South African Journal of Education
Copyright © 2003 EASA
Vol 23(4) 249 – 253

Understanding the structure of data when planning for analysis: application of Hierachical Linear Models

Mbithi wa Kivilu
Surveys, Analysis, Modelling & Mapping (SAMM), Human Sciences Research Council, Private Bag X41, Pretoria, 0001 South Africa
JMKivilu@hsrc.ac.za

Human beings and other living creatures tend to exist within organisational structures, such as families, schools, and business organisations. In an educational system, for example, students exist within a hierarchical social structure that can include classroom, grade level, school, school district and country. Data obtained from such social structures are hierarchical. It is critical that social scientists understand the structure of the data because it dictates the statistical techniques to be used for analysis and interpretation. For example, analysing hierarchical data using the conventional General Linear Models (GLMs) may result in inaccurate inferences being drawn from the data. A thorough understanding of the data in terms of structure, type of variables and relationships being investigated needs no further emphasis. Statistically valid inferences are drawn from data that have been carefully collected and subjected to the appropriate statistical techniques. Attention should also be paid to the underlying assumptions of a particular statistical technique. Use of Hierarchical Linear Models (HLMs) in analysing social research has several advantages. The problem of unit of analysis is avoided and data are no longer aggregated or disaggregated resulting in accurate and reliable estimation of each level effects. Furthermore, all estimated effects are adjusted for individual level and group level influence on the outcome variable. The only drawback of applying HLMs is that this requires an advanced level of sophistication in statistics.

Introduction

Central to any successful research endeavour is the planning of the research process in such a way that the key phases of the research are clearly defined. This planning is done during the development of the research design. Research design is developed to align the pursuit of a research goal with the practical considerations and limitations of the project. Emphasis is placed upon the fact that design and planning are directly related to the degree of structure and control in the research project. Selltiz, Jahoda, Deutsch and Cook (1965:50), in their classical book on research methodology, define research design as the arrangement of conditions for collecting and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure. According to Mouton and Marais (1996), the aim of a research design is to plan and structure a given research project in such a manner that the eventual validity of the research findings is maximised. Typical research decisions made in research design include: choice of the research area or topic, problem formulation, conceptualisation and operationalisation of variables and relationships, sampling, and data collection, analysis and interpretation of data. Available resources and potential limitations that would hamper the maximisation of validity of the research findings are critical considerations when making decisions in research design.

Although the choice of a research topic in contract research, by definition, is not in the hands of the researchers, they have a greater responsibility to ensure that the formulation of the problem is as objective and critical as possible. Contract research does not, however, imply that the organisation requesting the research determines the direction thereof, and the manner in which data are collected and analysed at the stage of entering into the contract.

The validity of the inferences drawn from a research study will depend on how well each of the phases is implemented and the quality of the resulting data. One aspect of a research design that receives little attention from social scientists is the structure of social science data. It is critical that the researcher understands the structure of the data because it dictates the statistical techniques used for analysis and interpretation. It could also be argued that the statistical techniques used for testing research hypotheses could dictate the way the data is structured for analysis.

The structure of data from social organisations

Human beings and other living creatures tend to exist within organisational structures, such as families, schools, business organisations, churches, towns, districts, provinces and countries. In education, for example, students exist within a hierarchical social structure that can include classroom, grade level, school, school district, province and country. Workers, on the other hand exist within production or skill units, businesses and sectors of the economy, as well as geographic regions. Health care workers and patients exist within households and families, medical practices and facilities (a doctor's practice, or hospital), districts, provinces and countries. Many other social communities exhibit hierarchical (also referred to as multilevel) data structures as well. A hierarchical educational data structure with four levels is illustrated in Figure 1. Learners at level 1 are nested within classroom at level 2, which are nested within schools at level 3 and finally schools are nested within district at level 4.

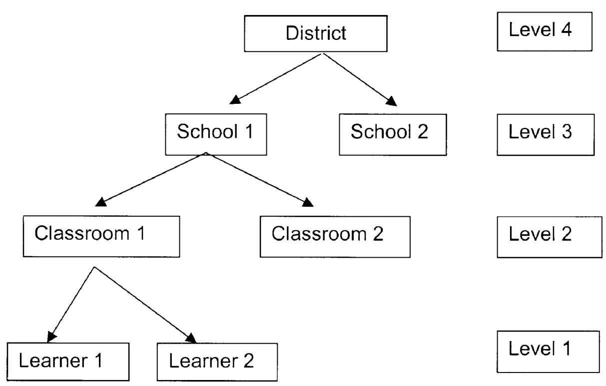

Figure 1 Hierarchical structure of educational data

Bryk & Raudenbush (1992) also discuss two other types of data hierarchies that are less obvious: repeated-measures data and meta-analytic data. Data repeatedly gathered on an individual is hierarchical as all the observations are nested within individuals. Whilst there are other adequate procedures for dealing with this sort of data, the assumptions relating to them are rigorous. Also, when researchers are engaged in the task of meta-analysis, or analysis of a large number of existing studies, it should become clear that subjects, results, procedures, and experimenters are nested within studies.

Despite the prevalence of hierarchical structures in behavioural and social research, past studies have often failed to address them adequately in the data analysis phase. This neglect has reflected limitations in conventional statistical techniques for the estimation of linear models with nested structures. In social sciences research, these limitations have generated concerns about aggregation bias, misestimated precision, and the "unit of analysis problem". The relevance of linear hierarchical models in the analysis of data with nested structure is discussed next.

Assumptions in statistical techniques

Most statistical methods such as the General Linear Models (GLM) (e.g. Multiple Regression and Analysis of Variance (ANOVA)) rely upon certain assumptions about the variables used in the analysis. When these assumptions are not met, the results may not be valid, resulting in a Type I error[1] or Type II error,[2] or over- or under-estimation of significance or effect size(s) (Osborne, Christensen & Gunter, 2001). In other words, these techniques are not robust to the violation of the assumptions. As Pedhazur (1997:33) notes, "Knowledge and understanding of the situations when violations of assumptions lead to serious biases, and when they are of little consequence, are essential to meaningful data analysis". However, as Osborne, Christensen and Gunter (2001) observe, few articles report having tested assumptions of the statistical tests they rely on when drawing their conclusions. This creates a situation where we have a rich literature in education and social science, but we are forced to call into question the validity of many of these results, conclusions, and assertions, as we have no idea whether the assumptions of the statistical tests were met. A discussion of some of the critical assumptions of GLMs is presented because Hierarchical Linear Models (HLMs) are developed using some of the principles of GLMs.

Several assumptions of GLMs are "robust" to violation (e.g. normal distribution of errors), and others are fulfilled in the proper design of a study (e.g. independence of observations). These and other assumptions that are not robust to violation and that researchers can deal with if violated are the focus of this article. More specifically, the article focuses on the assumptions of independence, linearity, reliability of measurement and normality.

Why is a hierarchical data structure an issue?

Hierarchical, or nested data present several problems for analysis. First, people or creatures that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population. For example, students in a particular third-grade classroom are more similar to each other than to students randomly sampled from the school district as a whole, or from the national population of third-graders. This is because students are not randomly assigned to classrooms from the population, but rather are assigned to schools based on geographic factors. Thus, students within a particular classroom tend to come from a community or community segment that is more homogeneous in terms of morals and values, family background, socio-economic status, race or ethnicity, religion, and even educational preparation than the population as a whole. Furthermore, students within a particular classroom share the experience of being in the same environment, the same teacher, physical environment and similar experiences, which may lead to increased homogeneity over time. As a result, the average correlation (expressed in intra-class correlation) between variables measured on students from the same schools will be higher than the average correlation between variables measured on students from different schools. Standard statistical tests lean heavily on the assumption of independence of the observations. If this assumption is violated (as is usually the case in hierarchical data) the estimates of the standard errors of conventional statistical tests are much too small and this results in many spuriously significant results.

The problem of dependencies between individual observations also occurs in survey research, when the sample is not taken at random but cluster sampling from geographical areas is used instead (Hox, 1995:7). For similar reasons as in the school example, respondents from the same geographical areas might be more similar to each other than respondents from different geographical areas. The result is again estimates of standard errors that are too small, leading to spurious significant results.

In most hierarchical data structure, we have not only the clustering of individuals within groups, but we also have variables measured at all available levels. Hierarchical linear models, also known as multilevel models, are designed to analyse variables from different levels simultaneously, using a statistical model that includes the various dependencies (Hox, 1995:7).

Unit of analysis

How to deal with cross-level data?

It is often the case that a researcher especially in education is interested in understanding how environmental variables (e.g. teaching style, teacher behaviours, class size, class composition, district policies or funding, or even provincial or national variables) affect individual outcomes (e.g. achievement, attitudes and retention). But given that outcomes are gathered at the individual level, and other variables at classroom, school, district, province, or nation level, the question arises as to what the unit of analysis should be, and how to deal with the cross-level nature of the data.

Units of analysis are determined by the research themes or questions. Most common units of analysis used in behavioural and social research are individual human beings, groups, organisations, artefacts and dimension of time. These are discussed as follows:

- Individuals: Individual human beings are probably the most common typical objects of research in the social sciences. Data from individuals are used to aggregate to group levels.
- Groups: A group possesses characteristics that are not necessarily applicable to the behaviour of individuals, for example, families, gangs, census blocks and couples.
- Organisations: Research in this unit focuses on the unique qualities of the social organisations such as organisational structure, lines of authority and promotional policy.
- Social artefacts: These include products of human behaviour such as social objects (e.g. paintings, buildings and songs) and social interactions (e.g. marriage ceremonies, family violence, adolescence delinquency and prostitution).
- Dimension of time: Data collected over time is analysed using the time as one of the variables, for example, in longitudinal studies (cohort and panel studies) and cross-sectional studies.

Threats to validity associated with the unit of analysis

Variables can be measured directly at their natural level, for example, at the school level we may measure school size and student composition (proportion of students by race), and at the student level measure of attitude and academic achievement. In addition we may convert variables from one level to another by aggregation or disaggregation. Aggregation means that the variables at a lower level are transferred to a higher level, for instance, by computing the school mean score from the individual student's test scores. Disaggregation means moving variables to a lower level, for instance by assigning to all students a variable that reflects the composition of the school they belong to, such as gender or race.

Historically, analyses of hierarchical data have led to analysis approaches that move all variables by aggregation or disaggregation to one single level of interest followed by ordinary multiple regression or ANOVA, or some other 'standard' analysis method. Analysing variables from different levels at one single common level creates two different sets of problems as discussed by Hox (1995):

1. *Statistical problem*

One set of problems is statistical. If data are aggregated, the result is that different data values from many sub-units are combined into

---

1 Type I error is the probability of rejecting the null hypothesis when it is true.

2 Type II error is the probability of failing to reject the null hypothesis when it is false.

fewer values for fewer higher-level units. Using school data one would aggregate individual student data up to the level of the classroom, school or district. Thus, we could talk about the effect of teacher or classroom characteristics on average classroom achievement. However, there are several problems with this approach that include:

- That much (80–90%) of the individual variability on the outcome variable is lost, which can lead to dramatic under- or over-estimation of observed relationships between variables (Bryk & Raudenbush, 1992); and
- The outcome variable changes significantly and substantively from individual achievement to average classroom achievement.

On the other hand, if data are disaggregated, the result is that a few data values from the small number of super-units are 'blown up' into values for a much larger number of sub-units. Using the school data, one strategy would be to assign classroom or teacher characteristics to all students (i.e. to bring the higher-level variables down to the lower level). The problem with this approach, again, is non-independence of observations, as all students within a particular classroom assume identical scores on a given variable.

Ordinary statistical tests treat all these disaggregated data values as independent information from this much larger sample. The proper sample size for these variables is of course the number of higher-level units. Using the higher number of disaggregated cases for the sample size leads to significance tests that reject the null-hypothesis far more often than the nominal alpha level suggests. In other words: investigators come up with a lot of spurious significant results. Both these strategies prevent the researcher from disentangling individual and group effects on the outcome of interest. As neither one of these approaches is satisfactory the hierarchical linear modelling becomes necessary.

2. *Conceptual problem*

The other set of problems encountered is conceptual. If the analyst is not very careful in the interpretation of the results s/he may commit the fallacy of the wrong level, which consists of analysing the data at one level, and drawing conclusions at another level. Probably the best known fallacy is the ecological fallacy, which refers to a situation where aggregated data is interpreted at the individual level. It is also known as the 'Robinson effect' after Robinson (1950). Robinson argued that an ecological correlation (correlation between the aggregated variables) is almost certainly not equal to its corresponding individual correlation. This has consequences the other way as well; drawing inferences at a higher level from analyses performed at a lower level is just as misleading; this error is known as the atomistic fallacy. A different but related fallacy is known as 'Simpson's paradox' (Lindley & Novick, 1981). Simpson's paradox refers to the problem that completely erroneous conclusions may be drawn if grouped data, drawn from heterogeneous populations, are collapsed and analysed as if they came from a single homogeneous population.

The following example illustrates the problem of ecological fallacy in data interpretation. Suppose we have the data of school mean scores in the last Senior Certificate Examination in South Africa. Assume that our interest is in the academic performance of schools that have a high proportion of female candidates. We have data at our disposal on the performance patterns in the different schools and also school statistics on the demographic composition of the schools. According to our analysis we find that among the top 10% of the schools, there was a higher proportion of schools with higher proportion of female candidates. One would be inclined to conclude that female candidates were more likely to perform better than their male counterparts. In doing this we would be in danger of committing the ecological fallacy. It might just as well have been the male candidates in the schools with higher proportion of female candidates did equally well. The problem here is related to the fact that we used school as our unit of analysis to arrive at conclusions about the behaviour of individual candidates.

Another conceptual problem associated with the unit of analysis problem is referred to as the reductionistic tendencies. The term *reductionistic tendencies* is used to refer to the situations where researchers tend to consider and present only those ex-planations and interpretations which are embedded in discipline- specific variables (Mouton & Marais, 1996:42). The problem arises when one of these approaches is given more prominence at the expense of the others.

The following example illustrates the problem of reductionistic tendencies in data interpretation. Suppose a panel of researchers drawn from disciplines such as psychology, economics, health, anthropology and geography were asked to explain why there is a higher rate of infection of preventable diseases in Sub-Saharan Africa than any other part of the Globe. Psychologists may attribute it to health seeking behaviour, health specialists may consider lack of immunisation or environmental causes, anthropologist may consider the cultural practices, economists may look at income of the households while geographers may consider migration patterns. All these explanations are plausible in their own discipline in explaining the phenomenon and therefore should be given relatively equal attention.

Solutions to both the statistical and conceptual problems associated with the issue of the unit of analysis have been proposed in literature (e.g. Mouton & Marais, 1996:42):

- A critical awareness of the unit of analysis when conclusions are reached about the data. Since ecological fallacy involves a threat to inferential validity, claims made in the conclusions reached must be supported by the data or information collected; and
- A critical awareness of the limitations of the scope of a given discipline in explaining a given phenomenon. The limitations of any given single discipline make it desirable that interdisciplinary strategies be used. By involving specialists from other disciplines, the probability of reductionism is, to some extent reduced.

Assumption of normality

The GLMs assume that the dependent variable follows a normal distribution. Non-normally distributed variables which are skewed and have large kurtosis with substantial outliers can distort relationships and significance tests. There are several pieces of information that are useful to the researcher in testing this assumption:

- Visual inspection of data plots,
- Skewness and kurtosis indices give researchers information about normality, and
- Statistical tests provide inferential statistics on normality (Osborne, Christensen & Gunter, 2001). Outliers can be identified either through visual inspection of histograms or frequency distributions, or by converting data to z-scores (or Q-Q plots).

Bivariate/multivariate data cleaning procedures can also be important (Tabachnick & Fidell, 2001:139) in establishing normality in multiple regression analysis. Most regression or multivariate statistics texts (e.g. Pedhazur, 1997; Tabachnick & Fidell, 2001) discuss the examination of standardised or studentised residuals, or indices of leverage. Analyses by Osbome (2001) show that removal of univariate and bivariate outliers can reduce the risk of Type I and Type II errors, and improve accuracy of estimates.

Removal of outlier (univariate or bivariate) is straightforward in most statistical software packages. However, it is not always desirable to remove outliers. Transformations (e.g. square root, log, or inverse), can improve normality, but complicate the interpretation of the results, and should be used deliberately and in an informed manner.

Assumption of a linear relationship between the dependent and independent variable(s)

Standard GLMs can only accurately estimate the relationship between dependent and independent variables if the relationships are linear in nature. As there are many instances in the social sciences where non-linear relationships occur, it is essential to examine analyses for non-linearity. If the relationship between independent variables and the dependent variable is not linear, the results of the analysis will under-estimate the true relationship. According to Hox (1995), this under-estimation carries two risks: increased risk of a Type II error for that independent variable and in the case of multiple regression, an increased risk of Type I errors (over-estimation) for other independent variables that share variance with that independent variable.

Three primary ways to detect non-linearity have been suggested by Pedhazur (1997), Cohen and Cohen (1983), and Berry and Feldman (1985). The first method is the use of theory or previous research to inform current analyses. However, as many prior researchers have probably overlooked the possibility of non-linear relationships, this method is not foolproof. A preferable method of detection is examination of residual plots (plots of the standardised residuals as a function of standardised predicted values, readily available in most statistical software).

The third method of detecting curvilinearity is to routinely run GLM analyses that incorporate curvilinear components (squared and cubic terms) or utilising the non-linear regression option available in most statistical packages. It is important that the non-linear aspects of the relationship be accounted for in order to best assess the relationship between variables.

Assumption that variables are measured without error (reliably)

The nature of our educational and social science research means that many variables we are interested in are also difficult to measure, making measurement error a particular concern. In simple correlation and regression, unreliable measurement causes relationships to be under-estimated increasing the risk of Type II errors. In the case of multiple regression or partial correlation, effect sizes of other variables can be over-estimated if the covariate is not reliably measured, as the full effect of the covariate(s) would not be removed. This is a significant concern if the goal of research is to accurately model the "real" relationships evident in the population. Although in psychometrics, it is assumed that reliability estimates (Cronbach alpha) of 0.7–0.8 are acceptable (e.g. Nunnally, 1978), Osborne, Christensen and Gunter (2001) reported that the average alpha reported in top educational psychology journals was 0.83, measurement of this quality still contains enough measurement error to make correction worthwhile.

Correction for low reliability is simple, and widely disseminated in most texts on regression, but rarely seen in the literature. Authors should correct for low reliability to obtain a more accurate picture of the "true" relationship in the population, and, in the case of multiple regression or partial correlation, to avoid over-estimating the effect of another variable. Since "the presence of measurement errors in behavioural research is the norm rather than the exception" and "reliabilities of many measures used in the behavioural sciences are, at best, moderate" (Pedhazur, 1997:172); it is important that researchers be aware of accepted methods of dealing with this issue.

In multiple regression, for example, with each independent variable added to the regression equation, the effects of less than perfect reliability on the strength of the relationship becomes more complex and the results of the analysis more questionable. With the addition of one independent variable with less than perfect reliability each succeeding variable entered has the opportunity to claim part of the error variance left over by the unreliable variable(s). The apportionment of the explained variance among the independent variables will thus be incorrect. The more independent variables added to the equation with low levels of reliability the greater the likelihood that the variance accounted for is not apportioned correctly. This can lead to erroneous findings and increased potential for Type II errors for the variables with poor reliability, and Type I errors for the other variables in the equation. Obviously, this gets increasingly complex as the number of variables in the equation grows.

How do hierarchical models work?

One goal of this article is to explain the concept of hierarchical modelling and explicate the need for the procedure in analysing hierarchical data. It cannot fully communicate the meaning and procedures needed to actually perform a hierarchical analysis. The reader is encouraged to refer to Bryk and Raudenbush (1992).

HLMs appear in diverse literature under a variety of titles. In sociological research, they are referred to as multi-level linear models (Goldstein, 1987; Mason, Wong & Entwisle, 1983). In biometric applications the terms mixed-effects models and random-effect models are used (Elston & Grizzle, 1962; Laird & Ware, 1982). They are called random-coefficient regression models in econometrics literature (Rosenberg, 1973; Longford, 1993) and in statistical literature they are often referred to as covariance components models (e.g. Longford, 1987).

In hierarchical data, such as found in educational organisations, school effects and certain learner characteristics require simultaneous exploration of relationships at the within- and between-school levels. Early school effects research relied primarily on single-level multiple regression models at either the learner level or the school level. These designs failed to adequately model the hierarchical structure of the learner-to-class-to-school relationships. Treating data as if they were all at the same unit of analysis has implications for statistical validity in that it has led researchers to misleading conclusions about the effects (or non-effect) of various aspects of the school environment on learner attributes.

The HLM approach allows for explicit modelling of effects at the various levels of the hierarchy. All estimated effects are adjusted for the individual-level and group-level influence on the dependent variables. HLM is a regression-like technique that proceeds as follows: A learner level linear regression model is estimated for each school to predict learners' measure of performance using learners' characteristics. Simultaneously, at the school level, a regression model is defined using school characteristics to estimate the parameters obtained at the learner-level. Conceptually, HLM entails an estimation of regressions of regression results, except that the equations at each level are estimated at the same time and the variance at one level is taken into account in estimating the next level (Raudenbush *et al*., 2000). In addition, HLM allows the examination of the correlation of school characteristics with the between-learner characteristics.

The general HLM models

HLMs can be developed for hierarchical data with two or three levels. The three-level model consists of three sub-models, one for each level. For example, if the research problem consists of data on students nested within classrooms and classrooms nested within schools, the level 1 model will represent the relationships among the student-level variables, the level 2 model will capture the influence of class level factors, and the level 3 will incorporate school level effects. The equations for a three-level HLM are presented in an Appendix.

A number of computer programmes are available for analysis of both linear and non-linear models found in hierarchical data . For example, HLM 5 (Raudenbush *et al*., 2000), VARL (Longford, 1990) and ML3 (Prosser, Rasbash & Goldstein, 1991). Only the HLM programme is discussed in this article. Certain specifications of the data sets have to be adhered to in order to apply the HLM 5 programme to the data. Although the interpretation of some of the content of the HLM 5 output, such as the regression coefficients, are not different from that of the standard multiple regression, there are a lot of features that are specifically modelled for hierarchical data.

Conclusion

Statistically valid inferences are drawn from data that have been carefully collected and subjected to the appropriate statistical techniques. A thorough understanding of the data in terms of structure, type of variables and relationships being investigated need no further emphasis. The use of HLMs in analysing social research has several advantages. The problem of unit of analysis is avoided and data is no longer aggregated or disaggregated resulting in accurate and reliable estimation of each level effects. Furthermore, all estimated effects are adjusted for individual-level and group-level influence on the outcome variable. The only draw back of applying hierarchical linear models is that it requires advanced level of sophistication in statistics. Advice on statistical techniques, that are appropriate for analysis of various types of data structures, could be sought from research methodologists and statisticians available in research organisations and higher education institutions.

References

Berry WD & Feldman S 1985. Multiple regression in practice. *Sage University Paper Series on Quantitative Applications in the Social Sciences*, Series No. 07-050). Newbury Park, CA: Sage.

Bryk AS & Raudenbush SW 1992. *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Cohen J & Cohen P 1983. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Draper D 1995. Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioural Statistics*, 20:115-147.

Elston RC & Grizzle JE 1962. Estimation of time response curves and their confidence bands. *Biometrics*, 18:148-159.

Goldstein H 1987. *Multilevel Models in Educational and Social Research*. London: Griffin.

Hoffman DA & Gavin MB 1998. Centering decisions in hierarchical linear models: Implications for research in organisations. *Journal of Management*, 24:623-641.

Hox JJ 1995. *Applied multilevel analysis*. Amsterdam: TT-Publikaties. [Online] Available url: http://www.ioe.ac.uk/multilevel/.

Laird NM & Ware H 1982. Random-effects models for longitudinal data. *Biometrics*, 38:963-974.

Lindley DV & Novick MR 1981. The role of exchangeability in inference. *Annuals of Statistics*, 9:45-58.

Longford NT 1987. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74:817-827.

Lonford NT 1990. VARL. *Software for Variance Component Analysis with Nested Random Effects* (*Maximum Likelihood*). Educational Testing Service, Princeton, NJ.

Longford NT 1993. *Random Coefficient Models*. Oxford: Clarendon Press.

Mason WM, Wong GM, & Entwisle B 1884. Contextual analysis through the multilevel linear model. In: S Leinhardt (ed.). *Socilogical Methodology, 1983-84*. San Francisco: Jossey-Bass.

Mouton J & Marais HC 1996. *Basic concepts in the methodology of the social sciences*. Revised edn. Pretoria: Human Sciences Research Council.

Nezlek JB & Zyzniewski LE 1998. Using hierarchical linear modelling to analyze grouped data. *Group Dynamics*, 2:313-320.

Nunnally JC 1978. *Psychometric theory,* 2nd edn. New York: McGraw-Hill.

Osborne JW 2000. Advantages of hierarchical linear modelling. *Practical Assessment, Research & Evaluation*, 7. [Online] Available url: http://ericae.net/pare/getvn.asp?v=7&n=1.

Osborne JW 2001. A new look at outliers and fringeliers: Their effects on statistic accuracy and Type I and Type II error rates. Unpublished manuscript, Department of Educational Research and Leadership and Counsellor Education, North Carolina State University.

Osborne JW, Christensen WR & Gunter J 2001. Educational psychology from a statistician's perspective: A review of the power and goodness of educational psychology research. Paper presented at the national meeting of the *American Education Research Association* (*AERA*), Seattle, WA.

Pedhazur EJ 1997. *Multiple regression in behavioral research*. Harcourt Brace: Orlando, FL.

Prosser R, Rasbash J & Goldstein H 1991. *ML3 Software for Three-Level Analysis. Users' Guide.* London: Institute of Education, University of London.

Raudenbush SW 1995. Reexamining, reaffirming, and improving application of hierarchical models. *Journal of Educational and Behavioral Statistics*, 20:210-220.

Raudenbush S, Bryk A, Cheong YF & Congdon R 2000. *HLM 5. Hierarchical linear and nonlinear modelling*. Illinois: Scientific Software International.

Robinson WS 1950. Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15:351-357.

Rosenberg B 1973. Linear regression with randomly dispersed parameters. *Biometrika*, 60:61-75.

Selltiz C, Jahoda M, Deutsch M & Cook SW 1965. *Research methods in social relations*. Revised edn. New York: Holt, Rinehart & Winston.

Tabachnick BG & Fidell LS 2001. *Using multivariate statistics,* 4th edn. Needham Heights, MA: Allyn and Bacon.

Appendix

The general three-level model

Formally there are $i = 1,..., n_{jk}$ level–1 units (e.g. students), which are nested within each of $j = 1,...,J_k$ level–2 units (e.g. classrooms), which in turn are nested within each of $k = 1,...,K$ level–3 units (e.g. schools).

Level–1 model

In the level–1 model we represent the outcome for case i within level–2 unit j and level–3 unit k as follows:

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}X_{1jk} + \beta_{2jk}X_{2jk} + ... + \beta_{pjk}X_{pjk} + \mathbf{e}_{ijk} \\ &= \beta_{0jk} + \Sigma\beta_{ijk}X_{pjk} + \mathbf{e}_{ijk} \end{aligned}$$

where

β_{pjk} (p = 0,1,....P) are level–1 coefficients,

X_{pjk} is a level–1 predictor p for case i in level–2 unit j and level–3 unit k, $\mathbf{e}_{ijk}$ is the level–1 random effect, and σ^2 is the variance of $\mathbf{e}_{ijk}$, that is, the level–1 variance.

We assume that the random term $\mathbf{e}_{ijk} \sim N(0, \sigma^2)$, that is, it is normally distributed with mean zero and variance σ^2.

Level–2 model

Each of the β_{pjk} coefficients in the level–1 model becomes an outcome variable in the level–2 model:

$$\begin{aligned} \beta_{pjk} &= \alpha_{p0k} + \alpha_{plk}W_{1jk} + \alpha_{p2k}W_{2jk} + ... + \alpha_{pQk}W_{Qjk} + \mathbf{r}_{pjk} \\ &= \alpha_{p0k} + \Sigma\alpha_{pQk}W_{qjk} + \mathbf{r}_{pjk} \end{aligned}$$

where

α_{pqk} (q = 0,1,....Q_p) are level–2 coefficients,

W_{qjk} is a level–2 predictor, and $\mathbf{r}_{pjk}$ is a level–2 random effect.

We assume that, for each unit j, the vector ($\mathbf{r}_{0jk}, \mathbf{r}_{1jk},..., \mathbf{r}_{Pjk}$) is distributed as multivariate normal where each element has a mean of zero and the variance of $\mathbf{r}_{pjk}$ is:

Level–3 model

Each of the level–2 coefficients, α_{pqk}, defined in the level–2 model, becomes an outcome variable in the level–3 model:

$$\begin{aligned} \alpha_{pqk} &= \gamma_{pq0} + \gamma_{pql}V_{1k} + \gamma_{pq2}W_{2jk} + ... + \gamma_{pqS}V_{Sk} + \mathbf{u}_{pqk} \\ &= \gamma_{pq0} + \Sigma\gamma_{pqs}V_{sk} + \mathbf{u}_{pqk} \end{aligned}$$

where

γ_{pqs} (s = 0,1,....S_{pq}) are level–3 coefficients,

V_{sk} is a level–3 predictor, and $\mathbf{u}_{pqk}$ is a level–3 random effect.

We assume that, for each level–3 unit, the vector of level–3 random effects (the $\mathbf{u}_{pqk}$ terms) is distributed as multivariate normal, with each having a mean of zero and with covariance matrix $\mathbf{T}_{\alpha}$.