Investigating Relationship Between Mental Health and Annual Income using Logistic Regression

STA304H1, Fall 2020

Boyu Cao Jiayi Yu Yijia Liu Ziyue Yang

October 19$^{\text{th}}$, 2020

Abstract

This is an abstract. (.)

Introduction

A topic that intrigues us is the relationship of annual income to subjective well-being of individuals. One of the measurement of well-being is reported mental health, which will be the focus of our analyses. Research shows that there has been increasing concern about the impact of the global economic recession on mental health (Sareen J 2011). We will investigate the relationship between the annual income and mental health status alongside with various factors, including gender, marital status, age. Our assumption is that these factors are playing essential roles to bringing effects on one's mental health.

Particularly, we built and fitted a **logistic regression model** (based on Page 154, Wu, Thompson), and carried out a binary dependent variable Y (i.e. a dummy variable having value of either 0 or 1) indicating the samples' mental health status. We estimated coefficients for each feature in order to predict the binary mental health outcome.

Data

Throughout, regression models will be illustrated using data from the twenty fourth cycle of the **General Social Survey (GSS)** on time-stress and well-being, collected in 2010. The csv data of selected features is available at https://github.com/yangzi33/sta304-ps3.

The target population for this survey includes all persons 15 years of age and older in Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut, as well as full-time residents of institutions. Computer assisted telephone interviewing was used to collect data for GSS, and households were reached by calling a series of randomly-generated phone numbers.

In the GSS, mental health is assessed by a question asking "In general, would you say your mental health is: . . . ", with provided answers "Excellent, Very Good, Good, Fair, Poor", or "Don't know" for those who are unsure about answering. Of the total of 15,390 respondents in the survey, 203 did not choose to answer, and 45 answered "Don't know". Therefore, analyses in our report are based on 15,142 respondents providing valid answers to this question.

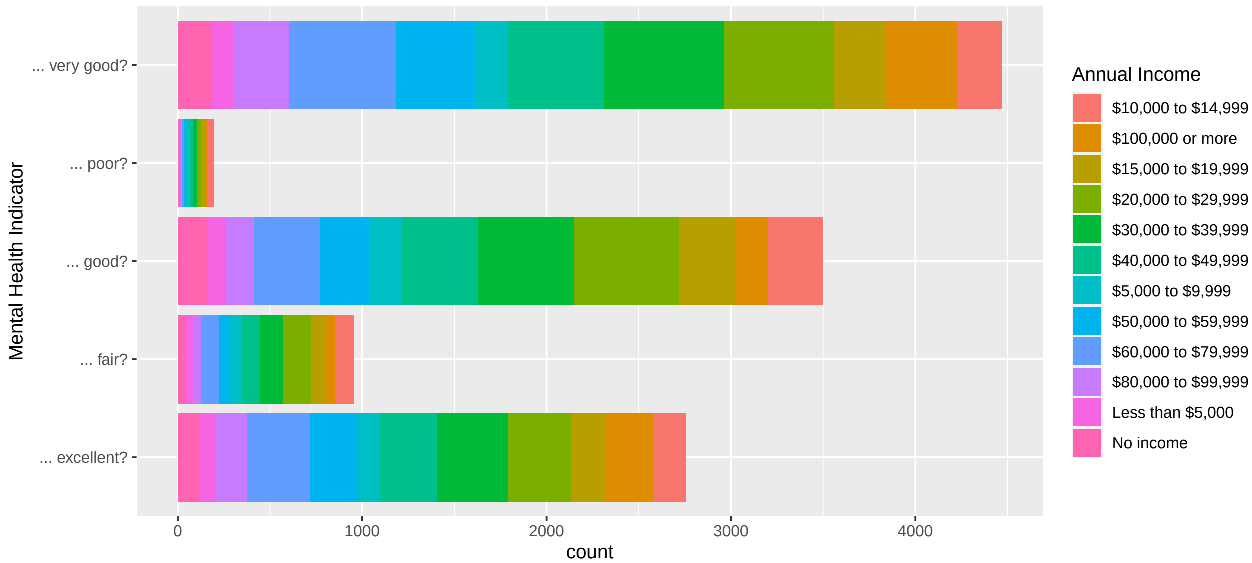
Due to the fact that there are merely four values, we treat them as a binary variable, by coding value 1 for those who answered "Excellent" or "Very Good", and 0 otherwise. The mean of this binary variable is 0.608 or 7224/11873, the proportion of those who are "mentally healthy" in the sample, denoted p, which can be viewed as the probability that a randomly selected sample will be "mentally healthy".

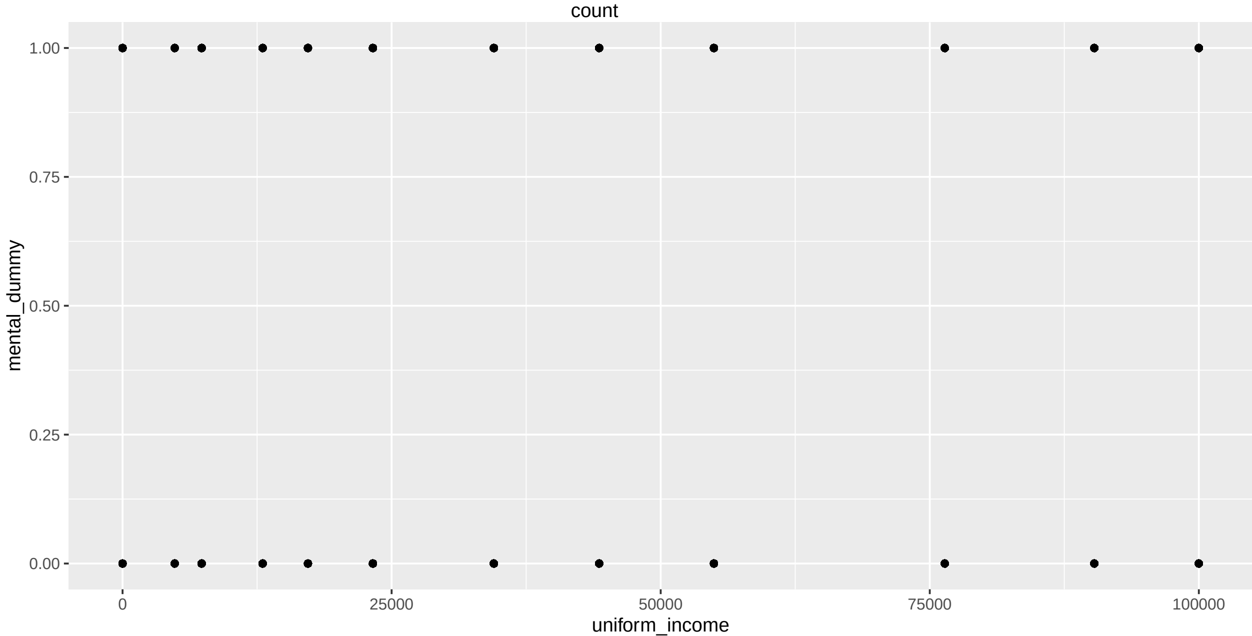
Our focus will be modeling mental health as a function of various variables. Two of the independent variables are treated as interval: age, income, while gender and marital status will be dummies. We will examine the relationship between mental health and marital status. Additionally, since the income groups are grouped by income intervals, which makes it challenging to perform logistic regression. Hence for each sample, we generate a uniformly random number that lies between the income interval of the sample using the `runif()` function.

The following plots illustrates the distribution of each age group

Figure 1. Proportion of Mental Health Status Grouped by Annual Income

Source: Cleaned Dataset.





Model

We choose to use a **Logistic Regression Model** (Page 154, Wu, Thompson) on annual income and mental health.

Specifically, suppose that the value of mental health measure is denoted Y, and

- the amount of annual income is denoted X_1,

- each of the six age groups are denoted $X_2, X_3, X_4, X_5, X_6, X_7$, respectively in an increasing order,
- marital statuses: Living common-law, Married, Separated, Single, Widowed are denoted $X_8, X_9, X_{10}, X_{11}, X_{12}$, respectively.
- Male gender is denoted X_13.

Furthermore, we have an intercept term β_0, and slope terms $\beta_1, \beta_2, \beta_3, \dots, \beta_{13}$ that we wish to predict throughout.

With p as the dependent variable, we write

$$p = \mathbb{P}(Y = 1) = \frac{\exp(\beta_0 + \sum_{i=1}^{13} \beta_i X_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{13} \beta_i X_i)}. \tag{1}$$

for some interception term β_0 and slope terms $\beta_1, \beta_2, \dots, \beta_n$.

Then, we perform a linearization on p using a logit transformation defined as

$$\text{logit}(p) := \log\left(\frac{p}{1-p}\right), \tag{2}$$

such that the logistic regression becomes

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_{13} X_{13}, \tag{3}$$

based on which we examine the relationship between mental health and income. We wish to estimate model by estimating the values of the intercept term and the slope terms as $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n$.

Results

The following table contains a summary of fitted linearized logistic model:

```
##
## Call:
## glm(formula = mental_dummy ~ uniform_income + age_10 + marital +
##     gender, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9116  -1.2710   0.8351   1.0114   1.4056
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   5.277e-01  1.107e-01   4.765 1.89e-06 ***
## uniform_income                1.123e-05  7.356e-07  15.271  < 2e-16 ***
## age_1025 to 34               -5.831e-01  9.825e-02  -5.935 2.93e-09 ***
## age_1035 to 44               -8.403e-01  9.935e-02  -8.458  < 2e-16 ***
## age_1045 to 54               -8.924e-01  9.804e-02  -9.102  < 2e-16 ***
## age_1055 to 64               -7.001e-01  9.868e-02  -7.095 1.30e-12 ***
## age_1065 to 74               -4.076e-01  1.054e-01  -3.866 0.000111 ***
## age_1075 years and over      -6.825e-01  1.168e-01  -5.846 5.04e-09 ***
## maritalLiving common-law      7.904e-02  9.230e-02   0.856 0.391832
## maritalMarried                2.659e-01  7.038e-02   3.778 0.000158 ***
```

```
## maritalSeparated              -2.393e-01  1.209e-01  -1.980 0.047735 *
## maritalSingle (Never married)  1.162e-02  8.326e-02   0.140 0.889010
## maritalWidowed                 3.527e-02  9.819e-02   0.359 0.719466
## genderMale                    -5.273e-02  4.079e-02  -1.293 0.196066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15897  on 11872  degrees of freedom
## Residual deviance: 15537  on 11859  degrees of freedom
## AIC: 15565
##
## Number of Fisher Scoring iterations: 4
```

From the table, we have the following estimated coefficients:

- $\hat{\beta_0} = 0.5277$,
- $\hat{\beta_1} = 1.123 \times 10^{-5}$,
- $\hat{\beta_2} = -0.5831$,
- $\hat{\beta_3} = -0.8403$,
- $\hat{\beta_4} = -0.8924$,
- $\hat{\beta_5} = -0.7001$,
- $\hat{\beta_6} = -0.4076$,
- $\hat{\beta_7} = -0.6825$,
- $\hat{\beta_8} = 0.0790$,
- $\hat{\beta_9} = 0.2659$,
- $\hat{\beta_{10}} = -0.2393$,
- $\hat{\beta_{11}} = 0.0116$,
- $\hat{\beta_{12}} = 0.0353$,
- $\hat{\beta_{13}} = -0.0527$,

Discussion (.)

Weaknesses and Potential Improvements

Why Logistic Model over Linear Model?

Outcome Variable

One of the major weaknesses in our model is that we do not know the specific amount of annual income of samples, but merely the range, which makes the feature `annual income` categorical, making it difficult to fit a regression model. We managed to avoid the problem by generating a random value that lies between the income interval for each sample; nonetheless,

References (.)

- General Social Survey (GSS) on Family (cycle 31) (2017).

- Bürkner P. C. (2017). brms: An R Package for Bayesian Multilevel Models using Stan. Journal of Statistical Software. 80(1), 1-28. doi.org/10.18637/jss.v080.i01

- Bürkner P. C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal. 10(1), 395-411. doi.org/10.32614/RJ-2018-017

- Carpenter B., Gelman A., Hoffman M. D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., and Riddell A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software. 76(1). 10.18637/jss.v076.i01
- 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/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor
- Sareen J, Afifi TO, McMillan KA, Asmundson GJG. Relationship Between Household Income and Mental Disorders: Findings From a Population-Based Longitudinal Study. Arch Gen Psychiatry. 2011;68(4):419–427. doi:10.1001/archgenpsychiatry.2011.15

Appendix (. Optional)