Analysis Using Open Data

1.

Excessive drinking will not only harm one's own health, but also affect the family relationship of the alcoholic. According to CDC Dietary Guidelines for Alcohol, drinking is related to various dangers like traffic accidents, violence, sexual risk behaviors, and many diseases (Centers for Disease Control and Prevention, 2020). Therefore, detecting what the drinking groups' characteristics are is crucial for social interference. This paper tries to build up a model using demographic information to simulate drinking conditions in states.

To get a big picture of overall drinking rates in the United States, this paper uses state-level data. The drinking-related data is extracted from Health Data 2015, and the demographic information is extracted from American Community Survey because they provide authority state-level data. According to a CDC report, the male has twice as common as women to drink alcohol (Centers for Disease Control and Prevention, 2019). Lower income and educational level people consume more alcohol drinks per year while higher-income and educational-level peopleore commonly drink (Centers for Disease Control and Prevention, 2019). Therefore, drinking may relate to gender, education level, income level, and this study complement other possible variables like race and employment status.

After determining what variables are necessary, we choose the variables with closet meaning and the most straightforward data unit. Because they are downloaded as CVS documents, we open them in Excel, skim them, delete some noise format, and save it as xlsx. document. Then input one spreadsheet in STATA, clean, rename, and save it as dta. document. Input another spreadsheet, clean, rename, and combine it with the former dta. document using the same FIPS or state name. After that, we get the full dta. document, which contains both the alcohol drinking population and demographic information in states. Finally, we convert various population data into ratio data by normalizing them.

2.

Metadata is something that describes data, organizes information, and makes it easier to find and use. From libraries in the past to email, phone communications, social media accounts, web pages, and more, metadata is all around daily life.

The ACS and Health Data 2020 databases we use both generated statistical metadata through collection and processing. Then they use descriptive metadata to describe the basic information of these objects and publish it on the Internet for people to find and use. For ACS 2019 (1-Year Estimates) dataset, its data source is U.S. Census Bureau. It provides a data dictionary for users, including data introduction, subject definition, instruction for applying statistical testing, examples for approximating margins of error, and accuracy of the data. When clicking on a specific variable, it shows the variable details with percent base, aggregation method, the formula used to compute this variable, and relevant documentation. It is more convenient to use variable detail to check its definition and formulas. For Health Data 2020 Release, its data source is County Health Rankings & Roadmaps. It has general notes, data quality, and subject definitions.

For ACS and Health Data 2020, users select their interest variables through filters. In contrast, Open Data Network needs users to select and obtain data through API and code. Users who have coding background may prefer Open Data Network because it is like an aggregation of datasets of various sources, and they can select, even clean the data while downloading. The ACS and Health Data are better for me because it looks more instinct and does not need coding knowledge. The only trouble is that you need to know what data source you can look for target variables and need to use Excel or Stata to merge and clean datasets after downloading.

3.

Alcohol addiction is not like a switch that can be flipped on in a second but rather an unconscious path. Therefore, it can be hard to find someone or yourself is the potential for alcohol abuse. It makes the data analysis crucial. Research could help change stereotypes about alcoholics and make people more aware of the risks to those around them.

**Question**: What kinds of the demographic characteristic may be indicators of drinking adults in around 2020?

We use the "State" variable to combine two datasets and then use the metadata (Drink, TotalPop, Male, Black, Age116\_19, DropSchool, Labor, Unemployed, Poverty) to produce ratio variables (MaleRate, BlackRate, DropRate, UnemRate, PovRate, DrinkRate). In the following analysis, we only use these normalized ratio variables.

The definition used for "Drink" is the percentage of adults drink more than 4 (women) or 5 (men) alcoholic beverages at once in the past 30 days. Alternatively, drink more than 1 (women) or 2 (men) beverages per day on average. The former kind of drinker is called an excessive drinker, and the latter is a heavy drinker.

Table 1 Variable Name

|  |  |  |
| --- | --- | --- |
| Name | Label | Unit |
| FIPS | FIPS code |  |
| NameofArea | Name of Areas |  |
| State | State Postal Abbreviation |  |
| Drink | Percent Drinking Adults (Persons 18 Years and Over) | Percentage |
| TotalPop | Total Population | universe |
| Male | Total Population: Male | universe |
| Black | Total Population: Black or African American Alone | universe |
| Age16\_19 | Civilian Population 16 to 19 Years: | universe |
| DropSchool | Civilian Population 16 to 19 Years: Not High School Graduate, Not Enrolled /Drop | universe |
| Labor | Civilian Population in Labor Force 16 Years and Over | universe |
| Unemployed | Civilian Population in Labor Force 16 Years and Over: Unemployed | universe |
| Poverty | Population for Whom Poverty Status Is Determined | universe |
| MaleRate | Male population rate | ratio |
| BlackRate | Black population rate | ratio |
| DropRate | Drop or not enroll high school population rate | ratio |
| UnemRate | Unemployment rate | ratio |
| PovRate | Poverty rate | ratio |
| DrinkRate | Rate of drinking adults | ratio |

**Analysis**

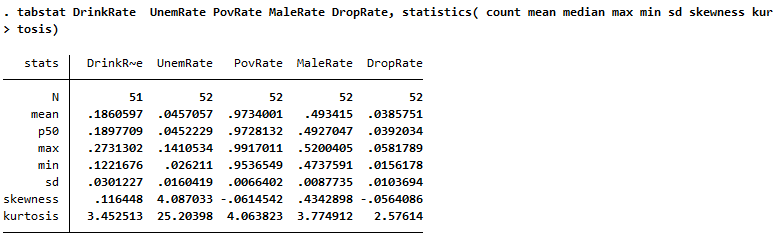


Figure Statistics

To be perfectly normal distribution, the skewness should equal to 0 and the kurtosis should be 3. The skewness of UnemRate is much higher than 0 and the kurtosis is much higher than 3. So, it is not normally distributed. While other variables have small skewness close to 0 and kurtosis close to 3. Therefore, variables except UnemRate may be quite normally distributed.

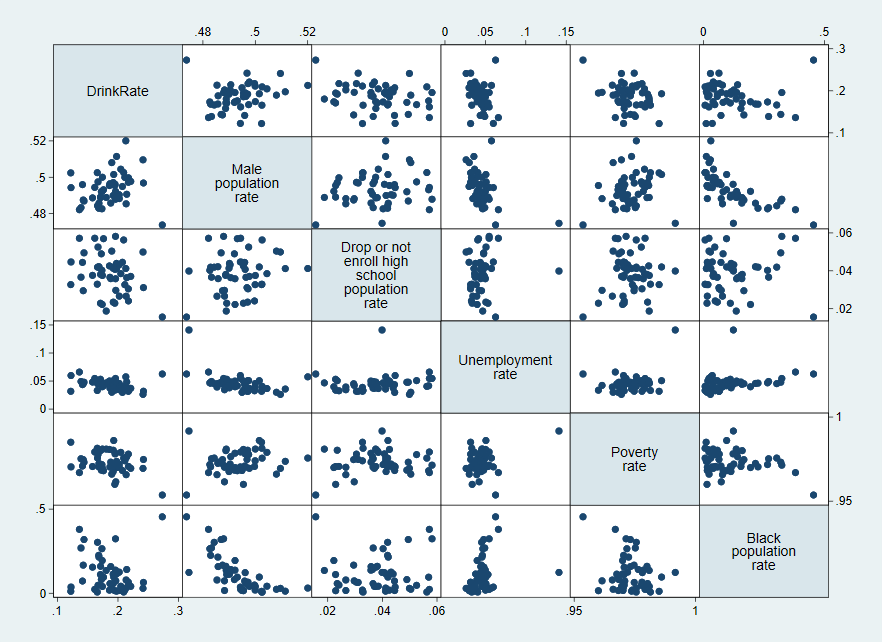


Figure Scatter Plots

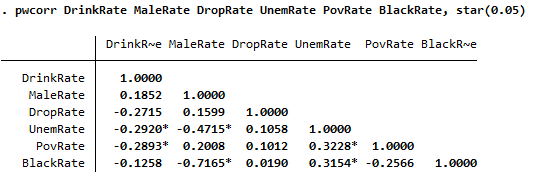


Figure Pairwise Correlation

According to the scatter plots and pairwise correlation table, there may be some relationship between DrinkRate (drink rate) and UnemRate (unemployed rate), DrinkRate (drink rate) and PovRate (poverty rate). DrinkRate has the least relation to BlackRate, so we drop this variable.

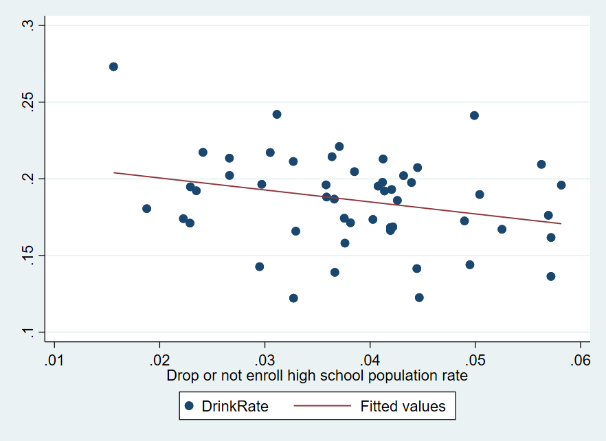
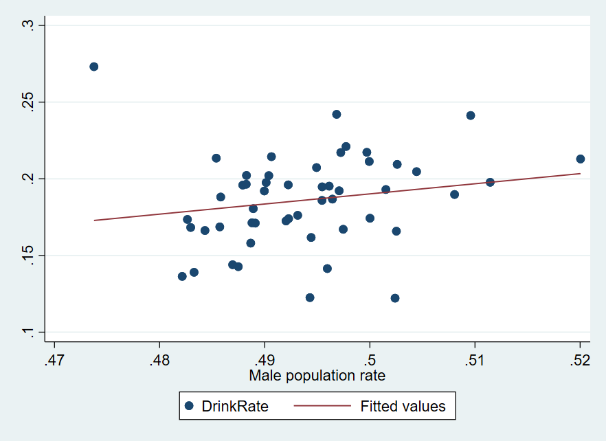


Figure DrinkRate vs. MaleRate Figure 5 DrinkRate vs. DropRate

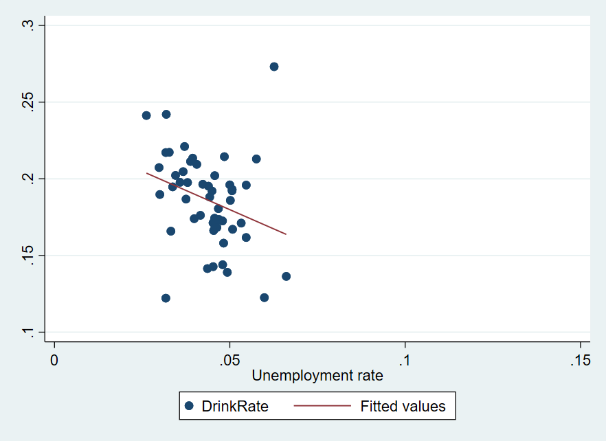
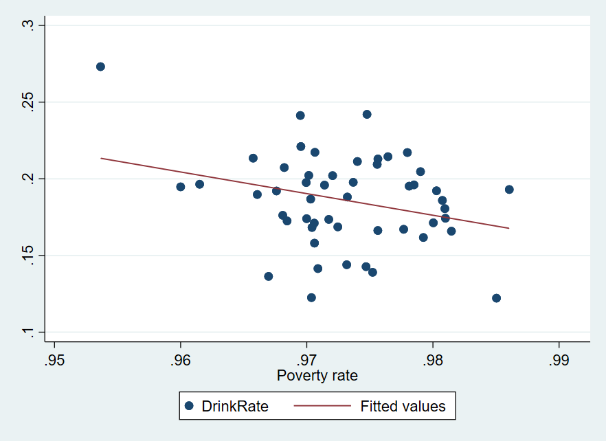


Figure 6 DrinkRate vs. PovRate Figure 7 DrinkRate vs. UnemRate

Adding the fitted value line can help us detect the relationship of variables. According to the graph, DrinkRate may have positive linear relation with MaleRate, negative relation with PovRate and UnemRate. Because these are linear relation, so we use linear regression model.

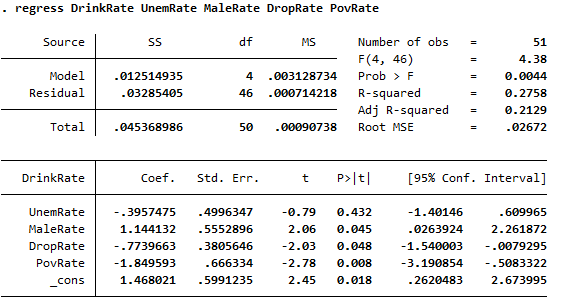


Figure 8 Regression Model 1

The Prob>F = 0.004, which is smaller than 0.01, is a good start. The R-squared is only 0.27, meaning this model can explain 27% of real data. The Root MSE is small shows no problem. However, the UnemRate has the biggest p-value, which exceeds 0.05 a lot. So, it may be better to preclude this variable.

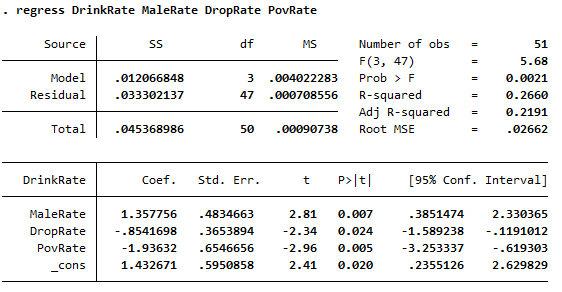


Figure 9 Regression Model 2

The second model has only three explanatory variables, but its R-squared=0.26, which can explain 26% of real data. The Prob>F = 0.0021 is still smaller than 0.01, and it is good. All the explanatory variables have a p-value smaller than 0.05 now. They are related to DrinkRate. The MaleRate positively related to DrinkRate, the DropRate, and PovRate negatively related to it.

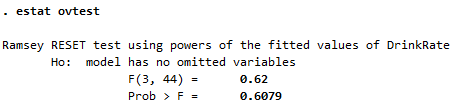


Figure 10 Omitted Variable Check

According to the estat ovtest, the Prob>F=0.6, which exceeds 0.05 so as cannot reject the Ho. There is no omitted variable in this model.

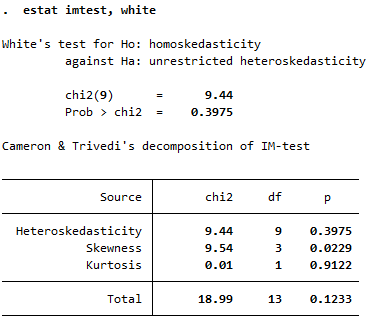


Figure 11 White Test

Using the White test, the Prob>chi2=0.39, which is much bigger than 0.05, cannot reject the Ho. There is no heteroskedasticity in the model.

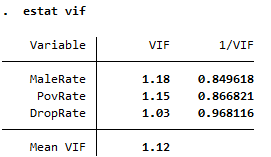


Figure 12 VIF Test

The VIF of every explanatory variable is smaller than 5, so there is no multicollinearity.

After checking all these potential problems in the model, it is reasonable to say drinking is positively related to the male population and negatively related to the drop high school and poverty population.

This finding is interesting because we often see images of drunk people living in run-down cabins in movies, and it is easy to associate low-income people with alcoholism. However, the fact is that people living in poverty and without a high school degree are not as likely to binge drink as others might expect. The study revealed that well-off families might be at greater risk of binge drinking, a warning against complacency. It also confirms the CDC report that people with higher income and educational background are more commonly drink.

# References

Centers for Disease Control and Prevention. (2019, December). *Alcohol and Public Health*. Retrieved from Binge Drinking: https://www.cdc.gov/alcohol/fact-sheets/binge-drinking.htm

Centers for Disease Control and Prevention. (2020, December). *Alcohol and Public Health*. Retrieved from Dietary Guidelines for Alcohol: https://www.cdc.gov/alcohol/fact-sheets/moderate-drinking.htm