STAT 8330: **Homework 3**

Due on September 28, 2020 (10:00pm)

1. JWHT (2013), Chap 6, Prob 9

2. **Boston housing analysis**: 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 in Section 6.5.3 of ISLR. 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 ISLR 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. **Lake analysis**: Consider the data in the file `lakes_DA3.Rdata` on the class website. These data are lake attributes, and the variables contain in the dataset are described in more detail below. In general, they contain measurements of lake quality metrics for 1188 lakes, as well as land use and climate variables pertaining to the location surrounding the lake.

   - *tp*: Total phosphorous in a sample of lake water.
   - *tn*: Total nitrogen in a sample of lake water.
   - *chla*: Chlorophyl in a sample of lake water.
   - *secchi*: A measure of lake clarrity; depth (measured in meters) at which a secchi disc is no longer visible. The clearer the lake, the deeper the disc will be visible.
   - *lat*: Latitude
   - *long*: Longitude
   - *lake_area*: Area of lake (measured in hectares)

- *mean_depth*: Average lake depth
- *max_depth*: Maximum depth within lake
- *iws_urban*: Percent of urban land in surrounding catchment
- *iws_ag*: Percent of agriculural land in surrounding catchment
- *iws_pasture*: Percent of pasture in surrounding catchment
- *iws_forest*: Percent of forest in surrounding catchment
- *iws_wetland*: Percent of wetland in surround catchment
- *mean_annual_temp*: Average annual temperature
- *mean_winter_temp*: Average winter temperature
- *mean_spring_temp*: Average spring temperature
- *mean_summer_temp*: Average summer temperature
- *mean_fall_temp*: Average fall temperature
- *mean_annual_precip*: Average annual total precipitation
- *mean_winter_precip*: Average winter total precipitation
- *mean_spring_precip*: Average spring total precipitation
- *mean_summer_precip*: Average summer total precipitation
- *mean_fall_precip*: Average fall total precipitation

Your task is to build the best predictive model you can for secchi based on the attributes with the methods given below. Specifically, build the best model you can to predict `log(secchi)`. Report the best model among the choices (1) ridge regression, (2) lasso regression, (3) PC regression, and (4) PLS regression. Your write-up should include a brief description of how you chose between the four methods, how you selected the variables in the model, and how/when you used cross-validation. Your final results should be presented for your best overall model of choice, using a 5-fold cross-validation with a random seed of (1). Make sure to clearly indicate which variables were selected for your best model.

4. Recall from the lecture notes that one could write a general regularized regression criteria as:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|^q,$$

for $q \geq 0$, where $q = 1$, $q = 2$ correspond to the lasso, and ridge regression, respectively. We typically do not "estimate" or tune q. Here, we investigate this parameter via a simulation example. Simulate data as follows (note, we are not including an intercept here):

```
> set.seed(1)
> x1 = rnorm(1000)
> x2 = rnorm(1000)
> x3 = rnorm(1000)
> x4 = rnorm(1000)
> x5 = rnorm(1000)
> e = .1*rnorm(1000)
> beta.truth = c(1.3,.01,-1.2,-.02,.6)
> x = cbind(x1,x2,x3,x4,x5)
> y = x%*%beta.truth + e
```

Consider the objective function as given above:

```
> myRSSgen <- function(beta,x,y,lam,q){
+     sum((y-x%*% beta)^2) + lam*sum((abs(beta))^q)
+ }
```

Now, use the general "optim" function in R to find the βs that minimize this for a given λ and q. E.g., for $\lambda = 10$ and $q = 2$,

```
> optim(rep(0,ncol(x)),myRSSgen,method='CG',x=x,y=y,lam=10,q=2)
```

(a) Use the "optim" function and your objective function to get the OLS estimates for the betas. What value of λ and q did you use?

(b) For $\lambda = 10$, investigate what happens to the parameter estimates as you vary q from 0 to 4. Discuss your findings.

(c) Use 5-fold cross validation to select λ for the ridge-regression case. Plot the CV error as a function of λ and report the parameter estimates for your selected λ on the full training data.

(d) Repeat part (c) but for the Lasso case.

(e) Do a grid search on λ and q to try to find the values that minimize the CV error. Plot the CV error as a function of these parameters via a contour plot. What is your optimal choice of these parameters? How do the β estimates compare to the ridge and lasso case from above?

(f) Overall, does it seem to make a difference to consider different values of q in this case? Discuss.