Stat 462/862 Assignment 1

(Due on Oct 6, 2015, hard copy, in the class)

1. **Grocery retailer.** Consider the data file Grocery.txt. A large, national grocery retailer tracks productivity and costs of its facilities closely. Data were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1), the indirect costs of the total labor hours as a percentage (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and the total labor hours (Y). (In Grocery.txt, the columns from left to right are Y, X_1, X_2, X_3.)

   (a) Create a pairwise scatter plots for dependent and independent variables. Show the plot and make comment on the plot.

   (b) Fit the linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$. Show the table of the fitted model: coefficients estimation, their standard deviation, z-score, and p-values. Show R^2 and the estimation $\hat{\sigma}^2$.

   (c) Use best subset (C_p), forward variable selection, and backward variable selection to find a smaller model with the best fits. Show the results of the fits as in (b).

   (d) Use F-test to check whether the model returned from (c) significantly different ($\alpha = 0.05$) from the complete model in (b).

   (e) Obtain the prediction of mean response, its associated prediction error and $100(1-\alpha)\%$ confidence interval based on the model your proposed (part (c)) for the new input $X_1 = 32000, X_2 = 7.5, X_3 = 1$.

2. **Lasso.** In Problem 1, add five more predictor variables $Z_1 = X_1X_2, Z_2 = X_1X_3, Z_3 = X_2X_3, Z_4 \sim N(30, 30), Z_5 \sim N(7, 1)$. Use the Lasso to analyze the data, and interpret your result.

3. **Housing.** For the data file housing.txt, let 'MEDV' be dependent variables and others be independent variables. Apply the ridge regression to the dataset.

   (a) What is the estimated coefficients for the fitted model with the parameter $\lambda = 2$.

(b) Fit a ridge regression with $\lambda = 1, 3, 5, 7, 9, 11$. For each value of λ, obtain the corresponding estimated coefficient vector. Comment on the behavior of each of the estimated coefficients as λ increases.

(c) Choose a value of λ based on generalized cross-validation.

4. (only for graduate students) For the ridge regression, show that

$$\hat{\beta}^{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}' + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}.$$