ESTIMATING BOOTSTRAP CONFIDENCE INTERVALS FOR TWO-LEVEL MODELS

Erik Meijer
Frank M. T. A. Busing
Rien van der Leeden
Leiden University

Multilevel models are generally estimated using maximum likelihood ML methods. Confidence intervals are then obtained straightforwardly from the estimates and the information matrix. Frequently, however, data are not normally distributed and sample sizes are not large, in which case the ML confidence intervals may not be adequate. Bootstrap confidence intervals may be useful alternatives in these cases. In this paper, bootstrap confidence intervals will be developed for multilevel models and it will be shown in a small simulation study that in some cases the performance of the bootstrap confidence intervals is better than the performance of the ML intervals.

Key words: multilevel analysis, hierarchical linear model, maximum likelihood, nonnormality, resampling.

1 INTRODUCTION

In social and behavioral science research, multilevel analysis has by now been established as a powerful technique for the analysis of hierarchically structured data. In hierarchical data the assumption of independence is violated because nested membership relations exist among the units of observation. Appropriate modelling of the resulting intra-class dependency is a major keystone of multilevel analysis. For an extensive discussion of theory and application of multilevel analysis we refer to Bryk and Raudenbush (1992), Longford (1993) and Goldstein (1995).

Multilevel analysis involves fitting hierarchically formulated linear (regression) models, mostly referred to as multilevel models. Estimation in these models usually relies on maximum likelihood ML methods. The various computer programs for multilevel analysis employ versions of full information FIML and restricted maximum likelihood REML methods. Two vital assumptions underlying ML theory are that (a) the residuals are independently distributed, usually following a multivariate normal distribution, and (b) the sample size is (sufficiently) large. More specifically, the attractive properties of FIML estimators—consistency, (asymptotic) efficiency, (asymptotic) normality, confidence intervals obtainable from the information matrix—are derived from the supposition that the sample size goes to infinity and the variables follow the assumed distribution. In practice, however, these assumptions will frequently be met only approximately, which may lead to biased estimators and incorrect standard errors (Busing, 1993) and, consequently, incorrect confidence intervals.

The bootstrap is a general approach to obtain confidence intervals . It has proven to be a method that yields satisfactory results in small sample situations under minimal assumptions (see, e. g., Efron, 1982; Hinkley, 1988). Thus, it seems useful to consider bootstrap confidence intervals in multilevel models for cases where the assumptions necessary for ML methods are violated. In this paper, we will show that under nonnormality the ML confidence intervals may be incorrect if the assumptions are not met. It will be shown that bootstrap confidence intervals may be more accurate in those cases.

The application of the bootstrap to multilevel models is not straightforward. Depending upon the nature of the data and the assumptions one is willing to make, there

are several possibilities, each with its own associated problems. For bootstrap confidence intervals, usually two choices have to be made. The first is the *resampling method* and the second is the *type of confidence interval*, given the resampling method.

In section 2, the choice of resampling method will be discussed. Section 3 discusses the different types of confidence interval. In section 4, a small simulation study will be discussed, in which the ML confidence intervals will be compared with several types of bootstrap confidence intervals. Section 5 contains the discussion.

2 BOOTSTRAPPING MULTILEVEL MODELS

The general idea of the bootstrap is to use the empirical distribution of the data to estimate properties of the estimators. In practice, this usually means that a Monte Carlo simulation study is performed, in which new (bootstrap) samples are drawn with replacement from the original data. For each of the bootstrap samples, the parameters are estimated and the properties of the estimates are used as estimators of the properties of the original estimator. In this section and the next, these statements will be elaborated within the framework of confidence intervals for two-level models. Extension of the principles to more levels is straightforward.

The two-level model

We assume data are obtained from N individuals nested within J groups, with group j containing N_j individuals. The Level-1 units are the individuals and the Level-2 units are the groups. In the context of multilevel analysis it is customary to specify linear regression models for each level in the data separately. Consequently, at Level-1, for each group j ($j = 1, \ldots, J$), the within-group model is given by

$$y_j = Z_j \beta_j + \varepsilon_j, \tag{1}$$

where $y_j = (y_{1j}, \ldots, y_{N_j j})'$ is the vector containing values on an outcome variable, Z_j is an $N_j \times q$ matrix with explanatory variables (including the constant, i. e. a unit vector coding the mean), β_j is a $q \times 1$ vector of regression coefficients, and $\varepsilon_j = (\varepsilon_{1j}, \ldots, \varepsilon_{N_j j})'$ is a vector with residuals.

We assume β_j to be a vector of *random* regression coefficients. Suppose group level variables exist that explain part of the variation in these random coefficients. Then, at Level-2, the between-group model is given by

$$\beta_j = W_j\gamma + u_j, \tag{2}$$

where W_j is a $q \times p$ matrix with explanatory variables (including the constant) obtained at the group level, γ is a $p \times 1$ vector containing fixed coefficients and u_j is a $q \times 1$ vector with residuals.

Usually, it is assumed that $\varepsilon_j \sim N(0, \sigma_\varepsilon^2 I_{N_j})$ and $u_j \sim N(0, \Theta)$, where σ_ε^2, the variance of the Level-1 residuals, is an unknown (scalar) parameter, and Θ, the covariance matrix of the Level-2 residuals, is a (symmetric) matrix of unknown parameters. The parameters can then be estimated by ML methods and confidence intervals can be obtained from the parameter estimates and the estimated standard errors, which are easily derived from the information matrix (see, e. g., Bryk & Raudenbush, 1992, chapter 10). If the error terms are not normally distributed, the same symbols (i. e., σ^2 and Θ) will be used for the (co)variance parameters of the error terms.

Resampling the two-level model

In order to make the bootstrap succeed, the simulation should reflect the properties of the stochastic model that is assumed to have generated the data. Therefore, a resampling scheme for multilevel models must first of all take into account the hierarchical structure of the data, that is, the fact that observations are subject to intra-class dependency. Another aspect of the stochastic model is whether the explanatory variables are considered fixed (design) variables or random variables. Van der Leeden, Busing, and Meijer (1995) discuss various resampling methods for multilevel models with fixed or random explanatory variables.

In this paper, we will assume that the explanatory variables are random variables. The corresponding bootstrap method is called *cases bootstrap* by Van der Leeden et al. (1995).

The resampling procedure is as follows:

1. Draw a sample of size J with replacement from the *Level-2 units*, that is, draw

a sample j_k^*, $k = 1, \ldots, J$ (with replacement) of Level-2 unit numbers with a corresponding set of scores on the Level-2 variables $W_{j_k^*}$.

2. For each k, draw a sample of entire cases, with replacement, from (the original) Level-2 unit $j = j_k^*$. Then, for each k, we have a set of data (y_{ik}^*, Z_{ik}^*), $i = 1, \ldots, N_{j_k^*}$. Obtain the bootstrap sample $(y_{ik}^*, Z_{ik}^*, W_{j_k^*})$ by linking these data to the proper set of scores $W_{j_k^*}$.

3. Compute estimates for all parameters of the two-level model.

4. Repeat Steps 1–3 B times.

An alternative formulation is: (1) draw one entire Level-2 unit (y_j, Z_j, W_j), containing N_j Level-1 cases, with replacement; (2) from this Level-2 unit, draw a bootstrap sample (y_j^*, Z_j^*, W_j^*) of size N_j with replacement; (3) repeat steps 1 and 2 J times; (4) compute all parameter estimates for the two-level model; (5) repeat Steps 1–4 B times.

The above procedure shows that for the cases bootstrap each observed response y_{ij} keeps joined together with the observed scores on the explanatory variables in Z_{ij} and W_j.

3 TYPES OF CONFIDENCE INTERVALS

In this section, we will discuss a number of different types of bootstrap confidence intervals for a parameter θ with true value θ_0. The performance of these intervals will be compared to the intervals that are derived from the asymptotic covariance matrix of the estimators under the normality assumption. We will only discuss two-sided intervals. One-sided intervals are defined analogously. The intended nominal coverage of the confidence interval will be denoted by $1 - \alpha$, so that the probability that the interval contains the true parameter value should be approximately $1 - \alpha$.

Notation

Before we introduce the different bootstrap confidence intervals, we will introduce some useful notation.

Let $\Phi(z)$ be the standard normal distribution function. Then z_α is the α-th quantile of the standard normal distribution, $z_\alpha = \Phi^{-1}(\alpha)$.

Let the distribution function of the estimator $\widehat{\theta}$ be $H(\theta)$, that is, $H(\theta) = \Pr(\widehat{\theta} \leq \theta)$. A consistent estimator of this distribution function is obtained from the B bootstrap replications $\widehat{\theta}^*_b$, $b = 1, \ldots, B$, of $\widehat{\theta}$:

$$\widehat{H}(\theta) \equiv \frac{\#\{b : \widehat{\theta} \leq \theta\}}{B} . \tag{3}$$

Bootstrap normal confidence interval

If the assumptions of the model, including the normality assumptions, hold, then the estimators are asymptotically normally distributed with a certain covariance matrix, derived from the likelihood function (Magnus, 1978; Busing, Meijer, & Van der Leeden, 1994, Appendix A). The usual confidence intervals are therefore

$$\left[\widehat{\theta} + z_{\frac{1}{2}\alpha} \widehat{\text{se}}_N(\widehat{\theta});\ \widehat{\theta} + z_{1-\frac{1}{2}\alpha} \widehat{\text{se}}_N(\widehat{\theta}) \right], \tag{4}$$

where $\widehat{\text{se}}_N(\widehat{\theta})$ is the estimator of the asymptotic standard deviation of $\widehat{\theta}$, derived from normal theory.

Under mild regularity conditions, the estimators are asymptotically normally distributed, even if the random terms in the model are not. In that case, $\widehat{\text{se}}_N$ may not be a consistent estimator of the standard deviation of the estimators of the variance components, although it is still consistent for the fixed parameters. This suggests replacing $\widehat{\text{se}}_N$ in (4) by a bootstrap estimator. This gives the *bootstrap normal* confidence interval

$$\left[\widehat{\theta} + z_{\frac{1}{2}\alpha} \widehat{\text{se}}_B(\widehat{\theta});\ \widehat{\theta} + z_{1-\frac{1}{2}\alpha} \widehat{\text{se}}_B(\widehat{\theta}) \right], \tag{5}$$

in which $\widehat{\text{se}}_B$ is the bootstrap estimator of the standard deviation of $\widehat{\theta}$. Alternatively, one might use

$$\left[\widehat{\theta}_B + z_{\frac{1}{2}\alpha} \widehat{\text{se}}_B(\widehat{\theta});\ \widehat{\theta}_B + z_{1-\frac{1}{2}\alpha} \widehat{\text{se}}_B(\widehat{\theta}) \right], \tag{6}$$

where $\widehat{\theta}_B$ is the bootstrap bias-corrected estimator of θ.

The bootstrap normal confidence interval relaxes the assumption of normality of the data, but still heavily relies on the asymptotic normality of the estimators. In finite samples, however, the estimators may not be approximately normally distributed (Busing, 1993).

Percentile interval

The idea behind this interval is quite different from the idea behind the bootstrap normal interval. It was stated above that $\widehat{H}(\theta)$ is a consistent estimator of the distribution function of $\widehat{\theta}$. Therefore, an asymptotic $1-\alpha$ confidence interval can be obtained by taking the relevant quantiles from $\widehat{H}$, which leads to the interval

$$\left[\widehat{H}^{-1}(\tfrac{1}{2}\alpha);\ \widehat{H}^{-1}(1-\tfrac{1}{2}\alpha)\right]. \tag{7}$$

The percentile interval does not rely on the asymptotic normality of $\widehat{\theta}$. Its coverage performance in finite samples is, however, frequently not very well, because the end points of the interval tend to be a little biased.

Bias-corrected percentile (BC)

The BC interval was introduced to correct for some bias in the endpoints of the percentile interval (7). We will only give the formula here. See Efron (1982, section 10.7) for the argument leading to this interval. Let

$$z_0 = \Phi^{-1}\left[\widehat{H}(\widehat{\theta})\right].$$

The BC interval is now

$$\left[\widehat{H}^{-1}\left(\Phi(2z_0+z_{\frac{1}{2}\alpha})\right);\ \widehat{H}^{-1}\left(\Phi(2z_0+z_{1-\frac{1}{2}\alpha})\right)\right]. \tag{8}$$

Note that if $\widehat{\theta}$ is equal to the median of the $\widehat{\theta}^*$, then $\widehat{H}(\widehat{\theta}) = 0.5$, $z_0 = 0$ and the BC interval (8) coincides with the percentile interval (7).

Percentile-t

The percentile-t interval (or bootstrap-t interval) takes the bootstrap normal interval (5) as its starting point. That interval is based on the idea that

$$\Pr\left(\widehat{\theta}+z_{\frac{1}{2}\alpha}\widehat{\mathrm{se}}_B(\widehat{\theta}) \le \theta_0 \le \widehat{\theta}+z_{1-\frac{1}{2}\alpha}\widehat{\mathrm{se}}_B(\widehat{\theta})\right) \longrightarrow 1-\alpha, \tag{9}$$

because $\widehat{\theta}$ is asymptotically normally distributed and $\widehat{\mathrm{se}}_B(\widehat{\theta})$ is a consistent estimator of its standard deviation. In finite samples, however, the distribution of $\widehat{\theta}$ may not be approximately normal (Busing, 1993). Therefore, instead of using quantiles of the standard normal distribution, using bootstrap quantiles may give more accurate results.

To derive the necessary bootstrap quantiles, let us rewrite (9) into the following form:

$$\Pr\left(z_{\frac{1}{2}\alpha} \leq \frac{\theta_0 - \widehat{\theta}}{\widehat{\mathrm{se}}_B(\widehat{\theta})} \leq z_{1-\frac{1}{2}\alpha} \right) \longrightarrow 1 - \alpha. \tag{10}$$

The quantiles of the normal distribution have to be replaced by quantiles of the distribution of

$$\frac{\theta_0 - \widehat{\theta}}{\widehat{\mathrm{se}}_B(\widehat{\theta})}.$$

These are estimated by quantiles of the bootstrap distribution of

$$\frac{\widehat{\theta} - \widehat{\theta}^*}{\widehat{\mathrm{se}}^*_B(\widehat{\theta}^*)}.$$

Let

$$\widehat{G}(t) = \frac{\#\left\{ b : \dfrac{\widehat{\theta} - \widehat{\theta}^*_b}{\widehat{\mathrm{se}}^*_{B,b}(\widehat{\theta}^*_b)} \leq t \right\}}{B},$$

and let $\widehat{t}_{\frac{1}{2}\alpha}$ and $\widehat{t}_{1-\frac{1}{2}\alpha}$ be the $\frac{1}{2}\alpha$-th and $1 - \frac{1}{2}\alpha$-th quantile of $\widehat{G}$, respectively, that is, $\widehat{t}_{\frac{1}{2}\alpha} = \widehat{G}^{-1}(\frac{1}{2}\alpha)$ and $\widehat{t}_{1-\frac{1}{2}\alpha} = \widehat{G}^{-1}(1 - \frac{1}{2}\alpha)$. The percentile-$t$ interval is obtained by replacing $z_{\frac{1}{2}\alpha}$ by $\widehat{t}_{\frac{1}{2}\alpha}$ and $z_{1-\frac{1}{2}\alpha}$ by $\widehat{t}_{1-\frac{1}{2}\alpha}$ in (5) and is thus

$$\left[\widehat{\theta} + \widehat{t}_{\frac{1}{2}\alpha} \widehat{\mathrm{se}}_B(\widehat{\theta});\ \widehat{\theta} + \widehat{t}_{1-\frac{1}{2}\alpha} \widehat{\mathrm{se}}_B(\widehat{\theta}) \right]. \tag{11}$$

It is necessary to have an estimate $\widehat{\mathrm{se}}^*_{B,b}(\widehat{\theta}^*_b)$ of the standard deviation of $\widehat{\theta}^*_b$ for each bootstrap resample b. This is usually obtained by performing a small bootstrap within each bootstrap resample. So, for example, $B = 1000$ bootstrap samples are drawn with replacement from the original sample and within each sample $b = 1, \ldots, B$, $B_2 = 25$ samples are drawn with replacement from the bootstrap sample. From the B_2 samples, $\widehat{\mathrm{se}}^*_{B,b}(\widehat{\theta}^*_b)$ is obtained.

This means that $B * B_2$ bootstrap samples have to be drawn and $B * B_2$ times the estimator $\widehat{\theta}$ has to be computed. In the example, this amounts to $1000 * 25 = 25\,000$ bootstrap samples and 25 000 times computing the estimator.

4 SIMULATION STUDY

In this section we evaluate the performance of the methods for constructing bootstrap confidence intervals in a Monte Carlo study. The study consists of repeatedly generating

samples from a known "population" model, performing a multilevel analysis, constructing confidence intervals and, finally, comparing the results with the "population" values.

Design

Data were generated for a two-level model that contains one predictor variable at each level. This model is given by

$$\begin{aligned} y_{ij} &= \beta_{1j} + \beta_{2j} X_{ij} + \varepsilon_{ij}, \\ \beta_{1j} &= \gamma_{11} + \gamma_{12} W_j + u_{1j}, \\ \beta_{2j} &= \gamma_{21} + \gamma_{22} W_j + u_{2j}. \end{aligned} \tag{12}$$

Predictor variables were drawn from a standard normal distribution using the polar Box-Müller method (Box & Müller, 1958). Uniform deviates were obtained with the `RANLUX` pseudo-random number generator (Lüscher, 1994). `RANLUX` was also used to obtain random numbers for other distributions. The parameters of the model were set to the following values: the fixed parameters γ were set to 1.0, the Level-2 variance components Θ_{11}, Θ_{12}, and Θ_{22} were set to 2.0, $0.5\sqrt{2.0}$ and 1.0, respectively, and the Level-1 variance component σ_ε^2 was set to 8.0. These values correspond to a conditional intraclass correlation of 0.2 and an intercept-slope correlation of 0.5. To evaluate the performance of the bootstrap methods adequately it was decided to simulate a realistic case in which the assumptions are violated to a certain extent. Therefore, we used a moderately small sample size (especially at Level-2), and a severely skewed distribution for the residuals. Specifically, the Level-2 sample size (J) was set to 20 and the Level-1 sample size (N_j) was drawn from a normal distribution with a mean of 10 and a variance of 2, and rounded to the nearest integer value. This resulted in a total sample size (N) of approximately 200. The residuals (u_{1j}, u_{2j} and ε_{ij}) were generated from a lognormal distribution with a skewness of 5.0, set in deviation of its mean (expectation equals 0.0) and unit normalized (variance equals 1.0).

The simulation procedure can be summarized by the following steps:

1. generate a two-level dataset following Equation (12),

2. use `ML` and the different methods to compute confidence intervals,

3. save estimates,

4. repeat step 1–3 R times.

R, the number of, what we have called, "macro replications" was set to 1000, which is expected to be sufficiently large to obtain reliable sampling distributions.

In the second step, parameter estimates, standard errors, and confidence intervals were computed with the MLA computer program (Busing et al., 1994; Busing, Meijer, & Van der Leeden, 1995). This program has been developed primarily for research on resampling methods in two-level models. MLA provides ML estimates of parameters and standard errors. The bootstrap confidence intervals are based on the FIML estimates. For the simulation procedure this means that in step 2, that is, within each macro replication, there are a number of bootstrap or, what we have called, "micro" replications B. B was set to 1000, which is expected to be sufficiently large to obtain accurate bounds for the confidence intervals.

Computations

The generated confidence intervals were evaluated by inspection of the coverage percentage, the proportion of cases where the true value is covered by the estimated confidence interval. For ML the usual confidence intervals were used, that is, the parameter estimate plus or minus 1.96 times the standard error estimate. The bootstrap confidence intervals were directly computed with the MLA program.

Results

Due to improper or non-converged solutions some results were invalid. Before further computations these results were removed from the analysis.

The true coverage percentage in Table 1 is 95%. The results for the fixed components indicate acceptable coverage for both conventional FIML and REML confidence intervals and bootstrap confidence intervals. Underestimated coverage can be seen for the grand mean γ_{11} (FIML= .88, REML= .89 and Normal= .88), and fixed component γ_{12} (Percentile-t= .90). Overall, the bias-corrected percentile method performs slightly better than the other methods.

Table 1. Coverage percentages confidence intervals

	Fixed components			
	γ_{11}	γ_{12}	γ_{21}	γ_{22}
FIML	.88	.93	.93	.94
REML	.89	.95	.94	.95
Normal	.88	.95	.95	.97
Percentile	.93	.97	.97	.98
BC-percentile	.94	.96	.96	.97
Percentile-t	.94	.90	.94	.95
	Variance components			
	σ_ϵ^2	θ_{11}	θ_{12}	θ_{22}
FIML	.40	.49	.67	.50
REML	.40	.56	.71	.56
Normal	.62	.80	.90	.79
Percentile	.72	.87	.89	.79
BC-percentile	.86	.54	.89	.77
Percentile-t	.95	.56	.85	.71

True value = 95 %

In case of the variance components, the conventional FIML and REML confidence intervals fail as expected, giving raise to conservative hypothesis testing. The bootstrap methods perform much better than the conventional methods, and for the Level-1 variance component the Percentile-t method performs even perfect. However, the BC-percentile and Percentile-t method show instable performances, as can be seen from the results for the Level-2 variance component θ_{11}.

5 DISCUSSION

The bootstrap confidence intervals, as introduced in this paper, may provide additional tools for hypothesis testing in cases where the sample size is small or the data are skewed. For the fixed components of the multilevel model, the surplus value of the bootstrap confidence intervals is small compared to the conventional methods, despite

the known underestimation of the standard errors. The bootstrap confidence intervals are particularly useful for the variance components, where a serious improvement may be expected compared to the conventional methods.

Future research should first concentrate on stabilizing current results, and then search for other refinements. Restricted maximum likelihood estimation can be used in bootstrap estimation, other bootstrap resampling methods can be used for confidence interval estimation and finally, other bootstrap confidence interval types, like the accelerated BC percentile interval or the variance stabilized bootstrap-t interval, can be used.

References

Box, G. E. P., & Müller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 29*, 610–611.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierachical linear models: applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Busing, F. M. T. A. (1993). *Distribution characteristics of variance estimates in two-level models; A Monte Carlo study* (Tech. Rep. No. PRM 93-04). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Busing, F. M. T. A., Meijer, E., & Van der Leeden, R. (1994). *MLA. Software for multilevel analysis of data with two levels. User's guide for version 1.0b* (Tech. Rep. No. PRM 94-01). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Busing, F. M. T. A., Meijer, E., & Van der Leeden, R. (1995). The `mla` program for two–level analysis with resampling options. In T. A. B. Snijders, B. Engel, J. C. Van Houwelingen, A. Keen, G. J. Stemerdink, & M. Verbeek (Eds.), *Toeval zit overal* [randomness is everywhere] (pp. 37–57). Groningen: iec ProGAMMA.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia: SIAM.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Hinkley, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society B, 50,* 321–337.

Longford, N. T. (1993). *Random coefficient models.* Oxford: Clarendon Press.

Lüscher, M. (1994). A portable high–quality random number generator for lattice field theory simulations. *Computer Physics Communications, 79,* 100–110.

Magnus, J. R. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics, 7,* 281–312.

Van der Leeden, R., Busing, F. M. T. A., & Meijer, E. (1995). *Bootstrap methods for two-level models* (Tech. Rep. No. PRM 95-04). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

ASSUMPTIONS, ROBUSTNESS, AND ESTIMATION METHODS IN MULTIVARIATE MODELING

Edited by

Joop Hox
Edith de Leeuw

TT-publikaties

1998, TT-Publikaties Amsterdam
Plantage Doklaan 40
NL-1018 CN Amsterdam

ISBN 90-801073-6-0
NUGI 659