Model building issues in multilevel linear models with endogenous covariates

Leonardo Grilli and Carla Rampichini

Dipartimento di Statistica "Giuseppe Parenti"
Università di Firenze
grilli@ds.unifi.it, carla@ds.unifi.it

knemo

KNEMO'06 Conference Anacapri, 3-6 September 2006

Motivation

- Endogeneity is a common problem in applied works
- We explore level 2 endogeneity in linear random effects models, i.e. random effects correlated with covariates
- Measurement error stemming from the use of cluster means is overlooked in the literature

Outline

- Endogeneity in multilevel models
- The linear random intercept model
- Between-cluster slope and measurement error: correction for consistent estimates
- Simulations
- Conclusions

The linear random intercept model

$$Y_{ij} = \alpha + \beta X_{ij} + v_j + e_{ij}$$

- i=1,2,…,n_j elementary (level 1) index
- j=1,2,…,J cluster (level 2) index Y_{ij} response
- X_{ij} level 1 covariate
- v_j level 2 errors, or random effects
- e_{ji} level 1 errors
- Examples:
  - Panel data (typically: n_j small, J large)
  - Clustered cross-section data (typically: n_j large, J small)

Level 2 endogeneity

- Level 2 endogeneity arises when

$$E(v_j \mid X_{ij}) \neq 0$$

➔ standard estimators are inconsistent for β

Note that $Cov(v_j, X_{ij}) \neq 0 \Rightarrow E(v_j \mid X_{ij}) \neq 0$

Sources of endogeneity

- Omission of relevant regressors at any level
- Measurement error in the covariates
- Self-selection
- Simultaneity

The data generating model

- X_{ij} must be treated as random
- The hierarchical framework requires to specify how varies between and within clusters, assume a variance component model

$$X_{ij} = X_j^B + X_{ij}^W$$

- Under the assumptions

X1 X_j^B are iid with mean μ_X and variance $\tau_X^2 > 0$

X2 X_{ij}^W are iid with zero mean and variance $\sigma_X^2 > 0$

X3 $X_j^B \perp\!\!\!\perp X_{ij}^W, \ \forall i, j$

The Overall model

$$Y_{ij} = \alpha + \beta^W X_{ij}^W + \beta^B X_j^B + u_j + e_{ij}$$

- β^W within effect, β^B between effect
- in general, $\beta^W \neq \beta^B$
- Assume:
  - Independent clusters
  - Two-stage sampling
  - Unbalanced design

Assumptions on the errors

In the Overall data generating model:

- X^W and X^B are exogenous
- Errors at different levels are independent
- At each level, errors are i.i.d.

Example: university effectiveness

- Y_{ij} observed income of the *i*-th student of the *j*-th school one year after graduation
- X_{ij} observed grade of such a student
- $X_{ij} = X^B_j + X^W_{ij}$
  - X^B_j school mean grade
  - $X^W_{ij} = X_{ij} - X^B_j$ student deviation from school mean
- The model decompose the total effect of X on Y:
  - β^B school effect on the income
  - β^W student effect on the income
- u_j school level residual (effectiveness)
- e_{ji} student level residual

An example …cont'd

Between and Within effects are conceptually different and sometimes have opposite signs!

<u>In this example (University of Florence data)</u>:

- β^W >0 ➔ *within a school*, students with higher grade have higher income
- β^B <0 ➔ schools giving higher grades show lower average income (e.g. Humanities)

In many settings we expect $\beta^W \neq \beta^B$

Nature of level 2 endogeneity

- Overall model alternative parametrization

$$Y_{ij} = \alpha + \beta^W X_{ij} + \delta X_j^B + u_j + e_{ij}$$

$$\delta = \beta^B - \beta^W$$

- If X^B is omitted➔ it is included in level 2 error

$$Y_{ij} = \eta + \beta^W X_{ij} + v_j + e_{ij}$$

$$v_j = \delta(X_j^B - \mu_X) + u_j$$

$$E(v_j) = 0 \quad Var(v_j) = \tau_{Y|X}^2 = \delta^2 \tau_X^2 + \tau_{Y|X^B X^W}^2$$

Omission of X^B

Note that: $Cov(v_j, X_{ij}) = Cov(v_j, X_j^B) = \delta\tau_X^2$ — Between variance of X (assumed >0)

➔ X exogenous iff δ=0

⬇

level 2 endogeneity can be interpreted as:

- a wrong constraint on the slopes, i.e. $\beta^W=\beta^B=\beta$

$Y_{ij}=\alpha+\beta^W X_{ij}^W+\beta^B X_j^B+u_j+e_{ij}$ → $Y_{ij}=\alpha+\beta X_{ij}+v_j+e_{ij}$

- the omission of a relevant regressor, i.e. X^B

$Y_{ij}=\alpha+\beta^W X_{ij}+\delta X_j^B+u_j+e_{ij}$ → $Y_{ij}=\alpha+\beta X_{ij}+v_j+e_{ij}$

Correct for level 2 endogeneity

Fit the Overall model

$$Y_{ij}=\alpha+\beta^W X_{ij}^W+\beta^B X_j^B+u_j+e_{ij}$$

or

$$Y_{ij}=\alpha+\beta^W X_{ij}+\delta X_j^B+u_j+e_{ij}$$

BUT X^B and X^W are UNOBSERVABLE!

What is observed?

Replace

- X^B with the cluster means $\bar X_j=\frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$
- X^W with the deviations $\tilde X_{ij}=X_{ij}-\bar X_j$

Observable split $X_{ij}=\bar X_j+\tilde X_{ij}$

Note that X^B and X^W are measured with error:

$$\bar X_j = X_j^B+\bar X_j^W \qquad \tilde X_{ij}=X_{ij}^W-\bar X_j^W$$

Working overall model

$$Y_{ij}=\alpha+\beta^W\tilde X_{ij}+\beta^B\bar X_j+z_j+e_{ij}$$

$$z_j=u_j-\delta\bar X_j^W$$ measurement error cancel out iff δ=0

Note that $\tilde X_{ij}$:

- is exogenous, i.e. $E(z_j\mid\tilde X_{ij})=-\delta E(\bar X_j^W\mid\tilde X_{ij}^W)=0$
- and uncorrelated with cluster mean: $Cov(\bar X_j,\tilde X_{ij})=0$

⬇

β^W consistently estimated

Estimable between slope

- The cluster mean is endogenous!

$$Cov(z_j,\bar X_j)=-\delta\sigma_X^2/n_j$$

- We consistently estimate $\beta^B_{cm}\neq\beta^B$
- In the balanced case:

$\lambda_X\in$ (0,1) reliability of X

$$\beta^B_{cm}=\lambda_X\beta^B+(1-\lambda_X)\beta^W \quad ,$$

- The model is correctly specified, but the measurement error causes endogeneity!!!

In the balanced case:
recover β^B using the estimated β^B_{cm} and λ_X

Reliability

$$\lambda_X=\frac{Var(X_j^B)}{Var(\bar X_j)}=\frac{\tau_X^2}{\tau_X^2+\sigma_X^2/n}=\left(1+\frac{1}{(\tau_X^2/\sigma_X^2)n}\right)^{-1}, \quad \lambda_X\in(0,1)$$

λ_X is an increasing function of :

- the cluster size n (sample design)
- the variance ratio τ_X^2/σ_X^2 (model parameters)

It is usual to find values far from 1, e.g.

$$\lambda_X=0.67 \text{ if } \begin{cases} n=2 & \tau_X^2=\sigma_X^2 \quad \text{(panel)} \\ n=20 & \tau_X^2=0.1\sigma_X^2 \quad \text{(cross-section)} \end{cases}$$

Estimate τ^2_X and σ^2_X by standard ANOVA methods

Variance correction

Estimated residual level 2 variance is inflated!

$$Var(z_j) = \tau^2_{Y|X^B X^W} + \lambda_X \delta^2 \frac{\sigma^2_X}{n}$$

true level 2 residual variance

Recover the 'true' variance subtracting this factor from the estimated variance

measurement error bias may be more serious on τ^2 than on δ!!!

In summary …

Working model	$Y_{ij} = \eta + \beta^W X_{ij} + v_j + e_{ij}$	$Y_{ij} = \alpha + \beta^W \widetilde{X}_{ij} + \beta^B \overline{X}_j + z_j + e_{ij}$
Omission of a regressor	yes	no
Measurement error	no	yes
Level 2 error cov	$Cov(v_j, X_{ij}) = \delta\tau^2_X$	$Cov(z_j, \tilde{X}_{ij}) = 0$, $Cov(z_j, \overline{X}_j) = -\delta\sigma^2_X / n_j$
Consistent β^W	No for n small	yes
Consistent β^B	no	yes (correction required)

"true" model

$$Y_{ij} = \begin{cases} \alpha + \beta^W X^W_{ij} + \beta^B X^B_j + u_j + e_{ij} \\ \alpha + \beta^W X_{ij} + \delta X^B_j + u_j + e_{ij} \end{cases}$$

Simulation: MC means on 1000 replications

$\delta=\beta^B-\beta^W$	$Y_{ij}=\alpha+\beta X_{ij}+\ldots$		$Y_{ij}=\alpha+\beta^W X_{ij}+\delta\overline{X}_j+\ldots$		
	β	τ^2_Y	β^W	δ	τ^2_Y
-2	0.61	3.63	1.01	-1.34	2.35
-1.5	0.62	2.27	1.00	-1.00	1.75
-1	0.71	1.52	1.01	-0.68	1.34
-0.5	0.84	1.12	1.00	-0.33	1.09
0	1.00	1.00	1.00	0.00	1.00
0.5	1.16	1.12	1.00	0.33	1.09
1	1.29	1.51	0.99	0.67	1.34
1.5	1.38	2.28	1.00	1.01	1.75
2	1.40	3.59	1.00	1.34	2.36

+ endogeneity (negative δ) / + endogeneity (positive δ)

$\lambda_X=2/3$, $\beta^W=1$, $\tau^2_Y=1$
$n=2$, $J=100$

What happens when cluster size n→∞

$\delta=\beta^B-\beta^W$	$Y_{ij}=\alpha+\beta X_{ij}+\ldots$		$Y_{ij}=\alpha+\beta^W X_{ij}+\delta\overline{X}_j+\ldots$		
	β	τ^2_Y	β^W	δ	τ^2_Y
-2	1	5	1	-2	1
-1.5	1	3.25	1	-1.5	1
-1	1	2	1	-1	1
-0.5	1	1.25	1	-0.5	1
0	1	1	1	0	1
0.5	1	1.25	1	0.5	1
1	1	2	1	1	1
1.5	1	3.25	1	1.5	1
2	1	5	1	2	1

+ endogeneity (negative δ) / + endogeneity (positive δ)

$\lambda_X=1$, $\beta^W=1$, $\tau^2_Y=1$
$n\approx50$, $J=100$

Alternative RE estimators

- Other methods for consistent estimation of β^W exist (e.g CIGLS and IV) $Y_{ij} = \alpha + \beta^W X_{ij} + v_j + e_{ij}$

→ level 2 variance is net (adjusted for X)

- If the interest is also on β^B, the overall model must be used $Y_{ij} = \alpha + \beta^W X_{ij} + \delta\overline{X}_j + z_j + e_{ij}$

▸ level 2 variance is net (adjusted for X)
▸ BUT attention to measurement error!

Concluding remarks

- In general RE models are better than FE also in presence of level 2 endogeneity
- More simulations are needed to study efficiency issues
- Many extensions need further investigation:
  - model with two or more covariates (problems of model selection)
  - model with random slopes
  - non-linear models.

References

Ebbes, P., Bockenholt, U. and Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models, *Statistica Neerlandica*, 58, 161–178.

Fielding, A. (2004). The Role of the Hausman Test and whether Higher Level Effects should be treated as Random or Fixed. *Multilevel Modelling Newsletter*, 16(2), 3–9.

Kim, J. S. and Frees, E. W. (2006). Omitted variables in multilevel models. *Psychometrika*, in press.

Rettore, E. and Martini, A. (2001). Constructing league tables of service providers when the performance of the provider is correlated to the characteristics of the clients. In *Processi e metodi statistici di valutazione*, Proceedings of the Conference of the Italian Statistical Society, Roma.

Rice, N., Jones, A. M. and Goldstein H. (2002). Multilevel models where the random effects are correlated with the fixed predictors. Manuscript available at: http://www.mlwin.com/hgpersonal/cigls.pdf.

Snijders, T. A. B. and Berkhof, J. (2006). Diagnostic checks for multilevel models. In Jan de Leeuw (Ed.), *Handbook of Multilevel Analysis*. New York: Springer, to appear.

Thanks for your attention!

Estimable slope

In model : $Y_{ij} = \alpha + \beta X_{ij} + v_j + e_{ij}$

the OLS estimable slope is actually

$$\beta = \beta^W + \rho_X \delta = \rho_X \beta^B + (1-\rho_X)\beta^W$$

$$\rho_X = \tau_X^2 / \left(\sigma_X^2 + \tau_X^2\right)$$

i.e. a mixture of the between and within slope, depending on the value of the ICC of X

Endogeneity test

- The endogeneity Hausman test is equivalent to the Wald test H_0: δ=0 in the Working Overall model $Y_{ij} = \alpha + \beta^W X_{ij} + \delta \overline{X}_j + z_j + e_{ij}$
- The estimable parameter is $\delta_{cm} = \lambda_x \delta$
- The test H_0: δ_{cm}=0 has a lower power

⬇

do the test using $\delta = \delta_{cm} / \lambda_x$

FE versus RE

- Standard solution for endogeneity in panel literature is the FE estimator (consistent for β^W)

but $\tilde{Y}_{ij} = \beta^W \tilde{X}_{ij} + \varepsilon_{ij}$

  - It doesn't allow any cluster level covariate
  - It can be quite inefficient because the number of parameters grow with the number of clusters

- If interest is only on β^W, a better solution is the Within RE model $Y_{ij} = \alpha + \beta^W \tilde{X}_{ij} + s_j + e_{ij}$

→ level 2 variance is gross (unadjusted for *X*)