STAT 8330: **Homework 4**

Due on October 8, 2020: 5pm

1. JWHT (2013), Chap 7, Prob 1

2. **Boston housing analysis revisited**: This question uses the `Boston` (housing) dataset from the `MASS` library as we saw in the linear regression lab in Section 3.6.2 in JWHT (2013). We are interested in predicting `nox` (nitrogen oxides concentration in parts per million) from `dis` (the weighted mean of distances to five Boston employment centers). Note, some parts of this question are similar to question 9 in Chapter 7, but some parts are a bit different.

   (a) 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.

   (b) Plot the polynomial fits for a range of different polynomial degrees from 1 to 10, and report the associated residual sum of squares.

   (c) Perform cross-validation to select the optimal degree for the polynomial, and explain your results. Note, show the CV results for each degree (1 - 10).

   (d) Use the `bs()` function to fit a regression spline to these data using 3-6 degrees of freedom using knots at uniform quantiles. Use cross-validation to select the best fit in terms of RSS. Report the chosen degree of freedom, associated RSS, and plot the fit for the best model.

   (e) Repeat (4) but use the natural splines.

   (f) Now fit the data using smoothing splines with cross-validation to select the smoothing level. Report your chosen value for λ and the resulting RSS.

   (g) Use the `loess()` function to fit these data. Describe how you choose the "span" for this fit. Report the RSS for the model. [Note: If you want to extrapoloate, you must include `control=loess.control(surface="direct")`]

3. **Lakes analysis revisited:** Using the `lakes_DA3.Rdata` data from Problem 3 of Homework 3, find the best **two-variable GAM** you can. There really is no one right answer (although try to beat your Homework 3 results!) You can use the data in any way you want to predict `log(secchi)`. For your answer, indicate (a) how you chose your two predictor variables and (b) how you chose your best model (give details about choice for smoothing parameters). Report the 5-fold cross-validation based MSE performance, and compare to that obtained in Homework 3.

4. **Levee analysis:** Consider the data set in `mmr_levee.txt` related to levee failures on the lower Mississippi river from Flor et al. (2010; Engineering Geology). The data set contains the following columns of data:

Column	Description
1	Failure (1=Yes, 0 = No) [RESPONSE]
2	Year
3	River Mile
4	Site underlain by coarse-grain channel fill (sediment)
5	Borrow pit indicator
6	Meander location (1=Inside bend, 2=outside bend, 3=chute, 4=straight)
7	Channel width
8	Floodway width
9	Constriction factor
10	Land cover type (1=open water, 2=grassy, 3=agricultural, 4=forest)
11	Vegetative buffer width
12	Channel sinuosity
13	Dredging intensity
14	Bank revetement

Use GAM to develop the best model you can to classify levee failure. Clearly describe your approach and your results. You may use any information to make your case that your model is reasonable.