An Instrumental variable consistent estimation procedure to overcome the problem of endogenous variables in multilevel models

Neil H Spencer[1], Antony Fielding[2]
[1]University of Hertfordshire, [2]University of Birmingham
e-mail: n.h.spencer@herts.ac.uk

Introduction

It is not unusual for a multilevel model to contain fixed effect explanatory variables that can be regarded as endogenous. This will happen if the variables are subject to the same unmeasured influences as the dependent response. These influences will be incorporated in the random effect disturbance terms. Thus, these variables are no longer exogenous and are not independent of the random effects in the model. In such circumstances, a basic model assumption is not met and obtaining consistent estimators of the parameters is not straightforward. Many standard multilevel procedures (e.g. Iterative Generalised Least Squares: IGLS) rely on the independence of regressors and model disturbances for their consistency properties (as, indeed, do single level procedures). Here, we present a modelling strategy based on instrumental variables and introduce an *MLwiN* macro that provides consistent estimators.

For exemplification, we consider a simple two level model with endogenous variables. Fielding (1998) introduces a dataset drawn from children in primary schools of the City of Birmingham Local Education Authority. Data are available on a range of school and pupil characteristics. The responses are the results of Key Stage 1 (KS1) tests, and we wish to relate one of these test results to gender, age of the child in months, and the results of baseline tests carried out when the child entered the reception classes in the school. A model such as this may be used to examine the progress children are making between reception and KS1 in different schools. For pupil i in school j and where we have just one baseline test we may write:

$$KS1TEST_{ij} = \beta_0 + \beta_1 GENDER_{ij} + \beta_2 AGE_{ij} + \beta_3 BASELINE_{ij} + u_j + \varepsilon_{ij} \quad (1)$$

The term u_j in model (1) represents a random school effect and ε_{ij} is a within school random pupil effect. The endogeneity in this model arises because the baseline test may be supposed to be related to the random pupil effect through the existence of important unmeasured and unmeasureable influences acting at this lowest level of the hierarchy (e.g. home circumstances). These influences are incorporated in the disturbance ε_{ij} but may also influence baseline test performance. It is also possible that there are some influences that make the baseline test related to the school effect u_j. The common influences may be such things as the locality in which the school is situated and from which the pupils generally come.

Overcoming the problem of endogeneity

Solutions to the problem of inconsistency caused by endogenous regressors, particularly when they are thought to relate to higher level effects such as u_j, have been proposed by Kiviet (1995) and Rice et al. (1998). Kiviet uses a bias corrected version of the least squares dummy variable estimator (LSDV). Rice et al. use conditioned iterative generalised least squares (CIGLS). The first of these approaches suffers from a problem that the bias correction applied may actually increase the bias in some circumstances. Neither approach can easily cope with the case where the level 1 (pupil) random effect is correlated with regressors. It is this latter situation on which we mainly focus. The difficulties caused by endogenous regressors in the context of generalised linear models for count data are also discussed by Crouchley and Davies (1999).

A frequently used method of overcoming such endogeneity problems in single level models is to use instrumental variable techniques. We adapt these techniques to cover multilevel random effects models within the framework of the *MLwiN* IGLS estimation procedures. The possibility is mentioned briefly and independently by Rice et al. (1998). Spencer (1997) also suggests such an approach for repeat testing in educational situations where explanatory variables are lagged versions of the response. A supplementary multilevel model for the endogenous explanatory variable is constructed using fixed effect explanatory variables that are assumed exogenous and independent of the random part of model (1). We stress that the existence of such variables and the adequate collection of data on them are a necessary pre-requisite. Predictions of the endogenous variable values for each child are then obtained from the fixed parts of the supplementary models. These predicted values, being independent of the random part of the model of interest (1), are then used as instruments.

Armed with data on the original set of regressors of model (1) and the set of instruments (being the original regressor set with endogenous variables replaced by their instruments), we estimate the fixed effect parameters in model 1 (see, e.g., Bowden & Turkington (1984)). This provides us with consistent estimates of the fixed parameters but at this stage adequate estimates of their standard errors are not available.

The next stage, then, is to obtain estimates of the random part of model (1). This is done by using *MLwiN* procedures to create constraints on the fixed parameters. They are forced to be equal to those calculated from the instrumental variable procedure. The resulting estimates of the random part of the model will be consistent (Goldstein, 1986) and can then be used to obtain standard errors of the instrumental variable fixed effects estimates.

MLwiN macros called IV have been written to implement this procedure. They are available from the authors on request or can be downloaded from the Birmingham web page www.bham.ac.uk/economics/staff/tony.htm

Application

We now use the example of Fielding (1999) discussed above to demonstrate the use and performance of the instrumental variable method embodied in the macros. Simulation results are also available (Spencer and Fielding, 2000). The particular response variable used is the Mathematics test at Key Stage 1, standardised to have mean zero and unit variance.

The seven baseline tests available in the data (various forms of Mathematics and English tests) are inevitably highly correlated and so the first principal component (accounting for 60% of the variation) was used. The supplementary multilevel model for the endogenous principal component score had a similar structure to model (1) with intercept random effects for school and pupils. Fixed effect explanatory variables included pupil's ethnicity, first language and attendance at nursery school. The ones used were, on investigation, related to ability and therefore to baseline test scores. However, none appeared to have an influence on progress. It is unlikely, therefore, that they are correlated with disturbances in target model (1). Predictions of the principal component of baseline scores from this model were thus thought to provide an instrument that was free of the problem of dependence on the disturbances in the original progress model of interest.

Table 1 shows estimates of the fixed parameters (and estimated standard errors) of the adapted model (1) obtained with and without the consistent instrumental variable estimation procedure (IV). It is noticeable that the influence of gender and baseline testing decreases and that of age increases (indeed almost doubles) when the consistent procedure is applied.

Table 1: Results with and without instrumental variable procedures

	Without IV		**With IV**	
Coefficient for	**Estimate**	**s. e**	**Estimate**	**s. e**
Intercept	-0.0671	0.0520	0.0353	0.0611
Male gender dummy	0.102	0.0244	0.0758	0.0335
Centred age in months	0.0145	0.00379	0.0281	0.00828
Baseline 1st Principal Component	0.314	0.00775	0.211	0.0540

It is well known that if good instruments for the endogenous variables cannot be found, then the resulting estimates, although consistent, may be quite imprecise. In some cases standard errors can become so large as to make results uninterpretable.

The estimated standard errors produced by the IV procedures are substantially higher than those produced without it. It is a matter of judgement whether the price in imprecision is worth paying to secure the promise of consistency.

Conclusion

A solution to the problem of inconsistency caused by the presence of endogenous variables in a multilevel model has been proposed, based on instrumental variable procedures. The implementation of the consistent estimation method suggested has been made possible using the flexible macro facilities of *MLwiN*. An illustration of the method has been presented and the results contrasted with those produced when the problem of heterogeneity is ignored. It is possible that the availability of further background data might have further improved the precision of the estimates. Sound planning in data collection is therefore important.

References

Bowden, R.J. and Turkington, D. A. (1984). *Instrumental Variables*, Cambridge, CUP.

Crouchley, R. and Davies, R. B. (1999). A Comparison of Population Average and Random Effects Models for the Analysis of Longitudinal Count Data with Baseline Information, *Journal of the Royal Statistical Society*, Series A, **162** (3), 331-348.

Fielding, A. (1999). Why Use Arbitrary Points Scores? Ordered Categories in Models of Educational Progress. *Journal of the Royal Statistical Society*, Series A, **162** (3), 303-328.

Goldstein, H. (1986). Multilevel Mixed Linear Model Analysis using Iterative Generalised Least Squares. *Biometrika* **73**, 43-56.

Kiviet, J. F. (1995). On Bias, Inconsistency and Efficiency of Various Estimators in Dynamic Panel Data Models. *Journal of Econometrics*, **68**, 53-78.

Rice, N., Jones, A. and Goldstein, H. (1998). Multilevel Models where the Random Effects are Correlated with the Fixed Predictors: A Conditioned Iterative Generalised Least Squares Estimator (CIGLS). *Multilevel Modelling Newsletter*, **10** (1), 7-11.

Spencer, N. H. (1998) Consistent Parameter Estimation for Lagged Multilevel Models, *University of Hertfordshire Business School Technical Report 1*, UHBS 1998:19.

Spencer, N. H. and Fielding, A. (2000). A Comparison of Modelling Strategies for Value-Added Analyses of Educational Data, *University of Hertfordshire Business School Technical Report 2*, UHBS 2000:7.