Important Factors that Affect our Feelings about Life

Yuchen Cong, Jingwen Deng, Ruoxi Guan, Yuwei Sun

Oct. 19, 2020

Abstract

We measure our feelings of life very subjectively; there are so many real-life factors, such as income, education, job and personal health, which could affect us. We evaluate how we feel about our life and consider it as to whether we are satisfied with our current lives. This report analyzes factors that could influence feelings of life score from respondents of the Canadian General Social Survey Dataset. During the analysis, we produce tables of data summary and graphs between related variables. A Multiple Linear Regression model is introduced with significant factors that could affect feelings of life scores. We find out the factor of income level, health level, mental health level, and marital status are likely related to the feelings of life score.

Introduction

In this report, our goal is to discover important factors that could affect the feelings of life scores collected by the Canadian General Social Survey Dataset. We deeply look into this dataset and find some interesting variables that could relate to our analysis object. After performing tables and figures among different variables, we filter out those who have less effect. Then we decide to explore more with the variables that contribute significantly to the feelings of life score. Since there are multiple variables chosen, we set up a Multiple Linear Regression model to see if the selected variables give a good prediction to our objective. Through our analysis throughout this report, we expect to find out the important factors that improve people's feelings of life to help those who are disappointed in their lives and give a hint to them about how they could have a better feeling of life. We used R markdown throughout the project.

Data

- The 2017 Canada General Social Survey dataset is used throughout this project. We obtained the dataset from the U of T library. This dataset was conducted from February 2nd to November 30th, 2017, a sample survey with sectional design. The **target population** for the 2017 GSS included 15 years of age and older in Canada, excluding the full-time institution residents and residents of the Yukon, Northwest Territories and Nunavut. The target sample size for 2017 was 20000, but there are 20602 respondents. The dataset was collected via **computer-assisted telephone interviews**. Those who refused to participate were re-contacted to explain the significance of the survey. The total response rate was 52.4% and the number of variables in this dataset was 81.
- **Stratified random sampling** was used in the 2017 GSS so that each of the ten provinces was divided into strata. The population is first sliced into homogeneous groups before the sample is selected. Then, a simple random sampling without replacement of records was performed next in each stratum. The **survey frame** consists of a list of telephone numbers in use and the Address Register (AR). Non-response was not permitted for questions required for weighting, so the 2017 GSS used a "three-stage non-response adjustment" (Appendix#1) to drop the non-responding telephone numbers.
- **Strengths and Weaknesses**: In order to reduce non-sampling errors in the survey and monitor the quality of the data, quality assurance measures were introduced at each step of the data collection and processing cycle. A telephone survey is a more direct approach resulting in a good response rate. Also, it can ensure a proper understanding of respondents by clarifying questions on the phone. However,

there are some limitations to the data. For example, some respondents did not have the patience or time to complete such a long survey on the phone, so that some responses were not completed. Also, not all respondents were willing to talk about their living conditions to a stranger on the phone, so they may choose to lie, affecting the responses' accuracy. Telephone surveys are very time consuming to administer and can be expensive when aiming for large samples. Furthermore, this kind of survey requires the survey administrator to be highly skilled to avoid bias.

- **Data Cleaning Process**: We choose the data which will affect the feelings of life score by two conditions. Firstly, the chosen variables from the dataset can not contain too many NA terms since the NA terms will influence the test's accuracy. After that, we will choose the variables to depend on life common sense. After our discussion, we pick the age, sex, region, total children, education level, levels of income of respondent, marital status, self-rated health level, self-rated mental health level as our variables. They all directly or indirectly reflect on the quality of people's life or their life pressure. As we all know, whether a person can harvest happiness, whatever from spiritual or material terms, determines their feelings of life scores. Then, we redivided the 'marital_status' variable into three categories: "Married", "Single" and "Other", and also categorized 'income_respondent' into "Less than $25,000", "$25,000 to $74,999", and "$75,000 and more". Besides, observations that are blank or contain words like "NA" and "Don't know" are not selected.

Model

Model Equation

$$
\begin{aligned}
Y_i = \beta_0 &+ \beta_1 \cdot X_{Income,\ 75{,}000\ and\ more} + \beta_2 \cdot X_{Income,\ Less\ than\ 25{,}000} \\
&+ \beta_3 \cdot X_{Marital\_Status,\ Other} + \beta_4 \cdot X_{Marital\_Status,\ Single} \\
&+ \beta_5 \cdot X_{Sex,\ Male} + \beta_6 \cdot X_{Health,\ Fair} \\
&+ \beta_7 \cdot X_{Health,\ Good} + \beta_8 \cdot X_{Health,\ Very\ Good} \\
&+ \beta_9 \cdot X_{Health,\ Excellent} + \beta_{10} \cdot X_{Mental\_Health,\ Fair} \\
&+ \beta_{11} \cdot X_{Mental\_Health,\ Good} + \beta_{12} \cdot X_{Mental\_Health,\ Very\ Good} \\
&+ \beta_{13} \cdot X_{Mental\_Health,\ Excellent} + \beta_{14} \cdot X_{Num\_Child} + \epsilon_i
\end{aligned}
$$

The "feelings_life" variable is chosen to be our response variable here, and it is a numerical variable with scaling from 0 to 10. A larger number represents a better feeling of life. A multiple linear regression (MLR) model is generated to investigate how our predictors influence people's feelings of life. In our MLR model, "income_respondent", "marital_status", "sex", "self_rated_health", "self_rated_mental_health", and "total_children" are chosen as predictors to predict the feelings of life score. Since "income_respondent", "marital_status", "sex", "self_rated_health", and "self_rated_mental_health" are categorical variables, we use them as a factor so that we have serval dummy variables corresponding to each categorical predictor.Since the response variable is a numeric, we can fit a MLR model. We do not fit GLM due to the fact that the response variable is not binary. Interpretation of the model, model checks and diagnostics issues are discussed in the later section.

Results

Tables

- Table 1 gives the first few observations of our dataset after the cleaning process. It clearly shows the variables we chose that could affect respondents' feelings of life score, which is what we are interested in.

Table 1: Data Visualization after Cleaning Process

Age	Sex	Region	Feelings of Life Score	Total Children	Education Level	Levels of Income of Respondant	Marital Status	Self Rated Health Level	Self Rated Mental Health Level
52.7	Female	Quebec	8	1	Low	$25,000 to $74,999	Single	Excellent	Excellent
51.1	Male	Prairie region	10	5	Low	Less than $25,000	Married	Good	Good
63.6	Female	Ontario	8	5	High	$25,000 to $74,999	Married	Very good	Good
80.0	Female	Prairie region	10	1	Low	$25,000 to $74,999	Married	Very good	Very good
28.0	Male	Quebec	8	0	High	Less than $25,000	Other	Good	Good
63.0	Female	Quebec	9	2	Low	Less than $25,000	Married	Excellent	Very good

- Table 2 is a summary table of different income levels, including each level's proportion among all data and averages of feelings of life score, age and number of children respondents have.

Table 2: Summary of Different Levels of Income

Levels of Income of Respondant	Proportion of Different Levels of Income	Feelings of Life Score	Average Age	Average Total Children
Less than $25,000	0.327	7.897	50.840	1.617
$25,000 to $49,999	0.300	8.084	54.090	1.713
$50,000 to $74,999	0.190	8.199	52.011	1.673
$75,000 to $99,999	0.099	8.303	50.030	1.676
$100,000 to $ 124,999	0.041	8.312	51.373	1.695
$125,000 and more	0.043	8.513	53.841	1.822

Figures

- Figure 1 shows the relationship between the respondent's feelings of life score with different income levels. However, except for those who earn $125,000 and more, respondents in other income levels have very small differences in median and Q3, the 75th percentile of the boxplot.

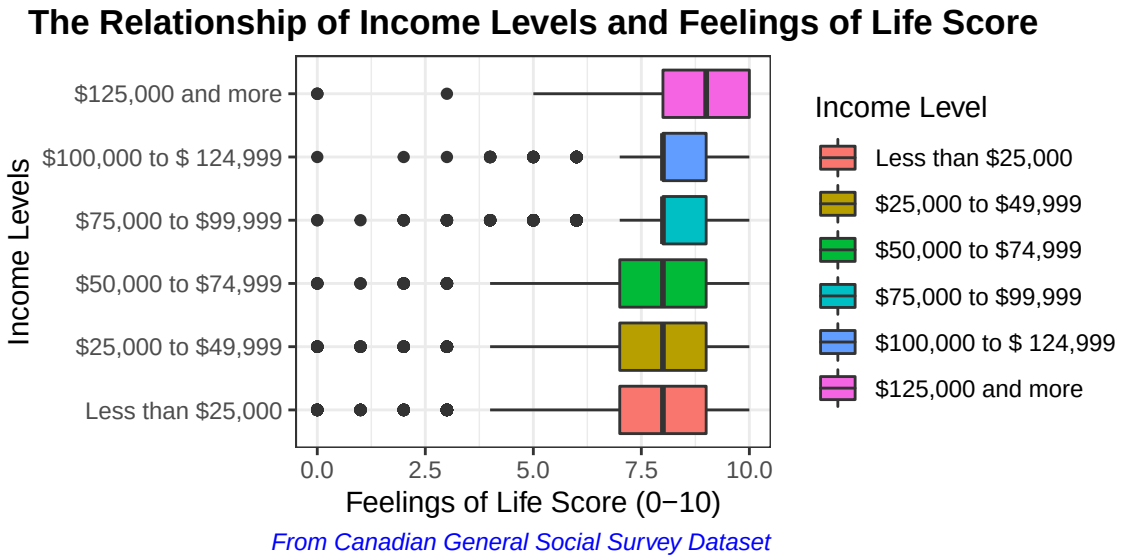

Figure 1: Feelings of Life Score VS. Income of Respondent

- Figure 2 gives us the relationship between the feelings of life score and respondents' self-rated health level. The boxplots form the shape of a ladder. The median score of respondents who have an "Excellent" health level is greater than or equal to the third quartile of "Good" and "Very good" health levels. Even though many respondents with "Poor" health levels have lower feelings of life scores, the median of these scores is almost 6 out of 10.

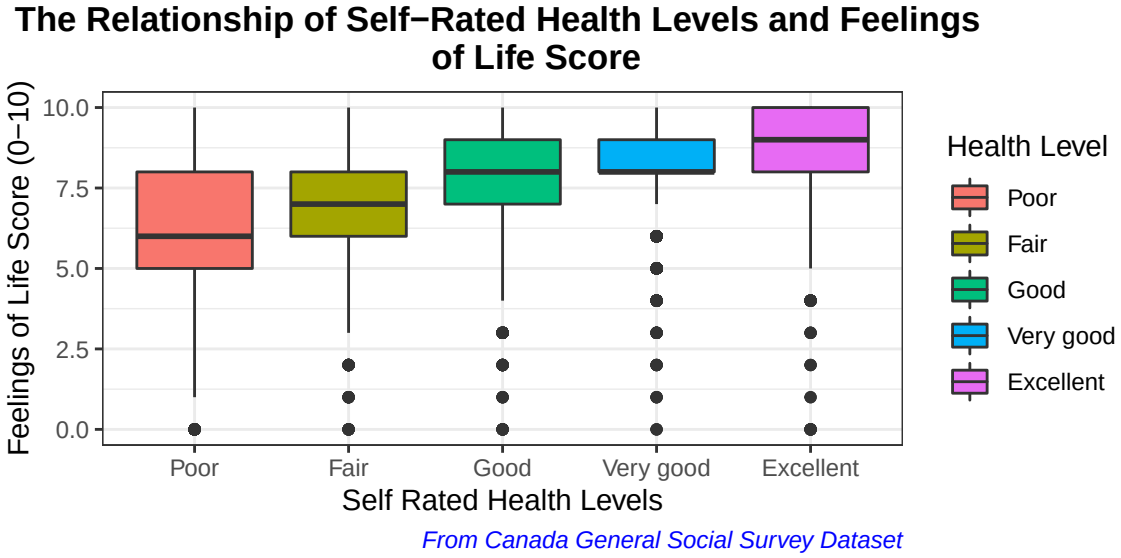

Figure 2: Feelings of Life Score VS. Self-Rated Health Levels

- Figure 3 is a graph of feelings of life score related to respondents' self-rated mental health levels. It has a more obvious ladder shape, which is similar to Figure 2. We notice that the third quantile of respondents with "Poor" mental health levels is much lower than respondents with "Poor" health level. Also, the median drops to nearly 5 out of 10.

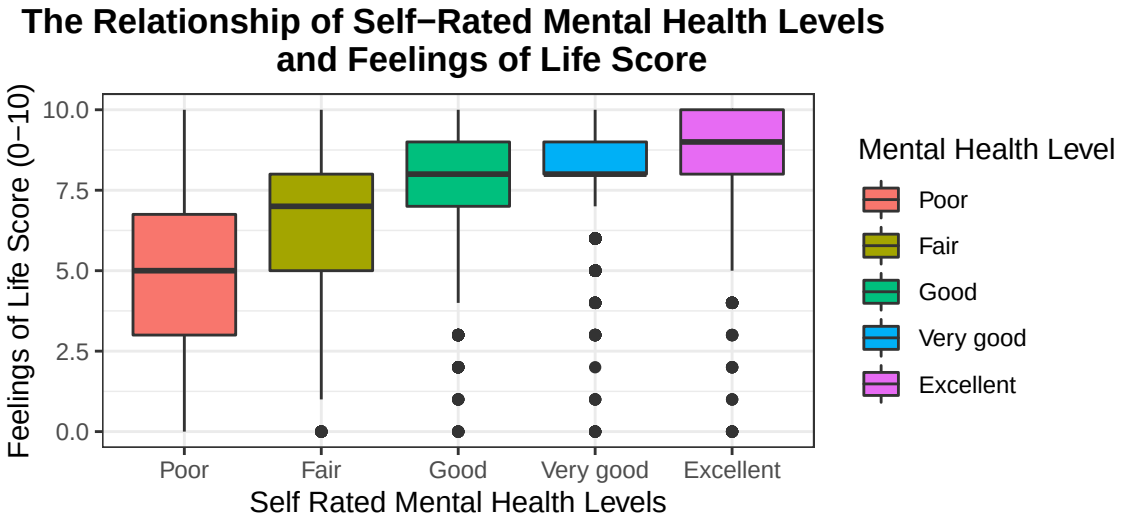

Figure 3: Feelings of Life Score VS. Self-Rated Mental Health Levels

Model

- Fitted Model Equation

$$\begin{aligned}\hat{Y}_i = {} & 4.314614 + 0.023244 \cdot X_{Income,\ 75{,}000\ and\ more} + 0.029977 \cdot X_{Income,\ Less\ than\ 25{,}000} \\ & - 0.311006 \cdot X_{Marital\_Status,\ Other} - 0.450532 \cdot X_{Marital\_Status,\ Single} \\ & - 0.113203 \cdot X_{Sex,\ Male} + 0.815864 \cdot X_{Health,\ Fair} \\ & + 1.163934 \cdot X_{Health,\ Good} + 1.373821 \cdot X_{Health,\ Very\ Good} \\ & + 1.563505 \cdot X_{Health,\ Excellent} + 1.279750 \cdot X_{Mental\_Health,\ Fair} \\ & + 2.322968 \cdot X_{Mental\_Health,\ Good} + 2.821931 \cdot X_{Mental\_Health,\ Very\ Good} \\ & + 3.273919 \cdot X_{Mental\_Health,\ Excellent} + 0.075379 \cdot X_{Num\_Child}\end{aligned}$$

- Summary Table

Observations	19936
Dependent variable	Feelings_Life
Type	OLS linear regression

F(14,19921)	577.94
R^2	0.29
Adj. R^2	0.29

	Est.	S.E.	t val.	p
(Intercept)	9.15	0.03	287.86	0.00
as.factor(Income)$75,000 and more	0.02	0.03	0.86	0.39
as.factor(Income)Less than $25,000	0.03	0.02	1.33	0.18
as.factor(Marital)Other	-0.31	0.03	-10.08	0.00
as.factor(Marital)Single	-0.45	0.02	-20.18	0.00
as.factor(Sex)Male	-0.11	0.02	-5.65	0.00
as.factor(Health)Fair	-0.75	0.04	-18.05	0.00
as.factor(Health)Good	-0.40	0.03	-12.86	0.00
as.factor(Health)Poor	-1.56	0.06	-26.50	0.00
as.factor(Health)Very good	-0.19	0.03	-6.49	0.00
as.factor(Mental_Health)Fair	-1.99	0.05	-42.57	0.00
as.factor(Mental_Health)Good	-0.95	0.03	-32.85	0.00
as.factor(Mental_Health)Poor	-3.27	0.09	-38.35	0.00
as.factor(Mental_Health)Very good	-0.45	0.03	-16.97	0.00
Num_Child	0.08	0.01	10.87	0.00

Standard errors: OLS

Discussion

Table and Figure

- From Table 2, we can see that most respondents earn less than $25,000, and this portion of people has the lowest average of feelings of life score. We also find that as the level of income goes in increasing order, the average of feelings of life score rises as well. However, the averages of respondents' age and the number of children respondents do not significantly differ in different income levels.
- Figure 1 clearly shows that the more the respondent earns, the higher feelings of life score the respondent would have. Even though each income level's average feelings of life score seems pretty high and equal, a large portion of people in the lower-income range level has the feelings of life score less than the average score.
- From Figure 2, we can easily tell that respondents with good health tend to have higher feelings of life scores. Figure 3 is very similar to Figure 2 but with more significant differences. In Figure 2, the 75th percentile of "Poor" and "Fair" health levels are about the same. However, in Figure 3, the Q3 of the "Poor" health level is much smaller than the Q3 of the "Fair" health level. And the difference between the medians of these two levels almost double. This implies that mental health influences more on our feelings of life. Indeed, the higher the mental health level, the greater the feelings of life score would be.

Model

- After generating our model, we first check our model assumptions. From our plot graph(Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage), we see that there exist a few extreme observations. Still, the MLR model assumptions are reasonably satisfied in general.
- Summarizing the model, we find out that it is significant at an overall level, and all covariates are significant except the income predictor. We discover that the self-rated mental health level and self-rated health level have a substantial impact on the feelings of life score since health is an important factor affecting the quality of life. If people's self-rated mental health level is excellent and other factors remain constant, their feelings of life score will increase 3.27 and very good mental health level increases 2.82; good mental health level increases 2.32. However, a fair mental health level only increases 1.28, and a poor mental health level increases 0. Moreover, based on our model, we surprisingly discovered that the number of children of a person also affects a person's feelings of life; one more child will increase a person's feelings of life score by 0.075. It is also shown that on average, a male's feelings of life score

will be 0.11 lower than a female's feelings of life score. This phenomenon may be due to the stress and expectation that are given to them by modern society. Marital status will also influence one's feelings of life. Married people tend to have higher feelings of life score. The adjusted R-squared (0.2883) of our MLR model means that our model has explained approximately 30% of our response variable's variation, the feelings life score.

Weaknesses

- **Questionnaire Weaknesses**: For the questionnaire, we have noticed that all the questions are very simple, only ask one thing at a time, which is good. However, some questions do not have a balance across the response options. For example, the variable "self_rated_health", has six response options: "Don't Know", "Poor", "Fair", "Good", "Very good" and "Excellent". The problem is that three of the responses are positive, with only one option for a negative response. Also, Excellent and Very Good have a similar meaning, so respondents may not easily distinguish between. A better one can be "Don't Know", "Very poor", "Poor", "Fair", "Good", and "Very good".
- **Dataset Weaknesses**: Because participants are motivated to answer survey questions, response bias reinforces the characteristics and behaviours preferred by the entire society while rejecting undesirable characteristics and behaviours. For instance, sometimes, people are more likely to overestimate their mental health levels. Some people may also choose a higher income range out of self-esteem, while others may tend to select a lower income range for privacy or other reasons.
- **Model Weaknesses**: Since some variables are chosen based on our intuition, they may not be good at fitting our MLR model, leading to the prediction's inaccuracy. Moreover, most of the selected variables are categorical variables, so the model's fitness will be affected when we treat them as numeric variables. Also, some further cleaning processes, such as recategorizing variables and removing "N/A" observations, may also affect the result. Some relationships are not being captured. Besides plotting the residual plots, we can notice some residual plot patterns, so there are some violations of the assumption. This may be due to the wrong variable or model selections and the dependence of variables.

Next Steps

- We can do some improvements to our model. For example, the transformation process of variables can be applied before fitting a model so that our model assumption would be more satisfied. Using another statistical method to fit another model or add other significant variables as predictors can also be considered.
- From the news, we can notice an increasing number of people who committed suicide due to depression. After finishing this project, we can identify those people who have low feelings of life score or are likely to have worse outcomes so that some people can be targeted for interventions to improve these outcomes and reduce the costs of themselves/government.
- We analyzed a few variables using MLR to see how they affect the feelings of life score; we could continue looking into each variable and our objective. Finding out what factor is the most influential one, then we should explore more on that variable.

References

- Cite R : R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Cite "dplyr" : Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Cite "ggplot2" : H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. Cite "knitr" : Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.
- Cite "jtools" : 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>.
- Cite "kableExtra" : Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra
- "Developing and Implementing Surveys." Developing and Implementing Surveys, artsengage.initiatives.qld.gov.au/images/and-implementing-surveys-fact-sheet.PDF.
- Life Satisfaction. (n.d.). Retrieved October 19, 2020, from http://www.oecdbetterlifeindex.org/topics/life-satisfaction/
- "Public Use Microdata File Documentation and User's Guide." General Social Survey, The Minister Responsible for Statistics Canada, Apr. 2020, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss3

Appendix

1. Three-stage non-response adjustment: Numbers with some auxiliary information available, numbers with supplemental information from various sources available to Statistics Canada and numbers with no auxiliary information are the three types of non-responding telephone numbers. The first adjustments were made for complete non-response, and it was done independently within each stratum. The second adjustment was made for non-response with auxiliary information, which was used to respond to the model propensity. The last adjustments were made for partial non-response. The second and third adjustments were done independently within each wave.
2. Code and data supporting this analysis is available at: "Github link" https://github.com/jingwennnn/Important-Factors-that-Affect-our-Feelings-about-Life