Exploring the factors that highly correlate with Stress Level

Xingyu pu, Geng Li, Yuchwn Wu, Zhihuan Shao

Oct.15, 2020

Abstract

This paper uses data from General Social Survey (GSS) on Canadians at Work and Home in 2016, which is a sample survey with cross-sectional design and conducted from August 2nd to December 23rd 2016. A logistic mathematical model will be used in this analysis, along with a few simple data analyzing techniques, to better illustration our findings as well as implications of this analysis. The results show that whether participants had high stress level or not is positively correlated with alcohol consumption, hours of working per week, personal income, and eating habits, and negatively correlated with smoking habits, however, observations with alcohol consumption and smoking habits were not significant. In conclusion, the longer the working hours per week are, the higher the income is, and the poorer the eating habit is, the higher chance of getting high stress level.

Introduction

General Social Survey (GSS) on Canadians at Work and Home in 2016, which is a sample survey with cross-sectional design and conducted from August 2nd to December 23rd 2016. The target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. This survey aimed at taking a comprehensive look at the way Canadians live by incorporating the realms of work, home, leisure, and overall well-being, and thus knowing more about the lifestyle behaviors of Canadians that impact their health and well-being both in the workplace and at home.

The Goal of this analysis is to explore a few potential factors that contribute to the stress level of participants, we used part of the survey data to find the correlations between those factors that could solely or cross influence the stress level, including but not limited to income level, life style, and working conditions. As more and more people nowadays are accompanied by increased stress level, our designing idea is to explore and understand the stress level of modern families and what aspect of daily life can affect and contribute to the increasing stress in people. We hope this report can help people better understand how we could help alleviate the current highly stressful society, and what aspect people should avoid to escape from severe stress.

Data

The data comes from GSS on Canadians at Work and Home in 2016, which contains various parts of daily lives and working conditions of Canadians in the year of 2016. We chose this year because it was the actual first year for our PM Justin Trudeau to proceed various policies since he won the election on Oct 2015. The high unemployment rate before 2016 had a steady control since he had taken the office, and gradually decrease afterwards. Thus, a survey regarding living condition and employment condition in 2016 could partially reflect the national situation at that time point, how people felt about the new government, how people felt about living pressures and working environment, as well as how various factors can affect the daily lives of the normal civilians.

The survey contains a lot of aspects, from working conditions, immigration conditions, incomes, to living styles, nutrition awareness, and physical and mental health. It thoroughly collected the data to reflect almost every aspect of daily lives and provided great opportunities to explore the connections among those factors.

We are mostly interested in the correlations between stress level and other factors, and exploring how those factors, for example, living styles and working hours, can contribute to the stress level of participants. To simplify the outcome, we divided the stress level into 2 categories - Not stressful and Stressful. Any participants that picked “Not at all stressful” or “Not very stressful” will be categorized to Not stressful, and the remaining ones will be categorized to Stressful. Since most answers are also categorical, it might make less sense to keep too many categories, thus, we simplified that into simply 2 categories. Other factors will be ranked numerically, with larger factors indicating a worse condition (e.g. poorer eating habits, or longer working hours).

Potential drawbacks: As mentioned above, most of the data are categorical, and not continuous, thus this could make a lot of analysis methods a little bit hard to implement. Lack of detailed parameters could results in larger bias, as well as larger error or extremely concentrated data distribution.

Model

The model we used is logistic model, and it will be used to predict whether a few factors can lead to stress in participants. Logistic model is best used to model a binary dependent variable, and can be used to measure the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. Comparing to other models, for example, SLR, logistic model is stronger at processing categorical data as opposed to SLR. However, due to the nature of the data itself, a Bayesian model might not provide more information as the prior distributions are not easily observed, thus a simple logistic model to predict the outcome is used in this report.

The model we implemented is shown as below:

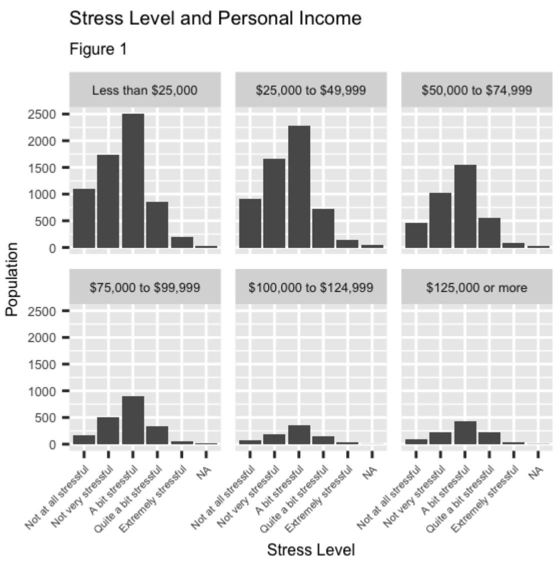
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{drink} + \hat{\beta}_2 X_{smoke} + \hat{\beta}_3 X_{hours} + \hat{\beta}_4 X_{income} + \hat{\beta}_5 X_{eating}$$

There are totally 5 factors that we chose to explore the connections between them and stress level, including alcohol consumption, smoking habit, working hours per week, personal income, and eating habit. This model will be built to predict whether a person with the corresponding conditions mentioned above will end up with stress or not. From our results, the finalized formula can be written as below:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.549 + 0.015X_{drink} - 0.039X_{smoke} + 0.240X_{hours} + 0.060X_{income} + 0.223X_{eating}$$

Results

First, let’s take a look at a couple of figures showing the connections between stress level and some factors



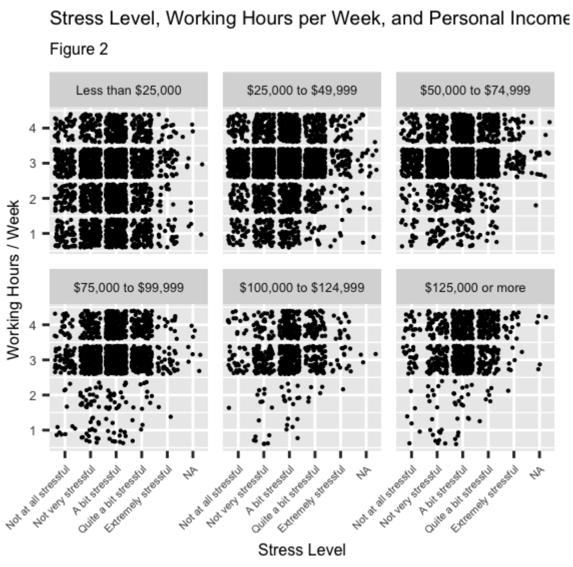


Figure 1 demonstrated the connections between stress level and personal income. It is not hard to observer that most people from all income levels were experiencing moderate stress level, and most people are with lower incomes (less than $25,000 per year). However, with the income level increased (more personal annual income), the concentrations shift to higher stress level. This may indicate that people with higher incomes has higher possibilities of experiencing higher stress level.

Similarly, when adding another components to the analysis, looking at stress level and income and hours of working per week, as shown in Figure 2.

Figure 2 gives us more information regarding the distribution of population in stress level, incomes, and hours worked per week. People with higher incomes tended to work longer, and more working hours per week seemed to contribute to a higher stress level.

Nevertheless, we would like to explore the impact on stress level with even more factors, and that is when logistic model is used. We divided the stress level into simply 2 categories, Stressful or not, and use 5 factors to predict whether a person could have stress with corresponding conditions, such as lower incomes or longer working hours per week.

The model summaries are shown below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5488120	0.0841431	-6.522361	0.0000000
drinking	0.0149842	0.0122891	1.219304	0.2227287
smoking	-0.0388763	0.0298964	-1.300368	0.1934750
hours_worked_week	0.2396693	0.0250858	9.554000	0.0000000
income	0.0603253	0.0149119	4.045455	0.0000522
eat_habit	0.2227777	0.0215136	10.355209	0.0000000

As shown above, the model fitted showed a positive correlation with alcohol consumption, hours working per week, personal income, and eating habits, and a negative correlation with smoking. This model predicted that higher level of alcohol consumption, longer working hours per week, higher incomes, and poorer eating habits, could all contribute to higher chance of getting stressful, and surprisingly, smoking could decrease the chance. However, observations with smoking habit and alcohol consumption is not significant, thus those cannot be regarded as principal components when exploring connections with stress level.

However, as the survey was conducted by random sampling, we still need to do a population correction. Based on the census in 2016, there were approximately 35.15 million people in Canada, and the survey sample pool is 19609. Thus, we need to correct the model based on the population.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5488120	0.0843990	-6.502588	0.0000000
drinking	0.0149842	0.0124397	1.204551	0.2284000
smoking	-0.0388763	0.0301640	-1.288831	0.1974814
hours_worked_week	0.2396693	0.0253563	9.452063	0.0000000
income	0.0603253	0.0149055	4.047175	0.0000522
eat_habit	0.2227777	0.0214766	10.373029	0.0000000

Discussion

Discussion on Survey Design: Taking a look at the questionnaire itself, it is quite elegantly laid out and designed. Most areas that involves various aspects of daily lives, activities, cultural influences, working conditions etc. had been included. Although some categories, such as incomes and hours per week can be even more detailed divided, since from the results, take personal income for example, most people earned less than 25,000 per year, however, less than 25,000 is still a comparatively large area. Knowing better of details could give more information on how those aspects can affect daily lives.

As for the methodologies, the target population included all persons 15 years of age and older in Canada, excluding a few indigenous and hard-to-reach people, and a simple random sample without replacement was implemented, using a combination of telephone and address register to contact participants. There existed a bias that only people with telephone access was samples, however, people with telephone were much harder to reach, this trade-off did not heavily impacted the results while keeping the cost low.

Discussion on the Data Analysis: From the analysis above, we can conclude that stress level is positively correlated with alcohol consumption, hours working per week, personal income, and eating habits, and negatively correlated with smoking habits. Among all of those, hours worked per week and eating habits have the most impact on stress level with highest slopes, and surprisingly, income is not one of the biggest impact factors that influences levels of stress in people. This findings reinforced our hypothesis that various aspects of working conditions and daily lives can contributed to stress level in people, especially how long people had been working per week and people’s eating habits in 2016.

Weaknesses

However, the analysis is not without drawbacks.

1. The model we used to fit the data was actually not very good. The full data will be shown in Appendix, and the residuals for the logistic model were quite high. Other models could potentially increase the accuracy of prediction if implemented well.
2. Simply dividing the stress level into 2 categories could be a little bit of rusty, since the stress levels were much more complicated.
3. A correlation within independent variables were not diagnosed, a potential multicollinearity could still occur with current variables

Next Steps

Our next steps will include using alternative models to evaluate the accuracy of predicting, for example, a Bayesian logistic model can be used, or a linear model with mixed effects can also be used. The multicollinearity should also be evaluated in future work, thus it wouldn’t affect the outcome from model prediction. Also, exploring more aspects and their relations with stress level is also worth looking into, for example, time spent with families, or sports habits, those could potentially affect stress level as well. This could provide more information regarding how should we deal with stress, what other aspects should we pay attention to in order to avoid severe stress etc.

References

1. H. Plecher, Unemployment rate in Canada 2021, Apr 28, 2020, https://www.statista.com/statistics/263696/unemployment-rate-in-canada/ (https://www.statista.com/statistics/263696/unemployment-rate-in-canada/)
2. General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home
3. Population and Dwelling Count Highlight Tables, 2016 Census, Statistics Canada, https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Comprehensive.cfm (https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Comprehensive.cfm)

Appendix

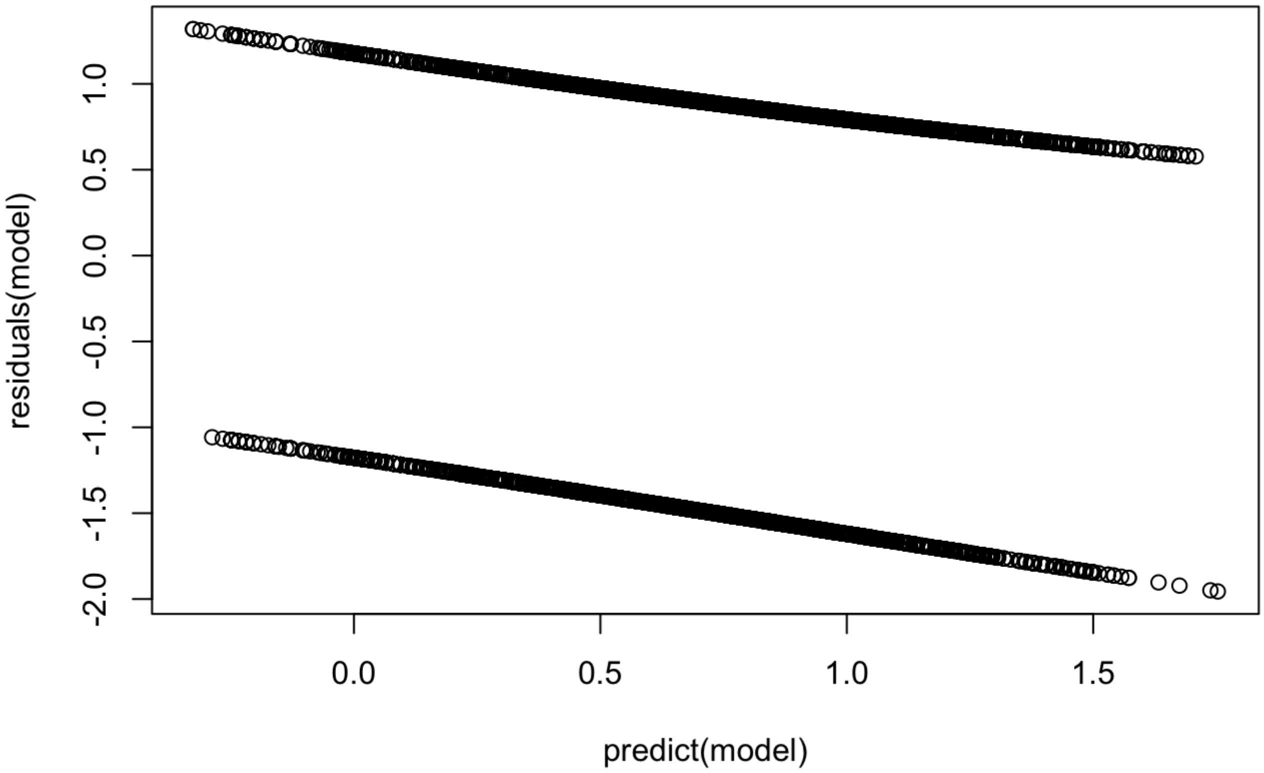
Code and data supporting this analysis is available at:

https://github.com/xingyupu/PS2 (https://github.com/xingyupu/PS2)

Model summary:

```
## 
## Call:
## glm(formula = stress ~ drinking + smoking + hours_worked_week + 
##     income + eat_habit, family = "binomial", data = df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9558  -1.3750   0.7998   0.9013   1.3189  
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)    
## (Intercept)       -0.54881    0.08414  -6.522 6.92e-11 ***
## drinking           0.01498    0.01229   1.219    0.223    
## smoking           -0.03888    0.02990  -1.300    0.193    
## hours_worked_week  0.23967    0.02509   9.554  < 2e-16 ***
## income             0.06033    0.01491   4.045 5.22e-05 ***
## eat_habit          0.22278    0.02151  10.355  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 15379  on 12155  degrees of freedom
## Residual deviance: 15112  on 12150  degrees of freedom
##   (7453 observations deleted due to missingness)
## AIC: 15124
## 
## Number of Fisher Scoring iterations: 4
```

Residual plots:


residuals(model)
predict(model)