Kelvyn Bladen

Stats 6100 – Dr. Cutler

Project Testing the Elastic Net & GroupLasso Against Lasso

Introduction

The purpose of this project is to test and expound on many of the claims made by Hui Zou and Trevor Hastie regarding the comparability of their Elastic Net to the Lasso technique for regression analysis. In their 2004 paper, "Regularization and Variable Selection via the Elastic Net", they suggest that the Elastic Net surpasses the Lasso when the data is highly collinear or when the number of predictors in a dataset far exceed the number of observations. They also claim that in addition to finding a method that can deal with the problems above, their 'goal is to find a new method that works as well as the lasso whenever the lasso does the best.' If this is true, it could be proposed that the Elastic Net may potentially replace the Lasso technique in statistical analysis. Using the Hald Cement, United States Air Pollution, Pulse Rate, SAS Leukemia, and Iowa Housing datasets, I aim to perform side-by-side comparisons of these two methods with the intent to demonstrate whether it is possible to replace the Lasso with the Elastic Net. If my results show that it is not always recommended to utilize the Elastic Net over the Lasso, I intend to offer conditions for when the Lasso is preferred with a clear explanation of its advantages over the Elastic Net.

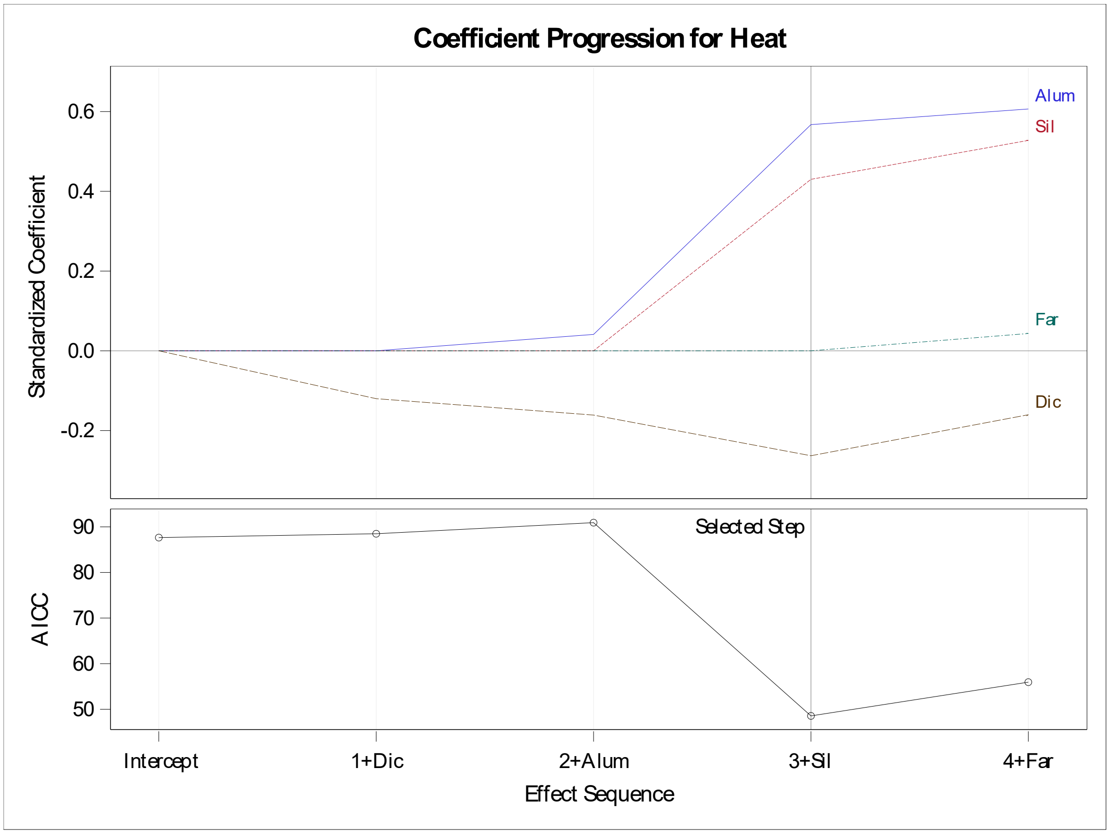
In addition to this primary research question, I also decided to compare the results from GroupLasso models to the Lasso and Elastic Net for each of these datasets. My objective is to assess their results to better understand the comparative advantages and disadvantages of the GroupLasso. Along with this, I hope to consider the sensibility of possibly merging the capabilities of the Elastic Net and GroupLasso together to potentially offer greater predictive accuracy.

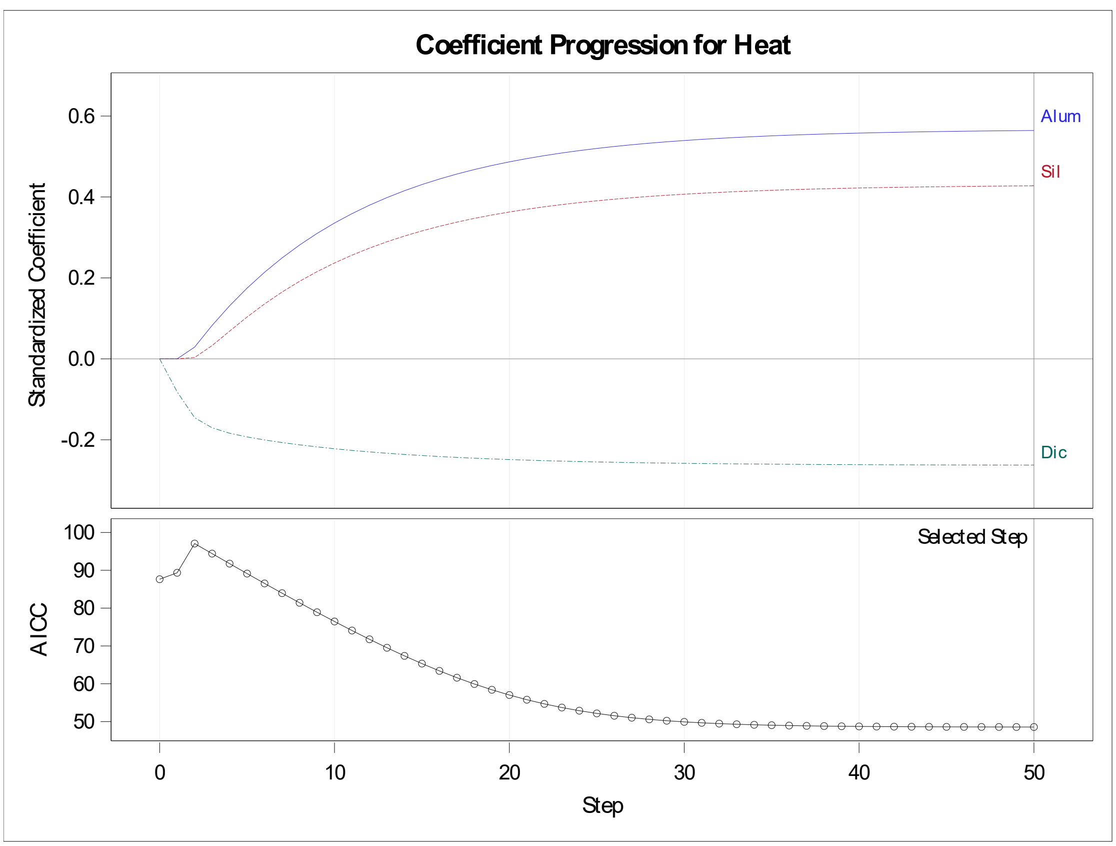
Hald Cement Data

The Hald Cement data is a small set with only 13 observations and 4 predictor variables. This set is characterized by extremely high negative collinearity. The Lasso is often dominated by ridge regression in instances such as this, especially since there is no reason to remove variables. The Elastic Net can be thought of as a methodology operating between the Lasso and Ridge regression. However, the results for the Elastic Net and Lasso are identical in this example, down to the Root Mean Squared Error of 2.309 and adjusted R-Squared of .9764. The GroupLasso method also has the same adjusted R-Squared value of .9764. But when compared to the other two techniques, I found that its RMSE was 2.311 and the coefficients are slightly different (see tables below). These differences are negligible and quite insufficient to suggest that the GroupLasso is inferior in this example. However, the fact that these values differ shows some algorithmic differences that should be acknowledged. The plots below show that the Lasso and Elastic Net consider only 4 steps: one per variable. The GroupLasso is shown to consider up to 50 steps which provide a smoothing technique where magnitudes gently change at each step. This fundamental difference between the Lasso and the GroupLasso could be substantial in future analyses.

Lasso & Net Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	71.670371
Alum	1	1.450851
Sil	1	0.415774
Dic	1	-0.236466

GroupLasso Parameter Estimates	
Parameter	**Estimate**
Intercept	71.830534
Alum	1.443165
Sil	0.413345
Dic	-0.235995


Coefficient Progression for Heat
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
Alum
Sil
Far
Dic
AICC
90
80
70
60
50
Selected Step
Intercept
1+Dic
2+Alum
3+Sil
4+Far
Effect Sequence


Coefficient Progression for Heat
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
Alum
Sil
Dic
AICC
100
90
80
70
60
50
Selected Step
0
10
20
30
40
50
Step

US Air Pollution Data

Next, I looked at the United States Air Pollution data which has 60 observations, 11 predictors, and possesses high collinearity like the Hald dataset. In this case, the collinearity is between two positively correlated variables: log(HC pollution) and log(Nox pollution). The effect is such that these variables' coefficients have opposite signs and high magnitudes when they ought to have the same direction and lower magnitudes. For this analysis, I used AICC selection as well as Cross-Validation. Since the AICC consistently outperformed the Cross-Validation R-Squared by about .03, I will just address the AICC results. In this example, the Lasso and Elastic Net again provide completely identical results, while the GroupLasso is slightly different but yields a model that is fundamentally the same. Selected variables and their coefficients are shown below. The corresponding adjusted R-Squared values are respectively .5920 and .5900. The models each were able to reduce the collinearity because of their selected factors, but it would be preferred to keep each of the pollution variables. Even though other techniques are likely preferred for this dataset, it is important to note that the Elastic Net and GroupLasso again match the Lasso.

Lasso & Net Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	1069.359045
Education	1	-18.115289
NonWhite	1	3.211678
lSO2	1	9.565873

GroupLasso Parameter Estimates	
Parameter	**Estimate**
Intercept	1068.136748
Education	-17.939264
NonWhite	3.190276
lSO2	9.428357

Run Pulse Data

The pulse rate data is another set we examined in class. Like the air pollution data, this set is characterized by somewhat high positive collinearity between the variables RunPulse and MaxPulse. This collinearity causes their coefficients to incorrectly have opposite signs. Using the same seed on the Lasso and the Elastic Net provided identical results from these methods. In this case, a model with 4 variables and adjusted R-Squared of .7810. However, when the seeds differed, the two methods often gave slightly different results usually characterized by the inclusion or removal of the Weight variable. The GroupLasso again provides slightly different results with the same variables but an adjusted R-Squared of .7801 being affected by marginally different coefficient estimates (see below). Unlike the pollution data, these Lasso models work very well in their ability to effectively remove the collinearity problem and maintain a relatively strong accuracy. Thus, I have seen that the Elastic Net and GroupLasso method effectively match the Lasso when it works well compared to other methods (Pulse Data) and when it struggles compared to other methods (Pollution Data).

Lasso & Net Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	105.633579
RunTime	1	-2.758209
Age	1	-0.209678
Weight	1	-0.021307
RunPulse	1	-0.102638

GroupLasso Parameter Estimates	
Parameter	**Estimate**
Intercept	105.057372
RunTime	-2.757396
Age	-0.205843
Weight	-0.019698
RunPulse	-0.101104

Leukemia Data

The Leukemia Microarray data has 7129 predictors but only 72 observations. It has already been noted that the Elastic Net is specifically designed for situations where the number of predictors far exceed the number of observations. Much of the analysis for this data has already been performed in class, but I will include the GroupLasso to it and reiterate some of the previously found results.

Starting with the set that is split into training and validation data, it was found that the Lasso provided a model with 34 terms and an adjusted R-Squared of .9849. This model is limited by the number of observations in the dataset and cannot have more terms than observations. After running the GroupLasso, I found that it had an identical R-Squared to the Lasso and the exact same set of terms were selected in both models. Again, the coefficient estimates were not identical, but the differences were negligible. However, I observed that computation time for the GroupLasso was sizably longer than for the Lasso. Consequently, this may be a liability when considering much larger datasets. Meanwhile, the Elastic Net chooses a model with about 54 terms and yields an adjusted R-Squared of 1.00. This model is preferred for its accuracy and its ability to select more variables than the Lasso, which seems to suggest that if variable retention is important, the Elastic Net may have a strong advantage.

Using the combined dataset and Cross-Validation, I found that the Lasso now yields a model with 42 terms and an adjusted R-Squared of .9664. The GroupLasso is extremely comparable with the same 42 terms and an adjusted R-Squared of .9688. The Elastic Net now has 88 variables with an adjusted R-Squared of 1.0573. Thus, the effectiveness of the Elastic Net is again demonstrated. For further study, the analysis of this dataset with the Elastic Net is discussed in greater detail in the previously mentioned paper by Zou and Hastie.

Categorical Iowa Housing Data

The Iowa housing data is possibly the most compelling of all the sets in my analysis. Unlike the other datasets, this housing data includes 39 categorical variables with its 26 continuous predictors. Consequently, the results for my three models are not as consistent as with other datasets. The adjusted R-Squared values for the Lasso, GroupLasso, and Elastic Net are respectively .7325, .8685, and .8064.

The success of the GroupLasso algorithm here is relatively simple. It has the capability to include categorical factors in its model, while it can be observed that the other methods do not retain any such variables. In theory, it should be possible for the Lasso and Elastic Net to consider categorical predictors by generating dummy variables associated with all but one level of each factor. Originally, I suspected that the GroupLasso was simply an expansion on the Lasso method to accomplish this objective. However, observations from the Hald analysis show that the algorithm has several other differences. It seems that the GroupLasso is characterized by the condition that either all levels of a categorical variable are included or none of them are. Consequently, I believe that the GroupLasso technique can be expanded in some way to work for the Elastic Net and could potentially be revised to use subsets of categorical variables rather than entire sets of levels.

The differences between the Lasso and Elastic Net values are surprising. The models below show that neither method includes any categorical variables, but the Elastic Net has a much broader selection of predictors. Vital take-aways from this include two prominent facts. First, these 5 additional predictors may offer greater insights about factors affecting housing prices. Second, they may also be providing unnecessary clutter to the model for only a small improvement in predictive accuracy. This second point is crucial to my research question of whether the Elastic Net is always preferred to the Lasso. However, before making a definite conclusion I will revise the housing data to only consider continuous variables and re-do the analysis.

Lasso Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	-25882
GarageCars	1	10360
KitchenQual_Ex	1	13089
OverallQual	1	18102
TotalSF	1	30.131408

Elastic Net Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	-463815
BsmtFinType1_GLQ	1	676.648370
ExterQual_TA	1	-2858.079827
GarageCars	1	6046.334028
KitchenQual_Ex	1	27189
LAln	1	5268.915149
OverallQual	1	15958
TotalSF	1	40.630089
YearBuilt	1	87.767978
YearRemodAdd	1	106.511181

GroupLasso Effects:	Intercept BsmtExposure BsmtFullBath ClyTile ExterQual Fireplaces GarageCars GarageFinish2 GLQ KitchenQual LAln MasVnrArea OverallQual OverallCond PcntBsmtUF PConc TotRmsAbvGrd TotalSF YearBuilt YearRemodAdd

Numeric Iowa Housing Data

A subset of the housing data that includes only numeric predictors yields a dataset of 1460 observations and 26 predictors. An examination of collinearity showed that all the variance inflation factors were below 5. Consequently, this dataset exemplifies the main conditions for which the Lasso should work well since there is little to moderate collinearity and the number of observations exceed predictors. I re-ran my 3 models to compare their variable selections and corresponding accuracy. Adjusted R-Squared values for the Lasso, GroupLasso, and Elastic Net were .7232, .8190, and .8032. However, the Lasso selected only 3 variables, while the

GroupLasso and Elastic Net both chose identical sets of 18 variables. This certainly expands on my observations from the analysis of the categorical datasets. The additional 15 variables+ could be useful to someone interested in evaluating specific aspects of a home to find direction, magnitude, and significance of individual predictors. But, from an interpretation and simplification standpoint, it is somewhat ridiculous to use 6 times the number of predictors in the Lasso to increase the R-Squared by .08 or .09. Most individuals interested in a quick and simple predictive equation would certainly prefer the Lasso model over the GroupLasso and Elastic Net in this case.

Conclusion

My objective in this line of research was to show whether the Elastic Net could entirely replace the Lasso and better understand why this could be possible or not. I also set out to better comprehend the relationship of the GroupLasso to the Lasso and Elastic Net, with the intent to consider the sensibility of combining the GroupLasso and Elastic Net techniques.

The cement, pollution, and pulse datasets show tremendous consistency between the three regularization methods. The Elastic Net demonstrated that it could match the Lasso coefficients and R-Squared exactly when both methods select the same set of variables. I suspect that this holds true always, but this is beyond my current research question. The GroupLasso was shown to provide exceptionally comparable results with no apparent advantages or disadvantages despite being quite different computationally. The leukemia dataset highlights the comparative benefits offered by the Elastic Net, since the number of predictors far exceed the number of observations. Notably, the GroupLasso appears to match the Lasso results when this condition exists, and only continuous variables are available. However, the leukemia analysis does illuminate the runtime shortcoming of the GroupLasso for vast datasets.

With these as an important background, the Iowa housing analysis provides a great deal of the essential information to my project. When running the Lasso and Elastic Net on its numeric variables, I found that their respective R-Squared values of .7232 and .8032 demonstrate a feature that I suspect may be always true: The Elastic Net offers equal or greater predictive power than the Lasso. If this is the case, then programmers with maximum accuracy as their goal may have strong reason to replace the Lasso with the Elastic Net. However, when interpretation and simplicity are the primary purpose, the Lasso's selection of 3 variables stands in stark contrast to the 18 chosen by the Elastic Net. Consequently, it is my conclusion that the Elastic Net is not always preferred over the Lasso. It seems to me that the defining conditions for selecting which model to use are less about data characteristics and more about statistical aspects. In general, the Elastic Net ought to be preferred for its accuracy and broader application, but when simplicity and interpretation are important points of consideration, the Lasso is a better option.

My observations of the full Iowa housing data analysis provide the final discussion points of this project. The GroupLasso was shown in that analysis to provide greater predictive power by including factorial variables in its model selection. As previously indicated, some algorithmic adjustments, to allow for the creation of dummy variables, in the Elastic Net could expand its application to include categorical predictors into a model of even greater accuracy. An important caveat to this idea is that several classification and machine learning techniques are already well equipped to handle categorical data with impressive accuracy. However, if an interpretable linear model may have some merit, there is ample reason to discover or create an algorithm capable of combining the capacities of the Elastic Net and GroupLasso together into a new regularization and variable selection method.