Modeling Drug and Alcohol Dependence by Racial Identity

Taylor Francisco

5/3/2022

Data Analysis: Prof. Mike He

Quantitative Methods of Social Science

Introduction

Growing up as a Native American woman, I frequently encountered situations in which my identity and my community were erased and misrepresented. As a child, this meant learning in school about the ‘generous Indian’ happily sharing their feast with the pilgrims and then coming home to learn that the creation of Navajo fry bread came from rations given in concentration camps during the relocation of Indigenous peoples in the era of Manifest Destiny. It meant seeing the ‘drunken Indian’ or the ‘savage’ on TV and seeing absolutely no resemblance to my family. As I have grown older, I’ve learned that the Navajo experience, like many indigenous experiences, is one inexorably marked by erasure and misrepresentation. This erasure has come in many forms over the centuries from the *physical* erasure of our people to the erasure of our culture and language. Today, this erasure is a lot more insidious—Native Americans are being erased by the systemic exclusion of Indigenous peoples in statistical representations.

Virtually every study I read breaks down race/ethnicity into four broad categories: White, Black, Hispanic, and Other. I am that ‘Other’, and my community is directly harmed when we are erased from this process of information gathering. The issues facing Native Americans, such as the disproportionately high rates of suicide and potentially higher rates of substance dependence, which directly impacts the allocation of resources and development of life-saving policies or programs. Hundreds of years of mistreatment and misrepresentation have led to a population struggling disproportionately with trauma. As such, I am curious to interrogate how, in the United States, race influences rates of dependence on illicit drugs or alcohol. Are Native Americans impacted differently from various variables than say how White or Asian Americans are?

In order to study this question, I will be using the 2019 National Survey on Drug Use and Health (NSDUH) data with the main dependent variable being the respondent's illicit drug or alcohol dependence. I will mainly be regressing this by race, as my main independent variable, particularly focused on the comparisons between Native Americans and Asian Americans.

Hypothesis:

1) Native Americans will be more likely to have alcohol or illicit drug dependence than any other race. This is likely modulated by socio-economic status of the respondents.
2) Native Americans are more impacted by lack of access to proper health care and resources (i.e. health insurance, education, etc..) than Asian Americans are.

Description of Data Set and Variables

The data I decided to use was the 2019 National Survey on Drug Use and Health (NSDUH), one of the largest and most representative surveys regarding drug use and abuse in the US. This survey has been collected every year since 1990 and collects information about various drug, alcohol, and tobacco use, as well as basic demographic information. (Center for Behavioral Health Statistics and Quality, 2020) It also collects data regarding mental health and previous experiences with treatment. It samples from the US population of civilian and noninstitutionalized people aged 12 or older. Interviews, in both English and Spanish, take place face-to-face in the respondent's home and more sensitive questions are answered through a computer interface (2020). It uses a stratified state sampling regions methods using information from the Census to randomly select participants from all across the country (2020). The large sample taken each year makes it pretty representative of the population it is sampling from, but cannot be generalized to those under 12 or those who are in the military or in jail. This large survey is funded and supported by the Substance Abuse and Mental Health Services Administration (SAMHSA) and U.S. Department of Health and Human Services (2020).

My main dependent variable I will be trying to model is, **DPPYILLALC**, which I recoded to be named dependence. This is a binary variable of whether a person was defined as having dependence (1) on any illicit drug or alcohol, or not (0). Dependence was determined by if a respondent positively endorsed three or more of the following criteria, for alcohol or any drug including inhalants, stimulants, marijuana, etc.:

> 1. Spent a great deal of time over a period of a month getting, using, or getting over the effects of the substance.
> 2. Unable to keep set limits on substance use or used more often than intended.
> 3. Needed to use substance more than before to get desired effects or noticed that using the same amount had less effect than before.
> 4. Unable to cut down or stop using the substance every time he or she tried or wanted to.
> 5. Continued to use substance even though it was causing problems with emotions, nerves, mental health, or physical problems.
> 6. Reduced or gave up participation in important activities due to substance use.

This roughly approximates whether an individual would be considered having a substance use disorder based off of the DSM.

My main independent variable, **NEWRACE2**, formats itself similarly to that of the latest Census questions. For 2019 they recoded this variable to combine both Hispanic identity and ethnicity together (1 = NonHisp White, 2 = NonHisp Black/Afr Am, 3 = NonHisp Native Am/AK Native, 4 = NonHisp Native HI/Other Pac Isl, 5 = NonHisp Asian, 6 = NonHisp more than one race, 7 = Hispanic). I further recoded it to be named race and made it into an unordered factor, removing NAs.

To further dive into what variables may influence the relationship between race and drug/alcohol dependence, I used several other control and independent variables that are listed below:

IRSEX- The sex a respondent reported as being, not able to leave blank. Recoded as a factor named sex. Used as a control variable. (1 = Male, 2 = Female)
CATAG6- The age of the respondent broken into 6 categories. Recoded as age. Used as a control variable. (1 = 12-17 Years Old, 2 = 18-25 Years Old, 3 = 26-34 Years Old, 4 = 35-49 Years Old, 5 = 50-64 Years Old, 6 = 65 or Older)
HEALTH- How the respondent reported their general health to be. Recoded scale to be more interpretable and removed missing answers (1 = Poor, 2 = Fair, 3 = Good, 4 = Very good, 5 = Excellent)
EDUHIGHCAT- The highest level of education completed, broken into 5 bins. Recoded to be edu, removing last category that included children younger than 6. (1 = Less high school, 2 = High school grad, 3 = Some coll/Assoc Dg, 4 = College graduate)
ANYHLTI2- Binary of whether the respondent had any type of insurance (1) or not (2), removing missing answers and renamed insurance.
POVERTY3- Poverty level in % of US Census Poverty Threshold relative to the income of the respondent. Recoded to be named poverty (1 = Living in Poverty, 2 = Income Up to 2X Fed Pov Thresh, 3 = Income More Than 2X Fed Pov Thresh)
SPDYR- Whether the respondent has significant psychological distress in the past year according to questions from the K6 Distress Scale. I recoded this to be called distress. (0 = No (K6<13), 1 = Yes (K6>=13))

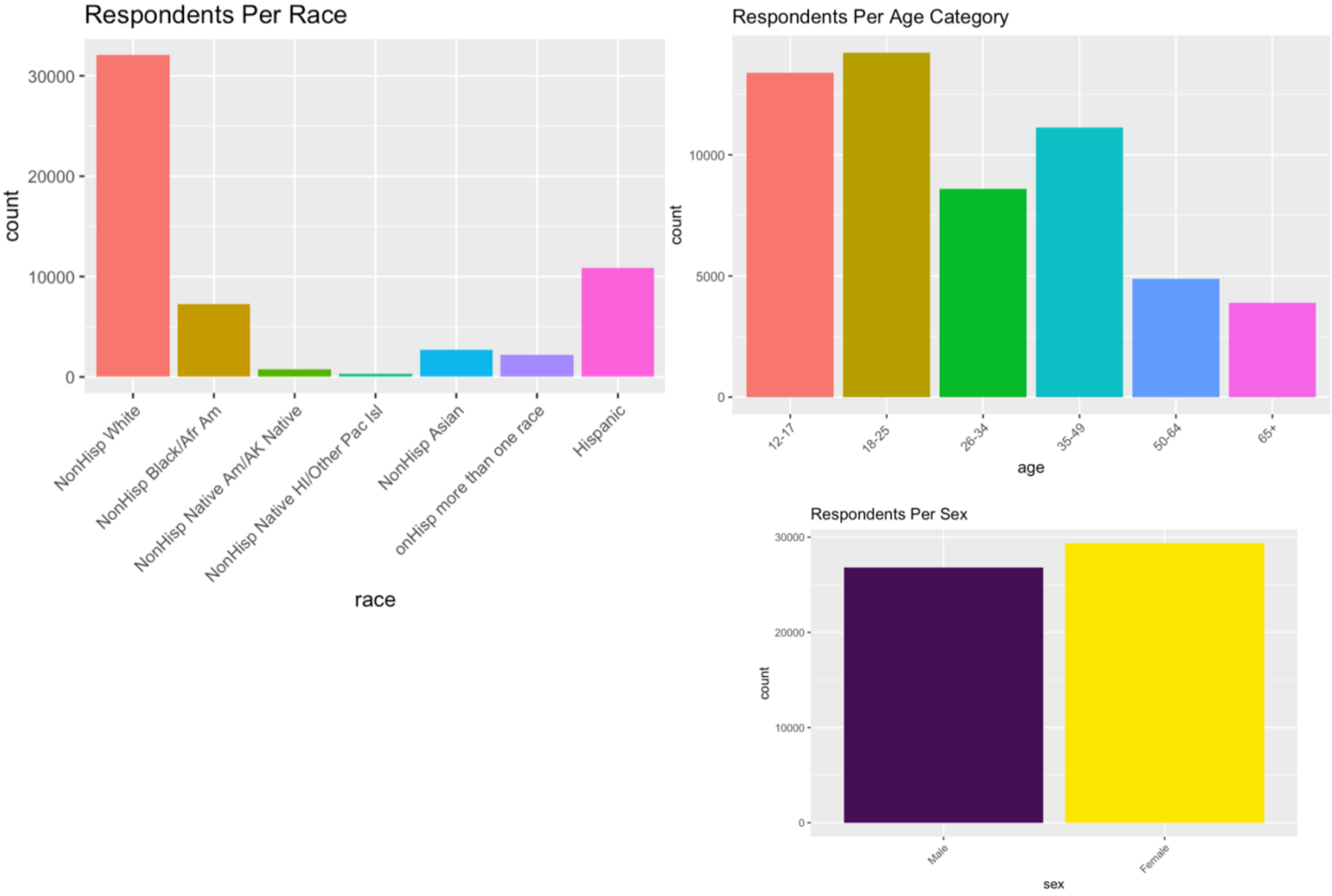
Descriptive Statistics

Table 1: Descriptive Statistics of Variables

variable	n	mean	sd	median	min	max	skew	kurtosis
race*	56136	2.72	2.45	1	1	7	0.967	-0.88
age*	56136	2.85	1.54	3	1	6	0.46	-0.84
sex*	56136	1.52	0.50	2	1	2	-0.09	-1.99
HEALTH	56127	3.76	0.97	4	1	5	-0.45	-0.35
distress	42739	0.18	0.38	0	0	1	1.67	0.79
insurance	55709	1.10	0.30	1	1	2	2.61	4.84
poverty	55609	2.43	0.78	3	1	3	-0.90	-0.78
edu	56136	3.29	1.29	3	1	5	-0.15	-1.08
dependence	56136	0.06	0.23	0	0	1	3.82	12.56

As race and age are ordinal variables, many of the above descriptive statistics are meaningless. The median can give a little insight into the distribution of these variables however, as race's median is 1, this indicates that at least the first half of data are Non-Hispanic White. Similarly, for sex the median being 2 indicates that less than half of the data are men. To get a clearer idea of race, age, and sex distributions a table with counts appears below.

Table 2: Histograms of Factored Variables



As you can see, there is a lot of Non-Hispanic White people as expected by the above median. The other races have much less respondents, which may reflect the distribution of these racial categories in the US, though more analysis would be needed to verify that. Sex appears to be close to 50/50, with slightly more Females than Males. Age seems more equally distributed than race across categories, though there appears to be much less people in the 65+ group, making the histogram of these categories look skewed to the right, though the skew provided above, 0.46, is not far off from the ideal 0. The kurtosis though is much different from the ideal 3 of a normally distributed variables, with a value of -0.84.

The health variable has a median of 3.76, which indicates that the average person reported their health to be between good and very good. The standard deviation is 0.97, which indicates that the variation in this variable is around 1, or a full category, meaning most of the respondents answered between 2.76-4.76, or roughly 2-5 (fair and excellent). The skew is small, but is slightly to the left, and the kurtosis is -0.35, which may indicate that this variable is not perfectly normally distributed.

Distress had a mean of 0.18, meaning the average score was closer 0, which is also supported by the median being 0 indicating at least half of the respondents reported having 0 (or no significant psychological distress the past year). The standard deviation was 0.38, meaning most of the respondents had a score close to the mean. These aren't the most useful descriptive statistics for a binary variable, but since I am including this as a potential variable and it may not be interesting to delve into further, these statistics give a good enough idea of the variable to play around with it in future models.

Poverty had a mean of 2.43, so the average response was between being in the top level of income and the middle income. The sd is relatively small 0.78, which makes sense as the scale is just 1-3, not leaving a lot of room for significant variation. The median being 3 suggest that more than half of the respondents reported having an income more than two times the federal poverty threshold. Similarly, to distress, this isn’t the clearest picture of this variable but gives me enough information to explore it as a possible independent variable.

Education had a mean of 3.29, so the average respondent has some college/an associate degree. It has a slightly larger SD of 1.29, meaning there is a bit of variation in the data set. The skew is -0.15 and kurtosis is -1.08, which is roughly close to what you would expect in a normally distributed variable.

Dependence has a mean of 0.06, with a standard deviation of 0.23. This indicates that there is not a lot of variation in responses and the average is closer to 0. The median is 0, meaning at least half of the respondents did not have a dependence on drugs or alcohol. Again, since this is a binary variable, looking at the proportion of responses would give us a more accurate idea of the spread of the data.

Table 3: Proportion of Respondent’s Per Category of Dependence

	No dependence (0)	**Dependence (1)**
Proportion of Respondents	0.94	0.06

From the above table we can see that the vast majority, 94%, of respondents were not classified as having a dependence on drugs or alcohol, with only 6% of respondents being classified as being dependent. This seems like this is accurate, as the majority of people do not have a significant issue with alcohol or drugs. However, since there is a large disparity between the two outcomes, a linear probability model would not be an ideal model to use for this dataset. Thus, a logistic regression will be a better initial model to explore the relationship between race and alcohol and drug dependence.

Another note is that the smallest N in the dataset is in the distress variable, meaning it had several missing variables- having only 42739 respondents. To get a general idea of how many variables can be used to create a model with this N, I will use the general rule of thumb of having around 10-20 observations per variable included. This means I can potentially use up to 42739/20 ≈2136 of variables, though using less will avoid creating an over-fitted model. Additionally, some races have less respondents than this, so to accurately describe my group of interest (Native Americans who have a count of 752), I should try to stick to 752/20 ≈37 independent variables. With all the variables above, including dummy variables, I have approximately 15 explanatory variables, which is within this general accepted range.

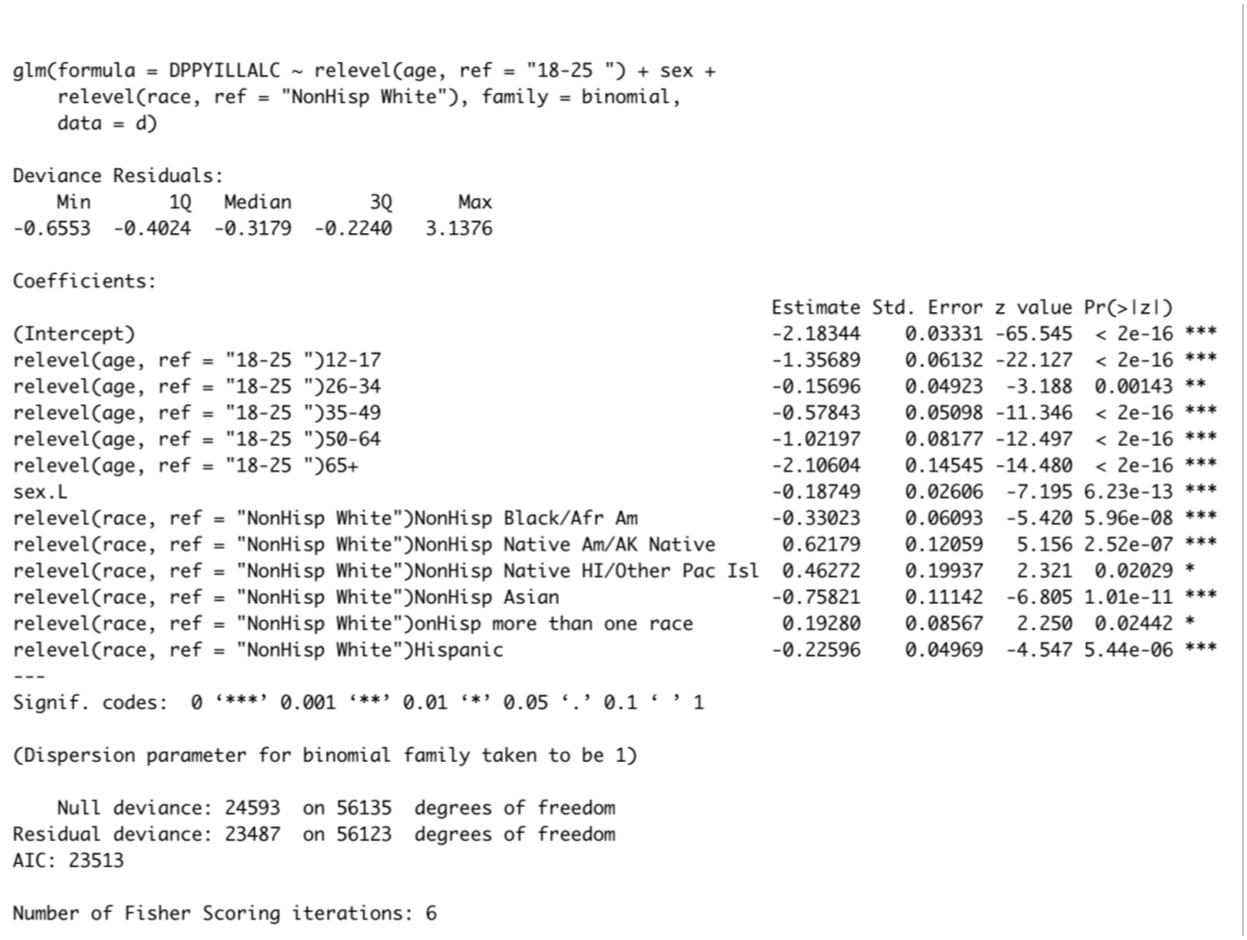
Initial Models

To model this binary outcome, I will start off with a simple logistical regression to see if there is an initial association between race and dependence, including some control variables. One of the assumptions to check again for this model is that I have a large enough sample size for the least frequent outcome, which in this case is Dependence, which is 0.06. Using the rule of thumb of 20 observation per variable and having 12 variables in this model: (20*12)/0.06 = 4000, which is less than the N we have in this model of 56136. Another simple assumption is that the observations are independent from each other, which is true in this dataset are from one individual at one time, without sampling from the same household, to avoid any dependency on one observation with another. While there are a few other assumptions, I will assume they are valid until the final model and examine them then once the final independent variables are chosen.

The model I began with was: $dependence = a + \beta_1 age + \beta_2 sex + \beta_3 race$

I created dummy variables for age, sex, and race as they are all categorical variables. I used white males between 18-25 as the reference group.

Table 4: Logistic Regression Model 1



```
glm(formula = DPPYILLALC ~ relevel(age, ref = "18-25 ") + sex +
    relevel(race, ref = "NonHisp White"), family = binomial,
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6553  -0.4024  -0.3179  -0.2240   3.1376

Coefficients:
                                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                     -2.18344    0.03331 -65.545  < 2e-16 ***
relevel(age, ref = "18-25 ")12-17                               -1.35689    0.06132 -22.127  < 2e-16 ***
relevel(age, ref = "18-25 ")26-34                               -0.15696    0.04923  -3.188  0.00143 **
relevel(age, ref = "18-25 ")35-49                               -0.57843    0.05098 -11.346  < 2e-16 ***
relevel(age, ref = "18-25 ")50-64                               -1.02197    0.08177 -12.497  < 2e-16 ***
relevel(age, ref = "18-25 ")65+                                 -2.10604    0.14545 -14.480  < 2e-16 ***
sex.L                                                           -0.18749    0.02606  -7.195 6.23e-13 ***
relevel(race, ref = "NonHisp White")NonHisp Black/Afr Am        -0.33023    0.06093  -5.420 5.96e-08 ***
relevel(race, ref = "NonHisp White")NonHisp Native Am/AK Native  0.62179    0.12059   5.156 2.52e-07 ***
relevel(race, ref = "NonHisp White")NonHisp Native HI/Other Pac Isl  0.46272    0.19937   2.321  0.02029 *
relevel(race, ref = "NonHisp White")NonHisp Asian               -0.75821    0.11142  -6.805 1.01e-11 ***
relevel(race, ref = "NonHisp White")onHisp more than one race    0.19280    0.08567   2.250  0.02442 *
relevel(race, ref = "NonHisp White")Hispanic                    -0.22596    0.04969  -4.547 5.44e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24593  on 56135  degrees of freedom
Residual deviance: 23487  on 56123  degrees of freedom
AIC: 23513

Number of Fisher Scoring iterations: 6
```

According to this model, net of race and sex, a 12-17 person decreases their log odds of being dependent on drugs/alcohol by 1.36, in reference to people aged 18-25. Similarly, 26-34 year odds decrease their log odds by 0.15, 35-49 decreased by 0.58, 50-64 decreased by 1.02, and those 65+ decreased by 2.1, net of race and sex. All of these decreases were statistically significant. Females have a 0.19 decrease of their log odds of dependence compared to males, controlling for all other variables, which is statistically significant.

Net of sex and age, a Non-Hispanic Black person decreases their log odds of dependence by 0.33 compared to Non-Hispanic White people. Non-Hispanic Native Americans/Alaskan

Natives have the largest increase of any race, with an increase of 0.62 log odds of dependence, controlling for all other variables. Native Hawaiians and Pacific Islanders increase their log odds by 0.46, net of all other variables. Non-Hispanic Asian people have a decreased log odds by 0.76, controlling for all other variables. Those who have two or more races increase their log odds by 0.19, net of all other variables. Lastly, Hispanic people decrease their log odds by 0.23 net of all other variables. All of the differences in race's log odds were found to be statistically significant with a p value less than 0.01.

Table 5: Odds Ratios of Model 1

Variable	Odd Ratio
(Intercept)	0.112652906913512
12-17	0.26
26-34	0.85
35-49	0.56
50-64	0.36
65+	0.12
sex	0.83
NonHisp Black/Afr Am	0.72
NonHisp Native Am/AK Native	1.86
NonHisp Native HI/Other Pac Isl	1.59
NonHisp Asian	0.47
NonHisp more than one race	1.21
Hispanic	0.80

Log odds however, are difficult to interpret, so I transformed the log odds to Odd ratios by taking the exponential of each logistic regression coefficient. The above table shows the resulting odd ratios, where an odd ratio of 1 indicates that as that variable increases the odds of dependence is as likely to occur as it is to not occur. An odd ratio above 1 indicates that dependence is more likely to occur, below 1 dependence is less likely to occur. This gives similar information to the results above. For example, Native Americans has the highest odd ratio of 1.86 (controlling for all other variables) and with sex (net of all variables) having a smaller odds ratio of 0.86, indicating that being female makes you less likely to have dependence compared to being male, while being Native American, you are more likely to have a dependence compared to a white person.

The most interpretable version of the above model is obtaining the predicted probabilities of a subset of a combination of variables. As an example, the below table shows the predicted probability for Black, White, and Native Americans between the ages of 18-25, both male and female. The rest of the combinations can be found in the Appendix. Compared to all other combinations, the type of person with the highest predicted probability, is a 18-25 Native American/Alaskan Native Male, with a 19% probability of having dependence.

Table 6:Example Predicted Probabilities of Model 1

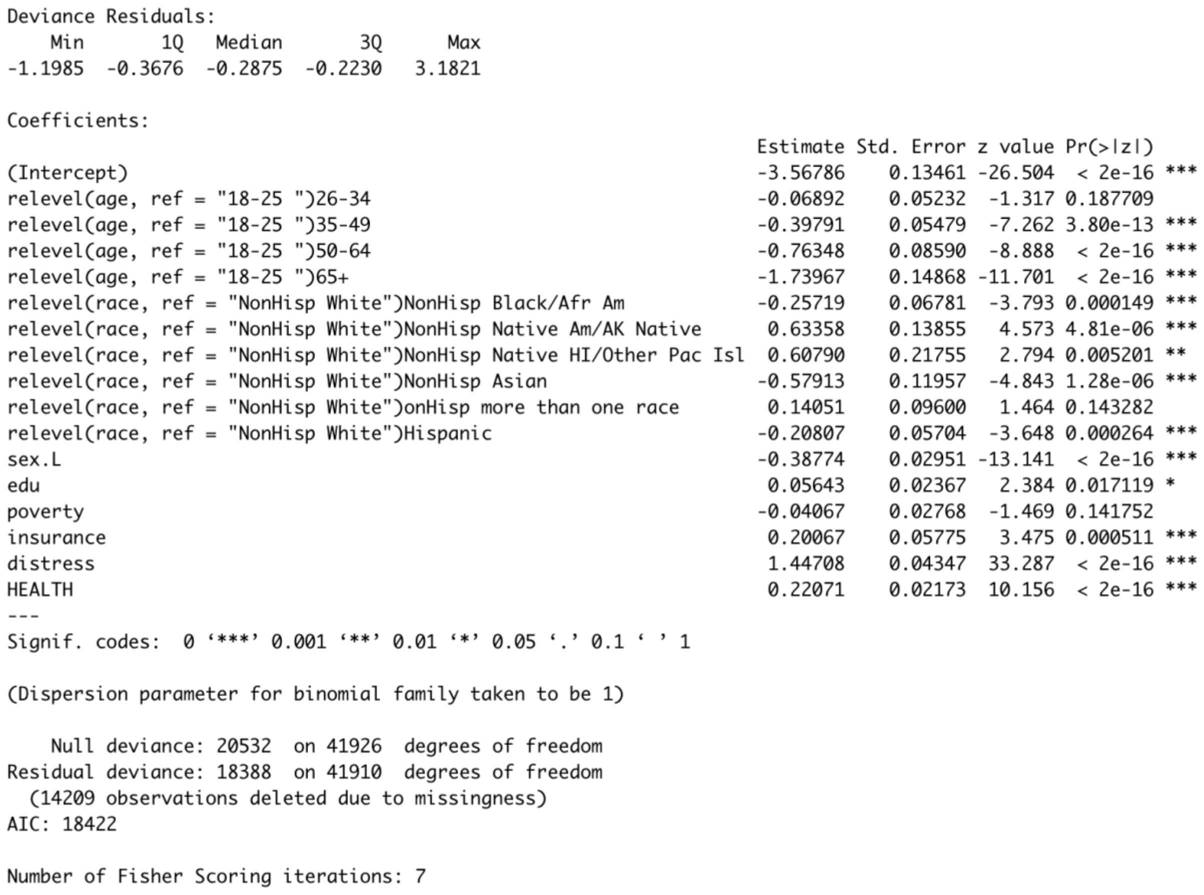
age	**race**	**sex**	**PredictedProb**
18-25	NonHisp White	Male	0.11
18-25	NonHisp Black/Afr Am	Male	0.08
18-25	NonHisp Native Am/AK Native	Male	0.19
18-25	NonHisp White	Female	0.09

18-25	NonHisp Black/Afr Am	Female	0.07
18-25	NonHisp Native Am/AK Native	Female	0.16

While this first model supports my first hypothesis that Native Americans are more likely to be dependent on alcohol/drugs, it does not give much insight into how various other SES factors impact this relationship. In order to get a better sense of this, I will add a few more independent variables to the model: distress, health, insurance, education, and poverty. This will be modeled as: : $dependence = a + \beta_1 age + \beta_2 sex + \beta_3 race + \beta_4 distress\ \beta_5 health + \beta_6 insurance + \beta_7 education + \beta_8 poverty$

To double check that this is not exceeding the sample size, with now 17 variables: (20*17)/0.06 = 5667, which is still less than the N we have in this model of 42739 (due to the added distress variable).

Table 7: Logistic Regression Model 2



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1985  -0.3676  -0.2875  -0.2230   3.1821

Coefficients:
                                                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                  -3.56786    0.13461 -26.504  < 2e-16 ***
relevel(age, ref = "18-25 ")26-34                            -0.06892    0.05232  -1.317 0.187709
relevel(age, ref = "18-25 ")35-49                            -0.39791    0.05479  -7.262 3.80e-13 ***
relevel(age, ref = "18-25 ")50-64                            -0.76348    0.08590  -8.888  < 2e-16 ***
relevel(age, ref = "18-25 ")65+                              -1.73967    0.14868 -11.701  < 2e-16 ***
relevel(race, ref = "NonHisp White")NonHisp Black/Afr Am     -0.25719    0.06781  -3.793 0.000149 ***
relevel(race, ref = "NonHisp White")NonHisp Native Am/AK Native  0.63358    0.13855   4.573 4.81e-06 ***
relevel(race, ref = "NonHisp White")NonHisp Native HI/Other Pac Isl  0.60790    0.21755   2.794 0.005201 **
relevel(race, ref = "NonHisp White")NonHisp Asian            -0.57913    0.11957  -4.843 1.28e-06 ***
relevel(race, ref = "NonHisp White")onHisp more than one race  0.14051    0.09600   1.464 0.143282
relevel(race, ref = "NonHisp White")Hispanic                 -0.20807    0.05704  -3.648 0.000264 ***
sex.L                                                        -0.38774    0.02951 -13.141  < 2e-16 ***
edu                                                           0.05643    0.02367   2.384 0.017119 *
poverty                                                      -0.04067    0.02768  -1.469 0.141752
insurance                                                     0.20067    0.05775   3.475 0.000511 ***
distress                                                      1.44708    0.04347  33.287  < 2e-16 ***
HEALTH                                                        0.22071    0.02173  10.156  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20532  on 41926  degrees of freedom
Residual deviance: 18388  on 41910  degrees of freedom
  (14209 observations deleted due to missingness)
AIC: 18422

Number of Fisher Scoring iterations: 7
```

Firstly, looking at the Akaike Information Criterion (AIC) of the first model (23513) compared to the second model (18422), the second model is much smaller, indicating that it is a better model. I cannot evaluate whether this difference is significant however, as the models used different sample sizes, as the distress variable had more data points missing, giving it a smaller N. However, this is a large enough difference for me to continue with this model, as it explains more variance and can give us insight into what other factors impact dependence in the different races.

Compared to the first model, every race category has the same direction of relationship, though now respondents with two or more races are not statistically significant from white respondents, though the other groups are still statistically significant. Similarly, all the age categories have the same direction, with only the 26-34 group dropping its significance compared to 18–25-year-olds. Sex is also relatively unchanged in direction and significance.

When looking at the added variables, we see that with every increase in education category, a respondent decreases their log odds of dependence by 0.06, net of all other variables. For insurance, with the change from having insurance to not having insurance, a person increases their log odds of dependence by 0.20, controlling for all other variables. Net of all other variables, with the change from not having any significant psychological distress to having distress, increases their log odds of dependence by 1.45. With every increase in health category, net of all other variables, a respondent increases their log odds of dependence by 0.22. All of these differences were statistically significant. The only added variable not found to be significant, was poverty, which had a small positive direction of log odds.

Looking at the odd ratios of this model, we can get a clearer idea of what this model suggests.

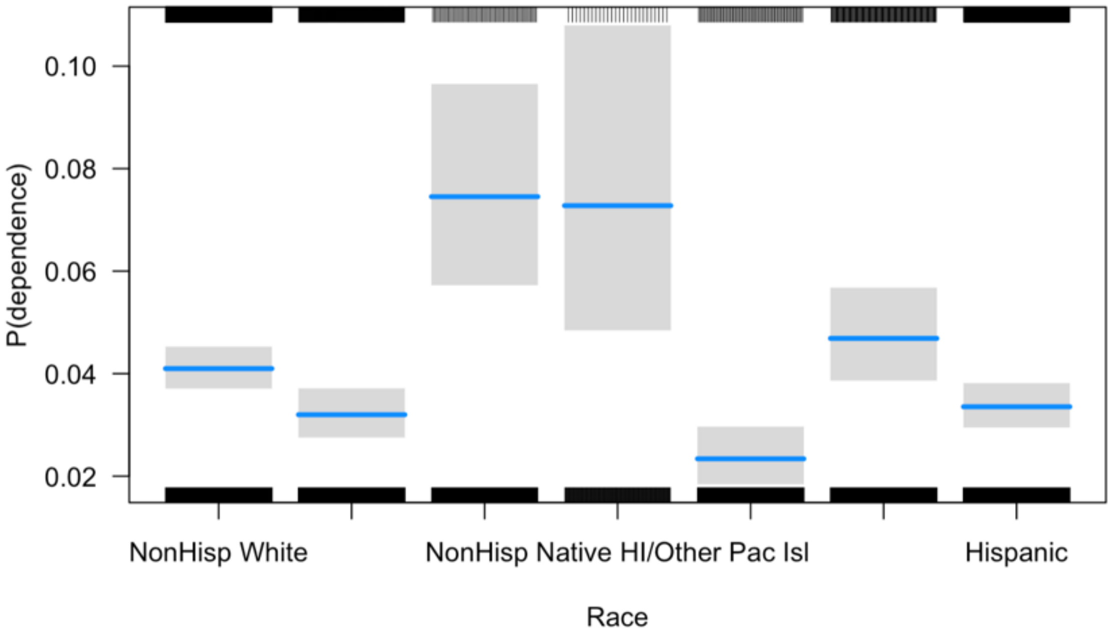
Table 8: Odd Ratios of Model 2

Variables	Odd Ratio
(Intercept)	0.03
26-34	0.93
35-49	0.67
50-64	0.47
65+	0.18
NonHisp Black/Afr Am	0.77
NonHisp Native Am/AK Native	1.88
NonHisp Native HI/Other Pac Isl	1.84
NonHisp Asian	0.56
NonHisp more than one race	1.15
Hispanic	0.81
sex	0.68
edu	1.06
poverty	0.96
insurance	1.22
distress	4.25
HEALTH	1.25

While the values have changed slightly from the first model for race, age, and sex, they are still in the same direction. Notably though, Native Americans/Alaskan Natives still have the highest odds ratio compared to the other races and age categories, again indicating that they have the highest likelihood of having dependence.

For the new variables, education has an odds ratio of 1.06 larger, which is close to 1 indicating that for each increase in category, they have a roughly equal chance of being dependent or not, with slightly more likelihood that they would be dependent, net of all other variables. Regarding insurance, net of all other variables, going from having insurance to not you have a 1.22 larger odd ratio (proportionally). Health has a similar 1.25 larger odds ratio with each category increase in general health, net of all other variables. Lastly, going from not having significant psychological distress to having distress, your odds ratio is 4.25 larger, meaning you are much more likely to have a dependence than not. These were all the variables that were found to be significant in the model, so those are the ones I will interpret. For following models, I will drop poverty as it does not seem to be contributing to the model, and there are other variables in the model that are proxies for SES, such as education or access to insurance.

Table 9: Predicted Probabilities of Race in Model 2



As there are a substantial number of possible combinations of 17 different variables, rather than having a table with the predicted probabilities of each possible combination, the above figure shows the predicted probabilities by Race. Each tick is in the order of how the variable was coded so (1= White, 3= Native American, 6= Asian, etc.). The grey bounds represent the confidence interval of the predicted probability, which is marked by the blue lines. The dark areas are representative of where the data lies (as in you either can or cannot be White). This figure again confirms that Native Americans have the highest predicted probability of dependence, evaluated at the mean of dependence, net of other factors.

While both models found significant associations between the independent variables I selected and dependence, they are still insufficient. One large potential issue is multicollinearity among my independent variables. Having too many variables that explain essentially the same things only increases variance in terms of noise, not unique variance that an independent variable brings to explain the total variance of the model. It also has the potential to render some variables as statistically insignificant. I assume that variables such as health, insurance, education, and poverty could be substantially correlated. To potentially solve this, creating a scale to represent SES will be helpful. Additionally, these models are simple additive regressions, and does not take into account potential interactions between variables. This may be leaving out synergistic effects of variables, for example if perhaps education and race interact to substantially increase risk of dependence, compared to other racees. This is key to addressing my second hypothesis of whether some variables interact with Native Americans in a different way than it does with other races.

Final Models

I first examined if there was in fact collinearity in my added dependent variables from my second model by checking Cronbach's alpha and running a Variance inflation factor (VIF) test.

Both suggested that my variables were not too collinear. The results of the VIF are below and the alphas can be found in the appendix.

Table 10: VIF of Model 2

variable	GVIF	Df	GVIF^(1/(2*Df))
age	1.18	4	1.02
race	1.12	6	1.01
sex	1.05	1	1.03
edu	1.26	1	1.120
poverty	1.19	1	1.09
insurance	1.08	1	1.04
distress	1.14	1	1.06
health	1.15	1	1.07

None of the variables have a VIF of 2.5 or more, so none of them have significant collinearity. Thus, creating a scale is not necessary.

To address the lack of interactions, I iterated through all the possible combinations of interactions, for example putting an interaction between race and age or race and distress. After all of those iterations, I chose the model with the lowest Akaike Information Criterion (AIC).

The final model for this paper ended up being:

$$dependence = a + \beta_1 age + \beta_2 race + \beta_3 sex + \beta_4 edu + \beta_5 health + \beta_6 distress + \beta_7 poverty + \beta_8 insurance + \beta_9 race * poverty + \beta_{10} sex * edu + \beta_{11} health * distress$$

This model had an AIC of 18407, which when comparing to the second model I created, was statistically better than that second model (Deviance= 30.616, F = 3.827, P = 0.0001644). The table below shows the coefficients of the final model.

Table 11: Coefficients of Final Model

Variable	coefficient
(Intercept)	-3.58
26-34	-0.06

35-49	-0.39
50-64	-0.73
65+	-1.74
NonHisp Black/Afr Am	-0.80
NonHisp Native Am/AK Native	1.39
NonHisp Native HI/Other Pac Isl	-0.37
NonHisp Asian	-0.27
NonHisp more than one race	-0.034
Hispanic	-0.46
poverty	-0.09
sex	-0.44
edu	0.06
health	0.27
distress	1.72
insurance	0.20
NonHisp Black/Afr Am:poverty	0.25
NonHisp Native Am/AK Native:poverty	-0.42
NonHisp Native HI/Other Pac Isl:poverty	0.43
NonHisp Asian:poverty	0.13
NonHisp more than one race:poverty	0.07
Hispanic:poverty	0.11

sex:edu	0.02
health:distress	-0.11

In terms of addressing my initial question, there are a few key coefficients to discuss: the overall trends in race alone and the interaction between race and poverty, both with a focus on the Native American/Alaskan Native groups.

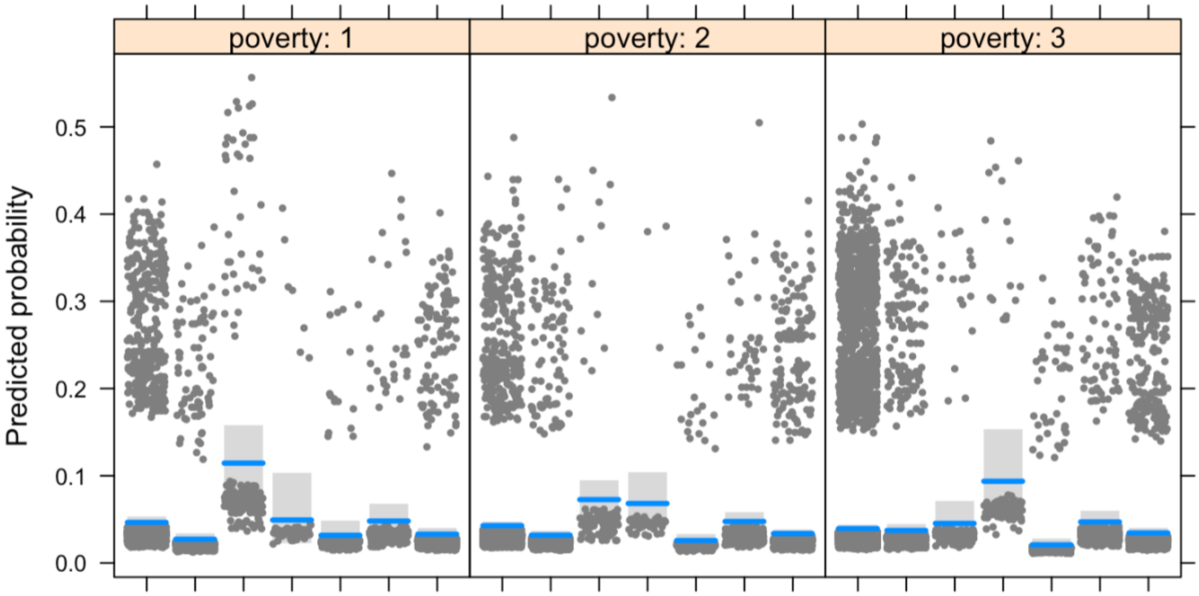
In this model, only Black, Native American, and Hispanic categories are statistically significant, so I will discuss those groups alone. Net of all other variables, a Non-Hispanic Black person decreases their log odds of dependence by 0.80 compared to Non-Hispanic White people. Controlling for all other variables, Hispanic people decrease their log odds of dependence by 0.46 compared to Non-Hispanic White People. Non-Hispanic Native Americans/Alaskan Natives increase their log odds of dependence by 1.39, controlling for all other variables. This again is the largest change in log odds of all the racial categories.

In regards to the interactions, only Black and Native American groups were statistically significant. For Black respondents, they increase their log odds of dependence for each category increase in poverty (i.e., the higher economic class they are in), relative to White respondents, net of all variables. Controlling for all other variables, Native American/Alaskan Native respondents, they decrease their log odds of dependence for each category increase in poverty (i.e., the higher the economic class), relative to White respondents.

To get a better understanding of this main comparison of how a SES factor (poverty level) impacts Native American's dependence differently than other races, the below figure shows the predicted probability of dependence broken down by both race and poverty level.

Again from an overall glance, you can see the third group, Native Americans/Alaskan Natives have the highest predicted probability, except for in the poverty 3 case. More interestingly, you can clearly see how most of the other races do not change much based off of their poverty category. Native Americans/Alaskan Natives however, decrease steadily as the category increases (i.e., as they are coming out of poverty). Notably, however, this group also has the largest confidence interval, suggesting that this estimate is not as accurate as the other racial categories. This is likely due to the small number of Native American/Alaskan participants. This however, is hard to improve as due to historical oppression, the percentage of Native Americans/Alaskan Natives in the US is incredibly small.

Table 12: Predicted Probabilities of Dependence based off of Race and Poverty



As a reminder, the ticks from left to right are labeled as (1 = NonHisp White, 2 = NonHisp Black/Afr Am, 3 = NonHisp Native Am/AK Native, 4 = NonHisp Native HI/Other Pac Isl, 5 = NonHisp Asian, 6 = NonHisp more than one race, 7 = Hispanic).

Cook’s distance was calucated in the appendix, and no values exceeding 1, so there are likely no severe outliers impacting this model, indicating that my model satisfies this assumption of a logistical regression.

While I think this model is illuminating and does support my hypothesis, I think there are several drawbacks. Firstly, this does not suggest causation at all, just a correlation. Since this is one time point, we cannot look at individuals before and after they have been determined to be dependent on alcohol or drugs. In the future, it would be interesting to find a dataset that follows individuals and try to isolate a causal relationship of what makes someone dependent or not on drugs or alcohol. Although my main interest in race, which cannot be randomly assigned, it would be very difficult to find a way to show causation of race on rate of dependence. I also think that a significant limitation to my model is the low N of Native Americans, which perhaps can be improved through either Bayesian statistics to create priors to get a better sense of the overall trend. Or additionally, if future versions of this dataset over-sampled Native Americans in the future so that more powerful statistical analysis can be done.

Conclusion

Overall, my initial hypotheses were supported. Native Americans are in fact more likely to be dependent on drugs or alcohol than any other race. Additionally, it seems that young Native American men are at the highest risk out of the entire population. My model also supports my hypothesis that other variables impact Native Americans differently than it does the general population. As shown in my final figure, decreasing poverty (going up in my scale of poverty) has a significant impact on the probability of dependence in Native Americans. This indicates that specific public health policies targeted at decreasing poverty in Native American communities can be a substantial protective factor against drug and alcohol use disorder, more so than in other communities. This also highlights the larger issue of importance of including Native Americans in analyses. If I were to have collapsed the racial categories to be White, Black, Hispanic, and Other this critical difference would have been masked. Especially considering that would mean combining Native Americans with the Asian respondents who generally were at the lowest risk of dependence- averaging these groups would eliminate any of the interesting deviations these groups have compared to their White counterparts.

As aforementioned, this final model is still not perfectly ideal. In the future there are a few adjustments I would like to do to improve accuracy. Firstly, I would want to explore more advance dimension reduction techniques such as PCA and factor analysis. While I did not find significant correlations between my variables, I think using these more advanced technique could identify relationships I did not see and would allow for me to use more of the variables provided in the dataset, and turn them into an index that would better represent complex variables, such as access to resources.

I would also like to incorporate Bayesian Statistics to incorporate priors about how often self-reported data is inaccurate when it comes to reporting drug use. This is particularly important, as I know in many Native American communities drug use is especially taboo, making self-reporting drug use is more likely to be inaccurate. I think also separating my regression into different regression for Native Americans and those of a comparison group, perhaps Asian Americans as they have a similar number of individuals. This would allow for more nuanced comparisons of what variables impact Native Americans differently than how they impact other groups.

References

Center for Behavioral Health Statistics and Quality. (2020). 2019 National Survey on Drug Use and Health Public Use File Codebook, Substance Abuse and Mental Health Services Administration, Rockville, MD

Appendix