**DEPARTMENT OF STATISTICS**

**STATS 784** Data Mining

# Laboratory 1 Monday July 31

Lab reports are due Wed Aug 2. Mail to Alan by 5pm.

This lab involves a prediction exercise.

In the R package ISLR, there is a data set **`Smarket`** containing data on 1250 daily percentage returns for the S&P 500 stock index between 2001 and 2005, as well as some other variables. (A percentage return is the percentage difference between today's price to yesterday's price.) The data are

| | |
|---|---|
| **Year** | The year that the observation was recorded |
| **Lag1** | Percentage return for previous day |
| **Lag2** | Percentage return for 2 days previous |
| **Lag3** | Percentage return for 3 days previous |
| **Lag4** | Percentage return for 4 days previous |
| **Lag5** | Percentage return for 5 days previous |
| **Volume** | Volume of shares traded (number of daily shares traded in billions) |
| **Today** | Percentage return for today |
| **Direction** | A factor with levels *Down* and *Up* indicating whether the market had a positive or negative return on a given day |

Is it possible to predict today's percentage return from the other variables? (Excluding Direction, you can ignore this). Conventional wisdom says not – otherwise statisticians would all be rich, which is obviously not the case. What do you think?

Suggestion: construct a linear predictor and estimate the prediction error. Compare the error with random guessing – this corresponds to predicting a return of zero.