STATISTICAL ANALYSIS AND EXPLORATORY DATA ANALYSIS OF MENTAL HEALTH AND DEPRESSION USING NHANES DATA

Authors:

Sravani Mahankali

Sai Rupesh Kavuturi

Vaishnavi Gannavaram

Department Name:

Department of Health Data Science,

Saint Louis University – School of Medicine

Course Name:

HDS-5310-02 Analytics and Statistical Programming,

Instructor:

Dr. Deepika Gopukumar,
Assistant Professor
Department of Health and Clinical Outcomes Research

Date:
05/02/2022

BACKGROUND

This study analyses whether there is an impact of each factor such as annual income, marital status, and gender on depression in adults or not using the 2013-2014 data of mental health – depression screener data file from NHANES data sets. This study has significance as it is important to study the effect of poverty on the mood and mental health of a person, the role of marriage in being happy and healthy, and gender involvement in mental peace.

The dependent variable in the data set is feeling down/depressed/hopeless (DPQ020), as these are measured or tested in this research. The demographic data which are age (RIAGENDR), gender (RIDAGEYR), marital status (DMDMARTL), and household income (INDHHIN2) are the independent variables in the dataset as they might influence the results.

This study can help us understand depression in a much better way. Also, dealing with depression becomes easier if we get to know the impact of factors causing depression. Additionally, we can use this sample data as an indicator for understanding the mental-health status of the rest of the adult population for future references. Since the project examines and highlights the relationship between poverty, marital life, gender, and mental health symptoms, it underlies wider health inequalities.

There are many studies that have been conducted to know the trends of depression, the extent of burden of depression, and the prevalence of mood disorders among the US adult population [1, 2, 3, 4, 10]. Also, there were some studies conducted to know the correlation between depression and age among adolescent girls with different tools [5]. However, this study checks the impact of various socio-economic factors on the mental health in adults of the US population, and the exploratory data analysis for a better understanding of the problem using R studio.

Studies like these help us in understanding and handling mental health problems like depression. Health care workers may find new ways to treat such disorders as they can get a better view of factors affecting or causing the disorder. Moreover, there is a chance of inventing some new mental health status detecting tests with the help of AI and a proper understanding of symptoms and causes of a disorder.

METHODOLOGY

Software and Dataset Used

The software used in this study is RStudio, and the Mental Health - Depression Screener dataset is downloaded from the NHANES official website as an XPT file. It is converted into the CSV file using RStudio [7].

Data Cleaning

The data will be filtered by age, and the adults (aged above 20 years) initially, and then the independent variables, age (RIDAGEYR), gender (RIAGENDR), marital status (DMDMARTL), household income (INDHHIN2), and the outcome variable i.e., feeling down/depressed/hopeless (DPQ020) will be selected. The variables will be cleaned by removing NA values and recoding the variables.

The gender (RIAGENDR) variable will be coded as `1` for ‘Male’ and `2` for ‘Female’ with no other gender categories included. So, there is no further need to recode. The missing values are dropped using the ‘drop_na()’ function.

The age variable (RIDAGEYR), the adult population (the population who are aged above 20 years) will be filtered. The age will be converted into intervals using the split function as 20-29, 30-39, 40-49, 50-

59, 60-69, 70-79, and >=80. The codebook indicates there are no missing values, so the ages 0 to 19 will be recoded as 'NA' and dropped. This makes us take only the adult population into consideration [13].

The marital status (DMDMARTL) variable is already coded from 1 to 6, 77, and 99 where 77 is the "Refused" and 99 is "Don't Know", which will be recoded to 'NA' and dropped along with 4406 missing values in the data [13].

The annual household income variable (INDHHIN2) is numerical data. It has the values spread across different numbers ranging from 0 to 99 as per the income of the respondents. So, we will make the values that are not appropriate to the 'NA' character. This is done using the recode factor function. We recode the values of `77` and `99` to NA and then drop those values. This completes the data cleaning process and makes the variable ready to analyze with no missing values. Likewise, the dependent variable (DPQ020) will be cleaned to take out the missing values and the factors that are not considered in the analysis will be recoded to 'NA'. This is done by re-coding the values of `0` and `3` as 'Yes' and 'No' respectively and all others to 'NA'.

Descriptive Statistics

Descriptive statistics for the independent variables namely, age (RIAGENDR), gender (RIDAGEYR), marital status (DMDMARTL), household income (INDHHIN2) and the outcome variable i.e., feeling down/depressed/hopeless (DPQ020) are planned to be calculated, after performing some visualizations such as creating histograms for every variable to check for skewness and kurtosis. If the variables are normal, then the mean and variance will be calculated, as the mean value would be a good representation of the middle of the data. Otherwise, the median and interquartile range will be calculated for the variables using a non-normal function [7]. All these statistics will be produced as a table using CreateTableOne or KableExtra packages.

Data Visualization

The visualizations namely, Scatterplots, and Boxplots are planned to be plotted. Scatterplots might reveal some important correlations between the independent variables namely, age (RIAGENDR), gender (RIDAGEYR), marital status (DMDMARTL), household income (INDHHIN2) and the outcome variable i.e., feeling down/depressed/hopeless (DPQ020). Also, Boxplots would reveal the outliers, if any, and the central tendency measures in the independent variables. This will be useful in getting some important correlations and distributions of the data.

Statistical Tests

Two statistical tests, namely, Chi-square test and t-test are planned to compute for the independent variables (RIAGENDR, RIDAGEYR, DMDMARTL, INDHHIN2), and the outcome variable (DPQ020) for the research.

Chi-squared Test:

The chi-square test of independence evaluates whether there is an association between the categories of the two variables [11]. Therefore, the Nominal (Categorical variable) that will be considered is Marital Status (DMDMARTL), and the outcome Variable is feeling down/ depressed/ hopeless (DPQ020). The relationship between marital status and depression will be analyzed through the chi-square test. Before performing the test, the assumptions, namely the variables must be nominal or ordinal, the expected values should be 5 or higher in at least 80% of groups, if the expected values are not greater than 5, then Cochran's Q-test can be considered, and the independence of observations will be checked. Then effect sizes are computed and interpreted to understand the strength of a significant chi-squared relationship by performing Cramer's V test, the formula is V = square root($\chi 2/(n(k-1))$). If the chi-squared statistic is large, then observed values are different from the expected values which in turn suggests a relationship between variables. If a warning such as "Chi-squared approximation may be incorrect"

appears, it means that the smallest expected frequency is lower than 5, then Fisher's exact test will be performed [12].

t-Test:

After EDA (Exploratory Data Analysis) if the data is found to be normally distributed, we continue with the t-test. If not, we transform the data to square root, cube root, log, and inverse and take the normally distributed transformation. As the data has one dependent and one independent variable which is categorical and continuous, we plan to do an independent sample t-test or two-sample t-test. This test is done to compare the means of two independent groups to determine whether there is statistical evidence that the associated population means are significantly different [9]. If the t-test fails, then the alternate tests which in this case will be the Mann Whitney U test or the Wilcoxon Signed Rank Test will be performed [6]. The only assumptions for carrying out a Mann-Whitney test are that the two groups must be independent and that the dependent variable is ordinal or numerical. For t-tests, the effect size statistic is Cohen's d. It will be computed when the test results are statistically significant and can be computed for each type of t-test using a slightly different formula.

Regression Analysis

Binary Logistic Regression:

As the outcome variable is binary after the removal of other groups from the variable during data cleaning, the binary logistic regression will be conducted. The binary logistic regression predicts the probability that a person is in one of the categories of the outcome variable (DPQ020) [6]. Firstly, the model will be predicted separately by taking the outcome variable and every individual variable and at last, the model will be predicted by taking all the independent variables and the outcome variable. We will compare the models and decide which model is good for predicting depression. The formula for the logistic regression is $\sigma(t)=1/1+(e^{(-t)})$, and after substituting the regression model for t we will get $p(y)=1/1+e^{(-(\beta 0+\beta 1x1+...))}$. Here, p(y) is the probability of y given t which is the sum of y- intercept ($\beta 0$) and the product of all the predictors and their coefficients. It can also be predicted if any of the independent variables are useful for prediction or not [8]. The odds ratios, model significance, and model fit will be calculated along with the R-squared to compute the percent correctly predicted by the model using the predicted probabilities, or fitted values, for each of the observations and comparing these probabilities to the true value of the outcome. If the R-squared value is less than 0.5, then the outcome would be 0 or no depression, else the outcome would be 1 [6]. Apart from these, the sensitivity and specificity of the models will be interpreted to know whether the model is better at predicting people with the outcome or people without the outcome. In the end, the assumptions of the binary logistic regression model will be analyzed. The assumptions include the independence of observation, and linearity which would be observed by graphing the log-odds of the outcome against each continuous predictor to see if the relationship is linear, also there should be no perfect multicollinearity which is checked using the generalized variance inflation factor (GVIF). The multicollinearity assumption will fail if the model has GVIF values greater than 2.5, else the assumption will pass. After this, model diagnostics will be performed to check for any observations that are having an unusual impact on the model. Firstly, the standard residuals will be calculated to check for any outliers, DF betas and cooks' distance are used to find influential values. Finally, all the models will be compared using the likelihood ratio test [6].

REFERENCES

[1] Cao, C., Hu, L., Xu, T., Liu, Q., Koyanagi, A., Yang, L., ... & Smith, L. (2020). Prevalence, correlates, and misperception of depression symptoms in the United States, NHANES 2015–2018. Journal of affective disorders, 269, 51-57.

[2] Yu, B., Zhang, X., Wang, C., Sun, M., Jin, L., & Liu, X. (2020). Trends in depression among Adults in the United States, NHANES 2005–2016. Journal of Affective Disorders, 263, 609-620.

[3] Brooks, J. M., Titus, A. J., Bruce, M. L., Orzechowski, N. M., Mackenzie, T. A., Bartels, S. J., & Batsis, J. A. (2018). Depression and handgrip strength among US adults aged 60 years and older from NHANES 2011–2014. *The journal of nutrition, health & aging*, *22*(8), 938-943.

[4] García-Velázquez, R., Jokela, M., & Rosenström, T. H. (2019). "The varying burden of depressive symptoms across adulthood: results from six NHANES cohorts." ("The varying burden of depressive symptoms across adulthood ...") Journal *of affective disorders*, *246*, 290-299.

[5] Shen, Y., Varma, D. S., Zheng, Y., Boc, J., & Hu, H. (2019). Age at menarche and depression: results from the NHANES 2005–2016. *PeerJ*, *7*, e7150.

[6] Harris, J. K. (2020). Statistics With R: Solving Problems Using Real-World Data (1st ed.). SAGE Publications, Inc.

[7] NHANES 2013–2014 Questionnaire Data. (2016). CDC. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2013

[8] Wikipedia contributors. (2022, May 1). Logistic regression. Wikipedia. https://en.wikipedia.org/wiki/Logistic_regression

[9] *Solved Part I: Independent Samples T-tests: This test ....* (n.d.). Retrieved from https://www.chegg.com/homework-help/questions-and-answers/part-independent-samples-t-tests-test-compares-means-two-independent-groups-determine-whet-q52452092

[10] The varying burden of depressive symptoms across adulthood .... (n.d.). Retrieved from https://pubmed.ncbi.nlm.nih.gov/30594042/

[11] What is a Chi-Square Test and Why Do We use it? | SURESH, A. (n.d.). Codementor. Www.codementor.io. https://www.codementor.io/@abhirajsuresh/what-is-a-chi-square-test-and-why-do-we-use-it-1365snyolr

[12] -Square Test of Independence in R./ Stats and R, Antoine Soetewey, 27 Jan. 2020, https://statsandr.com/blog/chi-square-test-of-independence-in-r/.

[13] NHANES 2013-2014 Demographic data. CDC. https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/DEMO_H.htm