Statistics 745

Group Assignment 4

1. **Group Assignment:** Let X be orthogonal (i.e. $X'X = I$). Denote $\hat{\beta}^{(ls)}$ as the least squares estimator. In this case $\hat{\beta}^{(ls)} = X'y$. Also, the ridge regression estimator is $\hat{\beta}^{(ridge)} = \frac{X'y}{1+\lambda}$. The lasso also has a closed form in this case. Show that the solution to the lasso is given as:

$$\hat{\beta}^{(lasso)} = \text{sign}(\hat{\beta}_j^{(ls)})(\mid \hat{\beta}_j^{(ls)} \mid -\gamma)^+.$$

   where sign returns the sign of its input, $(t)^+ = \max(0, t)$ and γ is determined so that the condition of $\sum \mid \hat{\beta}_j \mid = t$.

2. **Individual Assignment:** Let $X_1, \cdots, X_3$ be Standard Normal variables with $n = 100$. Let $\beta = (2, 1, -1, -2)$ and generate $\epsilon \sim N(0, 0.9I)$. Now generate $y = X\beta + \epsilon$.

   Define a matrix $Z_q = [Z_1, \cdots, Z_q]$ where $Z_i \sim N(0, 1)$. Our data matrix $X = [X_1, X_2, X_3, Z_1, \cdots, Z_q]$. We consider the LASSO as q increases using the full X data to model y. Use the *lars* package in **R** to fit the lasso. Plot the fit, to get the path plot. You will need to figure out how to interpret this plot.

   Discuss where variables X_1 to X_3 come in. Also, discuss the paths. Use 10-fold CV to estimate the number of variables and comment on the results. Simulate the following situations:

   (a) Run $q = 0$ first compare variable order. Also does the LASSO take any out of the active set.
   (b) Run $q = 10$.
   (c) Run $q = 50$
   (d) Run $q = 75$.
   (e) Run $q = 90$.
   (f) Run $q = 99$.