To print higher-resolution math symbols, click the
Hi-Res Fonts for Printing button on the jsMath control panel.

CMSC 478 — Fall 2018 — C. S. Marron
Lab 7: Regularization and Dimension Reduction

Data Description

In this lab, you will work with the `College` dataset of college admissions data. The dataset is part of the `ISLR` package; if you have not already installed the package, you may download the CSV file of the `College` dataset. Use `?ISLR::College` in R to see a description of the dataset.

Exercises

> In all of the exercises, you will be trying to predict the number of applications received, `Apps`, using all of the other variables. The training and test sets created in Exercise 1 should be used for all of the exercises.

Exercise 1: Compare Linear Regression to the regularization methods, Ridge Regression and The Lasso.

1. Split the data into a training set and a test set.
2. Fit a linear model on the training set using least squares and report the test error obtained.
3. Fit a ridge regression model on the training set, with λ chosen using cross-validation. Report the test error obtained.
4. Fit a lasso model on the training set, with λ chosen using cross-validation. Report the test error obtained along with the number of non-zero coefficient estimates.
5. Get the coefficients for the lasso model with the best λ values. What do you notice?
6. Compare your results. Of the three, which model is best?
7. Plot the test error with respect to λ for both ridge regression and the lasso. Do you notice any relationship?

Exercise 2: Compare the two rank reduction methods, PCR and PLS.

1. Fit a PCR model on the training set with M chosen by cross-validation. Report the test error obtained along with the value of M selected.

2. Fit a PLS model on the training set with M chosen by cross-validation. Report the test error obtained along with the value of M selected.

3. Is one of the methods preferable? Discuss how the results for PCR and PLS compare.

Exercise 3: Comment on the results obtained for *all* of the models. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?