Towards the prediction of Time Stress Level among people above 15 in Canada

STA304: Surveys, Sampling and Observational Data

Jiahan Deng, Zhaonan Liu, Shijie Min, Tianyi Zhang,

2020/10/16

Abstract

Recently, many people have experienced time stress at least once. Each type of person might experience stress for different reasons. Students might be struggling with too much homework or tests or exams within a short time while parents might not have enough time to take care of their children or finish daily housework. A dataset was obtained with possible cause of time stress, along with with other factors such as sex, education background, and employment status conducted from January to December 2010. A logistic model could be built to discover how different factors could affect the level of time stress. Then people can use these results to help them prevent possible stress from daily life.

Introduction

The goal is to find out what factors might influence the level of time stress, and how those variables affect it.The dataset was obtained from the Social and Aboriginal Statistics Division, the General Social Survey. There were 1580 variables researched in the survey, but 8 of them has been chosen. Focusing on the impact of respondent's sex, age, education background, and employment status to the level of time stress. In this analysis, the response variable "TCS_Q190" represent the respondent's stress status due to lack of time along with other possible predictors such as age, sex, education background, employment status, number of children in house, etc. There are in total 15391 observations conducted from January to December 2010. In this report, we will build a logistic model to predict the level of time stress, and find out how different factors could affect the level of time stress.

Data

The dataset was obtained from the Social and Aboriginal Statistics Division, the General Social Survey. The dataset originally contained 1580 variables researched in the survey with 15391 observations, using stratified sampling method.(Béchard, 2011)[1]. Each province is a strata and the team collected the survey through telephone. The target population includes all persons above

15 years old in Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut, and full-time residents of institutions. (Béchard, 2011)[1]. The target sample for Cycle 24 was 22,000, while the final sample size (respondents) was 15,390. The stratification being carried out selected 10 provinces which were divided into strata, many Census Metropolitan Areas were considered as strata and in total there were 27 strata being chosen. The sampling method is good because it covers as many people as possible, but it has a few drawbacks like cannot cover households without telephones or people who only have cellular telephone service. The questionnaire is effective because it covers many concepts of possible causes for stress and examines people's life with many details which is good for research. However, too many answers might make researchers harder to find the right variable.

The total number of variables in the main file is 1580. The dataset is sufficient because it contains a large number of observations, and there are a lot of attributes for people to choose and analyze. However, this dataset only contains citizens above 15 years old in Canada. We might not be able to say that only people above 15 might have different levels of time-stress. Moreover, the dataset contains limited information just for people in Canada, people with different demographic backgrounds might have different feelings towards time-stress. Last but not least, people's ability to withstand stress might be different, so people with similar attributes might still have different levels of time-stress.

Variable Selection For the response variable "TCS_Q190" is renmaed to "Stress". And among all 1580 variables, 8 possible predictor variables has been selected and for the response variable "Stress" which is the answer to the question "Do you often feel under stress when you don't have enough time?". The predictors was chosen from different perspectives.
For general information, we have chosen 3 variables as predictors:
1. AGEGR10: it is the age group of the respondent in groups of 5. We choose it because ages might be a significant factor for stress.And it is renamed to be "Age_group"
2. SEX: it is the sex of responendent. We would like to find out whether one sexuality is easier to get stress than the other. And it is renamed to be "Sex"
3. CHRTIME6: it is the number of respondent's children living in the household(any age or marital status). As children can be a cause for stress , we would like to analyze if the number of children can influence the stress level. And it is renamed to be "Household_size"
For Main Activity and Education of respondents, we choose four related questions:
4. MAR_Q134: it is whether the respondent looked for a job in the last four weeks. It is possible that the depression of no work or the anxiety of waiting for a response can cause more stress. And it is renamed to be "Find_job"
5. EQR_Q150: It is whether the respondent goes to further schooling beyond elementary or high school. In our opinion, education level might be one influential factor for stress and we would like to analyze it deeper later. And it is renamed to be "Further_schooling"
6. MAR_Q381: It is whether the respondent has more than one job in the latest week. And it is remaned to be "Morejob"
For time use in leisure, we choose 2 variables:
7. DVSPORT: It is the duration(in minute) of the respondent spends on sport each day. Sports can be a good way of healing from stress, so people who spend more time on sport might experience lower levels of stress.However, the data only gives categorical results of "No time spend doing this activities". So, the result has been modified into 'no time' and 'some time'. And the variable is renamed to "Activity".

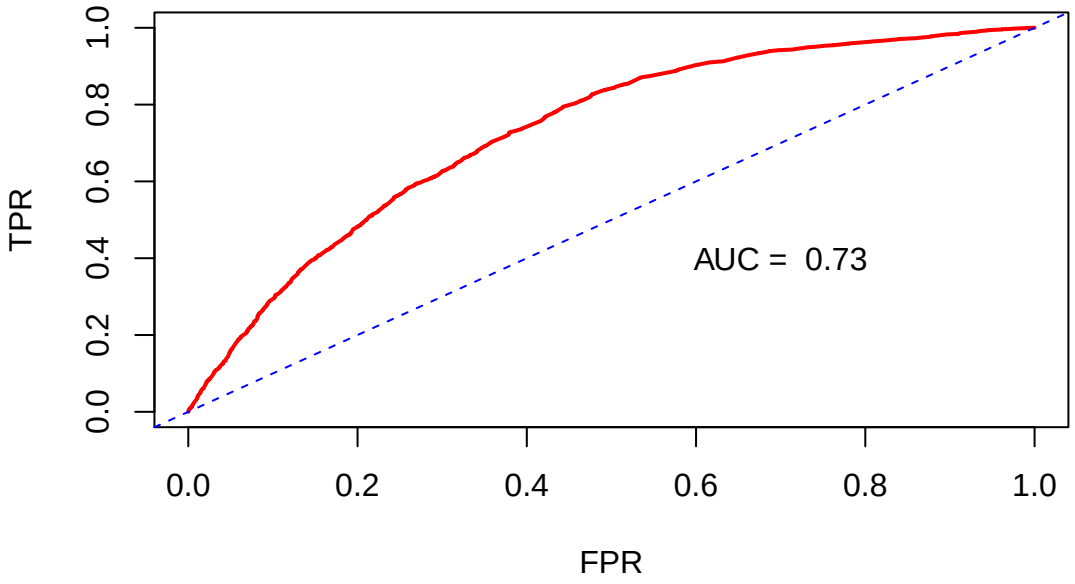
8. DVMEDIA: It is the duration(in minute) the respondent spends on media or communications. Media or communications are very effective when dealing with negative emotions, so it is possible that spending more time on media or communications can reduce the level of stress. However, the data only gives categorical results of "No time spend doing this activities". Therefore, the result has been modified into 'no time' and 'some time'. And the variable is renamed to "Media".

Result

Coefficients Table

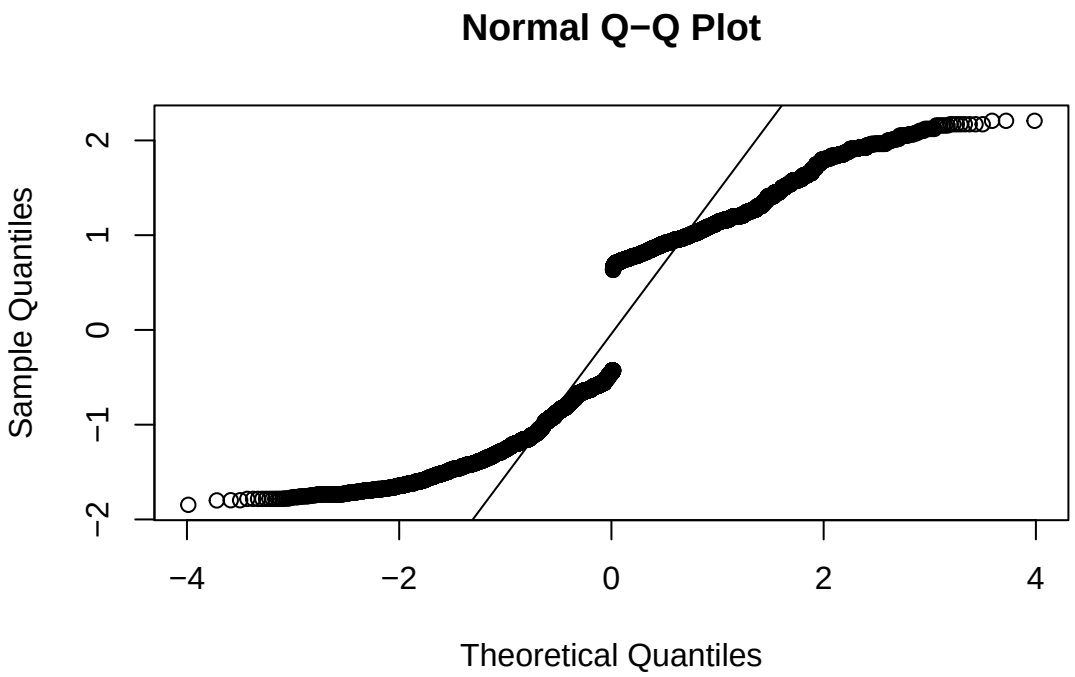
Variables	Coefficient	p-value
Intercept	0.90782	$<2e^{-}16$
Age group25 to 34	0.07037	$<2e^{-}16$
Age group35 to 44	0.11155	$<2e^{-}16$
Age group45 to 54	-0.29262	$<2e^{-}16$
Age group55 to 64	-0.89643	$<2e^{-}16$
Age group65 to 74	-1.75681	$<2e^{-}16$
Age group75 years and over	-2.24761	$<2e^{-}16$
SexMale	-0.57349	$<2e^{-}16$
Household sizeFour household members	0.06336	$<2e^{-}16$
Household sizeOne household members	-0.25927	$<2e^{-}16$
Household sizeSix household members	-0.12037	$<2e^{-}16$
Household sizeThree household members	-0.05799	$<2e^{-}16$
Household sizeTwo household members	-0.14074	$<2e^{-}16$
Find jobnotask	0.02420	$<2e^{-}16$
Further schoolingYes	0.25436	$<2e^{-}16$
MorejobYes	0.14105	$<2e^{-}16$
ActivitySometime	-0.08712	$<2e^{-}16$
MediaSometime	-0.10860	$<2e^{-}16$

From this table, it is clear that all predictor variables have p-value smaller than 0.05, which means they are significant to the response variable. Thus this might be a precise model and the ROC curve and Calibration plot are helpful to check its performance.

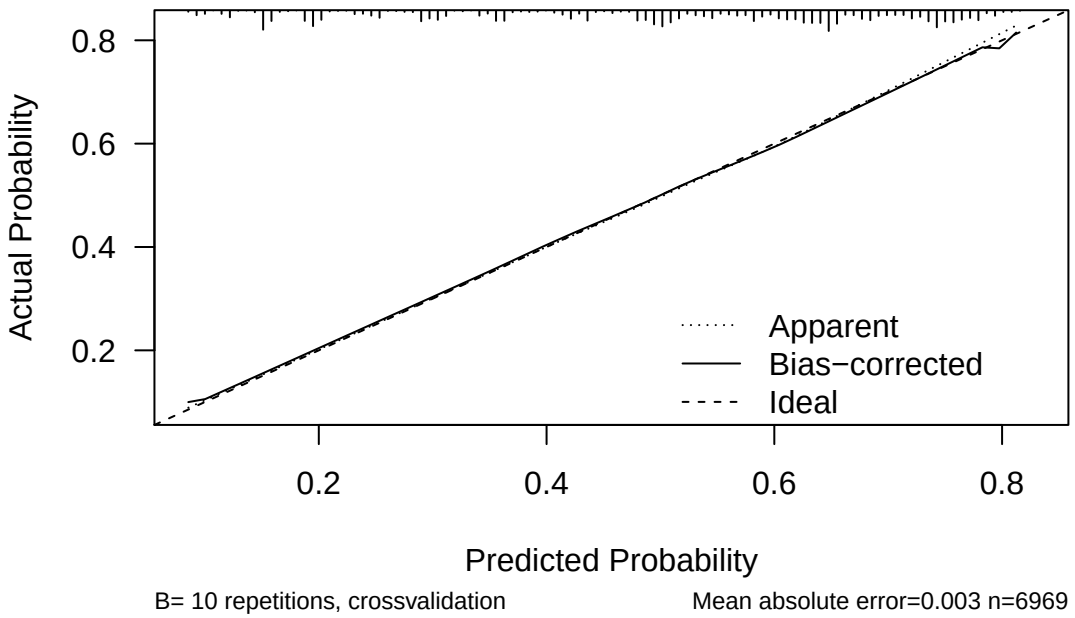



```
## Area under the curve: 0.7309
```

The ROC Graph The area under the curve(AUC) is 0.73. This means that the logistic regression model can discriminate stress or not stress 73% of the time. For more acturate digits, the calculated the area under the curve is 0.7309. Since AUC closer to 1 is better, the result of 0.73 is not a bad performance.



The Normal Q-Q Plot The Normal Q-Q plot helps to check the model assumptions. From this graph, there is a gap in values, probably because the NA observation that was deleted from the original dataset. A normal distribution in Q-Q plot should show a straight line, which is different from this graph; instead of a normal distribution, it is more like a bimodal distribution.



```
##
## n=6969   Mean absolute error=0.003   Mean squared error=2e-05
## 0.9 Quantile of absolute error=0.006
```

The Calibration Plot A Calibration plot can also used in verifying the performance of a model. This plot shows the bias-corrected line is almost the same with the expected diagonal 45-degree line, only the ends of the tails deviate from the expected diagonal line. Therefore, the model does perform very well in the prediction.

Discussion

Final Model $\log\left(\frac{p}{1-p}\right) = 0.908 + 0.070X_{AgeGroup25to34} + 0.112X_{AgeGroup35to44} - 0.293X_{AgeGroup45to54} - 0.896X_{AgeGroup55to64} - 1.757X_{AgeGroup65to74} - 2.248X_{AgeGroup75+} - 0.573X_{SexMale} - 0.259X_{Household1} - 0.141X_{Household2} - 0.058X_{Household3} + 0.063X_{Household4} - 0.120X_{Household6+} + 0.024X_{NotFindJob} + 0.254X_{FutherSchoolingYes} + 0.141X_{MoreJob} - 0.087X_{SomeActivity} - 0.109X_{SomeMedia}$

This represents that the odds of having time-stress for older than 75 years old respondent is exp(-2.248) times the odds of patients 25 to 34 years old. In other words, the odds of having time-stress from people older than 75 are smaller than people from 25 to 34 years old. This not surprising because elders are usually retired from their work and having a steady income(pension). Moreover, their children had grown up and could take care of them. However, people from 25 to 34 years old are just starting their career path or starting a family. They are beginning to gaining responsibility for their babies and even their parents. Thus they have a higher possibility of experiencing time-stress than the elders.

The model also shows the odds of having time-stress for males exp(0.029) times the females' odds. This means that men are better at dealing with time-stress. And American psychological association once did a study of gender and stress, which said: "Women are more likely than men (28 percent vs. 20 percent) to report having a great deal of stress"(American Psychological Association, 2012)[3]. This can also prove the result of this model that females have a higher possibility of experiencing time-stress than males.

And as assumed earlier, the model concludes that the odds of people who do sports and activity are exp(-0.087) times the odds of people who do not do any activity. A study from Anxiety and Depression Association of America also said: "Exercise and other physical activity produce endorphins—chemicals in the brain that act as natural painkillers(...)in turn reduces stress. (Physical Activity Reduces Stress)[4]", which proves the result of the model is correct. Therefore, people who do sports and activities are less likely than people who do not do any activity to report a time-stress.

This model shows the odds of people who are finding a job is exp(0.024) times the odds of people who are not finding a job; the odds of people who are having more than one job is exp(0.141) times the odds of people who only have one job; the odds of people who spend time on social media and communication is exp(-0.109) times the odds of people who does not communicate, etc.

Limitation

Based on the dataset itself, there are some limitations. The first one is the sampling error. The sample cannot represent the whole population, then there is a non-sampling error. Since there are many missing values in the dataset, it differences from the estimates to the true values. For example, the variable MAR_Q390("how many days a week did/do you usually work") was planned to be a predictor, but the dataset contains too many missing values and is not sufficient for the final model.Thus, the variable was dropped from the list.
Still, errors can be various. Not only by sampling errors, but respondents also made mistakes when taking the survey since they were required to answer so many questions. Answers being recorded into the CATI systems might be wrong. Moreover, errors happened through phases and some of them only can be decreased but not eliminated.
The error occurs closely related to people themselves. In addition, this dataset only contains citizens above 15 years old in Canada. It might not be able to say that only people above 15 are facing time stress. Different age intervals should be included. Also, the dataset contains information just for people in Canada, so the model cannot represent people with different demographic backgrounds. Last but not least, people's ability to withstand stress might be different, so people with similar attributes might still have different response of time-stress.

Next Step

The next step being taken to improve throughout this specific research is to use more variables to see if there is a more accurate model. Since this report only use 8 variables as predictors, it might not be very accurate when estimating the stress. Then for further research improvement, more changes can be applied to the survey and improve the accuracy of the model.
Since there are few errors and issues occur while performing analysis on this survey, the first step to take is to work on the perfection of the dataset by reducing the number of missing data or extra variables. A more efficient way to analyze this data is to divide the survey into broader sections. From the massive data set, the experimenter wanted to include different causes for stress, so it includes extra details. However, the section division is not clear enough for people who take this survey. Thus, changing the section division would be easier to analyze and lead to more accurate to prediction.
A follow-up study after this research could be on time stress between people of different ages. To be specific, the survey could be divided into two parts for adults(>18) and teenagers(<=18). It could be mainly focused on job and family for adults group, social interaction and education for teenagers. This could result in more precise models.

Reference

[1] Béchard, M. (2011, December). 2010 GSS Time Stress and Well-Being(Time Use) Public Use Microdata File. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss24v5/gss24main/en/more_doc/GSSC24V5ENgid.pdf
[2] "Welcome to My.access – Please Choose How You Will Connect." My.access - University of Toronto Libraries Portal, Canadian General Social Surveys, 2010, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda3+gss24mv5.
[3] "Gender and Stress." American Psychological Association, American Psychological Association, 2012, www.apa.org/news/press/releases/stress/2010/gender-stress.

[4] "Physical Activity Reduces Stress." Anxiety and Depression Association of America, ADAA, adaa.org/understanding-anxiety/related-illnesses/other-related-conditions/stress/physical-activity-reduces-st.