Drug Use by Age

CMSE 381 Final Project

Mira Ghazali

71

INTRODUCTION

Our perception of drug use is continuously changing in the United States; this, in turn, makes it very difficult to study the frequency and usage of drugs on a particular generation. However, a study was conducted in 2012 to explore the drug and alcohol habits of the baby boomer generation. Using the data provided by this study, we will look at different patterns and correlations within the information using modeling methods such as Cross-validation and the Bootstrap method. Particularly, we will be examining the relationships that may exist between marijuana use and use of hallucinogenic drugs.

An article published in 2002 from the John Hopkins Bloomberg School of Public Health revealed that there was epidemiologic evidence to support the notion that young marijuana smokers would be more open to try hallucinogenic drugs. This could be due to the nature of the social environment that marijuana smokers find themselves in, or due to other internal factors.

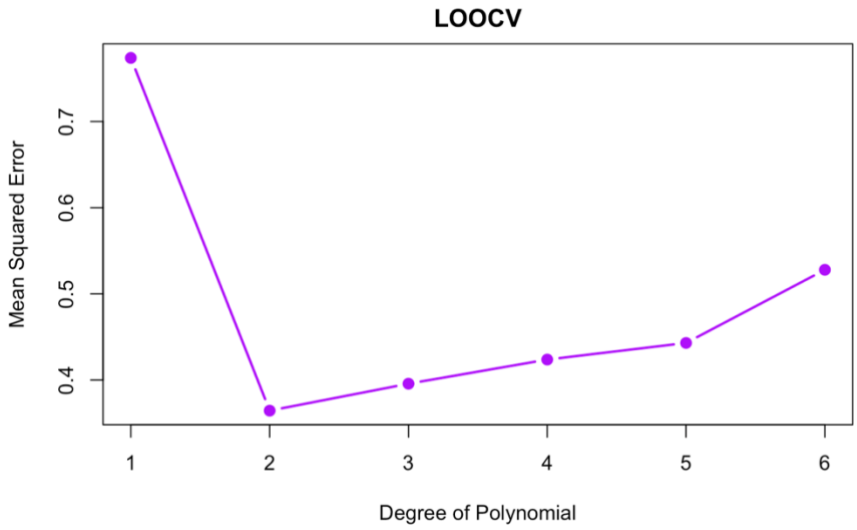
THE VALIDATION SET APPROACH

We first split our data into training and testing data: this allows us to build our model with one subset (training) of the data and then apply it to our testing data. In doing so, we can calculate our testing error and thus, preventing overfitting from the model. This approach provides evidence to support that a model using a quadratic function will perform better than a linear or cubic one.

LEAVE-ONE-OUT CROSS-VALIDATION (LOOCV)

Although a computationally expensive method to perform, Leave-one-out Cross-Validation (LOOCV) can reliably produce an estimate of our model's performance. This method would be appropriate to use on our dataset as it is good for small data and relatively inexpensive models to fit. When performing LOOCV on our Drug Use by Age dataset, we can see that the largest drop

in our mean squared error occurs between the linear model and the quadratic model. After the second degree, the MSE stays roughly constant.

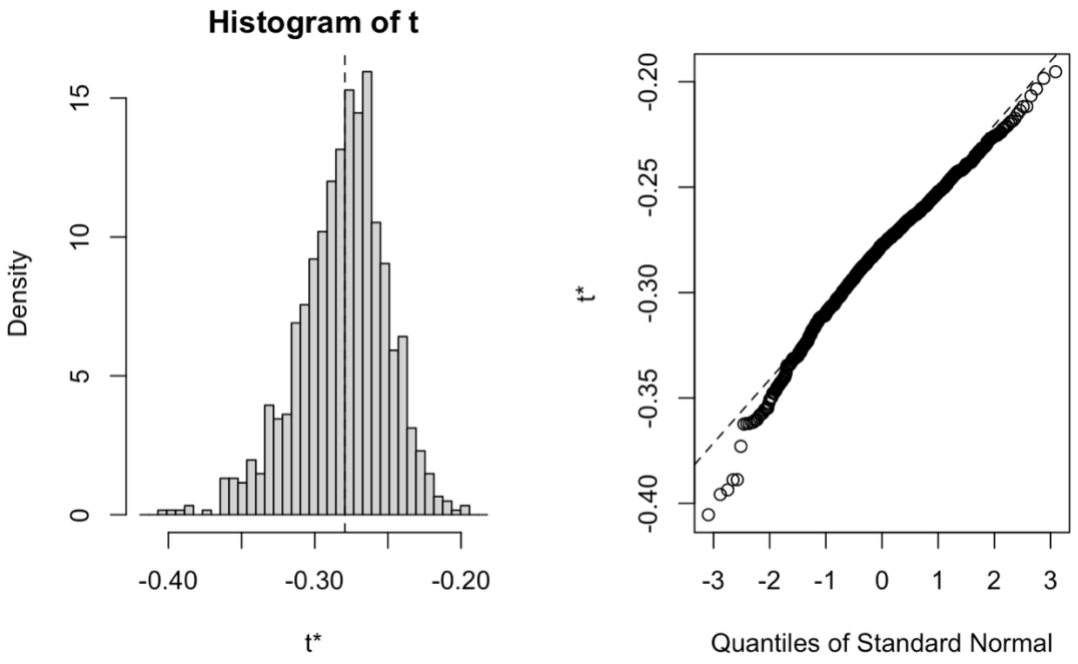


THE BOOTSTRAP

The Bootstrap is a resampling method that uses random sampling (with replacement) to determine the accuracy of sample estimates. We first create an alpha function that calculates the value of alpha that minimizes the variance:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

where $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$, and $\sigma_{XY} = Cov(X, Y)$. With this, we can generate a histogram of 1,000 bootstrap samples from our single data set. Having a large number of bootstrap samples being generated allows us to know with great accuracy the estimates of our parameters.



t is equivalent to α*

CONCLUSION

This project tackled the questionable correlation between marijuana users and their potential use of hallucinogenic drugs. This method used resampling methods such as Leave-one-out Cross-validation and Bootstrapping in order to reveal some sort of connect between the two behaviors. These results showed that there is no definitive clear link between the two, but further research and a deeper analysis must be conducted before declaring with certainty that a relationship does not exist.

REFERENCES

[1] https://fivethirtyeight.com/features/how-baby-boomers-get-high/

[2] https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/

[3] https://en.wikipedia.org/wiki/Bootstrapping_(statistics)#Approach

[4] https://www.jhsph.edu/news/news-releases/2002/marijuana-hallucinogen.html