STAT 425 — Fall 2018

Homework 6

(Posted on Fri Nov. 16; Due Thurs Nov. 29 at 9:30 AM)

Please submit your assignment *on paper*, following the Guidelines for Homework posted at the course website. (Even if correct, answers might not receive credit if they are too difficult to read.) Remember to include relevant computer output.

1. Consider the Misner rainfall data discussed in class (misner.dat is posted in the Homework folder on Compass). Consider a quadratic regression model using time, time^2, rain, rain^2, and the interaction of time and rain. Perform variable selection using some of the methods discussed in class to find a good model that explains yield.

2. Use the `seatpos` dataset with `hipcenter` as the response and all other variables as possible predictors. Fit a linear model and investigate the diagnostics. Find appropriate transformations for both the response and the covariates to come up with a good linear model. Use any variable selection technique to find a sparse (small) model based on the transformed model.

3. Describe the ridge regression estimator for a general design matrix X (not necessarily orthogonal design) and derive its bias vector and variance matrix. Show that the variance matrix of the ridge estimator is "smaller" than that of the least squares estimator.

4. We would like to compare Least squares, Ridge and Lasso estimators using synthetic data. Follow the below steps to generate a synthetic dataset with $n = 100; p = 20$ and

$$\beta_{20\times 1} = (1, -1, 1, 0, \cdots, 0).$$

   (a) Generate n draws from a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma_{20\times 20} = 0.7I + 0.3J,$$

   where I is the identity matrix and J is the matrix of all 1's. Define $X_{n\times p}$ to be the matrix with these n draws as its rows.

   (b) Generate n random errors $\epsilon_{n\times 1}$ from $N(0, 1)$ and define $Y = X\beta + \epsilon$.

   (c) Now obtain the Least squares, Ridge and Lasso estimates using the data $\{Y, X\}$ (by using λ based on cross validation). Compute the Mean Squared Errors (MSE) for each estimator as

$$MSE(\hat{\beta}) = \sum_{i=1}^{p} (\hat{\beta}_i - \beta_i)^2.$$

   Comment on the performance of the different methods.

   Repeat steps (a) - (c) to obtain 100 synthetic datasets and average the values of the MSEs.