CDA Final Project

Joshua Freeman, Coco Kusiak, and Luke Toomey

12/12/2017

Contents

Introduction

Motivation

Depression and anxiety disorders are the most common mental illnesses in the United States ("Depression Statistics" 2017). Without proper treatment, these conditions can become chronic diseases and lead to increased risk for mortality ("Mental Health" 2016). Although many treatments are available for these conditions unfortunately, only about 37% of those with anxiety seek treatment ("Depression Statistics" 2017). Having just one depressive episode leaves the afflicted person with a 50% of experiencing another ("Mental Health" 2016). We set out to see what is predictive of having poor mental health in the average American.

The Data

Results

Missing Data

Automatic Model Selection Methods

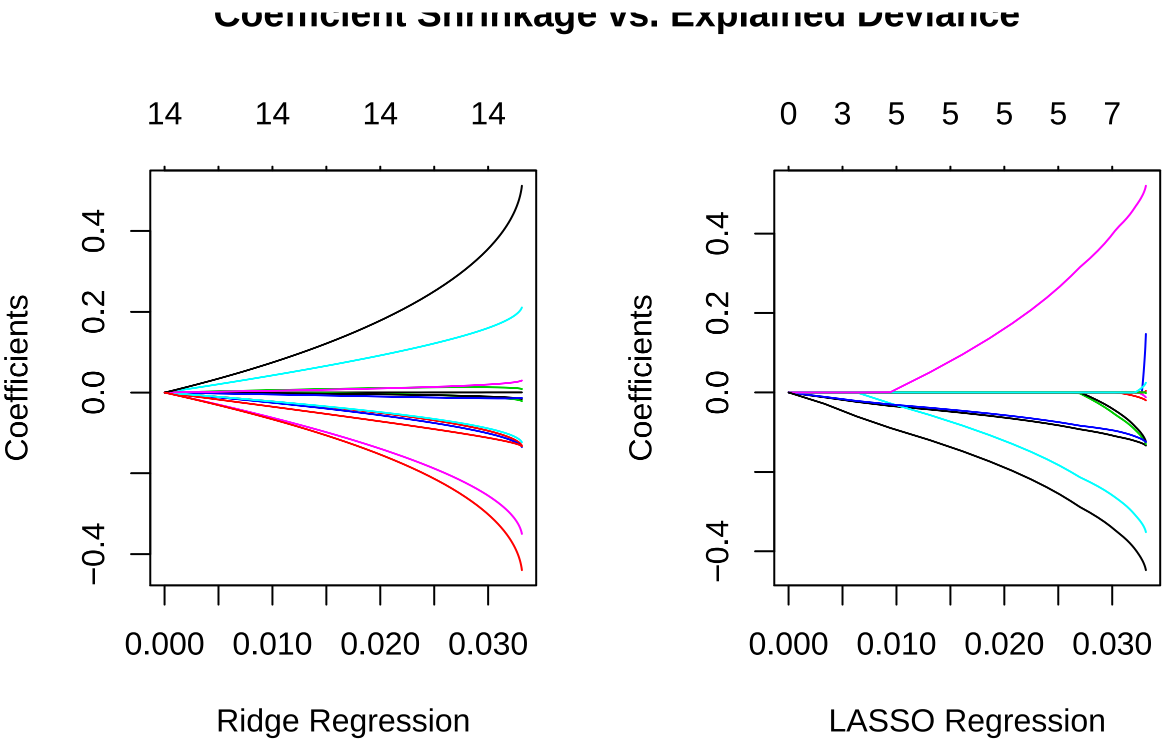
Introduction

The goal of my project is to test the performance of the different model selection algorithms ridge, LASSO, and stepwise regressions. LASSO and Ridge regressions are both forms of coefficient regularization. In addition to minimizing the residual sums of squares, these methods penalize covariates in our model based on some constraints. For ridge regression this can be written as: $\sum_{j=1}^{p} \beta_j^2 < c$ with p = # of predictors and c a constant. For LASSO, this can be written as $\sum_{j=1}^{p} |\beta_j| < c$ ("Regularization: Ridge Regression and Lasso" 2006). The important distinction between these two methods is that the LASSO method acutally drops predictors out of the model while the ridge method only shrinks coefficients close to 0.

Stepwise regression adds or removes variables based on which improves the model's AIC. Forward selection begins with no predictors and tests which one additional variable will improve the model. This continues until AIC improvements stop. Backwards selection begins with all available predictors and removes them one-by-one until AIC improvements stop ("Stepwise Regression" 2016). For the purpose of this study, we will focus on a combination of both forward and backward selection.

Step 1: Performance Assessment Based on Observed Data

To begin, I fit these three algorithms on the full data.



The plot above show the coefficient shrinkage as a function of the model's fraction of explained deviance. Each curve represents a predictor. At the far right, all predictors are unpenalized. As you move to the left, the explained deviance decreases as the coefficients shrink closer to 0. The numbers along the upper x-axis represent the number of predictors remaining in the model. The Ridge model maintains all 14 variables even when the explained deviance is essentially zero. At this point in the LASSO plot, 0 predictors remain. The Ridge method shrinks coefficients very *close* to 0, but the LASSO method drops some coefficients exactly to 0, removing them completely from the model.

	variable	Stepwise	LASSO	Ridge
1	binge	0.43	0.48	0.27
2	bmi	0.02	0.00	-0.01
3	exer30	-0.09	-0.10	-0.09
4	male	-0.50	-0.41	-0.23
5	sleep	-0.11	-0.13	-0.07

As shown on the previous page, most of the coefficient estimates are similar between the algorithms. However, for `bmi` stepwise and ridge estimate oppisite directions of association with the probability of having poor mental health, while the LASSO method drops it out of the model completely.

Next, I run 10-fold cross validation for each of these selection and the results are shown below.

	Mean.AUC	SD.AUC
Stepwise	0.57	0.00
LASSO	0.52	0.00
Ridge	0.50	0.00

As can be seen above, the stepwise method yields the highest area under the curve across the 10 iterations. This model will be used in subsequent analyses.

Step 2: Simulation Study

Conclusions

Random Effect Models

References

"Depression Statistics." 2017. *Anxiety and Depression Association of America.* https://adaa.org/about-adaa/press-room/facts-statistics.

"Mental Health." 2016. *CDC 24/7: Saving Lives, Protecting People.* https://www.cdc.gov/mentalhealth/basics/mental-illness/depression.htm.

"Regularization: Ridge Regression and Lasso." 2006. *Stanford Department of Statistics.* http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf.

"Stepwise Regression." 2016. *NCSS Statistical Software.* http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf.