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

Brief Take-aways

- talk about important variables and their direction of association

Missing Data

Automatic Model Selection Methods

Introduction

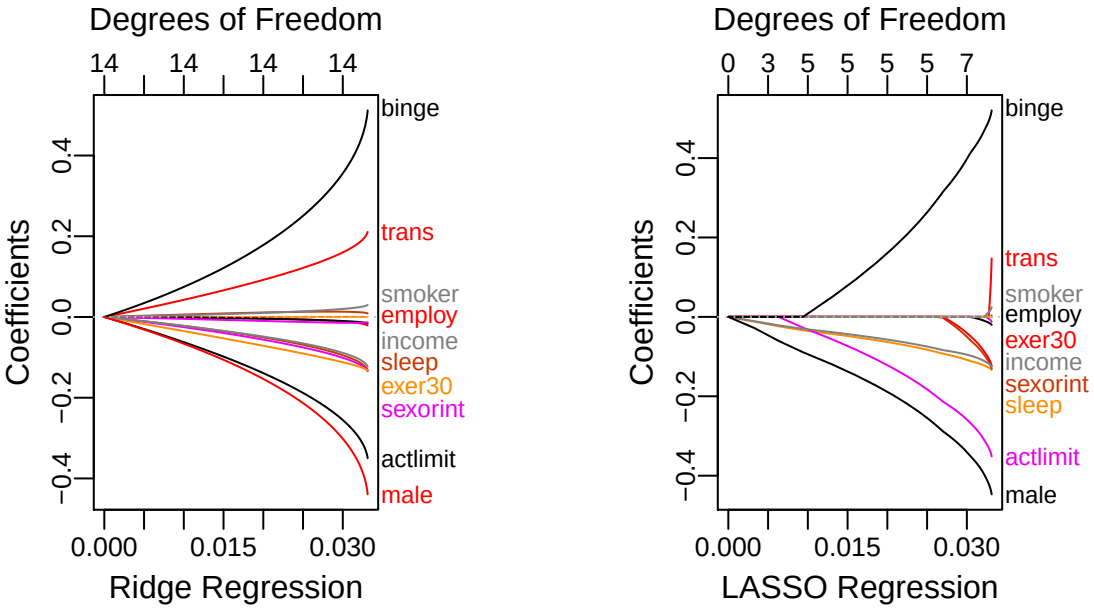
The goal of my project is to test the performance of the different model selection algorithms of ridge, LASSO, and stepwise regression. LASSO and Ridge regressions are both forms of coefficient regularization. In addition to minimizing the residual sums of squares, these methods penalize covariates in our model based on some constraints. For ridge regression this can be written as: $\sum_{j=1}^{p} \beta_j^2 < c$ with p = number 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 begins with a set of candidate predictors and adds or removes variables based on improvements to 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. This can be done usind the the `both` option in the

Step 1: Performance Assessment Based on Observed Data

Specifications

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	0.4266	0.4802	0.2678
2	bmi	0.0201	0.0000	-0.0129
3	exer30	-0.0860	-0.0953	-0.0946
4	male	-0.5044	-0.4078	-0.2276
5	sleep	-0.1112	-0.1250	-0.0732

As shown above, most of the coefficient estimates are similar between the algorithms. However, for `bmi` the stepwise and ridge methods estimate oppisite directions of association with the probability of having poor mental health, while the LASSO method drops it out of the model completely. However, across the three methods, it does not seem to have a large effect on mental health. `Exer30` has consistent estimates through the algorithms. Each estimate a coefficient of about -0.10 with $e^{-0.10} = 0.90$. This can be interpreted to mean that those who have exercised within the past 30 days have a 10% lower odds of having experienced poor mental health.

Again the ridge model keeps all 14 variables. The stepwise models also maintains all of the variables except for `metro` and `trans`. The LASSO menthod keeps only 10, removing `bmi`, `actlimit`, `race`, `state`, and `trans`.

Performance

The previous analyses were trained on the entire data. To test each algorithm's performace, a 10-fold cross validation is run. This method divides the data into 10 partitions. For each partition, the mehtods are trained on 90% of the data and tested on the remained 10%. For each model and each iteration, a receiver operating characteristic curve is created. The average area under the curve for model is display above. We use this metric to as our measure of performance. The results are displayed below.

	Mean.AUC	SD.AUC
Stepwise	0.57000	0.00165
LASSO	0.51600	0.00136
Ridge	0.50300	0.00126

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

The next step in this investigation is running a simulation study. A simulation study is a computer-based experiment involving randomly generated samples of the data (Wolfson 2014). We will do this by sampling n times from the original data. We will then generate new mental health values using the model specified by the stepwise selection algorithm outlined above. This means that in this case, we will know the true outcome generating process. This model icludes factors levels for every state, employment, and income level, thus an abridged version is written below:

$$logit(mental) = 1.10 - 0.29(actlimit = 1) + 0.42(binge) + 0.02(bmi) - 0.21(employ = 2) + 0.38(employ = 3)$$

$$-0.50(male) - 0.11(sleep) + 0.21(trans = 1)$$

In order to test how the performance of these algorithms depend on sample size, we reran them on samples of size 10,000 and 100,000. The means and standard deviations across these iterations are displayed below.

Note: These datasets include all 14 variables originally available for the models to choose from. This means that this includes `metro` and `trans` which were not variables used to generate outcomes from the stepwise model.

	Method	Mean.10.000	SD.10.000	Mean.100.000	SD.100.000
1	LASSO	0.50140	0.00209	0.51070	0.00328
2	Ridge	0.50000	0.00000	0.50000	0.00001
3	Stepwise	0.57050	0.00669	0.57050	0.00165

Conclusions

Our analyses have consistently shown stepwise regression to have the highest performance in terms of prediction for this dataset. It steadily resulted in the highest area under the curve.

Stepwise selection has been criticized since the stronger computing power has enabled us to use more computationally intensive methods such as LASSO selection. Stepwise's main advantages are that it is easy to implement and easy to understand(AdamO 2015). However, it is associated with a lot of disadvantages as well. Firstly, the stepwise approach is highly dependent on its initial set of candidate predictors. It is known to fall victim to finding spurious assocations (gung 2012). Because it makes decisions at every step, it can make choices that are "locally optimal, but suboptimal in general" (user20160 2016). These arbitrary choices can cause "severe biases in the resulting multivariable model fits while losing valuable predictive information from deleting marginally significant variables" (AdamO 2015). Thus overall, this method can result in strong biases and deceptive findings.

So why has the stepwise algorithm outperformed the others with this data? This could perhaps be because although there is a large sample size, there are only 14 predictors to parse through. LASSO regression typically does better than stepwise and ridge regression methods when the number of predictors, $p > n$. This is not the case in this study. Often a concern for stepwise regression is overfitting which is also likely to occur when $p > n$. Here though, this is not a worry. However, it is not completely surprising that in this context, the stepwise selection yields different results that the others becuase its algorithm is based on AIC rather than the penalization restraints used by LASSO and ridge regressions.

Overall, although this study found the stepwise algorithm to have the best performance, I would not say this method is the best selection by any means. Just because its model performed well here, it does not mean it will in other contexts. One huge disadvantage to stepwise is the computing time necessary with large sample sizes. For samples sized below 1,000 and 10 iterations, it fit in just under a minute. However, with sizes over 5,000, the time proliferated. The simulations with $i = 100$ and samples of 100,000 observations, the stepwise model took over 12 hours to run. Thus for larger samples and more iterations stepwise selection would be relatively infeasible in comparison to LASSO and rigde regression which took less than 10% of the time. It seems that each method may have its place in different situations. For example, ridge outperforms both LASSO and stepwise when the effects are better predicted by a combination of many weak predictors (user20160 2016). In conclusion, although stepwise regression performs well with these data, it is not the best model selection algorithm across the board and multiple methods should always be implemented and compared whenever it is practical to do so.

Random Effect Models

References

AdamO. 2015. "Superiority of Lasso over Forward Selection/Backward Elimination in Terms of the Cross Validation Prediction Error of the Model." Cross Validated. https://stats.stackexchange.com/q/89219.

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

gung. 2012. "Algorithms for Automatic Model Selection." Cross Validated. https://stats.stackexchange.com/q/20856.

"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.

user20160. 2016. "What Are the Advantages of Stepwise Regression?" Cross Validated. https://stats.stackexchange.com/q/218223.

Wolfson, Julian. 2014. "Simulation Studies & Reproducibility." *PUBH 8400: Research Skills in Biostatistics*.