STATISTICS FOR COMPLEX PROBLEMS:
THE MULTIVARIATE PERMUTATION APPROACH
AND RELATED TOPICS

in honor of the 70th birthday of Fortunato Pesarin

June 14–15, 2010, Padua

Measurement error induced by sample cluster means in multilevel models

Leonardo Grilli & Carla Rampichini

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

Dipartimento di Statistica "Giuseppe Parenti"
Università di Firenze

A simple random intercept model

- We focus on data with a two-level hierarchy
  - i is the level 1 index (e.g. pupil)
  - j is the level 2 index denoting the clusters (e.g. school)
- Let us consider a random effects model where the response Y (e.g. pupil's final score) depends on a single regressor X (e.g. pupil's prior score)

$$Y_{ij} = \alpha + \beta_{TOT} X_{ij} + u_j + e_{ij}$$

- This model assumes that the between and within effects are equal (i.e. the slope of X is the same both between clusters and within clusters) …
- … but in practice this assumption is too restrictive
- The between effect minus the within effect is known as the **contextual effect** (an important effect in several research fields): this simple model assumes that the contextual effect is null

The contextual random intercept model

- The contextual effect is estimated by adding the ***sample cluster mean*** as a further regressor

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta \bar{X}_j + u_j + e_{ij}$$

β_W within-clusters slope

β_B between-clusters slope

$\delta = \beta_B - \beta_W$ *contextual coefficient*

- However, if the clusters are not fully observed (i.e. the observed units are just a sample of all the units of the cluster in the population)
  → then the sample cluster mean is just an estimate of the population cluster mean (***measurement error***) → then the estimator of δ is downward biased (***attenuation***)

Measurement error of the cluster mean

- The measurement error issue raised by the use of the sample cluster mean is an instance of ***regressor/random-effect correlation***, or ***level 2 endogeneity*** :
  - **Ebbes P., Bockenholt U. and Wedel M. (2004)** Regressor and random-effects dependencies in multilevel models, *Stat. Neerlandica*, 58, 161-178.
  - **Kim J. S. and Frees E. W. (2007)** Multilevel Modeling with Correlated Effects. *Psychometrika*,72, 505–533.
- However, the instance is peculiar since the measurement error model is known and the variances can be estimated

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

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 and X_{ij}^W are independent

The population cluster mean X_j^B is measured with error by the sample cluster mean $\bar{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} = X_j^B + \bar{X}_j^W$

$$Var(\bar{X}_j) = \tau_X^2 + \frac{\sigma_X^2}{n_j}$$

Consequences of measurement error

- In the contextual model, the use of the sample cluster mean in lieu of the population cluster mean has the following consequences:
  - The intercept is biased (but the intercept is usually not of interest)
  - The within slope is unbiased
  - **The contextual coefficient is attenuated**
  - The level 1 variance is unbiased
  - **The level 2 variance is inflated**

  } The ICC is inflated
- The attenuation of the contextual coefficient is a well known fact
- However, the inflation of the level 2 variance (and the consequent inflation of the ICC) is often neglected

Attenuation of the contextual coefficient

Measurement-error-attenuated contextual coefficient $\delta_m = \lambda_X \delta$

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

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

A. The variance ratio of X τ_X^2/σ_X^2 (between-clusters variance on within-clusters variance)

B. the cluster size n (for simplicity we assume a balanced hierarchy)

Remark: the reliability of X depends on both population parameters (the variance ratio) and sampling design (the cluster size)

Measurement error vanishes iff $\delta=0$, anyway δ_m is close to δ when $\lambda_X \approx 1$

λ_X can be far from 1, e.g.

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

Inflation of the level 2 variance

- In the population the covariate X is the sum of a between component X^B and a within component X^W → the population model is conditioned on X^B and X^W → the residual variances of Y are denoted as

$$\text{level 1 variance} \quad \sigma^2_{Y|X^B X^W}$$
$$\text{level 2 variance} \quad \tau^2_{Y|X^B X^W}$$

- In the contextual model with the sample cluster mean, the level 1 variance is unbiased, but **the level 2 variance is inflated**

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

- The bias of the level 2 residual variance of Y depends on
  - the reliability of X
  - the contextual coefficient
  - the level 2 variance of X

A simulation to show the biases

- Monte Carlo means (1000 replications)

True values:
$\beta_W = 1$
δ = see table
variances = 1

Data structure:
J=1000 n=2

Reliability of X:
$\lambda_X = 0.67$

$\delta = \beta_B - \beta_W$	$Y_{ij} = \alpha + \beta_w X_{ij} + \delta \bar{X}_j + \dots$ β_W	δ	$\tau^2_{Y\|X^B X^W}$
-2	1.00	-1.33	2.32
-1.5	1.00	-1.00	1.75
-1	1.00	-0.67	1.33
-0.5	1.00	-0.33	1.08
0	1.00	0.00	1.00
0.5	1.00	0.33	1.09
1	1.00	0.67	1.34
1.5	1.00	1.00	1.76
2	1.00	1.33	2.34

Correcting the measurement error biases

- ***Post-estimation correction via the reliability***: fit the contextual model with the sample cluster mean and then correct the estimates using the estimated reliability of X
  - Our approach (the simplest approach: we investigate its properties and provide a valuable extension to the case of sampling from clusters of finite size)
- ***Structural equation approach***: fit a SEM with 1) *an equation for the main model* (with the cluster mean as a latent variable) and 2) *an equation for the sample cluster mean* (as an indicator of the latent population counterpart)
  - **Croon M.A. and van Veldhoven M.J.P.M (2007)** Predicting Group-Level Outcome Variables From Variables Measured at the Individual Level: A Latent Variable Multilevel Model. *Psychological Methods*, 12, 45–57.
  - **Lüdtke O., Marsh H.W., Robitzsch A., Trautwein U., Asparouhov T. and Muthén B. (2008)** The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies. *Psychological Methods*, 13, 203–229.

Post-estimation correction via the reliability

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

1. Fit the contextual model with the sample cluster mean to estimate:

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

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

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

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

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

Post-estimation correction via the reliability

- Pro
  - simple procedure
  - applied after running standard multilevel software (no need to use software for IRT or SEM)
  - applicable to results published by other researchers
  - can exploit an estimate of the reliability from external data
  - can be extended to the case of sampling from clusters of finite size
- Contra
  - exact only for balanced designs (but performs well also in unbalanced d.)
  - difficult to apply when there are many regressors (unless uncorrelated)
  - difficult to extend to non-linear models
  - the corrected estimators have larger sampling variances

Every correction method *reduces the bias* at the cost of increasing the sampling variance → Here we derive approximate expressions for the sampling variance and MSE of the estimator of the contextual effect

In terms of MSE, correcting the estimate using the reliability is often convenient (e.g. in the previous simulation setting it is convenient for $|\delta| \geq 0.25$

Structural equation approach

- Pro
  - efficient (full model fitted with ML)
  - can be applied to unbalanced designs, models with many covariates, non-linear models
- Contra
  - need specific SEM software (currently it is easily fitted by Mplus)
  - poor performance with few clusters (say < 30)
  - the values of X within each cluster are assumed to be iid → the approach is strictly appropriate only for random sampling from infinite-size clusters

When level 1 observations are sampled from **finite-size clusters**, the SEM approach ***over-estimates the contextual coefficient***

This bias increases with the *within-cluster sampling fraction* (bias is relevant for fractions over 10%)

Sampling from clusters of finite size

- Variance of the sample cluster mean (in case of random sampling of level 1 observations within each cluster):

$$Var(\bar{X}_j) = \begin{cases} \tau_X^2 + \dfrac{\sigma_X^2}{n} & \text{clusters of infinite size} \\ \tau_X^2 + \dfrac{\sigma_X^2}{n}\dfrac{N-n}{N-1} & \text{clusters of finite size} \end{cases}$$

Within-cluster sampling fraction

$$\text{shrinkage factor of the within variance} = \frac{N-n}{N-1} \simeq 1 - \frac{n}{N}$$

- The reliability of X is defined accordingly as

$$\lambda_X^f = \frac{\tau_X^2}{\tau_X^2 + \dfrac{\sigma_X^2}{n}\dfrac{N-n}{N-1}}$$

Larger than the standard reliability

→ if we correct the contextual coefficient with the standard reliability we get an upward bias

Estimating the finite-clusters reliability

- In the case of sampling from clusters of finite size it is not enough to adjust the reliability formula: **also the estimator of the level 2 variance needs a modification**
- Indeed, both ML and REML underestimate the level 2 variance: this is clear from the ANOVA method-of-moments formulae

$$\hat{\sigma}_X^2 = [\text{sample within variance}]$$

$$\hat{\tau}_X^2 = [\text{sample between variance}] - \frac{\hat{\sigma}_X^2}{n}$$

- The estimator of the level 2 variance should be modified as follows

$$\hat{\tau}_{X,f}^2 = [\text{sample between variance}] - \frac{\hat{\sigma}_X^2}{n}\frac{N-n}{N-1}$$

- The finite-clusters reliability is estimated as follows

plug $\hat{\sigma}_X^2$ and $\hat{\tau}_{X,f}^2$ into the expression of λ_X^f

A simulation study: standard reliability vs finite-clusters reliability

Data structure: J=200 n=10

True values: β_W=1 δ= 1 λ_X= 0.67

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta \bar{X}_j + u_j + e_{ij}$$

N	n/N	λ_X^f	MC Mean $\hat{\delta}_m$	MC Mean $\hat{\delta}_c$	MC Mean $\hat{\delta}_c^f$	MSE $\hat{\delta}_m$	MSE $\hat{\delta}_c$	MSE $\hat{\delta}_c^f$
10	1.00	1.000	1.003	2.247	1.003	0.0265	1.6950	0.0265
20	0.50	0.792	0.804	1.361	1.031	0.0595	0.2062	0.0376
40	0.25	0.722	0.725	1.163	1.016	0.0965	0.0896	0.0441
100	0.10	0.688	0.689	1.058	1.009	0.1153	0.0554	0.0434
200	0.05	0.677	0.678	1.032	1.010	0.1231	0.0511	0.0475
1000	0.01	0.669	0.669	1.030	1.003	0.1297	0.0477	0.0492

Within-cluster sampling fraction

No correction

Correction via standard reliability

Correction via finite-clusters reliability

Our work is published as a downloadable WP:

Grilli L. and Rampichini C. (2009) Measurement error in multilevel models with sample cluster means. *Electronic Working Papers 6/2009*, Department of Statistics - University of Florence.

A refined version is currently under revision for publication in a journal: ask us for a pdf copy of the new version!

Other approaches

- An EM algorithm where the population cluster mean is treated as a missing value
  - **Shin Y. and Raudenbush S.W. (2010).** A Latent Cluster-Mean Approach to the Contextual Effects Model with Missing Data. *Journal of Educational and Behavioral Statistics*, 35, 26–53.
- A two-step approach where 1) the population mean is estimated via EB (Empirical Bayes), and 2) the contextual model is fitted with the EB estimate in lieu of the cluster mean
  - **Kuha J., Skrondal A. and Fisher S.**, Work in Progress