Ishwarya Ravikumar

Bekzat Alish

Ramnath Kamakoti

Abstract

Prediction of days of alcohol consumption in a month based on past alcohol consumption and demographic characteristics

Prediction of Days of Alcohol Consumption in a Month

MSBA 6120 STATISTICS - GOLD COHORT - TEAM 10

Table of Contents

[Background 2](#_Toc490928327)

[Dataset 2](#_Toc490928328)

[Missing Data 2](#_Toc490928329)

[Assumptions 4](#_Toc490928330)

[Weighting by population 4](#_Toc490928331)

[Oversampling of the young-adult population 4](#_Toc490928332)

[Stability 4](#_Toc490928333)

[Scope of the study 4](#_Toc490928334)

[Model Fine Tuning 6](#_Toc490928335)

[First iteration 6](#_Toc490928336)

[Second iteration 6](#_Toc490928337)

[Third iteration 6](#_Toc490928338)

[Relationships of interest 7](#_Toc490928339)

[Final Model and its Interpretation 7](#_Toc490928340)

[How much variation of the data is being explained by the model 8](#_Toc490928341)

[Uses of the study 9](#_Toc490928342)

[Limitations of the study 9](#_Toc490928343)

[Truthfulness of respondents 9](#_Toc490928344)

[Cross Sectional data collection 9](#_Toc490928345)

[Target population of the survey 10](#_Toc490928346)

[Appendix 10](#_Toc490928347)

[Data Management 10](#_Toc490928348)

[R Code Output 11](#_Toc490928349)

[Data Manipulation and Factoring Categorical Variables 11](#_Toc490928350)

[Relationships of interest 11](#_Toc490928351)

[Checking for Categorical Variable - Age 11](#_Toc490928352)

[Checking for Categorical Variable - Sex 12](#_Toc490928353)

[Model – Initial Model 12](#_Toc490928354)

[Model – First Iteration 12](#_Toc490928355)

[Model – Second Iteration 13](#_Toc490928356)

[Model – Third Iteration 14](#_Toc490928357)

[Final model 15](#_Toc490928358)

[References 19](#_Toc490928359)

# 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).

Many 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.

# 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 etc.

## Missing Data

Given the size of the dataset, researchers who collected the data made multiple adjustments to it 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.

|  |  |  |
| --- | --- | --- |
| **Variable Description** | **Range** | **Type** |
| Number of days had 5+ drinks past 30 days | 0 − 30 | interval |
| Number of days had 4+ drinks in the same occasion past 30 days | 0 − 30 | interval |
| Number of days had a drink past 30 days | 0 − 30 | interval |
| Number of drinks past 30 days | 0 − 90 | interval |
| How many hours worked last week | 1 − 60 | interval |
| Number of days skipped work past 30 days | 0 − 30 | interval |
| Gender | 0 − 1 | nominal |
| Age category | 0 − 3 | ordinal |
| Education level | 0 − 4 | nominal |
| 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. The coding dictionary for these variables are given in the appendix.

Invalid data

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 the top four variables in Table 1. 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”, 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.

# Assumptions

## Weighting by population

The sampling process is weighted by the population of the state that it is being collected from. Within each state, sampling strata called State Sampling Regions (SSRs) were formed. These SSRs were based on their population. The eight large sample states (Top 8 states by population) were divided into 48 SSRs each; the remaining States were divided into 12 SSRs each. Therefore, the partitioning of the population of the United States formed 900 SSRs.

## Oversampling of the young-adult population

The sampling distribution was not uniform across all age groups. The survey process over sampled the young adult population to derive more insights into them. The surveys were distributed equally among three age groups (12-18, 18-25 ,26 and higher). This implies that the surveys were taken more by young adults than people above the age of 26.

## Stability

The process of collecting data on these variables is the survey that respondents took to answer these questions. Earlier, the respondents were voluntary but for 2012 onwards get paid an amount of $30. Also, the distribution of SSRs change with changing population distributions across the states. Some variables are added and removed from the dataset based on feedback to encourage honesty and confidentiality. These changes happen year over year and the survey process is kept constant h the year that the survey is taken on and so for this reason we are assuming that the surveying process is stable. But since we are considering only 2012, the sample collection process is stable across 2012. Since this data contains inputs for other drugs such as cigarettes, tobacco, cocaine etc. we can expect some errors here.

# Scope of the study

Even though we have a multitude of inputs for various drugs, we are only concerned about the inputs related to alcohol and the core demographic characteristics.

With that said, the model is to predict the alcohol consumption of a person in a single day – to be exact, it predicts the number of days a person has five or more drinks in a single day. The model considers only a single month (past 30 days) as the relevant period. Even though, there is data for a single year period, it would be more relevant to predict only the monthly data since it is more current and less susceptible to change.

The variables we considered for our input model are given below.

|  |  |
| --- | --- |
| Variables | Reasoning |
| Number of days had 4+ drinks in the same occasion past 30 days | This is a strong indicator of the level of alcohol consumption of a person |
| Number of days had a drink past 30 days | To check if that person drinks frequently |
| Number of drinks past 30 days | Indicator of how much that person drinks |
| How many hours worked last week | Work is a factor which may encourage or discourage drinking more frequently depending on the nature of work |
| Number of days skipped work past 30 days | This indicates that the person is not regular in their work |
| Gender | This also may be a factor in drinking |
| Age | Young people tend to drink more |
| Education level | Education may be a deterrent to drinking |
| Family income | Income may or may not be a factor in alcohol consumption |

The model was fine-tuned by considering the following factors.

* Level of significance in the regression model – If the variable was determined to be not significant in the regression model, it was removed from the calculation
* Contribution to variation of the model – If the variable did not contribute much to the variability explanation, it was removed from the model
* Error in the model – Variables were added/removed if it reduced the error in the model
* Method of variable calculation – Since some of the variables were recoded or imputed, they would be first considered for removal, compared to other variables, if they did not contribute to the variance explanation. The model considers mostly variables which were obtained as direct inputs from the correspondent.

# Model Fine Tuning

The model initially considered all the variables mentioned above and a regression model was run on the data to determine the number of days with at least 5 drinks in a day

**No. of days with at least 5 drinks consumption in a day** = β0 + β1 \* No. of drinks in a day + β2 \* No. of days with at least one drink + β3 \* Hours worked last week + β4 \* No. of days skipping work + β5 \* No. of days with 4 or more drinks + β6 \* Sex + β7 \* Age + β8 \* Income + β9 \* Education + ε

## First iteration

Removing variables with low significance

With this, we first removed variables which are being insignificant. After cleaning the data, these values were found to be insignificant statistically, hence removed them from the model.

* Hours worked last week
* Income

## Second iteration

Removing variables not contributing to variance explanation or error reduction

With this, we considered removing which are not contributing much to the variance explanation and to the model’s error reduction

* No. of days skipping work

## Third iteration

Removing variables which are not directly obtained

Some variables included in the model were obtained by imputation or recoding. If they do not contribute much to the variance or are statistically insignificant, they were removed from the model

* Education

## Relationships of interest

The model now just considered 5 variables from the initial 9 variables

**No. of days with at least 5 drinks consumption in a day** = β0 + β1 \* No. of drinks in a day + β2 \* No. of days with at least one drink + β3 \* No. of days with 4 or more drinks + β5 \* Sex + β6\* Age + ε

We considered the degree of the relationship between

* variable depicting no. of days with 4 or more drinks and no. of days with at least one drink – approximately 64%
* same variable (no. of days with 4 or more drinks) and no. of drinks in a day – approximately 50%

Even though this model explained **80% of the variation**, since the variable depicting no. of days with 4 or more drinks is having strong collinearity with the other two independent variables, it was removed from the model.

Also by having this variable in the model, it tended to include people who drink more (at least 4 drinks in a day). This might bias the study and then, it may not be applicable to the general population. So, this was one more reason to remove the variable from the model

# Final Model and its Interpretation

**No. of days with at least 5 drinks consumption in a day** = β0 + β1 \* No. of drinks in a day + β2 \* No. of days with at least one drink + β3 \* Sex + β4\* Age + ε

This is the final model, after removing the variables which were considered insignificant or having strong relationship with other variables or not obtained directly from the respondents

**No. of days with at least 5 drinks consumption in a day** = 0.188 + 0.315 \* No. of drinks in a day + 0.226 \* No. of days with at least one drink + β3 \* Sex + β4\* Age + ε

So, keeping all the other variables constant,

* If a person drinks one more drink in a day than usual (β1 variable), the number of days when that person will drink 5 or more drinks will increase by 0.315 days
* If a person drinks one more day than usual (β2 variable), the number of days when that person will drink 5 or more drinks will increase by 0.226 days
* For this particular case -0.0075 is the decrease in usual number of drinks in the past thirty days when moving from 12-18 years to the 18-25 years category
* If a person is a female (variable is based on male) (β4 variable), the number of days when that person will drink 5 or more drinks will decrease by 0.415 days

# How much variation of the data is being explained by the model

In a simple linear regression, it is a measure of the correlation between the dependent and independent variable, squared to remove the sign.   And in multiple regression it is a cumulative measure that represents the correlation between all the variables involved in the model, that is between the dependent and the independent variables and between the independent variables themselves [2].

* Let us consider a clinical trial of a drug as an example to consider how much explanation of the variable is significant. As with most drugs the effectiveness of the drug varies as per the individual who is undergoing the clinical trial
* So, the amount of variance in effectiveness that can be explained by the model is very low and consequently has a very low variation. But the model variables might be significant
* A result like this could potentially save many lives in the long run and be worth millions of dollars if the model results in the drug’s approval for widespread use (Business, F. S)
* Even when variation is low, highly significant variables still indicate a real relationship between the significant predictors and the response variable. The coefficients estimate the trends while variation represents the scatter around the regression line. The interpretations of the significant variables are the same for both high and low variation models. So, low variation values are problematic only when you need precise predictions
* In some cases, it’s possible that additional predictors can increase the true explanatory power of the model. However, in other cases, the data contain an inherently higher amount of unexplainable variability with each added independent variable to the model.
* And in our case, we are considering human behavior and that too the behavior of alcoholics, so it is normal to expect a degree of unexplained randomness attached to the model (Frost, J., 1970).
* The point of this analysis was to see whether there was a relationship between alcoholism and a multitude of other factors both concerning the present behavior and past demographics of respondents, however small the relationship maybe. Hence, we are not placing much emphasis on the variation explained obtained for the final model obtained in our analysis

# Uses of the study

* Predict alcohol consumption of a person in a month
* Potential indicators of alcoholism and their degree of effect
  + Education level, age category, etc.

# Limitations of the study

The NSDUH data is the annual estimate of drug use patterns among civilian members of non-institutionalized American society. It is designed to be highly private and confidential to encourage honest reporting of delicate and sensitive issues like drug-use. But since this is data about human behavior, there are certain limitations to the data collection. So, there are some points to note on the insights that can be gained from this endeavor. The limitations include

## Truthfulness of respondents

The drug use is self-reported data whose validity is directly proportional to the honesty and the memory of the respondents and there is no way to verify the truthfulness of their claims. Though the surveys were designed to encourage honesty there could be some wrong information that would have crept into the data.

## Cross Sectional data collection

Respondents were interviewed only once and there was no follow up in the subsequent years. Insights that could be made about alcohol were only those based on data obtained in one cross sectional slice of time rather than the changing drug habits of one person over time.

## Target population of the survey

The respondents were chosen from the non-institutionalized civilian population of the American states. This criterion excluded people serving in the army, people in prisons and persons in institutions such as hospitals, treatment centers etc. Close to just less than 2% of population was being excluded and this would have introduced a significant skew to the data, since the study is about human behavior

# Appendix

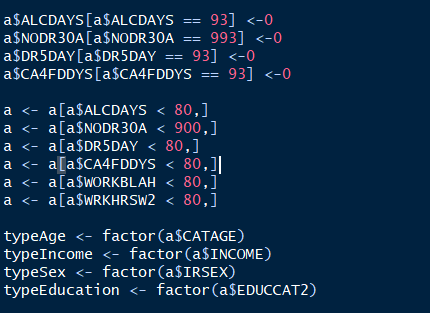
## Data Management

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

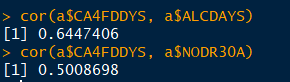
aGender: 0 = *male*, 1 = *female*. bAge: 0 = *12-17 years old*, 1 = *18-25 years old*, 2 = *26-34 years old*, 3 = *35 or older*.cEducation: 0 = *Less than high school*, 1 = *High school graduate*, 2 = *Some college*, 3 = *College graduate*, 4 = *12 to 17 years old*. dFamily income: 0 = *less than $20,000*,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

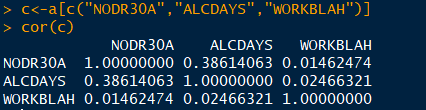
## R Code Output

### Data Manipulation and Factoring Categorical Variables

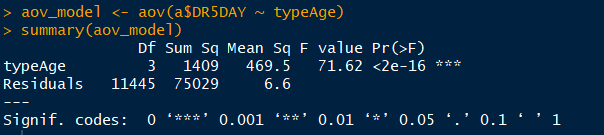


### Relationships of interest

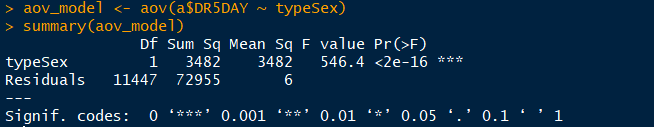




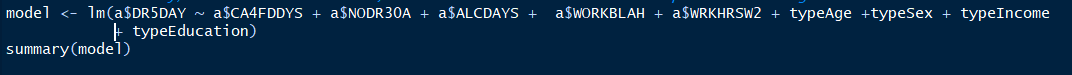
### Checking for Categorical Variable - Age

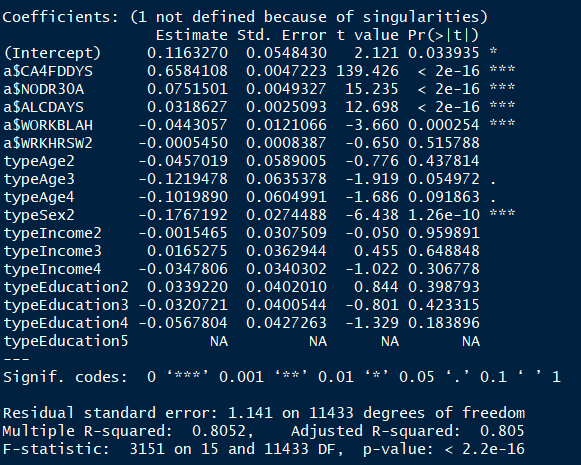


### Checking for Categorical Variable - Sex



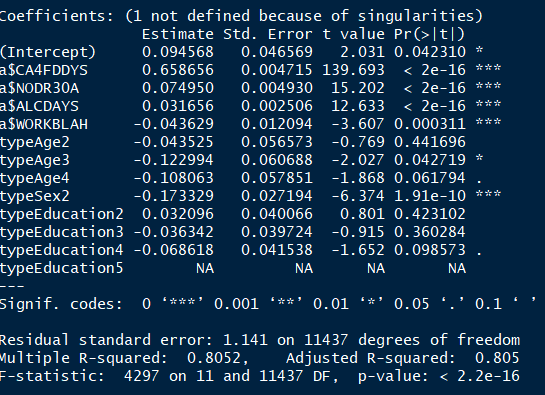
### Model – Initial Model





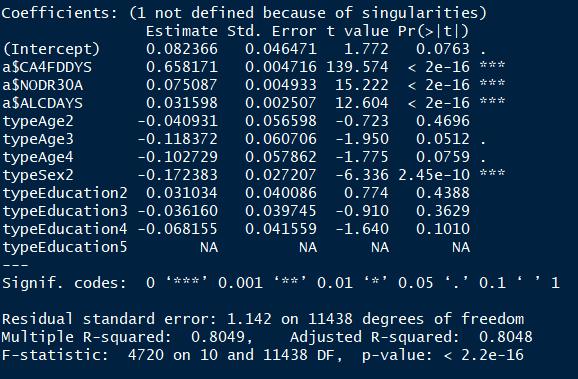
### Model – First Iteration

Screen Clipping



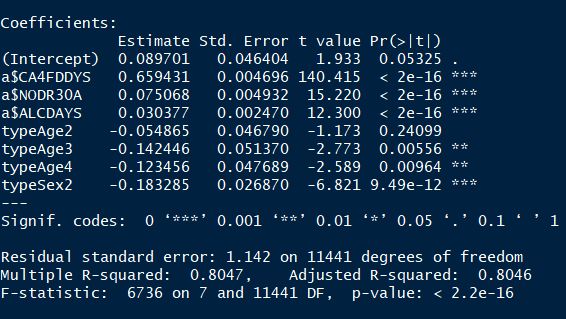
### Model – Second Iteration

Screen Clipping



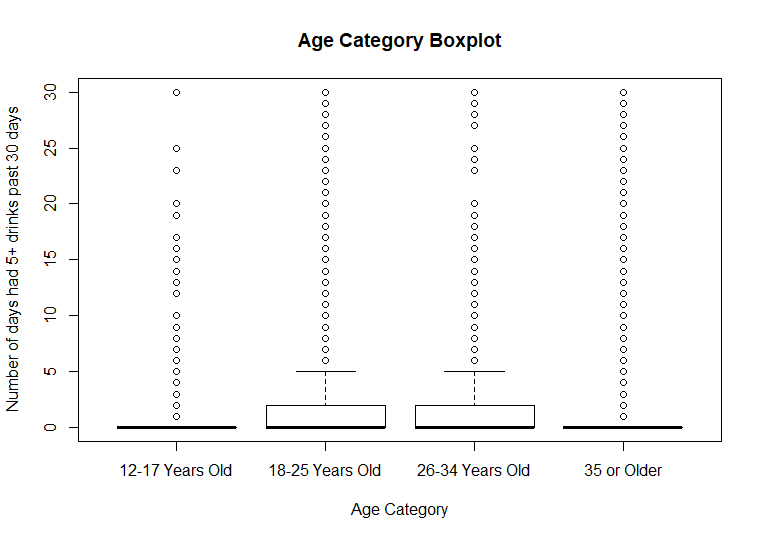
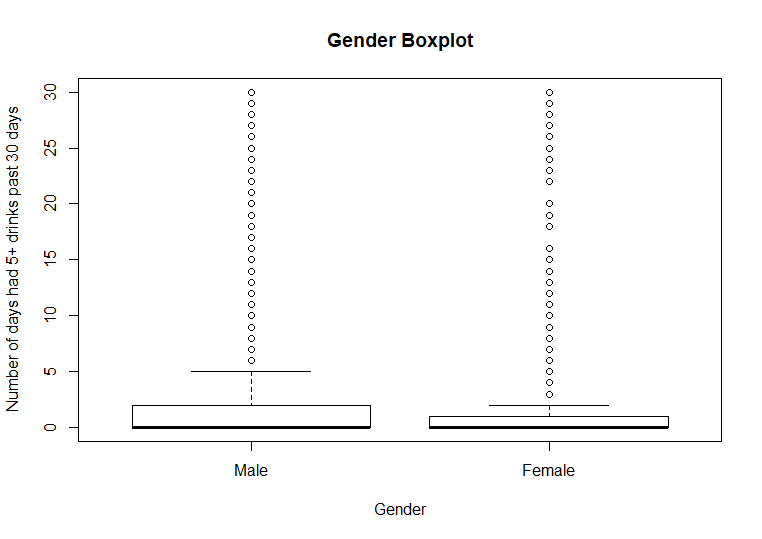
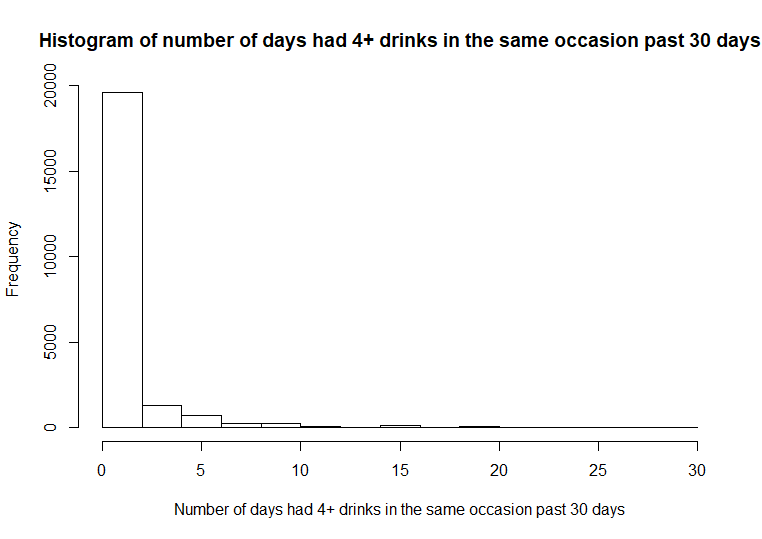
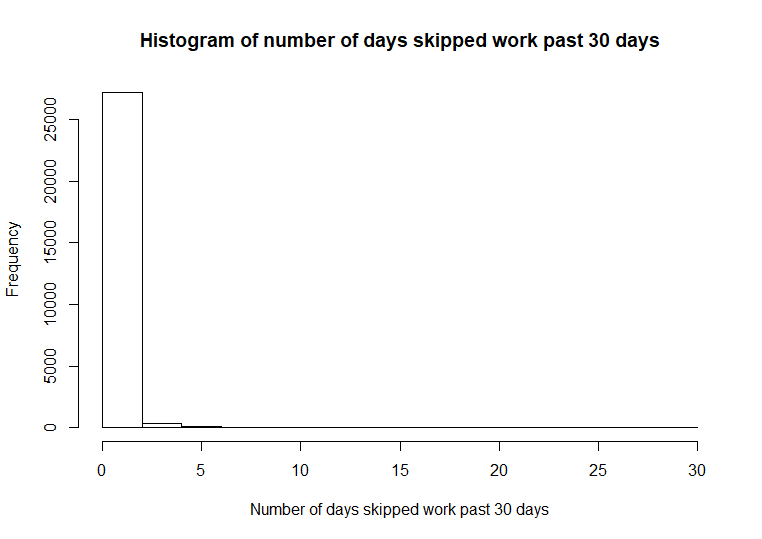
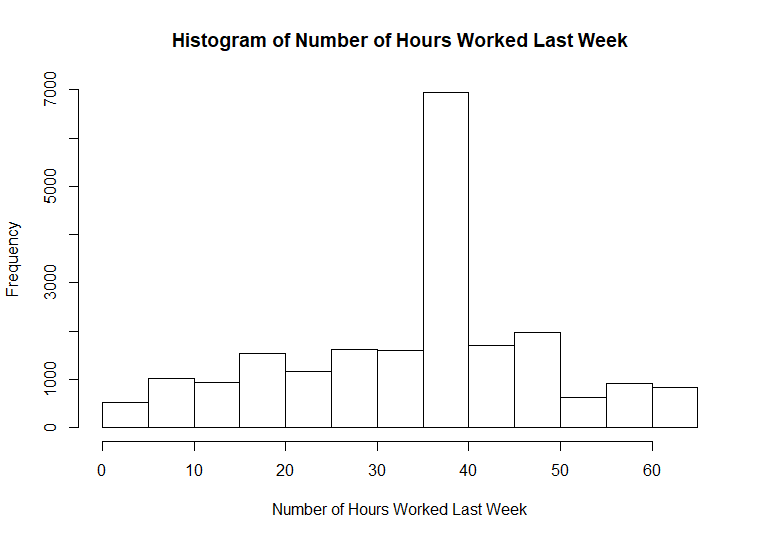
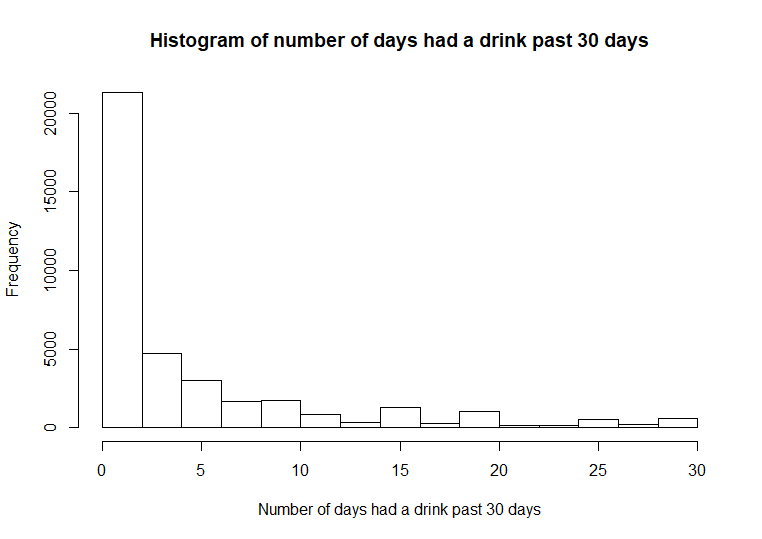
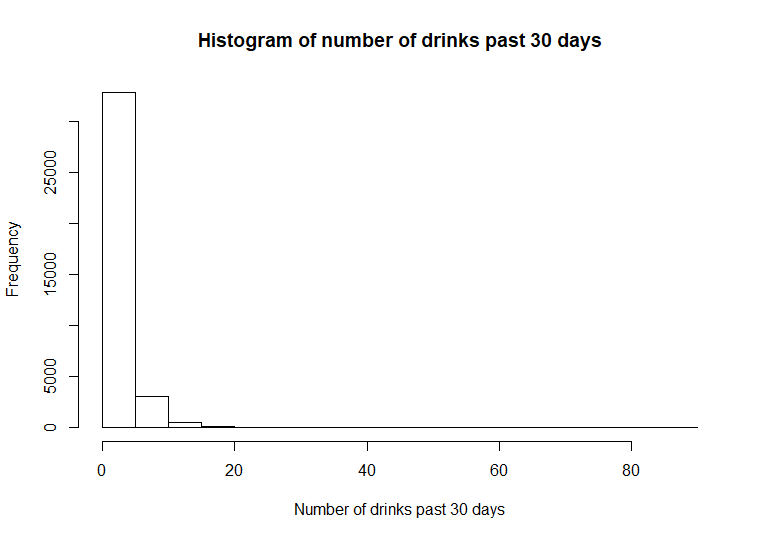
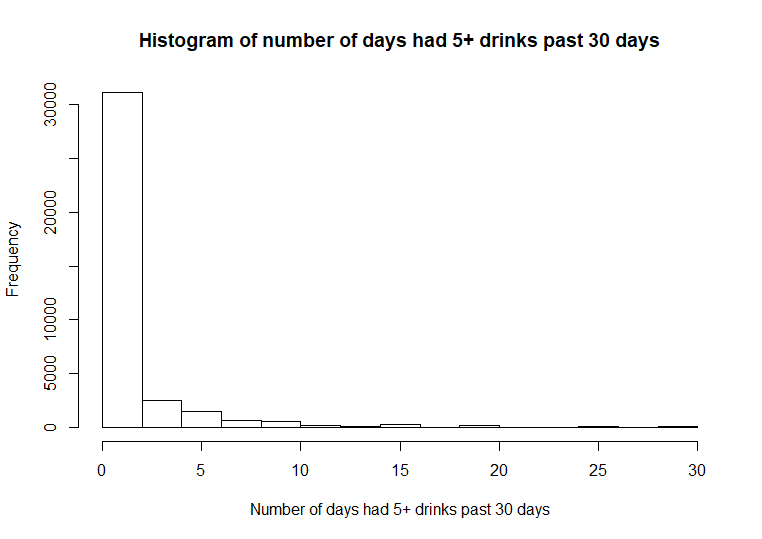
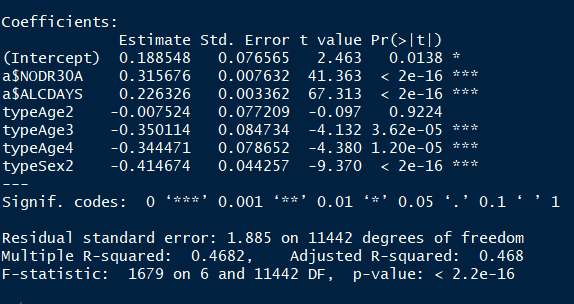
### Model – Third Iteration

Screen Clipping



### Final model

Screen Clipping



# References

Business, F. S. (n.d.). Https://people.duke.edu/~rnau/rsquared.htm. Retrieved from <https://people.duke.edu/~rnau/rsquared.htm>

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

Frost, J. (1970, June 12). How to Interpret a Regression Model with Low R-squared and Low P values. Retrieved from http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

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