Analysis of Potential Risk Factors of Depression

Ruoqiuyan Zhang (1003926718)

08/12/2020

The code and data supporing this analysis is available at https://github.com/RuoqiuyanZhang/STA304-Final-Project-RqyZhang

Abstract

As depression is widely concerned and worth analyzing using statistical methods, we will analyze the potential risk factors related to the cause of depression using ordinal logistic regression model to perform model estimation and prediction. Data comes from the National Health and Nutrition Examination Study (NHANES). The causal inference would be made, including the appropriateness of predictors and direction of the associations. We find out that individual's age, gender, whether or not in poverty, and whether taking hard drugs are closely related to the self-reported depression. There are more thorough discussion about the topic at the end.

Keywords

Depression, Causal Inference, Ordinal Logistic Regression, Observational Study, Poverty, Hard Drugs

Introduction

Mental health problems have been widely discussed and concerned worldwide in recent years. Depression is a common disorder worldwide, affecting over 264 million people, and it is a major contributor to the global burden of disease.(WHO, 2020) For the most serious part, depression can lead to suicide, and suicide is the second leading causes of death in young-aged individuals (15-29 years old). As a result of that, in May 2013, a resolution was passed in a World Health Assenmbly calling for a comprehensive and coordinatd response to mental disorders at the country level.(WHO, 2020) Some instruments including both interview and self-report are the measures of depression, which are important for the diagnose and treatment.(APA, 2019) Examples of those instruments are Beck Depression Inventory (BDI), Center for Epidemiologic Studies Depression Scale (CES-D), and Hamilton Depression Rating Scale (HAM-D).(APA, 2019) Treatments are medications that would vary depending on the type of depression and patients' severity.

The reason why depression can greatly affect people's everyday life, and even lead to suicide, is partly because it can result from a complex interaction of social, psychological and biological factors, and finding the root causes sometimes are not an easy task. According to previous researches, some of the risk factors are genetics, abuse, and some adverse life events, such as unemployment and divorce, and so on.(WebMD, 2019) Statistical analysis can be used to find potential risk factors from a number of observational data, so that we can make causal inference about this particular problem, and regarding this, we can better treat depression and improve people's mental condition.

One cleaned and selected observational data will be used to investigate the potential risk factors which have causal link to depression. In the Methodology section (Section 2), details about the data and the model will be described, and the data will be put into model for model estimation and prediction, and further draw causal inference. Result of the model estimation and prediction will be presented in Section 3, and conclusions draw from causal inference and more thorough discussion regarding the results will be presented in Section 4.

Methodology

- Data

The dataset used here is data from the US National Health and Nutrition Examination Study from the NHANES package in R. This is the survey and observational data collected by the US National Center for Health Statistics. The main goal of this survey is to analyze individual's health and nutrition status in the long term since the early 1960's. Data cleaning and variable selection were performed to get the interested dependent variable and independent variables out. Here provided a Table 1, which includes the baseline characteristics of the selected data

```
##
##                        Overall
##   n                     1582
##   Depressed (%)
##      None               1238 (78.3)
##      Several             238 (15.0)
##      Most                106 ( 6.7)
##   Gender = male (%)      802 (50.7)
##   Age (mean (SD))      41.83 (14.74)
##   Poverty (mean (SD))  2.74 (1.71)
##   HardDrugs = Yes (%)   275 (17.4)
```

The selected data "small.nhanes" includes 1582 observations and 6 variables in total. The dependent variable is the self-reported number of days where participant felt down, including three categories: None of the days, Several days, and Most of the days. There are four risk factors taken into consideration, which are participants' gender, age, ratio of family income to poverty guidelines (smaller numbers indicate more poverty), and whether participants are in hard drugs, like have tried cocaine, crack cocaine, heroin or methamphetamine. For categorical variables, Table 1 includes the numbers and percentages, and for numerical variables, Table 1 includes their mean and standard deviation.

- Model

The model chosen for the analysis is the ordinal logistic regression model. Since the dependent variable is a categorical variable, the logistic regression model is more apt. Furthermore, there are three categories in the dependent variable, and there is a clear ordering in the categories, so the ordinal logistic regression would be the best approach.

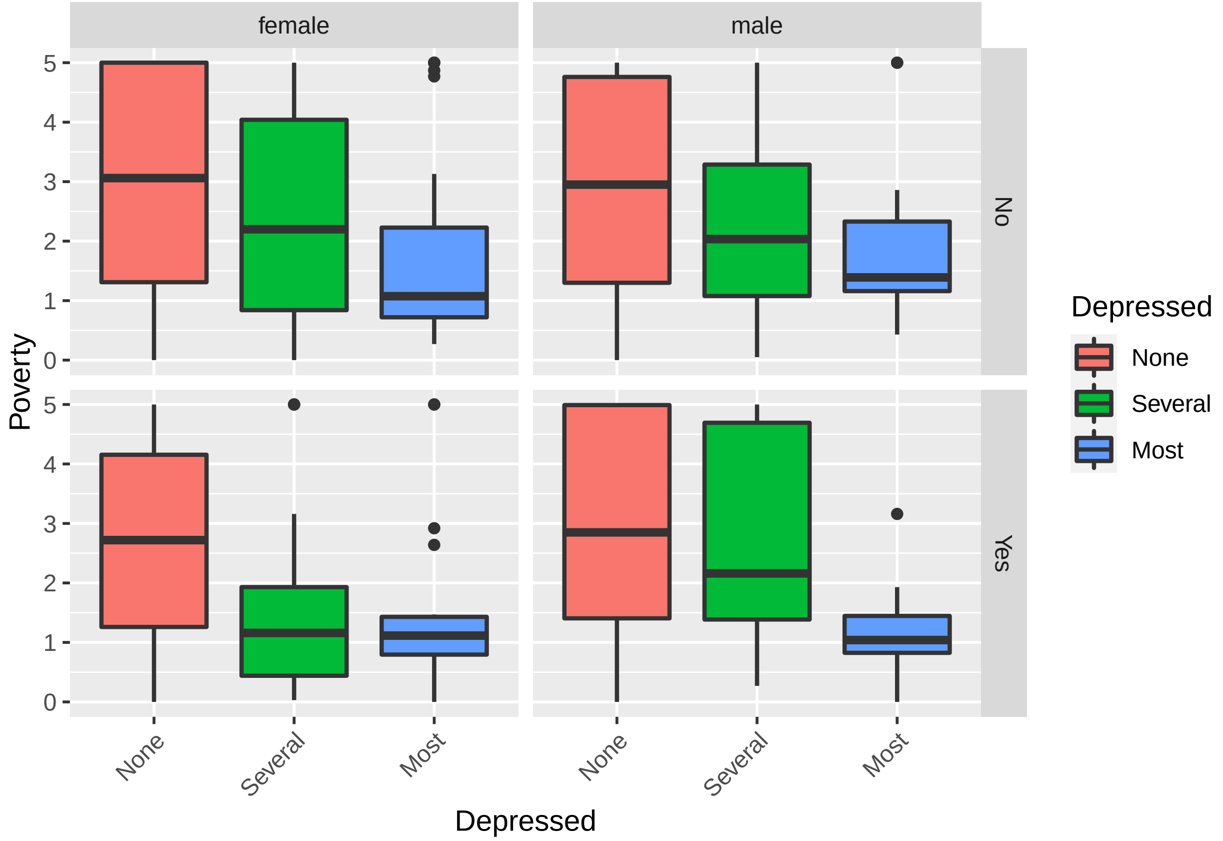
Let Y be the ordinal outcome and J be the total number of categories of the dependent variable. The mathematical formulation of the ordinal logistic regression model is given below:

$$logit\ [P(Y \leq j)] = \alpha_j - (\beta_1 X_{Age} + \beta_2 X_{Gender} + \beta_3 X_{Poverty} + \beta_4 X_{HardDrugs}) where\ j = 1, ..., J-1$$

Here, j is the level of an ordered category with J levels, such that: j = 1 refers to "None" j = 2 refers to "Several" j = 3 refers to "Most". $P(Y \leq j)$ is the cumulative probability of Y less than or equal to a specific category $j = 1, ..., J-1$.

Results

Based off the simple comparison of characteristics from Table 1, we first analyze the distribution of individual-reported depression status across the poverty ratio, gender and whether they are on hard drugs. From this plot, we can have a clearer idea of the relationship between depression status and the potential risk factors.



After fitting the ordinal logistic regression model, we can have the estimated model as:

$$logit\ (P(Y \leq 1)) = 0.988 - 0.011 * Age - (-0.324) * Gender - (-0.322) * Poverty - 0.848 * HardDrugs$$

$$logit\ (P(Y \leq 2)) = 2.419 - 0.011 * Age - (-0.324) * Gender - (-0.322) * Poverty - 0.848 * HardDrugs$$

where 0.988 is the intercept for None|Several, which means having “None of the days” depressed versus having “Several days” or “Most of the days” depressed. Similarly, 2.419 is the intercept for Several|Most, which means having “None of the days” or “Several days” depressed versus having “Most of the days” depressed. 0.011, -0.324, -0.322, and 0.844 are the coefficient of the predictor variables age, gender as male, ratio of poverty, and taking hard drugs, respectively.

Furthermore, we predicted the probabilities of certain individual’s depression status when considering those predictor variables. For example, if we set the test person as a 20-year-old male, not taking hard drugs, and the ratio of poverty is 4.5, the probability of this person has none of the days in depression is 0.926 and has most of days in depression is 0.019. However, if we only change the characteristics by setting the test person as a 50-year-old female, taking hard drugs, and the ratio of poverty is 1.5, the probability of this person has none of the days in depression decreases to 0.507 and has most days in depression increases to 0.188.

Discussion

- Summary

In summary, we analyzed the potential risk factors that would cause or worsen the self-reported depression status by fitting the data from the NHANES into the ordinal logistic regression model. Logistic model would help us investigate the causal relationship of depression with those factors. Since the cause of depression involves many complicated risk factors such as adverse life events, age, drug use, we introduced four predictor variables into the model, which are gender, age, ratio of poverty, and whether the individual is on hard drugs or not. We have the model estimation and some new data prediction base on the model.

- Conclusion

We estimated that the person taking hard drugs is associated with a higher likelihood of being depressed. The t-value is greater than 2 and p-value is less than 0.05, therefore it is statistically significant at the 5% level. Additionally, with one unit increase in ratio of poverty the log of odds of getting depressed decreases by 0.322. With one unit increase in age, the log of odds of being depressed increases by 0.011. Mathematically, the intercept "None|Several" corresponds to $logit\ (P(Y \leq 1))$. It can be interpreted as the log odds of having "None of the days" depressed versus having "Several days" or "Most of the days" depressed. Together with the results from the new data prediction, we can conclude that, with increase in age, decrease in ratio of poverty, as a female and taking hard drugs, a person would have a higher likelihood of being depressed.

- Weakness and Next Steps

Although the dataset includes four variables, there are still many factors not taken into considerations, and the interplay between those factors are very important but not easy to find. The selected data comes from the survey in 2012, which is not the data that could reflect the recent depression status. Especially in year 2020, the COVID-19 pandemic and the lockdown would affect people's mental health greatly, but the analysis cannot apply to the current situation. The model chosed is ordinal logistic regression model, which has both advantages and disadvantages. Logistic regression is easy to implement and interpret, and it provides not only the appropriateness of the predictors, but only the direction of association, so that causal inference is easy to make. However, the limitation is the assumption of linearity between dependent and independent variables, and it is hard to obtain complex relationships using logistic regression.

In order to get more comprehensive results, in the next step, we may find more related risk factors in the model. The mental health problem during COVID-19 pandemic is worth analyzing, so we can find more recent observational data before and during the pandemic to make causal inference about the effect of lockdown, economic recession, and other factors related to people's likelihood of being depressed during pandemic. In addition, multivariate linear regression model can also used to avoid the disadvantages of using logistic regression model.

References

"Depression." World Health Organization, World Health Organization, www.who.int/news-room/fact-sheets/detail/depression.

"Depression Assessment Instruments." Clinical Practice Guideline for the Treatment of Depression, American Psychological Association, Aug. 2019, www.apa.org/depression-guideline/assessment.

"Depression Rick Factors: Genetics, Grief, Abusive Relationships, and Other Major Events." Edited by Smitha Bhandari, WebMD, WebMD, 5 Sept. 2019, www.webmd.com/depression/guide/depression-are-you-at-risk.