

A Random Effects Model for Longitudinal Data with Measurement Errors

Un modello a effetti casuali per dati longitudinali con errori di misura

Michela Battauz, Ruggero Bellio, Enrico Gori

Dipartimento di Scienze Statistiche, Università degli Studi di Udine

e-mail: battauz@dss.uniud.it

Riassunto: In questo lavoro si considera un modello multilivello per la misura del contributo che le classi apportano al livello di apprendimento degli studenti (valore aggiunto). La specificazione considerata presenta il livello di apprendimento al tempo iniziale tra le variabili esplicative ed il livello di apprendimento nei tempi successivi come variabile di risposta. Inoltre, il livello iniziale è associato ad un coefficiente casuale. Il livello di apprendimento, tuttavia, è una variabile misurata con errore ed i metodi proposti in letteratura per la correzione dell'errore non sono direttamente estendibili al caso multilivello in cui la variabile affetta da errore entra nella parte casuale del modello. Questo lavoro propone un metodo per correggere per l'errore di misura nel caso in esame ed una applicazione a dati provenienti da alcune scuole medie della Lombardia.

Keywords: measurement error, multilevel models, value added models.

1. Introduction

Analyses of student achievement level focus on value-added indicators. The response variable is given by the score in a given test of interest, while the predictors may include measures of the student baseline level, together with social and economical individual characteristics. The data typically present a hierarchical structure, as students are grouped into classes and schools. Multilevel models are the natural choice for this kind of data, with random effects employed to describe class and school effects. Snijders and Bosker (1999) provide detailed explanations of this subject.

An important feature of multilevel-based analyses of achievement levels is the fact that both the response and the baseline level are measured with error. The latter problem is particularly serious, as it is well known that the use of covariables measured with error can lead to biased estimations of model parameters and to mistaken inferences. While several authors have addressed this point (see Carroll *et al.*, 1995), corrections for measurement errors are seldom adopted in educational data analyses. The reason is probably that methods adjusting for measurement errors are not straightforward to apply with correlated data. Moreover, the methods proposed in the literature can rarely be applied when measurement errors affect the random part of the model, which is often the case of value-added analyses.

The aim of this work is to propose a measurement error adjustment suitable for multilevel models, with an application to educational data from the Italian region Lombardia.

2. Models and Methods

We denote by ξ_{tij} the achievement level of student i in class j at time t , with $t = 0, \dots, T$. A typical model used in the literature for the estimation of the added value is the growth model

$$\xi_{tij} = \beta_0 + \beta_1 t + b_{0j} + b_{1j} t + z_{ij}^T \gamma + \varepsilon_{tij}, \quad (1)$$

where b_{0j} and b_{1j} are random effects that follow a bivariate normal distribution, ε_{ij} is an independent normally distributed error term, and z_{ij} is a vector of covariables measured without error. The growth model can be estimated when more than two observations per student are available and when data present a hierarchical structure.

Model (1) tries to explain the achievement level as a function of observed covariates and class effects. This seems a strong assumption, as the achievement level is likely to depend on the past history of the student. For this reason, we prefer to avoid this hypothesis and condition on the achievement level at the baseline period. More precisely, the proposed model takes the following form

$$\begin{aligned} \xi_{tij} &= \xi_{0ij} + \Delta_{ij} t, \\ \Delta_{ij} &= (\alpha + b_{j0}) + (\beta + b_{j1}) \xi_{0ij} + z_{ij}^T \gamma + \varepsilon_{ij}, \end{aligned} \quad (2)$$

that is

$$\xi_{tij} = \xi_{0ij} + [(\alpha + b_{j0}) + (\beta + b_{j1}) \xi_{0ij} + z_{ij}^T \gamma + \varepsilon_{ij}] t. \quad (3)$$

Rearranging the terms in (3) leads to

$$\frac{\xi_{tij}}{t} = (\alpha + b_{j0}) + \left(\beta + \frac{1}{t} + b_{j1} \right) \xi_{0ij} + z_{ij}^T \gamma + \varepsilon_{ij}. \quad (4)$$

Model (4) presents a variable measured with error, ξ_{0ij} , as a regressor and this introduces bias in parameter estimation. It is then necessary to correct for the measurement error by considering the following measurement equation

$$Y_{tij} = \xi_{tij} + u_{tij}, \quad u_{tij} \sim N(0, \sigma_t^2). \quad (5)$$

The measurement error variances σ_t^2 , will be considered as known in the following. This is not a restrictive assumption, as when the achievement levels are estimated by means of the Rasch model, such variances are typically available as a by-product of the procedure (Fisher and Molenaar, 1995).

When the response is a surrogate variable (see Carroll *et al.*, 1995, §13.2), accounting for measurement error simply leads to a decomposition of the level 1 residuals in two parts

$$\begin{aligned} \frac{Y_{tij}}{t} &= (\alpha + b_{j0}) + \left(\beta + \frac{1}{t} + b_{j1} \right) \xi_{0ij} + z_{ij}^T \gamma + \varepsilon_{ij} + \frac{u_{tij}}{t}, \\ &= (\alpha + b_{j0}) + \left(\beta + \frac{1}{t} + b_{j1} \right) \xi_{0ij} + z_{ij}^T \gamma + \varepsilon_{tij}^*. \end{aligned} \quad (6)$$

where $\varepsilon_{tij}^* = \varepsilon_{ij} + u_{tij}/t$, that has variance equal to $\sigma_\varepsilon^2 + \sigma_t^2/t^2$.

Consider now the covariate error. The method we propose to correct the bias follows a structural approach. In particular, we assume a normal distribution for the latent covariate

$$\xi_{0ij} \sim N(\mu_\xi, \sigma_\xi^2). \quad (7)$$

For parameter estimation we propose an EM algorithm (Robert and Casella, 1999, §5.3), treating the random effects and the latent covariate as missing data. The algorithm requires the specification of the distribution of the missing data given the observed data. This can not be obtained in closed form, as the response variable is not normally distributed. The product between the random slope and the latent covariate in (6), in fact, makes the response to be not normal. We propose then to use a Monte Carlo approximation of the average complete-data log-likelihood given the observed data. In particular, we used a Gibbs sampling algorithm (e.g. Robert and Casella, 1999) to obtain a random sample from the distribution of the missing data given the observed ones. The Gibbs sampling, in fact, is particularly appealing in the present case because both the distribution of the random effects given the latent covariate and the observed data, and the distribution of the latent covariate given the random effects and the observed data, are multivariate normal. The EM algorithm, however, do not automatically provide standard errors of the parameter estimates. To this end, here we used the Supplemented EM algorithm (Meng and Rubin, 1991).

3. Application

The longitudinal model defined by equations (4) and (5) has been applied to educational data coming from some middle schools of the Italian region Lombardia. The subject considered is Mathematics. For each student three observations are available. Table 1 summarizes the coefficient estimates obtained without accounting for the measurement error (*Naive*) and applying the method described above (*Structural*). As a comparison we present also the values obtained by applying the SIMEX method (see Carroll *et al.*, 1995), which corrects covariate measurement errors following a functional approach. We verified that all the variance components are significantly larger than zero with all the three methods. We notice that not accounting for the measurement error leads to an underestimation of the coefficient of the achievement level at time zero. The two error-adjusting methods produce very similar coefficients for this variable that results greater but still significantly smaller than zero. This means that on average high proficiency students grow less than low proficiency ones. Both the SIMEX and the structural method lead to a downwards correction of all the fixed coefficients. The standard deviation of the random intercept is also corrected downwards with both the methods, but the structural model leads to a larger difference. The correction of the standard deviation of the random slope is controversial, as the SIMEX method produces a upward correction, while the structural method reduces the coefficient. Both the procedures increase the correlation of the two random slopes and correct downwards the standard deviation of the residual.

4. Conclusion

The specification chosen to model the achievement growth over time plays a crucial role in educational studies. Here we have proposed an approach conditional on the baseline

level. It may be important to explore further specifications, including alternative forms accounting for nonlinearity of growth. In every case, measurement error adjustment remains a crucial issue.

Table 1: *Naive, SIMEX and Structural estimates (standard errors).*

	<i>Naive</i>	<i>SIMEX</i>	<i>Structural</i>
Intercept	0.300 (0.161)	0.781 (0.170)	0.626 (0.146)
Achievement level at time 0 (differential effect)	-0.242 (0.025)	-0.066 (0.033)	-0.058 (0.024)
Female	0.046 (0.032)	0.045 (0.032)	0.037 (0.030)
Mother education level: (relative to primary school)			
middle school	0.168 (0.010)	0.113 (0.102)	0.123 (0.092)
high school	0.232 (0.102)	0.164 (0.105)	0.176 (0.094)
university	0.166 (0.109)	0.097 (0.113)	0.124 (0.100)
Father education level: (relative to primary school)			
middle school	0.153 (0.091)	0.088 (0.093)	0.110 (0.086)
high school	0.239 (0.093)	0.141 (0.096)	0.183 (0.088)
university	0.303 (0.100)	0.188 (0.101)	0.228 (0.093)
Number of books at home: (relative to 0-10)			
11-25	0.074 (0.113)	0.019 (0.114)	0.009 (0.103)
26-100	0.204 (0.107)	0.085 (0.108)	0.122 (0.097)
101-200	0.219 (0.108)	0.084 (0.108)	0.162 (0.098)
> 200	0.379 (0.109)	0.215 (0.109)	0.265 (0.099)
Variance parameters			
σ_{b0}	0.430	0.425	0.363
σ_{b1}	0.160	0.198	0.110
ρ	0.562	0.787	0.897
σ_{ε}	0.713	0.516	0.443

References

- Carroll R.J., Ruppert D., Stefanski L.A. (1995) *Measurement Error in Nonlinear Models*, Chapman and Hall, London.
- Robert, C. P., Casella, G. (1999) *Monte Carlo Statistical Methods*, Springer, New York.
- Fisher, G. H., Molenaar, I. W., eds. (1995) *Rasch Models: Foundation, Recent Developments, and Application*, Springer, New York.
- Meng X. M., Rubin C. B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *Journal of the American Statistical Association*, 86, 899–909.
- Snijders T. A. B., Bosker R. J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, Sage Publications, London.