Homework 4

#for nicer table output  
library(knitr)  
library(tidyverse)

# Conceptual Problems

Section 7.9

* Book Problem 3 (you can sketch these by hand and take a picture or use R/python/etc. to do so). The problem is reproduced below.

Suppose we fit a curve with basis functions . (Note that equals 1 for and 0 otherwise.) We fit the linear regression model

and obtain estimates . Sketch the estimated curve between and . Note the intercepts, slopes, and other relevant information.

* Book Problem 4 (you can sketch these by hand and take a picture or use R/python/etc. to do so). The problem is reproduced below.

Suppose we fit a curve with basis functions . We fit the linear regression model

and obtain estimates . Sketch the estimated curve between and . Note the intercepts, slopes, and other relevant information.

* Book Problem 5. The problem is reproduced below.

Consdier two curves, and , defined by

where represents the th derivative of .

1. As , will or have the smaller training RSS?
2. As , will or have the smaller test RSS?
3. For , will or have the smaller training and test RSS?

# Implementation Problems

1. This question uses the variables dis (the weighted mean of distances to five Boston employment centers) and nox (nitrogen oxides concentration in parts per 10 million) from the Boston data. We will treat dis as the predictor and nox as the response.
   1. Use the poly() function to fit a cubic polynomial regression to predict nox using dis. Report the regression output, and plot the resulting data and polynomial fits.
   2. Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.
   3. Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.
   4. Fit a smoothing spline model to predict nox using dis using four degrees of freedom. How did you choose the knots? Plot the resulting fit.
   5. Now fit a smoothing spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.
   6. Perform cross-validation or another approach in order to select the best degrees of freedom for a smoothing spline on this data. Describe your results.
2. This question relates to the College data set in the ISLR2 library.
   1. Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.
   2. Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.
   3. Evaluate the model obtained on the test set, and explain the results obtained.
   4. For which variables, if any, is there evidence of a non-linear relationship with the response?