THE RESULTS OF A PERFORMANCE TEST: A MULTILEVEL ANALYSIS

O. Giambalvo, A. M. Milito, and A. M. Oliveri
Università degli Studi di Palermo, Italy
milito@unipa.it

The aim of the paper is to analyse the results of a performance test, created to evaluate how well a group of middle school pupils learned statistics, using multilevel analysis. The results show importance of the classroom/teacher and the school on the learning process.

INTRODUCTION

Evaluation is a problem which those involved in the field of didactics constantly face. Indeed, an analysis of the results is the only way to evaluate whether or not a particular teaching/learning process has been effective and, consequently, where to intervene. Various statistical approaches can be used in an evaluation and it is often quite useful to analyse the same data using different methodologies.

During the 2001 school year, a performance test was administered to 1,545 middle school pupils in four Italian cities. They were part of a project concerning the experimentation of new didactic strategies for statistical learning. The students were divided into two groups, each with different didactic conditions Data Oriented Approach, or DOA, and DOA with Cooperative Learning Method or CL. The performance test was made up of 37 sub-items regarding the statistical notions introduced in the classroom. Various analyses were carried out on these data using descriptive statistical methodology (Milito-Marsala, 2002).

The aim of this study is to analyse the results of the performance test by using a multilevel analysis, a methodology which quickly illustrates the relationship between variables while taking into account the hierarchical structure of the data.

THE MULTILEVEL MODEL

As already mentioned, among the various statistical approaches available, a multilevel model was chosen as it lends itself to hierarchical data and finds its natural application in the field of didactics. The early work on a multilevel approach (Aitkin *et al*. 1981) was precisely about educational data. [Two expository volumes appeared in the early 1990's. The one by Bryk and Raudenbush (1992) discusses 2- and 3-level linear multilevel models with applications, especially for educational data and repeated measure designs.] Now this method and its extensions are beginning to be widely applied, not only in education, but also in epidemiology, geography, child growth, to mention a few

The aim of a multilevel analysis is modelling links among the explicative and response variables, taking into account the hierarchical structure of the data, not randomly considered but naturally or suitably identified. The model also estimates the effects of the different levels of clustering on the response-variable vector. In a multilevel approach we refer to a hierarchy as consisting of *units* grouped at different *levels* (students may be the level-1 units, clustered within classrooms and schools, which in turn may be the second and third levels). The existence of such data hierarchies is neither accidental nor ignorable.

Multilevel models may be used, above all, with *i*) sample survey data (the population structure is viewed as being of potential interest in itself while the survey design can be used to collect and analyze data about the higher level units in the population); *ii*) repeated measure data (the same individuals or units are measured on more than one occasion); *iii*) event history models (considering "time to failure" as a variable); *iv*) discrete response data (proportion or percentage). Another application which is particularly important is where measurements are missing by design rather than at random.

The multilevel models can be classified, according to the included parameter typology, in fixed effects models (ANOVA type), variance component models and random slope models.

For the 2-level random slope model (Goldstein, 2003):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + (v_{0j} + v_{1j} x_{ij} + e_{0ij})$$

$$\text{var}(e_{0ij}) = \sigma^2_{e0}$$

where: β_{0j} e β_{1j} are random variables: $\beta_{0j} = \beta_0 + v_{0j}$, $\beta_{1j} = \beta_1 + v_{1j}$

where v_{0j}, v_{1j} are random variables with parameters:

$$E(v_{0j}) = E(v_{1j}) = 0$$

$$\text{var}(v_{0j}) = \sigma^2_{u0}, \quad \text{var}(v_{1j}) = \sigma^2_{u1}, \quad \text{cov}(v_{0j}, v_{1j}) = \sigma_{u01}$$

$$\text{var}(e_{0ij}) = \sigma^2_{e0}$$

We have expressed the response variable y_{ij} as the sum of a fixed part and a random part within the brackets. The random variables are referred to as 'residuals' and in the case of a single level model the level-1 residual e_{0ij} becomes the usual linear model residual term. To make the model symmetrical so that each coefficient has an associated explanatory variable, we can define a further explanatory variable for the intercept β_0 and its associated residual, u_{0j}, namely x_{0ij}, which takes the value 1.0. For simplicity this variable may often be omitted.

The difference between a multilevel model and a standard linear model of regression or analysis of variance type is the presence of more than one residual term, and this implies that special procedures are required to obtain satisfactory parameter estimates.

The interpretation of results is quite similar to that of a regression model.

RESULTS

The fitting of a statistical model starts from building a theoretical one and selecting variables to explain the variation measured on students with respect to the administered performance test.

The hierarchical nature of collected data is quite clear: we dealt with 1,541 students (level-1 units) clustered within 75 classrooms (or teachers, level-2 units) clustered within 49 schools (level 3), finally clustered within 4 towns (level 4). The fourth level, composed of very few units, was excluded from the final model.

Different teaching strategies (variable label: *experimentation*, categories: DOA = 0, CL = 1) and sex (categories: male = 0 and female = 1), measured at the students' level, were considered as predictors of results on the overall performance test (variable label: *ztest*). Although we assumed there was an effect caused by the explanatory variables, we had no pre-existing idea on the direction of such a relationship (better marks associated with DOA or CL? Better marks for boys or girls?). Different effects on results were hypothesized, also depending on classrooms and schools. Since different teaching strategies were in fact chosen by teachers, differences among classrooms and schools were to be expected.

A preliminary analysis addressed the hypothesis of no effect of *sex* and *experimentation* on the results of the performance test. The sample was reduced to 1,541 units depending on listwise deletion. The difference between the means were significant only for *sex* (*Z = 2.21, p-value = 0.03*), not for *experimentation* (*Z = 1.56, p-value = 0.12*). Besides this, the interaction between the two variables was studied but provided no significant results (*experimentation within sex 0*: *Z = 1.796, p-value = 0.07, experimentation within sex 1*: *Z = 0.4, p-value = 0.69*).

The variable *classrooms* had an influence on the response. The ANOVA test provided a significant *F* statistic ($F_{74;1466} = 7.86$, *p-value = 0.00*). 25.3% of the overall response variance was attributed to differences among classrooms.

The same results held true for the variable *school* ($F_{48;1492} = 9.56$, *p-value = 0.00*). Such results seemed to fit multilevel model incorporating parameters for (and the over-dispersion caused by) level-2- and level-3 units. Such a model can be expressed in the following terms (Model 1):

$$\begin{aligned} \text{ztest}_{ijk} &= \beta_{0jk} + e_{ijk} \\ \beta_{0jk} &= \beta_0 + v_{0k} + u_{0jk} \\ v_{0k} &\sim \text{N}(0, \sigma^2_{v0}) \\ u_{0jk} &\sim \text{N}(0, \sigma^2_{u0}) \\ e_{ijk} &\sim \text{N}(0, \sigma^2_e) \end{aligned} \qquad (1)$$

the *i* index (i=*1,2, ...,1541*) indicating students, *j=1,2, ...,75* indicating classrooms and *k=1,2,.....49*, schools.

Model 1 provided the following estimates (produced by means of the Iterative Generalized Least Squares procedure: Goldstein, 2003) for the overall mean of the variable *ztest*, the variance accounted for respectively by schools, classrooms and students (within brackets the standard errors):

β_0 = -0.017(0.073), σ^2_{v0} = 0.145(0.059), σ^2_{u0} = 0.113(0.041), σ^2_e = 0.752(0.028)

The *-2*loglikelihood* statistic was 4079.626.

The introduction of the explanatory variables *sex* and *experimentation* (Model 2) improved the fitting of the model, the regression equation of which is reported below:

$$\text{ztest}_{ijk} = \beta_{0jk} + \beta_1 \text{experim}_{ijk} + \beta_2 \text{sex}_{ijk} + e_{ijk} \qquad (2)$$

The likelihood ratio test (LRT) = 4079.626 – 4072.649 = 6.977. As we know, the statistic follows a χ^2 distribution, with $\nu = 2$ degrees of freedom, and was not significant (*p-value = 0.06*).

The regression coefficient β_1 was not significant (*z=0.01*, *p-value=0.5* for one-sided test). So it was possible to exclude the variable *experimentation* and to fit a simpler model (Model 3). Such a model, synthesized below in usual equation terms, provided the following estimates:

$$\text{ztest}_{ijk} = \beta_{0jk} + \beta_1 \text{sex}_{ijk} + e_{ijk} \qquad (3)$$

β_0 = -0.077(0.076), β_2 =0.119(0.045), σ^2_{v0} =0.147(0.059), σ^2_{u0} =0.112(0.040), σ^2_e = 0.748(0.028).

The *-2*loglikelihood* statistic was the same as Model 2: 4072.649. Now the LRT was compared to a $\chi^2_{\nu=2}$ and was finally significant (*p-value = 0.008*).

Up to this point we assumed variations among groups depending only on the intercepts. What if we admitted that the regression lines also had different slopes? The question was whether to fit a *random slope* model (Model 4) or not.

The fitting of Model 4, expressed as follows,

$$\begin{aligned} \text{ztest}_{ijk} &= \beta_{0jk} + \beta_{1jk} \text{sex}_{ijk} + e_{ijk} \\ \beta_{0jk} &= \beta_0 + v_{0k} + u_{0jk} \\ \beta_{1jk} &= \beta_1 + v_{1k} + u_{1jk} \\ \begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} &\sim \text{N}(0, \Omega_v) \; : \; \Omega_v = \begin{bmatrix} \sigma^2_{v0} & \\ \sigma_{v01} & \sigma^2_{v1} \end{bmatrix} \\ \begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} &\sim \text{N}(0, \Omega_u) \; : \; \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \\ e_{ijk} &\sim \text{N}(0, \sigma^2_e) \end{aligned} \qquad (4)$$

provided a LRT = 15.364, that was compared to a $\chi^2_{\nu=5}$ and was significant (*p-value = 0.009*). Yet Model 4 proved to be less parsimonious than Model 3: LRT = 4072.649 - 4064.262 = 8.387, $\nu = 4$ (*p-value* = 0.08). There was enough evidence to consider Model 3 as the best fitting one.

CONCLUSION

By using a multilevel model we were able to examine the relationship between the results obtained by the students on the performance test and some explanatory variables, including the estimate of the variance components attributed to various hierarchical levels (classrooms and schools). The results of those analyses, influenced by the absence of specific information regarding level-2 and level-3 variables, did not demonstrate relevant differences between didactic conditions, while the variable *sex* seemed to assume a certain relevance, showing better performance on average for girls. Moreover, although the model explains only 25% of the overall variance, it did demonstrate a significant effect according to *classroom* (level 2) and to *school* (level 3).

We can, therefore, hypothesize that it is the teacher (i.e. the classroom) that influence the learning of statistics, together with the schools which participated in the project. That may be due to the fact that the experimentation design allowed the teachers using DOA to assign group work, making the two didactic conditions quite similar, at least in this respect

The results of our analysis confirm what other studies (Milito-Marsala, 2002) have already found regarding the minor differences between the two didactic conditions and the relevance of the variable *teacher* (Giambalvo, 2001).

ACKNOWLEDGEMENTS

This paper is the result of the collective work of the authors. However, the specific parts of the paper can be attributed as follows: O. Giambalvo “The Multilevel Model”; A.M. Milito “Introduction” and “Conclusion”; A. M. Oliveri “Results.”

REFERENCES

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of Royal Statistic Society*, 144, 148-61.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Newbury Park: Sage.

Giambalvo, O. and Milito, A. M. (2003). The implicative analysis on a performance test. *Proceedings of the 54th Session of the International Statistical Institute*, (pp. 402-403), Berlin.

Giambalvo, O. (2001). L’effetto docente sulla valutazione dell’insegnamento della statistica nelle scuole media inferiori. *Atti del convegno intermedio della SIS*, Roma, (pp. 303-306), June 4-6.

Goldstein, H. (2003). *Multilevel Statistical Models*. London: Edward Arnold.

Milito, A. M. and Marsala, M. R. (2002). *Insegnare ed Apprendere – La statistica a scuola* (a cura di). Quaderni del Dipartimento di Metodi Quantitativi per le Scienze Umane N. 1, Palermo.