Note: This is an expanded version of the Appendix to: Joop Hox (2002). *Multilevel Analysis. Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates. Any corrections to the text in the book are in red. Additions to the book text are set in a different font. These data are made available for use as example data or exercises. (May 10, 2002)

Appendix A: Data and Stories

This appendix describes the data used for the examples in *MULTILEVEL ANALYSIS. TECHNIQUES AND APPLICATIONS*. Some of the example data sets 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. 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 AMOS, LISREL and M*PLUS*. System files and setups using these packages, where present, will also be made available on the Internet (currently at: http://www.fss.uu.nl/ms/jh). 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.

Note that in most files the unit and group identification variables are defined as numeric. If these files are to be read into HLM directly, the type of the identification variables must be changed into *string* first. 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.

POPULARITY DATA

The popularity data in the file POPULAR 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 1-7. The explanatory variables are pupil gender (boy=0, girl=1) 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 9 (chapter 9 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 11. 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.

GPA DATA

The GPA 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 (HLM, Prelis) use this format. Other multilevel software packages (MLwiN, MixReg, SAS) require that the separate measurement occasions are different data records. The GPA data, arranged in this 'flat' data format are in the data file GPAFLAT. 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. This analysis is referred to in chapter 5, but not presented. The incomplete data are in files GPAMISS and MISFLAT.

Note that to use these data in a program like MLwiN that expects a separate line of data for each measurement occasion, the data must be written out to produce such a 'flat' file (cf. the GPA data files). See the discussion of aggregation and disaggregation added to the end of the data description.

CHILDREN'S VOCABULARY DATA

The children's vocabulary growth data are longitudinal data, 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 Bryk and Raudenbush (1992) and in the HLM user's manual. The data are a combination of the data from two different studies. Since the two studies use different time periods between vocabulary measurements, and there are some other missing data as well, the result is a very unbalanced data set. This data set is used in chapter 5 to illustrate the issues in longitudinal modeling using 'real time' instead of fixed occasions. The data are in file VOCAGRWT. They contain an identification variable for the measurement occasion, the child's age in months at that occasion, and the vocabulary size at that measurement occasion. Since the vocabulary size at age zero is undefined, it makes sense to center the age variable on some sensible time point. The data set includes a variable 'age12', the age centered on 12 months, and 'age12sq' which is the square of 'age12'. Time invariant covariates are the child's gender (0= boy, 1= girl) and the amount of maternal speech measured at one specific occasion. These variables are not used in this book; Bryk and Raudenbush (1992) model the vocabulary growth using these variables as well. In the course of analyzing these data for this book, I found that they are an interesting example, because they are difficult to analyze, most probably because of their small sample size and small variances at time zero. MLwiN (vs. 1.10) has problems analyzing the more complicated models, HLM (vs. 5.02) does not. This shows that with marginally sufficient data (small samples, complex models) the details of the software implementation can become important. With sufficiently large data sets, such occurrences are rare; Kreft, de Leeuw and van der Leeden (1994) found almost no differences in their review of five multilevel software packages.

THAILAND EDUCATION DATA

The Thailand education data 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 the missing data have to be dealt with before these data are transported to a multilevel program.

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 9.

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 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 7.

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 SOCSFLAT for the multilevel analysis. In SOCSFLAT 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.

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.

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.

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 11 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 11.

VAN PEET DATA

The van Peet data are 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 12 to illustrate multilevel factor analysis. There are two files: PEETCOMP contains the complete data for 187 children from 37 families, and PEETMIS contains in addition the incomplete data. This data file is interesting because the data set is actually rather small for a SEM analysis, which shows up in (small, insignificant) negative variance estimates. The data file contains the additional variable gender, which is not used in the analyses in this book.

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.

Appendix B:
Aggregating and Disaggregating in SPSS

A common procedure in multilevel analysis is to aggregate individual level variables to higher levels. In most cases, aggregation is used to attach to higher level units (e.g., groups, classes, teachers) the mean value of a lower level explanatory variable. However, other aggregation functions may also be useful. For instance, one may have the hypothesis that classes that are heterogeneous with respect to some variable differ from more homogeneous classes. In this case, the aggregated explanatory variable would be the group's standard deviation or the range of the individual variable. Another aggregated value that can be useful is the group size.

In SPSS, aggregation is handled by the procedure *aggregate*. This procedure produces a new file that contains the grouping variable and the (new) aggregated variables. In SPSS/Windows *aggregate* is available in the DATA menu. A simple syntax to aggregate the variable IQ in a file with grouping variable GROUPNR is as follows:

```
GET FILE `indfile.sys'.
AGGREGATE OUTFILE='aggfile.sys'/BREAK=groupnr/
  meaniq=MEAN(iq)/stdeviq=SD(iq).
```

Disaggregation means adding group level variables to the individual data file. This creates a file where the group level variables are repeated for all individuals in the same group. In SPSS, this can be accomplished by the procedure JOIN MATCH, using the

so-called TABLE lookup. Before JOIN MATCH is used, the individual and the group file must both be sorted on the group identification variable. In SPSS/Windows JOIN MATCH is available in the DATA menu. For instance, if we want to read the aggregated mean IQ and IQ standard deviation to the individual file, we have the following setup:

```
JOIN MATCH FILE='indfile.sys'/
  TABLE='aggfile.sys'/BY groupnr/MAP.
```

The example below is a complete setup that uses aggregation and disaggregation to get group means and individual deviation scores for IQ:

```
GET FILE `indfile.sys'.
SORT groupnr.
SAVE FILE `indfile.sys'.
AGGREGATE OUTFILE='aggfile.sys'/PRESORTED/BREAK=groupnr/
  meaniq=MEAN(iq)/stdeviq=SD(iq).
JOIN MATCH FILE='indfile.sys'/
  TABLE='aggfile.sys'/BY groupnr/MAP.
COMPUTE deviq=iq-meaniq.
SAVE FILE `indfile2.sys'.
```

In this setup I use the AGGREGATE subcommand PRESORTED to indicate that the file is already sorted on the BREAK variable groupnr, because this saves computing time. The subcommand MAP on the JOIN MATCH procedure creates a map of the new system file, indicating from which of the two old system files the variables are taken. In this kind of `cutting and pasting' it is extremely important to check the output of both AGGREGATE and JOIN MATCH very carefully to make sure that the cases are indeed matched correctly.

It should be noted that the program HLM contains a built-in procedure for centering explanatory variables. The program MLwiN has a procedure to add group means to the individual data file, and commands to create centered and group-centered variables. In most other programs

A particular form of disaggregation is when we have a file with repeated measures, with repeated measures represented by separate variables. Many programs nead data where each measurement occasion is seen as a separate row of data, with time invariant

variables repeated in the new data file. The GPA data are a good example. To create the 'flat' data file needed I used the following SPSS syntax:

```
GET FILE 'd:\data\gpa.sav'.
WRITE OUTFILE 'd:\data\gpaflat.dat' RECORDS=6/
      student '  0  ' gpa1 job1 sex highgpa /
      student '  1  ' gpa2 job2 sex highgpa /
      student '  2  ' gpa3 job3 sex highgpa /
      student '  3  ' gpa4 job4 sex highgpa /
      student '  4  ' gpa5 job5 sex highgpa /
      student '  5  ' gpa6 job6 sex highgpa .
EXECUTE.
DATA LIST FILE 'd:\ data\gpaflat.dat' FREE /
      student occasion gpa job sex highgpa .
SAVE OUTFILE 'd:\ data\gpaflat.sav' .
DESCRIPTIVES ALL.
```

This syntax fist writes out the data in ASCII format, and then reads these into SPSS using the DATA LIST command using a different structure. The final command DESCRIPTIVES is use to check if all variables have plausible values.

A complication arises if the original data file has missing values. These are often coded in SPSS as system missing values, which are written out as *spaces*. When the command DATA LIST 'filename' FREE / is used in SPSS, these are read over in after the first such occurrence all other variables have incorrect values. To prevent this, we need to insert a command that recodes all system missing values into a real code, and after creating the flat data file records which contain such missing value codes must be removed. To create the 'flat' data file needed from the incomplete data set GPAMISS I used the following SPSS syntax:

```
GET FILE 'd:\data\gpamiss.sav' .
RECODE gpa1 to job6 (SYSMIS=9).
WRITE OUTFILE 'd:\joop\Lea\data\misflat.dat' RECORDS=6/
      student '  0  ' gpa1 job1 sex highgpa /
      student '  1  ' gpa2 job2 sex highgpa /
      student '  2  ' gpa3 job3 sex highgpa /
      student '  3  ' gpa4 job4 sex highgpa /
      student '  4  ' gpa5 job5 sex highgpa /
      student '  5  ' gpa6 job6 sex highgpa .
EXECUTE.
DATA LIST FILE 'd:\ data\misflat.dat' FREE /
      student occasion gpa job sex highgpa .
COUNT out=gpa job (9).
SELECT IF (out=0).
SAVE OUTFILE 'd:\ data\misflat.sav' .
```