PSTAT 126 Project

Jeff Shen, Kenneth Villatoro, Christopher Hong, Gavin Wolfe, Omer Randhawa

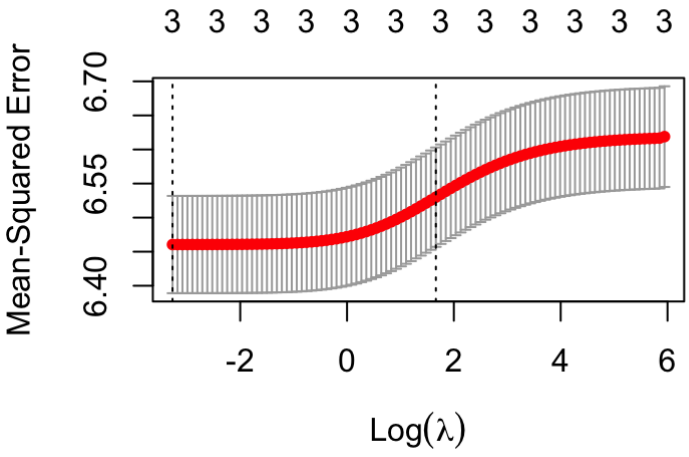
2023-06-05

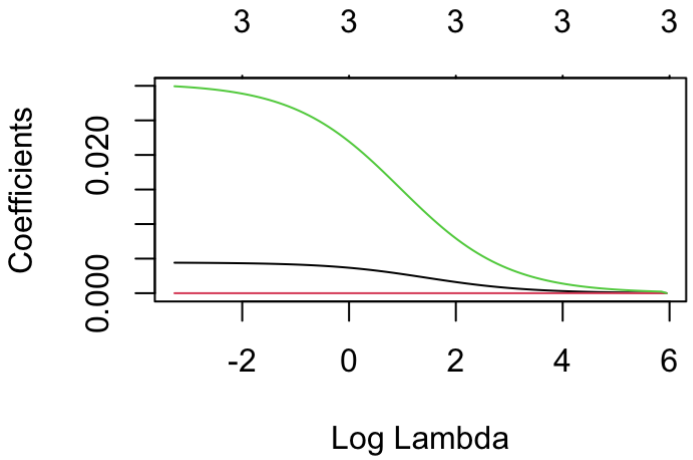
STEP FOUR

Introduction

In this project, we will analyze the relationship between income and education using the adult.csv dataset. Rather than simply comparing income levels and education levels, we will aggregate the data by calculating the percentage of individuals with income higher then 50k for each education level. This will allow us to create a quantitative variable for the proportion of individuals in each education level who earn a high income, which we can then use to explore the relationship between education and income. The dataset of interest comes from the UCI Machine Learning Repository and is available on Kaggle as well at: https://www.kaggle.com/datasets/uciml/adult-census-income (https://www.kaggle.com/datasets/uciml/adult-census-income). It contains information about individuals from the 1994 Census database, including demographic variables such as age, education, marital status, occupation, and more.

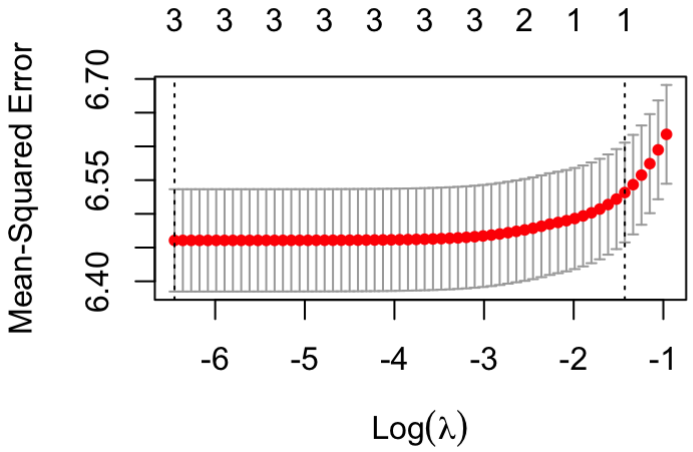
RR and LASSO






```
## [1] "Best Lambda:  0.0381072514054624"
```

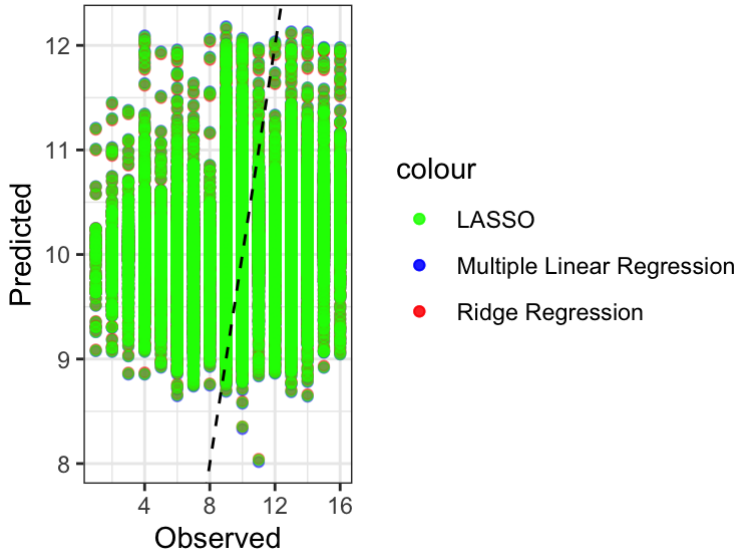
```
## [1] "R^2:  0.0241178571995193"
```




```
## [1] "Best Lambda:  0.00157459635341683"
```

```
## [1] "R^2:  0.0241178571995193"
```

MLR, RR, LASSO



Conclusion

The data collected above provides an insight in how the Ridge Regression and LASSO regression are implied in order to show the types of shrinkage methods used in order to optimize the coefficients for prediction. We see that for Ridge and LASSO regression, the R Squared value for both of these methods is 0.0004896063, thus implying a low variability among the coefficients when shrunk towards 0 (ridge) or at 0 (LASSO). The graph showing the comparison between Multiple Linear Regression, Ridge Regression, and LASSO regression shows that the observed value and the predicted value are almost nearly identical as the abline shown on the plot is nearly vertical.

Bootstrapping

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = income_data, statistic = r_squared, R = 100, formula = y ~
##     x)
##
##
## Bootstrap Statistics :
##       original  bias    std. error
## t1* 0.02412269       0           0
```

```
## [1] "All values of t* are equal to  0.0241226877537391"
```

```
## [1] "All values of t are equal to  0.0241226877537391 \n Cannot calculate confidence
intervals"
```

```
## NULL
```

The reason we chose to use the Bootstrap model is because this model is a non parametric estimation method ideal for when the distribution of a statistic is unknown or complicated. In our dataset, final weight is classified as a complicated statistic being estimated from multiple parameters. Luckily, the bootstrap method does not ask for specific distribution methods, so finding the correlation between for example education and final weight can be achieved. The output of our bootstrap gives us a better understanding of this comparison as opposed to other methods.

Bootstrapping is mainly the re sampling of the data provided in order to create more stimulated samples. The purpose of this is to calculate standard errors, t values, and create confidence for a wide set of stimulated samples. It also calculates how variable the model parameters due to the small changes in data values. This affects the regression coefficients and the variation of the parameters.

The technical specifications for this model can be broken down into a few steps. Firstly being how many samples needed to be performed in order to generate an ideal bootstrap statistic. Next for each sample, we need to specify the sample size. We chose to do 400, thanks to the professor's recommendations as to a quality sample size in class. The next specification is that the bootstrap model calculates a statistic of interest for each sample. Lastly, the mean is calculated from each sample statistic. Without these specifications, the bootstrap statistic cannot be generated correctly. Our bootstrap sample generated what seems to be an appropriate statistic of interest. In our case a t value of 0.024 was generated.