**Predicting LSD, Psilocybin, and Ecstasy/Molly use based on another drug usage**

Andrew Pype

Project Category: https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age

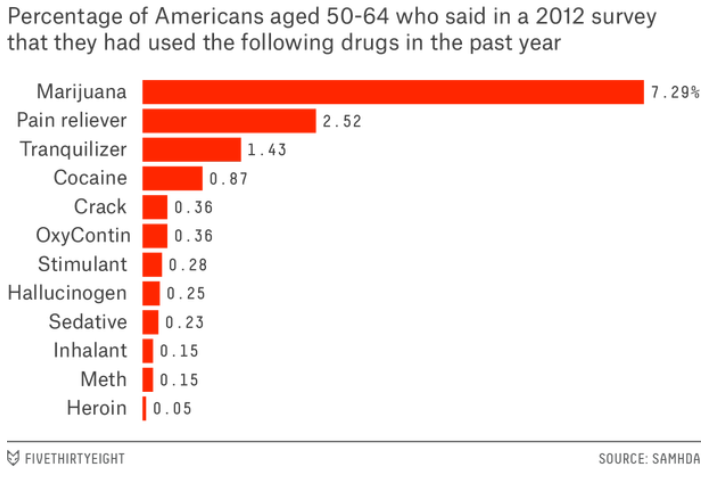
Introduction:

When growing up and learning more about the world as a kid, we are warned of many different situations. Typically, drugs and alcohol are mention frequently. A gateway drug is a drug that may not be addictive but can lead to other addictive drugs. The dataset used in this analysis came from the National Survey on Drug Use and Health 2016 [2]. It contains a lot of statistical information regarding the use of illicit drugs and their addictiveness. I narrowed down this dataset to include only data pertaining to the usage of cigarettes, alcohol, marijuana, cocaine, LSD, psilocybin, and ecstasy/molly. The goal is to predict which parameters have the most influence in LSD and ecstasy/molly use. I use regression models such as multiple linear regression and logistic regression to classify which drugs are a gateway to harder drugs such as the hallucinogens listed.

Related Work:

The Substance Abuse and Mental Health Data Archive (SAMHDA) is a dataset from 2012 and was used for reporting baby boomer illicit drug usage since their rates have be higher over the past decade [1]. The dataset is similar in terms of collecting data and using it to evaluate how patterns of drug use develop with age. Although not the same dataset, it pertains to similar results.

The article performs simple data analysis including listing various statistics about the older generation of interest and their drug habits. Questions they answer include; “Are their patters of use different from other age groups? How similar are people with the baby-boomer cohort when it comes to drug use?” [1]. To no one surprise, alcohol was the most common drug used in the boomer age range (50 to 64), roughly 67% of boomers use alcohol. Marijuana was next which is common is all age groups. The next highest drug use was pain relievers which is mostly due to the effect of growing old. Below shows a graph made by the authors of ‘How Baby Boomers Get High’ [1], showing the percentage of baby boomers’ usage of a given drug in the past year.



‘How Baby Boomers Get High’ [1]

Comparing the boomer data to younger generations can shed a lot about how gateway drugs operate as one grows older. It turns our that younger boomers have a higher rate of painkiller use than older boomers. What causes these trends are still unknown but looking towards the younger generations and their drug use offers insight to these questions.

Dataset:

The trend of drug use is of interest; pertaining to which common drugs influence the use of harder drugs in the future. The dataset used, NSDUH-2016-DS0001, is similar to the dataset used in the boomer analysis but is much larger. This means we can pull a lot more information on what drugs and their abuse led to other drug abuse.

The NSDUH-2016-DS0001 dataset is massive, 56897 observations over 2668 features. This set is much too large to work with and contains data that is not of interest for the sake of this project. I decided to cut down to variables only pertaining to cigarettes, alcohol, marijuana, cocaine, LSD, psilocybin, and ecstasy/molly; 56897 observations over 44 features.

The features included mainly pertain to questions such as; Have you ever tried [given drug]? How often…. Features such as these can tell a lot about how a single drug can influence the chances of trying another drug. By looking into how often a given drug is taken, can tell us if an addiction has formed. These features were chosen because they are the most common, meaning more people will have tried it thus given me more data to work with.

Methods:

1. Multiple Linear Regression:

Multiple linear regression is used because it is a relatively simple model and can tell us how a single predictor can influence other predictors as well. It seems as though alcohol use and marijuana use probably influence one another, and many other drugs too. This is why multiply linear regression is useful. The equation below shows the form of multiple linear regression.

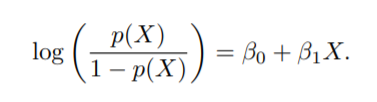
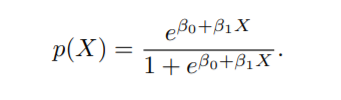


ISLR, eq 3.19 [3]

The number is features is denoted by p, each predictor denoted as Xp is association with a variable, βp, describing the relationship between the variable and the response Y. To estimate these coefficients, a minimization of the multiple least squares estimate takes place. Overall finding the best coefficients to represent a given multiple linear regression model.

1. Logistic Regression

This type of regression is relevant to classification in the sense that we model the response to belong in a particular category. For the sake of the project, these categories are either yes, have tried a given drug or no, have not tried a given drug. This type of regression models the probability (ranging from 0 to 1) of being in a given category. Since we care about the probability, the function must return values from 0 to 1, hence the logistic function, see eq 4.2. To denote one unit increase of X, it represents an increase in the log odds of p(X) by β1.



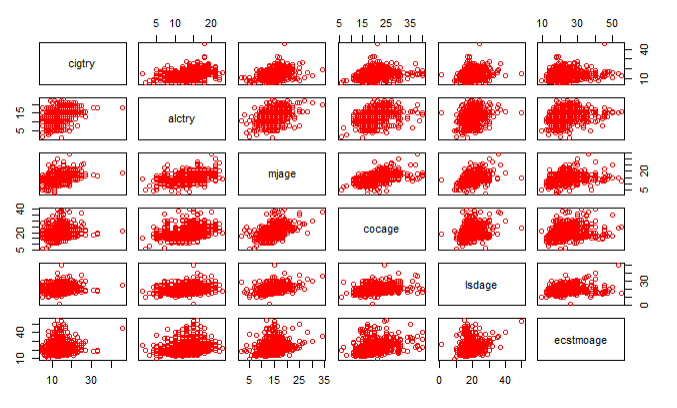
ISLR, eq 4.2 & 4.4 [3]

The parameters are estimated through the method of maximizing the likelihood function. Fitting non-linear methods can be done through different means but the likelihood function is a general approach.

Results:

Multiple Linear Regression:

Multiple linear regression was performed first, to try and demonstrate which factor are relevant enough to contribute to predicting LSD age and Ecstasy/molly age. Psilocybin was not chosen for the section because the dataset did not contain a feature for Psilocybin age. I used a simple hypothesis model to determine the important coefficients. This simply means that the p value produced by the fit is smaller than 0.05, than that coefficient contributes to the model.



The dataset had to be slimmed down to only include variables that answered yes or no for each drug of interest. Above shows a simple point plot of each age variable regarding a given drug. It is important to see how these variables influence one another before any model is applied.

LSD Age Predictor Values

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate Value | Error | P value |
| Intercept | 6.37 | 0.522 | < 2e-16 |
| Cigarette Age | 0.0383 | 0.0265 | 0.15 |
| Alcohol Age | 0.0358 | 0.323 | 0.267 |
| Marijuana Age | 0.298 | 0.3603 | <2e-16 |
| Cocaine Age | 0.229 | 0.0231 | <2e-16 |
| Ecst/Molly Age | 0.116 | 0.0148 | <2e-16 |

We can see that only marijuana, cocaine and ecstasy/molly age are significant enough to affect the age at which someone is likely to try LSD. To explain what the estimated value represents, think of it as a single value increase in Marijuana Age will result in a 0.298 increase in LSD Age as long as the other features are held constant. To understand it in terms of age, one would be more likely to try LSD as you grow older.

Ecstasy/ Molly Age Predictor Values

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate Value | Error | P value |
| Intercept | 7.823 | 0.846 | <2e-16 |
| Cigarette Age | -0.0514 | 0.0423 | 0.225 |
| Alcohol Age | -0.0329 | 0.0514 | 0.522 |
| Marijuana Age | 0.0296 | 0.0585 | 0.613 |
| Cocaine Age | 0.421 | 0.0365 | <2e-16 |
| LSD Age | 0.294 | 0.0376 | 8.12e-15 |

When the linear fit is done fore Ecstasy/Molly age, we see that the only significant terms is cocaine age and LSD age. Cigarettes, alcohol and marijuana play no role in our model to determine Ecstasy/Molly age.

In terms of both multiple linear regression model, they made sense. Cigarettes and alcohol are much more common drug used by society. Therefore, it would make sense that more people from the dataset would try those things and not go any further. Another interesting aspect is that marijuana age plays a role in the LSD model and not the Ecstasy/Molly model. This might be due to marijuana having hallucinating affects so one might be more likely to try another hallucinogen where ecstasy/molly is more of a stimulant.

Logistic Regression:

Logistic regression was then applied onto a dataset containing the statistics of whether a given person tried the drug (a simple yes or no question). I used this binary dataset to predict how likely one is to try LSD or Ecstasy/Molly.

The variables relevant enough to predict whether someone would try LSD includes if they tried cigarettes, marijuana, cocaine, psilocybin, and ecstasy/molly. All but alcohol was included in the model due to the hypothesis test. The model was the tested against the true values of whether someone was likely to try LSD. Similarly, a model was fit for ecstasy/molly were the significant variables were all but psylocibin.

LSD Logistic Regression Prediction

|  |  |  |
| --- | --- | --- |
| Predict | No | Yes |
| No | 1696 | 1109 |
| Yes | 1860 | 3444 |

Ecstasy/Molly Logistic Regression Prediction

|  |  |  |
| --- | --- | --- |
| Predict | No | Yes |
| No | 1693 | 879 |
| Yes | 2108 | 3429 |

The LSD model contained a testing error of 0.366 while the Ecstasy/Molly model had a test error rate of 0.368. This does not provide a good indication that trying the ‘gateway’ drugs will result in the use of LSD or ecstasy/molly.

Conclusion:

The models used provided nice descriptions of how a single drug may influence the use of another. The multiple linear regression showed us which variables where useful for the age prediction of LSD and ecstasy/molly. Those results made a lot of sense because cigarettes and alcohol were not significant enough. The logistic regression model was decent but maybe a poor model to evaluate this dataset. In the future I wish to try other models, possibly some more advanced ones, and see which performs the best for drug use. Overall, the data analysis gave a lot of answers as to which drugs influence the use of another drug.

References:

1. <https://fivethirtyeight.com/features/how-baby-boomers-get-high/>
2. <https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2016-nsduh-2016-ds0001-nid17185>
3. An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (ISLR)