

## GLOSSARY

# A glossary for multilevel analysis

A V Diez Roux

Multilevel analysis has recently emerged as a useful analytical technique in several fields, including public health and epidemiology. This glossary defines key concepts and terms used in multilevel analysis.

**M**ultilevel analysis, originally developed in the fields of education, sociology, and demography, has received increasing attention in public health and epidemiology over the past few years. This glossary defines key terms and concepts in multilevel analysis. The intent is to provide conceptual explanations of basic concepts, particularly those that are fundamental, that have been used inconsistently or that lend themselves to confusion. Selected terms and concepts more broadly related to the presence of multiple levels of organisation (such as group level variables and inferential fallacies) are also included. Although the glossary often refers to individuals nested within groups, multilevel analysis is applicable to a broad range of situations involving units at a lower level (or micro units) nested within units at a higher level (or macro units) (including for example, persons nested within studies as in meta-analysis, and measures over time nested within individuals as in the analysis of repeat measures). References to terms that have their own specific entry are in SMALL CAPITALS.

### AGGREGATE DATA

Term used to refer to data or variables for a higher level unit (for example, a group) constructed by combining information for the lower level units of which the higher level unit is composed (for example, individuals within the group). Examples of aggregate data include summaries of the properties of individuals comprising a group, for example, the percentage of persons in a neighbourhood with complete high school or the mean income of state residents. Implicit in most uses of the term aggregate data is the idea that aggregate variables are merely summaries of the properties of lower level units and not measures of higher level properties themselves (although this is not necessarily true in all cases, see DERIVED VARIABLES).

### ATOMISTIC FALLACY

The fallacy sometimes present when drawing inferences regarding variability across groups (or the relation between group level variables) based on individual level data, or more generally, the fallacy of drawing inferences regarding variability across units defined at a higher level based on data collected for units at a lower level. The atomistic fallacy arises because associations between

*J Epidemiol Community Health* 2002;**56**:588–594

two variables at the individual level may differ from associations between analogous variables measured at the group level. For example, a study of individuals may find that increasing individual level income is associated with decreasing coronary heart disease mortality. If it is inferred from these data that at the country level, increasing per capita income is associated with decreasing coronary heart disease mortality, the researcher may be committing the atomistic fallacy (because across countries, increasing per capita income may actually be associated with *increasing* coronary heart disease mortality). The sources of the atomistic fallacy are similar to those of the ECOLOGIC FALLACY. In the atomistic fallacy, the conceptual model being tested corresponds to the higher level, but the data are collected for a lower level.<sup>1,2</sup> The atomistic fallacy has sometimes been referred to as the INDIVIDUALISTIC FALLACY.<sup>3,4</sup>

### COMPOSITIONAL EFFECTS

When inter-group (or inter-context) differences in an outcome (for example, disease rates) are attributable to differences in group composition (that is, in the characteristics of the individuals of which the groups are comprised) they are said to result from compositional effects.<sup>5</sup> On the other hand, when group differences are attributable to the effects of GROUP LEVEL VARIABLES or properties, they are said to result from CONTEXTUAL EFFECTS.

### CONTEXTUAL ANALYSIS

An analytical approach originally used in sociology to investigate the effect of collective or group characteristics on individual level outcomes.<sup>4,6,7</sup> In contextual analysis, group level predictors (often constructed by aggregating the characteristics of individuals within groups) are included together with individual level variables in standard regressions with individuals as the units of analysis (CONTEXTUAL EFFECTS MODELS). This approach permits the simultaneous examination of how individual level and group level variables are related to individual level outcomes. It thus allows for macro processes that are presumed to have an impact on individuals over and above the effects of individual level variables.<sup>6</sup> The terms “contextual analysis” and MULTILEVEL ANALYSIS have sometimes been used synonymously,<sup>8–10</sup> and both approaches are similar in allowing the investigation of how group level (or macro) and individual level (or micro) variables (as well as their interactions) are related to individual level outcomes. However, MULTILEVEL MODELS are more general than the original contextual models in that (1) they allow (and account for) the possibility of residual correlation between individuals within groups; and (2) they allow examination of between group variability and the factors associated with it. In contrast,

Correspondence to:  
Dr A V Diez Roux, Division  
of General Medicine, 622  
West 168th Street, PH9  
East Rm 105, New York  
NY 10032, USA;  
ad290@columbia.edu

contextual models often do not account for residual correlation (although they can be modified to do so) and do not allow the examination of inter-group variability or of the factors associated with it (see also **VARIANCE COMPONENTS**).

### CONTEXTUAL EFFECTS

Term generally used to refer to the effects of variables defined at a higher level (usually at the group level) on outcomes defined at a lower level (usually at the individual level) after controlling for relevant individual level (lower level) confounders. The term is most often used to refer to the effect of a **DERIVED GROUP LEVEL VARIABLE** (for example, mean neighbourhood income) on an individual level outcome (such as blood pressure) after controlling for its individual level namesake (for example, individual level income).<sup>6, 11</sup> However, “contextual effects” is also sometimes used to refer to the effects of group level variables generally be they **DERIVED VARIABLES** or **INTEGRAL VARIABLES**, and can apply to any situation involving lower level units nested within higher level units (for example, contextual effects of country characteristics on disease rates for small areas, contextual effects of tissue characteristics on cell biology). Contextual effects are sometimes contrasted with **COMPOSITIONAL EFFECTS**.<sup>5</sup>

### CONTEXTUAL EFFECTS MODELS

Regression models with individuals as the units of analysis that include both group level and individual level variables as predictors of individual level outcomes. Traditional contextual effects models are equivalent to multilevel models in which all coefficients are modelled as fixed (that is, no error terms are included in the group level or level 2 equations, see **MULTILEVEL MODELS**). See **CONTEXTUAL ANALYSIS**.

### CONTEXTUAL VARIABLES

See **DERIVED VARIABLES** and **GROUP LEVEL VARIABLES**.

### CROSS LEVEL EFFECTS

Term used to refer to the main effects of higher level variables (for example, group level variables) on outcomes at a lower level (for example, individual level outcomes) as well as to modifications of the effects of lower level (individual level) variables by higher level (group level) variables (see **CROSS LEVEL INTERACTION**).<sup>12</sup> Examples include the effect of country level income inequality on individual level self reported health (effect of a higher level variable on outcomes at a lower level), and the presence of stronger associations between individual level income and self reported health in the presence of high country level income inequality (modifications of the effects of lower level variables by higher level variables). The term “ecological effects” has sometimes been used as a synonym for “cross level effects”.<sup>12</sup>

### CROSS LEVEL INFERENCE

The drawing of inferences regarding factors associated with variability in the outcome at one level based on data collected at another level (for example, drawing inferences regarding relations between individual level variables based on group level associations, or vice versa). See **ECOLOGIC FALLACY** and **ATOMISTIC FALLACY**.

### CROSS LEVEL INTERACTION

Refers to the interaction between higher level and lower level variables—that is, to modification of the effects of lower level variables by characteristics of the higher level units to which the lower level units belong (or vice versa).<sup>5, 12</sup> For example, if the relation between individual level income and blood pressure differs by neighbourhood characteristics (that is, neighbourhood and individual level variables interact), there is said to be a cross level interaction. In multilevel models

whenever group specific estimates of the effect of a lower level variable are modelled as a function of higher level (group level) variables (as in equation (3) under the entry for **MULTILEVEL MODELS**), a cross level interaction appears in the final model ( $\gamma_{it}, C_j, I_{ij}$  in equation (4) under **MULTILEVEL MODELS**).

### DERIVED VARIABLES

A type of **GROUP LEVEL VARIABLE** constructed by mathematically summarising the characteristics of individuals in the group (for example, means, proportions, or measures of dispersion, such as, percentage of persons with incomplete high school, mean income, standard deviation of the income distribution).<sup>11, 13</sup> Some derived variables have no individual level analogue (for example, standard deviation of the income distribution) and therefore necessarily refer to group level constructs. Others (for example, mean neighbourhood income) do have individual level analogues (for example, individual level income), but may provide information on group level constructs, distinct from their individual level namesake. The mean of the dependent variable in the group (for example, proportion infected in a study of the causes of infection) can be thought of as a special type of derived variable.<sup>14</sup> Although derived and **INTEGRAL VARIABLES** are sometimes presented as conceptually distinct, they are closely interrelated. Derived variables often operate by shaping certain integral properties of the group. For example, the composition of a group may influence the predominant types of interpersonal contacts, values, and norms or may shape organisations or regulations within the group that affect all members.<sup>15</sup> The terms “analytical variables” and “aggregate variables” have been used as synonyms for “derived variables”. The term “contextual variables” has also been used as a synonym for “derived variables”<sup>14</sup> although it is sometimes used to refer to **GROUP LEVEL VARIABLES** generally.<sup>6, 13</sup>

### ECOLOGICAL FALLACY

The fallacy sometimes present when drawing inferences at the individual level (that is, regarding relations between individual level variables) based on group level data. The ecological fallacy arises because associations between two variables at the group level (or ecological level) may differ from associations between analogous variables measured at the individual level. These differences between individual level and group level associations were first described for correlation coefficients<sup>16</sup> but may also be present for other measures of association such as regression coefficients.<sup>11, 17</sup> More generally, the fallacy may occur whenever data for units at a higher level are used to draw inferences regarding factors associated with variability across units at a lower level—that is, when the conceptual model being tested corresponds to the lower level, but the data are collected for a higher level.<sup>1, 2</sup> Suppose a researcher finds that at the country level, increasing per capita income is associated with increasing mortality attributable to traffic accidents. If he/she infers that at the individual level, increasing personal income is associated with increasing motor vehicle related mortality, she may be committing the ecological fallacy, because within countries, motor vehicle related mortality may always be lower in high income than in low income persons. In the case of regression coefficients, the sources of the ecological fallacy include (1) the lack of information on constructs pertaining to a lower level of organisation; and (2) the failure to realise that a variable defined and measured at one level of organisation may tap into a different construct than its namesake at another level.<sup>18</sup>

### EMPIRICAL BAYES ESTIMATES

Estimates of parameters for a given group or higher level unit (for example, estimates of group specific intercepts or slopes, such as  $b_{0j}$  and  $b_{1j}$  in equation (1), under **MULTILEVEL MODELS**)

obtained by combining information from the group itself with information from other similar groups investigated.<sup>10 19 20</sup> This is particularly useful when estimating parameters for a group with few within group observations. These estimates are “optimally” weighted averages that combine information derived from the group itself with the mean for all similar groups. The weighted average shifts the group specific estimate (derived using data only for that particular group) towards the mean for similar groups. The less precise the group specific estimate and the less the variability observed across groups, the greater the shift towards the overall group mean. Thus, the estimate for a given group is based not only on its own data but also takes into account estimates for other groups and the characteristics groups share.<sup>20</sup> Empirical Bayes estimates of parameters for a given group can be derived from multilevel models using estimates of the group level errors (for example,  $U_{0j}$  and  $U_{1j}$ , see MULTILEVEL MODELS) for that particular group. Empirical Bayes estimates are also sometimes referred to as “shrinkage estimates” because they “shrink” the group specific estimate towards the overall mean (although in fact when the overall mean is greater than the group specific estimate, the “shrunk” or empirical Bayes estimate may actually be greater than the group specific estimate). In public health, empirical Bayes estimation can be used, for example, to derive improved estimates of rates of death or diseases for small areas with few observations,<sup>21</sup> or to estimate rates of different health outcomes for individual providers (hospitals, physicians, etc).<sup>22</sup> In other applications (which do not involve the structure of individuals within groups although they are analogous to it), empirical Bayes estimates of regression coefficients have been used to obtain improved estimates of associations in studies investigating the role of multiple exposures.<sup>23</sup>

### ENVIRONMENTAL VARIABLES

In the context of ecological studies and multilevel analysis, the term “environmental variables” has sometimes been used to refer to group level measures of physical or chemical exposures. Environmental variables, so defined, have been proposed as a “type” of GROUP LEVEL VARIABLE, distinct from DERIVED VARIABLES and INTEGRAL VARIABLES.<sup>11</sup> These variables are not derived by aggregating the characteristics of individuals but they do have group level and individual level analogues (for example, days of sunlight in the community and individual level sunlight exposure information). In contrast with derived and integral variables, which may be used as indicators of group level constructs, group level environmental variables are used exclusively as proxies for individual level exposures (which may be more difficult to measure for logistic or methodological reasons), rather than as indicators of a group level property, which is conceptually different from the analogous measure at the individual level.

### FIXED EFFECTS/FIXED COEFFICIENTS

Regression coefficients (intercepts or covariate effects) that are not allowed to vary randomly across higher level units (see MULTILEVEL MODELS). For example, in the case of persons nested within neighbourhoods, two options are available for modelling the effects of neighbourhood. One option is to include a dummy variable for each neighbourhood. In this case the neighbourhood coefficients are modelled as fixed (sometimes called “fixed effects”). Another option is to assume that the neighbourhoods in the sample are a random sample of a larger population of neighbourhoods and that the coefficients for the “neighbourhood effect” vary randomly around an overall mean (for example, as reflected by  $U_{0j}$  in equation 2 under the entry for MULTILEVEL MODELS). In this case, the neighbourhood effects are modelled as random (sometimes called “random effects”, see RANDOM EFFECTS MODELS). In the same example, the coefficients for individual level covariates can also be modelled

as fixed or random. For example, if the relation between individual level income and blood pressure is not allowed to vary randomly across neighbourhoods, the coefficient for individual level income is fixed (“fixed coefficient”). On the other hand, if the coefficient for individual level income is allowed to vary randomly across neighbourhoods around an overall mean effect (as reflected by  $U_{1j}$  in equation 3 under the entry for MULTILEVEL MODELS), the coefficient for income is modelled as random (sometimes called a “random coefficient”, see RANDOM COEFFICIENT MODELS). Although the terms “fixed effects” and “fixed coefficients” are sometimes distinguished as noted above, they are often used interchangeably. Fixed effects models or fixed coefficient models are models in which all effects or coefficients are fixed. See also RANDOM EFFECTS/RANDOM COEFFICIENTS.

### GROUP LEVEL VARIABLES

Term used to refer to variables that characterise groups. The terms group level variables, macro variables and ecological variables are often used interchangeably.<sup>2 6 11 14 24</sup> Group level variables may be used as proxies for unavailable or unreliable individual level data (for example, when neighbourhood mean income is used as a proxy for the individual level income of individuals living in the neighbourhood) or as indicators of group level constructs (for example, when mean neighbourhood income is used as an indicator of neighbourhood characteristics that may be related to individual level outcomes independently of individual level income). It is the second usage (as indicators of group level constructs) that is of particular interest in multilevel analysis. Group level variables have been classified into two basic types.<sup>11 13 24</sup> DERIVED VARIABLES and INTEGRAL VARIABLES. Two additional types of group level variables, STRUCTURAL VARIABLES<sup>13</sup> and ENVIRONMENTAL VARIABLES<sup>11</sup> are sometimes distinguished. The term contextual variables has been used as a synonym for group level variables generally<sup>6 13</sup> although it is sometimes reserved for derived group level variables.<sup>11 14</sup>

### HIERARCHICAL (LINEAR) MODELS

See MULTILEVEL MODELS

### INDIVIDUAL LEVEL VARIABLES

Term used to refer to variables that characterise individuals and refer to individual level constructs (for example, age or personal income).

### INDIVIDUALISTIC FALLACY

Term used as a synonym for the ATOMISTIC FALLACY. May sometimes also be used as a synonym for the PSYCHOLOGISTIC FALLACY.

### INTEGRAL VARIABLES

A type of GROUP LEVEL VARIABLE. Integral variables differ from DERIVED VARIABLES (another type of group level variable) in that they are not summaries of the characteristics of individuals in the group. Integral variables have no individual level analogues and necessarily refer to group level constructs. Examples of integral variables include the existence of certain types of laws, political or economic system, social disorganisation, or population density.<sup>11 13</sup> Integral variables have also been referred to as primary or global variables.

### INTRAClass CORRELATION

A measure of the degree of resemblance between lower level units belonging to the same higher level unit or cluster.<sup>25</sup> In the case of individuals nested within groups (for example, neighbourhoods), the intraclass correlation measures the extent to which values of the dependent variable are similar for individuals belonging to the same group. It can be thought

of as the average correlation between values of two randomly drawn lower level units (for example, individuals) in the same, randomly drawn higher level unit (for example, neighbourhood). It can also be defined as the proportion of the variance in the outcome that is between the groups or higher level units. In the case of a simple random intercept model, the intraclass correlation coefficient is estimated by the ratio of population variance between groups ( $\tau_{00}$ ) to the total variance ( $\tau_{00} + \sigma^2$ ).<sup>25</sup> (see MULTILEVEL MODELS) The estimation of the intraclass correlation coefficient in models including random covariate effects, or in the case of non-normally distributed dependent variables, is more complex and not always straightforward.

### MARGINAL MODELS

See POPULATION-AVERAGE MODELS.

### MIXED MODELS

Term used to refer to models that contain a mixture of FIXED EFFECTS (or fixed coefficients) and RANDOM EFFECTS (or random coefficients). In mixed models some of the regression coefficients (intercepts or covariate effects) are allowed to vary randomly across higher level units but others are not (see MULTILEVEL MODELS). Thus mixed models can be thought of as a particular case of the more general multilevel models (although the term is also occasionally used as a synonym of multilevel models generally). Sometimes the term mixed models is also used to encompass models that account for correlation between lower level units (for example, individuals) within higher level units (for example, neighbourhoods) in other ways—that is, by modelling the correlations or covariances themselves rather than by allowing for random effects or random coefficients.<sup>26</sup> These models (which are not multilevel models) have also been called covariance pattern models,<sup>26</sup> marginal models, or POPULATION AVERAGE MODELS.

### MULTILEVEL ANALYSIS

An analytical approach that is appropriate for data with nested sources of variability—that is, involving units at a lower level or micro units (for example, individuals) nested within units at a higher level or macro units (for example, groups such as schools or neighbourhoods).<sup>5 10 19 24 25 27–30</sup> Multilevel analysis allows the simultaneous examination of the effects of group level and individual level variables on individual level outcomes while accounting for the non-independence of observations within groups. Multilevel analysis also allows the examination of both between group and within group variability as well as how group level and individual level variables are related to variability at both levels. Thus, multilevel models can be used to draw inferences regarding the causes of inter-individual variation (or the relation of group and individual level variables to individual level outcomes) but inferences can also be made regarding inter-group variation, whether it exists in the data, and to what extent it is accounted for by group and individual level characteristics. In multilevel analysis, groups or contexts are not treated as unrelated but are conceived as coming from a larger population of groups about which inferences want to be made. Multilevel analysis thus allows researchers to deal with the micro-level of individuals and the macro-level of groups or contexts simultaneously.<sup>5</sup>

Multilevel analysis has a broad range of applications in many situations involving nested sources of random variability such as persons nested within neighbourhoods,<sup>5 30</sup> patients nested within providers,<sup>31</sup> meta analysis (observations nested within sites)<sup>19 32</sup> longitudinal data analysis (repeat measurements over time nested within persons),<sup>28 33 34</sup> multivariate responses (multiple outcomes nested within individuals),<sup>5</sup> the analysis of repeat cross sectional surveys (multiple observations nested within time periods),<sup>35</sup> the examination of

geographical variations in rates (rates for smaller areas nested within regions or larger areas)<sup>36</sup> and the examination of interviewer effects (respondents nested within interviewers).<sup>37</sup> Multilevel analysis can also be used in situations involving multiple nested contexts<sup>19 28</sup> (for example, multiple measures over time on individuals nested within neighbourhoods) as well as overlapping or cross classified contexts (for example, children nested within neighbourhoods and schools).<sup>38</sup> The statistical models used in multilevel analysis are referred to as MULTILEVEL MODELS<sup>25 28 29</sup> or hierarchical linear models.<sup>19 39</sup>

### MULTILEVEL MODELS

The statistical models used in MULTILEVEL ANALYSIS.<sup>19 25 28 29</sup> The terms “hierarchical models” and “multilevel models” are often used synonymously. These models (or variants of them) have previously appeared in different literatures under a variety of names including RANDOM EFFECTS MODELS OF RANDOM COEFFICIENT MODELS<sup>40–42</sup> “covariance components models” or “variance components models”,<sup>43 44</sup> and MIXED MODELS.<sup>26</sup> A simplified example for the case of a normally distributed dependent variable, a single individual level (lower level unit) predictor and a single group level (higher level unit) predictor is provided below. Analogous models can be formulated for non-normally distributed dependent variables.<sup>10 28 39 45</sup>

In the case of multilevel analysis involving two levels (for example, individuals nested within groups), the multilevel model can be conceptualised as a two stage system of equations.

In the first stage (level 1), a separate individual level regression is defined for each group or higher level unit.

$$(1) Y_{ij} = b_{0j} + b_{1j}I_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$Y_{ij}$  = outcome variable for  $i^{\text{th}}$  individual in  $j^{\text{th}}$  group

$I_{ij}$  = individual level variable for  $i^{\text{th}}$  individual in  $j^{\text{th}}$  group

$b_{0j}$  is the group specific intercept

$b_{1j}$  is the group specific effect of the individual level variable

Individual level errors ( $\varepsilon_{ij}$ ) are assumed to be independent and identically distributed with a mean of 0 and a variance of  $\sigma^2$ . The same regressors are generally used in all groups, but regression coefficients ( $b_{0j}$  and  $b_{1j}$ ) allowed to vary from one group to another.

In a second stage (level 2), each of the group or context specific regression coefficients defined in equation (1) ( $b_{0j}$  and  $b_{1j}$  in this example) are modelled as a function of group level (or higher level) variables.

$$(2) b_{0j} = \gamma_{00} + \gamma_{01}G_j + U_{0j} \quad U_{0j} \sim N(0, \tau_{00})$$

$$(3) b_{1j} = \gamma_{10} + \gamma_{11}G_j + U_{1j} \quad U_{1j} \sim N(0, \tau_{11}) \\ \text{cov}(U_{0j}, U_{1j}) = \tau_{10}$$

$G_j$  group level variable

$\gamma_{00}$  is the common intercept across groups

$\gamma_{01}$  is the effect of the group level predictor on the group specific intercepts

$\gamma_{10}$  is the common slope associated with the individual level variable across groups

$\gamma_{1i}$  is the effect of the group level predictor on the group specific slopes

The errors in the level 2 equations ( $U_{0j}$  and  $U_{1j}$ ), sometimes called “macro errors”, are assumed to be normally distributed with mean 0 and variances  $\tau_{00}$  and  $\tau_{11}$  respectively.  $\tau_{01}$  represents the covariance between intercepts and slopes. Thus, multilevel analysis summarises the distribution of the group specific coefficients in terms of two parts: a “fixed” part that is common across groups ( $\gamma_{00}$  and  $\gamma_{01}$  for the intercept, and  $\gamma_{10}$  and  $\gamma_{11}$  for the slope) and a “random” part ( $U_{0j}$  for the intercept and  $U_{1j}$  for the slope) that is allowed to vary from group to group (see also FIXED COEFFICIENTS and RANDOM COEFFICIENTS).

By including an error term in the group level equations (equations (2) and (3)), these models allow for sampling variability in the group specific coefficients ( $b_{0j}$  and  $b_{1j}$ ) and also for the fact that the group level equations are not deterministic (that is, the possibility that not all relevant macro-level variables have been included in the model). The underlying assumption is that group specific intercepts and slopes are random samples from a normally distributed population of group specific intercepts and slopes, or alternatively, that the macro errors are exchangeable—that is, that the residual variation in group specific coefficients across groups is unsystematic.<sup>10</sup>

An alternative way to present the model fitted in multilevel analysis is to substitute equations (2) and (3) in (1) to obtain:

$$Y_{ij} = \gamma_{00} + \gamma_{01}C_j + \gamma_{10}I_{ij} + \gamma_{11}C_jI_{ij} + U_{0j} + U_{1j}I_{ij} + \varepsilon_{ij}$$

The model includes the effects of group level variables ( $\gamma_{01}$ ), individual level variables ( $\gamma_{10}$ ) and their interaction ( $\gamma_{11}$ ) on the individual level outcome  $Y_{ij}$ . These coefficients ( $\gamma_{01}$ ,  $\gamma_{10}$  and  $\gamma_{11}$ ), which are common to all individuals regardless of the group to which they belong are often called the FIXED COEFFICIENTS (OR fixed effects). The model also includes a random intercept component ( $U_{0j}$ ), and a random slope component ( $U_{1j}$ ). The values of these components vary randomly across groups, and hence  $U_{0j}$  and  $U_{1j}$  referred to as the RANDOM COEFFICIENTS (OR random effects). The parameters of the above equations (fixed effects, random effects, variances of the random effects, and residual variance) are simultaneously estimated using iterative methods. The level 1 and level 2 variances ( $\sigma^2$ ,  $\tau_{00}$ ,  $\tau_{11}$ ,  $\tau_{01}$ ) are called the (CO)VARIANCE COMPONENTS.

Many variants of the more general model illustrated above are possible. For example, only group specific intercepts ( $b_{0j}$ ) may be modelled as random (these models have also been called RANDOM EFFECTS MODELS). When covariate effects ( $b_{1j}$  in the example above) are modelled as random these models have also been called RANDOM COEFFICIENT MODELS. When some of the coefficients are fixed and other are random these models have also been called “mixed effects models” or simply MIXED MODELS. When all coefficients are modelled fixed (no random errors are included in level 2 equations) these models are reduced to traditional CONTEXTUAL EFFECTS MODELS. Multilevel models can also account for multiple nested contexts (or levels)<sup>19–28</sup> allowing fixed and random coefficients to be associated with variables measured at different levels of the data hierarchy being analysed. Multilevel models can also be modified to allow for non-hierarchical, overlapping or cross classified contexts (for example, children simultaneously nested within neighbourhoods and schools).<sup>38</sup>

### NON-INDEPENDENCE OF OBSERVATIONS

Refers to situations in which dependent variables for observations at a lower level nested within the same higher level unit (or cluster) are correlated, even after measured characteristics

are taken into account. For example, two persons from the same neighbourhood may tend to have more similar blood pressure levels than two persons from different neighbourhoods, even after measured individual and neighbourhood characteristics are taken into account. In the case of repeat measures on individuals over time, two blood pressure measurements on the same person may tend to be more similar than two measures on different persons even after relevant covariates are taken into account. One reason for this correlation may have to do with the omission of important higher level variables that observations within the same higher level unit share. This residual correlation violates the assumption of independence of observations underlying usual regression approaches. Ignoring this correlation may lead to incorrect inferences. Efficiency of estimation may also be reduced.<sup>40</sup> Multilevel models account for potential residual correlation by modelling intercepts and regression coefficients as random (for example, by allowing for macro level errors,  $U_{0j}$  and  $U_{1j}$  in second level equations, see MULTILEVEL MODELS).

### POPULATION-AVERAGE MODELS

Models that account for correlation between lower level units within higher level units (or clusters) by modelling the correlations or covariances themselves rather than by allowing for random effects or random coefficients as MULTILEVEL MODELS do.<sup>40–46</sup> These correlations are taken into account in the estimation of regression coefficients and their standard errors. Different correlation structures (describing within cluster or within higher level unit correlations) can be specified. “Population-average models” are also referred to as “marginal models”<sup>40–46</sup> or “covariance pattern models”.<sup>26</sup> Whereas multilevel models model the dependent variable conditional on the random effects (or random coefficients), population-average models model the marginal expectation of the dependent variables across the population (in a sense, “averaged” across the random effects). For this reason, marginal models have also been called “population-average” models (as a way to contrast them with SUBJECT SPECIFIC random effects models).<sup>46</sup> The Generalised Estimating Equation (GEE) approach is one approach to fitting marginal models.<sup>46</sup>

Population-average models model the population-average response as a function of covariates without explicitly accounting for heterogeneity across higher level units.<sup>46</sup> In contrast, MULTILEVEL MODELS investigate and explain the source of group to group variation (and of the within group correlation) by modelling group specific regression coefficients as a function of group level variables plus random variation. Therefore, although population-average models account for the correlation between outcomes within higher level units, the source of this correlation is not directly investigated (the correlation, and sometimes higher level effects themselves, are viewed as nuisance parameters that must be taken into account but are not of direct interest). Therefore, population-average models do not allow examination of group to group variation, of the group level or individual level variables related to it, or of the degree of variation present between and within groups, as multilevel models do (see VARIANCE COMPONENTS). Differences between both types of models also have consequences for the interpretation of regression coefficients: in the multilevel model, the regression coefficient estimates how the response changes as a function of covariates conditional on the random effects; in the marginal model, the coefficient expresses how the response changes as a function of covariates “averaged” over group to group heterogeneity (or group random effects).<sup>40–46</sup> In the case of continuous dependent variables these coefficients are mathematically equivalent, but in the case of non-normally distributed variables (for example, logistic models) the marginal parameter values will usually be smaller in absolute value than their random effects analogues.<sup>46–47</sup>

### PSYCHOLOGISTIC FALLACY

An inferential fallacy that may arise from the failure to consider group characteristics in drawing inferences regarding the causes of variability across individuals<sup>1,2</sup>—that is, assuming that individual level outcomes can be explained exclusively in terms of individual level characteristics. Although the level at which data are collected may fit the conceptual model being investigated (that is, individual level), important facts pertaining to other levels (that is, group level) may have been ignored.<sup>1,2</sup> For example, a study based on individuals might find that immigrants are more likely to develop depression than natives. But suppose this is only true for immigrants living in communities where they are a small minority. A researcher ignoring the contextual effect of community composition might attribute the higher overall rate in immigrants to the psychological effects of immigration or to genetic factors, ignoring the importance of community level factors and thus committing the psychologistic fallacy.<sup>1</sup> The term “psychologistic fallacy” is not entirely appropriate because the individual level factors used to explain the outcome are not always exclusively psychological.<sup>2</sup> Although the term “individualistic fallacy” may appear more adequate, it has also been used as a synonym for the related but distinct ATOMISTIC FALLACY.<sup>3,4</sup> See also SOCIOLOGISTIC FALLACY.

### RANDOM COEFFICIENT MODELS

Term originally used for models in which the regression coefficients corresponding to covariates in the model are treated as random rather than fixed<sup>19,26</sup> (that is, models containing RANDOM COEFFICIENTS, see for example  $b_{ij}$  in the entry for MULTILEVEL MODELS). Traditional random coefficient models do not include higher level (or group level) predictors in the group level equations for the covariate effects (that is, in a traditional random coefficient model, equation (3) would be  $b_{ij} = \gamma_{i0} + U_{ij}$ ).<sup>19</sup> Thus random coefficient models can be thought of as a particular case of the more general MULTILEVEL MODELS. However, the term random coefficient models is sometimes used more broadly to refer to multilevel models generally. See also RANDOM EFFECTS MODELS.

### RANDOM EFFECTS/RANDOM COEFFICIENTS

Regression coefficients (intercepts or covariate effects) that are allowed to vary randomly across higher level units (that is, are assumed to be realisations of values from a probability distribution) (see MULTILEVEL MODELS). For example, in the case of persons nested within neighbourhoods, neighbourhood effects can be assumed to vary randomly around an overall mean (random effect, see RANDOM EFFECTS MODELS). Similarly, the effect of personal income on individual health may be allowed to vary randomly across neighbourhoods (random coefficient, see RANDOM COEFFICIENT MODELS). Although the terms “random effects” and “random coefficients” are sometimes distinguished as noted above, they are often used interchangeably. The use of random effects or random coefficients is especially appropriate when the higher level units (or groups) can be thought of as random samples from a larger population of units (or groups) about which inferences wish to be made. See also FIXED EFFECTS/FIXED COEFFICIENTS.

### RANDOM EFFECTS MODELS

Term originally used for models in which differences across groups (or other classification system) are treated as random rather than fixed<sup>19,26</sup> (that is, models containing RANDOM EFFECTS). For example, in the case involving individuals nested within neighbourhoods, a model treating neighbourhood differences as fixed would include all neighbourhoods represented in the sample as a set of dummy variables in a regression equation with individuals as the units of analysis (see FIXED COEFFICIENTS). In contrast, a random effects model

would treat neighbourhood differences as realisations from a probability distribution—that is, neighbourhood intercepts would be allowed to vary randomly across neighbourhoods following a probability distribution (see MULTILEVEL MODELS). An underlying assumption is that the neighbourhoods in the study are a random sample from a larger population of neighbourhoods about which inferences wish to be made. Random effects models can be thought of as a particular case of the more general MULTILEVEL MODELS in which only intercepts are allowed to vary randomly across groups (that is, random intercept models). Sometimes, however, the term random effects models is used more broadly to refer to MULTILEVEL MODELS generally (that is, models that allow for both random intercept and random covariate effects). See also RANDOM COEFFICIENT MODELS.

### RESIDUAL CORRELATION

See NON-INDEPENDENCE OF OBSERVATIONS.

### SOCIOLOGISTIC FALLACY

An inferential fallacy that may arise from the failure to consider individual level characteristics in drawing inferences regarding the causes of variability across groups.<sup>1,2</sup> Although the level at which data are collected may fit the conceptual model being investigated (that is, group level), important facts pertaining to other levels (that is, the individual level) may have been ignored.<sup>1</sup> Suppose a researcher finds that communities with higher rates of transient population have higher rates of schizophrenia, and he/she concludes that higher rates of transient population lead to social disorganisation, breakdown of social networks, and increased risk of schizophrenia among all community inhabitants. But suppose that schizophrenia rates are only increased for transient residents (because transient residents tend to have fewer social ties, and individuals with few social ties are at greater risk of developing schizophrenia). That is, rates of schizophrenia are high for transient residents and low for non-transient residents, regardless of whether they live in communities with a high or a low proportion of transient residents. If this is the case, the researcher would be committing the sociologistic fallacy in attributing the higher schizophrenia rates to social disorganisation affecting all community members rather than to differences across communities in the percentage of transient residents. See also PSYCHOLOGISTIC FALLACY.

### STRUCTURAL VARIABLES

A type of GROUP LEVEL VARIABLE that refers to relations or interactions between members of a group,<sup>13</sup> for example, characteristics of social networks within the group or patterns of contacts or interactions between members of the group. Structural variables are sometimes considered a subtype of INTEGRAL VARIABLES.<sup>12,18</sup>

### SUBJECT SPECIFIC MODELS

Term used to refer to RANDOM EFFECTS/RANDOM COEFFICIENT MODELS (OF MULTILEVEL MODELS generally) in order to contrast them with POPULATION-AVERAGE models. “Subject specific” is used because the term was originally developed in the context of longitudinal data analysis,<sup>36</sup> where individuals or subjects are the higher level units and repeat measures are the lower level units. In this case, the fixed effects coefficients derived from a random effects, random coefficient, or multilevel model are conditional on person level (or person specific) random effects, hence the term “subject specific”. More generally, they can be thought of as “higher level unit” specific (or cluster specific), because they are conditional or higher level unit (or cluster specific) random effects. For example, in the entry for MULTILEVEL MODELS, the estimate of  $\gamma_{01}$  is conditional on group level random effects (as reflected by the presence of  $U_{0j}$  and  $U_{1j}$ ).

## VARIANCE COMPONENTS

Using multilevel models the total variance in individual level outcomes (or lower level outcomes generally) can be decomposed into variance within and between groups (or higher level units generally). For example, the variance in blood pressure across individuals can be decomposed into variance within and between neighbourhoods. These components are referred to as variance components. The ability to estimate the variance components (which provide important information on the variability in the outcome between and within groups) is a key feature of multilevel models, and what distinguishes multilevel models from traditional CONTEXTUAL EFFECTS MODELS and POPULATION-AVERAGE MODELS. For this reason, multilevel models have also sometimes been referred to as variance component or covariance component models. See also MULTILEVEL MODELS.

## REFERENCES

- Riley MW. Special problems of sociological analysis. In: *Sociological research I: a case approach*. New York: Harcourt, Brace, and World, 1963:700–25.
- Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *Am J Public Health* 1998;**88**:216–22.
- Alker HR. A typology of ecological fallacies. In: Dogan M, Rokkam S, eds. *Social ecology*. Boston: The MIT Press, 1969:69–86.
- Scheuch EK. Social context and individual behavior. In: Dogan M, Rokkam S, eds. *Social ecology*. Boston: The MIT Press, 1969:133–55.
- Duncan C, Jones K, Moon G. Context, composition, and heterogeneity: using multilevel models in health research. *Soc Sci Med* 1992;**46**:97–117.
- Blalock HM. Contextual-effects models: theoretical and methodological issues. *Ann Rev Sociol* 1984;**10**:353–72.
- Iversen G. *Contextual analysis*. Newbury Park, CA: Sage, 1991.
- Hermalin A. The multilevel approach: theory and concepts. The methodology for measuring the impact of family planning programs on fertility. *Population studies Addendum Manual IX*. New York: United Nations, 1986;**66**:15–31.
- van den Eeden P, Huttner HJ. Multi-level research. *Curr Sociol* 1982;**30**:1–178.
- DiPrete TA, Forristal JD. Multilevel models: methods and substance. *Annu Rev Sociol* 1994;**20**:331–57.
- Morgenstern H. Ecologic studies in epidemiology: concepts, principles, and methods. *Annu Rev Public Health* 1995;**16**:61–81.
- Blakely TA, Woodward AJ. Ecological effects in multi-level studies. *J Epidemiol Community Health* 2000;**54**:367–74.
- Lazarsfeld PF, Menzel H. On the relation between individual and collective properties. In: Etzioni A, ed. *A sociological reader on complex organizations*. New York: Holt, Rinehart, and Winston, 1971:499–516.
- Susser M. The logic in ecological: I. The logic of analysis. *Am J Public Health* 1994;**84**:825–9.
- Valkonen T. Individual and structural effects in ecological research. In: Dogan M, Rokkam S, eds. *Social ecology*. Boston: The MIT Press, 1969:53–68.
- Robinson WS. Ecological correlations and the behavior of individuals. *Am Sociol Rev* 1950;**15**:351–7.
- Piantadosi S, Byar DP, Green SN. The ecological fallacy. *Am J Epidemiol* 1988;**127**:893–904.
- Diez-Roux AV, Schwartz S, Susser E. Ecologic studies and ecologic variables in public health research. In: *The Oxford textbook of public health*. Volume 2. London: Oxford University Press, 2002:493–508.
- Bryk AS, Raudenbush SW. *Hierarchical linear models: applications and data analysis methods*. Newbury Park: Sage, 1992.
- Rice N, Jones A. Multilevel models and health economics. *Health Econ* 1997;**6**:561–75.
- Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987;**43**:671–81.
- Thomas N, Lonford N, Rolph J. Empirical Bayes methods for estimating hospital-specific mortality rates. *Stat Med* 1994;**13**:889–903.
- Witte JS, Greenland S, Haile RW, et al. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology* 1994;**5**:612–21.
- Von Korff M, Koepsell T, Curry S, et al. Multi-level research in epidemiologic research on health behaviors and outcomes. *Am J Epidemiol* 1992;**135**:1077–82.
- Snijders TAB, Bosker RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: Sage, 1999.
- Brown H, Prescott R. *Applied mixed models in medicine*. New York: Wiley, 2000.
- Mason W, Wong G, Entwisle B. Contextual analysis through the multilevel linear model. In: Leinhardt S, ed. *Sociological methodology*. San Francisco: Josey Bass, 1983–1984: 72–103.
- Goldstein H. *Multilevel statistical models*. New York: Halsted Press, 1995.
- Kreft I, deLeeuw J. *Introducing multilevel modeling*. London: Sage, 1998.
- Diez-Roux AV. Multilevel analysis in public health research. *Annu Rev Public Health* 2000;**21**:171–92.
- Sixma HJ, Spreeuwenberg PM, Pasch MAVd. Patient satisfaction with the general practitioner: a two-level analysis. *Med Care* 1998;**36**:212–29.
- Hedeker D, Gibbons R, Davis J. Random regression models for multicenter clinical trials data. *Psychopharmacol Bull* 1991;**27**:73–7.
- Rutter C, Elashoff R. Analysis of longitudinal data: random coefficient regression modelling. *Stat Med* 1994;**13**:1211–31.
- Cnaan A, Laird NM, Slator P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med* 1997;**16**:2349–80.
- DiPrete T, Grusky D. The multi-level analysis of trends with repeated cross-sectional data. *Sociol Methodol* 1990;**20**:337–68.
- Langford I, Benthall G, McDonald A. Multi-level modelling of geographically aggregated health data: a case study of malignant melanoma mortality and uv exposure in the European community. *Stat Med* 1998;**17**:41–57.
- Hox JP, de Leeuw ED, Kreft IGG. The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In: Biemer PP, Lyberg LE, Mathiowetz NA, et al, eds. *Measurement errors in surveys*. New York: Wiley, 1991.
- Goldstein H. Multilevel cross-classified models. *Sociol Methods Res* 1994;**22**:364–75.
- Wong G, Mason W. The hierarchical logistic regression model for multilevel analysis. *J Am Stat Assoc* 1985;**80**:513–24.
- Diggle PJ, Liang KY, Zeger SL. *Analysis of longitudinal data*. New York: Oxford University Press, 1994.
- Laird NM, Ware H. Random effects models for longitudinal data. *Biometrics* 1982;**38**:963–74.
- Longford NT. *Random coefficient models*. Oxford: Clarendon, 1982.
- Dempster AP, Rubin DB, Tsutakawa RK. Estimation in covariance components models. *J Am Stat Assoc* 1981;**76**:341–56.
- Searle SR, Casella G, McCulloch CE. *Variance components*. New York: Wiley, 1992.
- Morris C, Christiansen C. Fitting Weibull duration models with random effects. *Lifetime Data Anal* 1995;**1**:347–59.
- Zeger S, Liang K, Albert P. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988;**44**:1049–60.
- Burton P, Gurrin L, Sly P. Extending the simple linear regression model for correlated responses: an introduction to generalized estimating equations and multi-level mixed modeling. *Stat Med* 1998;**17**:1261–91.