CHAPTER

14

Multilevel Regression and Multilevel Structural Equation Modeling

Joop J. Hox

Abstract

Multilevel modeling in general concerns models for relationships between variables defined at different levels of a hierarchical data set, which is often viewed as a multistage sample from a hierarchically structured population. Common applications are individuals within groups, repeated measures within individuals, longitudinal modeling, and cluster randomized trials. This chapter treats the multilevel regression model, which is a direct extension of single-level multiple regression, and multilevel structural equation models, which includes multilevel path and factor analysis. Multilevel analysis was originally intended for continuous normally distributed data. This chapter refers to recent extensions to non-normal data but does not treat these in detail. The end of the chapter presents some statistical issues such as assumptions, sample sizes, and applications to data that are not completely nested.

Key Words: Multilevel model, mixed model, random coefficient, cluster sampling, hierarchical data

Introduction

Social and behavioral research often concerns research problems that investigate the relationships between individuals and the larger context in which they live, such as families, schools, or neighborhoods. Similarly, longitudinal data are becoming more common, where individuals are followed for a period of time to observe and model their development. Multilevel models and software have been introduced to combine in a statistically sound way variables defined at the individual and the group level. These models were discussed in the educational and sociological research literature in the 1980s and described in monographs in the early 90s by, for example, Bryk and Raudenbush (1992) and Goldstein (1987). For an exhaustive review of the older multilevel literature, *see* Hüttner and Van den Eeden (1995). The monographs by Bryk and Raudenbush and by Goldstein are mathematically oriented; more introductory level handbooks appeared later—for example, Bickel (2007), Hox (2002), and Snijders and Bosker (1999).

Although multilevel modeling was initially discussed mostly in the context of individuals within groups, the model was rapidly extended to longitudinal and repeated measures data. The translation is simple—one just needs to replace individuals within groups with measurement occasions within individuals, and restructure the data from the conventional multivariate ("wide") structure to a stacked ("long") multilevel structure. This application was already described by Goldstein (1987). As it turns out, multilevel modeling of longitudinal data is a very powerful approach, because it enables a very flexible treatment of the metric of time, and it deals naturally with incomplete data resulting from incidental dropout and panel attrition. Just as multilevel analysis of individuals within groups does not assume that the group sizes are equal, multilevel analysis of repeated measures within individuals does not assume that all individuals have the same number of measures.

A more recent development is the introduction of multilevel structural equation modeling (SEM). Structural equation models are more flexible than (multilevel) regression models. Regression models assume predictor variables that are perfectly reliable, which is unrealistic. Structural equation models do not make that assumption, because they can include a measurement model for the predictor or outcome variables. In addition, they can model more complicated structures, such as indirect effects in a mediation analysis.

This chapter treats the multilevel regression model as applied to individuals within groups and as applied to measurement occasions within individuals. It follows with a description of (multilevel) SEM for measurement occasions within individual and for mediation analysis. Next, some issues are discussed concerning assumptions and sample sizes. The chapter ends with a brief discussion.

Multilevel Regression Modeling: Introduction and Typical Applications

Individuals Within Groups

The multilevel regression model for individuals within groups is often represented as a series of regression equations. For example, assume that we have data from pupils in classes. On the pupil level, we have an outcome variable, "pupil popularity." We have two explanatory variables on the pupil level, pupil gender (0 = boy, 1 = girl) and pupil extraversion, and one class level explanatory variable teacher experience (in years). There are data on 2,000 pupils in 100 classes, so the average class size is 20 pupils. The data are described and analyzed in more detail in Hox (2010) and available on the web (www.joophox.net).

The lowest level regression equation predicts the outcome variable as follows:

$$popularity_{ij} = \beta_{0j} + \beta_{1j} gender_{ij} + \beta_{2j} extraversion_{ij} + e_{ij}. \quad (1)$$

In this regression equation, β_{0j} is the intercept, β_{1j} is the regression slope for the dichotomous explanatory variable gender, β_{2j} is the regression slope for the continuous explanatory variable extraversion, and e_{ij} is the usual residual error term. The subscript j is for the classes ($j = 1...J$) and the subscript i is for individual pupils ($i = 1...n_j$). The major difference with the usual regression model is that we assume that each class has a different intercept β_{0j}, and different slopes β_{1j} and β_{2j}. This is indicated in the equation by attaching a subscript j to the regression coefficients. The residual errors e_{ij} are assumed to have a normal distribution with a mean of zero and some variance that is estimated. This chapter uses σ_e^2 to denote the variance of the lowest level residual errors.

Because the regression coefficients of the individual-level variables vary across classes, the next step is to explain this variation using explanatory variables at the second or class level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \mathit{Teacher\ Exp}_j + u_{0j}, \quad (2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \mathit{Teacher\ Exp}_j + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21} \mathit{Teacher\ Exp}_j + u_{1j}. \quad (3)$$

Equation 2 predicts the average popularity in a class (the intercept $\beta_{0\mathrm{j}}$) by the teacher's experience. The equations under Equation 3 state that the *relationship* (as expressed by the slope coefficients β_j) between the popularity and the gender and extraversion of the pupil depends on the amount of experience of the teacher. The amount of experience of the teacher acts as a *moderator variable* for the relationship between popularity and gender or extraversion; this relationship varies according to the value of the moderator variable.

The *u*-terms u_{0j}, u_{1j}, and u_{2j} are residual error terms at the class level. These are assumed to have means of 0 and to be independent from the residual errors e_{ij} at the individual (pupil) level. The variance of the residual errors u_{0j} is specified as $\sigma_{u_0}^2$, and the variances of the residual errors u_{1j} and u_{2j} are specified as $\sigma_{u_1}^2$ and $\sigma_{u_2}^2$. The *covariances* between the residual error terms are denoted by σ_{u01}^2, σ_{u02}^2, and σ_{u12}^2 and are generally *not* assumed to be 0.

Using standard multilevel regression software, we can estimate a series of models. Table 14.1 presents three models of increasing complexity. The first model is the intercept-only model, which allows us to calculate the intraclass correlation ρ as $\rho = \frac{\sigma_{u0}^2}{\sigma_{u_0}^2 + \sigma_e^2}$. For the popularity data, the intraclass correlation is 0.36, which is relatively large. Model 2 adds the predictor variables, with a random slope for pupil extraversion (the variance of the slope for pupil gender is 0 and therefore omitted from the model). The last model adds the cross-level interaction to explain the variation of the extraversion

Table 14.1. Models for the Pupil Popularity Data

Model:	Intercept-only	Main effects	With interaction
Fixed part	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	5.08(0.09)	0.74(0.20)	−1.21(0.27)
Pupil gender		1.25(0.04)	1.24(0.04)
Pupil extraversion		0.45(0.02)	0.80(0.04)
Teacher experience		0.09(0.01)	0.23(0.02)
Extra*T.exp			−0.03(0.003)
Random part			
σ_e^2	1.22(0.04)	0.55(0.02)	0.55(0.02)
σ_{u0}^2	0.69(0.11)	1.28(0.28)	0.45(0.16)
σ_{u2}^2		0.03(0.008)	0.005(0.004)
σ_{u02}		−0.18(0.05)	−0.03(0.02)
Deviance	6327.5	4812.8	4747.6

slope; after this interaction is included, the variance of this slope is no longer significant, as determined by a likelihood ratio test.

The interpretation of the main effects model (second model) in Table 14.1 is that girls and more extraverted pupils tend to be more popular. The significant variance for the slope of extraversion (σ_{u2}^2 in the random part) indicates that the effect of extraversion varies across classes. The interaction model (model 3) models this variance with an interaction between extraversion and teacher experience. The negative sign of the regression coefficient for this interaction indicates that the effect of extraversion on popularity is smaller with more experienced teachers. The interpretation of direct effects in the presence of a significant interaction is delicate; in general, it is recommended to support such interactions by drawing a graph using the observed range of the interacting variables (Aiken & West, 1991; Hox, 2010).

When predictor variables are added to the model, the resulting decrease in the residual error variance is often interpreted as explained variance. This interpretation is not quite correct, as Snijders and Bosker (1999) have shown. Table 14.1 illustrates this: when the predictors are added, the unexplained variance at the second level actually appears to increase. In this specific instance, this is the result of the changes in the random part, where a slope variance is added. This completely changes the model. When the random part is left unaltered, adding predictors generally results in decreasing residual error variances, and these are often interpreted as explained variance (Raudenbush & Bryk, 2002). Nevertheless, negative explained variances can and do occur, and interpreting decrease in variance as explained variance is at best an approximation (Hox, 2010).

Measurement Occasions Within Individuals

Longitudinal data, or repeated measures data, can be viewed as multilevel data with repeated measurements nested within individuals. Multilevel analysis of repeated measures is often applied to data from large-scale panel surveys. In addition, it can also be a valuable analysis tool in a variety of experimental designs—for example, intervention studies with an immediate and a later final follow-up measurement, where incomplete data resulting from attrition are common.

The example is a data file compiled by Curran (1997) from a large longitudinal data set. The data are a sample of 405 children who were within the first 2 years of entry to elementary school. The data consist of four repeated measures of both the child's antisocial behavior and the child's reading recognition skills. In addition, at the first measurement occasion, measures were collected of emotional support and cognitive stimulation provided by the mother. Other variables are the child's gender and age and the mother's age at the first measurement occasion. There was an appreciable amount of panel dropout: all 405 children and mothers were interviewed at measurement occasion 1, but on the three subsequent occasions the sample sizes were 374, 297, and 294. Only 221 cases were interviewed at all four occasions. These data have been analyzed extensively in Hox (2010) and can also be obtained from the web (www.joophox.net).

The multilevel regression model for longitudinal data is a straightforward application of the multilevel regression model described earlier. It is also written as a sequence of models for each level. At the lowest, the repeated measures level, we have:

$$Y_{ti} = \pi_{0i} + \pi_{1i} T_{ti} + \pi_{2i} \chi_{ti} + e_{ti}, \tag{4}$$

where Y_{ti} is the outcome variable for subject i at measurement occasion t, T_{ti} is a time indicator for the measurement occasion, and X_{ti} is some other time-varying predictor variable. The regression intercept and slopes are commonly denoted by π_i, so at the individual level we can again use β for the regression coefficients. In our example, the outcome variable could be reading skill, the time indicator could be 0, . . ., 3 for the four measurement occasions, and the time-varying predictor could be antisocial behavior. The intercept and slopes in Equation 4 are assumed to vary across individuals. Just as in two-level models for individuals within groups, this variation can be explained by adding individual level predictors and cross-level interaction effects:

$$\begin{aligned} \pi_{0i} &= \beta_{00} + \beta_{01} Z_i + u_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11} Z_i + u_{1i} \\ \pi_{2i} &= \beta_{20} + \beta_{21} Z_i + u_{2i} \end{aligned} \tag{5}$$

By substitution, we get the single equation model:

$$\begin{aligned} Y_{ti} = \beta_{00} &+ \beta_{10} T_{ti} + \beta_{20} \chi_{ti} + \beta_{01} Z_i \\ &+ \beta_{11} T_{ti} Z_i + \beta_{21} \chi_{ti} Z_i \\ &+ u_{1i} T_{ti} + u_{2i} \chi_{ti} + u_{0i} + e_{ti} \end{aligned} \tag{6}$$

Table 14.2 presents a sequence of models for these data, predicting reading skill from the available predictor variables, omitting non-significant effects.

The interpretation of Table 14.2 is that there is an increase in reading skill over time. Relatively older children and children that are cognitively stimulated have better reading skill. Children vary in the speed at which reading skill increases, which is partially explained by interactions with their age and cognitive stimulation. Relatively older children increase their reading skill less fast, and children who are cognitively stimulated increase faster.

A comparison of the intercept-only model with the model that includes measurement occasion shows the anomaly mentioned earlier; adding occasion results in an increase of the second level variance, hence in negative explained variance. The reason was also mentioned earlier, interpreting changes in the variance terms as explained variance is questionable. The variance decomposition in the intercept-only model depends on the assumption of random sampling at all available levels. In longitudinal panel designs, the sampling at the lowest level follows a very specific scheme, and as a result the occasion level variance is overestimated and the individual level variance is underestimated (for details, *see* Hox, 2010). The pragmatic approach is to use as a null-model a model with measurement occasion properly specified, which in Table 14.2 is the model that includes occasion with a random slope.

Two important advantages of multilevel modeling of longitudinal data should be mentioned. As is clear from the reading skill example, incomplete data resulting from missed measurement occasions are no special problem. In the stacked ("long") data file, the rows corresponding to missed occasions are simply left out, and the analysis proceeds as usual. Given the large fraction of missing data in these data, this is a major advantage. An even more important advantage is that an analysis using repeated measures MANOVA, with listwise deletion of incomplete cases, assumes that missing data are missing completely at random (MCAR), an unlikely assumption. Multilevel analysis assumes missing at random (MAR), which is a much weaker assumption (*see* chapter 27, this volume?).

The second advantage of multilevel modeling for longitudinal data is the flexible treatment of time. Because time is included in the model as a time-varying predictor, we can attempt to specify the metric of time in ways that are more accurate than counting the measurement occasion. In our example, it appears theoretically sounder to use the actual age of the child at each measurement occasion as the time variable. It is more accurate, because it reflects real observed age differences rather than just measurement occasions, and in contrast to measurement occasion, it does have a theoretical interpretation. Table 14.3 highlights the differences between these two metrics of time.

When we use the real age rather than the measurement occasions, which are spaced 2 years apart, we halve the scale of the time variable. Thus, for the age slope, we obtain values that are precisely half the

Table 14.2. Multilevel Models for Longitudinal Data Reading Skill

Model	**Intercept-only**	**Add occasion**	**Occasion varying**	
Fixed part				
Predictor	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	4.11(0.05)	2.70(0.05)	2.70(0.05)	−3.28(0.42)
Occasion		1.10(0.02)	1.12(0.02)	2.23(0.24)
Child age				0.80(0.06)
Cogn. Stim.				0.05(0.01)
Occasion*Child age				−0.19(0.03)
Occasion*Cogn. Stim.				0.02(0.01)
Random part				
σ_e^2	2.39(0.11)	0.46(0.02)	0.35(0.02)	0.35(0.02)
σ_{u0}^2	0.30(0.08)	0.78(0.07)	0.57(0.06)	0.30(0.04)
σ_{u1}^2			0.07(0.01)	0.06(0.01)
σ_{u01}			0.06(0.02)	0.11(0.02)
r_{u01}			0.29(0.13)	0.86(0.18)
Deviance	5051.8	3477.1	3371.8	3127.9

Table 14.3. Comparing Occasion and Child's Age for Longitudinal Data Reading Skill

Model				
Fixed part	**Occasion**	**Occasion varying**	**Child age**	**Child age varying**
Predictor	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	2.70(0.05)	2.70 (0.05)	2.19 (0.05)	2.16(0.04)
Occasion	1.10(0.02)	1.12 (0.02)	–	–
Child age	–	–	0.55 (0.01)	0.56(0.01)
Cogn. Stim.				
Occasion*Child age				
Occasion*Cogn. Stim.				
Random part				
σ_e^2	0.46(0.02)	0.35 (0.02)	0.45 (0.02)	0.36(0.02)
σ_{u0}^2	0.78(0.07)	0.57 (0.06)	0.65 (0.06)	0.17(0.05)
σ_{u1}^2		0.07 (0.01)		0.02(0.003)
σ_{u01}		0.06 (0.02)		0.05(0.001)
r_{u01}		0.29 (0.13)		0.88(0.30)
Deviance	3477.1	3371.8	3413.9	3226.8
AIC	3485.1	3383.8	3421.9	3238.8

values of the occasion slope. But the child level variances are quite different. When models are nested, meaning that we can proceed from one model to the next by adding (or deleting) terms, the model change can be tested using a test on the deviances of the models. However, replacing the predictor variable measurement occasion by actual age does not lead to nested models. Provided the dependent variable and the number of cases remain the same (which implies no additional missing values induced by using age), we can compare such models using the Akaike Information Criterion (AIC). The AIC (Akaike, 1987) is calculated as the deviance minus twice the number of estimated parameters; models with a lower AIC are considered to be better. Thus, the values of the AIC in Table 14.3 suggest that using the actual age results in better models. For a general discussion of these issues, *see* Willett and Singer (2003). A more detailed analysis of the reading skill data using child age as the metric of time can be found in Hox (2010), which also discusses the AIC and related indices in more detail.

Multilevel Structural Equation Modeling: Introduction and Typical Applications

Structural equation models are a very flexible family of models that allow estimation of relationships between observed and latent variables, direct and indirect effects, and assessment of the fit of the overall model. Conventional SEM software can be tricked to estimate two-level models by viewing the two levels as two groups and using the multigroup option of conventional software (Muthén, 1994). The approach outlined by Muthén is a limited information method. Mehta and Neale (2005) have described how general multilevel models can be incorporated in SEM, and how these models can be estimated by conventional SEM software. Using conventional SEM software requires incredibly complicated set-ups, but recent versions of most SEM software incorporates handle these complications internally and have special multilevel features in their command language, which make it easier to specify multilevel models.

Latent Curve Modeling

An interesting structural equation model for panel data is the latent curve model (LCM), sometimes referred to as the latent growth model (LGM). In the LCM, the measurement occasions are defined by the factor loadings in the measurement model of the latent intercept and slope factors. Figure 14.1

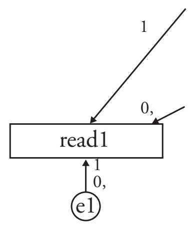

Figure 14.1 Path diagram for the intercept + slope model for reading skill.

shows the path diagram of a simple LCM for the reading skill data. The loadings of the intercept factor are all constrained to 1, and the loadings of the slope factor are constrained to 0, 1, 2, and 3, successively. Thus, the loadings of the slope factor specify the four measurement occasions. The means of the intercept and slope factors are equal to the estimates of the intercept and slope in the corresponding multilevel model, and the variances are equal to the variances of the intercept and slope in multilevel regression.

It can be shown that the LCM and the multilevel regression model for longitudinal data are identical. That does not mean that there are no differences between the two approaches. For example, in SEM, it is trivial to use the intercept and slope factors in a GCM as predictors of some distant outcome. This is very difficult in a multilevel regression model. On the other hand, in multilevel regression software, it is trivial to extend the model with additional levels, whereas most current multilevel SEM software can deal with only two levels. In addition, in multilevel regression, the time variable is a predictor variable, which makes it easy to use the actual child ages rather than the measurement occasions (recent versions of SEM software like Mplus and Mx allow varying time-points as well but still have issues with widely varying numbers of measurement occasions). However, as MacCallum et al. have phrased it: "A wide range of models have equivalent representations in either framework" (MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997, p. 246). The most important conclusion to draw from the comparison of GCM using a multilevel versus a structural equation approach is that these models are fundamentally the same but generally have a different representation in dedicated multilevel or SEM software. Hence, differences between these two approaches are more apparent than real (Bollen & Curran, 2006).

When the model presented in Figure 14.1 is estimated using conventional SEM software, the output highlights one important difference between the multilevel regression and the SEM approach. The

SEM analysis produces the same estimates as the multilevel regression, but it also produces a global model test and several goodness-of-fit indices. The global chi-square test rejects the model ($\chi^2(5) = 174.6, p < 0.001$), and the fit indices indicate a very bad fit Comparative Fit Index (CFI) = 0.78, Root Mean Square Error of Approximation (RMSEA) = 0.29 [95% CI 0.25–0.33]). This is important information that the multilevel regression approach does not provide. Further exploration of the model shows that the latent curve is decidedly nonlinear. If the slope loadings for readings 3 and 4 are estimated freely, then they are estimated as 1.6 and 2.1, respectively, which is quite different from the linear constraints of 2.0 and 3.0. The resulting model shows an excellent fit ($\chi^2(3) = 4.3, p = 0.23$, CFI = 1.00, RMSEA = 0.03 [95% CI 0.00–0.11]). A more detailed multilevel regression analysis of these data in Hox (2010), using the actual child ages, also finds a strongly nonlinear curve.

Multilevel Structural Equation Modeling

The LCM is a real multilevel model, where the latent factors represent the random regression coefficients of the multilevel regression model, but it can be specified as a conventional single level structural model. Multilevel structural equation models in general need the aforementioned extensions in the SEM software to be estimated easily. Multilevel structural equation modeling assumes sampling at the individual and the group level, with both within-group (individual level) and between-group (group level) variation and covariation. In multilevel regression modeling, there is one dependent variable and several independent variables, with independent variables at both the individual and group level. At the group level, the multilevel regression model includes random regression coefficients and error terms. In the multilevel SEM, the random intercepts are second-level latent variables, capturing the variation in the means of the observed individual level variables. Some of the group level variables may be random slopes, drawn from the first level model, but other group level variables may be variables defined only at the group level, such as group size.

Mehta and Neale (2005) explain how multilevel SEM can be incorporated into conventional SEM. By viewing groups as observations, and individuals within groups as variables, they show that models for multilevel data can be specified in the full-information SEM framework. Unbalanced data—that is, unequal numbers of individuals within groups—are handled the same way as incomplete data in modern SEM estimation methods. So, in theory, multilevel SEM can be specified in any SEM package that supports FIML estimation for incomplete data. In practice, specialized software routines are used that take advantage of specific structures of multilevel data to achieve efficient computations and good convergence of the estimates. Extensions of this approach include extensions for categorical and ordinal data, incomplete data, and adding more levels. These are described in detail by Skrondal and Rabe-Hesketh (2004).

In two-level data, the observed individual level variables are modeled by:

$$\begin{aligned} \mathbf{y}_W &= \boldsymbol{\Lambda}_W \boldsymbol{\eta}_W + \boldsymbol{\varepsilon}_W \\ \boldsymbol{\mu}_B &= \boldsymbol{\mu} + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_B + \boldsymbol{\varepsilon}_B, \end{aligned} \tag{7}$$

where μ_B are the random intercepts for the variables $\mathbf{y}_W$ that vary across groups. The first equation models the within-groups variation, and the second equation models the between-groups variation and the group level means. By combining the within and between equations, we obtain

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_W \boldsymbol{\eta}_W + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_B + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_W. \tag{8}$$

In Equation 8, μ is the vector of group level means, Λ_W is the factor matrix at the within level, Λ_B is the factor matrix at the between level, and ε_W and ε_B are the residual errors at the within and the between level. With the exception of the notation, the structure of Equation 8 follows that of a random intercept regression model, with fixed regression coefficients (loadings) in the factor matrices Λ and a level-one and level-two error term. By allowing group level variation in the factor loadings, we can generalize this to a random coefficient model. The model in Equation 8 is a two-level factor model, by adding structural relationships between the latent factors at either level, we obtain a two-level SEM.

Multilevel SEMs are often estimated in separate steps. First, the intraclass correlations of the variables are inspected. If they are all small—for example, smaller than 0.05—the between-group variance is small and there may be no need for a complex group level model. The dependency in the data can be dealt with using standard analysis methods for cluster samples. If the between-group variances are considerable, then an investigation of the between structure is warranted. In general, because the sample size at the individual level is generally much larger than the sample size at the group level, the analysis is started with an analysis of the within structure. Standard analysis methods for clustered samples can

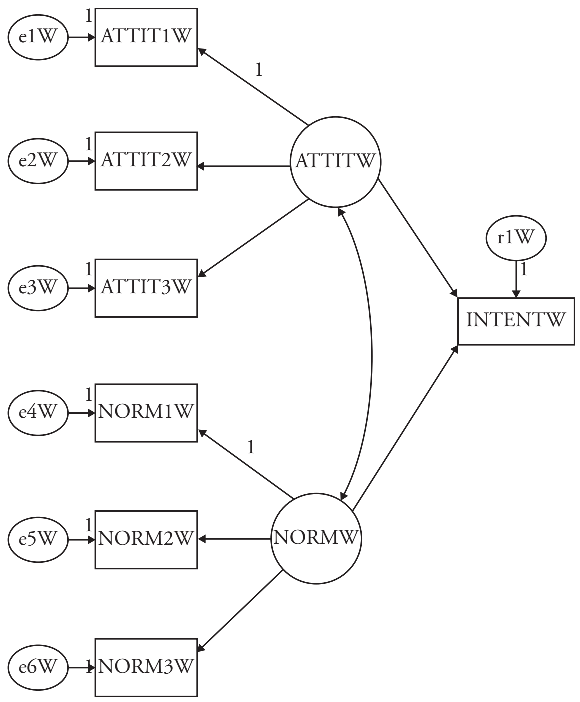

Figure 14.2 Multilevel model for group level intervention.

be used here, such as the complex sample analysis methods used in survey research (cf. de Leeuw, Hox, & Dillman, 2008), which are implemented in, for example, Mplus. Next, the between structure is investigated in a two-level model with the within-structure fully specified.

Figure 14.2 depicts a two-level model that contains both observed and talent variables at both levels. It represents a model based on the theory of reasoned action (Ajzen & Fishbein, 1980) that predicts behavior from intention toward that behavior, and intention is in turn predicted from attitudes and social norms concerning that behavior. The attitudes and norms are latent factors, each indicated by three observed variables. In general, unless the intraclass correlation is 0, all observed variables exist at both the individual and the group level. Note that the variables that are observed variables at the individual level are latent variables at the group level; these latent variables represent the group-level variation of the intercepts. There is one variable that exists only at the group level. The variable *expcon* represents some experimental intervention at the group level, aimed at changing the attitude toward the behavior. If the groups are assigned at random to the intervention or the control condition, then this example represents a group randomized trial.

Example data were generated directly from the model, for 100 groups of 10 subjects each and intraclass correlations of around 0.10, which is relatively

Table 14.4. Unstandardized Factor Loadings (Standard Errors) for Attitude and Norms

Attit1	1.00*	–
Attit1	1.00(0.06)	–
Attit1	0.98(0.05)	–
Norm1	–	1.00*
Norm2	–	0.98(0.05)
Norm3	–	0.98(0.05)

Note: * indicates constrained for identification. Correlation between factors estimated as 0.50 (within) and 0.69 (between).

high but not unusual. All variables are continuous; to simplify the modeling the intervention variable is ordered categorical with five categories.

The model depicted in Figure 14.2 is estimated using Mplus (Muthén & Muthén, 1998–2010). The program reports the intraclass correlations for all observed variables; these range from 0.15 to 0.22. Thus, multilevel modeling of these data is justified. The fit of the model is excellent ($\chi^2(44) = 26.7$, $p = 0.98$, CFI = 1.00, RMSEA = 0.00), which is unsurprising because the example data were generated from this model.

The model illustrates some issues that occur more generally in two-level SEM. First, we have a measurement model that specifies how attitude and norms are measured by the observed variables. Because the measurement model is the same at both levels, the question arises if we can impose equality constraints on the factor loadings across the two levels. If we impose these four constraints, the chi-square increases by 3.748, which with four degrees of freedom is not significant ($p = 0.44$). Table 14.4 presents the unstandardized factor loadings after imposing the equality constraints. Because we have established that there is measurement equivalence across the two levels, we can proceed to calculate the intraclass correlations for the two latent factors. If we specify a model without the intervention variable, then the intraclass correlation for attitude is 0.20 and for norms 0.16. The intraclass correlation for attitude is inflated because part of the variance in attitude is caused by the group-level intervention. If we analyze the model including the intervention variable, then the intraclass correlation for attitude is estimated as 0.17; this could be interpreted as a partial intraclass correlation, disregarding the variance in attitude caused by the intervention.

Table 14.5. Direct and Indirect Paths from Intervention to Behavior, Group Level

Dependent	Independent (mediating) variables path coefficient (standard error)		
	Intervention	Attitude	Intention
Attitude	0.52 (0.12)	–	–
Intention	0.34 (0.08)	0.64 (0.09)	–
Behavior	0.25 (0.07)	0.49 (0.08)	0.75 (0.06)

In addition to the inclusion of latent variables, SEM allows estimating and testing indirect effects. In our example, the effect of the intervention on the behavior is mediated at the group level by attitude and intention. Table 14.5 shows the standardized direct and indirect effects of the paths leading from the intervention to behavior, at the group (between) level.

The group-level explained variances are 0.27 for attitude, 0.74 for intention, and 0.57 for behavior. Predictably, the effect of the intervention becomes smaller when the chain of mediating variables becomes longer. The explained variance of the intervention on the attitude is 0.27, which translates to a correlation of 0.52—in Cohen's (1988) terms, a large effect size. In this example, the mediation is entirely at the group level. It is possible to model mediation chains where the group-level intervention affects individual-level variables (latent or observed) that in turn affect group or individual level outcomes. Especially when random slopes are involved, multilevel mediation is a complex phenomenon, and I refer to MacKinnon (2008) for a thorough discussion of the details.

Methodological and Statistical Issues

Assumptions

Multilevel regression and SEM make the same assumptions as their single-level counterparts. So, multilevel regression analysis assumes perfectly measured predictor variables, linearity of relationships, normal residual errors, homoscedasticity, and independence conditional on the grouping variables in the model. In addition, it assumes that the residual errors at the separate levels are independent. Structural equation modeling can incorporate a measurement model; thus, there is no assumption that variables are measured without measurement error, but otherwise the assumptions are much the same.

Investigating potential violations of assumptions is more complicated in multilevel models than in their single-level counterparts. For example, if there are random slopes in the model, then at the group level there is a set of residuals that are generally assumed to have a multivariate normal distribution. Investigating the normality assumption here implies investigating all residuals. In addition, the model itself is more complex. For example, Bauer and Cai (2009) have shown that if a nonlinear effect is not modeled as such, then this misspecification may show up as an entirely spurious variance parameter for a slope or a spurious cross-level interaction effect. Wright (1997) has shown that in multilevel logistic regression, sparse data resulting from skewed distributions or small samples may result in spurious overdispersion (a variance larger than implied by the underlying binomial distribution). So, investigating assumptions is both more difficult and more important in multilevel models. Specialized multilevel software such as HLM and MLwiN incorporate many procedures for investigating assumptions that are specific to multilevel regression models, but more general software like SAS, SPSS, or Mplus for multilevel SEM do not incorporate such features and rely completely on the ingenuity of the researcher to devise diagnostic checks.

Sample Size

In multilevel modeling, the most important limitation on sample size is generally the second or higher level, because the higher level sample sizes are usually smaller than the lower level sample sizes. Eliason (1993) recommends a minimum sample size of 60 when maximum likelihood estimation is used. In multilevel modeling, this would apply to the highest level. Maas and Hox (2005) have found that in multilevel regression modeling, a highest level sample size as low as 20 may be sufficient for accurate estimation, provided that the interest is in the regression coefficients and their standard errors. If the interest is in the variance estimates, then the higher level sample sizes must be much larger, and Maas and Hox have recommended at least 100 groups (although 50 groups may suffice for small models). Multilevel SEM are fundamentally based on the within-group and between-group covariance matrices, and hence it is not surprising that the recommendation for the accurate estimation of higher level variances in multilevel regression carries over to SEM: at least 100 groups are recommended, but in small models 50 groups may suffice (Hox, Maas, & Brinkhuis, 2010).

Unequal sample sizes at any of the levels are not a problem, as the model does not assume equal sample sizes at all. Missing values resulting from missing occasions or panel dropout can be dealt with easily in longitudinal models. However, incomplete data at the higher level are more difficult to handle. Structural equation software is sometimes able to analyze incomplete data directly using full information maximum likelihood procedures, but most multilevel software does not have such provisions. Multiple imputation is an option, but the problem is that the imputation model must also be a multilevel model. Van Buuren (2011) has discussed incomplete multilevel data in detail.

Small sample sizes at the lowest level do not pose a problem by themselves. For example, multilevel models have proven valuable in analysis of dyadic data, such as couples or twins (Atkins, 2005). Even groups of size 1 are fine, provided other groups are larger. However, small groups present some limitations, especially to the complexity of the within-groups (individual level) model. A model with a random intercept and one random slope is just identified, and more complex models cannot be estimated (Newsom, 2002). For a recent review of multilevel models for dyadic data, *see* Kenny and Kashy (2011).

Further Important Issues

In multilevel modeling, predictor variables are sometimes centered on some value. Centering on a single value, usually the grand mean of the predictor variable, poses no special problems. It facilitates estimation—especially when multicollinearity is present—and makes the interpretation of interactions easier. Centering predictor variables on their respective group means is different. Group mean centering totally changes the meaning of the model and should be used with caution. In particular, group mean centering removes all information about the group means from the model. Adding the group means as predictor variables to the model solves that issue, but the resulting model is still fundamentally different from a model that incorporates the original uncentered predictor variables. Enders and Tofighi (2006) have discussed these issues in detail and have provided some guidelines for when group mean centering is appropriate.

Effect sizes are somewhat problematic in multilevel models. In general, calculating explained variance is not different from calculating explained variance in similar single-level models. However, in multilevel modeling, one would want to be able to establish how much variance is explained at each of the available levels. This turns out to be problematic. Simply using the reduction in residual variance when predictor variables are added as suggested in Raudenbush and Bryk (2002) does not work, as this procedure can result in impossible values such as negative explained variances (Hox, 2010; Snijders & Bosker, 1994). There have been several proposals to cope with this problem (Roberts, Monaco, Stovall, & Foster, 2011; Snijders & Bosker, 1999), but these tend to be complicated and to have their own problems. In the end, Hox (2010) has recommended using the simple method (Raudenbush & Bryk, 2002), in combination with grand mean-centered predictors, and interpreting the resulting values as indicative, rather than mathematical truth.

In regression and SEM, the interest is often mostly on the fixed coefficients—that is, the regression coefficients, factor loadings, and path coefficients. Their significance can be tested using their standard errors. In latent growth models and in multilevel SEM, there is often considerable substantive interest in the variance components as well—for example, in testing whether the higher level variances are significant. Testing variances using the standard error is generally not a very accurate approach, because variances do not have a normal distribution. For significance testing, the recommended method is comparing a model that includes the variance component with a model that does not include it, using a likelihood ratio test or the equivalent deviance difference test (cf. Berkhof & Snijders, 2001). Establishing correct confidence intervals for variance components is possible using multilevel bootstrap methods (Goldstein, 2011) or Baesian approaches (Hamaker & Klugkist, 2011).

The multilevel regression and the multilevel SEM were originally developed for continuous and (multivariate) normal variables. Both have been extended to include non-normal variables, such as dichotomous, ordered categorical, or count variables. With such variables, estimation problems tend to occur. For multilevel logistic regression, estimation procedures have been developed based on Taylor series linearization of the nonlinear likelihood. These methods are approximate, and in some circumstances (such as the combination of small groups and high intraclass correlations) the approximation is not very good. Numerical methods that maximize the correct likelihood are superior (Agresti, Booth, Hobart, & Caffo, 2000), but they can be computationally intensive, especially in models that contain a large number of random effects. For such models, Bayesian estimation procedures are attractive. Some general software for multilevel modeling, such as MLwiN and Mplus, include Bayesian estimation options. General Bayesian modeling software such as (Win)BUGS can be used for multilevel modeling, but these require more complicated setups. For an introduction to Bayesian multilevel modeling, I refer the reader to Hamaker and Klugkist (2011), and a detailed discussion including setups in BUGS is given by Gelman and Hill (2007).

Conclusion

Multilevel models are increasingly used in a variety of fields. Initially these models were viewed as a means to properly analyze hierarchical data, with individual cases or measures nested within larger units such as groups. Although such applications still abound, applications have come to include models for cross-classified hierarchical data, dyadic data, network analysis, meta-analysis, and spatial modeling. The common characteristic of these models is that they contain complex relationships involving random effects. Multilevel analysis is a tool that allows great flexibility in the actual modeling, which is why it is an attractive option in analyzing these complex models.

Future Directions

As has been noted, multilevel models can be specified as simple SEM that can be analyzed using standard structural equation software. In practice, this leads to model set-ups that are unwieldy, and recent SEM software has incorporated special features to accommodate multilevel models. At the time of writing, multilevel structural equation software has practical limitations, such as a limited number of levels, convergence problems, or long execution times. As software development continues, structural equation software will outgrow these limitations.

A difficult problem in actual research is often obtaining a large enough sample on the higher levels; the maximum likelihood estimation method requires a reasonable sample size to be accurate. Bayesian methods are promising in this respect—they tend to be more stable with smaller sample sizes and will always generate parameter values that are within their proper boundaries. However, Bayesian methods are still undergoing rapid development, and standard software lags behind in their implementation. As standard software (as opposed to specialized Bayesian software such as WINBUGS) develops to incorporate Bayesian methods (at the time of writing already available in the software MLwiN and Mplus), it is expected that their use will increase.

A problem that still awaits a good solution is incomplete multilevel data, including missing data at the higher levels. In SEM, estimation methods have been developed that provide parameter estimates based on the incomplete data themselves; no listwise or pairwise deletion or imputation of missing values is involved. Estimation methods for multilevel models generally lack this flexibility. In addition, multilevel multiple imputation must be considered to be in its infancy. Given the requirement that the imputation model must be at least as complex as the analysis model, developing proper procedures for multilevel multiple imputation is a daunting task.

Glossary of Key terms	
Between groups	Model for the structure at the group level. Usual term in two-level SEM to refer to the group (second) level. As three- and more-level SEM develops, this term is becoming unclear, and better replaced with a reference to the level of interest (e.g., class or school level).
Cross-level interaction	Higher level variables may have a direct effect on the outcome in a multilevel model, or they may affect the effects of lower level variables on the outcome. This is generally modeled by an interaction between a higher level and a lower level predictor variable.
Fixed effect, fixed coefficient	Regression coefficients (including factor loadings and path coefficients) that do not vary across higher level units.

Glossary of Key terms (Continued)	
Intraclass correlation	The estimate of the similarity in the population between individuals belonging to the same group. Also defined as the proportion of variance (in the population) at the group level.
Mixed model	A model that contains both fixed effects and random effects.
Multilevel model	A model that contains variables defined at different levels of a hierarchically structured population. Other terms used are hierarchical linear model, mixed model and random coefficient model. Although these models are not identical, in practice these terms are often used interchangeably.
Random effect, random coefficient	Regression coefficients (including factor loadings and path coefficients) that are assumed to vary across higher level units. They are generally assumed to have a normal distribution with a mean of zero and some variance that is estimated.
Variance component	Generally used to refer to the higher level variances and covariances of the varying coefficients. In multilevel analysis of longitudinal data specific structures are sometimes assumed for the variances and covariances over time.
Within groups	Model for the structure at the lowest level. Usual term in two-level SEM to refer to the individual (first) level. As three- and more-level SEM develops, this term is becoming unclear and better replaced with a reference to the level of interest (e.g., individual or measurement-occasion level).

An extended glossary to key terms used in multilevel regression modeling is presented by Diez Roux (2002).

Symbols used	
β_{pj}	Regression coefficient for variable p varying at the level indicated by j
γ_p	Fixed regression coefficient for variable p
π	Regression coefficient for time or measurement occasion indictor
u_{pj}	Residual error term for variable p varying at the level indicated by j
σ_u^2	Variance of residual error u
σ_e^2	Variance of lowest level residual error e
Λ	Factor matrix, subscript B or W indicates Between/Within level
η	Factor score, subscript B or W indicates Between/Within level

References

Agresti, A., Booth, J.G., Hobart, J.P., & Caffo, B. (2000). Random effects modeling of categorical response data, *Sociological Methodology, 2000*, *30*, 27–80.

Aiken, L.S., & West, S.G. (1991). *Multiple regression: Testing and interpreting interaction*. Newbury Park, CA: Sage.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Akaike, H. (1987). Factor analysis and the AIC. *Psychometrika*, *52*, 317–332.

Atkins, D.C. (2005). Using multilevel models to analyze couple and family treatment data: Basic and advanced issues. *Journal of Family Psychology*, *19*, 98–110.

Bauer, D.J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, *34*, 97–114.

Berkhof, J. & Snijders, T.A.B. (2001). Variance component testing in multilevel models. *Journal of Educational end Behavioral Statistics*, *26*, 133–152.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: Guilford Press.

Bollen, K.A. & Curran, P.J. (2006). *Latent curve models*. New York: Wiley.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

van Buuren, S. (2011). Multiple imputation of multilevel data. In J.J. Hox & J.K. Roberts (Eds). *Handbook of Advanced Multilevel Analysis* (pp. 173–196). New York: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Curran, P.J. (1997). *Supporting Documentation for Comparing Three Modern Approaches to Longitudinal Data Analysis: An Examination of a Single Developmental Sample*. Retrieved, May 2012 from http://www.unc.edu/ ~curran/srcd.html.

Diez Roux, A.V. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health*, *56*, 588–594.

Eliason, S.R. (1993). *Maximum Likelihood estimation*. Newbury Park, CA: Sage.

Enders, C.K., & Tofighi, D. (2006). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*, 121–138.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Goldstein, H. (1987). Multilevel models in educational and social research. London: Griffin/New York: Oxford University Press.

Goldstein, H. (2011). Bootstrapping in multilevel models. In J.J. Hox & J.K. Roberts (Eds.). *Handbook of advanced multilevel analysis*. (pp. 163–172). New York: Routledge.

Hamaker, E.L., & Klugkist, I. (2011). Bayesian estimation of multilevel models. In J.J. Hox & J.K. Roberts (Eds.). *Handbook of advanced multilevel analysis*. (pp. 137–162). New York: Routledge.

Hox, J.J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J.J. (2010). *Multilevel analysis: Techniques and applications. Second edition.* New York: Routledge.

Hox, J.J., Maas, C.J.M., & Brinkhuis, M.J.S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170.

Hüttner, H.J.M., & Van den Eeden, P. (1995). The multilevel design. London: Greenwood Press.

Kenny, D.A., & Kashy, D.A. (2011). Dyadic data analysis using multilevel modeling. In J.J. Hox & J.K. Roberts (Eds.). *Handbook of advanced multilevel analysis*. (pp. 335–370). New York: Routledge.

de Leeuw, E.D., Hox, J.J. & Dillman, D.A. (Eds.) (2008). *The International Handbook of Survey Methodology*. New York: Taylor & Francis.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 85–91.

MacCallum, R.C., Kim, C., Malarkey, W.B., & Kiecolt-Glaser, J.C. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*, 215–253.

Mackinnon, D.P. (2008). Introduction to statistical mediation analysis. New York: Taylor & Francis.

Mehta, P.D., & Neale, M.C. (2005), People are variables too: multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284.

Muthén, B.O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398.

Muthén, L.K. and Muthén, B.O. (1998–2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Newsom, J.T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling*, *9*, 431–447.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models*. Second Edition. Thousand Oaks, CA: Sage.

Roberts,, J.K., Monaco, J.P., Stovall, H., & Foster, V. (2011). Explained variance in multilevel models. In J.J. Hox & J.K. Roberts (Eds.). *Handbook of advanced multilevel analysis*. (pp. 219–230). New York: Routledge.

Singer, J.D., & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, United Kingdom: Oxford University Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Snijders, T.A.B., & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*, 342–363.

Snijders, T.A.B., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Wright, D.B. (1997). Extra-binomial variation in multilevel logistic models with sparse structures. *British Journal of Mathematical and Statistical Psychology*, *50*, 21–29.