The data for this practical come from a study by Dr Peter Lunn (Economic and Social Research Institute, Dublin), undertaken in 2006, and were collected by government officials in the Republic of Ireland as part of an investigation into the nation’s general health. The government’s stated objective was (and is) to introduce measures which promote the taking of regular exercise and therefore it was necessary to assess the factors which affect continued participation in sport. The sample comprised 3081 individuals, selected at random from the population. The variables are given below.

- **Plays**: 1 = currently participates in sport, 0 = does not play sport
- **Gender**: M = male, F = female
- **Age**: Age in years
- **Income**: Weekly income in Euros
- **Education**: A = none, B = primary, C = junior cert, D = leaving cert, E = diploma, F = degree, G = postgrad
- **Parents**: Both = both played, Moth = mother played, Fath = father played, Neith = neither played

The data are in the file `sport.txt`. Note that three of the variables are categorical and two are not, and that there are some missing values. You should note further that the factors for **Education** could, arguably, be ordered; should you want to order them you will need to arrange this by using `ordered()`.

The objective is to produce a written report on the factors which influence continued participation in sport. You should include a quantitative interpretation of your model and show clearly how the odds-ratios are affected by the influential explanatory variables.

Since the response is binary and three of the explanatory variables are categorical, scatterplots will not be productive. In setting about the analysis you might like to try the following steps.

(i) Obtain contingency tables of **Plays** against each of the categorical explanatory variables. In which cases does the response pattern appear to differ, thereby indicating variables which exert important influence?

(ii) Look at boxplots of **Age** and **Income**. Does it look as if any transformations are needed?

(iii) Obtain the correlation matrix for the explanatory variables. What does it tell you?

(iv) Fit, in turn, eight separate logistic regression models, each with just a single explanatory variable. Which variables seem to be the most important individually? Now perform a multiple logistic regression of **Plays** on all explanatory variables: which seem important? Can any categories be pooled? Should there be any interaction terms? What do diagnostic checks tell you?

(v) Fit your final model model. What does it tell you in qualitative terms? Remember that the main purpose in fitting the model was to identify important factors affecting continued participation in sport. Do you think that all the relevant data has been recorded by the researchers?

(vi) Produce a written report.