Homework 2

Due on 03/12/2024

In this exercise, we explore the application of nonlinear models to analyze the "College" dataset, comprising statistics from 565 US colleges as reported in a past 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 to predict out-of-state tuition (`Outstate`) using the percentage of alumni who donate (`perc.alumni`) as the only predictor, across a range of degrees of freedom. Plot the model fits for each degree of freedom. Describe the observed patterns that emerge with varying degrees of freedom. Select an appropriate degree of freedom for the model and plot this optimal fit. Explain the criteria you used to determine the best choice of degree of freedom.

(b) Train a multivariate adaptive regression spline (MARS) model to predict the response variable. Report the regression function. Present the partial dependence plot of an arbitrary predictor in your model. Report the test error.

(c) Construct a generalized additive model (GAM) to predict the response variable. Does your GAM model include all the predictors? For the nonlinear terms included in your model, generate plots to visualize these relationships and discuss your observations. Report the test error.

(d) In this dataset, would you favor a MARS model over a linear model for predicting out-of-state tuition? If so, why? More broadly, in general applications, do you consider a MARS model to be superior to a linear model? Please share your reasoning.
   CV needed