

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 <u>cross-section</u> settings are rare (Borgoni & Billari, 2002; Bellio & Gori, 2003; Grilli & Rampichini, 2004)
- No systematic study on sample selection in multilevel models

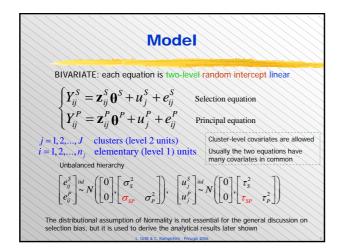
The selection problem

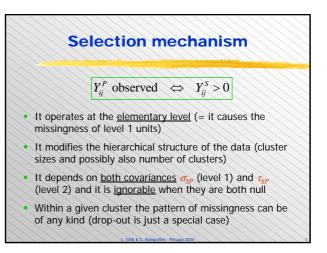
Sample selection in a multilevel model is more complex than in a single-level model:

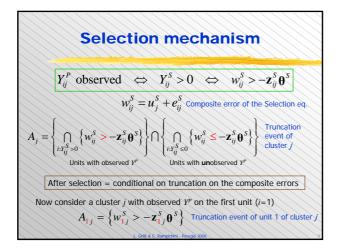
- the selection process can act at different hierarchical levels, giving rise to a wide variety of patterns
- the variance-covariance structure is often of primary interest, so it must be carefully assessed how it is affected by selection
- the selection process modifies the hierarchical structure (number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

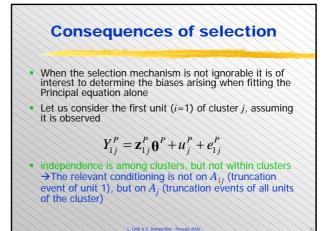
Scope of analysis

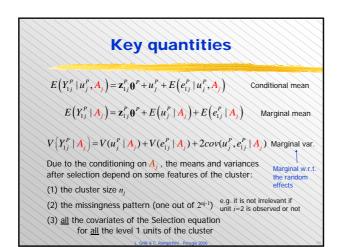
- We consider sample selection in a two-level random intercept linear model
- Our analysis is quite general in several respects:
 the selection mechanism is driven by unobserved factors
 - (errors) at both hierarchical levels
 the errors determining the selection are <u>distinct</u> from the errors determining 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

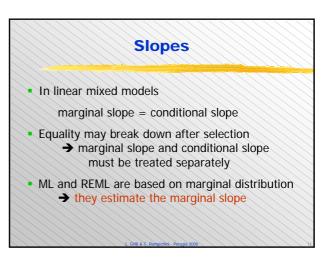




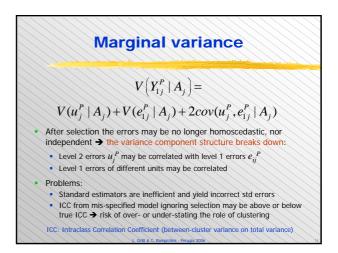


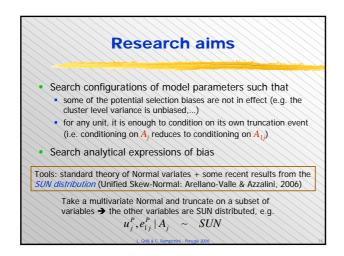




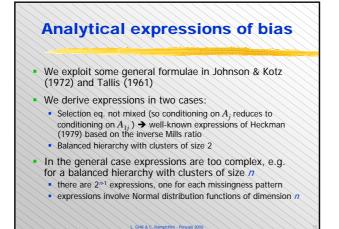


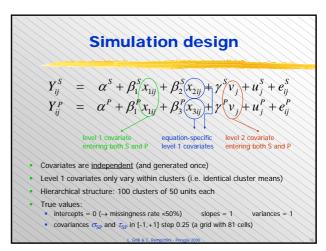
Marginal slope
$\frac{\partial E\left(Y_{1j}^{P} \mid A_{j}\right)}{\partial z_{k1j}} = \theta_{k}^{P} + \frac{\partial E\left(u_{j}^{P} \mid A_{j}\right)}{\partial z_{k1j}} + \frac{\partial E\left(e_{1j}^{P} \mid A_{j}\right)}{\partial z_{k1j}}.$ Slope after set. Slope before set. level 2 bias level 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





	Case 2	Case 3
Selection eq. cluster var τ_s^2	>0	>0
Level 2 cov. $ au_{SP}$	≠0	0
Level 1 cov. $\sigma_{\scriptscriptstyle SP}$	0	≠0
Reduction to one- element truncation A_{ij}	no	во
Bias on slope	$\partial E(u_j^R A_j) / \partial z_{k1j}$	$\partial E(e_{1j}^{P} A_{j})/\partial z_{k1j}$
$e_{1j}^p \perp u_j^p A_j$	yes	yes
$e_{ij}^{p}\perp e_{ij}^{p}\mid A_{j}$	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





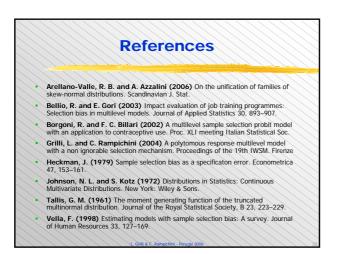
$\sigma_{SP} \setminus \tau_{SP}$	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00
-1.00	0.78	0.79	0.78	0.79	(0.79)	0.79	0.79	0.79	0.78
-0.75	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
-0.50	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95
-0.25	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
0.00	1.00	1.00	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

$\sigma_{SP} \setminus \tau_{SP}$	-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	H.I	1.18	1.25	1.31
-0.50	0.72	0.81	0.88	0.95	1.01	1.07	1.1	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

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)





$\tau_{SP} \mid \tau_{SP}$	-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	1.1	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

			ope of		0010		1	
				100000000000000000000000000000000000000				
\rightarrow	\rightarrow	\rightarrow				\rightarrow		\rightarrow
-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00
1.23	1.23	1.22	1.22	1.22	1.22	1.22	1.21	1.21
1.18	1.17	1.17	1.17	1.16	1.16	1.16	1.16	1.16
1.12	1.12	11.1	1.11	111	1.11	1.10	1.10	1.10
1.06	1.06	1.06	1.06	1.06	1.05	1.05	1.05	1.05
1.01	1.01	1.00	1.00	1.00	1.00	0.99	0.99	0.99
0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.93
0.90	0.90	0.89	0.89	0.89	0.89	0.89	0.88	0.88
0.84	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.82
0.79	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.77
e has onl	ly within-	cluster va	riation			\sim	\sim	
$z_{ii} =$	$\overline{\overline{z}}_i$ +	(Z _{ii})	$-\overline{z}_{i}$		MC	means	on 1000	runs
			1 2 1					
	$\begin{array}{c} 1.23 \\ 1.18 \\ 1.12 \\ 1.06 \\ 1.01 \\ 0.95 \\ 0.90 \\ 0.84 \\ 0.79 \\ e \text{ has on } \\ ij \\ Be \end{array}$	1.23 1.23 1.18 1.17 1.12 1.12 1.06 1.06 1.01 1.01 0.95 0.95 0.90 0.90 0.84 0.84 0.79 0.78 e has only within	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$

20		parameter	\square		y T	
σ_{sp}	τ_{sp}	purumeter	2	5	10	50
\sim	\square	σ_{R}^{2}	1.4	-0.2	-0.0	-0.1
	0.5	τ_P^2	-7.1	-3.2	-3.3	1,1
<u> </u>		β_1^p	-7.0	-3.7	-2.0	-0.5
		() yP	-8.0	-6.4	-5.4	-2.5
	\sim	σ_P^2	-5.7	-4.7	-5.3	-5.4
05		τ_P^2	1.9	0.1	0.5	-0.1
0.5	l v	β_1^{ρ}	-11.3	-10.4	-10.5	-10.
		× P	-8.9	-9.4	-10.8	<u>1</u> 1.
\square	\sim	σ_P^2	-3.5	-5.8	-5.1	-5.5
		τ_p^2	-14.4	-11.4	-12.4	-10.
0.5	0.5	β_1^p	-16.9	-14.0	-12.4	-41.
		1 yr	-16.8	-16.8	-16.7	-14.

elementary	cluster level co	ovariance	
level cov	$\tau_{sp} \neq 0$	$\tau_{SP} = 0$	
	$E(e_{ij}^p u_j^p, A_j)$	$E(e_{ij}^{p} A_{j})$	Slope biased due to correlation at level 1
$\sigma_{\rm SP} \neq 0$	$E(u_j^p A_j) + E(e_{ij}^p A_j)$ Var $(u_j^p + e_{ij}^p A_j)$	$E(e_{1j}^{p} A_{j})$	 Marginal = conditional
	$Var(u_j^{p} + e_{1j}^{p} A_j)$	$\tau_p^2 + Var(e_{1j}^p A_j)$	Errors at different levels are independent
$\sigma_{\rm SP}=0$	$ \begin{array}{c} 0\\ E(u_j^p \mid A_j)\\ Var(u_j^p \mid A_j) + \sigma_p^2 \end{array} $	$\begin{array}{c} 0 \\ 0 \\ \tau_p^2 + \sigma_p^2 \end{array}$	• Errors at level 1 e_{ij}^{p} are <u>not</u> independent, except when the Selection eq. is not mixed ($\tau_{s}^{2} = 0$)
	Slope biased due to correlation at level 2		ICC over-estimated if the Selection eq. is not mixed
	Marginal × conditional		
	Errors at different levels are independent		
	• Errors at level 1 e_{ij}^{P} are independent	(to conditi	litioning on A_j reduces oning on A_{ij} only when eq. is not mixed ($r_s = 0$)
	ICC under-estimated	1192	