Logistic Regression Analyisis of Binge Drinking

Kimberly Bond

December 04, 2025

Contents

Summary

Identifying factors that encourage or discourage heavy binge drinking could lead to a better understanding of who is most susceptible to frequent binge drinking, which is often a significant indicator of alcoholism. The particular factors of interest for their possible impact on frequent binge drinking include biological sex, age, general health, mental health, education, income, employment status, vaping use, tobacco and nicotine use, and marijuana use. By analyzing the Substance Abuse and Mental Health Services Administration's (SAMHSA) National Survey on Drug Use and Health (NSDUH) [Citation 1], a data set containing 47,273 observations, a survey-weighted logistic regression model and a survey-weighted cross-validated lasso logistic regression model were fit. These models found that heavy alcohol use was most strongly associated with co-occurring substance use. Tobacco users had over three times the odds of heavy drinking, and marijuana users had more than twice the odds, compared to non-users. Females and individuals in the "other" work-status category had significantly lower odds of heavy drinking, as did those reporting poorer overall health. Income, education, vaping, and past-year major depression were not significant predictors after adjustment. Overall, substance-use behaviors were the dominant factors associated with heavy alcohol consumption in the model.

Introduction

Finding factors related to substance abuse has long been a key area of interest, and the study of which is now much more feasible thanks to extensive surveys and datasets like SAMHSA's NSDUH. In their research, *Logistic Regression Model of Demographic Predictors and Confounders of Binge Alcohol Use Among Adults with Major Depression* [Citation 2], Areen Omari focuses on identifying demographic factors that increase the likelihood of alcohol abuse among adults with a lifelong history of major depressive episodes. Their analysis of the NSDUH dataset found that age and marital status significantly predicted binge drinking among individuals with major depressive episodes. Additionally, they found that adults under the age of 50, with a college degree, never married, divorced/separated, and with a high-middle income level or higher, were at higher risk for binge alcohol use. While this information is incredibly useful for identifying individuals who may be susceptible to frequent binge drinking and thus alcohol abuse based on demographics, further information could be gained regarding adults who participate often in binge drinking and their patterns of use with other substances and other variables. Thus, the main factors of interest in this analysis will be biological sex, age, general health, mental health, education, income, employment status, vaping use, tobacco and nicotine use, and marijuana use.

From initial analysis, 5.03% of the adult survey respondents participated in heavy binge drinking in the past month (2,950 adults). Age groups 18–25 and 35–49 have much higher binge drinking occurrence in the past month compared to age groups 26–34 and 50 or more years old. Additionally, 55.9% of binge drinkers were male versus female. In terms of other substance use, 35% of binge drinkers also vaped, 60.7% used any type of tobacco or nicotine, and 43.9% used marijuana in the past month. For health, 14.9% of binge drinkers had a major depressive episode in the past year, and the majority of heavy binge drinkers qualified themselves as very good or good health. Lastly, 44.6% of those who participated in heavy binge drinking had an annual income greater than $75,000 per year, and 58.6% had full-time employment. Key patterns are shown in the exploratory analysis plots (Figure~1A–C). This study aims to gain further insight into these observations by exploring the predictive capabilities of the described predictors and examining the association between increased probability of heavy binge drinking and their presence.

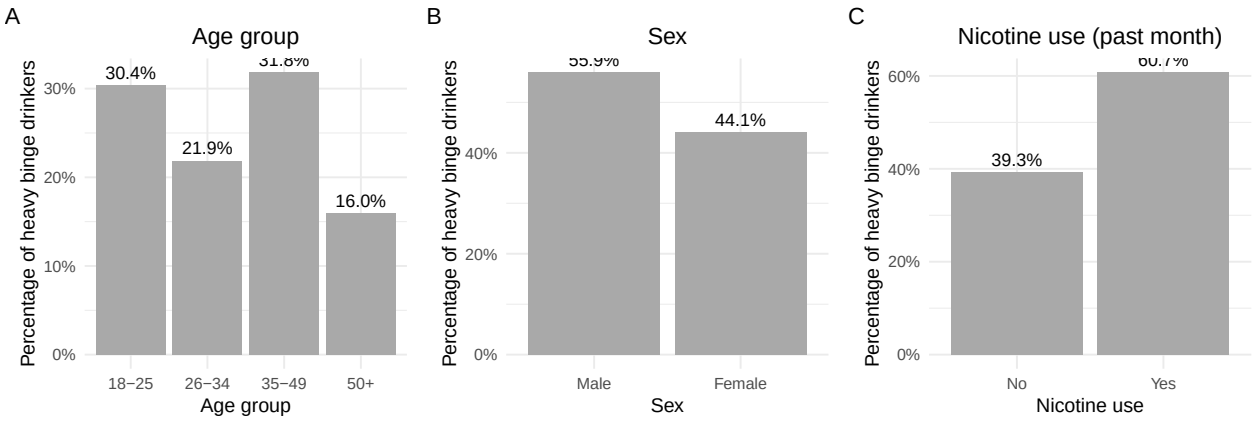

Figure 1: Proportion of heavy binge drinkers by (A) age group, (B) sex, and (C) nicotine use in the past month.

Methods

Data Source and Description

The analytic sample was drawn from the National Survey on Drug Use and Health (NSDUH) and included demographic, socioeconomic, behavioral, and mental-health indicators relevant to substance-use risk. Age was categorized into five groups: 12–17 (n = 11,334; 19.33%), 18–25 (n = 14,136; 24.11%), 26–34 (n = 9,621; 16.41%), 35–49 (n = 12,845; 21.91%), and 50+ (n = 10,697; 18.24%). Note that individuals under 18 were excluded from the analysis to be consistent with the focus on adults. Sex was recorded using the IRSEX classification. Self-reported overall health ranged from excellent to poor: Excellent (12,453; 21.24%), Very Good (21,501; 36.67%), Good (17,813; 30.38%), Fair (5,902; 10.07%), and Poor (930; 1.59%). Educational attainment (IREDUHIGHST2) spanned from less than fifth grade to college graduate or higher, with the largest groups being high-school graduates or GED holders (13,004; 22.18%), those with some college but no degree (9,379; 16.00%), and college graduates or higher (15,963; 27.23%). Annual household income was grouped into four levels: Less than $20,000 (9,443; 16.11%), $20,000–49,999 (14,956; 25.51%), $50,000–74,999 (8,829; 15.06%), and $75,000 or more (25,405; 43.33%). Employment status among respondents aged 18+ included full-time employment (22,901; 39.06%), part-time employment (7,159; 12.21%), unemployment (3,108; 5.30%), and other/non-labor-force status (14,131; 24.10%), while youth aged 12–17 were coded as missing.

Behavioral risk variables captured past-month substance use. Nicotine vaping (NICVAPMON) was reported by 7,542 individuals (12.86%), while 51,091 (87.14%) reported no past-month use. Tobacco or nicotine use from any source (TOBVNICMON) occurred in 13,247 individuals (22.59%), compared with 45,386 non-users (77.41%). Marijuana use in the past month (MRJMON) was endorsed by 9,955 respondents (16.98%), with 48,678 (83.02%) reporting no use. Mental-health indicators included the presence of a major depressive episode in the past year, collected separately for adults (AMDEYR) and youths (YMDEYR). After converting ineligible age groups to NA, 5,003 adults (8.53%) and 1,790 youths (3.05%) met criteria for a depressive episode, while 40,656 adults (69.34%) and 9,127 youths (15.57%) reported none. The primary outcome variable, heavy alcohol use in the past 30 days (HVYDRKMON), was relatively rare, with 2,950 respondents (5.03%) classified as heavy drinkers compared to 55,683 (94.97%) who reported no heavy use.

Table 1: Chi-Square Tests of Association Between Predictors and Heavy Binge Drinking

variable	statistic	p.value	sig
sex	150.549582	0.0000000	***
age_cat	88.335067	0.0000000	***
health	75.079514	0.0000000	***
income	2.983686	0.3941473	
work_status	156.426428	0.0000000	***
vape_use	1072.304071	0.0000000	***
tob_use	1897.497521	0.0000000	***
mj_use	1176.488933	0.0000000	***
mde_adult	47.426285	0.0000000	***
educ	33.555701	0.0002196	***

Exploratory Data Analysis

Following the initial identification of the analytic variables of interest from the NSDUH codebook, preliminary data cleaning was conducted, including removal of limited missing values and the recoding of ordered factors. Exploratory data analysis was performed to characterize the distributions of the selected predictors and their bivariate relationships with heavy binge drinking in the past month. Data cleaning, EDA function creation, and general debugging of all code was assisted by GenerativeAI [Citation 3]. These descriptive findings are summarized in the introduction and supported by the visualizations presented therein.

To formally examine the association between each predictor and the binary outcome, a series of bivariate statistical tests were performed. Pearson Chi-Square Tests of Association were used for categorical variables. Additionally, Cochran–Armitage trend tests were used to assess linear trends across ordered categories.

The Chi-Square tests indicated statistically significant associations (p-value < 0.01) between heavy binge drinking and several predictors, including sex, age group, overall health, employment status, past-month vaping, past-month tobacco or nicotine use, past-month marijuana use, and past-year major depressive episode status. Trend tests provided evidence of monotonic trends only for age (p-value $= 0.003944$) and health (p-value < 0.0001), whereas education and income did not exhibit significant linear trends. A summary of the Chi-Square test results is presented in Table~1.

Modeling

Model Selection

With a better understanding of the relationships between heavy binge drinking and the predictor variables, a survey-weighted logistic regression model was fit, including all selected predictors and incorporating the survey design variables VEREP (variance primary sampling unit) and VESTR (variance stratum) as recommended for the NSDUH dataset and any large, complex survey data set [Citation 4]. This survey-weighted model adjusts the variance of the estimators such that $Var(\hat{\beta}) = DesignEffect * StandardError^2$. This full model indicated strong overall evidence of association between the predictors and heavy binge drinking, with particularly strong effects for sex, tobacco use, and marijuana use (see Table~2). Using `svyglm` with a quasibinomial family, ordinal predictors

Table 2: Survey-weighted logistic regression coefficients for heavy binge drinking in the past 30 days.

Term	Estimate	Std. Error	p-value	Sig
sexFemale	-0.3305384	0.0647517	< 0.0001	***
age_cat.L	0.1165640	0.0746584	0.127202	
age_cat.Q	-0.0315353	0.0825747	0.704781	
age_cat.C	-0.1380733	0.0586758	0.024197	*
health.L	-0.4070422	0.1723856	0.023747	*
health.Q	-0.3683876	0.1382320	0.011451	*
health.C	-0.0104089	0.0950187	0.913378	
health^4	0.0192877	0.0819152	0.815187	
work_statusPart-time	-0.1630230	0.1259576	0.203814	
work_statusUnemployed	-0.1526994	0.1027850	0.146085	
work_statusOther	-0.3783447	0.1054770	0.000986	***
vape_useYes	0.0364907	0.1046964	0.729466	
tob_useYes	1.1715826	0.0963311	< 0.0001	***
mj_useYes	0.8344553	0.0749304	< 0.0001	***

were modeled using orthogonal polynomial contrasts (linear, quadratic, cubic, and higher-order components). These higher-order terms were used primarily to assess predictive performance; for purposes of inference and interpretation, the focus is on the linear components.

Automated model selection processes, such as step-wise AIC/BIC model selection, could not be used for this model because a lack of assumptions met (weights and stratification are included in the survey-weighted model), so ANOVA was used to reduce the model. An initial ANOVA on the full model showed that MDE_ADULT, INCOME. and EDUC did not add anything to the model. By running subsequent Likelihood Ratio Tests on a reduced model removing each of the variables, it was found each of variables for major depressive episode in the past year, income, and education did not add anything to the model (LRT P-Value for all tests > 0.3), and thus were removed to reduce complexity. The final model coefficients are summarized in Table~2, and the corresponding model specification can be seen as (not including the survey weighting):

$$\begin{aligned}\operatorname{logit}\big(P(Y=1 \mid X)\big) = \beta_0 &+ \beta_1 \,\text{sex}_{\text{Female}} \\ &+ \beta_2 \,\text{age\_cat.L} + \beta_3 \,\text{age\_cat.Q} + \beta_4 \,\text{age\_cat.C} \\ &+ \beta_5 \,\text{health.L} + \beta_6 \,\text{health.Q} + \beta_7 \,\text{health.C} + \beta_8 \,\text{health}^4 \\ &+ \beta_9 \,\text{work}_{\text{Part-time}} + \beta_{10} \,\text{work}_{\text{Unemployed}} + \beta_{11} \,\text{work}_{\text{Other}} \\ &+ \beta_{12} \,\text{vape\_use}_{\text{Yes}} + \beta_{13} \,\text{tob\_use}_{\text{Yes}} + \beta_{14} \,\text{mj\_use}_{\text{Yes}} + \epsilon\end{aligned}$$

$$\text{where } \operatorname{logit}\big(P(Y=1 \mid X)\big) = \log\left(\frac{P(Y=1 \mid X)}{1 - P(Y=1 \mid X)}\right)$$

and Y = 1 represents the presence of heavy binge-drinking in the past month.

Model Evaluation

In addition to the survey-weighted logistic regression used for inference and interpretation, a cross-validated, survey-weighted lasso logistic regression model was fit to address the class imbalance

Table 3: Classification performance metrics for the survey-weighted logistic regression model and the cross-validated lasso logistic regression model.

Model	Threshold	Accuracy	Sensitivity	Specificity	AUC
Survey-weighted logistic	0.047	0.692	0.717	0.691	0.747
CV-lasso logistic	0.335	0.692	0.691	0.716	0.749

between heavy binge drinkers and non–heavy drinkers and to evaluate potential improvements in predictive performance. Under the original survey-weighted model, a Youden-optimal threshold for classifying heavy binge drinking was estimated at 0.05297912, whereas the corresponding threshold for the cross-validated lasso model was 0.3522646. Receiver operating characteristic (ROC) curves were constructed for both models, and each achieved an area under the curve (AUC) of approximately 0.75 (see Figure~2).

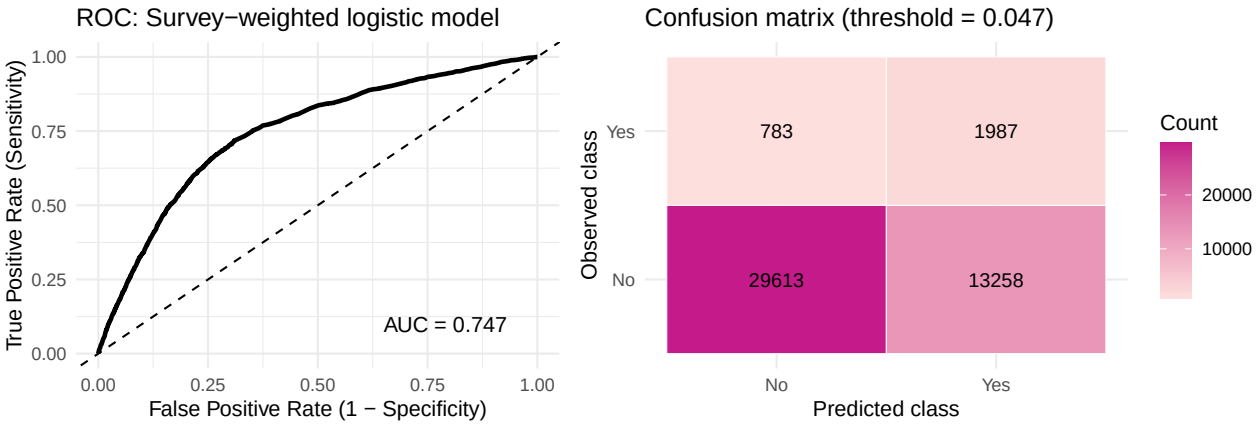

Figure 2: ROC curve (left) and confusion matrix (right) for the survey-weighted logistic regression model of heavy binge drinking.

Thsi cross-validated weighted lasso produced extremely similar AUC, accuracy, sensitivity, and specificity (Table~3). Coefficients from the two models were broadly consistent in direction, with similar patterns of positive and negative associations across predictors. Given the desire to preserve valid inference under the complex survey design and the performance simlarities between the two models, subsequent analyses and interpretation focus on the survey-weighted logistic regression model rather than the cross-validated lasso model.

Because of the rather low prediction accuracy of the model, likely due to the small portion of the study that was identified as having participated in heavy binge drinking, although the models show strong integrity for the interpretation of coefficients to identify trends and patterns, it would not be recommended to use these models in a predictive sense for any one individual.

Results

Key Significant Predictors

Sex. Females had significantly lower odds of heavy alcohol use compared to males (odds ratio [OR] = 0.72, 95% confidence interval [CI]: 0.63–0.82, $p < 0.001$). This indicates that, holding other

factors constant, women are substantially less likely than men to engage in heavy drinking.

Health status. There was a significant negative linear trend across self-reported health categories ($p = 0.0237$). Higher values of the health variable correspond to poorer perceived health, and the OR less than 1 (OR = 0.67, 95% CI: 0.47–0.97) suggests that individuals reporting worse health had decreased odds of heavy alcohol use.

Work status. Participants in the "Other'' employment category showed significantly lower odds of heavy alcohol use relative to full-time workers (OR = 0.68, 95% CI: 0.55–0.85, $p < 0.001$). Neither part-time employment nor unemployment was a significant predictor in the adjusted model.

Substance use behaviors

Tobacco or nicotine use and marijuana use emerged as the strongest predictors of heavy alcohol consumption. Tobacco use was associated with more than a threefold increase in the odds of heavy drinking (OR = 3.23, 95% CI: 2.68–3.92, $p < 0.001$). Marijuana use was also strongly associated with higher odds (OR = 2.30, 95% CI: 1.98–2.68, $p < 0.001$). In contrast, vaping was not significantly associated with heavy alcohol use ($p = 0.73$).

Non-significant Predictors

Age did not show a significant linear trend, although the point estimate suggested a slight positive association with heavy drinking (OR = 1.11, $p = 0.127$), but it did show a significant non-linear trend, specifically a cubic trend, suggesting a cubic relationship between heavy binge drinking and age.

Overall, the results indicate that co-occurring substance use (tobacco and marijuana) is the strongest correlate of heavy drinking in this population. Sex, health status, and nonstandard work arrangements also play meaningful roles. Socioeconomic indicators (income, education, and employment status excluding the "Other'' category) did not show significant associations after controlling for other variables.

Conclusion

This study examined demographic, socioeconomic, health, and substance-use–related predictors of heavy alcohol consumption using a multivariable logistic regression model. The findings highlight that co-occurring substance-use behaviors, particularly tobacco use and marijuana use, are the strongest correlates of heavy drinking, even after adjusting for demographic and socioeconomic factors. Females, individuals reporting poorer health, and those in nonstandard work-status categories demonstrated significantly lower odds of heavy alcohol use. Other factors, including income, education, age, depression history, and vaping, were not significant predictors in the adjusted model.

Overall, the model suggests that heavy alcohol consumption is most closely intertwined with other substance-use patterns, reinforcing the importance of integrated prevention efforts that consider clusters of risky behaviors rather than treating alcohol use in isolation.

Table 4: Adjusted odds ratios for heavy binge drinking from the survey-weighted logistic regression model.

Predictor	Odds ratio	95% CI (lower)	95% CI (upper)	p-value	Sig
Sex: Female	0.72	0.63	0.82	< 0.0001	***
Age Category (linear trend)	1.12	0.97	1.31	0.127202	
Age Category (quadratic trend)	0.97	0.82	1.15	0.704781	
Age Category (cubic trend)	0.87	0.77	0.98	0.024197	*
Health (linear trend)	0.67	0.47	0.94	0.023747	*
Health (quadratic trend)	0.69	0.52	0.92	0.011451	*
Work status: Part-time	0.85	0.66	1.10	0.203814	
Work status: Unemployed	0.86	0.70	1.06	0.146085	
Work status: Other	0.68	0.55	0.85	0.000986	***
Vape use (yes)	1.04	0.84	1.28	0.729466	
Tobacco/nicotine use (yes)	3.23	2.65	3.92	< 0.0001	***
Marijuana use (yes)	2.30	1.98	2.68	< 0.0001	***

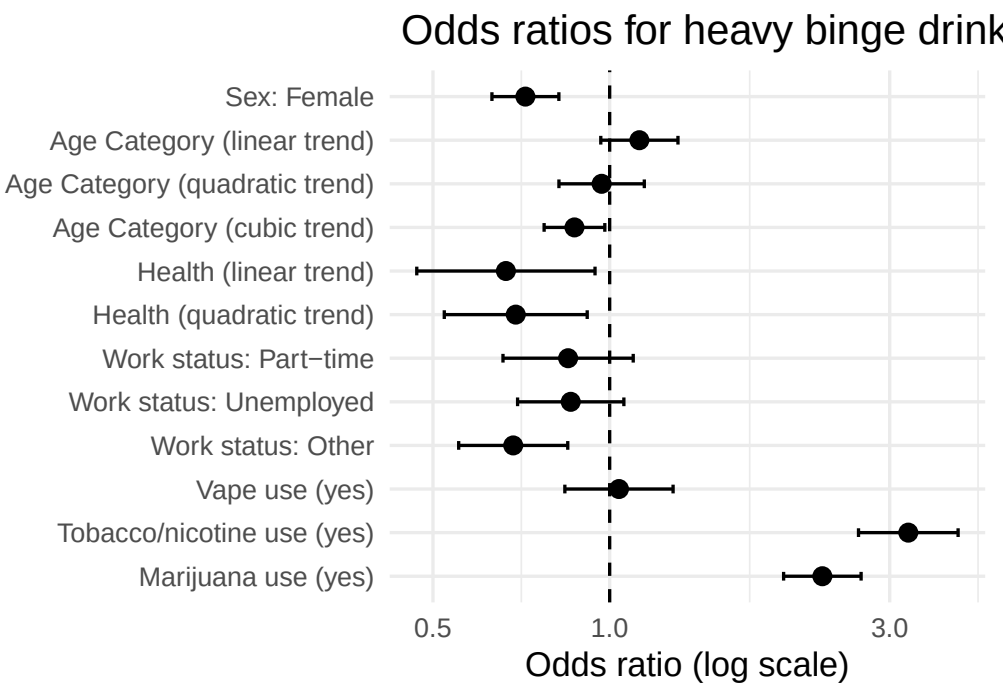

Figure 3: Forest plot of adjusted odds ratios and 95% confidence intervals for heavy binge drinking from the survey-weighted logistic regression model.

Future Directions and Limitations

Although this analysis provides meaningful insight into factors associated with heavy alcohol use, several limitations should be acknowledged. The cross-sectional design prevents conclusions about causality, and unmeasured confounders—such as stress, social environment, or family history—may influence both substance use and drinking behaviors. In addition, some predictors were modeled only as linear trends, which may not fully capture nonlinear or threshold effects. Future studies should explore interaction effects, alternative model structures (e.g., nonlinear or penalized approaches), and longitudinal datasets to better understand temporal relationships between co-occurring substance use and heavy drinking. Incorporating richer psychosocial variables and validating the model in independent populations would further strengthen the generalizability and depth of these findings.

References

- [1] U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (2018). National Survey on Drug Use and Health 2016 (NSDUH-2016-DS0001). Retrieved from https://www.samhsa.gov/data/
- [2] Omary A. Logistic Regression Model of Demographic Predictors and Confounders of Binge Alcohol Use Among Adults with Major Depression. Int J Ment Health Addict. 2022 Apr 28:1-15. doi: 10.1007/s11469-022-00808-y. Epub ahead of print. PMID: 35502437; PMCID: PMC9047467.
- [3] Generative AI: As stated, ChatGPT helped with creating data cleaning and EDA functions, as well as just generally troubleshooting all code. Grammarly was used throughout the writing process to grammar check and ensure quality writing. When compiling the final r-markdown file, ChatGPT was used to ensure everything compiled correctly and all figures were referenced correctly. It appears that it changed some wording in the markdown paragraph areas to make it "better" while it was adjusting the formatting, but the entire message is the same as my original work and only wording/style was changed, not content.
- [4] Dey D, Haque MS, Islam MM, Aishi UI, Shammy SS, Mayen MSA, Noor STA, Uddin MJ. The proper application of logistic regression model in complex survey data: a systematic review. BMC Med Res Methodol. 2025 Jan 22;25(1):15. doi: 10.1186/s12874-024-02454-5. PMID: 39844030; PMCID: PMC11752662.