Topics covered this week

Statistics 207
Winter Quarter, 2015

Tuesday, Feb 17	Partial least squares.
Thursday, Feb 19	Lasso, Logistic regression (chaps 14.1-14.3, Appl. Lin. Stat. Models).

Homework 6 (due on Thursday, Feb 26)
Problems. [From App. Lin. Stat. Mod.] 14.9, 14.11, 14.14, 14.19
and the problems given below.

[You may form a group of 2 students (including yourself) registered in this course. Only one work per group needs to be submitted. Please write down the names of the group-members on the first page. The first page of submitted homework should contain the names of the students in the group, but all work should start from page 2.]
Note that we will not grade all the problems.

Problem 5. Refer to the Apartment data given in the last homework. First standardize all the variables. Use the R package "pls" to answer the questions given below.
(a) Obtain the cross-validation scores for partial least squares with 0,. . . ,5 components. Write down the loadings of the first three components. Also obtain the R^2 and the adjusted-R^2 values for each of these six models. Plot the scores of the first three components against each other. Comment on your findings.
(b) Use the information in part (a) to carry out sequential F-tests to decide how many components you should keep. If this result not consistent with the one obtained by the CV criterion, make a decision on the number of components to keep.
(c) For the model you have decided on in part (b), obtain the estimated beta parameters, plot the observed against the fitted Y-values, residuals against the fitted values and the histogram of the residuals. Summarize your findings.

Problem 6. This problem also uses the Apartment data. Standardize all the variables. Use the R package "glmnet".
(a) Use the command cv.glmnet to plot the CV criterion against the penalty (or log(penalty)). Also obtain the value of the penalty at which the CV is the smallest.
(b) For the value of the penalty at which the CV is smallest, obtain the estimated beta parameters, plot the observed against the fitted values, residuals against the fitted and the histogram of the residuals. Summarize your findings.

Problem 7. (This problem will not be graded) Ridge regression can be obtained using a Bayesian framework. Suppose we have the model $Y = X\beta + \varepsilon$, where

Y is $n \times 1$, X is $n \times p$, β is $p \times 1$ and ε is $n \times 1$. Assume that ε consists of iid $N(0, \sigma^2)$ variables. Now assume that the components of β are iid $N(0, \tau^2)$.
(a) Obtain the joint distribution of Y and β. Denote this by $L(\beta)$ in the manner of the usual likelihood.
(b) Let $l(\beta) = -2\log(L(\beta))$. Maximizing L with respect to β is equivalent to minimizing l with respect to β. Re-express τ^2 as σ^2/k and then show that the minimum of l is attained at $\hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y$.

Problem 8. This problem has the same set-up as in the last one and assume that all the variables have been standardized. Let $\lambda_1 \geq \lambda_2 \geq \cdots$ be the eigenvalues of $X^T X$ with $e_1, e_2, \ldots$ the corresponding orthonormal eigenvectors. You are given the following information about the model: $n = 25, p = 5, \sigma^2 = 2.5$ and

j	1	2	3	4	5
λ_j	19	3	1	0.7	0.3
$e_j^T \beta$	0.8	0.3	0.2	0.2	0.1

Let $\hat{\beta}(k)$ be the ridge regression with penalty $k > 0$, i.e., $\hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y$.
(a) Use the computer to find the value of k at which $D(k) = E[||\hat{\beta}(k) - \beta||^2]$ is minimized.
(b) Use the computer to find the value of k at which $L(k) = E[||X\hat{\beta}(k) - X\beta||^2]$ is minimized.
(c) Compare the minimum value of $D(k)$ (i.e., $\min_{k>0} D(k)$) to $D(0)$. Recall that $k = 0$ is the least squares case. Do the same for L. Comment on your findings.