Finding the Factors That Have A Significant Impact on Canadians' Feelings About Life As A Whole

Boxiang Tang, Xinyu Tan, Muxin Zhang, Shiyun Tang

October 19th

Abstract

We cannot deny that society continues to evolve and undergo rapid changes, and thus it is not unusual that people nowadays are under tremendous pressure. As more and more people have been diagnosed with mental health disorders, the self-rated well-being of Canadians remains an ongoing interest. In particular, we are interested in how some specific family-related factors affect Canadians' feelings of life as a whole. In this report, we will build a linear regression model between Canadian's feelings about life as a whole and their age, number of children, family income, average number of hours worked per week and age at first marriage. According to our model, it is family income that has the most significant effect on Canadians' feelings about life. Age and age at first marriage can also influence Canadians' feelings about life, but the effect is minor. Therefore, we conclude that in order to effectively improve people's well-being, the Canadian government should strive for a more prosperous economy so that there will be an increase in family income among all Canadians.

Introduction

Feelings about life as a whole is something that is very abstract and mysterious. Each individual may have a distinct standard of judging to what extent his/her life is satisfying. Despite the fact that feelings about life can be extremely subjective, there is no doubt that people today all share the same aspiration, which is to improve their general feelings about life. Therefore, our goal of the study is to determine what common factors can have an effect on feelings about life as a whole for all Canadians. The conclusion can in turn shed light on what kind of people tend to feel good about life and what kind of people tend to feel bad about life. With our model, people who have bad feelings about life and are more vulnerable to getting mental health disorders can be detected. They can then receive mental health support from professionals. There will be less people suffering from mental illnesses and the country will become a better place.

In this report we will explore how Canadians' feelings about life as a whole are determined by five variables, which are their age, number of children, family income, average number of hours worked per week and age at first marriage. We chose "age" because we wish to see if feelings about life can vary with increased life experience as one gets older. We chose "number of children", for it is both rewarding and demanding to raise a child, and thus we wish to see whether an increase in the number of children gives rise to more happiness or more burden. In addition, we also chose "family income" and "average number of hours worked per week", which are two variables that are likely to be positively and negatively correlated with one's feelings, respectively. We also chose "age at first marriage". Going into marriage means one needs to take care of not only him/herself but also someone else. The earlier one gets married, the greater overall responsibility one takes in his/her lifetime. On the other hand, getting married also means sharing joy and happiness with another person. Thus we assume that the age when the first marriage started can affect one's feelings about life either positively or negatively.

In the following sections, we will first provide some general information of the dataset we chose, the reasons why we chose it and its strengths and drawbacks. We will then introduce the model and interpret the results of it. Based on the results, we will comment on the effectiveness of our model and how it contributes to our goal, which is to determine what family-related factors can have an impact on Canadians' feelings about life as a whole. Last but not least, we will discuss if there are any limitations of the data, drawbacks of the survey and weaknesses of our analysis.

Data

The dataset we use for this report is from the General Social Survey (GSS), 2017: Family. The main objective of the study is to have a deeper understanding of Canadian families as of 2017 and to get some ideas about what Canadian families will be like in the future. The stratified sampling method (STRS) is applied for choosing the sample. Under this method, the population is first sliced into homogeneous groups called strata. Units are then selected at random from each stratum (i.e., Simple Random Sampling). These units together form the whole sample. Notice that each stratum is entirely different from one another, but within the same strata, elements are homogenous. Applying STRS effectively reduces bias and variabilities of the results. However, this method can be time-consuming and cause waste of resources. For the 2017 GSS Survey, the basis for stratum dividing is the list of "Census Metropolitan Areas" (i.e., St. John's, Halifax, Saint John, Montreal, Quebec City, Toronto, Ottawa, Hamilton, Winnipeg, Regina, Saskatoon, Calgary, Edmonton and Vancouver); also, there are three more strata formed by the rest of areas that are not on the list. All data was collected by computer assisted telephone interview format (CATI), which means all respondents were sent to fully trained interviewers to complete the survey via centralized telephone. The target population is all noninstitutional persons who are 15 or older in Canada except for residents of the Yukon, Northwest Territories, and Nunavut. The frame population is the people on the lists of telephone numbers in use that are available to Statistics Canada from various sources and the members on the Address Register (AR), which is the list of all dwelling within ten provinces (Beaupré, 2020).

There are various reasons why we chose the dataset. Firstly, it is the most recent data we could access from the GSS. The world is constantly changing, and so is the data collected. We use the dataset from 2017 so that our analysis of the data is still valuable and can shed light on the present situation. By choosing the 2017 dataset, we can also minimize the inaccuracy of our conclusion caused by data overage. Secondly, the dataset is very informative with a huge amount of analyzable variables. This makes it highly selective, and thus we could easily define what we want to study. Thirdly, all the variables and information of the data are readable and easy to understand. In this way, we can quickly decide what our topic of study and variables of interest are.

However, there are also several drawbacks to the dataset. Firstly, there are a great number of categorical variables but few numerical variables, making it hard for us to build estimation models and draw statistical graphs. Secondly, not only does the dataset contain information about respondents, but it also contains information about their spouse, children and parents. Thus the whole dataset looks quite messy, which makes it challenging to extract desired data. Thirdly, there is a lot of missing data (marked as "NA") in the dataset. Therefore, the statistical strength of our study would be reduced, our estimation could be biased, and our conclusion might also be invalid.

Model

We will fit and describe a linear regression model between Canadians' feelings about life as a whole and their age, number of children, family income, average number of hours worked per week and age at first marriage. Since Canadians' feelings about life is what we are interested in (i.e., the variable we wish to predict), it is the response variable. The other five variables, which act as predictors (i.e., their values are given, and we put them into our model to get the estimated value of the response variable), are explanatory variables.

The linear regression model is a statistical model which relates the numerical response variable to all explanatory variables by a linear equation Y = ax1+bx2+…+C. Notice that our response variable is categorical; to make it numerical, we divide feelings about life into eleven levels and assign numerical values 0-10 to each level in an ascending order (i.e., the higher the number, the better the feeling). Then we can treat the response variable as a numerical variable and compare different feelings based on their assigned numerical values. Thus the linear regression model perfectly fits our demand for estimating a numerical value that corresponds to a specific level of feelings given the value of each predictor.

The mathematics behind the model we use is as follows: Y= b0+b1x1+b2x2+b3x3+b4x4+b5x5, where b0 is the value of y-intercept which can be interpreted as the expected value of Y when xi = 0 for i = 1,. . . ,5 (Note: usually, it does not have any practical meaning), and bi's are coefficients representing change in the expected value of Y for every one unit of increase in xi's for i=1,. . . ,5.

If the model we build is effective, we are then able to estimate one's feelings about life as a whole given his/her age, number of children, family income, average number of hours worked per week and age at first marriage.

```
##
## Call:
## svyglm(formula = feelings_life ~ age + total_children + age_at_first_marriage +
##     as.factor(average_hours_worked) + as.factor(income_family),
##     design = ucla.design, family = "gaussian")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4701 -0.7804  0.0981  1.2423  3.1443
##
## Coefficients:
##                                                     Estimate Std. Error t value
## (Intercept)                                         6.549267   0.704950   9.290
## age                                                 0.008555   0.003425   2.497
## total_children                                      0.035824   0.023345   1.535
## age_at_first_marriage                              -0.014822   0.006328  -2.342
## as.factor(average_hours_worked)0.1 to 29.9 hours    0.994481   0.656570   1.515
## as.factor(average_hours_worked)30.0 to 40.0 hours   0.867299   0.654554   1.325
## as.factor(average_hours_worked)40.1 to 50.0 hours   0.900709   0.660519   1.364
## as.factor(average_hours_worked)50.1 hours and more  0.991548   0.664378   1.492
## as.factor(income_family)$25,000 to $49,999          0.127665   0.133158   0.959
## as.factor(income_family)$50,000 to $74,999          0.377414   0.134774   2.800
## as.factor(income_family)$75,000 to $99,999          0.488016   0.142848   3.416
## as.factor(income_family)$100,000 to $ 124,999       0.769033   0.154031   4.993
## as.factor(income_family)$125,000 and more           0.867845   0.134462   6.454
##                                                    Pr(>|t|)
## (Intercept)                                         < 2e-16 ***
## age                                                0.012577 *
## total_children                                     0.125035
## age_at_first_marriage                              0.019246 *
## as.factor(average_hours_worked)0.1 to 29.9 hours   0.129995
## as.factor(average_hours_worked)30.0 to 40.0 hours  0.185295
## as.factor(average_hours_worked)40.1 to 50.0 hours  0.172814
## as.factor(average_hours_worked)50.1 hours and more 0.135719
## as.factor(income_family)$25,000 to $49,999         0.337786
## as.factor(income_family)$50,000 to $74,999         0.005147 **
## as.factor(income_family)$75,000 to $99,999         0.000646 ***
## as.factor(income_family)$100,000 to $ 124,999      6.40e-07 ***
## as.factor(income_family)$125,000 and more          1.32e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.593 on 2311 degrees of freedom
## Multiple R-squared:  0.04413,    Adjusted R-squared:  0.03917
## F-statistic: 8.892 on 12 and 2311 DF,  p-value: < 2.2e-16
```

Results

$FeelingsLife = \beta_0 +$
$\beta_1 Age +$
$\beta_2 TotalChidren +$
$\beta_3 AgeOfFirstMarriage +$
$\beta_4 AverageHoursWorked(0.1-29.9) +$
$\beta_5 AverageHoursWorked(30-40) +$
$\beta_6 AverageHoursWorked(40.1-50) +$
$\beta_7 AverageHoursWorked(50.1-) +$
$\beta_8 IncomeFamily(25000-49999) +$
$\beta_9 IncomeFamily(50000-74999) +$
$\beta_1 0 IncomeFamily(75000-99999) +$
$\beta_1 1 IncomeFamily(100000-124999) +$
$\beta_1 2 IncomeFamily(125000-)$

where $\beta_0 = 6.549267$, $\beta_1 = 0.008555$, $\beta_3 = -0.014822$, $\beta_9 = 0.377414$, $\beta_{10} = 0.488016$,$\beta_{11} = 0.769033$, $\beta_{12} = 0.867845$, and $\beta_{2,4,5,6,7,8} = 0$ since their p-values are greater than 0.05. Here $\beta_0 = 6.549267$ means that when the values of all five explanatory variables are 0, our estimated value for feelings about life is 6.549267; however, this number does not have any practical meaning here. For all those numerical values of β_i, where i = 1, . . . , 12, they represent the value of change when there is a one unit increase in their assigned variables.

In this model, average_hours_worked and income_family are categorical variables and they are also dummy variables.

For the model we have a low R-square which is 0.04413, and a low p-value which is 2.2e-16 (p-value <= 0.05). It means the model might not make accurate predictions but the predictors we used have an effect on the independent variable. Due to the low p-value, the feelings of life do change when the five independent variables change.

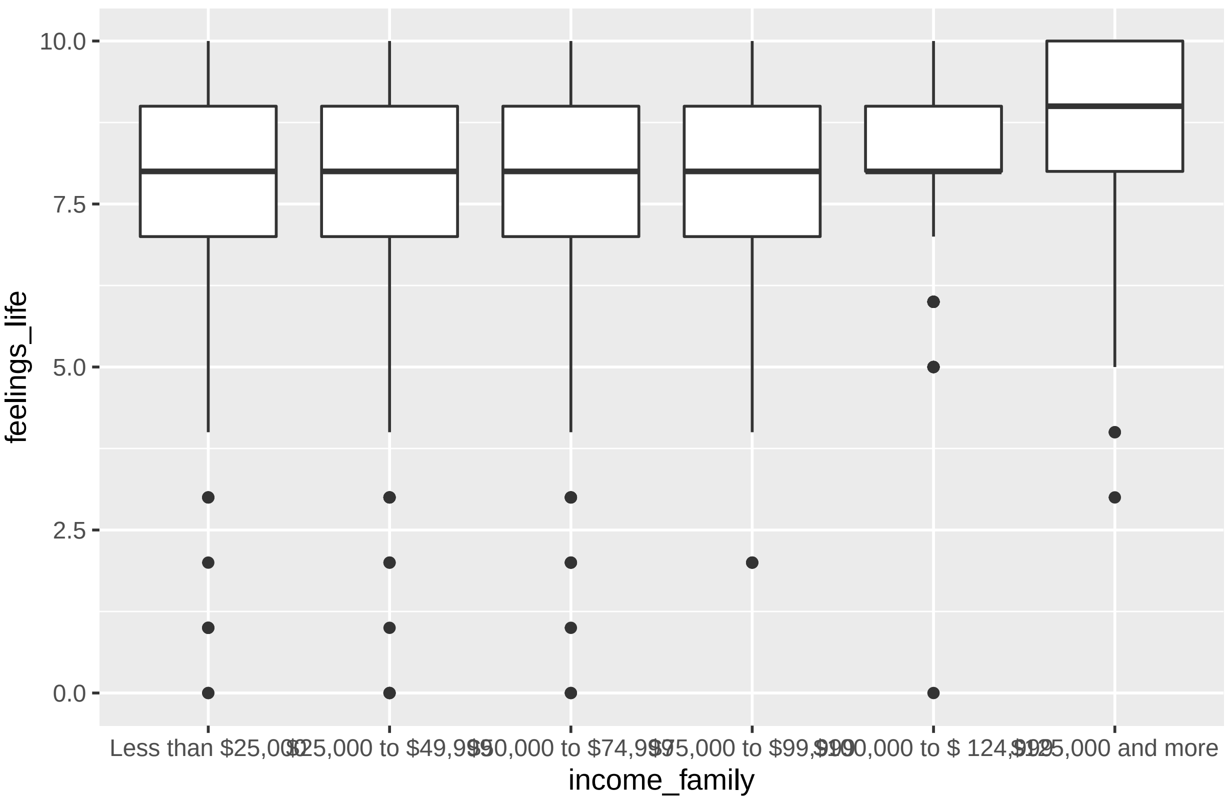
The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable.

For the numerical variables, we can see that the number of children does not have a significant impact on feelings of life because its p-value is 0.12577, which is bigger than 0.05. In contrast, the p-values of both age and age at first marriage are smaller than 0.05, which means we have strong evidence to reject the null hypotheses. Hence these two variables do affect people's feelings of life.

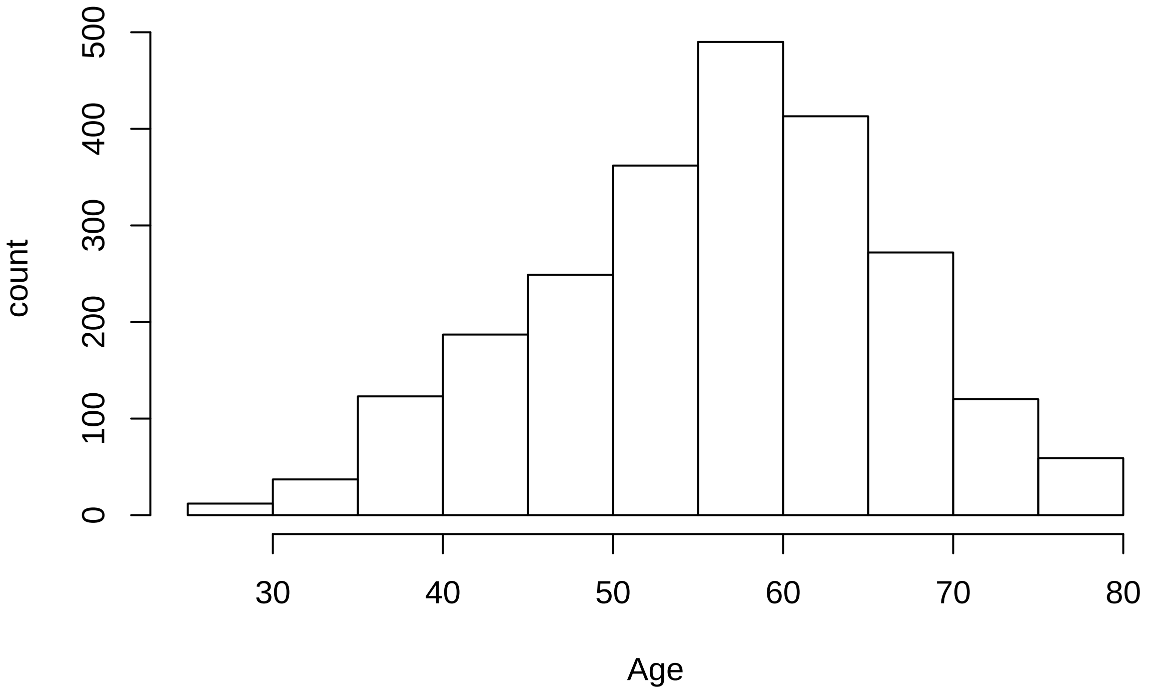
For the categorical variables, since all of the p-values for average work hours are bigger than 0.05, we can say there is no association between the change of average work hours and the feelings of life. For the family income variable, we have a small enough p-value for income over \$50000, which means we have enough evidence to reject the null hypothesis. Thus we conclude that feelings of life do increase when family income increases given it is larger than \$50000.

To sum up, the factor that has a major influence on people's feelings of life is income. We can see from the summary table that the estimate of β increases as the income level goes up. The boxplots(figure 1) shows the difference in Canadians' feelings of life. There are many smaller outliers which indicate that the distribution of our sample is both highly tailed and highly skewed. Also, we can see from the boxplots that each plot has a longer lower whisker which represents a left- skewed distribution. Generally, the median value for our boxplots group is approximately 8.125 (i.e., 7.5 + 2.5/4). However, notice that Canadians feel good about their life qualities on average (with a mean level of feelings of 8.10) regardless of the amount of family income. The left-skewed histogram (figure 2) also illustrates a high satisfaction of life quality overall. On the other hand, age and age at first marriage have a relatively small influence on people's subjective well-being. In particular, age at first marriage is negatively correlated to the feelings of life.

Boxplot between income_family and feelings_life(figure 1)



Histogram of Respondents Age(figure 2)



Discussion

As mentioned in the Results section, our model shows that while family income plays a significant role in influencing Canadians' feelings about life as a whole, age and age at first marriage have a trivial influence on people's feelings of life. Number of children and average hours worked per week might not affect people's feelings of life. Hence the original goal of our study, which is to determine what factors generally can have an effect on feelings about life for all Canadians, is achieved. Besides, it is conceivable that there should be some other variables, which we had not taken into account, influencing people's feelings about life as a whole. In this sense, our study has laid a foundation for future studies of the same topic.

The conclusion that family income affects Canadians' feelings about life most gives insight into what the Canadian government can do in order to generally improve people's well-being. The key is to boost the economy so that there are more employment opportunities and better salaries. The government should also take into consideration the fact that people who have a low family income are likely to have bad feelings about life and thus may be at higher risk of developing mental illnesses. The government should provide those people with mental health support whenever necessary. As time goes by, there will be fewer and fewer cases of mental health disorders and the country will become more harmonious and a better place.

Weaknesses

The weaknesses of our study can be categorized into different types, which are data drawbacks, methodological flaws, non-response issues and explanatory variables' ineffectiveness, respectively.

For drawbacks of the data itself, as we have mentioned in the Data section, there are many categorical variables but few numerical variables, too much information and a lot of missing data. These together make it difficult for us to extract useful data, build appropriate models and make statistically sound conclusions.

For the methodology of the survey, we have identified two potential weaknesses. Firstly, there can be undercoverage bias. Since the interview is conducted via telephone, people who belong to the target population but do not have a telephone can never be selected by the researchers. Furthermore, people who belong to the target population but are not in the sampling frame (i.e., available lists of telephone numbers and lists of addresses) could never be selected either. There are also flaws in the design of the survey questions and the implementation of the survey. Firstly, the survey is too lengthy and thus respondents will possibly lose patience as the interview progresses, which in turn may result in a decrease in the accuracy of responses. Secondly, some aspects of the survey are quite personal and involve the respondents' privacy. This can prevent the respondents from telling the truth. Thirdly, despite the fact that the interviewers are fully trained, there might still be communication barriers occuring. The respondents might then misunderstand what is being asked and provide incorrect information. Fourthly, there can be over reporting for some questions. In other words, respondents might tend to provide inaccurate information to make themselves seem advantaged. Fifthly, using the stratified sampling method to obtain the sample population is both costly and time-consuming. Lastly, the proportion of number of strata in CMA's drastically exceeds that in non-CMA's so the data may not be representative of people in non-CMA's.

We have also noticed that the response rate of the survey is only 52.4%, which is not high enough for us to be confident that our conclusion can be applied to the population.

Last but not least, it has been brought to our attention that some of the explanatory variables we chose may not be effective enough in predicting people's feelings about life as a whole. For example, many of the respondents who are single and thus generally have a lower family income than married respondents still have good feelings about life. In this sense, respondents' personal income may be a more effective predictor than family income is. In addition, many respondents do not have any wedding experiences but still feel good about life, which implies that the variable "age at first marriage" may play an insignificant role in influencing people's feelings about life. Finally, different people have different personal experiences, which largely affect their perspectives and feelings about life. Therefore, it may be the personal experiences that essentially influence people's feelings about life rather than age.

Next Steps

In order to improve the quality of the data, a follow-up survey that mainly focuses on numerical variables and asks the exact values rather than intervals can be conducted. We can build a new linear regression model, which will undeniably better contribute to our goal.

In order to improve the methodology of the survey, we can first make the number of strata in CMA's proportional to that in non-CMA's. With such a fair redistribution, we can get a model that is representative of both people in CMA's and people in non-CMA's. We can formulate simpler survey questions, making it easier and more comfortable for respondents to give responses. Then there will be less missing data. We can change the way of distributing the survey. Instead of reaching the sample population via telephone, we can print the survey on pieces of paper and distribute them. In this way, more people who are in the target population could be involved. Also, we do not need interviewers anymore, and thus issues of communication barriers could be fixed. We can use the simple random sampling method, rather than the stratified sampling method, for obtaining the sample population, which saves time and money and makes it easier for data collection.

In order to increase the response rate, we can make the survey shorter, clearer and more readable so that more people would like to participate.

In order to make our model more useful, we can replace the aforementioned non-effective explanatory variables with other variables that are potentially more effective. We then build a new model, compare it with the previous model and decide if the new one is more useful. We can repeat this procedure several times and find the best one.

References

1. Beaupré, P. (2020). General Social Survey, Cycle 31: Families, Public Use Microdata File Documentation and User's Guide. Authority of the Minister responsible for Statistics Canada.
2. Alexander, R., & Caetano, S. (2020). Data Cleaning Code.
3. Wickham, H. et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
4. Lumley, T. (2020). Survey: Analysis of Complex Survey Samples. R package version 4.0.
5. Wickham, H., & Henry, L. (2020). tidyr: Tidy Messy Data. R package version 1.0.2. https://CRAN.R-project.org/package=tidyr
6. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
7. Wickham, H., Hester, J., & Francois, R. (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

Link

https://github.com/SHIYUN-TANG/STA304