STA 450S/4000S: Homework #2
Due March 2, 2004

1. *Project*: Your project involves finding a data set of interest, and analysing it. This means deciding which questions might be of interest for this data, and using one or more of the techniques from this course to answer the questions. Often the data sets themselves have questions provided at the same site. With this homework, please submit a brief (one paragraph or less) description of the data set you plan to analyse. Include the web site address where the data may be obtained, or another complete reference if the data is not on the web.

   Your data set should have a large-ish number of cases (at least 100, probably not more than 5000), a single response (output) which may be continuous or categorical, and several features (inputs). Preferably with no missing values on any case.

   While not required for this homework, if you have preliminary thoughts on how you might analyse the data feel free to submit this now for feedback.

2. *Ex. 4.5 from HTF* Consider a logistic regression problem based on data (y_i, x_i), where y_i is 0 or 1 and $x_i \in R$, i.e. there is just one input. Write the likelihood function for the Bernoulli model with $\log\{p_i/(1-p_i)\} = \beta_0 + \beta_1 x_i$, and give the equations for the maximum likelihood estimates of β_0 and β_1. What will the solution to these equations be if all the x_i's for which $y_i = 0$ are less than some fixed value x_0 and all the x_i's for which $y_i = 1$ are greater than x_0?

3. *The heart data* Carry out a linear discriminant analysis of the heart data, using `chd` as the class variable. Compare the results to that obtained by logistic regression. The data is in `"/u/reid/heart.data"`, and the logistic regression code is in the slides from the Jan 28 lecture.

4. *Optional (bonus) for 450; required for 4000* In Section 5.2.3 of HTF, smoothing is used to reduce the number of inputs. Reproduce the analysis depicted in Figure 5.5 and describe how the red curve in the lower figure helps in understanding the regularity in the data.