Recent working status and self-rated feelings of life positively influence mental health based on 2017 GSS

Jingyi Chen, Tianxiao Ma, Boyu Sheng, Qiyue Zhang

October 19, 2020

Abstract

In our research, we investigated the 2017 General Social Survey(GSS)[7] on the family to deeply study the relationship between mental health and other two factors: self-rated feelings of life and recent working status. Because those two factors respectively represent both subjective and objective reasons that may affect people's mental state. Our research discovered that both of those two factors positively correlated to their mental status. The better people feel about their lives, the better people rate their own mental status. And for people who have worked in the past week, the less likely they would have a low mental status quote as well. Such findings suggest that more entertainment and improving the working conditions for people may lower the number and proportion of people who suffer from mental problems.

Introduction

Mental illness has always been a critical issue that bothers Canadians for a long period of time. In any given year, 1 in every 5 Canadians suffers from different mental illness or other mental problems. (Canadian Mental Health Association, 2016)[2] Mental health, in a broad sense, can be interpreted differently by different subjects such as psychology, biology, sociology and etc. Inside the GSS, Canadian government focused the issues mainly on social behaviour. Instead of studying the genetic and biological effects of mental problems, social interpretation highlights the external factors of adults in their daily life such as depression, anxiety, etc. Insomnia, distraction and etc are all the main minor outcomes as a result of mental illness.More seriously, Homicide is one of the worst outcomes of mental illness, in 2017 police reported 680 homicide cases across Canada, 48 more than in 2016. (Allen, 2018)[1] So it is essential for a responsible government to figure out a way to improve citizens' mental health in order to reduce social turbulence.

In the paper, we analyse the data from the GSS in the year of 2017. GSS is designed and introduced by the government of Canada in order to fully understand every aspect of daily life of people who are over 15 years old. After some filtering, we chose two main factors that would possibly affect people's mental health the most to discuss: 1. Feelings of one's life. As this factor is a comprehensive assessment of one's depression, anxiety, happiness or just normality. 2. Recent working status. The reason for choosing this factor is because that working can provide people with basic needs and entertainment and it's strongly correlated with one's mood. From those,we can find out the correlations between each of the factors and give suggestions to the government on how to improve people's mental state and help the Canadian society be more stable.

After analyzing the data, we have found out that both factors have positive correlations with one's mental state by the figures. Feelings of life have a stronger effect on people's mental status than their working status. The standard errors of both factors are so low that provide the result a stable and reliable evidence. However, there is also bias in this survey, such as young people who do not need to worry about working and senior people who live on pensions may care less about their current working status and will affect the result and the survey only investigated the people who had access to telephone. Thus in the next section, we will analyze the data in detail and fully discuss the problem and weaknesses that we encountered in this research.

Data

Data originally adapted from the 2017 General Social Survey (GSS) on the Family.[7]

2017 GSS on Families is objective to gather data on social trends in order to monitor changes in the living conditions and well-being of Canadians over time; and, to provide information on specific current social policy issues or emerging interest.

The target population for 2017 GSS 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. The target sample size (i.e. the desired number of respondents) for the 2017 GSS was 20,000 while the actual number of respondents was 20,602.

Data for 2017 GSS on Families was collected from February 2 to November 30, 2017 via computer assisted telephone interviews (CATI). Calls being made from approximately 9:00 a.m. to 9:30 p.m. Mondays to Fridays. Interviewing was also scheduled from 10:00 a.m. to 5:00 p.m. on Saturdays and 1:00 p.m. to 9:00 p.m. on Sundays. Respondents were interviewed in the official language of their choice. Proxy interviews were not permitted.

Survey frame for 2017 GSS on Families uses the redesigned GSS frame, which integrates data from sources of telephone numbers (landline and cellular) available to Statistics Canada and the Address Register (AR). The sampling unit is defined as groups of telephone numbers.

In 2017 GSS, each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records was next performed in each stratum.

The overall response rate for the 2017 GSS was 52.4%. Total non-response was handled by adjusting the weight of households who responded to the survey to compensate for those who did not respond.

Although interviews were instructed to obtain a completed interview with the randomly selected member of households as the interviewers even re-contacted up to two more times with households who refused to participate at first, and explain the importance of the survey to encourage their participation. However, non-sampling errors still occurred as a weakness when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information.

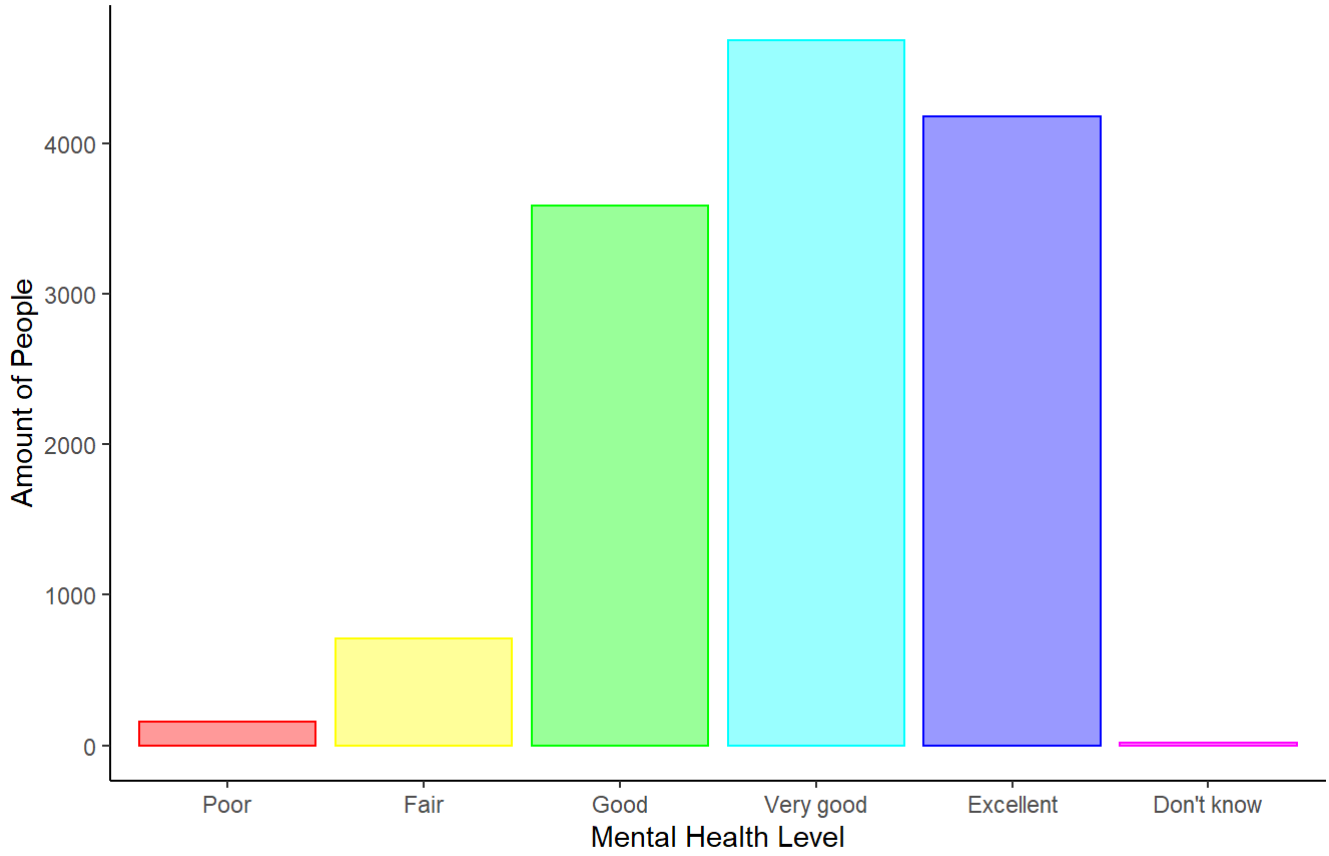
The dataset used in the paper was adapted from gss.csv created by Rohan Alexander and Sam Caetano Since we are interested in the relationship between respondents' mental health levels with their recent working status and self-rated feeling of life, we filtered data from the 2017 GSS data frame after cleaning, and formed a new data frame. The new data frame contains three variables, `feelings_life`, `self_rated_mental_health`, `worked_last_week`, with NAs removed.

Table1: Variables in the Dataset used for the Model

Variables	**Description**
feelings_life	Feelings about life as a whole
self_rated_mental_health	Self rated mental health
worked_last_week	Worked as a job or business last week

Table 1 above displays the variables in the dataset used for the model. We choose the categorical variable `worked_last_week` instead of `average_hours_worked` as we want to focus on the factor working status, whether the respondent has a job to work recently, instead of how long they had worked.

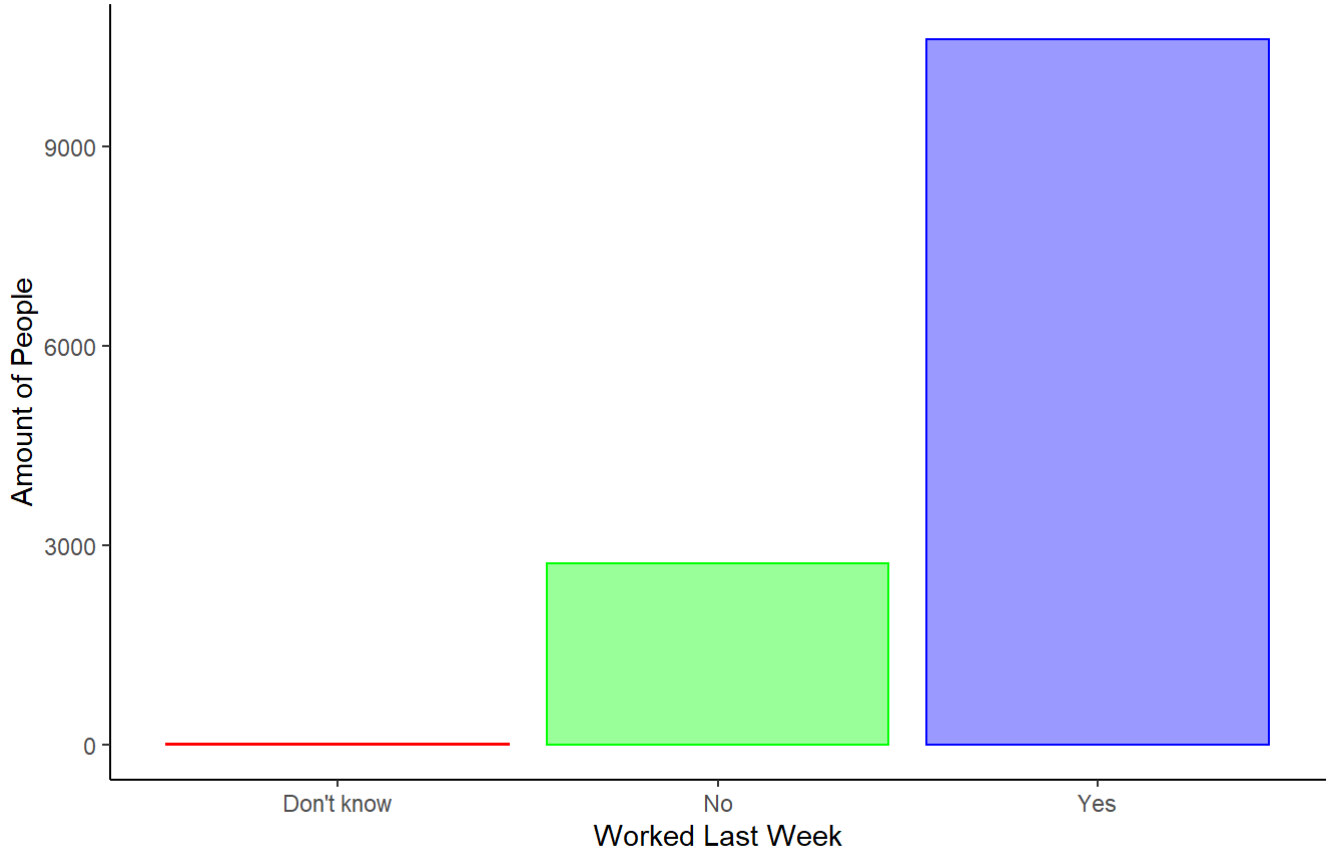
Figure 1: Barplot for Mental Health



Source: Adaption from 2017 GSS

The barplot in Figure 1 shows the distribution of mental health levels rated by the respondents themselves. The majority of the respondents feel positive about their psychological well-being. However, only precious few people are having issues with their mental health.

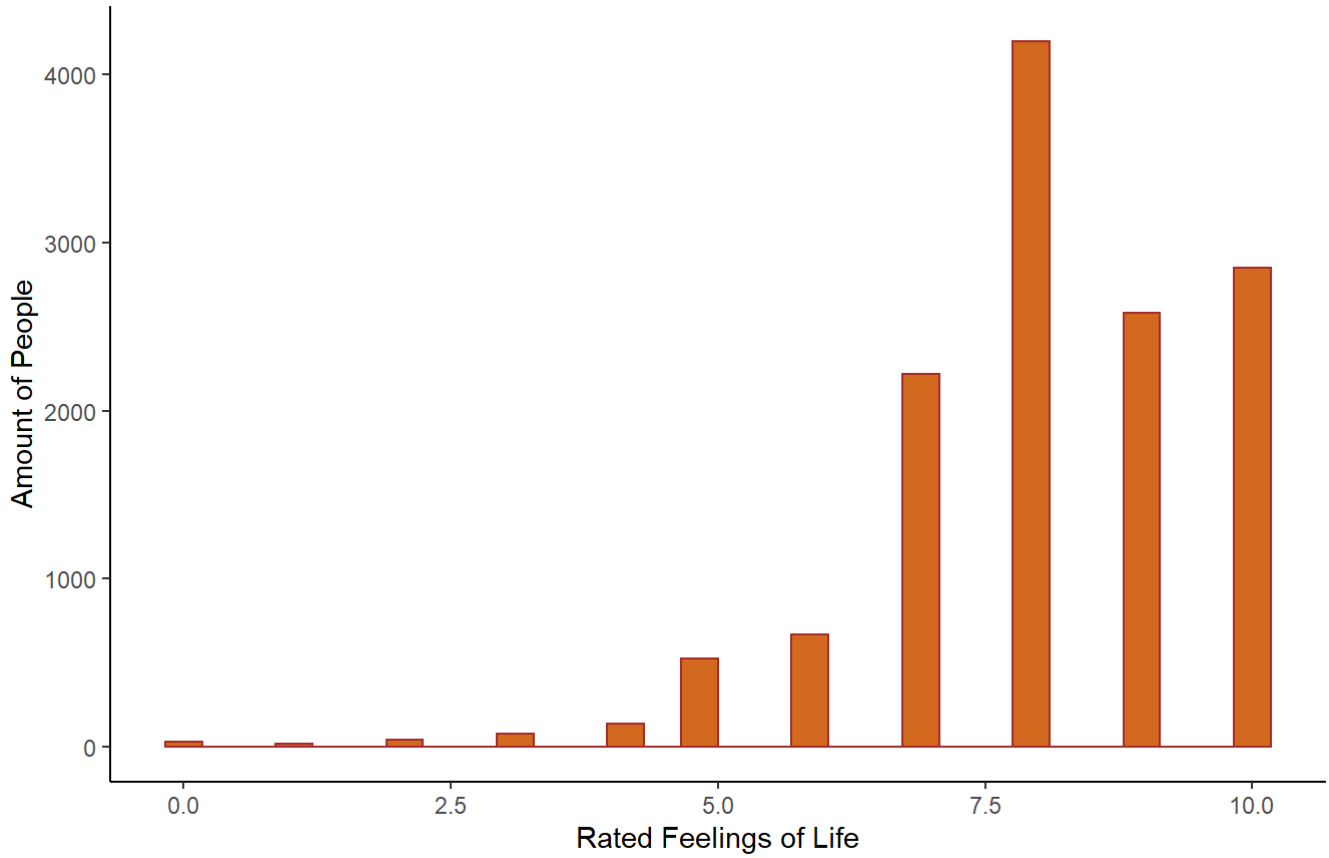
Figure 2: Barplot for Work



Source: Adaption from 2017 GSS

The barplot from Figure 2 gives the distribution of the work status of the respondents, which provides the number of people who are working is more than tripled the number of people who are not.

Figure 3: Histogram for Feelings of Life



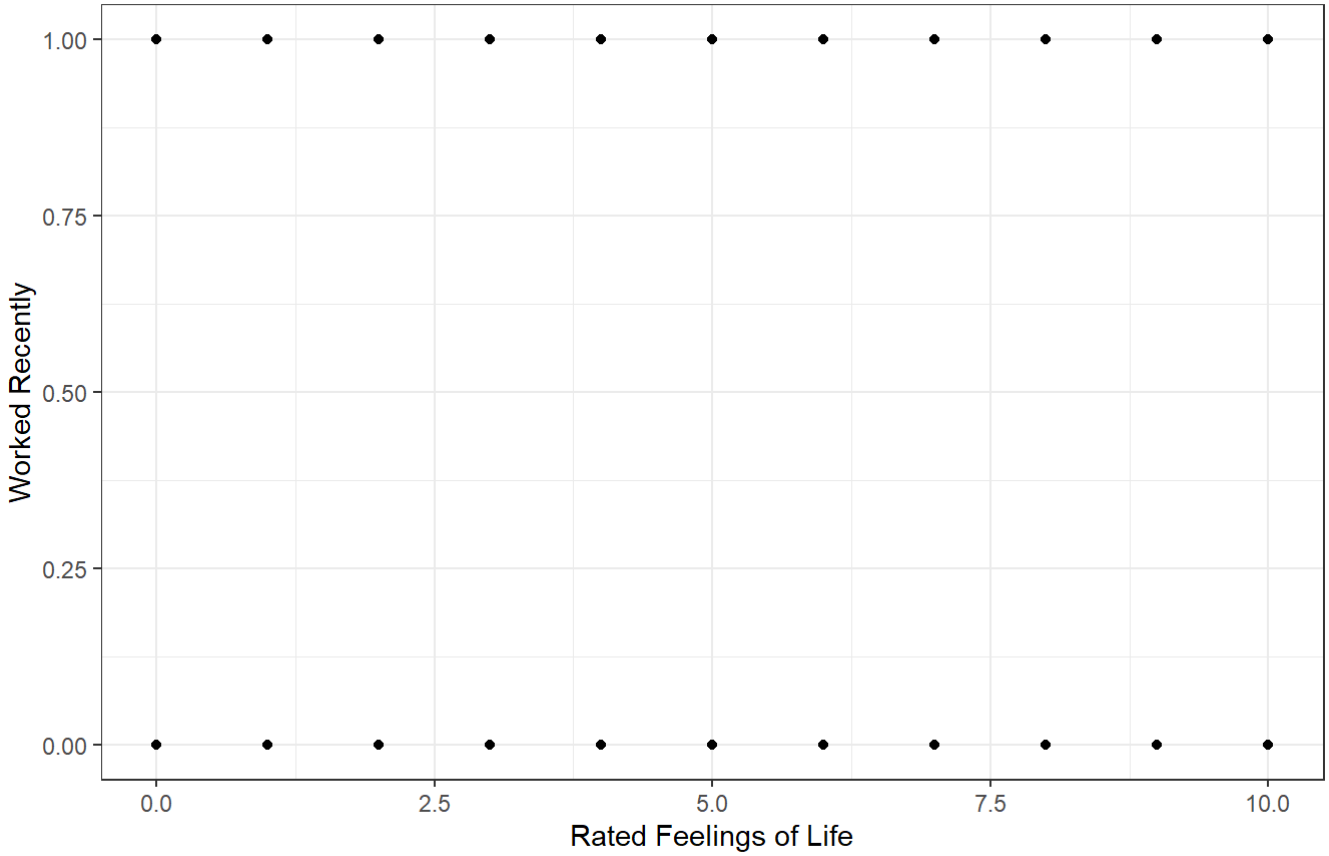
Source: Adaption from 2017 GSS

The left skewed histogram given in Figure 3 contributes to the distribution of the feelings of life rated by the respondents themselves, which shows people tend to feel very satisfied with their lives on average.

To design the model, we create two dummy variables `worked_recently` and `mental_levels` for `worked_last_week` and `self_rated_mental_health` respectively. For `worked_recently`, 0 represents not worked last week and 1 represents worked last week. For `mental_levels`, mental health levels are rated from

1 to 5, which indicates from poor to excellent respectively.

Figure 4: Scatterplot of Rated Feelings of Life and Worked Recently



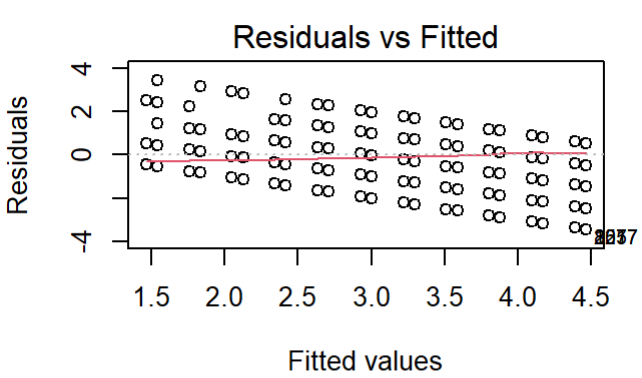
Source: Adaption from 2017 GSS

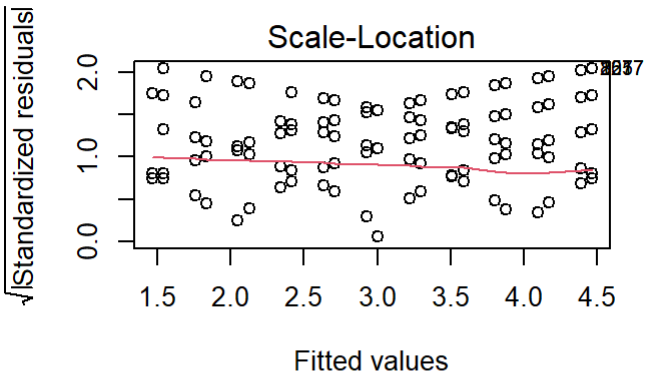
The scatterplot provided in Figure 4 suggests there is no relationship between the variable feelings of life and work status, which means we can use them as two factors to explain the mental health level of the respondents.

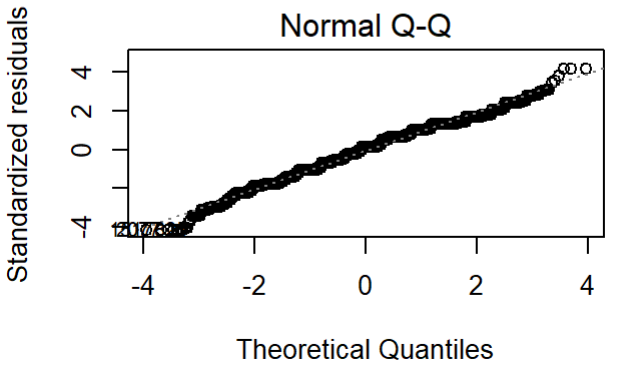
Model

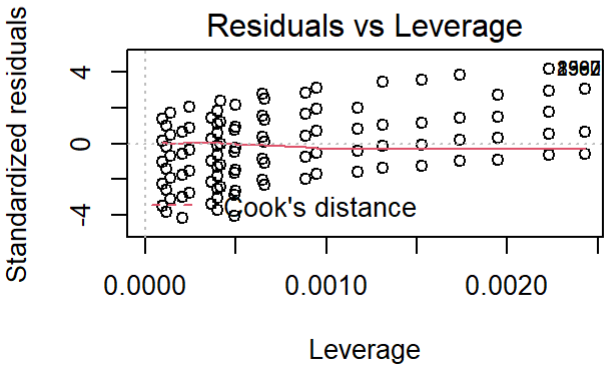
#Result

```
##
## Call:
## lm(formula = mental_levels ~ worked_recently + feelings_life,
##     data = clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4652 -0.5880  0.1196  0.5348  3.4590
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.464563   0.041014  35.709   <2e-16 ***
## worked_recently 0.076428   0.017913   4.267    2e-05 ***
## feelings_life   0.292423   0.004716  62.008   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8323 on 13310 degrees of freedom
## Multiple R-squared:  0.2261, Adjusted R-squared:  0.226
## F-statistic:  1945 on 2 and 13310 DF,  p-value: < 2.2e-16
```









```
##
## Call:
## glm(formula = mental_levels ~ worked_recently * feelings_life,
##     data = clean)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.723e-13  -1.712e-13  -1.712e-13  -1.710e-13  -1.694e-13
##
## Coefficients:
##                                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)                   1.000e+00  1.565e-14  6.39e+13   <2e-16 ***
## worked_recently               2.920e-15  1.821e-14  1.60e-01    0.873
## feelings_life                 2.204e-25  1.910e-15  0.00e+00    1.000
## worked_recently:feelings_life -1.318e-16  2.216e-15 -5.90e-02    0.953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.920186e-26)
##
##     Null deviance: 0.0000e+00  on 13312  degrees of freedom
## Residual deviance: 3.8865e-22  on 13309  degrees of freedom
## AIC: -744959
##
## Number of Fisher Scoring iterations: 1
```

Discussions Section

How features enter the model

In 2017 GSS, questionnaires like ‘feelings about life as a whole’, ‘Self rated mental health’, ‘Worked as a job or business last week’ occur to determine living conditions and well-being of Canadians over time. In the paper, we aim to analyze how mental health levels are influenced by feelings about life as a whole(feelings_life) and worked last week. People’s mental health, and feelings about life are subjective and can be affected significantly by things that have happened recently. Using data such as the annual income may not be too accurate to describe the respondent’s mental health at the time of answering the survey. As a result, we decide to use worked_last_week as one predictor variable since it happened recently. Holding the assumption the person’s feeling is stable during the time answering, then the answer of ‘feelings about life as a whole’ and ‘Self rated mental health’ should have a small variance. Thus feelings about life as a whole is picked as the second predictor variable. worked_last_week is a categorical variable, which is not ideal when adding to our model as a predictor, so we recode it into a dummy variable. The dummy variable has value 1 when the respondent answered ‘Yes’ for the question ‘worked as a job or business last week’, and we assign value 0 and set it as a reference group when the respondent answered ‘No’. Self rated mental health is a categorical variable, we recode it into a numerical variable and drop the ‘Don’t know’ answer to better fit a linear regression model. The original answers in the survey are categorized by “Poor” , “Fair”, “Good”, “Very good”, “Excellent” , corresponding to 1,2,3,4,5 after recoding accordingly. The reason for dropping ‘Don’t know’ answers is it’s complicated to predict the respondent’s actual mental health level. Maybe we can include the respondent may fall into the ‘poor’ group, but response error may influence the overall accuracy of the model.

Alternative Model:

Since Self rated mental health is a categorical variable, we could transform it into a dummy variable, such as assign 0 to the dummy variable if the mental health is ‘Poor’ or ‘Fair’ and used as a reference group, assign 1 otherwise. In this way, we could use a logistic regression model to analyze how mental health levels are influenced by feelings about life as a whole(feelings_life) and worked last week. However, the disadvantages of using a logistic model in this case are weighted more than its advantages,a multiple linear regression model fits better in this case. Advantages: logistic regression fits better when removing attributes that are unrelated to the response variable as well as attributes are correlated to each other. Additionally, logistic regression can be implemented easily. Disadvantages: Logistic regression gives a discrete result but linear regression gives a continuous result. In this case, defining mental health as a dummy variable is not a good fit, it is too vague and inaccurate. We don’t want to assign 1 to represent people who rank their mental health as “good”, ‘very good’ and ‘excellent’, and 0 otherwise. The result will be like ‘people worked last week are mentally stable since they fall more in value 1’. A detailed response variable using ranks is always better to indicate how good/bad mental health status is.

Multiple Linear Regression Model:

We are interested in the relationship of respondents’ mental health levels associated with their last week’s working status and self-rated feeling of life in 2017 GSS. Since we have two explanatory variables(last week’s working status and self-rated feeling of life) to predict the outcome of a response variable(mental health level). The statistical technique, Multiple Linear relationship with statistical software R studio helped us to run the model. Based on the survey content.

Model Assumptions:

Homoscedasticity, that is the error term has a constant size across independent variables Independence of observations: R square for independent variables(mental_levels and worked_recently) is 0.001519443, the extremely small value allows us to conclude the absence of multicollinearity between two independent variables. shows that there are no hidden relationships among independent variables. Normality: The data follows a normal distribution. Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

Diagnostic Plots analyze

From the diagnostic plots (figure 5), the upper left plot shows Residuals vs Fitted. Used to check the Linearity. We can observe the smooth horizontal line without a distinct pattern. This suggests that we can assume a linear relationship between the predictors and the outcome variables. The lower left plot is a normal Q-Q plot used to examine whether the residuals are normally distributed. Observed residuals points follow the straight dashed line, so our model satisfies normality assumption. The upper right Scale-Location plot is used to check the homogeneity of variance of the residuals (homoscedasticity).We can observe a horizontal line with equally spread points, so it is a good indication of homoscedasticity. The lower right Residuals vs Leverage plot is used to identify influential cases, that is outliers and high leverage points. We can observe the plot highlights the top 3 most extreme points with standardized residuals around 4.

The Regression Equation

$$y = \beta_0 + \beta_1 x_1 + \beta 2 x_2 + \epsilon$$

where b0, b1, and b2 are regression coefficients. X1 and X2 are regression coefficients defined as: X1 is a dummy variable, if the respondent worked last week then X1 = 1; X1 = 0, otherwise.Note that respondents who didn’t work last week are the reference group, in the case X1= 0 X2 represents the respondent’s self-rated feeling of life score collected in the 2017 GSS survey. Y represents respondent’s mental health level collected

in the 2017 GSS survey, where “Poor” , “Fair”, “Good”, “Very good”, “Excellent” corresponds to 1,2,3,4,5 accordingly. The independent error term, epsilon represents the difference between the expected respondent’s mental health level at a particular time and mental health level that was actually observed. The error term follows a normal distribution with mean 0 and constant variance σ^2

Interpreting Regression Coefficients

$$\hat{y} = 1.464563 + 0.076428x_1 + 0.292423x_2$$

Where ŷ is the predicted value of mental health rank, x1 is an indicator variable to define if the respondent worked last week, X2 represents the respondent’s self-rated feeling of life score collected in the 2017 GSS survey. and b0, b1, and b2 are regression coefficients in the estimated model with values 1.464563, 0.076428 and 0.292423 accordingly.

Null Hypothesis(H0) : β1 = β2 = 0 Alternative hypothesis(H1): at least one βi ≠ 0, i = 1, …, k

$$\hat{beta}_0 = 1.464563$$

shows that the predicted mental health level is 1.464563 when the respondent didn’t work last week and has 0 self-rated feeling of life score.

$$\hat{beta}_1 = 0.076428$$

in this model, after effects of self-rated feeling of life score are taken into account,that mental health rank is 0.076428 higher when the respondents who worked last week as opposed to respondents who didn’t work last week.

$$\hat{beta}_2 = 0.292423$$

In this model, for every unit increase in self-rated feeling of life score, the predicted mental health levels rank increased by 0.292423 when the respondent didn’t work last week. b1 has p-value 2e-05 while b2 has p-value 2e-16,both suggest that beta1 and beta 2 are needed in a model with all the other predictors included. We can reject our null hypothesis, two predictor variables in our model are significant predictors of a respondent’s mental health level.

Why is this Model Important?

The model has R square value 0.226, it gives around 22.6% of variation in mental levels explained by the regression model with predictors, feelings of life score and last week’s working status.It’s reasonable and accepted in this case because we only have two independent variables, R square will be larger by adding more independent variables. In addition, self-rated mental health is subjective and can be affected by many other factors. The standard errors for the slopes of 0.018 (last week’s working status) and 0.004 (feelings_life) are both small compared with the slopes themselves, so the coefficient estimates are fairly precise.

Weakness and next step

1. Since the sample was obtained based on a combination of landline and cellular telephone numbers. People who are 15 years of age or older without access to telephone are omitted. More methods should be introduced in order to attract people from different social groups and the target population should be more inclusive. For young people, more surveys can be given out in person inside a mall, school or other entertainment facilities. For middle-aged people, email is always a great option as most people who work in the office are used to communicating with the internet. For seniors, it’s more effective to use

mail as it's more traditional. Giving out surveys inside the supermarket and casinos with incentives is a great way to attract senior respondents too. This method can make sure the final result is more inclusive.

2. The mental health data is self-quoted which is relatively subjective and different people all have different standards of their own mental situations so that the data was relatively inaccurate. Some people with serious mental problems may not notice or some people underestimated their mental states because of temporary frustrated emotions. The next step should focus on reducing the data which are easily affected by emotions. Data should be more objective instead of subjective. In this particular case, the government should work with the public hospital and distribute free mental diagnoses and give people a complete quantitative report based on their mental states so that it can be more precise to evaluate the mental status of the respondents.

3. The option " feelings of life" needs to be more in detail and has to have longevity. Detail means people are recommended to name some of the aspects that they focus more in their life. Some people care about money and others may care more about fame so that a detailed data can perform a more accurate correlation with mental health. So more options should be added regarding which particular aspect do the respondents focus more on the feeling of your life. Longevity means the "feeling of life" of the respondents has to be effective for a relatively long period of time instead of the one moment when they are taking the survey. Because people may be affected by contingencies, for example: winning lottery tickets(small amounts) before the survey or feeling down because of the death of relatives or pets. Those feelings are temporary but may change the answer of the survey. So the survey can be more specific and set a time period of their feeling of life. Such as: " How do you feel about your life in the past 10/5 years and in the future?" This method can lower the effect and bias of temporary sentiment to the answer.

Reference

1. Allen, M. 2018 "Police-reported crime statistics in Canada, 2017." Juristat. Statistics Canada Catalogue no. 85-002-X.

2. Canadian Mental Health Association. (2016). Mental Health Statistics, from https://ontario.cmha.ca/wp-content/uploads/2016/10/CMHA-Mental-health-factsheet.pdf (https://ontario.cmha.ca/wp-content/uploads/2016/10/CMHA-Mental-health-factsheet.pdf)

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

4. Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr (https://CRAN.R-project.org/package=readr)

5. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (https://www.R-project.org/).

6. Rohan Alexander and Sam Caetano(2020), gss.csv

7. Statistics Canada. (2017). General social survey (GSS), 2017: Cycle 31, Oct 19, 2020. [Public use microdata file and codebook]. Ottawa, ON: Statistics Canada. Retrieved from http://www.odesi.ca (http://www.odesi.ca)

8. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Appendix

Link to dataset and code used for the paper: https://github.com/rubytianxiaoma/STA304-PS3 (https://github.com/rubytianxiaoma/STA304-PS3)

please refer to Mental health.Rmd for the code and target_nona.csv for dataset used