Homework 3 Statistics 8330 (DAIII; Wikle) September 13, 2016

DUE: September 22, 2016

Instructions: Write your answers clearly on a separate answer sheet. To minimize the grading effort, ONLY report the results that are asked for. Please type your answers using a word processor.

1. Problem 9 in Chapter 6.

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). We are interested in predicting the per capita crime rate in this data set.

   (a) Begin by making a training set and validation set as done on page 248 of Lab 5. Make sure you use the `set.seed(1)` command before you create the training set so that your answer is comparable to mine.

   (b) Fit a linear model using least squares on the training set, and report the test error (test MSE) obtained.

   (c) Use the procedure in Section 6.5.3 of Lab 5 to obtain the number of variables that should be in the model according to cross-validation. Given this number of variables, get the best subset of variables using the training data set. Given these variables, report the test MSE.

   (d) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report λ and the test error obtained.

   (e) Fit a lasso model on the training set, with λ chosen by cross-validation. Report λ and the test error obtained, along with the number on non-zero coefficient estimates. Which variables have parameters estimated to be zero, if any?

   (f) Fit a PCR model on the training set, with M chosen by cross-validation. Report M and the test error obtained.

   (g) Fit a PLS model on the training set, with M chosen by cross-validation. Report M and the value of the test error obtained.

   (h) Comment on the results from the steps above. How accurately can we predict per capita crime rate? Is there much difference among the test errors resulting from these approaches? Which variables seem to be the most important as predictors? Is there anything you would do differently if you were analyzing these data again?

3. Consider the data in the file `student-mat.csv` on the Blackboard website. These data attributes are described in the file `student.txt`. These data correspond to math scores of students in Portugal and associated attributes – for more information, see

   `http://archive.ics.uci.edu/ml/datasets/Student+Performance`

   Your task is to build the best predictive model you can for the scores based on the attributes with the methods given below. Specifically,

(a) Build the best model you can to predict the final grade (G3) but DO NOT use G1 and G2 as a predictor. Report the best model using (1) ridge regression, (2) lasso regression, (3) PC regression, and (4) PLS regression. Your write-up should include a brief description of how you selected the variables in the model and how/when you used cross-validation. Your final results should be presented for each model, each using a 5-fold cross-validation with a a random seed of (1.0); you should also present only the *best* model for each methodology (and, summarize which is the best one overall). [Please indicate which variables were selected for your best models.]

(b) Repeat the model building process in part (a) but this time you can also use G1 and G2 as predictors. (Again, you can *only* use the 4 methods described here.

(c) Describe any particular difficulties you had with this modeling.

(d) Include the R code in an Appendix that can duplicate your best model results for part (a).

Note, you can read the data with the following R command:

```
data1=read.table("student-mat.csv",sep=";",header=TRUE)
```