A multivariate multilevel model for the analysis of TIMMS & PIRLS data

European Congress of Methodology
July 23 - 25, 2014 - Utrecht

Leonardo Grilli[1], Fulvia Pennoni[2],
Carla Rampichini[1], Isabella Romeo[2]

[1] Department of Statistics, Informatics, Applications 'G. Parenti' - University of Florence, e-mail: grilli@disia.unifi.it, rampichini@disia.unifi.it

[2] Department of Statistics and Quantitative Methods - University of Milano-Bicocca, e-mail: fulvia.pennoni@unimib.it, isabella.romeo@unimib.it

July 24th, 2014

TIMSS AND PIRLS SURVEYS

TIMSS and PIRLS are **large scale assessment surveys** held by the International Association for the Evaluation of Educational Achievement (IEA).

Rutkowski, L., Gonzalez, E., Joncas, M. von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. Educational Researcher

- **TIMSS** (Trends in International Mathematics and Science Study): at *fourth and eighth* grades every four years since 1995;
- **PIRLS** (Progress in International Reading Literacy Study): at *fourth* grade every five years since 2001.

In 2011 - for the first time - TIMSS and PIRLS cycles coincided.

The **TIMSS&PIRLS 2011 Combined International Database** concerns *fourth grade* students and collects data from questionnaires administrated to students, parents, teachers, and school principals.

TIMSS & PIRLS 2011 DATA

Sample design

Two stage (according to the hierarchical structure):

- schools are first sampled proportionally to their size (number of students)
- then 1 or 2 classes are randomly sampled and all students are interviewed

Martin, M. O., Mullis, I. V. S. (2012). Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Italian TIMSS&PIRLS 2011 sample:

- 4125 students nested in 239 classes nested in 202 schools
- 483 teachers (each class may have from one to three teachers)

Plausible Values

Rotating scheme of item administration → each student answers a subset of items in order to

- minimize testing burden
- ensure accurate population estimates

⇒ For any student, the total score is missing and replaced by five multiple imputations (Plausible Values)

Plausible Values (PVs)

- *WHAT are PVs?*: they are random draws from the distribution of the total score derived from an IRT model. Mislevy (1991) Randomization-based inference about latent variables from complex samples, Psychometrika.
- *HOW using PVs?*: run separate analyses with each PV and combine the results through multiple imputation procedures. Rubin (1987) Multiple imputation for nonresponse in sample surveys.

OBJECTIVES OF THE ANALYSIS

Using TIMSS&PIRLS 2011 data for Italy, we aim to

- explore the relationships among performances in the three subjects: Reading, Math and Science
- analyse the determinants of the achievement at different hierarchical levels (students and classes)
- perform effectiveness analysis at class level

⇒ We need a model that is both **multilevel** (students in classes) and **multivariate** (Reading, Math and Science)

Remark: to the best of our knowledge, all reports and papers exploit multilevel models for a single outcome - no multivariate modelling!

THE MULTIVARIATE MULTILEVEL MODEL

Features

- the three scores on Reading, Math and Science are a joint outcome
- the level 2 is represented by classes (instead of schools) since several factors act at the class level (e.g. peer effects)
- the school is not added as level 3 since in most schools only one class was sampled (however, *cluster-robust standard errors* are used)

Advantages

- estimating the (residual) correlations between pairs of outcomes at both hierarchical levels
- testing whether the effects of the covariates are identical across outcomes (e.g. differences between males and females are the same in Reading and Math?)

MODEL EQUATION

We specify the following *multivariate two-level* model:

$$Y_{mij} = \alpha_m + \boldsymbol{\beta}_m \mathbf{x}_{mij} + \boldsymbol{\gamma}_m \mathbf{w}_{mj} + u_{mj} + e_{mij}$$

- outcome m with $m = 1, 2, 3$ (1: Reading, 2: Math, 3: Science)
- student i with $i = 1, \ldots, n_j$
- class j with $(j = 1, \ldots, 239)$
- $\mathbf{x}_{mij}$ vector of student-level covariates
- $\mathbf{w}_{mj}$ vector of class-level covariates (also including covariates at higher level, e.g. school or province)
- u_{mj} class-level errors
- e_{mij} student-level errors

Remark: the model allows for outcome-specific covariates, e.g. the experience of the teacher

Model errors: covariance matrices

Student-level errors: $\mathbf{e}'_{ij} = (e_{1ij}, e_{2ij}, e_{3ij})$ Class-level errors: $\mathbf{u}'_j = (u_{1j}, u_{2j}, u_{3j})$

- $\mathbf{e}_{mij}$ indep. across students, $\mathbf{u}_{mj}$ indep. across classes
- $\mathbf{e}_{mij}$ independent from $\mathbf{u}_{mj}$
- $\mathbf{e}_{mij}$ and $\mathbf{u}_{mj}$ multivariate normal with zero means

Covariance matrix at student level

Covariance matrix at class level

$$Var(\mathbf{e}_{ij}) = \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{pmatrix} \quad Var(\mathbf{u}_j) = \boldsymbol{T} = \begin{pmatrix} \tau_1^2 & \tau_{12} & \tau_{13} \\ & \tau_2^2 & \tau_{23} \\ & & \tau_3^2 \end{pmatrix}$$

$\mathbf{Y}_{ij} = (Y_{1ij}, Y_{2ij}, Y_{3ij})'$ has residual covariance matrix $\mathbf{\Sigma} + \boldsymbol{T}$.

We tried several alternative specifications (e.g. heteroscedastic class-level errors) but with no significant improvement of the fit

Grilli L., Rampichini C. (2014) Specification of random effects in multilevel models: a review. Quality & Quantity (to appear)

Model fitting

- **Estimation sample**: 3741 students in 237 classes (indeed, 284 students and 2 classes have been excluded due to missing values in the covariates)
- **Estimation method**: maximum likelihood
- **Plausible values**: estimation is performed separately for each of the five plausible values and then results are combined using Multiple Imputation (MI) formulas (Rubin, 1987)
- **Software**: `mixed` and `mi` commands of Stata 13

Next steps

- results from the null model
- model selection
- results from the final model

Results from the null model

Decomposition of the correlation matrix:

	Correlations									*% Between class of (co)variances*		
	Within class			*Between class*			*Total*					
Subject	Read	Math	Scie	Read	Math	Scie	Read	Math	Scie	Read	Math	Scie
Read	1.00			1.00			1.00			19.8		
Math	**0.71**	1.00		**0.93**	1.00		0.76	1.00		29.5	28.8	
Science	**0.81**	**0.74**	1.00	**0.97**	**0.98**	1.00	0.85	0.81	1.00	28.2	35.0	29.4

- Correlations among outcomes are higher **between classes** rather than **within classes**

RESULTS FROM THE NULL MODEL (CONT.)

Decomposition of the correlation matrix:

	Correlations									*% Between class of (co)variances*		
	Within class			*Between class*			*Total*					
Subject	Read	Math	Scie	Read	Math	Scie	Read	Math	Scie	Read	Math	Scie
Read	1.00			1.00			1.00			**19.8**		
Math	0.71	1.00		0.93	1.00		0.76	1.00		29.5	**28.8**	
Science	0.81	0.74	1.00	0.97	0.98	1.00	0.85	0.81	1.00	28.2	35.0	**29.4**

- **Reading** has the lowest percentage of class-level variance, maybe because it is the subject more influenced by the student background characteristics

Model selection strategy

The selection process in principle requires fitting the multivariate model repeatedly, each time combining the estimates with MI

To speed up the process, we adopt two simplifications:

- the outcomes are analyzed separately by means of **univariate multilevel models**, retaining covariates being significant in at least one of the univariate models
- the estimation is carried out using only the **first plausible value** (underestimated standard errors ⇒ conservative selection of the covariates)

Covariates are added in the following *hierarchical order*:

student, teacher, class, school, province

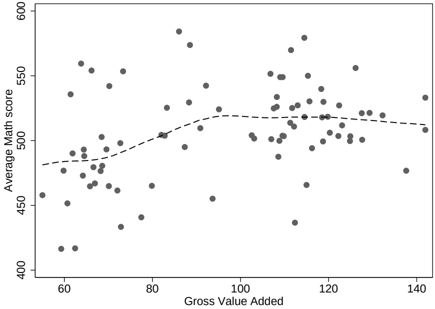
Remark: we center continuous covariates at their sample grand means, and we do not center student-level covariates at their class-level means

Gross Value Added (GVA)

We control for differences in wealth across Italy by means of the *per capita* Gross Value Added (GVA) at market prices in 2011.

The GVA is measured for each of the 110 Italian provinces, ranging from 45 to 142 (national average = 100).

The relationships between the achievement scores and the GVA are explored through local polynomial regression (see the plot for Math)



- The line for GVA< 100 (national average) has a significant positive slope,
- the line for GVA> 100 is nearly flat and the slope is not significantly different from zero. $\Rightarrow$ We constrain to zero the slope of the second line of the spline (i.e. GVA> 100).

Selected covariates

Student characteristics

- **Gender**
- **Language spoken at home**
- **Pre-school**
- **Home resources for learning** [1]
- **Early literacy/numeracy tasks** [2]

[1] It is derived from items on the number of books and study supports available at home and parents' levels of education and occupation (Martin & Mullis, 2013).

[2] It is derived from the parents' responses to how well their child could do some early literacy and numeracy activities when he/she began primary school (Martin & Mullis, 2013).

Teacher characteristics

- Gender
- Years been teaching
- Degree

Class and School characteristics

- % Students attended pre-school
- % Language spoken at home is not Italian
- Average of home resources for learning
- Average of Early literacy/numeracy tasks
- Average of home resources for learning
- Average of Early literacy/numeracy tasks
- School is safe and orderly
- School with Italian students $>90\%$ [1]
- $<$ 10% of students has a low SES [1]
- School is located in a big area[1]
- Area with more than 50.000 inhabitants [1]
- 6 days of school per week
- **Adequate environment and resources** [1]
- **GVA** [2]

[1] Declared by the school principal

[2] *per capita* Gross Value Added (GVA) at market prices in 2011 (proxy of the school socio-economic context)

ESTIMATED REGRESSION COEFFICIENTS

Estimates and robust standard errors of the selected multivariate multilevel model (MI combined results)

	Read		Math		Science		Test *F*
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	*p*-value
Intercept	531.73	3.57	514.99	4.25	531.47	3.92	**0.0006**
Student covariates							
Female	2.92	2.41	-11.96	3.05	-10.64	2.28	**0.0000**
Language at home is not Italian	-22.57	3.12	-14.94	3.27	-23.74	3.53	**0.0161**
Pre-school	8.85	3.01	8.46	2.51	10.91	3.15	0.6386
Home resources for learning	14.04	0.84	10.64	0.84	13.23	0.93	**0.0009**
Early literacy/numeracy tasks	7.24	0.77	10.07	0.76	6.53	0.83	**0.0051**
School covariates							
Adequate environment & resources	5.28	1.92	8.61	3.19	7.00	2.96	0.1950
Province covariates							
GVA (below 100)	0.45	0.15	0.48	0.21	0.55	0.20	0.3983

Joint test *F*

Test *F* for the equality of regression coefficients among the three outcomes:

$H_0 : \beta_{Read} = \beta_{Math} = \beta_{Science}$

Except for Pre-school, student-level covariates have significantly different effects

ESTIMATED REGRESSION COEFFICIENTS (CONT.)

	Read		Math		Science		Test *F*
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	*p*-value
Intercept	531.73	3.57	514.99	4.25	531.47	3.92	0.0006
Student covariates							
Female	**2.92**	2.41	**-11.96**	3.05	**-10.64**	2.28	0.0000
Language at home is not Italian	-22.57	3.12	-14.94	3.27	-23.74	3.53	0.0161
Pre-school	8.85	3.01	8.46	2.51	10.91	3.15	0.6386
Home resources for learning	14.04	0.84	10.64	0.84	13.23	0.93	0.0009
Early literacy/numeracy tasks	7.24	0.77	10.07	0.76	6.53	0.83	0.0051
School covariates							
Adequate environment & resources	5.28	1.92	8.61	3.19	7.00	2.96	0.1950
Province covariates							
GVA (below 100)	0.45	0.15	0.48	0.21	0.55	0.20	0.3983

Gender

Females have a significantly lower performance in Math and Science, but not in Reading.

ESTIMATED REGRESSION COEFFICIENTS (CONT.)

	Read		Math		Science		Test *F*
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	*p*-value
Intercept	531.73	3.57	514.99	4.25	531.47	3.92	0.0006
Student covariates							
Female	2.92	2.41	-11.96	3.05	-10.64	2.28	0.0000
Language at home is not Italian	**-22.57**	3.12	**-14.94**	3.27	**-23.74**	3.53	0.0161
Pre-school	8.85	3.01	8.46	2.51	10.91	3.15	0.6386
Home resources for learning	**14.04**	0.84	**10.64**	0.84	**13.23**	0.93	0.0009
Early literacy/numeracy tasks	**7.24**	0.77	**10.07**	0.76	**6.53**	0.83	0.0051
School covariates							
Adequate environment & resources	5.28	1.92	8.61	3.19	7.00	2.96	0.1950
Province covariates							
GVA (below 100)	0.45	0.15	0.48	0.21	0.55	0.20	0.3983

Read&Science vs Math

Family background covariates have a similar effect on Read and Science, as opposed to Math

⇒ the abilities required for Science seem to be closer to those for Read

Likely, this is a consequence of the way Science is taught in Italian primary schools.

Estimated regression coefficients (cont.)

	Read		Math		Science		Test *F*
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	*p*-value
Intercept	531.73	3.57	514.99	4.25	531.47	3.92	0.0006
Student covariates							
Female	2.92	2.41	-11.96	3.05	-10.64	2.28	0.0000
Language at home is not Italian	-22.57	3.12	-14.94	3.27	-23.74	3.53	0.0161
Pre-school	8.85	3.01	8.46	2.51	10.91	3.15	0.6386
Home resources for learning	14.04	0.84	10.64	0.84	13.23	0.93	0.0009
Early literacy/numeracy tasks	7.24	0.77	10.07	0.76	6.53	0.83	0.0051
School covariates							
Adequate environment & resources	5.28	1.92	8.61	3.19	7.00	2.96	0.1950
Province covariates							
GVA (below 100)	**0.45**	0.15	**0.48**	0.21	**0.55**	0.20	0.3983

Gross Value Added (GVA)

- The effect of GVA is modelled by a **linear spline with a single knot in 100** (the national average) ⇒ GVA has a significant effect only for provinces below the national average, with no significant difference across outcomes.
- For the province with the lowest value of GVA (55) the effect is minus 22.5 points.

EXPLAINED VARIANCES AND RESIDUAL ICC'S

The **proportions of variance explained by the final model with respect to the null model** are higher at class level:

- the within-class variances reduce by 15% for the three outcomes
- the between-class variances reduce by 33% for Reading, 20% for Math and 26% for Science

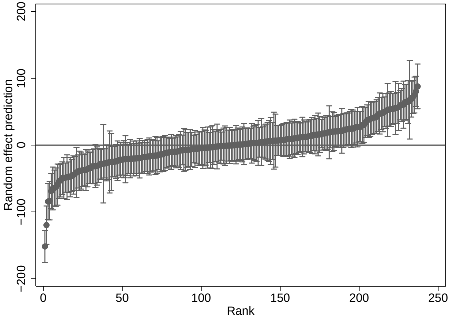
⇒ compositional and contextual effects are more relevant for the achievement in Reading

The **residual ICC**'s are quite high: 16% for Reading, 28% for Math and 27% for Science ⇒ relevant unobserved class-level factors

The correlations among outcomes are similar to those in the null model.

EMPIRICAL BAYES RESIDUALS

The level 2 error (*class random effect*) u_{mj} is the contribution of class j to the achievement of students in outcome m (it may be interpreted in terms of *effectiveness*)



Empirical Bayes residuals for Math with 95% confidence intervals

- *good* classes (CI above 0): students on average achieve substantially more than expected on the basis of the covariates
- *poor* classes (CI below 0): students on average achieve substantially less than expected

The residuals give an indication or further territorial differences not captured by GVA:

- in North-West *good* classes prevail on *poor* classes, while in the Centre the pattern is reversed (we tried to add geographical dummies in the fixed part of the model ⇝ not significant)
- in the South there are high percentages of both *good* and *poor* classes ⇒ greater variability of achievement (we tried to specify heteroscedastic random effects ⇝ not significant)

FINAL REMARKS

- Outcomes in Reading, Math and Science from large-scale assessment surveys are usually studied one by one (univariate multilevel models)
- Using the Italian subset of the TIMSS&PIRLS 2011 combined dataset, we performed a joint analysis of achievement in Reading, Math and Science by means of a *multivariate multilevel* model ⇒ the multivariate approach allowed us to obtain the following findings:
  - *estimating correlations among outcomes*: we found that correlations at class level are higher than correlations at student level (so high that the three outcomes yield the same results of school/class effectiveness)
  - *testing for differential effects of covariates on the outcomes*: we found that background covariates have similar effects on Reading and Science, as opposed to Math; moreover, females have a lower performance in Math and Science, but not in Reading

FINAL REMARKS (CONT.)

- We accounted for territorial differences in wealth through the Gross Value Added (GVA) at province level (instead of adding dummy variables for geographical areas → more interesting interpretation)
- The class-level Empirical Bayes residuals allowed us to identify *good* and *poor* classes and to point out further territorial patterns concerning both the mean and the variance (e.g. greater variability of achievement in the South of Italy, not included in the model due to lack of statistical significance)

thanks for your attention :-)