R000223115 Short Final Report

1. Introduction: the effect of measurement errors in variables

It is widely accepted that educational measurements are made with error. Much research using OLS and multilevel modelling (MLM) regression ignores this even though measurement error in the analysis of educational effects can change conclusions. This project has looked at one way of dealing with measurement error in multilevel modelling.

Most error correction methods in regression applications are moment based (see Fuller, 1987) and until recently based on single-level regression models, though substantial recent progress has been made on multilevel models (Woodhouse *et al.,* 1996; Goldstein, 1995). Theory has been developed mainly for the situation where errors are normally distributed, but also for multinomial misclassification. More general models have not been widely considered.

2. The bootstrap

One possibility for handling more general models is the bootstrap (see for example Davison and Hinkley, 1997). The bootstrap is increasingly used to approach problems where analytic solutions are not readily available. In the errors-in-variables situation, it may be valuable to utilise the bootstrap both to adjust for biases, and to produce confidence intervals. Compared with the standard moments methods, the bootstrapping procedures potentially offer wider applicability with less reliance on parametric distributional assumptions.

Building on some preliminary investigations by Hutchison (1999), the current work used bootstrap bias-correction methods to allow for measurement error in multilevel models. A range of models for measurement error was included, normally distributed continuous error terms, multinomial misclassification, random slopes and sampling errors in aggregated variables. Simulated data and real data from research projects already carried out at the NFER were used. Topics investigated included

a) which of the existing bootstrap methods is most effective in correcting regression-type models?

b) what do we condition on?

c) how many replications are required for confidence intervals for coefficients?

d) determining the number of bootstrap iterations required for convergence

e) how do we estimate standard errors for the iterated procedure?

3. What kind of bootstrap?

Using bootstrapping to correct for bias is theoretically simple. First one estimates the model using a technique which for some reason is biased. In this project, the bias arises because of errors in the variables. Then, using an iterative bootstrapping procedure, one attempts to estimate a model which will give rise to the observed model when the biased procedure is used, i.e. in this project, after measurement error is added. This is the bias-corrected estimate.

There are two main types of bootstrap paradigms, whole case and residuals resampling. In the situation where it is used for bias-correction for a model, it is necessary to use the latter. Within the residuals approach, two types are distinguished, empirical residuals, based on the discrepancy between actual and predicted outcome, and modelled residuals (Carpenter *et al.*, 2000; Hutchison, 1999). Empirical residuals generally have the attractive quality that they are based directly on the data, and are less dependent on the fit of the data to some particular statistical model. This advantage is less evident in the measurement errors situation, since the empirical residual is contaminated by the measurement error. The project devised shrinkage techniques for allowing for this, and applied them to the two level variance components model with normal error and compared the results with those from a residual modelling approach.

4. Shifting sands - what do we condition on?

In carrying out residuals bootstrapping, it is necessary to condition on some values of the predictor variables. By definition of course one does not know the true values of these, so some kind of imputation is required. The project investigated transformations of existing data, and concluded that unbiased estimates of individual values did not in general give unbiased estimates of the quantities required for regression. With the aid of Professor Goldstein of the University of London Institute of Education, we devised methods for transformations of data for single-level problems. These did not readily extend to the multilevel situation, and we decided to concentrate on reproducing the sufficient statistics (i.e. the generalised SSP matrix) rather than individual values on a case by case basis.

Problems investigated included:

single continuous independent variable measured with error;
two correlated continuous independent variables, one measured with error;
aggregated group-level effects;
errors in variables and random slopes;

single dichotomous independent variable with misclassification.

The majority of our investigations were carried out on simulated data, which gave the opportunity to investigate the success or otherwise of the methods. The program MLWiN (Rasbash *et al.*, 2000) was the central part of this. In general we found that:

a) The substantial majority of the bias in the fixed effects was removed by one iteration of the process, though usually up to four iterations were necessary for the results to stabilise.

b) Fixed effects were better reproduced than random ones, especially level 2 random effects. This finding is fairly standard in this type of area.

c) These results are asymptotic, with 200 level 2 units. Results for smaller samples, of 50 schools, while still removing the preponderance of the biases, are less effective in reproducing the generating distribution.

d) For the normally distributed variance components case with normally distributed errors, modelled and empirical residuals gave the same results

e) A method of allowing for misclassifications in categorical independent variables has worked successfully on a simple example with only one independent variable, but a more general method of allowing for misclassification in independent variables has not yet been shown to be unbiased.

f) All methods of adjusting for measurement error, including existing ones based on the method of moments, assume that the measurement error in the observed data can be adequately represented by its moments. Especially for the relatively small sample numbers at higher levels, this may not hold.

5. Estimating standard errors and confidence intervals

Methods of estimating standard errors and confidence intervals were considered, and it was decided that the only satisfactory method of estimating standard errors using the bootstrapping paradigm is to replicate the complete set of analyses, in other words to bootstrap the bootstrap.

This was potentially extremely cpu-intensive, and a short-cut method was devised. In this we carried out a large number of bootstrap replications (resamples) with a relatively small number of analyses within each iteration. We were then able to use a multilevel model to separate the systematic between-resamples component of variation from the within-resamples noise. A range of techniques was investigated for this.

6. Conclusions and next steps

The technique investigated here has proved very promising as a method of correcting for measurement error. It has proved applicable to a wide range of applications both on simulated and actual data. Where it has been possible to verify it, the results have corresponded with the known generating distribution.

This project has been conducted in close contact with the Multilevel Models project of London University Institute of Education, who together with the Fellows group of former ALCD Fellows have proved a valuable sounding board and source of suggestions.

We aim to extend the range of applications to other error mechanisms, for example to censored distributions and multiple independent variables including misclassified categorical variables together with continuous variables.

One problem with this investigation is that we have only looked at asymptotic results. Further research is needed to unearth how these procedures may be modified to be used with smaller samples, especially smaller than the 100 - 200 units at level 2 generally used in this project.

7. Publications and conference papers

Paper presented to the American Statistical Association Annual Conference 2000.
Paper presented to the European Conference on Educational Research, 2000.
Paper presented to the Amsterdam Multilevel Models Conference, 2001.
Other papers are currently in preparation for publication.

8. References

CARPENTER, J., GOLDSTEIN, H. & RASBASH, J. (1999). ‘A nonparametric bootstrap for multilevel models’, *Multilevel Modelling Newsletter.*
DAVISON A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.
FULLER, W.A. (1987). *Measurement Error Models.* London and New York: Wiley.
GOLDSTEIN, H. (1995). *Multilevel Statistical Models, Second Edition.* Kendall’s Library of Statistics, 3. London: Arnold.

HUTCHISON, D. (1999). The effect of group-level influences on pupils' progress in reading. Doctoral thesis submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy of the University of London.

RASBASH, J., HEALY, M., CAMERON, B. & CHARLTON, C. (2000). MLWiN; v1.10.1006 (Computer Program).

WOODHOUSE, G., YANG, M., GOLDSTEIN, H., RASBASH, J. and PAN, H. (1996). 'Adjusting for measurement error in multilevel analysis', *Jr. Roy. Statist. Soc. (A)*, **159**, 2, 201 - 12.