

A Monte-Carlo study to evaluate value-added models for institutions' rankings

Bruno Arpino and Roberta Varriale

Abstract The aim of this paper is to assess the quality of the ranking of institutions obtained with different types of widely used random and fixed effects value-added models. Through a Monte Carlo simulation study we assess the robustness of the ranking obtained in the presence of different model misspecifications and data structures. Coherently with the well-established literature, we find that it is quite hard to obtain a reliable ranking of the whole effectiveness distribution, while it is possible to identify institutions with extreme performances under various experimental conditions. Multilevel models where the between and within cluster components of first-level covariates are distinguished perform significantly better than both multilevel models where the two effects are set to be equal and fixed effect models. We also find that the estimated rankings are of poor quality when the effectiveness distribution does not follow a symmetric and unimodal distribution. For these situations we plan to explore simple data transformations, as the Box-Cox, and non-parametric methods for the estimation of random effects that can help to obtain a better ranking.

1 Introduction

In many fields of applied research and policy evaluation it is of interest to rank institutions with regard to their performance. This is a widely addressed issue in the educational literature, where it is often required to rank schools or universities with respect to their effectiveness in order to implement a system of actions and counteractions aimed at improving the general quality of the performed activities. Schools

Bruno Arpino

Department of Decision Sciences and “Carlo F. Dondena”, Center for Research on Social Dynamics, Bocconi University, Via Roentgen 1, 20136 Milano, Italy, e-mail: bruno.arpino@unibocconi.it

Roberta Varriale

Department of Statistics “G.Parenti”, University of Florence, viale Morgagni Viale Morgagni, 59 - 50134 Firenze, Italy e-mail: roberta.varriale@ds.unifi.it

effectiveness, as measured by one or more indicators, depends not only on the characteristics of the institution but also on those of their components (students). As a consequence, simple descriptive measures are not satisfactory since they do not net out the institution compositional effects. With reference to the general problem of ranking schools, the seminal work by [1], successively discussed by [4], describes the advantage of using multilevel regression models compared to one-level models. Subsequently, many other applied works in the same context of analysis used similar methodologies (see e.g., [10], [9], [2], [5], [15]). The quality of the ranking obtained through multilevel models depends on the validity of the assumptions underlying the multilevel regression model, that are similar to those used in ordinary multiple regression analysis, such as homoscedasticity and normal distribution of the residuals. While some Monte Carlo simulation studies have been carried out in order to evaluate the robustness of multilevel models with respect to the parameter estimates and their standard errors in case of violations of these assumptions (see, as an example, [6], [14]), we focus on the effect of different model misspecifications on the ranking quality. Furthermore, [11] developed an estimator to derive school ranking by using aggregated data and compared the performance of their estimator with the performance of OLS on aggregate data and multilevel models on disaggregate data. Our focus is, instead, on comparing multilevel and fixed effect models with disaggregate data.

2 Ranking in multilevel linear models

In the paper, the issue of ranking schools will be used as an illustrative example. Let Y_{ij} be a performance indicator, such as a test result, measured on student i attending school j , for $i = 1, \dots, n_j$ and $j = 1, \dots, J$. Our aim is to rank the schools with respect to the performance indicator in order to identify and reward the best-performing institutions. A general approach to obtain such a ranking is through regression analysis having Y_{ij} as dependent variable, while the characteristics of the institutions and those of their components are included as covariates. Within this general framework, fixed and random effects (multilevel models) approaches can be used. To illustrate how it is possible to derive a ranking of schools that adjusts for student characteristics using multilevel techniques, we will consider a simple example of a random intercept linear regression model with two level-1 covariates:

$$Y_{ij} = \alpha + \beta_1^{TOT} X_{1ij} + \beta_2^{TOT} X_{2ij} + e_{ij} + u_j \quad (1)$$

where X_{1ij} and X_{2ij} are level-1 covariates, β_1^{TOT} and β_2^{TOT} are the regression coefficients measuring the total effect of the covariates on the outcome variable, α is a constant, and e_{ij} and u_j are the level-1 and level-2 errors. In all the paper, with α and β we indicate fixed unknown parameters. The usual assumptions of multilevel linear regression models are made [12]: exogenous covariates, uncorrelated errors, normality and homoskedasticity of the level-1 and level-2 error distributions. In a mul-

tilevel context, the relationships at the cluster level, measured by the between-cluster effects, can be very different from the relationships at the micro level, measured by the within-cluster effects ([7], [13]). The regression model in equation (1) “mixes” the two types of relationships and its estimated total regression coefficients, β^{TOT} , are an average of the between-cluster and the within-cluster effects. From model 1, we can obtain purely between-schools (β^B) effects of the covariates by aggregating the response and exploratory variables at the school level:

$$Y_{.j} = \alpha + \beta_1^B X_{1.j} + \beta_2^B X_{2.j} + e_{.j} + u_j \quad (2)$$

where the dot is the common notation to denote that the variable has been averaged over the corresponding index. In this model, all the information on the within-schools variability is ignored.

If we are interested exclusively on the within-schools effects (β^W), we can subtract model 2 from 1, obtaining:

$$Y_{ij} - Y_{.j} = \alpha + \beta_1^W (X_{1ij} - X_{2.j}) + \beta_2^W (X_{2ij} - X_{2.j}) + (e_{ij} - e_{.j}). \quad (3)$$

The same within-school estimates can be obtained by replacing the random effect u_j in equation (1) with a fixed intercept α_j for each group j ([8]), obtaining:

$$Y_{ij} = \beta_1^W X_{1ij} + \beta_2^W X_{2ij} + e_{ij} + \alpha_j. \quad (4)$$

In order to simultaneously estimate both the between and within-cluster effects, we can combine model 2 and 3:

$$Y_{ij} = \alpha + \beta_1^W (X_{1ij} - X_{1.j}) + \beta_1^B X_{1.j} + \beta_2^W (X_{2ij} - X_{2.j}) + \beta_2^B X_{2.j} + e_{ij} + u_j. \quad (5)$$

When the within and between effects are equal for each covariates, $\beta_r^W = \beta_r^B$ ($r = 1, 2$), the models in equation (5) and (1) are equivalent. To derive a ranking of schools using a two-level model with students clustered into schools we can give a value-added interpretation to the models we introduced above. In random effects models in equation (1) and (5), the level-2 residuals, u_j , can be interpreted as a measure of the residual effect of the schools on the outcome variable measured at student level, after the effect of the independent variables included in the model has been controlled for. Through the ranking of the errors u_j we obtain our ultimate goal, which is the ranking of schools j . There are two methods commonly used to assign values to u_j : the maximum likelihood estimation and the Empirical Bayes prediction; the former treats u_j as an unknown fixed parameter, while the latter treats it as a random variable ([13]). Let define the total residuals for models in equation (1) and (4) as $\xi_{ij} = e_{ij} + u_j$ and the predicted errors, $\hat{\xi}_{ij}$, as $\hat{\xi}_{ij} = Y_{ij} - \hat{Y}_{ij}$. In a two-level random intercept model, the ML estimates of u_j can be obtained as the sample mean of $\hat{\xi}_{ij}$ for each cluster ([12]). The Empirical Bayes predictions u_j^{EB} are obtained as the mean value of the posterior distribution of u_j , where the prior is usually a normal distribution with zero mean and the estimated variance of u_j plugged in ([8]). In two-level random intercept linear models there is a simple rela-

tionship between the estimates obtained with these two approaches: $\hat{u}_j^{EB} = \hat{s} * \hat{u}_j^{ML}$, where $\hat{s} = \tau^2 / (\tau^2 + \sigma^2 / n_j)$ is the so-called shrinkage factor that ranges from 0 to 1 and causes the empirical Bayes prediction to be shrunk toward 0. When the data structure is balanced (constant number of first-level units in each cluster) the values assigned by the two methods to u_j are equal up to a constant and, as a consequence, the rankings of u_j (and of j) are the same. Since in the simulation studies we present in this paper we use balanced data structures, we will only focus on ML estimates of u_j and we will refer to these simply as $\hat{u}_j$. Since in real word examples unbalanced data structures are the norm and the shrinkage effect can be substantial on the ranking of u_j , we plan to extend the current work in that direction. In particular, we will contrast ML and EB results to assess the effect of shrinkage.

3 Simulation procedure and preliminary results

The setup of the simulation study builds on the setup used by [6], where the authors focus on the issue of bias and efficiency of fixed and random effects estimators, while we focus on the quality of the u_j ranking. Our study consists in 5 main steps:

1. we generate the data representing the performance of schools through a multi-level regression linear model with different types of distribution of the level-2 error term, u_j ;
2. we obtain the true ranking of u_j (and j) for each true distribution of u_j ;
3. we estimate the model parameters through three different estimation methods;
4. for each estimation method we obtain, as explained in section 2, the maximum likelihood estimates of u_j , and we rank them (and, consequently, we rank the clusters j);
5. for each estimation method, we compare the true and estimated rankings of u_j .

3.1 Steps 1 and 2

As the first step of our simulation study, we generate two-level balanced data structures, where the overall sample size N is determined by the product of the number of clusters, nc , and the fixed cluster size, cs . In the data generating model we use two level-1 covariates, X_1 and X_2 , which are allowed to vary both within and between clusters. Finally, the outcome is generated accordingly to equation (5).

We use five different processes to generate the “true” distribution of the error terms u_j :

- a. homoschedastic, Normal (N.Hom);
- b. heteroschedastic, Normal (N.Het);
- c. homoschedastic, asymmetric (Chi-Square with one degree of freedom, CHI);

- d. homoschedastic, symmetric and bimodal (50:50 mixture of two normal distributions, BIM);
- e. homoschedastic, symmetric and heavy-tailed (t-Student with two degree of freedom, STU).

The first specification (a.) conforms to the usual assumptions; the other four specifications imply different forms of model misspecification when the normality and homoskedasticity assumptions are used in the estimation procedure. In all cases, the values sampled from the assumed distribution of u_j are used to obtain the true ranking of j , indicated with RT_j . Besides the error term distribution, three other conditions have been varied: (i) number of clusters ($nc = 30, 50, 100$), (ii) cluster size ($cs = 5, 30, 50$) and (iii) the intraclass correlation coefficient value, $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$, ($\rho = 0.1, 0.2, 0.3$); as mentioned by [6], these values are commonly found in the applied educational literature. The residual variance at level 2 is determined as a consequence of the imposed values of ρ and of the residual variance at level 1, which is set to 0.5. The intercept term α in equation (5) is set to 0; the other regression coefficients are set to 1, except β_1^B that is set to 1.5: in this way, the between- and the within-effects of the first covariate are different. Finally, since we also want to evaluate the consequences of assuming the equality of the covariates' between- and within-effects, we run a set of simulations where the difference between the two effects for both covariates is varied and takes values -1, 0 and +1.

3.2 Steps 3 and 4

In order to estimate the model parameters and subsequently obtain the ranking of u_j , we use and compare three different estimation models (see section 2):

$$M1. Y_{ij} = \alpha + \beta_1^W (X_{1ij} - X_{1\cdot j}) + \beta_1^B X_{1\cdot j} + \beta_2^W (X_{2ij} - X_{2\cdot j}) + \beta_2^B X_{2\cdot j} + u_j + e_{ij} \quad (6)$$

$$M2. Y_{ij} = \alpha + \beta_1^{TOT} X_{1ij} + \beta_2^{TOT} X_{2ij} + u_j + e_{ij} \quad (7)$$

$$M3. Y_{ij} = \beta_1^W X_{1ij} + \beta_2^W X_{2ij} + \sum_{j=1}^J \alpha_j D_j + e_{ij} \quad (8)$$

where D_j in the third equation represent a dummy variable taking on value 1 if the unit belong to cluster j and 0 otherwise. M1 and M2 are two-level models corresponding to equations 5 and 1. M3 is a fixed effect model corresponding to equation 4. Each model is used to obtain the estimates of the cluster effects u_j and, consequently, the estimated ranking of j (RE_j); M1 and M2 by means of the maximum likelihood estimates of u_j and M3 by means of the estimated α_j coefficients. In the estimation of Models M1 and M2 we make the usual assumptions on the error terms (section 2) and, therefore, these models are misspecified if the true u_j distribution is not normal and/or not homoschedastic (cases from b.-e.). In this case, we aim at assessing the implications on the ranking quality of these misspecifications by comparing M1 and M2 with respect to model M3. Moreover, as we explained in section

2, the ranking obtained through the three models is expected to be different if β_r^W and β_r^B ($r = 1, 2$) coefficients are different.

3.3 Step 5

As the last step of our Monte Carlo experiment, we evaluate the quality of the estimated rankings obtained with different model conditions comparing the closeness of RE_j (estimated ranking) to RT_j (true ranking). For each experimental condition we generate 1000 simulated data sets and we evaluate two measures of the ranking quality. As a first measure, we compute the average Spearman correlation coefficient $\rho_{RT,RE}$ between the true and the estimated distribution of the level-2 residuals over the 1000 simulated data sets. For each replication, $\rho_{RT,RE}$ ranges from -1 to +1 and the closer it is to 1, the more similar the rankings are. The average Spearman correlation coefficient also ranges from -1 to +1 and shows how well a model performs in the classification of the level-2 units. The Spearman correlation coefficient is an overall measure and is affected by the difference between the estimated ranking positions of each unit with respect to the true positions. A property of the index is that the impact of a misclassification increases quadratically with the distance between the true and estimated position and only the absolute value of this distance matters. For example, consider the ranking of the school j out of 100 schools. Let consider the two situations whit $RT_j = 100$; $RE_j = 90$ and $RT_j = 50$; $RE_j = 40$: in both cases, the mistake in the classification of the school j has the same weight in the index, even if the two mistakes can be considered different from a substantive point of view. In the first case a school with an extreme performance is ranked 10 positions below and this could make a difference in the assignment of a price or a sanction; in the second case, it probably makes little difference for the school j to be ranked in the 50th or 40th position. In order to evaluate the quality of the estimated ranking separately for groups of institutions we use the classification matrixes. The idea is to divide the true and the estimated distributions of u_j into intervals defined by the deciles of the respective distributions and to evaluate, for each interval, the percentage of correctly classified schools over the 1000 simulated data sets.

4 Preliminary results and discussion

As a first results, our study shows that the ranking obtained throught multilevel models is reliable only for extreme institutions. Consistently with multilevel literature (see, i.e., [3]), we find that it is easier to reliably rank the institutions with extreme performances but it is hard to precisely rank the institutions with average performances. However, we do not think this is a reason to abandon the approach because extremely “bad” and “good” performing institutions are usually the most interesting for researchers and policy makers. Second, the effect of non-normal errors

at the second level can be detrimental also to rank extreme institutions. In particular, a highly asymmetric distribution (e.g., Chi-square) of second-level residuals implies a good ranking quality only of one tail of the distribution and a rather poor quality of the other tail. A bi-modal distribution, on the contrary, produces a low ranking quality for both tails. This highlights the importance of testing for normality of residuals' distribution. With respect to the data structure, large sample sizes help to increase the ranking quality, while the number and size of clusters, *per se*, play a less important role. We also find that a large ICC facilitates the ranking. Third, the assessment of ranking quality depends on the research/policy goals. While the performance of multilevel models in individuating the true best-performing (Top) institutions can be unsatisfactory in specific conditions, the probability of classifying a true Non Top institution as a Top institution is always very low for all the setups we used in our simulation study. On the other hand, the error probability associated to miss-classifying a true Top institution as Non Top is sometimes too high. Finally, discrepancies in the between and within effects of covariates is a crucial point for the quality of the ranking. In all experimental situations, the multilevel model with cluster means, that allows the between and within effects of the covariates to be different, performs much better than the others. Only when the between and within effects are equal for all the covariates, the three models perform very similarly. These results highlight the importance, also for clusters' ranking, to take into account that within-cluster and between-cluster relationships can be very different when dealing with multilevel data structures. These results are in line with similar remarks made in the multilevel literature, for example by [7] and [12].

Concluding, on the basis of our results two "best-practice" can be drawn. First, with the respect to the estimation strategy, it is preferable to start from a more general random effect model like M1, that allows to test for differences in the between and within covariates' effects. If there are no discrepancies, a simpler model M2 or M3 can be employed. Secondly, the suitability of model adjusted ranking should be judged by policy makers on the basis of the losses associated with the different potential errors, which in turn depend on the policy goals. For example, if ranking is implemented to identify and reward the best schools or sanction the worst schools, then policy makers might find misclassifying non-extreme schools as more costly. Our results suggest that the probability of these mistakes is low and the model adjusted ranking is, in this case, satisfactory and useful. On the contrary, if, for example, the accountability system aims at investing on the worst schools to improve their performances, then missing investments due to misclassify extreme schools as non-extreme are more costly. The probability of these misclassifications is higher and a model adjusted ranking approach could not be satisfactory.

As already mention, the current work will be extended to deal also with unbalanced data structures. Furthermore, we plan to extend the current simulation work in several other directions. First of all, we will explore data structures with larger number of clusters and bigger cluster sizes. This is driven by the necessity to consider data structures found in the empirical literatures on multilevel models in different reserach fields (medical, sociological, etc.). Another important development of the work consists in the proposal of simple data transformations (Box-Cox type) and

non-parametric methods for the estimation of random effects, as tools to improve the ranking quality in the presence of non-normal 2-level residuals.

References

1. Aitkin, M., Longford, N.: Statistical modelling in school effectiveness studies (with discussion). *Journal of Royal Statistical Society A*. **149**, 1–42 (1986)
2. Chiandotto, C., Varriale, R.: Un modello multilivello per l'analisi dei tempi di conseguimento del titolo nell'Ateneo fiorentino. In: Crocetta, C. (Ed.) *Modelli di analisi della transizione universit-lavoro*, 7. Cleup, Padova (2005)
3. Goldstein, H., Healy, M.J.R.: The graphical presentation of a collection of means. *Journal of Royal Statistical Society A*. **158**, 175–177 (1995)
4. Goldstein, H., Rasbash, J., Yang, M., Woodhouse G., Pan H., Nuttall D., Thomas S.: A multilevel analysis of school examination results. *Oxford Review of Education*. **19**, 425–433 (1993)
5. Jürges, H., Sneider, K.: Fair ranking of teachers. *Empirical Economics*. **32**, 411–431 (2007)
6. Maas, C.J.M., Hox, J.J.: Robustness issues in multilevel regression analysis. *Statistica Neerlandica*. **58(2)**, 127–137 (2004)
7. Neuhaus, J.M., Kalbfleisch, J.D.: Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*. **54**, 638–645 (1998)
8. Rabe-Hesketh, S., Skrondal, A.: *Multilevel and longitudinal modeling using Stata*. Stata Press, College Station, TX (2005)
9. Rampichini, C., Grilli, L., Petrucci, A.: Analysis of university course evaluations: from descriptive measures to multilevel models. *Statistical Methods and Applications*. **13(3)**, 357–373 (2004)
10. Raudenbush, S.W., Willms, J.D.: The estimation of school effects. *Journal of Educational and Behavioral Statistics*. **20**, 307–335 (1995)
11. Richter, F.G.C., Brorsen, B.W.: Aggregate versus disaggregate data in measuring school quality. *J Prod Anal*. **25**, 279–289 (2006)
12. Skrondal, A., Rabe-Hesketh, S.: *Generalized latent variable modeling*. Chapman and Hall, Boca Raton (2004)
13. Snijder, T., Bosker, R.: *An Introduction to Basic and Advanced Multilevel Modeling*. Sage, London (1999)
14. Solaro, N., Ferrari, P.A.: Robustness of parameter estimation procedures in multilevel models when random effects are MEP distributed. *Statistical methods & applications*. **16**, 51–67 (2007)
15. Wössmann, L.: Efficiency and equity of European education and training policies. *International Tax and Public Finance*. **15**, 199–230 (2008)