Multivariate multilevel modelling of student achievement data

Modelli multilivello multivariati per l'analisi dell'apprendimento degli studenti

Leonardo Grilli, Fulvia Pennoni, Carla Rampichini and Isabella Romeo

Abstract The paper analyses the Italian sample of the TIMSS&PIRLS 2011 Combined International Database, containing the results of the assessment on Reading, Mathematics and Science for a sample of 4th grade pupils, alongside with several variables about the pupils, their teachers and their schools. This is the first time that TIMSS and PIRLS surveys are conducted on the same sample, thus allowing to jointly analyse the achievement for the considered subjects. We propose a multivariate multilevel model, with pupils nested within classes, to explore the determinants of achievement in the three subjects.

Abstract *L'articolo analizza il campione italiano dell'archivio internazionale congiunto TIMSS&PIRLS 2011, contenente i risultati della valutazione in Lettura, Matematica e Scienze per studenti di quarta elementare, e altre variabili sulle caratteristiche di studenti, insegnanti e scuole. Questa è la prima volta che le indagini TIMSS e PIRLS sono condotte sullo stesso campione, permettendo quindi di analizzare congiuntamente i risultati nelle tre materie. Proponiamo un modello multilivello multivariato, nel quale i bambini sono raggruppati in classi, per individuare le determinanti dei risultati nelle tre materie.*

Key words: hierarchical linear model, large-scale assessment data, school effectiveness, secondary data analysis

Leonardo Grilli, Carla Rampichini
Department of Statistics, Informatics, Applications 'G. Parenti', University of Florence,
e-mail: grilli@disia.unifi.it, e-mail: rampichini@disia.unifi.it

Fulvia Pennoni, Isabella Romeo
Department of Statistics and Quantitative Methods, University of Milano-Bicocca,
e-mail: fulvia.pennoni@unimib.it, e-mail: isabella.romeo@unimib.it

1 Introduction

We aim to investigate the determinants of student achievement by exploiting a data set combining the large-scale assessment surveys TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study). The TIMSS&PIRLS 2011 Combined International Database is released by the International Association for the Evaluation of Educational Achievement (IEA) [1]. This is the first time that the TIMSS and PIRLS cycles coincide, thus allowing to assess the same students in three subjects, enabling to explore the relationships among performances in Reading, Mathematics and Science.

In the official report [2] the three results are analyzed separately, without exploiting relationships among them. Even research papers usually analyse the results subject by subject. We overcome this limitation by considering the three scores on Reading, Math and Science as a joint outcome, measured at pupil level, with pupils nested within classes. This multivariate analysis allows us to decompose the residual variances and covariances among the three scores into a pupil-level and class-level components. Moreover, the multivariate model allows the covariates to have different effects on the three responses. Preliminary results show that the three scores are highly correlated, in particular at class level. Moreover, the proportion of variability of scores at class level is relevant, thus calling for an analysis of teacher effects and contextual factors, such as family background, peer effects, school resources and economic well-being of the geographical area of the school. In particular, for the last factor, we consider the gross value-added at the province level.

The remainder of this paper is organized as follows: Section 2 provides a description of the examined dataset, Section 3 illustrates the proposed model and Section 4 summarizes the main findings.

2 Data

In this paper we consider data from the Combined International Database which includes data on 32 countries from both TIMSS 2011 and PIRLS 2011. TIMSS is an international assessment of mathematics and science at the fourth and eighth grades, conducted every four years since 1995. PIRLS provides information on trends in reading literacy achievement of fourth grade students every five years since 2001. The data at hand are derived from instruments administered to the students of fourth grade, their parents, their teachers, and their school principals. The combined dataset includes only pupils that responded to both assessments, with a marginal reduction in sample size.

The Italian TIMSS&PIRLS combined sample includes 4125 pupils nested in 239 classes, that in turn are nested in 202 schools. Each class may have from one to three teachers, with different combinations of subjects. The total number of teachers is 483, many classes have two teachers. As explained in [1], achievement scales were produced for Reading, Mathematics, and Science. Each scale has an international

mean of 500 points and a standard deviation of 100 points. For each achievement scale, the TIMSS and PIRLS 2011 Combined International Database provides five separate estimates of each student's score, known as *plausible values*. Variability between plausible values encapsulates the uncertainty inherent in the scale estimation process [4].

In order to adjust the observed achievement scores we consider a subset of the many variables available in the combined data set. The choice of this subset is driven by theoretical arguments, exploratory data analysis and previous studies [2]. The variables at student level include gender, pre-school and language spoken at home as well as home background context questionnaire scales [2]. The variables at teacher level include gender, age class, education and years been teaching. Class variables are defined as averages or proportions of the corresponding student-level covariates. Among the school variables there are those derived from student-level variables and those measured at school level such as those concerning the environment order and security as well as some other school characteristics declared by the school principal. As a proxy of the socio-economic context where the schools operate, we consider the *per capita* Gross Value Added (GVA) at market prices in 2011 which is an index number measured for each of the Italian provinces (Italy = 100). This index varies a lot among provinces, ranging from 55 to 142.

3 Model specification

Let Y_{mij} be the value of the m-th score for the i-th pupil of the j-th class, with $m = 1,2,3$ (1=Reading, 2=Math, 3=Science), $i = 1,\ldots,n_j$, $j = 1,\ldots,J$. The number of pupils of the j-th class is denoted with n_j, thus $\sum_{j=1}^{J} n_j = N$ is the total number of pupils.

Official reports [2] consider pupils as level 1 units and schools as level 2 units. However, in our model the level 2 is represented by classes since several factors act at the class level (e.g. the teacher), thus merging different classes could obscure some sources of variability. The schools could be considered as a third hierarchical level with its own random effect. However, in the Italian sample only 37 schools out of 202 have more than one class, thus it is not feasible to add school-level random effects. Nevertheless, the school characteristics are included in the model as covariates and the correlation between classes of the same school is accounted for by computing robust standard errors using the formula for clustered observations. We specify the following multivariate two-level model [3] for response m of pupil i in class j:

$$Y_{mij} = \alpha_m + \boldsymbol{\beta}_m \mathbf{x}_{mij} + \boldsymbol{\gamma}_m \mathbf{w}_{mj} + \boldsymbol{\delta}_m \mathbf{z}_{mj} + u_{mj} + e_{mij} \tag{1}$$

where $\mathbf{x}_{mij}$ is the vector of pupil's covariates, $\mathbf{w}_{mj}$ is the vector of teacher covariates, and $\mathbf{z}_{mj}$ is the vector of class-level covariates, including both class and school characteristics. All the vectors have the subject index m since, in principle, they can include subject-specific covariates. Pupil-level errors e_{mij} are assumed inde-

pendent across pupils, and class-level errors u_{mj} are assumed independent across classes. The errors e_{mij} are independent from the errors u_{mj}. Moreover, pupil-level errors $\mathbf{e}'_{ij} = (e_{1ij}, e_{2ij}, e_{3ij})$ and class-level errors $\mathbf{u}'_j = (u_{1j}, u_{2j}, u_{3j})$ are assumed to be multivariate normal with zero means. The basic specification assumes constant covariance matrices for the errors, i.e. pupil-level errors and class-level errors have covariance matrices, respectively:

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

Therefore, the response vector $\mathbf{Y}_{ij} = (Y_{1ij}, Y_{2ij}, Y_{3ij})'$ has residual covariance matrix $\boldsymbol{\Sigma} + \boldsymbol{T}$. The class-level residual u_{mj} represents the unobserved factors at class level for subject m. Since the school level is omitted, u_{mj} also accounts for school unobserved factors. Moreover, since each teacher operates in only one class, u_{mj} also collects unobserved factors of the teacher.

In order to exploit estimation algorithms for models with nested random effects, we view the multivariate two-level model (1) as a single-outcome three-level model with subjects at level 1, pupils at level 2 and classes at level 3. To this end we expand the data set, considering three observations for each pupil, one for each of the three scores.

4 Main findings

The model selection process is based on the first plausible value, whereas the results for the final model (not reported in this manuscript) will exploit the five plausible values by combining the results through multiple imputation.

The analysis is based on the multivariate two-level model (1) with student-level variance-covariance matrix and class-level variance-covariance matrix (2). Model fitting is performed by means of the `mixed` command of the Stata software.

The model selection process begins with the null model, which has three means and twelve variance-covariance parameters. The results of the null model (Table 1) are summarized in terms of correlations among subjects and class-level proportions. The *within class* and *between class* correlation matrices are derived from the corresponding covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{T}$ of equation (2), respectively, whereas the total correlation matrix is derived from the total covariance matrix $\boldsymbol{\Sigma} + \boldsymbol{T}$. Note that the null model decomposes the observed variances and covariances, thus the estimated total covariance matrix is approximately equal to the sample covariance matrix. Table 1 shows that the three scores are highly correlated, in particular at class level. The rightmost matrix in Table 1 reports the percentage of variances and covariances at the class level, namely each element of $\boldsymbol{T}$ is divided by the corresponding element of $\boldsymbol{\Sigma} + \boldsymbol{T}$. For example, the percentage of class level variance for Reading is 19.9, obtained as $100 \times \hat{\tau}_1^2/(\hat{\sigma}_1^2 + \hat{\tau}_1^2)$. Note that, Reading has the lowest

Table 1 Two-level multivariate null model on subject scores: correlation matrix decomposition. TIMSS&PIRLS 2011, Italy.

	Correlations									*% Var/Cov*		
	Within class			*Between class*			*Total*			*between class*		
Subject	Read	Math	Science	Read	Math	Science	Read	Math	Science	Read	Math	Science
Read	1.00			1.00			1.00			19.9		
Math	0.71	1.00		0.93	1.00		0.76	1.00		28.9	28.0	
Scie	0.81	0.74	1.00	0.97	0.98	1.00	0.85	0.81	1.00	28.0	34.9	29.9

percentage of class-level variance, maybe because it is the subject more influenced by the student background characteristics. As for the class-level covariance, Math and Science show the highest proportion, probably because these subjects are typically taught by the same teacher. The proportion of variability of scores at class level is relevant, thus calling for an analysis of contextual factors. We add the covariates in the following hierarchical order: student characteristics, teacher variables, class and school variables and GVA by province. In order to enhance the interpretability of the intercept, we centered continuous covariates at their sample grand means, except for the index GVA which is centered at the national average. We fit models including both the student-level covariates and their class means. Note that, contrary to the official reports [2], we do not center student-level covariates at their class-level means, thus the coefficients of class-mean covariates represent the contextual effects. None of the contextual effects turns out to be significant, namely the final model does not include class-level means.

At student level, all the considered variables turn out to be significant. The teacher variables considered are not significant, indicating homogeneity between classes of these characteristics. Furthermore, among the class contextual factors that have been considered, none is significant in the multivariate model. At school level, the only significant variable is the adequate environment and resources. This highlights the importance of school context in explaining student performance.

Table 2 reports the estimates and robust standard errors of the selected multivariate multilevel model. The units with missing data in a variable are not included the analysis. Therefore 284 students and 2 classes of the total are not considered in the analysis. The final model is then fitted on 3741 students and 237 classes.

The intercepts in Table 2 represent the average scores for the baseline student (male, language spoken at home is Italian, no pre-school, and all the other covariates are set at mean values). The performance of the baseline student is beyond the international mean of 500 in all the considered subjects, though the average score in Math is substantially lower than the average scores in Reading and Science. All the covariates have the expected signs and are significant for all the considered subjects, except for being female, which has a negative effect for Math and Science, but no significant effect for Reading. At school level, disposing of adequate environment and resources helps to reach a higher score on all subjects, especially for Math.

The socio-economic context is measured by the index GVA, whose effect is modelled by a linear spline with a single knot in 100 (the national average). It turns out that GVA has a significant effect only for provinces below the national average.

Table 2 Multivariate multilevel model of equations $(1)-(2)$: estimated regression coefficients (TIMSS&PIRLS 2011, Italy, 1st plausible value).

	Read		Math		Science	
	Coef.	Robust s.e.	Coef.	Robust s.e.	Coef.	Robust s.e.
Intercept	530.53	3.32	516.71	3.67	533.01	3.74
Student						
Female	3.04	2.27	-14.16	2.23	-11.27	2.21
Language spoken at home is not Italian	-22.13	2.98	-14.71	2.88	-21.95	2.68
Pre-school	10.44	2.56	8.37	2.48	10.65	2.60
Home resources for learning	14.01	0.77	10.61	0.78	13.32	0.80
Early literacy/numeracy tasks	7.45	0.70	10.23	0.72	6.09	0.73
School						
Adequate environment and resources	5.36	1.73	7.95	3.10	6.79	2.85
Province						
GVA (below 100)	0.46	0.15	0.55	0.20	0.61	0.20

The effect of GVA is about half point in the score for each point in the index: considering that the index has a minimum at 55, the largest effect of GVA is about $-45 \times 0.5 = -22.5$ (actually, the effect is lower for Reading and higher for Math and Science).

The analysis will be completed by the inspection of model residuals, which can be interpreted as effectiveness measures, in order to compare schools and detect outlying institutions

Acknowledgements The research has been supported by the grant *"Finite mixture and latent variable models for causal inference and analysis of socio-economic data"* (FIRB - Futuro in ricerca) funded by the Italian Government (RBFR12SHVV).

References

1. Foy, P.: TIMSS and PIRLS 2011 user guide for the fourth grade Combined International Database. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College (2013)
2. Foy, P., O'Dwyer, L.M.: Technical Appendix B. School Effectiveness Models and Analyses. in: TIMSS and PIRLS 2011 Relationships report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College (2013)
3. Snijders, T.A.B., Bosker, R.J.: Multilevel analysis: an introduction to basic and advanced multilevel modelling (2nd ed). SAGE Publications, Inc (2011)
4. Wu, M.: The role of plausible values in large-scale surveys. Studies in Educational Evaluation 31, 114–128 (2005)