Multilevel modeling of heterogeneity in math achievements: different class- and school-effects across Italian regions

Indagine delle differenze nei livelli di apprendimento in matematica in Italia mediante modelli a effetti misti

Tommaso Agasisti, Francesca Ieva and Anna M. Paganoni

Abstract Catching the differences in educational attainments between groups of students and across schools is becoming increasingly interesting. With the aim of assessing the extent of these differences in the Italian educational system, the paper applies multilevel modeling to a dataset containing detailed information of students' math attainments at grade 6 in 2011/12, provided by the Italian Institute for the Evaluation of Educational System.

Abstract *Obiettivo del lavoro è indagare i determinanti delle differenze nei livelli di apprendimento in matematica nelle diverse scuole d'Italia. A tal fine vengono impiegati modelli multilivello con raggruppamento per classe e scuola, stratificando sulle aree geografiche. I dati oggetto di studio sono i risultati dei test di matematica degli studenti di prima media nell'anno 2011/2012, rilevati dall'Istituto Nazionale per la Valutazione del Sistema educativo di Istruzione e di formazione.*

Key words: Multilevel models, School effectiveness, Education

1 Introduction

The institutional organization of the Italian educational system is based on strong assumptions about its equality purposes, among which a key role is assigned to the

Tommaso Agasisti
DIG - Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milano (Italy) e-mail: tommaso.agasisti@polimi.it

Francesca Ieva
Name, Department of Mathematics, Università degli Studi di Milano, Milano (Italy) e-mail: francesca.ieva@unimi.it

Anna M. Paganoni
MOX - Modeling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Milano (Italy) e-mail: anna.paganoni@polimi.it

presumption that all schools provide similar educational standards. Recent aggregate data provided by the Italian Institute for the Evaluation of Educational System (hereafter, Invalsi) show that it is not the case, and that a significant portion of variance in students' test scores is attributable to structural between-schools differences. This evidence is accompanied by a specific feature of the Italian educational system, namely a strong difference in educational attainment and results in different geographical macro-areas [1], with students in Northern Italy obtaining (on average, and all else equal) higher scores than their counterparts in Central and Southern part of the country. While the determinants of this gap are still not completely clear, the empirical evidence illustrates that also between-schools differences are stronger in the South than in the North. In this perspective, a study of school effects on achievement for the different areas of the country seems worthy of specific attention. Additionally, it is important to investigate the relationship between achievement and variables measuring students and schools' characteristics, together with estimates of the relative weights of the two levels of grouping (classes and schools). This study is inserted into a stream of the applied statistics literature, which uses multilevel models to investigate the relative impact of different sets of observable variables on students' achievement. Some studies used these methods in measuring specific phenomena, such as the differences between performances of native and immigrant students (see [3], among others).

2 Data

The Math Score Invalsi database collects the achievement in math tests of pupils attending the first year of junior secondary school. Several information are provided at pupil, class and school level. A complete description of these variables is reported in [2]. The outputs (MS, i.e., the score in the Mathematics standardized test administered by Invalsi) are expressed as ”cheating-corrected” scores (CMS). There is also the score in the Math test at grade 5 (CMS5), which is used as a control in the multilevel model to specify a Value-Added (VA) estimate of the school's fixed effect. In fact, it is well known from the literature that education is a cumulative process.The empirical analysis can be then better conducted in a VA fashion, namely considering the role of variables statistically correlated with test scores. In a cross-section setting, like the one of this paper, it is then important to include prior achievement among independent variables; in this case, we have information about the test score of the $i-th$ individual in the prior academic year and we use it when estimating the model parameters. However, the procedure of matching individual data longitudinally at student level is new in Italy, and still undergoing improvements - the main problems are related to the transmission of information from schools to the Ministry. Unfortunately, for the year under scrutiny this procedure led to the loss of around half of the observations (precisely, 46.5%). The database consists of $509,360$ records, within $25,922$ classes and $5,311$ schools. They represent the entire population of children from the first year of junior secondary schools in Italy. If we consider only statistical units with no missing information, the database reduces to $259,757$ records, within $18,761$ classes and $4,119$ schools.

3 Models, Methods and Results

The output of interest in our analyses is the CMS of students attending the first year of junior secondary school. It is a normalized score ranging from 0 to 100, with median equal to 46.94 and mean value (std.dev.) equal to 61.05 (17.74). For each geographical area $R = \{Northern, Central, Southern\}$ we fit the following model:

$$y_{ij}^{(R)} = \beta_0^{(R)} + \sum_{k=1}^{K} \beta_k^{(R)} x_{kij}^{(R)} + b_j^{(R)} + \varepsilon_{ij}^{(R)} \tag{1}$$

$$b_j^{(R)} \sim \mathscr{N}(0, \sigma_b^{2(R)}) \quad \varepsilon_{ij}^{(R)} \sim \mathscr{N}(0, \sigma_\varepsilon^{2(R)}) \tag{2}$$

$$\hat{b}_j^{(R)} = \gamma_0^{(R)} + \sum_{l=1}^{L} \gamma_l^{(R)} z_{lj}^{(R)} + \eta_j^{(R)} \qquad \eta_j^{(R)} \sim \mathscr{N}(0, \sigma_\eta^{2(R)}) \tag{3}$$

$i = 1, \dots, n_j^{(R)}$ and $j = 1, \dots, J^{(R)}$. Table 1 shows the resulting estimates. It is worth noting the difference in Percentage of Variation captured by Random Effects (also called Variance Partitioning Coefficient, VPC) over the three geographical areas. VPC is obtained as the proportion of random effects variance over the total variation, i.e., $\frac{\sigma^2_{b^{(R)}}}{\sigma^2_{b^{(R)}} + \sigma^2_{\varepsilon^{(R)}}}$. The findings highlight that the educational production functions look quite different across the three geographical areas.

Looking at the estimates of the schools' effects $b_j^{(R)}$s, they are characterized by a greater variability in the Southern area. Figure 1 shows the distributions of the random effects estimated by fitting model (1) to the North, Center and South database respectively. They reflect the differences in variation we appreciated from computing PVRE in Table 1.

A further aspect that is interesting is to provide some empirical evidence about the main characteristics of the schools that exert a positive/negative effect on students' achievement. A potential approach for this purpose is to investigate substantially which are the main feature that can ”explain” (in a correlational, not causal way) the schools' effect $b_j^{(R)}$, $j = 1, \dots, J^{(R)}$. Once the model in (1) is fitted to the data concerning each geographical area, we try to model the estimates of the random effects by means of suitable school-level covariates. To this aim, the Lasso regression is an efficient variable selection algorithm. The penalization parameter is chosen by cross-validation techniques.

Table 2 shows the resulting models selected by Lasso regression.

Even if the collinearity issue can be addressed by using penalized regression techniques, the amount of unexplained variability remains high. This is probably due to the unobserved variables like those that reflect the kind of activities which are un-

Table 1 ML estimates for model (1). Asterisks denote different levels of significance: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$

Fixed effects	NORTH	CENTER	SOUTH
Intercept	1.157***	7.914***	16.833***
Female	-1.695***	-2.659***	-2.141***
1st generation Immigrant	-0.623***	-0.590	0.436
Late-enrolled student	-2.566***	-1.794***	-3.933***
ESCS *	1.943***	2.428***	3.181***
Student NOT living with both parents	-1.216***	-1.335***	-1.485***
CMS5	0.700***	0.571***	0.387***
Random effects	NORTH	CENTER	SOUTH
σ_b	3.645	4.510	7.354
σ_ε	12.434	13.527	14.622
PVRE	7.91%	10%	20.18%
Size	NORTH	CENTER	SOUTH
Number of Observations	130,256	46,529	82,972
Number of Groups (schools)	1,843	712	1,564

* Economic and Social Cultural Status

dertaken within classes of each school, together with those at school level. In other words, part of the school effect is actually driven by differences between classes of the same school, so exploring the variance between-classes (within-school) can add explanatory power to our empirical analysis.

We denote by y_{ijk} the attainment at stage 6 in mathematics (CMS) of pupil i, $i = 1,\dots,n_{lj}^{(R)}; n^{(R)} = \sum_{l,j} n_{lj}^{(R)}$, in class l, $l = 1,\dots,L_j^{(R)}; L^{(R)} = \sum_{k} L_j^{(R)}$, in school j, $j = 1,\dots,J^{(R)}$. We then fit a three-level random effects model. The simplest such

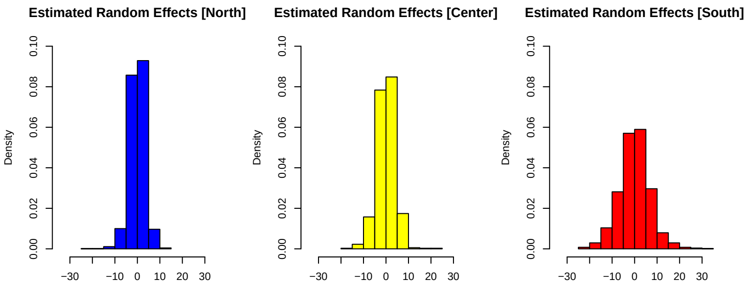

Fig. 1 Distribution of the Random Effects arising from fitting model (1) to the databases of data concerning students and schools of Northern area (left panel), Central area (central panel) and Southern area (right panel).

Table 2 ML estimates for Lasso regression model.

LASSO Model Coefficients	NORTH	CENTER	SOUTH
Intercept	-0.6996	-3.5284***	-2.2368$^{\cdot}$
mean school ESCS		0.9171$^{\cdot}$	1.9452***
% Female	0.0312*	0.0627**	0.0686**
% 1st generation immigrants	-0.0601**	0.0547$^{\cdot}$	0.1383**
% Early-enrolled		-0.1958*	-0.1585**
% Late-enrolled	-0.0713**		-0.2474***
Number of students	0.0027*	0.0050*	0.0118***
Istituto Comprensivo			0.0085***
Private	-0.7481**	-2.570**	

model allows the regression intercept to vary randomly across classes and schools. So for each geographic area $R = Northern, Central, Southern$, we have

$$y_{ilj}^{(R)} = \beta_0^{(R)} + \sum_{k=1}^{K} \beta_k^{(R)} x_{kilj} + b_j^{(R)} + u_{lj}^{(R)} + \varepsilon_{ilj}^{(R)} \tag{4}$$

$$b_j^{(R)} \sim \mathcal{N}(0, \sigma_{School}^{2^{(R)}}) \quad u_{lj} \sim \mathcal{N}(0, \sigma_{Class}^{2^{(R)}}) \quad \varepsilon_{ilj} \sim \mathcal{N}(0, \sigma_{\varepsilon}^{2^{(R)}}) \tag{5}$$

where x_{ilj} is the value of the $k-th$ predictor variable at student's level, $\beta^{(R)} = (\beta_0^{(R)}, \ldots, \beta_k^{(R)})$ is the $(K+1)$-dimensional vector of parameters referred to the R-th geographical area to be estimated and $\varepsilon_{ilj}^{(R)}$ is the zero mean gaussian error. The random effects $u_{lj}^{(R)}$ for the l-th class within the j-th school and b_j for the j-th school are assumed to be independent of any predictor variables that are included in the model. The results (see Table 3) show some interesting elements. First, part of the variance that was explained at school level, now is attributed to differences between classes, nevertheless variance between schools is still higher in the South than in the North. Of particular interest is the estimated variance between classes, which is substantial in all of the three areas, highlighting that not only the chosen school matters, but also the specific class attended by the students. Such an effect is even more marked in the South (where the variance between classes is much higher than between schools) suggesting the presence of sorting phenomena (or different educational quality) even within each school, that can explain some unobserved components of school effects. Table 3 illustrates another interesting feature of the geographical gap, as the "class-effect" is again higher in the South than in the North of the country, suggesting that in that area not only the chosen school matters, but also the class that the student attends has a higher and significative effect on the student's test scores.

Table 3 ML estimates for model (4).

Fixed effects	NORTH	CENTER	SOUTH
Intercept	0.797***	7.305340***	16.524***
Female	-1.683***	-2.638***	-2.165***
1st generation Immigrant	-0.637**	-0.377	0.389
Late-enrolled stud.	-2.466***	-1.827***	-3.791***
ESCS	1.879***	2.268***	2.676***
noMF	-1.182***	-1.256***	-1.276***
MS5	0.706***	0.581***	0.391***
Random effects	NORTH	CENTER	SOUTH
σ_{School}	3.13	3.58	5.77
σ_{Class}	3.68	5.19	8.17
σ_{ε}	12.00	12.75	12.86
Size	NORTH	CENTER	SOUTH
Number of Observations	130,256	46,529	82,972
Number of Groups (schools)	1,843	712	1,564
Number of Groups (classes)	8,615	3,485	6,661

4 Conclusions

The paper empirically shows that the differences in the determinants of student achievement in the three macro-areas of the country are so profound that it is impossible to specify a single empirical model for investigating them; as a consequence, this study promotes the idea of using three different models, one for each area. Another major message from this paper is that the ”school effect” is actually very heterogeneous and very dependent upon specific students and schools’ characteristics. We believe that describing the school effects diversity is useful for policy purposes, as it reduce the emphasis on the ”average” effects, and instead stimulates policy makers and school administrators to look at specific circumstances that can facilitate or impede the influence of schools on students’ experiences and results.

Acknowledgements We are grateful to Invalsi for having provided the original dataset.

References

1. Agasisti, T., Vittadini, G.: Regional economic disparities as determinants of students’ achievement in Italy. Research in Applied Economics, **4** (1), 33–53 (2012)
2. Agasisti, T., Ieva, F., Paganoni, A.M.: Heterogeneity, school-effects and achievement gaps across Italian regions: further evidence from statistical modeling. Submitted (2014) [online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/07-2014.pdf
3. Steele, F., Vignoles, A., Jenkins, A.: The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach. JRSS - A, **170** (3), 801–824 (2007)