New insights into the Conditioning and Gain Score approaches in multilevel analysis

Alcune riflessioni su approccio condizionato e approccio alle differenze nell'analisi multilivello .

Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto and Carla Rampichini

Abstract We consider the issue of estimating the effect of a treatment variable on student achievement when a pre-test is available, taking into account the hierarchical structure of the data, with students nested into schools. The treatment variable can be either at student level or at school level. This effect can be estimated alternatively by adjusting for the pre-test score, i.e. conditioning, or by using the difference between post-test and pre-test scores, namely the gain score. The performance of the two approaches depends on pre-test reliability and validity of the common trend assumption. We derive approximated analytical results and we compare the two approaches via a simulation study.

Abstract *Questo lavoro affronta il problema della stima dell'effetto di uno specifico intervento (trattamento) sull'apprendimento degli studenti, nel caso in cui si disponga di una misura dell'abilità sia prima che dopo il trattamento, tenendo in considerazione la natura gerarchica dei dati, con gli studenti raggruppati in scuole. Il trattamento può essere sia a livello di singolo studente che a livello di scuola. Considerando il caso in cui l'assegnazione al trattamento non è casuale, esistono due approcci alternativi per la stima dell'effetto di interesse. Un primo approccio si basa sul condizionamento rispetto al test di abilità effettuato prima del trattamento, mentre il secondo approccio considera il guadagno, cioè la differenza tra i punteggi al test prima e dopo il trattamento. La validità dei due approcci dipende dalla affidabilità del test prima del trattamento come misura dell'abilità dello studente e dall'ipotesi di effetto comune dell'abilità nel determinare il punteggio ai due test. In questo lavoro, mostriamo alcuni risultati teorici e i risultati di uno studio di simulazione per il confronto tra questi due approcci.*

Key words: Achievement tests, Random effects model, Treatment effect.

Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence e-mail: bruno.arpino@unifi.it, silvia.bacci@unifi.it, leonardo.grilli@unifi.it, raffaele.guetto@unifi.it, carla.rampichinii@unifi.it

1 Introduction

In the education literature, student achievement is typically measured using multilevel models [1], with students nested into schools. In this contribution we aim to assess the effect of a specific treatment at student level (e.g. an individual support program) or at the school level (e.g. a given school policy). We consider a setting where student achievement is measured by means of a standardized test in two occasions, one before the treatment (pre-test), and the other one after the treatment (post-test). Two main methodological approaches have been proposed to estimate the treatment effect in this setting. The first approach consists in estimating the effect of the treatment on the post-test score, conditionally on the pre-test score (*conditioning* approach). The second approach considers the difference between the post-test score and the pre-test score as response variable, the so called *gain score* approach.

These two approaches give unbiased estimates under different assumptions, which are difficult to evaluate in observational studies. The debate on which of the two methods has to be preferred is still ongoing. Recently, Kim and Steiner [2] reconsidered the choice between the two approaches using graphs to illustrate the conditions under which a method has to be preferred. Despite the important contribution of this and several other studies that we review next, this literature has overlooked the fact that often test scores are collected in data sets with a multilevel structure, like in the educational setting.

We extend the results of Kim and Steiner [2] in two directions. First, we consider the most frequent case of a binary treatment variable rather than a continuous one. Second, we carry out the comparison between the two approaches when data have a multilevel structure, like in education setting, relying on a two level linear model. The multilevel setting has additional characteristics, playing a role in the comparison of the two approaches. In particular, the treatment can be either at student level or at school level. Moreover, the model includes random effects and it may also include cluster means.

2 Conditioning and gain score approaches

We consider individuals ($i = 1,\dots,n_j$) nested into clusters ($j = 1,\dots,J$), for example students nested into schools. Let Y_{1ij} and Y_{2ij} be continuous variables describing the observable scores on the pre-test and the post-test. These scores are error prone measures of a latent ability A_{ij}. The difference $G_{ij} = Y_{2ij} - Y_{1ij}$ is the gain score.

We are interested in assessing the effect of a treatment Z_{ij} on the post-test score Y_{2ij}, taking into account that the unobservable ability A_{ij} acts as a confounder affecting both Z_{ij} and Y_{2ij}. We assume that the two scores are generated by the following random intercept models:

$$Y_{1ij} = \mu_1 + \beta_1 A_{ij} + \psi_1 \bar{A}_j + u_{1j} + \lambda_1 e_{ij} \tag{1}$$

$$Y_{2ij} = \mu_2 + \beta_2 A_{ij} + \psi_2 \bar{A}_j + \tau Z_{ij} + u_{2j} + \lambda_2 e_{ij} + v_{ij}, \tag{2}$$

where $\bar{A}_j$ is the cluster-mean ability, μ_1 and μ_2 are intercepts, β_1 and β_2 are the within effects of the ability on the pre- and post-test scores, respectively, whereas ψ_1 and ψ_2 are the corresponding contextual effects. The treatment effect of interest is τ. The random variables u_{1j} and u_{2j} are the level 2 errors, while e_{ij} is a common level 1 error, and v_{ij} is a level 1 error specific to the post-test. All errors are assumed to be normally distributed with 0 mean and constant variances. Without loss of generality, we assume that both the individual ability and the error terms are normally distributed with mean 0 and unit variance, that is, $E(A_{ij}) = E(e_{ij}) = E(u_j) = 0$, and $Var(A_{ij}) = Var(e_{ij}) = 1$. Then, $Var(Y_{1ij}) = \beta_1^2 + \sigma_u^2 + \lambda_1^2$. When $\lambda_1 = 0$ the ability A_{ij} is measured without error by the pre-test Y_{1ij}. The gain score is defined as $G_{ij} = Y_{2ij} - Y_{1ij}$. The data generating path defined by equations (1) and (2) is represented in Figure 1. In order to estimate the treatment effect τ it is necessary to rely on some assumptions on the unobserved confounding due to the ability A_{ij}.

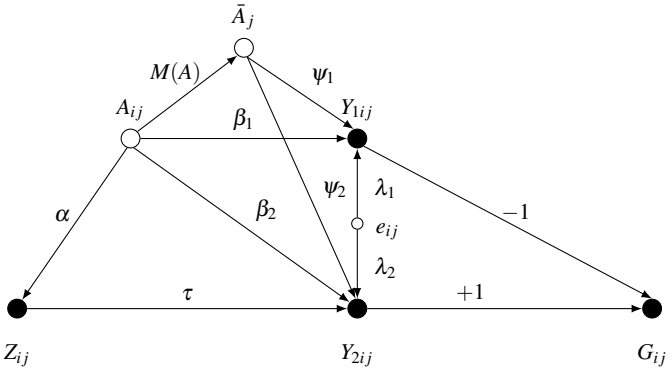

Fig. 1 *Path diagram for conditioning and gain score approaches; different effect of ability at within- and between-levels.*

In particular, according to the conditioning approach, we use the pre-test score Y_{1ij} as a proxy of the latent ability A_{ij}. Thus the conditioning model is specified as follows:

$$Y_{2ij} = \mu_2 + \beta_2 Y_{1ij} + \psi_2 \bar{Y}_{1j} + \tau Z_{ij} + u_{2j} + \varepsilon_{ij}. \tag{3}$$

The treatment effect τ is correctly estimated from model (3) if A_{1ij} is measured without error by Y_{1ij}, i.e. when λ_1 in equation (2) is equal to zero.

On the other hand, considering equations (1) and (2), we obtain $E(G_{ij}) = Y_{2ij} - Y_{1ij} = (\mu_2 - \mu_1) + (\beta_2 - \beta_1)A_{ij} + (\psi_2 - \psi_1)\bar{A}_j + \tau Z_{ij}$. Thus, under the common trend assumption, i.e. $\beta_2 = \beta_1$ and $\psi_2 = \psi_1$, the gain score is unaffected by the ability and the treatment effect τ can be estimated without bias from the following model:

$$G_{ij} = \tilde{\mu} + \tau Z_{ij} + \tilde{u}_j + \tilde{\varepsilon}_{ij}. \tag{4}$$

3 Main results

Considering a binary treatment, we derived the bias formula for the conditioning approach using model (3) without random effects, e.g. for the OLS estimator of τ, thus extending the results of Kim and Steiner [2]. We checked by a simulation experiment that this formula is a good approximation of the bias for the GLS estimator of τ in the random effects model (3).

The simulation study is based on 1000 data sets, each of them being composed of $10,000$ individuals uniformly distributed in 100 groups. Values of parameters used to generate data mimic the structure of Invalsi data [3]. In particular, the true value of τ (treatment effect) is equal to 2.

Table 1 displays the main results of the simulation study. Nine different configurations are taken into account, which distinguish for: (i) absence ($\lambda_1 = \lambda_2 = 0$) or presence of the measurement error, only on the pre-test ($\lambda_1 \neq 0$) and also on the post-test ($\lambda_2 \neq 0$), and (ii) validity of the common trend assumption, at level 1 ($\beta_2 = \beta_1$), at level 2 ($\psi_1 = \psi_2$), or at both levels. For each configuration, we show the mean of the estimated treatment effects and the corresponding relative error.

Table 1 Simulation study: Means of estimated treatment effect ($\hat{\tau}$) and corresponding relative error ($\%err$), individual-level treatment: conditioning and gain score approaches.

Conf.	Measurement error	Common trend level 1	Common trend level 2	Conditional $\hat{\tau}$	Conditional %err	Gain $\hat{\tau}$	Gain %err
1	no ($\lambda_1 = \lambda_2 = 0$)	yes	yes	2.0	0.0	2.0	0.0
2	no ($\lambda_1 = \lambda_2 = 0$)	no	yes	2.0	0.0	9.0	348.1
3	no ($\lambda_1 = \lambda_2 = 0$)	no	no	2.0	0.0	9.0	348.5
4	yes ($\lambda_1 = 6; \lambda_2 = 0$)	yes	yes	3.9	95.5	2.0	0.0
5	yes ($\lambda_1 = 6; \lambda_2 = 0$)	yes	no	3.9	95.5	2.0	0.0
6	yes ($\lambda_1 \neq \lambda_2$)	yes	yes	3.0	47.5	2.0	0.0
7	yes ($\lambda_1 = \lambda_2 = 6$)	yes	yes	2.0	0.0	2.0	0.0

Our simulations show that the results of Kim and Steiner generalize to a multilevel setting with some adjustments and further assumptions. Indeed, the treatment effect is correctly estimated using the conditioning approach when the pre-test is measured without error (configurations 1, 2, 3), whereas the gain score approach gives unbiased estimates when the common trend assumption is satisfied (configurations 4, 6, 7). Thus, under the common trend assumption, the gain score approach has to be preferred to the conditioning one in presence of measurement error on the pre-test, even if the reliability is quite high (say around 0.85).

As a further peculiarity, the conditioning approach provides satisfactorily results when the measurement error acts both on the pre-test and the post-test, but at the same extent (configuration 7). In addition, when the treatment is at level 1 (i.e., the target of the treatment are the single students), as in the simulation study here presented, its effect is correctly estimated using the gain score approach if the common trend assumption holds at level 1, even if it is violated at level 2 (configuration 5).

For the future development of this work, we intend to extend the simulation study to consider when the treatment acts at level 2 (i.e., the target of the treatment are the schools). We expect that in such a situation the common trend assumption at level 2 is crucial.

References

1. Goldstein, H.: Multilevel Statistical Models, 4th ed., Wiley (2010)
2. Kim, Y., Steiner, P.M.: Gain Scores Revisited: A Graphical Models Perspective. Sociological Methods & Research. (2019) doi: 10.1177/0049124119826155
3. Martini, A.: L'effetto scuola (valore aggiunto) nelle prove Invalsi 2018. Tech. Rep., Invalsi (2018)