Taylor Keating
June 2021

Analysis of Self-Reported Depression and Levels of Alcohol Use in Young Adults

Abstract:

Alcohol use remains commonplace in social settings and is an easily accessible method for coping with problems. For many young adults, alcohol use gets introduced in university, right as they are developing their own identities and methods for dealing with stress. The goal of this study was to explore the association between self-reported levels of depression and alcohol use in young adults. We used a cross-sectional sample from the National Longitudinal Survey of Youth (1997), selecting the year 2006. The responses to questions on depression and alcohol use for the 7460 participants were analyzed using ordinal logistic regression models with self-reported categorical depression as the dependent variable and levels of drinking as the predictors of interest. We found a weak positive association between self-reported feelings of depression and the number of days of drinking in the past month in American young adults of the same age, sex, and income level (OR=1.083 of feeling more depressed per 5 day difference in the number of days of drinking last month, 95% CI of 1.035 to 1.131). As well, we found a stronger positive association between self-reported feelings of depression and the number of days of drinking at least 5 drinks in the past month in American young adults of the same age, sex, and income level (OR= 1.217 of feeling more depressed per 5 day difference in the number of days of drinking at least 5 drinks last month, 95% CI of 1.131 to 1.307).

Introduction:

Background:

Many people deal with stress in their lives, and drinking alcohol is a common method for coping. This is because alcohol is fairly inexpensive, very easy to access, and is present in many social settings. It is also an important public health concern because alcohol use often begins in university, right at the time when many young adults are becoming independent and figuring out their identities. If the first method of stress and anxiety management that young adults are taught is alcohol use, they are not being set up for success. Alcoholism is known to lead to conditions such as high blood pressure, stroke, heart disease, liver cirrhosis, pancreatitis, and many more. This study will focus on the relationship between alcohol use and mental health.

Objectives:

The overall objective of this study is to explore the relationship between alcohol use and mental health. Specifically, the study will look into the relationship between alcohol use and depression in young adults. The following two scientific questions will be addressed: Among 21-26 year old Americans, what is the association between the number of days of alcohol drinking in the past month and the odds of feeling more depressed in the past month, adjusting for age, sex, and yearly income. Among 21-26 year old Americans, what is the association between the number of days of drinking at least 5 drinks in the past month and the odds of feeling more depressed in the past month, adjusting for age, sex, and yearly income.

Taylor Keating
June 2021

Methods:

Study Design:

This study will look at a cross-sectional sample of one year from a longitudinal cohort survey study. The original study is called the National Longitudinal Survey of Youth (1997)[1] in which a sample of Americans were interviewed at least every other year from 1997 until present day (2021) and asked about many topics including education, employment, relationships, health, and substance use. This study will be considering the sample of participants who completed the survey interview in 2006.

To attempt to answer the objectives, several questions relating to alcohol use and depression were identified. The questions selected pertaining to alcohol use are: Have you drank alcohol in the last year? In the last 30 days, on how many days did you drink an alcoholic beverage and on how many days did you drink at least 5 alcoholic beverages? In the last 30 days, on the days you drank alcohol, how many drinks did you usually have? The question selected pertaining to depression was: How much of the time during the last month have you felt so down in the dumps that nothing could cheer you up? The possible answers to this question are: All of the time, most of the time, some of the time, and none of the time.

Setting:

The surveys were conducted using a computer-assisted personal interview and attempted to be completed in person with an interviewer. In circumstances in which the interviewee was not able to meet in person, the interview was conducted over the phone.

Participants:

Study participants were obtained by using a cross-sectional sampling of U.S. residents between the ages of 12 and 17 years old in 1997, as well as an oversampling of Hispanic or Latino and Black individuals. The cross-sectional sample comprised of 6,748 participants and the oversampling comprised of 2,236 participants for a total of 8,984. During the 2006 survey interview, the cohort was between the ages of 21 and 26 years old. 1425 individuals were not able to be contacted for the 2006 survey, so only 7559 observations were obtained. As well, 99 of these respondents were prisoners in an insecure environment at the time (2006), so they were excluded from the study analysis. Therefore, the total number of observations from the 2006 survey was 7460.

Variables:

Next we will describe adjustments made pertaining to the alcohol use questions. We first looked at the responses to whether the subjects had drank an alcoholic beverage in the last year. We saw that 5639 said yes, 1756 said no, and 65 did not answer. The 1756 subjects that said no were then not asked about the number of days they drank in the past 30 days and the number of days they drank at least 5 drinks in the past 30 days (because they had not drank at all in the past year). Therefore, we assigned these 1756 subjects the value 0 for the number of days they drank in the past 30 days and the number of days they drank at least 5 drinks in the past 30 days.

	Variable	Details	Possible Answers
Outcome	Depression	How much of the time during the last month have you felt so down in the dumps that nothing could cheer you up?	• All of the time • Most of the time • Some of the time • None of the time
Primary Predictor	Days of Drinking per month	In the last 30 days, on how many days did you drink an alcoholic beverage?	Numerical: 0-30
Secondary Predictor	Days of Heavy Drinking per month	In the last 30 days, on how many days did you drink at least 5 alcoholic beverages?	Numerical: 0-30
Other Covariates	Age	Age of participant in 2006 (in years)	22-26 years
	Sex	Sex of participant	Male, Female
	Yearly Income (USD)	During 2005, how much income did you receive from wages, salary, commissions, or tips from all jobs, before deductions for taxes or for anything else?	Numerical: >=0

It should be noted that the income variable is very positively skewed, as the age of participants in the 2006 survey were between the ages 22 and 26 years. Therefore, many participants will have an annual income of $0 USD, but a few participants will have incomes that are higher than $50,000 USD. Therefore, the income variable will be log-transformed (log(income + 1)).

Statistical Methods:

Primary Analysis:

Scientific question is: Among 21-26 year old Americans, what is the association between the number of days of alcohol drinking in the past month and the odds of feeling more depressed in the past month, adjusting for age, sex, and yearly income.

This question will be answered by fitting the following ordinal logistic regression[2] model to the data, with the categorical self-reported depression in the last month as the outcome, the number of days of alcohol drinking in the last month as the predictor of interest, and adjusting for covariates age, sex, and income (log-scaled).

$$\log\big(Odds(depressed \leq j)\big) = \beta_{j0} + \beta_1(days\_drank) + \beta_2(age) + \beta_3(sex) + \beta_4 log(income + 1)$$

- $depressed$ = self reported depression in last month
- $\text{j} = \begin{cases} \text{1 if none of the time} \\ \text{2 if some of the time} \\ \text{3 if most of the time} \\ \text{4 if all of the time} \end{cases}$
- $days\_drank$ = number of days of drinking in the last month

- *age*= age of participant in 2006 (in years)
- *sex*= 1 if participant is male, 0 if female
- *income*= annual income from previous year in USD

I will test the hypothesis that comparing groups that differ in the number of days of drinking in the last month by 1 day, the odds ratio of being "more depressed" (i.e. at least some of the time vs none of the time, at least most of the time vs less than most of the time, and at least all of the time vs less than all of the time) is equal to 1, after adjusting for age, sex, and income. This is equivalent to testing that $\beta_1 = 0$.

Due to the proportional odds assumption, the odds ratio is constant across all categories of depression.

Secondary Analysis:

Scientific question is: Among 21-26 year old Americans, what is the association between the number of days of drinking at least 5 drinks in the past month and the odds of feeling more depressed in the past month, adjusting for age, sex, and yearly income.

This question will also be answered by fitting an ordinal logistic regression model to the data as in the primary analysis, but now using the number of days of drinking at least 5 drinks in the last month as the predictor of interest.

I will test the hypothesis that comparing groups that differ in the number of days of drinking at least 5 drinks in the last month by 1 day, the odds ratio of being "more depressed" (i.e. at least some of the time vs none of the time, at least most of the time vs less than most of the time, and at least all of the time vs less than all of the time) is equal to 1, after adjusting for age, sex, and income.

Multiple-Comparisons Correction:

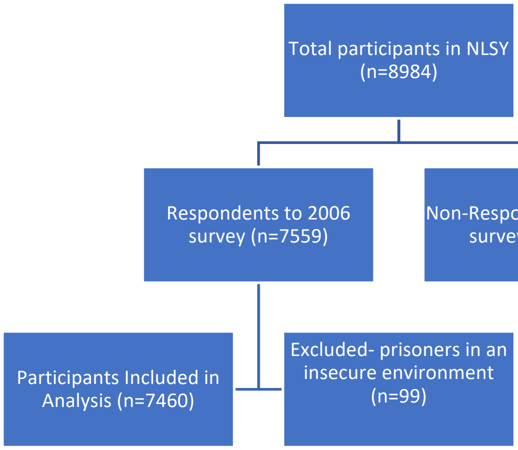
Since I will be testing two separate hypotheses, I will use the Bonferroni correction and test each of the hypotheses at the 0.025 level in order to obtain an overall 0.05 significance level.

Missing Data:

Only complete cases will be used in the analysis. The observations with missing data in the outcome variable of depression will be compared against that from the complete cases in order to see if there might be any bias in the observations that are missing data.

Results:

Participants:

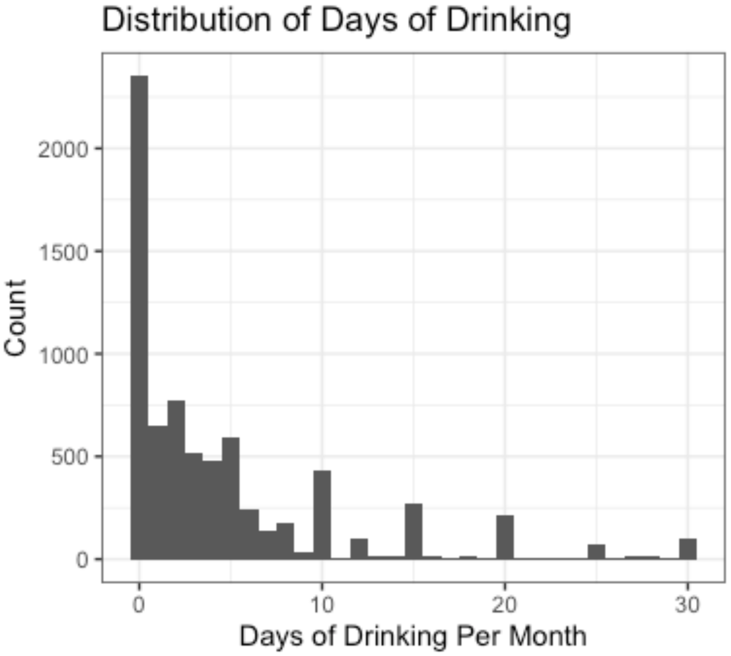

Total participants in NLSY (n=8984)
Respondents to 2006 survey (n=7559)
Non-Respondents to 2006 survey (n=1425)
Participants Included in Analysis (n=7460)
Excluded- prisoners in an insecure environment (n=99)

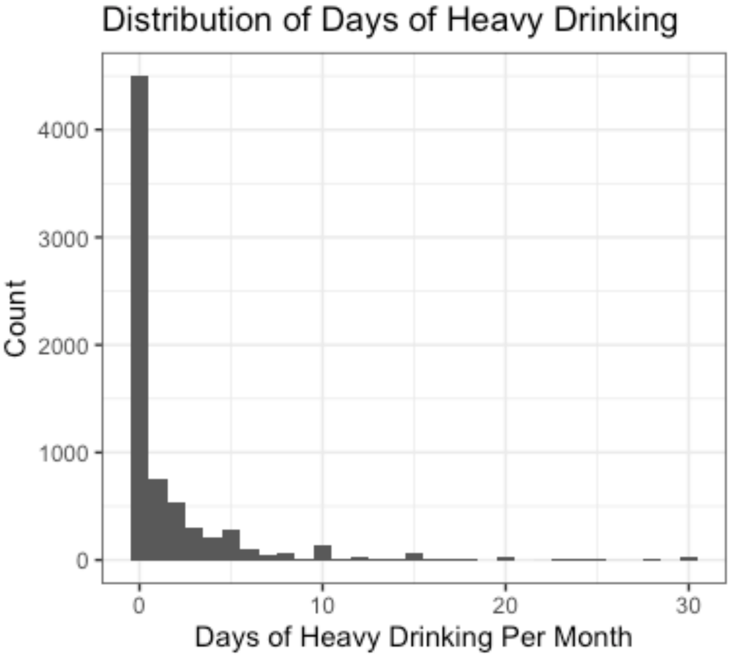
The remaining 7460 participants were included in the analysis. However, participants had the option to answer “Refusal” or “Don’t Know” on any of the questions given. These two answers will be treated as NA’s for the purposes of the analysis. Please see sections *Descriptive Data* and *Outcome Data* to see the summaries and the number of NA’s from each of the survey questions used.

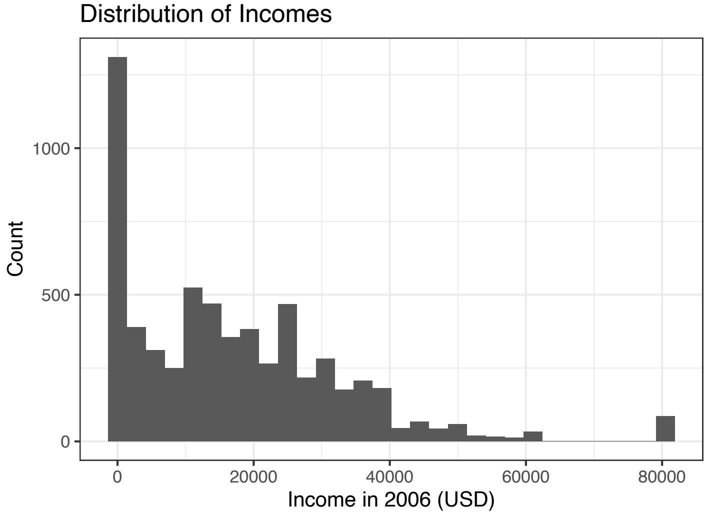
Descriptive Data (Study Participant Characteristics):

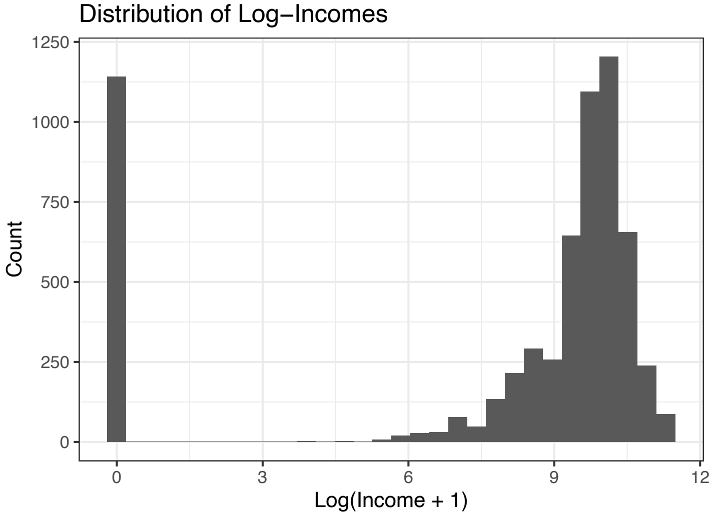
Table 1- Summary of Covariates

	Variable				
	Age	**Male Sex**	**Income**	**Days Drinking**	**Days Heavy Drinking**
Number of Observations	7460	7460	6177	7292	7181
Number Missing	0	0	1283	168	279
Mean	23.95	0.5	16978	4.72	1.68
Std Dev	1.39	0.5	15686	6.43	3.78
Min	22	0	0	0	0
Median	24	0	15000	2	0
Max	26	1	80471	30	30









From **Table 1**, we see that there are 1283 missing data from income in 2006, 168 missing data from the number of days drinking in the past 30 days, and 279 missing data from the number of days drinking at least 5 drinks in the past 30 days. We notice that participants were all between the ages of 22 and 26, with an even distribution through this range. As well, half of the participants are male. We notice that both of the responses to the drinking questions are skewed, with most of the participants answering between 0 and 10 days per month for both questions, but with answers all the way up to 30 days per month. Looking at the income, we see that the median is $15,000 and the max is $80,471. Since this income variable is skewed greatly, we have decided to include it in the model with a log-transform (log(income + 1)).

Outcome Data:

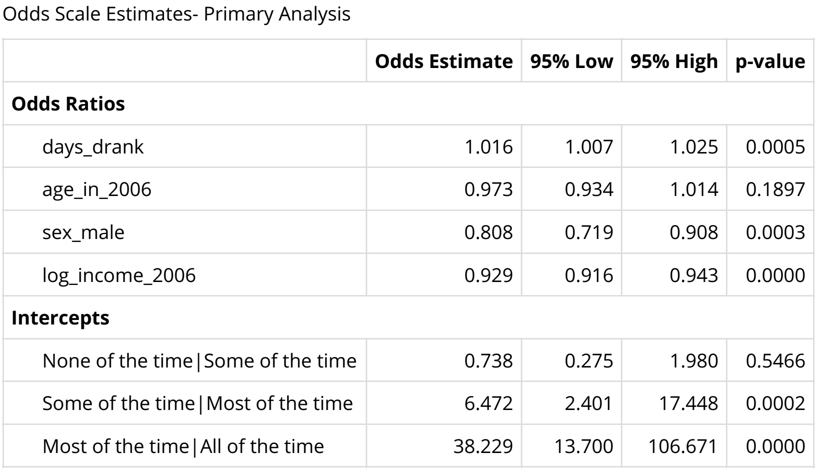
Table 2- Summaries of Drinking for Each Depression Category

Depression	Number Obs	Mean Days Drinking per Month	StdDev Days Drinking per Month	Mean Days Heavy Drinking per Month	StdDev Days Heavy Drinking per Month
None of the time	5199	4.69	6.26	1.59	3.63
Some of the time	1748	4.95	6.72	1.92	4.06
Most of the time	269	5.08	8.06	2.34	5.24
All of the time	60	3.12	6.48	1.14	2.60
Total	7276	4.76	6.45	1.69	3.81
NA	184	2.85	5.18	0.76	2.07

There are 184 (of 7460) missing data from the question on depression in 2006. The majority of participants answered to being depressed "None of the time" (5199/7460) in the past month, and the count of answers decreases in order of the categories "Some of the time" (1748/7460), "Most of the time" (269/7460), and "All of the time" (60/7460). We notice a general trend of higher mean days drinking and mean days heavy drinking per month for groups of participants that answered to with higher categories of depression. However, participants who answered "All of the time" have the smallest mean days drinking and mean days heavy drinking per month, breaking this general trend.

Main Results:

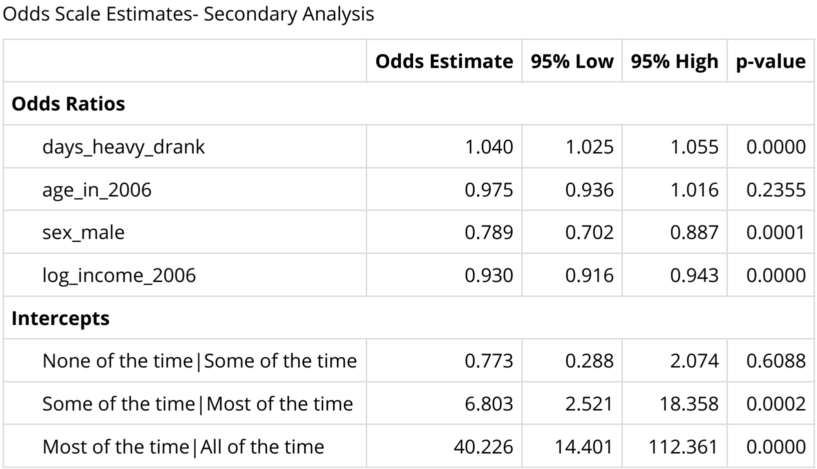
We fit an ordinal logistic regression model to the data with days of drinking per month as the primary predictor, as described in Statistical Methods. From the primary analysis, we estimate that when comparing groups of the same age, sex, and income but that differ in the number of days of drinking in the last month by 1 day, the group that drinks more will have 1.016 times the odds of feeling "more depressed" in the last month (95% CI of 1.007, 1.025). We reject the null hypothesis at the 0.025 level that this odds ratio is 1 (p=0.0005). This per-day odds ratio corresponds to an odds ratio of 1.083 for groups differing by 5 days of drinking in the last month (95% CI of 1.035, 1.131). Below are the estimated odds ratios for the entire model.



Odds Scale Estimates- Primary Analysis

	Odds Estimate	95% Low	95% High	p-value
Odds Ratios				
days_drank	1.016	1.007	1.025	0.0005
age_in_2006	0.973	0.934	1.014	0.1897
sex_male	0.808	0.719	0.908	0.0003
log_income_2006	0.929	0.916	0.943	0.0000
Intercepts				
None of the time|Some of the time	0.738	0.275	1.980	0.5466
Some of the time|Most of the time	6.472	2.401	17.448	0.0002
Most of the time|All of the time	38.229	13.700	106.671	0.0000

We fit an ordinal logistic regression model to the data with days of heavy drinking per month as the primary predictor, as described in Statistical Methods. From the secondary analysis, we estimate that when comparing groups of the same age, sex, and income but that differ in the number of days of heavy drinking in the last month by 1 day, the group that drinks more will have 1.040 times the odds of feeling "more depressed" in the last month (95% CI of 1.025, 1.055). We reject the null hypothesis at the 0.025 level that this odds ratio is 1 ($p<0.0001$). This per-day odds ratio corresponds to an odds ratio of 1.217 for groups differing by 5 days of heavy drinking in the last month (95% CI of 1.131, 1.307). Below are the estimated odds ratios for the entire model.



Odds Scale Estimates- Secondary Analysis

	Odds Estimate	95% Low	95% High	p-value
Odds Ratios				
days_heavy_drank	1.040	1.025	1.055	0.0000
age_in_2006	0.975	0.936	1.016	0.2355
sex_male	0.789	0.702	0.887	0.0001
log_income_2006	0.930	0.916	0.943	0.0000
Intercepts				
None of the time|Some of the time	0.773	0.288	2.074	0.6088
Some of the time|Most of the time	6.803	2.521	18.358	0.0002
Most of the time|All of the time	40.226	14.401	112.361	0.0000

Note that for both of the above analyses, feeling "more depressed" is feeling depressed: at least some of the time vs none of the time, at least most of the time vs less than most of the time, and at least all of the time vs less than all of the time.

Discussion:

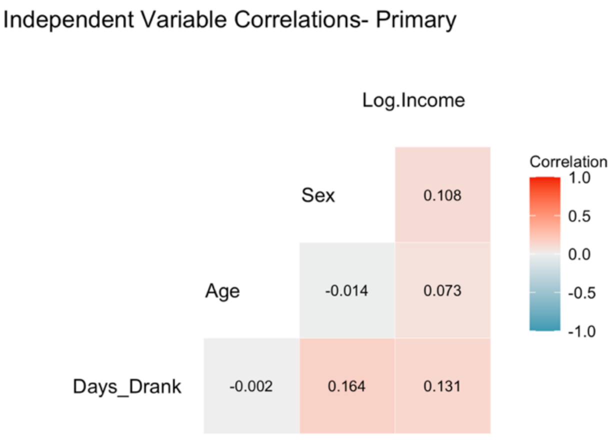
Assumptions[6]:

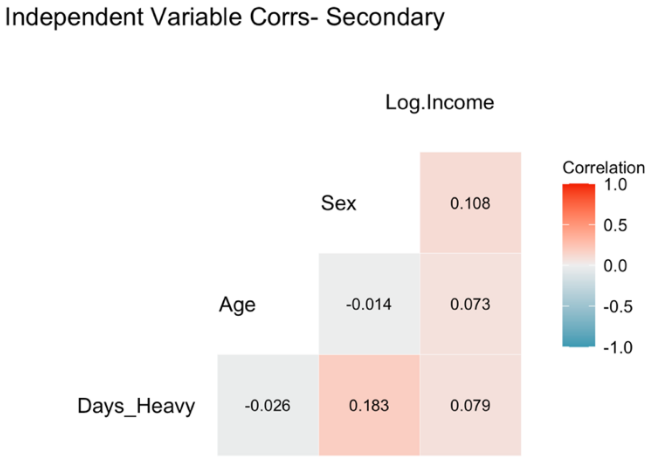
Independent Observations: Each observation is a participant's survey responses from the National Longitudinal Survey of Youth (1997). Specifically, only the survey responses from the 2006 survey are taken. The survey aimed to take a random sample of American youth between the ages of 12 and 17 in 1997, and we are only considering one year of survey responses. These observations would generally be assumed to be independent, however it should be noted that in the survey design they did allow for multiple participants from the same household to be surveyed. This could affect the independence of observations, but could not be tested.

Dependent Variable is Ordered: This assumption is not violated, as the dependent variable is self-reported depression in the last month, broken down into the categories "None of the Time", "Some of the Time", "Most of the Time", and "All of the Time." Clearly these are ordered categories.

Independent Variables are Continuous or Categorical: The independent variables in these analyses were age, sex, income, days of drinking in the last month, and days of heavy drinking in the last month. Age, income, and the drinking variables are all continuous while sex is binary (categorical). Therefore, this assumption is not violated.

No Multi-Collinearity: The plots below looks at the correlations between each of the independent variables in the primary and secondary analyses. We see that none of the independent variables appear to be highly correlated with each other. The highest correlation between any two independent variables is between sex and days of drinking in the last month, but the correlation is only 0.164 in the primary and 0.183 in the secondary analysis. Therefore, the assumption of no multi-collinearity does not seem to be violated. As well, we present the variance inflation factors (VIF) for both the primary and secondary analyses. In both models, the VIF for the covariates are just over 1, indicating very little multicollinearity.





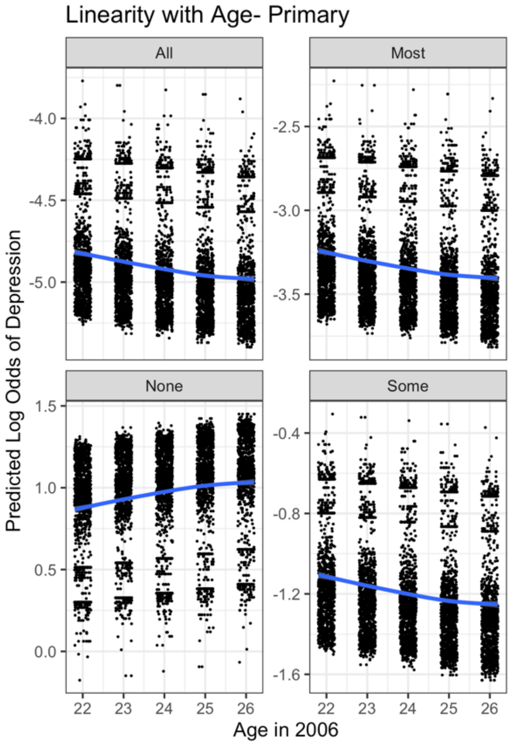
VIF- Primary Analysis

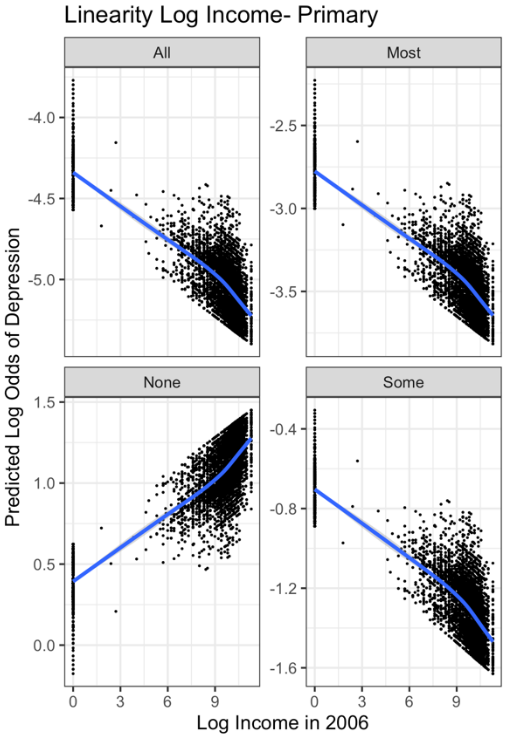
	VIF
days_drank	1.0424
age_in_2006	1.0053
sex_male	1.0379
log_income_2006	1.0304

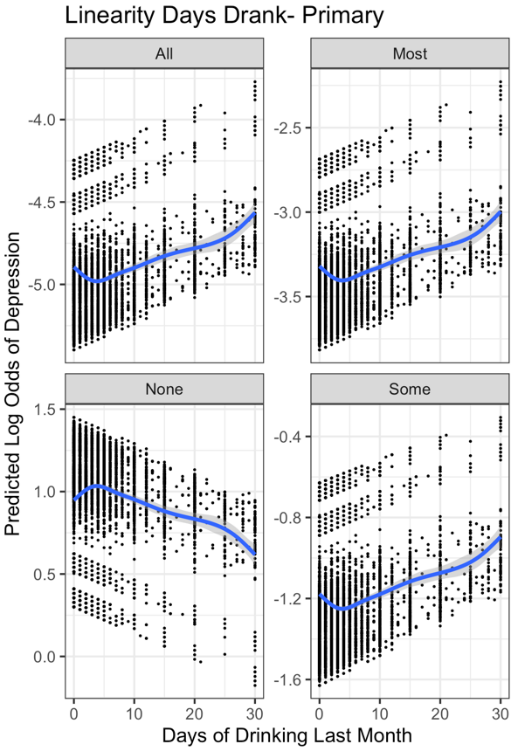
VIF- Secondary Analysis

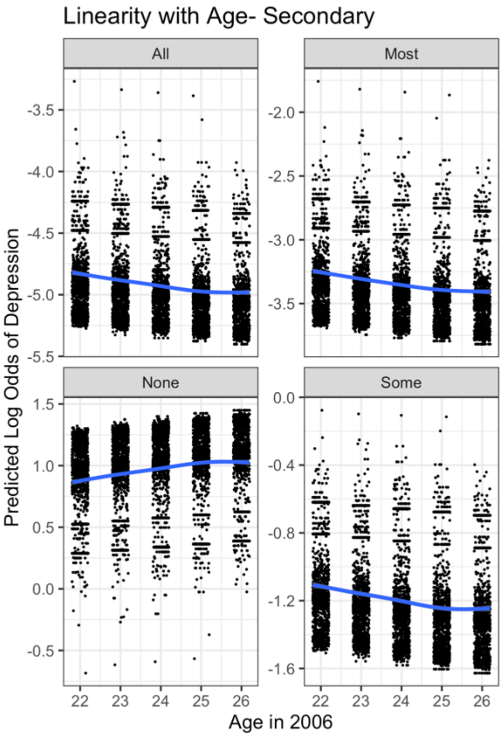
	VIF
days_heavy_drank	1.0409
age_in_2006	1.0062
sex_male	1.0458
log_income_2006	1.0214

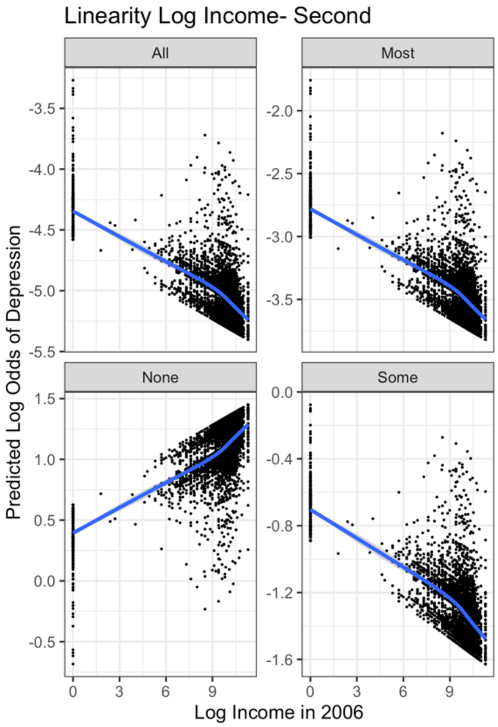
Linearity of Log-Odds: Ordinal logistic regression models assume linearity of independent variable and log odds. We have plotted the fitted log-odds of depression against the continuous independent variables of age in 2006, log income in 2006, and days of drinking / days of heavy drinking for both the primary and secondary analyses. We see that the fitted log odds and the continuous independent variables appear to have a fairly linear relationship in both the primary and secondary analyses. Therefore, it is fair to assume that this assumption is not violated.

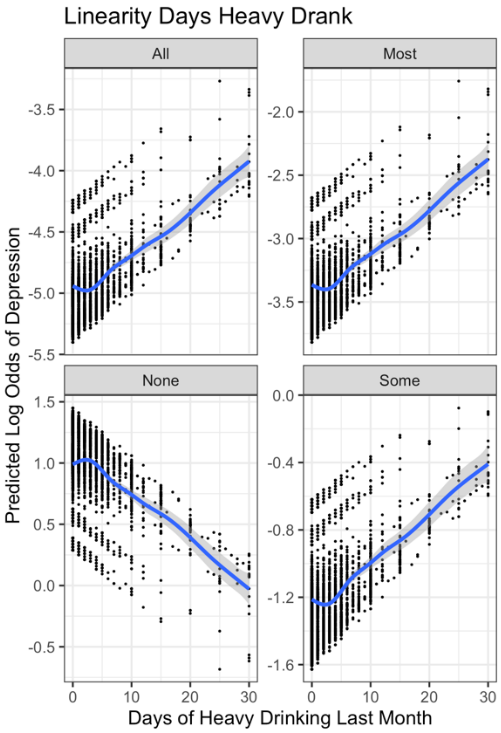




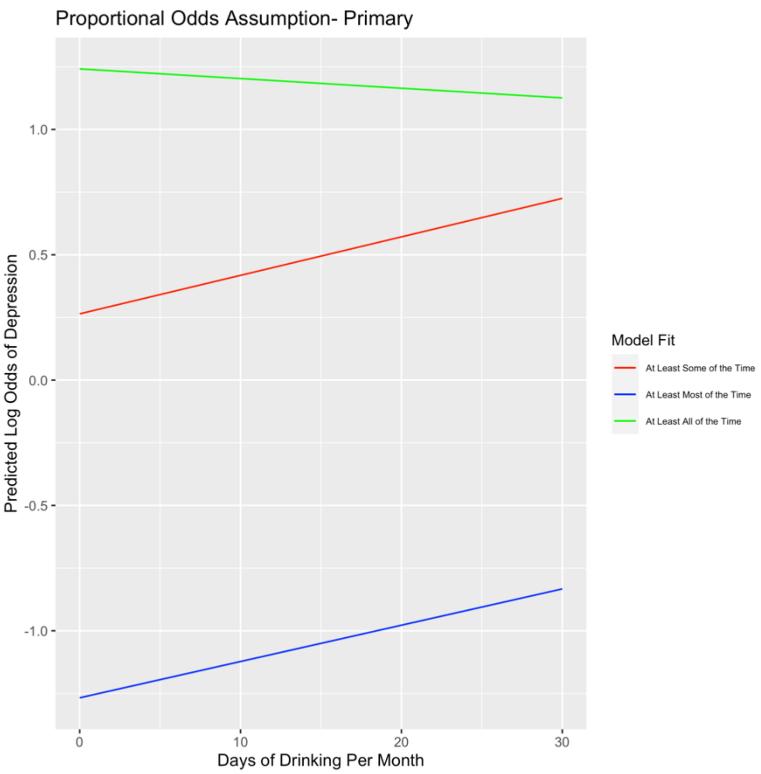


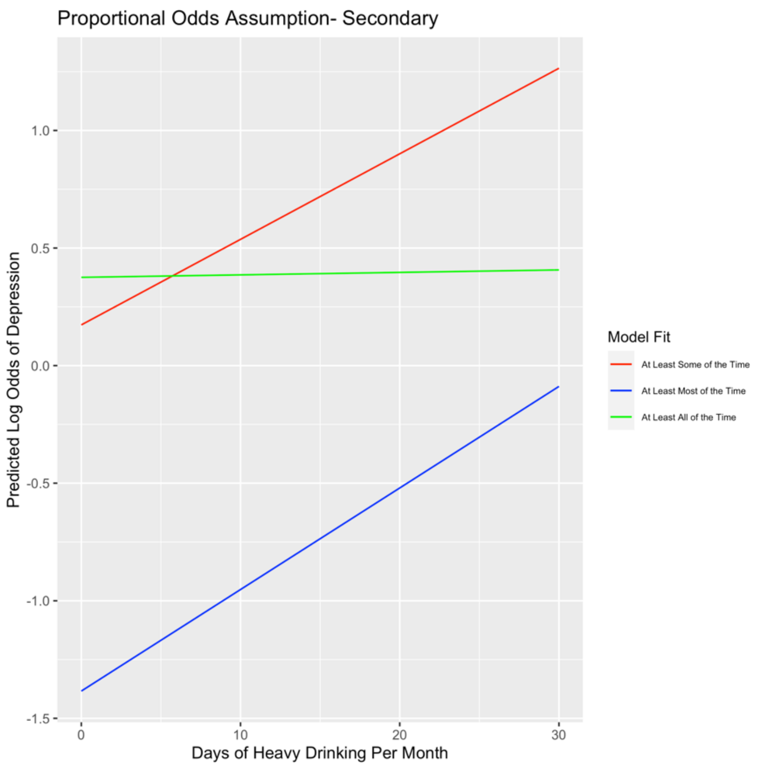






Proportional Odds Assumption: One of the key assumptions underlying the ordinal logisitic regression models fit is the proportional odds assumption. It assumes that comparing groups that differ in the number of days of drinking in the last month by 1 (or differ in the number of days of heavy drinking in the last month by 1), our estimated odds ratio is constant across all categories of depression. In order to test this assumption, we first performed a Brant test[5]. In the primary analysis, we reject the null hyptothesis that the proportional odds assumption holds ($p<0.001$). In the secondary analysis, we reject the null hypothesis that the proportional odds assumption holds ($p<0.001$). In order to further visualize the violation of this assumption, we have fit binary logistic regression models at each of the depression category cutpoints for both the primary and secondary analyses. Below are plots showing the predicted log odds of depression from the models against the days of drinking per month (for the primary analysis) and against the days of heavy drinking per month (for the secondary analysis). It appears that in both analyses, the binary logistic regression model fit using depression greater than or equal to "All of the time" as the cutpoint produces a slope estimate that is inconsistent with the models from the other two cutpoints. These plots indicate that the major violation in the proportional odds assumption is due to observations that had depression "All of the Time" in the last month. We argue that since only 60 out of 7460 participants had depression "All of the Time" in the last month, we should not entirely discount the estimates produced from the ordinal logisitic regression models. Both the primary and secondary analyses provide meaningful insight into the general relationship between alcohol use and depression.





Large Sample Size: In this analysis, our sample size is 7460 participants. As well, we have 4 independent variables in both the primary and secondary analyses. We note that the least frequent outcome of our dependent variable is feeling depressed “All of the Time” in the last month, which was observed 60 times. Since we have greater than 10 observations of this least frequent outcome per independent variable included, our sample size should be large enough for valid inference using these ordinal logistic regression models.

Limitations:

Missing data bias: Missing data in the outcome depression question could lead to bias in the estimated association if there was a systematic difference between the participants who were included in the analysis and those who did not answer the question on depression. It is very possible that on average, the participants who chose not to answer the question on depression had higher levels of depression in the last month than participants who did answer the question. From **Table 2**, we see that there were 184 (2.5%) participants who did not select a category in the question on depression. We notice that these participants had a mean number of days drinking per month of 2.85 and a mean number of days heavy drinking per month of 0.76. The participants that did answer the depression question had a mean number of days drinking per month of 4.76 and a mean number of days heavy drinking per month of 1.69. These differences are a concern, and could make our estimated association between depression and drinking falsely high if the participants who chose not to answer in reality had higher depression levels as well as these lower levels of drinking. This concern is somewhat eased by the fact that only 2.5% of the participants chose not to respond for the outcome variable.

Time of Year of questionnaire response: The time of year when participants were surveyed could be an unmeasured confounder in this analysis. This is because the time of year could impact the drinking levels as well as the feelings of depression for participants. Ideally we would adjust for the time of year when participants answered the survey, however this information was not recorded.

Selection bias: The survey attempted to obtain a representative sample from the target population. However, bias could still arise in selecting participants for the survey, as well as in the respondents who were able to be contacted for the survey each year (specifically for 2006 in this analysis). The National Longitudinal Survey of Youth (1997) performed stratified multistage area probability sampling, which attempted to obtain a representative sample of the target population defined by race, income, and region.

Self-report bias: Self-report bias including social desirability bias[3] could have an impact on the results from this study. Due to the self-reporting nature of the questionnaire, it is very likely that the levels of alcohol use[3a] as well as the levels of depression[3b] are underreported due to social desirability bias. Since both of these variables could be underreported, it is not clear how the estimate of the association between depression and alcohol use would be affected. It is also not possible to account for this bias in the analyses.

Recall bias: Recall bias[4] can be a major limitation in survey studies. However, the survey questions chosen to be used in the analysis eliminate some of the recall bias by asking participants to answers question about information “in the last 30 days”. This fairly short and recent time period for questions surrounding alcohol use and depression limit the recall bias somewhat.

Key Results:

From this study, we found an overall weak positive association between self-reported feelings of depression in the past month and the number of days of drinking in the past month in American young adults between the ages of 22 to 26, after adjusting for age, sex, and income. We also found a slightly stronger positive association between self-reported feelings of depression in the past month and the number of days of drinking at least 5 drinks in the past month in American young adults between the ages of 22 to 26, after adjusting for age, sex, and income. It should be noted that there were some limitations in the analysis including potential missing data bias, as well as selection bias and social desirability bias through the survey design. As well, we found that the proportional odds assumption used in the ordinal logistic regression models was violated, especially due to the observations with the highest levels of depression. The study still provides meaningful insights into the overall positive associations between self-reported feelings of depression and levels of drinking in the past month in American young adults, while keeping in mind that this association may not be generalizable to individuals with higher levels of depression.

Statistical Software:

All of the analyses were carried out using R version 4.0.2. The ordinal logistic regression model was fit using the polr function[2] from the MASS package in R. The Brant test concerning the proportional odds assumption was performed using the brant function[5b] from the brant package in R.

References:

1. National Longitudinal Survey of Youth (1997)-
    a. Bureau of Labor Statistics, U.S. Department of Labor. National Longitudinal Survey of Youth 1997 cohort, 1997-2017 (rounds 1-18). Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. Columbus, OH: 2019.
        i. https://www.nlsinfo.org/content/cohorts/NLSY97
2. Ordered Logistic Regression
    a. Introduction to SAS. UCLA: Statistical Consulting Group.
        i. from https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/(accessed March-June, 2021).
        ii. https://stats.idre.ucla.edu/r/faq/ologit-coefficients/
    b. Introduction to SAS. UCLA: Statistical Consulting Group.
        i. from https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/(accessed March-June, 2021).
        ii. https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/
    c. https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5
3. Self-Report/ Social Desirability biases
    a. Davis CG, Thake J, Vilhena N. Social desirability biases in self-reported alcohol consumption and harms. Addict Behav. 2010 Apr;35(4):302-11. doi: 10.1016/j.addbeh.2009.11.001. Epub 2009 Nov 10. PMID: 19932936.
        i. https://pubmed.ncbi.nlm.nih.gov/19932936/#:~:text=Results%3A%20Both%20studies%20show%20consistent,likely%20to%20report%20risky%20drinking.
    b. Hunt M, Auriemma J, Cashaw AC. Self-report bias and underreporting of depression on the BDI-II. J Pers Assess. 2003 Feb;80(1):26-30. doi: 10.1207/S15327752JPA8001_10. PMID: 12584064.
        i. https://pubmed.ncbi.nlm.nih.gov/12584064/#:~:text=One%20problem%20in%20identifying%20and,depressive%20symptoms%20more%20than%20women.
4. Recall Bias
    a. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc*. 2016;9:211-217. Published 2016 May 4. doi:10.2147/JMDH.S104807

        i. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4862344/#:~:text=of%20confirmation%20bias.-,Self%2Dreporting%20bias,questions%20without%20his%2Fher%20interference.

5. Brant Test

    a. Brant, Rollin. “Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression.” *Biometrics*, vol. 46, no. 4, 1990, pp. 1171–1178. *JSTOR*, www.jstor.org/stable/2532457. Accessed 20 May 2021.

        i. https://www.jstor.org/stable/2532457?seq=1Ordinal

    b. https://rpubs.com/rslbliss/r_logistic_ws

6. Logistic Regression Assumptions

    a. https://medium.com/evangelinelee/ordinal-logistic-regression-on-world-happiness-report-221372709095#:~:text=The%20assumptions%20of%20the%20Ordinal,No%20multi%2Dcollinearity.

    b. https://www.statisticssolutions.com/assumptions-of-logistic-regression/