

An Introduction to the Use of Linear Models With Correlated Data

Benoît Laplante

Centre interuniversitaire d'études démographiques
Institut national de la recherche scientifique
Montreal, Quebec, Canada

Benoît-Paul Hébert

Institut national de la recherche scientifique
Montreal, Quebec, Canada

Abstract

Correlated data originate when observations are not selected independently because of sampling design, study design (especially panel studies), or spatial distribution of the population. In these situations, conventional methods for estimating the parameters of linear models are inappropriate, and the conventional methods for estimating the variances of these estimates may yield biased results. These two problems are different, but they are related. This paper provides an introduction to the problems caused by correlated data and to possible solutions to these problems. First, we present the two problems and try to specify the relations between the two as clearly as possible. Second, we provide a critical presentation of random effects, mixed effects and hierarchical models that would help researchers to see their relevance in other kinds of linear models, particularly the so-called measurement models.

Résumé

On obtient des données corrélées lorsque les observations ne sont pas sélectionnées de manière indépendante soit à cause du plan d'échantillonnage, soit à cause du plan d'enquête (surtout dans les études à passage répété), soit à cause de la répartition de la population dans l'espace. Dans de tels cas, les méthodes usuelles d'estimation des paramètres des modèles linéaires ne sont pas appropriées et les méthodes usuelles d'estimation des erreurs-types de ces estimés peuvent produire des résultats biaisés. Ces deux problèmes sont différents mais reliés. Nous proposons une introduction aux problèmes créés par les données corrélées et à leurs solutions. Nous présentons tout d'abord les deux types de problèmes en tentant d'éclaircir au mieux les relations entre les deux. Nous proposons ensuite une présentation critique des modèles à effets aléatoires, à coefficients aléatoires, mixtes et hiérarchiques qui devrait permettre aux chercheurs de mieux comprendre les liens qui les unissent à d'autres formes du modèle linéaire, en particulier les modèles de mesure.

Key Words: Correlated data, cluster sampling, random effects models, measurement models

Introduction

Correlated data originate in situations where observations in a sample are not selected independently of each other. This may happen in various settings, most of which are not unfamiliar to demographers.

The lack of independence may be a consequence of the sampling design. Selection of many individuals within a sampled unit, as in cluster sampling, is clearly a violation of the assumption of independence in the selection of observations and creates a situation where individuals selected from the same cluster are likely to be more alike than people selected from different clusters.

The lack of independence may also be the consequence of the design of the study. Studies that involve multiple observations on the same individuals, as in repeated measures design or panel surveys, generate observations that are correlated. Observations made on the same individual over time are more likely to provide similar information than observations on different individuals. This situation is quite similar to that of time series in econometrics.

The lack of independence may also be the consequence of the spatial distribution of the population under study. Researchers interested in any phenomenon in which a process involves a population and its location in space are likely to deal with samples in which the similarity between individuals is related to the distance between their location in space. These situations may arise in ecology, geography, urban studies, as well as in geology and other earth sciences.

In all these situations, conventional methods for estimating the parameters of linear models are inappropriate because they are all based on the assumption of independence between observations. And, the conventional methods for estimating the variances of these estimates, on which confidence intervals are based, are biased.

All these situations have a common fact. The selection of the primary sampling units (PSU) – that is, the clusters from which several individuals are selected on whom repeated measurements are made or the areas in which specimens are collected – is done at random within a population of clusters, individuals or areas. But, the collection of information from individuals is conditioned on their being the members of the PSU, and hence not random. This creates a situation where, if there are reasons to believe that there may be systematic differences between the PSU, the estimates of linear models will be affected by the sampling process.

The two problems are different in their origin and formal treatment, but they are nevertheless conceptually related and, from the researcher's perspective, they happen in the same setting and have to be dealt with at the same time. Over the last fifteen years or so, the increase in the power of computers has stimulated the development of software that can implement solutions to these problems. The possibility of handling these problems has led to complex survey designs on which studies, especially longitudinal studies, are based that collect correlated data. The availability of such data has stimulated the development of new statistical tools. Greene (1997) notes that in econometrics, panel data modelling has been one of the two most important areas of development since the beginning of the eighties. Social scientists are more and more likely to deal with the problems arising from correlated data. This is not only because there are more and more of such data but also because publication of studies based on such data in reputed journals is likely to become more and more demanding about the treatment of these data. It seems then that social scientists have no choice but to get themselves familiar with these problems and their solutions.

This, however, is more easily said than done. Part of the problem is that, outside economics, social scientists have seldom been trained in the mathematics needed to understand the many aspects of the problems and of the solutions. Reading articles or textbooks and software manuals about these problems and solutions

requires, among other things, a working knowledge of probability theory, calculus and matrix algebra, as well as a good understanding of maximum likelihood estimation and some familiarity with Bayesian estimation. The latter is especially important if one wants to come to terms with hierarchical modelling, one of the approaches that propose solutions to the problems of correlated data and random selection of values of variables. Another part of the problem is that work on correlated data is being done in very different domains and from perspectives that are sometimes hard to reconcile. This makes getting a comprehensive view of the field quite problematic, which alone is sufficient to deter benevolent researchers from using the methodological tools they know they should use.

In this article, first we will try to provide a clear outline of the two problems caused by correlated data and of the solutions to these problems. Separate treatments of these problems may be found elsewhere, but here, we will try to present them in an integrated fashion and make the relations between the two as clear as possible. Second, we will provide a critical presentation of random effects, mixed effects and hierarchical models that should help researchers to relate these with other kinds of special linear models, notably measurement models, and to clear the confusion that surrounds them at times. Towards this end, we shall review the basic principles of sampling (including stratification and clustering with special emphasis on intraclass correlation, design effects and on the use of weighting) and of robust variance estimator to obtain correct point estimates and standard errors. We will then look at the relations between prediction, measurement error and random effects, and their relevance in the context of correlated data. We will then look at various types of linear models developed from the idea of random effects: random-effects models, random-coefficients models, mixed models, and hierarchical models. Finally, we will look at a special case of correlated data modelling: panel data modelling.

Papers dealing with methodological topics usually include examples. This one does not, for the following reason. Complex surveys have been around for a while and longitudinal surveys are becoming more common. The techniques needed to handle the problems of statistical inferences arising from complex surveys have been known for a long time, although they are just starting to become widely available. The models needed to correctly analyse correlated data are more recent and are still being developed. However, most Canadian researchers cannot yet use these techniques and models on Canadian data. Statistics Canada, which manages most of the large-scale social and demographic surveys conducted in Canada, considers that including the information on PSU membership in its public use microdata files would create an intolerable risk of confidentiality breach. Access to this information is therefore restricted to deemed employees of Statistics Canada. This situation should change soon: Canadian researchers should be given access to confidential data through a network of university based regional data access centres that will become operational in 2001. Canadian researchers should therefore soon start to

be able to use the techniques and models described in this article. The authors became interested in these techniques and methods when they understood that they would have to use them and teach them in the near future. Until now, however, they have only been able to use them on foreign data. Examples using Canadian data are scarce, but real research using Canadian data should become common in the coming years.

Issues Related to Sample Design

Independent Observations: Simple Random Sampling and Stratification

A sampling procedure in which each member of the population has the same probability of being selected into the sample produces a probabilistic sample. The simple random sample is the best known of probabilistic sample designs. However, many, if not most, of the sampling procedures used in social sciences do not use such a simple scheme. Phone surveys typically use a two-stage selection strategy in which households are selected through their phone numbers and then one member of the household is randomly selected. Assuming that each household has the same probability of being selected in the first stage, it is obvious that the probability of being selected into the sample is not the same for everyone: this probability decreases with the number of people sharing the same telephone number. The resulting sample will therefore include proportionally more people living alone than there are in the population, less people living in two-people household than there are in the population, even less people living in three-people household than there are in the population, and so on. This problem can be handled in different ways, but the simplest and the most common is by weighting the selected individuals according to the size of the household they belong to. Such a simple weighting procedure corrects the sample and gives it a structure that is, theoretically and usually, identical to that of the population from which it has been drawn. All point estimates and other statistics computed from such a sample are unbiased.

Furthermore, it is possible to use non-proportional sampling designs to improve the precision of statistics computed from survey data. Whenever there are reasons to believe that the variance of a given variable is small within the categories of another variable, whereas the differences between the means of the first variable computed within the categories of the second are large, the standard error of the mean can be reduced by using non-proportional sampling. However, taking full advantage of this property requires the use of computation formulas that are not implemented in conventional software packages.

Non-independent Observations and Correlated Data: Cluster Sampling

As we have just seen, it is possible, in certain circumstances, to improve the precision of estimates by using a given sample design. In many cases, however, sample designs that depart from the simple random sample may actually decrease the power of the sample and thus increase the standard errors of any estimates that can be computed from the data. Most sample designs that depart from the simple random design involve some form of clustering. Whereas strata are categories of the population to which the design allocates different sampling probabilities and thus creates a series of smaller simple random samples, a cluster sample is a sample in which individual cases are selected because they belong to the same sampled unit. If the clusters are households, the sampling probabilities of the spouses are related. Once the household is sampled, both spouses will belong to the sample, whereas none of them will be included in the sample if the household is not selected. In such a design, it is clear that individual cases are not selected independently from each other. Households may be selected independently from each other, but not spouses. In real life, things may get much more complicated. Many large-scale samples rely on some form of geographical clustering. Such designs typically call first for a sampling of geographical areas, then for a sampling of several smaller areas within the areas selected during the first stage, then for a sampling of several buildings within each smaller area, and then for a selection of several households within the buildings. In such cases, it is quite clear that all stages of the design but the first add some form of dependence between the sampling probabilities of the individuals. Such samples are not as good as simple random or stratified samples as far as statistical inference is concerned, because they lack the property on which all the statistical theory of inference relies: independence of the observations. There are several ways, however, to use them to make inferences, but there is no way to extract from them as much information as there would be in a simple random sample of the same size because they do not contain as much information in the first place.

Variance Components and Intraclass Correlation

Using cluster samples for statistical inference requires an understanding of the sources of variance of the variables measured in such samples. In a simple random sample, all the variance stems from the differences between the sampled individuals, computed as deviations from the sample mean. In a cluster sample, the variance comes in part from the differences between the means of the sampled units, say households, and in part from the deviations of the individuals from the mean of the sampled unit to which they belong. The variance computed from the differences between the means of the sampled units is similar to the variance computed from the differences between the sampled individuals in a simple random sample. But, the variance computed from the deviations of the individuals from the mean of the sampled unit to which they belong is a

different thing. It contains no usable information for purposes of statistical inference because it does not come from random sampling; the individuals are selected because they are related to each other. Thus, in a cluster sample, the variance of any given variable has to be broken down into two components: the variance arising from the random sampling process and the variance that comes from the fact that individuals are selected because they are related to each other.

The computation of these variance components is similar to the computations performed in one-way analysis of variance. The between-clusters variance is equivalent to the between-groups variance of anova, except that it is interpreted as the part of the total variance that comes from the random sampling process. The within-cluster variance is equivalent to the within-group variance of anova except that it is interpreted as the portion of the total variance that does not come from a random sampling process but rather from the homogeneity of the cluster, that is the similarity of individuals belonging to the same selected unit.

As in analysis of variance, the estimation of the two variance-components begins by computing the ‘within’ and of the ‘between’ sums of squares. The within sum of squares is the sum of squares of all the deviations of the individual values on the dependent variable from the mean of this variable within the cluster to which they belong:

$$SSW = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \quad (1)$$

The between sum of squares is the sum of square of the deviations of the means of the dependent variable in each cluster from the overall mean:

$$SSW = \sum_{i=1}^k (y_i - \bar{y}_{..})^2 \quad (2)$$

The estimate of the residual, or error, variance is simply the within sum of squares divided by its number of degrees of freedom, which equals the product of the number of clusters and the number of individuals by cluster minus one. In the case of unbalanced data, that is when the number of individuals varies from cluster to cluster, the estimator has to be modified, but the logic remains the same. The estimate of the residual error is also known as the mean square error:

$$\hat{\sigma}_e^2 = \frac{SSW}{k(n-1)} = MSE \quad (3)$$

One should recall that whereas it is possible to compute an estimate of the residual or error variance directly from the within sum of squares, it is not possible to compute directly an estimate of the model variance from the between sum of squares, or in this case, an estimate of the variance due to the random sampling process. The mean square computed from the between sum of squares is not an estimate of the model variance because it is the sum of the (estimate of the) model variance and of the (estimate of the) error variance. To get an estimate of the model variance, one must therefore subtract the error variance from the mean square computed from the between sum of squares.

$$MSB = \frac{SSB}{k-1} \quad (4)$$

$$\hat{\sigma}_m^2 = \frac{SSB}{k-1} - MSE = MSB - \hat{\sigma}_e^2 \quad (5)$$

Once the estimates of the model and error variances are computed, it is possible to estimate the proportion of the total variance that comes from the clustered sampling design.

$$\hat{\rho} = \frac{\hat{\sigma}_m^2}{\hat{\sigma}_m^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_m^2}{\hat{\sigma}^2} \quad (6)$$

The ratio of the between clusters variance to the total variance is the proportion of the total variance that comes from the differences between the clusters. It is therefore also the proportion of the variance that does not come from the differences between individuals who belong to the same cluster. This quantity, known as the intraclass correlation coefficient and usually denoted by ρ (rho), cannot be negative and varies between 0 and 1. A high value of ρ implies that most variance comes from the differences between the clusters and that the individuals belonging to the same cluster are very similar. A low value of ρ implies that most of the variance comes from the differences between the individuals belonging to the same cluster. For this reason, ρ can be interpreted as the rate of homogeneity of the elements within clusters (Kish, 1965:170).

Robust Estimates of Variances

Unbiased point estimates of various statistics, from means to parameter estimates of linear models, can be computed from stratified or clustered data using conventional formulas as long as the individual observations are weighted in a way that makes the sample isomorphic to the population it was drawn from. However, for reasons that should be clear by now, conventional formulas for the computation of the variances (or standard errors) of statistics do not provide reliable results when they are used on stratified data or on clustered data.

For several decades, a common practice among social science researchers who use data from complex surveys has been to normalize weights in such a way that they make their samples isomorphic to the population while retaining the actual size of the sample. It was believed that this rescaling provided a good approximation of the power of the sample. In fact, this form of rescaled weighting provides conservative estimates of the variances for stratified samples, but systematically underestimates the variances when used in clustered samples. In other words, even when using weights that make the sample isomorphic while retaining its original size, conventional formulas provide estimates that are likely to be too large when they are computed from stratified data and almost certainly far too small if they are computed from clustered data. Therefore, using these formulas with stratified samples is tolerable while not optimal, whereas using them with clustered samples is risky, if not a sure way to disaster.

Complex sample designs that use both stratification and one or several levels of cluster selection make things even more complicated, even for the standard errors of statistics as simple as means and proportions. Things become quite intractable when one wants to compute standard errors for the estimates of the parameters of linear models.

This problem is similar, up to a point, to the problem created by statistics whose sampling distribution is not known or impractical. To make statistical inferences using these statistics, one needs to find a way to gather some knowledge of the distribution of the statistic without being able to deduct this knowledge by purely analytical means.

There are basically two strategies to circumvent this problem:

- The first strategy is to put aside the formulas for the variances and standard errors of statistics, simply build an empirical distribution of estimates of these coefficients, then compute the variance and standard deviations of these empirical distributions and finally use these as estimates of the appropriate variances and standard errors.

- The second strategy is to find a more mathematical solution to the problem.

There are three common methods that implement the first strategy; the balanced repeated replication, the jackknife repeated replication, and the bootstrap repeated replication. Although it can be adapted to estimate variances from clustered samples, the balanced repeated replication method (BRR) is basically designed to handle stratified samples and requires, at least in principle, that exactly two individuals be selected from each stratum. It is not commonly used to estimate the variances of the estimates of linear models.

The jackknife and bootstrap methods are quite similar. Jackknife estimates of the variance of a statistic are computed by calculating the statistic with its usual estimator once in each of the pseudo samples that can be created from the original sample by deleting one different observation at a time. The data can thus be used to create up to as many different pseudo samples as there are observations in the original sample (the maximum number of possible pseudo samples is simply the number of possible combinations of $n-1$ observations that can be drawn without replacement in a population of size n , that is $n!/(n-1)!(n-(n-1))!$, which reduces to n .) Once the statistic has been computed from the different pseudo samples, its empirical variance can be computed using a formula very similar to the common estimator of variance. One will recall that the estimator of the variance of a variable from a simple random sample is

$$\hat{\sigma}_y^2 = \left(\frac{1}{n-1} \right) \sum_{i=1}^n (y_i - \bar{y})^2 \quad (7)$$

whereas the estimator of the variance of the mean is simply

$$\hat{\sigma}_{\bar{y}}^2 = \frac{\hat{\sigma}_y^2}{n} = \left(\frac{1}{n(n-1)} \right) \sum_{i=1}^n (y_i - \bar{y})^2 \quad (8)$$

The jackknife estimator of the variance of the mean computed from n pseudo samples is

$$\hat{\sigma}_{\bar{y}}^2 = \left(\frac{n-1}{n} \right) \sum_{j=1}^n (\bar{y}_j^* - \bar{\bar{y}}^*)^2 \quad (9)$$

where the star symbolises statistics computed from the pseudo samples.

The same logic can be used to estimate robust variances of any statistic, including coefficients of linear models and statistics for which there is no easy or simple way to implement an estimator of the variance, such as the median.

The general formula of the jackknife estimator of the variance of any statistic θ is

$$\hat{\sigma}_{\theta}^2 = \left(\frac{n-1}{n} \right) \sum_{j=1}^n (\theta_j^* - \bar{\theta}^*)^2 \quad (10)$$

The bootstrap method is quite similar to the jackknife method. One has again to start with a sample of size n and an estimator. With the bootstrap method, one creates as many pseudo samples of size n as appropriate by drawing n cases with replacement from the original sample. The pseudo samples will contain duplicate cases. One then estimates the statistic θ using the data from each of the pseudo samples and uses these values to compute the bootstrap variance using the bootstrap estimator of variance, which is nothing but the estimator of the variance from simple random data:

$$\hat{\sigma}_{\theta}^2 = \left(\frac{1}{k-1} \right) \sum_{j=1}^k (\theta_j^* - \bar{\theta}^*)^2 \quad (11)$$

where k is the number of pseudo samples. The appropriate number of pseudo samples, or replications, varies according to the needs of the researcher and the kind of statistics for which a variance estimate is needed. Depending on these, the appropriate number of replications may vary from 50 to 200 if one wishes simply to estimate the standard error whereas as many as 1000 may be needed to estimate 95% confidence intervals.

It should be stressed that whereas the jackknife and bootstrap estimators of the variance of a statistic are more accurate estimators of the variance of a statistic than the conventional estimators, the point estimates that could be computed using the jackknife or bootstrap replication methods are not better point estimates of the statistic itself than the estimate computed from the original sample.

Although we did not stress it up to now, the presentation we just made of the jackknife and bootstrap estimators assumes simple random sampling. It is therefore well suited for the computation of the variance of a statistic whose sampling distribution is not known, but which was computed using data from a simple random sample. However, what we need are variances of statistics computed from samples not created using simple random selection. Fortunately, the solution to the problem is quite simple: to compute correct estimates of the variance of statistics using data from complex samples, one simply has to make sure that the pseudo samples used by the replication algorithms in their computations are samples of PSU and not samples of individuals. The

computation of the correct variances of the statistics then becomes straightforward.

Whereas the balanced repeated replication, the jackknife and bootstrap methods all use pseudo samples to compute estimates of the variances of statistics, it is possible, at least in certain circumstances, to compute robust estimates of the variance of statistics using a more mathematical approach to the problem (the second strategy). Practically, this approach relies on the use of an estimator of the variance that goes by many names among which the Taylor series method, the linearization method, the first-order Taylor series linearization method, and the Hubert/White/sandwich estimator method are the most common. The mathematical introduction to this estimator is a bit demanding (see Binder (1983), Cochran (1977), Fuller (1975), Godambe (1991), Kish and Frankel (1974), Särndal et al. (1992), Shao (1996), and Skinner (1989)).

Issues Related to Model Design

Prediction Error

Equation 12 below presents a simple regression model. In such a model, the actual dispersion of the observed dependent variable Y is modelled as a deterministic and a stochastic process. The deterministic process models the expected value of Y conditional on the independent variable X . The modelling of the expected mean accounts for some of the actual dispersion of Y . The remaining dispersion is assumed to be produced by a Gaussian stochastic process that is represented by ε and is usually interpreted as prediction error. The ε represents the dispersion of the observed dependent variable Y around its expected mean conditional on the values of the independent variable X . This is the most conventional form of regression model.

$$Y = \alpha + \beta X + \varepsilon \quad (12)$$

Measurement Error

At least in the social sciences, most users of linear models have never been taught that historically, regression and related models are derived from measurement theory and that the error or residual that appears in these models was originally thought of as measurement error. From such a perspective, Equation 12 does not represent, say, the influence of education (X) on income (Y). Rather, the expected value of Y is the best point estimate of the true value of something that would be repeatedly measured with an instrument whose outcomes are affected by random imprecision (ε) and a systematic bias (α).

These two interpretations of the basic linear model have been around for a long time and are still usually kept apart. Educational studies and psychometrics, for instance, pay a great deal of attention to measurement theory whereas social sciences are basically interested in causal modelling. However, the two perspectives can be merged: one can devise models in which prediction error and measurement error coexist. Equation 13 is such a model.

$$Y = \alpha + \beta(X + \delta) + \varepsilon \quad \text{or} \quad Y = \alpha + \beta X + \beta\delta + \varepsilon \quad (13)$$

Y is a dependent variable and X an independent variable, α is the conventional origin of the regression equation, ε is the prediction error of Y conditional on X , but here, we do not assume that X was measured exactly but rather that it was measured with some imprecision. Practically speaking, we consider that the observed value of X is not its true value, but the sum of the true value and some disturbance. In psychometrics or measurement theory, the proportion of the variance of X that would come from the true X would be referred to as its reliability. In a survey, for instance, X could be any independent variable. If it were income, the measurement error could be the consequence of the reluctance of people to disclose their personal income, or their ignorance of the household income, or the imprecision generated by rounding or categorisation of the income into wide income categories. If it were sex, the imprecision would come most likely from coding errors or presumably rare events such as transsexual respondents, respondents playing games with the interviewer or interviewers intentionally tampering data.

Of course, once one has assumed that some or all independent variables should be thought of as measured with error, there is no reason to assume that dependent variables are measured without errors. Thinking of a model such as Equation 14 becomes therefore unavoidable.

$$(Y + \psi) = \alpha + \beta(X + \delta) + \varepsilon \quad (14)$$

In this model, both the dependent and the independent variables are assumed to be measured with error. In equation 14 as in Equation 13, ε still represents prediction error.

The inclusion of measurement errors in regression models typically increases the size and significance of the estimated effects of the independent variables because all of the covariance between the dependent and independent variable is imputed to the ‘reliable’ portion of their variances. In other words, in any real situation, the estimate of β in Equation 14 should be different from the estimate of β in Equation 12. Models that include measurement errors as well as prediction errors have become common over the last twenty years mainly through the development of structural equation modelling, also known as

covariance structure analysis, through the work of Karl Gustav Jöreskog and Peter Bentler and the available software, LISREL and EQS respectively. However, the models commonly used in covariance structure analysis are different from Equation 14. Typical covariance structure analysis models make a conceptual distinction between the prediction and the measurement parts of the model and are usually of the form

$$\eta = \beta\xi + \psi, \quad y = \eta + \varepsilon \quad \text{and} \quad x = \xi + \delta \quad (15)$$

In such a model, ψ is the prediction error of the dependent variable η , whereas ε is the measurement error of the observed dependent variable and δ is the measurement error of the observed independent variable. However, in this model, the regression (or structural) coefficient β multiplies only the ‘true’ latent independent variable ξ and not the measurement error δ of its observed counterpart x . The regression coefficient from a model that takes measurement errors into account in this way is known as a disattenuated regression coefficient.

Random-effects Models

The distinction between measurement theory and causal modelling and their respective interpretation of the regression equation is not a new idea, but the combination of the two perspectives in the same models is a rather recent development. Another old statistical idea that has received new attention and generated new developments in recent years is the distinction between fixed effects and random effects.

The categorical variables that are used as factors in an analysis of variance may be classified into two groups: those for which the entire population of categories is represented in the sample and those for which the sample contains only a sample of the possible categories. Sex is a variable that belongs to the first category. In a sample of the population, we will find men and women, and only men and women. The variable has only two categories, and both are present in the sample.

Studies where the effect of the experimenter or the interviewer is taken into account provide a classical example of the second type of variables. Although the idea may seem confusing at first, experimenters and interviewers are persons who could have been replaced by other experimenters or other interviewers. Thus, they are sampled and can be seen as a random sample drawn from a population of experimenters or interviewers. Suppose that interviewers make up an independent variable in an analysis of sample data. Then, we create a strange situation. We have a sample of interviewees drawn from the population under study and a sample of the categories of this particular independent variable

(interviewers), whose categories we happen to have sampled from the population of all the people who could have been hired to conduct the interviews. Using that kind of variable raises a new question: Would the estimate of the coefficient or the size of the effect of the interviewer variable have been the same if we had drawn a different sample of experimenters or interviewers? This question is equivalent to considering the estimates of the coefficient of the interviewer variable itself to be a source of random error. In other words, when the categories of one of the independent variables are sampled from a population of categories, the variance of the dependent variable must be decomposed into three types of variances: a) variance explained by the model, b) residual variance unexplained by the model, and c) variance due to the sampling of the categories of one of the independent variables.

The simplest way to implement this idea is to assume that the effects of all the other independent variables in the linear model are the same for all categories of the sampled independent variable and to consider that all the differences between categories can be modelled as differences between their intercepts. Such a model can be implemented easily by using a series of dichotomies to represent the different categories of the sampled independent variables and using ordinary least square estimates. Because this method does not really model the randomness of the differences between the categories but simply represents the differences between them as a series of ordinary fixed effects, it is known as a *fixed-effects* model. Precisely because these effects are assumed to be fixed and estimable, although they arise from a random sampling process, the estimates of these models are conditioned on the sample from which they are estimated. This limits the generalization of their results to the population. In other words, the estimates of such a model are truly sampled from a population of estimates that vary according to the sampled values of the independent variables.

Another way to implement the randomness of the selection of the values of an independent variable is to still consider that all the differences between categories can be modelled as differences between their intercepts, but now include the randomness of the selection process.

$$Y = \alpha_i + \beta X + \varepsilon \text{ or } Y = (\alpha + \delta) + \beta X + \varepsilon \text{ or } Y = \alpha + \beta X + \delta + \varepsilon \quad (16)$$

In such a model, the variance of the dependent variable is really broken down into three distinct components: the variance explained by the deterministic part of the model (that is, the variance explained by the independent variable X), the residual variance of the prediction error ε , and the variance of the intercept α or the variance of the random component δ of the intercept. In Equation 16, δ represents what is known, depending on the setting, as the cluster effect or the panel effect.

Practically speaking, random-effects models are similar to their conventional counterpart. Conceptually, the main difference lies in this. The variance-

covariance matrix of the residuals is no more a square matrix with a single value on the diagonal (σ_ε^2 , the variance of the residuals that is equal for all values of the dependent variable under the assumption of homoscedasticity) and zeros everywhere else (under the assumption of independence of the residuals). It is replaced by a square matrix of a slightly more complex structure, a matrix whose elements are matrices (Equation 17). The size of the matrix is the number of clusters or panels. Each diagonal element of the matrix is a matrix whose size is the number of elements (individuals in the same family, observations on the same individual, etc.) in the cluster or panel. All off-diagonal elements of the matrix are null matrices, that is, matrices of zeros.

$$\mathbf{V} = \begin{bmatrix} \mathbf{\hat{U}}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\hat{U}}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ & & \vdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\hat{U}}_n \end{bmatrix} \quad (17)$$

$$\mathbf{U}_i = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_v^2 & \sigma_v^2 & \sigma_v^2 & \cdots & \sigma_v^2 \\ \sigma_v^2 & \sigma_\varepsilon^2 + \sigma_v^2 & \sigma_v^2 & \cdots & \sigma_v^2 \\ & & \vdots & & \\ \sigma_v^2 & \sigma_v^2 & \sigma_v^2 & \cdots & \sigma_\varepsilon^2 + \sigma_v^2 \end{bmatrix} \quad (18)$$

Each of the diagonal submatrices has the same structure as in Equation 18. All diagonal elements are the sum of the variance of the prediction error term (σ_ε^2), which is the same for all individuals or observations, and of *the variance of the cluster or panel effects* (σ_v^2), which is assumed to be constant for all clusters or panels. The off-diagonal elements of the submatrices are equal to the variance of the cluster or panel effects. Thus, the total unexplained variance is the sum of the variance of the prediction error and of the variance of the cluster or panel effects, and the covariance between the error terms within a cluster or a panel is simply the variance of the random cluster or panel effects.

One should note that although both the prediction error and the within-cluster or within-panel effect are assumed to be uncorrelated across clusters or panels, the two variances are assumed to be the same for every case and within each cluster or panel. In other words, the variance of the within-cluster or panel effects is assumed to be the same for all panels, and the correlation among the errors within each panel is assumed to be the same within each panel.

In the context of random effects models or in that of generalized linear models, models using a variance-covariance matrix with a structure corresponding to (17) and (18) are often referred to as using a ‘compound symmetry correlation matrix’ or an ‘exchangeable correlation matrix.’

Random-coefficients Models

From a conceptual perspective, there is no special reason other than convenience to limit the randomness to the intercept. Any independent variable can be thought as having a random effect rather than a fixed effect. In other words, models such as

$$Y = \alpha + \beta_i X + \varepsilon \text{ or } Y = \alpha + (\beta + \zeta)X + \varepsilon \text{ or } Y = \alpha + \beta X + \zeta X + \varepsilon \quad (19)$$

are perfectly possible and, indeed, are known as random-coefficient models. Such models imply that the actual values of the independent variable X have been randomly sampled from a population of values, that the effect of X on Y varies across the values of X , and that the estimate of β (in this context, the *average* effect of X on Y), would have been different had we got a different random sample of the values of X . This makes sense if we are willing to assume that X is some form of idiosyncratic characteristic that cannot be reduced to a more general trait and that the effect of X on Y varies across the values of X following a given distribution, almost always the normal distribution with mean 0. If X were a general trait or could be replaced or ‘explained’ by one or several general traits, it would be preferable to replace X by these traits and build a more explicit model. Or, if there were reasons to believe that the effect of X on Y varied in some systematic rather than random way, it would be more meaningful to model this variation as an interaction rather than a random effect.

An example may help to grasp these subtleties. Let us imagine a survey in which we are, again, interested in the effect of the interviewers. In our previous attempt at modelling this effect, we considered that the influence of the interviewers on the data collection was merely on the measurement of the dependent variable. This is a very disputable assumption, though. If each interviewer was likely to induce some form of systematic measurement bias, at least using a survey questionnaire in which there is no clear distinction between what will be a dependent or an independent variable, there is no reason to assume that the effect will be on the measurement of any variable in particular. In other words, the interviewer effect is really a measurement problem, the kind that should be dealt with using covariance structure analysis, and not a random effects problem. However, it is possible that the difference between interviewers were not in the way they recorded the dependent variable, but more in the way they recorded the relationship between the dependent variable and some independent variable. For instance, one can imagine that some interviewers had harder times

at getting answers to more complex questions from less educated people, thus creating an artificial relation between complex questions and education as well as other potentially independent variables that are related to education such as income. These differences would not be spread all over the sample, but rather located only in the portions of the sample interviewed by those presumably less experienced interviewers. In a situation like this, the differences between the interviewers could not be modelled properly as a difference in the intercept of the model, but should be modelled as a difference in the effects of education on whatever the complex questions were used to measure the dependent variable. This could be modelled using a random coefficient for education.

Towards Hierarchical Models

Let's now borrow an example from Goldstein (1995), this time about a political survey. We have information on party preference. Along with other information, we know the constituency of the voter. There are sound reasons to believe that party preference may vary according to sex, age, education and income, but also according to constituency. In the context of random-effects models, sex, age, education and income are 'fixed' characteristics because there are no reasons to believe that the values of these variables in our sample are only a sample of the possible values of these variables. They are general characteristics, well-defined variables that can be used to measure individuals on a given scale or to classify them according to given criteria. Constituency is a rather different thing. Formally, it can be considered as a variable because it can be used to categorize people in an exhaustive and exclusive way. However, it is not something that measures individuals on a scale or classifies them according to some substantive and meaningful concept such as language, ethnicity or religion. As a variable, it is barely more meaningful than a name or postal code. Certainly, many differences across constituencies can be explained by the composition of the population living in the constituencies and can therefore be accounted for using 'fixed' variables and interactions among these if so needed. However, there may be differences across constituencies that cannot be reduced to 'fixed' characteristics. One way of addressing this problem is to consider these irreducible differences as residual variance and to throw them in the error term of the model. Another way is to consider that as long as these remaining differences can be related to constituencies, the model should be estimated by taking advantage of them. Hierarchical models are basically random-coefficients models used for taking advantage of the information contained in groupings of individuals that are not necessarily sampled as are categories in random-effects models, but that cannot be reduced to fixed characteristics either.

Boyle & Lipman (1998) provide an example of the use of hierarchical models with Canadian survey data (National Longitudinal Survey of Children and Youth). Their study of child problem behaviour assessed the effects of factors

measured at three nested levels: the children, their family and their neighbourhood.

**Final Remarks:
Relations Between Random Effects,
Clusters and Complex Sample Designs**

The relations between random effects, clusters and complex sample designs tend to be confusing. A random effect arises when the model includes an independent variable for which the data contain only some values of a population of values. A model in which we want to include the interviewer (who collected the data) as an independent variable will contain a random effect because the data could have been collected by other interviewers. And, the data would probably have been different, not because they would have come from a different sample of the population of individuals (we are talking of different interviewers, not different interviewees), but because different interviewers would have collected the data in a different manner. The data would likely show some correlation within the groups of individuals interviewed by the same interviewer. It is not because these individuals have anything special in common (the sample is assumed to be random and the individuals independently selected) but because each interviewer is assumed to have a distinct but not otherwise specified way of collecting the data.

A proper treatment of such situations involves the use of linear models that allow for the inclusion of the random effect. The random effect may be specified in various ways. Because the random effect creates groups within the sample whose data are thought to be correlated, the specification of the random effect is a form of the specification of a correlation or more generally of dependence among observations.

Clustering is a sampling technique by which the individuals are not selected independently. A simple and classical example is a sample selected by using all the members of randomly selected families. In such a sample, only the families are selected at random and independently. Such a sample contains less information than a simple random sample of the same size, because it is expected that individuals will show more similarities with their near relatives than with other individuals. The data within the families are likely to be correlated, not because the individuals have been submitted to some common procedure as it is the case with the interviewer effect, but because they have not been selected independently. The data will likely show some correlation within the families. However, because this correlation is a consequence of the sample design and not one of the measured independent variables, it is commonly handled as a sample design problem and not as modelling problem. Practically, the estimates of the variances and covariances of the estimates are corrected using one of several common correction procedures such as jackknife repeated

replication, bootstrap repeated replication, or Taylor series method (also known as the first-order Taylor-series linearization method, the Hubert/White/sandwich estimator, the delta method and the propagation of variance). Such correction methods allow not only for simple two-stage cluster designs but also for more complex survey designs involving strata as well as clusters. Software packages implementing these correction methods usually do so with estimation routines in order to provide robust estimates of the standard errors of the estimates, but also use the information on the sample design (i.e., the sampling weights) to compute unbiased estimates of the linear model coefficients as well.

Correlated data may arise in other circumstances. Repeated measurements are one of the most common of these. In such cases, information is usually collected from a random sample of individuals through several interviews. The observations coming from the same individual are not independent and the data are correlated. Formally, there is little difference between this situation and that of a clustered sample in which all the members of randomly sampled families would have been selected. However, repeated measurements data are commonly analysed using random-effects and random-coefficients models. The main reason for this choice is that the sample consists of a set of related observations that would have been different if different individuals had been sampled and thus, this situation can be related to the case of the interviewer effect that generates a random effect. There is an important difference, however, between the case of the interviewer effect and that of repeated measurements: the individuals interviewed by the same interviewers were all selected independently, whereas the observations on the same individual are not independent. Thus, one way to obtain parameters estimates of a linear model that takes correlation between repeated measurements on the same individuals into account is to include individual-specific effects in the model and to assume that these effects are randomly distributed.

Another way to estimate linear models for repeated measurements that take the correlation between the observations into account is to use the *generalized estimating equations* (GEE) method (Liang and Zeger, 1986). Models estimated with this method do not include cluster- or individual-specific effects (they are ‘population-averaged’ models), but can deal with different correlation structures within the clusters.

Some types of surveys combine the problems created by clustering and by random selection of the categories of some independent variable. These are very common in educational studies. Samples of children selected from randomly selected schools and classes within schools are clustered samples that create the problems we have examined before. Estimates should be computed using the appropriate weights, and variances of the estimates should be computed using a robust procedure that takes into account the weighting scheme. But they also create another kind of problem when the classes and the schools – the sampled units – to which the children belong are used as factors in an analysis of

variance, a regression or any other linear model. Unlike sex, which has only two categories, schools and classes are numerous and, by design, the sample can contain only a fraction of the population of the schools and classes and thus, only a fraction of the possible values of the variable ‘School’ and ‘Class’ are present in the sample. In such a situation, the values of the variables that are to be used in the linear model have been sampled at random. They should be dealt with accordingly, using random-effects models, random-coefficients models, mixed-effects models, or hierarchical models.

Appendix — Computer Packages

This appendix lists several statistical programs that offer some capabilities in the estimation of robust variances from complex survey data, random-effects models and hierarchical or multilevel models respectively. The choice arises from the authors’ usage of various programs in their own research and is therefore far from exhaustive. The sections on robust estimates of variance and hierarchical models list a couple of standalone programs and mention the capabilities of two relatively well known statistical packages: SAS and STATA. SPSS is not discussed further simply because it does not handle complex survey data, does not compute robust estimates of variance and handles random effects solely within the context of its general linear model procedure which practically limits the analyses to multivariate analysis of variance and conventional regression. We made no effort to assess what is offered by other statistical packages (e.g. SYSTAT), semi-specialized programs (e.g. LIMDEP) or very general programs (S-PLUS).

Robust Estimates of Variance

SUDAAN from Research Triangle Institute (www.rti.org): SUDAAN offers Taylor series linearization (using the generalized estimable equation approach in regression models), jackknife, and balance repeated replication robust variance estimators. Release 7.5 offers descriptive statistics plus several linear models: regression, logistic regression, multinomial logistic regression and Cox semi-parametric proportional hazards model. May be used with SAS or as a standalone product.

WesVar from Westat (www.westat.com): WesVar offers jackknife and balance repeated replication robust variance estimators. Version 4 includes descriptive statistics plus several linear models: analysis of variance, regression, logistic regression, multinomial logistic regression. Standalone program that reads SPSS system files and SAS transport files.

SUDAAN and WesVar have been developed by research companies who needed these programs to perform their business. Research Triangle Institute is

mainly active in medical research, whereas Westat is more oriented towards survey research.

Version 8.1 of SAS/STAT, from SAS Institute Inc. (www.sas.com), includes two procedures that use Taylor series linearization to estimate the variances of estimates for means and regression from complex survey data. Two caveats however. 1) SAS Institute makes available a macro, JACKBOOT that allows either jackknife or bootstrap replication methods to estimate standard errors. However, the samples drawn by this macro during each replication step are samples of individuals instead of samples of clusters. The macro is therefore suited to compute the variances of statistics computed using simple random samples, but whose variance cannot be computed using a known formula; it is not suited for the estimation of the variance of statistics computed using stratified or clustered samples. 2) In the version 8.1 of SAS, PROC PHREG, the procedure that estimates the Cox semi-parametric proportional hazards model, comes with an option that computes the Lin and Wei (1989) robust estimates of variance. These estimates of variance are robust in the sense that they allow valid statistical inferences when the model being estimated is misspecified. Unless the way they are computed is modified to take cluster membership into account—which is not the case in PHREG—they rely on the simple random sampling assumption and, therefore, they do not allow valid statistical inferences from clustered samples.

STATA 7, from Stata Corporation (www.stata.com), includes procedures that use Taylor series linearization to estimate the variances of estimates for descriptive statistics, two-way tables and several forms of the linear model: regression, logistic regression, probit, multinomial logistic regression, Poisson regression, censored regression as well as a few others including Cox semi-parametric proportional hazards model. In the last case, it uses the Lin and Wei (1989) robust variance estimates, but allows taking clustering into account. STATA also includes procedures that allow the Taylor, bootstrap, and jackknife methods with almost any of its statistical models. Contrary to the SAS macro, these procedures include options that take the structure of complex samples into account in the computation (for the Taylor method) or in the replications (bootstrap and jackknife methods).

Random Effects Models

SAS/STAT PROC MIXED, a component of SAS, is basically a multivariate analysis of variance program that handles a wide variety of form of structures of correlation among residual errors. ‘Mixed models’ is just a generic name for models that may include both fixed and random effects. Within the framework of MANOVA, PROC MIXED allows random effects models and random coefficients models. It may be used for repeated measures models. SAS Institute makes available two macros, GLIMMIX and NLINMIX, that may be used to

expand the possibilities of PROC MIXED. GLIMMIX uses PROC MIXED to fit generalized linear mixed models, whereas NLINMIX uses PROC MIXED and PROC NLIN to fit nonlinear mixed models. SAS/ETS includes a procedure, PROC TSCS, designed to estimate econometric models using time series cross-sectional data, which are basically random effects regression models.

STATA 7 offers a wide variety of procedures to handle various forms of random effects models, specially in the context of panel data: random-effects tobit, probit, logistic regression, complementary log-log regression, Poisson regression, as well as a few others. Interestingly, most of these procedures offer robust standard errors estimates computed with the Taylor linearization method as an option, at least for some of the models they estimate. As mentioned above, the software includes procedures that allow the Taylor, bootstrap, and jackknife methods with almost any of its statistical models. Most serious limitation: STATA 7 has little capabilities for the estimation of random coefficients models. In principle, random coefficients models could be estimated using the user provided hierarchical generalized linear models described below.

Hierarchical or Multilevel Models

Hierarchical models and multilevel models are two names that have become widely spread. Both refer to the same kinds of models, which can be described as random coefficients models developed within the framework of regression, rather than that of analysis of variance, to handle problems pertaining to the modelling of correlated data arising from clustered sampling. The two most widely recognized standalone programs for the estimation of these models are MLWin, developed by a team lead by Harvey Goldstein from the Institute of Education of the University of London (www.ioe.ac.uk/mlwin) and HLM, developed by Stephen Raudenbush, Anthony Bryk and Richard Congdon and distributed by Scientific Software International (www.ssicentral.com). Each program can estimate a variety of specific linear models and provide robust estimates of standard errors. Detailing the differences between the two programs goes beyond the scope of this appendix.

Donald Hedeker and Robert D. Gibbons of the University of Illinois at Chicago have developed a suite of programs (MIXOR, MIXREG, MIXNO, MIXPREG, and MIXGSUR) for mixed-effects linear regression, mixed-effects logit or probit models for binary or ordinal outcomes, mixed-effects logistic regression for nominal outcomes, mixed-effects Poisson regression, and mixed-effects grouped-time survival analysis. The suite is available free of charge at <http://www.uic.edu/~hedeker/mix.html>.

Three STATA users, Sophia Rabe-Hesketh, Andrew Pickles and Colin Taylor, have committed themselves to the endeavour of writing a procedure for the estimation of hierarchical (or multilevel) generalized linear models. This

procedure is distributed by Stata Corporation as “gllamm6”. The procedure has been written following McCullagh and Nelder (1989) presentation of generalized linear models (which has very little to do with the much more conventional “general linear model” approach we refer to in discussing the limited capabilities of SPSS). Specific models can be estimated by combining one of several distributions for the random component of the model (in gllamm6, these are binomial, gaussian, gamma, or Poisson) with one of several links (identity, log, logit, reciprocal or probit). The procedure does not seem to provide robust standard errors of the estimates; in principle, it should be possible to use the procedure with the robust variance estimation procedures of STATA.

References:

- Binder, D. A. 1983. “On the variances of asymptotically normal estimators from complex surveys,” *International Statistical Review* 51: 279-292.
- Boyle, M. H., and E. L. Lipman. 1998. *Do places matter? A multilevel analysis of geographic variations in child behaviour in Canada*, Applied Research Branch, Human Resources Development Canada, Working Paper no W-98-16E.
- Burton, P., L. Gurrin, and P. Sly. 1998. “Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling,” *Statistics in Medicine* 17: 1261-1291.
- Carlin, J. B., R. Wolfe, C. Coffey and G. C. Patton. 1999. “Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: Prevalence and incidence of smoking in an adolescent cohort,” *Statistics in Medicine* 18: :2655-2679.
- Cnaan, A., N. M. Laird and P. Slasor. 1997. “Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data,” *Statistics in Medicine*, 16: 2349-2380.
- Cochran, W. G.. 1977. *Sampling techniques*, 3rd edition. New York, NY: John Wiley and Sons.
- Fuller, W. A. 1975. “Regression analysis for sample survey,” *Sankhyā*, Series C 37: 117-132.
- Godambe, V. P. ed. 1991. *Estimating Functions*. Oxford: Clarendon Press.

An Introduction to the Use of Linear Models with Correlated Data

- Goldstein, H. 1995. *Multilevel Statistical Models* (2nd edition), London: Edward Arnold.
- Kish, L. and M. R Frankel. 1974. "Inference from complex samples," *Journal of the Royal Statistical Society*, B 36: 1-37.
- Lin, D. Y. and L. J. Wei. 1989. "The robust inference for the Cox proportional hazards model," *Journal of the American Statistical Association*, 84: 1074-1078.
- Liang, K.-Y., and S. L. Zeger. 1986. "Longitudinal data analysis using generalized linear models," *Biometrika*, 73: 13-22.
- Nelder, J. A. 1998. "A large class of models derived from generalized linear models," *Statistics in Medicine* 17: 2747-2753.
- McCullagh P. and J. A. Nelder 1989. *Generalized Linear Models, second edition*. Boca Raton: Chapman and Hall/CRC.
- Rabe-Hesketh, S., A. Pickles, and C. Taylor. 2000. "Generalised, linear, latent and mixed models," *Stata Technical Bulletin* 53: 47-57.
- Särndal, C. E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York, NY: Springer-Verlag.
- Shao, J. 1996. "Resampling methods for sample surveys (with discussion)," *Statistics* 27: 203-254.
- Skinner, C. J. 1989. "Introduction to Part A" in C. J. Skinner, D. Holt, and T. F. Smith, eds, *Analysis of Complex Surveys*, (p.23-58). New York, NY: John Wiley and Sons.
- Sullivan, L. M., K. A. Dukes and E. Losina. 1999. "An introduction to hierarchical linear modelling," *Statistics in Medicine* 18: 855-888
- Tepping, B. J. 1968. "Variance estimation in complex surveys," *American Statistical Association proceedings on social statistics section*: 11-18.