Report on the Life Feeling Level of Canadians and the Influence Factors

Based on the Canada General Social Survey(2017)

Author: Runqi Bi, Guangyu Du, Qiu Jin, Xingkun Yin

Date: 19 October 2020

Abstract

This report will focus on if some specific factors will influence individuals' life feeling, and if so, how does them influence the life feeling level respectively so that we can know how to help the Canadians live better. We used the data from the Canadian general social survey (GSS), which was collected by Statistics Canada in the year 2017. In general, we conclude that all selected elements, other than individual income, will have a positive influence on life feeling level while the effect of personal income is ambiguous. That means, with the individual income increases, the life feeling level will not necessarily increase.

Introduction

Canada has always been one of the countries who has the highest average life feelings level in the world. Our research team wants to determine if different factors can influence life feelings level, and if they do, then how? These factors can also be used as a guideline for new government policies regarding the improvement of the general life quality of Canadian. With knowing the answer to this question, we can help the Canadian to improve their living quality and life feeling level. The data we used is from Statistics Canada, they do general social surveys(GSS) annually, and for our report, we will focus on the year 2017. Further information and the details of the GSS(2017) will be discussed in the following part.

To do better analysis, we did not use all provided factors that in the GSS. In contrast, we picked specific elements that we think is related to the life feeling level the most. To determine which variable should be chosen, we used the **Akaike information criterion**(AIC). The AIC is an estimator of in-sample prediction error and the relative quality of statistical models for a given set of data. It is a helpful tool that can help us to decide the appropriate model. After calculation, we determined that six factors should be taken into account, which is: self-rated health, self-rated mental health, family income, personal income, education level, and total children. These factors are considered as explanatory variables, and the life feeling level is the response variable. To fit the variables, we used a **Multiple Linear Regression Model**(MLR). MLR is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

In general, the model shows us almost all elements will have a positive effect on the life feeling level more or less, except the personal income. A higher personal income does not necessarily lead to a higher life feeling level.

The structure of this research includes[1]:

1. General Data and Selected Data
2. Model
3. Results
4. Result Discussion
5. Next Steps and Future Works

[1]* Full code and data supporting this analysis is available at githublink: https://github.com/troyyxk/GSS_Report

Data

In this part, we want to introduce more about the GSS data we used for analyzing. We will discuss the process of data collection, the strengths and weakness of the data, and its primary objectives.

Key Features, Strengths, and Weakness

There are two primary objectives of the General Social Survey:

1. Collect data on social trends, which will help monitor changes in Canadians' living conditions and well-being over time; and
2. Provide information on specific social policy issues of current or emerging interest.

In the past few decades, our understanding of Canadian families has greatly improved. However, the future of the family is still uncertain. How many families are there in Canada? What are their characteristics and socioeconomic conditions? What does a family look like at different stages of life? How common are single-parent families or single-parent families? Family Questions GSS will provide answers to these and many other questions.

Data collection

The **target population** for the 2017 GSS included all persons who are at least fifteen-years-old in Canada[2].

A **stratified sampling method** is used for sampling, and each of the ten provinces is divided into different levels based on geographic area. A total of 27 strata was formed, including ten more groups formed by non-CMA[3] areas of each of the ten provinces. A **simple random sample** without replacement of records was next performed in each stratum.

The **sample frame** was created with two different components:

1. Lists of phone numbers in use (both landline and cellular numbers) available to Statistics Canada from various sources[4];
2. The Address Register (AR): List of all apartments in the ten provinces.

The Address Register (AR) is used to group all the phone numbers associated with the same effective address. If multiple phone numbers are attached to one record, the first phone number is considered the best phone number to reach the household[5]. For the 2017 GSS, 91.8% of the selected telephone numbers reached eligible households which have at least one person who is 15 years old or older. Non-eligible households will not be included in further research.

The **target sample size**, which is also the desired number of respondents for the 2017 GSS was 20,000 while the actual number of respondents was 20,602. The data was collected via computer-assisted telephone interviews (CATI). All interview calls are conducted from approximately 9:00 am to 9:30 pm. Monday to Friday. The interview time is also scheduled from 10:00 am to 5:00 pm. on Saturdays and from 1:00 pm to 9:00 pm. Those who initially refused to participate will be contacted up to two more times to explain the importance of the survey and encourage them to participate. For the situation where the interviewer is inconvenient to call, an appointment is arranged to call back at a more convenient time. In the absence of a household, many callbacks were made. The overall response rate of GSS in 2017 was 52.4%.

The primary source of **non-sampling errors** in the survey is the impact of non-response on the survey results. The degree of non-response ranges from partial non-response[6] to non-response at all. Total non-response happened when the interviewer was unable to contact the respondent; hence no family member could provide information, or the interviewee refused to participate in the survey. By adjusting the weight of the households who responded to the survey, the households who did not respond were compensated. In most cases, when

[2]Excluding: 1. Residents of the Yukon, Northwest Territories, and Nunavut; and 2. Full-time residents of institutions.
[3]A non-CMA area is small urban areas with a population of less than 100,000.
[4]Telephone companies, censuses, Etc.
[5]Sort by the source and type of phone numbers; fixed phone first, mobile phone last
[6]Failure to answer one or several questions

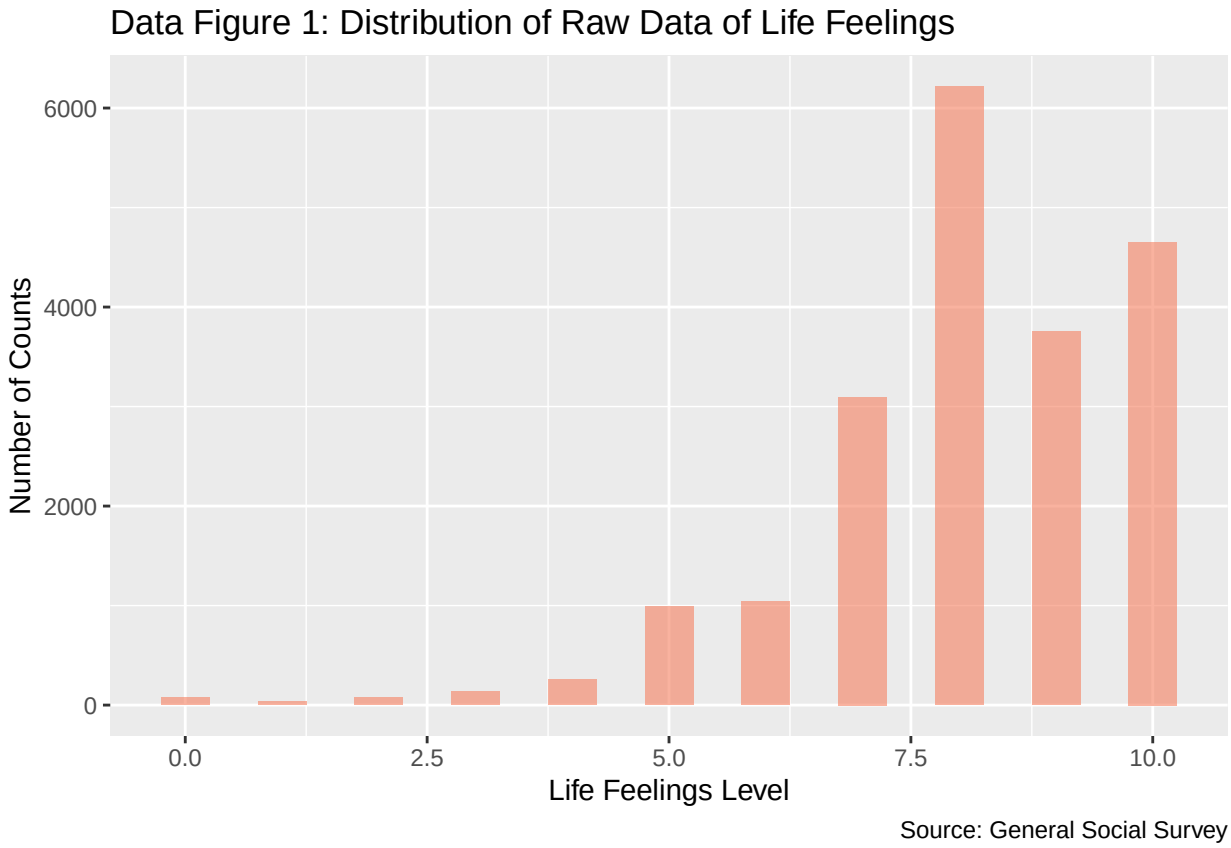
the interviewee does not understand or misinterpret the question, refuses to answer the question, or cannot recall the requested information, the survey will be partially unanswered.

An Overview on Raw Data

Table 1: Summary of Raw Data

Total Observation	Total Variables	Year of Observation
20602	81	2017

The following is a plot of the raw data for the feeling of life. There exist an obvious left-skewed pattern on the plot, which means the majority of Canadian people are satisfied with their life with a high rating on life feelings.



From the table below, we can tell the average of life feeling score reaches 8.09 out of 10, with a minimum rating of 0 and a maximum rating of 10. Also notice that from the raw data, there are 271 samples which did not give a feedback on life feeling. We will omit these samples when selecting data.

Table 2: Summary of Life Feelings base on Raw Data

Min	Median	Average	Max	Count of NAs
0.00	8.00	8.09	10.00	271

Data selection

Before analyzing starts, we need to process the raw data by cleaning it. By data cleaning[7], we cleaned up the data, renamed the variables, removed NA values and modified some string values into numeric values.

In order to select proper explanatory variables that have a strong relationship with the response variable, which is life feelings in our research, we built up several testing models and determined the importance of

[7]* Data cleaning code is supported by Rohan Alexander and Sam Caetano, see full citation in reference

each explanatory variable base on the value of its corresponding coefficient.

The result of tests shows that the age and gender of the sample have little relationship with his/her life feeling since from the linear regression model with only age and gender as the explanatory variables, the values of corresponding coefficients are extremely small. Similarly, we found out that the education level, income level of both family and respondent, physical and mental health level, number of children are significant variables related to the life feeling of samples. Therefore, we select these attributes as the explanatory variables in this study.

Questionnaire

The questionnaire design is overall very clear and efficient. However, some questions may not receive consistent answers and therefore have bias within them. Let's look at the question "In general, would you say your mental health is. . . ?" at position 679. This question asks people to report their own mental health, but people's self ratting on mental health varies a lot depending on the mood. An emotional person may be very motivated at the beginning of the day and rate himself "Excellent" mental health but if the day does not go smoothly, she might get a little depressed and report his mental health with a "Poor" status. This problem is less obvious on questions like "In general, would you say your health is. . . ?" at position 678 which probably contains less bias in its response as people's health status is typically more consistent than and external factors like mood usually do not have a direct short-term impact on people's view on their own health condition.

Model

Multi-linear regression model is chosen as the model of this research, since we are studying on a continuous response variable and we assume that the variables are following approximate normal distribution base on the central limit theorem. However, we were concerned that too many explanatory variables could be over-fitting the model. In order to find the model that best fit our data, we generated three models with different amount of explanatory variables.

Model 1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

Model 2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6 + \epsilon$$

Model 3:

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + +\beta_6 X_6 + \epsilon$$

where:
Y : life feelings (continuous variable);
X_1 : number of total children (continuous variable);
X_2 : self rated health level (indicator variable);
X_3 : self rated mental health level (indicator variable);
X_4 : education level (indicator variable);
X_5 : respondent income level (indicator variable);
X_6 : family income level (indicator variable);
b_i : corresponding coefficients of each variables, where b_0 is the intercept.

Then we compare the AIC of each model and select the best fitted model with the smallest AIC. Model 1 has an AIC of 13178.36, Model 2 has an AIC of 13199.04 and Model 3 has an AIC of 13433.8. Therefore, Model 1 is the best fit model for our data.

The software we use is Rstudio and lm (linear regression) module.

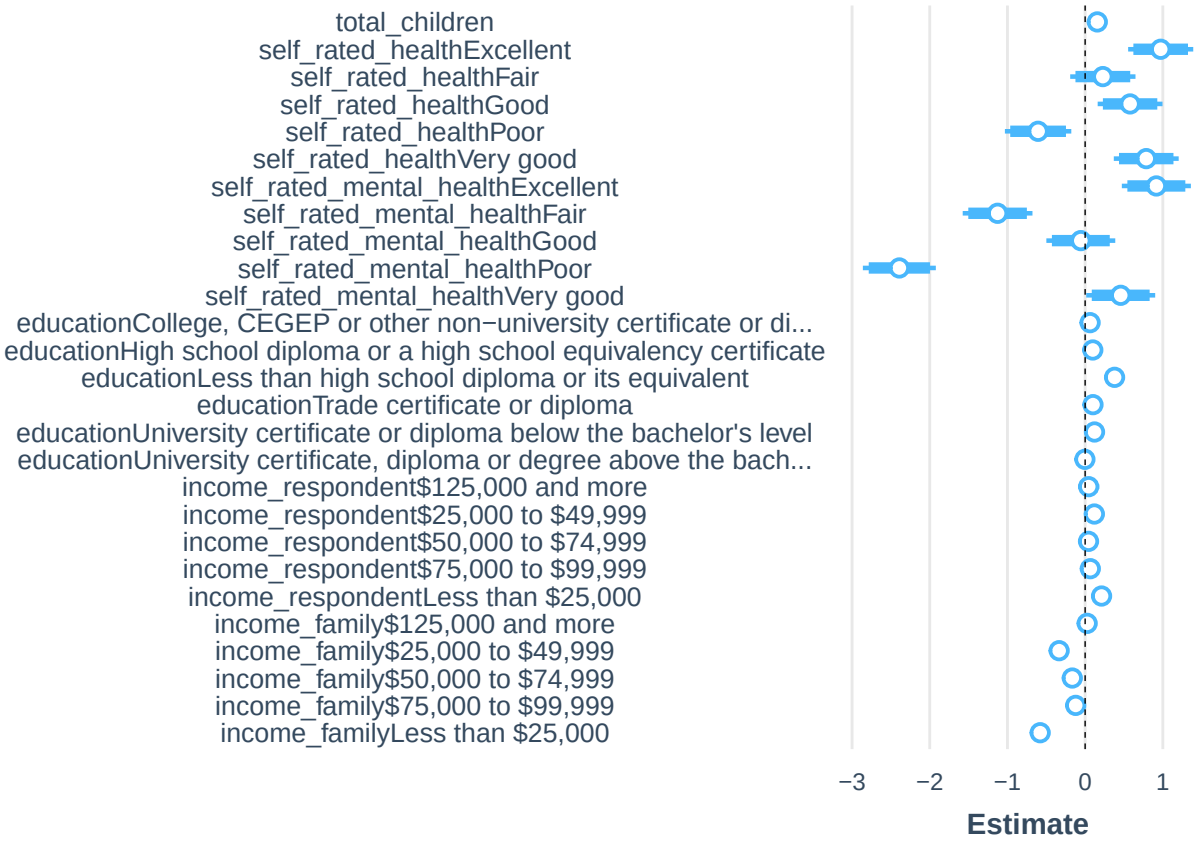
Table 1: Estimates of Variables and Intercept

	x
(Intercept)	6.9211847
total_children	0.1054778
self_rated_healthExcellent	0.9726130
self_rated_healthFair	0.2279409
self_rated_healthGood	0.5787963
self_rated_healthPoor	-0.6053973
self_rated_healthVery good	0.7865648
self_rated_mental_healthExcellent	0.9167382
self_rated_mental_healthFair	-1.1269634
self_rated_mental_healthGood	-0.0554792
self_rated_mental_healthPoor	-2.3915627
self_rated_mental_healthVery good	0.4579032
educationCollege, CEGEP or other non-university certificate or di...	0.0638997
educationHigh school diploma or a high school equivalency certificate	0.0985531
educationLess than high school diploma or its equivalent	0.3790626
educationTrade certificate or diploma	0.1009626
educationUniversity certificate or diploma below the bachelor's level	0.1209088
educationUniversity certificate, diploma or degree above the bach...	-0.0028596

	x
income_respondent$125,000 and more	0.0458859
income_respondent$25,000 to $49,999	0.1187998
income_respondent$50,000 to $74,999	0.0435575
income_respondent$75,000 to $99,999	0.0668430
income_respondentLess than $25,000	0.2114685
income_family$125,000 and more	0.0261981
income_family$25,000 to $49,999	-0.3345919
income_family$50,000 to $74,999	-0.1653278
income_family$75,000 to $99,999	-0.1218040
income_familyLess than $25,000	-0.5786692

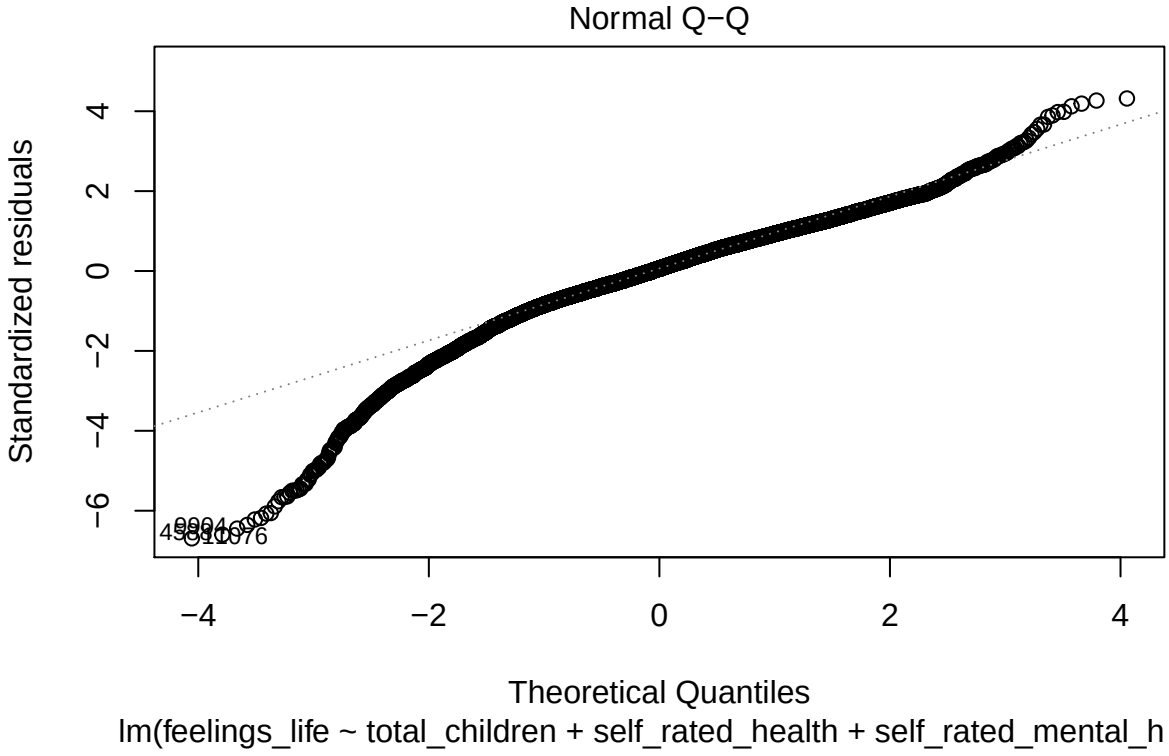
By observing the model and the above table, we can clearly notice that health has a great influence on people's happiness index. Among them, thinking that one's mental health is very poor has an inverse correlation with the impact of life feeling level, there are as much as about -2.39. And think that the estimate of their mental health is fair is around -1.13 for life feeling level, which is also very high. On the contrary, the estimate that thinks one's overall health is excellent is 0.97, which is positively correlated with the life happiness index. By observing the estimate data, we can also see that when the family income is too low, when it's less than 25,000, it has a great impact on the life feeling. Its estimate is about -0.58, which is negatively correlated. The estimate for a family income between 25,000 and 49,999 is around -0.33. Although it is not as low as a family income of less than 25,000, it is also significant data. Regarding academic qualifications, it is surprising to see that for people with lower academic qualifications than high school, the estimate for life feeling is 0.38, which is a positive correlation. By observing the estimates of the rest of the academic qualifications, it is found that they have no obvious influence on life feeling.

Model Figure 1: Estimates of all explanatory variables with Confidence Intervals.



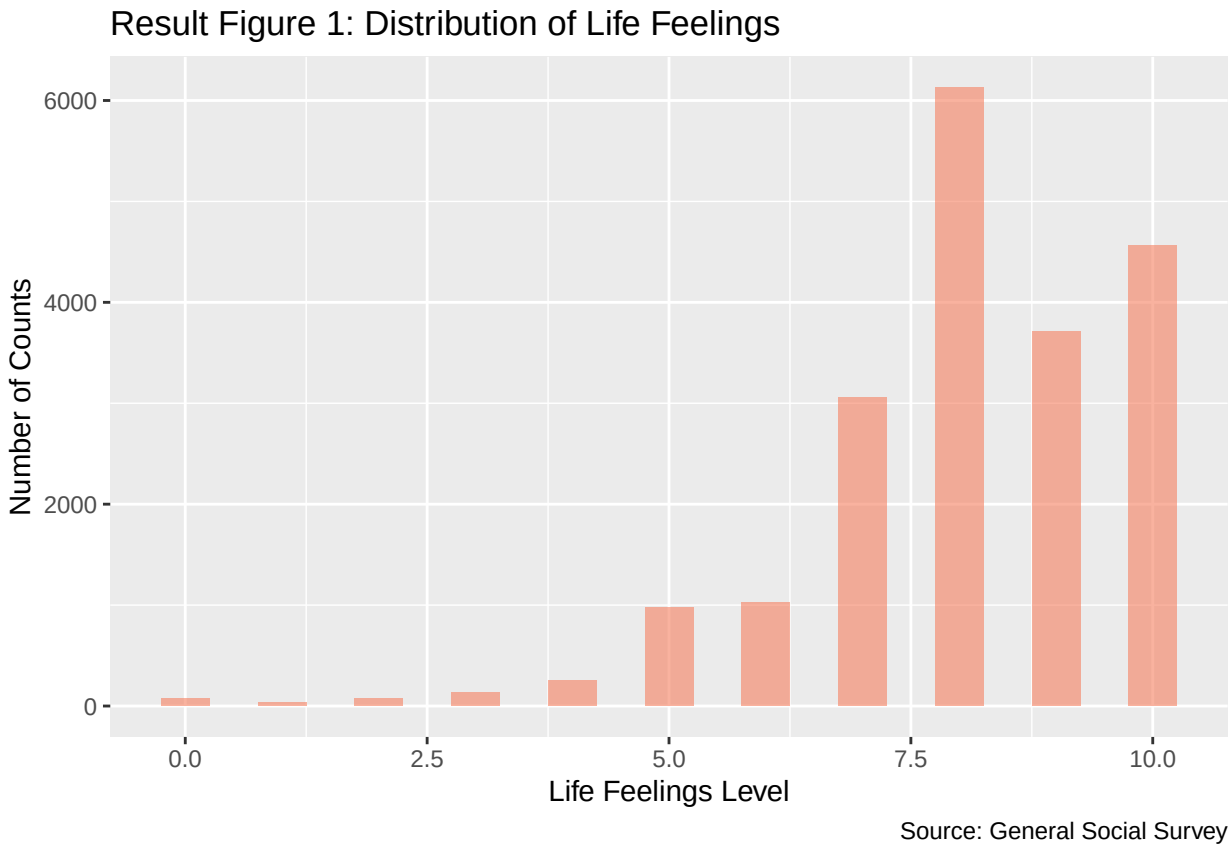
The graph above is the visualization of all the estimates of our model with 90% confidence intervals. The further the estimate is away from zero, the stronger influence will be made to sample's life feeling with one unit change of the corresponding variable. For example, a sample with poor self-rated mental health will decrease his/her life feeling score by approximately 2.3.

Model Figure 2: Checking normality of the model using qq-plot.

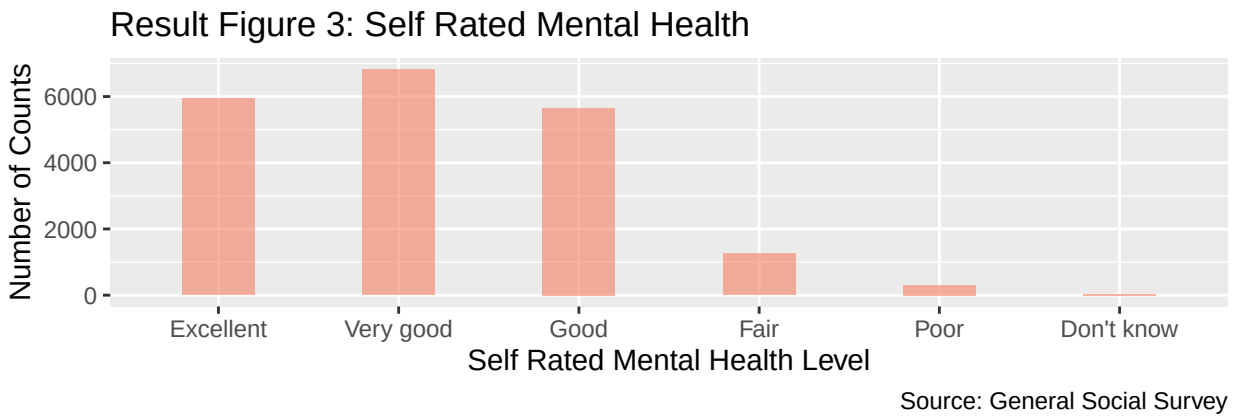


QQ plot is a great tool for checking the normality of linear regression models. The model would follow normal distribution if most of the data gathered along the normal line. From the normal qq-plot of the model above, we can tell most of the data are following the normal line, which means our data satisfy the normality assumption. Therefore, the model we generated fits the data moderately.

Results

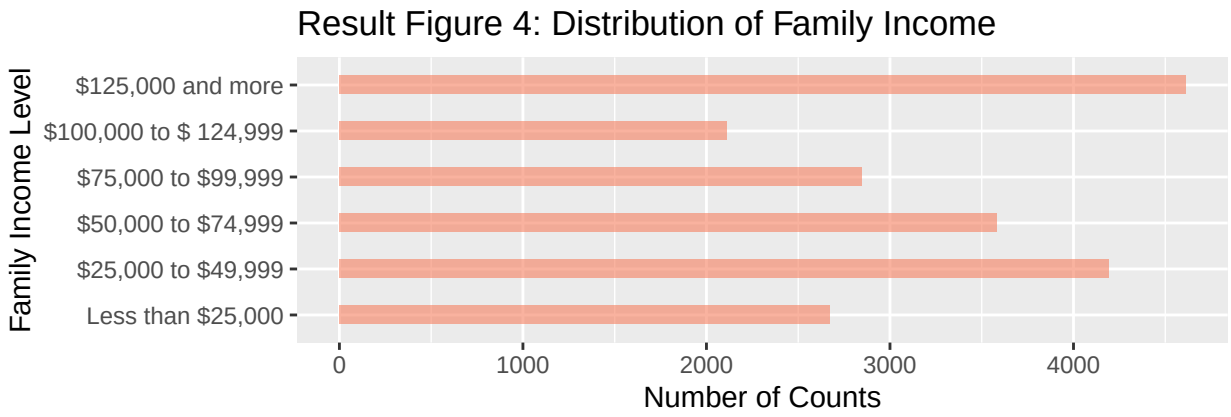


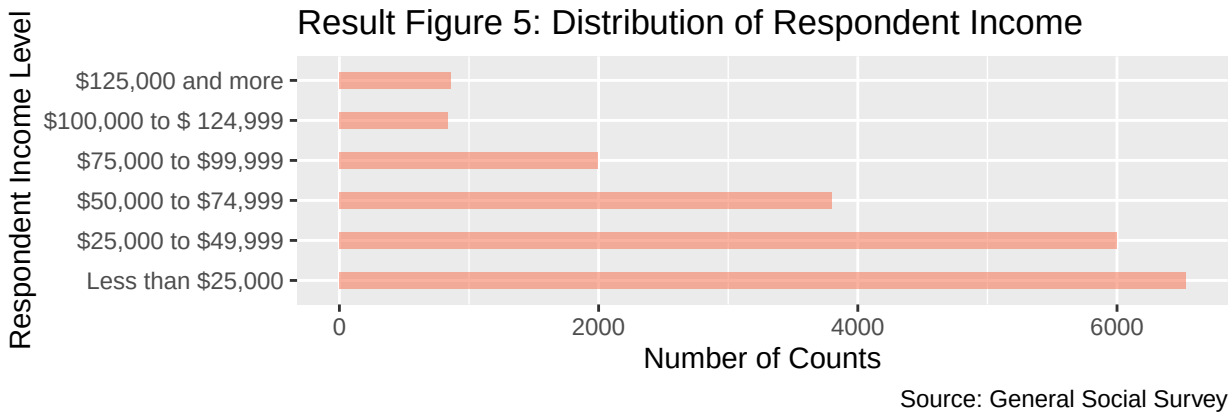
Overall, from Figure 1, we can find that the life feelings level of most people is more than 7.5, and only a few people are less than 5. It is worth noticing that the second highest score for life feelings level is 10, and a very small number of people rate their life feelings level 0.



Source: General Social Survey

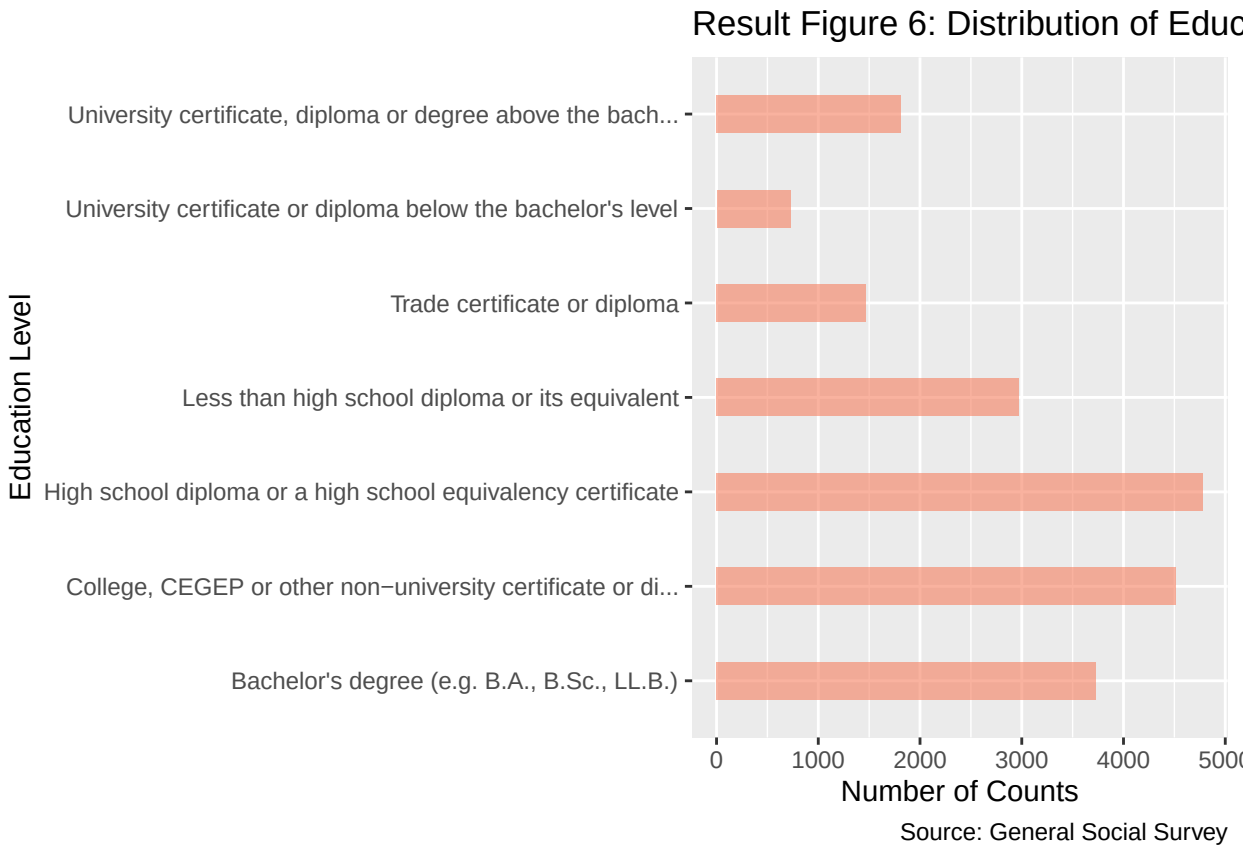
From Figure 2 and Figure 3, we can find that most people think that their overall health and mental health are fair or higher. This result is very similar to people's rate of life feelings level. People who think they are healthy are the majority, and those who are satisfied with life are also the majority.





Source: General Social Survey

The results shown in Figure 4 and Figure 5 are interesting. We can see that the distribution of household income is relatively even, compared to the level of personal income. The income of most households is higher than 125,000, but other income levels are not far below the highest level. However, looking at the personal income distribution shown in Figure 5, we can see that the higher the income, the smaller the proportion of people. The highest proportion of people is from the lowest income level, less than 25,000.



Through Figure 6, we can observe that a large proportion of people have a high school degree or below. At the same time, people with college degree or university degree also account for a lot of people. In general, in terms of academic qualifications, there is no obvious tendency. The number of people with low academic qualifications is similar to high academic qualifications.

Result Discussion

This survey collected data from all Canadian above age 15 excluding residents of the Yukon, Northwest Territories, and Nunavut and full-time residents of institutions. This is a stratified survey with each province as its own strata. As our focus is on the general life feeling, we ignore the provincial factor.

Weakness

Subjective questions like "In general, would you say your mental health is...?" as explained above in the data section are one of the main sources of bias in this dataset in our opinion. Therefore, we try to minimize the use of subjective questions. Among the six factors and their corresponding questions we used, only the self-rated mental health status is a rather subjective question. Self-rated health status may be ambiguous and people could argue it is somewhat subjective but we believe it is a rather objective question as the nature of human health for the most parts does not vary a lot. The other factors, number of children, education level, respondent income and family income.

Findings from result

The result above shows some interesting points.

The total number of children shows a slight positive correlation with the feeling of life with a coefficient of 0.1. This is against our prediction as we thought the more children the respondent has, the more tired the respondent would be and more negative one would feel about life.

The correlation between health and feeling towards life is as expected as the better respondent rate his own health, the higher the life feeling is. All indicators of self-rated health explanatory variables(excellent, good and fair) except "poor" show positive relation with the feeling about life. However, the relationship between self-rate mental health and feeling towards life is confusing. We expected the correlation between self-rated mental health and feeling about life to be the same as the previous one but the explanatory variable of "good" and "fair" show a negative correlation. We expect this has something to do with the bias explained above as self-rated mental health can be easily affected by external factors like weathers and moods.

Education level does not have a clear relationship with the feeling about life as well. People with a degree less than high school have a positive coefficient of 0.37 while other factors are around +- 0. with the highest level of education, a diploma of the above bachelor has the lowest coefficient of -0.003.

We include both individual income and family income and individual income. Individual income of all ranges shows a positive relationship with the feeling towards life which all but the higher family level show a negative correlation with the feeling about life. This is counter-intuitive to look at at first but this might revive a deeper story. Looking at Figure.4 and Figure.5, we see that their distributions are different. The number of respondents grows as the individual income shrinks. The family income shows no clear pattern. We suspect this is because of the only one person in the family work or the income of the working family members are very unequal. We can further suspect that gender inequality plays a hidden role here. The incongruence between different family members with their role in the family together with potential gender inequality creates the difference between the family income V.S. feeling about life relationship and individual income V.S. feeling about life relationship.

Applying the Model Back to Reality

With the finding from this model, people can rethink about the value of education which the economical cost and opportunity cost of post-secondary education keeps growing in this society.

This would also work as a guideline for future government policies. The government might want to encourage families to have more babies as this improve feelings towards life but more importantly, the Canadian government should look into the inequality of income among different family members and potential gender inequality to be improved.

Confounding Variable

Other variables may include factors like if the respondent lives in a rented property or the respondent's own property. People who live in their own living space are generally expected to live happier as they do have pressure to pay rent and have more choices with their monthly income.

Status of immigration may also affect life feeling level. The status of immigration has a direct effect on a person's future. A high life feeling level usually comes with a promising future. An unstable future would lead to possible depression and a poor life feeling level.

Meal preparation also shows some insights into the respondent's life and ultimately life feeling. It is an objective question about who prepares the meal in a family and would reveal the role the respondent plays in his or her own family. Together with the family income factor, we would potentially find out how the role one plays in one's own family affects happiness.

Next Step and Future Works

For the next step, we are interested in looking into the inconsistency between the family income V.S. feeling about life relationship and individual income V.S. feeling about life relationship and the hidden problem behind it.

Appendix

For readers from UofT, you may download the data from the library.[8] To do that: 1. Go to: http://www.chass.utoronto.ca/
2. Data centre –> UofT users or http://dc.chass.utoronto.ca/myaccess.html
3. Click SDA @ CHASS, should redirect to sign in. Sign in.
4. Continue in English.
5. Crtl F GSS, click
6. Click "Data" on the one you want. This code applies to 2017.
7. Click download
8. Select CSV data file, data definitions for STATA (gross, but stick with it for now).
9. Can select all variables by clicking button next to green colored "All". Then continue.
10. Create the files, download and save

[8]* Data download instruction is supported by Rohan Alexander and Sam Caetano, see full citation in reference

References

1.Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

2.Alexander Rohan, & Caetano Sam (2020, October 07). Gss_cleaning.R. STA304 Lecture, University of Toronto.

3.Canada, CHASS data centre, University of Toronto, Faculty of Arts &Sciences. (2020. General social survey on Family (cycle 31), 2017. ON: University of Toronto.

4.Hadley Wickham (2020). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.0. https://CRAN.R-project.org/package=forcats

5.Hadley Wickham, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse."Journal of Open Source Software,4(43), 1686. doi:10.21105/joss.01686.

6.Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.0. https://CRAN.R-project.org/package=dplyr

7.Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

8.JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com. Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL https://bookdown.org/yihui/rmarkdown.

9.Lionel Henry, Hadley Wickham and Winston Chang (2020). ggstance: Horizontal 'ggplot2' Components. R package version 0.3.4. https://CRAN.R-project.org/package=ggstance

10.Long JA (2020). *jtools: Analysis and Presentation of Social Scientific Data.* R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>.

11.R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

12.Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

13.Statistics Canada(2020). General Social Survey. Cycle 31 : Families Public Use Microdata File Documentation and User's Guide. Catalogue no. 45250001 Issue no. 2018001

14.Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN978-1498716963 Yihui Xie (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