Factors Related to Canadian Current Situation on Mental Health Condition

Binqi LI(1004871123), WenXuan Zhai(1004831890), Ziqing Lyu(1005022093), Xue Shan(1004173262)

October 19th, 2020

1. Abstract

Considering the development of our current society and its improvement in our living standards, one of the most concerning topics is mental health. Taking into account that mental health could impact various aspects of our lives, this kind of physical issue is mainly generated based on an individual's life encounter. With the provided dataset, it attracts our attention to investigate the mental health distribution among the population and what life conditions specifically have an impact on one's mental health.

With the given dataset GSS31, Our investigation will be conveyed through r-studio in developing models that eliminate unrelated factors, further selecting potential society-related variables, to construct a regression model to certify the correlation.Hopefully, our developed model and discovery can provide people with a more meaningful and comprehensive insight to learn more about Canadians' current situation of mental health.

2. Introduction

Our primary goal in this topic is to analyze factors that potentially affect Canadian mental health. These factors are divided into two categories: living environment and personal ability. Living environments specifically correspond to variables conducted by family factors which include living arrangements, and a number of marriages. Personal ability is the factors relating to an individual's social and personal condition including age, sex, region, and respondent income level. Based on the selected condition above, the goal for this project is to testify our model to finally against our null hypothesis of suggesting non-correlated relation between mental health and our proposed factors.

To further testify our initial proposal, various kinds of plots will be drawn to help visualize the current mental health situation within factors in Canada. Based on our model analysis results on the data, utilizing p-value to evaluate our hypothesis of whether those living conditions determine the mental well-being of individuals, hopefully, our project could obtain a more significant distribution of the Canadian mental health situation. More detail will be illustrated in the next few sections, following the sequence of understanding the given database and further evidence on the response of social well-being will be applied to fulfill our goal.

3. Data

3.1 Introduction to the Dataset

To conduct our projection, we mainly operate our data based on the dataset “General Social Survey Cycle 31: Family”, a cross-sectional design-based survey that took place among 10 provinces of Canada during 2017, investigating issues related to Canadian’s living status and life satisfaction.

3.2 Methodology and Sampling Approach

The 2017 GSS target population is composed of individuals who are at the age of 15 and order with the exclusion of specific regions(Yukon, Northwest Territories, and Nunavut) and individuals who are full-time residents of the institution. To further execute sampling, the method stratification is used to separate 10 provinces into strata, as well as to operate the method of computer-assisted telephone interviews(CATI) and reach out to the frame population under-recorded telephone number and the address register within strata to fulfill the survey. In addition, the record survey frame is a simple random sample without replacement in each stratum.

The phone interviewer was mainly calling from 9:00am-9:30am on weekdays and participated longer hours from 10:00am to 5:00pm on Saturdays, and 1:00pm to 9:00pm on Sundays in which scheduled the most possible available time in reaching out to the interviewees. During the interview, two or more attempts were made to the interviewees who initially refused to respond, the further explanation and encouragement were made to the participants underscoring the importance of contributing to the survey. Moreover, appointments were organized for interviewers to call back for some are not available in the first call. Interviewers were also making sure to record as many participants as possible in making numbers of calls back to unanswered household calls. The non-response survey results greatly contributed to the non-sampling error. The total non-response is placed into the household who did respond to the survey.

The result turns out that 91.8% of the given telephone number is reached out successfully and finally made up to an actual number of 20,602 responses with the addition of 602 observations than the primarily designed target sample size.

Table 1: Number and Percentage of Survey Participants

```
                        Number       %
Household non-response 14687.0    37.4
Response(CATI)         20602.0    52.4
```

3.3 Key Features

2017GSS is the dataset that took a comprehensive view in reflecting the contemporary well being of families in Canada, as the contents of the survey followed the latest social issue(ex. Mental health, feelings of life) along with related life-standard variables(ex.income, gender, age) which conform our main purpose of investigating correlations and potential factors influencing the people's self-evaluation on mental health.

3.4 Weakness of Dataset and Questionnaire

It is noticeable that there are many missing values under several variables. For example, the question related to financial support for kids resulted in up to 97.8% of skips and only 545 participants provided valid answers. The small sampling population may result in a great inaccuracy in representing the target population. Therefore, this type of variable account to a large proportion of missing value would not be the priority to take further investigation.Thus, it is worth to take consideration of how the question should be conveyed in order to reduce skips.

The survey has conducted on many latest social condition among individuals providing comprehensive information reflecting one's well-being under 81 questions, it is a time-consuming process in fulfilling the whole survey, the interviewee may lose their patience as the time of survey exceed their potential self-preset time, which leads to debase on accuracy among individuals as more questions are being asked.

3.5 Visualize the Data

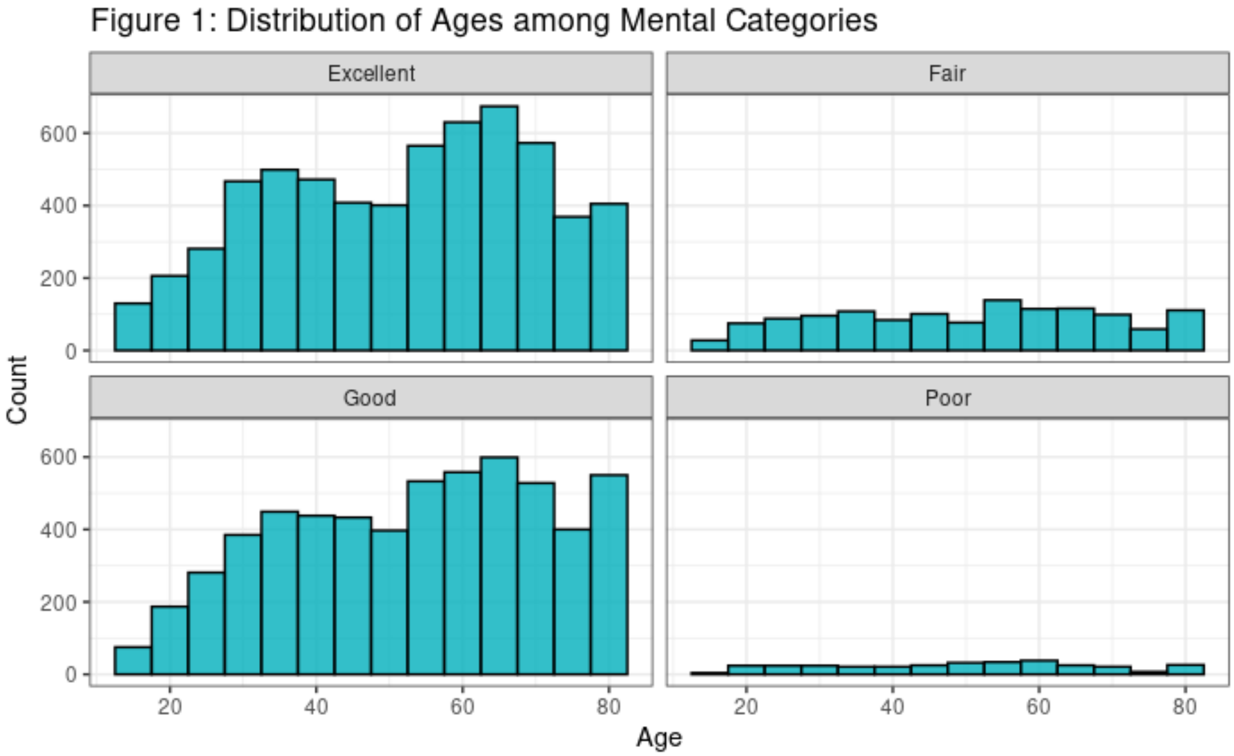
Figure 1: Distribution of Ages among Mental Categories

The histogram demonstrates the distribution of survey participants' different self-evaluated mental health based on ages 15 years and older. Four graphs illustrate the separate distribution of different answers. The histogram is composed of both the y-axis and x-axis which reflect the number of counts and ages respectively.

Based on the presentation of the graph above, the positive responded mental health question is distributed in a similar trend. The age with respect to the number of counts demonstrating an unequally distributed left-skewed histogram peaked at age of 60 to 70 without outlier under positive self-aware mental health. Negative self-rated mental health including answers of "Fair" and "Poor", however, demonstrating an equally distributed histogram without an evident peak and outlier. It is noticeable that the number of counts towards negative self-rated mental health is far less than the number who felt satisfied with their mental health.

The comparison between the number of distributions towards self-evaluation mental condition demonstrates a great proportion of our samples which further indicates that the majority of our target population are under good mental health conditions. The right-skewed histogram suggests that people felt positively towards their mental conditions as the age gets older, senior at the age from approximately 60 to 70 years old shown to have a highly distributed answer of "excellent" and "good" for this survey. One possible assumption is that people at this age mostly enjoy their retirement, there is less chance for them from suffering under financial crisis, emotional attacks, and pressure from work.

Figure 2: Distribution of Mental Health of People Having Different Income Ranges

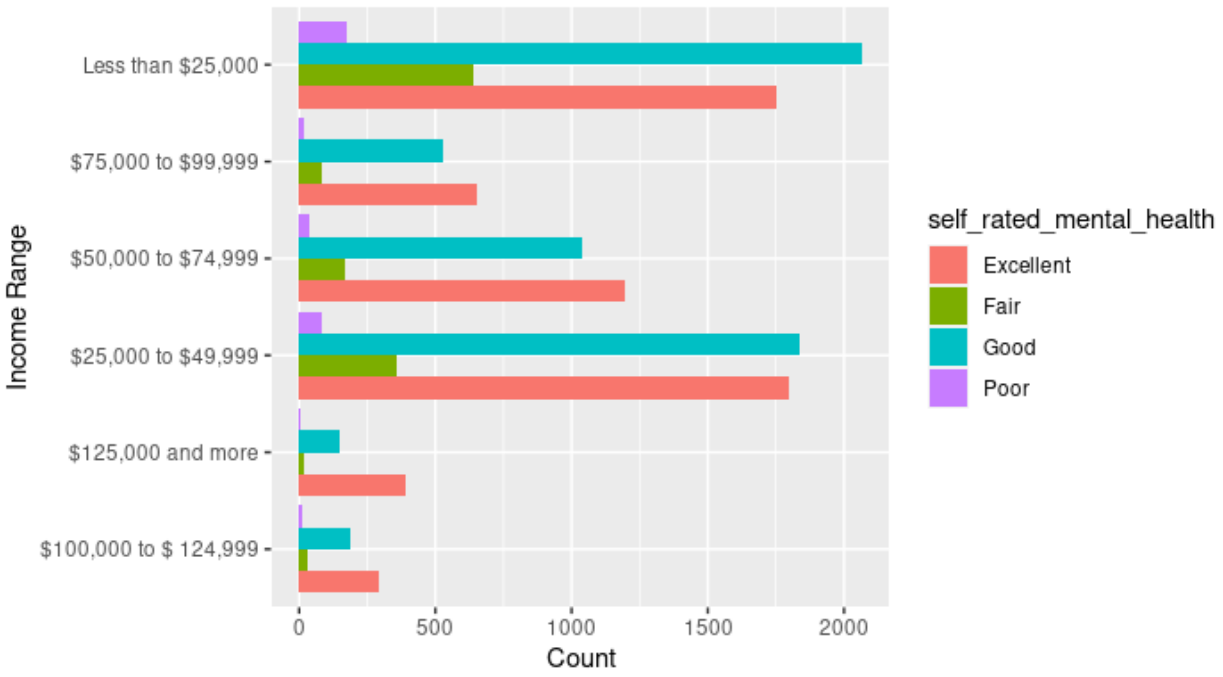

Figure 2 demonstrates a bar plot reflecting self-rated mental health interviewees based on their income range. As the x-axis and y-axis is the number of counts and income range respectively, the mental health condition is distinctive with each other using different colors. The annual income is strictly divided into 6 ranges from less than $25,000 up to 125,000 according to the GSS31 codebook.

The bar plot above shows that most positive self-rated counts of “Excellent” and “Good” are in the income range between $25,000 and less and income between $50,000 to $74,999.The income range between $25,000 to $49,999 also demonstrates a positive count exceeds 1000. However, the negative self-evaluation on mental health indicates a similar trend, most negative counts devoted to the range from $25,000 and less, followed up by the income range between $50,000 and $74,999.

It is noticeable that participants for this survey are majority from the income range of under $25,000 and the income range between 25,000 and 49,999. Those represent a vast majority of income conditions among regular Canadian as they have a positive attitude towards its mental condition. However, this model can not convey the idea that the higher income class will have a higher rate of having positive self-evaluated mental health since we are not calculating the percentage within each range.

6. Model

First and foremost, the response variable is reclassified into either good or bad mental health conditions, since it could ease people’s subjective impression of their mental state and deviation in the same classification. It is vague to distinguish the difference between “good” and “very good” for participants and there is no such tough standard for them to reference. Hence, a clear classification of two opposite options of mental state might help for further analysis of model prediction.

In order to seek a relationship between the potential factors selected and a binary response variable about mental health, a logistic regression model could fit the design of this study. Moreover, It is using a stratified sampling method based on different regions according to the process of distributing this survey. From the survey package, it provides softwares to build a logistic model by adjusting standard errors based on the sampling method using finite population correction. In this case, it is the population of each region. On the other hand, it is confirmed that the average distances from each observation to the regression line are relatively larger when using a general logistic model instead of regress from the survey package. Thus, it ensures a higher level of accuracy by using the svyglm() function.

As mentioned, a combination of predictors is selected including two numerical variables and four categorical variables from the two aspects including life environment and personal abilities. “Age” and “Sex” are dependent on the external world whereas people will be able to choose the “Region” they are staying in. However, “Living Arrangement “, “Income Respondent”(before tax), and “Number of Marriages” are relatively diverse parameters in real life, meaning that people’s decisions may depend on various occasions.

Accounts for all the predictors selected, the full model did not explain well for mental health, especially reflected in the sections on the living arrangement and the participant’s gender. The p-value from the summary of the full model indicates that the probability of the predictors has a meaningful coefficient by rejecting the null hypothesis i.e. there is no relationship between this specific predictor and the response variable. It displays that gender is extremely not significant as well as some types of living arrangements according to the significant level of 0.05. This might be because social influences on humans do not depend on gender as gender equality is widely introduced especially in a deeply culturally diverse country like Canada. The living arrangement might also depend on the participant's economic capacity and marital status which could already be explained from the other two variables.

Finally, by getting rid of the ineffectual predictors, the final model for predicting the state of Canadian mental health is as the following, noting that all selected independent variables are significant:

log(p/1-p) = $\widehat{\beta_0}$ + $\widehat{\beta_1}$age + $\widehat{\beta_2}$regionBritish Columbia + $\widehat{\beta_3}$regionOntario + $\widehat{\beta_4}$regionPrairie region + $\widehat{\beta_5}$regionQuebec + $\widehat{\beta_6}$income_respondent$125,000 and more + $\widehat{\beta_7}$income_respondent25,000 to 49,999 + $\widehat{\beta_8}$income_respondent50,000 to 74,999 + $\widehat{\beta_9}$income_respondent75,000 to 99,999 + $\widehat{\beta_{10}}$income_respondentLess than 25,000 + $\widehat{\beta_{11}}$number_marriages

Overall the model performance is very good; we can predict the probability of the respondent's state on mental health by using all of the variables in our model.

7. Results

Table 2: Summary Table for Final Model on predicting state of Mental Health

```
# A tibble: 12 x 5
   term                                estimate std.error statistic  p.value
   <chr>                                  <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                         -0.162    0.0640       -2.53 1.15e- 2
 2 age                                  0.00393  0.000665      5.91 3.54e- 9
 3 regionBritish Columbia              -0.109    0.0368       -2.96 3.08e- 3
 4 regionOntario                       -0.260    0.0299       -8.71 3.34e-18
 5 regionPrairie region                -0.0944   0.0325       -2.91 3.63e- 3
 6 regionQuebec                        -0.609    0.0326      -18.7  3.90e-77
 7 income_respondent$125,000 and more  -0.555    0.0748       -7.41 1.31e-13
 8 income_respondent$25,000 to $49,999  0.494    0.0554        8.91 5.93e-19
 9 income_respondent$50,000 to $74,999  0.296    0.0573        5.17 2.35e- 7
10 income_respondent$75,000 to $99,999  0.227    0.0616        3.68 2.32e- 4
11 income_respondentLess than $25,000   0.756    0.0552       13.7  2.16e-42
12 number_marriages                    -0.102    0.0190       -5.39 7.08e- 8
```

Taking a closer look at the final model, interpretations could be expressed as the following:

The intercept of -0.1616772 means when age is "15", region is "regionAtlantic region", income is "income_respondent100,000 to 124,999", number of marriages is "0", log odds of "self rated mental health" is good equals to -0.1616772. When the respondent is from region British Columbia for instance, then regionBritish Columbia = 1, otherwise regionBritish Columbia = 0, and it carries a coefficient of about -0.108966 indicating that holding all other explanatory variables in the model fixed, when the region changes from Atlantic region to British Columbia, the log odds of self rated mental health is good will decrease by 0.108966. For income levels, when the respondent's income changes from

"100,000 to 124,999" to "125,000 and more", holding all other explanatory variables in the model fixed, log odds of self rated mental health is good will decrease by 0.554773.

Also, for simply numerical variables like "Age", it shows that holding all other explanatory variables in the model fixed, for every 1 increase in age(above 15 years old), log odds of self rated mental health is good will decrease by 0.003929.

8. Discussion

Our model contributes to providing a reference to all people that are concerned about Canadians mental health issues. We primarily analyze mental health issues regarding aspects that include the distribution of age, income level, Canadian regional distribution, and marriage situation. For example, based on statistical results, it can be noticed that compared to young people, the elderly have higher happiness indexes and have healthier mental health conditions. While considering income, people with lower incomes and upper-middle incomes are more likely to have an inactive mental health status. Therefore, we can infer that life pressure originated by age, and income level is very likely to cause mental health problems. Another considered factor includes the number of marriages. From the bar plot, we can observe that people married one time or 0 times have a healthier mental state. This corresponds to our model that the number of marriages and mental health status has a relatively strong negative relationship. Lastly, Ontario has the largest population compared to other provinces in Canada, this might be the reason that it has the greatest proportion of people with good mental health.

It is worth noting that the data only contain the population in Canada, which is not necessarily applicable to other countries. Each country has its own government regularization on the residents and social habits which may alter an individual's mental status from those factors. Furthermore, everyone has different criteria for evaluating themselves, and stating overall mental health at a moment may depend on other temporary changes in life routines. Thus, "self-rated mental health" is subjectively decided by the respondent's personal criteria.

9. Weaknesses

Considering our survey on mental health, there are some deficiencies present in it. One of the most apparent issues is data selection. For example, the definition of "good" or "Poor" was very subjective. Moreover, the participants' responses may not be accurate enough since medical certificates were not provided. Therefore, it's not possible to determine whether or not they would have mental issues.

Moreover, we did not exclude Canadians with innate mental problems, those number of individuals might be inherited with bad mental health from family members. Thus, the social-related factors would not be the trigger for one's negative mental health result.

One weakness accorded to our visualized data is that it can only display the actual condition based on the data and It would not help to conclude a specific relation within factors. For example, our histogram indicates an age range between 60 to 70 have the most positive self-rated counts, it may not possibly lead to a result of age from 60 to 70 has the most positive mental health. One possible assumption is that the number of individuals who answered this question is the majority in this age range.

On the other hand, some improvements could be made in our survey. For example, to investigate factors that impact mental health, we could consider studying family factors and making the research question more specific. We can also focus our investigation on participants who have rated themselves in a poor condition to provide them with more significant help based on our findings.

10. Next Steps

In order to improve this investigation on the mental health of Canadian people, it may require more balanced data. To clarify, some regions from this dataset have relatively large differences in population surveys; if the number of people surveyed between regions can be reduced, it may be helpful for further analysis. Secondly, narrowing the scope of the survey and investigating more data in one region to analyze it for a specific area, may also result in more accurate conclusions. Then, specific treatment could be delivered to a specific region, making the whole investigation more meaningful and practicable.

Also, models could be built using multilinear logistic regression from the raw data (Using the original four or six categories of self-rated mental health) or Bayesian inference to seek for more accurate results. Finally, some other variables could be taken into considerations such as daily sleep hours, wake-up/sleep time, etc., which may enrich the entire dataset and help in predicting the main goal of this study on mental health.

11. References

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

2. T. Lumley (2020) “survey: analysis of complex survey samples”. R package version 4.0.

3. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

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

5. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. https://broom.tidymodels.org/, http://github.com/tidymodels/broom

6. General social survey on Family (cycle 31), 2017， https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.html

7. General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide, https://sda-artsci-utoronto-

ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

8. 2017 General Social Survey: Families Cycle 31 Public Use Microdata File PUMF https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_Codebook.pdf

9. Code for access the variables present in the data framework: https://www.geeksforgeeks.org/accessing-variables-of-a-data-frame-in-r-programming-attach-and-detach-function/

10. Chambers, J. M. and Hastie, T. J. (1992) Statistical Models in S. Wadsworth & Brooks/Cole. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/factor

11. Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/levels

12. %in% operator in R. (2020, September 15). Retrieved October 20, 2020, from https://www.datasciencemadesimple.com/in-operator-in-r/

13. Two Way Tables. (n.d.). Retrieved October 20, 2020, from https://www.cyclismo.org/tutorial/R/tables.html