STAT 2131:
Applied Statistical Methods I
HW #5
Due Thursday, November 11th

1. Refer to the data set CH09PR10.txt (from the KNNL book). A personnel officer in a governmental agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical positions in the agency. For purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. The scores on the four tests (X_1, X_2, x_3, X_4) and the job proficiency score (Y) for the 25 employees were recorded.

   (a) Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each of the predictor variables? Are any serious multicollinearity problems evident? Explain.

   (b) Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

   (c) Consider only the four first order terms of X_1, X_2, X_3, X_4, find the best subset regression models according to the adjusted R^2 criterion and the AIC criterion.

   (d) Using forward stepwise regression, find the best suset of predictor variables to predict job proficiency. Use α limits of 0.05 and 0.1 for adding or deleting a variable.

   (e) How does the best subset model in part (d) compare to that in part (c)?

2. We have looked at using ordinary least squares/maximum likelihood estimation for the simple linear regression model. This problem considers an alternative estimation procedure. For simplicity, we will assume that the variance σ^2 is known.

   You observe outcomes y_i from $i = 1, \ldots, n$ subjects and assume that the data follow the linear model

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i$$

   where x_i are known deterministic covariates, β_0 and β_1 are unknown deterministic parameters, and ϵ_i are independent and identically distributed mean zero Gaussian random variables with *known* variance 1. Given some $\lambda > 0$, you decide to estimate β_1 with $\tilde{\beta}_1$ that minimizes the penalized sum-of-squares

$$\text{PSS}_\lambda(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda\beta_1^2$$

   so that

$$\left(\tilde{\beta}_0, \tilde{\beta}_1\right) = argmin_{\beta_0,\beta_1\in\mathbb{R}}\text{PSS}_\lambda(\beta_0, \beta_1).$$

(a) What is $\tilde{\beta}_1$ for a given $\lambda > 0$?

(b) Show that $\tilde{\beta}_1$ is biased for a given $\lambda > 0$.

(c) Compare the variance of $\tilde{\beta}_1$ with the variance of the OLS $\hat{\beta}_1$. Is one variance always smaller than the other? If so, prove it. If not, under what conditions is $var(\tilde{\beta}_1) < var(\hat{\beta}_1)$?