Simple bootstrap for linear mixed effects under model misspecification

Katarzyna Reluga* and Stefan Sperlich†

Abstract

Linear mixed effects are considered excellent predictors of cluster-level parameters in various domains. However, previous work has shown that their performance can be seriously affected by departures from modelling assumptions. Since the latter are common in applied studies, there is a need for inferential methods which are to certain extent robust to misspecifications, but at the same time simple enough to be appealing for practitioners. We construct statistical tools for cluster-wise and simultaneous inference for mixed effects under model misspecification using straightforward semi-parametric random effect bootstrap. In our theoretical analysis, we show that our methods are asymptotically consistent under general regularity conditions. In simulations our intervals were robust to severe departures from model assumptions and performed better than their competitors in terms of empirical coverage probability.

Keywords: linear mixed model; mixed effect; robust inference; small area estimation; simultaneous interval.

*Division of Biostatistics, School of Public Health, University of California, Berkeley, U.S.A. E-mail: katarzyna.reluga@berkeley.edu.
†Stefan Sperlich is a Professor at the Geneva School of Economics and Management, University of Geneva, Switzerland. E-mail: stefan.sperlich@unige.ch.

The authors gratefully acknowledge support from the Swiss National Science Foundation for the projects 200021-192345 and P2GEP2-195898.

We would like to thank D. Flores, W. Gonzalez Manteiga, E. López-Vizcaíno, D. Morales, T. Schmid, N. Salvati and S. Ranjbar for helpful discussions.
1 Introduction

Linear mixed models are frequently used for modelling hierarchical and longitudinal data. Within this modelling framework, population parameters are represented using fixed regression parameters, whereas the extra between-cluster variation is captured by cluster-specific random effects. We consider bootstrap methods for statistically valid inference for mixed effects which are linear combinations of fixed and random effects. Mixed effects are considered excellent predictors of cluster-level parameters in various domains, e.g. small area estimation, ecology or medicine (cf. monographs of Verbeke and Molenberghs [2000], Jiang [2007], Rao and Molina [2015]).

Further inference on mixed parameters heavily depends on model and distributional assumptions. Bootstrap methods have been introduced to partially relax this reliance and approximate in a flexible way functions of the estimators and predictors. Although they could be derived analytically using model-dependent large sample theory, the application of the latter often leads to inaccurate results in finite samples, and it is typically not robust to model misspecifications (cf., Chatterjee et al. [2008], Reluga et al. [2021b]).

The family of bootstrap methods for clustered data is rich, and the extensive reviews are provided by Field and Welsh [2007], Chambers and Chandra [2013] and more recently Flores-Agreda and Cantoni [2019]. All essential procedures can be classified into three broad categories: bootstrapping by resampling clusters and observations within clusters (Davison and Hinkley [1997], McCullagh [2000]), bootstrapping by random weighting of estimating equations (Field et al. [2010], Samanta and Welsh [2013], O'Shaughnessy and Welsh [2018]), and bootstrapping by resampling predictors of random effects and/or residuals (Davison and Hinkley [1997]). The latter is referred to as a random effect bootstrap and can be further subcategorized into parametric versions (Butar and Lahiri [2003], Hall and Maiti [2006a, 2006b], Chatterjee et al. [2008]) and semiparametric versions (Carpenter et al. [2003], Hall and Maiti [2006a], Lombardía and Sperlich [2008], Opsomer et al. [2008]). Regardless of the category they belong to, the main goal of all bootstrap schemes is to construct the empirical estimates which faithfully reproduce some features of the true data generating mechanism. There exist a range of criteria to evaluate the quality of bootstrap schemes for clustered data. In the context of inference for mixed parameters, the existing literature focuses on bootstrap estimation of the mean squared error which boils down to the accurate approximation of the first few moments (see, e.g. Butar and Lahiri [2003], Hall and Maiti [2006a, 2006b], Chatterjee et al. [2008]). In our work, we assess the ability of bootstrap methods to reproduce cumulative distribution functions of some continuous functions of mixed effects which are used in the subsequent steps of statistical inference. At this place we need to emphasize that our goal is not to compare the performance of all existing procedures to select an optimal scheme with respect to a predefined criterion. Even though such a
comparison in the context of mixed effects has not been attempted yet and it could be an interesting direction for further research, it requires a careful definition of the optimality criterion which is beyond the scope of this manuscript.

In this article, we construct statistical tools for cluster-wise and simultaneous inference for mixed parameters under model misspecification using simple, semiparametric random effect bootstrap as in Carpenter et al. (2003) and Opsomer et al. (2008). We show that our bootstrap scheme successfully reproduces cumulative distribution functions of studentized and maximal statistics which are the core elements of our inferential tools. We thus generalize the work of Reluga et al. (2021b) who develop inferential tools for linear mixed effect once the modelling assumptions are satisfied. Our theory applies to the construction of intervals and testing procedure. In our analysis, we show that our methods are asymptotically consistent under general regularity conditions. In simulations our intervals were robust to severe departures from model assumptions and performed better than their competitors in terms of empirical coverage probability. Our bootstrap-based inference is complementary to other techniques handling model misspecifications and dealing with outliers, such as robust inference (Chambers and Tzavidis, 2006; Sinha and Rao, 2009) or estimation using data transformation (Rojas-Perilla et al., 2020).

2 Inference on linear mixed effects

Consider a response vector \( y \in \mathbb{R} \) modelled by \( y = X\beta + Zu + e \) where \( X \in \mathbb{R}^{n \times (p+1)} \), \( Z \in \mathbb{R}^{n_j \times q} \) are known full column rank design matrices for fixed and random effects, vector \( \beta \in \mathbb{R}^{p+1} \) contains fixed effects, whereas random effects \( u \in \mathbb{R}^{q} \) and errors \( e \in \mathbb{R}^{n} \) are assumed to be mutually independent and identically distributed with \( \text{var}(e) = G \) and \( \text{var}(u) = R \). We focus on the model of Laird and Ware (1982)

\[
y_j = X_j\beta + Z_j u_j + e_j, \quad j = 1, \ldots, m,
\]

where \( y_j \in \mathbb{R}^{n_j} \), \( X_j \in \mathbb{R}^{n_j \times (p+1)} \), \( Z_j \in \mathbb{R}^{n_j \times q_j} \), \( e = (e_1, e_2, \ldots, e_m)^T \), \( u = (u_1, u_2, \ldots, u_m)^T \). We denote the total sample size with \( n \), the number of clusters with \( m \) and \( n = \sum_{j=1}^{m} n_j \) where \( n_j \) is the number of observations in the \( j \)th cluster. Furthermore, \( G \) and \( R \) are block-diagonal with blocks \( G_j = G_j(\delta) \in \mathbb{R}^{q_j \times q_j} \) and \( R_j = R_j(\delta) \in \mathbb{R}^{n_j \times n_j} \) which depend on variance parameters \( \delta = (\delta_1, \ldots, \delta_h)^T \). Let \( E(y) = X\beta \) and \( \text{var}(y) = V = R + ZGZ^T \) where \( V \) is a block-diagonal with blocks \( V_j = R_j + Z_jGZ_j^T \). Under normality of random effects and errors, \( y_j \sim N(X_j\beta, V_j) \) and \( y_j|u_j \sim N(X_j\beta + Z_ju, G_j) \). The methods of maximum likelihood and restricted maximum likelihood are often used to obtain an estimator \( \hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_h)^T \) (see, for example, Verbeke and Molenberghs, 2000, Chapter 5). In contrast, \( \beta \) and \( u \) are estimated and predicted using two-stage techniques. In particular, in the
first stage one can use maximum likelihood, estimating equations of [Henderson (1950)] or h-likelihood of [Lee and Nelder (1996)] to obtain best unbiased linear estimator $\hat{\beta} = \beta(\delta) = (X^tV^{-1}X)^{-1}X^tV^{-1}y$ and the best unbiased linear predictor $\hat{u}_j = u_j(\delta) = G_jZ_j^tV^{-1}(y_j - X_j\hat{\beta})$. In the second stage, we replace $\delta$ with $\hat{\delta}$ which results in empirical best unbiased linear estimator $\hat{\beta} = \beta(\hat{\delta})$, and empirical best unbiased linear predictor $\hat{u}_j = u_j(\hat{\delta})$. Our goal is to develop valid inferential tools for general cluster-level parameters $\theta_j = k_j^T\beta + l_j^T u_j, j = 1, \ldots, m,$ (2)

Individual confidence interval $I_{j,\alpha}$ at $1 - \alpha$-level for $\theta_j$ in (2) is a region which satisfies $P(\theta_j \in I_{j,\alpha}) = 1 - \alpha$. To construct $I_{j,\alpha}$, it is enough to find a critical value which is a high quantile from the distributions of statistic $t_j$, that is $q_{j,\alpha} = \inf\{a \in \mathbb{R} : P(t_j \leq a) \geq 1 - \alpha}\}$. We can use a similar strategy to construct simultaneous confidence intervals $I_{\alpha}$ at $1 - \alpha$-level which satisfies $P(\theta_j \in I_{\alpha} \forall j \in [m]) = 1 - \alpha$, where $[m] = \{1, \ldots, m\}$. Let $q_\alpha = \inf\{a \in \mathbb{R} : P(M \leq a) \geq 1 - \alpha\}$ be a high quantile from the distribution of statistic $M$. We thus have

$I_{j,\alpha} : \{\hat{\theta}_j \pm q_{j,\alpha} \times \hat{\sigma}_j\}, \quad I_{\alpha} = \bigtimes_{j=1}^m I_{j,\alpha}^s, \quad I_{s,\alpha}^s : \{\hat{\theta}_j \pm q_\alpha \times \hat{\sigma}_j\},$ (4)

and it follows that $I_{\alpha}$ covers all mixed effects with probability $1 - \alpha$ (see [Reluga et al., 2021a,b] for more details on the importance of maximal statistic in the simultaneous inference for mixed parameters). Due to the central limit theorem, $q_{j,\alpha}$ is often replaced by a high quantile from the standard normal distribution or the Student’s t-distribution. The relation between confidence intervals and hypothesis testing allows us to define modified statistics $t_j$ and $M$ that can be used to carry out hypothesis testing. More specifically, let
A ∈ ℝ^{m'×m}, θ_H = (θ_{H_1}, θ_{H_2}, \ldots, θ_{H_m}) = Aθ ∈ ℝ^{m'} and c = (c_1, c_2, \ldots, c_{m'}) = ℝ^{m'} be a vector of some constants, with m' ≤ m. Then consider the testing hypotheses

\[ H_0_j : θ_{H_j} = c_j \quad \text{vs.} \quad H_1 : θ_{H_j} ≠ c_j, \quad \text{(individual test)}, \]
\[ H_0 : θ_H = c \quad \text{vs.} \quad H_1 : θ_H ≠ c \quad \text{(multiple test)}. \]

To obtain test statistics for tests in (5) and (6), we need to simply replace \( θ_j \) by \( c_j \) in the definition of test statistic \( t_j \) in (3), that is

\[ t_{H_j} = \frac{\hat{θ}_{H_j} - c_j}{\hat{σ}_j}, \quad t_{H_0j} = \frac{\hat{θ}_{H_j} - θ_{H_j}}{\hat{σ}_j}, \quad M_H = \max_{j=1,\ldots,m'} |t_{H_j}|, \quad M_{H_0} = \max_{j=1,\ldots,m'} |t_{H_0j}|, \]

where \( t_{H_0j} \) and \( M_{H_0} \) are for retrieving the critical values. Tests using statistics \( t_{H_j} \) and \( M_H \) reject \( H_0j \) and \( H_0 \) at the \( α \)-level if \( t_{H_j} ≥ q_{H_0j,α} \) and \( M_H ≥ q_{H_0,α} \) where \( q_{H_0j,α} = \inf\{a ∈ ℝ : P(t_{H_j} ≤ a) ≥ 1 - α\} \) and \( q_{H_0,α} = \inf\{a ∈ ℝ : P(M_{H_0} ≤ a) ≥ 1 - α\} \).

Construction of the studentized statistics in (3) requires the estimation of \( \hat{σ}_j^2 \). The most common measure to assess the variability of the prediction is the mean squared error \( \text{MSE}(\hat{θ}_j) = E(\hat{θ}_j - θ_j)^2 \), where \( E \) denotes the expectation with respect to model (1). Nevertheless, following Chatterjee et al. (2008), a simpler choice of \( \hat{σ}_j^2 = l_j(G_j - G_jZ_jV_j^{-1}Z_jG_j)l_j \) which accounts for the variability of \( θ_j \) without accounting for the estimation of \( β \) or \( δ \) lead to the most satisfactory numerical results. Simulation results showing finite sample performance of the intervals constructed using other variability estimators can be found in our Supplementary Material.

3 Inference robust to misspecifications by semiparametric bootstrap

We present a bootstrap scheme to construct individual and simultaneous intervals which are robust to model misspecifications. Denote bootstrap generated observations by

\[ y^* = X\hat{β} + Zu^* + e^*, \]

where \( e^* \) and \( u^* \) are bootstrap replica of the random components in the model. We further set \( δ^* = \hat{δ}, V^* = \hat{V}, G^* = \hat{G} \) and define \( \hat{β}^* = β(δ^*) = (X'V^{-1}\bar{X})^{-1}X'V^{-1}y^*, \hat{u}_j^* = u_j(δ^*) = G_jZ_jV_j^{-1}(y_j^* - X_j\bar{β}^*). \) In addition, let \( \hat{δ}^* \) be an estimated version of \( δ^* \) obtained by regressing \( y^* \) on \( X \). Then we have \( \hat{β}^* = β(\hat{δ}^*) \) and \( \hat{u}_j^* = u_j(\hat{δ}^*). \) Bootstrap mixed effects are thus defined as

\[ θ_j^* = k_j^Tβ^* + l_j^T u_j^*, \quad \hat{θ}_j^* = θ_j(δ^*) = k_j^T\hat{β}^* + l_j^T \hat{u}_j^*, \quad \hat{θ}_j^* = θ_j(\hat{δ}^*) = k_j^T\hat{β}^* + l_j^T \hat{u}_j^*. \]
The bootstrap versions of the statistics of interest in (3) are given by
\[ t_j^* = \frac{\hat{\theta}_j^* - \theta_j^*}{\hat{\sigma}_j^*}, \quad M^* = \max_{j=1,\ldots,m} |t_j^*|. \] (8)

We use statistics in (8) to construct bootstrap equivalents of intervals in (4), that is
\[ q_{j,\alpha}^* = \inf \{ a \in \mathbb{R} : P(t_j^* \leq a) \geq 1 - \alpha \}, \quad I_{j,\alpha}^* = \{ \hat{\theta}_j \pm q_{j,\alpha}^* \times \hat{\sigma}_j \}, \quad j = 1, \ldots, m, \] (9)
\[ q_{\alpha}^* = \inf \{ a \in \mathbb{R} : P(M^* \leq a) \geq 1 - \alpha \}, \quad I_{\alpha}^* = \bigcap_{j=1}^m I_{j,\alpha}^{*s} = \{ \hat{\theta}_j \pm q_{\alpha}^* \times \hat{\sigma}_j \}. \] (10)

The most popular choice is to use a parametric bootstrap and draw \( e^* \) and \( u^* \) from a postulated normal distribution with estimated variance parameters. In contrast, we use a semiparametric bootstrap method introduced by Carpenter et al. (2003) and generalised by Opsomer et al. (2008). The empirical performance of this bootstrap scheme for fixed parameters has been studied by Chambers and Chandra (2013). The goal is to mimic the data generating process in model (1). Before writing down explicitly the bootstrap algorithm, we provide some motivation behind it. Let \( \tilde{y} = X\hat{\beta} = X(X^TV^{-1}X)^{-1}X^TV^{-1}y = Hy, \quad \tilde{e} = y - X\hat{\beta} - Z\hat{u} = (I - ZGZ^T)^{-1}(I - H)y = RV^{-1}(I - H)y \) and \( \hat{e} = y - X\hat{\beta} - Z\hat{u}. \) Then, by some algebraic transformations we have \( I - ZGZ^T = RV^{-1}, \) which leads to \( \text{var}(\tilde{u}) = GZ^T\{V^{-1}(I - H)\}ZG \) and \( \text{var}(\hat{e}) = R\{V^{-1}(I - H)\}R. \) Thus, we should re-scale \( \hat{e} \) and \( \hat{u} \) before sampling with replacement to avoid the effects of shrinkage (Morris, 2002). Centring, that is subtracting the empirical mean, is also advisable to assure that the empirical re-scaled residuals have mean zero. This suggests sampling from \( \hat{e}_{sc} \) and \( \hat{u}_{sc} \) defined as follows
\[ \hat{e}_{sc} = \hat{e}_s - \bar{e}_s, \quad \bar{e}_s = \frac{\sum_{i=1}^n \hat{e}_{si}}{n}, \quad \hat{e}_s = [R\{V^{-1}(I - H)\}]^{-1/2}\hat{e}, \]
\[ \hat{u}_{sc} = \hat{u}_s - \bar{u}_s, \quad \bar{u}_s = \frac{\sum_{i=1}^n \hat{u}_{sj}}{m}, \quad \hat{u}_s = [GZ^T\{V^{-1}(I - H)\}Z]^{-1/2}\hat{u}. \]

The algorithm to obtain bootstrap quantiles and construct intervals in (9) and (10) is:

**A semiparametric random effects bootstrap algorithm**

1. Obtain consistent estimators \( \hat{\beta} \) and \( \hat{\delta}. \)

2. For \( b = 1 \) to \( b = B: \)
   
   (a) Obtain vectors \( u^* \in \mathbb{R}^m, e^* \in \mathbb{R}^n \) by sampling independently with replacement from \( \hat{u}_{sc} \) and \( \hat{e}_{sc}. \)
(b) Generate sample \( y^* = X\hat{\beta} + Zu^{*(b)} + e^* \) in (7) and obtain \( \theta_j^* \), \( j = 1, \ldots, m \).

(c) Fit LMM to bootstrap sample from the previous step.

(d) Obtain bootstrap estimates \( \hat{\delta}^* \), \( \hat{\beta}^* \), \( \hat{\theta}_j^* \) and \( M^* \), \( j = 1, \ldots, m \).

3. Estimate critical values \( q_{j, \alpha}^*, q_\alpha^* \) by the \( \left\{ (1 - \alpha) B \right\} + 1 \)th order statistics of \( t_j^* \) and \( M_j^* \), \( j = 1, \ldots, m \).

4. Construct bootstrap intervals as indicated in (9) and (10). Fisher consistency of \( \hat{\delta}^* \) and \( \hat{\beta}^* \) obtained using semiparametric bootstrap in the above algorithm has been proved by Carpenter et al. (2003). In Lemma 1 and 2 we show the consistency of statistics \( t_j^* \) and \( M_j^* \).

**Lemma 1** (Consistency of \( t_j^* \)). Let \( F_{t_j}(a) = P(t_j < a) \), \( F_{t_j}^*(a) = P(t_j^* < a) \) be the cumulative distribution functions of statistics \( t_j \), \( t_j^* \) defined in (3) and (8). If the regularity conditions in Appendix 1 are satisfied, then we have in probability

\[
\sup_{a \in \mathbb{R}} \left| F_{t_j}(a) - F_{t_j}^*(a) \right| \to 0.
\]

**Proof.** Without loss of generality, we assume that the sequence of estimators \( t_j \) converges to a continuous distribution function \( F \). A standard way of proving the consistency of bootstrap procedure in Lemma 1 (see, for example, Van der Vaart, 2000, Chapter 23) is to show that, for every \( a \) \( F_{t_j}(a) \to F(a) \) in distribution and \( F_{t_j}^*(a) \to F(a) \) given the original sample size in probability. Let \( \hat{\theta}^* = (\hat{\beta}^*, \hat{\theta}^*) \) and \( E^* \) be a bootstrap operator of the expected value. Then \( t_j \) and \( t_j^* \) can be written as \( t_j = f(\vartheta, \hat{\vartheta}, u_j) \) and \( t_j^* = f(\vartheta, \hat{\vartheta}^*, u_j^*) \), respectively for a continuous and a differentiable function \( f \). Consider a general score equation \( s_n(\vartheta) \) defined in Appendix and its bootstrap equivalent \( s_n^*(\vartheta) = \sum_{j=1}^m \sum_{i=1}^{n_j} \psi(y_{ij}, \vartheta) \) with \( y \) replaced by \( y^* \). It follows that \( E^*\{s_n^*(\vartheta)\} = 0 \) at \( \vartheta = \hat{\vartheta} \) which yields the consistency of the sequence of bootstrap estimators \( \hat{\vartheta}^* \). The consistency of random effects under random effect bootstrap was proved by Field and Welsh (2007) under Condition 4 in Appendix which is in alignment with results of Jiang (1998). We thus have that \( \sqrt{n}(\hat{\theta}_j - \theta_j^*) \) and \( \sqrt{n}(\hat{\theta}_j - \theta_j) \) converge to the same distribution. Final consistency result follows by Slutsky’s lemma.

Corollary 1 ensures the consistency of the individual confidence intervals.

**Corollary 1** (Consistency of \( I_{j, \alpha}^* \)). Lemma 1 implies that under the same assumptions

\[
P(\theta_j \in I_{j, \alpha}^*) \to 1 - \alpha.
\]
Proof. The proof follows along the same line as Lemma 23.3 in Van der Vaart (2000). By Lemma 1, the sequences of distribution functions $F_{t_j}$ and $F_{t_j^*}$ converge weakly to $F$, which implies the convergence of their quantile functions $F_{t_j}^{-1}$ and $F_{t_j^*}^{-1}$ at every continuity point. We thus conclude that $q_{j,\alpha}^* = F_{t_j^*}^{-1}(1 - \alpha) \to F^{-1}(1 - \alpha)$ almost surely, and

$$P(\theta_j \geq \hat{\theta}_j - \hat{\sigma}_j q_{j,\alpha}^*) = P\left(\frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}_j} \leq q_{j,\alpha}^*\right) \to P\{t_j \leq F^{-1}(1 - \alpha)\} = 1 - \alpha$$

which completes the proof. \[\square\]

The consistency of $M^*$ does not follow from Lemma 1 by the delta method, because max function is not differentiable. Instead, Lemma 2 provides a heuristic proof based on results known from the extreme value theory.

Lemma 2 (Consistency of $M^*$). Let $M$ and $M^*$ be as defined in (3) and (8). If the regularity conditions in Appendix are satisfied and Lemma 1 holds, then we have in probability

$$\sup_{a \in \mathbb{R}} |F_M(a) - F_{M^*}(a)| \to 0.$$

Proof. Observe that $F_M(a) = P(M < a) = P(t_1 \leq a, \ldots, t_m \leq a, -t_1 \leq a, \ldots, -t_m \leq a)$. Since $t_j, j = 1, \ldots, m$ are asymptotically independent and identically distributed, we have an approximation $F_M(a) \approx \prod_{j=1}^{2m} F_j(a)$ with $F_j(a)$ some proper, non-degenerate distributions. By classical results in extreme value theory (Beirlant et al., 2004; Embrechts et al., 2013), we can assume that there exist sequences of re-normalizing constants $\{b_j > 0\}$, $\{c_j\}$ such that $P\{(M_j - c_j)/b_j \leq a\}$ converges to a non-degenerate distribution function $H(a)$ as $j \to \infty$, i.e., the $F_j(a)$ belong to the max-domain of attraction of some non-degenerate, continuous distribution $H(a)$. The consistency of $F_{M^*}(a)$ follows by evoking the properties of the random effects bootstrap and the arguments used in the proof of Lemma 1. \[\square\]

Corollary 2. Lemma 2 implies that under the same assumptions

$$P(\theta_j \in I_{\alpha}^* \forall j \in [m]) \to 1 - \alpha.$$

Proof. The proof follows now along the same lines as in Corollary 1 with statistic $t_j$ replaced by $M$. \[\square\]

Similarly as in case of intervals, we can use semiparametric bootstrap to approximate critical values $q_{H_0,\alpha}$ and $q_{H_0,\alpha}$ for tests in (5) and (6). Thanks to the relation between intervals and test, the consistency proof for intervals applies also for testing procedures with some changes of the notation (cf. Reluga et al., 2021a).
We carry out numerical simulation studies to evaluate finite sample properties of our bootstrap intervals. In all scenarios we generate outcomes from a linear mixed effect model in (1) with a fixed and a random intercept, and a uniformly distributed covariate, that is, we set \(x_{ij1} = 1, \ z_{ij} = 1, \ x_{ij2} \sim U(0, 1)\). We consider three types of sample sizes to mimic joint asymptotics: in setting 1 we have \(m = 25, \ n_j = 5\), in setting 2: \(m = 50, \ n_j = 10\), and in setting 3: \(m = 75, \ n_j = 15\). Furthermore, in each simulation, errors and random effects are drawn from one of the following distributions: standard normal, Student’s t with 6 degrees of freedom, or chi-square with 5 degrees of freedom. The distributions are always centred to zero and re-scaled to variances \(\text{var}(e_{ij})\) and \(\text{var}(u_j)\). We compare the performance of our individual and simultaneous intervals in (9) at the \(\alpha = 0.05\) level obtained using semiparametric bootstrap, parametric bootstrap as in Chatterjee et al. (2008) and Reluga et al. (2021b) as well as intervals constructed using large-sample asymptotic approximations, that is, with a \((1 - \alpha/2)\) and \((1 - \alpha/2m)\) quantiles from normal distributions (the latter by Bonferroni correction). We employ following criteria to assess the performance of intervals: empirical coverage probability for individual and simultaneous intervals, that is, \(\text{Cov}_{\text{ind}} = 1/mS \sum_{j=1}^{m} \sum_{s=1}^{S} 1\{\theta_j^{(s)} \in I_{j,a}^{*}\}\) and \(\text{Cov}_{\text{sim}} = 1/S \sum_{s=1}^{S} 1\{\theta_j^{(s)} \in I_{a}^{*} \forall j \in [m]\}\); average widths of the intervals \(\text{Width} = 1/mS \sum_{j=1}^{m} \sum_{s=1}^{S} \rho_j^{(s)}\) and \(\text{VarWidth} = 1/m(S - 1) \sum_{j=1}^{m} \sum_{s=1}^{S} (\rho_j^{(s)} - \bar{\rho}_j)^2\); all of them over \(S = 1000\) simulation runs, where \(\rho_j^{(s)} = 2q_j^{(s)} \hat{\sigma}_j^{(s)}\), \(\bar{\rho}_j = \sum_{s=1}^{S} \rho_j^{(s)}/S\) and \((\cdot)\) stands for the pair \(j, \alpha\) for individual intervals and for \(\alpha\) for simultaneous intervals.

Table 1 displays the numerical performance of individual intervals for mixed effect \(\theta_j\) in (2). In this case, the performance of all methods seems to be similar – the distribution of errors and random effects does hardly affect the empirical coverage, even for the intervals derived asymptotically. Our simulations indicate a surprisingly strong robustness to distributional misspecifications and the application of bootstrapping seems superfluous in this setting. The situation changes dramatically in Table 2 which shows numerical performance of simultaneous intervals. In this case, the results are similar for all methods only when errors and random effects are normally distributed (cf. results in Reluga et al., 2021b). Regardless of the distribution of errors and/or random effects, the performance of intervals obtained using semiparametric bootstrap is superior to other methods. In fact, their application leads to serious undercoverage even for large sample sizes under departures from normality. Furthermore, the average length of semiparametric bootstrap intervals is not excessively wide in comparison to other methods. We can thus conclude that the ap-
Table 1: Empirical coverage, width and variance of widths of individual intervals at $\alpha = 0.05$ level. $S_1$, Setting 1; $S_2$, Setting 2, $S_3$, Setting 3; M, Method; A, asymptotic; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

The application of our semiparametric bootstrap-based method leads to a satisfactory numerical performance even under considerable departures from normality. In comparison to other robust techniques, it does not involve robust estimation or any data transformation, which is extremely appealing for practitioners.

5 Discussion

Linear mixed effects are popular to predict cluster-level parameters in various domains. Yet, the underlying assumptions which should guarantee their satisfactory numerical performance are often violated in practice. We studied to what extent the application of a simple bootstrapping scheme might mitigate the negative effects of distributional misspecifications without the need to reach for more advanced techniques such as robust estimation or data transformation. Our numerical study confirms that mixed effects are fairly robust to such misspecification unless they undergo complex transformations. This is particularly
Table 2: Empirical coverage, width and variance of widths of simultaneous intervals at \( \alpha = 0.05 \) level. \( S_1 \), Setting 1; \( S_2 \), Setting 2, \( S_3 \), Setting 3; M, Method; A, asymptotic; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

interesting for their application in small area estimation in which mixed effects are often used in nonlinear poverty indicators [Rojas-Perilla et al., 2020] for which the application of semiparametric bootstrap inference could be particularly beneficial.

Appendix 1

Regularity conditions

We adopt some regularity conditions from Shao et al. (2000) and Reluga et al. (2021b). Let \( \vartheta = (\beta, \delta), \hat{\vartheta} = (\hat{\beta}, \hat{\delta}) \) and \( \vartheta_0 \in \Theta \subset \mathbb{R}^{p+n+1} \) be the true parameter value. We assume that

1. Score equation \( s_n(\vartheta) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \psi(y_{ij}, \vartheta) \) is well defined if: (a) \( s_n(\vartheta) \) is continuous and differentiable for each fixed \( y \), (b) \( E\{s_n(\hat{\vartheta})\} = 0 \) at \( \vartheta_0 \), (c) \( \vartheta_0 \) is an interior point of \( \Theta \) and the estimator \( \hat{\vartheta} \) is an interior point of the neighborhood of \( \vartheta_0 \).
2. $\lim \inf \lambda[n^{-1}\text{var}\{s_n(\vartheta)\}] > 0$ and $\lim \inf \lambda[n^{-1}E\{\nabla s_n(\vartheta)\}] > 0$ where $\nabla s_n(\vartheta) = \frac{\partial \psi(\vartheta)}{\partial \vartheta}$ and $\lambda[A]$ indicates the smallest eigenvalue of matrix $A$.

3. There exists $b > 0$ such that $E \|\psi(y_{ij}, \vartheta)^{2+b}\| < \infty$, and $E(h_N(y_{ij})^{1+b})$ in a compact neighbourhood $N$, where $h_C(y_{ij}) = \sup_{\vartheta \in N} \|\nabla s_n(\vartheta)\|$.  

4. Convergence: $m \to \infty$, $n \to \infty$.

5. $V_j(\delta)$ has a linear structure in $\delta$, $j = 1, \ldots, m$.

Conditions 1–3 ensure that one can use the score equation $s_n$ to estimate fixed parameters $\vartheta$ up to a vanishing term. Condition 4 is required to ensure the convergence of mixed effect predictors, whereas Condition 5 implies that the second derivatives of $R_j$ and $G_j$ are 0. The assumption of $m \to \infty$, which is common in small area estimation literature once the modelling assumptions are satisfied (cf. Reluga et al., 2021b, in the context of simultaneous inference), must be replaced by the joint asymptotics in Condition 4 to ensure the convergence of cumulative distributions functions of mixed effects under departures from normality (cf. Jiang, 1998). Nevertheless, this assumption is important only for the theoretical derivations – in practice bootstrap intervals perform well for a sample size as small as $n_j = 5$ (cf., results in Tables 1-2).

Appendix 2

Additional simulation results

In this section, we present additional simulations results using different MSE estimators. Analytical MSE can be decomposed as follows

$$
\text{MSE}(\hat{\theta}_j) = \text{MSE}(\tilde{\theta}_j) + E\left(\hat{\theta}_j - \tilde{\theta}_j\right)^2 + 2E\left\{ (\tilde{\theta}_j - \theta_j)(\hat{\theta}_j - \tilde{\theta}_j) \right\} \\
= g_{1j}(\delta) + g_{2j}(\delta) + g_{3j}(\delta) + 2E\left\{ (\tilde{\theta}_j - \theta_j)(\hat{\theta}_j - \tilde{\theta}_j) \right\},
$$

(11)

where $\text{MSE}(\tilde{\theta}_j)$ accounts for the variability of $\theta_j$ when the variance components $\delta$ are known. It particular, $g_{1j}$ accounts for the variability of $\theta_j$ for known $\beta$, $g_{2j}$ for the estimation of $\beta$, $g_{3j}$ quantifies the square difference between $\hat{\theta}_j$ and $\tilde{\theta}_j$. There exists a vast literature to estimate it (see, for example, Rao and Molina, 2015). The last term in (11) disappears
under normality of errors and random effects. Let $b_j^T = k_j^T - o_j^T X_j$ with $o_j^T = l_j^T G_j Z_j V_j^{-1}$. Under linear mixed model, the analytical estimator of variability $\text{mse}_L(\hat{\theta}_j)$ reduces to

$$\text{mse}_L(\hat{\theta}_j) = g_{1j}(\hat{\delta}) + g_{2j}(\hat{\delta}) + 2g_{3j}(\hat{\delta}),$$

and $g_1$, $g_2$ and $g_3$ are defined in expression (12):

$$g_{1j}(\delta) = l_j^T (G_j - G_j Z_j V_j^{-1} Z_j G_j) l_j,$n$$

$$g_{2j}(\delta) = b_j^T \left( \sum_{j=1}^m X_j^T V_j^{-1} X_j \right)^{-1} b_j,$$

$$g_{3j}(\delta) = \text{tr} \left\{ (\partial o_j^T / \partial \delta) V_j (\partial o_j^T / \partial \delta)^T V_A(\hat{\delta}) \right\},$$

(12)

where $V_A(\hat{\delta})$ the asymptotic covariance matrix. In addition, $E \{ \text{mse}_L(\hat{\theta}_j) \} = \text{MSE}(\theta_j) + o(m^{-1})$. First, we complete the numerical results from Section 4 by considering additional simulation scenarios. Tables 3-4 show the numerical results with $\hat{\sigma}_j^2 = g_{1j}$.

| $e_{ij}$ | $u_j$ | M | S1 | S2 | S3 | S1 | S2 | S3 | S1 | S2 | S3 |
|----------|-------|---|----|----|----|----|----|----|----|----|----|
| N(0.5)   | N(1)  | A | 948 | 949 | 949 | 1191 | 856 | 705 | 6 | 1 | 1 |
| P        |       | S | 948 | 948 | 947 | 1201 | 855 | 702 | 7 | 1 | 1 |
| A        |       | P | 949 | 948 | 947 | 1202 | 855 | 702 | 7 | 1 | 1 |
| S        |       | A | 952 | 949 | 950 | 1529 | 1137 | 952 | 16 | 3 | 3 |
| t_6(1)   | t_5(0.5) | S | 953 | 948 | 949 | 1552 | 1138 | 951 | 20 | 4 | 4 |
| P        |       | P | 963 | 947 | 951 | 1627 | 1133 | 948 | 62 | 4 | 4 |
| A        |       | A | 949 | 949 | 950 | 1614 | 1183 | 981 | 22 | 3 | 3 |
| t_6(0.5) | t_5(0.5) | S | 949 | 946 | 946 | 1633 | 1181 | 977 | 26 | 4 | 4 |
| P        |       | P | 951 | 951 | 949 | 1647 | 1182 | 978 | 27 | 4 | 4 |
| A        |       | A | 949 | 949 | 950 | 1614 | 1183 | 981 | 22 | 3 | 3 |
| $\chi_5(1)$ | $\chi_5(0.5)$ | S | 949 | 946 | 946 | 1633 | 1181 | 977 | 26 | 4 | 4 |
| P        |       | P | 951 | 951 | 949 | 1647 | 1182 | 978 | 27 | 4 | 4 |
| A        |       | A | 949 | 949 | 950 | 1614 | 1183 | 981 | 22 | 3 | 3 |

Table 3: Empirical coverage, width and variance of widths of individual intervals at $\alpha = 0.05$-level, $\sigma_j^2 = g_{1j}$. $S_1$, Setting 1; $S_2$, Setting 2; $S_3$, Setting 3; M, Method; A, asymptotic; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.
Coverage Length Variance of length
\( e_{ij} \) \( u_j \) \( M \ S_1 \ S_2 \ S_3 \ S_1 \ S_2 \ S_3 \ S_1 \ S_2 \ S_3 \)

|            | Coverage | Length | Variance of length |
|------------|----------|--------|---------------------|
| \( N(0.5) \) | \( N(1) \) | A 932 | 955 | 944 | 1879 | 1438 | 1224 | 14 | 2 | 1 |
| \( t_6(1) \) | \( t_6(0.5) \) | S 945 | 956 | 947 | 1919 | 1443 | 1225 | 17 | 2 | 1 |
| \( \chi^2_5(1) \) | \( \chi^2_5(0.5) \) | P 946 | 961 | 946 | 1924 | 1445 | 1225 | 16 | 2 | 1 |
| \( t_6(0.5) \) | \( \chi_5(1) \) | A 916 | 926 | 920 | 2411 | 1909 | 1653 | 40 | 8 | 3 |

Table 4: Empirical coverage, width and variance of widths of simultaneous intervals at \( \alpha = 0.05 \)-level, \( \sigma_j^2 = g_{1j} \). \( S_1 \), Setting 1; \( S_2 \), Setting 2, \( S_3 \), Setting 3; \( M \), Method; \( A \), asymptotic; \( S \), semiparametric bootstrap; \( P \), parametric bootstrap. All numerical entries are multiplied by 1000.

Alternatively, one could estimate MSE using bootstrap. The most straightforward bootstrap estimator is \( \text{MSE}^\ast(\hat{\theta}^\ast_j) = E^\ast\left(\hat{\theta}^\ast_j - \theta^\ast_j\right)^2 \) which might be approximated by

\[
\text{MSE}^\ast_{B_1}(\hat{\theta}^\ast_j) \approx \text{mse}^\ast_{B_2}(\hat{\theta}_j) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}^\ast_{j(b)} - \theta^\ast_{j(b)}\right)^2 ,
\]

(13)

and \( \hat{\theta}^\ast_{j(b)} \), \( \theta^\ast_{j(b)} \) as defined in Section 3, calculated from the \( b^{th} \) bootstrap sample. Tables 5-6 display the performance of individual and simultaneous intervals constructed using \( \text{MSE}^\ast_{B_1} \). As we can see, a general trend is the same as in case of \( \sigma_j^2 = g_{1j} \), that is there is not much different between the performance of parametric and semiparametric bootstrap individual intervals, but this changes dramatically if we consider simultaneous intervals.

We can define several other bootstrap estimators. For example, \( \text{MSE}^\ast_{3T} \) directly approximates each term in (11) by bootstrap, that is

\[
\text{MSE}^\ast_{3T}(\hat{\theta}^\ast_j) = \text{MSE}^\ast_{B_2}(\hat{\theta}_j) + E^\ast\left(\hat{\theta}^\ast_j - \tilde{\theta}^\ast_j\right)^2 + 2E^\ast\left\{\left(\hat{\theta}^\ast_j - \theta^\ast_j\right)\left(\tilde{\theta}^\ast_j - \hat{\theta}^\ast_j\right)\right\} .
\]

(14)

Tables 7-8 display the performance of individual and simultaneous intervals constructed using \( \text{MSE}^\ast_{3T} \).
Table 5: Empirical coverage, width and variance of widths of individual intervals at $\alpha = 0.05$-level, $\sigma_j^2 = \text{MSE}^*_B(\hat{\theta}_j)$. $S_1$, Setting 1; $S_2$, Setting 2, $S_3$, Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

It is well known that $\text{MSE}^*_B$ leads to estimators with bias of order $O(m^{-1})$. To obtain a bias of order $o(m^{-1})$, Butar and Lahiri (2003) advocate approximating only intractable terms in (11) by bootstrap. Specifically, with $g_1(d(\cdot))$ and $g_2(d(\cdot))$ as defined in (12), one takes

$$\text{MSE}^*_{SPA}(\hat{\theta}_j) = 2 \left\{ g_{1j}(\hat{\delta}) + g_{2j}(\hat{\delta}) \right\} - E^* \left\{ g_{1j}(\hat{\delta}^*) + g_{2j}(\hat{\delta}^*) \right\} + E^* \left( \hat{\theta}_j^* - \tilde{\theta}_j^* \right)^2$$

$$+ 2E^* \left\{ (\hat{\theta}_j^* - \theta_j^*)(\hat{\delta}_j^* - \tilde{\delta}_j^*) \right\} ,$$

where the last term is zero under normality. Tables 9-10 display the performance of individual and simultaneous intervals constructed using $\text{MSE}^*_{SPA}$. In contrast, Hall and Maiti (2006a) propose a bias reduction with the aid of a double-bootstrap $\text{MSE}^*_{B_2}(\hat{\theta}_j) = E^* (\hat{\theta}_j - \hat{\theta}_j^*)$. In this bootstrapping scheme, for each sample $b$ we must generate $c =$

| $e_{ij}$ | $u_j$ | Coverage | Length | Variance of length |
|---------|-------|----------|--------|-------------------|
|         |       | $S_1$    | $S_2$  | $S_3$  | $S_1$ | $S_2$ | $S_3$ |
| $N(0.5)$ | $N(1)$ | S 943 947 947 1181 852 701 7 1 1 1 |        |        |        |        |
|         |       | P 944 947 946 1181 852 701 7 1 1 1 |        |        |        |        |
| $N(1)$ | $N(0.5)$ | S 946 946 947 1521 1129 948 15 3 1 1 |        |        |        |        |
|         |       | P 939 946 946 1488 1128 947 15 3 1 1 |        |        |        |        |
| $t_6(0.5)$ | $t_6(1)$ | S 943 946 948 1175 853 702 14 2 1 1 |        |        |        |        |
|         |       | P 943 945 948 1170 849 700 13 2 1 1 |        |        |        |        |
| $t_6(1)$ | $t_6(0.5)$ | S 946 945 948 1501 1126 947 18 4 2 1 |        |        |        |        |
|         |       | P 954 949 950 1560 1137 952 21 3 1 1 |        |        |        |        |
| $\chi_5(0.5)$ | $\chi_5(1)$ | S 945 948 949 1201 861 706 15 2 1 1 |        |        |        |        |
|         |       | P 945 944 945 1590 973 74 24 4 2 1 |        |        |        |        |
| $\chi_5(1)$ | $\chi_5(0.5)$ | S 943 944 945 1590 973 74 24 4 2 1 |        |        |        |        |
|         |       | P 943 948 948 1591 976 74 24 4 2 1 |        |        |        |        |
| $t_6(0.5)$ | $\chi_5(1)$ | S 940 944 945 1173 699 14 2 1 1 |        |        |        |        |
|         |       | P 946 949 948 1172 849 701 13 2 1 1 |        |        |        |        |
| $\chi_5(0.5)$ | $t_6(1)$ | S 943 946 947 1591 978 706 24 4 1 1 |        |        |        |        |
|         |       | P 943 947 826 1588 1174 1534 24 5 2 1 |        |        |        |        |
| $t_6(1)$ | $\chi_5(0.5)$ | S 943 944 945 1590 973 74 24 4 1 1 |        |        |        |        |

It is well known that $\text{MSE}^*_B$ leads to estimators with bias of order $O(m^{-1})$. To obtain a bias of order $o(m^{-1})$, Butar and Lahiri (2003) advocate approximating only intractable terms in (11) by bootstrap. Specifically, with $g_1(d(\cdot))$ and $g_2(d(\cdot))$ as defined in (12), one takes

$$\text{MSE}^*_{SPA}(\hat{\theta}_j) = 2 \left\{ g_{1j}(\hat{\delta}) + g_{2j}(\hat{\delta}) \right\} - E^* \left\{ g_{1j}(\hat{\delta}^*) + g_{2j}(\hat{\delta}^*) \right\} + E^* \left( \hat{\theta}_j^* - \tilde{\theta}_j^* \right)^2$$

$$+ 2E^* \left\{ (\hat{\theta}_j^* - \hat{\theta}_j)(\hat{\delta}_j^* - \tilde{\delta}_j^*) \right\} ,$$
Table 6: Empirical coverage, width and variance of widths of simultaneous intervals at \( \alpha = 0.05 \)-level, \( \sigma_j^2 = \text{MSE}_{B1}^*(\hat{\theta}_j^*) \). \( S_1 \), Setting 1; \( S_2 \), Setting 2, \( S_3 \), Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

1, ..., \( C \) bootstrap samples (in practice, \( C = 1 \) works quite well), where

\[
\theta_j^{**} = k_j^T \beta^{**} + l_j^T u_j^{**}, \quad \tilde{\theta}_j^{**} = \theta_j(\hat{\delta}^{**}) = k_j^T \tilde{\beta}^{**} + l_j^T \tilde{u}_j^{**}, \quad \hat{\theta}_j^{**} = \theta_j(\hat{\delta}^{**}) = k_j^T \hat{\beta}^{**} + l_j^T \hat{u}_j^{**}.
\]

We can thus consider double bootstrap bias-corrected MSE estimator which is defined as follows

\[
\text{MSE}_{BC}^*(\hat{\theta}_j^*) = 2\text{MSE}_{B1}^*(\hat{\theta}_j^*) - \text{MSE}_{B2}^{**}(\hat{\theta}_j^{**}) .
\]

Tables [1][2] display the performance of individual and simultaneous intervals constructed using \( \text{MSE}_{BC}^* \).

To sum up, the performance of individual and simultaneous intervals is not strongly affected by the choice of the estimator of \( \sigma_j^2 \). The most important factors in the performance
| \(e_{ij}\)   | \(u_j\) | \(\text{Coverage}\) | \(\text{Length}\) | \(\text{Variance of length}\) |
|---|---|---|---|---|
| | | \(S_1\) | \(S_2\) | \(S_3\) | \(S_1\) | \(S_2\) | \(S_3\) | \(S_1\) | \(S_2\) | \(S_3\) |
| \(N(0.5)\) | \(N(1)\) | S | 943 | 947 | 947 | 1181 | 852 | 701 | 7 | 1 | 1 |
| | | P | 944 | 947 | 946 | 1181 | 852 | 701 | 7 | 1 | 1 |
| \(N(1)\) | \(N(0.5)\) | S | 946 | 946 | 947 | 1521 | 1129 | 948 | 12 | 3 | 1 |
| | | P | 939 | 946 | 946 | 1488 | 1128 | 947 | 15 | 3 | 1 |
| \(t_6(0.5)\) | \(t_6(1)\) | S | 943 | 946 | 948 | 1175 | 853 | 702 | 14 | 2 | 1 |
| | | P | 943 | 945 | 947 | 1170 | 849 | 700 | 13 | 2 | 1 |
| \(t_6(1)\) | \(t_6(0.5)\) | S | 946 | 945 | 948 | 1501 | 1126 | 947 | 18 | 4 | 2 |
| | | P | 954 | 949 | 950 | 1560 | 1137 | 952 | 21 | 3 | 1 |
| \(\chi_5(0.5)\) | \(\chi_5(1)\) | S | 941 | 944 | 945 | 1200 | 859 | 705 | 15 | 2 | 1 |
| | | P | 945 | 948 | 949 | 1201 | 861 | 706 | 15 | 2 | 1 |
| \(\chi_5(1)\) | \(\chi_5(0.5)\) | S | 943 | 944 | 945 | 1590 | 1173 | 974 | 24 | 4 | 2 |
| | | P | 943 | 948 | 948 | 1591 | 1176 | 976 | 24 | 4 | 2 |
| \(t_6(0.5)\) | \(\chi_5(1)\) | S | 943 | 944 | 945 | 1590 | 1173 | 974 | 24 | 4 | 2 |
| | | P | 943 | 948 | 947 | 1591 | 1176 | 706 | 24 | 4 | 1 |
| \(\chi_5(0.5)\) | \(t_6(1)\) | S | 940 | 944 | 945 | 1173 | 849 | 699 | 14 | 2 | 1 |
| | | P | 946 | 949 | 948 | 1172 | 849 | 701 | 13 | 2 | 1 |
| \(t_6(1)\) | \(\chi_5(0.5)\) | S | 943 | 946 | 947 | 1591 | 1179 | 978 | 24 | 5 | 2 |
| | | P | 943 | 947 | 826 | 1588 | 1174 | 1534 | 24 | 4 | 249 |

Table 7: Empirical coverage, width and variance of widths of individual intervals at \(\alpha = 0.05\)-level, \(\sigma_j^2 = MSE_3^{\ast} (\hat{\theta}_j^\ast)\). \(S_1\), Setting 1; \(S_2\), Setting 2, \(S_3\), Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

of our method is the statistic we are trying to estimate and the appropriate bootstrap method.

References

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons.

Butar, F. B. and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112(1):63–76.

Carpenter, J. R., Goldstein, H., and Rasbash, J. (2003). A novel bootstrap procedure
| $e_{ij}$ | $u_j$ | Coverage | Length | Variance of length |
|---------|--------|----------|--------|-------------------|
|         |        | $S_1$    | $S_2$  | $S_3$  | $S_1$ | $S_2$ | $S_3$ |
| $N(0.5)$ | $N(1)$ | S 930 947 941 | 1861 | 1430 | 1219 | 17 3 2 |
|          |        | P 928 948 938 | 1865 | 1432 | 1218 | 16 3 2 |
| $N(1)$  | $N(0.5)$ | S 922 946 945 | 2386 | 1895 | 1646 | 28 7 3 |
|          |        | P 912 940 944 | 2357 | 1896 | 1645 | 35 6 3 |
| $t_6(0.5)$ | $t_6(1)$ | S 906 940 944 | 1923 | 1502 | 1276 | 58 17 8 |
|          |        | P 874 918 909 | 1846 | 1427 | 1217 | 32 6 2 |
| $t_6(1)$ | $t_6(0.5)$ | S 926 940 938 | 2340 | 1973 | 1715 | 63 25 12 |
|          |        | P 862 916 913 | 2323 | 1888 | 1643 | 60 10 4 |
| $\chi_5(0.5)$ | $\chi_5(1)$ | S 910 915 937 | 2009 | 1555 | 1316 | 63 12 4 |
|          |        | P 885 855 897 | 1895 | 1448 | 1227 | 36 6 2 |
| $\chi_5(1)$ | $\chi_5(0.5)$ | S 919 922 931 | 2618 | 2094 | 1799 | 90 19 7 |
|          |        | P 899 868 892 | 2515 | 1975 | 1696 | 57 10 4 |
| $t_6(0.5)$ | $\chi_5(1)$ | S 919 922 931 | 2618 | 2094 | 1799 | 90 19 7 |
|          |        | P 899 868 914 | 2515 | 1975 | 1227 | 57 10 2 |
| $\chi_5(0.5)$ | $t_6(1)$ | S 920 918 936 | 1963 | 1534 | 1301 | 56 11 4 |
|          |        | P 898 874 899 | 1850 | 1427 | 1217 | 31 6 2 |
| $t_6(1)$ | $\chi_5(0.5)$ | S 916 931 940 | 2599 | 2071 | 1772 | 94 34 16 |
|          |        | P 874 907 824 | 2511 | 1973 | 1535 | 59 11 249 |

Table 8: Empirical coverage, width and variance of widths of simultaneous intervals at $\alpha = 0.05$-level, $\sigma_j^2 = MSE_3^{*}(\hat{\theta}_j^*)$. $S_1$, Setting 1; $S_2$, Setting 2; $S_3$, Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):431–443.

Chambers, R. and Chandra, H. (2013). A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics*, 22(2):452–470.

Chambers, R. L. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93(2):255–268.

Chatterjee, S., Lahiri, P., and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Ann. Statist.*, 36(3):1221–1245.
| $e_{ij}$ | $u_j$ | Coverage | Length | Variance of length |
|---------|------|----------|--------|-------------------|
|         |      | M        | $S_1$  | $S_2$  | $S_3$  | $S_1$  | $S_2$  | $S_3$  | $S_1$  | $S_2$  | $S_3$  |
| $N(0.5)$ | $N(1)$ | S 943 947 947 | 1181 | 852 701 | 7 1 1 |
|         |      | P 944 947 946 | 1181 | 852 701 | 7 1 1 |
| $N(1)$  | $N(0.5)$ | S 946 946 947 | 1521 | 1129 948 | 12 3 1 |
|         |      | P 939 946 946 | 1488 | 1128 947 | 15 3 1 |
| $t_6(0.5)$ | $t_6(1)$ | S 943 946 948 | 1175 | 853 702 | 14 2 1 |
|         |      | P 943 945 947 | 1170 | 849 700 | 13 2 1 |
| $t_6(1)$  | $t_6(0.5)$ | S 946 945 948 | 1501 | 1126 947 | 18 4 2 |
|         |      | P 937 945 947 | 1467 | 1123 945 | 26 4 2 |
| $\chi_5(0.5)$ | $\chi_5(1)$ | S 941 944 945 | 1200 | 859 705 | 15 2 1 |
|         |      | P 945 948 949 | 1201 | 861 706 | 15 2 1 |
| $\chi_5(1)$  | $\chi_5(0.5)$ | S 943 944 945 | 1590 | 1173 974 | 24 4 2 |
|         |      | P 943 948 948 | 1591 | 1176 976 | 24 4 2 |
| $t_6(0.5)$  | $\chi_5(1)$ | S 943 944 945 | 1590 | 1173 974 | 24 4 2 |
|         |      | P 943 948 947 | 1591 | 1176 706 | 24 4 1 |
| $\chi_5(0.5)$ | $t_6(1)$ | S 940 944 945 | 1173 | 849 709 | 14 2 1 |
|         |      | P 946 949 948 | 1172 | 849 701 | 13 2 1 |
| $t_6(1)$  | $\chi_5(0.5)$ | S 943 946 947 | 1591 | 1179 978 | 24 5 2 |
|         |      | P 943 947 830 | 1588 | 1174 1538 | 24 4 248 |

Table 9: Empirical coverage, width and variance of widths of individual intervals at $\alpha = 0.05$-level, $\sigma^2 = MSE^*_{SPA}$. $S_1$, Setting 1; $S_2$, Setting 2; $S_3$, Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

Field, C., Pang, Z., and Welsh, A. H. (2010). Bootstrapping robust estimates for clustered data. *Journal of the American Statistical Association*, 105(492):1606–1616.

Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390.
| $e_{ij}$ | $u_j$ | Coverage | Length | Variance of length |
|--------|-------|----------|--------|-------------------|
|        |       | M | $S_1$ | $S_2$ | $S_3$ | $S_1$ | $S_2$ | $S_3$ | $S_1$ | $S_2$ | $S_3$ |
| $N(0.5)$ | $N(1)$ | S | 934 | 956 | 946 | 1864 | 1433 | 1221 | 16 | 2 | 1 |
|         |       | P | 930 | 951 | 942 | 1868 | 1435 | 1221 | 15 | 2 | 1 |
| $N(1)$  | $N(0.5)$ | S | 924 | 945 | 945 | 2390 | 1899 | 1650 | 26 | 5 | 2 |
|         |       | P | 910 | 947 | 947 | 2360 | 1900 | 1649 | 33 | 5 | 2 |
| $t_6(0.5)$ | $t_6(1)$ | S | 908 | 944 | 949 | 1927 | 1506 | 1280 | 56 | 16 | 7 |
|         |       | P | 877 | 918 | 918 | 1850 | 1430 | 1220 | 30 | 5 | 2 |
| $t_6(1)$  | $t_6(0.5)$ | S | 924 | 947 | 949 | 2436 | 1978 | 1719 | 61 | 24 | 11 |
|         |       | P | 860 | 921 | 919 | 2327 | 1892 | 1646 | 58 | 9 | 3 |
| $\chi_5(0.5)$ | $\chi_5((1)$ | S | 910 | 917 | 938 | 2014 | 1560 | 1320 | 62 | 11 | 3 |
|         |       | P | 886 | 859 | 907 | 1898 | 1451 | 1230 | 34 | 5 | 2 |
| $\chi_5(1)$  | $\chi_5(0.5)$ | S | 921 | 918 | 935 | 2623 | 2100 | 1805 | 87 | 17 | 6 |
|         |       | P | 898 | 878 | 901 | 2519 | 1980 | 1700 | 54 | 8 | 3 |
| $t_6(0.5)$  | $\chi_5(1)$ | S | 921 | 918 | 935 | 2623 | 2100 | 1805 | 87 | 17 | 6 |
|         |       | P | 898 | 878 | 919 | 2519 | 1980 | 1230 | 54 | 8 | 2 |
| $\chi_5(0.5)$ | $t_6(1)$ | S | 921 | 919 | 939 | 1968 | 1539 | 1306 | 54 | 10 | 3 |
|         |       | P | 897 | 878 | 904 | 1854 | 1430 | 1220 | 29 | 5 | 2 |
| $t_6(1)$  | $\chi_5(0.5)$ | S | 914 | 933 | 944 | 2604 | 2078 | 1778 | 92 | 32 | 15 |
|         |       | P | 880 | 902 | 833 | 2515 | 1977 | 1538 | 56 | 9 | 249 |

Table 10: Empirical coverage, width and variance of widths of simultaneous intervals at $\alpha = 0.05$-level, $\sigma^2 = MSE_{SPA}^*$. $S_1$, Setting 1; $S_2$, Setting 2; $S_3$, Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

Flores-Agreda, D. and Cantoni, E. (2019). Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. *Computational Statistics & Data Analysis*, 130:1–17.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Ann. Statist.*, 34(4):1733–1750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *J. R. Statist. Soc. B*, 68(2):221–238.

Henderson, C. R. (1950). Estimation of genetic parameters. *Biometrics*, 6(1):186–187.

Jiang, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statistica Sinica*, 8(1):861–885.
\begin{table}
\centering
\begin{tabular}{|l|l|l|l|l|l|l|l|l|l|l|l|}
\hline
$e_{ij}$ & $u_j$ & Coverage & Length & Variance of length \\
\hline
 & & M & $S_1$ & $S_2$ & $S_3$ & $S_1$ & $S_2$ & $S_3$ & $S_1$ & $S_2$ & $S_3$ \\
\hline
$N(0.5)$ & $N(1)$ & S & 943 & 947 & 947 & 1181 & 852 & 701 & 7 & 1 & 1 \\
 & & P & 944 & 947 & 946 & 1181 & 852 & 701 & 7 & 1 & 1 \\
$N(1)$ & $N(0.5)$ & S & 946 & 946 & 947 & 1521 & 1129 & 948 & 12 & 3 & 1 \\
 & & P & 939 & 946 & 946 & 1488 & 1128 & 947 & 15 & 3 & 1 \\
t$_6(0.5)$ & t$_6(1)$ & S & 943 & 946 & 948 & 1175 & 853 & 702 & 14 & 2 & 1 \\
 & & P & 943 & 945 & 947 & 1170 & 849 & 700 & 13 & 2 & 1 \\
t$_6(1)$ & t$_6(0.5)$ & S & 946 & 945 & 948 & 1501 & 1126 & 947 & 18 & 4 & 2 \\
 & & P & 937 & 945 & 947 & 1467 & 1123 & 945 & 26 & 4 & 2 \\
$\chi_5(0.5)$ & $\chi_5(1)$ & S & 941 & 944 & 945 & 1200 & 859 & 705 & 15 & 2 & 1 \\
 & & P & 945 & 948 & 949 & 1201 & 861 & 706 & 15 & 2 & 1 \\
$\chi_5(1)$ & $\chi_5(0.5)$ & S & 943 & 944 & 945 & 1590 & 1173 & 974 & 24 & 4 & 2 \\
 & & P & 943 & 948 & 948 & 1591 & 1176 & 976 & 24 & 4 & 2 \\
t$_6(0.5)$ & $\chi_5(1)$ & S & 943 & 944 & 945 & 1590 & 1173 & 974 & 24 & 4 & 2 \\
 & & P & 943 & 948 & 947 & 1591 & 1176 & 706 & 24 & 4 & 1 \\
$\chi_5(0.5)$ & t$_6(1)$ & S & 940 & 944 & 945 & 1173 & 849 & 699 & 14 & 2 & 1 \\
 & & P & 946 & 949 & 948 & 1172 & 849 & 701 & 13 & 2 & 1 \\
t$_6(1)$ & $\chi_5(0.5)$ & S & 943 & 946 & 947 & 1591 & 1179 & 978 & 24 & 5 & 2 \\
 & & P & 943 & 947 & 817 & 1588 & 1174 & 1542 & 24 & 4 & 256 \\
\hline
\end{tabular}
\caption{Empirical coverage, width and variance of widths of individual intervals at $\alpha = 0.05$-level, $\sigma^2 = \text{MSE}_{BC}^*$. $S_1$, Setting 1; $S_2$, Setting 2, $S_3$, Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.}
\end{table}

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B*, 58(4):619–656.

Lombardía, M. J. and Sperlich, S. (2008). Semiparametric inference in generalized mixed effects models. *J. R. Statist. Soc. B*, 70(5):913–930.

McCullagh, P. (2000). Resampling and exchangeable arrays. *Bernoulli*, pages 285–301.
Table 12: Empirical coverage, width and variance of widths of simultaneous intervals at \( \alpha = 0.05 \)-level, \( \sigma^2 = \text{MSE}^*_{BC} \). \( S_1 \), Setting 1; \( S_2 \), Setting 2, \( S_3 \), Setting 3; M, Method; S, semiparametric bootstrap; P, parametric bootstrap. All numerical entries are multiplied by 1000.

Morris, J. S. (2002). The blups are not “best” when it comes to bootstrapping. *Statistics & Probability Letters*, 56(4):425–430.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008). Nonparametric small area estimation using penalized spline regression. *J. R. Statist. Soc. B*, 70:265–286.

O’Shaughnessy, P. and Welsh, A. H. (2018). Bootstrapping longitudinal data with multiple levels of variation. *Computational Statistics & Data Analysis*, 124:117–131.

Rao, J. N. K. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Reluga, K., Lombardía, M.-J., and Sperlich, S. (2021a). Simultaneous inference for em-

| \( e_{ij} \) | \( u_j \) | Coverage | Length | Variance of length |
|---|---|---|---|---|
| \( N(0.5) \) | \( N(1) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 926 | 940 | 942 | 1873 | 1439 | 1224 | 22 | 7 | 4 |
| \( N(1) \) | \( N(0.5) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 911 | 941 | 948 | 2365 | 1905 | 1654 | 45 | 12 | 8 |
| \( t_6(0.5) \) | \( t_6(1) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 876 | 922 | 908 | 1854 | 1434 | 1223 | 38 | 9 | 5 |
| \( t_6(1) \) | \( t_6(0.5) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 864 | 907 | 907 | 2332 | 1897 | 1651 | 70 | 17 | 9 |
| \( \chi_5(0.5) \) | \( \chi_5(1) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 897 | 878 | 887 | 2525 | 1986 | 1705 | 68 | 18 | 9 |
| \( \chi_5(1) \) | \( \chi_5(0.5) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 884 | 859 | 890 | 1903 | 1455 | 1234 | 42 | 10 | 5 |
| \( t_6(0.5) \) | \( \chi_5(1) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 897 | 878 | 913 | 2525 | 1986 | 1234 | 68 | 18 | 5 |
| \( \chi_5(0.5) \) | \( t_6(1) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 893 | 866 | 902 | 1859 | 1434 | 1223 | 37 | 9 | 5 |
| \( t_6(1) \) | \( \chi_5(0.5) \) | M | \( S_1 \) | \( S_2 \) | \( S_3 \) | S \( _1 \) | S \( _2 \) | S \( _3 \) |
| P | 868 | 900 | 825 | 2521 | 1983 | 1543 | 71 | 18 | 5 |
pirical best predictors with a poverty study in small areas. *J. Am. Statist. Ass.*, To appear(ja):1–33.

Reluga, K., Lombardía, M. J., and Sperlich, S. A. (2021b). Simultaneous inference for linear mixed model parameters with an application to small area estimation. *arXiv:1903.02774*.

Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N. (2020). Data-driven transformations in small area estimation. *J. R. Statist. Soc. A*, 183(1):121–148.

Samanta, M. and Welsh, A. H. (2013). Bootstrapping for highly unbalanced clustered data. *Computational Statistics & Data Analysis*, 59:70–81.

Shao, J., Kübler, J., and Pigeot, I. (2000). Consistency of the bootstrap procedure in individual bioequivalence. *Biometrika*, 87(3):573–585.

Sinha, S. K. and Rao, J. (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37(3):381–399.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer.