CMSC 478 — Fall 2018 — C. S. Marron
Lab 8: Non-Linear Methods

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 the following exercises, you will be trying to predict out-of-state tuition costs, `Outstate`, using a subset of the other variables. The training and test sets created in Exercise 1 should be used for all of the exercises.

Exercise 1: Split the data into training and test sets and use forward stepwise selection to identify a model that uses a subset of the predictors.

1. Split the data into a training set of size 500 and a test set of size 277.

2. Compute the forward stepwise selection fit on the training data.

3. Use an appropriate statistic (AIC, BIC, etc.) to select a satisfactory subset of the predictors. Plot the statistic and indicate the optimal number of predictors on the plot. You will use only these predictors in the subsequent exercises.

Exercise 2: Fit a GAM on the training data.

1. Fit a GAM on the training data using smoothing splines and the predictors selected in Exercise 1.

2. Plot the relationship between each predictor and the response, including the standard error. For which variables, if any, is there evidence of a non-linear relationship with the response?

3. The spline for `perc.alumni` is nearly linear. Use ANOVA to determine which of the following three models is most appropriate: a GAM that *excludes* `perc.alumni`, a GAM that uses a linear function for `perc.alumni`, and a GAM that uses a smoothing spline for `perc.alumni`. For each model, use smoothing splines for the predictors other than `perc.alumni`.

4. Evaluate the best GAM model on the test data and explain the results obtained.

5. Can you improve the model by increasing or decreasing the degrees of freedom in the spline?

Construct at least two other GAM models and determine which gives the smallest test error.

Exercise 3: Fit a linear model and compare prediction results.

1. Fit a linear model with `Outstate` as the response and the predictors selected from Exercise 1.
2. Evaluate the linear model on the test data; explain the results and compare to the results for the GAM model.