International Workshop on

Statistical Latent Variable Models in Health Sciences

Perugia, 6-8 September 2006

Selection Bias

in Multilevel Models

Leonardo Grilli
grilli@ds.unifi.it

Carla Rampichini
carla@ds.unifi.it

Department of Statistics

University of Florence

Outline

Aim: understanding the consequences of **sample selection** in **multilevel linear models**

- selection mechanisms in multilevel models
- the bivariate random intercept linear model
- consequences of selection
  - theoretical results (in some special instances)
  - simulation study (in more complex cases)
- future research

The selection problem

- Sample selection arises when an outcome Y^P (P = principal) is observed conditionally on another variable, e.g. $Y^S > 0$ (incidental truncation)
- Selection is present in many settings, e.g. wage can be observed only for employed people
- Problems arise if the selection mechanism depends on unobserved variables correlated with the errors terms

The selection problem

- Consequences of selection and remedies are well established in standard (single-level) models and in *random effects models for panel/longitudinal data* (Vella, 1998)
- Applications in multilevel cross-section settings are rare (Borgoni & Billari, 2002; Bellio & Gori, 2003; Grilli & Rampichini, 2004)
- No systematic study on sample selection in multilevel models

The selection problem

Sample selection in a multilevel model is more complex than in a single-level model:

- the selection process can act at different hierarchical levels, giving rise to a wide variety of patterns
- the variance-covariance structure is often of primary interest, so it must be carefully assessed how it is affected by selection
- the selection process modifies the hierarchical structure (number of clusters and cluster sizes), a feature that is relevant in the estimation phase (estimation algorithms, asymptotic approximations, power of the tests)

Scope of analysis

- We consider sample selection in a
  two-level random intercept linear model
- Our analysis is quite general in several respects:
  - the selection mechanism is driven by unobserved factors (errors) at both hierarchical levels
  - the errors determining the selection are distinct from the errors determining the outcome (though they are allowed to be the same)
  - the missingness pattern is arbitrary
  - the analysis concerns the effect of selection on the properties of the model, rather than on specific estimators

Model

BIVARIATE: each equation is two-level random intercept linear

$$\begin{cases} Y_{ij}^S = \mathbf{z}_{ij}^S \boldsymbol{\theta}^S + u_j^S + e_{ij}^S & \text{Selection equation} \\ Y_{ij}^P = \mathbf{z}_{ij}^P \boldsymbol{\theta}^P + u_j^P + e_{ij}^P & \text{Principal equation} \end{cases}$$

$j = 1, 2, \ldots, J$ clusters (level 2 units)
$i = 1, 2, \ldots, n_j$ elementary (level 1) units

Unbalanced hierarchy

Cluster-level covariates are allowed

Usually the two equations have many covariates in common

$$\begin{bmatrix} e_{ij}^S \\ e_{ij}^P \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_S^2 & \\ \sigma_{SP} & \sigma_P^2 \end{bmatrix}\right), \quad \begin{bmatrix} u_j^S \\ u_j^P \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_S^2 & \\ \tau_{SP} & \tau_P^2 \end{bmatrix}\right)$$

The distributional assumption of Normality is not essential for the general discussion on selection bias, but it is used to derive the analytical results later shown

Selection mechanism

$$Y_{ij}^P \text{ observed} \iff Y_{ij}^S > 0$$

- It operates at the elementary level (= it causes the missingness of level 1 units)
- It modifies the hierarchical structure of the data (cluster sizes and possibly also number of clusters)
- It depends on both covariances σ_{SP} (level 1) and τ_{SP} (level 2) and it is ignorable when they are both null
- Within a given cluster the pattern of missingness can be of any kind (drop-out is just a special case)

Selection mechanism

$$Y_{ij}^P \text{ observed} \iff Y_{ij}^S > 0 \iff w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S$$

$$w_{ij}^S = u_j^S + e_{ij}^S$$ Composite error of the Selection eq.

$$A_j = \left\{ \bigcap_{i:Y_{ij}^S > 0} \left\{ w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S \right\} \right\} \cap \left\{ \bigcap_{i:Y_{ij}^S \le 0} \left\{ w_{ij}^S \le -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S \right\} \right\}$$ Truncation event of cluster j

Units with observed Y^P — Units with **un**observed Y^P

After selection = conditional on truncation on the composite errors

Now consider a cluster j with observed Y^P on the first unit (i=1)

$$A_{1j} = \left\{ w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S \right\}$$ Truncation event of unit 1 of cluster j

Consequences of selection

- When the selection mechanism is not ignorable it is of interest to determine the biases arising when fitting the Principal equation alone
- Let us consider the first unit (i=1) of cluster j, assuming it is observed

$$Y_{1j}^P = \mathbf{z}_{1j}^P \boldsymbol{\theta}^P + u_j^P + e_{1j}^P$$

- independence is among clusters, but not within clusters
  →The relevant conditioning is not on A_{1j} (truncation event of unit 1), but on A_j (truncation events of all units of the cluster)

Key quantities

$$E\left(Y_{1j}^P \mid u_j^P, A_j\right) = \mathbf{z}_{1j}^P \boldsymbol{\theta}^P + u_j^P + E\left(e_{1j}^P \mid u_j^P, A_j\right)$$ Conditional mean

$$E\left(Y_{1j}^P \mid A_j\right) = \mathbf{z}_{1j}^P \boldsymbol{\theta}^P + E\left(u_j^P \mid A_j\right) + E\left(e_{1j}^P \mid A_j\right)$$ Marginal mean

$$V\left(Y_{1j}^P \mid A_j\right) = V(u_j^P \mid A_j) + V(e_{1j}^P \mid A_j) + 2cov(u_j^P, e_{1j}^P \mid A_j)$$ Marginal var.

Marginal w.r.t. the random effects

Due to the conditioning on A_j , the means and variances after selection depend on some features of the cluster:

(1) the cluster size n_j

(2) the missingness pattern (one out of 2^{nj-1}) — e.g. it is not irrelevant if unit i=2 is observed or not

(3) all the covariates of the Selection equation for all the level 1 units of the cluster

Slopes

- In linear mixed models

  marginal slope = conditional slope

- Equality may break down after selection
  ➔ marginal slope and conditional slope must be treated separately
- ML and REML are based on marginal distribution
  ➔ they estimate the marginal slope

Marginal slope

$$\underbrace{\frac{\partial E\left(Y_{1j}^P \mid A_j\right)}{\partial z_{k1j}}}_{\text{Slope after sel.}} = \underbrace{\theta_k^P}_{\text{Slope before sel.}} + \underbrace{\frac{\partial E\left(u_j^P \mid A_j\right)}{\partial z_{k1j}}}_{\text{level 2 bias}} + \underbrace{\frac{\partial E\left(e_{1j}^P \mid A_j\right)}{\partial z_{k1j}}}_{\text{level 1 bias}}.$$

- The two components of bias add up, they may have same signs or opposite signs (and even cancel out)
- The bias is null if covariate z_k is not in the Selection equation, since A_j does not contain z_k (but if covariate z_k is correlated with others the estimable slope may be biased anyway)
- The effect of a covariate varies from unit to unit:
  - The estimable slope is an average
  - Possible to end with an incorrect specification with random slopes

Marginal variance

$$V\left(Y_{1j}^P \mid A_j\right) =$$
$$V(u_j^P \mid A_j) + V(e_{1j}^P \mid A_j) + 2cov(u_j^P, e_{1j}^P \mid A_j)$$

- After selection the errors may be no longer homoscedastic, nor independent ➔ the variance component structure breaks down:
  - Level 2 errors u_j^P may be correlated with level 1 errors e_{ij}^P
  - Level 1 errors of different units may be correlated
- Problems:
  - Standard estimators are inefficient and yield incorrect std errors
  - ICC from mis-specified model ignoring selection may be above or below true ICC ➔ risk of over- or under-stating the role of clustering

ICC: Intraclass Correlation Coefficient (between-cluster variance on total variance)

Research aims

- Search configurations of model parameters such that
  - some of the potential selection biases are not in effect (e.g. the cluster level variance is unbiased,...)
  - for any unit, it is enough to condition on its own truncation event (i.e. conditioning on A_j reduces to conditioning on A_{1j})
- Search analytical expressions of bias

Tools: standard theory of Normal variates + some recent results from the *SUN distribution* (Unified Skew-Normal: Arellano-Valle & Azzalini, 2006)

Take a multivariate Normal and truncate on a subset of variables ➔ the other variables are SUN distributed, e.g.

$$u_j^P, e_{1j}^P \mid A_j \sim SUN$$

Three cases where selection causes biases, but things are not so bad...

		Case 2	*Case 3*
Selection eq. cluster var τ_S^2		>0	>0
Level 2 cov. τ_{SP}		≠0	0
Level 1 cov. σ_{SP}		0	≠0
Reduction to one-element truncation A_{1j}		no	no
Bias on slope		$\partial E(u_j^P \mid A_j)/\partial z_{k1j}$	$\partial E(e_{1j}^P \mid A_j)/\partial z_{k1j}$
$e_{1j}^P \perp u_j^P \mid A_j$		yes	yes
$e_{ij}^P \perp e_{i'j}^P \mid A_j$		yes	no
Bias on level 1 v. σ_P^2		no	downward
Bias on level 2 v. τ_P^2		downward	upward
Bias on ICC_P		downward	upward

Analytical expressions of bias

- We exploit some general formulae in Johnson & Kotz (1972) and Tallis (1961)
- We derive expressions in two cases:
  - Selection eq. not mixed (so conditioning on A_j reduces to conditioning on A_{1j}) ➔ well-known expressions of Heckman (1979) based on the inverse Mills ratio
  - Balanced hierarchy with clusters of size 2
- In the general case expressions are too complex, e.g. for a balanced hierarchy with clusters of size n
  - there are 2^{n-1} expressions, one for each missingness pattern
  - expressions involve Normal distribution functions of dimension n

Simulation design

$$Y_{ij}^S = \alpha^S + \beta_1^S x_{1ij} + \beta_2^S x_{2ij} + \gamma^S v_j + u_j^S + e_{ij}^S$$
$$Y_{ij}^P = \alpha^P + \beta_1^P x_{1ij} + \beta_3^P x_{3ij} + \gamma^P v_j + u_j^P + e_{ij}^P$$

level 1 covariate entering both S and P (x_{1ij}); equation-specific level 1 covariates (x_{2ij}, x_{3ij}); level 2 covariate entering both S and P (v_j)

- Covariates are independent (and generated once)
- Level 1 covariates only vary within clusters (i.e. identical cluster means)
- Hierarchical structure: 100 clusters of 50 units each
- True values:
  - intercepts = 0 (→ missingness rate ≈50%) slopes = 1 variances = 1
  - covariances σ_{SP} and τ_{SP} in [-1,+1] step 0.25 (a grid with 81 cells)

Estimates of level 1 variance σ_P^2

$\sigma_{SP} \setminus \tau_{SP}$	-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

MC means on 1000 runs

Estimates of level 2 variance τ_P^2

$\sigma_{SP} \setminus \tau_{SP}$	-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

MC means on 1000 runs

Future work on sample selection

- Understanding
  - Linear mixed models with random slopes
  - Non-linear mixed models, e.g. logit
  - Other selection mechanisms, e.g. cluster-based selection
- Diagnostic tools
- Solutions (two-equation models, instrumental variables, sensitivity analysis)

Questions and further material

- Email: **grilli@ds.unifi.it**
- Web: **www.ds.unifi.it/grilli**

Thanks for your attention!

References

- **Arellano-Valle, R. B. and A. Azzalini (2006)** On the unification of families of skew-normal distributions. Scandinavian J. Stat.
- **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. Proc. XLI meeting Italian Statistical Soc.
- **Grilli, L. and C. Rampichini (2004)** A polytomous response multilevel model with a non ignorable selection mechanism. Proceedings of the 19th IWSM. Firenze
- **Heckman, J. (1979)** Sample selection bias as a specificaton error. Econometrica 47, 153–161.
- **Johnson, N. L. and S. Kotz (1972)** Distributions in Statistics: Continuous Multivariate Distributions. New York: Wiley & Sons.
- **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.

Estimates of slope of level 2 covariate γ^P

$\sigma_{SP} \setminus \tau_{SP}$	-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

MC means on 1000 runs

Estimates of slope of level 1 covariate β_1^P

$\sigma_{SP} \setminus \tau_{SP}$	-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

This covariate has only within-cluster variation

In general $z_{ij} = \bar{z}_j + (z_{ij} - \bar{z}_j)$

Between variation — Within variation

MC means on 1000 runs

MC mean percentage bias on 1000 replications for different data structures (J= 100, n_j= 2, 5, 10, 50)

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

elementary level cov	cluster level covariance: $\tau_{SP} \neq 0$	cluster level covariance: $\tau_{SP} = 0$
$\sigma_{SP} \neq 0$	$E(e_{1j}^P \mid u_j^P, A_j)$ $E(u_j^P \mid A_j) + E(e_{1j}^P \mid A_j)$ $Var(u_j^P + e_{1j}^P \mid A_j)$	$E(e_{1j}^P \mid A_j)$ $E(e_{1j}^P \mid A_j)$ $\tau_P^2 + Var(e_{1j}^P \mid A_j)$
$\sigma_{SP} = 0$	0 $E(u_j^P \mid A_j)$ $Var(u_j^P \mid A_j) + \sigma_P^2$	0 0 $\tau_P^2 + \sigma_P^2$

For $\sigma_{SP} \neq 0$, $\tau_{SP} = 0$:

- Slope biased due to correlation at level 1
- Marginal = conditional
- Errors at different levels are independent
- Errors at level 1 e_{ij}^P are not independent, except when the Selection eq. is not mixed ($\tau^2_S = 0$)
- ICC over-estimated if the Selection eq. is not mixed

For $\sigma_{SP} = 0$, $\tau_{SP} \neq 0$:

- Slope biased due to correlation at level 2
- Marginal ≠ conditional
- Errors at different levels are independent
- Errors at level 1 e_{ij}^P are independent
- ICC under-estimated

The conditioning on A_j reduces to conditioning on A_{1j} only when the Selection eq. is not mixed ($\tau^2_S = 0$)