BOOTSTRAPPING THE EFFECT OF MEASUREMENT ERRORS IN VARIABLES ON REGRESSION RESULTS.

Dougal Hutchison, Jo Morrison, Rachel Felgate, National Foundation for Educational Research
Dougal Hutchison, NFER, The Mere, Upton Park, Slough, Berkshire, SL1 2DQ

Key words: Measurement error, Bootstrapping, Multilevel models

1. Introduction: Measurement errors in variables, their effect on regression results and how to allow for it.

It has been shown that taking account of measurement error in the analysis of educational effects can change results, for example reversing conclusions (Goldstein, 1979) or creating apparent effects where none exist (Hutchison, 1999a, 1999b). Fuller (1987) gives a comprehensive account of methods for dealing with errors of measurement in OLS regression models. Methods for allowing for measurement error in multilevel data have been described in Goldstein (1995). This paper describes a new method of allowing for error in multilevel regression by using bootstrapping procedures. We illustrate this method on a simple two level model with two independent variables.

A two-level linear model for y_{ij} and true or 'latent' values x_{ij}, z_{ij}, where *i,j* refers to the i^{th} level-1 unit within the j^{th} level-2 unit is given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + u_j + e_{ij} \qquad (1.1).$$

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

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

z_{ij} is considered to be measured without error in this example.

The 'true' or latent values x_{ij} and y_{ij} in (1.1) are observed with measurement error m_{ij}, η_{ij} giving observed values X_{ij} and Y_{ij} where

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

$$Y_{ij} = y_{ij} + \eta_{ij} = \beta_1 x_{ij} + \beta_2 z_{ij} + u_j + e_{ij} + \eta_{ij} \qquad (1.2)$$

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

$$E(m_{ij}) = E(\eta_{ij}) = E(x_{ij}) = E(y_{ij}) = 0;$$

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

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

These are standard assumptions in this type of work, as defined by Goldstein, (1995): for examples of other assumptions, see Fuller (1987). Theory has been developed in this area mainly for the situation where errors are normally distributed, but also for multinomial misclassification (Fuller, 1987; Goldstein, 1995). More general models have not been widely considered, though Woodhouse (1998) has looked at the effect of errors in variables on slopes.

2. The use of the bootstrap to correct for biases

There are two main uses for bootstrapping techniques, estimation of sampling distributions and standard errors, and correction of biases. We discuss the use of the bootstrap for sampling distributions and standard errors in a later section. Bootstrapping techniques can be used to correct for biases in estimation techniques, using an iterative procedure. We illustrate on model 1.1, 1.2 above.

Stage 1

Regress Y on observed $\underline{X} = (X, Z)$ and find $\underline{\hat{\gamma}_0} = (\hat{\gamma}_{11}, \hat{\gamma}_{21})$, observed coefficient.

Stage 2

Simulate $\hat{Y}$, using $\underline{\beta_0} = (\hat{\beta}_{10}, \hat{\beta}_{11}) = \underline{\gamma_0}$, and an estimated value of $\underline{x} = (\hat{x}, \hat{z})$ to be determined.

Add measurement error to $\underline{\hat{x}}$ to give $\underline{\hat{X}}$.

Regress simulated $\hat{Y}$ on $\underline{\hat{X}}$ and find $\underline{\gamma_{1b}}$, observed coefficient.

Do this a large number B of times, and find $\underline{\gamma_1}$ the mean of the $\underline{\gamma_{1b}}$

Estimate bias $\underline{\hat{b}_1}$ by $\hat{b}_1 = (\hat{\gamma}_1 - \hat{\gamma}_0)$.

Estimate $\underline{\hat{\beta}_1}$ as $\underline{\hat{\beta}_0} - \underline{\hat{b}_1}$

Stage 3

Repeat stage 2, starting at $\underline{\hat{\beta}_1}$.

Keep iterating until process converges.

3. Example: Correcting for Measurement Error: Hierarchical Model with Two Predictor Variables: Simulated data

We consider two predictor variables, one of which is subject to measurement error. So we have the variable x_{ij}, measured with error by X_{ij}

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

and the variable $Z_{ij} = z_{ij}$, assumed measured without error. The variables x, z are correlated with correlation ρ.

The aim is, given the observed Variance-covariance matrix C, to produce two variables, $\hat{x}, z$ which have the error-corrected Variance-covariance matrix c. All that is required is to generate a pair of variables that provide the appropriate V-C matrix. The variables individually do not need to be only linear transformations of the corresponding observed variables.

A basic data set of 5000 cases of 200 level 2 units with 25 level 1 units in each, was then created, according to the model

$$y_{ij} = \beta_1 x_{ij} + \beta_2 z_{ij} + u_j + e_{ij} \quad : \quad \beta_1 = \beta_2 = 1$$

$$E(y_{ij}) = E(x_{ij}) = E(z_{ij}) = E(u_i) = E(e_{ij}) = 0$$

$$C(x_{ij}, u_i) = C(x_{ij}, e_{ij}) = C(z_{ij}, u_{ij}) = C(z_{ij}, e_{ij}) = 0,$$

$V[u_j] = \sigma_u^2$, $V[e_{ij}] = \sigma^2$.

x, z are correlated ρ_{xz}

We added measurement error m_{ij} to x_{ij} to give X_{ij}. y_{ij}, X_{ij}, z_{ij} form the basic data set under investigation in each set of analyses.

Analyses are carried out for ratio $\sigma_w^2/(\sigma_w^2 + \sigma_b^2)$ 0.10, 0.20, 0,30 where σ_w^2, σ_b^2 are within- and between-level 2 variances for x, z, and values of x-reliability from 1.0 to 0.65 and values of ρ_{xz}=0.10, 0.40, 0,80

We carry out the procedure separately on Level 1 and Level 2.

Level 1 variance.

We estimate $\sigma_{\tilde{X}}^2, \sigma_{\tilde{Z}}^2, \sigma_{\tilde{X}\tilde{Z}}$ at level 1. The program MLWiN (Rasbash *et al, 2000)* was used for this.

Since we are not looking at level-2 error, only the level 1 VC matrix $\tilde{C} = \begin{bmatrix} \sigma_{\tilde{X}}^2 & \sigma_{\tilde{X}\tilde{Z}} \\ \sigma_{\tilde{X}\tilde{Z}} & \sigma_{\tilde{X}}^2 \end{bmatrix}$ needs to be corrected.

The matrix $\tilde{c} = \begin{bmatrix} \rho_1 \sigma_{\tilde{X}}^2 & \sigma_{\tilde{X}\tilde{Z}} \\ \sigma_{\tilde{Z}\tilde{X}} & \sigma_{\tilde{Z}}^2 \end{bmatrix}$, where ρ_1 is the level-1 reliability is taken as the target for the simulation.

The command MRAN in MLWiN is now used on matrix $\tilde{c}$ to create the level-1 part $(\hat{x}, \hat{z})$ of the simulation data set. The resultant data is only equal to the required quantity in expectation and is subject to sampling fluctuation. The variables are transformed to make it precisely equal. This is done multiplying the data $(\hat{x}, \hat{z})$ by $\tilde{M}\tilde{L}^{-1}$, where $\tilde{L}$ is the Choleski decomposition of $V[\tilde{X}, \tilde{Z}]$, and $\tilde{M}$ is the Choleski decomposition of $V[\rho_1 \tilde{X}, \tilde{Z}]$. Variables will be $\hat{\tilde{x}}, \hat{\tilde{Z}}$.

Similarly we created the Level-2 data, $\hat{\bar{x}}, \hat{\bar{Z}}$.

Add $\hat{\bar{x}}, \hat{\tilde{x}}$ and $\bar{Z}, \tilde{Z}$ to form the total $\hat{x}, \hat{Z}$. Measurement error was added to $\hat{x}$ give $\hat{X}, \hat{Z}$.

The bootstrap procedure as in section 2 was used to estimate β_1, β_2. In each set of analyses, a large number of cases were simulated according to the given model. Sets of analyses using 2000 replications were carried out. Ten iterations were used to investigate the convergence of the procedure.

Figure 3.1 shows one example of the convergence of the process for a correlation $\rho(x,z) = 0.8$. The true value for both $x-$ and $z-$coefficients is 1.0. It can be seen that at the first iteration, the coefficient of x is below the 'true' value, and that of z is above. From about iteration 4, values stabilise to values slightly below 1.0. . However, one would not expect the process to converge precisely to this value because of the random quantities introduced in generating the original data.

c) We have been working with up to 2000 **replications** within each iteration.

This suggests a total of 2000 resamples
each including * 5 iterations
each including *2000 replications
20 000 000 bootstrap analysis (per problem).

Impractical at this stage for real analyses!

Can this be improved?

Within each **Resample** the **Iterations** and the **Replications** have different aims.

The **iterations** reduce the (expected) bias in the estimate. Graphs so far with constructed data have shown that the bias is effectively removed in four iterations for the simple model considered here and that the variation remaining is in the nature of oscillation rather than bias-correction
The **replications** reduce the variance. Graphs comparing 500 and 2000 replications show that there is less variation in the latter.

If we have a large number of equivalent resamples, then this should provide a large set of estimates (with a number of replications giving rise to variation within each). If we know when the iterations have converged, then we can focus attention on the replications within each converged analysis. This is a multilevel structure (replications within iterations within resample). We could view it in this way, as a three-level structure. However, in the middle level, the iterations are not exchangeable if the procedure has not converged. Consequently we would prefer to focus on a single iteration at or beyond the convergence stage. This would give us a two level model (replications within converged-iteration-within-resample). It should be possible to feed these results into a multilevel model.

In fact we shouldn't need a very large number of replications within each, since the multilevel structure may mean that we can handle a degree of variation. This would have the drawback that we wouldn't necessarily know that the iterations have converged. It would be necessary to have some kind of prior idea of the total number of iterations under a wide range of resamples. Alternatively one could run a larger number of iterations than strictly necessary, and examine the convergence behaviour.

This would potentially give a two-level model for β_{bcd}, the d^{th} replication within the c^{th} iteration of the b^{th} resample.

$$\beta_{bcd} = \beta_0 + \beta_{bc} + e_{bcd} \qquad (4.1)$$
$$V[\beta_{bc}] = \sigma_\beta^2, V[e_{bcd}] = \sigma_e^2$$

$sqrt\{V[\beta_{bc}]\}$ 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 re-inflated to give the appropriate variance. Then the percentiles of these partially reinflated residuals could be used to give percentiles of the distributions.

Example of implementation
This is the theory. We next describe how this was implemented for the whole case resampling. As before, we have two predictor variables, one of which is subject to measurement error. So we have the variable x_{ij}, measured with error by X_{ij}

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

and the variable $Z_{ij} = z_{ij}$, assumed measured without error. The variables x, z are correlated with correlation $\rho = 0.8$ in this example.

We create a basic data set of 5000 cases, as in Section 3 above. 2000 replications were carried out in each analysis.

Results of a set of simulations are shown in Table 4.1

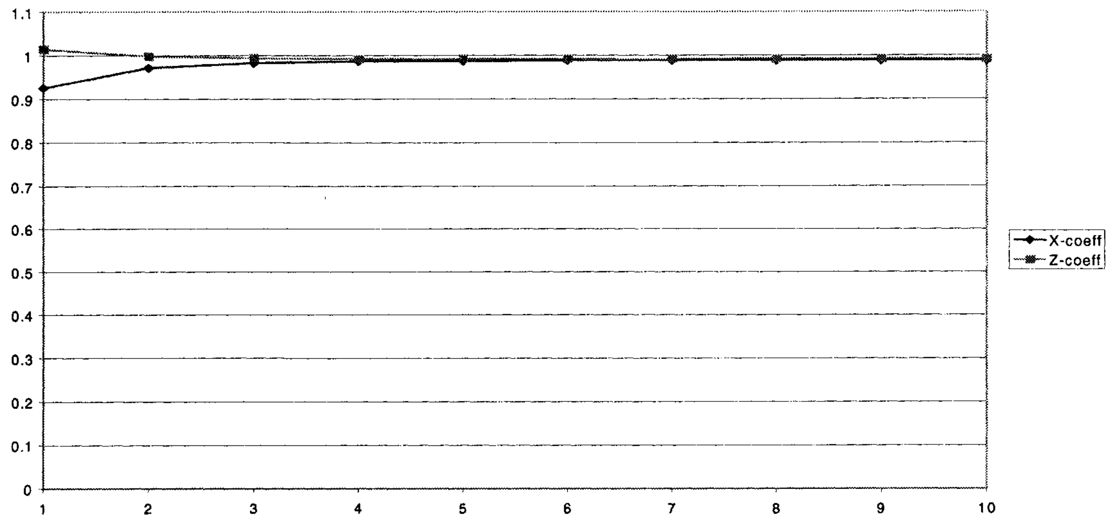
Table 4.1: Results of Bootstrap Estimation of Std Error

Variable	Estimate	Standard Error	Generating value
X	0.99	0.015	1
Z	0.99	0.011	1
L-2 Var	4.54	.76	4
L-1 Var	25.93	1.08	25
5000 resamples			

The standard error of the X-coefficient is estimated as 0.015, and that for the Z-coefficient is rather smaller at 0.012. This would be expected, since there is no measurement error in Z.

The estimated values of the variance components are rather higher than the generating values, especially the L-2 var as a proportion of the actual.

Fig 3.1: Convergence of X,Z coefficients:N=5000;Corr=0.8



4 Estimating Standard Errors by the bootstrap

How can the standard error of the result be estimated? Remember that the eventual solution is the result of a series of iterations, with a large number of bootstrap replications at each one.

a) One obvious possibility could be to use the standard deviation of the coefficients in the replications at the final iteration. However this would be providing the standard error of the γ-coefficients (i.e. the uncorrected regression) rather than the β – coefficients (i.e. the corrected regression, which is what we would be seeking).

b) The amount of oscillation between successive iterations is another suggestion. This has the same problem.

c) Kuk (1995) produced a formula for the standard errors of coefficients under this type of approach. This requires a differentiable formula connecting β and γ, which would not be available readily, since the two are connected by a bootstrap procedure.

Desiderata for an estimate of standard error.

a) It should not contain anything that is a feature of the procedure being used to estimate the coefficients. For example, the number of bootstrap replications should not affect it.

b) It should not contain any of the uncertainty in estimating the true value of x from X. It would thus be the value of the estimated standard error of the coefficients at whatever values are chosen for the independent variables.

Suggested means of proceeding.
Produce a bootstrap replication of the original sample, and carry out the procedure on the resample.
Replicate many times. *i.e.* bootstrap the bootstrap. There are two types of possibility for the 'outer' bootstrap

a) Whole case resampling

b) Residuals resampling. This can be either parametric or non-parametric (Carpenter *et al*, 1999; Hutchison, 1999a).

Here we present results using whole case resampling. Research continues on residuals resampling

This means that we have three levels of looping. Some kind of convention on nomenclature is obviously necessary.

a) **Resamples** from the original (actual or generated) data set. This is what we have described as the 'outer' bootstrap above. The literature suggests that it would be necessary to take of the order of 2000 resamples to get reliable estimates of the percentiles, confidence intervals, etc.

b) **Iterations to convergence** within each resample. Work so far suggests that something like 4-5 iterations would be required. (This may be larger on more complex problems).

5. Conclusions

The results presented in this paper have provided an example of a potentially generalisable procedure for estimating the sampling behaviour of multilevel regression models under measurement error.

The next steps will aim

- to use residual resampling procedures (model-based simulations or non-parametric residuals)
- to consider other error distributions
- to examine more complicated models

References

CARPENTER, J., GOLDSTEIN, H. & RASBASH, J. (1999) 'A nonparametric bootstrap for multilevel models'. *Multilevel Modelling Newsletter.*

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.

---------------- (1987). *Multilevel Models in Educational and Social Research.* London: Griffin.

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

HUTCHISON, D. (1999b) '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.

PLEWIS, I. (1985). *Analysing Change: Measurement and Explanation Using*
Longitudinal Data. Chichester: John Wiley and Sons.

RASBASH, J., HEALY, M., CAMERON, B, & CHARLTON, C. (2000) MLWiN v1.10.006 (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.