Bekzat Alish, Ishwarya Ravikumar, Ramnath Kamakoti

Shawn Curley

MSBA 6120 Introduction to Statistics for Data Scientists

Project Report

08.19.2017

Alcohol Study

1. Introduction

A. Background

According to the U.S. Department of Health and Human Services about 18 million of adult Americans suffer from alcohol dependence. Excessive alcohol drinking cost the United States around $249 billion in 2010 due to drinker’s loss in productivity, health care expenses, criminal activities, and driving accidents (CDC).

A number of statistical studies explored the relationship between alcohol use and other various variables. Individuals with lower level of education were found to more likely develop alcohol abuse than individuals with higher level of education (Crum, Helzer, & Anthony, 1993). Age at first use of alcohol was shown to powerfully predict long-term alcohol abuse. Increasing ages of first alcohol trials were associated with decreasing levels of alcohol dependence (Grant, & Dawson, 1997). Increasing levels of income were shown to have a positive relationship with alcohol usage (Keyes, & Hasin, 2008).

There have been lots of studies on alcohol dependence, but the biggest limitations of most of those studies were the small sample sizes and an inability to generalize the results to general population. The purpose of this study is to extend the body of research on alcohol use by examining the significance of the relationship between different socio-demographic, alcohol related factors and high level of alcohol consumption using a huge dataset sampled from all over the US. The hypothesis is that there is a relationship between those factors and an increased consumption of alcohol.

B. Dataset

This study uses data from the National Survey on Drug Use and Health collected in 2012 by the U.S. Department Health and Human Services. The survey data was collected through a multistage area probability sample for each of the 50 states and the District of Columbia. 55,268 people participated in the survey, and the data was collected on 3,120 variables, including drug consumption variables related to alcohol, tobacco, marijuana, cocaine and etc.

Given the size of the dataset, researchers who collected the data made multiple adjustments to it in order to make the dataset suitable for statistical research purposes. A portion of survey responses did not correspond to a valid interview response or were incomplete. Such responses were imputed into “bad data”, “don’t know”, and “refused to answer” variable levels which were treated as missing data in the analysis in this study. Some sociodemographic variables missed data for some observations, so they were imputed using an imputation procedure called predictive mean neighborhood (PMN). Certain variables were recoded from other variables. For example, age category variable is recoded into broader age category variables with more number of levels.

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| Variable # | Variable Description | Range | Type |
| 1 | Number of days had 5+ drinks past 30 days | 0 − 30 | interval |
| 2 | Number of days had 4+ drinks in the same occasion past 30 days | 0 − 30 | interval |
| 3 | Number of days had a drink past 30 days | 0 − 30 | interval |
| 4 | Number of drinks past 30 days | 0 − 90 | interval |
| 5 | How many hours worked last week | 1 − 60 | interval |
| 6 | Number of days skipped work past 30 days | 0 − 30 | interval |
| 7 | Gender | 0 − 1 | nominal |
| 8 | Age category | 0 − 3 | ordinal |
| 9 | Education level | 0 − 4 | nominal |
| 10 | Family income | 0 − 3 | ordinal |

We are predicting one interval variable, which is the number of days a person had 5 or more drinks in the past 30 days before taking the survey. The predictor variables include variables of all three types – interval, nominal, and ordinal, that range from quantities and number of days of alcohol consumption to sociodemographic characteristics such as income levels and education level achieved. For modeling purposes, we treated the variables the following way.

If a survey respondent “did not use alcohol in the past 30 days”, we assigned a value of zero for all alcohol consumption variables that include variables 1-4. If survey respondents never tried alcohol in their lives, we removed them from the dataset due to the introduction of a potential bias into the statistical analysis. For all variables, we treated responses such as “don’t know”, “refused to answer”, and etc. as missing data and removed those observations from the dataset.

As it was discussed before, sociodemographic variables in the dataset in this study do not have missing data due to imputation procedures. Out of the four variables used in our modeling process, only gender variable was not imputed. Age category variable was recoded from an imputed age variable by the researchers who collected the dataset. They also recoded education level and family income level variables by decreasing the amount of levels in the original imputed variables.

2. Analyses

3. Conclusion

References

Center for Disease Control and Prevention (CDC). (n.d.). *Excessive Drinking is Draining the U.S. Economy*. Retrieved from https://www.cdc.gov/features/costsofdrinking/

Crum, R., Helzer, J., & Anthony, J. (1993). *Level of education and alcohol abuse and dependence in adulthood: A further inquiry*. Retrieved from https://jhu.pure.elsevier.com/en/publications/level-of-education-and-alcohol-abuse-and-dependence-in-adulthood--3

Grant, B., & Dawson, D. (1997). *Age at Onset of Alcohol Use  
and Its Association with DSM-IV  
Alcohol Abuse and Dependence:  
Results from the National Longitudinal  
Alcohol Epidemiologic Survey*. Retrieved from http://ac.els-cdn.com/S0899328997900092/1-s2.0-S0899328997900092-main.pdf?\_tid=fd17a766-822b-11e7-93ac-00000aacb35f&acdnat=1502851290\_1e21e657873138969bf481778d600697

Keyes, K., & Hasin, D. (May 20, 2008). *Socio-economic status and problem alcohol use: the positive relationship between income and the DSM-IV alcohol abuse diagnosis*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1360-0443.2008.02218.x/full

U.S. Department of Health and Human Services. (n.d.). *Alcoholism and Alcohol Abuse: MedLine Plus*. Retrieved from https://medlineplus.gov/alcoholismandalcoholabuse.html

U.S. Department of Health and Human Services. (October 30, 2013). *National Survey on Drug Use and  
Health, 2012*. Retrieved from http://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/34933

Appendix

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| Variable # | Variable Description | Range | Type |
| 1 | Number of days had 5+ drinks past 30 days | 0 − 30 | interval |
| 2 | Number of days had 4+ drinks in the same occasion past 30 days | 0 − 30 | interval |
| 3 | Number of days had a drink past 30 days | 0 − 30 | interval |
| 4 | Number of drinks past 30 days | 0 − 90 | interval |
| 5 | How many hours worked last week | 1 − 60 | interval |
| 6 | Number of days skipped work past 30 days | 0 − 30 | interval |
| 7 | Gendera | 0 − 1 | nominal |
| 8 | Age categoryb | 0 − 3 | ordinal |
| 9 | Education levelc | 0 − 4 | nominal |
| 10 | Family incomed | 0 − 3 | ordinal |

aGender: 0 = *male*, 1 = *female*. bAge: 0 = *12-17 years old*, 1 = *18-25 years old*, 2 = *26-34 years old*, 3 = *35 or older*.cEducation: 0 = *Less than high school*, 1 = *High school graduate*, 2 = *Some college*, 3 = *College graduate*, 4 = *12 to 17 years old*. dFamily income: 0 = *less than $20,000*, 1 = *$20,000 - $49,999*, 2 = *$50,000 - $74,999*, 3 = *$75,000 or more*.