Module 4 | Assignment: Regularization

Trang Tran

CPS, Northeastern University

ALY6015 | Intermediate Analytics

Professor Steve Morin

May 14, 2023

Introduction

We are looking at a college dataset from ISLR library that contains a large number of US Colleges from the 1995 issue of US News and World Report. The dataset has 18 variables with 777 observations. The variables dictionary attached as below:

Figure 1: Data Dictionary

`Private`
A factor with levels `No` and `Yes` indicating private or public university

`Apps`
Number of applications received

`Accept`
Number of applications accepted

`Enroll`
Number of new students enrolled

`Top10perc`
Pct. new students from top 10% of H.S. class

`Top25perc`
Pct. new students from top 25% of H.S. class

`F.Undergrad`
Number of fulltime undergraduates

`P.Undergrad`
Number of parttime undergraduates

`Outstate`
Out-of-state tuition

`Room.Board`
Room and board costs

`Books`
Estimated book costs

`Personal`
Estimated personal spending

`PhD`
Pct. of faculty with Ph.D.'s

`Terminal`
Pct. of faculty with terminal degree

`S.F.Ratio`
Student/faculty ratio

`perc.alumni`
Pct. alumni who donate

`Expend`
Instructional expenditure per student

`Grad.Rate`
Graduation rate

In this week's analysis, we focus on building regularization models using Ridge and Lasso techniques. The goal is to predict the 'Grad.Rate' variable using these models. Regularization methods such as Ridge and Lasso help us mitigate overfitting and improve the generalization performance of our models.

Regularization Analysis

To begin, we split the dataset into a training set (train_df) and a test set (test_df) with a ratio of 70/30 by using the sample() function. The train and test sets have 543 and 234 observations respectively.

Ridge Regression

For the Ridge Regression models, an alpha value of 0 was used, which corresponds to Ridge Regression. Cross-validation was performed to estimate the optimal lambda values, and the results were visualized through a plot [figure 2]. Table 1 summarizes the lambda.1se and lambda.min values, along with their corresponding logarithmic values and RMSE values of the train and test sets through 2 fitting Ridge models. We fit the Ridge regression models against the training set respectively based on both lambda.1se and lambda.min values.

The lambda.1se and lambda.min values showed a notable difference. Furthermore, Figure 2 displays the cross-validation plot, which provides insights into the relationship between the lambda values and the mean cross-validated error. This plot helps in understanding how significantly different lambda values affect the model's performance and can assist in selecting an appropriate lambda value.

The difference between the lambda.1se and lambda.min values suggests that the regularization effects of Ridge Regression can vary significantly depending on the chosen lambda value. This implies that different levels of shrinkage and model complexity can be highly achieved by adjusting the lambda parameter.

Ridge test	λ	Log	RMSE_train	RMSE_test	RMSE difference
lambda.min	2.347	0.853	12.567	12.994	0.427
lambda.1se	26.365	3.272	13.314	13.066	-0.248

Table 1: Ridge test results

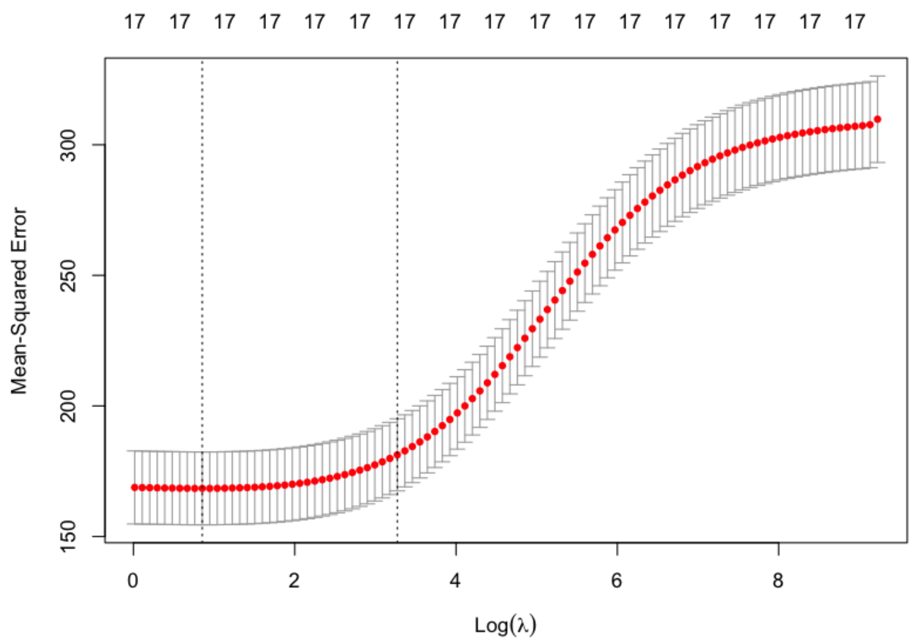


Figure 2: CV plot of Ridge regression

Figures 3 and 4 below represent the coefficients' result when fitting Ridge regression models against the training set.

Model 1: lambda.1se and alpha = 0

```
18 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept) 44.2459562178752
PrivateYes   2.7149197959226
Apps         0.0002038426698
Accept       0.0001622568708
Enroll      -0.0000001493571
Top10perc    0.0750116119249
Top25perc    0.0779706181780
F.Undergrad -0.0000431359957
P.Undergrad -0.0006867769853
Outstate     0.0004168618001
Room.Board   0.0012352701541
Books       -0.0004404550717
Personal    -0.0018403380590
PhD          0.0403536691307
Terminal     0.0165833125273
S.F.Ratio   -0.1750412907090
perc.alumni  0.1504594190886
Expend       0.0000689037003
```

Figure 3: Coefficients matrix of model 1

Model 2: lambda.min and alpha = 0

```
18 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept) 39.70948618567
PrivateYes   4.26030505018
Apps         0.00070672323
Accept       0.00025117845
Enroll      -0.00020706664
Top10perc    0.07405929634
Top25perc    0.13375982269
F.Undergrad -0.00004685279
P.Undergrad -0.00115378075
Outstate     0.00061526316
Room.Board   0.00198931740
Books       -0.00058728780
Personal    -0.00290342939
PhD          0.08487560256
Terminal    -0.07488589418
S.F.Ratio   -0.17003185797
perc.alumni  0.28157822355
Expend      -0.00029627547
```

Figure 4: Coefficients matrix of model 2

Even though there is not a big difference between the 2 coefficients' matrices, the Ridge model with lambda.1se value shows better performance on prediction when comparing the RMSE differences in Table 1 (-0.248 < 0.427). In terms of generalization error, this model has a good fit since the difference is close to 0.

LASSO Regression

An alpha value of 1 was used for the LASSO Regression models. We continue to do the same way with the Ridge part above. Cross-validation of LASSO Regression was performed in figure 5, estimating optimal lambda values. Table 2 provides lambda.1se and lambda.min values, along with their logarithmic equivalents and RMSE values on train and test sets. The lambda values demonstrated a slight difference, indicating potential variations in the level of regularization. The CV plot exhibits a half-U-shaped curve, with the lowest point indicating the log of lambda.min value that strikes the right balance between bias and variance.

Figure 5: CV plot of LASSO regression

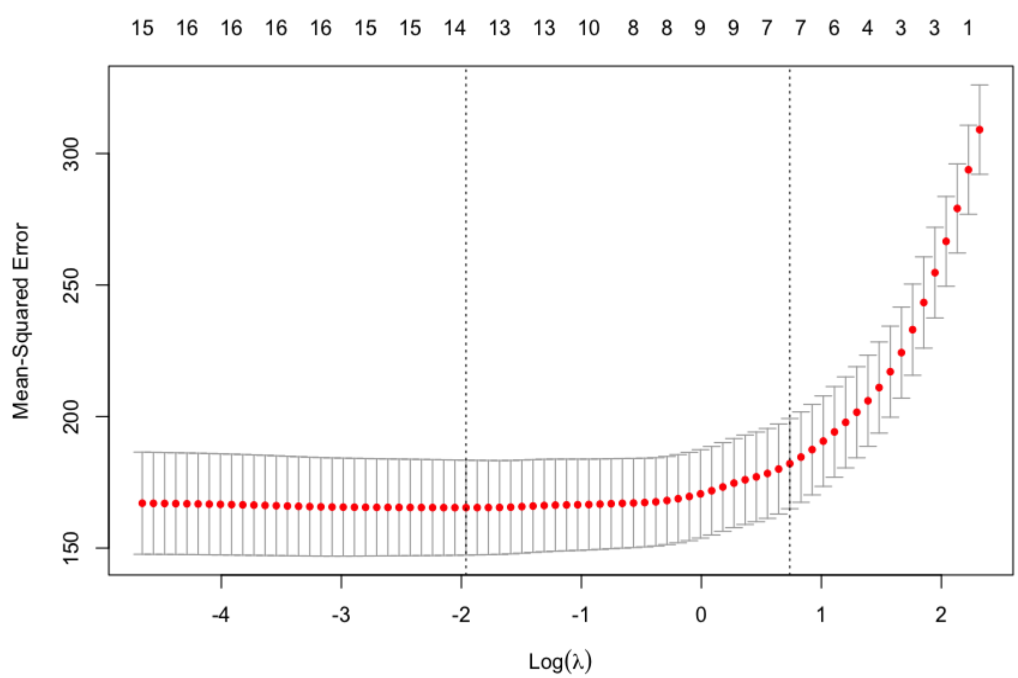


Table 2: LASSO test results

LASSO test	λ	Log	RMSE_train	RMSE_test	RMSE difference
lambda.min	0.141	-1.961	12.518	13.079	0.561
lambda.1se	2.089	0.737	13.334	13.029	-0.305

We also run the two different LASSO models on lambda.min and lambda.1se values, which are shown in Figures 6 and 7. With the lambda.1se penalty term, there are 8 non-zero coefficients, while the number is 15 with the lambda.min value. Table 3 is also created to show the comparison between all LASSO and Ridge models altogether.

Model 3: lambda.1se and alpha = 1

```
18 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept) 39.7604482310
PrivateYes   .
Apps         .
Accept       .
Enroll       .
Top10perc    0.0228706102
Top25perc    0.1312181692
F.Undergrad  .
P.Undergrad -0.0002378434
Outstate     0.0009294806
Room.Board   0.0010576317
Books        .
Personal    -0.0010718785
PhD          .
Terminal     .
S.F.Ratio    .
perc.alumni  0.2325486726
Expend       .
```

Figure 7: Coefficients matrix of model 3

Model 4: lambda.min and alpha = 1

```
18 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept) 37.6731701137
PrivateYes   4.6074968920
Apps         0.0009748666
Accept       .
Enroll      -0.0002358432
Top10perc    0.0153962723
Top25perc    0.1748259948
F.Undergrad  .
P.Undergrad -0.0012511084
Outstate     0.0006788258
Room.Board   0.0020511673
Books        .
Personal    -0.0029312387
PhD          0.0955276981
Terminal    -0.0980886508
S.F.Ratio   -0.1422377776
perc.alumni  0.3164431834
Expend      -0.0003550326
```

Figure 6: Coefficients matrix of model 4

Both LASSO models performing variable selection by eliminating irrelevant features also present quite good performance on prediction when compared to the Ridge models. The LASSO lambda.1se model has the best performance among the four models since the RMSE difference is close to 0 and its nonzero coefficient count is only 8 - the smallest.

Table 3: Ridge and LASSO comparison

Model	LASSO (L1) alpha = 1				Ridge (L2) alpha = 0			
	nonzero coefficient count	train rmse	test rmse	rmse difference	nonzero coefficient count	train rmse	test rmse	rmse difference
lambda.min	15	12.518	13.079	0.561	18	12.567	12.994	0.427
lambda.1se	8	13.334	13.029	-0.305	18	13.314	13.066	-0.248

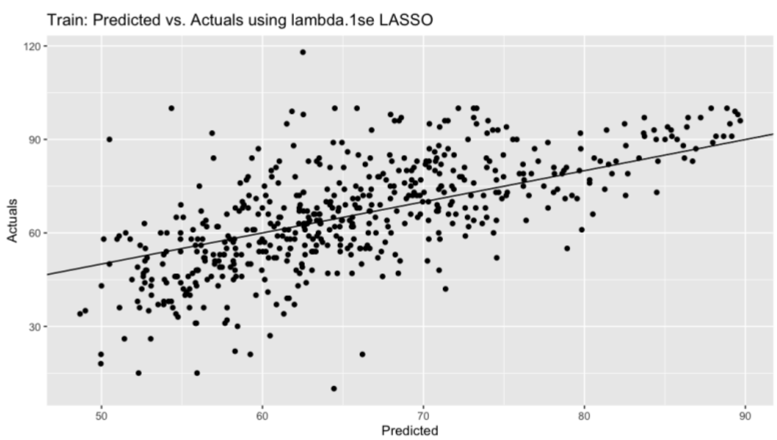


Figure 8: LASSO lambda.1se model on train set

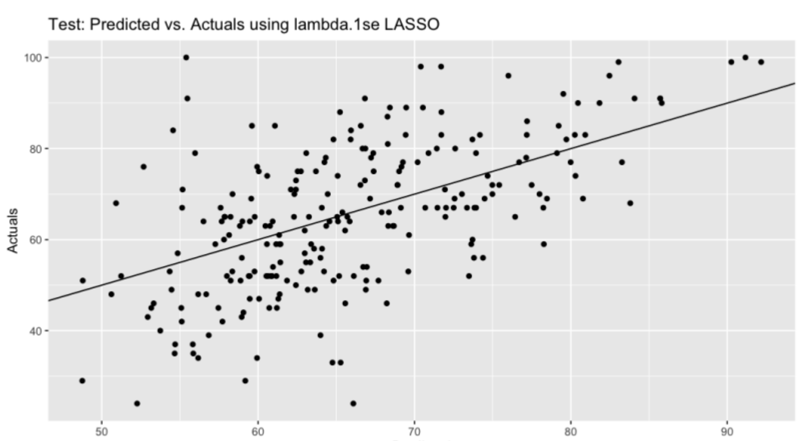


Figure 9: LASSO lambda.1se model on test set

Stepwise selection

In this part, we fit another method: the both-sided stepwise selection model. The result is captured in Figure 10. There are a total of 13 nonzero coefficients, in which only the “Accept” variable is not statistically significant (p-value > 0.05).

Figure 10: Fit Stepwise model

```
Call:
lm(formula = Grad.Rate ~ Private + Apps + Accept + Top25perc +
    P.Undergrad + Outstate + Room.Board + Personal + PhD + Terminal +
    perc.alumni + Expend, data = train_df)

Residuals:
    Min      1Q  Median      3Q     Max
-51.842  -7.123  -0.586   6.963  53.423

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept) 35.0093351  3.9355603   8.896 < 0.0000000000000002 ***
PrivateYes   5.4253539  2.0151104   2.692             0.007320 **
Apps         0.0017183  0.0005165   3.327             0.000939 ***
Accept      -0.0011039  0.0007517  -1.469             0.142508
Top25perc    0.1787948  0.0382949   4.669         0.0000038419 ***
P.Undergrad -0.0013048  0.0004127  -3.162             0.001659 **
Outstate     0.0007828  0.0002661   2.942             0.003403 **
Room.Board   0.0020948  0.0006616   3.166             0.001634 **
Personal    -0.0029227  0.0008886  -3.289             0.001072 **
PhD          0.1517769  0.0669276   2.268             0.023745 *
Terminal    -0.1535561  0.0722936  -2.124             0.034127 *
perc.alumni  0.3310725  0.0567440   5.834         0.0000000094 ***
Expend      -0.0004790  0.0001559  -3.073             0.002227 **
```

After that, we make predictions with the model on the test set and start to compare the best model from glmnet in parts 1 and 2 (the LASSO labda.1se model) with the model from the stepwise feature selection. By getting absolute error values on both models, we want to test the differences between two means of two sets.

H0: means are equal

H1: means are not equal (claim)

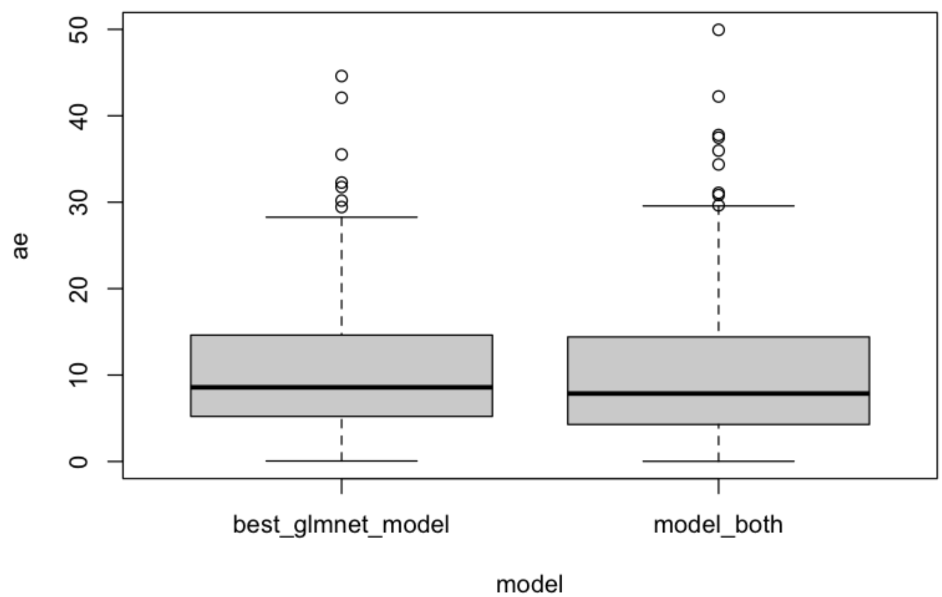
Figure 11: Wilcoxon test

```
	Wilcoxon rank sum test with continuity correction

data:  x and y
W = 28813, p-value = 0.3268
alternative hypothesis: true location shift is not equal to 0
```

Figure 11 shows the result of a nonparametric test called Wilcoxon. The P-value is 0.3268, higher than the significant level of 0.05. So, we fail to reject the null hypothesis and conclude that there is not enough evidence to support the claim. Now let’s check the boxplot of absolute error means of two models in Figure 12 as follows:

Figure 12: Boxplot of absolute error means



We would choose the model with the lower absolute error mean on this plot, which is the model of stepwise selection. This model performs slightly better than the LASSO lambda.1se model in terms of prediction error but does not have a statistically significant difference.

Conclusion

In this analysis., we evaluated and compared several models, including Ridge regression and LASSO models with lambda.min and lambda.1se values, and a stepwise selection model. Both LASSO models showcased good prediction performance by effectively selecting relevant features and outperformed the Ridge models. Specifically, the LASSO lambda.1se model exhibited superior performance with a minimal difference in RMSE and the lowest count of non-zero coefficients (8).

Furthermore, we conducted a comparison between the best model from glmnet (LASSO lambda.1se model) and the stepwise selection model in terms of prediction performance. The statistical test revealed insufficient evidence to support the claim that there is a significant difference in the means of the two sets of absolute errors. Upon examining the boxplot, we observed that the stepwise selection model demonstrated slightly better prediction error performance, although the difference was not statistically significant.

References

1. ISLR. Dataset. Retrieved May 03, 2023. https://rdrr.io/cran/ISLR/man/College.html#heading-0
2. Bluman, A. G. (2018). Chapters 9 & 13. In *Elementary statistics Book*. McGraw-Hill.