Homework 4 Statistics 8330 (DAIII; Wikle) September 27, 2016

DUE: October 4, 2016

Instructions: Write your answers clearly on a separate answer sheet. To minimize the grading effort, ONLY report the results that are asked for or any information you believe needs to clarify your answers. Please type your answers using a word processor.

1. Problem 1, Chapter 7 in the JWHT (2013) book.

2. 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). For problems 1-7, 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.

   (g) Use the `loess()` function to fit these data. Describe how you choose the "span" for this fit.

3. Consider the newspaper data in problem 3 of Homework 2. Fit this model with a GAM using 5-fold cross-validation (don't use interactions in the model). Plot the results. For which variables, if any, is there evidence of a nonlinear relationship with the response? Does this model fit better than the linear regression model in Homework 2?

4. 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.

5. Use a GAM to fit the student test score data from HW 3 (prob 3). You can use the data in any way you want to predict the final grade (G3), but do not use (G1) and (G2) as predictors for this problem. Compare to your best model from HW 3 (again, not using G1 and G2). For your answer, just indicate your best model (give details about model components and choice for smoothing parameters), its CV-based MSE performance, and the associated CV-based MSE for the best model from HW 3.