Estimating student learning value-added models from repeated cross-sections

Dalit Contini and Elisa Grand

Abstract. In this paper we address the problem of estimating school achievement value-added models from repeated cross-sections. We use data of the INVALSI standardized assessment on primary and lower secondary school 2010.

Introduction

The recent development of standardized assessments of student learning has provided the basis for a novel research strand on educational inequalities, moving the focus from educational attainment to learning achievement. Given the cross-sectional nature of these surveys, performance differentials across socio-demographic groups are investigated at specific stages of the schooling career. Yet, since learning processes are cumulative, greater knowledge of how these differentials build over the schooling career would help designing effective educational policies to contrast inequalities.

Quite surprisingly, there are only few contributions, and not very convincing, tackling this issue. We aim at filling this gap: in this paper we address the problem of estimating school achievement value-added models from repeated cross-sections. Consistently with a simple learning accumulation model, we propose a strategy that allows to "link" two surveys, and we apply it to INVALSI data on 5th and 6th grades. By explicitly addressing the issue of measurement error (due to the substitution of true lagged values with proper estimates), we obtain consistent estimates of the parameters of the model of interest even with another source of measurement error (test scores imperfectly measuring achievement), also affecting genuine panel data.

[1] Dalit Contini, University of Torino; dalit.contini@unito.it
Elisa Grand, Collegio Carlo Alberto; elisa.grand@carloalberto.org

2 Model

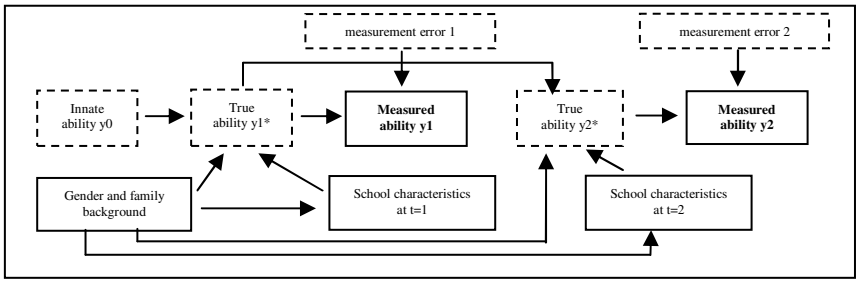
Consider two cross-sectional surveys assessing learning at times t=1 and t=2. A stylized but fairly comprehensive model of learning process and observed performance scores, consistent with the idea of a cumulative process where abilities build up over time, is depicted in Figure 1.[1] Our aim is to estimate the following models:

$$y_{1i} = \mu_1 + \beta_1' \boldsymbol{x}_i + \varepsilon_{1i} \quad (1)$$

$$y_{2i} = \mu_2 + \gamma y_{1i} + \beta_2' \boldsymbol{x}_i + \varepsilon_{2i} \quad (2)$$

where y_1 and y_2 represent performance scores, and $\boldsymbol{x}$ is a vector of socio-demographic individual variables. Our interest rests on *total* effects of socio-demographic variables given by direct and indirect effects through school features, so we do not include school features in the models. Error terms include a random component with the usual properties and measurement error; ε_1 also captures innate ability. β_1 and β_2 assess the degree of inequality. β_2 measures how inequality develops over time. Explanatory variables, which are time invariant characteristics, could be the same for models (1)-(2).

Figure 1: A stylized dynamic model of performance scores



3 Estimation strategy

Model (1) is estimated using the cross-sectional survey at t=1 (CS1). The issue is how to estimate (2) in the absence of genuine longitudinal data. Consider individuals in CS2: even if their own lagged scores y_1 are not observed, by exploiting CS1 we can obtain y_1 for different but "similar" children. A simple strategy would be to randomly select for each child in CS2 a child sharing the same characteristics in CS1, and use his score in place of true y_1. This strategy leads to biased results because y_1 is affected by (large) measurement error; conventional methods to correct for measurement error (Fuller, Hidiroglou; 1978) are not appropriate, because the error depends on both true and observed values. We could estimate instead a model for cell means, defined as groups of similar individuals. The advantage is that measurement error is substantially reduced, but correlation with lagged performance is not fully eliminated. Moffitt (1993) and Verbeek, Vella (2005) discuss the estimation of linear dynamic panel data models

[1] The assumption that ability at t=1 does not affect school features at t=2 is untenable if t=2 is after tracking. This version of the model does not apply to upper secondary school tests (e.g. PISA).

obtained by substituting unobserved genuine lagged values with OLS estimates derived from previous cross-sections. The resulting measurement error will be uncorrelated with lagged performance; however, to produce consistent estimates it must be uncorrelated also with the other explanatory variables. Whether this condition is met depends crucially on the nature of the dynamic model and of the model employed to predict lagged values.

Our case is relatively simple. However, if the set of independent variables is identical for y_1 and y_2, model (2) is not identified when substituting y_1 with $\hat{y}_1$. We must find a variable w, acting like an instrument, that affects y_1, but is unrelated to y_2 given y_1. The first problem is that w must be observed also at time t=2, so natural candidates such as variables describing school features at t=1 cannot be used. As the literature reports that elder children perform better than their younger peers, we use the month of birth as an instrument. We assume that the "true" model for y_1 is:

$$y_{1i} = \mu_1 + \beta_1' \boldsymbol{x}_i + \delta w_i + \varepsilon_{1i} \quad (3)$$

By substituting y_1 with its estimate obtained by (3) for the same $\boldsymbol{x}$ and w, we then estimate model (2), which, expressed in terms of $\hat{y}_1$, becomes:

$$y_{2i} = \mu + \gamma \hat{y}_{1i} + \beta_2' \boldsymbol{x}_i + (\gamma \hat{\varepsilon}_{1i} + \varepsilon_{2i}) \quad (4)$$

Measurement error $\hat{\varepsilon}_1$ is not of the *classical error in variables* type. Thanks to the properties of OLS, this error is uncorrelated with $\hat{y}_1$; given the simple structure of the model, it is also uncorrelated with $\boldsymbol{x}$. As $\boldsymbol{x}$ variables are all time-invariant, ε_2 is independent of all explanatory variables including the lagged score. Thus, under the above assumptions OLS estimates of (4) are consistent. Yet, the corresponding standard errors will be larger than with genuine panel data.

4 Data and empirical analysis

We use data of the standardized learning assessment administered in 2010 by INVALSI to the entire student population of 5th and 6th graders. Tests cover the domains of Italian and math, and follow the experience of international assessments. A questionnaire recording personal information, including family composition and home possessions is submitted to students, while school-boards provide information on parental background. School teachers are normally in charge of test administration. To control for cheating, a random sample of classes (~ 30,000 students) have taken the tests under the supervision of external personnel, representing a benchmark to evaluate performance scores at the population level. In the present work we use this sample data.

Performance is measured by the share of correct answers, varying between 0 and 1. We use two measures of socio-economic status: the number of books at home and a composite index (ESCS), derived from data on home possessions, parental education and occupation. We also include gender and dummies for geographical area to investigate territorial differentials. Since immigrant background children are often held back to earlier grades (so the month of birth could be meaningless), in this paper we focus on Italian students, leaving migrant/native gaps for further analyses. We exclude Italian students that repeated a school year in 5th grade. Being endogenous, variable repetition is not included in model (4), as it would capture part of the effects of interest.

Results are summarized in Table 1. We report estimates of model (1) in the first

column, of a cross-sectional model for y_2 in the second, of model (4) in the third. In a cross-sectional perspective, if we exclude geographical area, the effects of socio-demographic factors do not change much over time. When it comes to the value-added model, given 5th grade performance, socio-economic status affects 6th grade scores in Italian only mildly but does not affect math scores. Similar results hold for gender. Lagged score coefficients are quite large, especially for math. Overall, these results support the thesis that learning is particularly important during the early stages of the schooling career. Yet, the finding that, given social background and ability in primary school, children in the first year of lower secondary school (only one year later) living in the South still perform substantially worse than their peers in the North, suggests the existence of a large territorial divide in the quality of schooling at both primary and lower secondary school levels.

Table 1: Estimates of models for test scores with cross-sectional data and pseudo-panel data

	Italian			*Mathematics*		
Variable	5th grade	6th grade	6th grade	5th grade	6th grade	6th grade
Costant	0.658***	0.583***	0.114*	0.615***	0.494***	-0.111*
Birth mon.	-0.003***			-0.003***		
Books[1]	0.024***	0.021***	0.003	0.022***	0.023***	0.001
ESCS	0.033***	0.032***	0.008***	0.030***	0.033***	0.003
Female	0.008***	0.005***	-0.001	-0.031***	-0.029***	0.003
North East	-0.005	0.000	0.004	-0.013***	0.008**	0.021***
Centre	-0.026***	-0.017***	0.002	-0.022***	-0.027***	-0.004
South	-0.036***	-0.041**	-0.014***	-0.009***	-0.044***	-0.034***
Islands	-0.063***	-0.079**	-0.031***	- 0.040***	-0.084***	-0.042***
Lagged y_1			0.738***			1.018***
R squared	0.132	0.178	0.181	0.091	0.137	0.141
N° obs.	26616	31612	31550	27333	31594	31562

1. Books: 0=0-10; 1=11-25; 2=26-100; 3=101-200; 4=>200. *p-value<0.05; **p-value<0.01; ***p-value<0.005

Finally, we may decompose performance gaps at t=2 into a component that can be ascribed to previous ability differentials and a component representing new effects developed between 5th and 6th grade. Consider a change from (Books=1, ESCS=-1.5) to (Books=5, ESCS=+1.5). For Italian scores, the gap in 6th grade, amounting to 18 percentage points, can be split as follows: 14.2 points are attributed to previous performance, 3.8 to a new effect developed during the first year of middle school. As another example, compare North-West and Islands: of the 7.9 percentage points gap in 6th grade, almost 4.7 points depend on previous performance and 3.2 to a new effect.

References

1. Fuller, W.A., Hidiroglou, M.A.: Regression estimation after correcting for attenuation, Journal of the American Statistical Association, 73, 99-104 (1978)
2. Moffitt, R.: Identification and estimation of dynamic models with a time series of repeated cross-sections, Journal of Econometrics, 59, 99-123 (1993)
3. Verbeek, M., Vella, F.: Estimating dynamic models from repeated cross-sections, Journal of Econometrics, 127, 83-102 (2005)