

# **IOM 530: Applied Modern Statistical Learning Methods**

## **Assignment 3 (Due 9/26/2008)**

### **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/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 4.6.1 (Lab: Logistic Regression) 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 3 (Due 9/26/2008)

### 1. Logistic Regression

Refer to the **Weekly** data set, which is part of the **ISLR** package. This data is similar in nature to the **Smarket** data from the lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

- a. Produce some numerical and graphical summaries of the **Weekly** data. Do there appear to be any patterns or correlations?
- b. Use the full data set to perform a logistic regression with **Direction** as the response and the five lag variables plus **Volume** as predictors. Use the `summary` function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones? Interpret the coefficients of the statistically significant predictors.
- c. Compute the confusion matrix and the misclassification error rate. Explain what the confusion matrix is telling you about the types of mistakes made by the logistic regression.
- d. Now fit the logistic regression model using a training data period from 1990 to 2008, with **Lag2** as the only predictor. Compute the confusion matrix and the misclassification error rate for the testing data (i.e., the data from 2009 and 2010).