Università di Milano Statale

10 dicembre 2012

MULTILEVEL MODELLING FOR VALUE ADDED ANALYSIS IN EDUCATION

LEONARDO GRILLI

EMAIL: GRILLI@DS.UNIFI.IT
WEB: WWW.DS.UNIFI.IT/GRILLI

UNIVERSITÀ DI FIRENZE

DIPARTIMENTO DI STATISTICA "G. PARENTI"
(01.01.2013 → DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI)

Outline

SECTION 1

Value-added analysis in education: definitions and estimation via multilevel models

Effectiveness

- The effectiveness of an organization is the degree of achievement of its institutional targets
  - ABSOLUTE (absolute effectiveness or impact analysis): evaluation of interventions, e.g. a specific training course
  - RELATIVE (relative or comparative effectiveness): comparison among many institutions

Types of effectiveness in education

The educational process yields multiple outcomes → many measures of effectiveness

- Internal effectiveness:
  - Dropout (1=Yes, 0=No)
  - Duration of studies (time to the degree)
- External effectiveness:
  - Occupational status after degree (1=Yes, 0=No)
  - Duration of unemployment (time to first job)
  - Wage or job satisfaction

The stakeholders (government, management, students) give different weights to the outcomes according to their preferences → the evaluation system should avoid summarizing the various kinds of effectiveness into a single overall indicator

Defining effectiveness in education

- For ***educational institutions (schools, universities)*** the effectiveness cannot be defined in absolute terms, but only <u>with respect to the effects on the students</u>
- In economic terms, the customers (students) are also *inputs of the production function* of the educational institution
- The effects on the students are affected by the features of the students themselves: how to make a fair assessment?

Hanushek E (1986) The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature* 24:1141–1177
Special issue of the Journal of Econometrics (2004): The econometrics of higher education

Value added

- The analysis of the educational process is difficult → the quality of educational institutions is usually measured via an **input/output** approach:
  - the process is a *black-box*
  - the output (*outcome*) is evaluated in the light of the input → effectiveness = **value added** by the school

VALUE-ADDED = ACTUAL OUTCOME
minus
EXPECTED OUTCOME GIVEN THE INPUT

Braun H and Wainer H (2007) Value-Added Modeling. In: Rao, C.R., Sinharay, S. (eds.) Handbook of Statistics 26, Psychometrics, pp. 475–501. Elsevier.
Special issue of the J. of Educational and Behavioral Statistics (2004)
Special issue of Education, Finance and Policy (2009)

Need for value added analysis

- Empirical research has found that the differences in student outcomes across schools are due
  - mainly to differences in student prior achievement and socio-economic background
  - for a minor part to differences in school factors such as teachers ability, organization…
- Thus comparing the unadjusted outcomes is markedly unfair and a value added approach is needed

Type A and B effectiveness

- **Type A**: performance of the institution adjusted for the features of the students

  → **to inform school choice**

- **Type B**: performance of the institution adjusted for the features of the students ***and*** for the context (e.g. socio-economic composition of enrolled students, resources, local labour market)

  → **for accountability**

Raudenbush SW & Willms JD (1995) The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307-335.

Statistical issues

- The statistical models for assessing the relative effectiveness of educational institutions must face two main issues:
  - **Adjustment**: the measures must be adjusted for the features of the students and, possibly, for the context (necessary for a fair comparison)
  - **Quantification of uncertainty**: the measures must be accompanied by error bars (necessary to make assessments properly supported by empirical evidence)

The *raw rankings* (often called 'League Tables') ignore both issues:
Goldstein H & Spiegelhalter DJ (1996) League tables and their limitations: statistical issues in comparisons of institutional performances. *JRSS A*, 159, 385-443

Statistical issues (cont.)

Adjustment & Quantification of uncertainty

⇓

Regression models

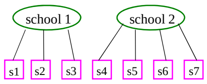
But *standard* models are not suitable!

- INACCURATE MODELLING: Standard models are unable to represent some key features, e.g. non-uniform effects (varying slopes)
- INACCURATE INFERENCE: Standard models make unrealistic assumptions on the variance-covariance structure (independence among observations, while the results of the students of the same school tend to be positively correlated) → poor quantification of uncertainty (usually *confidence intervals are too short*, and *tests have type I error rates higher than the nominal level*)

SECTION 2

Review of multilevel modelling

Multilevel structure, multilevel model



level 2 (cluster, group): school (index *j*)

level 1 (unit, individual): student (index *i*)

$$y_{ij} = \underbrace{\alpha + \beta x_{ij} + \gamma w_j}_{\text{Fixed part}} + \underbrace{u_j + e_{ij}}_{\text{Random part}}$$

This is an example of multilevel model, called 'random intercept model' since each school has its own intercept $\alpha + u_j$ that randomly varies among schools

The two-level linear model

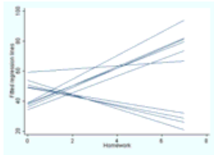
(one covariate at level 1)

Sample of J schools (from a population of schools)

Level 1 model

Equation for the *j*-th school:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \qquad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$$



each school has its own slope and intercept

The two-level linear model (cont.)

(one covariate at level 1)

Level 2 model

Each school has a couple of "parameters" (intercept & slope)

Assumption: the "parameters" (β_{0j}, β_{1j}) are iid random variables with a bivariate Normal distribution in the population of schools

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ & \sigma_{u1}^2 \end{bmatrix} \right)$$

(β_{0j}, β_{1j}) independent from e_{ij}

The Normal distribution is the "default" since it has nice properties and works well in many cases. Other choices are possible, such as a different continuous parametric family or an arbitrary discrete distribution

The two-level linear model (cont.)

(one covariate at level 1)

Level 1 model: $y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$ $\qquad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$

Level 2 model: $\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{cases}$ $\qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ & \sigma_{u1}^2 \end{bmatrix} \right)$

Combined model: $y_{ij} = \underbrace{\gamma_{00} + \gamma_{10} x_{ij}}_{\text{Fixed part}} + \underbrace{u_{1j} x_{ij} + u_{0j} + e_{ij}}_{\text{Random part}}$

Fixed part $(parameters: \gamma_{00}, \gamma_{10})$ Random part $(parameters: \sigma_e^2, \sigma_{u0}^2, \sigma_{u1}^2, \sigma_{u01})$

The two-level linear model (cont.)

(one covariate at level 1)

General case – Random (intercept and) slope model:

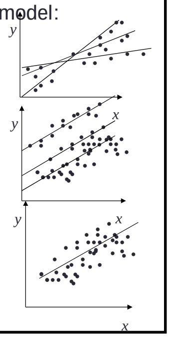
$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{1j} x_{ij} + u_{0j} + e_{ij}$$

Special case 1 – Random intercept model:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + e_{ij} \qquad \cancel{u_{1j}} \; (\sigma_{u1}^2 = 0)$$

Special case 2 – Ordinary regression model:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + e_{ij} \qquad \begin{cases} \cancel{u_{1j}} \; (\sigma_{u1}^2 = 0) \\ \cancel{u_{0j}} \; (\sigma_{u0}^2 = 0) \end{cases}$$



The two-level linear model

(one covariate at level 1 + one covariate at level 2)

- In the two-level model with level 1 covariates, the differences among clusters are accounted for (due to the random effects), but they are not explained!
- Often the key research question is: *why clusters differ in their mean level and slope*? The response can be obtained by including level 2 covariates
- Level 2 covariates represent features of the clusters useful to
  - define a model for the level 1 parameters (β_{0j}, β_{1j})
  - and so reduce the level 2 variances $(\sigma_{u0}^2, \sigma_{u1}^2)$
- Example: W is a binary variable coded
  1=public school; 0=private school

The two-level linear model (cont.)

(one covariate at level 1 + one covariate at level 2)

Level 1 model: $y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$

Level 2 model: $\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j} \end{cases}$

Here it becomes clear why the γ have a double index

Combined model:

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij}$$ Fixed part

$$+ u_{0j} + u_{1j} x_{ij} + e_{ij}$$ Random part

Random intercept model for school effects

Features of the student

Features of the school/context

Outcome of the student

School random effect

Student residual

i = student

j = school

$$y_{ij} = \alpha + \boldsymbol{\gamma}_x \mathbf{x}_{ij} + \boldsymbol{\gamma}_w \mathbf{w}_j + u_j + e_{ij}$$

$$y_{ij} - \left(\alpha + \boldsymbol{\gamma}_x \mathbf{x}_{ij} + \boldsymbol{\gamma}_w \mathbf{w}_j\right) = u_j + e_{ij}$$

Actual outcome

Expected outcome given student and school/context features

The random effect u_j is the *school value added*, or effectiveness. It is a *residual* term → its meaning depends on which covariates are in the model

Random intercept model: Type A and B effects

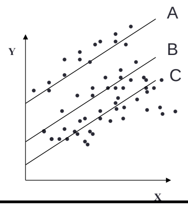
Type A effect of school *j*

$$y_{ij} = \alpha + \boldsymbol{\gamma}_x \mathbf{x}_{ij} + \boldsymbol{\gamma}_w \mathbf{w}_j + u_j + e_{ij}$$

Both effects are uniform (same effect for all the students)

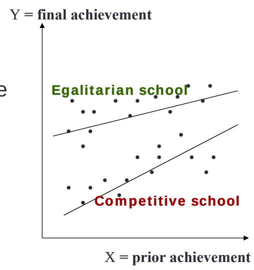
Type B effect of school *j*

- Constant slopes → parallel regression lines
- Unique ranking of the schools
  - ranking on Type A effects to inform potential students
  - ranking on Type B effects for accountability



From uniform to varying school effects

- Uniform effects are often a restrictive assumption
- Typically a given school practice has more or less impact on student learning depending on the kind of student under consideration:
  - **Egalitarian schools** try to reduce the gap in the prior achievement
  - **Competitive schools** tend to boost the initial differences
- In statistical terms: competitive schools have an higher slope on prior achievement

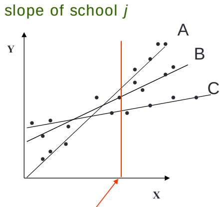


Random slope model for school effects

$$y_{ij} = \left(\alpha + u_{0j}\right) + \left(\gamma_x + u_{1j}\right) x_{ij} + e_{ij}$$

Random intercept of school j Random slope of school j

- Random slopes → crossing regression lines
- Varying effects → different school effects, depending on student characteristics
- No unique ranking of the schools → different rankings conditionally on student characteristics



1. Define student profiles
2. Build rankings by profile

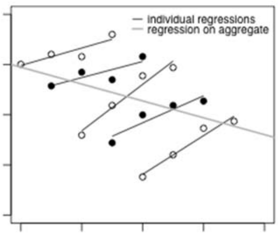
Within and between effects – ecological fallacy

1) $y_{ij} = \ldots + \gamma_{total} x_{ij} + \ldots$

2a) $y_{ij} = \ldots + \gamma_{within} (x_{ij} - \bar{x}_{.j}) + \gamma_{between} \bar{x}_{.j} + \ldots$

2b) $y_{ij} = \ldots + \gamma_{within} x_{ij} + (\gamma_{between} - \gamma_{within}) \bar{x}_{.j} + \ldots$

Statistically equivalent (2a and 2b)



If within and between slopes are different → model 1 is wrong (the total slope is a weighted average of within and between slopes)

Example: regress Y='employability' on X='graduation mark' for a sample of graduates clustered into schools

The contextual effect

- **Compositional variable** = cluster-level variable obtained by summarizing the within-cluster distribution of an individual-level variable
- The most important compositional variable is the cluster mean
  - E.g. if X = prior score, with pupils nested into schools → the school mean of the prior score is a compositional variable measuring the quality of the educational environment (peer effects)
- In a model with both the individual variable X and its cluster mean, the slope of the cluster mean is the **contextual effect**

$$y_{ij} = \ldots + \gamma_{within} x_{ij} + \delta \bar{x}_{.j} + \ldots \qquad \delta = \gamma_{between} - \gamma_{within}$$

- It is the additional effect of the school mean of X on Y that is not accounted for by the individual level X (usually X is prior score or Socio-Economic Status)
- The estimate of the contextual effect of X will partially encompass the effects of all school level variables that are correlated with X such as peer influences, school climate, allocation of resources, organizational and structural features of schools

Omitting the contextual effect produces level 2 endogeneity

Assume the true model is

$$y_{ij} = \alpha + \gamma x_{ij} + \delta \bar{x}_{.j} + u_j + e_{ij}$$

where δ is not null (= there is a contextual effect = between and within slopes are different)

Suppose the model is specified without the cluster mean

$$y_{ij} = \alpha^* + \gamma^* x_{ij} + u_j^* + e_{ij}^*$$

$$u_j^* = \delta \bar{x}_{.j} + u_j \quad \Rightarrow \quad \mathrm{cov}\left(u_j^*, x_{ij}\right) \neq 0$$

There is level 2 endogeneity (the random effect is correlated with the covariate) → the estimate of the slope of x_{ij} is biased

Solutions to level 2 endogeneity

- Two solutions:
  A. Replace the random effects with fixed effects
  B. Keep the random effects but add the cluster mean as a regressor (Mundlak 1978; Hausman & Taylor 1981)
- The famous Hausman specification test (routinely used to check for level 2 endogeneity in panel models) is just a test for the equality of between and within slopes, i.e.

$$H_0: \ \delta = 0$$

- A common misconception: thinking that when the Hausman test rejects the null hypothesis one is forced to use solution A. Indeed: also solution B is feasible!

The fixed effects model

$$y_{ij} = \gamma x_{ij} + \alpha_j + e_{ij}$$

random effects u_j replaced by parameters α_j

Thus no distributional assumptions !!!

- INTERPRETATION: all the variation between clusters (*including contextual effects*) is absorbed by the fixed effects α_j → the covariates can only explain the variation within clusters, thus the slope γ is not the *total* effect, but the *within* effect (in panel data the corresponding estimator is known as the *fixed effects estimator*)
- The *fixed effects model* is the standard choice in Econometrics (contrary to most fields, e.g. Epidemiology, Sociology, Demography, Psychometrics …)

Rivkin S.G., Hanushek E.A., Kain J.F. (2005) Teachers, schools, and academic achievement. Econometrica, 73, 417-458.

The fixed effects model: pros and cons

- PROS:
  - No distributional assumptions on the cluster effects → need not worry about homoscedasticity, normality, correlation between random effects and covariates
  - Feasible *even with very few clusters* (e.g. 5 clusters)
- CONS:
  - Impossible to use cluster-level covariates (due to perfect collinearity): a dramatic limitation when the research question concerns the effect of cluster-level covariates!
  - Loss of efficiency (since number of fixed effects = number of clusters)
  - Inefficient estimation of cluster effects (for example, if a cluster has two units its fixed effect is estimated with just two observations)

Use of the model

Once a suitable model is fitted the results can be used to

- **Analyse the associations** among the outcome and the explanatory variables
- **Predict the outcome** for a given student in a given school

$$\hat{Y}_{ij} = \hat{\alpha} + \hat{\gamma} \underbrace{x_{ij}}_{\text{student}} + \underbrace{\hat{\delta} \bar{x}_{j} + \hat{u}_j}_{\text{school}}$$

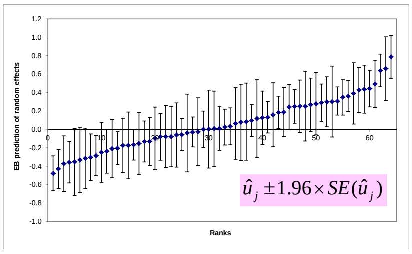
The university could build a system where the student plug-in her characteristics and obtain the predicted outcome for every school

- **Rank the schools** according to effectiveness (using school-level residuals)

Caterpillar plot of the residuals for comparing each cluster with the mean

The residuals are ordered and endowed with 95% confidence bars (+/- 1.96 the *comparative standard errors*)

The width of the error bar of a given cluster depends of its size



For the aim of pair-wise comparisons (e.g. Goldstein 2011), the intervals must be shorter:

$$\hat{u}_j \pm 1.39 \times SE(\hat{u}_j)$$

Outcomes and models

The nature of the outcome determines the kind of multilevel (mixed) model

Outcome	*Model*
Continuous (e.g. wage)	linear
Binary (e.g. dropout)	logit, probit
Count (e.g. credits)	Poisson
Time (e.g. time to degree)	duration

Generalized **L**inear **M**ixed **M**odels

Resources on multilevel modelling

- Snijders T.A.B. and Bosker R.J. (2011) Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. 2nd edition. Sage.
- Raudenbush S.W. and Bryk A.S. (2002) Hierarchical linear models : applications and data analysis methods. Sage.
- Hox J. (2010) Multilevel analysis : techniques and applications. Erlbaum. 2nd ed.
- Rabe-Hesketh S. and Skrondal A. (2008) Multilevel and Longitudinal Modeling Using Stata (2nd Edition). Stata Press.
- Gelman and Hill (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge Univ. Press.

- ENCYCLOPAEDIA ENTRY (5 pages) by Tom Snijders downloadable from http://stat.gamma.rug.nl/MultilevelAnalysis.pdf
- A POPULAR 3-page ARTICLE by Harvey Goldstein on Significance 2007, vol 4(3) [download from www.cmm.bristol.ac.uk/team/HG_Personal]
- WEB of Centre for Multilevel Modelling: www.cmm.bristol.ac.uk
- WEB of Statistical Computing at UCLA: www.ats.ucla.edu/stat

SECTION 3

Illustration: school 'league tables' in England

Education and examinations in England

- The English educational system
  - Age 6, beginning of primary school
  - Age 11, *end of primary school* (**KS 2**: Key Stage 2 examination)
  - Age 16, *end of secondary school* (**GCSE**: General Certificate of Secondary Education)
- GCSE is composed by at least 8 examinations with grades from A* (scored 58) to G (scored 16)
- The ***National Pupil Database*** contains – for each pupil – the results on all key examinations plus several individual characteristics `http://www.bris.ac.uk/cmpo/plug/npd/`

School performance indicators in England

- Published by the *Department for Children, Schools and Families* with the main purpose of informing school choice `www.dcsf.gov.uk/performancetables`
- History of *school performance indicators* in England
  - 1992: **raw** (only final score)
  - 2002: **value-added** (final score adjusted for prior score)
  - 2006: **contextual value-added** (final score adjusted for prior score and school-mean prior score)
- From 2006 uncertainty is considered (95% confidence intervals are published)

The analysis of Leckie and Goldstein (2009)

Leckie G, Goldstein H (2009) The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society – Series A*, 172, 835-851.

- Data on the 2007 cohort – GCSE score (age 16) adjusted for KS 2 score (age 11)
- They criticize the addition in 2006 of *compositional variables* (such as the school-mean KS 2 score) since the main purpose of the indicators is to inform school choice (Type A effectiveness)
- Indicators based on 2007 are used by parents in 2009 to enrol pupils who will get their GCSE in 2014, thus **7 years apart** !!
  - This kind of uncertainty is completely ignored in the confidence intervals
  - The issue is relevant: the literature on the stability of school effects shows that value-added measures are not strongly correlated over time (the correlation for a 5-year lag is about 0.6)

Random intercept value-added model for normalized GCSE score – 2007 cohort (sample from England: 48504 pupils in 266 schools) Leckie and Goldstein (2009)

VARIABLE	ESTIMATE (SE)
Fixed part	
Constant	-0.071 (0.014)
KS 2 average point score	0.681 (0.005)
KS 2 average point score (squared)	0.043 (0.003)
KS 2 average point score (cubed)	-0.026 (0.001)
Female	0.184 (0.006)
Age within cohort	-0.009 (0.001)
Free school meals	-0.182 (0.010)
Special educational needs	-0.373 (0.009)
English as an additional language	0.326 (0.019)
Ethnicity (ref: white British)	
White non-British	0.096 (0.023)
Black Caribbean	0.071 (0.028)
Black African	0.194 (0.031)
Indian	0.143 (0.027)
Pakistani	0.026 (0.028)
Chinese	0.383 (0.057)
Other ethnic group	0.067 (0.016)
Neighbourhood social depriv. IDACI	-0.119 (0.004)
Random part	
Between-school variance (level 2)	0.046 (0.004)
Within-school variance (level 1)	0.397 (0.003)

Prior score: normalized (= transformed to a standard Normal) KS 2 average score → cubic polynomial

Pupil-level characteristics

Percentage of between-school variance: 10.4%

Compositional variables: to adjust or not to adjust?

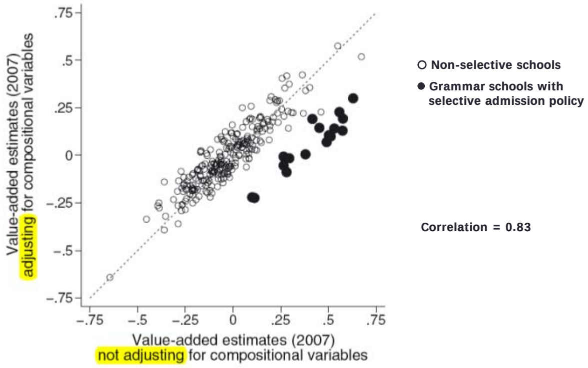
- Compositional variables, such as the school-mean prior score, aim at measuring the quality of enrolled pupils (peer effects)
- Should the value-added models adjust for them?
  - To inform school choice: NO
  - For accountability: in general YES … but questionable if the quality of the students is correlated with the quality of the school

	Normal school	Excellent school
Mean prior score	50	70
Mean final score	80	90
Mean progress	30	20

Adjusting for the quality of enrolled students unduly penalizes the best schools IF THEY ATTRACT BETTER STUDENTS BECAUSE OF THEIR HIGH REPUTATION

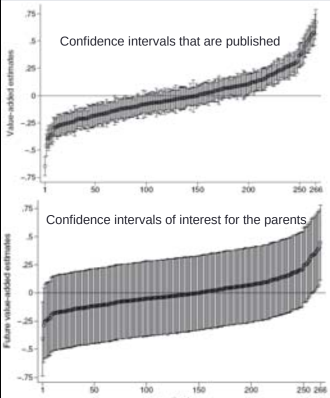
Leckie and Goldstein (2009) value-added model for normalized GCSE score

School effects adjusting vs not adjusting for school compositional variables



Leckie and Goldstein (2009) value-added model for normalized GCSE score

School effects with 95% confidence intervals



Uncertainty on CURRENT value-added (year 2007 based on cohort 2007)

168 schools (63.2%) are significantly different from the overall average

Uncertainty on FUTURE value-added (year 2014 based on cohort 2007)

Only 9 schools (3.4%) are significantly different from the overall average

SECTION 4

Final remarks

The value added approach: summary

- The value added (VA) approach is a powerful tool to analyse the factors related with student achievement and to identify outlying schools with extremely bad or good performances
- VA is a great improvement over the analysis of raw achievement scores; however, it has several limitations:
  1. VA analysis is not enough to understand why schools are more or less effective (field investigations are needed)
  2. Studies of school effects are quasi-experiments (students are not randomly assigned to schools!) → causal conclusions are questionable
  3. A satisfactory adjustment for the input factors requires several good-quality covariates
  4. Measurement error in the covariates (especially prior achievement) may yield biased estimates
  5. It is difficult to fully account for all the uncertainty
  6. It is difficult to communicate the results to a non specialized audience

Research activity in Italy on the evaluation of education (mainly about universities)

- Chiandotto B, Grilli L, Rampichini C (Eds) (2005) Valutazione dei processi formativi di terzo livello: contributi metodologici, Collana Valmon n. 12, Università di Firenze. `http://valmon.ds.unifi.it`
- Boero G. and Staffolani S. (Eds) (2006) Performance accademica e tassi di abbandono. Un'analisi dei primi effetti della riforma universitaria. CUEC, Cagliari
- Fabbris L (Ed) (2007) Effectiveness of University Education in Italy: Employability, Competences, Human Capital, Heidelberg: Springer-Verlag.
- Capursi V, Ghellini G (Eds) (2008) Dottor Divago. Discernere, valutare e governare la nuova università. Franco Angeli.
- Bini M, Monari P, Piccolo D, Salmaso L (Eds) (2009), Statistical methods for the evaluation of educational services and quality of products. Physica-Verlag.

Including: Grilli L. & Rampichini C. (2009) *Multilevel models for the evaluation of educational institutions: a review.* Download from `http://www.ds.unifi.it/grilli/papers.htm`