# Drug Use by Age Analysis

## Final Report

### 1 Introduction

Drugs are a very common substance in our world. Here in the US, alcohol is legal to drink at 21, marijuana has been made legal in over 10 states, including Michigan, and Oregon just made it so someone cannot be punished for a small amount of any drug no matter the kind. We will look at a dataset that gives us the information of drug users by there age. This project aims at fitting a model to best represent the likeliness of people, within an age group, using drugs. The methods performed are linear regression, ridge regression and lasso regression. The features of number of users, alcohol users, marijuana users and cocaine users will be used to predict the likeliness someone uses drugs at a given age.

### 2 Related Work

There are many repositories on Github that can be viewed, however most of these are just basic analysis of the data set. All through this course, (CMSE 381) we have learned various ways of manipulating and reducing data. We have learned many methods of fitting data and predicting variables. This project is a continuation of this course and the aim is to further enrich and apply the tools I have learned into a dataset.

### 3 Dataset

The dataset for this study comes from *The National Survey on Drug Use and Health from the Substance Abuse and Mental Health Data Archive.* There are 13 drugs across 17 age groups, also given is the sample size of the survey with respect to age groups.

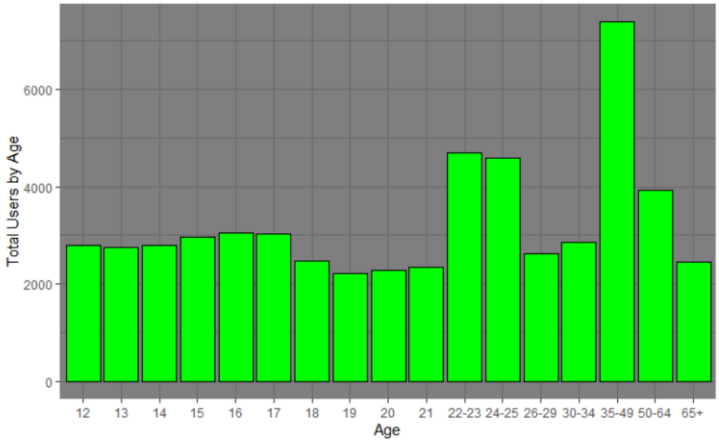

**Fig 1:** *The image shows the total number of drug users by Age group (n-samples)*

When cleaning up the data I noticed some missing values where related predictors were zero, here I had to fill in those missing values. When visualizing the data there was an extra column filled with NA values, this column would mess up the model if not taken care of. I removed the column and proceeded to look through the data. Originally, Age was a factor in our dataset, this was then changed to a numeric value. This order is: age 12 = 1, age 13 = 2, etc. the age groups 22-23 and so on are represented as the numbers 11 – 17. The data was split into a training and testing set, where the training is 70% of the original data and the testing is the remaining 30%.

### 3.1 Feature Selection

Many feature selection methods were used to find the best predictors of the data. When choosing features, we look for features that are the most significant to reduce our variance. First method was fitting a linear model over the entire dataset. The results weren't shocking when alcohol came out as the most significant predictor, the next best predictor was marijuana use. And the other predictors were not significant at all.

The second method was fitting a ridge regression, to penalize our model and decrease bias. The model performed had a high $R^2$ score and reduces the Root Mean Squared Error (RMSE) from the linear model. The next method was performing a Lasso regression on the data which adds a penalty term. This eliminates some coefficients and centers the data around a central point. This method produced a very high RMSE and a lower $R^2$ score than the previous method.

Finally, the features chosen were alcohol use, marijuana use, cocaine use and pain reliever use. These features showed to be the most significant in the data and showed moderately high correlation.

| | age | number.users | alcohol.use | marijuana.use | cocaine.use | pain.releiver.use |
|---|---|---|---|---|---|---|
| age | 1.0000000 | 0.3818455 | 0.6832030 | 0.1997115 | 0.3835299 | 0.2272118 |
| number.users | 0.3818455 | 1.0000000 | 0.8555999 | 0.3628856 | 0.4347998 | 0.5312816 |
| alcohol.use | 0.6832030 | 0.8555999 | 1.0000000 | 0.6370241 | 0.7460313 | 0.7388744 |
| marijuana.use | 0.1997115 | 0.3628856 | 0.6370241 | 1.0000000 | 0.9111500 | 0.9598851 |
| cocaine.use | 0.3835299 | 0.4347998 | 0.7460313 | 0.9111500 | 1.0000000 | 0.9346423 |
| pain.releiver.use | 0.2272118 | 0.5312816 | 0.7388744 | 0.9598851 | 0.9346423 | 1.0000000 |

**Fig 2:** *The image shows the correlation between each predictor.*

## 4 Methods

Linear regression was used as a base model and used for picking features. After linear regression, then came ridge using L2 regularization and Lasso using L1 regularization. All these tests were performed under various packages in RStudio. Cross-validation was used to determine the best lambda values for regularization.

### 4.1 Linear Regression

Linear regression was used to determine the relationship between predictors. For this prediction I wanted to look at the percentage of which a specific age group will consume one of the four drugs. The regression line falls under the simple form of $f(x) = B_0 + B_1 * x$. The error produced is the distance from the data point to the line.

### 4.2 Ridge Regression

Ridge regression utilizes L2 regularization in which it will add and L2 penalty which equals the square of the magnitude of coefficients. From there the coefficients are shrunk by the same factor which means no coefficients will be eliminated. Ridge regression is good to use because it will decrease the variance and produce better results.

### 4.3 Lasso Regression

Unlike Ridge Regression, Lasso regression utilizes the L1 regularization which equals the absolute value of the magnitude of coefficients. L1 regularization can lead to sparse models with few coefficients because the penalty will remove some coefficients. A larger penalty will produce coefficients closer to zero, which is good for a simple model. This is a good model to use because it can be easier to interpret results using lasso because it removes some coefficients. Lasso regression algorithm is $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j) + \lambda \sum_{j=1}^{p}|\beta_j|$. The goal is to minimize the sum of squares constraint, the closer to zero means there is a regression model that is easier to interpret.

## 5 Experiment and Discussion

Root Mean Squared Error and Adjusted $R^2$ score were used to evaluate each model. The data was split into 70% training and 30% validation. This was done to test the models by comparing the predicted values from the model to the validation set. The results are listed below.

**Linear Regression:**

Train: RMSE = 0.18, $R^2$ = 0.96 Test: RMSE = 0.12, $R^2$ = 0.93

**Ridge Regression:**

Train: RMSE = 0.29, $R^2$ = 0.9828325 Test: RMSE = 0.39, $R^2$ = 0.9828254

**Lasso Regression:**

Train: RMSE = 1.04, $R^2$ = 0.78 Test: RMSE = 1.55, $R^2$ = 0.72

The model that had the lowest variance was the Linear model based on the RMSE scores being the closest to zero. Linear and Ridge performed the best whereas Lasso performed the worst since the $R^2$ score is significantly lower than that of Linear and Ridge. As for the model that performed the best it is clear to see that Ridge regression with L2 regularization fitted the data the best and we can say with over 98% certainty that the model is accurate. Shrinking the coefficients without removing any was the better way of fitting this data.

## 6 Conclusions and Future Work

This project aimed at finding the likeliness an age group is to consume drugs. Using the tools learned in CMSE 381 including Linear regression, Ridge regression, Lasso regression, cross-validation and more. These models were used to determine the best results for Root Mean Squared Error and Adjusted $R^2$. initial fitting of the model leads me to conclude that most predictors are unnecessary which would lead to high model variance and overfitting of the data. Ridge regression was deemed necessary as it would reduce the coefficients of our model for a better fit. This was proven when Ridge regression performed the best out of all three models.

For future works it would be interesting to do a Principal Component Analysis to see which features are closely related and to determine the differences between the two components. Another interesting thing to be looked at is the frequency rates given in the data. From the frequencies we can ask questions like, “How often are 13-year-olds smoking weed?” or “What is the Increase of frequency from each age group?” These could be interesting things to look at and discuss because from the frequency we can start thinking about drug addiction. Another analysis that could be done is K-means clustering in which we could classify groups based on the data points and distinguish different types of drug users.

**References**


1. Stephanie. “Ridge Regression: Simple Definition.” *Statistics How To*, 6 Feb. 2021, www.statisticshowto.com/ridge-regression/.
2. Stephanie. “Lasso Regression: Simple Definition.” *Statistics How To*, 16 Sept. 2020, www.statisticshowto.com/lasso-regression/.
3. Swaminathan, Saishruthi. “Linear Regression - Detailed View.” *Medium*, Towards Data Science, 18 Jan. 2019, towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86.
4. Moody, James. “What Does RMSE Really Mean?” *Medium*, Towards Data Science, 6 Sept. 2019, towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e.
5. Frost, Jim, et al. “How To Interpret R-Squared in Regression Analysis.” *Statistics By Jim*, 13 Apr. 2021, statisticsbyjim.com/regression/interpret-r-squared-regression/.