# Bootstrapping for multilevel models

**Multilevel Models Project**

**Working paper**

09 December 1998

## The bootstrap sample

Consider a simple random sample of *n* observations $x_1,....x_n$ from which we wish to estimate a population quantity, say a mean or median. We choose an estimator, say $\bar{x}$ and we wish to estimate features of the distribution of this estimator, say its standard error or population quantiles. The simplest *nonparametric* bootstrap is obtained as follows.

A single bootstrap sample is obtained by sampling randomly (i.e. according to the assumed mechanism which generated the observations), with replacement, *n* observations from the original sample. Denote this by $X^* = \{x_1^*,....,x_n^*\}$. Then we can obtain *B* of these bootstrap samples, $X^{*1}, X^{*2},...., X^{*B}$.For each of these we calculate our estimate, say of the mean, and each of these is referred to as a bootstrap *replicate*. It is such replicates which are used for inference.

## Standard error estimates for bootstrap functions

For a set of replicates we can calculate the standard error or quantiles to make inferences. Thus, we can get an *estimate* of the standard error of the mean simply by calculating the standard deviation of the bootstrap replicates for the sample mean. Note, however, that this is only an estimate: it is based on a finite sample of bootstraps. As the sample size tends to infinity so this becomes more accurate and approximates the *ideal bootstrap estimate.* If $\theta^*$ is the bootstrap estimator then the ideal bootstrap estimate for the standard error is the square root of $E_F(\theta^* - \theta)^2$, where *F* is the distribution function for the data.

Note, however, that even as the number of bootstrap replications tends to infinity, the estimate of the population density function which is used to generate the bootstrap samples is the empirical 'plug in' one derived from the observations by placing mass points (e.g. equal probabilities) at each one. Thus, with nonparametric bootstrapping, we do not have exact inference. This does not carry over to the parametric case where the assumed population distribution is used for sampling. In some situations the nonparametric bootstrap can perform very badly, for example in small or moderate samples where the statistic of interest is the smallest or largest value, say of a set of higher level residuals in a multilevel model.

In practice, we would normally wish to stop generating bootstrap replicates when the running estimate of the quantity of interest (the standard deviation in this case) 'settles down' to a predetermined accuracy - for example in terms of the coefficient of

variation of the bootstrap estimate - when it reaches a certain value. The coefficient of variation depends on the underlying distribution so will often not be useful when that is unknown. Clearly we require a general practical stopping rule. What will be particularly important is visual inspection of the updated histogram and smoothed density function.

A further consideration, as with all statistical analysis, is the detection of rogue values or 'outliers', in this case individual replicates. Density displays and box and whisker plots are useful diagnostic tools here. We should be careful about discarding extreme values, and as an alternative we can use robust estimators of the standard error of replicates. One such would be

$$\frac{\hat{\theta}^{*(\alpha)} - \hat{\theta}^{*(1-\alpha)}}{2z^{(\alpha)}} \qquad (1)$$

Where $z^{(\alpha)}$ is the $100\alpha^{\text{th}}$ percentile of the standard Normal distribution. Unless the distribution of the bootstrap quantity $\theta$ is Normal this is biased - nevertheless, it is consistent by virtue of the Central Limit Theorem, and this same result allows us to use approximate Normal theory distribution for calculating confidence intervals for the parameter function of interest. Inspection of the bootstrap density function and the use of Normal plots will show how good the Normal approximation is in any particular case.

In some cases, for example estimating a mean or a set of regression coefficients, the standard error of a bootstrap sample can be obtained analytically, depending only on functions of covariates (e.g. a cross product matrix) and the variance, or residual variance of the observations which is obtained from the original analysis. This does not carry over directly to the multilevel case, where the standard errors are functions of the parameters, but we can study the accuracy of these estimated standard errors via the bootstrap replications.

## Bootstraps for complex data structures

We shall use as an illustration a 2-level variance components model and the JSP data set (Woodhouse, 1996).

$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij} \qquad (2)$$

where the explanatory variables are 8 year maths score and gender and the response is 11 year maths score. We have simulated the response from the model results given in Woodhouse (1996) to ensure an approximately Normal distribution. Based on 887 pupils in 48 schools the IGLS estimates are given in table 1.

| **Table 1. JSP 2 level variance components model parameter estimates (IGLS).** | |
|---|---|
| *Fixed* | |
| Intercept | 16.06 (0.93) |
| 8 year maths | -0.17 (0.37) |
| Gender | 0.58 (0.033) |
| *Random* | |
| Level 2 | 4.61 (1.32) |
| Level 1 | 29.32 (1.43) |
| | |

We now consider drawing a bootstrap sample. We first consider nonparametric versions. To do this we need to decide whether we are going to sample complete units or just residuals. In general it seems that we would wish to use the latter, since mostly we are concerned with conditional inference, i.e. fixing the explanatory variables. In some situations, however, such as survey samples, it is more natural to think of all the variables as generated randomly so that complete unit selection is to be preferred.

The process of selecting a bootstrap sample corresponds to the supposed probabilistic mechanism which generated the data. This can be modelled as the selection of a simple random sample of school residuals according to the density *f(U)* and within each school a sample of students according to *f(E)*. As we shall see, this is appropriate for the parametric bootstrap but raises difficulty in the nonparametric case.

## Nonparametric Bootstrap

In the nonparametric complete unit bootstrap suppose we sample, with replacement, a random sample of schools. Suppose we number these 1,…,n . Then, if we sample with replacement from the students associated with the *j*th school, this will in general lead to variable total numbers (*N*) of students across bootstrap samples. This procedure, however, does retain the data structure. The variability of *N* will add some noise to our estimates but for moderate or large sample sizes this will be negligible and the procedure will be consistent. For higher level models we obtain a consistent bootstrap by sampling just the higher level units with replacement.

Another possibility is to sample level 1 units directly. Having selected a random sample of the required size we then sort into their actual level 2 units. This leads to a variable number of level 2 units and a variable number of level 1 units per level 2 unit, but retains the overall number of level 1 units. This procedure, however, pays no attention to the sample structure since each level one unit is sampled independently so that the within-unit correlation structure is not preserved. Likewise, if we sample level 2 units and then sample level 1 units from within each level 2 unit, the *joint* probability of selection for two level 1 units within a level 2 unit is the product of their separate selection probabilities. This is also the case for two level 1 units from different level 2 units - in the balanced case these joint probabilities are $1/n^2$, where *n* is the size of each level 2 unit. Thus, again the within-unit correlation structure is not preserved. In both these cases the independent selection of level 1 units will tend

to add precision to the estimation of level 2 effects and so overestimate the level 2 variation in the bootstrap samples. We note that the same considerations apply if the level 1 units are selected with replacement, sorted into their level 2 units and then these level 2 units selected with replacement.

These same considerations apply when sampling empirical (estimated) residuals in addition to the other problems which occur as follows. For sampling residuals within a nonparametric bootstrap, we work with the estimated (posterior) residuals (possibly after centring them to ensure they have zero means). The following procedure retains the sample structure. Sample with replacement the level 2 residuals, one for each level 2 unit. For each level 2 unit, sample with replacement the required number of level 1 residuals associated with that same bootstrap sampled level 2 residual. The required number is the number in the original data set for that level 2 unit. Note that in some cases this will mean sampling more level 1 residuals (with replacement) than there actually are associated with the chosen level 2 residual. The reason for this is that the level 1 and level 2 residual estimates are correlated and we need to preserve this correlation structure in our bootstrap sampling. Note that both level 1 and level 2 residuals are 'shrunken': the variance of each is less than the population variances, but the correlation between them ensures that the variance of the sum is equal to the total residual variance. In the variance components case this is equivalent to sampling the raw residuals for each chosen level 2 unit and then further sampling with replacement from these raw residuals to achieve the required number.

For each bootstrap sample we then carry out the estimation of the parameters of the model. The results of doing this, sampling residuals, for 500 bootstraps for the JSP data is given in Table 2. A difficulty with this procedure, however, is that the amount of shrinkage is correlated with the school size. Thus, the larger level 2 residuals will also tend to have the largest number of level 1 residuals so that these will be given greater weight in the estimation. This violates the assumption that the random errors provided for the bootstrap should be independent of the unit sizes. This will tend to lead to an upward bias and this is confirmed in Table 2 (linked level 1 residuals). The alternative procedure of selecting level 1 residuals from the overall set of level 1 residuals will tend to reduce both level 2 and level 1 variances as is also shown in Table 2 (unlinked level 1 residuals).

A final possibility is to select, for each school, a set of linked level 1 + level 2 residuals and attaching these to the same number of sets of fixed variables by selecting these with replacement from each school. This, however, destroys the sample structure and again leads to overestimation of the level 2 variation.

## Parametric Bootstrap

In the parametric case, we sample first the level 2 residuals from the (estimated) level 2 distribution, in this case a simple Normal distribution. Then we sample level 1 residuals from the (estimated) level 1 distribution. The structure is preserved since, according to the model assumptions the distributions are independent across levels. The procedures extend naturally to the random coefficient case.

| **Table 2. Results of 500 bootstrap replications for three bootstrap procedures for the JSP data in Table 1. Mean of bootstraps (s.d. in brackets - estimating model s.e.)** | | | | |
|---|---|---|---|---|
| | Complete case - random sample of level 1 units | Sampling complete level 2 units only | Posterior residuals - unlinked | Posterior residuals - linked |
| | | | | |
| *Fixed coefficient* | | | | |
| Intercept | 16.03 (1.09) | 16.11 (0.91) | 16.09 (0.95) | 15.97 (1.40) |
| Gender | -0.18 (0.38) | -0.14 (0.38) | -0.18 (0.36) | -0.18 (0.36) |
| 8 year maths | 0.58 (0.034) | 0.55 (0.033) | 0.58 (0.032) | 0.58 (0.032) |
| *Random* | | | | |
| Level 2 variance | 6.34 (1.09) | 4.46 (1.00) | 3.11 (0.95) | 6.62 (1.40) |
| Level 1 variance | 27.75 (1.38) | 29.27 (1.46) | 28.12 (1.35) | 26.69 (1.96) |

We notice that in none of these cases do we obtain satisfactory estimates of the random parameters, for the reasons already discussed. The fixed parameters and their standard errors are, however, well estimated except where we have tried to preserve the distributions of level a and level 2 units.

Table 3 shows the results of a fully parametric bootstrap obtained by simulating from the estimated random parameters of the model.

| **Table 3. JSP 2 level variance components model parameter estimates (IGLS) and 500 parametric bootstraps** | | |
|---|---|---|
| *Fixed* | Fitted model (s.e.) | Bootstrap (s.d.) |
| Intercept | 16.06 (0.93) | 16.06 (0.93) |
| 8 year maths | -0.17 (0.37) | -0.17 (0.37) |
| Gender | 0.58 (0.033) | 0.58 (0.033) |
| *Random* | | |
| Level 2 | 4.61 (1.32) | 4.46 (1.29) |
| Level 1 | 29.32 (1.43) | 29.20 (1.36) |
| Note that the RIGLS estimates for the level 2 and level 1 variances are 4.76 and 29.30 respectively. | | |

We see now that the bootstrap estimates are very close to those from the fitted model and similar to those from the complete level 2 nonparametric bootstrap in Table 2, although the standard deviation for the level 2 variance in the latter case appears to be an underestimate. The RIGLS level 2 estimate which corrects for the ML bias is higher by an amount which is the difference between the fitted estimate and the bootstrap one, implying that the bootstrap accurately corrects for the bias in the ML estimate of this parameter. This leads onto the topic of bias correction

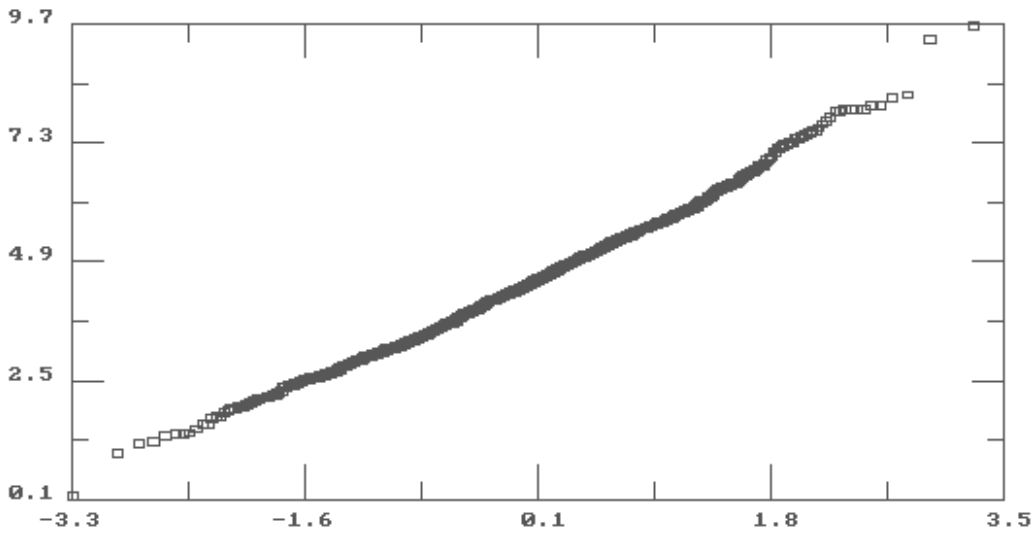
## Bootstrap bias correction

If the bootstrap estimate of a parameter (or other function of the data) is $\theta^*$ then the bias in the estimate $\hat{\theta}$ is $\theta^* - \hat{\theta}$. Thus the bias corrected estimate is $2\hat{\theta} - \theta^*$. In some models the bias of the estimation procedure is a function of the parameter values, so that a simple bias correction will be an approximation only and an iterative procedure will be necessary. This is the case for generalised linear multilevel models with nonlinear link functions and a worked example is given by Goldstein (1996). Because a bias corrected estimate of a parameter may have greater variability it is also useful routinely to check the *accuracy* of a bias corrected estimate by using it (or a set of these) as the basis for another set of bootstrap replications. As a rule of thumb, and as a default, 500 bootstrap replications should be used for bias correction.

## Confidence intervals

For many purposes the Normal approximation for the bootstrap replications is adequate and we can use the estimated standard errors for constructing confidence intervals (and significance tests). For example Figure 1 is a Normal plot based on 1000 bootstrap replications for the level 2 variance from Table 3.

**Figure 1. Normal score plot for level 2 variance bootstrap replications of Table 3.**




In general, however, we may not be able to rely upon the Normal approximation (although studying plots such as Figure 1 should help in making a decision in any particular case). In this case the simplest procedure is to use the empirical bootstrap distribution by simply reading off the $100\alpha$ - percentile points, interpolating where necessary. Call these $\hat{\theta}^{*(\alpha_1)}$, $\hat{\theta}^{*(\alpha_2)}$ where in the standard symmetrical case $\alpha_2 = 1 - \alpha_1$ and the coverage is $2\alpha$. This does, however, require a large number of replicates, as a rule of thumb 2000 can be used for a 95% interval.

Where there may be biases a better interval is the bias corrected one computed as follows. Define

$$\alpha_1 = \Phi(2\hat{z}_0 + z^{(\alpha)})$$
$$\alpha_2 = \Phi(2\hat{z}_0 + z^{(1-\alpha)}) \quad (3)$$
$$\hat{z}_0 = \Phi^{-1}(\frac{\#\{\hat{\theta}^* < \hat{\theta}\}}{B})$$

where $\Phi$ is the standard Normal cumulative distribution function and *B* is the number of bootstrap replicates.

For the iterated bootstrap, if we denote the final bias-corrected estimate by $\hat{\theta}_c$ then the percentage points are given by

$$\alpha_1 = \Phi(\hat{z}_0 + z^{(\alpha)})$$
$$\alpha_2 = \Phi(\hat{z}_0 + z^{(1-\alpha)}) \quad (4)$$
$$\hat{z}_0 = \Phi^{-1}(\frac{\#\{\hat{\theta}^* < \hat{\theta}_c\}}{B})$$

This makes the assumption, however, that the parameters from the (biased) bootstrap replicates have the same variability as those derived from an unbiased procedure. In general this will not be the case where the bias is a function of the underlying true values. Suppose the functional relationship between the bias-corrected value and the biased one is given by $\theta_c = g(\theta^*)$. If we apply this transformation to the final set of bootstrap replicates $\{\theta^*\}$, with the median fixed at $\hat{\theta}_c$, then we can use these transformed values for inference. This relationship can be estimated, for example, by simulating samples from a range of values covering the random parameter space of interest. This space will include values beyond the range generated by the procedure because of the variability of individual replicates. For multiparameter models this may be complicated. An alternative approximation is, for each parameter, to regress the sequence of bias corrected values on the bootstrap replicate means, using a low order polynomial relationship and extrapolating to cover the range of the bootstrap replicates. This line is then used to transform the final set of bootstrap replicates. If we wish to estimate the standard deviation of the bias corrected replicates then we can fit a straight line and use the slope of this line to scale the standard deviation as calculated from the actual bootstrap replicates. Kuk (1995) considers a simple scaling correction which is effectively just the ratio $\hat{\theta}_c / \hat{\theta}^*$, that is the slope of the line passing through the origin and the final estimates. This is the one used by *MLwiN* . Some simulations of these procedures would be useful.

In studying the Normality of a bootstrap set of replicates it may also be useful to look at (Normal) kernel density estimates for varying window sizes.

It is also useful to be able to study running estimates of the bootstrap parameters of functions, including standard errors (bootstrap standard deviations) and percentile estimates so that visual inspection can be used to determine when to stop bootstrap sampling. In this case it is useful to compute, and plot, an estimate of the coefficient of variation, based e.g. on an estimate of standard deviation taken from the previous, say, 100 running estimates and the current running estimate.

## Bootstrap likelihood

The likelihood, considered as a function of a parameter $\theta$, is proportional to

$$L(\theta) = \prod_i p(y_i|\theta) \qquad (4)$$

where $i$ indexes the data units. The *partial likelihood* based on a parameter estimate $\hat{\theta}$ rather than the data $\{y_i\}$ can be approximated by a bootstrap as follows. We consider the parametric bootstrap.

The first stage is to generate $B_1$ bootstrap replications to produce bootstrap parameter estimates: label the set of the parameter of interest $S_1=\{\hat{\theta}^*_1,..,\hat{\theta}^*_b,..,\hat{\theta}^*_{B_1}\}$. For each replication (i.e. from the parameter estimates associated with each replication) we generate a second stage bootstrap set of replicates giving the set of interest $S_{2b}=\{\hat{\theta}^{**}_{b1},....,\hat{\theta}^{**}_{bB_1}\}$. For $S_{2b}$ we estimate the (Normal) kernel density $\hat{p}(t|\hat{\theta}^*_b)$ as a function of $t$ and evaluate it at $t=\hat{\theta}$. Because the set $S_{2b}$ was generated from a population with parameter $\hat{\theta}^*_b$, $\hat{p}(\hat{\theta}|\hat{\theta}^*_b)$ is an estimate of the partial likelihood of $\theta$ at $\theta=\hat{\theta}$. We thus have estimates of the likelihood for all the values in $S_1$ and we can use a suitable smoother (such as LOESS) to plot the likelihood function. In fact, for the region of interest a simple polynomial, or fractional polynomial function may be adequate. From this function we can obtain the maximum and by plotting *-2log(likelihood)* we can obtain confidence intervals using the asymptotic chi squared approximation.

This can be extended to more than one parameter (the estimates for all the parameters are available from the bootstrap replications), but this will then involve smoothing in more than one dimension, although again we may be able to achieve a satisfactory smoothing via an additive function of polynomials.

*Harvey Goldstein*