The Study of How Family Status Positively Affects Life Satisfaction

Shuxian Cao, Xueying Fu, Yichen Su, Yiyang Huang

October 19, 2020

Contents

1 Information

Topic: The Study of How Family Status Positively Affects Life Satisfaction

Author: Shuxian Cao, Xueying Fu, Yiyang Huang, Yichen Su

Date: October 19, 2020

Code and data supporting this analysis is available at: https://github.com/YiyangHuang028/STA304-Problemset-2.git

2 Abstract

The marriage rate declined over the last few years, and more people refuse to get married due to various reasons. The target population of our study is the entire Canadian population, and our study aims at analyzing how family status affects life satisfaction and giving a prediction of life satisfaction of each individual based on their gender, marital status, living arrangement and total number of children they have. The study is based on the General Social Survey conducted in 2017 in Canada, and we use multiple linear regression to build our model. The result suggests that married people generate a higher level of life satisfaction, especially found in married women and those who live with their spouse.

3 Introduction

In the past, people were expected to get married at a certain age, and women were expected to give birth when they are young. We are now living in an open society where each individual has the right to make their own moral decision (Mason, 2018). Some people don't want to get married due to various reasons and this leads to a decline in marriage rates during the past few years.

The study conducted by Pew Research Center among unmarried adults suggested that 14% of them said they don't want to get married, and 27% are "not sure if they want to get married" (Parker & Stepler, 2017). According to the survey of U.S. adults conducted in 2019, only 16% of men and 17% of women think being married is essential for living a fulfilling life, and only 16% of men and 22% of women see having children as essential (Horowitz, Graf & Livingston, 2019). Some people avoid marriage because it carries too many expectations and responsibilities, which is totally acceptable.

This study analyzed the impact of family status on life satisfaction. We would like to know how marriages and children affect feelings of life in Canada. The study is based on the General Social Survey, Cycle 31: Families, conducted in 2017. Plots and tables are used to visualize the relationship among gender, marital status, living arrangement, total children and feelings of life. Multiple linear regression is used to give an estimation of feelings of life given certain variables. The result can be used as a reference for those unmarried adults who are uncertain about marriage.

4 Data

The dataset that is used in this report is retrieved from the 2017 General Social Survey(GSS). The objective of the survey is to see the life changes for Canadian families over time as well as provide information for current interest of Canadians on some policies. The target population of GSS is 15 years old and older people in Canada. The raw dataset `ggs_original2017` originally contains 20602 observations and 81 variables. It also contains a great amount of missing values.

Stratified sampling and simple random sampling were used for the survey, and stratas were divided upon different provinces. The survey was done through telephones, and the telephone numbers were collected from "Statistic Canada" and the "Address Register". Nearly half of the people responded to the call. For those who did not respond, their records of the survey were dropped. Since the data are collected from making calls, the cost of the survey is much less than a face-to-face interview.

According to the questionnaire, the strength is that there are many various types of questions, and therefore, the information on people would be very detailed. Nevertheless, most of the variables are non-numeric so it is difficult to make a fitted model for categorical variables. Also, making calls through telephones may not be a good choice since there are less Canadians who have a landline at home. Also, telephone surveys may increase response bias.

In this report, we will mainly be focusing on discovering how family status influences people's life satisfaction. Thus, our essential variables are gender, marital status, living arrangement, and total numbers of children in a family. After the data cleaning process, the dataset contains 20306 observations and 8 variables. For the `living_arrangement`, which is a categorical variable, we combined similar categories and reduced them to six different categories instead of twelve.

5 Model

$$\begin{aligned}Y_i = \beta_0 + \beta_1 X_{male,i} + \beta_2 X_{MS_{LCL},i} + \beta_3 X_{MS_{Married},i} + \beta_4 X_{MS_{Seperated},i} + \beta_5 X_{MS_{Single},i} + \beta_6 X_{MS_{Widowed},i} \\ + \beta_7 X_{LA_{C.only},i} + \beta_8 X_{LA_{others},i} + \beta_9 X_{LA_{S.only},i} + \beta_{10} X_{LA_{parents},i} + \beta_{11} X_{LA_{SandC},i} + \beta_{12} X_{TotalC,i} + e_i\end{aligned}$$

To analyze population's life satisfaction and provide a good estimation, we chose to use multiple linear regression(MLR). In comparison to single linear regression, MLR is useful to determine "the relative influence of one or more predictor variables to the criterion value" as well as "outliers and anomalies" (Weedmark, 2018). Unlike the logistic model, MLR can be used to predict a numeric variable, which is exactly what we need in this study. Another strength of multiple linear regression is that it is easy to interpret, which means it can be understood by most of the population. We put this into consideration because if people can understand how we get the result, they are more likely to be convinced.

In this study, MLR allows us to model the relationship between 4 independent predictor variables and one responding variable. The responding variables that we need to analyze is `feelings_of_life`, which is a numeric variable that measures the life satisfaction of each individual. Some possible factors that affect life satisfaction are `sex`(gender), `marital_status`, `living_arrangement`, and `total_children`. The model is run by R.

`sex`(gender) is a categorical variable with 2 categories: male and female. In general, men and women think differently. They play different roles in marriages and they have different needs for life. We expect men and women to give different scores towards life, so we put `sex` as a feature in our model.

`marital_status` is a categorical variable with 6 categories: divorced, living common-law, married, separated, single and widowed. Social connections and relationships usually bring happiness to people. Marriage plays an important role in our life and it may affect life satisfaction . Also, the goal of our study is to analyze how family status influences life satisfaction. Thus, we think marital status is a key feature that we need to include in our model.

`living_arrangement` is a categorical variable which originally contains 12 categories. Home is a place where we recharge ourselves. The environment and the people we live with decide our mood and quality of life to a large extent. Some of the categories under this variable are similar so we combined some of them and finally reduced it to 6(alone, spouse only, children only, with child(ren) and spouse, with parents) in order to get a concise model.

`total_children` is a numeric variable with integer input from 0 to 7. Being a parent can be stressful. It carries a lot of responsibilities especially when your children are still young and it might affect an individual's life satisfaction. Thus, we put `total_children` as a feature in our model.

The model assumption is checked by plotting standard residual plot of all predictor variables and fitted values. According to Figure 5, the dots are evenly spread out and there is no obvious pattern observed. Thus, we can conclude that there are no obvious model violations (independent and constant variance). In the normal QQ plot, some of the dots do not follow the straight line which suggests that the normality is not fully satisfied. For example:

$$\hat{Y_i} = 7.631 - 0.105 X_{male,i} + 0.276 X_{MS_{LCL},i} + 0.463 X_{MS_{Married},i} - 0.355 X_{MS_{Seperated},i} - 0.033 X_{MS_{Single},i} + 0.227 X_{MS_{Widowed},i}$$

$$-0.0966 X_{LA_{C.only},i} + 0.0744 X_{LA_{others},i} + 0.368 X_{LA_{S.only},i} + 0.361 X_{LA_{parents},i} + 0.23 X_{LA_{SandC},i} + 0.046 X_{TotalC,i}$$

However, multiple linear regression has a few weaknesses. The model is built based on the mean of the variables, so it can be easily affected by outliers which makes the estimation to be inaccurate. Also, if the predictor variables are correlated to each other, the result can also be affected.

(See notations in Appendix A)

6 Results

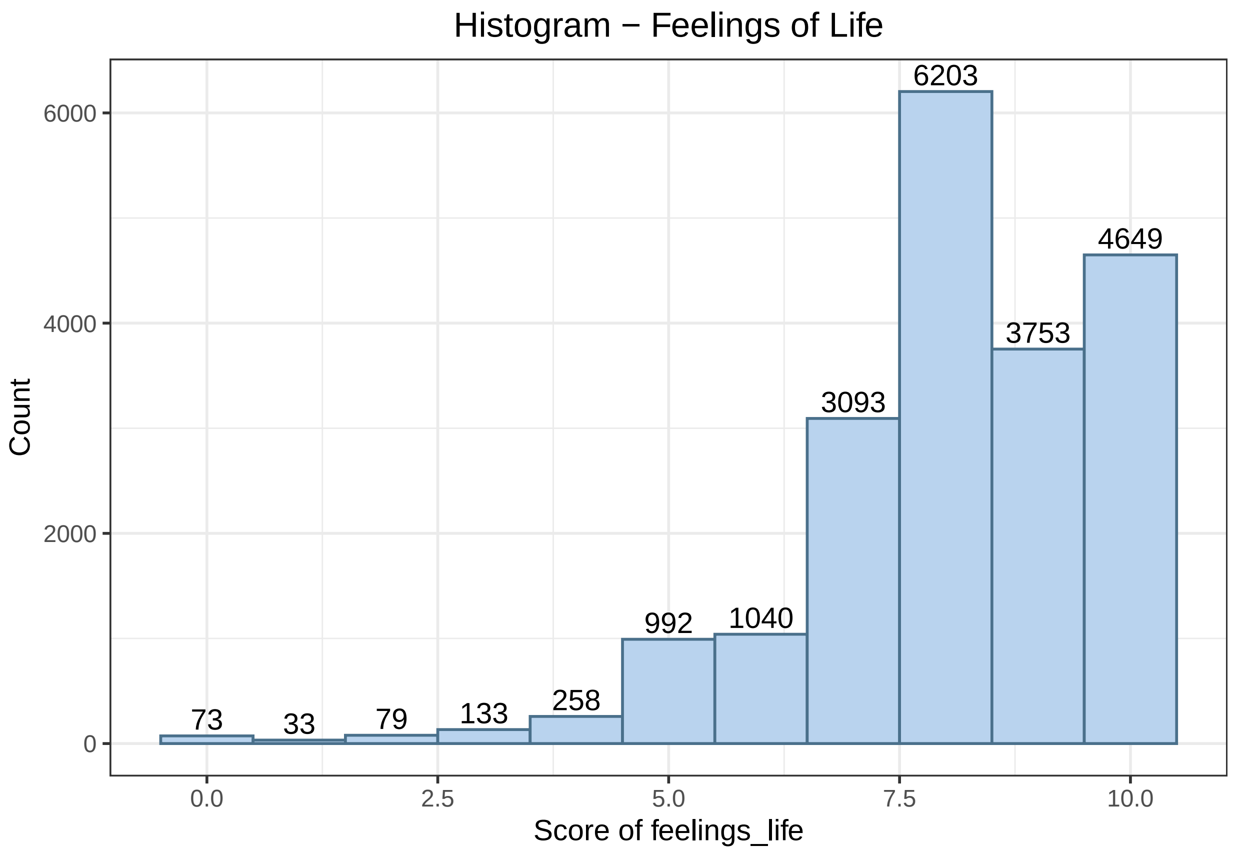

Figure 1: Histogram - Feelings of Life in Counts

Table 1: Summary - Feelings of Life by Marital Status

Marital Status	Mean Score	Median Score	Standard Deviation	Interquartile Range
Divorced	7.679	8	1.909	2
Living common-law	8.224	8	1.427	1
Married	8.433	8	1.414	2
Separated	7.317	8	1.994	3
Single, never married	7.672	8	1.750	2
Widowed	7.952	8	1.841	2

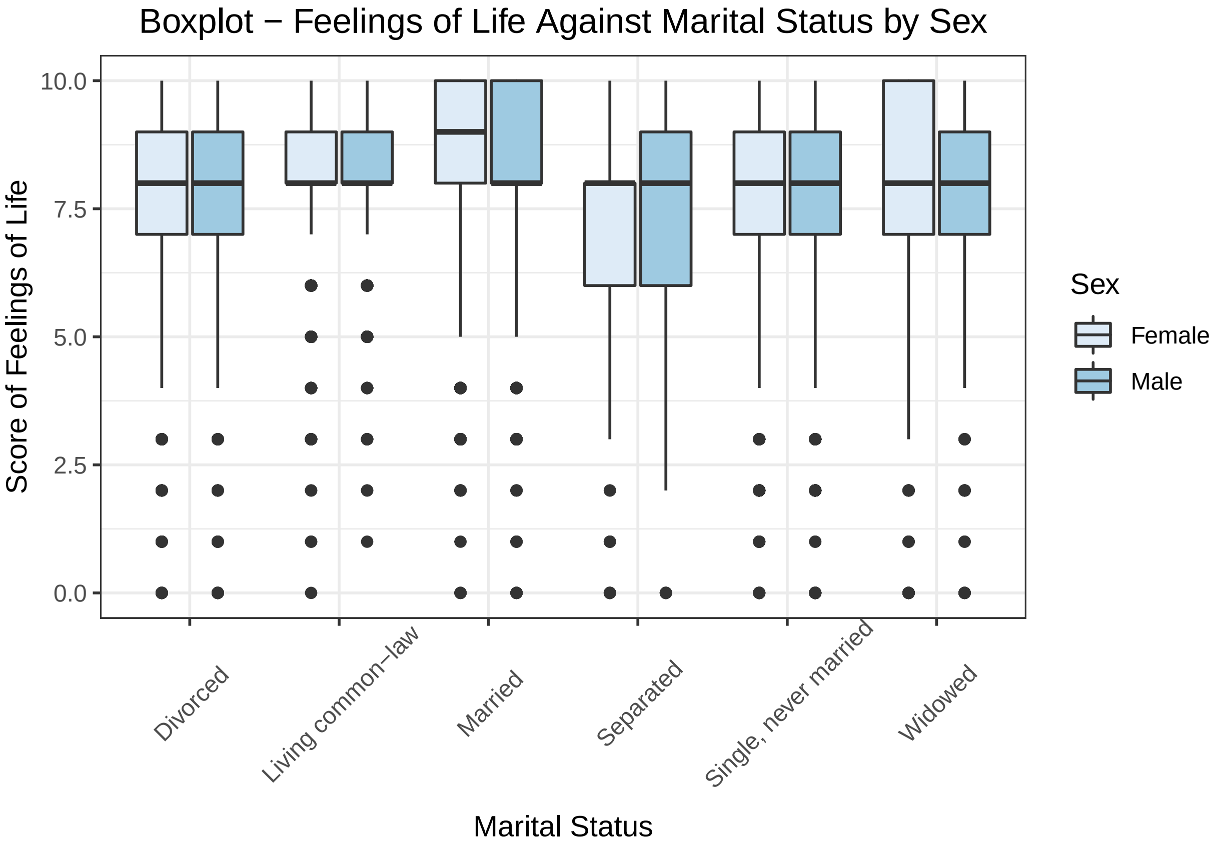

Figure 2: Relationship between Feelings of Life and Marital Status by Sex

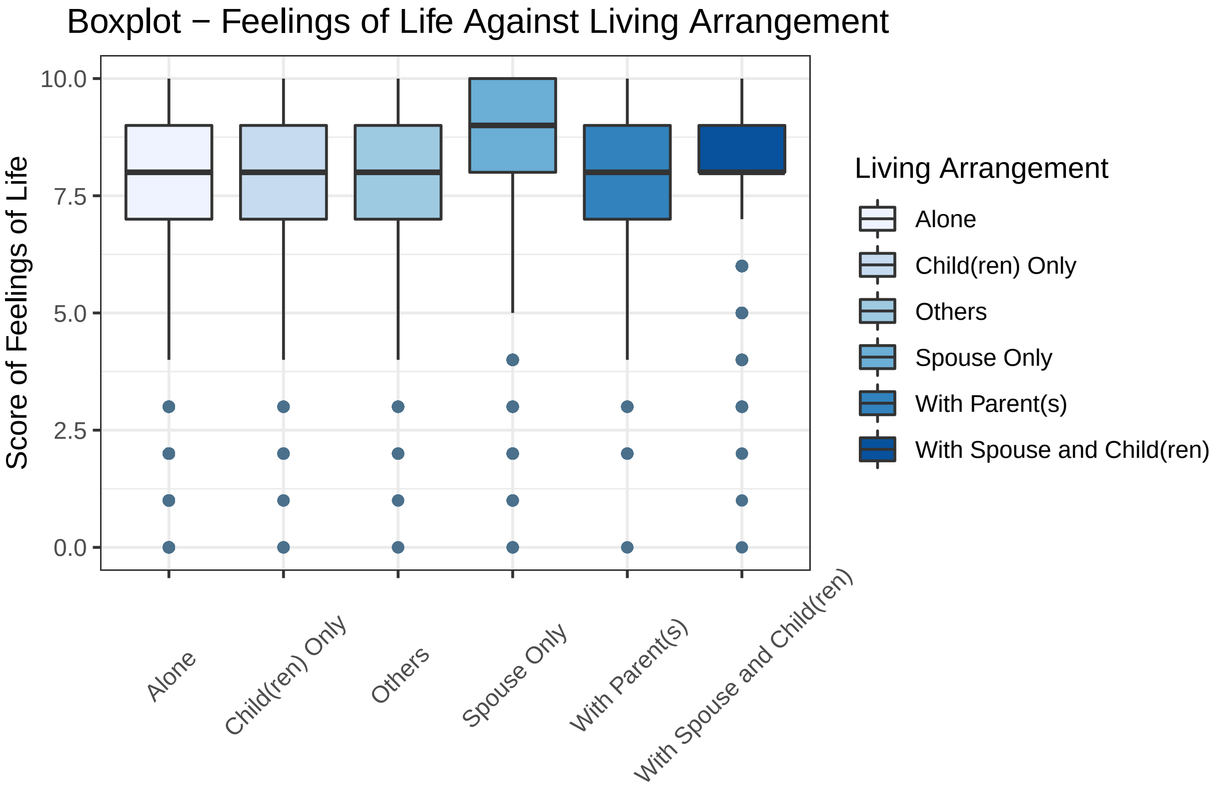

Figure 3: Relationship between Feelings of Life and Living Arrangement

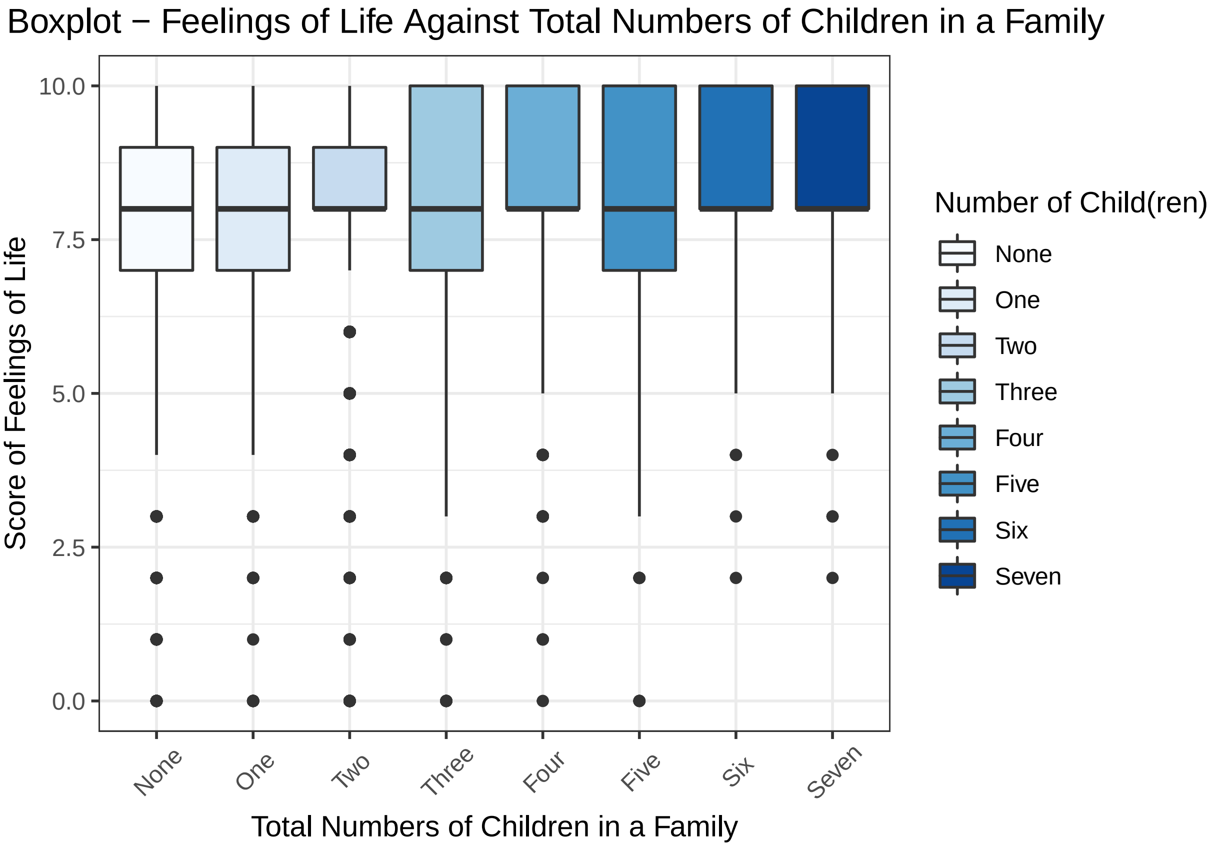

Figure 4: Boxplot - Relationship between Feelings of Life and Total Numbers of Children in a Family

Model 1: Linear Regression Model

```
##
## Call:
## lm(formula = feelings_life ~ sex + marital_status + living_arrangement_type +
##     total_children, data = gss_selected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5980 -0.5986  0.1453  1.3573  2.8802
##
## Coefficients:
##                                                    Estimate Std. Error t value
## (Intercept)                                        7.631117   0.044307 172.232
## sexMale                                           -0.105166   0.023077  -4.557
## marital_statusLiving common-law                    0.275573   0.088136   3.127
## marital_statusMarried                              0.462590   0.082241   5.625
## marital_statusSeparated                           -0.355094   0.074934  -4.739
## marital_statusSingle, never married               -0.032551   0.049353  -0.660
## marital_statusWidowed                              0.226786   0.054269   4.179
## living_arrangement_typeChild(ren) Only            -0.096628   0.055963  -1.727
## living_arrangement_typeOthers                      0.074420   0.055914   1.331
## living_arrangement_typeSpouse Only                 0.367733   0.077945   4.718
## living_arrangement_typeWith Parent(s)              0.361327   0.051310   7.042
## living_arrangement_typeWith Spouse and Child(ren)  0.230228   0.079008   2.914
## total_children                                     0.045535   0.008916   5.107
##                                                   Pr(>|t|)
## (Intercept)                                        < 2e-16 ***
## sexMale                                           5.21e-06 ***
## marital_statusLiving common-law                    0.00177 **
## marital_statusMarried                             1.88e-08 ***
## marital_statusSeparated                           2.16e-06 ***
## marital_statusSingle, never married                0.50955
## marital_statusWidowed                             2.94e-05 ***
## living_arrangement_typeChild(ren) Only             0.08424 .
## living_arrangement_typeOthers                      0.18321
## living_arrangement_typeSpouse Only                2.40e-06 ***
## living_arrangement_typeWith Parent(s)             1.96e-12 ***
## living_arrangement_typeWith Spouse and Child(ren)  0.00357 **
## total_children                                    3.30e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.6 on 20293 degrees of freedom
## Multiple R-squared:  0.05403,    Adjusted R-squared:  0.05347
## F-statistic: 96.59 on 12 and 20293 DF,  p-value: < 2.2e-16
```

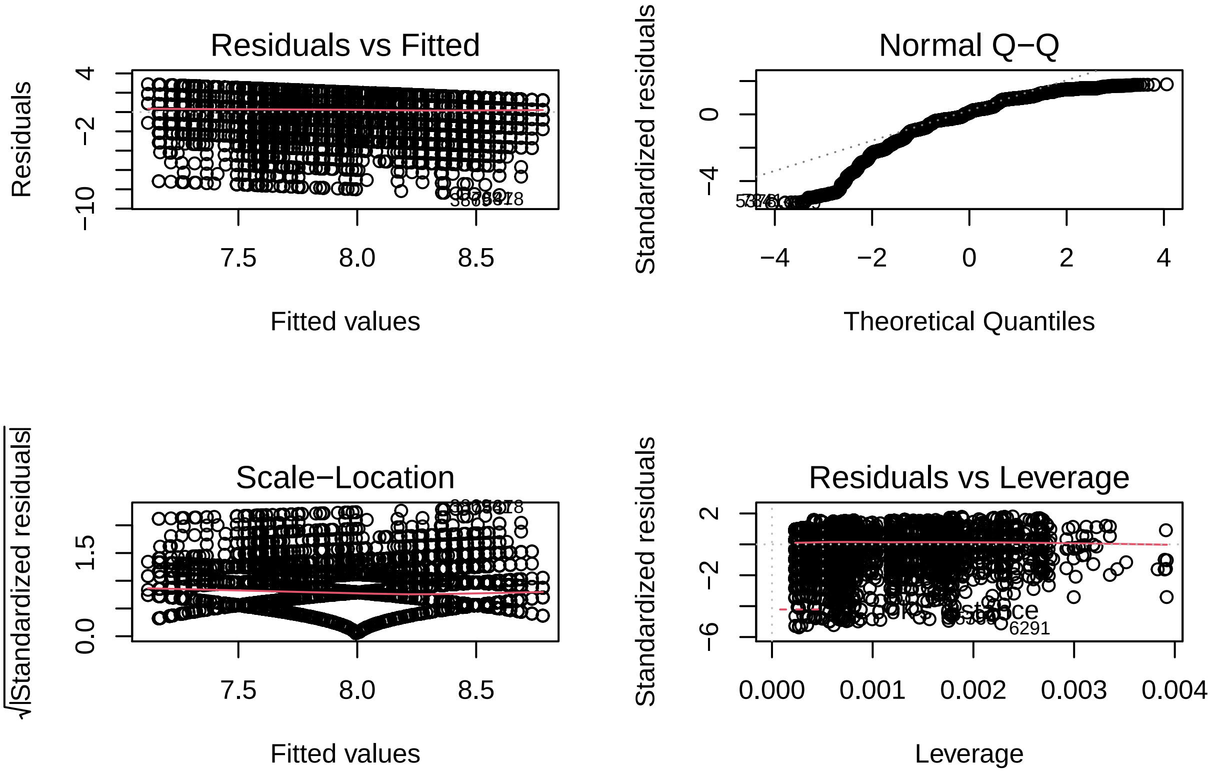

Figure 5: Linear Regression Model Check

7 Discussion

We would first want to look at how the score of the feeling of life is distributed.

Figure 1 demonstrates a histogram based on the score of the feeling of life, which has a left-skewed and multimodal distribution. So, we would expect the mean of the score is greater than the median. We can see that most people score their feeling of life around 7.5 to 8, and only a small number of people rate less than 5.0. Therefore, we have the general idea that most people are highly satisfied with their lives, but only a minority of people have extremely low life satisfaction. We will further discuss different family conditions that impact people's feelings of life.

Then, We would take a look at the summary table of the feelings of life grouped by the marital status.

From Table 1, the mean satisfaction score among all marital status is all above 7. And we can see that people who are married and living common-law have a relatively high score of around 8.3. And the people who are separated from their partner have the lowest mean score. Therefore, most people who have significant others have higher scores of satisfaction.

Also, people who are married and living common-law have the smallest standard deviation of around1.4 and IQR value between 1 and 2. This means that the spread of the scores is more concentrated at the high level of score compared to people while another marital status. Therefore, it presents that marriage has a positive effect on people's feelings of life. Interestingly, for those people who are widowed, we expect them to have a low level of satisfaction. Instead, their satisfaction with life is even higher than those who experience divorce or separation.

We would further look at how marital status impacts the feeling of life between men and women.

Figure 2 shows that the all marital status toward the score of feelings of life have a left-skewed distribution. And for male, the median score of the feelings of life are almost identical between all marital status. However, for females, we can see that the married female has an especially higher median level of life satisfaction compared to those in other status. It showed that the change in marital status has relatively less impact on the men compared to the females.

Both female and male who are living common law have the smallest spread, which means the scores are more concentrated at a range of values. In contrast, males who are separated and females who are widowed have a relatively large IQR. It describes people's scores of satisfaction are distributed and varied in a wide range of values.

The outliers are also shown in all marital status. And people who are living common-law have a significantly large number of outliers. It is also interesting to know that 25% of females who are widowed score their feeling of life as the maximum value 10. Although a group of women who are widowed and seem to have an unfortunate life, they can still be very fulfilled in their life.

We now want to look at how the living arrangement would affect the score of the feeling of life. First, we would reorganize the living arrangement into six main categories which are Alone, only Child(ren), only spouse, with parents(s), with spouse and child(ren), and others.

Figure 3 shows all living arrangements with respect to the score of the feeling of life have left-skewed distribution. We can see that the median score for people living with a spouse only is higher than any other living arrangement. This shows that living with parents or children would affect people's feelings of life. For people who live with only children or only parents, we can observe a similar median value and IQR value. This presents that living with only children or only parents has a similar impact on their score of life satisfaction. Regarding people who live with both spouses and children, we can see that they have a smaller IQR compared to others. This suggests that their variation of satisfaction score is smaller and their range of score is concentrated at a high level around 8 to 9.

After looking at the living arrangement and marital status, we would also want to see if the total number of children in a family would differ in people's scores of feelings of life.

In Figure 4, we can see that people with any number of children in a family have a 25 % quartile greater than the score of 6.5, which shows that there are only less than 25% of people who have the score of feelings of life lower than 6.5. For people who have six to seven children in a family, the score is generally high above 8 and the score of the IQR of the score is also small. This suggests that most people with six or seven children would rate a high score of life satisfaction and have less variation and spread of the score.

Lastly, we would want to know if there is a linear relationship between our predictors (sex, martial sttus, living arrangement, and total number of children) and the response feelings of life.

In linear regression model, which notated as **Model 1**, since the p-value for most predictors are smaller than 0.05 ,this suggests that we have a significant linear relationship between the predictors and response. The interpretation for the model that has four predictors:

- marital status (married): The additional average scores of feelings of life is 0.46 if having people who are married, holding other factors constant.
- living arrangement (spouse only): the average difference in the score of feelings of life is 0.37 having living arrangement to be With Spouseonly, and others holding other variables constant (we would have similar interpretation as above for other categorical variables)
- total children: we can see the average increase of 0.46 scores of feeling of life for an one unit increase in the total number of children.

Overall, we demonstrate some data analysis and diagnosis of the gss dataset. Since the mean and median of the score of feelings of life are 8.09 and 8 (See Appendix B), we can see that most people have a high score of feelings of life. The results show that married people will generate a higher level of life satisfaction, especially found in married women. Similarly, for people who live only with a spouse, they are more likely to establish a better life satisfaction. And also based on our regression model, we can conclude a positive relationship between marriage and the feelings of life. Although people who live with a spouse are more satisfied with life than those who live with both children and a spouse, we can see that having more children in a family also provides a higher level of life satisfaction. This might be related to those people who are widowed but having children because people who are windowed also rate a higher score of life satisfaction. However, we might need to establish further analysis. It is important to do this analysis because the family condition closely links to our life, which is closely related to people's feelings of life. Therefore, the study could also provide a general overview for those people who are unmarried.

Besides, since the dataset is conducted from the telephone survey, it will involve potential bias that the survey might only attract a certain group of people. Some people might not be willing to answer sensitive questions that are contained in the survey. Therefore, the dataset contains lots of missing value which we have removed before we conduct the model and analysis.

8 Weaknesses & Next Steps

8.1 Weaknesses

For `living_arrangement`, we categorized them into 6 groups instead of 12 in order to make an easier plot for us to interpret. Therefore, we are only able to see the behavior of joint variables instead of seeing the behavior for each of them separately. Then, if all the variables have the same trends for their datas, we may have the problem of multicollinearity (Frost, 2020). Additionally, some of the variables in the joint variables may have affected the others if they do not have similar trends for their datas. In such cases, our model may not be accurate enough. Also, we did not take outliers and influential points into consideration which may also affect our result.

8.2 Next Steps

Our next step is to find a better option to handle categorical variables. For example, we can apply an anova or manova table for analyzing categorical variables (Katerina, 2018), and we can also use variance inflation factor to avoid multicollinearity. Additionally, we improved the methodology of doing the survey. The main methodology of taking this survey is through telephone. Since only 63% Canadians have a landline at home (In CBC News, 2018), a survey through phone number or online might be a better choice. The result of the survey can be used as a reference for those adults who are uncertain about marriages. They may not be that anxious about marriage if they are aware that marriage can actually improve the level of life satisfaction.

9 References

9.1 GSS Dataset & User Guide

Ruus, L. (2017). General social survey on Family (cycle 31), 2017. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User's Guide. (2020). In Statt=istics Canada. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc /GSS31_User_Guide.pdf

9.2 Other References

Bache, S. M. & Wickham, H (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. Retrieved from https://CRAN.R-project.org/package=magrittr

Department of Statistics, Columbia University. (n.d.). Colors in R. In stat.columbia.edu. Retrieved from http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

Frost, J. (2020). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. Retrieved from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Horn, B. (2013). RColorBrewer Palettes. In Applied R Code. Retrieved from http://applied-r.com/rcolorbrewer-palettes/

Horowitz,J.M., Graf, N & Livingston, G. (2019). Marriage and Cohabitation in the U.S.. Retrieved from https://www.pewsocialtrends.org/2019/11/06/marriage-and-cohabitation-in-the-u-s/

Holtz, Y. (2018). Ordering boxplots in base R. In R Graph Gallery. Retrieved from https://www.r-graph-gallery.com/9-ordered-boxplot.html

Just 56% of Alberta households still have a landline - the lowest rate in the country | CBC News. (2018). Retrieved from https://www.cbc.ca/news/canada/calgary/alberta-landline-use-drops-below-56-per-cent-1.4946343

Katerina. (2018). Working with categorical data. Retrieved from https://medium.com/whats-your-data/working-with-categorical-data-c338122b9521

Kassambara. (2020). "Ggplot Title, subtitle and caption". In Datanovia. Retrieved from https://www.datanovia.com/en/blog/ggplot-title-subtitle-and-caption/.

Mason, R. (2018). Misconceptions of the Open Society. Retrieved from https://wespeakfreely.org/2018/06/05/misconceptions-open-society/

Parker, K. & Stepler, R. (2017). As U.S. marriage rate hovers at 50%, education gap in marital status widens. Retrieved from https://www.pewresearch.org/fact-tank/2017/09/14/as-u-s-marriage-rate-hovers-at-50-education-gap-in-marital-status-widens/

R Core Team. (2020). R: A language and environment for statistical computing.R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Schork, J. (n.d.). Change Legend Title in ggplot2 (2 Example Codes) | Modify Text of ggplot Legends. In Statistics Globe. Retrieved from https://statisticsglobe.com/change-legend-title-ggplot-r

Schork, J. (n.d.). Rotate ggplot2 Axis Labels in R (2 Examples). In Statistics Globe. Retrieved from https://statisticsglobe.com/rotate-ggplot2-axis-labels-in-r

Weedmark, D. (2018). The Advantages & Disadvantages of a Multiple Regression Model Retrieved from https://sciencing.com/advantages-disadvantages-multiple-regression-model-12070171.html

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. In Springer-Verlag New York, 2016.

Wickham, H., François, R., Henry, L. & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. Retrieved from https://CRAN.R-project.org/package=dplyr

Xie, Y. (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Xie, Y. (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Xie, Y. (2016). bookdown: Authoring Books and Technical Documents with R Markdown. Chapman and Hall/CRC. ISBN 978-1138700109

Xie, Y. (2020). bookdown: Authoring Books and Technical Documents with R Markdown. Rpackage version 0.21.

Xie, Y. (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. In R package version 1.30.

10 Appendix

10.1 Appendix A

Y_i represents the dependent variable that we need to estimate - feelings of life (life satisfaction).

X_i represents the predictor variables in our model.

For X_1 to X_i, each of the X_i takes the value 0 or 1:

$X_{male,i}$ is 1 if the input is Male and 0 otherwise;

$X_{MS_{LCL},i}$ is 1 is the input is Living Common-law;

$X_{MS_{Married},i}$ is 1 if the input is Married and 0 otherwise;

$X_{MS_{Seperated},i}$ is 1 if the input is Separated and 0 otherwise;

$X_{MS_{Single},i}$ is 1 if the input is Single, never married and 0 otherwise;

$X_{MS_{Widowed},i}$ is 1 if the input is Widowed and 0 otherwise;

$X_{LA_{C.only},i}$ is 1 if the input is Child(ren) only and 0 otherwise;

$X_{LA_{others},i}$ is 1 if the input is Others and 0 otherwise;

$X_{LA_{S.only},i}$ is 1 if the input is Spouse Only and 0 otherwise;

$X_{LA_{parents},i}$ is 1 if the input is With Parent(s) and 0 otherwise;

$X_{LA_{SandC},i}$ is 1 if the input is With Spouse and Child(ren) and 0 otherwise;

$X_{TotalC,i}$ represents the input value of total children.

β_0 is the constant term which represents the y-intercept at time zero. It is the value of y when every $X_i = 0$.

β_i is the regression coefficient:

β_1 represents the average difference in Y between male and female holding other variables constant

β_2 represents the average difference in Y between having marital status to be living common-law and others holding other variables constant

β_3 represents the average difference in Y between having marital status to be married and others holding other variables constant

β_4 represents the average difference in Y between having marital status to be Separated and others holding other variables constant

β_5 represents the average difference in Y between having marital status to be Single, never married and others holding other variables constant

β_6 represents the average difference in Y between having marital status to be Widowed and others holding other variables constant

β_7 represents the average difference in Y between having living arrangement to be Child(ren) only and others holding other variables constant

β_8 represents the average difference in Y between having living arrangement to be others and other groups holding other variables constant

β_9 represents the average difference in Y between having living arrangement to be Spouse Only and others holding other variables constant

β_{10} represents the average difference in Y between having living arrangement to be With Parent(s) and others holding other variables constant

β_{11} represents the average difference in Y between having living arrangement to be With Spouse and Child(ren) and others holding other variables constant

β_{12} represents the average difference in Y for each one unit change in total children holding other variables constant.

10.2 Appendix B

Table 2: Summary - Feelings of Life

Mean Score	Median Score	Standard Deviation	Interquartile Range
8.094	8	1.645	2

10.3 Appendix C

Code and data supporting this analysis is available at: https://github.com/YiyangHuang028/STA304-Problemset-2.git