1. Introduction: measurement errors in variables and their effect

This project has looked at one method of allowing for measurement error in multilevel modelling. In much research in education and other areas, OLS and multilevel modelling (MLM) regression tend to treat observations as if they were made without error. Yet it is widely accepted that this assumption does not hold universally, and it has been shown that taking account of measurement error in the analysis of educational effects can change conclusions. For example, in an analysis of the educational attainment of children aged 11, Goldstein (1979) showed how the sign of a coefficient can be reversed when a correction is made for measurement error, while Hutchison (1999, 2000) showed that measurement error can give rise to apparent group-level effects where none exist.

Most methods of dealing with measurement error in regression applications are moment based (see Fuller, 1987) and based on classical single-level regression models. Most educational research data has a hierarchical structure, and is more appropriately analysed by multilevel models, and substantial progress has been made on allowing for measurement error in these (Woodhouse *et al.*, 1996; Goldstein, 1995). Theory has been developed in this area 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 Efron and Tibshirani, 1993; Davison and Hinkley, 1997). The bootstrap is an approach to estimating sampling variances, confidence intervals, and other properties of statistics, as well as for correcting estimates for bias. It aims to mimic the sampling behaviour in a population by taking further samples from an already drawn sample. It is increasingly used to approach problems where analytic solutions are not readily available, either because the available distributions may not fit the data, or where available solutions are asymptotic and the problem is a small sample one, or simply because analytic solutions are too complicated to envisage. In this situation it may be valuable to utilise the bootstrap both to adjust for biases in the estimation of coefficients of the errors-in-variables model, and to produce confidence intervals. Compared with the standard moments methods of correcting for measurement error, the bootstrapping procedures potentially offer a wider range of applicability with less reliance on parametric distributional assumptions in many instances.

Hutchison (1999) has considered the application of bootstrapping methods to MLM with error. He considered two main sampling paradigms applied to regression, the **resampling residuals model** and the **resampling units model,** as well as a bias correction procedure. Among the residuals models, he compared conditioning upon observed and reliability corrected estimates of independent variables, with neither emerging a clear favourite. Among whole case resampling, he compared three methods, of which sampling entire higher level units was the best option. This research was of a preliminary nature, and while it provided useful ground-clearing information about the applicability of the bootstrap method to multilevel errors-in-variables models, it had a number of limitations. These were:

a) The bootstraps took as their starting point an already unbiased estimate.

b) Numbers of replications (500), while generally sufficient to provide estimates of standard errors, were insufficient to provide estimates of confidence intervals. Substantially more replications are required to provide estimates of confidence intervals, compared with the number required for standard errors (Efron, 1994).

c) Not all of the methods used were completely unbiased, even asymptotically.

The investigation described here built on this work. It used bootstrap bias-correction methods to allow for measurement error in multilevel models, especially in the independent variables. A range of models for measurement error was included, normally distributed continuous error terms, multinomial misclassification, random slopes and sampling errors in aggregated variables. The data sets used for this purpose included simulated data and real data from research projects already carried out at the NFER. Topics investigated included:

a) Bootstrap methods to correct regression-type models for a variety of measurement error models, as suggested above. Which of the existing bootstrap methods is most effective?

b) How many replications are required for confidence intervals for coefficients via bootstrap methods?

c) Determining the number of bootstrap iterations required for convergence.

d) Estimation of standard errors for the iterated procedure. The Kuk (1995) procedure does provide an expression for standard errors, but this is not relevant in the errors-in-variables case.

3. What kind of bootstrap?

The principle of using bootstrapping to correct for bias is 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. after measurement error is added. This is the bias-corrected estimate.

To take one very simple example, a two-level linear variance components model for true or 'latent' values x_{ij} and y_{ij} is given by

$$y_{ij} = \beta x_{ij} + u_j + e_{ij} \qquad (3.1).$$

$$Cov(u_{j'}, u_j) = Cov(e_{ij}, e_{i'j}) = Cov(u_j, e_{ij}) = 0$$

$$E(y_{ij}) = E(x_{ij}) = E(u_j) = E(e_{ij}) = 0 \ \text{var}(u_j) = \sigma_u^2; \text{var}(e_{ij}) = \sigma^2$$

The 'true' or latent values x_{ij} and y_{ij} in (3.1) are observed with measurement error m_{ij}, η_{ij} giving observed values X_{ij} and Y_{ij} where

$$X_{ij} = x_{ij} + m_{ij}$$

$$Y_{ij} = y_{ij} + \eta_{ij} = \beta x_{ij} + u_j + e_{ij} + \eta_{ij} \qquad (3.2)$$

$$Cov(m_{ij}, m_{i'j}) = Cov(m_{ij}, \eta_{ij}) = 0;$$

$$E(m_{ij}) = E(\eta_{ij}) = 0; \ \text{var}(m_{ij}) = \sigma_m^2; \text{var}(\eta_{ij}) = \sigma_\eta^2$$

m_{ij}, η_{ij} are independent of x_{ij}, y_{ij}

Attempting to estimate β using standard MLM procedures will give a biased estimate γ.

There are two main types of bootstrap paradigms, namely whole case and residuals resampling. In many applications either is acceptable. In the situation where the bootstrap is used for bias-correction for a model, it is necessary to use a residuals approach. 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 $e_{ij} - m_{ij}\beta$ is contaminated by the measurement error. Empirical estimates are not generally available for measurement error, virtually by definition, and it is necessary to use a modelling approach for these. 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 (see Appendix A and Section 6, **Estimating standard errors and confidence intervals**, below).

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. In the measurement error situation by definition of course one does not know the true values of these, so some kind of imputation is required. Reliability-corrected values were considered an attractive prospect as they were the best predictor of the value of the underlying x. Hutchison (1999) used observed and reliability-corrected values (X and ρX, respectively) for continuous variables with normal errors. While he observed little difference between the two, neither was completely satisfactory, both giving biased values for the total sums of squares even in an OLS regression. This was less important in a simple bivariate regression, since simple scale adjustments to the distribution of the random error term enabled an unbiased result for β, but led to substantial complications when other correlated independent variables were introduced into the equation.

The project started by considering OLS regression, with the aim of working from the simple case, and sorting out some of the problems before attacking the multilevel model. It was then realised that the aim was to predict the SSP matrix, rather than the individual values, and initially, we considered using $\rho^{\frac{1}{2}}X$. This worked well for simple bivariate OLS regression, but fell down when additional independent variables were introduced. Professor Goldstein of the London University Institute of Education (personal communication) suggested a solution which involved Choleski transformations of the original data to give a data set with the error-corrected SSP matrix. This worked well for OLS regression but did not generalise readily to the multilevel situation.

The GLS SSP matrix is given by Block Diagonal $\lfloor S_1, S_2, \ldots S_j, \ldots \rfloor$ where

$$S_j = \begin{vmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{vmatrix}$$

for a level 2 unit containing three level 1 units. It became clear that simple transformations of the existing data did not readily reproduce the multilevel structure, in particular the correlation between lower level units within a higher level unit.

The alternative approach was to get away from transformation of individual cases, and to aim rather to reproduce the sufficient statistic. The eventual tactic adopted was to estimate the variance covariance matrices of level 1 and level 2 components of independent variables, and to produce

random components corresponding to these matrices using the MLWiN (Rasbash *et.al.*, 2000) procedure MRAN. These were then added to produce independent variables with the desired properties (Appendix A).

This could be extended to error in aggregated level 1 variables. Measurement error in these was a combination of the aggregated error in the level 1 variables, and sampling from complete level 2 units. This latter was straightforwardly achieved by taking a bootstrap sample from level 2 units (Appendix B).

The question of dealing with misclassifications in categorical independent variables was more problematic. In this situation the observed value of a p-category variable may be represented by a p-category vector $A_{ij} = [0,0,0,...1,..0]$ with zeroes in all positions except for the observed value. The corresponding true value is x_{ij}. Categorical variables may be treated as a set of dichotomous variables, so we focused on dichotomous variables. Following our experience with continuous variables, we concentrated on reproducing the sufficient statistic, rather than transforming individual values of the variables. The usual approach in moments methods (Fuller, 1987; Goldstein, 1995) of correcting for measurement error is first to transform the independent variable by multiplying by $\hat{x}_{ij} = \mathrm{K}^{-1}A^{T}{}_{ij}$ where $\mathrm{K} = \begin{bmatrix} r & 1-s \\ 1-r & s \end{bmatrix}$

Unfortunately this does not lend itself readily to simulation procedures as $\hat{x}_{ij} = \mathrm{K}^{-1}A^{T}{}_{ij}$ is not generally an integer. Two methods were investigated. The first method was suggested by Professor Goldstein. It involved taking the value $\hat{p}_j = \hat{x}_j = \mathrm{K}^{-1}A^{T}{}_j$ for the j^{th} group, and producing $Int[N_j\hat{q}_j]$ zeroes and $Int[N_j\hat{p}_j]$ ones, and drawing the remaining element with the appropriate probability to give the correct overall proportion. This approach worked well (Appendix C) for this simple application.

However, we realised that it did not offer the possibility of readily modelling the situation where there was more than one independent variable, and a more complex model was devised. In this, the level 1 and level 2 components of variation in x were estimated, and a distribution of level 2 elements drawn. For each level 2 unit, N_j pairs of [0,1] units were drawn. The zero was weighted by q_j and the one by p_j. Theoretically this reproduces the observed SSP matrix, and the error variance, and offers the possibility of including more than one independent variable. Unfortunately simulations do not appear to converge to give produce unbiased results (Appendix D). Investigation will continue until an explanation is found.

5. Results of bias-correction simulations

The methods outlined above were tested out by simulations on a variety of problems. We generated data according to a specified model, added measurement error, and attempted to reproduce the original generating mechanism starting from the observed data and knowledge of the measurement error rule. The problems we investigated included:

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

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 (see Section 2 above) gave the same results.

e) As noted above, the proposed general method of allowing for misclassification in categorical independent variables has not yet been shown to be unbiased.

f) All methods of adjusting for measurement error, including the existing ones based on the method of moments, rely on the assumption that the realisation of the measurement error incorporated 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 (Appendix D).

6. Estimating standard errors and confidence intervals

The stability of these results can be investigated by bootstrapping procedures. The results are then manipulated to estimate standard errors and confidence intervals. We originally considered the possibility of standard errors directly from the bias correction runs, for example by oscillations between iterations of the convergence process, or between analysis replications when the result had converged. Further consideration convinced us that these would be indicators of variation in the uncorrected regression rather than the corrected one. 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 (Appendix A).

If we want to get sufficient resamples to estimate distributions and confidence intervals, then a minimum of 2,000 resamples would be required. Carrying out the entire bias correction procedure on all 2,000 resamples would mean an extremely large number of analyses. The project was able to suggest a less cpu-intensive method. Adopting the following convention for the three levels of looping, we have

a) **Resamples** from the original (actual or generated) data set.

b) **Iterations to convergence** within each resample.

c) **Replications** within each iteration.

In this method 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. This gives a two-level model for β_{bcd}, the d^{th} replication within the c^{th} iteration of the b^{th} resample.

$$\beta_{bcd} = \beta_0 + \beta_{bc} + e_{bcd} \qquad (6.1)$$

$$V[\beta_{bc}] = \sigma^2_\beta, V[e_{bcd}] = \sigma^2_e$$

$sqrt\{V[\beta_{bc}]\}$ could be taken as an estimate of the standard error of the estimate of β_0. A normal approximation to confidence intervals could be taken from the highest level variation. For a more general result, shrunken top-level residuals could be partially reinflated to give the appropriate variance (Appendix A).

In contrast to the bias-correction procedure, this offered the possibility of using whole cases bootstrap for the outer bootstraps. Also in contrast, we did not know the true values of the standard errors, so we could only compare the different types of bootstrap procedure. Results were similar for different methods, though it is possible that whole cases bootstrapping gives slightly larger standard errors than the two residuals resampling methods.

7. 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. (Some problems with one of the methods for dichotomous independent variables are still apparent. Resolving these is our next priority.)

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.

At present these procedures have been programmed on a one-off basis using the MLWiN programming language. We would aim to be able to produce macros to carry out at least some parts of these procedures.

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.

8. 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.

9. Appendices.

Appendix A. Paper presented to the American Statistical Association Annual Conference 2000
Appendix B. Paper presented to the European Conference on Educational Research, 2000
Appendix C Working paper SRAG 01/01.
Appendix D Working paper SRAG 02/01.

10. References

CARPENTER, J., GOLDSTEIN, H. and 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.

EFRON, B. (1994) 'Missing data, imputation, and the bootstrap', *J. of the American Statistical Association,* **89**, 463 - 475.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

FULLER, W.A. (1987). *Measurement Error Models.* London and New York: Wiley.

GOLDSTEIN, H. (1979). 'Some models for analysing longitudinal data on educational attainment', *J. of the Royal Statistical Society*, **142,** 3, 407 - 42.

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.

HUTCHISON, D. (2000). 'When is a compositional effect not a compositional effect?' Submitted to *Jr Educ Behav Stats.*

KUK, A. (1995). 'Asymptotically unbiased estimation in generalized linear models with random effects', *J. of the Royal Statistical Society*, **B**, 2, 395 - 407.

RASBASH, J., HEALY, M., CAMERON, B. and 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.