Estimating the influence of life satisfaction

What are some potential factors influencing one's life satisfaction

Xiao Bai Yichun Zhang Hailan Huang

20/03/2022

Abstract

Canadian General Social Survey (GSS) is designed to gather data on social trend and provide information about Canadian's well-being and social conditions for improving people's daily life. In the 2017 GSS, 'how do you feel about life as a whole' were aksed, and we conducted study about people's feelings about life using Canadian general social survey data using a multiple linear regression model. Supplimentarily, we also conducted hypothesis test using bootstrap method to see if people have less life satisfaction if they rent their dwelling. As a result, we found that age, dwelling, health, mental health, working hour and education level significantly affect people's overall feeling about life, which means people should pay more attention on these potential factors to improve individual's well-being as a whole.

Contents

Code and data are available at[1]

[1] https://github.com/XiaoBai-blip/STA304-Paper-3

1. Introduction

Life satisfaction and other subjective well-being measures have been of considerable importance in gerontology. According to Livni (2018), people is likely to maximize their satisfaction with themselves and with their lives throughout their whole life time. Based on ("Life Satisfaction," n.d.), people tend to rank their general life satisfaction on average as 6.7 as they are given a life satisfaction scale from 0 to 10. This value is relatively large in OECD countries. However, in other countries such as Colombia, Greece, Korea, these countries tend to have a low level of overall life satisfaction with general score below 6. Existing theories and empirical findings state that life satisfaction has been significant related to one's achievement goals (Wang et al. 2017). Ideologically, the research about people's life satisfaction, in this case, evaluated as "feeling about life as a whole" is rooted in 18th century, when people's life started to realize as of existence of life itself other than the service of King or God. Self-actualization and happiness started to become central values. Starting from 19th century, the public's conviction about best society established as one that provide greatest happiness for the greatest number, and countries were inspired to develop the Welfare State. Social statistics were developed to record the extent of progress achieved. In the 1970s, researches and politicians started to search for suitable indicators of non-economic welfare evaluation (Veenhoven 1996). In this paper, we are used 2017 Canadian General Social Survey (GSS) to study the factors that affect people's life satisfaction. Respondents were asked: "using a scale of 0 to 10 where 0 means"Very dissatisfied" and 10 means "Very satisfied", how do you feel about your life as a whole right now?"

We conducted a Linear Regression Model between feelings about life as a whole and age group of respondent (groups of 10), number of respondent's children in household - any age/marital status, full-time/part-time job, income of respondent - total (before tax), dwelling - owned or rented, self rated health, self rated mental health, number of weeks employed - past 12 months, average number of hours worked per week, province of residence of the respondent, marital status of the respondent, education - highest certificate, diploma or degree, and living arrangement of respondent's household (12 categories), where feelings about life as a whole is the dependent variable which we want to dig deep into and the rest are the independent variables which we supposed that they might be influential to the dependent variable feelings about life as a whole. Then we conducted a Hypothesis Test on if the self-rated mental health is related to whether dwelling of the respondents is owned or rented, and conducted a Confidence Interval Analysis using the bootstrap method with a confidence interval of 90% of these two groups.

After the first Linear Regression Model, we found that age group of respondent (groups of 10), dwelling - owned or rented, self rated health, self rated mental health, and education - highest certificate, diploma or degree are the influential factors to feelings about life as a whole. Using these variables, we conduct a new Linear Regression Model, especially focused on the variables of age groups of 15 to 24 years and 75 years and older, owned and rented a household, self rated health and self rated mental health, and education level less than high school diploma or its equivalent. The result of the Hypothesis Test tells that the p-value is less than 0.05, so we reject the null hypothesis, and believe that there is different in mean value of owned or rented. The Confidence Interval Analysis shows the confidence intervals for the two groups are significantly different.

Key words: 2017 General Social Survey (GSS); Family; Feelings about life as a whole; Age group of respondent (groups of 10); Dwelling - Owned or rented; Self rated health; Self rated mental health; Education - Highest certificate, diploma or degree.

The analysis will be conducted in R (R Core Team 2020), and the package we will use is tidyverse (Wickham et al. 2019). All graphs will be created using function ggplot2 (Wickham 2016). The packages knitr (Friendly et al. 2020) are also used to generate the R markdown report.

2. Data

2.1 Data Sources

We found the data from statistics Canada, under the surveys and statistical programs category. We downloaded the data and loaded them into RStudio for our own analysis. We point our focus on what factors affect people's felling about life, in other word, how satisfied people are about their life. The GSS data is the result of general social survey (GSS) on families, conducted in Canada, 2017. 421 questions were asked and over 20000 people responded. The GSS is recognized for its ability to test and develop innovative ideas that solve current or upcoming difficulties in a cost-effective and timely manner. It was open for participation from February 2nd until November 30th, 2017. Then, in each stratum, a basic random sample was taken without the use of record replacement, and the process was repeated. In the 2017 GSS, computer-assisted telephone interviews were employed to collect data. The respondents were questioned in their native language. Interviewers were supposed to conduct a thorough interview with a random household member. The survey has been divided into two sections. The first component contains current phone numbers, while the second contains a list of all occupied residences. Province-specific strata were allocated to each individual survey record. Then, in each stratum, a basic random sample was taken without the use of record replacement, and the process was repeated. The architecture of GSS was developed with the use of data from the United States Census, administrative files, and billing files. Increased coverage as compared to the random number dialling frame that came before it (though over coverage and under coverage may still exist). The respondents from the eleven provinces were contacted via phone. Homes without telephones were not included in the sample. Adjusted (weighted) survey estimates were used to account for all members of the target population, including those who were not surveyed in the first place. Many individuals do not respond to queries about their income, and when they do, the figures they provide are often guesses rather than precise figures. You are not required to make questions about this content due to the hyperlinks. Personal tax records (T1 Family File or T4 Information Slip) have been connected to General Social Survey data from 2017. (Cycle 31). Two critical pieces of information are required to establish a connection: the responder's social security number, surname, name, date of birth or age and gender, and the respondent's home address.

2.2 Data cleanning and data overview

In this study, respondents were asked about their age, education, family origins, leaving from parental home, conjugal history, intentions of forming unions, children, health and well-being, dwelling, spouse or partner and etc. Results were summarized in a CSV file, and variables were coded. A codebook can be found from the Statistics Canada website to check what each code can be represented. Along with numbers that are no longer operational and any other non-functioning lines, telephone numbers belonging to corporations, institutions, or other out-of-scope residences are all instances of telephone numbers that are not included in this study's sample. All phone numbers that are not relevant to the procedure are simply removed from the sample, leaving just relevant entries. The initial weight assigned to all records included in the scope remains unchanged from the previous phase. Three categories were established to classify non-responding phone numbers: those with some more information (for example, a complete list of household members), those with further data accessible from other sources, and those with no additional data available at all. Making non-response changes included three processes: To account for the total absence of responsiveness, the initial phase was altered (i.e., households for which no auxiliary information was available). This was accomplished layer by layer. Later, if appropriate, changes were made based on additional data acquired from sources made available to Statistics Canada during the second stage. Additional data was supplied to them in order to construct a model of how these families would act. Changes were made in the third stage to allow for partial non-response. Additional information was submitted in order to construct a model predicting how likely these families were to react. Each wave's last two changes were done independently.

For the authentication of the survey, options like "refuse to answer", "I don't know" are included, but we excluded them from our study. Since we are interested in quantitative evaluations for measuring people's felling about life and factors that affect people's felling about life, we filter out "refusal", "don't know" and "valid skip" each time when we conduct our study. Questions in the survey vary from partner situation, children, intention of fertility to dwelling and health, and since we are interested in people's felling about life, we selected relevant variables including:

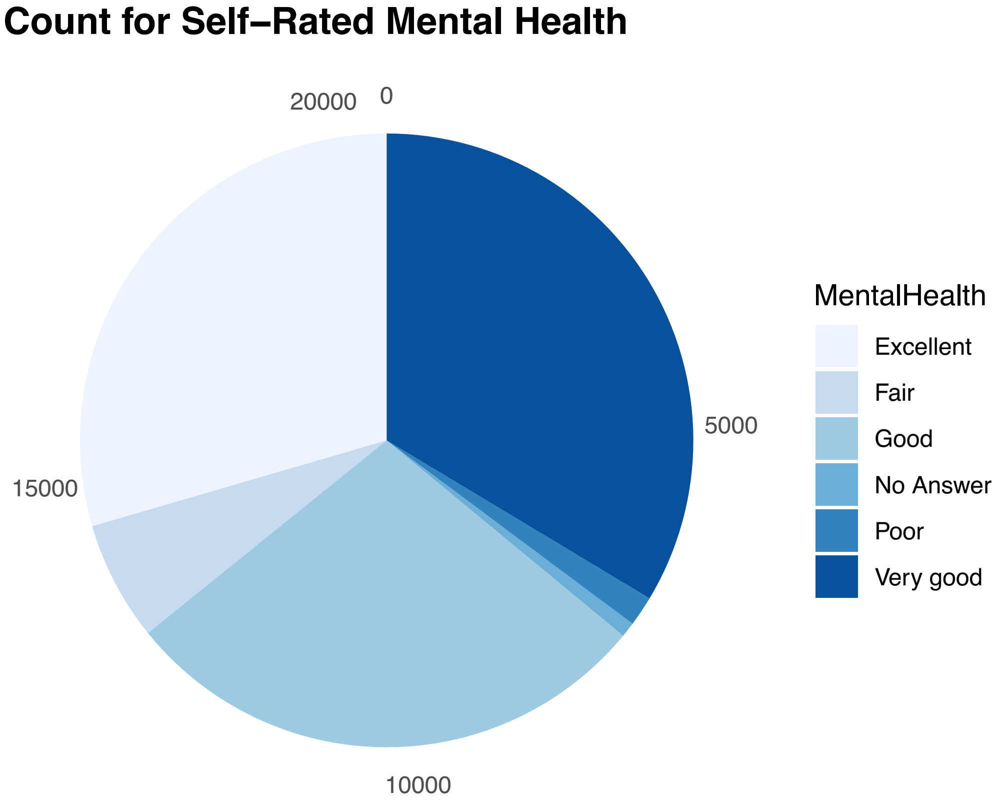
table 1	code	meaning
1	AGEGR10	Age group of respondent (groups of 10)
2	CHRINHDC	Number of respondent’s children in household - Any age/marital status
3	RTO_101	Full-time/Part-time job
4	TOE_240	Permanent/Not permanent job
5	TTLINCG2	Income of respondent - Total (before tax)
6	ODR_10	Dwelling - Owned or rented
7	SRH_110	Self rated health
8	SRH_115	Self rated mental health
9	NWE_110	Number of weeks employed - Past 12 months
10	UHW_16GR	Average number of hours worked per week
11	PRV	Province of residence of the respondent
12	MARSTAT	Marital status of the respondent
13	EHG3_01B	Education - Highest certificate, diploma or degree
14	LIVARR12	Living arrangement of respondent’s household (12 categories)

2.3 Data Summary

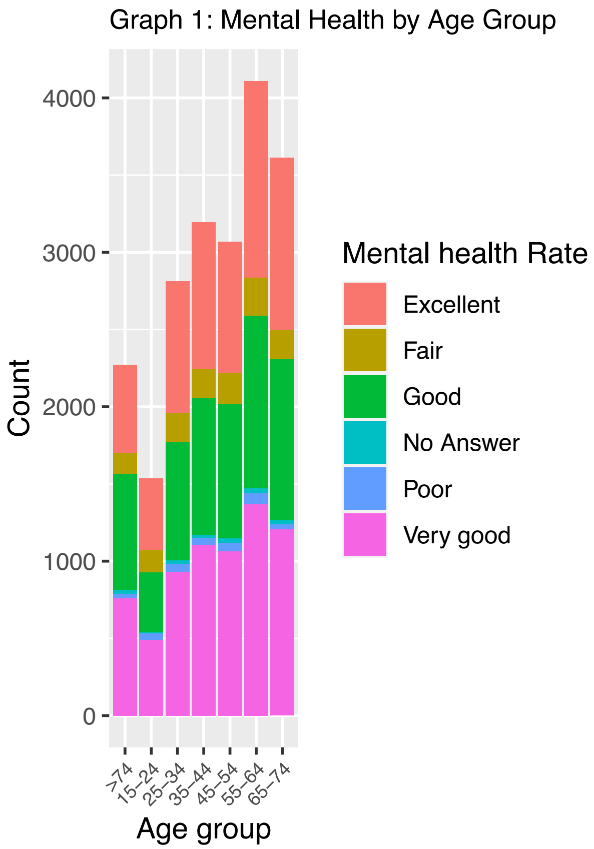
To clean the data, first we selected relevant variables. It makes sense that age, number of children, job type, income, dwelling, health, mental health, employment situation and marital status of the respondent affect their overall life satisfaction. The original data collect from the survey used code to represent people's life satisfaction. People's life satisfaction is scaled 1 - 10, where 1 present high satisfaction and 10 represent low life satisfaction. To reduce the uncleanness due to the coding of variables, we mutated variables into categories. For people's age, the original data used number 1 - 7 to represent different groups. To reduce to confusion of variable types, we also mutated the variables into different characteristic group, like "15 to 24 years", etc. The distribution and variability of relevant variables are listed:

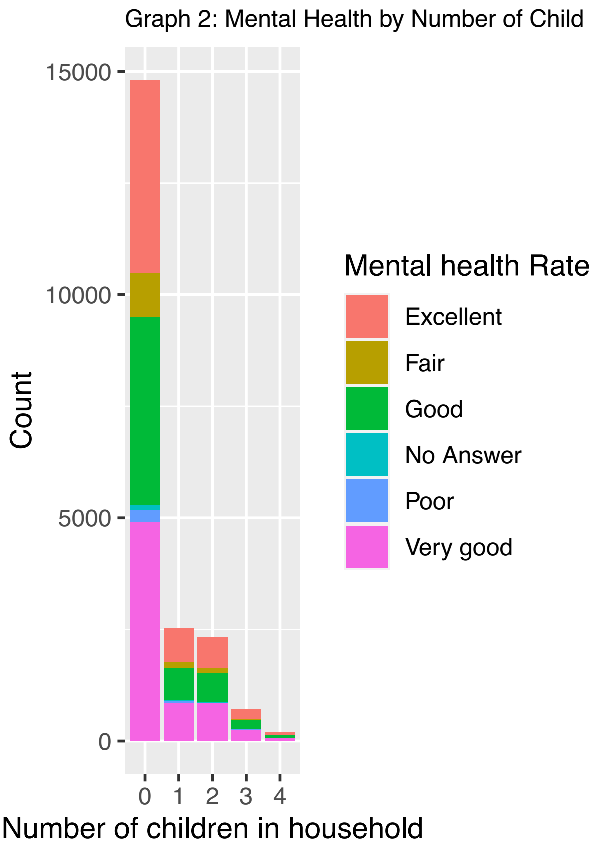
Table 2	representation	Storage	Original
AGEGR10	age of respondents	1	15 to 24 years
		2	25 to 34 years
		...	...
		7	75 years and older
CHRINDHC	number of children	0	No Child
		1	One child
		...	...
		4	Four or more children
RTO101	Job type	1	full time
		2	part time
ttlincg2	income	1	Less than $25,000
		2	25,000 to 49,999
		3	50,000 to 74,99
		...	...
		6	125,000 and more
ODR_10	Ownership of dwelling	1	Owned by you or a member of this household
		2	Rented
MARSTAT	Marital Status	1	Married
		2	Living Common-law
		3	Widowed
		4	45 to 54 years
		5	55 to 64 years }
		...	...
EHG3_01B	Education Level	1	Less than high school
		2	High school diploma or equivalent
		3	Trade certificate or diploma
		4	College, CEGEP or other
		...	...
LIVARR12	resident situation		
		...	...

2.4 Graphs

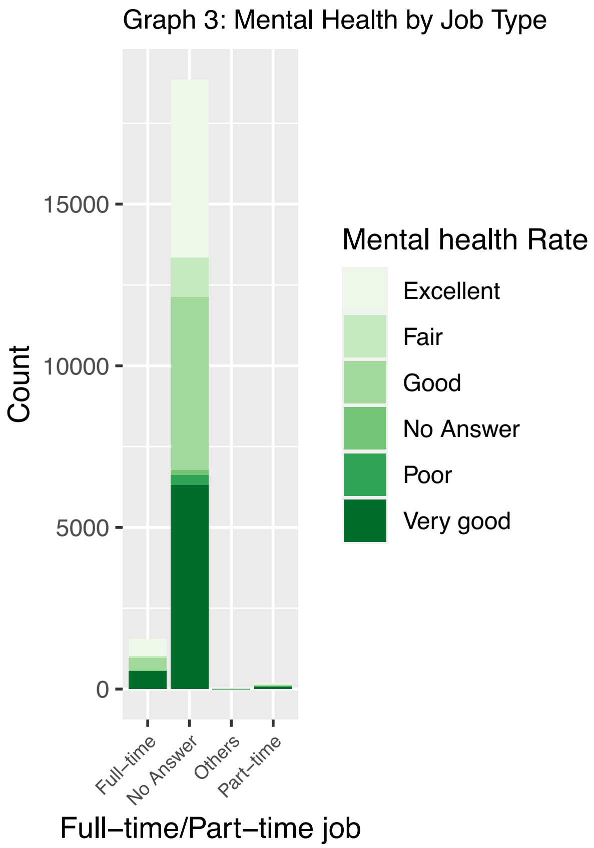


The pie chart shows the result of a survey in which 20,620 people were asked about their self-rated mental health. Respondents were given five different levels of choices for mental health condition, representing “excellent”, “very good,”good“,”fair" and “poor”. From the pie chart, it is clear that the majority of the respondents thought they do not have any problem with their mental health. Specifically, the proportion of respondents that report as “excellent”, “very good” or “good” level of mental health looks like the same, with just a tiny difference between them. Over one fourth of participants self-check their mental health as “very good”. Only a small number of participants report their health condition as “fair” or “poor”. There are also a small proportion of respondents did not show their answers, some of them do not have an idea of how to check their mental health status, others refuse to tell or skip this question in purpose. As a concluded, since most of our respondents’ age is between 24 and 74 years old, they are expected to have a clear understanding of their self-consciousness for mental health. If that is the case, the result of self-rated mental health should be accurate. In addition, the result also shows that most of the participants have a good mental health condition. If there are more people participating this survey, we can probably expect to a larger proportion of people report their mental health as “good”.

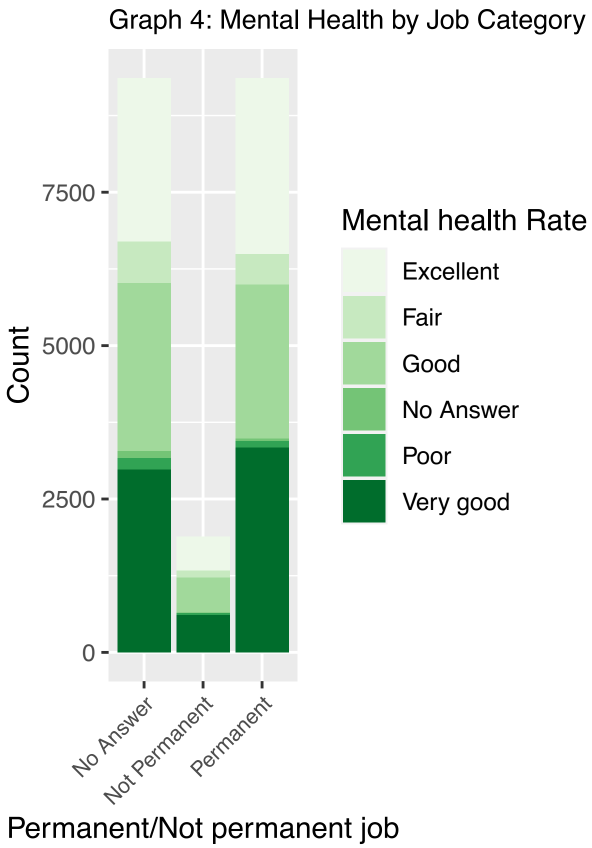




The first bar chart shows the mental health condition according to different age groups. The count is the total number of respondents and is represented by the height of the bars (vertical axis). Look at the height of different colored bars separately, each color represents one scale of mental health level. Mental health level "excellent" is shown in red, "very good" is shown in yellow, "good" is in green color, etc., which is the same as the one in previous pie chart. Overall, it can be seen that more than 4000 of our respondents in the age group 55-64, which accounts for the largest amount among all age groups. The number of respondents with mental health "excellent", "very good" and "good" have about the same proportion among all seven age groups, with a slightly 100 counts difference among these three groups. Only around 0-250 respondents reported that they have a "fair" or "poor" mental health throughout different age groups. The second graph indicates the mental health condition according to participants family's fertility circumstance. We can see that more than 12,000 respondents do not have child, and the number of people have one or two children are equal with only few respondents have four children by now. Moreover, we assume that the mental health problem might get worse as people have more children. It is because we thought that parents with more children are likely to be more stressful than parents have only one or two kids. However, the result is not what we expected as the graph reveals that the three different scales (excellent/very good/good) of mental health status literally has almost the same number of respondents.



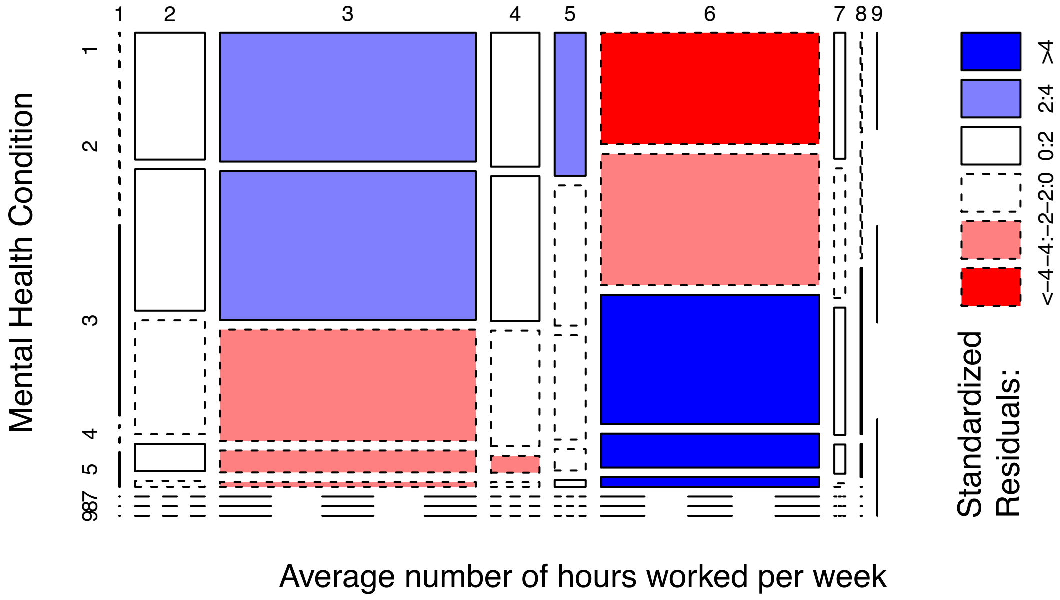
Graph 3: Mental Health by Job Type



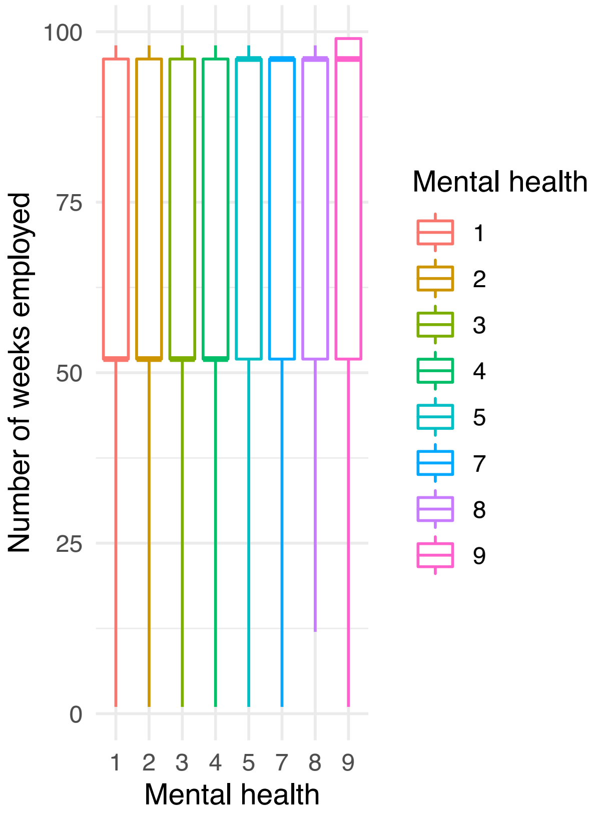
Graph 4: Mental Health by Job Category

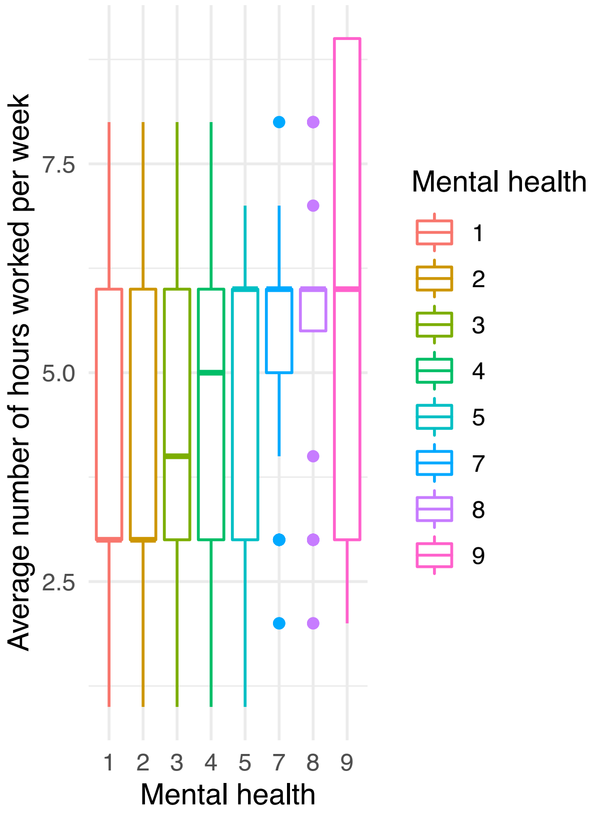
The fourth graph indicates the respondents' job type, of how many of them currently having a permanent or not permanent job. Based on the graph, we can see that half of participants have a permanent job, and only around 2000 respondents have non-permanent (seasonal/temporary/casual/term) job. However, it is surprisingly to see that within all 20,602 people participating this survey, almost half of them chose to skip this question. This phenomenon might make our further analysis difficult to process because as we remove those who skip this question, there are only around 10,000 respondents left. A sample size that is too small might reduce the power of the study and increases the margin of error, which can render the further study meaningless. Similar problem arises when people respond to full time or part time job they have (graph 3), where almost all respondents chose to valid skip this question for some reason. Only 2,000 people report that they are having a full-time job. Since the sample size for actual responded answer is too small, we may pay more attention to this variable if we need it for further analysis. Otherwise, our result might not be as accurate as we expected.

Graph 5: Mental Health by Number of Hours Worked

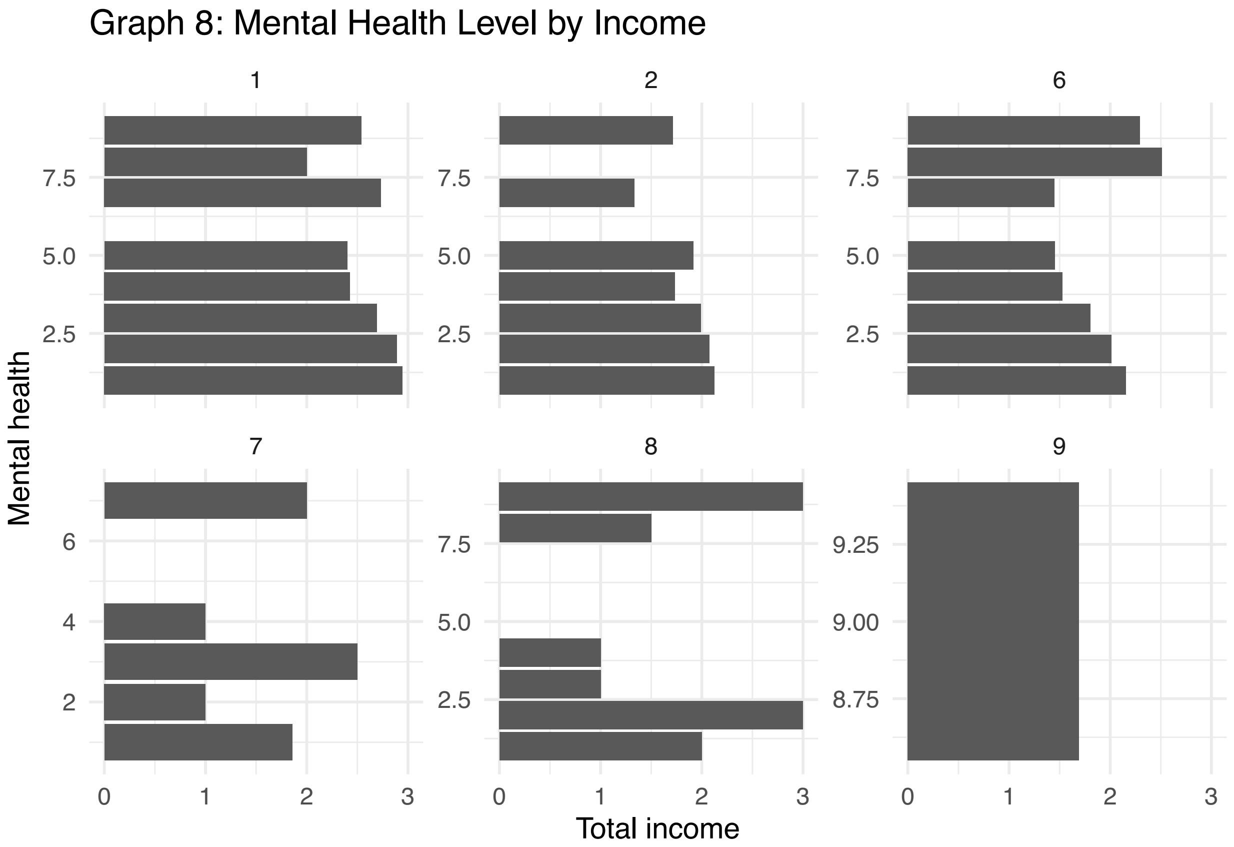


A mosaic plot is used for visualizing data from two or more categorical variables. The categorical variables are first put in order. Then, each variable is assigned to an axis. In our analysis, we want to explore the mental health of respondents according to their number of hours worked per week. At the left edge of the first variable we first plot "Mental Health condition," meaning that we divide the data vertically in eight blocks: the block with number "1" corresponds to "excellent" mental health, "2" corresponds to "very good" mental health and so on as we stated in variable description section. One immediately sees that proportion of mental health condition with "excellent", "very good" and "good" are distributed equally. We then applies the second variable "Average number of hours worked per week" to the top edge. The nine vertical columns therefore mark the nine values of that variable (0 hour, 0.1 to 29.9 hours, 30.0 to 40.0 hours, and so on). These columns are of variable thickness, because column width indicates the relative proportion of the corresponding value on the population. 30.0 to 40.0 hours plainly represents the largest group for respondents with all different degrees of mental health condition. The number of respondents refuse to answer or skip this question is also seen to have been marginal.


Graph 6: Number of weeks employed with different mental health


Graph 7: Working hours with different mental health

We also used mapping to plot the relationship between mental and average number of hours worked per week, as well as the relationship between between mental health and the total number of employment. As can be observed from the graph, for all level of self-rated mental health, the number of weeks employed has the same spread and distribution. The median of all number of weeks employed given the mental health level is at around 75, with minimum reaching 0 and maximum at around 100. This means people's mental health does not quite relate to people's number of weeks employed. On the other hand, people have different level of self-rated mental health seems to have different spread of average number of hours worked per week. For people who rated a low level of mental health, the distribution seems different. For people who rated their health to be really low, they have more long-time working cases. Also, the median of working hours worked per week seems to increase when the mental health level decreases.



We grouped our data by job type and self-rated mental health, and summarized the income accordingly. For people who do not have permanent job, they have mental health to be rated from 1 to 5 more frequently. Around 75% of not permanent job people have mental health rated from 1 to 5. For people who have permanent job, they have mental health rated varies more. The y axis of the bar plots are the income. As can be observed, permanent job tend to have more high income people.

3. Method and Model

A hypothesis expresses a scientific inquiry in terms of a suggested value for a parameter in a probability model. The method of hypothesis testing is one of establishing proof via falsification. It is composed of two critical components: a null hypothesis and a counterfactual hypothesis. The null hypothesis is a specified value for the parameter that specifies the hypothesis we are attempting to refute. Although it is often expressed as a single number, it might be composite. It is referred to as H0. The alternative hypothesis is the one whose truthfulness we are attempting to demonstrate. It is often characterized by a value or collection of values for the parameter that differs from the null hypothesis's value. T test is the difference between group (means) divided by the normal variability within groups. If t is large, the difference between groups is much bigger than the normal variability with groups, which means the two groups are significantly different from each other. If, on the other hand, t is small, the difference between groups is much smaller than the normal variability with groups, which means the two groups are not significantly different from each other. Rejection of H0 happens when $plessthan\alpha$ or the test statistics falls into rejection region. It is shown with very small p-value. When p-value is less than 0.001, there is very strong evidence to reject H0 and support Ha. We use 2 types of error to distinguish between the null hypothesis and the alternative hypothesis, type 1 and type 2 error. The first type of error occurs when the null hypothesis is wrongly rejected. The second type of error occurs when the null hypothesis is wrongly not rejected. Our assumptions for doing this test is that sample independent. The parameter of doing a hypothesis test can be mean, median, proportion and so on. Suppose $X, Y \sim N(\mu, \delta^2)$ with the same δ unknown, and let $x = x_i = X_i(\omega)$ be a lid n sample from X. let $y = y_i = Y_i(\omega)$ be a lid n sample from Y. Let $\mu_{x,y}$ be the (resp) supposed mean of (resp) X, Y from observing (resp.) x, y. Then we have the null hypothesis, $H_0 : \mu_x = \mu_y$.

For this research, I will do a confidence interval analysis using the bootstrap method with a confidence interval of 90%. Our result should provide us with a range within which we may be 90% certain about the real value of our parameter. A confidence interval represents a range of potential true parameter values. More exactly, it approximates a sequence of values that is likely to include an unknown actual parameter. Confidence levels are expressed as a percentage that is proportional to the width of each confidence interval. This % indicates our confidence that the findings accurately reflect the underlying population parameter, based on the bond's luck and your random sample. As I intend to account for 90% of probable outcomes, this indicates that 1- is 90%. The bootstrap sample distribution is normal, and a 90% confidence interval denotes the middle 90%. We will be using Bootstrap in this investigation. Bootstrap is a statistical technique that use random sampling with replacement to infer the sample distribution of a population. The term "resamples" refers to these repeated experiments. The empirical bootstrap is a sort of bootstrap that samples from the sampling distribution of an estimate without defining the data distribution. Because we do not know the distribution of our initial data, we will employ empirical bootstrap in this investigation. Additionally, the z-distribution requires knowledge of the population standard deviation, while the t-distribution requires just the sample standard deviation. We will use the t distribution in our study since we do not know the real value. Additionally, since we do not know the real value of the parameter, we will use a sample estimator to estimate it. Nonetheless, each estimator is unique since each sample is unique (selected at random). In this scenario, the confidence interval may be used to calculate how often an interval will contain the real parameter.

The multiple linear regression (MLR) is used to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. The general form for multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

Where β_i represents the coefficients need to estimated, $x_i, i = 1...p$ is the predictor variables, y is the response variable, ϵ is the error term. Mathematically, β_0 measures where the line intercept y-axis, and $\beta_1 - \beta_p$ measures the slope of the line. More practically, β_o is the value of y when x equals the zero, and $\beta_1 - \beta_p$ is the average change of y when x increase by 1.

4 Result

4.1 Dwelling type VS. Feelings about life

We want to study if the self-rated satisfaction is related to whether dwelling of the respondents is owned or rented. Our hypothesis is that people rented their dwelling have the same self-rated mental health as people who own the dwelling. In this case, H_0 is mean of self-rated mental health for people who rent house. We filtered out people who rent dwelling and calculated their mean self-rated mental health. Our null hypothesis is that there is no difference between people's feeling about life with different dwelling condition. More specifically, we think people have same self-rated mental health for people who rent dwelling and people who own the dwelling. Test statistic is obtained with the formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Then, we use this formula to get p-value: $p - value = P[T \geq t]$ or $P[Z \geq z]$ P-value is less than 0.05, so we reject the null hypothesis, and think there is different in mean value of owned or rented.

4.2 Factors affecting feelings about life

Since our study focus on analyzing potential factors that may influence one's feeling about life, we conduct a regression based on these factors. According to Healthcare (n.d.), feeling anxious around our financial, housing or work situation can make our mental health worse. Being unemployed can dislodge our sense of purpose and may make it difficult to maintain self-confidence. Therefore, we choose variable related to one's financial status to identify its relation with one's feeling. Our assumption for that variable is that they might have a positive correlation. In addition, we also consider age factors as we think age is likely to be one of the largest factor resulting on the variation of feeling towards life. Based on ("Aging and Depression" 2012) , there is evidence that some natural body changes associated with aging may increase a person's risk of experiencing depression. Thus, we chose two different age groups to see if both of them have impact on individual's feeling. Moreover, we also take education level into account as we assume that people with higher education level tend to be more optimistic towards life than those experienced less education. In summary, factors we think that are most influential to feelings about life are:

1. age 25 to 34 years
2. age 75 years and older
3. Owned by you or a member of this household
4. Rented
5. Self Rated Health
6. Self Rated Mental Health
7. Education - Highest certificate, diploma or degree

, and we will then conduct a multiple regression based on these explanatory variables to see if there exist some sort of relationship with feeling about life.

New Model

The new model we built based on these variables is:

$$Feeling\hat{about}life = \hat{\beta_0} + \hat{\beta_1} * Agegroup + \hat{\beta_2} * Dwelling +$$

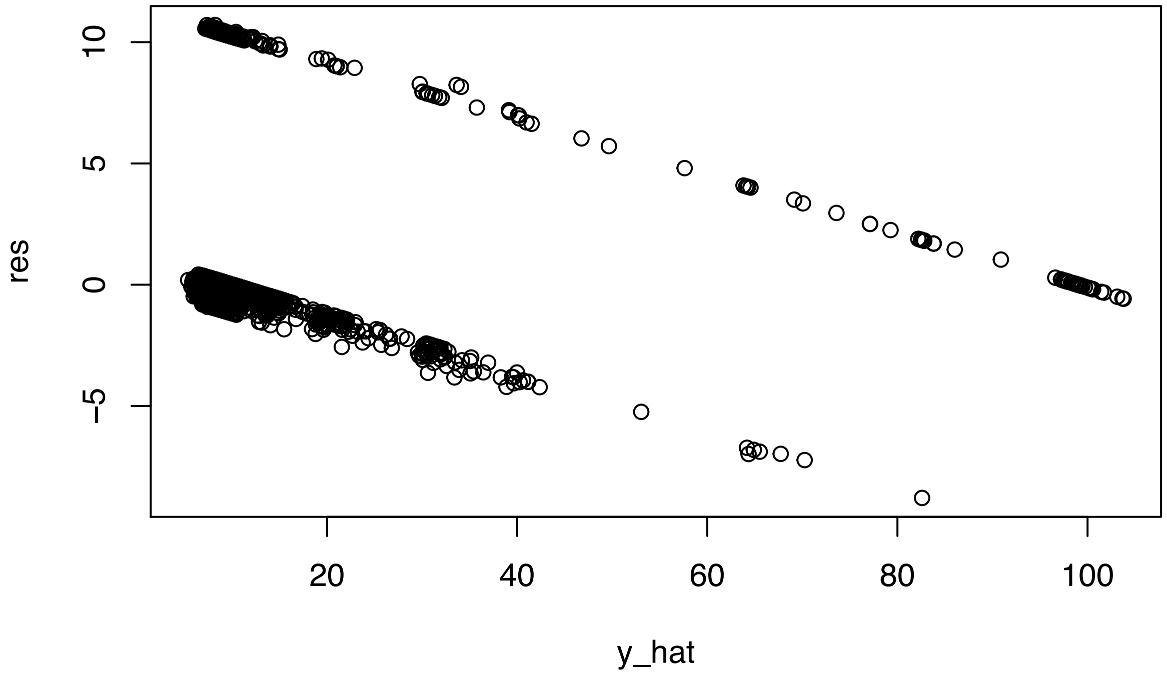
$$\hat{\beta_3} * Selfratedhealth + \hat{\beta_4} * Self-ratedmentalhealth + \hat{\beta_5} * Employweek + \hat{\beta_6} * Workinghours + \hat{\beta_7} * Education$$

4.2.1 Use Residual Plots to identify potential violations against model assumptions

Before we run the result, we need to test the validation of this regression, and this is based on the statistical theory about assumptions of linear regression model. The linear regression model has four assumptions:

linearity, uncorrelated errors, constant variance and normality. As we saw when we were deriving the unbiasedness and the covariance of our estimator, we use the assumptions many times to obtain our results. When all the model assumptions are satisfied, we can then be sure that the estimators will behave in a nice way and have all these lovely properties. However, if even one assumption is violated, this can have a large impact on how we can use our estimates.

We can use residual plot to determine whether there are violations of model assumptions. Residual plots allow us to visually inspect the model assumptions. Moreover, we work with residual plots because the data can sometimes be too noisy to see model violations clearly. There are three main types of residual scatter plots that we use: residuals versus predictor plots, residuals versus fitted values plots, and normal qq plots. Both residuals versus predictor and residuals versus fitted value plots can be used to assess whether our first three assumptions hold. We can check by observing from the residuals plot and if there is no discernible pattern seen in the residual's plots, then the assumptions hold. In other words, to satisfy the assumptions, residual verses fitted predictors plot should not have any pattern or large clusters of residuals. Since residual versus predictors plot can be used when the predictors are numerical, we can only apply residuals versus fitted plot to see if the three assumptions are satisfied. Our plot seems to violate the independence and linearity assumption as we can see there is a pattern shown in this plot and there are two clusters of residuals. This means the result of our regression might not be as accurate as we expected. Moreover, since we cannot use residual versus predictors to check three assumptions, the result we got might not be accurate as well.



4.2.2 Normal Q-Q Plot

To check the normality with residual plot, we do so by using a QQ plot. Normality is verified by using a QQ plot which computes quantiles from the residuals and plot them against the standard Normal quantities. We are expected to see a straight diagonal string of points in the plot with minimal deviations at the ends. However, our plot clearly does not satisfy this assumption as points are not distributed along the diagonal. It looks like follows a bimodal pattern. This means our regression model may violate normality assumption.

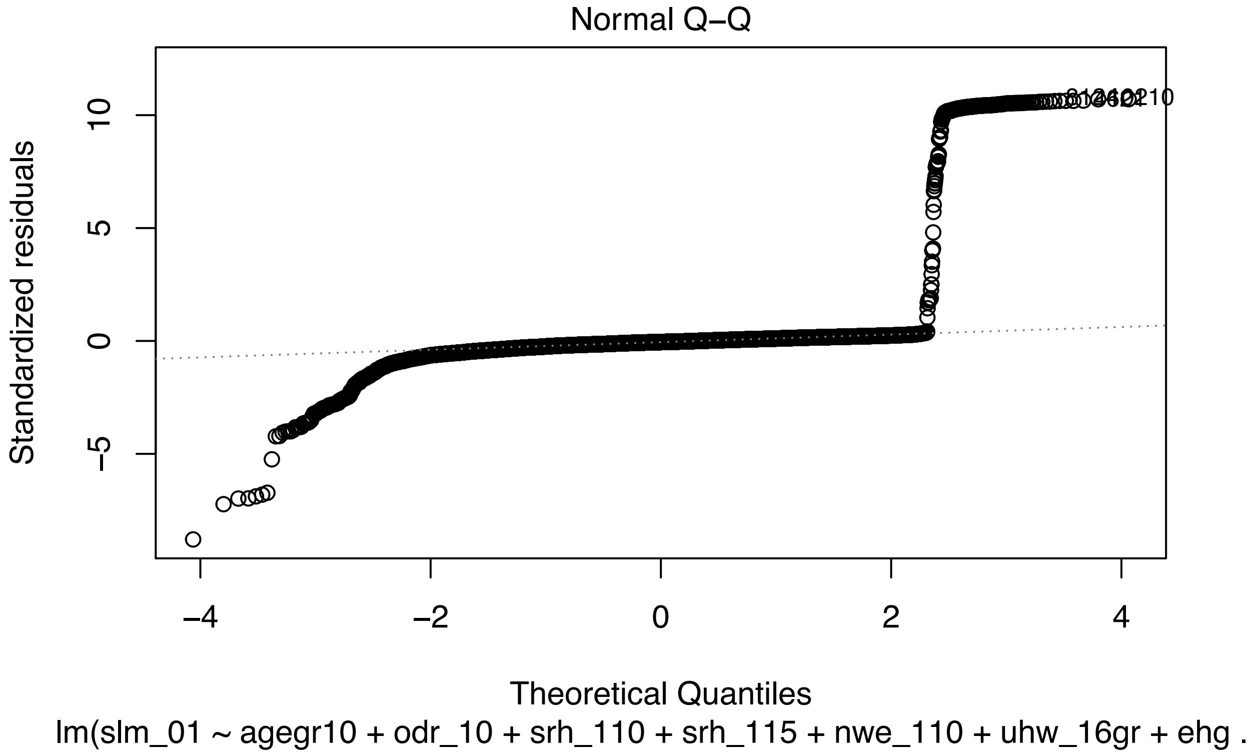


Table 3	Estimate	Std. Error	t value
Intercept	age of respondents	1	15 to 24 years
agegr1025 to 34 years	0.244664	0.289595	0.845
agegr1075 years and older	1.835031	0.315521	5.816
odr_10Owned by you	-5.448789	1.018864	-5.348
odr_10Rented	-5.862449	1.021734	-5.738
srh_110Excellent	-10.680222	1.149512	-9.291
srh_110Fair	-11.294780	1.153213	-9.794
srh_110Good	-11.012880	1.143060	-9.635
srh_110Poor	-11.155725	1.177934	-9.471
srh_110VeryGood	-10.974938	1.144415	-9.590
srh_115Excellent	-21.382963	1.146437	-18.652
srh_115Fair	-22.721808	1.163311	-19.532
srh_115Good	-22.061033	1.143552	-19.292
srh_115Poor	-22.160372	1.236175	-17.927
srh_115VeryGood	-21.932058	1.144297	-19.166
nwe_1102	-0.579957	1.223522	-0.474
nwe_1103	-0.243392	1.224064	-0.199
nwe_1104	0.244015	0.989532	0.247
uhw_16gr0.1 to 29.9 hours	-2.590675	1.767085	-1.466
uhw_16gr30 to 40 hours	-2.418833	1.762162	-1.373
uhw_16gr40.1 to 50 hours	-2.474529	1.772914	-1.396
uhw_16gr50.1 hours and more	-1.949973	1.779697	-1.096
ehg3_01bHigh school diploma or a high school equivalency certificate	0.195949	0.190894	1.026
ehg3_01bUniversity certificate, diploma or degree above the bachelor's level	0.179743	0.243066	0.739

P-values and coefficients in regression analysis work together to be used for showing which relationships in the model are statistically significant and the nature of those relationships. The p-values for the coefficients indicates whether these relationships are statistically significant. The sign of a regression coefficient can tell us whether there is a positive or negative correlation between each independent variable and the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase.

Our regression output can be seen in table 3, where all the coefficients of our predictors and their p value are listed. Based on the table we can see that the — predictor variables are statistically significant because their p-values are less than α. On the other hand, — is not statistically significant because its p-value is greater than the usual significance level of 0.05.

The coefficients describe the mathematical relationship between each independent variable and the dependent variable. Since our predictors are all categorical, the concept that $beta_o$ is the mean of the group that constitutes the reference level for both (all) categorical variables, giving the end interpretation that the regression coefficient is the difference in mean of the two categories.It can be seen that the coefficient of variable "age" with age group 25 - 34 is 0.245 while the coefficient with age group 75 years and old.

5. Discussion

5.1 Survey Methology

The poll is conducted over a six to twelve month period, with an average interview length of 40 to 45 minutes. The two primary methodologies are random digital dialing (RDD) and computer-assisted telephone interviewing (CATI) (CATI). The 2017 GSS on Families is the first time the new GSS frame has been utilized, which incorporates data from sources of telephone numbers (landline and cellular) available to Statistics Canada, as well as the Address Register (AR). Additionally, the sample unit is stated in terms of groups of telephone numbers rather than individual numbers. The survey frame was created by fusing two distinct sets of data. Statistics Canada has access to a variety of listings of cellular and landline telephone numbers in use published by various organizations (telephone companies, Census of population, etc.). The Address Register (AR) is a database of all homes in each of Canada's ten provinces. The Address Register (AR) was used to bring together all telephone numbers associated with the same genuine address. Around 86 percent of available telephone numbers were affiliated with the AR. It is possible for the records produced as a result of this linkage to have several telephone numbers (grouped by the address). It was decided to include in the frame the remaining 14% of telephone numbers that were unrelated to the AR2. As a result of the integration of these two components, the survey frame was formed. The purpose of using all telephone numbers (both AR-connected and non-AR-connected) was to ensure that all residences with telephone numbers were appropriately covered. When a record had several telephone numbers, the numbers were sorted according to the source of the telephone number and the kind of telephone number (landline telephone numbers first and cellular telephone numbers last). When it came to contacting the homeowner, it was widely accepted that the first telephone number supplied was the best. Please bear in mind that for the remainder of this article, the word "record" will refer to the collection of telephone numbers that comprise our sample unit on the survey frame.

5.2 Findings

In this paper, we focus on examining individuals' feelings about life as a whole in related to possible family-correlated factors. In order to get this done, we used different methods and developed multiple models. We read through the General Social Survey - Cycle 31 : Families - Public Use Microdata File Documentation and User's Guide and the 2017 General Social Survey: Families Cycle 31 - Public Use Microdata File codebook. Then we selected some variables that we thought and believed might be useful and relative to the topic we are focusing on. Our dependent variable is feelings about life as a whole. Our independent variables are age group of respondent (groups of 10), number of respondent's children in household - any age/marital status, full-time/part-time job, income of respondent - total (before tax), dwelling - owned or rented, self rated health, self rated mental health, number of weeks employed - past 12 months, average number of hours worked per week, province of residence of the respondent, marital status of the respondent, education - highest certificate, diploma or degree, and living arrangement of respondent's household (12 categories). To find out which of these independent variables are influential to the dependent variable, we used three different methods: Linear Regression Model, Hypothesis Test, and Confidence Interval Analysis. We also conducted many graphs and tables on some variables to illustrate our contents and delivery useful information more detaily.The first thing we learned after this paper is that age group of respondent (groups of 10), dwelling - owned or rented, self rated health, self rated mental health, and education - highest certificate, diploma or degree are the influential factors to feelings about life as a whole are the most influential factors to feelings about life as a whole.

Weakness, Potential and Future

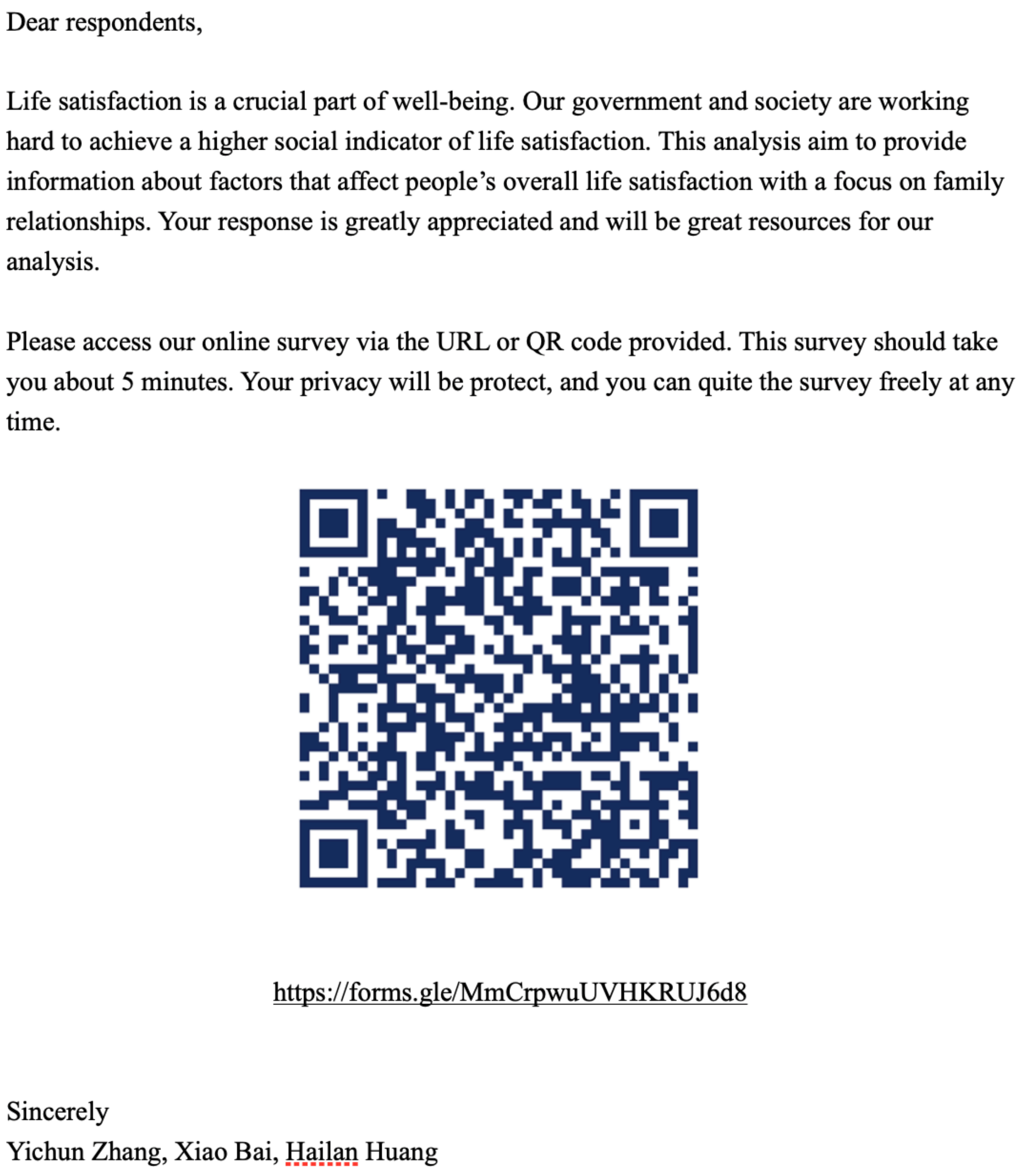
In order to check the accuracy of our Linear Regression Model, we created a normal Q-Q plot of the new model which shows if residuals are normally distributed. It is illustrated that around one third of the residuals are off the straight dashed line. Overall, it is not good enough to say that residuals are normally distributed since it is heavy tailed. The reason for this weakness of the Linear Regression Model might be that there are too many categorical variables in this 2017 GSS dataset, selecting one suitable variable is not hard, but when it comes to many suitable variables need to be carefully selected, we sometimes might overlook some details and make mistakes. Therefore, it is hard to create a perfect model to demonstrate our analysis. To round things up, we may have ignored the issue of potential variable bias in our study. Outside of one's finances and whether or not one has children, there are many other elements that might have a big influence on one's degree of life satisfaction, such as one's education, sexual orientation and religion. We may have overestimated the relevance of the variables that we included by removing these components from our model. This is because the predictor variables are no longer distributed independently of the error term, leading to incorrect conclusions.

We can include more variables from the GSS dataset codebook to set up the models and see if there will be a possible better model than the ones we created previously in this paper. Also, we can introduce other methods to fit a model, such as cross validations and box-cox transformations.One limitation of the 2017 GSS dataset is that its target population only included all persons 15 years of age and older in Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut, and full-time residents of institutions ("2017 General Social Survey: Families Cycle 31 Public Use Microdata File Pumf" 2017). Therefore, the dataset is considered to be conservative. These groups that are excluded from the survey's target population might even have more family issues and higher significant contributions to different levels of feelings about life as a whole. We can also conduct a follow-up survey on testing respondents' feelings or satisfaction level on life. In the follow-up survey, we can conclude some open-ended questions based on our paper's topic. From the follow-up survey, we can find out if the variables we selected in the above models are actually influential, if not, this survey can help us on improving future research area. We can also investigate on other possible factors mentioned in the follow-up survey which are not included in our above models.Thus, a follow-up survey will be very necessary for us to proceed in the future. With this, we can learn more on the family and feelings about life as a whole topics, and dig deeper into these areas.

The conclusions were obtained via the use of multinomial regression and then compared to the existing outcomes. More rigorous statistical methods should be used to identify significant independent variables. When too many independent variables are included, the issue of omitted variable bias is compounded by a problem of multi-collinearity, which increases the number of independent variables. Because multicollinearity increases the standard error of our estimations, there is a trade-off between omitted variable bias and multicollinearity. Accumulating meaningful sets of independent variables for our research may require using either the Akaike or Bayesian information criteria. The most important metric is the mean time between visits, which is unaffected by other variables.

6. Appendix

Survey: https://forms.gle/MmCrpwuUVHKRUJ6d8



Dear respondents,

Life satisfaction is a crucial part of well-being. Our government and society are working hard to achieve a higher social indicator of life satisfaction. This analysis aim to provide information about factors that affect people's overall life satisfaction with a focus on family relationships. Your response is greatly appreciated and will be great resources for our analysis.

Please access our online survey via the URL or QR code provided. This survey should take you about 5 minutes. Your privacy will be protect, and you can quite the survey freely at any time.

https://forms.gle/MmCrpwuUVHKRUJ6d8

Sincerely
Yichun Zhang, Xiao Bai, Hailan Huang

7. Reference

"2017 General Social Survey: Families Cycle 31 Public Use Microdata File Pumf." 2017. *Statistics Canada.*

"Aging and Depression." 2012. *American Psycological Association.*

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2020. *Lahman: Sean "Lahman" Baseball Database.* https://CRAN.R-project.org/package=Lahman.

Healthcare, Northern. n.d. "What Factors Affect Our Mental Health? Understanding the Social, Physiological and Environmental Impact."

"Life Satisfaction." n.d. *OECD Better Life Index.* https://www.oecdbetterlifeindex.org/topics/life-satisfaction/.

Livni, Ephrat. 2018. "A Nobel Prize-Winning Psychologist Says Most People Don't Really Want to Be Happy." *QUARTZ.*

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Veenhoven, Ruut. 1996. "THE Study of Life Satisfaction."

Wang, Wangshuai, Gong Sun, Zhiming Cheng, and Xin-an Zhang. 2017. "Achievement Goals and Life Satisfaction: The Mediating Role of Perception of Successful Agency and the Moderating Role of Emotion Reappraisal." *Psicologia: Reflexão E Crítica.*

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.