Using hierarchically linear models to analyze multilevel data

Kreft, Ita G. G.

Using Hierarchically Linear Models to Analyze Multilevel Data

Von Ita G.G. Kreft

Vom 2. bis 5. Juli dieses Jahres fand bei ZUMA der Workshop "The Analysis of Hierarchically Nested Data" statt, der von Dr. Ita G.G. Kreft, California State University, Los Angeles durchgeführt wurde. Mit dieser Veranstaltung wurde die Reihe der ZUMA-Workshops fortgesetzt, die in den letzten Jahren Themen der Mehrebenenanalyse zum Gegenstand hatten. Unter Mehrebenenanalyse ist jedes statistische Verfahren zu verstehen, mit dem Beziehungen zwischen Einheiten oder Variablen unterschiedlichen Aggregationsniveaus statistisch überprüft werden kann. Traditionell haben der Gegensatz von Makro- und Mikrosoziologie und die Idee homologer Beziehungen zwischen Daten der Makro- und der Mikroebene die Ansätze der Mehrebenenanalyse dominiert. Die Grenzen dieser Vorstellung sind in der Literatur in einer Fülle von Beispielen zum sogenannten ökologischen Fehlschluß demonstriert worden, wonach die Verwendung von Aggregatdaten zur Ableitung individueller Beziehungen teilweise extrem irreführend sein kann. Es sind allerdings auch Modellansätze bekannt, in denen umgekehrt die Verwendung von Aggregatdaten bei der Parameterschätzung von Mikromodellen gegenüber einer Schätzung mit Hilfe von Mikrodaten überlegen ist. Abseits von diesem Mikro-Makro-Puzzle sind in der Mehrebenanalyse in den letzten Jahren statistische Modelle und die dazugehörige Software entwickelt worden, in denen der gemeinsame Einfluß von Mikro- und Makrovariablen auf abhängige Mikrovariablen statistisch stringent formuliert werden kann. Die ersten Modelle dieser Art sind auch unter dem Namen "Kontextmodelle" bekannt geworden. Der methodische Fortschritt gegenüber früheren Ansätzen besteht nun darin, daß nicht nur die fixen Effekte von Einflußgrößen der Makroebene modelliert werden, sondern darüber hinaus auch zufällige Makroeffekte zugelassen sind. Mit anderen Worten: Die üblichen individuellen Fehlerausdrücke der linearen Modelle als Substitut für die unsystematischen zufälligen Einflüsse ungemessener Variablen werden in einer spezifischen Weise um analoge Fehlerterme der Makroebene erweitert; man gelangt damit zu speziellen Varianzkomponentenmodellen. Der nachfolgende Artikel von Ita G.G. Kreft gibt eine Einführung in Spezifikation und Anwendung dieses Modelltyps. *(Der Herausgeber)*

1. Why new techniques? An example from educational research

Evaluating the effectiveness of large-scale experiments, for instance in education, involves the analysis of *hierarchical data structures*. Educational data are often hierarchical because pupils are in schools, schools are in districts, districts are in counties, and counties are in states. In a large-scala research project we usually have information about two or more of the levels involved, for instance: variables describing individuals (such as intelligence, school career, and family background), variables describing the schools (school type, schools in a special program, curricula offered), and perhaps variables describing districts or countries (available resources). It is well known that analysis of these variables on any of these levels separately can be seriously misleading, (for an overview see Burstein 1980, and Kreft 1987). More satisfactory would be to construct models and techniques which simultaneously take information of all levels into account. But in order to be able to do this, some serious statistical problems have to be solved, like problems in hardware and software that were unsolvable until recently. In the last few years however, a number of papers in the statistical and methodological literature have directly attacked the problem of analyzing variables measured at different levels of a hierarchy. (See Mason/Wong/Entwistle 1985; Aitkin/Longford 1986; Goldstein 1986; Raudenbush/Bryk 1986; and De Leeuw/Kreft 1986). These investigators work with basically the same model known as the *hierarchical linear model*, the *random coefficient model*, or the *Bayesian linear model*. All models deal with the problem of analyzing nested data collected under non-experimental conditions. These nested data have the same type of structure as the above educational example. Multilevel data analysis techniques are available in several software packages. We refer to Mason/Wong/Entwisle (1985 program GENMOD), Aitkin/Longford (1986, program VARCL), Goldstein (1986, program ML3) and Raudenbush/Bryk (1986, program HLM). The packages treat the data the way they are collected; at two or more levels.

2. Some uses of hierarchically nested data in research

Clustered samples are very common in all types of research, especially in education and sociology. Observations are sampled or observed within certain groups, which may be students within schools, employees within industries, or labor strikes within regions.

Two of the most prominent hierarchical data sets in education are the High School and Beyond (HSB) data set, and the Second International Mathematics Study (SIMS) data set. The HSB data were first used by Coleman/Hoffer/Kilgore (1982). It exists of a random sample of high schools in the USA, which are sampled from within the two sectors: The public and the private. Many student level variables are included, as well as teacher level and school level variables. The SIMS data are collected by the International Association for the Evaluation of Educational Achievement. One description and analysis can be found in Burstein/Kim/Delandshere (1988), which exists of random samples of classes that are collected from within several countries. And furthermore contains student level variables such as pre- and post mathematics test scores, background variables of the student, and school level and country level variables. This list of variables surveyed in both studies is running into the hundreds, with the emphasis on factors that influence student achievement on any level of the hierarchy.

Large hierarchical data sets, such as the two mentioned above, have several major advantages over single level data sets. Data sets with student level variables only are restricted to relations between students, while data sets with school or class level data only are restricted to relations between types of schools or types of classes. Making cross level inferences from school level analysis to students, or from a student level analysis to schools, is since Robinson's famous article in 1950 not considered to be a valid way to proceed. The danger here is ecological fallacy (see Robinson 1950; Kreft/De Leeuw 1988). Only data sets that contain measurements of both levels can be used in a multilevel data analysis and analyzed at both levels, which allows a testing of cross level interactions as well.

Furthermore, examples of hierarchically nested data can be found in other areas besides education. A well known example is growth curve analysis, where measurements obtained at different time points are the lower level observations, nested within individuals. A hypotheses that can be tested here is, if growth curves are equal for all individuals or differ for different types (e.g. gender or race differences). Another example of hierarchically nested data are vignet studies, which are portraits judged by different people. A typical research question in such a case is to ascertain how subjective the judges are in their judgments, and if these subjective variances can be related to between-judge difference in background and/or personality (see Hox/Kreft/Hermkens 1991). Vignets are the lower level observations, nested within judges. Research that studies interview bias is another example of hierarchically nested data. Interviewers are the context on the higher level, while the interviewees are nested within interviewers. An important research question from this field is: "What is the influence of the interviewer on the answers given by the interviewee?", meaning in technical terms if cross-level interaction effects are present. For instance, does the race (or gender) of the interviewer make a difference? Has the interviewer a different effect on the interviewee if both are of the same gender (or race) compared to a situation where both are of a different gender (or race)?

Multilevel analysis, with data that contains measurements from different levels of the hierarchy, allows researchers to separate the total variance of the dependent variable into two (in two level analysis) or three (in three level analysis) orthogonal variances: the within and the between strata variation. Single level data analysis estimates the total variance, while neglecting any grouping of the data, so that this analysis ignores either the variation at the group level by doing an individual level analysis, or the variation at the individual level by doing a group level analysis. Since both levels may have a significant influence on the dependent variable, it is often important to be able to analyze both levels at the same time. Analyzing both levels separately ignores another important effect, which is the interaction of the individual and the context, namely the cross-level interaction. Only a multilevel analysis can effectively deal with all level effects, while at the same time testing for cross level interactions. The full potential multilevel data can only be used when multilevel data analysis is applied.

The most important advantage of multilevel data sets analyzed in a multilevel way is, that it allows researchers to answer questions, that could not be addressed with the traditional linear models such as multiple regression or analysis of (co)variance. For instance, given the information that 55% of the variation in student achievement can be explained by individual student characteristics such as Intelligence (IQ), ability, SES of parents, gender and aspiration level (Alexander/Cook/MCDill 1978 in the USA, and Van Herpen/Smulders 1980 in the Netherlands), the remaining 45% variation may be explained by outside factors such as "significant others" (see the models used by Sewell/Haller/Portes 1969), peer group influences (see Webb 1982, 1984), school policies like streaming (see Oakes 1985), or school organization (see Coleman et al. 1982). Obviously, student level analysis cannot test the influence of the school or the class characteristics, while a school level analysis cannot test the influence of the individual student charateristics. Analysis of (co)variance can test for group effects, correcting for individual differences, while random effects Anova could be used when the groups studied are a sample of all possible groups (as schools in most educational research are), instead of a fixed number of treatments. But both methods have their shortcomings; for one, they limit the number of groups that can be used in the analysis. More importantly, they do not allow to test what makes some school significantly different from others. What can be tested in these models is if schools differ significantly, but not why they are different. Both traditional techniques (ANCOVA and multiple regression) have limitations in their relation to student learning when individual student-effects have to be separated from school-effects.

The following example shows how the multilevel model can be used to test two contradictory theories. One theory claims that the achievement of students cannot be directly influenced by the school, while another theory states that it can. The latter propostion stimulates research that investigate the effects of different organizations and different teaching styles. The other hypotheses considers the impact of schools or teachers largely illusory, since it claims that the unexplained 45% of the variance can only be due to student's personal attributes, that inhibit or facilitate good use of what the school environment has to offer. In this theory, the school environment is considered as a proxy for the character and motivation of the student (Hauser/Sewell/Alwin 1976, following Sewell/Haller/Portes 1969). In the criticism of school effectiveness research, for instance Hauser (1970), group effects are considered spurious effects and artifacts of inadequately controlled individual effects. Of course the same may be true for the individual effects, as being an artifact of inadequately controlled group effects. For testing opposing claims traditional models are rendered useless. Using a hierarchical linear model allows for a better control over the effects of both levels of observation, the level of the school and the level of the student. Since within each context the same model is fitted, significant variation in outcomes between models over contexts can show that contexts are different. The assumption that we need to make is, that the omitted variables work in the same way within context and can be context specific. Furthermore, we can assume that omitted variables are randomly influencing individuals within contexts, but may have different effects across contexts. The effect of omitted variables is summarized in the error term (as it is in traditional regression models), with the usual assumption that the errors have a normal distribution with a variance and a mean of zero within each context. The model can test both, the within and between variation of, for instance, school climate. This is nicely illustrated in Rowan/Raudenbush/Kang (1991).

A traditional way of analyzing nested data was multiple regression, which ignored the fact that observations are grouped. The observations are treated as if belonging to the same domain, although measured at different levels. For instance teacher, school and student variables are used as predictors of student achievement in a single level path model, and the level of analysis is defined by the dependent variable at the student level. In such instances, where the student is the unit of analysis, school and teacher effects are disaggregated to the student level. As a result, the significance test associated to the coefficients of the disaggregated variables are biased. In any case, the standard errors of the higher level estimates will be too small. In all regression procedures standard errors of the coefficients are calculated based on the number of observations. Although the number of schools may be smaller than the number of individual students, the larger number is used in the calculation of each standard error of the coefficients in the path model. For class- and school-variables the correct number of observations is mostly smaller than the number of students, resulting in standard errors that are under estimated.

3. The intra class correlation

Students in the same class tend to be more alike than students in different classes, due to selection processes and shared history, so that this closeness in space produces a correlation between observations. While in cases where simple random sampling is used, each observation adds new and independent information to the data. In hierarchically clustered samples the observations are sampled from within the same stratum. As a result these observations are not independent of each other, but repeat more or less the same information. The assumption of independence of observations, an assuption of the traditional linear model, is violated. Thus we have to assume that the individual error terms in the multiple regression model, the **e** 's, or error terms, within the same context are no longer uncorrelated. Error terms of a linear model are defined as containing (random) measurement error and the (random) effects of omitted variables. Variables not included in the model are assumed to have a random instead of a systematic effect. In data where observations are clustered within contexts this is a questionable assumption, since observations close in time or space share experiences due to living in the same time period. The unmeasured effects of time period or context are more likely to be specific than random. For example, take two neighborhoods in Los Angeles, one very rich: Bel-air, the other poor: Watts. We can reasonably assume that unmeasured effects of the neighborhood are more specific than random, since the neighborhood probably has a systematic effect on the behavior of the children growing up there. This effect may be in part general and effect all children alike, and in part interactive, affecting different children differently. We assume that children growing up in Watts tend to react more like each other than like the children in Bel-Air. This causes an intra class correlation.

Using fixed effects linear models, that do not specify this intra-class correlation properly, or assumes a random effect of omitted context variables, is introducing bias and unreliable results. Working with incompletely specified models inevitably leads to a loss of efficiency in the estimates. The standard errors and the hypothesis tests lean heavily on distributional assuptions, much more so than point estimators do. In addition, there is more than one source of error (both at the individual and at the group level). The standard significance test in fixed effects models ignores the error at the second level by only using the source of error at the individual level. As a result, the tests based on the standard errors and the explained variance are less reliable here.

4. A model that decomposes the total variance into a within and a between group variance

The analysis model for the analysis of hierarchically nested data is based on the concept of a micro-model, defined separately for each macro-unit. It is a linear model, with individual- and group-level regressors or predictors, and an individual-level dependent variable, with different assumptions than the traditional linear models. The assumptions of the multilevel analysis or random coefficient model are more realistic and based on the way the data are collected, which is a stratified sample of groups and individuals within those groups in non-experimental conditions.

The basic random coefficient model is (random variables are still in bold face),

$$\mathbf{Y}_{ij} = \mathbf{a}_j + \mathbf{b}_j X_{ij} + \mathbf{e}_{ij}. \qquad (1)$$

Index i is again used for individuals and index j for groups. Variable $\mathbf{a}_j$ is the random intercept, $\mathbf{b}_j$ is the random slope, and $\mathbf{e}_{ij}$ is the disturbance term. We assume that $\mathbf{e}_{ij}$ has expectation zero. All $\mathbf{e}_{ij}$ are independent of each other. The variance of $\mathbf{e}_{ij}$ is equal to σ^2. $\mathbf{Y}_{ij}$ is the dependent variable score of an observation i within a context j, while X_{ij} is the individual-level predictor or regressor score of the same observation. The assumptions are: random coefficients (constant and slopes) at group level; the slopes are independent from each other but correlated with the random group level intercepts. Error terms are correlated within groups, which is based on the fact that we do not assume that a group member's development and performance is independent of the experiences and acts of the other members. Slopes and intercepts are correlated only if they belong to the same group, while disturbances of both levels are uncorrelated.

The next step in the modeling process is to specify the properties of the random slopes and intercepts. We first split them up in fixed components and disturbances. These disturbances are on type group level, have expectation zero, as usual, and they are independent of the individual-level disturbances $\mathbf{e}_{ij}$. So

$$\mathbf{a}_j = \gamma_{00} + \mathbf{g}_j \qquad (2)$$

$$\mathbf{b}_j = \gamma_{10} + \mathbf{h}_j. \qquad (3)$$

From (2) and (3) it is clear that the variance in the dependent variable (Y) is divided up in a number of components. A grand mean and other fixed effects

(γ_{00} and γ_{10}), the first level error variance of $\mathbf{e}_{ij}$ and one or more second level error variances of $\mathbf{g}_j$ and $\mathbf{h}_j$, since in terms of variance components both the intercept $\mathbf{a}_j$ and the slope coefficients $\mathbf{b}_j$ are random at the second level. The grand mean effect is γ_{00}, while $\mathbf{g}_j$ (the error term) illustrates the deviance of each context from this overall mean. The same is true for the slope γ_{10} which is the mean slope, of fixed part, while estimate $\mathbf{h}_j$ represents the random part or the error for the different context around this mean slope.

The variances are called variance components, each being a variance in its own right.

Therefore this kind of model is sometimes referred to as variance components models (see Longford 1986) or error components model. The model contains fixed effects ($\gamma_{00} + \gamma_{10}$) and random effects ($\mathbf{g}_j$, $\mathbf{h}_j$ and $\mathbf{e}_{ij}$).

The variance of the disturbance of $\mathbf{e}_{ij}$ is σ^2, that of $\mathbf{g}_j$ is τ^2_j, that of $\mathbf{h}_j$ is γ^2_j, and the covariance of $\mathbf{g}_j$ and $\mathbf{h}_j$ is ρ_j. If we substitute the decompositions of the random intercepts (equation 2 and 3) in equation (1) we obtain

$$\mathbf{Y}_{ij} = \gamma_{00} + \gamma_{10}\mathbf{X}_{ij} + \mathbf{g}_j + \mathbf{h}_j\mathbf{X}_{ij} + \mathbf{e}_{ij}. \quad (4)$$

In above model (4), no macro effects are present, but only the multiple sources of error. The concern with regard to comparing contexts here is, whether they show significant differences in the mean outcomes (the intercepts) and/or in the relation between dependent and independent variables (the slopes). As such, the simple model above can be used as a preliminary analysis to assess if statistical significant differences in context exists, which can later be modeled as functions of one or more context characteristics. If no such differences can be found in this basic mean model, no relationships will show up in a more complete model with macro level variables, since no systematic variability over contexts is present. For instance, if $\mathbf{h}_j$ (the variation at the second level for slopes) is equal to zero, no variation of the slopes between contexts exists in the data. This implies that γ_j and ρ_j are zero as well. Thus

$$\mathbb{E}(\mathbf{Y}_{ij}) = a_j + b_j\mathbf{X}_{ij}, \quad (5)$$

$$\mathbb{COV}(\mathbf{Y}_{ij}, \mathbf{Y}_{kj}) = \tau_j^2 + \delta^{ik}\sigma_j^2, \quad (6)$$

where δ^{ik} is the Kronecker delta. Modeling a variance of the intercept (or slopes) is modeling a correlation among residuals within the same unit. It follows that the correlation between the disturbances of individuals in the same schools, when only a random intercept is present, is equal to $\tau_j^2/(\tau_j^2 +$

σ^2). To define the same correlation when random slopes are also present is more complex.

If g_j turns out to be close to zero as well, we have again the traditional fixed linear model (1). The ML (or GLS) estimates converge to OLS estimates when the ratio between the individual level error variance and the total error variance (individual and between group error variance) converge to one. To put this another way, if all variation in the dependent variable is explained by the within group relations, no variation is left to be explained between groups. This is the case when no significant between group error variance (g_j) exists. The OLS procedure corresponds to a model in which this ratio is assumed to be one. The estimation procedures generalize the one step and two step procedures used in Boyd/Iversen (1979) and put them on a more solid statistical basis. The estimation of the parameters in random coefficient models is complex, especially when the n_j's are not the same in each group. Several methods are used here, such as Generalized Least Squares (GLS), Maximum Likelihood (ML) and weighted Least Squares. Discussion of the difference in this respect in the software packages for analysis of hierarchically nested data are in Kreft et al. 1990.

5. An example

If the available data are generated under experimental conditions, standard statistical methods can be applied. Most observational research however comes from complex situations in real life, and as an illustration we use the example of Kreft/De Leeuw (1988). In this study a contextual analysis is applied, in which the reading test scores of students within elementary school classes is predicted by gender and the percentage of girls in the class. The equation is

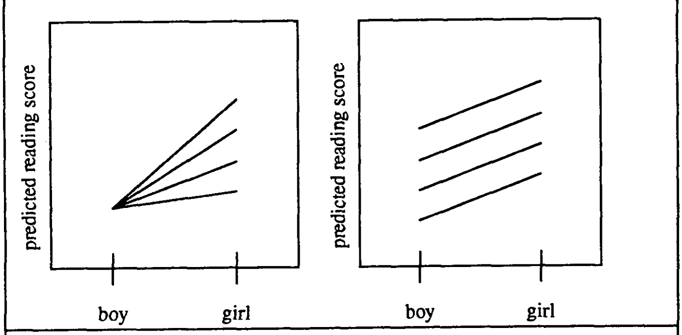
$$\mathbf{Y}_{ij} = \mathbf{a} + \mathbf{b}_1 \mathbf{X}_{ij} + \mathbf{b}_2 \mathbf{X}_{.j} + \mathbf{e}_{ij} . \qquad (7)$$

In this analysis gender is measured at two different levels: the individual gender (X_{ij}) and the class level gender ($X_{.j}$), the percentage of girls in each class. The results of this contextual analysis was surprising, since the gender effects at the two levels have opposite signs, girls are better readers, but classes with a high percentage of girls do worse (see 8). The solution in (8) is given in standardized scores. Gender is scored 1=boy, 2=girl. The positive sign for the coefficient for X_{ij} means that girls are better readers than boys, while the negative sign for the coefficient for $X_{.j}$ implies that the higher the percentage of girls is in a class, the lower the mean reading scores is of that class.

$$Y_{ij} = 0.12\, X_{ij} - 1.46\, X_{.j} + e_{ij}\,. \qquad (8)$$

There are two explanations for above results possible. The first is, that in classes where boys are in the majority girls do better, or are rated higher by their teachers. The other explanation is, that in classes where girls are in the majority, both boys and girls both do worse than in classes where boys are in the majority.

Both explanations imply that different processes are present in different classes depending on the percentage of girls. Is so, than the total analysis above, which ignores grouping of students, does not reflect that fact that each class may have its own best fitting line, as is shown in the two figures below.



Figures 1 and 2: regression lines for various school classes.

Figure 1 shows four lines of four (hypothetical) classes. The lines start at the same place, but the effect of gender is in some classes stronger than in others. The difference between girls and boys in reading scores is larger in the class that produced the top regression line than in the class that produced the bottom line. In Figure 2 we have another situation. All lines are parallel, meaning that the effect of gender on reading scores is in all classes equally strong and in favor of girls. The difference between classes is in mean achievement or intercept, since the lines do not start equally high. In the situation of Figure 2, the mean reading level of classes is not equal. In Figure 1 the slopes of the lines differ, in Figure 2 the intercepts differ.

This may be the situation, but more likely is that intercepts and slopes both differ over the classes. Heterogeneity of relations within classes is largely ignored in the traditional contextual model (8). But heterogeneity can explain why the same effects, but measured at different levels of the hierarchy, can have an opposing nature. In both figures this can be shown. Imagine drawing a line through the means of the classes (equivalent to a school class level analysis); this line will be almost orthogonal to the lines present in the pictures. For both examples in the Figures 1 and 2 a pooled regression will lead to biased estimates, but the direction of the bias cannot be identified a priori; it can go either way.

In a multilevel analysis one can allow to let intercepts and/or slopes be different over contexts in the following way:

$$\mathbb{Y}_{ij} = \mathbb{a}_j + \mathbb{b}_j \mathbb{X}_{ij} + \mathbb{e}_{ij}. \qquad (9)$$

If the intercept is allowed to differ over classes, while adding a second level variable (here the percentage of girls: $\mathbb{X}_{.j}$) to explain this variation, we get (10). This is the situation as pictured in Figure 2.

$$\mathbb{a}_j = \gamma_{00} + \gamma_{01} \mathbb{X}_{.j} + \mathbb{g}_j. \qquad (10)$$

If the slopes are allowed to differ over classes, partly as a function of the percentages of girls in the class ($X_{.j}$), we get (11).

$$\mathbb{b}_j = \gamma_{10} + \gamma_{11} X_{.j} + \mathbb{h}_j. \qquad (11)$$

In (11) the γ_{11} effect is called a cross level interaction. This cross level interaction is shown in Figure 1, where the positive effect of the first level variable "gender" is negatively influenced by the second level variable "the percentages of girls in the class".

Substituting (10) and (11) into (9) results in (12), the equation of a random coefficient model. In (12) we see a multilevel model with two random coefficients, a random intercept (a) and a random slope (b).

$$\mathbb{Y}_{ij} = \gamma_{00} + \gamma_{01} X_{.j} + \mathbb{g}_j + (\gamma_{10} + \gamma_{11} X_{.j} + \mathbb{h}_j) X_{ij} + \mathbb{e}_{ij}. \qquad (12)$$

Multiplying and rearranging fixed and random effects leads to (13).

$$\mathbb{Y}_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} X_{.j} + \gamma_{11} X_{.j} X_{ij} + (\mathbb{e}_{ij} + \mathbb{g}_j + \mathbb{h}_j X_{ij}). \qquad (13)$$

The complicated error term (between brackets) consists of an individual error term (with the within class variance δ^2), and error terms related to the intercept and the slope. The variances of the last two terms (τ^2 and γ_j^2) together are the between variances. Since the slope variance is related to the X variable, the between variance can no longer be defined as a single value. The value will differ with different values of X. For the same reason the intra class correlation, defined earlier at page 51 as $\tau_j^2/(\tau_j^2+\sigma_j^2)$ cannot be calculated either.

Notes

Suggested Readings: For an overview see Burstein 1980, and Kreft 1987. For a collection of technical articles see Bock (Ed.), 1988. See Raudenbush/Willms (Eds.), 1991 for a collection of papers which apply multilevel models in different fields of education. For articles that explain the basic model see the original articles by: Mason/Wong/Entwistle 1985; Aitkin/Longford 1986; Goldstein 1986; Raudenbush/Bryk 1986; and De Leeuw/Kreft 1986. On these articles the four existing software packages are based.

For information on the available software packages for multilevel analysis, contact the following authors:
GENMOD: Mason, University of California, Hilgard Avenue 405, Los Angeles, CA 90024.
VARCL: Longford, Educational Testing Service, Princeton, New Jersey.
ML3: Goldstein, Institute of Education, University of London, WC1N 1AZ, Great Britain.
HLM, Raudenbush, College of Education, Michigan State University, East Lansing, Michigan 48824.

References

Aitkin, M./Longford, N., 1986: "Statistical modeling issues in school effectiveness studies", Journal of the Royal Statistical Society, 149, Part 1: 1-43.

Alexander, K.L./Cook, M./Mcdill, E.L., 1978: "Curriculum tracking and educational stratification: some further evidence." American Sociological Review, 43: 47-66.

Bock, R.D. (Ed.) 1988: Multilevel analysis of educational data. San Diego: Academic Press.

Boyd, L.H./Iversen, G.R., 1979: Contextual analysis: concepts and statistical techniques. Belmont, CA: Wadsworth.

Burstein, L., 1980: The analysis of multilevel data in educational research in evaluation. Review of Research in Education, 8: 158-233.

Burstein, L./Kim, K.S./Delandshere, G., 1988: Multilevel investigations of systematically varying slopes: Issues, Alternatives and Consequences. In: Bock, R.D. (Ed.) Multilevel analysis of educational data. San Diego, Academic Press.

Cochran, W.G., 1977: Sampling Techniques (third edition). New York, Wiley and Sons.

Coleman, J.S./Hoffer, Th./Kilgore, S., 1982: High school achievement: public, catholic and private schools compared. New York: Basis books Inc.

De Leeuw, J./Kreft I.G.G., 1986: Random coefficient models. Journal of Educational Statistics, 11, 1: 55-85.

Goldstein, H., 1986: Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika, 73: 43-56.

Hauser, R.M./Sewell, W.H./Alwin, D.F., 1976: "High school effects on achievement." In: W.H. Sewell/R.M. Hauser/D.L. Featherman, (Eds). Schooling and achievement in American Society. New York: Academic Press.

Hox, J.J./Kreft, I.G.G./Hermkes, P.L.J., 1991: Multilevel models for factorial surveys. Sociological Methods and Research, 19, 4: 493-511.

Kreft, I.G.G., 1987: Methods and Models for the Measurement of School effects. Dissertation, University of Amsterdam.

Kreft, I.G.G./De Leeuw, E.D., 1988: The See-Saw effect: A multilevel problem? Quality and Quantity, 22: 127-137.

Kreft, I.G.G./De Leeuw, J./Kyung-Sung K., 1990: Comparing four different Statistical Packages for Hierarchical Linear Regression; GENMOD, HLM, ML2 and VARCL. Center for the Study of Evaluation. CSE Technical Report No. 11.

Longford, N.T., 1986: Statistical modeling of data from hierarchical structures using variance component analysis. Program developed at the center for applied statistics, Lancaster University.

Mason, W.M./Wong, G.Y./Entwisle, B., 1984: Contextual analysis through the multilevel linear model. Sociological Methodology, 72-103.

Oakes, J. 1985: How schools structure inequality. New Haven, CT: Yalen University Press.

Raudenbush, St.W./Bryk, A.S., 1986: A hierarchical model for studying school effects. Sociology of Education, 59: 1-17.

Raudenbush, St.W./J. Douglas Willms (Eds.) 1991: Schools, Classrooms and Pupils. International Studies of Schooling from a Multilevel Perspective. San Diego, Academic Press.

Robinson, W.S., 1950: "Ecological correlations and the behavior of individuals." American Sociological Review 15: 351-357.

Rowan, B./Raudenbush, St.W./Kang, K., 1991: School climate in secondary schools. In: Raudenbush and Willms (Eds.). Schools, Classrooms and Pupils. International Studies of Schooling from a Multilevel Perspective. San Diego, Academic Press.

Van Herpen, L.W./Smulders, R.H.M., 1980: The relation between the professional status of the parents and success in high school of their children. (Dutch: Sociale beroepsgroep en schoolkeuze). CBS Select 1: 117-133.

Webb, N.M., 1982: Group composition, group interaction and achievement in cooperative small groups. Journal of Educational Psychology, 74: 475-484.

Webb, N.M., 1984: Stability of Small Group Interaction and Achievement over Time. Journal of Educational Psychology, 76, 2: 211-224.