STAT 425
Case Study #2
Carrie Mecca, Charlie Marcou, Jessie Bustin

Introduction

As the proliferation of data increases, the usefulness of statistical modeling also increases. While more data would seem to always be a net positive when statisticians build models for either analysis or prediction, using more variables and, as such, more data might not always be ideal. The decision of whether more variables are worth the added complexity leads us to the model selection process. The value of model selection is pivotal for modeling through the noise and producing useful models. In the following executive summary, we will use a criteria-based model selection strategy and two shrinkage methods to predict crime rates for 47 states in the year 1960. We will then compare training and testing root mean square error (RMSE) across our three selected models and weigh the strengths and weaknesses of each.

Methodology

To begin our analysis, we checked the data for completeness and duplicate entries and then proceeded to complete a modest exploratory data analysis. The data is comprised of 47 observations with 15 variables pertaining to state demographics that we could use for modeling crime rates. One of the variables was a binary variable indicating whether the state is in the south. Depending on the chosen technique moving forward, this variable may be handled differently from the numeric predictors. We checked correlation among the numeric predictors and found three sets of highly correlated variables. For each set, we noted the variable that was the most highly correlated with the target. From the variables not as highly correlated with the target, we created a list of variables for potential removal. After this initial cleaning, validation, and exploration, we proceeded with our model selection and shrinkage techniques.

For our criteria-based approach, we utilized a Leaps and Bounds algorithm to reduce the number of predictor variables to fit a linear model. Our decision to use this technique was based on the small number of predictors in our data. This algorithm returns the global optimum compared to greedy algorithms that choose the best algorithm at each step instead of finding the global best. In a more computationally heavy scenario where many features are involved, a greedy algorithm could be used instead. For this model, we input the binary predictor as a factor variable so it would be treated categorically versus numerically. We also removed the set of variables that we identified in our exploratory analysis as causing collinearity issues. From the algorithm's output, we identified the combination of variables that produced the lowest BIC. A model was fit (*Figure 2)* with the eight variables identified, and then training and testing RMSE were calculated (*Figure 1)*.

For our second technique, we performed principal component regression (PCR). For this method, we kept all variables because collinear predictors can appear together in components without issue. With the training data, we utilized leave-one-out cross-validation to select the optimal number of principal components. We chose leave-one-out due to the small nature of the dataset, but would have used k-fold with larger data. We noted that even with all 15 principal components, only about 72% of the variance in our target variable was explained. Our process identified eight to be the ideal number of principal components. We evaluated the loadings or the proportions of each variable used in each component, but did not notice a discernable pattern. To wrap up our principal component regression, we calculated training and testing root mean square error for a linear regression with our eight most powerful components as the predictors (*Figure 1)*.

For our final method, we used a LASSO penalty in a linear regression to perform model selection, introducing bias in an effort to decrease variance. We again opted to remove the highly correlated variables. Ridge regression could have been utilized if collinear variables were included, but we valued the selection abilities of the LASSO penalty. We utilized cross-validation to tune the coefficient for the LASSO penalty, lambda. *Figure 4* shows a plot mapping the cross-validated mean square errors by conducting a grid search across the possible lambda values. We found an ideal lambda that selected four predictors, all of which had also appeared as desirable in our Leaps and Bounds selection. The coefficients for our selected lambda can be seen in *Figure 3*. We then calculated the RMSE from the training and testing sets (*Figure 1*).

Figure 1 shows a table of the errors we computed for each model selection or shrinkage technique. The feature selection in Leaps and Bounds returned the lowest training and testing RMSE. We do want to note that if we used all 15 variables with Leaps and Bounds, this does not remain true. By removing a subset of highly correlated variables, we saw a noticeable improvement to the model selected by Leaps and Bounds. In this scenario, PCR did not improve the training and testing RMSE. Perhaps if many more demographic variables were included, each component could capture underlying groupings in the data more efficiently. This could lead to each component accounting for more variability in the target. If simplicity is preferred, however, the LASSO penalized model with four predictors could be useful despite having the highest training and testing RMSE.

Conclusion

Across the three techniques we implemented to perform model selection, we recommend the linear model selected through Leaps and Bounds with the reduced variable set. However, prior to recommending this model for analysis, inference, or prediction, it would be important to check model diagnostics. While PCR and LASSO returned higher testing RMSEs, that does not diminish their usefulness. PCR might prove more interpretable and accurate if more variables were included and LASSO produced the most simplistic model and adjusting lambda could add or subtract complexity as desired. This process confirmed that having more predictor variables does not necessarily equate to better model performance.

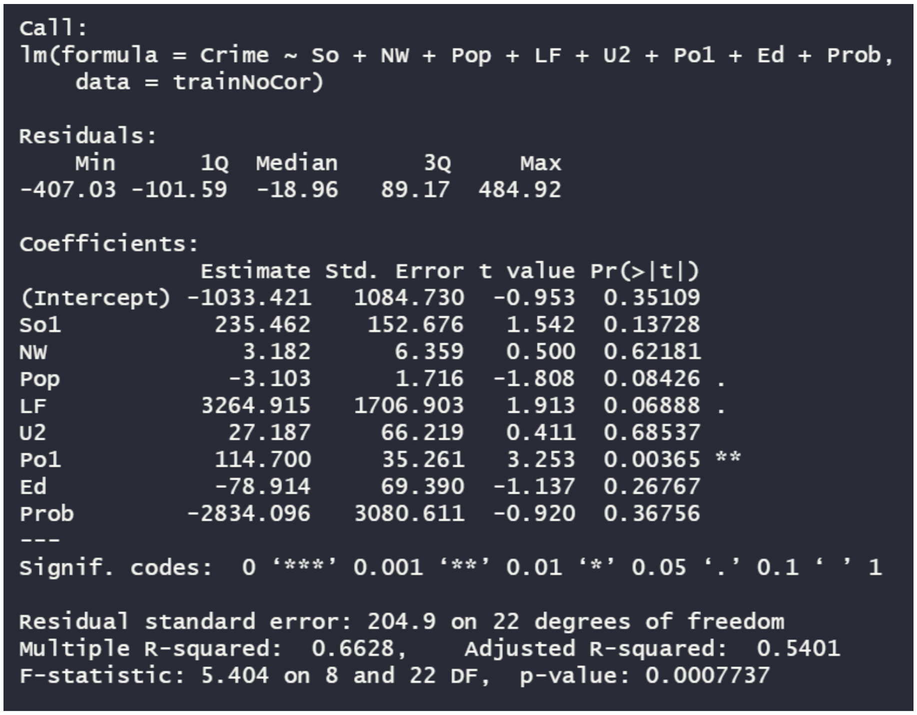
Appendix

Appendix A - Figures

Figure 1 - Summary Table of Models

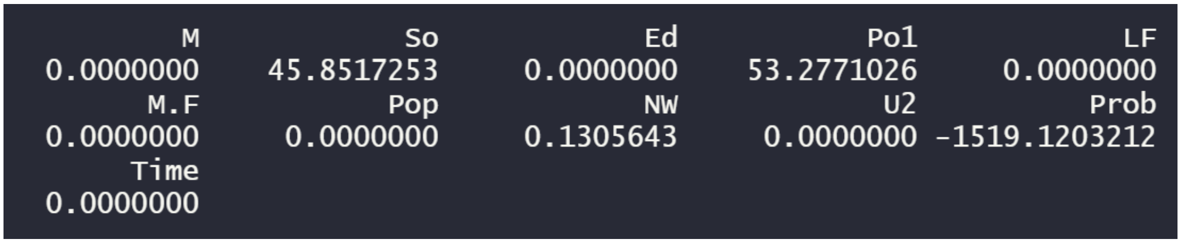
Model	Training RMSE	Testing RMSE	Variables Selected
Leaps & Bounds	172.6025	368.0904	8
Principal Component Regression	179.2845	373.6876	8 Principle Components
LASSO	213.8298	399.2151	4

Figure 2 - Leaps & Bounds Selected Model



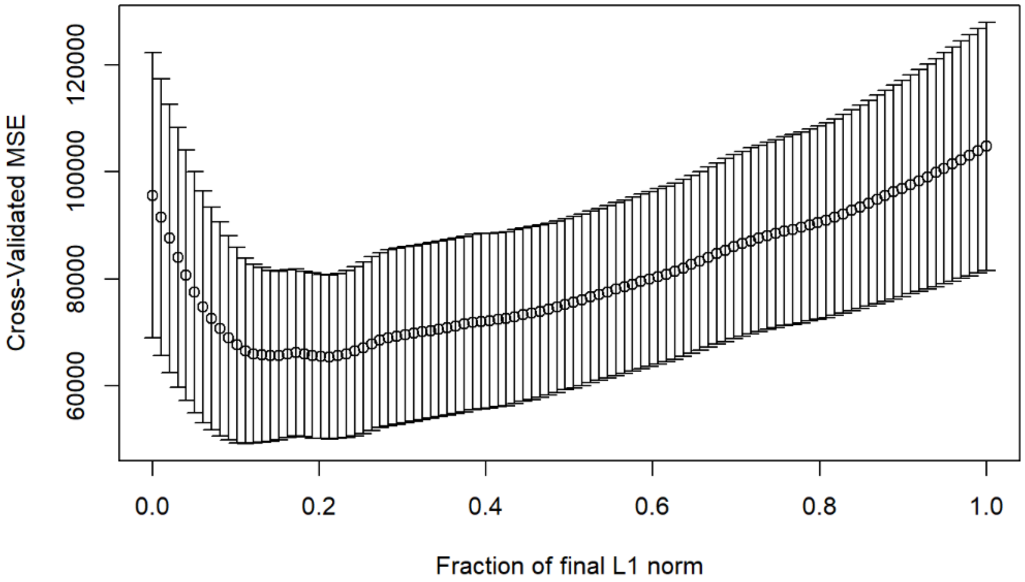
```
Call:
lm(formula = Crime ~ So + NW + Pop + LF + U2 + Po1 + Ed + Prob,
    data = trainNoCor)

Residuals:
    Min      1Q  Median      3Q     Max 
-407.03 -101.59  -18.96   89.17  484.92 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1033.421   1084.730  -0.953  0.35109   
So1           235.462    152.676   1.542  0.13728   
NW              3.182      6.359   0.500  0.62181   
Pop            -3.103      1.716  -1.808  0.08426 . 
LF           3264.915   1706.903   1.913  0.06888 . 
U2             27.187     66.219   0.411  0.68537   
Po1           114.700     35.261   3.253  0.00365 **
Ed            -78.914     69.390  -1.137  0.26767   
Prob        -2834.096   3080.611  -0.920  0.36756   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 204.9 on 22 degrees of freedom
Multiple R-squared:  0.6628,    Adjusted R-squared:  0.5401 
F-statistic: 5.404 on 8 and 22 DF,  p-value: 0.0007737
```

Figure 3 - Coefficients for LASSO Model



```
            M            So            Ed           Po1            LF 
    0.0000000    45.8517253     0.0000000    53.2771026     0.0000000 
          M.F           Pop            NW            U2          Prob 
    0.0000000     0.0000000     0.1305643     0.0000000 -1519.1203212 
         Time 
    0.0000000 
```

Figure 4 - LASSO CV Errors by Lambda



Appendix B - Project Logistics

The dataset, associated R code, a PDF of code output, and all project documentation can be found on github. Details about files are provided in the README file.

https://github.com/jbustinUIUC/stat425_casestudies

R Libraries Required for Reproducibility:
tidyverse, MASS, leaps, pls, lars

Questions pertaining to data collection and data sourcing can be directed to the Fall 2022 teaching staff led by Dr. Chronopoulou for STAT 425 at The University of Illinois - Champaign Urbana.

Further information about the analysis and model build can be directed to the project team.
jbustin2@illinois.edu - Jessie Bustin
cmarcou@illinois.edu - Charlie Marcou
cmecca2@illinois.edu - Carrie Mecca