CHAPTER 6: REGRESSION ANALYSIS

The data set used to illustrate regression analysis in Chapter 6 is given in the file *girls.txt*, the equivalent data set for boys, *boys.txt*, and also the data set on hedonism, *hedonism3.txt*, are included for you to analyse. Each dataset is also given in SPSS format (*.sav).

The SPSS syntax file (*girls.sps*) reproduces some of the regression analyses presented in the chapter. The SPSS output is given in *girls.spo* and *girls.pdf*.

You are encouraged both to repeat the analyses described in Chapter 6 and to try out new regression analyses. Also you could investigate regression diagnostic techniques which are designed to help in assessing whether a particular regression model is appropriate. [These are not covered in the book due to lack of space.]

Girls and boys test scores data

The files *girls.txt* and *boys.txt* (*girls.sav* and *boys.sav*) contain the following variables (in this order):

GCSE — score for GCSE English

WORDS — number of words written in ten minutes. To get the variable SPEED (the number of words per tenth of a minute) divide WORDS by 100. [Note in the book we regress GCSE on SPEED, the number of words per tenth of a minute, in order that the regression coefficient should not be too small.]

JOINS — an indicator variable taking the value 1 if the student has a problem with joined up writing and 0 otherwise.

CATsVS — Cognitive Ability Test score (Verbal Stanine), which is the score on a multiple choice style test taken at about age 12 designed to test verbal ability.

CATsNVS — Cognitive Ability Test score (Non-Verbal Stanine), which is the score on a multiple choice style test taken about age 12 designed to test non-verbal ability.

WARNING: The girls and boys files contain only complete cases; any students for whom any data were missing were omitted from the data set. This might give rise to bias if such students were atypical.

Hedonism data

The file *hedonism3.txt* (and in SPSS format *hedonism3.sav*), is a sub-set of the data in *hedonism.txt* used in Chapter 12 to illustrate multilevel modelling. There twenty countries are analysed, here we consider only three.

The data are from the 2002-03 European Social Survey (ESS). The dependent variable is a measure of hedonism, one of ten human values. Further details on value theory and how it is operationalised in the ESS can be found at http://essedunet.nsd.uib.no/cms/topics/1/

WARNING: The data files include cases with missing values. In the text file, these are given numeric codes (see below) which you need to declare in whatever software package you use for the analysis. In the SPSS data file, missing values have been declared as 'system missing' which means they will be automatically excluded from any analysis.

The file contains the following variables (in this order):

COUNTRY	a categorical variable coded "A" for respondents from Austria, "B" for respondents from Belgium and "C" for respondents from the Czech Republic.
AGE	age in years (missing values coded 999 in the *.txt* file)
GENDER	an indicator variable taking the value 1 for female and 0 for male respondents (missing values coded 9 in the *.txt* file)
INCOME	household income (missing values coded 99 in the *.txt* file)
EDYRS	years in education (missing values coded 99 in the *.txt* file)
HEDSCORE	hedonism score which is designed as a measure of the relative importance of hedonism to an individual in their whole value system.