STA244 **4/23/2003**

Final Data Analysis Assignment

Due 4/24/2009 by 4pm in 244 box in 211 Old Chem.

You may ask question of the TA or Prof; otherwise all work should be your own.

1. Refer to the data in Christensen Exercise 14.10. Please turn in a typed solution that addresse the questions below. You should have at most 2 pages of text and tables, and at most 4 pages including figures.

   (a) Conduct an exploratory data analysis. What does this suggest about relationships among the variables?

   (b) Fit the full model and carry out a residual analysis. Does this and the EDA above suggest that you should transform any variables? Are there any outliers or influential points? Provide point estimates and 95% confidence intervals for each of the coefficients, with interpretations . Are there any indications of problems due to multicollinearity? (discuss)

   (c) Use a classical model selection method (forward, backward, or stepwise) to see if you could identity a simpler model. (see help(step)) for R implementations. What is the final model? Briefly interpret output.

   (d) Use the BAS package (or the BMA package) to explore the posterior distributions over all models (use the Zellner-Siow prior). What does this suggest about the most important models or most important variables?

   (e) (extra credit) Implement at least one of the shrinkage methods (ridge, lasso, or a fully Bayes approach). Summarize your findings and compare to the frequentist model selection and Bayesian Model averaging.

   (f) For each of the approaches above (full model, stepwise model selection, BMA, and optionally one of the shrinkage methods if you implemented that) carry out a cross-validation exercise, where you fit using all data but the *ith* case to obtain an out-of-sample prediction of the *ith* case, $\hat{Y}_i$. Do this for each of case, $i = 1, \dots, n$ and compute CV residual $Y_i - \hat{Y}_i$. Compute the cross-validation MSE[1]

   $$\text{CV-MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

   Also create side-by-side boxplots of the CV residuals for each method, and a scatterplot matrix of residuals under each method. The predict.bma function in

[1] This is leave-one-out cross validation, which often selects more complex models, a better approach is to do leave k-out CV.

BAS is similar to the predict function for linear models and will give predictions under BMA.

How do the methods compare?

2. It is well known that Bayesian procedures for model selection are biased. Are the estimators from a stepwise (or backwards/forwards) regression procedure unbiased for estimating the $E[\mathbf{Y}] = \boldsymbol{\mu}$? (You may assume that $\boldsymbol{\mu}$ is in a subspace that is contained in the column space of the design matrix of the full model, i.e. we have not omitted any predictor variables.