Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2007/10

Selection bias in linear mixed models

Leonardo Grilli, Carla Rampichini

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Selection bias in linear mixed models

Leonardo Grilli

Department of Statistics 'Giuseppe Parenti', Florence, Italy.

Carla Rampichini

Department of Statistics 'Giuseppe Parenti', Florence, Italy.

Summary. The paper investigates the consequences of sample selection in multilevel or mixed models, focusing on the random intercept two-level linear model under a selection mechanism acting at both hierarchical levels. The behavior of sample selection and the resulting biases on the regression coefficients and on the variance components are studied both theoretically and through a simulation study. Most theoretical results exploit the properties of Normal and Skew-Normal distributions. In the case of clusters of size two, analytic formulae of the bias are provided that generalize Heckman's formulae. The analysis allows to outline a taxonomy of sample selection in the multilevel framework that can support the qualitative assessment of the problem in specific applications and the development of suitable techniques for diagnosis and correction.

Keywords: clustered data, multilevel model, sample selection, Skew-Normal distribution, truncation.

1. Introduction

In many settings the statistical units are nested in hierarchical structures, such as pupils in schools or repeated measurements on a set of individuals, with *level 1 units* (pupils, repeated measurements) embedded in *clusters* or *level 2 units* (schools, individuals). This kind of structure often implies correlated responses at level 1, which can be taken into account by means of *mixed* models (Verbeke and Molenberghs, 2000; Goldstein, 2003), also known as *multilevel*, *random effects* or *variance components* models.

Observational studies are often affected by sample selection, that is the response variable of principal interest is observed conditionally on the value of another variable. For example, wage is observed only for people actually working. In regression analysis, including multilevel modelling, sample selection leads to biases if the selection mechanism depends on unobserved variables correlated with the model errors.

Starting from the work of Heckman in the Seventies (Heckman, 1979), the problem of selection bias has been thoroughly studied in the context of standard single-level models. See Vella (1998) for a general review and Puhani (2000) for a review on simulation studies on this topic. The issue of selection bias was tackled by several authors in the framework of panel data (Hausman and Wise, 1979; Wooldridge, 1995; Kyriazidou, 1997; Vella and Verbeek, 1999; Jensen et al., 2001), usually with reference to the linear mixed model. The same problem was considered also in the framework of longitudinal data in Biometrics (Wu and Carroll, 1988; Follmann and Wu, 1995; Saha and Jones, 2005).

However, even if the random effects models for panel or longitudinal data are an instance of mixed models, their specificity makes the extension of the results to the general multilevel setting not trivial, especially for the case of cross-sectional studies. A few examples of applied works dealing with sample selection in multilevel models are Borgoni and Billari (2002), Bellio and Gori (2003) and Grilli and Rampichini (2007). Anyway, the existing treatments of selection bias in mixed models do not provide a systematic discussion of the many types of selection mechanisms that can arise in a hierarchical framework, nor they discuss at length how the selection bias is affected by the parameters of the model and of the selection process. In fact, it is essential to recognize that the phenomenon of selection in a mixed model is much more complex than in a standard fixed effects model for the following reasons: (*i*) the selection process can act at different levels, giving rise to a wide variety of patterns; (*ii*) the model of interest is quite complex, as it is characterized not only by the regression coefficients, but also by the variance-covariance structure which is often of primary interest, so the effect of selection on the variance-covariance structure must be carefully assessed; (*iii*) the selection process modifies the hierarchical structure of the data (number of clusters and cluster sizes), a feature that is relevant in the estimation phase, as it influences the behavior of the estimation algorithms, the accuracy of the asymptotic approximations and the power of the tests.

The purpose of the work is to investigate the effects of sample selection in multilevel models. The treatment of the selection problem in the paper is quite general in several respects: (*i*) the selection mechanism is driven by unobserved factors (errors) at both levels; (*ii*) the errors determining the selection are distinct from the errors determining the outcome (though they are allowed to be perfectly correlated); (*iii*) the missingness pattern is arbitrary; (*iv*) the analysis concerns the effect of selection on the properties of the model, rather than on specific estimators.

The paper is organized as follows. Section 2 presents the model and the selection mechanism. Section 3 reports theoretical results on sample selection in mixed models, going from broad properties to analytical formulae. Section 4 reports evidence from a simulation study that illustrates the theoretical results and gives further insight into the topic. In Section 5 the main findings are summarized. Technical details are arranged in two appendices: Appendix A reviews some properties of Skew-Normal distributions and reports the proofs of three results of Section 3.2, while Appendix B shows how to derive the formulae of Section 3.4.

2. The model and the selection mechanism

2.1. The bivariate linear random intercept model

Let us denote the response variables as Y^S and Y^P, where S stands for $Selection$ and P for $Principal$, i.e. the variable of main interest. A selection mechanism is assumed such that Y^P is observed depending on the value of Y^S.

The model is made of a couple of linear equations:

$$\begin{array}{rcl} Y_{ij}^S & = & \mathbf{z}_{ij}^S\boldsymbol{\theta}^S + u_j^S + e_{ij}^S \\ Y_{ij}^P & = & \mathbf{z}_{ij}^P\boldsymbol{\theta}^P + u_j^P + e_{ij}^P, \end{array} \tag{1}$$

where $j = 1, 2, \ldots, J$ is the cluster (level 2) index and $i = 1, 2, \ldots, n_j$ is the elementary (level 1) index: for example, in a panel setting the level 1 units are waves and the level 2 units are individuals, while in a cross-section framework the level 1 units could be individuals and the level 2 units could be institutions or geographical areas. Moreover, $\mathbf{z}_{ij}$ are covariates at level 1 or level 2 and $\boldsymbol{\theta}$ are the corresponding regression coefficients. Each covariate may enter one or both equations. Finally, u_j are level 2 errors, also called random effects, while e_{ij} are level 1 errors.

Errors at different levels are assumed to be independent, while at each level the errors are assumed to be iid and independent of the covariates. Though the general discussion on the consequences of sample selection does not rely on distributional assumptions, the analytical developments will be based on multivariate Normality:

$$\left[\begin{array}{c} e_{ij}^S \\ e_{ij}^P \end{array}\right] \overset{iid}{\sim} N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma_S^2 & \\ \sigma_{SP} & \sigma_P^2 \end{array}\right]\right) \tag{2}$$

$$\left[\begin{array}{c} u_j^S \\ u_j^P \end{array}\right] \overset{iid}{\sim} N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \tau_S^2 & \\ \tau_{SP} & \tau_P^2 \end{array}\right]\right). \tag{3}$$

Assuming that only the sign of Y_{ij}^S is observable, as usual in selection models, the *Selection* equation is a binary probit, so σ_S^2 is not identified (the probit specification implies $\sigma_S^2 = 1$).

In the literature on panel data the model just outlined leads to the so-called *random effects* estimator, in contrast to the *fixed effects* estimator, which is associated with a variant of the model where the level 2 errors are treated as parameters (Wooldridge, 2002).

In the following the conditioning on the covariates is always implicit, so variances and covariances are in fact residual. The terms *marginal* and *conditional* are referred to the random effects.

The marginal variances and covariance are decomposed into level 2 and level 1 components:

$$\begin{array}{lclcl} \text{var}(Y_{ij}^S) & = & \text{var}(u_j^S) + \text{var}(e_{ij}^S) & = & \tau_S^2 + \sigma_S^2 \\ \text{var}(Y_{ij}^P) & = & \text{var}(u_j^P) + \text{var}(e_{ij}^P) & = & \tau_P^2 + \sigma_P^2 \\ \text{cov}(Y_{ij}^S, Y_{ij}^P) & = & \text{cov}(u_j^S, u_j^P) + \text{cov}(e_{ij}^S, e_{ij}^P) & = & \tau_{SP} + \sigma_{SP} \end{array}$$

and the marginal correlation among the responses is $(\tau_{SP}+\sigma_{SP})/\sqrt{(\tau_S^2+\sigma_S^2)(\tau_P^2+\sigma_P^2)}$. For each response, the Intraclass Correlation Coefficient (ICC) is the proportion of variance due to clustering:

$$ICC_S = \tau_S^2/(\tau_S^2 + \sigma_S^2) \tag{4}$$

$$ICC_P = \tau_P^2/(\tau_P^2 + \sigma_P^2). \tag{5}$$

2.2. The selection mechanism

Let the variable of interest Y^P be observed if and only if the value of the selection variable Y^S is greater than zero:

$$Y_{ij}^P \text{ is observed if and only if } Y_{ij}^S > 0\ . \tag{6}$$

This kind of selection operates at level 1, as it causes the missingness of single observations (even when σ_{SP} is null, as in many models for panel or longitudinal data). Note that within a given cluster the pattern of missingness can be of any kind ("non-monotone missingness"), while in many studies attention is restricted to the special case of drop-out or attrition, where missingness at a given occasion implies missingness at all subsequent occasions (Little and Rubin, 2002).

A selection mechanism that acts on the level 1 units modifies the hierarchical structure of the data in terms of the cluster sizes and possibly also in terms of the number of clusters. The probability that a whole cluster is eliminated depends on various factors, such as the size of the clusters, the power of selection (determined by the fixed part of the *Selection* equation), and the ICC of the *Selection* equation. Specifically, leaving other factors unchanged, an increase in the ICC of the *Selection* equation leads to an higher probability of whole cluster exclusion.

Other selection mechanisms are possible. For example, a selection mechanism acting at level 2 can be modelled through a *Selection* equation defined at level 2, so that all the level 1 units belonging to a cluster with $Y^S \leq 0$ have a missing Y^P. This kind of selection is simpler as it depends only on level 2 factors, though it is not possible in general to evaluate if it is more or less harmful than a mechanism that selects the level 1 units.

The selection mechanism generates missing data on Y^P. Since the focus of the analysis is on the *Principal* equation, the key point is whether the selection mechanism is ignorable (Little and Rubin, 2002).

In fact, under ignorable selection the analysis can be performed by fitting only the *Principal* equation, otherwise some procedure for selection-bias correction must be set up. Assuming a likelihood inference framework and the usual separability condition on the parameters, the selection mechanism here described is ignorable when both $u_j^S \perp\!\!\!\perp u_j^P$ and $e_{ij}^S \perp\!\!\!\perp e_{ij}^P$, i.e. under Normality when both covariances σ_{SP} and τ_{SP} are null: in this case the models for the *Selection* and *Principal* equations can be fitted separately, without any bias or loss of efficiency. The ignorable selection mechanism is MCAR when $\boldsymbol{\theta}^S = \mathbf{0}$ and MAR when $\boldsymbol{\theta}^S \neq \mathbf{0}$.

When the selection mechanism is not ignorable it is of interest to determine the biases arising when fitting the *Principal* equation alone.

3. Selection bias in the linear random intercept model

To investigate the consequences of the sample selection mechanism (6) on the *Principal* equation of model (1), first note that Y_{ij}^P is observed if and only if $w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S$, where

$$w_{ij}^S = u_j^S + e_{ij}^S$$

is the composite error of the *Selection* equation. Therefore, in the present context, the term “after selection” means “conditional on truncation on the composite errors w_{ij}^S”.

Let us consider the i-th level 1 unit of a certain cluster j, assuming that its response Y_{ij}^P is observed. To derive the properties of the *Principal* equation of model (1) after selection, the observations pertaining to other clusters are irrelevant, as independence is assumed among clusters. The relevant variables are thus the two errors in Y_{ij}^P, namely u_j^P and e_{ij}^P, plus all the composite errors determining selection in the cluster under consideration, namely $(w_{1j}^S, \ldots, w_{n_j j}^S)$.

Truncation is below for the level 1 units which are observed and above for the others. Therefore the set of truncation events of the whole cluster is:

$$A_j \;=\; \left\{ \bigcap_{h:Y_{hj}^P\ observed} \left\{ w_{hj}^S > -\mathbf{z}_{hj}^S \boldsymbol{\theta}^S \right\} \right\} \bigcap \left\{ \bigcap_{h:Y_{hj}^P\ missing} \left\{ w_{hj}^S \leq -\mathbf{z}_{hj}^S \boldsymbol{\theta}^S \right\} \right\}. \quad (7)$$

Note that A_j is a function of: (*i*) the cluster size; (*ii*) the missingness pattern of the cluster, with $2^{n_j - 1}$ admissible patterns (since the response Y_{ij}^P of the i-th unit is assumed to be observed); (*iii*) all the covariates of the *Selection* equation for all the level 1 units of the cluster, $\{\mathbf{z}_{1j}^S, \ldots, \mathbf{z}_{n_j j}^S\}$; (*iv*) the regression coefficients of the *Selection* equation $\boldsymbol{\theta}^S$. In general, each cluster has a different A_j.

For convenience, let us define also the truncation event for the i-th level 1 unit of the j-th cluster, which is assumed to be observed:

$$A_{ij} = \{w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S\}. \quad (8)$$

For the following discussion it is essential to realize that conditioning on A_{ij} implies a dependence on the features of the level 1 unit under consideration, while conditioning on A_j generates dependence on the features of the whole cluster the unit belongs to. In the following it will be shown that the relevant conditioning set is A_j, even if in a special case the conditioning on A_j is the same as the conditioning on A_{ij}.

3.1. Potential consequences of sample selection

To evaluate the effect of sample selection on the model for Y_{ij}^P, it is necessary to calculate the mean and variance of Y_{ij}^P after truncation, i.e. conditional on A_j.

The mean of Y_{ij}^P can be defined marginally or conditionally w.r.t. the random effects (the conditioning on the covariates being implicit). It is well known that in linear mixed models the marginal and conditional regression coefficients coincide, i.e. a change in a covariate has the same effect on the marginal mean and on the conditional mean of the response. However, such equivalence may break down due to sample selection, so the marginal and the conditional coefficients must be treated separately. Before proceeding it is essential to realize that, even if in mixed models the regression coefficients have a conditional interpretation, in the linear case the usual likelihood estimation methods (ML and REML) are based on the closed-form marginal distribution of the response, so the *estimated coefficients* are actually *marginal coefficients*.

Therefore, the derivation of the coefficients being estimated in the model affected by sample selection requires the calculation of the mean of Y_{ij}^P conditional on A_j, but marginal w.r.t. the random effects u_j^P:

$$E\left(Y_{ij}^P \mid A_j\right) = \mathbf{z}_{ij}^P \boldsymbol{\theta}^P + E\left(u_j^P \mid A_j\right) + E\left(e_{ij}^P \mid A_j\right). \tag{9}$$

The slope of the k-th covariate z_{kij}, marginal w.r.t. u_j^P, is

$$\frac{\partial E\left(Y_{ij}^P \mid A_j\right)}{\partial z_{kij}} = \theta_k^P + \frac{\partial E\left(u_j^P \mid A_j\right)}{\partial z_{kij}} + \frac{\partial E\left(e_{ij}^P \mid A_j\right)}{\partial z_{kij}}. \tag{10}$$

As in non mixed models, sample selection leads to covariate effects which vary from unit to unit, while the model assumes constant slopes, so the estimated slope of a covariate is an average of the slope on each unit.

The sum of the two right-most terms of (10) represents the selection bias, i.e. the difference between the actual coefficient in presence of selection and the corresponding "true" coefficient θ_k^P. The bias is given by the sum of a level 2 component and a level 1 component. If the two components have the same sign, their effects add up, otherwise they partially or totally cancel out.

Even when the selection mechanism is not ignorable, the bias on the coefficient of z_{kij} is null if z_{kij} is not present in the *Selection* equation, since in that case z_{kij} is not included in A_j and thus the derivatives in the right hand side of (10) are both null. Notwithstanding, if z_{kij} is correlated with other covariates that appear in both equations, its slope will be estimated with bias (Wooldridge, 2002).

Since the bias terms are functions of A_j, for a given level 1 unit the selection bias on the regression coefficients depends on the cluster size, on the missingness pattern, on the covariates $\mathbf{z}_{ij}^S$ of the unit under consideration and on the covariates $\mathbf{z}_{hj}^S$, $h \neq i$, of the other units of the cluster. Note that $E(e_{ij}^P \mid A_j)$ varies for each level 1 unit. On the contrary, $E(u_j^P \mid A_j)$ is the same for all the units of the cluster, but its derivative in (10) depends on the value of z_k, so it is different for each unit when z_k is a level 1 covariate, while it is constant within the cluster when z_k is a level 2 covariate.

The dependence of the mean of the errors on the covariates is known as endogeneity, so sample selection can be seen as a source of endogeneity.

In the multilevel framework it is customary to search for random slopes, i.e. slopes varying among clusters, so it is possible to find fictitious random slopes if the cluster component of the variability of the covariate effect is relevant. Therefore, ignoring selection may lead to an incorrect specification with random slopes.

Even if the estimated slopes are the marginal ones, it is instructive to look also at the conditional ones. The mean of Y_{ij}^P conditional on u_j^P and A_j is

$$E\left(Y_{ij}^P \mid u_j^P, A_j\right) \quad = \quad \mathbf{z}_{ij}^P \boldsymbol{\theta}^P + u_j^P + E\left(e_{ij}^P \mid u_j^P, A_j\right), \tag{11}$$

so the slope of the k-th covariate z_{kij} conditional on u_j^P is

$$\frac{\partial E\left(Y_{ij}^P \mid u_j^P, A_j\right)}{\partial z_{kij}} = \theta_k^P + \frac{\partial E\left(e_{ij}^P \mid u_j^P, A_j\right)}{\partial z_{kij}}. \tag{12}$$

Comparing expressions (12) and (10) it is clear that after selection the conditional and marginal slopes are different.

Contrary to non mixed models, where the residual variance usually is a nuisance, in multilevel models the variance structure is of primary interest. It is then relevant to assess the effect of selection on the model variances. The residual variance of Y_{ij}^P after truncation is

$$Var\left(Y_{ij}^P \mid A_j\right) = Var\left(u_j^P \mid A_j\right) + Var\left(e_{ij}^P \mid A_j\right) + 2Cov\left(u_j^P, e_{ij}^P \mid A_j\right). \tag{13}$$

After selection the variance component structure breaks down: in general, the errors of the *Principal* equation are no longer homoscedastic, nor independent, giving rise to inefficient estimators and incorrect standard errors. The crucial point is that the ICC is no longer defined after selection, since the decomposition of the variance of Y^P requires homoscedastic and independent errors. If the ICC of the *Principal* equation (5) is estimated from a misspecified model ignoring selection, one can reach false conclusions about the role of the hierarchical structure.

To summarize, sample selection in multilevel models modifies the distribution of the model errors, leading to a complex configuration where the basic model assumptions break down, causing bias in both slopes and variance components. Unfortunately, in general there are as many bias formulas as the number of admissible missingness patterns, so, even in a balanced hierarchy, it is not feasible to write down such formulae, except when the cluster size is small. In addition, the moments (9) and (13) have quite complex expressions for clusters of size greater than two. In Section 3.4 the expressions for the special case of a balanced hierarchy with $n_j = 2$ are shown.

In special cases some of the potential biases caused by sample selection do not operate and the calculation of means and variances becomes simple. In particular, the independencies reported in Table 1 imply important simplifications. The moments in Table 1 are derived by exploiting the fact that A_j is a function of $(w_{1j}^S, \ldots, w_{n_j j}^S)$ and the property that $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$ implies $g_1(\mathbf{x}_1) \perp\!\!\!\perp g_2(\mathbf{x}_2)$ for arbitrary random vectors and functions.

Independencies (a1) and (a2) characterize instances where truncation is irrelevant in certain respects, while independence (b) means that truncation does not corrupt the independence between errors at different levels in the *Principal* equation. Combining the independencies of Table 1 leads to important special cases. For example, when both (a2) and (b) hold the conditional and marginal slopes are equal.

Another interesting question is whether there exist non trivial cases where it is enough to condition on the truncation event of the unit under consideration, i.e. the distribution of the errors conditional on A_j is the same as the distribution conditional on A_{ij}. This would lead to simpler patterns and formulae.

To establish when the independencies of Table 1 hold and when the conditioning reduces to A_{ij}, it is necessary to investigate the joint distribution of the errors before and after selection.

Table 1. Independencies among the model errors and consequences on some of their moments

Independence	Moments
(a1) $e_{ij}^P \perp\!\!\!\perp (w_{1j}^S, \ldots, w_{n_j j}^S)$	$E\;\; e_{ij}^P \mid A_j \;\; = E\;\; e_{ij}^P \;\; = 0$ $Var\;\; e_{ij}^P \mid A_j \;\; = Var\;\; e_{ij}^P \;\; = \sigma_P^2$
(a2) $u_j^P \perp\!\!\!\perp (w_{1j}^S, \ldots, w_{n_j j}^S)$	$E\;\; u_j^P \mid A_j \;\; = E\;\; u_j^P \;\; = 0$ $Var\;\; u_j^P \mid A_j \;\; = Var\;\; u_j^P \;\; = \tau_P^2$
(b) $e_{ij}^P \perp\!\!\!\perp u_j^P \mid A_j$	$E\;\; e_{ij}^P \mid u_j^P, A_j \;\; = E\;\; e_{ij}^P \mid A_j$ $Cov\;\; u_j^P, e_{ij}^P \mid A_j \;\; = 0$

3.2. Independencies among the errors before and after selection

Let us consider the joint distribution of $(w_{ij}^S, \mathbf{w}_{(i)j}^S, u_j^P, e_{ij}^P, \mathbf{e}_{(i)j}^P)$, where $\mathbf{w}_{(i)j}^S = (w_{1j}^S, \ldots, w_{(i-1)j}^S, w_{(i+1)j}^S, \ldots, w_{n_j j}^S)$ and $\mathbf{e}_{(i)j}^P = (e_{1j}^P, \ldots, e_{(i-1)j}^P, e_{(i+1)j}^P, \ldots, e_{n_j j}^P)$. Under assumptions (2)-(3) the distribution before selection is multivariate Normal with zero means and the following covariance matrix:

$$
Var \left[\begin{array}{c} w_{1j}^S \\ \vdots \\ w_{ij}^S \\ \vdots \\ w_{n_j j}^S \\ \hline u_j^P \\ e_{1j}^P \\ \vdots \\ e_{ij}^P \\ \vdots \\ e_{n_j j}^P \end{array} \right] = \left[\begin{array}{ccccc|cccccc}
\tau_S^2 + \sigma_S^2 & \cdots & \tau_S^2 & \cdots & \tau_S^2 & \tau_{SP} & \sigma_{SP} & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\tau_S^2 & \cdots & \tau_S^2 + \sigma_S^2 & \cdots & \tau_S^2 & \tau_{SP} & 0 & \cdots & \sigma_{SP} & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\tau_S^2 & \cdots & \tau_S^2 & \cdots & \tau_S^2 + \sigma_S^2 & \tau_{SP} & 0 & \cdots & 0 & \cdots & \sigma_{SP} \\
\hline
\tau_{SP} & \cdots & \tau_{SP} & \cdots & \tau_{SP} & \tau_P^2 & 0 & \cdots & 0 & \cdots & 0 \\
\sigma_{SP} & \cdots & 0 & \cdots & 0 & 0 & \sigma_P^2 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & \sigma_{SP} & \cdots & 0 & 0 & 0 & \cdots & \sigma_P^2 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & \cdots & \sigma_{SP} & 0 & 0 & \cdots & 0 & \cdots & \sigma_P^2
\end{array} \right] \tag{14}
$$

In the following it is assumed that σ_S^2, σ_P^2 and τ_P^2 are strictly positive, while for τ_S^2 two cases are considered: (*i*) $\tau_S^2 > 0$, i.e. the *Selection* equation is mixed; (*ii*) $\tau_S^2 = 0$, i.e. the *Selection* equation is not mixed. Note that if $\tau_S^2 = 0$ then $\tau_{SP} = 0$.

For future reference, let us summarize the given assumptions as follows:

ASSUMPTION 1. *The random vector $(w_{ij}^S, \mathbf{w}_{(i)j}^S, u_j^P, e_{ij}^P, \mathbf{e}_{(i)j}^P)$ has a multivariate Normal distribution with zero means and covariance matrix (14), with $\sigma_S^2 > 0$, $\sigma_P^2 > 0$ and $\tau_P^2 > 0$.*

Under Normality the zero entries in the covariance matrix represent unconditional independencies. In particular:

$$e_{ij}^P \;\perp\!\!\!\perp\; u_j^P \tag{15}$$

$$e_{ij}^P \;\perp\!\!\!\perp\; \mathbf{w}_{(i)j}^S \tag{16}$$

3.2.1. Basic results on independencies among the model errors

The target now is to determine the values of the variance-covariance parameters for which the independencies of Table 1 hold.

Since the errors are jointly Normally distributed, the unconditional independencies can be read from the covariance matrix. Therefore, the cases where independencies (a1) and (a2) hold can be identified by looking at the covariance matrix (14):

RESULT 1. *Under Assumption 1, independence* (a1) $e_{ij}^P \perp\!\!\!\perp (w_{ij}^S, \mathbf{w}_{(i)j}^S)$ *holds if and only if* $\sigma_{SP} = 0$.

RESULT 2. *Under Assumption 1, independence* (a2) $u_j^P \perp\!\!\!\perp (w_{ij}^S, \mathbf{w}_{(i)j}^S)$ *holds if and only if* $\tau_{SP} = 0$.

In the light of (15), (a1) implies that e_{ij}^P is independent of all the other model errors, so it is independent of u_j^P also after truncation. Similarly, (a2) implies that u_j^P is independent of all the other model errors, , so it is independent of e_{ij}^P also after truncation. Therefore, if $\tau_{SP} = 0$ or $\sigma_{SP} = 0$, then $e_{ij}^P \perp\!\!\!\perp u_j^P \mid A_j$.

When the *Selection* equation is not mixed ($\tau_S^2 = 0$) then $\tau_{SP} = 0$ and the following Result holds:

RESULT 3. *Under Assumption 1 and* $\tau_S^2 = 0$*, independence* (b) $e_{ij}^P \perp\!\!\!\perp u_j^P \mid A_j$ *holds.*

When the *Selection* equation is mixed, $\tau_{SP} = 0$ or $\sigma_{SP} = 0$ are sufficient conditions for independence (b) $e_{ij}^P \perp\!\!\!\perp u_j^P \mid A_j$. However, it is not straightforward to prove that such conditions are also necessary since, given a set of jointly Normal random variables, after truncation on some components the joint distribution of the subset of the non truncated variables is no more Normal. Nevertheless, such distribution is a member of the Unified Skew-Normal (SUN) family of Arellano-Valle and Azzalini (2006), who proved some properties that can be used to derive the conditions for independence (b). In Appendix A the following Result is proved:

RESULT 4. *Under Assumption 1 and* $\tau_S^2 > 0$*, independence (b)* $e_{ij}^P \perp\!\!\!\perp u_j^P \mid A_j$ *holds if and only if either* $\tau_{SP} = 0$ *or* $\sigma_{SP} = 0$ *or both.*

To put another way, u_j^P is no longer independent of e_{ij}^P after selection if and only if $\tau_{SP} \neq 0$ and $\sigma_{SP} \neq 0$, i.e. selection depends on unobservables at both levels.

3.2.2. Relevant truncation events

When studying sample selection in mixed models, it is necessary to condition on the truncation events of all the units of the cluster, A_j. A basic question is whether there are cases where the conditioning on the truncation event of the unit under consideration, A_{ij}, is equivalent to the conditioning on A_j.

When the *Selection* equation is not mixed ($\tau_S^2 = 0$) $\tau_{SP} = 0$ and $\mathbf{w}_{(i)j}^S$ is independent of the remaining errors, so the distribution of u_j^P is not affected by truncation at all, while for the distribution of e_{ij}^P conditioning on A_{ij} is enough. This is stated in the following Result, where the symbol $\sim$ stands for *equal distributions*):

RESULT 5. *Under Assumption 1 and* $\tau_S^2 = 0$*,* $u_j^P \mid A_j \sim u_j^P$ *and* $e_{ij}^P \mid A_j \sim e_{ij}^P \mid A_{ij}$.

On the other hand, the properties of the SUN distribution allow to prove that when the *Selection* equation is mixed the following results hold (see Appendix A):

RESULT 6. *Under Assumption 1 and* $\tau_S^2 > 0$, $e_{ij}^P \mid A_j \sim e_{ij}^P \mid A_{ij}$ *if and only if* $\sigma_{SP} = 0$.

RESULT 7. *Under Assumption 1 and* $\tau_S^2 > 0$, $u_j^P \mid A_j \sim u_j^P \mid A_{ij}$ *if and only if* $\tau_{SP} = 0$.

Results 5 and 6 hold also conditionally on u_j^P.

When the *Selection* equation is mixed the only way to let u_j^P or e_{ij}^P be independent of $\mathbf{w}_{(i)j}^S$ after truncation on w_{ij}^S is to to remove the corresponding covariances: if $\tau_{SP} = 0$ then u_j^P is independent of the remaining variables, so its distribution is not affected by truncation at all; similarly, if $\sigma_{SP} = 0$ then e_{ij}^P is independent of the remaining variables, so its distribution is not affected by truncation at all.

To summarize, when the *Selection* equation is mixed either truncation is irrelevant or all the truncation events of the cluster must be considered.

3.2.3. Independence among the level 1 errors

Another kind of independence which may be corrupted after selection is the one among the level 1 errors e_{ij}^P of the same cluster.

When the *Selection* equation is not mixed ($\tau_S^2 = 0$) then for any i the couple (w_{ij}^S, e_{ij}^P) is independent of all the other model errors, yielding:

RESULT 8. *Under Assumption 1 and* $\tau_S^2 = 0$, *independence* $e_{ij}^P \perp\!\!\!\perp e_{i'j}^P \mid A_j$ *holds for any* $i \neq i'$.

In Appendix A the following result is proved:

RESULT 9. *Under Assumption 1 and* $\tau_S^2 > 0$, *independence* $e_{ij}^P \perp\!\!\!\perp e_{i'j}^P \mid A_j$ *holds for any* $i \neq i'$ *if and only if* $\sigma_{SP} = 0$.

It is easy to check that Results 8 and 9 also hold conditionally on (u_j^P, A_j), though Result 9 would be different when $\tau_{SP}^2 = \tau_S^2 \times \tau_P^2$. In such a case, called *shared parameter model* (Follmann and Wu, 1995), $u_j^P = u_j^S = u_j$, so conditioning on u_j makes the composite errors of the *Selection* equation independent, i.e. $w_{ij}^S \perp\!\!\!\perp w_{i'j}^S \mid u_j, \forall i \neq i'$. Therefore, conditional on u_j, one gets the same result as if the *Selection* equation was not mixed, namely $e_{ij}^P \perp\!\!\!\perp e_{i'j}^P \mid (u_j, A_j)$ $\forall i \neq i'$, even if $\sigma_{SP} \neq 0$.

Note that when $e_{ij}^P \not\perp\!\!\!\perp e_{i'j}^P \mid (u_j^P, A_j)$, the usual factorization of the conditional likelihood (e.g. Skrondal and Rabe-Hesketh, 2004) is no longer valid.

3.3. Selection bias: special cases

The results of the previous Section allow to establish the behavior of selection bias in the general case and three special cases of Table 2. The configuration $\sigma_{SP} = 0$ and $\tau_{SP} \neq 0$ is often assumed in models for panel or longitudinal data, but it is unrealistic in a cross-section setting.

First of all, when the *Selection* equation is mixed either truncation is irrelevant or all the truncation events must be considered. Sample selection causes a bias in the slopes of covariates entering both the *Principal* and the *Selection* equations. The bias has an additive structure with a cluster-level component depending on the covariance τ_{SP} among cluster-level errors and a subject-level component depending on the covariance σ_{SP} among subject-level errors. Moreover, the marginal slope is different from the conditional slope whenever $\tau_{SP} \neq 0$, i.e. in the general case and special case *b*.

Table 2. Consequences of selection for the general case and three special cases.

	General case	*Case a*	*Case b*	*Case c*
	$\tau_S^2 > 0$ $\tau_{SP} \neq 0$ $\sigma_{SP} \neq 0$	$\tau_S^2 > 0$ $\tau_{SP} = 0$ $\sigma_{SP} \neq 0$	$\tau_S^2 > 0$ $\tau_{SP} \neq 0$ $\sigma_{SP} = 0$	$\tau_S^2 = 0$ $\tau_{SP} = 0$ $\sigma_{SP} \neq 0$
one-element truncation A_{ij}	no	no	no	yes
Slope bias	$\frac{\partial E(e_{ij}^P \mid A_j)}{\partial z_{kij}} + \frac{\partial E(u_j^P \mid A_j)}{\partial z_{kij}}$	$\frac{\partial E(e_{ij}^P \mid A_j)}{\partial z_{kij}}$	$\frac{\partial E(u_j^P \mid A_j)}{\partial z_{kij}}$	$\frac{\partial E(e_{ij}^P \mid A_{ij})}{\partial z_{kij}}$
marginal slope = conditional slope	no	yes	no	yes
$e_{ij}^P \perp\!\!\!\perp u_j^P \mid A_j$	no	yes	yes	yes
$e_{ij}^P \perp\!\!\!\perp e_{i'j}^P \mid A_j$	no	no	yes	yes
Bias on σ_P^2	downward	downward	no	downward
Bias on τ_P^2	?	?	downward	no
Bias on ICC_P	?	?	downward	upward

As regards the structure of the errors, in cases *b* and *c* the key independencies $e_{ij}^P \perp\!\!\!\perp u_j^P$ and $e_{ij}^P \perp\!\!\!\perp e_{i'j}^P$ still hold after selection, so the errors decomposition $u_j^P + e_{ij}^P$ still implies the variance decomposition $Var(u_j^P + e_{ij}^P) = Var(u_j^P) + Var(e_{ij}^P)$, even if the errors are no more homoscedastic. In such cases it is straightforward to show the effect of selection on the variances. In fact, under Normality, truncation reduces the variances (e.g. Arellano-Valle and Azzalini, 2006), so $\sigma_{SP} \neq 0$ implies $Var(e_{ij}^P \mid A_j) < \sigma_P^2$ and $\tau_{SP} \neq 0$ implies $Var(u_j^P \mid A_j) < \tau_P^2$. Moreover, from Table 1 and Results 1 and 2, $\sigma_{SP} = 0$ implies $Var(e_{ij}^P \mid A_j) = \sigma_P^2$ and $\tau_{SP} = 0$ implies $Var(u_j^P \mid A_j) = \tau_P^2$. Consequently, in case *b* of Table 2 the ICC is overestimated, while in case *c* the ICC is underestimated.

In case *a* the independence $e_{ij}^P \perp\!\!\!\perp e_{i'j}^P$ does not hold after selection, while in the *general case* also the independence $e_{ij}^P \perp\!\!\!\perp u_j^P$ does not hold after selection. In both cases the variance decomposition $Var(u_j^P + e_{ij}^P) = Var(u_j^P) + Var(e_{ij}^P)$ does not hold and the ICC is meaningless.

A special instance of case *a* is when the *Principal* equation is not mixed. In such a case the true ICC is null, but the correlation among the level 1 errors after selection may lead to a significant ICC. This case is particularly interesting as it shows that the correlation among the observations may be entirely due to the selection process.

3.4. Analytic expressions of bias

Assuming multivariate Normality before truncation, the analytical expressions of the means and variances of the errors after selection can be derived through the formulae of Johnson and Kotz (1972) and Tallis (1961) as reported in Appendix B.

In the following we consider two cases: (*i*) the cluster size n_j is arbitrary, but the relevant truncation set has only one element; and (*ii*) the cluster size is $n_j = 2$ for all the clusters.

3.4.1. Bias when the *Selection* equation is not mixed

The only case where truncation is relevant and the truncation set A_j reduces to the one-element set A_{ij} is when the *Selection* equation is not mixed, i.e $\tau_S^2 = 0$ and thus $\tau_{SP} = 0$. In such a case the results of Section 3.2 imply:

$$E\left(Y_{ij}^P \mid A_{ij}\right) = \mathbf{z}_{ij}^P \boldsymbol{\theta}^P + E\left(e_{ij}^P \mid A_{ij}\right) \tag{17}$$

$$Var\left(Y_{ij}^P \mid A_{ij}\right) = \tau_P^2 + Var\left(e_{ij}^P \mid A_{ij}\right). \tag{18}$$

The mean and variance of e_{ij}^P given A_{ij} can be derived using formulae (28) and (29) of Appendix B, where $\mathbf{U} = e_{ij}^P$, $\mathbf{W} = w_{ij}^S$ and 'selection on $\mathbf{W}$' replaced with $A_{ij} = \{w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S\}$. Taking into account that $\tau_S^2 = 0$, the well-known expressions of Heckman (1979) are obtained:

$$E\left[e_{ij}^P \mid w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S\right] = \frac{\sigma_{SP}}{\sqrt{\sigma_S^2}} \cdot \lambda\left(\frac{-\mathbf{z}_{ij}^S \boldsymbol{\theta}^S}{\sqrt{\sigma_S^2}}\right) \tag{19}$$

$$Var\left[e_{ij}^P \mid w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S\right] = \sigma_P^2 - \frac{\left(\sigma_{SP}\right)^2}{\sigma_S^2} \cdot \delta\left(\frac{-\mathbf{z}_{ij}^S \boldsymbol{\theta}^S}{\sqrt{\sigma_S^2}}\right), \tag{20}$$

where $\lambda(x) = \phi(x)/\left[1 - \Phi(x)\right]$ is the *inverse Mills ratio* (the hazard function of the standard Normal distribution) and $\delta(x) = \lambda(x)\left[\lambda(x) - x\right]$ is its first derivative. It can be shown that $\lambda(x)$ and $\delta(x)$ are increasing functions, with $\lambda(x) > 0$ and $0 < \delta(x) < 1$.

From (17) and (19) the marginal effect of the k-th covariate z_{kij} after selection is:

$$\frac{\partial E\left(Y_{ij}^P \mid A_{ij}\right)}{\partial z_{kij}} = \theta_k^P - \frac{\sigma_{SP}}{\sigma_S^2} \cdot \delta\left(\frac{-\mathbf{z}_{ij}^S \boldsymbol{\theta}^S}{\sqrt{\sigma_S^2}}\right) \cdot \theta_k^S. \tag{21}$$

In Section 3.2.1 it was shown that when the *Selection* equation is not mixed the conditional and marginal means coincides, so in this case expression (21) is valid for the conditional effect too.

The sign of the bias of the slope is determined by the sign of the product $\sigma_{SP}\theta_k^S$, while the magnitude of the bias depends on: (*i*) the force of the selection mechanism, σ_{SP}/σ_S^2; (*ii*) the probability of missingness, determined by the linear predictor of the *Selection* equation (the lower the linear predictor, the higher the probability of missingness and the higher the value of the δ function); (*iii*) the slope of the covariate under consideration in the *Selection* equation, θ_k^S.

From (18) and (20) the variance of the response after selection is

$$Var\left(Y_{ij}^P \mid A_{ij}\right) \quad = \quad \tau_P^2 + \sigma_P^2 - \frac{\left(\sigma_{SP}\right)^2}{\sigma_S^2} \cdot \delta\left(\frac{-\mathbf{z}_{ij}^S \boldsymbol{\theta}^S}{\sqrt{\sigma_S^2}}\right).$$

Since δ ranges from 0 to 1, the selection reduces the level 1 variance and thus leads to overestimation of the ICC. The reduction of the level 1 variance depends on the magnitude, but not on the sign of σ_{SP}. Since δ is a function of the linear predictor of the *Selection* equation, selection induces heteroscedasticity at level 1.

3.4.2. Bias in the case of a cluster of size $n_j = 2$

In a cluster of size two, i.e. $n_j = 2$, the relevant missingness patterns are only two, namely both responses Y^P on the *Principal* equation are observed or only one is observed. Moreover, the formulae for the means of the truncated bivariate Normal distribution (Tallis, 1961) are reasonably manageable, allowing to generalize the well-known formula (21).

The generalized formulae are useful to study the bias analytically for a *balanced hierarchy*, i.e. when $n_j = 2$ for any cluster j. This case arises, for example, in panel studies with two waves and in some special cross-section applications, e.g. studies on twins.

Without loss of generality, formulae are shown for $i = 1$, namely Y_{1j}^P is assumed to be observed, while Y_{2j}^P can be observed (*Pattern 1*) or missing (*Pattern 2*). As shown in Appendix B, when both responses are observed (*Pattern 1*) the means of the errors of the *Principal* equation after truncation are

$$E\left[\left(\begin{array}{c} u_j^P \\ e_{1j}^P \end{array}\right) \mid w_{1j}^S > -\mathbf{z}_{1j}^S\boldsymbol{\theta}^S, w_{2j}^S > -\mathbf{z}_{2j}^S\boldsymbol{\theta}^S\right] = \left[\begin{array}{c} \frac{\tau_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{u,1}\left(\alpha_{1j},\alpha_{2j},\rho\right) \\ \frac{\sigma_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{e,1}\left(\alpha_{1j},\alpha_{2j},\rho\right) \end{array}\right] \quad (22)$$

and when only the first unit is observed (*Pattern 2*) the means are

$$E\left[\left(\begin{array}{c} u_j^P \\ e_{1j}^P \end{array}\right) \mid w_{1j}^S > -\mathbf{z}_{1j}^S\boldsymbol{\theta}^S, w_{2j}^S \leq -\mathbf{z}_{2j}^S\boldsymbol{\theta}^S\right] = \left[\begin{array}{c} \frac{\tau_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{u,2}\left(\alpha_{1j},\alpha_{2j},\rho\right) \\ \frac{\sigma_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{e,2}\left(\alpha_{1j},\alpha_{2j},\rho\right) \end{array}\right] \quad (23)$$

where $\rho = ICC_S = \tau_S^2/(\tau_S^2+\sigma_S^2) \geq 0$ is the ICC of the *Selection* equation, $\alpha_{ij} = -\mathbf{z}_{ij}^S\boldsymbol{\theta}^S/\sqrt{\tau_S^2+\sigma_S^2}$ is the standardized truncation point of unit ij, and the functions $\lambda_{u,1}, \lambda_{e,1}, \lambda_{u,2}, \lambda_{e,2}$ are bivariate generalizations of the inverse Mills ratio defined in formulae (30) and (31) of Appendix B. When $\rho = 0$ both the functions $\lambda_{e,1}$ and $\lambda_{e,2}$ reduce to the inverse Mills ratio.

When the *Selection* equation is not mixed ($\tau_S^2 = \rho = 0$) then $\tau_{SP} = 0$, so from equations (22) and (23) the mean of u_j^P after truncation is null and the mean of e_{1j}^P after truncation is equal to the classical expression (19) whichever the missingness pattern. Therefore, the bias can be read from expression (21).

When the *Selection* equation is mixed, i.e. $\rho > 0$, from (10) the effect of a covariate z_{k1j} on the marginal mean of Y_{1j}^P after truncation is

$$\frac{\partial E\left(Y_{1j}^P \mid A_j\right)}{\partial z_{k1j}} = \theta_k^P + \frac{\tau_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\frac{\partial \lambda_{u,\#}}{\partial z_{k1j}} + \frac{\sigma_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\frac{\partial \lambda_{e,\#}}{\partial z_{k1j}}, \quad (24)$$

where # denotes the missingness pattern (1 or 2). If the covariate under consideration is *at level 1*, the effect is

$$\frac{\partial E\left(Y_{1j}^P \mid A_j\right)}{\partial z_{k1j}} = \theta_k^P - \frac{\tau_{SP}}{\tau_S^2+\sigma_S^2}\frac{\partial \lambda_{u,\#}}{\partial \alpha_{1j}}\theta_k^S - \frac{\sigma_{SP}}{\tau_S^2+\sigma_S^2}\frac{\partial \lambda_{e,\#}}{\partial \alpha_{1j}}\theta_k^S. \quad (25)$$

However, if the covariate under consideration is *at level 2*, a change in the covariate affects both truncation points α_{1j} and α_{2j}, so the effect is

$$\frac{\partial E\left(Y_{1j}^P \mid A_j\right)}{\partial z_{k1j}} = \theta_k^P - \frac{\tau_{SP}}{\tau_S^2+\sigma_S^2}\left(\frac{\partial \lambda_{u,\#}}{\partial \alpha_{1j}} + \frac{\partial \lambda_{u,\#}}{\partial \alpha_{2j}}\right)\theta_k^S - \frac{\sigma_{SP}}{\tau_S^2+\sigma_S^2}\left(\frac{\partial \lambda_{e,\#}}{\partial \alpha_{1j}} + \frac{\partial \lambda_{e,\#}}{\partial \alpha_{2j}}\right)\theta_k^S. \quad (26)$$

A numerical analysis of the derivatives shows that, for any missingness pattern and any combination of $\rho > 0$, α_{1j} and α_{2j},

$$\frac{\partial \lambda_{u,\#}}{\partial \alpha_{1j}} + \frac{\partial \lambda_{u,\#}}{\partial \alpha_{2j}} > \frac{\partial \lambda_{u,\#}}{\partial \alpha_{1j}} > 0, \qquad 0 < \frac{\partial \lambda_{e,\#}}{\partial \alpha_{1j}} + \frac{\partial \lambda_{e,\#}}{\partial \alpha_{2j}} < \frac{\partial \lambda_{e,\#}}{\partial \alpha_{1j}}.$$

Therefore, regardless of the level of the covariate, the sign of the bias stemming from the level 2 correlation is determined by the product $\tau_{SP}\theta_k^S$, while the sign of the bias stemming from the level

1 correlation is determined by the product $\sigma_{SP}\theta_k^S$. However, for a given couple of values of τ_{SP} and σ_{SP}, with a level 2 covariate the bias stemming from the level 2 correlation is expanded, while the bias stemming from the level 1 correlation is attenuated.

Even in the balanced case, when $n_j = k > 2$ it is impractical to write down the formulae for the bias, since there are many different missingness patterns and the formulae become complex (the means after truncation are given by sums of k terms involving Normal distribution functions of order k and $k-1$).

Finally, it is worth to compare the two considered instances of analytical expressions for the bias on the slope: (*i*) *Selection* equation not mixed and an arbitrary hierarchy, leading to expression (21), and (*ii*) *Selection* equation mixed and a balanced hierarchy with $n_j = 2$, leading to expression (24). First of all, in case (*ii*) there are two additive sources of bias, one for each hierarchical level. Another basic difference is that in case (*i*) there is a single bias formula holding for any cluster, while in case (*ii*) there is a bias formula for each missingness pattern: as a consequence, the actual bias is an average across the patterns and thus depends on their frequencies. Moreover, in case (*ii*) the bias has different expressions for level 1 and level 2 covariates.

4. Simulation

Since simple analytical expressions of selection bias in mixed models exist only in special cases, the evaluation of the bias in more general cases requires simulation experiments.

4.1. Simulation design

The simulation study considers a two-level random intercept linear model, with vectors of covariates $\mathbf{z}_{ij}^S = (x_{1ij}, x_{2ij}, v_j)$ and $\mathbf{z}_{ij}^P = (x_{1ij}, x_{3ij}, v_j)$, and vectors of parameters $\boldsymbol{\theta}^S = (\beta_0^S, \beta_1^S, \beta_2^S, \gamma^S)'$ and $\boldsymbol{\theta}^P = (\beta_0^P, \beta_1^P, \beta_3^P, \gamma^P)'$:

$$\begin{array}{lcl} Y_{ij}^S & = & \beta_0^S + \beta_1^S x_{1ij} + \beta_2^S x_{2ij} + \gamma^S v_j + u_j^S + e_{ij}^S \\ Y_{ij}^P & = & \beta_0^P + \beta_1^P x_{1ij} + \beta_3^P x_{3ij} + \gamma^P v_j + u_j^P + e_{ij}^P . \end{array} \tag{27}$$

The covariate v is at level 2, i.e. it varies only between clusters, while x_1, x_2, x_3 are *purely within* level 1 covariates, i.e. their variation is only within clusters. In a mixed model, purely within covariates are needed to separate the within and between effects (Neuhaus and Kalbfleish, 1998). However, the simulation results can be used to evaluate also the bias on a level 1 covariate that varies both within and between clusters. Indeed, a general level 1 covariate, say z_{ij}, varying within and between clusters, can be written as the sum of two components: $z_{ij} = \overline{z}_j + (z_{ij} - \overline{z}_j)$, where $\overline{z}_j$ is the cluster mean (a level 2 covariate) and $(z_{ij} - \overline{z}_j)$ is the deviation from the cluster mean (a purely within covariate). The bias on the slope of z_{ij} is a mixture of the biases on the slopes of the two components depending on the proportion of variance of z_{ij} due to the clustering. Therefore, letting $v_j = \overline{z}_j$ and $x_{1ij} = (z_{ij} - \overline{z}_j)$, one can appreciate the bias on the general level 1 covariate z_{ij}.

Two of the covariates, x_1 and v, enter both equations, while the other covariates are equation-specific. In this way identification problems are avoided (Wooldridge, 2002).

The errors in model (27) are Normally distributed as in (2) and (3).

The data generation process requires the specification of several aspects: the distribution of the covariates, the parameter values and the hierarchical structure of the units.

The values of the covariates are drawn from independent standard Normal distributions. They are generated only once and used in all the experiments and replications. The level 1 covariates are centered around their sample cluster means to ensure that they are *purely within*.

The values of the *true* parameters used in the experiments are:

- *intercepts*: $\beta_0^S = 0$, $\beta_0^P = 0$;
- *slopes*: $\beta_1^S = \beta_2^S = \gamma^S = 1$, $\beta_1^P = \beta_3^P = \gamma^P = 1$;
- *variances*: $\sigma_S^2 = \tau_S^2 = 1$, $\sigma_P^2 = \tau_P^2 = 1$;
- *correlations*: given unit variances, the level 1 and level 2 correlations are equal to the corresponding covariances σ_{SP} and τ_{SP}. A two-dimensional grid of values is used by letting σ_{SP} and τ_{SP} vary in the interval [-1,+1] with a step of 0.25, for a total of 81 different combinations.

Note that the ICCs (4) and (5) are 0.5, meaning that the clustering of the units is quite relevant, though in a panel setting such a value is considered as moderate.

The value of β_0^S is crucial in determining the proportion of missing responses on the *Principal* equation. Fixing β_0^S to zero leads to a selection that excludes about half of the observations on Y^P. In the set of all the performed simulations the percentage of missingness ranges from 43% to 60%, with a mean of 51%.

Regarding the hierarchical structure of the data, a balanced design is assumed with a total of 5000 observations, arranged in 100 clusters with 50 observations per cluster. This data structure is typical of cross-sectional studies, e.g. in the educational setting. The role of the hierarchical structure is investigated through a specific set of simulations discussed at the end of the Section (see Table 7).

As already noted, a mechanism selecting the level 1 units destroys the balanced structure of the data and, in fact, the data used for the simulation study turn out to be highly unbalanced after selection. In this simulation design the probability of whole cluster exclusion is negligible, so after selection the number of clusters is unchanged.

The data generation process starts with the calculation of the linear predictors $\mathbf{z}_{ij}^S \boldsymbol{\theta}^S$ and $\mathbf{z}_{ij}^P \boldsymbol{\theta}^P$ for each level 1 unit. Then, for each Monte Carlo replication the following steps are performed:

(a) for each cluster the errors u_j^S and u_j^P are drawn from a bivariate normal distribution with zero means, variances τ_S^2 and τ_P^2 and correlation τ_{SP};

(b) for each level 1 unit the errors e_{ij}^S and e_{ij}^P are drawn from a bivariate normal distribution with zero means, variances σ_S^2 and σ_P^2 and correlation σ_{SP};

(c) for each level 1 unit the responses Y_{ij}^S and Y_{ij}^P are calculated as in (27): at the end of this step a data set unaffected by selection is obtained;

(d) for each level 1 unit the response on the *Principal* equation Y_{ij}^P is set to missing if the response on the *Selection* equation Y_{ij}^S is less or equal to zero: at the end of this step a data set affected by selection is obtained.

Given the grid of values for the correlations (see e.g. Table 5), 81 experiments are performed with 1000 replications each.

The estimates are obtained by fitting the *Principal* equation alone on the data affected by sample selection. The estimation method used in the simulation study is REML, with the ridge-stabilized Newton-Raphson algorithm implemented in the MIXED procedure of the SAS software (Littell et al., 2006).

Table 3. Monte Carlo means on 1000 replications of the estimates of σ^2_P for different values of the correlations at level 2 (τ_{SP}) and level 1 (σ_{SP}).

	τ_{SP}								
σ_{SP}	*-1.00*	*-0.75*	*-0.50*	*-0.25*	*0.00*	*0.25*	*0.50*	*0.75*	*1.00*
-1.00	0.78	0.79	0.78	0.79	0.79	0.79	0.79	0.79	0.78
-0.75	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
-0.50	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95
-0.25	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
0.50	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
0.75	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
1.00	0.79	0.78	0.78	0.78	0.79	0.78	0.78	0.79	0.79

4.2. Simulation results

The results of the simulation study outlined in Section 4.1 are shown in Tables 3 to 6 concerning, respectively, the variances σ^2_P and τ^2_P, and the slopes β^P_1 and γ^P. The estimates of the intercept β^P_0 are not reported as they are not of interest, while the estimates of β^P_3 are not reported as the covariate x_3 is independent of the others and does not enter the *Selection* equation, so the bias is null.

Table 3 shows that the estimate of the level 1 variance σ^2_P is affected only by the level 1 correlation σ_{SP}. The bias is downward and depends on the absolute value of σ_{SP}.

Table 4 shows that the estimate of the level 2 variance τ^2_P is affected by both correlations σ_{SP} and τ_{SP}. From the row with $\sigma_{SP} = 0$ it appears that the effect of τ_{SP} is to reduce the estimate of τ^2_P. On the contrary, from the column with $\tau_{SP} = 0$, it appears that the effect of σ_{SP} is to inflate the estimate of τ^2_P. The reason is that when $\sigma_{SP} \neq 0$ level 1 errors e^P_{ij} are no more independent after selection, and indeed the simulations show that they are positively correlated: since such correlation has a level 2 nature it inflates the level 2 variance.

When the correlations σ_{SP} and τ_{SP} are both different from zero, the bias on τ^2_P is not simply the sum of the two biases just outlined, because there is a third source of bias due to the lack of independence after truncation between the level 1 errors e^P_{ij} and the level 2 errors u^P_j. The simulations show that the correlation between e^P_{ij} and u^P_j is negative when σ_{SP} and τ_{SP} have the same sign, while it is positive when σ_{SP} and τ_{SP} have opposite signs. Such correlation has a level 2 nature and thus affects the estimate of τ^2_P as shown in Table 4. The bias on τ^2_P caused jointly by σ_{SP} and τ_{SP} is far more important than the bias caused by σ_{SP} alone or τ_{SP} alone.

As for the ICC, looking at Tables 3 and 4 it is apparent that: (*i*) when $\sigma_{SP} \neq 0$ and $\tau_{SP} = 0$ the ICC is overestimated; (*ii*) when $\sigma_{SP} = 0$ and $\tau_{SP} \neq 0$ the ICC is underestimated; (*iii*) when $\sigma_{SP} \neq 0$ and $\tau_{SP} \neq 0$ the bias on the ICC is upward if σ_{SP} and τ_{SP} have opposite signs, otherwise the direction of the bias depends on the specific values of the correlations σ_{SP} and τ_{SP} and the bias may even vanish, e.g. when $\sigma_{SP} = 0.50$ and $\tau_{SP} = 0.25$.

Table 5 shows that the estimate of the slope β^P_1 of the level 1 (purely within) covariate x_1 is affected by both correlations σ_{SP} and τ_{SP}. Looking at the row with $\sigma_{SP} = 0$ and the column with $\tau_{SP} = 0$, it appears that the direction of the bias depends on the sign of the correlation, while the magnitude depends on the absolute value of the correlation. A given value of correlation yields more bias when it is at level 1 rather than at level 2. When the correlations σ_{SP} and τ_{SP} are both different from zero, the bias on β^P_1 is simply the sum of the biases stemming from the two levels.

Table 6 refers to the slope γ^P of the level 2 covariate v. The pattern of the bias is analogous to the one just discussed for β^P_1. The bias on γ^P is slightly lower than the bias on β^P_1 when σ_{SP} and

Table 4. Monte Carlo means on 1000 replications of the estimates of τ_P^2 for different values of the correlations at level 2 (τ_{SP}) and level 1 (σ_{SP}).

	τ_{SP}								
σ_{SP}	*-1.00*	*-0.75*	*-0.50*	*-0.25*	*0.00*	*0.25*	*0.50*	*0.75*	*1.00*
-1.00	0.53	0.68	0.81	0.93	1.05	1.16	1.27	1.36	1.46
-0.75	0.62	0.74	0.85	0.94	1.02	1.11	1.18	1.25	1.31
-0.50	0.72	0.81	0.88	0.95	1.01	1.07	1.11	1.14	1.17
-0.25	0.81	0.88	0.93	0.97	1.00	1.03	1.04	1.05	1.05
0.00	0.93	0.95	0.98	0.99	1.00	0.99	0.99	0.96	0.93
0.25	1.04	1.05	1.05	1.02	1.00	0.97	0.93	0.89	0.82
0.50	1.17	1.16	1.11	1.06	1.00	0.95	0.90	0.81	0.72
0.75	1.31	1.26	1.18	1.12	1.03	0.94	0.84	0.74	0.62
1.00	1.46	1.37	1.27	1.16	1.05	0.94	0.80	0.68	0.53

Table 5. Monte Carlo means on 1000 replications of the estimates of β_1^P for different values of the correlations at level 2 (τ_{SP}) and level 1 (σ_{SP}).

	τ_{SP}								
σ_{SP}	*-1.00*	*-0.75*	*-0.50*	*-0.25*	*0.00*	*0.25*	*0.50*	*0.75*	*1.00*
-1.00	1.23	1.23	1.22	1.22	1.22	1.22	1.22	1.21	1.21
-0.75	1.18	1.17	1.17	1.17	1.16	1.16	1.16	1.16	1.16
-0.50	1.12	1.12	1.11	1.11	1.11	1.11	1.10	1.10	1.10
-0.25	1.06	1.06	1.06	1.06	1.06	1.05	1.05	1.05	1.05
0.00	1.01	1.01	1.00	1.00	1.00	1.00	0.99	0.99	0.99
0.25	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.93
0.50	0.90	0.90	0.89	0.89	0.89	0.89	0.89	0.88	0.88
0.75	0.84	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.82
1.00	0.79	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.77

Table 6. Monte Carlo means on 1000 replications of the estimates of γ^P for different values of the correlations at level 2 (τ_{SP}) and level 1 (σ_{SP}).

	τ_{SP}								
σ_{SP}	*-1.00*	*-0.75*	*-0.50*	*-0.25*	*0.00*	*0.25*	*0.50*	*0.75*	*1.00*
-1.00	1.29	1.27	1.25	1.25	1.24	1.22	1.21	1.20	1.19
-0.75	1.23	1.22	1.20	1.19	1.18	1.16	1.15	1.14	1.13
-0.50	1.17	1.16	1.15	1.13	1.12	1.10	1.09	1.08	1.07
-0.25	1.11	1.10	1.09	1.07	1.06	1.04	1.03	1.01	1.00
0.00	1.06	1.04	1.03	1.02	1.00	0.99	0.97	0.96	0.94
0.25	1.00	0.98	0.97	0.96	0.94	0.93	0.91	0.90	0.88
0.50	0.94	0.92	0.91	0.90	0.89	0.87	0.85	0.84	0.82
0.75	0.87	0.86	0.85	0.83	0.83	0.81	0.80	0.78	0.77
1.00	0.81	0.79	0.79	0.78	0.76	0.76	0.74	0.73	0.71

τ_{SP} have opposite signs, and slightly higher in the other cases.

A different bias for γ^P and β_1^P has implications in the decomposition of the overall slope of a level 1 covariate in the between and within slopes, because it modifies their relative sizes. To see that, let $z_{ij} = x_{1ij} + v_j$. In the simulation the within slope of z_{ij} is $\beta_1^P = 1$, while the between slope is $\gamma^P = 1$, so the overall slope of z_{ij} is one. After selection, the within and between slopes are no more equal due to the differential bias and the analyst may wrongly interpret this result as a genuine difference in the population rather than a consequence of sample selection.

Finally, Table 7 reports the mean percentage bias for different cluster sizes (n_j=2,5,10,50) and 3 combinations of values of the correlations at level 2 (τ_{SP}) and level 1 (σ_{SP}). For any considered hierarchy, the level 1 correlation induces more bias then the level 2 correlation on the slopes and the level 1 variance. The bias induced by the level 1 correlation is quite similar among the hierarchies, while the bias due to the level 2 correlation is substantially larger in hierarchies with small clusters.

The results of the simulation study confirm the general theoretical findings of Sections 3.1 and 3.3. Also the results for β_1^P and γ^P are in line with the analytical formulae (25) and (26) for the simple case $n_j = 2$, that is (*i*) γ^P is more biased than β_1^P when the selection acts through the level 2 correlation τ_{SP}, while (*ii*) β_1^P is more biased than γ^P when the selection acts through the level 1 correlation σ_{SP}. Result (*i*) is valid for any considered hierarchy, but result (*ii*) is no longer valid for hierarchies with larger n_j.

5. Final remarks

The paper has investigated the nature of sample selection in linear mixed models, focusing on the two-level random intercept case. Selection bias has been studied for both the variance components and the slopes, separately for between covariates and within covariates. Some results hold regardless of the distributional assumption, while other results are derived under Normality, partly exploiting the properties of the Unified Skew-Normal family of Arellano-Valle and Azzalini (2006). Analytic formulae have been provided in the case of a balanced hierarchy with clusters of size two, extending the well-known formulae of Heckman (1979). A simulation study has illustrated the theoretical results and has given some further insight in the topic. Most results are valid for an arbitrary hierarchy and apply to both panel and cross-sectional data.

The main findings can be summarized as follows:

- Sample selection in mixed models can assume several configurations as it depends on two sources: correlation at level 1, i.e. among errors at the individual level, and correlation at

Table 7. Monte Carlo mean percentage bias on 1000 replications of the estimates of the parameters for different data structures (J=100, n_j=2,5,10,50) and 3 combinations of values of the correlations at level 2 (τ_{SP}) and level 1 (σ_{SP}).

σ_{SP}	τ_{SP}	*parameter*	n_j			
			2	5	10	50
0	0.5	σ_P^2	1.4	-0.2	-0.0	-0.1
		τ_P^2	-7.1	-3.2	-3.3	-1.1
		β_1^P	-7.0	-3.7	-2.0	-0.5
		γ^P	-8.0	-6.4	-5.4	-2.7
0.5	0	σ_P^2	-5.7	-4.7	-5.3	-5.4
		τ_P^2	1.9	0.1	0.5	-0.1
		β_1^P	-11.3	-10.4	-10.5	-10.8
		γ^P	-8.9	-9.4	-10.8	-11.5
0.5	0.5	σ_P^2	-3.5	-5.8	-5.1	-5.5
		τ_P^2	-14.4	-11.4	-12.4	-10.4
		β_1^P	-16.9	-14.0	-12.4	-11.4
		γ^P	-16.8	-16.8	-16.7	-14.8

level 2, i.e. among errors at the cluster level. The effect on the estimates of the level 2 correlation depends on the hierarchical structure and it is weaker than the effect of the level 1 correlation, except for the level 2 variance. The two correlations have an additive effect on the slopes, but for the level 2 variance there is a strong interaction effect.

- The bias in the estimation of a parameter is an average of individual contributions; for any unit the contribution to the bias depends on the features of the cluster it belongs to, including the cluster size, the missingness pattern and the covariates of all the units of the cluster. The dependence on the features of the cluster is due to the correlation of the responses on the Selection equation: indeed, if the *Selection* equation is not mixed such dependence vanishes and the analysis of sample selection is a straightforward extension of the standard case.

- In mixed models the variance components are of primary interest and, indeed, sample selection can have serious consequences on them. The variance component structure breaks down, since after selection the errors are heteroscedastic and, in most cases, even not independent. The level 1 variance is underestimated if there is correlation at level 1, but it does not depend on the correlation at level 2. The behavior of the level 2 variance is more complex, as the bias depends on both correlations (level 1 and level 2). When the Selection equation is mixed and there is correlation at level 1, sample selection corrupts the independencies that underlie the decomposition of the variance into levels: this affects the level 2 variance, which can be seriously biased. Depending on the values of the two correlations, the level 2 variance can be overestimated or underestimated. This fact may also cause a serious bias in the Intraclass Correlation Coefficient, with a possible overrating or underrating of the role of the hierarchy.

- Other consequences of sample selection in mixed models concern the specification of the effect of level 1 covariates: (*i*) even if a covariate has a fixed slope, sample selection induces a variability of the slope at cluster level that can be wrongly interpreted as evidence of a random slope in the population; and (*ii*) sample selection causes different biases on the within and between slopes of a level 1 covariate, so after selection the difference among within and between slopes is not the same as in the population.

The analysis of sample selection shown in the paper is quite general as it investigates the model properties under an arbitrary selection mechanism. Anyway, the considered model is a special instance of mixed model in that it is linear and it does not have random slopes. Moreover, most results are based on Normality. Further research is needed to extend the analysis to more general mixed models.

The understanding of the nature and implications of sample selection is an essential step, but the applied researcher then needs reliable techniques to diagnose the selection bias and to correct it. In the mixed models framework most of the techniques to handle sample selection are specific for panel data (see e.g. Wooldridge, 2002). Much research is needed to develop effective tools for general mixed models.

Acknowledgements

We wish to thank Adelchi Azzalini for his advice on Skew-Normal distributions.

Appendix

A. Results derived from properties of the SUN distribution

Some properties of the SUN distribution

Let $(\mathbf{W}', \mathbf{U}')'$ be a random vector with multivariate Normal distribution and let $\mathbf{V}$ be a random vector with the distribution of $\mathbf{U}$ conditional on truncation on $\mathbf{W}$, namely $\mathbf{U}|\mathbf{W} > \mathbf{w}$, where $\mathbf{w}$ is the vector of truncation points. In this formulation all the variables in $\mathbf{W}$ are truncated below, but it is possible to allow for a subset of variables to be truncated above by multiplying them by -1 (this transformation only affects the sign of the covariances, not their magnitude).

In general $\mathbf{V}$ ∼SUN, where SUN stands for the Unified Skew-Normal distribution introduced by Arellano-Valle and Azzalini (2006). When truncation is on a single variable, i.e. $\mathbf{W}$ is scalar, the distribution of $\mathbf{V}$ reduces to the classical Skew-Normal distribution (Azzalini and Dalla Valle, 1996).

A couple of properties of the SUN distribution are useful for proving some results of Section 3.2:

(1) *Skewness of a component (subset of variables).* A given component of $\mathbf{V}$ can be either skewed or symmetric (i.e. with a regular multivariate Normal distribution): it is skewed if and only if the corresponding component of $\mathbf{U}$ is correlated with at least one component of $\mathbf{W}$.

(2) *Independence between two skewed components (subsets of variables).* A *necessary* condition for independence between two skewed components of $\mathbf{V}$ is the existence of at least one partition of $\mathbf{W}$ into two independent components. As a corollary, when $\mathbf{W}$ is scalar (truncation on a single variable) two skewed components cannot be independent.

A further interesting property, not used in the proofs of this Appendix, is the following: a sufficient and necessary condition for independence between a skewed component and a symmetric component of $\mathbf{V}$ is the nullity of their correlation coefficients.

Proof of Result 4 of Section 3.2.1

Sufficiency. If $\tau_{SP} = 0$ then u_j^P is independent of the remaining errors, so it is independent of e_{ij}^P also after truncation; similarly, if $\sigma_{SP} = 0$ then e_{ij}^P is independent of the remaining errors, so it is independent of u_j^P also after truncation.

Necessity. Let $\mathbf{U}' = (u_j^P, e_{ij}^P)$ and $\mathbf{W}' = (w_{ij}^S, \mathbf{w}_{(i)j}^S{}')$. If $\tau_{SP} \neq 0$ and $\sigma_{SP} \neq 0$, property (1) implies that both u_j^P and e_{ij}^P are skewed after truncation. In addition, when $\tau_S^2 > 0$, $\mathbf{W}$ cannot be partitioned into two independent components, so for property (2) the two components u_j^P and e_{ij}^P cannot be independent after truncation.

Proof of Results 6 and 7 of Section 3.2.2

Sufficiency. In both Results 6 and 7, the distributions being compared are trivially equal because of Results 1 and 2.

Necessity. Let $\mathbf{U}' = (u_j^P, e_{ij}^P, \mathbf{w}_{(i)j}^S{}')$ and $\mathbf{W} = w_{ij}^S$. Consider the case where the *Selection* equation is mixed, i.e. $\tau_S^2 > 0$. Then the component $\mathbf{w}_{(i)j}^S$ of $\mathbf{U}$ is correlated with w_{ij}^S, so for property (1) $\mathbf{w}_{(i)j}^S$ is skewed after truncation on w_{ij}^S. Moreover, when $\tau_{SP} \neq 0$ also u_j^P in $\mathbf{U}$ is correlated with w_{ij}^S, so both u_j^P and $\mathbf{w}_{(i)j}^S$ in $\mathbf{U}$ are skewed and for property (2) they are not independent after truncation on w_{ij}^S, so the distribution of u_j^P is modified by a further truncation on $\mathbf{w}_{(i)j}^S$. Similarly, when $\sigma_{SP} \neq 0$, the distribution of e_{ij}^P is modified by a further truncation on $\mathbf{w}_{(i)j}^S$.

Proof of Result 9 of Section 3.2.3

Sufficiency. If $\sigma_{SP} = 0$ then each error e_{ij}^P is independent of all the other model errors, so $e_{ij}^P \perp\!\!\!\perp e_{i'j}^P \mid A_j, \forall i \neq i'$.

Necessity. Let $\mathbf{U}' = (u_j^P, e_{1j}^P, \cdots, e_{n_j j}^P)$ and $\mathbf{W}' = (w_{ij}^S, \mathbf{w}_{(i)j}^S{}')$. If $\sigma_{SP} \neq 0$ for property (1) truncation on $\mathbf{W}$ makes each error e_{ij}^P skewed. Moreover, when the *Selection* equation is mixed ($\tau_S^2 > 0$) $\mathbf{W}$ cannot be partitioned into two independent components, so for property (2) two skewed components cannot be independent, thus $e_{ij}^P \not\!\perp\!\!\!\perp e_{i'j}^P \mid A_j, \forall i \neq i'$.

B. Formulae of means and variances of the errors after truncation

General formulae

Let $(\mathbf{W}', \mathbf{U}')'$ be a random vector with multivariate Normal distribution

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{U} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{WW} & \mathbf{\Sigma}_{WU} \\ \mathbf{\Sigma}_{UW} & \mathbf{\Sigma}_{UU} \end{bmatrix} \right)$$

and let $\mathbf{W}$ be affected by a general selection mechanism, for example truncation from below or above. Then the mean and variance of $\mathbf{U}$ after selection on $\mathbf{W}$ are (Johnson and Kotz, 1972)

$$E\left[\mathbf{U} \mid \text{selection on } \mathbf{W}\right] = \mathbf{\Sigma}_{UW}\mathbf{\Sigma}_{WW}^{-1}\widehat{\boldsymbol{\mu}}_W \quad (28)$$

$$Var\left[\mathbf{U} \mid \text{selection on } \mathbf{W}\right] = \mathbf{\Sigma}_{UU} - \mathbf{\Sigma}_{UW}\left(\mathbf{\Sigma}_{WW}^{-1} - \mathbf{\Sigma}_{WW}^{-1}\widehat{\mathbf{\Sigma}}_{WW}\mathbf{\Sigma}_{WW}^{-1}\right)\mathbf{\Sigma}_{WU}, \quad (29)$$

where $\widehat{\boldsymbol{\mu}}_W$ and $\widehat{\mathbf{\Sigma}}_{WW}$ are the mean and variance of $\mathbf{W}$ after selection, respectively.

If the type of selection is truncation from above or below, the distribution of $\mathbf{U}$ after selection on $\mathbf{W}$ belongs to the SUN family (see Appendix A). Even if Arellano-Valle and Azzalini (2006) derives the moment generating function of the SUN, the calculation of the means and variances after truncation is easier using formulae (28) and (29). Moreover, such formulae are apt to study the selection bias in a wider framework, since they hold even when: (*i*) the type of selection is other than truncation from above or below; and (*ii*) the distribution is not Normal, provided the regressions are linear and homoscedastic (Johnson and Kotz, 1972).

In the light of formulae (28) and (29), for computing the mean and variance of $\mathbf{U}$ after truncation on $\mathbf{W}$ the only difficult point is the calculation of $\widehat{\boldsymbol{\mu}}_W$ and $\widehat{\boldsymbol{\Sigma}}_{WW}$. The difficulty depends on the dimensionality of $\mathbf{W}$ and on whether the components of $\mathbf{W}$ are independent or not. Tallis (1961) derives the moment generating function of the standard multivariate Normal distribution when each component is truncated from below. In practice, $\widehat{\boldsymbol{\mu}}_W$ and $\widehat{\boldsymbol{\Sigma}}_{WW}$ have reasonably simple expressions only when the set of truncated variables $\mathbf{W}$ has one or two elements, or when its elements are independent. Note that the case of independent elements of $\mathbf{W}$ is not interesting in the present application to sample selection, because it always coincides with the case where the *relevant* truncation set has only one element: in fact, $\boldsymbol{\Sigma}_{WW}$ is the covariance matrix of the composite errors of the *Selection* equation, which are independent if and only if the *Selection* equation is not mixed ($\tau_S^2 = 0$), but in such a case the *relevant* truncation set has only one element (see Section 3.2.2).

Means of the errors after truncation for a cluster of size $n_j = 2$

In the case of a cluster of size two, i.e. $n_j = 2$, the relevant missingness patterns are only two, namely both responses Y_{1j}^P and Y_{2j}^P are observed or only one, say Y_{1j}^P, is observed. In the following, the general formula (28) is used to calculate the means of the model errors under the two relevant missingness patterns. To this end, let $\mathbf{U} = (u_j^P, e_{1j}^P)'$ and $\mathbf{W} = (w_{1j}^S, w_{2j}^S)'$. Moreover, let α_{1j} and α_{2j} be the standardized truncation points

$$\alpha_{ij} = \frac{-\mathbf{z}_{ij}^S \boldsymbol{\theta}^S}{\sqrt{\tau_S^2 + \sigma_S^2}}, \qquad i = 1, 2,$$

and let ρ be the correlation between w_{1j}^S and w_{2j}^S, which coincides with the ICC of the *Selection* equation

$$\rho = ICC_S = \frac{\tau_S^2}{\tau_S^2 + \sigma_S^2}.$$

In the following, $\Phi(\cdot)$ denotes the standard Normal distribution function and $\Phi_2(\cdot, \cdot; \rho)$ denotes the bivariate standard Normal distribution function with correlation ρ.

The matrices $\boldsymbol{\Sigma}_{UW}$ and $\boldsymbol{\Sigma}_{WW}^{-1}$ in formula (28) are:

$$\boldsymbol{\Sigma}_{UW} = \begin{pmatrix} \tau_{SP} & \tau_{SP} \\ \sigma_{SP} & 0 \end{pmatrix}$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_{WW}^{-1} &= \frac{1}{\sigma_S^2\,(2\tau_S^2 + \sigma_S^2)} \begin{pmatrix} \tau_S^2 + \sigma_S^2 & -\tau_S^2 \\ -\tau_S^2 & \tau_S^2 + \sigma_S^2 \end{pmatrix} \\ &= \frac{1}{(\tau_S^2 + \sigma_S^2)(1 - \rho^2)} \times \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \end{aligned}$$

so their product is

$$\boldsymbol{\Sigma}_{UW}\boldsymbol{\Sigma}_{WW}^{-1} = \frac{1}{(\tau_S^2 + \sigma_S^2)(1 - \rho^2)} \begin{pmatrix} \tau_{SP}(1 - \rho) & \tau_{SP}(1 - \rho) \\ \sigma_{SP} & -\sigma_{SP}\rho \end{pmatrix}.$$

Pattern 1: both Y_{1j}^P and Y_{2j}^P observed
When both responses of the *Principal* equation are observed 'selection on $\mathbf{W}$' stands for conditioning on $A_j = \{w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, w_{2j}^S > -\mathbf{z}_{2j}^S \boldsymbol{\theta}^S\}$. Moreover, from Tallis (1961) it follows

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_W &= E\left[\left(\begin{array}{c} w_{1j}^S \\ w_{2j}^S \end{array}\right) \mid w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, w_{2j}^S > -\mathbf{z}_{2j}^S \boldsymbol{\theta}^S\right] \\ &= \frac{\sqrt{\tau_S^2 + \sigma_S^2}}{\Phi_2\left(-\alpha_{1j}, -\alpha_{2j}; \rho\right)} \times \left[\begin{array}{c} \phi\left(\alpha_{1j}\right)\Phi\left(\frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}}\right) + \rho\phi\left(\alpha_{2j}\right)\Phi\left(\frac{\rho\alpha_{2j}-\alpha_{1j}}{\sqrt{1-\rho^2}}\right) \\ \phi\left(\alpha_{2j}\right)\Phi\left(\frac{\rho\alpha_{2j}-\alpha_{1j}}{\sqrt{1-\rho^2}}\right) + \rho\phi\left(\alpha_{1j}\right)\Phi\left(\frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}}\right) \end{array}\right]\end{aligned}$$

Using formula (28) the means of the errors of the *Principal* equation after truncation are

$$E\left[\left(\begin{array}{c} u_j^P \\ e_{1j}^P \end{array}\right) \mid w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, w_{2j}^S > -\mathbf{z}_{2j}^S \boldsymbol{\theta}^S\right] = \left[\begin{array}{c} \frac{\tau_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{u,1}\left(\alpha_{1j}, \alpha_{2j}, \rho\right) \\ \frac{\sigma_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{e,1}\left(\alpha_{1j}, \alpha_{2j}, \rho\right) \end{array}\right]$$

where

$$\begin{aligned}\lambda_{u,1}\left(\alpha_{1j}, \alpha_{2j}, \rho\right) &= \frac{\phi(\alpha_{1j})\Phi \quad \frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}} \quad +\phi(\alpha_{2j})\Phi \quad \frac{\rho\alpha_{2j}-\alpha_{1j}}{\sqrt{1-\rho^2}}}{\Phi_2(-\alpha_{1j},-\alpha_{2j};\rho)} \\ \lambda_{e,1}\left(\alpha_{1j}, \alpha_{2j}, \rho\right) &= \frac{\phi(\alpha_{1j})\Phi \quad \frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}}}{\Phi_2(-\alpha_{1j},-\alpha_{2j};\rho)} \qquad .\end{aligned} \tag{30}$$

Pattern 2: Y_{1j}^P observed, Y_{2j}^P missing
When the response is observed only for one unit, say Y_{1j}^P, 'selection on $\mathbf{W}$' stands for conditioning on $A_j = \{w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, w_{2j}^S \leq -\mathbf{z}_{2j}^S \boldsymbol{\theta}^S\}$. Posing $\widetilde{w}_{2j}^S = -w_{2j}^S$ the truncation set is equivalent to $\{w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, \widetilde{w}_{2j}^S > \mathbf{z}_{2j}^S \boldsymbol{\theta}^S\}$, so the formula of Tallis (1961) can still be applied reversing the signs of the correlation ρ and of the second truncation point α_{2j}, yielding

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_W &= E\left[\left(\begin{array}{c} w_{1j}^S \\ w_{2j}^S \end{array}\right) \mid w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, w_{2j}^S \leq -\mathbf{z}_{2j}^S \boldsymbol{\theta}^S\right] \\ &= E\left[\left(\begin{array}{c} w_{1j}^S \\ -\widetilde{w}_{2j}^S \end{array}\right) \mid w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, \widetilde{w}_{2j}^S > \mathbf{z}_{2j}^S \boldsymbol{\theta}^S\right] \\ &= \frac{\sqrt{\tau_S^2 + \sigma_S^2}}{\Phi\left(-\alpha_{1j}\right) - \Phi_2\left(-\alpha_{1j}, -\alpha_{2j}; \rho\right)} \times \\ &\quad \left[\begin{array}{c} \phi\left(\alpha_{1j}\right)\Phi\left(-\frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}}\right) - \rho\phi\left(\alpha_{2j}\right)\Phi\left(\frac{\rho\alpha_{2j}-\alpha_{1j}}{\sqrt{1-\rho^2}}\right) \\ -\phi\left(\alpha_{2j}\right)\Phi\left(\frac{\rho\alpha_{2j}-\alpha_{1j}}{\sqrt{1-\rho^2}}\right) + \rho\phi\left(\alpha_{1j}\right)\Phi\left(-\frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}}\right) \end{array}\right]\end{aligned}$$

where $\Phi\left(-\alpha_{1j}\right) - \Phi_2\left(-\alpha_{1j}, -\alpha_{2j}; \rho\right) = \Phi_2\left(-\alpha_{1j}, \alpha_{2j}; -\rho\right)$.

Using formula (28) the means of the errors of the *Principal* equation after truncation are

$$E\left[\left(\begin{array}{c} u_j^P \\ e_{1j}^P \end{array}\right) \mid w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S, w_{2j}^S \leq -\mathbf{z}_{2j}^S \boldsymbol{\theta}^S\right] = \left[\begin{array}{c} \frac{\tau_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{u,2}\left(\alpha_{1j}, \alpha_{2j}, \rho\right) \\ \frac{\sigma_{SP}}{\sqrt{\tau_S^2+\sigma_S^2}}\lambda_{e,2}\left(\alpha_{1j}, \alpha_{2j}, \rho\right) \end{array}\right]$$

where

$$\begin{aligned} \lambda_{u,2}\left(\alpha_{1j},\alpha_{2j},\rho\right) &= \frac{\phi(\alpha_{1j})\Phi \quad -\frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}} \quad -\phi(\alpha_{2j})\Phi \quad \frac{\rho\alpha_{2j}-\alpha_{1j}}{\sqrt{1-\rho^2}}}{\Phi(-\alpha_{1j})-\Phi_2(-\alpha_{1j},-\alpha_{2j};\rho)} \\ \lambda_{e,2}\left(\alpha_{1j},\alpha_{2j},\rho\right) &= \frac{\phi(\alpha_{1j})\Phi \quad -\frac{\rho\alpha_{1j}-\alpha_{2j}}{\sqrt{1-\rho^2}}}{\Phi(-\alpha_{1j})-\Phi_2(-\alpha_{1j},-\alpha_{2j};\rho)} \quad . \end{aligned} \tag{31}$$

References

Arellano-Valle, R. and A. Azzalini (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics 33*, 561–574.

Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution. *Biometrika 83*, 715–726.

Bellio, R. and E. Gori (2003). Impact evaluation of job training programmes: Selection bias in multilevel models. *Journal of Applied Statistics 30*, 893–907.

Borgoni, R. and F. C. Billari (2002). A multilevel sample selection probit model with an application to contraceptive use. In *Proceedings of the XLI meeting of the Italian Statistical Society*. Padova: CLEUP.

Follmann, D. and M. Wu (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics 51*, 151–168.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). New York: Oxford University Press.

Grilli, L. and C. Rampichini (2007). A multilevel multinomial logit model for the analysis of graduates' skills. *Statistical Methods and Applications, to appear*.

Hausman, J. and D. Wise (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica 47*, 455–473.

Heckman, J. (1979). Sample selection bias as a specificaton error. *Econometrica 47*, 153–161.

Jensen, P., M. Rosholm, and M. Verner (2001). A comparison of different estimators for panel data sample selection models. *University of Aarhus: Economics Working Paper No. 2002-1*.

Johnson, N. L. and S. Kotz (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley & Sons.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica 65*, 1335–1364.

Littell, R., G. Milliken, W. Stroup, R. Wolfinger, and O. Schabenberber (2006). *SAS for Mixed Models, Second Edition*. Cary: SAS Institute Inc.

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Neuhaus, J. M. and J. D. Kalbfleish (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics 54*, 638–645.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys 14*, 53–68.

Saha, C. and M. P. Jones (2005). Asymptotic bias in the linear mixed effects model under nonignorable missing data mechanisms. *Journal of the Royal Statistical Society B 67*, 167–182.

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

Tallis, G. M. (1961). The moment generating function of the truncated multinormal distribution. *Journal of the Royal Statistical Society, B 23*, 223–229.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources 33*, 127–169.

Vella, F. and M. Verbeek (1999). Two-step estimation of panel data models with censored endogenous variables and selection bias. *Journal of Econometrics 90*, 239–263.

Verbeke, G. and G. Molenberghs (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Wooldridge, J. (1995). Selection corrections for panel data models under conditional mean independece assumptions. *Journal of Econometrics 68*, 115–132.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

Wu, M. and R. Carroll (1988). Estimation and comparison of changes in the presence of informative censoring by modeling the censoring process. *Biometrics 44*, 175–188.