**STA 450S/4000S**: Homework #1
Due February 4, 2005

1. *Bayesian inference in the linear model:*
Suppose that the $n \times 1$ vector $Y$ follows a normal distribution with mean $X\beta$ and variance $\sigma^2 I$:

$$Y \sim N(X\beta, \sigma^2 I)$$

i.e. that

$$f(y \mid \beta) = \frac{1}{\sqrt{2\pi}\sigma^n} \exp\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\}.$$

Suppose that we assume a *prior* distribution for $\beta$ that is $N(0, \tau^2 I)$:

$$f(\beta) = \frac{1}{\sqrt{2\pi}\tau^p} \exp(-\frac{1}{2\tau^2}\beta^T\beta).$$

By Bayes theorem the *posterior* distribution of $\beta$, given $y$, is

$$f(\beta \mid y) = f(y \mid \beta)f(\beta) / \int f(y \mid \beta)f(\beta)d\beta.$$

Show that this posterior distribution is also normal, and give expressions for the mean and variance. What is the limit as $\tau^2 \to \infty$?
**4000 only**: Generalize this to unknown $\sigma^2$ and the priors $\beta \mid \sigma^2 \sim N(0, \sigma^2 I)$, $\sigma^2 \sim IG(\nu/2, \nu\tau^2/2)$, where $\nu$ and $\tau$ are assumed known. (A random variable $x$ follows the $IG(a, b)$ distribution if its density is given by $f(x) = \frac{b^a}{\Gamma(a)x^{a+1}} \exp(-\beta/x)$.)

2. *Exercise 3.5 of HTF*:
Consider the ridge regression problem

$$\hat{\beta}^{ridge} = \text{argmin}_\beta \left\{\Sigma_{i=1}^N (y_i - \beta_0 - \Sigma_{j=1}^p x_{ij}\beta_j)^2 + \lambda\Sigma_{j=1}^p \beta_j^2\right\}. \quad (1)$$

Show that this problem is equivalent to the problem

$$\hat{\beta}^c = \text{argmin}_{\beta^c} \left\{\Sigma_{i=1}^N (y_i - \beta_0^c - \Sigma_{j=1}^p (x_{ij} - \bar{x}_j)\beta_j^c)^2 + \lambda\Sigma_{j=1}^p \beta_j^{c2}\right\}.$$

Give the correspondence between $\beta^c$ and the original $\beta$ in (1). Characterize the solution to this modified criterion.

3. *Exercise 3.10 of HTF* **4000 only**: Show that the ridge regression estimate can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix $X$ with $p$ additional rows $\sqrt{\lambda}I$, and augment $y$ with $p$ zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients towards zero. This is related to the idea of *hints* due to Abu-Mostafa (1995), where model constraints are implemented by adding artificial data points that satisfy them.

4. *The abalone data*:
One of the regression data sets at the UCI machine learning repository is the *abalone data*. It can be accessed from within R as "/u/reid/abalone.data". There are 4177 cases, and 8 predictor variables (features). The output variable is the number of rings, which is equivalent to the age of the abalone. The goal is to use the features to predict the number of rings.

   (a) Choose 1000 cases at random to be your personal training data set. Choose another 1000 cases to be a validation set. The remaining 2177 cases are the test data set. (This is more elaborate than what was done with the prostate data.)

   (b) Fit a linear regression using all the predictors to the training data set.

   (c) Fit a linear model using ridge regression, for various values of the smoothing parameter.

   (d) The validation data set will be used to choose the best of the models in (a) and (b). For (a), use all possible subsets or stepwise regression to choose a best subsets linear model, as determined by the value of $(1/1000)\Sigma(y_i - \hat{y}_i)^2$ on the validation data. For (b), choose the value of $\lambda$ that gives the smallest value of $(1/1000)\Sigma(y_i - \hat{y}_i)^2$ on the validation data.

   (e) Compare the best fitting linear model to the best fitting ridge regression model on the test data.

   (f) **4000 only**: Include a comparison of the lasso and of principal components regression.

   You should present your conclusions and supporting evidence in a report that does not include computer code. The computer code that you used should be commented and included as an appendix.

Students in STA 450 are welcome to do the 4000 questions as bonus.