000	3rd Qu.:3.00
Max. :80.00	Max. :7.000	Max. :6.00

Table 3: Number of Observations in Income_family group

income_family	n
$100,000 to $ 124,999	2137
$125,000 and more	4661
$25,000 to $49,999	4256
$50,000 to $74,999	3653
$75,000 to $99,999	2890
Less than $25,000	2715

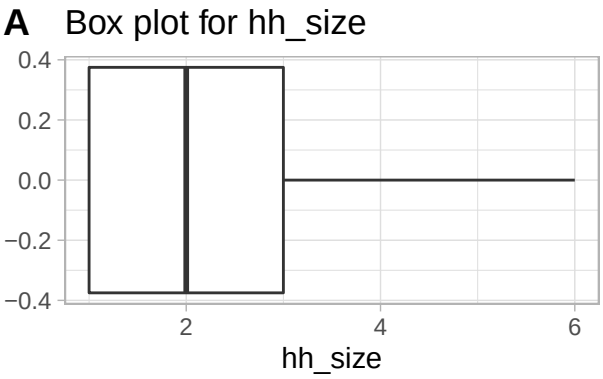

Figure 4.a

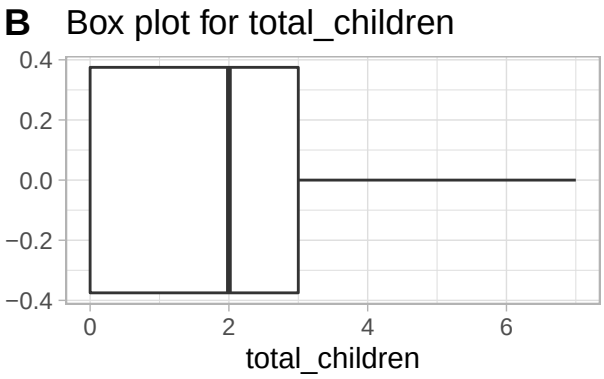

Figure 4.b

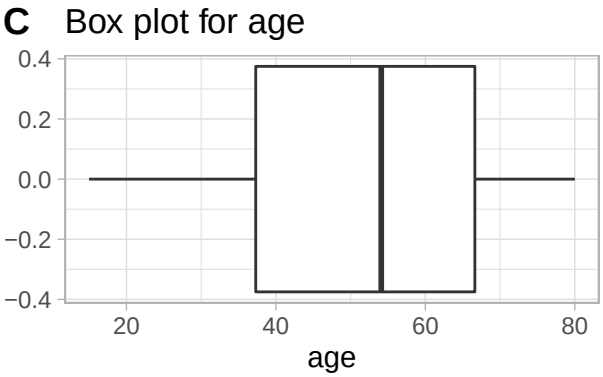

Figure 4.c

Appendix B

Details on questionnaire of the survey:

The 460 questions could be classified into 14 different categories, such as the entry component, family origins, conjugal history, children of respondents, and so on. The questions listed in the questionnaire followed a logical order. Respondents were first asked to provide the date of birth and then asked for other entry information. Since the targeting population of the general social survey were Canadians over 15 years old, the first question was conducive in increasing the efficiency of conducting the survey, as getting rid of the substandard respondents. In addition, the questionnaire was designed to contain questions asking conjugal history. Data such as current legal marital status of the respondent, the respondent's spouse or partner's date of birth and marital status prior to the union, and also the children's information for each union were obtained. The diverse data, which are impossible using other sources, allows analysis to be conducted from different perspectives.

The intended respondent might speak with the interviewer, but refuse to consent to respond, when they learned about the time needed to answer about 400 questions. It is also possible that they may consent to respond, and begin the interview, but are not able to finish all the questions due to the time limit. Therefore, the number of questions in the questionnaire might cause certain errors.

Appendix C

Details on methodology of dataset:

Stratification: Many of the Census Metropolitan Areas1 (CMAs) were each considered separate strata for St. John's, Halifax, Saint John, Montreal, Quebec City, Toronto, Ottawa, Hamilton, Winnipeg, Regina, Saskatoon, Calgary, Edmonton and Vancouver. All CMAs not on this list are located in Quebec, Ontario and British Columbia, with the exception of Moncton. Three more strata were formed by grouping the remaining CMAs (except Moncton) in each of Quebec, Ontario and British Columbia. Finally, the non-CMA areas of each of the ten provinces were also grouped to form ten more strata, for a total of 27 strata. Moncton was added to the non-CMA stratum for New Brunswick. For each province, minimum sample sizes were determined that would ensure certain estimates would have acceptable sampling variability at the stratum level.

Appendix D

Linearity between the Predictors and Response: Figure4: Here were three scatter plots which correspond to the linear relationship between numerical variables, specifically were `age`, `household size`, and `the number of the total childrens`, and the response variables. The first plot showed that the age had a strong linear relationship with the response variable. However, for the second and the third plots was a step function, fit to the qualitative variable household size and the total number of children.

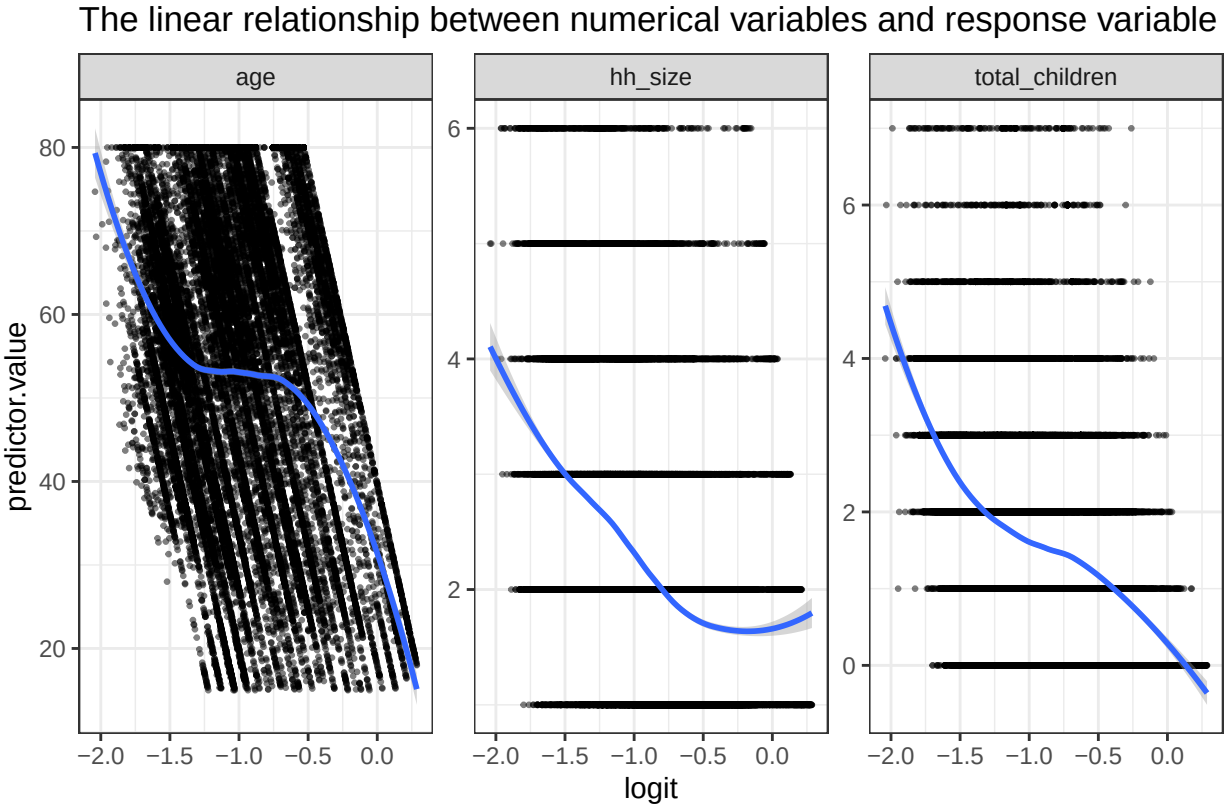

Figure 5

Appendix E

Influential Points and Cook's Distance:

Figure 5: This plot illustrated cook distances for leverage points of the samples, and picked out the three points with largest cook distances. These three points sit a far distance from the center of all sample predictors.

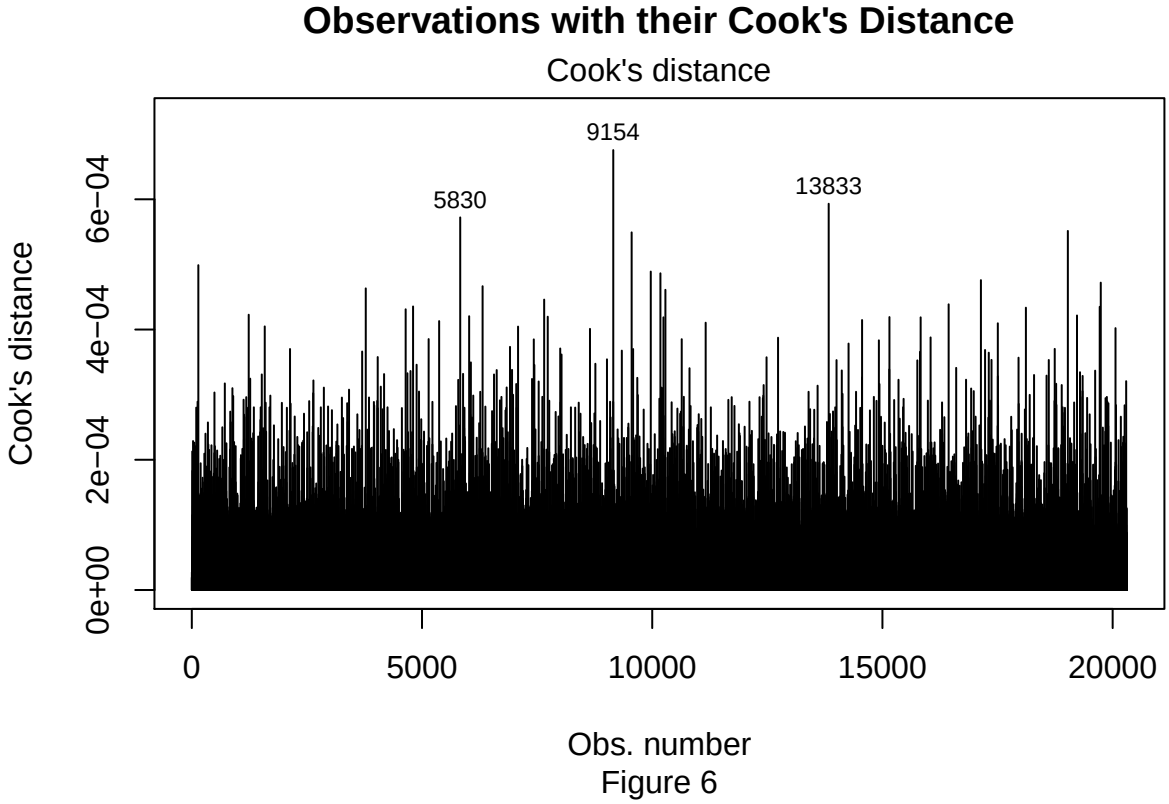

Figure 6

Table 4: This table displayed three points which have the three largest cook's distance in the entire dataset. These values demonstrated that all the points in the dataset are not influtial points that effect the model preforming.

Table 4: Three Points with Largeat Cook's Distance

index	.cooksd
5830	0.0005722
9154	0.0006756
13833	0.0005932

Figure 6: This is a plot of residuals versus predicted (or fitted) values for the GSS2017 datasets. A strong pattern in the residuals indicates non-linearity in the data. Here, the residual plot showed no discernible pattern which provides a strong indication of linearity in the data, for both satisfied and not so satisfied samples.

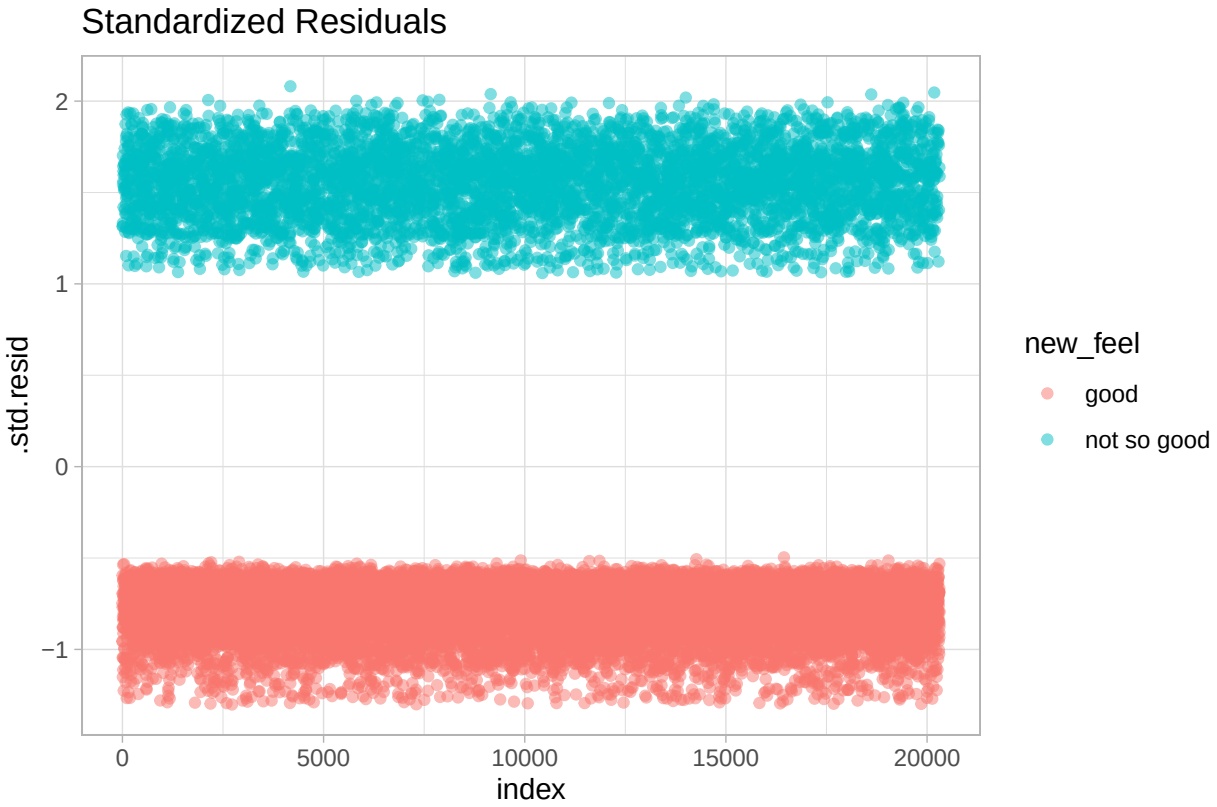

Figure 7

Appendix F

Variance Inflation Vector for Predictors:

Table 5: the variance inflation vectors for each predictor are listed. The result indicates that there is not multicollinearity between the variables, as the VIFs for predictors are all smaller than 5.

Table 5: Variance Inflation Vector

	GVIF	Df	GVIF^(1/(2*Df))
age	1.962239	1	1.400799
hh_size	1.742772	1	1.320141
total_children	1.536354	1	1.239497
as.factor(income_family)	1.192299	5	1.017744

Appendix G

Github Repo Link: https://github.com/neverknowhen/STA304