Homework 2

Due on 03/08/2023

In this exercise, we build nonlinear models using the "College" data. The dataset contains statistics for 565 US Colleges from a previous issue of US News and World Report. The response variable is the out-of-state tuition (`Outstate`). The predictors are

- `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
- `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

Partition the dataset into two parts: training data (80%) and test data (20%).

(a) Fit smoothing spline models using `perc.alumni` as the only predictor of `Outstate` for a range of degrees of freedom, as well as the degree of freedom obtained by generalized cross-validation, and plot the resulting fits. Describe the results obtained.

(b) Fit a generalized additive model (GAM) using all the predictors. Does your GAM model include all the predictors? Plot the results and explain your findings. Report the test error.

(c) Train a multivariate adaptive regression spline (MARS) model using all the predictors. Report the final model. Present the partial dependence plot of an arbitrary predictor in your final model. Report the test error.

(d) In this data example, do you prefer the use of MARS model over a linear model when predicting the out-of-state tuition? Why? For general applications, do you think MARS is a better approach compared to a linear model?