IOM 530: Applied Modern Statistical Learning Methods
Assignment 5 (Due 10/10/2008)

Guidelines for assignment submission:

1. Type each question before you answer it, and provide a clear separation between each part.
2. All relevant computer output should be provided unless noted otherwise.
3. Print your homework, and submit it at the beginning of the class. Make sure that it is stapled, and your name is typed on it.
4. Attach your R code as an Appendix. Make sure to provide comments on what your code is doing. Keep it clean and clear!
5. Note the main aim of this homework is to get practice doing logistic regression in R. Hence you shouldn't restrict yourself to only doing specifically what is asked. Anything else you might want to do to build a better logistic regression model would be welcome. The following question could be answered using a data set you have gathered. Some good sources for interesting data sets are
   http://www.econ-datalinks.org/search.html
   http://fisher.osu.edu/cgi-bin/DB_Search/db_search.cgi?setup_file=finance.setup.cgi
   http://fisher.osu.edu/fin/fdf/osudata.htm
   http://www.census.gov/epcd/www/recent.htm
   http://www.bized.ac.uk/dataserv/freedata.htm

   Warning: These websites are fascinating, with thousands of possible data sets to explore. You may find yourself ignoring your loved ones and your other classes just so that you can sneak back and spend more time exploring. This is unfortunately one of the dangers of taking a statistics class and is unavoidable! Seriously, real data sets often have real problems (that have nothing to do with R) associated with them. If you run into problems you may find it easier to work with the auto data.

6. You can refer to Section 5.3.1, 5.3.1, and 5.3.3 (Lab: Resampling Methods) in the course book and the facebook group for this class if you need some help with the R code necessary to finish this assignment.

IOM 530: Applied Modern Statistical Learning Methods
Assignment 5 (Due 10/10/2008)

Refer to the Default data set in the ISLR library. Recall from chapter 4 in the course book, that we used logistic regression to predict the probability of default using income and balance. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis so that you can reproduce the same results every time you run your code. (set.seed(1))

10-fold Cross Validation Approach

1. Using the 10-fold approach:
    a. We would like to fit a logistic regression model that uses income and balance to predict default. In order to estimate the test error of this model, you must perform the following steps:
        i. Fit a multiple logistic regression model using all observations
        ii. Using the cv.glm() function, compute the validation set error.
    b. Now consider a logistic regression model that predicts the probability of default using income, balance, and student. Estimate the validation set error for this model by repeating the same steps in part (a).
    c. Comment on whether or not including the student variable leads to a reduction in the test error rate.