Appendix A

Data and Stories

This appendix describes the data used for the examples in *Multilevel Analysis: Techniques and Applications, 2nd edition*. Some of the example data are real data; other data sets have been simulated especially for their use in this book. The simulated data sets have been constructed following some hypothetical but plausible real-world scenario. This appendix describes the various data sets, giving either a reference to the study where they come from, or the 'story' that has been used as a template to generate the data.

Data are currently available on the Internet in SPSS system-file and portable file format, and in addition in the format in which they were analyzed for the vook (e.g. HLM or MLwiN files). Most analyses in this book can be carried out by most of the available multilevel software. Obviously, there is a limit to the number of computer packages one can master. Most of the multilevel regression analyses in this book have been carried out in both HLM and MLwiN, and the multilevel SEM analyses have been carried out using LISREL and M*plus*. System files and setups using these packages, where present, will also be made available on the Internet (currently http://www.geocities.com/joophox). I invite users of other multilevel software to use these data for their own learning or teaching. I appreciate receiving data sets and setups that have been transferred to other software systems, so I can make them also available to other users.

The format of the variables is chosen in such a way that writing the variables out in ASCII format results in a file where all variables are separated by at least one space. This file can be read into other programs using the *free format* option. The data sets are described in the order that they are introduced in the book.

POPULARITY DATA

The popularity data in *popular2** are simulated data for 2000 pupils in 100 schools. The purpose is to offer a very simple example for multilevel regression analysis. The main outcome variable is the *pupil popularity*, a popularity rating on a scale of 1-10 derived by a sociometric procedure. Typically, a sociometric procedure asks all pupils in a class to rate all the other pupils, and then assigns the average received popularity rating to each pupil. Because of the sociometric procedure, group effects as apparent from higher level variance components are rather strong. There is a second outcome variable: pupil popularity as rated by their teacher, on a scale from 1-10. The explanatory variables are pupil gender (boy=0, girl=1), pupil extraversion (10-point scale) and teacher experience in years. The pupil popularity data are used as the main example in chapter two. It could also be used with both outcome variables as an example for the multilevel multivariate analysis in chapter 10 (chapter 10 uses the survey meta-analysis data for that purpose; a multivariate multilevel analysis of the popularity data is left as an exercise for the reader). It is also used as the vehicle to compare the different estimation and testing procedures described in chapter 3 and chapter 13. The popularity data have been generated to be a 'nice' well-behaved data set: the sample sizes at both levels are sufficient, the residuals have a normal distribution, and the multilevel effects are strong.

NURSES

The files *nurses.** contain 3-level data from a hypothetical study on stress in hospitals. The data are from nurses working in wards nested within hospitals. It is a cluster-randomized experiment. In each of 25 hospitals, four wards are selected and randomly assigned to an experimental and control condition. In the experimental condition, a training program is offered to all nurses to cope with job-related stress. After the program is completed, a sample of about 10 nurses from each ward is given a test that measures job-related stress. Additional variables are: nurse age (years), nurse experience (years), nurse gender (0=male, 1=female), type of ward (0=general care, 1=special care), and hospital size (0=small, 1=medium, 2=large). The data have been generated to illustrate 3-level analysis with a random slope for the effect of ExpCon.

GPA DATA

The GPA2 data are a longitudinal data set, where 200 college students have been followed 6 consecutive semesters. The data are simulated. In this data set, there are GPA measures on 6 consecutive occasions, with a JOB status variable (how many hours worked) for the same 6 occasions. There are two student-level explanatory variables: the gender (1= male, 2= female) and the high school GPA. These data are used in the longitudinal analyses in chapter 5, and again in the latent curve analysis in chapter 14. There is also a dichotomous student-level outcome variable, which indicates whether a student has been admitted to the university of their choice. Since not every student applies to a university, this variable has many missing values. The outcome variable ‘admitted’ is not used in any of the examples in this book.

These data come in several varieties. The basic data file is *GPA*. In this file, the 6 measurement occasions are represented by separate variables. Some software packages (e.g., Prelis) use this format. Other multilevel software packages (HLM, MLwiN, MixReg, SAS) require that the separate measurement occasions are different data records. The GPA data, arranged in this ‘long’ data format are in the data file *gpalong*. A second data set based on the GPA data is data where a process of panel attrition is simulated. Students were simulated to drop out, partly based on having a low GPA in the previous semester. This dropout process leads to data that are Missing At Random (MAR). A naive analysis on the incomplete date gives biased results. A sophisticated analysis using multilevel longitudinal modeling or SEM with the modern raw data likelihood (available in AMOS, M*plus* and MX, and in recent versions of LISREL) should give unbiased results. Comparing analyses on the complete and the incomplete data sets gives an impression of the amount of bias. The incomplete data are in files *gpamiss* and *gpamislong*. The GPAS data are labeled GPA2 because this file is not completely equal to the GPA data used in the first edition.

THAILAND EDUCATION DATA

The Thailand education data in file *thaieduc* are one of the example data sets that are included with the software HLM (also in the student version of HLM). They are discussed at length in the HLM user’s manual. They stem from a large survey of primary education in Thailand (Raudenbush & Bhumirat, 1992). The outcome variable is dichotomous, an indicator whether a pupil has ever repeated a class (0= no, 1= yes). The explanatory variables are pupil gender (0= girl, 1= boy), pupil pre-primary education (0 =no, 1= yes) and the school’s mean SES. The example in chapter 6 of this book uses only pupil gender as explanatory variable. There are 8582 cases in the file *thaieduc*, but school mean SES is missing in some cases; there are 7516 pupils with complete data.

Note that these missing data have to be dealt with before these data are transported to a multilevel program. In the analysis in chapter 6 they are simply removed using listwise deletion. However, the percentage pupils with incomplete data is 12.4%, which is too large to be simply ignored in a real analysis.

SURVEY RESPONSE META-ANALYSIS DATA

The survey response data used to analyze proportions in chapter 6 are from a meta-analysis by Hox & de Leeuw (1994). The basic data file is *metaresp*. This file contains an identification variable for each study located in the meta-analysis. A mode-identification indicates the data collection mode (face-to-face, telephone, mail). The main response variable is the proportion of sampled respondents who participate. Different studies report different types of response proportions: we have the completion rate (the proportion of participants from the total initial sample) and the response rate (the proportion of participants from the sample without ineligible respondents (moved, deceased, address nonexistent). Obviously, the response rate is usually higher than the completion rate. The explanatory variables are the year of publication and the (estimated) saliency of the survey's main topic. The file also contains the denominators for the completion rate and the response rate, if known. Since most studies report only one of the response figures, the variables 'comp' and 'resp' and the denominators have many missing values.

Some software (e.g., MLwiN) expects the *proportion* of 'successes' and the denominator on which it is based, other software (e.g., HLM) expects the *number* of 'successes' and the corresponding denominator. The file contains the proportion only, the number of successes must be computed from the proportion if the software needs that. The file *multresp* contains the same information, but now in a three-level format useful if the data are analyzed using the multivariate outcome, which is demonstrated in chapter 11.

STREET SAFETY DATA

A sample of 100 streets is selected, and on each street a random sample of 10 persons are asked how often they feel unsafe while walking that street. The safety is asked using three answer categories: 1 = never, 2 = sometimes, 3 = often. Predictor variables are age and gender, street characteristics are an economic index (standardized *Z*-score) and a rating of the crowdedness of the street (7point scale). File: *Safety*. Used in chapter 7 on ordinal data.

EPILEPSY DATA

The epilepsy data come from a study by Leppik et al. (1987). They have been analyzed by many authors, a.o. Skrondal and Rabe-Hesketh (2004). The data come from a randomized controlled study on the effect of an anti-epileptic drug versus a placebo. It is a longitudinal design. For each patient the number of seizures was measured for a two week baseline. Next, patients were randomized to the drug or the placebo condition. For four consecutive visits the clinic collected counts of epileptic seizures in the wto weeks before the visit. The data set contains the following variables: count of seizures, treatment indicator, visit nr., dummy for visit #4, log of age, log of baseline count. All predictors are grand mean centered. The data

come from the gllamm homepage at www.gllamm.org/books. Used in chapter 7 on count data.
(Leppik I.E., Dreifuss F.E., Porter R., Bowman T., Santilli N., Jacobs M., Crosby C., Cloyd J., Stackman J., Graves N., et al. (1987). A controlled study of progabide in partial seizures: methodology and results. *Neurology*, *37*, 963–968.)

FIRST SEX DATA

A data set from Singer and Willett's book on longitudinal data analysis (2003), from a study by Capaldi, Crosby and Stoolmiller (1996). 180 middle school boys were tracked from the 7th through the 12th grade, the outcome measure is when they had sex for the first time. At the end, 54 boys (30%) were still virgins. These observations are censored. File"*firstsex*, used as an example of (single level) survival data in chapter 800. There is ons dichotomous predictor variable, whether there has been a parental transition (0 if the boy lived with his biological parents before the data collection began.

SIBLING DIVORCE

Multilevel survival data analyzed by Dronkers and Hox (2006). The data are from the National Social Science Family Survey of Australia of 1989-1990. In this survey detailed information was collected included educational attainment of respondents, social and economic background, such as parental education and occupational status of the father, parental family size and family form and other relevant characteristics of 4513 men and women in Australia. The respondent answered all these questions also about his or her parents and siblings. The respondents gave information about at most three siblings, even if there were more siblings in the family. All sibling variables were coded in the same way as the respondents, and all data were combined in a file with respondents or siblings as the unit of analysis. In that new file respondents and siblings from the same family had the same values for their parental characteristics, but had different values for their child characteristics. The data file contains only those respondents or siblings that were married or had been married, and gave no missing values. File: *sibdiv*.

PUPCROSS DATA

This data file is used to demonstrate the cross-classified data with pupils nested within both primary and secondary schools. These are simulated data. One thousand pupils have gone to 100 primary and subsequently to 30 secondary schools. There is no complete nesting structure; the pupils are nested within the cross-classification of primary and secondary schools. The file *pupcross* contains the secondary school-achievement score, which is the outcome variable, and the explanatory pupil-level variables gender (0= boy, 1= girl) and SES. School-level explanatory variables are the denomination of the primary and the secondary school (0= no, 1 =yes). These data are used for the example of a cross-classified analysis in chapter 8.

SOCIOMETRIC SCORES DATA

The sociometric data are simulated data, intended to demonstrate a data structure where the cross-classification is at the lowest level, with an added group structure because there are several groups. The story is that in small groups all members are asked to rate each other. Since the groups are of different sizes, the usual data file organized by case in *socscors* has many missing values. The data are rearranged in data file *socslong* for the multilevel analysis. In *soclong* each record is defined by the sender-receiver pairs, with explanatory variables age and sex defined separately for the sender and the receiver. The group variable ‘group size’ is added to this file.

SCHOOL MANAGER DATA

The school manager data are from an educational research study (Krüger, 1994). In this study, male and female school managers from 98 schools were rated by 854 pupils. The data are in file *manager*. These data are used to demonstrate the use of multilevel regression modeling for measuring context characteristics (here: the school manager’s management style). The questions about the school manager are question 5, 9, 12, 16, 21 and 25; in chapter 9 of the book these are renumbered 1…6. These data are used only to demonstrate the multilevel psychometric analyses in chapter 9. They can also be analyzed using one of the multilevel factor analysis procedures outlined in chapter 12. The data set also contains the pupils’ and school manager’s gender (1= female, 2= male), which is not used in the example. The remaining questions in the data set are all about various aspects of the school climate; a full multilevel exploratory factor analysis is a useful approach to these data.

SOCIAL SKILLS META-ANALYSIS DATA

The social skills meta-analysis data in file *meta20* contain the coded outcomes of 20 studies that investigate the effect of social skills training on social anxiety. All studies use an experimental group/control group design. Explanatory variables are the duration of the training in weeks, the reliability of the social anxiety measure used in each study (2 values, taken from the official test manual), and the studies’ sample size. The data are simulated.

ASTHMA & LRD META ANALYSIS DATA

The asthma and LRD data are from Nam, Mengersen and Garthwaite (2003). The data are from a set of studies that investigate the relationship between children’s environmental exposure to smoking (ETS) and the child health outcomes asthma and lower respiratory disease (LRD). Available are the logged odds-ratio (LOR) for asthma and LRD, and their standard errors. Study level variables are average age of subjects, publication year, smoking (0 parents, 1 other in household), and covariate adjustment used (0=not, 1=yes).

There are two effect sizes, the logged odds ratio for asthma and lower respiratory disease (LRD). Only a few studies report both. Datafile: *AstLrd*.

ESTRONE DATA

The estrone data are 16 independent measurements of the estrone level of 5 post-menopausal women (Fears et al., 1996). The data file *estronex* contains the data in the usual format, the file *estrflat* contains the data in the format used for multilevel analysis. Although the data structure suggests a temporal order in the measurements, there is none. Before the analysis, the estrone levels are transformed by taking the natural logarithm of the measurements. The estrone data are used in chapter 13 to illustrate the use of advanced estimation and testing methods on difficult data. The difficulty of the estrone data lies in the extremely small sample size and the small value of the variance components.

GOOD89 DATA

The file *good89* (from Good, 1999, p. 89) contains the very small data set used to demonstrate the principles of bootstrapping in chapter 13.

FAMILY IQ DATA

The Family IQ data are patterned to follow the results from a study of intelligence in large families (van Peet, 1992). They are the scores on six subscales from an intelligence test. They are used in chapter 14 to illustrate multilevel factor analysis. The file FamilyIQ contains the data from 275 children in 50 families. The data file contain the additional variables gender and parental IQs, which are not used in the analyses in this book. Datafile: *FamIQ*.

GALO DATA

The GALO data in file *galo* are from an educational study by Schijf & Dronkers (1991). They are data from 1377 pupils within 58 schools. We have the following pupil level variables: father's occupational status *focc*, father's education *feduc*, mother's education *meduc*, pupil sex *sex*, the result of GALO school achievement test *GALO*, and the teacher's advice about secondary education *advice*. On the school level we have only one variable: the school's denomination *denom*. Denomination is coded 1= Protestant, 2= Nondenominational, 3= Catholic (categories based on optimal scaling). The data file *galo* contains both complete and incomplete cases, and an indicator variable that specifies whether a specific case in the data file is complete or not.