A misspecification test for finite-mixture logistic models for clustered binary and ordered responses[*]

Francesco Bartolucci
email: francesco.bartolucci@unipg.it

Silvia Bacci
email: silvia.bacci@unipg.it

Claudia Pigini
email: pigini@stat.unipg.it

Dipartimento di Economia, Università degli studi di Perugia

1 Abstract

Generalized Linear Mixed Models (GLMMs, Skrondal and Rabe-Hesketh, 2004; McCulloch et al., 2008) represent a very useful instrument for the analysis of clustered data, as they use random effects to account for the dependence between observations within the same cluster. A well known approach, that formulates in a flexible way the random effect distribution, is based on assuming a discrete distribution that leads to a finite-mixture model. The finite-mixture approach has some advantages over the normal approach. Mainly, it avoids integrating out the random effects, and a rather simple Expectation Maximization (EM) algorithm (Dempster et al., 1977) may be used instead. Moreover, the approach leads to a natural clustering of sample units that may be of main interest in certain relevant applications.

We propose a general test for misspecification of the discrete mixing distribution in logistic models with binary and ordered responses. We extend the approach developed by Tchetgen and Coull (2006) which is based on the comparison of CML and MML estimates for the fixed effects, as in the Hausman's test (Hausman, 1978); the difference between the two estimates is normalized on the basis of the estimated variance-covariance matrix of this difference. The test relies on the consistency of the CML estimator that is attained under mild distributional assumptions; essentially, the random effects must be constant within each cluster.

Our approach presents some novelties and peculiarities deriving from the finite-mixture nature of the models of interest. First of all, since none of the two estimators compared is ensured to be fully efficient, we use a generalized estimate of the variance-covariance matrix of the difference through a method adopted, in a related context, by Bartolucci et al. (2014). This also ensures stable results in small samples, while retaining the simplicity of the approach and its low computational complexity. Second, the proposed test may also be used to select the number of support points of the discrete distribution, which is alternative to commonly used selection

[*] Presented at the second internal meeting of the FIRB ("Futuro in ricerca" 2012) project "Mixture and latent variable models for causal-inference and analysis of socio-economic data", Rome (IT), January 23-24, 2015

criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC); this is a crucial issue in the use of the models of our interest. Third, an issue that is typically ignored is that one of the possible sources of misspecification is the dependence between the random effects and the observable covariates, that is, a problem of endogeneity. In the finite-mixture approach, a greater variety of methods to model this dependence is available with respect to the normal approach, and the proposed test has an important role in this regard.

The performance of the proposed test is evaluated through a Monte Carlo study that provides satisfactory results under different scenarios: we observe good size properties under the null hypothesis of correct specification of the number of support points of the random effect distribution. The simulation results suggest that: (*i*) when the number of classes is underspecified, rejection rates are particularly high especially when the random effect distribution is skewed and has a large variance; (*ii*) when the random effect distribution is continuous, the Hausman test tends to select a more parsimonious specification of the number of support points, with respect to standard selection criteria, especially with many units per cluster; (*iii*) in the presence of correlation of the random effects with the regression covariates, rejection rates are remarkably high even in very small samples and increase for higher correlation values.

The approach is also illustrated by three applications covering different settings, that is, multilevel data, longitudinal data, item responses. Interestingly, each application presents a different potentiality of the proposed approach. In the first application we obtain the same results of selection criteria such as BIC in terms of number of mixture components. In the second application, contrary to the BIC, the proposed test leads to the conclusion that a latent structure is not necessary, and then to a very parsimonious and easily interpretable model. In the third application, the proposed approach leads to reject all models in which the random effects are assumed to be independent of the covariates, considering therefore a form of endogeneity.

References

Bartolucci, F., Belotti, F., and Peracchi, F. (2014). Testing for time-invariant unobserved heterogeneity in generalized linear models for panel data. *Journal of Econometrics*, in press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, 46:1251–1271.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models.* Wiley.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling. Multilevel, Longitudinal and Structural Equation Models.* Chapman and Hall/CRC, London.

Tchetgen, E. J. and Coull, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika*, 93:1003–1010.