A Comparison of Bootstrap Methods for Mixed Model Analysis of Longitudinal Data

Xiao Wang[1], Mark Reiser[2], Jeanne Wilcox[3], Shelley Gray[4]
[1]Statistics and Data Corporation, 21 E 6th St Suite # 110 85281, Tempe, AZ,
[2]Arizona State University, SOMSS, PO Box 871804, Tempe, AZ,
[3]Arizona State University, DELAI, PO Box 871811, Tempe, AZ,
[4]Arizona State University, SAHS, PO Box 870102, Tempe, AZ

Abstract
Cluster bootstrap is the usually used method for bootstrapping clustered data. A longitudinal study often contains multiple levels. For example, an educational study may have two levels: student level and classroom level (students nested within classrooms). In this case, resampling may be done on either the student level or the classroom level. This paper compares these two cluster bootstrap methods with the parametric bootstrap method for standard errors, bias and confidence intervals of parameter estimates obtained under a two-level mixed model. Several Monte Carlo simulations are also performed, showing that the parametric bootstrap and cluster bootstrap at the classroom level are better methods comparing with the residual bootstrap and cluster bootstrap at the student level.
Key Words: cluster effect, Monte Carlo simulation, multilevel

1. Introduction

Bootstrap is an important statistical tool that can be used to estimate the properties of an estimator. It is usually used in complex situations where asymptotic approximations are difficult to compute or non-available (D.Boos 2003). Bootstrap standard error, bias and confidence intervals are often calculated to measure the accuracy of complex statistics, for example, the parameter estimates of linear regression model.

In history, the Quenouille-Tukey jackknife method preceded the bootstrap, which was shown as an approximation to the bootstrap via the delta method (B. Efron 1979). The bootstrap principle is simple (R. T. B. Efron 1986). In the real world, we have an observed random sample $X = (X_1,\ldots\ldots,X_n)$, which is sampled from an unknown probability distribution P, and the statistic of interest is a function of X: $\hat{\theta} = s(X)$. In the bootstrap world, we have an observed bootstrap sample $X^* = (X^*_1,\ldots\ldots,X^*_n)$, which is sampled from the empirical distribution $\hat{P}$ ($\hat{P}(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i),\ for\ A \sqsubseteq R$), and the statistic is $\widehat{\theta^*}=s(X^*)$. Instead of evaluating the statistical properties (bias, standard errors, etc) of $\hat{\theta}$ based on the sampling distribution of $\hat{\theta}$, we mimic this process by evaluating these properties of $\widehat{\theta^*}$ based on the bootstrap sampling distribution of $\widehat{\theta^*}$. The benefit of doing so is we don't actually need to compute the exact bootstrap sampling distribution of $\widehat{\theta^*}$, we can use Monte Carlo methods to obtain an approximation: draw B independent bootstrap samples $X^{*(1)},\ldots\ldots, X^{*(B)}$ from $\hat{P}$, compute $\widehat{\theta^{*(b)}}$ for each bootstrap sample, and finally compute the estimated bias, standard errors and confidence intervals from $\widehat{\theta^{*(1)}},\ldots\ldots,\widehat{\theta^{*(B)}}$. It is noticeable that we assume resampled cases in the bootstrap samples are *i,i,d*.

There are various kinds of methods to draw the bootstrap samples from clustered data (A. C.A.Field 2007), such as randomized cluster bootstrap, cluster bootstrap, two-stage bootstrap, random-effect bootstrap, residual bootstrap and parametric bootstrap. In this study, we use the parametric bootstrap and the cluster bootstrap. The cluster bootstrap is a simplified version of randomized cluster bootstrap, in which clusters are selected by simple random sampling with replacement, and no further permutation. When we treat each observation as a cluster in the horizontal data format, for example, student level in this study, the cluster bootstrap is simply case resampling bootstrap. The parametric bootstrap is similar with the nonparametric bootstrap, the difference is parametric bootstrap samples are taken from the estimated parametric model instead of the empirical distribution $\hat{P}$.

In this study, we investigate the relationship between the students' grades with factors including classroom, treatment and mother education level. There are 289 students in total, and each student was measured six times during the semester. A two-level mixed model (student level, classroom level) is fitted to the data using the lme4 package in R (Bates 2010).

The paper is organized as follows: In section 2, we make a brief description of the two-level random effects model. In section 3, we introduce the bootstrap methods we use, including cluster bootstrap with either cluster in student level (CID) or classroom level (TID) and parametric bootstrap. In section 4, we explain how to calculate the bootstrap bias, bootstrap standard errors and bootstrap confidence intervals (percentile method, Bca method and bootstrap-t method). Section 5 we carry out a Monte Carlo study to evaluate the performance of different bootstrap methods we use. Finally, in section 6, we apply different bootstrap methods for our real educational sample.

2. Linear Mixed-Effects Model Without Bootstrap

In this study, we would like to investigate the relationship between the students' grades (letter sound identification, lsrSIto) with several factors, such as the treatments (condition=0/1), mother education level (matedu_c), measurement time (time) and time square (timesq), student (CID) and classroom (TID). Students are nested within the classrooms, classrooms are randomly nested within the treatments (S.Gray 2011). The linear mixed model is:

$$Y_{ijkl} = \beta_0 + \beta_1 X_{1ijkl} + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 X_{2ijkl} + \beta_5 t_{ij} X_{1ijkl} + \beta_6 t_{ij}^2 X_{1ijkl} + b_{0i} + u_{kl} + b_{1i} t_{ij} + \varepsilon_{ijkl}$$

i = 1,2,…,n_i (n_i=number of students); j = 1,2,…6 (measurement time point); k = 1,2 (treatment); l = 1,2,…, n_l (n_l=classroom number); X_1: dummy variable for treatment level (1=condition1; 0=condition0); X_2: continuous variable for mother education level; t_{ij}: measurement time point; $b_{0i} \sim N(0,\sigma_1^2)$ random intercept effect for student i; $b_{1i} \sim N(0,\sigma_3^2)$ random slope effect for student i; $u_{kl} \sim N(0, \sigma_4^2)$ random intercept effect for classroom l within treatment k; $\varepsilon_{ijkl} \sim N(0, \sigma^2)$ residual.

If we consider the general form of linear mixed model as $\vec{Y_i}=X_i\,\vec{\beta}+Z_i\vec{b_i}+\vec{\varepsilon_i}$ (i=1,2,… n_i), where $\vec{Y_i}$ is a 6×1 vector. We assume the $\vec{Y_i}$ in the same classroom l within treatment k are correlated via the random effect u_{kl}. The variance-covariance matrix of $\vec{Y_i}$ is

Cov($\vec{Y}_i$)= $Z_i G_i Z_i' + R_i$, where $R_i = Cov(\vec{\varepsilon}_i) = \sigma^2 I$. $G_i = Cov(\vec{b}_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{13} & 0 \\ \sigma_{13} & \sigma_3^2 & 0 \\ 0 & 0 & \sigma_4^2 \end{pmatrix}$.

The intraclass correlation among students within classroom kl is defined as $\rho=\sigma_4^2/(\sigma_4^2 * \sigma^2)$. The REML likelihood function is: -2logL($\vec{\alpha}; K\vec{y}$)=In |KV($\vec{\alpha}$)K`|+($\vec{y} - X\hat{\vec{\beta}}$)` V($\vec{\alpha}$)$^{-1}$($\vec{y} - X\hat{\vec{\beta}}$)+(H-P)In(2π), where V($\vec{\alpha}$) is big variance covariance matrix. The covariance parameter vector $\vec{\alpha}$ is defined as ($\sigma_1^2, \sigma_3^2, \sigma_4^2, \sigma_{13}, \sigma^2$). H is the total observation number. K is selected as E(K$\vec{Y}$)=0, P=Rank(X).

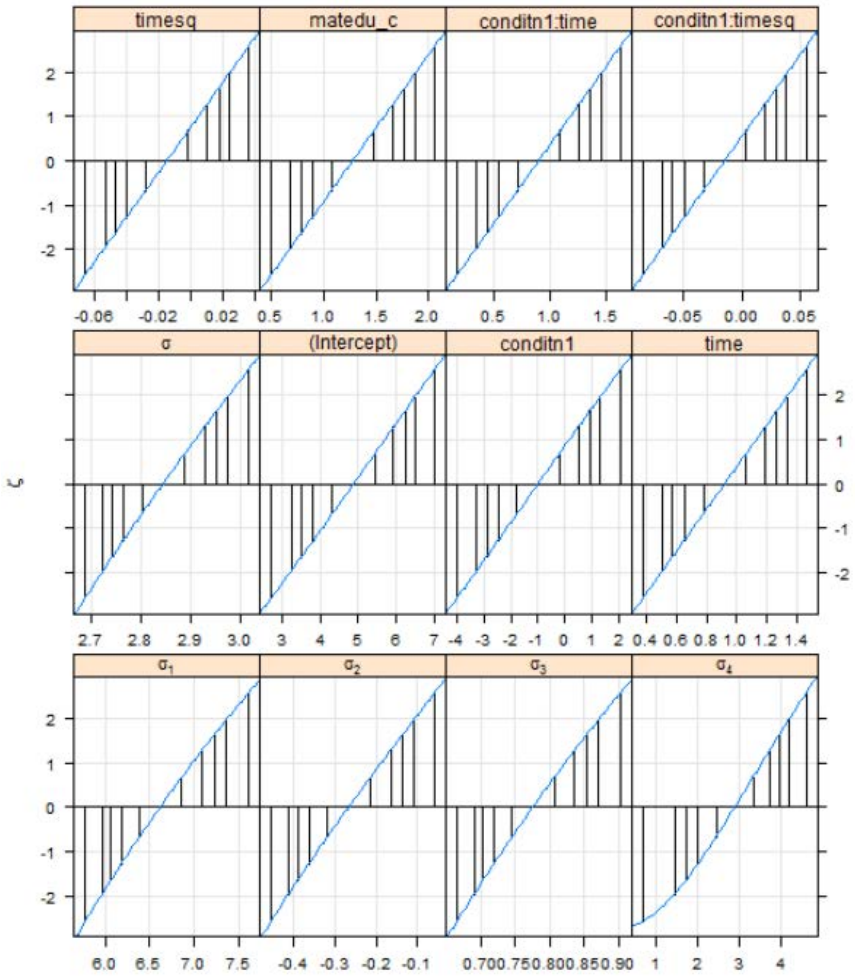

Figure 1: Signed square root, ζ, of the likelihood ratio test statistic for each of the parameters in the linear mixed model without bootstrap. The vertical lines are the endpoints of 50%, 80%, 90%, 95%, and 99% profile confidence interval derived from the REML test statistics. $\boldsymbol{\rho_{13} = \sigma_{13}/\sigma_1 * \sigma_3}$, which is the correlation.

We fit this model with lme4 package in R. The restricted maximum likelihood (REML) estimates and the profile confidence intervals are shown in Table 5 and Table 6. The lme4 package also provides the profile zeta plots for assessing the variability of the parameter estimates. The signed square root of the likelihood ratio test statistic, called ζ, is plotted versus the parameter value (Figure 1). A ζ value can be compared to the quantiles of the standard normal distribution, for example, a 95% profile confidence interval is -1.960< ζ<1.960 (Bates 2010). All the profile zeta plots are very close to straight lines, which means

the statistical inferences based on the parameters' estimates are reliable. Zero is contained in the 95% profile confidence intervals of condition1 (β_1), timesq (β_3), and condition1:timesq (β_6).

3. Cluster Bootstrap and Parametric Bootstrap Method

Bootstrapping data with more than one level is complicated. The non-parametric random-effect bootstrap and residual bootstrap are two commonly used methods, which have been fully discussed with either balanced data (A. C.A.Field 2007) or unbalanced data (Z. P. C.A.Field 2008). Both these two methods and the parametric bootstrap method generate the bootstrap samples after modelling, which assumes the model is correctly specified. Conversely, the cluster bootstrap method (case-resampling method) draws the bootstrap samples before modelling, thus is more robust to model misspecification. Of course, there are certain disadvantages of applying the cluster bootstrap method to multilevel data, for example, resampling at one level will usually destroy the natural hierarchy of the original data (Carpenter 2003).

3.1 Cluster bootstrap at the student (CID) level or the classroom (TID) level

Our dataset contains 1734 observations in the long format, which are clustered at two levels: (1) all observations from the same student, and (2) all observations from the same classroom. There are 289 unique students, each containing 6 observations (6 measurement times), and there are 92 unique classrooms, containing unequal observations.

Cluster bootstrapping at the student level: we draw samples of g = (student numbers) clusters independently with replacement. The bootstrap sample is the set of mg observations Y_{ij}^* (m=6, i=1,2,…,g), where the g m-vectors $(Y_{i1}^*, \dots, Y_{im}^*)$ are treated as $i.i.d$ distributions with probability $1/g$ on each of the g m-vectors $(Y_{i1}, \dots, Y_{im})$. In this case, we ignore correlation among students in the same classroom. It has been proven that under the random-effect model, the simple cluster bootstrap variances of several statistics, such as the sample total ($T = mg\bar{Y}_{..}$), between- and within-cluster sum of squares ($S_{B2} = m\sum_{i=1}^{g}(\bar{Y}_{i.} - \bar{Y}_{..})^2$, $S_{W2} = \sum_{i=1}^{g}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_{i.})^2$) and the covariance between the sum of squares are asymptotically correct as $g \to \infty$ with m fixed (A. C.A.Field 2007).

Cluster bootstrapping at the classroom level: we draw samples of g = (classroom numbers) clusters independently with replacement. The bootstrap sample contains $\sum_{i=1}^{g} m_i$ observations. It is noticeable that since our data is unbalanced at the classroom level, the new bootstrap sample size might not be the same as the original dataset.

In addition, we notice that the classrooms are randomly assigned with a unique treatment (condition=0/1) in the educational example. In order to reduce the distortion of the natural hierarchy of the original data, firstly we subset the original dataset into condition=0 group and condition=1 group, then do the cluster resampling separately in both groups, and finally combine the two subset samples into a new bootstrap sample.

3.2 Parametric bootstrap for random-effects models

Based on the random effects model described in section 2. We assume that the parameters $(\vec{\beta}, G_i, \sigma^2)$ have been estimated by the REML method, and we obtain the estimates $(\hat{\vec{\beta}}, \hat{G}_i, \hat{\sigma}^2)$. Then the parametric bootstrap proceeds as follows:

(1). Simulate $\varepsilon^*_{ijkl} \sim N(0, \hat{\sigma}^2)$, also simulate the $\vec{b}^*_i$ from the distribution $N(0, \hat{G}_i)$.
(2). Calculate the bootstrap sample data Y^*_{ijkl} by setting:

$$Y^*_{ijkl} = \hat{\beta}_0 + \hat{\beta}_1 X_{1ijkl} + \hat{\beta}_2 t_{ij} + \hat{\beta}_3 t^2_{ij} + \hat{\beta}_4 X_{2ijkl} + \hat{\beta}_5 t_{ij} X_{1ijkl} + \hat{\beta}_6 t^2_{ij} X_{1ijkl} + b^*_{0i} + u^*_{kl} + b^*_{1i} t_{ij} + \varepsilon^*_{ijkl}$$

(3). Refit the random-effects model to the new bootstrap sample data in step 2. Get the first set of bootstrap estimates of parameters $(\hat{\vec{\beta}}^*, \hat{G}_i{}^*, \hat{\sigma}^{2*})$.
(4). Repeat steps 1~3 B times to obtain B sets of parameter estimates. Calculate bootstrap standard errors, bias and confidence intervals.

The build-in function 'bootMer' in the R lme4 package performs model-based parametric bootstrap for mixed models. We perform two kinds of parametric bootstrap here based the option 'use.u=FALSE/TRUE': if 'use.u=FALSE', each simulation creates new values of both $\vec{b}^*_i$ and the i.i.d. ε^*_{ijkl} using *R rnorm()*. If 'use.u=TRUE', only the i.i.d. ε^*_{ijkl} are resampled, and the other random effects are fixed at their estimated values. The first one is labeled as 'parametric', and the second one is called 'residual' in the following tables. It is noticeable that the 'residual bootstrap' in the later sections is actually the 'parametric residual bootstrap', not the usual residual bootstrap.

4. Bootstrap Evaluation of the Parameter Estimates

4.1 The bootstrap estimate of standard error

We follow the notations in the Introduction section. Suppose the observed data X = $(X_1,\ldots\ldots,X_n)$ is sampled from an unknown probability distribution P, and the statistic of interest is a function of X: $\hat{\theta}$ = s(X), to which we will assign an estimated standard error. Let $\sigma(P)$ be the standard error of $\hat{\theta}$, indicating a function of the distribution P. we have: $\sigma(P) = [Var_P\{s(X)\}]^{1/2}$. We define the bootstrap estimate of standard error is $\hat{\sigma}_B = \sigma(\hat{P})$, where $\hat{P}$ as the empirical distribution.

In most cases, it is difficult to calculate the function $\sigma(P)$ and also the function $\sigma(\hat{P})$, however, since we notice that a bootstrap sample is just the random sample of size n drawn with replacement from the original data X = $(X_1,\ldots\ldots,X_n)$, we can use a Monte Carlo algorithm to approach to $\sigma(\hat{P})$. Assume the statistics calculated from each bootstrap sample are $\widehat{\theta^{*(1)}},\ldots\ldots,\widehat{\theta^{*(B)}}$, the sample standard deviation of all the $\widehat{\theta^{*(b)}}$s (b=1,2,…,B) is,

$$\hat{\sigma}^*_B = \left(\frac{\sum_{b=1}^{B} \left\{ \widehat{\theta^{*(b)}} - \widehat{\theta^{*(.)}} \right\}^2}{B-1} \right)^{1/2}$$

$$\widehat{\theta^{*(.)}} = \frac{\sum_{b=1}^{B} \widehat{\theta^{*(b)}}}{B}$$

It is easy to see that as B→∞, $\hat{\sigma}^*_B$ gets closer to $\hat{\sigma}_B = \sigma(\hat{P})$, it has been shown by Efron that the difference between $\hat{\sigma}^*_B$ and $\hat{\sigma}_B$ can be ignored once B is adequate large (50~200).

4.2 The bootstrap estimate of bias

For the non-bootstrap case, based on the definition of bias, the bias of the statistics of interest $\hat{\theta}$ = s(X) for estimating parameter μ is,

$$\xi = Bias(\hat{\theta}) = E_P R(X,P) = E_P\{s(X)\} - \mu(P)$$
$$where\ R(X,P) = s(X) - \mu(P)$$

E represents the expectation with the probability distribution of P. For example, s(X) can be the sample mean, $\mu(P)$ can be the true mean of distribution P. For the bootstrap case, the bootstrap estimate of bias is,

$$\hat{\xi}_B = \widehat{Bias}(\hat{\theta}^*) = E_{\hat{P}} R(X^*,\hat{P}) = E_{\hat{P}}\{s(X^*)\} - \mu(\hat{P})$$
$$where\ R(X^*,\hat{P}) = s(X^*) - \mu(\hat{P})$$

As we do for the bootstrap standard estimate, we can apply a Monte Carlo algorithm to approach to the $\hat{\xi}_B$, which is,

$$\hat{\xi}_B^* = \frac{\sum_{b=1}^{B}\widehat{\theta^{*(b)}}}{B} - \mu(\hat{P}) = \widehat{\theta^{*(\cdot)}} - \hat{\mu}(P) \qquad \text{As B}\to\infty,\ \hat{\xi}_B^* \to \hat{\xi}_B.$$

4.3 The bootstrap estimate of confidence interval

Once we have the B bootstrap samples, we can estimate the sampling distribution of the interested statistics, from which the bootstrap confidence intervals can be obtained. There are a variety of confidence interval types. The simplest case, standard confidence interval is calculated as,

$$\theta_{L(standard)} = \hat{\theta} - z_{\alpha/2}\hat{\sigma}_B^*, \quad \theta_{U(standard)} = \hat{\theta} + z_{\alpha/2}\hat{\sigma}_B^*$$

where $\hat{\theta}$ is the estimate of statistic of interest based on the original data, and $\hat{\sigma}_B^*$ is the bootstrap standard error. $z_{\alpha/2}$ represents the critical value of the standard normal distribution.

The percentile interval is calculated by the empirical quantiles of the bootstrap replications, $\widehat{\theta^{*(1)}}$,……,$\widehat{\theta^{*(B)}}$, that is,

$$\hat{P}\left(\widehat{\theta^*} \le \theta_{L(percentile)}\right) = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\widehat{\theta^{*(b)}} \le \theta_{L(percentile)}\right\} \approx \frac{1}{2}\alpha$$
$$\hat{P}\left(\widehat{\theta^*} \ge \theta_{U(percentile)}\right) = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\widehat{\theta^{*(b)}} \ge \theta_{U(percentile)}\right\} \approx \frac{1}{2}\alpha$$

The Bias Corrected Accelerated (BCA) method is a modification of the Bias Corrected percentile method, it adjusts for both bias and skewness of the bootstrap distribution. Define the cumulative bootstrap sample distribution of statistics $\widehat{\theta^*}$ is $\hat{G}$, that is, $\hat{G}(t) = \#(\widehat{\theta^{*(b)}} < t)/B$, (b=1,2,…,B, # represents the number counting). $\hat{G}^{-1}(\cdot)$ is the corresponding quantile function from bootstrap distribution. $\Phi^{-1}(\cdot)$is the quantile function from the standard normal distribution. The bias corrected percentile confidence interval (BC) is described as follows: $\hat{z}_0$ is the bias correction factor, $\hat{\theta}$ is the estimate of statistic of interest based on the original data. The BC confidence interval becomes the percentile interval when $\hat{z}_0 = 0$ (no bias):

$$\theta_{L(BC)} = \hat{G}^{-1}[\Phi(2\hat{z}_0 + z_{\alpha/2})]$$
$$\theta_{U(BC)} = \hat{G}^{-1}[\Phi(2\hat{z}_0 + z_{1-\alpha/2})]$$
$$where\ \hat{z}_0 = \Phi^{-1}\left[\hat{G}(\hat{\theta})\right] = \Phi^{-1}\left[\frac{\#(\widehat{\theta^{*(b)}} < \hat{\theta})}{B}\right]$$

Furthermore, we define $\hat{\theta}_{\_i}$ as the interested statistic calculated by deleting observation x_i from the original sample $(x_1,\dots,x_n)$, the average deleted statistic is defined as $\hat{\theta}_0 = \sum_{i=1}^{n}\hat{\theta}_{\_i}/n$, which is similar with the jackknife principle. The Bias Corrected Accelerated

(BCA) confidence interval is described as follows, $\hat{a}$ is the accelerated factor, the BCA confidence interval becomes the BC interval when $\hat{a} = 0$.

$$\theta_{L(BCA)} = \hat{G}^{-1}\{\Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}\left(\hat{z}_0 + z_{\alpha/2}\right)}\right]\}$$

$$\theta_{U(BCA)} = \hat{G}^{-1}\{\Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}\left(\hat{z}_0 + z_{1-\alpha/2}\right)}\right]\}$$

$$where\ \hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_0 - \hat{\theta}_{\_i}\right)^3}{6[\sum_{i=1}^{n}(\hat{\theta}_0 - \hat{\theta}_{\_i})^2]^{3/2}}$$

Finally, we consider the bootstrap-t confidence interval (A.Colin Cameron 2000). Instead of calculating the confidence interval of θ, we define $w = (\hat{\theta} - \theta)/\hat{\sigma}$, where $\hat{\sigma}$ is the estimated standard error of θ based on the original sample, and we want to calculate the bootstrap confidence interval of w. Similarly, we draw B bootstrap samples from the original sample, calculate $w_b^* = (\widehat{\theta^{*(b)}} - \hat{\theta})/\widehat{\sigma^{*(b)}}$, where $\widehat{\sigma^{*(b)}}$ is the standard error estimate based on the bootstrap sample b. Based on the calculated $w_1^*, \ldots, w_B^*$, we draw the percentile bootstrap-t confidence interval as below,

$$\hat{P}\left(w_b^* \le w_{L(percentile)}\right) = \frac{1}{B}\sum_{b=1}^{B} 1\{w_b^* \le w_{L(percentile)}\} \approx \frac{1}{2}\alpha$$

$$\hat{P}\left(w_b^* \ge w_{U(percentile)}\right) = \frac{1}{B}\sum_{b=1}^{B} 1\{w_b^* \ge w_{U(percentile)}\} \approx \frac{1}{2}\alpha$$

5. Monte Carlo Simulation

In order to examine the properties of our bootstrap methods, we conduct several Monte Carlo exercises for the two level linear mixed model below. Compared with the model in section 2, we exclude the mother education fixed effect:

$$Y_{ijkl} = \beta_0 + \beta_1 X_{1ijkl} + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 t_{ij} X_{1ijkl} + \beta_5 t_{ij}^2 X_{1ijkl} + b_{0i} + u_{kl} + b_{1i} t_{ij} + \varepsilon_{ijkl}$$

i = 1,2,…,n_i (n_i=number of students); j = 1,2,…6 (measurement time point); k = 1,2 (treatment); l = 1,2,…, n_l (n_l=classroom number); X_1: dummy variable for treatment level (1=condition1; 0=condition0); t_{ij}: measurement time point; $b_{0i} \sim N(0,\sigma_1^2)$ random intercept effect for student i; $b_{1i} \sim N(0,\sigma_3^2)$ random slope effect for student i; $u_{kl} \sim N(0, \sigma_4^2)$ random intercept effect for classroom l within treatment k; $\varepsilon_{ijkl} \sim N(0, \sigma^2)$ residual.

We generate *R*=500 Monte Carlo samples by using the parameter estimates as in Section 2. The random effect values are obtained by using *rmvnorm* and *rnorm* function in R. We assume there are 92 classrooms with 5 students in each classroom, each student is measured at six time points (time= 1, 2, 3.5, 6, 7.3, 8.8). We assign 49 out of 92 classrooms with treatment 0, and the other 43 classrooms with treatment 1. Totally, each Monte Carlo sample contains 2760 (92×5×6) observations.

For each Monte Carlo sample we generate *B*=1000 bootstraps by using the different methods discussed in section 3. Finally we get *R*=500 bootstrap confidence intervals (α=0.05). If the method is appropriate, the probability of the confidence intervals containing the corresponding true parameter values should be around 0.95 (for 95% confidence interval). It is noticeable that the model might not converge for some bootstrap samples, and the Bca confidence interval might not be able to calculate in some cases. So the total number of confidence intervals is less than or equal to *R*=500.

Table 1 and Table 2 show the Monte Carlo simulation results for different bootstrap methods with the normal distribution model. The shaded area in Table 1 and Table 2 are the numbers outside the theoretical interval $0.95 \pm 1.96 * \sqrt{\frac{0.05*0.95}{500}}$, which is (0.931, 0.969).

Table 1: Monte Carlo simulation results of the fixed effects parameters (model residual term normally distributed). R: Monte Carlo sample number. The scores in the table are the percentiles of confidence intervals containing the true parameter values. Good methods will have the values around 95%, since our significance level α=0.05.

Methods	Monte Carlo Simulations (Normal, R=500,α=0.05)					
	Intercept	conditn1	time	timesq	conditn1: time	conditn1: timesq
No bootstrap	95.60%	94.80%	96.20%	96.80%	94.00%	94.20%
Residual (Perc)	56.00%	54.60%	95.20%	97.80%	92.20%	93.00%
Parametric (Perc)	97.00%	95.80%	95.20%	94.80%	95.40%	95.00%
Cluster TID (Perc)	96.00%	93.80%	95.00%	94.40%	93.20%	92.80%
Cluster CID (Perc)	85.60%	85.60%	95.40%	96.20%	93.80%	94.80%
Residual (Bca)	55.40%	53.80%	95.00%	97.40%	91.80%	92.80%
Parametric (Bca)	96.40%	95.20%	96.00%	96.40%	95.00%	95.00%
Cluster TID (Bca)	95.80%	94.20%	93.00%	94.00%	93.20%	93.80%
Cluster CID (Bca)	85.20%	85.60%	95.00%	96.40%	94.00%	94.40%
Cluster TID (b_t)	96.40%	94.40%	95.40%	94.20%	92.80%	92.40%
Cluster CID (b_t)	77.20%	77.20%	95.60%	96.20%	93.40%	93.80%

Table 2: Monte Carlo simulation results of the random effects parameters (model residual term normally distributed). sdcor1(σ_1): random intercept effect for student; sdcor3(σ_3): random slope effect for student; sdcor4: random intercept effect for classroom; $sdcor2\ (\sigma_2) = \sigma_{13}/\sigma_1 * \sigma_3$, which is the correlation. sdcor: residual term.

Methods	Monte Carlo Simulations (Normal, R=500, α=0.05)				
	sdcor1	sdcor2	sdcor3	sdcor4	sdcor
No bootstrap	97.00%	96.00%	94.80%	96.00%	95.00%
Residual (Perc)	38.40%	32.20%	8.00%	37.60%	95.60%
Parametric (Perc)	96.40%	95.80%	94.80%	94.80%	95.80%
Cluster TID (Perc)	96.00%	94.40%	93.20%	93.60%	94.20%
Cluster CID (Perc)	28.80%	91.60%	93.40%	6.60%	93.80%
Residual (Bca)	2.14%	25.00%	0.00%	36.55%	96.20%
Parametric (Bca)	96.20%	96.00%	95.00%	95.40%	95.60%
Cluster TID (Bca)	95.40%	94.40%	94.20%	94.00%	94.00%
Cluster CID (Bca)	62.30%	96.20%	94.40%	0.00%	93.80%

We notice that for all fixed effect parameters and random effect parameters, the parametric bootstrap and cluster bootstrap (TID) methods have the Monte Carlo score around 95%, which means about 95% confidence intervals calculated by these two methods contain the true corresponding parameter values. However, the residual bootstrap and cluster bootstrap (CID) methods fails to achieve the 95% score for the fixed effects (intercept, condition1) and the random effects (sdcor1, sdcor2, sdcor3, sdcor4). This

simulation result indicates that clustering in higher level (classroom, TID) is better than clustering in lower level (student, CID); it is because that bootstrap at student level ignores correlation among the children in the same classroom. Residual bootstrap is much less accurate than the parametric bootstrap, since it only resamples the i.i.d. ε^*_{ijkl}, but not the other random effects. Different confidence interval calculation methods, the percentile, the Bca and bootstrap-t, do not have significant impacts on the final Monte Carlo simulation results.

Table 3: Monte Carlo simulation results of the fixed effects parameters (model residual term log normally distributed).

Methods	Monte Carlo Simulations (Log Normal, R=500,α=0.05)					
	Intercept	conditn1	time	timesq	conditn1: time	conditn1: timesq
Residual (Perc)	1.00%	8.80%	85.80%	88.00%	96.00%	96.40%
Parametric (Perc)	8.00%	27.80%	85.00%	83.60%	97.80%	96.60%
Cluster TID (Perc)	7.80%	22.00%	93.60%	93.60%	94.00%	94.60%
Cluster CID (Perc)	3.40%	12.40%	93.80%	93.20%	94.20%	94.40%
Residual (Bca)	1.20%	12.20%	85.80%	87.00%	95.80%	96.60%
Parametric (Bca)	7.60%	23.20%	84.00%	83.00%	97.00%	96.60%
Cluster TID (Bca)	8.20%	22.40%	92.60%	92.60%	93.60%	93.80%
Cluster CID (Bca)	4.40%	13.00%	93.20%	92.60%	93.60%	93.20%
Cluster TID (b_t)	8.20%	23.40%	92.40%	92.60%	93.20%	93.00%
Cluster CID (b_t)	1.60%	10.20%	94.20%	92.60%	93.80%	93.20%

Table 4: Monte Carlo simulation results of the random effects parameters (model residual term log normally distributed).

Methods	Monte Carlo Simulations (Log Normal, R=500, α=0.05)				
	sdcor1	sdcor2	sdcor3	sdcor4	sdcor
Residual (Perc)	23.60%	25.60%	12.00%	51.80%	0.00%
Parametric (Perc)	77.80%	72.00%	64.20%	96.00%	0.00%
Cluster TID (Perc)	94.00%	93.00%	91.20%	93.40%	0.00%
Cluster CID (Perc)	48.20%	91.60%	91.20%	5.20%	0.00%
Residual (Bca)	2.14%	25.00%	0.00%	36.55%	0.00%
Parametric (Bca)	77.60%	63.80%	70.40%	95.80%	0.00%
Cluster TID (Bca)	91.60%	90.66%	85.60%	93.40%	0.00%
Cluster CID (Bca)	62.30%	98.73%	87.20%	0.00%	0.00%

Furthmore, we also do 500 Monte Carlo simulations based on log normal distributed samples, which means in the model, $\varepsilon_{ijkl} \sim \text{LogN}(0, \sigma^2)$. We still use the parameter estimates as in Section 2, except we decrease the residual error estimate to the half of the original value. The log normal Monte Carlo results are shown in Table 3 and Table 4, most of the values are out of the theoretical interval (0.931, 0.969). However, there are still certain values calculated by cluster bootstrap method within the range. It is noticeable that the *sdcor* has 0.00% for all methods. In order to explain it, we perform another R=300 log normal Monte Carlo simulation and calculate the Monte Carlo bias of *sdcor*: Cluster CID (3.37), Cluster TID (3.38), Parametric (3.46), Residual (3.45). All methods have very large bias, which explain the 0.00% estimation power.

6. Application Results

In Section 2, we applied the simple linear mixed model to our educational data, however, we observed a lot of outliers of this data from the chi-square QQ plot (Figure 2). The next step, we applied the bootstrap methods and bootstrap bias, standard error, confidence interval calculations to our real educational dataset. The simple regression result without any bootstrapping is shown in Section 2. Here, we show the confidence interval results of the fixed effect parameters and random effect parameters in Table 5 and Table 6 correspondently.

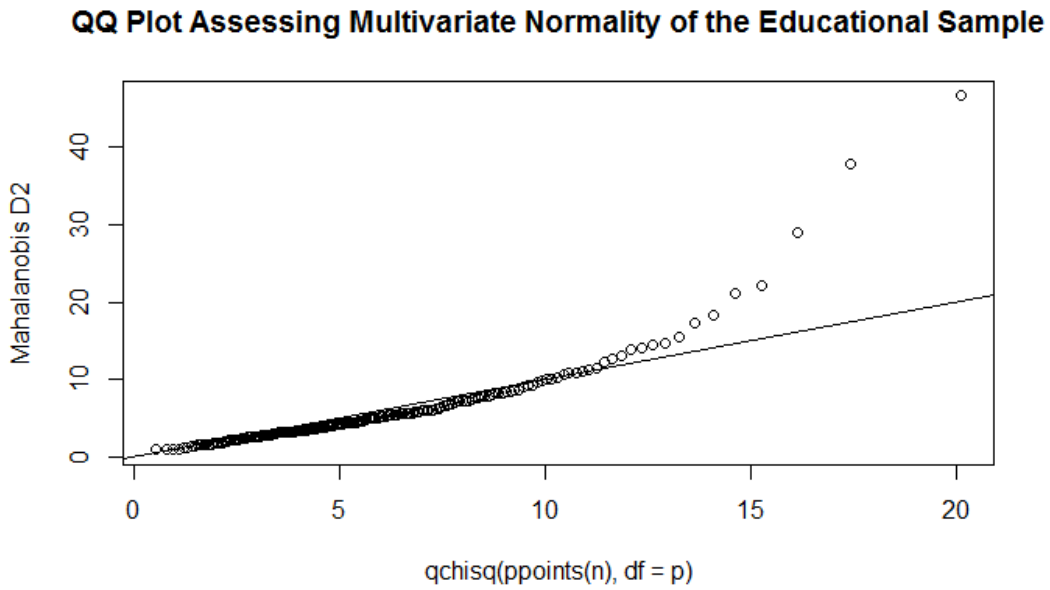

Figure 2: Chi-square QQ plot of the educational data. The multinormality of the educational data (lsrSIto1. lsrSIto2… lsrSIto6) is assessed by the chi-square QQ plot. Only the complete observations were used for making the plot, those with missing values were ignored

Under the null hypothesis H_0: $\beta_i=0$, we reject if zero is not within the calculated confidence interval. Different bootstrap methods and confidence interval calculation methods give us similar results (Table 5), for example, all methods suggest a significant linear time trend(time), however ,the quadratic time trend (timesq) is not significant. The distinct predictions are marked in bold in Table 5, for example, the cluster bootstrap (TID) indicates insignificant coefficient for condition1:time, however, the other methods suggest that it is significant. The intraclass correlation:

$$\rho = 3.016^2/(3.016^2 + 2.8476^2) = 0.5287.$$

It is interesting that the estimates of sdcor4, which is the random intercept effect for classroom, are very different between the cluster bootstrap (CID) and cluster bootstrap (TID) (Table 6). The cluster bootstrap method (CID) gives us the biggest variability between classrooms. No significant differences exist among different confidence interval calculation methods (percentile, Bca and bootstrap -t). The abbreviations are shown in Table 7.

Table 5: The 95% confidence intervals of the fixed effect parameters of the educational sample. The estimates (Est.) are the mediums of the bootstrap samples. The b_t, H0 is the bootstrap-t method with H0 imposed.

Methods	Fixed effects (95% CI) of the Educational Sample																				
	(Intercept)			conditn1			time			timesq			matedu_c			conditn1:time			conditn1:timesq		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
No bootstrap	4.871	3.24	6.493	-0.987	-3.28	1.321	0.92	0.51	1.33	-0.01	-0.05	0.024	1.283	0.685	1.869	0.899	0.34146	1.455	-0.015	-0.067	0.038
Residual (Perc)	4.783	3.92	5.625	-0.884	-2.01	0.247	0.95	0.56	1.35	-0.02	-0.06	0.022	1.28	1.142	1.407	0.857	0.3275	1.42	-0.011	-0.067	0.042
Parametric (Perc)	4.878	3.19	6.657	-0.972	-3.26	1.6001	0.92	0.48	1.34	-0.01	-0.05	0.027	1.3	0.724	1.929	0.896	0.35452	1.465	-0.015	-0.068	0.038
Cluster TID (Perc)	4.905	3.41	6.357	-1.001	-3.03	0.9749	0.9	0.26	1.69	-0.01	-0.08	0.045	1.293	0.678	1.858	0.904	**-0.018**	**1.85**	-0.015	-0.1	0.072
Cluster CID (Perc)	4.778	3.4	6.103	-0.883	-2.66	0.9946	0.92	0.41	1.43	-0.01	-0.06	0.032	1.432	0.888	2.019	0.906	0.17729	1.618	-0.016	-0.08	0.052
Residual (Bca)	4.783	4.15	5.821	-0.884	**-2.2**	**-0.04**	0.95	0.46	1.28	-0.02	-0.05	0.028	1.28	1.155	1.409	0.857	0.41749	1.488	-0.011	-0.07	0.034
Parametric (Bca)	4.878	3.2	6.638	-0.972	-3.3	1.4313	0.92	0.47	1.33	-0.01	-0.05	0.027	1.3	0.711	1.886	0.896	0.36623	1.461	-0.015	-0.067	0.038
Cluster TID (Bca)	4.905	3.33	6.302	-1.001	-2.97	0.9914	0.9	0.3	1.72	-0.01	-0.09	0.044	1.293	0.671	1.843	0.904	**-0.024**	**1.827**	-0.015	-0.103	0.071
Cluster CID (Bca)	4.778	3.59	6.213	-0.883	-2.9	0.7726	0.92	0.41	1.46	-0.01	-0.07	0.03	1.432	0.522	1.692	0.906	0.1711	1.607	-0.016	-0.078	0.057
Cluster TID (b_t,H0)	5.95	4.1	7.858	-0.858	-2.56	0.8567	4.29	1.25	7.64	-0.69	-4.1	2.348	4.338	2.211	6.472	3.197	**-0.061**	**6.648**	-0.548	-3.796	2.617
Cluster CID (b_t,H0)	4.804	3.42	6.049	-0.61	-1.92	0.6822	4.34	1.98	6.8	-0.7	-3.15	1.642	4.799	2.798	6.814	3.203	0.64204	5.603	-0.621	-2.933	2.086
Cluster TID (b_t)	0.039	-1.9	1.792	-0.013	-1.72	1.7643	-0.1	-3.3	3.38	0.067	-3.33	3.124	0.034	-1.94	1.962	0.019	-3.2461	3.502	0.0046	-3.257	3.193
Cluster CID (b_t)	-0.09	-1.5	1.223	0.0697	-1.18	1.402	-0	-2.5	2.4	0.033	-2.42	2.411	0.495	-1.35	2.456	0.026	-2.5985	2.536	-0.065	-2.365	2.627

Table 6: The 95% confidence intervals of the random effect parameters of the educational sample. The estimates (Est.) are the mediums of the bootstrap samples. sdcor1(σ_1): random intercept effect for student; sdcor3(σ_3): random slope effect for student; sdcor4: random intercept effect for classroom; *sdcor2* (σ_2)=$\sigma_{13}/(\sigma_1 * \sigma_3)$, which is the correlation. sdcor: residual term.

Methods	Random effects Parameter (95% CI)of the Educational Sample														
	sdcor1			sdcor2			sdcor3			sdcor4			sdcor		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
No bootstrap(Profile)	6.634	5.963	7.358	-0.271	-0.412	-0.109	0.78	0.691	0.87	3.016	1.455	4.192	2.8476	2.724	2.9749
Residual (Perc)	6.263	5.939	6.571	-0.178	-0.278	-0.076	0.679	0.617	0.735	3.072	2.669	3.423	2.842	2.723	2.9692
Parametric (Perc)	6.616	5.919	7.334	-0.27	-0.404	-0.118	0.776	0.686	0.862	2.914	1.498	4.067	2.844	2.721	2.964
Cluster TID (Perc)	6.629	5.839	7.358	-0.267	-0.412	-0.104	0.773	0.684	0.861	2.86	0.734	4.12	2.82	2.517	3.1678
Cluster CID (Perc)	5.769	4.932	6.424	-0.375	-0.533	-0.189	0.776	0.69	0.864	4.82	3.805	5.658	2.826	2.56	3.0977
Residual (Bca)	6.263	6.686	6.704	-0.178	-0.338	-0.263	0.679	NaN	NaN	3.072	2.544	3.318	2.842	2.739	2.9837
Parametric (Bca)	6.616	6.006	7.373	-0.27	-0.4	-0.108	0.776	0.692	0.866	2.914	1.745	4.146	2.844	2.729	2.9731
Cluster TID (Bca)	6.629	5.844	7.365	-0.267	-0.418	-0.119	0.773	0.69	0.874	2.86	1.476	4.312	2.82	2.562	3.1945
Cluster CID (Bca)	5.769	6.799	6.818	-0.375	-0.338	-0.065	0.776	0.696	0.872	4.82	3.001	3.002	2.826	2.593	3.1388

Table 7: Table of Abbreviations			
Definition	**grp**	**var1**	**var2**
sdcor1	**TID:CID**	**(Intercept)**	**<NA>**
sdcor2	**TID:CID**	**(Intercept)**	**time**
sdcor3	**TID:CID**	**time**	**<NA>**
sdcor4	**TID**	**(Intercept)**	**<NA>**
sdcor	**Residual**	**<NA>**	**<NA>**

The bootstrap bias and bootstrap standard errors are shown in Table 8 and Table 9. Since we don't know the true parameter value, the bias of the 'no bootstrap' cannot be calculated. Also, since the sampling distribution of variance estimates is in general strongly asymmetric, the standard error is not a good characterization of the uncertainty. Here we don't provide the Wald variance-covariance matrix of the variance-covariance parameters themselves. In accordance with the confidence interval results, the cluster bootstrap (CID) method gives the biggest bias (1.779) of sdcor4 (random intercept effect for classroom).

7. Conclusion

In this study, we compare several different bootstrap methods (parametric bootstrap, parametric residual bootstrap and clustering bootstrap) for clustered, hierarchical or multi-level data. The Monte Carlo simulations demonstrate that the parametric bootstrap and clustering bootstrap (at classroom level) perform better estimation than the residual bootstrap and cluster bootstrap (at student level). The application result shows that different bootstrap methods do make some differences on the final estimates, however different confidence interval calculation methods (percentile, Bca, bootstrap-t) give us similar results corresponding to each bootstrap method. The cluster bootstrap provides a very simple resampling scheme comparing with the more complex strategies, for example, random-effects bootstrap or wild bootstrap. It will be interesting to explore the number of clusters effect in our future study.

Table 8: Bootstrap Bias and Bootstrap Standard Errors of the Educational Sample (Fixed effects)														
Methods	(Intercept)		conditn1		time		timesq		matedu_c		conditn1: time		conditn1: timesq	
	SD	bias	SD	bias	SD	bias	SD	bias	SD	bias	SD	bias	SD	bias
No bootstrap	0.8312	NaN	1.1774	NaN	0.21134	NaN	0.0198	NaN	0.3001	NaN	0.28416	NaN	0.0269	NaN
Residual	0.4245	-0.085	0.5544	0.096345	0.20296	0.027441	0.0198	-0.002	0.068	-0.00347	0.2712	-0.0354	0.027	0.0024
Parametric	0.8677	0.0068	1.2257	0.014242	0.22218	0.000258	0.0206	0.0006	0.3002	0.01742	0.28597	-0.003	0.027	-0.0005
Cluster TID	0.7961	0.0207	1.0259	-0.01407	0.3674	0.00349	0.0329	-6E-04	0.2965	0.00353	0.49578	-0.0082	0.0449	0.0006
Cluster CID	0.7047	-0.089	0.9488	0.113648	0.27078	0.0012	0.0248	4E-05	0.2965	0.16346	0.36344	0.00198	0.0336	-0.0002

Table 9: Bootstrap Bias and Bootstrap Standard Errors of the Educational Sample (Random Effects)										
Methods	sdcor1		sdcor2		sdcor3		sdcor4		sdcor	
	SD	bias	SD	bias	SD	bias	SD	bias	SD	bias
No bootstrap	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Residual	0.15982	-0.3736	0.0511	0.0913	0.0302	-0.1001	0.1827	0.0527	0.0643	-0.0062
Parametric	0.3713	-0.0182	0.074	0.0003	0.0445	-0.0032	0.6791	-0.1026	0.0637	-0.0036
Cluster TID	0.39371	-0.0018	0.0765	0.0054	0.0471	-0.0064	0.8257	-0.2251	0.1605	-0.0218
Cluster CID	0.383	-0.903	0.0867	-0.1008	0.0446	-0.0024	0.4713	1.7791	0.1393	-0.0158

Acknowledgements

This research was supported by the U.S. Department of Education, Institute of Education Sciences Grant R324A110048. The opinions expressed in this presentation/article are those of the authors and no official endorsement by the IES should be inferred.

References

A.Colin Cameron, Jonah B.Gelbach and Douglas L.Miller. 2000. "Bootstrap-based improvements for inference with clustered errors." The review of economics and statistics 90 (No.3): 414-427.

B. Efron. 1979. "Bootstrap Methods: Another Look at The Jackknife." The annals of statistics 7 (No.1): 1-26.

B. Efron, R. Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." Statistical Science 1 (1): 54-75.

Bates, Douglas M. 2010. lme4: Mixed-effects modeling with R. Springer.

C.A.Field, A.H.Welsh. 2007. "Bootstrapping clustered data." J.R.Statist.Soc.B part3: 369-390.

C.A.Field, Zhen PANG and Alan H. WELSH. 2008. "Bootstrapping data with multiple levels of variation." The Canadian Journal of Statistics 521-539.

Carpenter, James R. 2003. "A novel bootstrap procedure for assessing the relationship between class size and achievement." Apl.Statist 52 (Part 4): 431-443.

D.Boos, Dennis. 2003. "Introduction to the Bootstrap World." Statistical Science 18 (No.2): 168-174.

S.Gray, M.J.Wilcox and. 2011. "Efficacy of the tell language and literacy curriculum for preschoolers with developmental speech and/or language impairment." Early Childhood Research Quarterly 26: 278-294.