Time spent with family, self-rated health, workload management seriously affects the stress level of young Canadians

Yifan Xu

2020/12/9

Abstract

There are multiple factors that affect the stress level of young Canadians. In this paper, we explored the Canadian General Social Survey (GSS), Cycle 30, 2016: Canadians at work and at home, and fitted a regression model based on factors related to the stress level with R. We found that the self-rated health, time spent with family and workload management seriously affect Canadians' mental health self assessment. This can not only roughly predict the young adults' stress level, but also provide people with a way to relieve stress, especially during the COVID-19 pandemic.

Keywords: Canadian well being; stress level; Sample survey without replacement ; Self-rated health ; logistic Regression analysis with R.

Introduction

Stress has always been a common problem among modern people, and different stress occurs at all stages of people's life. According to a survey of Canadians by Ipsos, almost half of Canadians said they feel severely stressed at least once a week, and more than 15% of them said they feel pretty stressed nearly every day. Among these people, compared with the information five years ago, it suggests that young adults undergo greater pressure in life than middle-aged and elderly people. (Ipsos (2020)) Their pressure may come from family, work and job, close relationships, or related to themselves.The present condition of young people's stress is worthy of attention. Especially in the current COVID-19 epidemic, new diseases have caused people to panic to a certain extent. At the same time, many measures have been taken to prevent the spread of the virus, such as keeping a safe distance during isolation and closing public entertainment venues, making them isolated and lonely, thereby increasing stress and anxiety. (Disease Control and Prevention (2020)) Therefore, we intend to find the main causes of stress among young adults, the approach to how to relieve their stress, and how to predict their stress level.

We utilized the dataset about Canadian General social survey(GSS), Cycle 30, 2016: Canadians at work and home, which was conducted from August 2 to December 23, 2016. It was a sample survey with a cross-sectional design and covered 10 provinces of Canada. Non-institutionalized personnel aged 15 and above were the target population. This survey provided information on social trends about working conditions, family status, time utilization, and how these factors affect the well beings of Canadians. (GSS (2016)) This report mainly examined which of these factors can strongly affect young adults' stress level and how to predict it. The main method we used is building a generalized linear model GLM (Hadley Wickham (2020) Wickham et al. (2019)) in R (R Core Team (2020)) to measure the relationship between the young adults' stress level and exploratory variables related to inducing or reduce the stress. We find that the time spent with family, self-rated health, workload management seriously affects the stress level of young Canadians.

Below is the data section, we will introduce the data set and focus on several variables we choose, then conduct the logistic regression analysis with R (R Core Team (2020)) and obtain the logistic regression model, which followed by multiple figures and tables in the result section. Finally in the discussion section, we will discuss our findings based on the model results as well as some weaknesses and opportunities for future work.

Data

In this paper, we utilized the dataset about Canadian General social survey(GSS), Cycle 30, 2016: Canadians at work and home. The data, user guide and code book can be found at:https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss30/gss30/more_doc/index.htm (It can be accessed by faculty, students, and staff member at University of Toronto only.) We also used the haven (Wickham and Miller (2020)) and janitor packages (Firke (2020)) in the data cleaning.

The 2016 GSS was conducted from August 2 to December 23, 2016, and it is a sample survey using a cross-sectional design. Its target population contains all non-institutionalized people over the age of 15 who live in 10 Canadian provinces, except residents from Yukon, Northwest Territories, and Nunavut and full-time institution residents. The frame of this survey was conducted with 2 different components: 1) lists of telephone numbers in use from Statistics Canada and 2) list of all houses in ten provinces from the Address Register (AR). The two main methods of this survey are stratified sampling and random sample without replacement (SSWR). First of all, every record from the survey frame was assigned to one stratum in its province. Next, the SSWR of these records was selected in each stratum. For each family, one person was chosen to complete the online survey or a telephone interview.(GSS (2016))

There were non-response cases in this survey, mainly because the respondents could not be connected, could not provide information, or refused to participate in the survey. In 2016, the response rate of GSS was approximately 50.8%. (GSS (2016)) In the process of processing records, there are mainly three ways to deal with a small number of incorrect records or records lacking information, which are completing them, make corrections, or calculated based on other information in the questionnaire. To be specific, according to whether there is some auxiliary information used to model response tendency for each household, there are three types of unresponsive telephone numbers: phone numbers without auxiliary information, phone numbers with some available auxiliary information, and phone numbers with auxiliary information available from Statistics Canada. The approach to deal with non-response is to compensate the non-responders by adjusting the weight of the households responding to the survey. For the first type of non-response, those unresponsive phone numbers were ignored and dropped. Such adjustment was made independently in each stratum. Then for the last 2 types, non-responses adjustments were made independently in their regions. (GSS (2016))

Since SSWR was chosen in this survey, it is inevitable that the estimates of the sampling survey would be affected by sampling errors, which is likely to cause certain biases in the final results. In order to reduce such errors as much as possible, the researchers used sample data to estimate a statistical measure of the standard error and the sampling error. (GSS (2016))

Among the hundreds of variables in this data set, we noticed the variable " SMG_Q01", which is used to measure the stress level of Canadians. Therefore, it is treated as the response variable in our research. Moreover, since we are concerned about the stress of young adults, we choose people 15 to 44 years old as the observations. And we choose the exploratory variables mostly from three aspects: life, work and personal situation, they are:

- TTLINCG2 Income - Personal income group (before tax)
- DOS_Q05 Level of satisfaction - Personal appearance
- SRH_110 Self rated health in general
- FAM_Q03 Level of satisfaction - Amount of time spent as a family
- WIR_Q01 Management of workload
- STJ_Q05 Match between current job and field of education or training
- ODA_Q01 Did not participate in outdoor activities - Past 12 months

Figures will show below: (we used the ggplot2 package (Wickham (2016)))

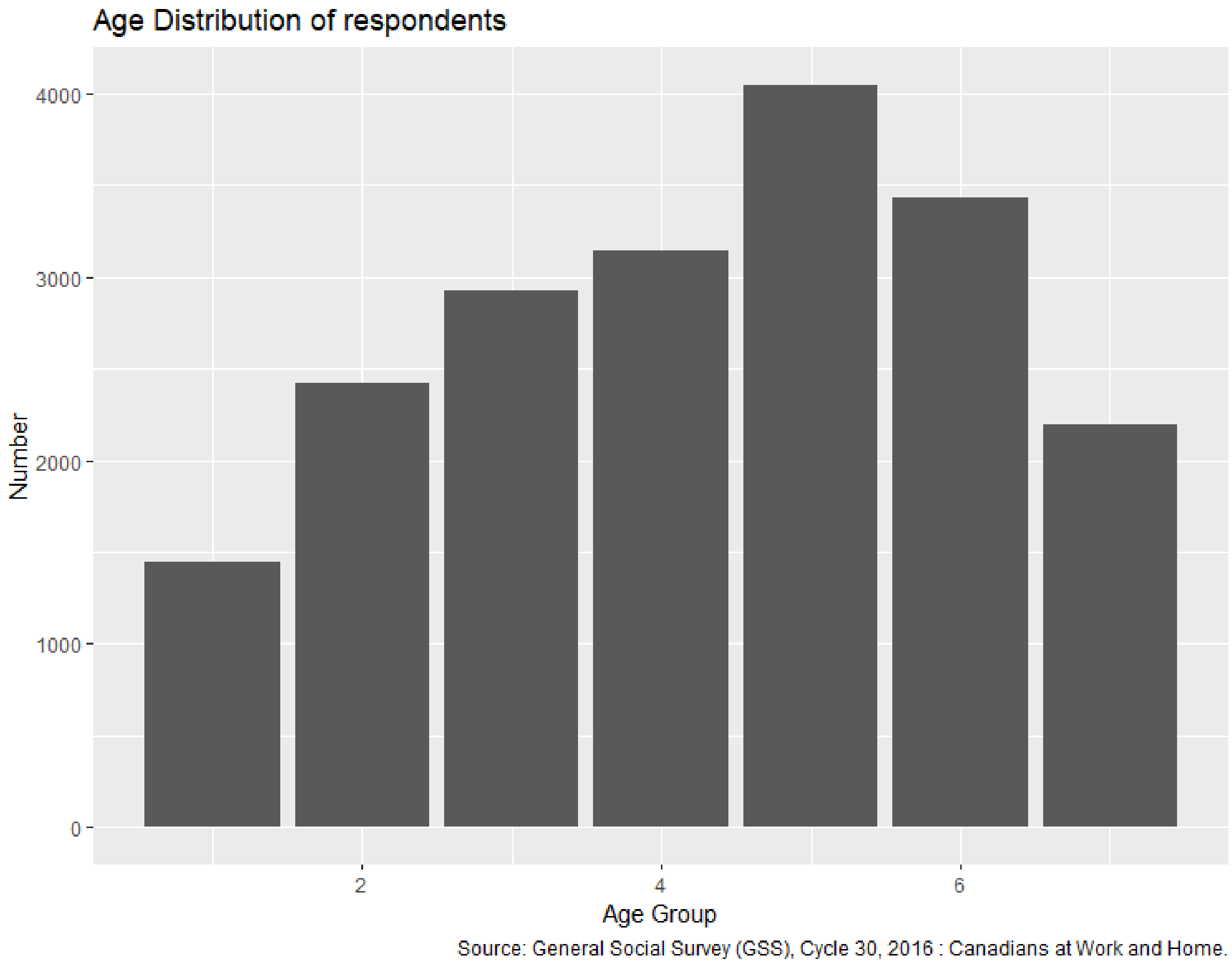

Age Distribution of respondents
Number
4000
3000
2000
1000
0
2
4
6
Age Group
Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 1:

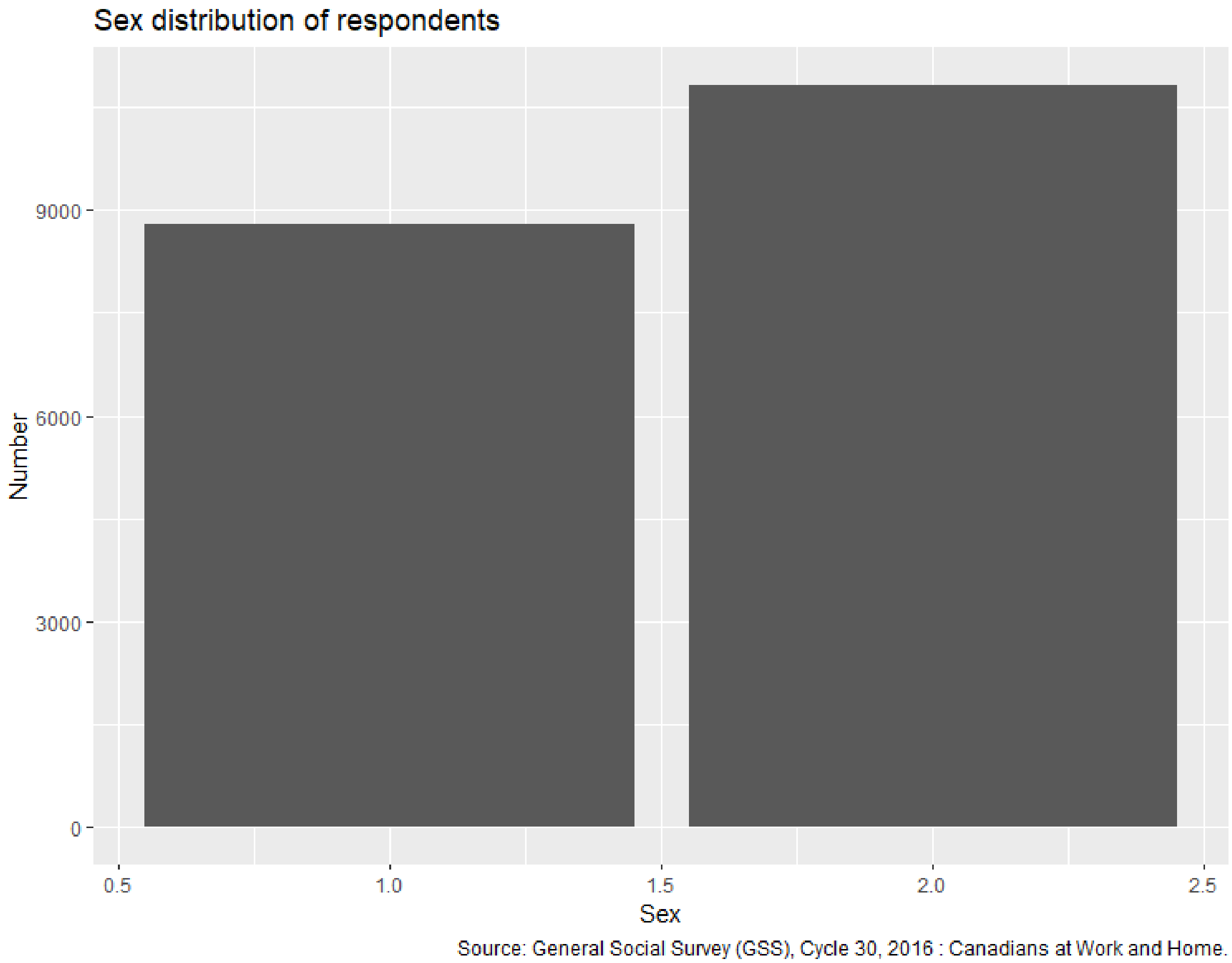

Sex distribution of respondents
Number
9000
6000
3000
0
0.5
1.0
1.5
2.0
2.5
Sex
Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 2:

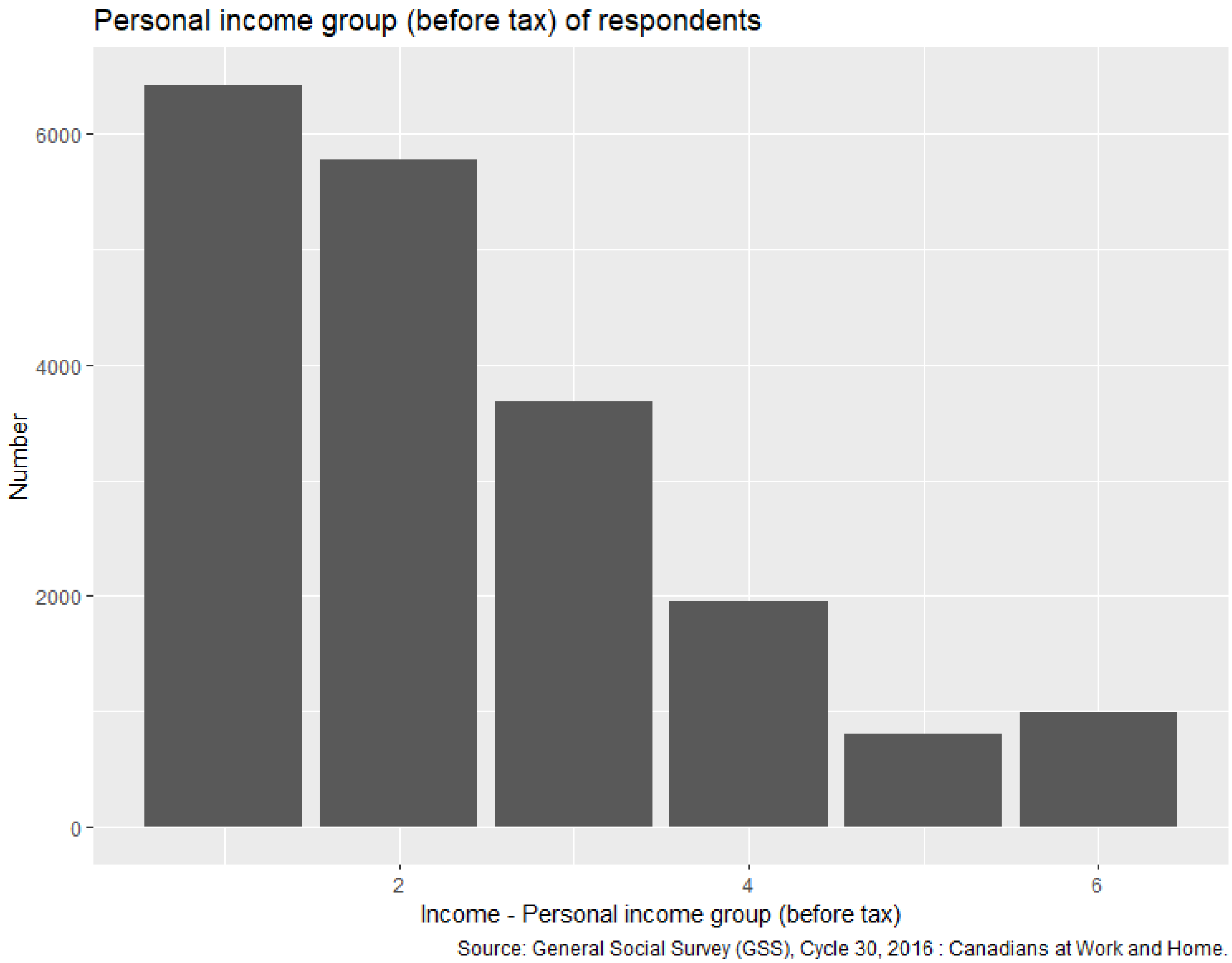
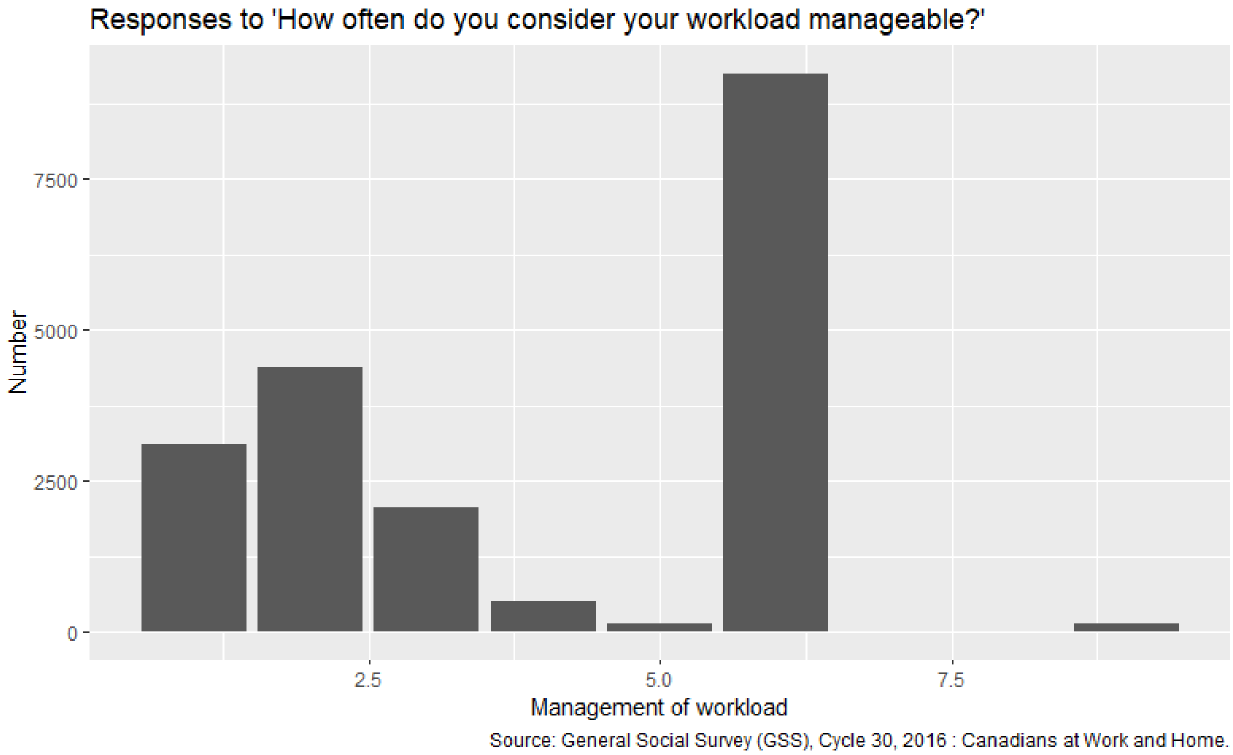
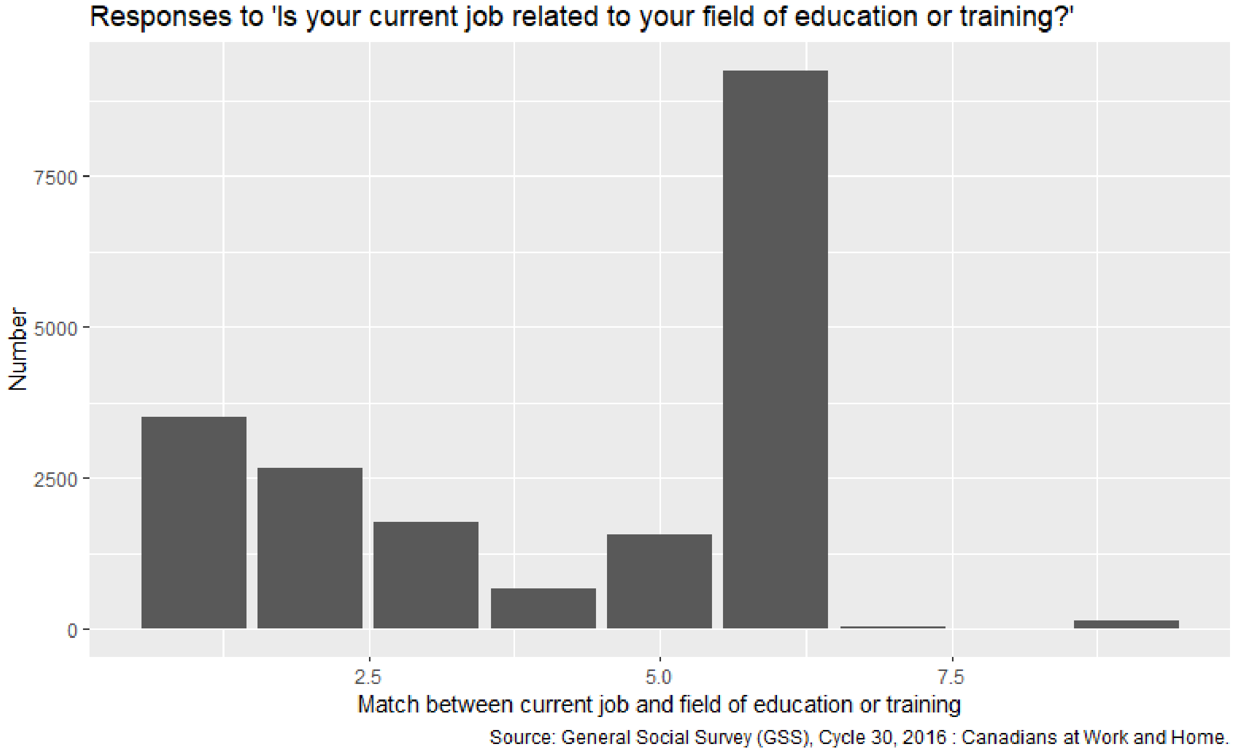

Personal income group (before tax) of respondents
Number
6000
4000
2000
0
2
4
6
Income - Personal income group (before tax)
Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 3:



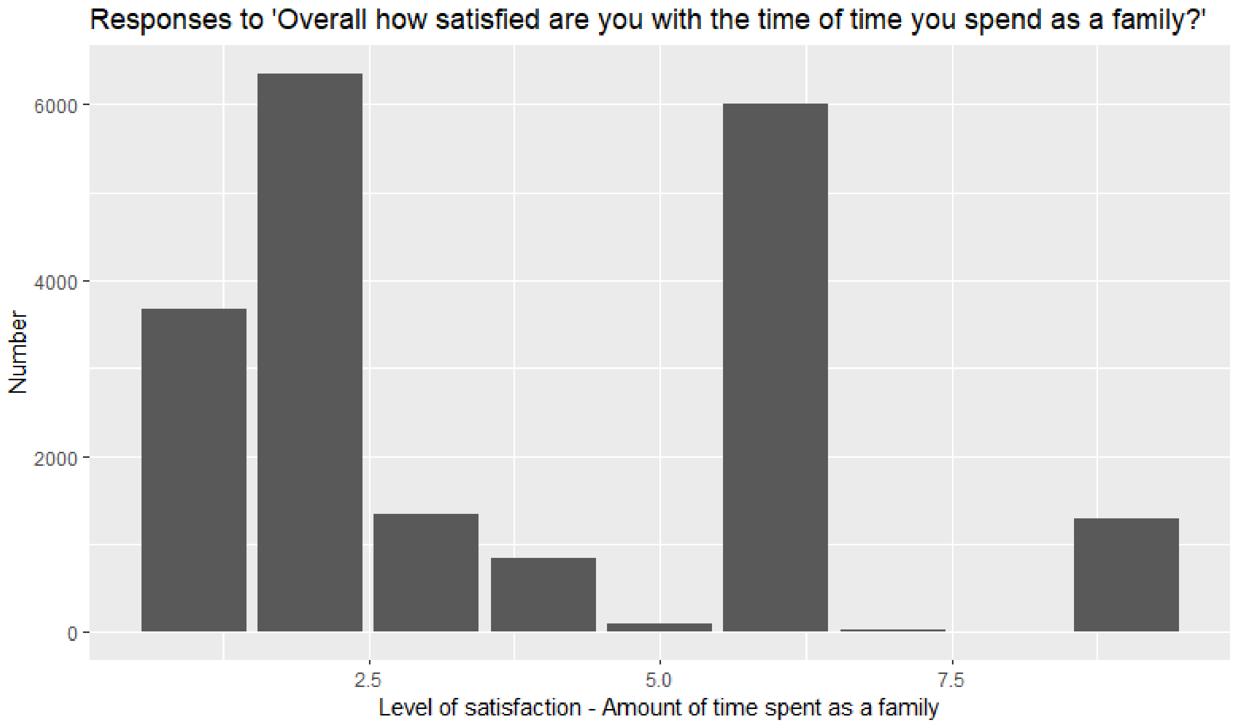
Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 4:



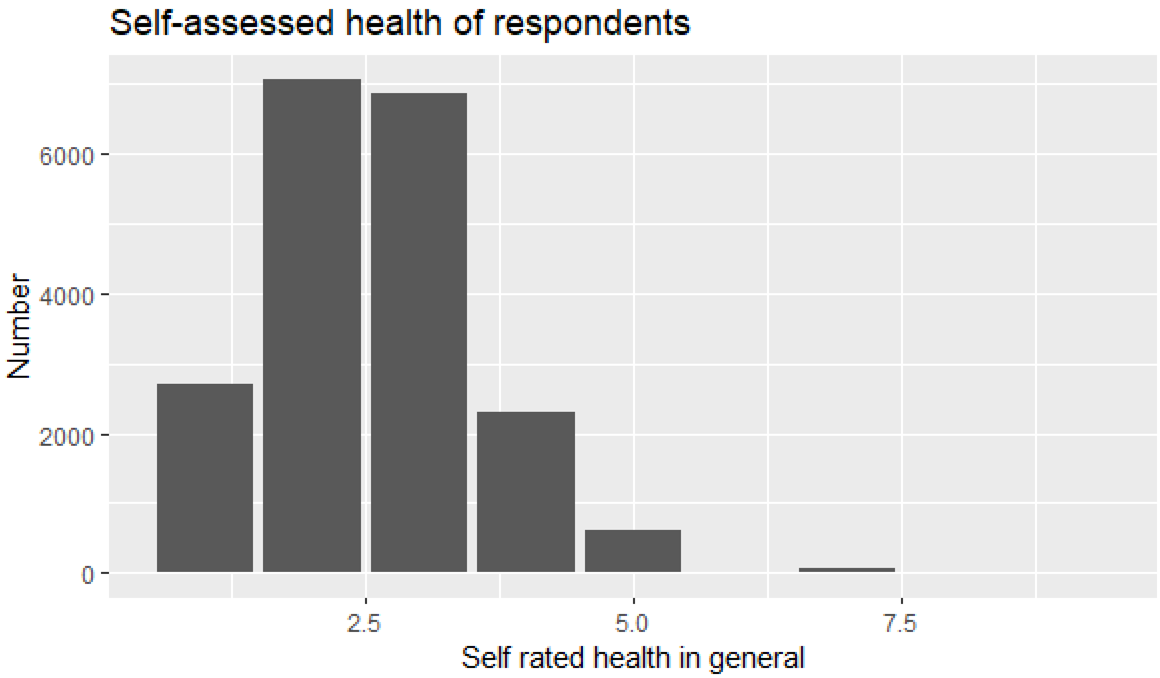
Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 5:



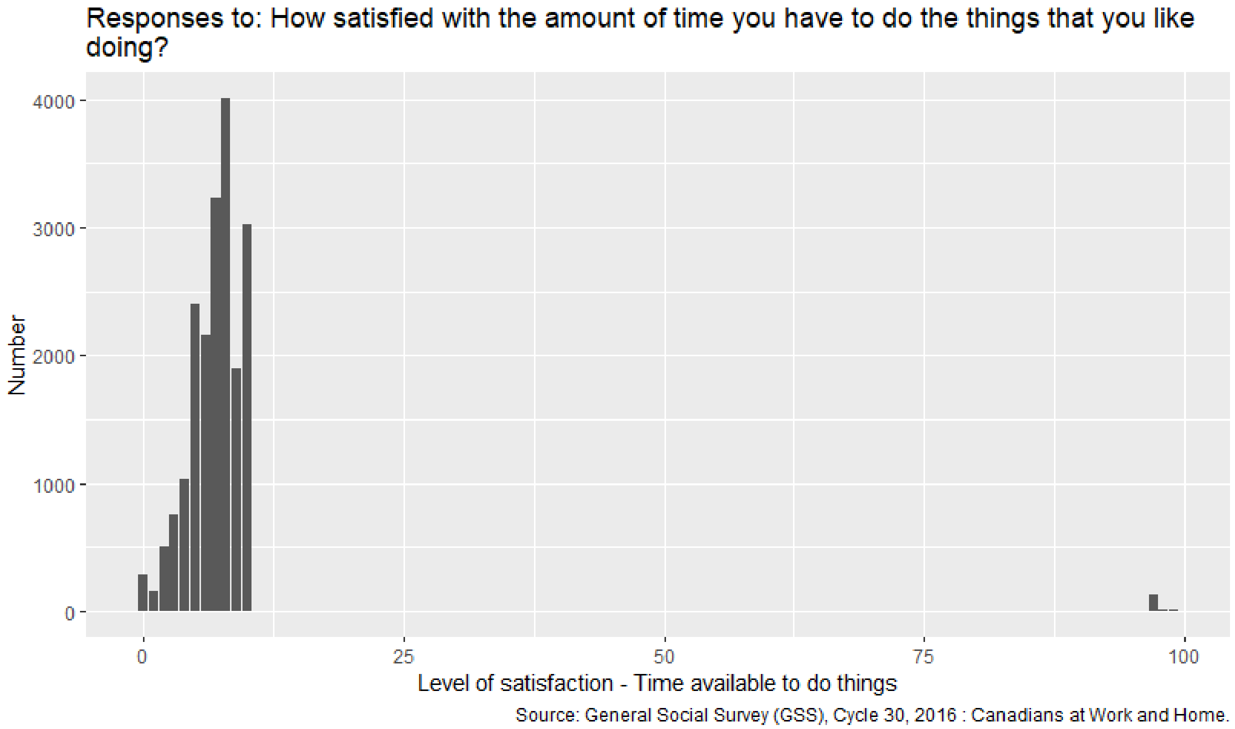
Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 6:



Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 7:



Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 8:

Since GSS intends to protect respondent privacy, the options for each survey question include "Valid skip", "Don't know", "Refusal", "Not stated", we deleted them using na.omit() function in R for better analysis. Moreover, as the answer options for multiple survey questions are somewhat similar, we merged these options to facilitate the analysis. For example, regarding the question "how satisfied are you with the amount of time you spend as a family?", the options "Very satisfied" and "satisfied" are combined, and "dissatisfied" and "Very dissatisfied" are combined.

In the next section of modeling, we analyzed these variables and how they affected the stress level among young adults. Meanwhile, we used R (R Core Team (2020)) to obtain a proper logistic regression model with good predictions on stress level among young adults.

Model

We plan to use a generalized linear regression model for prediction. Among all these 7 predictors, we find that time spent with family, self-rated health, workload management have the most predictive effect. Therefore, we obtain a GLM with these predictors using the function gml (we used the dplyr and tidyverse packages (Hadley Wickham (2020) Wickham et al. (2019)) in R (R Core Team (2020)).)

Before building the model, we randomly used 60% of the survey data as a training set and 40% as a test set for the model validation. The training set was used to build the model while the test set was to check our model performance, such as whether each predictor is still significant, and whether the estimated coefficients are similar.

Next, we build a full model with all the selected variables above, and then through deleting and selecting them, we find 4 variables strongly related to the stress level of young adults with effective predictive ability. They are:

- Health.rated: In general, would you say your health is. . . ? (Excellent/ Very good / Good / Fair / Poor)

- Workload.manage: How often do you consider your workload manageable? (Always / Sometimes / Rarely)
- Family time: Overall, how satisfied are you with the amount of time you spend as a family?" (Satisfied / General / Dissatisfied)

The final logistic regression model we obtain is:

$$log(\frac{p}{1-p}) = -1.3285 + 0.6635x_{family.time_{general}} + 1.2838x_{family.time_{dissatisfied}} + 0.2041x_{health.rated_{Verygood}}$$

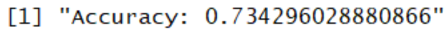
$$+0.5896x_{health.rated_{good}} + 1.4144x_{health.rated_{fair}} + 2.2444x_{health.rated_{poor}}$$

$$+1.3616x_{workload.manage_{sometimes}} + 2.2479x_{workload.manage_{rarely}}$$

Model Interpretation

For the right side of the model, -1.3285 is the intersection of the model. There are four exploratory variables. Each variable takes three to five different levels, and these x values can be taken as 0 or 1. For example, for the variable: “Family time: Overall, how satisfied are you with the amount of time you spend as a family?”, there are three different levels: “satisfied”, “general”, and “dissatisfied”. If a person is generally satisfied with the family time, then the coefficient should be set to 1, and the others to 0. If a person is satisfied with the family time, then all other coefficients should be set to 0.

For the left side of the model, P represents the possibility that an individual is under high stress. After getting the result on the right side of the model, we can obtain this probability by taking the exponent and a few steps of calculation.



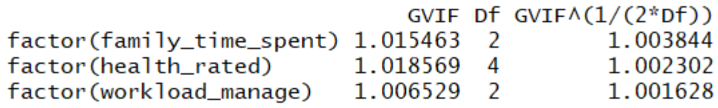
```
[1] "Accuracy: 0.734296028880866"
```

Figure 9: Accuracy of the model

When our model is applied to an independent new data set (test data set), the prediction accuracy of the model is nearly 73.43%. (Assuming the prediction is correct when the predicted probability is greater than 50%)

Results

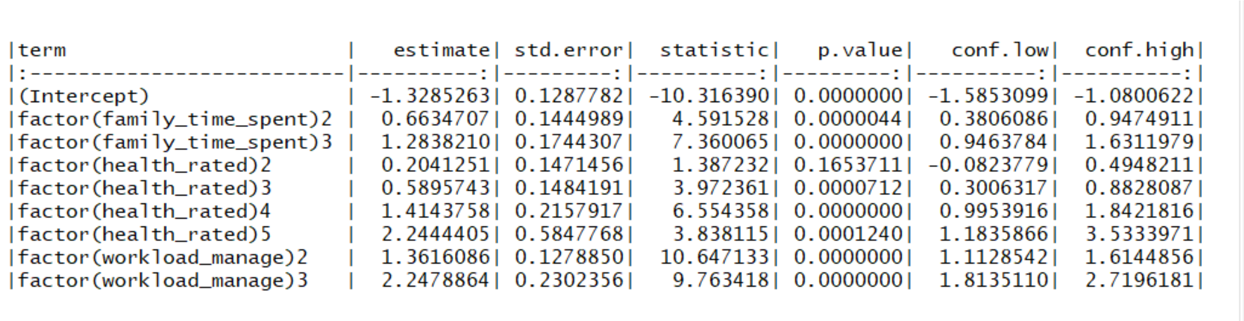
To ensure that there is no strong correlation between the predictor variables, we check the multicollinearity of each variable by their VIF values. (we used the Car package (Fox and Weisberg (2019)))



```
                              GVIF Df GVIF^(1/(2*Df))
factor(family_time_spent) 1.015463  2        1.003844
factor(health_rated)      1.018569  4        1.002302
factor(workload_manage)   1.006529  2        1.001628
```

Figure 10: VIF values of each predictor

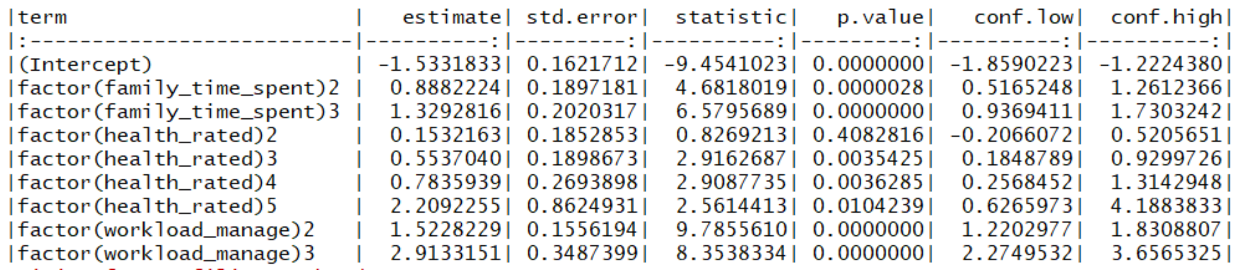
Figure 10 shows that their VIF values are less than 5, indicating that no multicollinearity occurs in the model.



```
|term                         |   estimate| std.error|  statistic|   p.value|   conf.low|  conf.high|
|:----------------------------|----------:|---------:|----------:|---------:|----------:|----------:|
|(Intercept)                  | -1.3285263| 0.1287782| -10.316390| 0.0000000| -1.5853099| -1.0800622|
|factor(family_time_spent)2   |  0.6634707| 0.1444989|   4.591528| 0.0000044|  0.3806086|  0.9474911|
|factor(family_time_spent)3   |  1.2838210| 0.1744307|   7.360065| 0.0000000|  0.9463784|  1.6311979|
|factor(health_rated)2        |  0.2041251| 0.1471456|   1.387232| 0.1653711| -0.0823779|  0.4948211|
|factor(health_rated)3        |  0.5895743| 0.1484191|   3.972361| 0.0000712|  0.3006317|  0.8828087|
|factor(health_rated)4        |  1.4143758| 0.2157917|   6.554358| 0.0000000|  0.9953916|  1.8421816|
|factor(health_rated)5        |  2.2444405| 0.5847768|   3.838115| 0.0001240|  1.1835866|  3.5333971|
|factor(workload_manage)2     |  1.3616086| 0.1278850|  10.647133| 0.0000000|  1.1128542|  1.6144856|
|factor(workload_manage)3     |  2.2478864| 0.2302356|   9.763418| 0.0000000|  1.8135110|  2.7196181|
```

Figure 11: Model summary built with the training data

(We used the knitr package (Xie (2014) Xie (2015) Xie (2020).))

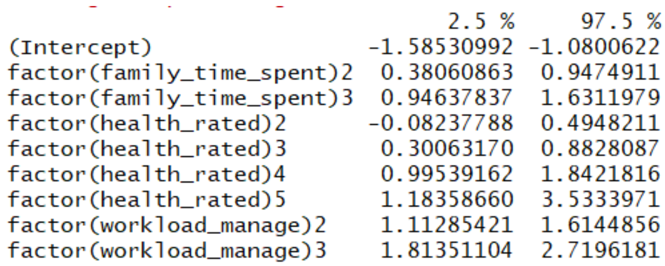


```
|term                         |   estimate| std.error|  statistic|   p.value|   conf.low|  conf.high|
|:----------------------------|----------:|---------:|----------:|---------:|----------:|----------:|
|(Intercept)                  | -1.5331833| 0.1621712| -9.4541023| 0.0000000| -1.8590223| -1.2224380|
|factor(family_time_spent)2   |  0.8882224| 0.1897181|  4.6818019| 0.0000028|  0.5165248|  1.2612366|
|factor(family_time_spent)3   |  1.3292816| 0.2020317|  6.5795689| 0.0000000|  0.9369411|  1.7303242|
|factor(health_rated)2        |  0.1532163| 0.1852853|  0.8269213| 0.4082816| -0.2066072|  0.5205651|
|factor(health_rated)3        |  0.5537040| 0.1898673|  2.9162687| 0.0035425|  0.1848789|  0.9299726|
|factor(health_rated)4        |  0.7835939| 0.2693898|  2.9087735| 0.0036285|  0.2568452|  1.3142948|
|factor(health_rated)5        |  2.2092255| 0.8624931|  2.5614413| 0.0104239|  0.6265973|  4.1883833|
|factor(workload_manage)2     |  1.5228229| 0.1556194|  9.7855610| 0.0000000|  1.2202977|  1.8308807|
|factor(workload_manage)3     |  2.9133151| 0.3487399|  8.3538334| 0.0000000|  2.2749532|  3.6565325|
```

Figure 12: Model summary built with the test data

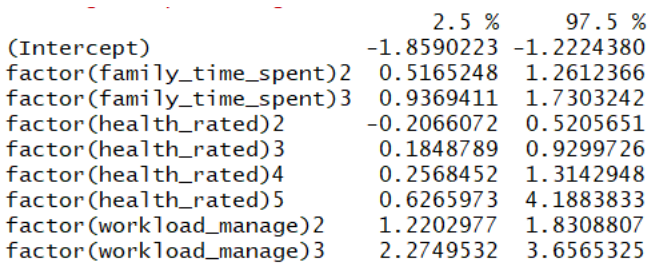
(We used the knitr package (Xie (2014) Xie (2015) Xie (2020)))

From Figure xx and xx, we compare the two models. Figure xx indicates that just one significant variable was lost, and the estimated coefficients of predictors for the two models are similar.



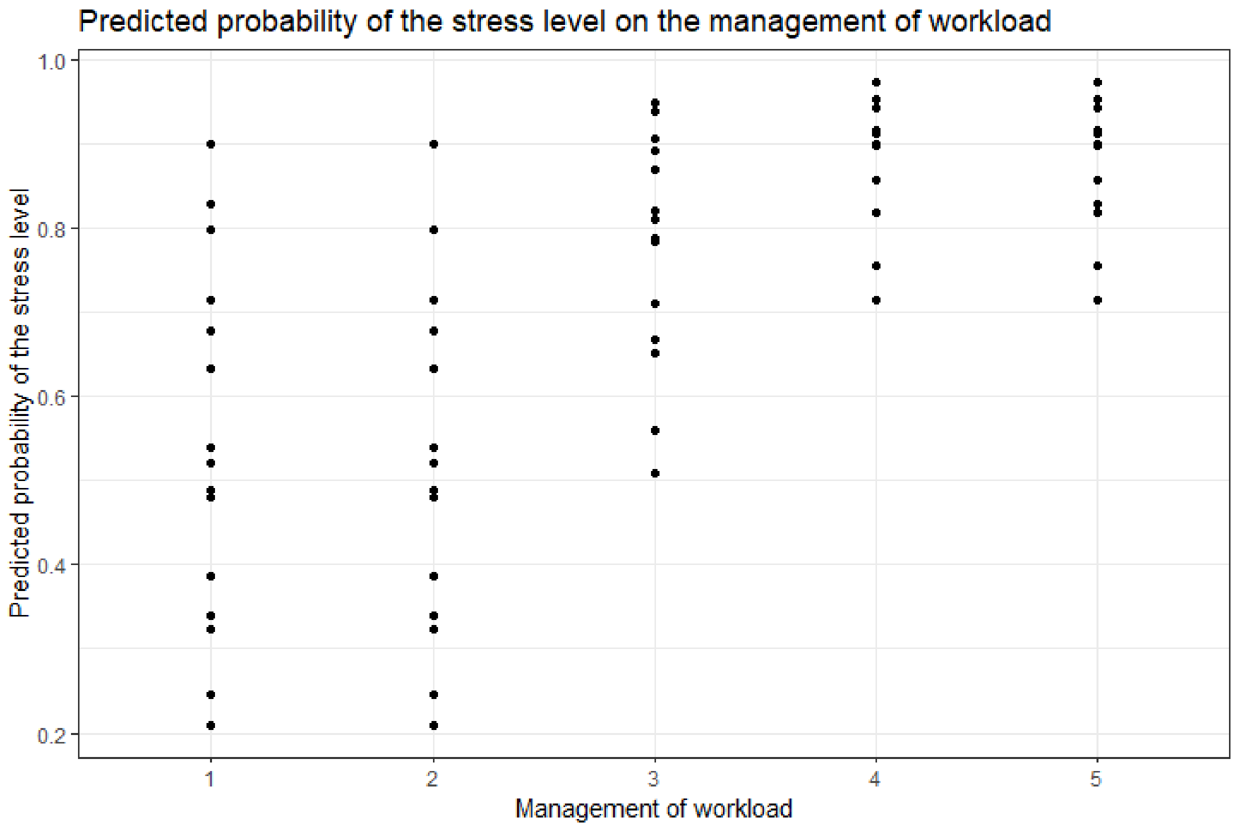
```
                                2.5 %     97.5 %
(Intercept)                -1.58530992 -1.0800622
factor(family_time_spent)2  0.38060863  0.9474911
factor(family_time_spent)3  0.94637837  1.6311979
factor(health_rated)2      -0.08237788  0.4948211
factor(health_rated)3       0.30063170  0.8828087
factor(health_rated)4       0.99539162  1.8421816
factor(health_rated)5       1.18358660  3.5333971
factor(workload_manage)2    1.11285421  1.6144856
factor(workload_manage)3    1.81351104  2.7196181
```

Figure 13: Confidence interval of model coefficients built with the training data



```
                                2.5 %     97.5 %
(Intercept)                -1.8590223 -1.2224380
factor(family_time_spent)2  0.5165248  1.2612366
factor(family_time_spent)3  0.9369411  1.7303242
factor(health_rated)2      -0.2066072  0.5205651
factor(health_rated)3       0.1848789  0.9299726
factor(health_rated)4       0.2568452  1.3142948
factor(health_rated)5       0.6265973  4.1883833
factor(workload_manage)2    1.2202977  1.8308807
factor(workload_manage)3    2.2749532  3.6565325
```

Figure 14: Confidence interval of model coefficients built with the test data



Source: General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home.

Figure 15: Predicted probability of the stress level on the workload manage

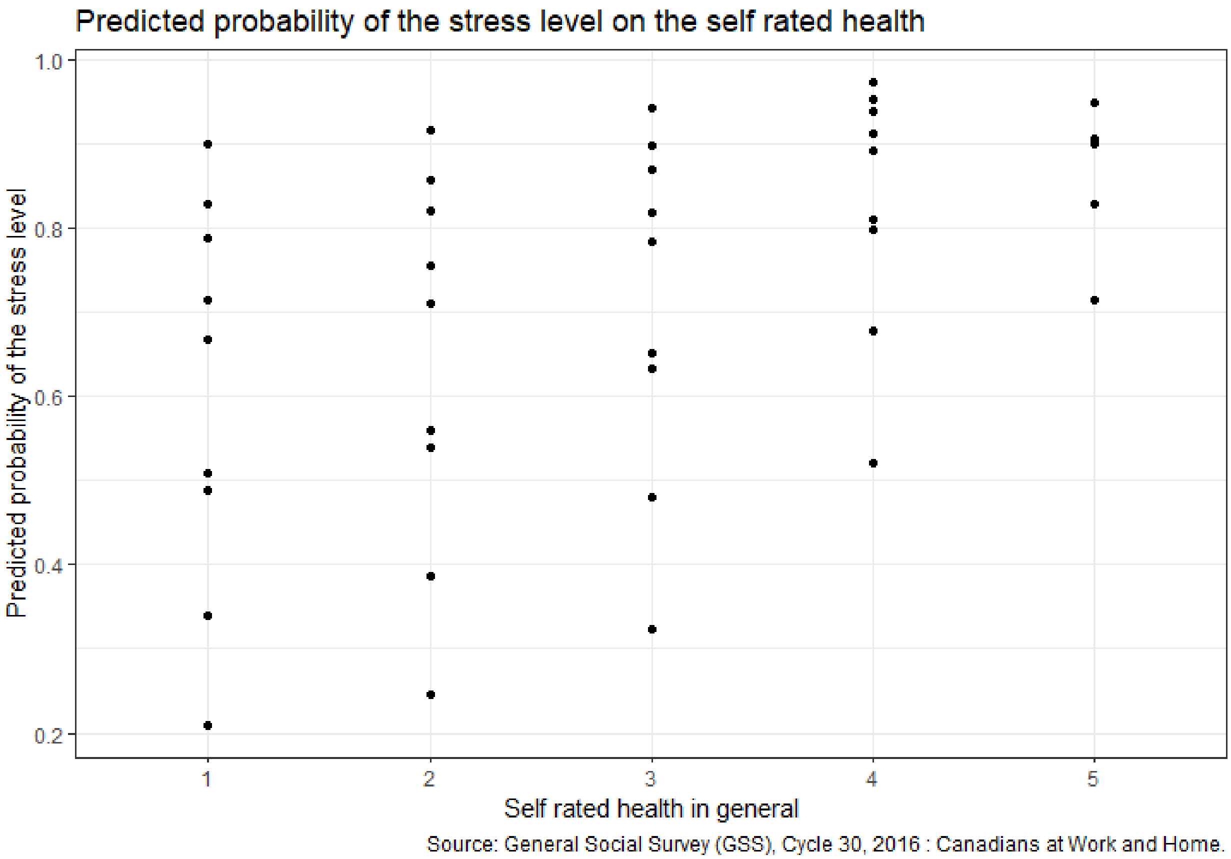


Figure 16: Predicted probability of the stress level on the self rated health

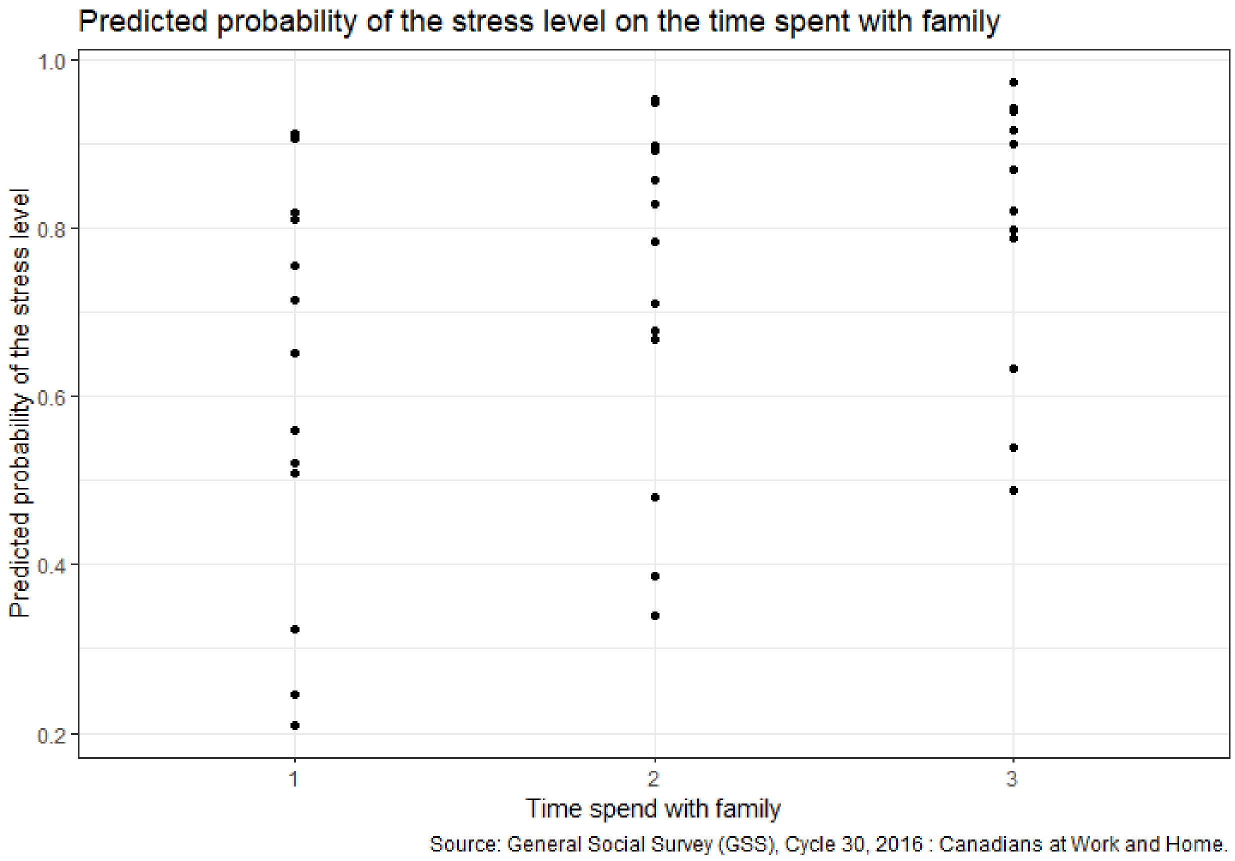


Figure 17: Predicted probability of the stress level on the time spent with family

Discussion

to be continued…

Weakness and Next Step

to be continued…

Appendix

Code for this analysis is available at: https://github.com/cici889/finalpaper

Reference

Disease Control, Centers for, and Prevention. 2020. "Coping with Stress." https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/managing-stress-anxiety.html.

Firke, Sam. 2020. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression.* Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

GSS. 2016. *General Social Survey (Gss), Cycle 30, 2016 : Canadians at Work and Home.* https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5221.

Hadley Wickham, Lionel Henry, Romain François. 2020. *Dplyr: A Grammar of Data Manipulation a Fast, Consistent Tool for Working with Data Frame Like Objects, Both in Memory and Out of Memory.* https://dplyr.tidyverse.org,%20https://github.com/tidyverse/dplyr.

Ipsos. 2020. "Stress Becoming a Way of Life for Canadian." https://www.ipsos.com/en-ca/stress-becoming-way-life-canadians.

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

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.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files.*

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr.* 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://yihui.org/knitr/.