Measurement error in multilevel models with sample cluster means

Leonardo Grilli & Carla Rampichini

grilli@ds.unifi.it rampichini@ds.unifi.it

Dipartimento di Statistica “Giuseppe Parenti”
Università di Firenze

SFC - CLADAG 2008
Caserta, 11-13 June 2008

Motivation /1

- Regression analysis with data from observational studies is often affected by the problem of endogeneity
- In multilevel (mixed) models, this problem can concern error terms at any level: level 1 (e.g. student), level 2 (e.g. school), level 3 (e.g. district) …
- We explore **level 2 endogeneity** in two-level models, i.e. random effects correlated with covariates, an issue well known in the setting of panel data due to the famous Hausman test (in panel data level 1 are waves, level 2 are subjects)
- This type of endogeneity arises from a wrong equality restriction on the between-cluster and within-cluster slopes

Motivation /2

- Solution: allow distinct between and within slopes through the addition of the cluster mean as a further covariate
- BUT the use of the sample cluster mean instead of the population cluster mean entails a **measurement error** that yields a biased estimator of the between-cluster slope
- Measurement error stemming from the use of cluster means is overlooked in the literature
- We propose a correction to obtain unbiased estimates and evaluate its performance

The framework

Assume a 2-level hierarchy with
- $j=1,2,\dots,J$ level 2 units (clusters)
- $i=1,2,\dots,n_j$ level 1 units
  - Panel (typically: J large, n_j small)
  - Clustered cross-section (typically: J small, n_j large)

We first focus on balanced designs ($n_j=n$) to obtain simple formulae, then we generalize

Consider two variables
- X_{ij} covariate at level 1
- Y_{ij} response at level 1

We want to study endogeneity issues in a random intercept model for $Y_{ij} \mid X_{ij}$ → we must specify a model also for X_{ij}

The data generating model for X

- We adopt a variance component model

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

<u>Assumptions</u>

- X_j^B iid with mean μ_X and variance $\tau_X^2 > 0$
- X_{ij}^W iid with mean 0 and variance $\sigma_X^2 > 0$
- $X_j^B \perp X_{ij}^W$ (independent components)

But X^B and X^W are unobservable!

X_j^B can be measured by the sample cluster mean $\bar{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} = X_j^B + \bar{X}_j^W$

X_{ij}^W can be measured by the deviation $\tilde{X}_{ij} = X_{ij} - \bar{X}_j = X_{ij}^W - \bar{X}_j^W$

The data generating model for $Y \mid X$

$$Y_{ij} = \alpha + \beta_W X_{ij}^W + \beta_B X_j^B + u_j + e_{ij} \qquad \text{(1a)}$$

$$Var(u_j) = \tau^2_{Y|X^B X^W}$$

- Model (2a) allows for different between and within effects

$\beta_B \neq \beta_W$ In many settings, between and within effects are conceptually different and may even have opposite signs, so it is important to distinguish among them

- <u>We assume</u>:
  - X^W and X^B are independent of the errors
  - Errors at different levels are independent
  - At both levels, iid errors (→ independent clusters)

The data generating model for $Y \mid X$

- Alternative parameterization of the data generating model

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta X_j^B + u_j + e_{ij} \qquad \text{(1b)}$$

$\delta = \beta_B - \beta_W$ is known as *contextual coefficient*

- In educational research the *contextual coefficient* is often found to be significant, meaning that the context has an effect on the individual outcomes.
- For example, if X is the prior student achievement, X^B is the school mean prior achievement, a proxy of the *quality of the context*. If the *contextual effect* is not null, two students with the same prior achievement will obtain different final achievements depending on the school attended.

Working model $\left(X_j^B \text{ measured by } \bar{X}_j\right)$

To avoid level 2 endogeneity we must include the cluster mean, as in model (1b). Since the population cluster mean X^B is unobservable, we measure it through the **sample cluster mean**:

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta \bar{X}_j + z_j + e_{ij} \qquad \text{(2)}$$

Due to **measurement error**, the sample cluster mean $\bar{X}_j$ is endogenous

$$Cov(z_j, \bar{X}_j) = -\delta\sigma_X^2/n$$

It can be shown that the within slope β_W is unbiasedly estimated, while the contextual coefficient δ is attenuated

The inclusion of the <u>sample cluster mean</u>
- avoids level 2 endogeneity due to *omission of a relevant regressor*
- but still entails level 2 endogeneity due to *measurement error*

Attenuation of the contextual coefficient δ

Measurement-error-attenuated contextual coefficient

$$\delta_m = \lambda_X \delta$$

Measurement error vanishes iff $\delta = 0$, i.e. $\beta_B = \beta_W$

Anyway δ_m is close to δ when $\lambda_X \approx 1$

Reliability coefficient

$$\lambda_X = \frac{Var(X_j^B)}{Var(\overline{X}_j)} = \frac{\tau_X^2}{\tau_X^2 + \sigma_X^2/n} = \left(1 + \frac{1}{(\tau_X^2/\sigma_X^2)n}\right)^{-1}$$

λ_X takes values in $(0,1)$ and is an increasing function of:

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

Values of λ_X can be far from 1, e.g.

$$\lambda_X = 2/3 \quad \text{if} \begin{cases} n = 2 & \text{and} \quad \tau_X^2 = \sigma_X^2 \quad \text{(e.g. panel)} \\ n = 20 & \text{and} \quad \tau_X^2 = 0.1\sigma_X^2 \quad \text{(e.g. cross-section)} \end{cases}$$

Measurement error correction via λ_X

The measurement error induced by the use of the sample cluster mean can be corrected **with the data at hand**

1. Use the *working* model to estimate:

$$\delta_m = \lambda_X \delta \qquad \text{(attenuated)}$$

$$\tau^2_{Y|X^B X^W,m} = (1-\lambda_X)\delta^2\tau_X^2 + \tau^2_{Y|X^B X^W} \qquad \text{(inflated)}$$

2. Estimate τ_X^2 and σ_X^2, and thus λ_X, by standard methods
3. Recover unbiased estimates:

$$\hat{\delta}_c = \hat{\delta}_m / \hat{\lambda}_X$$

$$\hat{\tau}^2_{Y|X^B X^W,c} = \hat{\tau}^2_{Y|X^B X^W,m} - (1-\hat{\lambda}_X)\hat{\delta}_c^2\hat{\tau}_X^2$$

Measurement error bias may be more serious on $\tau^2_{Y|X^B X^W}$ than on δ!

Simulation

1. Generate data under 'true' model (1b) with varying δ
2. Fit models A and B (MC means on 1000 replicates, REML)

$\delta = \beta_B - \beta_W$	**Model A without cluster mean (omitted regressor)** $Y_{ij} = \eta + \beta X_{ij} + \dots$		**Model B with cluster mean (measurement error)** $Y_{ij} = \alpha + \beta_W X_{ij} + \delta \overline{X}_j + \dots$		
	β	$\tau^2_{Y\|X}$	β_W	δ	$\tau^2_{Y\|X^B X^W}$
-2	0.61	3.57	1.00	-1.33	2.32
-1.5	0.62	2.26	1.00	-1.00	1.75
-1	0.70	1.49	1.00	-0.67	1.33
-0.5	0.84	1.11	1.00	-0.33	1.08
No endogeneity 0	1.00	1.00	1.00	0.00	1.00
0.5	1.16	1.12	1.00	0.33	1.09
1	1.30	1.50	1.00	0.67	1.34
1.5	1.37	2.28	1.00	1.00	1.76
2	1.39	3.59	1.00	1.33	2.34

+ endogeneity

+ endogeneity

True values: $\lambda_X = 2/3$, $\beta_W = 1$, $\tau^2_{Y|X^B X^W} = 1$

Data structure: $J = 1000$, $n = 2$

Even if $\delta \neq 0$, when the cluster size increases ($n \to \infty$ and thus $\lambda_X \to 1$):

- the slopes are unbiased in both models
- the residual cluster variance is unbiased in model B but inflated in model A

Variance and MSE of the corrected estimator

- The sampling variance of the corrected estimator of δ

$$Var(\hat{\delta}_c) = Var(\hat{\delta}_m / \hat{\lambda}_X)$$

can be easily computed using the Taylor approximation of the variance of a ratio (simulations show that the approximation is good)

- The correction cancels the bias, but inflates the sampling variance; simulations show that in most cases it is worthwhile in terms of MSE:

$$Var(\hat{\delta}_c) > Var(\hat{\delta}_m) \quad \text{but in most cases} \quad MSE(\hat{\delta}_c) < MSE(\hat{\delta}_m)$$

Correction via λ_X in unbalanced designs

- The reliability varies with the cluster size
  ➔ several reliability values
- How summarize them?
  reliability with average *n* vs **average reliability**

Simulations show that

- As the degree of unbalancedness increases:
  - stronger attenuation (lower *attenuation factor*)
  - the reliability with average *n* is constant, so it is not useful
  - the **average reliability** decreases
- The average reliability tends to be larger than the true attenuation factor, but in most cases the correction is satisfactory

Correction via λ_X when sampling from clusters of finite size

- Need to adjust the estimators of the variance components to account for finite population ➔ modify the reliability

Simulations show that

- the **modified reliability** is a good approximation of the attenuation of the contextual effect due to measurement error, thus the corrected estimator has a good performance.
- Failing to use the modified reliability leads to an overcorrection that becomes remarkable for sampling fractions of 0.25 or more.
- The corrected estimator of the contextual effect has a lower MSE than the uncorrected estimator, even if the gap diminishes as the sampling fraction increases.

Correction via λ_X: pros and cons

- Pros
  - very simple procedure
  - applied after running standard multilevel software (no need to use software for IRT or SEM)
  - easy to apply to results published by other researchers
  - with prior information on the ICC of the covariate, the amount of attenuation can be evaluated when planning the sampling design
- Cons
  - the sampling variance of the corrected estimator increases → need to evaluate if the correction is worthwhile in terms of MSE
  - exact only for balanced designs (even if quite good in most unbalanced designs)
  - difficult to apply when there are many regressors

The structural model approach

- The bias stemming from covariate measurement error can be avoided by fitting a structural model that includes a measurement model for the covariate via **simultaneous estimation** of:
  - measurement model for the covariate X
  - regression model for the response Y
- Main **advantages** of the structural model approach:
  - standard errors that account for measurement error, so the **inferential procedures** are correct, e.g. it is straightforward to perform a likelihood ratio test for the level 2 variance of *Y*
  - easy to **extend to complex models**, such as models with several covariates, random slopes and categorical responses

Structural model approach: simulations

- Some simulations show the performance of the ML estimation algorithm implemented in M*plus* (Muthén and Muthén, 2007).
- The structural estimator is more efficient than the reliability-adjusted estimator: e.g. for the sample design J = 200 and n = 10 the reduction of the MSE is about 5%.
- A detailed simulation study on the properties of the structural estimator is carried out by Lüdtke *et al.* (2007).

Table 9: *Structural model approach: MC mean, s.e. and MSE of $\widehat{\delta}_s$ for $\delta = 1$ and $\lambda_X = 0.667$ (1000 replications).*

				$\widehat{\delta}_s$			
n	J	τ_X^2	ICC	MC Mean	MC s.e.	$s.e.(\widehat{\delta}_s)\dagger$	MSE
2	1000	1.0	0.5000	0.9992	0.0778	0.0709	0.0060
10	200	0.2	0.1667	1.0006	0.2204	0.2134	0.0485
20	100	0.1	0.0909	1.0067	0.4453	0.4149	0.1981

True values: $\mu_X = 1$, $\sigma_X^2 = 1$; $\alpha = 0$, $\beta_W = 1$, $\delta = 1$, $\tau^2_{Y|X^BX^W} = \sigma^2_{Y|X^BX^W} = 1$

$\dagger$ MC mean of the s.e. calculated by M*plus*.

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.** (2007). Multilevel Modeling with Correlated Effects . *Psychometrika*, in press.
- **Lüdtke O., Marsh H.W., Robitzsch A., Trautwein U., Asparouhov T., Muthén B.** (2007) The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies. *Submitted paper*.
- **Neuhaus J.M. and McCulloch C.E.** (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society B*, 68, 859-872.
- **Snijders T. A. B. and Berkhof J.** (2007). Diagnostic checks for multilevel models. In Jan de Leeuw (Ed.), *Handbook of Multilevel Analysis*. New York: Springer, to appear.

Thanks for your attention!

CLADAG
CLASSIFICATION AND DATA ANALYSIS GROUP OF SIS

Your comments are welcome!

Ask the authors for a draft copy of the paper:

Leonardo Grilli grilli@ds.unifi.it
Carla Rampichini rampichini@ds.unifi.it