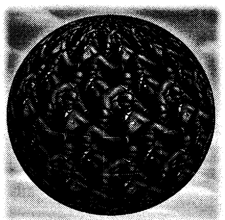

Multilevel Analysis

Joop J. Hox
Utrecht University, Utrecht, The Netherlands

Cora J. M. Maas
Utrecht University, Utrecht, The Netherlands

Glossary

centering Transforming a variable by subtracting the mean. In multilevel data, there are two ways of centering: subtracting the overall mean and subtracting group means. Centering on the overall mean is a straightforward linear transformation, whereas centering on the group mean leads to a radically different model.

cross-level interaction When a regression coefficient of an individual-level variable varies across groups, this variation can be modeled by introducing an interaction term of that variable with one of the group-level variables. Such cross-level interactions are common in multilevel analysis.

cluster sampling A procedure in which data are collected in a two-stage design, starting first with sampling of groups (clusters), followed by sampling of individuals within groups (clusters).

dependence The assumption in cluster sampling and multilevel data that the observations are not sampled independently of one another. The result of this assumption is strongly biased standard errors when standard statistical methods are used.

design effect The amount of bias in the standard errors introduced by having dependent observations.

intraclass correlation The expected correlation between individuals within the same group. When the intraclass correlation is greater than zero, the observations are not independent.

fixed part The part of the model equation that contains the regression coefficients.

generalized linear regression model A regression model that is used when the linearity and distributional assumptions of the linear regression model are not met, for example. when the dependent variable is categorical.

linear regression model A regressions model that describes the relationship between one response variable and one or more explanatory variables by calculating the best-fitting linear line (or plane).

logistic regression A generalized linear regression model for response variables that are dichotomous or proportions.

maximum likelihood (ML) The estimation method most commonly used in multilevel analysis. ML estimation produces estimated parameter values that make the probability of observing the data highest.

multilevel analysis of longitudinal data The analysis of longitudinal data or repeated measurements as multilevel data by viewing these as observations nested within individuals. The advantage is that the analysis deals easily with missing measurement occasions.

multilevel data Data that have a hierarchical or nested structure, usually individuals within groups.

random coefficients The regression coefficients of the lowest-level explanatory variables; these coefficients can vary across groups and part of this variation is assumed to be random (stochastic).

random part The part of the model equation that contains the residual error terms.

variance components The variances (and covariances) of the residual errors. In multilevel analysis, there are variance components at each distinct level.

Social research often concerns the relationship between individuals and the groups to which they belong. This leads to hierarchical or multilevel data structures, with individuals nested within the groups. Examples are educational research with pupils nested within classes nested within schools (a three-level data structure), cross-national studies with individuals nested within national units, and family research with members nested within families. Less obvious applications of multilevel modeling are longitudinal studies with measurement occasions nested within individuals and meta-analysis with subjects nested within studies. More specialized multilevel models have been developed that can incorporate nonnested hierarchical structures and multiple response variables. Multilevel modeling has become popular for the analysis of a variety of problems, going beyond the classical individuals-within-groups applications. This entry gives a brief summary of the reasons for using multilevel models and provides examples why these reasons are indeed valid reasons. Next, the multilevel model is introduced and illustrated with an empirical example. The extension to multilevel logistic regression is briefly discussed.

Introduction

Multilevel modeling is used in the analysis of data that have a hierarchical or clustered structure. Such data arise routinely in various fields, for instance in educational research in which pupils are nested within schools, in family studies in which children are nested within families, in medical research in which patients are nested within physicians or hospitals, and in biomedical research, for instance the analysis of dental anomalies in which teeth are nested within different people's mouths. Clustered data may also arise as a result of the specific research design. For instance, in large-scale survey research the data collection is usually organized in a multistage sampling design that results in clustered or stratified data. Another example is a longitudinal design, in which the data are a series of repeated measurements nested within individual subjects.

A crucial problem in the statistical analysis of hierarchically structured data is the dependence of the observations at the lower levels. Older approaches to the analysis of multilevel data simply ignore this problem and commonly perform the analysis by disaggregating all data to the lowest level and subsequently applying standard analysis methods. The magnitude of the statistical bias introduced by this approach can be illustrated by a simple example from sample surveys. Survey statisticians have long known that the extent to which samples are clustered affects the sampling variance and, hence, causes a bias in statistical significance tests. In his classic 1965 work, Kish defines the design effect (deff) as the ratio of the operating sampling variance to the sampling variance that applies to simple random sampling. Thus, deff is the factor with which the simple random sampling variance must be multiplied to provide the actual operating sampling variance. In simple cluster sampling with equal cluster sizes, deff can be computed by $\text{deff} = [1 + \rho(n_{\text{clus}} - 1)]$, where ρ is the intraclass correlation, and n_{clus} is the common cluster size. (The intraclass correlation ρ indicates the degree of similarity between respondents within the same cluster; the formula is presented in the next section.) It is clear that deff equals 1 only when either the intraclass correlation is zero or the cluster size is 1. In all other situations, deff is larger than 1, which implies that standard statistical formulas will underestimate the sampling variance and therefore lead to biased significance tests with an inflated Type I error rate.

The impact of cluster sampling on the operating α level is often large. For example, assume that we carry out a t test at a nominal α level of 0.05. If we have a cluster sample, with a small intraclass correlation of $\rho = 0.05$ and a cluster size of 10, the actual operating α level is 0.11. With larger intraclass correlations and larger cluster sizes, the operating α level increases rapidly. Consider the effect of cluster sampling in educational research, in which data are often collected from classes. Assuming a common class size of 25 pupils, and a typical intraclass correlation for school effects of $\rho = 0.10$, the operating α level is 0.29 for tests performed at a nominal α level of 0.05! Clearly, in such situations *not* adjusting for clustered data produces very misleading significance tests. In addition, for nonlinear models such as logistic regression, not only the standard errors, but also the regression coefficients themselves are biased.

If we have clustered data, the standard statistical tests can be adjusted using deff. However, multilevel modeling is more general. In most multilevel problems, we have not only clustering of individuals within groups, but we also have variables measured at all available levels. Combining variables from different levels in one statistical model is a different problem than estimating and correcting for design effects. Multilevel models are designed to analyze variables from different levels simultaneously, using a statistical model that includes the various dependencies. This leads to research into the direct effects and the interactions between variables that describe the individuals and variables that describe the groups, a kind of research that is now often referred to as multilevel research.

Multilevel research requires multilevel theories, an area that seems underdeveloped compared to the statistical and computational advances. If there are effects of the social context on individuals, these effects must be mediated by intervening processes that depend on characteristics of the social context. Multilevel models in general assume that the grouping criterion is clear and that variables can be assigned unambiguously to their appropriate level. In reality, group boundaries may be somewhat arbitrary and the assignment of variables is not always obvious and simple. In addition, if we have many variables at many levels, there is an enormous number of possible interactions between different levels. Ideally, a multilevel theory should specify which variables belong to which level and which direct effects and cross-level interaction effects can be expected. The common denominator in such theories is that they all postulate processes that mediate between individual variables and group variables, such as communication processes, social comparison processes, and the internal structure of groups.

The Multilevel Regression Model

The multilevel regression model is known in the statistical literature under a variety of names: hierarchical linear model, random coefficient model, variance component model, and mixed (linear) model. Most often it assumes hierarchical data, with one response variable measured at the lowest level and explanatory variables at all existing levels. Conceptually, the model is often viewed as a hierarchical system of regression equations. For example, assume we have data in J groups or contexts and a different number of individuals N_j in each group. On the individual (lowest) level we have the dependent variable Y_{ij} and the explanatory variable X_{ij}, and on the group level we have the explanatory variable Z_j. Thus, we have a separate regression equation in each group:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \qquad (1)$$

In Eq. (1) β_0 is the usual regression intercept, β_1 is the regression slope for the explanatory variable X, and e_{ij} is the residual term. The regression coefficients β carry a subscript j for the groups, which indicates that the regression coefficients may vary across groups. The variation in the regression coefficients β_j is modeled by explanatory variables and random residual terms at the group level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \qquad (2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j} \qquad (3)$$

Substitution of Eqs. (2) and (3) into Eq. (1) produces the single-equation version of the multilevel regression model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + u_{1j}X_{ij} + u_{0j} + e_{ij} \qquad (4)$$

In general, there will be more than one explanatory variable at the lowest level and also more than one explanatory variable at the highest level. Assume that we have P explanatory variables X at the lowest level, indicated by the subscript p ($p = 1, \ldots, P$), and Q explanatory variables Z at the highest level, indicated by the subscript q ($q = 1, \ldots, Q$). Then, Eq. (4) becomes the more general equation:

$$Y_{ij} = \gamma_{00} + \sum_p \gamma_{p0}X_{pij} + \sum_q \gamma_{0q}Z_{qj} + \sum_q \sum_p \gamma_{pq}Z_{qj}X_{pij} + \sum_p u_{pj}X_{pij} + u_{oj} + e_{ij} \qquad (5)$$

In Eq. (5), the γ are the usual regression coefficients, the u terms are residuals at the group level, and the e term represents the residual at the individual level. The regression coefficients are identified as the fixed part of the model because this part does not change over groups or individuals. The residual error terms are identified as the random or stochastic part of the model.

The assumptions of the most commonly used multilevel regression model are that the residuals at the lowest level e_{ij} have a normal distribution with a mean of zero and a common variance σ^2 in all groups. The second-level residuals u_{0j} and u_{pj} are assumed to be independent of the lowest level errors e_{ij} and to have a multivariate normal distribution with means of zero. Other assumptions, identical to the common assumptions of a multiple regression analysis, are fixed predictors and linear relationships. Most multilevel software assumes by default that the variance of the residual errors e_{ij} is the same in all groups. However, certain forms of heteroscedasticity can be explicitly modeled.

The estimation of parameters (regression coefficients and variance components) in multilevel modeling is generally done using the maximum likelihood (ML) method. The standard errors (SEs) generated by the ML procedure are asymptotic, meaning we need fairly large samples at all levels. These standard errors can be used to establish a p value for the null hypothesis that in the population a specific regression coefficient is zero. Thus, the significance of a regression coefficient can be tested by referring $Z = \beta/SE(\beta)$ to the standard normal distribution. The ML procedure also generates a value for the deviance that is based on the likelihood (the deviance equals −2 times the log-likelihood). In addition to the standard errors, the deviance can also be used to test parameters for significance. When two models are nested, which means that the smaller model can be obtained by removing terms from the larger model, the difference between the deviances of these two models has a chi-square distribution, with degrees of freedom being the difference in numbers of estimated parameters. This is useful for testing the significance of variance terms. The asymptotic Z test previously described is not optimal for testing variances. First, it assumes normality, and variances do not have a normal distribution. Second, testing the null hypothesis that a variance is zero is a test on the boundary of the parameter space (variances cannot be negative), where standard likelihood theory is no longer valid. The significance of a variance component can be tested by comparing the deviance of a model containing this parameter to the deviance of the same model without this one variance parameter. This value can be treated as a chi-square variate with one degree of freedom, and this can be used to test the significance of that variance component. It should be noted that Raudenbush and Bryk present a different chi-square test for variance components, which is not based on the deviance.

Two different likelihood functions are commonly used in multilevel regression analysis. The first is full maximum likelihood (FML). The second is restricted maximum likelihood (RML). The difference is that

RML maximizes a likelihood function that is invariant for the fixed effects. Because RML is more realistic, it should, in theory, lead to better estimates of the variance components, especially when the number of groups is small. Nevertheless, FML has one advantage over RML—because the likelihood is maximized over both the fixed and the random part, the difference between two deviances can be used to test for differences between two nested models that differ only in the fixed part (the regression coefficients). With RML, only differences in the random part (the variance components) can be tested this way.

The proportion of variance in the population explained by the grouping structure is indicated by the intraclass correlation ρ. The model used to estimate ρ is the model that contains no explanatory variables at all, called the intercept-only model:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{6}$$

Using this model, the intraclass correlation ρ is estimated by the equation:

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e} \tag{7}$$

where $\sigma^2_{u_0}$ is the variance of the second-level residuals u_{0j} and σ^2_e is the variance of the lowest level residuals e_{ij}.

Example of Multilevel Regression Analysis

Assume that we have data from school classes. On the pupil level, we have the outcome variable Popularity measured by a self-rating scale that ranges from 0 (very unpopular) to 10 (very popular). We have one explanatory variable Gender (0 = boy, 1 = girl) on the pupil level and one class level explanatory variable Teacher experience (in years). We have data from 2000 pupils from 100 classes, so the average class size is 20 pupils. The data are described and analyzed in more detail in Hox's 2002 handbook.

Table I presents the parameter estimates and standard errors for a series of models. Model M0 is the null model, the intercept-only model. The intercept-only model estimates the intercept as 5.31, which is simply the weighted average popularity across all schools and pupils. The variance of the pupil-level residuals is estimated as 0.64. The variance of the class-level residuals is estimated as 0.87. The intercept estimate is much larger than the corresponding standard error, and the calculation of the Z test shows that it is significant at $p < 0.005$. As previously mentioned, the Z test is not optimal for testing variances. If the second-level variance term is restricted to zero, the deviance of the model goes up to 6489.5. The difference between the deviances is 1376.8, with one more parameter in the intercept-only model. The chi-square of 1376.8 with one degree of freedom is also significant at $p < 0.005$. The intraclass correlation is $\rho = \sigma^2_{u0}/(\sigma^2_{u_0} + \sigma^2_e) = 0.87/(0.87 + 0.64) = 0.58$. Thus, 58% of the variance of the popularity scores is at the group level, which is very high. Because the intercept-only model contains no explanatory variables, the variances terms represent unexplained residual variance.

Model M1 predicts the outcome variable Popularity by the explanatory variables Gender and Teacher experience, with a random component for the regression coefficient of gender, and model M2 adds the cross-level interaction term between Gender and Teacher experience. We can view these models as built up in the following sequence of steps:

$$\text{Popularity}_{ij} = \beta_{0j} + \beta_{1j}\text{Gender}_{ij} + e_{ij} \tag{8}$$

Table I Multilevel Models for Pupil Popularity

Model	*M0: Intercept-only*	*M1: +Pupil gender and Teacher experience*	*M2: +Cross-level interaction*
Fixed part			
Predictor	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	5.31 (0.10)	3.34 (0.16)	3.31 (0.16)
Pupil gender		0.84 (0.06)	1.33 (0.13)
Teacher experience		0.11 (0.01)	0.11 (0.01)
Pupil gender Teacher exprience			−0.03 (0.01)
Random part			
σ^2_e	0.64 (0.02)	0.39 (0.01)	0.39 (0.01)
$\sigma^2_{u_0}$	0.87 (013)	0.40 (0.06)	0.40 (0.06)
$\sigma^2_{u_1}$		0.27 (0.05)	0.22 (0.04)
$\sigma_{u_{01}}$		0.02 (0.04)	0.02 (0.04)
Deviance	5112.7	4261.2	4245.9

In this regression equation, β_{0j} is the usual intercept, β_{1j} is the usual regression coefficient (regression slope) for the explanatory variable gender, and e_{ij} is the usual residual term. The subscript j is for the classes ($j = 1, \ldots, J$) and the subscript i is for individual pupils ($i = 1, \ldots, N_j$). We assume that the intercepts β_{0j} and the slopes β_{1j} vary across classes.

In our example data, the model corresponding to Eq. (8) results in significant variance components at both levels (as determined by the deviance-difference test). In the next step, we hope to be able to explain at least some of this variation by introducing class-level variables. Generally, we will not be able to explain all the variation of the regression coefficients, and there will be some unexplained residual variation—hence the name random coefficient model, the regression coefficients (intercept and slopes) have some amount of (residual) random variation between groups. Variance component model refers to the statistical problem of estimating the amount of random variation. In our example, the specific value for the intercept and the slope coefficient for the pupil variable Gender are class characteristics. A class with a high intercept is predicted to have more popular pupils than a class with a low value for the intercept. Similarly, differences in the slope coefficient for gender indicate that the relationship between the pupils' gender and their predicted popularity is not the same in all classes. Some classes may have a high value for the slope coefficient of gender; in these classes, the difference between boys and girls is relatively large. Other classes may have a low value for the slope coefficient of gender; in these classes, gender has a small effect on the popularity, which means that the difference between boys and girls is small.

The next step in the hierarchical regression model is to explain the variation of the regression coefficients β_{0j} and β_{1j} by introducing the explanatory variable Teacher experience at the class level. Model M1 models the intercept as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,\text{Teacher experience}_j + u_{0j} \qquad (9)$$

and model M2 models the slope as follows:

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\,\text{Teacher experience}_j + u_{1j} \qquad (10)$$

Equation (9) predicts the average popularity in a class (the intercept β_{0j}) by the teacher's experience. Thus, if γ_{01} is positive, the average popularity is higher in classes with a more experienced teacher. Conversely, if γ_{01} is negative, the average popularity is lower in classes with a more experienced teacher. The interpretation of Eq. (10) is more complicated. Equation (10) states that the *relationship*, as expressed by the slope coefficient β_{1j}, between the popularity and the gender of the pupil, depends on the amount of experience of the teacher. If γ_{11} is positive, the gender effect on popularity is larger with experienced teachers. On the other hand, if γ_{11} is negative, the gender effect on popularity is smaller with experienced teachers. Thus, the amount of experience of the teacher interacts with the relationship between popularity and gender; this relationship varies according to the value of the teacher experience.

The u terms u_{0j} and u_{1j} in Eqs. (9) and (10) are the residual terms at the class level. The variance of the residual u_{0j} is denoted by $\sigma^2_{u_0}$, and the variance of the residual u_{1j} is denoted by $\sigma^2_{u_1}$. The covariance between the residuals u_{0j} and u_{1j} is $\sigma_{u_{01}}$, which is generally *not* assumed to be zero.

Our model with one pupil-level and one class-level explanatory variable including the cross-level interaction can be written as a single complex regression equation by substituting Eqs. (9) and (10) into Eq. (8). This produces:

$$\begin{aligned}\text{Popularity}_{ij} = {} & \gamma_{00} + \gamma_{10}\,\text{Gender}_{ij} \\ & + \gamma_{01}\,\text{Teacher experience}_j \\ & + \gamma_{11}\,\text{Teacher experience}_j \times \text{Gender}_{ij} \\ & + u_{1j}\,\text{Gender}_{ij} + u_{0j} + e_{ij} \qquad (11)\end{aligned}$$

Note that the result of modeling the slopes using the class-level variable implies adding an interaction term and second-level residuals u_{1j} that are related to the pupil-level variable Gender. Model M2 is the most complete, including both available explanatory variables and the cross-level interaction term. The interaction term is significant using the Z test. Because we have used FML estimation, we can also test the interaction term by comparing the deviances of models M1 and M2. The deviance-difference is 15.3, which has a chi-square distribution with one degree of freedom and $p < 0.005$. Using a deviance-difference test on the second-level variance components in model M2, by restricting variance terms to zero and then comparing deviances, leads to the conclusion that all variance terms are significant and that the covariance term is not. This means, that not all residual variation in the intercept and slope can be modeled by the explanatory variables.

The interpretation of model M2 is straightforward. The regression coefficients for both explanatory variables are significant. The regression coefficient for pupil gender is 1.33. Because pupil gender is coded 0 = boy and 1 = girl, this means that on average the girls score 1.33 points higher on the popularity measure. The regression coefficient for teacher experience is 0.11, which means that for each year of experience of the teacher, the average popularity score of the class goes up with 0.11 points. Because there is an interaction term in the model, the effect of 1.33 for pupil gender is the expected effect for teachers with zero experience. The regression coefficient for the cross-level interaction is −0.03, which is small but significant. The negative value means that with experienced teachers, the advantage of being a girl is smaller than expected from the direct effects only. Thus, the difference between boys and girls is smaller with more experienced teachers. A comparison of the other results between the two models shows that the variance component for pupil gender goes down from 0.27 in the direct effects model (M1) to 0.22 in the cross-level model (M2). Hence, the cross-level model explains about 19% of the variation of the slopes for pupil gender.

The significant and quite large variance of the regression slopes for pupil gender implies that we should not interpret the estimated value of 1.33 without considering this variation. In an ordinary regression model, without multilevel structure, the value of 1.33 means that girls are expected to differ from boys by 1.33 points, for all pupils in all classes. In our multilevel model, the regression coefficient for pupil gender varies across the classes and the value of 1.33 is just the expected value across all classes (for teachers with zero experience). The variance of the slope is estimated in model M1 as 0.27. Model M2 shows that part of this variation can be explained by variation in teacher experience. The interpretation of the slope variation is easier when we consider their standard deviation, which is the square root of the variance, or 0.52 in our example data. The varying regression coefficients are assumed to follow a normal distribution. Thus, we may expect 95% of the regression slopes to lie between two standard deviations above or below their average. Given the estimated values of 1.33 (in model M2, for inexperienced teachers) or 0.84 (in model M1, average for all teachers) the vast majority of the classes are expected to have positive slopes for the effect of pupil gender. Figure 1 provides a graphical display of the slope variation, which confirms the conclusion that almost all class slopes are expected to be positive.

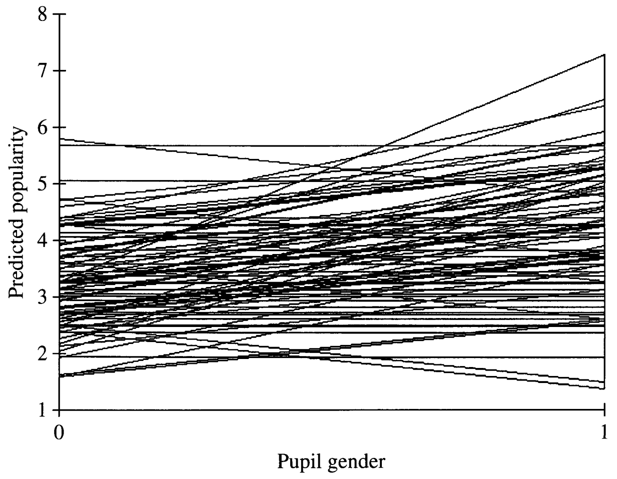

Figure 1 One hundred class slopes for pupil gender.

Analysis of Proportions and Binary Data

The multilevel regression model discussed so far assumes a continuous dependent variable and normal distributions for the residuals. When the response variable is a dichotomous variable or a proportion, both the assumptions of continuous scores and of normality are not met. In addition, the assumption of homoscedastic errors is violated.

The classical approach to the problem of nonnormally distributed variables and heteroscedastic errors is to apply a transformation to achieve normality and reduce heteroskedasticity, followed by a traditional multiple regression analysis. The modern approach to the problem of nonnormally distributed variables is to include the necessary transformation and the choice of the appropriate error distribution (not necessarily a normal distribution) explicitly in the statistical model. This class of statistical models is called generalized linear models. Generalized linear models are defined by three components: (1) a linear regression equation, (2) a specific error distribution, and (3) a link function which is the transformation that links the predicted values for the dependent variable to the observed values. If the link function is the identity function ($f(x) = x$) and the error distribution is normal, the generalized linear model simplifies to the ordinary multiple linear regression model.

Multilevel generalized linear models are described by Raudenbush and Bryk and by Goldstein. Estimating the parameters (regression coefficients and variance components) for such models is more complicated than ordinary multilevel analysis because the likelihood function used in the ML estimation is nonlinear. One approach to estimating such nonlinear models is to linearize the likelihood function. This results in an approximation to the likelihood, and as a result statistical tests based on the likelihood (such as the deviance-difference test) cannot be used. The second approach is to maximize the nonlinear likelihood itself. This is difficult, and therefore it is implemented only in some of the available software and only for a limited set of models. The link functions presently supported in most software are the logistic link function for binary (dichotomous) and binomial data (proportions), the logarithmic function for Poisson data, and the reciprocal link function for γ-distributed data.

The example presented in what follows concerns data from a meta-analysis of studies that compared face-to-face, telephone, and mail surveys on various indicators of data quality. One of these indicators is the response rate—the number of completed interviews divided by the total number of eligible sample units. Overall, the response rates differ among the three data collection methods. In addition, the response rates also differ across studies, which makes it interesting to analyze if study characteristics account for these differences.

These data have a multilevel structure. The lowest level is the condition-level, and the higher level is the study-level. There are three variables at the condition level: the number of completed interviews in that specific condition, the number of eligible respondents in that condition, and an explanatory categorical variable indicating the data collection method used. The categorical data collection variable has three categories: face-to-face, telephone, and mail. It is recoded into two dummy variables: a telephone-dummy and a mail-dummy; this makes the face-to-face condition the reference condition. We use one variable at the study level: the saliency of the questionnaire topic. We have 45 studies in which a total of 99 data collection conditions are compared.

The dependent variable is the response rate. This variable is a proportion—the number of completed interviews divided by the number of eligible respondents. Proportions are analyzed using logistic regression, which is a specific generalized linear model. The link function for binomial data and proportions is the logit function, which is defined as $\text{logit}(x) = \ln[x/(1-x)]$. The corresponding error function is the binomial distribution.

Let P_{ij} be the observed proportion respondents in condition i of study j. Although P_{ij} has a binomial distribution, $\text{logit}(P_{ij})$ has a distribution that is approximately normal, and so we use a linear regression equation at the lowest level. The simplest model, corresponding to the intercept-only model in ordinary multilevel regression analysis, is given by:

$$\text{logit}(P_{ij}) = \beta_{0j} \qquad (12)$$

Note that the usual lowest level error term e_{ij} is not included in Eq. (12). In the binomial distribution, the variance of the observed proportion depends only on the population proportion π. As a consequence, the lowest level variance is determined completely by the predicted value for P_{ij}, and it does not enter the model as a separate term.

The model in Eq. (12) can be extended to include an explanatory variable X_{ij} (e.g., the mail or face-to-face condition) at the condition level:

$$\text{logit}(P_{ij}) = \beta_{0j} + \beta_{1j}X_{ij} \qquad (13)$$

The regression coefficients β are assumed to vary across studies, and this variation is modeled by the study level variable Z_j in the usual second-level regression equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \qquad (14)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j} \qquad (15)$$

By substituting Eqs. (15) and (14) into Eq. (13), we get the single-equation version:

$$\text{logit}(P_{ij}) = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + u_{0j} + u_{1j}X_{ij} \qquad (16)$$

It should be kept in mind that the interpretation of the regression parameters is *not* in terms of the response proportions we want to analyze but instead in terms of the underlying variate defined by the logit transformation $\text{logit}(x) = \ln[x/(1-x)]$. The logit link function is nonlinear and transforms the proportions, which are between 0.00 and 1.00 by definition, into values that range from $-\infty$ to $+\infty$. For a quick examination of the analysis results, we can simply inspect the regression parameters. To understand the implications of the regression coefficients for the proportions we are modeling, we must transform the predicted values back to the original scale or transform the regression coefficients to odds ratios. This problem is not specific to multilevel logistic regression.

Table II presents the results for a sequence of three models: the intercept-only model, a model with the two condition dummies, and a model with the two condition dummies and the study-level variable Saliency. In the

Table II Multilevel Logistic Models for Survey Response

Model	*M0: intercept-only*	*M1: +conditions*	*M2: +saliency*
Fixed part			
Predictor	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	1.02 (0.13)	1.29 (0.14)	0.54 (0.22)
Telephone		−0.21 (0.09)	−0.19 (0.10)
Mail		−0.58 (0.16)	−0.56 (0.15)
Saliency			0.68 (0.17)
Random part			
$\sigma^2_{u_0}$	0.84 (0.17)	0.83 (0.19)	0.63 (0.14)
$\sigma^2_{u_1}$		0.26 (0.07)	0.27 (0.08)
σ_{u_2}		0.60 (0.20)	0.56 (0.18)

intercept-only model, the intercept γ_{00} is estimated as 1.02. As noted before, this refers to the underlying distribution established by the logistic link function and *not* to the proportions themselves. To determine the expected proportion, we must use the inverse transformation for the logistic link function, given by:

$$g(x) = \frac{e^x}{1 + e^x}$$

Using this inverse function, the estimated intercept of 1.02 translates back to an expected proportion of 0.73. The study-level variance is considerable and significant by the Z test. Because the estimation method used here is based on the linearization approach, the deviance is approximate and not available for the deviance-difference test. Hence, in this specific case the significance of the variance components is assessed using the Z test.

The next model adds the condition-level variables Telephone-dummy and Mail-dummy, assuming random regression slopes. In this model, the intercept represents the condition in which both explanatory variables are zero, which is the face-to-face condition. Thus, the value for the intercept in model M1 in Table II estimates the expected response in the face-to-face condition, which is 1.29. This corresponds to an expected response proportion of 0.78. The large negative values for the slope coefficients for the Telephone-dummy and Mail-dummy variables indicate that in these conditions the expected response is lower. To find out how much lower, we must use the regression equation to predict the response in the three conditions and transform these values (which refer to the underlying variate) back to proportions. For the telephone condition, which is coded by Telephone = 1 and Mail = 0, the regression equation reads: $Y = 1.29 - 0.21 = 1.08$, which transforms to an expected response proportion of 0.75. For the mail condition, which is coded by Telephone = 0 and Mail = 1, it reads $Y = 1.29 - 0.58 = 0.71$, which transforms to an expected response proportion of 0.67. The variance components for the regression coefficients are significant by the (approximate) Z test.

The final model includes the study-level variable Saliency. Compared to the earlier results, the regression coefficients are about the same, but the value for the intercept is different. This is not informative, because the intercept almost always changes when other variables are added to or deleted from the regression equation. In our case, including the study-level explanatory variable Saliency in the model causes the shift of the intercept value. Saliency is coded as 1 = very salient, 2 = somewhat salient, and 3 = not salient. The coded values for Saliency do not include the value 0. Hence, the estimated value of the intercept has no meaningful interpretation. The regression coefficient for Saliency is positive, indicating that the response rate increases when the study's topic is more salient.

The last logical step is to introduce interaction variables of Saliency with the two condition variables to model the random coefficients. In our example data, it turns out that this interaction variable does not explain any variation of the regression coefficients.

The random coefficient model leads to another interesting conclusion. In general, telephone and mail surveys obtain a lower response rate than face-to-face surveys. For instance, on the underlying scale, the regression coefficient for Telephone is −0.19 in the final model. However, this regression coefficient has a large variance across studies: $\sigma^2_{11} = 0.27$. The corresponding standard deviation is 0.52. Using the standard normal distribution, we can calculate that in 36% of similarly conducted studies this regression coefficient is expected to be larger than zero! It is instructive to see that, even if there is little doubt that *on the average* the telephone interview has a lower response rate than the face-to-face interview, there is still a chance that in a specific study we will find the opposite relation.

Further Topics

Extensions of the Multilevel Regression Model

The multilevel regression model is one attractive approach to analyzing longitudinal data. In this case, the hierarchical structure is viewed as measurement occasions nested within individuals. Extensions of the multilevel regression model are models for data that are not fully nested, such as cross-classified data, and models in which group membership may not be fully known. The nonlinear regression model discussed in the previous section has been extended to models for ordered or unordered categorical response variables and models for the analysis of counts. Finally, multilevel factor analysis and multilevel structural equation models are becoming available.

Software and Internet Resources

Multilevel analysis modules have appeared in most of the large statistical packages, such as SPSS, SAS, Stata, and SPLUS. Although these modules are quite powerful, specialized software for multilevel tends to have more analysis options and more coverage of the model extensions previously mentioned. The best-known specialized multilevel software are HLM and MlwiN. Don Hedeker provides a set of freeware programs for multilevel regression modeling. A 2001 review of some of these packages was given by De Leeuw and Kreft. The multilevel models project in London maintains a large homepage on multilevel modeling, with emphasis on their own product

MlwiN, but also including much general information. Their website also provides links to other multilevel websites, including one to Don Hedeker's freeware packages. There is also an ongoing review of all software that is able to analyze multilevel data. Finally, there is an active Internet multilevel discussion group.

See Also the Following Articles

Clustering • Internet Measurement • Maximum Likelihood Estimation • Misspecification in Linear Spatial Regression Models

Further Reading

Berkhof, J., and Snijders, T. A. B. (2001). Variance component testing in multilevel models. *J. Educ. Behav. Statist.* **26,** 133–152.

De Leeuw, E. D. (1992). *Data Quality in Mail, Telephone, and Face-to-face Surveys.* TT-Publikaties, Amsterdam.

De Leeuw, J., and Kreft, I. (2001). Software for multilevel analysis. In *Multilevel Modelling of Health Statistics* (A. H. Leyland and H. Goldstein, eds.), pp. 187–204. John Wiley, New York.

Iverson, G. R. (1991). *Contextual Analysis.* Sage, Newbury Park, CA.

Goldstein, H. (2003). *Multilevel Statistical Models.* Arnold, London.

Hosmer, D. W., and Lemeshov, S. (1989). *Applied Logistic Regression.* John Wiley, New York.

Hox, J. J. (2002). *Multilevel Analysis. Techniques and Applications.* Erlbaum, Mahwah, NJ.

Hox, J. J., and de Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face to face surveys. *Quality Quantity* **28,** 329–344.

Kish, L. (1965). *Survey Sampling.* John Wiley, New York.

Maas, C. J. M., and Snijders, T. A. B. (2003). The multilevel approach to repeated measures for complete and incomplete data. *Quality Quantity* **37,** 71–89.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman & Hall, London.

Multilevel Internet Discussion Group. http://www.jiscmail.ac.uk/lists/multilevel.html

Multilevel Models Project. http://www.multilevel.ioe.ac.uk

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models.* Sage, Thousand Oaks, CA.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components.* Wiley, New York.

Snijders, T. A. B., and Bosker, R. (1999). *Multilevel Analysis.* Sage, Thousand Oaks, CA.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.