CHAPTER 2

2 INITIAL CONCEPTS OF HIERARCHICAL DATA STRUCTURE

In this Chapter we introduce all the basic concepts of Hierarchical Data Structure analysis, the research areas where hierarchical structure is met, as well as all the possible statistical approaches to analyze hierarchical data structures in these various research areas.

The scope of this chapter is, starting from simple to most sophisticated models, to describe the limitations of the simplest models for analyzing hierarchical data and, therefore, to justify the need of using Multilevel Models for an effective analysis.

2.1 Types of Variables in Hierarchical Data Structure

In traditional data analysis techniques, where all data are measured at the same level, we simply refer to the “response” (dependent) variable, which is the variable to be examined, by using a number of “explanatory” (dependent) variables. In Multilevel Research, however, the situation is quite more complex, since variables can be defined at any level of hierarchy. In order to make clear to which level the measurements properly belong, we present a brief description of each type of variables (based on Hox, 1995) as well as a table with the relations between different types of variables, defined at different levels.

Some of the variables may be measured directly at their natural level; for example, at the school level we may measure school size and denomination, and at the students’ level intelligence and school success. In addition, we may move variables from one level to another by aggregation or disaggregation. Aggregation means that the variables at the lower level are moved to a higher level, for instance by computing the school mean of the students’ intelligence scores. Disaggregation means moving variables to a lower level, for instance by assigning to all pupils a variable that reflects the denomination of the school they belong to.

Table 2.1: Types of variables in Hierarchical Structure Data Models

Level:	1		2		3	etc
Variable Type:	Absolute	⇒	Analytical			
	Relational	⇒	Structural			
	Contextual	⇐	Global	⇒	Analytical	
			Relational	⇒	Structural	
			Contextual	⇐	Global	⇒
					Relational	⇒
					contextual	⇐

As shown in Table 2.1, at each level we have several types of variables, however some of them are related to each other. *Global* and *Absolute* variables refer only to the level at which they are defined, without reference to any other units or levels ('absolute variables' is simply the term used for global variables defined at the lowest level). A pupil's intelligence would be a global or absolute variable. *Relational* variables also refer to one single level; they describe the relationships of a unit to the other units at the same level. Many sociometric indices are relational variables. *Analytical* and *Structural* variables are measured by referring to the subunits at a lower level. Analytical variables refer to the distribution of an absolute or a global variable at a lower level, for instance to the mean of a global variable from a lower level. Structural variables refer to the distribution of relational variables at the lower level; many social network indices are of this type. Constructing an analytical or relational variable from the lower level data involves *aggregation* (indicated by ⇒): data on lower level units are aggregated into data on a smaller number of higher level units. *Contextual* variables, on the other hand, refer to the superunits; all units at the lower level receive the value of a variable for the superunit to which they belong at the higher level. This is called *disaggregation* (indicated by ⇐): data on higher level units are disaggregated into data on a larger number of lower level units. The resulting variable is called a *contextual* variable, because it refers to the higher level context of the units we are investigating.

2.2 Areas with Hierarchical Data Structure

Multilevel Analysis can be applied in situations where the existence of data hierarchies is neither accidental nor ignorable (Goldstein, 1995), in other words the hierarchical data structure is straightforward, or in cases where the applications of multilevel research are not so obvious. This chapter is just an introductory description of some research areas where Multilevel Analysis is commonly used. More detailed examples, when necessary, are presented in Chapter 4 of the thesis, both from educational and from other areas of interest:

Organizational Research: The case where individuals are nested within their organizations. An illustrative example was performed by Kreft et al. (1995), when data were collected on workers in 12 different industries. The hierarchical structure of such model is straightforward, since the workers (first-level units) are nested in industries (second-level groups). The response variable of interest was the income of the workers, measured at the first level of hierarchy and explanatory variables were measured at both levels. Also aggregated measures were used in the analysis. We won't go further to the techniques as well as the results of this particular study since the method used are "overlapped" in following examples.

Clinical Therapy and health research: Clinical psychology is another area where multilevel techniques can be a rational approach, especially in the evaluation of group therapy research. In such studies the patients (first-level objects) are very often gathered in particular therapy groups (second-level objects). In group therapy the type of therapy is an effect under the control of the researcher, but the group dynamics is not. Therapy groups are, at the outset, as much alike as chance can make them by randomly assigning clients to therapy groups, but they change over time. The interactions within each group depend on the dynamics of the group, which develops over time in unpredictable directions.

If the two types of group therapy administered are directive intervention and non-directive interventions, groups within the same treatment can become different, especially under the non-directive intervention treatment. The behaviour of each client starts to reflect the type of therapy as well as the specific dynamics that develops in the client's therapy groups. The interaction between group members makes clients in the same group more alike than clients in different groups. Consequently, the observations of group members can no longer be considered statistically independent.

Application of traditional statistical methods, such as ANCOVA, will fail to analyze correctly this kind of data, in the sense that it will ignore the intra-class correlation that develops over time, leading to an underestimation of the error variance of the estimated coefficients. The group dynamics cannot be modeled in a traditional ANCOVA model, nor can characteristics of the therapist. A multilevel model, on the contrary, will take care of the dependency of observations within groups, and also will model differences between groups by means of macro-level characteristics, such as different approaches by therapists and different group dynamics.

Health economics is another area of health research where multilevel techniques can be applied. We will focus more on this area in Chapter 4 of the thesis.

Twin Studies: A very special case of Clinical Research in which multilevel models can apply is the Twin Studies. This case is indeed special because we may have a large number of groups (2^{nd} level units) but all these groups are of size two. This is an uncommon case in hierarchical or clustered data where a large number of individuals (level-one units) are nested within a small number of common groups. On the other hand, within these small samples of two, we expect to have a high intra-class correlation. Since it is difficult to estimate a model within each group with only two observations, the statistical stability has to come from the number of groups. A Multilevel approach provides the advantage to introduce separate variables for the individuals in the pairs (1^{st} level variables) and variables which the members of the pairs have in common (2^{nd} level variables).

Repeated Measures and Growth curve analysis: Repeated-measures and growth curve analysis is a wide area of research in Statistical Analysis. Although it is not straightforward, longitudinal studies can be considered as an issue where Multilevel Analysis techniques can be performed. However, they form a special case in Multilevel Analysis, since the individuals are the 'macro level' instead of the 'micro level' as in common cases. In other word, here the occasions (measurements) are clustered within individuals that represent the level-two units with measurement occasions the level-one units. Such structures are typically strong hierarchies because there is much more variation between individuals in general, than between occasions within individuals who are, naturally, correlated. Intra-class correlation, in this case measures the degree to which behaviour of the same person is more similar to his/her own previous behaviour in comparison to behaviour of other people (Kreft, 1998). In the case of child growth, for example, once we have adjusted for the overall trend with age, the variance between successive measurements on the same individual is generally no more than 5% of the variation in height between children (Goldstein, 1995).

The major advantage of performing Multilevel Techniques in repeated measures analysis is that they can handily deal with unbalanced data structure with missing values. There are cases in practice where individuals are measured irregularly, some of them a great number of times and some perhaps only once. By considering such data as a general 2-level structure (measurements-the 1st level and individuals-the 2nd level) we can apply the standard set of multilevel modeling techniques while providing statistically efficient parameter estimation and at the same time presenting a simpler conceptual understanding of the data.
Later in the thesis we will focus more on application of multilevel analysis in the area of repeated measures, by presenting a rather special example of autobiographical memories.

Geographical information systems: Spatial Statistics is also an area where multilevel techniques can be easily adapted to analyze the measured data. Such cases are census data, election data, demographical studies and so on. It is obvious that in all these cases sites or individuals are nested within geographic regions, and thus the intra-class correlation comes from spatial autocorrelation. The measured variables can refer to, either geographical characteristics and information about the region (2nd level units), or characteristics of the individuals themselves (1st level units).

An illustrative example was presented by Courgeau and Baccaini (1998). In their work the migration flows of the 19 Norwegian regions is being examined, by using multilevel logit techniques. We will focus on this example with more details later on in the thesis.

Survey Research: Survey research is a wide area where multilevel techniques can be and are already applied in order to examine the, more or less, obvious hierarchical structure of the data measured. This concerns both the methological aspect of sampling design and the practical aspect of data collection through interviews.

The standard literature on surveys, reflected in survey practice, recognizes the importance of taking account of the clustering in complex sample designs. Thus, in a household survey, the first-stage sampling unit will often be a well-defined geographical unit. From those geographical areas, which are randomly chosen, further stages of random selection are carried out until the final households (the final sample) are selected. This is an obvious case of hierarchical data structure where respondents nested in the same geographical area will be more similar to each other than respondents from different areas. However in the analysis insofar the population structure, as it is mirrored in the sampling design, is seen as a 'nuisance factor'. Ignoring hierarchy in the analysis will cause estimates for standard errors that are too small, and 'spurious' significant results – the so-called 'design effect' in survey research. The most usual correction for design effects, taking into account the clustering of individuals within groups, is to compute the standard errors by ordinary analysis methods, estimate the intra-class correlation between respondents within clusters (geographical regions) and to employ a correction formula to the standard errors (Kish,1987).

Although these procedures developed to produce valid statistical inferences can be quite powerful (Skinner et al, 1989), they still do not allow for simultaneous analysis from variables taken from different levels, using a statistical model that includes the various dependencies. In other words, the multilevel modeling approach views the population structure as of potential interest in itself, so that a sample designed to reflect that structure is not merely a matter of saving costs as in traditional survey design, but can be used to collect and analyze data about the higher level units in the population. The subsequent modeling can then incorporate this information and obviate the need to carry out special adjustment procedures, which are built into the analysis model directly.

In chapter 4 we will focus more on an example of survey research taking into account the hierarchy of individuals nested within households nested within geographical areas.

Another aspect of survey research where multilevel techniques are applied is in the so-called 'Interviewer Effect" on the results a study. Whenever the sampling method of the survey is telephone or personal interviews, the interview is carried out at the respondents from a smaller number of interviewers. Even though it is not so obvious, this kind of survey forms a hierarchical structure where respondents (the 1st level units) are "nested" within their interviewers (the second-level of the analysis). It is logical, even though it is not desirable, that in many cases the interviewer characteristics affect the results of the survey, or in other words the respondents being interviewed by the same interviewer tend to have more similar answers. Multilevel analysis can detect such interviewers effect by introducing variables measured in both levels, as well as interactions between interviewers and respondents (cross-level interactions). An illustrative example performed by Hox (1994) will be presented later on in the thesis (Chapter 4).

Meta-Analysis: Meta-analysis or integrative analysis, as it is often called, is a quantitative approach to reviewing the research literature. The term Meta Analysis (Hedges & Olkin, 1985) refers to the pooling of results of separate studies, all of which are concerned with the same research hypothesis. The aim is to achieve greater accuracy than that obtainable from a single study and also to allow the investigation of factors responsible for between-study variation. In other words, the primary goal of meta-analysis is to generalize from a set of studies about a specific substantive issue, by statistically combining quantitative study outcomes from existing research on a particular question. The basic idea is to apply formal statistical methods to the results of a specific set of studies. The statistical approach is the one of the main characteristics that distinguishes meta-analysis from the more traditional narrative literature review (Bangert-Drowns, 1986). Each study typically provides an estimate for an 'effect', for example a group difference, for a 'common' response and the original data are unavailable for analysis. In general, the response measure used will vary, and care is needed in interpreting them as meaning the same thing. Furthermore, the scales of measurement will differ, so that the effect is usually standardized using a suitable within-study estimate of between-unit standard deviation.

Clearly, in a meta-analysis the most important preliminary question is, whether the results differ more from each other than corresponds to the random sampling variation that is expected given the studies' sample size. If the results do not differ more than is expected given the pure sampling error, they are called homogeneous, meaning that they come from a single population. In the next analysis step we would want to estimate the common value of the population parameter of interest. If the results differ more than expected given the pure sampling variation, they are called heterogeneous, meaning that they come from different populations. In this case, estimating the 'average' result is not the primary goal; instead, our goal becomes to analyze the excess variation as a function of the known study characteristics such as the age or sex composition of the sample, or methological characteristics such as the methological quality of the study.

There are various methods to analyze and combine separate study results, one of them, even not so profound, is multilevel analysis. The problem of combining the varying results from different studies has some similarity to the multilevel problem of combining the varying micro-models from different groups or contexts. If we had access to the original data of all the studies, we could analyze them using the hierarchical regression model. But in meta-analysis we generally do not have access to the row data. Still, the statistical problem looks familiar. In multilevel modeling we have a number of regression models computed in different contexts, and we want to estimate the expectation and the variability of the various regression coefficients, and draw conclusions based on all available information. In meta-analysis we have a number of statistics computed in different contexts, and we want to assess their average value and their variability, and again draw conclusions based on all available information. According to Raudenbush & Bryk (1992) meta-analysis may be viewed as a special case of the two-level hierarchical linear model. In each study, a within study model is estimated, and a second level or between study model is added to explain the variation in the within study parameters as a function of differences between the studies. The variability within the studies is considered to be sampling variability, which is known if the relevant sampling distribution and sample size are known. The variability between the studies reflects both sampling variance and systematic differences between the results of different studies. If the study level variance is significant, the studies' results are assumed to be heterogeneous, meaning that there are indeed systematic differences between the studies. If the study level variance is not significant, they are assumed to be homogeneous, meaning that the apparent differences between the studies are just sampling variance.

School Effectiveness Research: Schooling and more generally educational systems is an area where the hierarchical data structure is profound, since students (the 1st level units) are nested or clustered within schools/universities (the 2nd level units), which themselves may be clustered within education authorities or boards or geographical regions (the 3rd level units) and so on. In special cases, students form the 2nd level of hierarchy when repeated measures (the 1st level) are performed within the same student.

In one or another way, school and other institutions effectiveness has been widely examined by educational researchers who have been interested in comparing schools and other educational institutions, most often in terms of the achievements of their pupils. In other examples, the teachers effectives on students achievement is of primary interest, see Bennet (1976) and the analysis of different teaching styles, which was reconsidered by Aitkin et al. (1981) using more advanced statistical methodology. In other cases we are interested in studying the extent to which schools differ for different kinds of students, for example to see whether the variation between schools is greater for initially high scoring students than for initially low scoring students (Goldstein et al, 1993) and whether some factors are better at accounting for or 'explaining' the variation for the former students than for the latter. Moreover, there is often considerable interest in the relative ranking of individual schools, using the performances of their students after adjusting for intake achievements. The response variable in most of the cases discussed above is the students' performance, which is measured, for example, by an examination test.

Most of the traditional approaches for the analysis of such data have been carried out by researchers, such as regression analysis, sometimes by fitting a separate regression line within each group, or ANCOVA models that treat schools as fixed factor. All of them however, either ignore the hierarchy of the data structure at all, or fail to correct for the intra-class correlation within each group. Multilevel techniques stand for a straightforward solution, since multilevel analysis models introduce variables in all levels of hierarchy simultaneously, as well as interactions of the characteristics between levels. They can answer, therefore, to all the theoretical questions as stated in the previous paragraph, and, at the same time obtain statistically efficient estimates of regression coefficients, correct standard errors, confidence intervals and significance tests.

In Greece, although the school effectiveness and students' performance issue has been discussed widely, statistical approaches are very poor in literature and, when they exist, they are mainly constrained in descriptive presentation of the results (Centre of Development of Educational Policy, General Confederation of

Greek Workers (GSEE) 2009). Marouga (2004) used more sophisticated factor analysis and cluster analysis techniques in order to examine students' performance and preferences according to the Greek National Exams. Moreover, Kosmopoulou's dissertation (1998) is the only statistical project where multilevel techniques were performed in Greek educational data in order to assess school effectiveness and students' performance in the National Exams of 1990 and 1991. However, both projects referred before analyze educational data from a national educational system of student's Access in the National Universities and Technical Institutions that does not longer exist.

A more thorough examination of multilevel techniques approach in educational data and school effectiveness research will be performed later on in the thesis, by reviewing, presenting or applying representative examples.

2.3 Possible Approaches for Hierarchical Data Structure - Traditional Models to Random Coefficient Models

In this Chapter we present a number of variations on the ordinary linear model and on OLS regression that have been suggested to deal with hierarchically nested data. They vary from total or pooled regression, which completely ignores the between-group variation, to aggregate regression, which completely ignores the within-group variation. And, on another dimension, they vary from separate regressions for each group, with separate sets of regression parameters, to a single regression with only one set of parameters.

In many cases, however, it makes sense to take the group structure into account more explicitly. Forms of regression analysis, in which both individual and group level variables are used, are known as contextual analyses. In contextual analysis group membership is not neglected. The units of observation are treated as members of certain groups, because the research interest is in individuals as well as in their contexts. Traditional contextual models, the Cronbach model, the ANCOVA model and the various multilevel models decompose the variation in the data into a within and a between part, but each in their one way.

2.3.1 Models and formulae

In contextual analysis techniques the free parameters of the linear model are estimated based on the following model, where $\underline{y}$ is the response variable, x the explanatory variable at the individual level and z is the explanatory variable at the context level. The subscript i is for individual, and j is for context. The model is:

$$\underline{y}_{ij} = a_j + bx_{ij} + cz_j + \underline{\varepsilon}_{ij}. \qquad (2.1)$$

The $\underline{\varepsilon}_{ij}$ are disturbances, which are centered, homoscedastic and independent. This means they have expectation zero and constant variance σ^2. Generally, of course, there may be more than one explanatory variable on both levels. We will discuss more on model structure in the next Chapter by using even more appropriate formulas, since in this Chapter the main interest is to explain the differences between the various models.

Model (2.1) can be expressed in a slightly different way, that more clearly shows its structure. We write:

$$\underline{y}_{ij} = a_j + bx_{ij} + \underline{\varepsilon}_{ij} \qquad (2.2a)$$

$$a_j = a + cz_j \qquad (2.2b)$$

Equation (2.2b) shows that the contextual models of equation (2.1) are varying intercept models, i.e. regression models for each group which are linked because they have the same slope b and the same error variance σ^2. They differ, however, in their intercepts. The different contextual models we discuss in this chapter specify the relationship between the varying intercepts and the group-level variables in different ways.

In hierarchically nested data with two levels the variances and covariances of the observed variables can be divided into a between-group and a within-group matrix. This distinction of between and within variation of variables is not straightforward and differs from technique to technique. To explain the definition of regression coefficients in different models we make use of the notion of the correlation ratio. The correlation ratio is the percentage group variance of a variable, which can be explained as follows. Variables such as x as an explanatory variable and y as response variable can be divided into a between- and a within-group part. This induces a corresponding decomposition of the variances as

$$V_T(x) = V_B(x) + V_W(x) \qquad (2.3a)$$

and equally

$$V_T(y) = V_B(y) + V_W(y) \qquad (2.3b)$$

where the indices T, B and W denote total variance, between-group variance and within-group variance, respectively. Moreover, the total covariance between variables x and y can be divided in the same way into a within and a between part,

$$C_T(x,y) = C_B(x,y) + C_W(x,y) \qquad (2.3c)$$

where C denotes covariance.

The coefficients for regressions over the total sample b_T, between groups b_B and within groups b_W, can be defined by the variances within or between groups, compared to the total variance, as follows:

$$b_T = \frac{C_T(x,y)}{V_T(x)}, \qquad (2.4a)$$

$$b_B = \frac{C_B(x,y)}{V_B(x)}, \qquad (2.4b)$$

$$b_W = \frac{C_W(x,y)}{V_W(x)}, \qquad (2.4c)$$

These coefficients can be related to the correlation ratio η^2, defined for x and y in the following way:

$$\eta^2(x) = \frac{V_B(x)}{V_T(x)}, \qquad (2.5a)$$

$$\eta^2(y) = \frac{V_B(y)}{V_T(y)}. \qquad (2.5b)$$

The equations show group variation in the response variable as the percentage of the total variance in y declared between groups. This is at the same time the definition of the intra-class correlation. We will return to this measure later on in the thesis, since it is of great importance in Multilevel Analysis.

Also,

$$1 - \eta^2(x) = \frac{V_W(x)}{V_T(x)}, \qquad (2.6a)$$

$$1 - \eta^2(y) = \frac{V_W(y)}{V_T(y)}, \qquad (2.6b)$$

The proportion of variance within groups is equal to $1-\eta^2(x)$, and equal to the ratio of the within variance and the total variance.

We know from classical regression theory that the "best" estimate of b for the regression over the total sample, irrespective of group membership, is b_T. It can be shown that the estimate of b_T is a weighted composite of the between-group regression b_B and the within-group regression b_W, as we can see in the following equation:

$$b_T = \eta^2(x)b_B + (1-\eta^2(x))b_W \,. \qquad (2.7)$$

2.3.2 Total or Pooled regression

The first technique we discuss is a simple one. It is not a multilevel analysis, and in most cases not even a contextual analysis. We analyze the effect of the explanatory variable of the individual level on the response variable in a single regression for the total sample. No context variable is used; the fact that some individuals are in the same group and others are in different is not reflected in the model.

Executing a regression analysis over the total sample of individuals, ignoring group membership, is the same as ignoring the subscript j in equation (2.1). The model becomes

$$\underline{y}_{ij} = a + bx_{ij} + \underline{\varepsilon}_{ij} \,, \qquad (2.8)$$

where the $\underline{\varepsilon}_{ij}$ are independent, with mean zero and constant variance σ^2. For completeness, and for later comparisons, we also fit the corresponding null model, with only the intercept a and no explanatory variable. This null model is:

$$\underline{y}_{ij} = a + \underline{\varepsilon}_{ij} \,. \qquad (2.9)$$

A regression analyzing individual observations over the total group is called a total regression. The individual is the unit of analysis, the unit of sampling and the unit of decision-making. Using this analysis means that no systematic influence of groups on the response variable is expected, and all influences of the groups are incorporated in the error term of the model. The fact that the observations are nested within groups is disregarded, and assumed to be of no importance for the research question. In terms of the contextual model (2.2a), in the total regression the intercepts a_j are assumed to be equal for all groups j.

2.3.3 Aggregate regression

One rather crude way to take the grouping of the individuals into account is to do a regression over the group means, a so-called aggregated analysis. There is a priori no real reason to expect that regression coefficients from a total regression analysis and those from an aggregate regression analysis will be similar. In fact, it is easy to construct examples in which the differences between the two techniques will be very large.

For the analysis we form the means for the explanatory variable $x_{\bullet j}$, the means for the response variable $y_{\bullet j}$, and we fit the model

$$\underline{y}_{\bullet j} = a + bx_{\bullet j} + \underline{\varepsilon}_j, \quad (2.10)$$

where the bullet replaces the index for individuals i to indicate that the x and y are summed over individuals. As usual, it is assumed that $\underline{\varepsilon}_j$ has a mean of zero. The variance of $\underline{\varepsilon}_j$ is now, compared to the total model, $n_j^{-1}\sigma^2$, because it is a mean of n_j disturbances, each with variance σ^2. In this analysis we fit a weighted regression, with weights equal to n_j. The regression is heteroscedastic.

Clearly aggregate regression ignores all within-groups variation, and thus throws away a large amount of possibly important variance. At the same time, the standard errors of the regression coefficients normally become much larger, because they are based on only n_j observations. Aggregate regression equations must be interpreted carefully. From the prediction point of view, we can state predictions and merely draw conclusions for individuals, and actually making such statements on the basis of aggregated results. This is known as the "ecological fallacy" (Robinson, 1950).

2.3.4 The contextual model

The contextual model has been used widely in the past in research interested in the effect of group membership on individual behaviour. Typically in this type of analysis the group mean of an individual-level variable is used as a contextual variable. Together with the individual level characteristic x_{ij}, a characteristic of groups is defined as the average value of the groups' members $x_{\bullet j}$. The same measurement is used twice in the same regression, once as the original individual measurement, and once as the mean for each group. In other words, the characteristic is aggregated from individual to group level. The model is thus written as follows:

$$\underline{y}_{ij} = a_j + bx_{ij} + \underline{\varepsilon}_{ij} \quad (2.11a)$$

$$a_j = a + cx_{\bullet j} \quad (2.11b)$$

Substitution gives us the following equation for the contextual model:

$$\underline{y}_{ij} = a + bx_{ij} + cx_{\bullet j} + \underline{\varepsilon}_{ij} . \quad (2.12)$$

It turns out that the best estimate of b in equation (2.12) is b_W, while the best estimate of c is $b_B - b_W$. It can be shown (Duncan et al., 1966) that the within regression (b_W) is confounded with the between regression (b_B) in the estimation of the context effect.

Some more technical problems are present in this contextual model, one related to multicollinearity and one to the level of analysis. Multicollinearity is introduced in this analysis by the correlation of the individual variable and the group mean for this variable. The level of analysis is the individual, because the response variable is defined at the individual level. Performing a regression analysis at one level ignores the true hierarchically nested structure of the data, and treats the aggregated variable as if it was still measured at the first level. The contextual effect in this contextual model is merely the difference between b_B and b_W. It is clear that the individual and group effects are confounded in c, and as a result interesting and significant relationships can be distorted by this procedure.

2.3.5 The Cronbach model

The Cronbach Model (Cronbach & Webb, 1975) provides a clearer picture of the individual effect together with the group mean effect on the response variable. The individual variables are first centered around their respective group means, as in the following equation:

$$\underline{y}_{ij} = a + b_1(x_{ij} - x_{\bullet j}) + b_2(x_{\bullet j} - x_{\bullet\bullet}) + \underline{\varepsilon}_{ij}. \quad (2.13)$$

In equation (2.13) the centered individual scores $x_{ij} - x_{\bullet j}$ form a variable that is orthogonal to the variable formed by the centered group-level scores $x_{\bullet j} - x_{\bullet\bullet}$. Raw scores are thus transformed into deviation scores from the group mean. Centering explanatory variables in this model provides a convenient way of avoiding the problem of correlation between the two variables that are measurements for a characteristic at the two different levels. The two predictors in the Cronbach model are a centered individual characteristic and the centered group mean for this characteristic analyzed again with regression. Because the two predictors are orthogonal, the best estimate of b_1 is equal to b_W, and thus also to the estimate in the contextual model discussed previously. The difference compared to the contextual model is in the estimate for the contextual effect, where b_2 is now equal to b_B and thus equal to the effect of b_B in the aggregate model. Within and between effects are no longer confounded in the Cronbach model.

Although the collinearity problem of the correlation between the individual variable and its aggregated counterpart is solved in the Cronbach model, the significance tests are just as suspect as they are in the contextual model. In both contextual models discussed so far, the analysis is executed at the lower level. As a result the standard error for the coefficient of the group mean is underestimated. The result is an increase in the alpha level of the test of significance. The group mean has only as many independent observations as the number of groups. Since we have say k groups with nk observations each, the total number of observations on which the standard error is based is $k \times n_k$, instead of the correct number k. Another threat to the validity of the standard errors in the above contextual model is intra-class correlation in the sense that when intra-class correlation is present, the alpha level enhances.

2.3.6 Analysis of Covariance (ANCOVA)

Analysis of covariance is another traditional way of analyzing group data. Both levels are included in the model but not in equal roles. Individual-level explanatory variables are involved, as in regression models, but at the same time groups are allowed to differ in the intercepts. The ANCOVA model incorporates both quantitative and qualitative variables and therefore has a mixed character. It is a regression model, with dummy variables to code group membership. While the regression model enables us to access the effect of quantitative factors, ANCOVA enables us to model qualitative factors.

ANCOVA is a technique with a somewhat different purpose from contextual analyses. It evaluates the effect of groups, correcting for pre-existing differences among these groups. With this technique we can study if the groups are equal in the response variable, corrected for the differences in the amount of a first- level variable. Such an analysis would tell us if groups differ in average response, and which group "scores", on average, the best. In ANCOVA the individual effects are neglected, or considered as noise, and the emphasis is on the group effect.

The individual variable(s) functions as covariate (s), while the grouping is used as the important factor in the design. Because the model was originally developed for designed experiments, groups in ANCOVA are considered to be different treatment categories. The equation for the analysis of Covariance is

$$\underline{y}_{ij} = a_j + bx_{ij} + \underline{\varepsilon}_{ij}. \quad (2.14)$$

Different values for a_j mean that some groups have higher "starting values" for the response variable than others. The assumption in ANCOVA, that all groups have the same slope (the b in the model), means that we assume that the relation between the first-level explanatory variable and response variable is the same for all groups. We can see that equation (2.14) is the same as (2.2a) and that (2.2b) is missing. There is no additional structure imposed on the a_j; they can take all possible values.

Since ANCOVA expresses the differences between k groups using all $k-1$ degrees of freedom, the model provides an upper limit on the amount of variance potentially attributable to overall differences in contexts. In contrast to the traditional contextual model in equation (2.1), ANCOVA cannot tell us which characteristics of the context explain the differences between them. The only thing it shows is how large the overall group effect is, by giving a measure of the explained between-group variation of the intercepts.

The chief advantage of ANCOVA is that it has greater predictive power than the traditional contextual models, as in equations (2.2a) and (2.2b). ANCOVA accounts for all variability between the context means, and not only for variability related to a context-specific explanatory variable, as in contextual models. At the same time the specificity is strength of the contextual model, because it identifies important group characteristics. Most researchers consider the analysis of (co)variance useful as an estimate of the composite group effect preliminary to contextual analysis. It is true that where the a_j in a covariance analysis adds little explained variance, we know from the outset that none of the context characteristics can explain much additional variance of the response variable in subsequent models. But that is only true if variation among contexts is studied in relation to the intercepts, the main effects. But more and more research is dedicated to studying differences among contexts in relationships between explanatory variables and response variable, the b-coefficients in model (2.14). The assumption of ANCOVA that each of the k explanatory variables, or covariates, has the same relation with the response variable over all groups is unrealistic. Each group may need its own unique solution, and its own unique relation between the response variable and the explanatory variable.

2.3.7 Moving from one single-level to multilevel-model techniques

By now, we have discussed some of the traditional ways of analyzing grouped data that consist of two levels, an individual one and a contextual one. The data analysis in these models is always executed at one single level, which can be either the individual or the context level. Analyses executed at the individual level can still be different in the way they handle the between variation. As a result, different regression estimates for the contextual effect are observed among models. From the discussion of the models and the different results, we see that we are in need of a more general model. We need a model that treats the data at the level they are measured, and can answer research questions about the influence of all explanatory variables on the response variable, irrespective of the level in the hierarchy at which they are measured, or to which they are aggregated. Such models are multilevel models, i.e. the Varying Coefficients Models and their modern version, the Random Coefficient (RC) models. Varying coefficient models are also known as the "slopes-as-outcomes" approach.

Varying coefficients or "slopes-as-outcomes"

Traditional strategies for analyzing group data are several forms of regression analysis. The basic equation defining this linear model is

$$\underline{y}_{ij} = a_j + b_j x_{ij} + \underline{\varepsilon}_{ij}, \qquad (2.15)$$

which is similar to equation (1.2a) discussed previously. In equation (2.15) x is again the individual explanatory variable, and y the response variable. The a_j are intercepts and the b_j are slopes. We use the plural form, since instead of the usual single intercept and single slope, separate ones are estimated for each context. To indicate that fact in the formula the subscript j is added to the coefficients a and b. Thus subscript j refers to contexts and subscript i to individuals. The $\underline{\varepsilon}_{ij}$ is the usual individual error term, with an expectation (mean) of zero and a variance of σ^2. In equation (2.15) only $\underline{\varepsilon}_{ij}$ and $\underline{y}_{ij}$ are random variables. Later, when we move away from varying to random coefficient models, the a_j and b_j will be underlined too, implying that a_j and b_j are also random variables.

Within the traditional fixed effects linear framework the 'slopes-as-outcomes' approach can be considered a multilevel analysis approach. This approach is the first step toward modern multilevel modeling. A linear model with individual-level explanatory variables and an individual-level response variable estimates separate parameters within each group, allowing each context to have its own micro model. Three hypothetical situations of varying coefficients analysis are used for comparison:

- A situation with varying intercepts only;
- A situation with varying slopes only; and
- A situation with varying intercepts and varying slopes.

In the first case, the groups' regression lines are parallel, meaning that the slope of the regression of y on x is equal for each group. But the lines start at different points, showing that the overall mean level for y is different from context to context. Unequal intercepts mean that some groups "score" better than others on the response variable, after the amount of the explanatory variable is taken into account. This situation resembles an ANCOVA solution already discussed previously, where unequal intercepts but equal relationships (or parallel lines) between x and y are assumed.

The second case represents a situation where all regression lines for groups start at the same point, thus having the same intercept. But the regression of y on x

is stronger in some groups, resulting in different slopes. The steeper the slope, the stronger the relationship between x and y.

The third case exhibits a more realistic situation where both intercepts and slopes in the regression model differ. This is an example of where the 'slopes-as-outcomes' approach is most valuable. Each group is allowed to have its own unique solution, which may be a more realistic situation than forcing groups to have some or all features in common.

All three situations show that different intercepts and/or slopes are estimated for each context, representing the first step in the slopes-as-outcomes approach. In subsequent steps parameter estimates for intercepts and slopes are used as response variables in macro-level regressions together with macro-level explanatory variables.

Another name sometimes used for this type of analysis is 'two-step-analysis', because in a first step the individual, or micro-level, parameters are estimated within each context and used in a second step as response variables, predicted by macro-level variable(s). In both steps Ordinary Least Squares (OLS) is the estimation method. The following equations show the second step, which is at the macro level

$$a_j = c_0 + c_1 z_j, \qquad (2.16a)$$

$$b_j = d_0 + d_1 z_j, \qquad (2.16b)$$

where a_j and b_j are the regression coefficients for intercept and slope respectively. The number of observations in each step can be different. In the micro analyses of the first step the number of observations varies for each group. In the macro analyses, with either intercepts or slopes as response variable, the number of observations is equal to the number of groups. m groups produce m different slopes b_j and m different intercepts a_j. The macro equations produce macro intercepts and slopes which are c_0 and d_0 and c_1 and d_1 respectively in (2.16a) and (2.16b). The same equations show the group-level variable z is used to explain the variation among intercepts and slopes. z can either be a global variable or an aggregated variable, as were determined in previous chapter.

The 'slopes-as-outcomes' approach is promising and a potentially good way to find interesting features in the data, features that were previously ignored. But the approach has a practical disadvantage; it requires a separate analysis for each context. Separate analyses for each group may be the best way to represent its group in its uniqueness, but with a large number of groups, this method is hardly feasible, not parsimonious and ignores the fact that groups also may have things in common. The model of the ‘slopes-as-outcomes’ approach has also some other drawbacks. First, the error structure is not specified correctly, which makes the p-values for the parameter estimates questionable. Secondly, the regression coefficients obtained in the first step are not equally efficient: some may have large standard errors and some small ones. This is not accounted for in the second step. Each coefficient is weighted equally.

An alternative way is the extension of this approach to RC models, which will be discussed in the next section. This approach combines the conceptually interesting features of the ‘slopes-as-outcomes’ approach with the statistical advantage of parsimony, and the practical advantage of taking into account not only the uniqueness of each group but also what they have in common.

The Random Coefficient (RC) Model

The RC model is conceptually based on the ‘slopes-as-outcomes’ model. One difference between the two models is that the RC model does not estimate coefficients for each context separately, although each context is allowed to differ from the other contexts in intercept, in slope(s), or in both. A single model is estimated from which the groups are allowed to deviate. From a “graphical” point of view, in the RC approach, a single (solid) regression line is calculated with two other (dashed) lines on either side of it. These two lines capture the variation of the groups from the average line, corresponding with the variance in the ‘fixed but varying coefficient’ cases in the previous section.

The three situations discussed in the previous chapter are compared with similar ones referring to RC models, in order to show similarities and differences between models. These three situations are again:

- A situation with varying intercepts only;
- A situation with varying slopes only; and
- A situation with varying intercepts and varying slopes.

In the first case where intercepts vary but slopes are the same, this is reflected in a variance around the regression line, which is regular and equal for all values of x.

In the second case, where intercepts do not vary but slopes do, the space around the regression line is not equal for all values of x. This is to be expected in

RC models, because variation in slopes is related to values of x, the explanatory variable. The higher the value of x, the larger the spread around the mean (regression) line.

Finally, in the third case where both slopes and intercepts are different for the groups, the variation around the regression line is produced by the combination of the variance of the slope, the variance of the intercept and the covariance between the two. The variation in slopes is related to values of x, as is the covariance between the variances of intercept and slope. The total variance around the line is the sum of all three (co)variances. As a result, the pattern of the variation of the groups around the average line is irregular, with a minimum and a maximum at certain values of x. If the variation around the average (regression) line, as indicated by the values for the variances, is large, we say that the single line does not represent all groups equally well. Since the regression line is an 'average', we know by the value of the dispersion or variance of the coefficients that some groups are above the line, while others are below it. If, on the other hand, the variances of the intercept and slope are small, the line is close to equal for all groups. A single-level regression analysis would then represent the relationship in this data equally well. The groups can differ either in intercept, in slope, or in both. In RC models each coefficient has its own variance, allowing groups to be unique. Uniqueness for each context is translated into the extent of the deviation of a group from the overall regression line. This deviation (or error) can be used to calculate the posterior means, which are separate values for intercepts and slope(s) for separate contexts, very similar to the 'slope-as-outcomes' approach.

In the previous paragraphs we have described the principles of RC modeling, as well as the differences between RC models and 'slopes-as-outcomes' models. Next we formalize the same principles in equation form. We have shown that the coefficient estimates for separate contexts are represented as varying around the overall regression line. As a result coefficients in RC models consist of two parts: a mean or fixed part, and a variance or random part. The random part is represented by a macro variance, showing the deviation from the overall solution. This variance is referred to as macro-level variance, because the coefficients differ from each other at the macro or context level. The equation of the random model starts with the familiar regression equation, where we underline random variables as before:

$$\underline{y}_{ij} = \underline{a}_j + \underline{b}_j x_{ij} + \underline{\varepsilon}_{ij} . \qquad (2.17)$$

Index i is again used for individuals and index j for groups. $\underline{y}_{ij}$ is the score on the response variable of an observation i within a context j., while x_{ij} is the individual level explanatory variable of the same observation. The variable $\underline{a}_j$ is the random intercept, $\underline{b}_j$ is the random slope, and $\underline{\varepsilon}_{ij}$ is the disturbance term. We assume that $\underline{\varepsilon}_{ij}$ has expectation zero. All $\underline{\varepsilon}_{ij}$ are independent of each other. The variance of $\underline{\varepsilon}_{ij}$ is equal to σ^2.

Note that the underlining of a and b in equation (2.17) is a new feature signifying random coefficients. Observe that this underlying is the only difference between this equation and equation (2.15) for the 'slopes-as-outcomes' model.

The models discussed so far have fixed coefficients. In RC models coefficients can be either fixed or random. The choice between random and fixed coefficients can be made separately for each coefficient in an analysis based on an RC model. Coefficients in RC models are estimated as a main effect with a variance around it. This variance represents the deviation of contexts from that overall or main effect. To specify the properties of the random coefficients, we define them as fixed components plus disturbances. These disturbances are at the group level. They have expectation zero, as usual, and they are independent of the individual-level disturbances $\underline{\varepsilon}_{ij}$. The macro-level equations express the properties of the random slope and intercept in terms of overall population values plus error, as specified in the following macro equations:

$$\underline{a}_j = \gamma_{00} + \underline{u}_{0j}, \qquad (2.18a)$$

$$\underline{b}_j = \gamma_{10} + \underline{u}_{1j}. \qquad (2.18b)$$

The macro-level errors $\underline{u}_{0j}$ and $\underline{u}_{1j}$ in (2.18a) and (2.18b) indicate that both the intercept γ_{00} and slope γ_{10} vary over contexts. The grand mean effect in (2.18a) is γ_{00} while $\underline{u}_{0j}$, the macro-error term, measures the deviation of each context from this overall or grand mean. In the same manner the grand slope estimate across all contexts is γ_{10}, while $\underline{u}_{1j}$ represents the deviation of the slope within each context from the overall slope, as in equation (2.18b). For the gammas the subscript is defined as follows: the fist index is the number of the variable at the micro level, the second represents the number of the variable at the macro level. Hence, γ_{st} is the effect of the macro variable t on the regression coefficient of micro variables s. Zero signifies the intercept, that is to say, the variable with all values equal to +1, either at the micro level or at the macro level. For instance, γ_{00} is the effect of the macro-level intercept on the micro-level coefficient of the intercept. Note that equations (2.18a) and (2.18b) display the model coefficients $\underline{a}_j$ and $\underline{b}_j$ as a function of two components: a fixed component γ_{00} and γ_{10} respectively, and a random component $\underline{u}_{0j}$ and $\underline{u}_{1j}$ respectively, where $\underline{u}_{0j}$ has variance τ_{00}, $\underline{u}_{1j}$ has variance τ_{11}, while $\underline{u}_{0j}$ and $\underline{u}_{1j}$ have covariance τ_{01}.

The elements in the matrix T in equation (2.19) summarize the variance components of an RC model with a random intercept and one random slope and indicate the extra parameters that are estimated in RC models. The τ parameters show the degree to which the groups differ from the overall line.

$$T = \begin{array}{c} \\ \underline{u}_{0j} \\ \underline{u}_{1j} \end{array} \begin{array}{c} \begin{array}{cc} \underline{u}_{0j} & \underline{u}_{1j} \end{array} \\ \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \end{array} \qquad (2.19)$$

To show that the separate equations are not really separate, but part of the model, we substitute the separate equations (2.18a) and (2.18b) into equation (2.17) resulting in

$$\underline{y}_{ij} = (\gamma_{00} + \underline{u}_{0j}) + (\gamma_{10} + \underline{u}_{1j})x_{ij} + \underline{\varepsilon}_{ij}\,. \qquad (2.20)$$

Expanding and rearranging terms yields

$$\underline{y}_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + (\underline{u}_{0j} + \underline{u}_{1j}x_{ij} + \underline{\varepsilon}_{ij}). \qquad (2.21)$$

The rearranging of the terms yields an equation that looks a bit more organized. The fixed effects (gammas) are together and the micro error $\underline{\varepsilon}_{ij}$ and the two macro errors $\underline{u}_{0j}$ and $\underline{u}_{1j}x_{ij}$ are also collected together (in parentheses). The result is a single equation that resembles a traditional regression equation, except for the error terms in parentheses. It is already mentioned that the macro-level variance of the slope (the variance of $\underline{u}_{1j}$) is related to the values of x, as in equation (2.21), the error term, in parentheses, depends on the variable x.

The uniqueness of each context is expressed in these macro errors (the $\underline{u}$ s) which are the deviances from the overall solution. Solutions based on this model no longer produce unique regression lines for each context, such as in the 'slopes-as-outcomes' approach. The result of the RC analysis is a single regression line as an overall solution. Groups fluctuate around this average line. The parameters of the line are the gammas in the above equation, also called the fixed effects. The random effects or macro variances are $\underline{u}_{0j}$ and $\underline{u}_{1j}x_{ij}$. If these variances are significantly different from zero we say that context effects are present.

The next step in the RC approach is to add a new second-level variable to the analysis model, in order to 'explain' the variation in the coefficients for slope and intercept. By adding a macro-level variable z, the variation among groups in general (in the intercepts) or in particular (in the slopes) may disappear. If that works we say that the macro-level variable 'explains' the variation among groups.

As in the 'slopes-as-outcomes' approach, we can choose to model the intercept variance or the slope variance. What we could not do in the 'slopes-as-outcomes' approach was model both variances in the same step. We will show how the model can be extended by fitting macro variances together. All parameters are estimated in a single model, instead of fitting two different macro models as in the 'slopes-as-outcomes' approach.

Out task is to add an explanatory macro-level variable z_j that can account for the explanation of the intercept variance as well as the slope variation among groups. We relate the macro-level variable to the intercept and slope by changing the equations (2.18a) and (2.18b) respectively to:

$$\underline{a}_j = \gamma_{00} + \gamma_{01} z_j + \underline{u}_{0j}, \qquad (2.22a)$$

$$\underline{b}_j = \gamma_{10} + \gamma_{11} z_j + \underline{u}_{1j}. \qquad (2.22b)$$

By fitting this model we assume that intercepts vary as a function of the macro-level explanatory variable z_j plus a random fluctuation, which is represented in the macro-error term $\underline{u}_{0j}$ in (2.22a) and, at the same time, we create and introduce an interaction of the micro-level variable x_{ij} with the macro-level variable z_j. Substituting the new macro equations for the slope (2.22b) and for the intercept (2.22a) into the basic equation (2.17), we produce the single equation:

$$\underline{y}_{ij} = \gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} x_{ij} z_j + (\underline{u}_{0j} + \underline{u}_{1j} x_{ij} + \underline{\varepsilon}_{ij}). \qquad (2.23)$$

The difference between model (2.23) and (2.21) is in the estimation of two more parameter coefficients, γ_{01} and γ_{11}, while the rest stay the same.

We will discuss in more extent about the estimation methods for all the parameters of the RC model (fixed and random) in the next Chapter of the thesis.

2.3.8 Assumptions and Differences for the Linear Models – A Brief Summary

Table 2.2 summarizes the differences between the models discussed in this Chapter, two traditional linear models, regression and ANCOVA, and two multilevel linear models, ‘slopes-as-outcomes’ and random coefficients. Most models in Table 2.2 are fixed effects linear models, while the RC model is the only random effects linear model. Within the fixed models the choice is to allow intercepts to be equal (2.24a) or different (2.24b):

$$a_1 = a_2 = \cdots = a_m, \qquad (2.24a)$$

$$a_1 \neq a_2 \neq \cdots \neq a_m. \qquad (2.24b)$$

Equation (2.24a) applies to the total regression model, where group membership is ignored, and all contexts are assumed to have the same effect on individuals. ANCOVA models assume unequal intercepts over contexts, as in equation (2.24b).
Linear models can also differ in what they assume concerning slope coefficients. Slopes can also be assumed to be equal or unequal over contexts. Equal slopes are assumed in the analysis of variance model, where a pooled within slope is estimated, as in equation (2.25a):

$$b_1 = b_2 = \cdots = b_m, \qquad (2.25a)$$

$$b_1 \neq b_2 \neq \cdots \neq b_m. \qquad (2.25b)$$

Random and varying coefficient models allow slopes to differ, as in equation (2.25b).
RC and ‘slopes-as-outcomes’ model allow researchers to assume that coefficients within contexts vary systematically as a function of the context. Different intercepts together with different slopes can be fitted.

ANCOVA and regression are based on a more restrictive model than the two multilevel models. Multilevel models are more general, because some restrictions are lifted and more parameters are estimated. While more general models allow more freedom than restricted models, they are at the same time less parsimonious.

The equations of RC model presented previously show that this model is an intermediate solution between a totally restricted one, such as standard regression that ignores the context, and a totally unrestricted one, such as the ‘slopes-as-outcomes’ approach that takes the context too literally. In the 'slopes-as-outcomes' approach all contexts are treated as separate entities as if they have nothing in common, while in the total regression approach contexts are treated as if they are the same and interchangeable. The RC model is also statistically in between the two extremes. The RC model estimates fewer fixed parameters than the 'slopes-as-outcomes' approach, but RC models estimate more parameters than are estimated in the total regression model.

Table 2.2: Assumptions of Traditional Linear Models and Multilevel Models

Model	Intercepts	Slopes
Traditional linear regression	Equal	Equal
ANCOVA	Unequal	Equal
'Slopes-as-outcomes'	Unequal	Unequal
Random Coefficients (RC)	Unequal	Either equal or unequal

2.4 Conclusions of the Chapter

As we can conclude from the thorough discussion of the Chapter, despite the attempts to introduce more sophisticated models in order to analyse hierarchical data (Total or Pooled Regression, Aggregated Regression, Contextual Model, Cronbach Model, ANCOVA, "Slopes-as-Outcomes" Model), all these attempts fail to describe effectively the hierarchy of the data. On the other hand, Multilevel Models seem to be more effective and, therefore, the theoretical aspects of these models will be elaborated more in the following Chapter.