Statistical Methods & Applications (2022) 31:551–585
https://doi.org/10.1007/s10260-021-00581-7

ORIGINAL PAPER

Recycled two-stage estimation in nonlinear mixed effects regression models

Yue Zhang[1] · Ben Boukai[1]

Received: 14 June 2020 / Revised: 3 May 2021 / Accepted: 4 July 2021 /
Published online: 29 August 2021

Abstract

We consider a re-sampling scheme for estimation of the population parameters in the mixed-effects nonlinear regression models of the type used, for example, in clinical pharmacokinetics. We provide a two-stage estimation procedure which resamples (or *recycles*), via random weightings, the various parameter's estimates to construct consistent estimates of their respective sampling distributions. In particular, we establish under rather general distribution-free assumptions, the asymptotic normality and consistency of the standard two-stage estimates and of their resampled version and demonstrate the applicability of our proposed resampling methodology in a small simulation study. A detailed example based on real clinical pharmacokinetic data is also provided.

Keywords Resampling · Random weights · Hierarchical nonlinear models · Random effects

Mathematics Subject Classification MSC code1 · MSC code2 · more

1 Introduction

Hierarchical mixed-effects nonlinear regression models are widely used nowadays to analyze complex data involving longitudinal or repeated measures which are often arising in pharmacokinetics or from medical, biological and other similar applications (see for example Davidian and Giltinan (2003)). In such studies, the sampling units are often "subjects" drawn from the relevant population of interest

✉ Yue Zhang
yzhangmath@gmail.com

Ben Boukai
bboukai@iupui.edu

[1] Department of Mathematical Sciences, IUPUI, Indianapolis, Indiana 46202, USA

whereby statistical inference, primarily for the estimation of various model parameters, is being sought, primarily on certain characteristics of the underlying population of interest. In that context, the hierarchical nonlinear model can be considered as an extension of the ordinary nonlinear regression models constructed to handle and 'aggregate' data obtained from several individuals. Modeling this type of data usually involves a 'functional' relationship between at least one predictor variable, x, and the measured response, y. As it often the case, the assumed 'functional' model between the response y and the predictor x, is based on some on physical or mechanistic grounds and is usually nonlinear in its parameters. For instance, in pharmacokinetics, a typical (compartmental) model of drug's concentration in the plasma is obtained from a set differential equations reflecting the nonlinear time-dependency of the drug's disposition in the body. Figure 1 below, illustrates such plasma concentrations profiles for a group of $N = 12$ patients (the sample), each observed at $n = 11$ time points following the administration of the drug under study (the *Theophylline* study, see for example Boeckmann et al. (1994), Davidian and Giltinan (1995) and also Sect. 5.3 for more details about this well-known data set).

The primary aim of such pharmacokinetic studies with data as depicted in Fig. 1, is to make, based on the N patients' data, generalizations about the drug disposition in the population of interest to which the group patients belongs. Therefore such studies require a valid and reliable estimation procedure of the population's

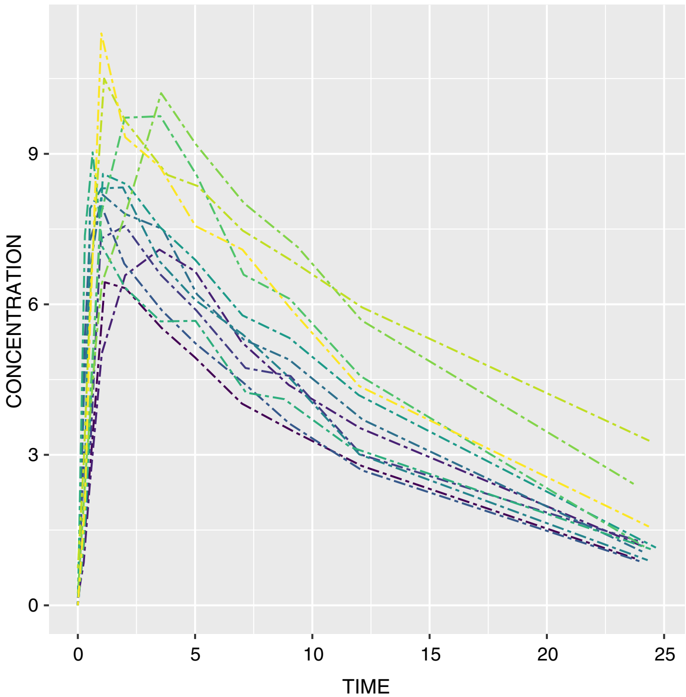

Fig. 1 The *Theophylline data*—drug plasma concentrations (ng/mL) profiles of $N = 12$ patients recorded over time (hr)

“typicaland” variability values for each of the underlying pharmacokinetic parameters (e.g.: the ‘typical’ rates of absorption, elimination, and clearance)– usually reflecting the population (hierarchical) distribution of the relevant model’s parameters. In that context, there are three basic types of ’typical’ population pharmacokinetic parameters. Some are viewed as *fixed-effect* parameters which quantify the population average kinetics of a drug; others represent inter-individual *random-effect* parameters, which quantify the typical magnitude of inter-individual variability in pharmacokinetic parameters and the intra-individual random-effect parameter which quantifies the typical magnitude of the intra-individual variability (the experimental error).

The basic hierarchical linear regression model for pharmacokinetics applications was pioneered by Sheiner et al. (1972), which accounted for both types of variations; of within and between subjects. The nonlinear case received widespread attention in later developments. Lindstrom and Bates (1990) proposed a general nonlinear mixed effects model for repeated measures data and proposed estimators combining least squares estimators and maximum likelihood estimators (under specific normality assumption). Vonesh and Carter (1992) discussed nonlinear mixed effects model for unbalanced repeated measures. Additional related references include: Mallet (1986), Davidian and Gallant (1993a); Davidian and Giltinan (1993b, 1995).

In all, the standard approach for inference in hierarchical nonlinear models is typically based on full distributional assumptions for both, the intra-individual and inter-individual random components. The most commonly used assumption is that both random components are considered to be normally distributed. However, this can be a questionable assumption in many cases. Our main results in this work offer a more generalized framework that does not hinge on the normality assumption of the various random terms. In fact, the rigorous asymptotic results we obtained are established only with minimal moments conditions on the random errors and random effect components of the underlying model and thus could be construed as a distribution-free approach.

One simple approach for estimation in such hierarchical ’population’ models is the so-called two-stage estimation method. At the first stage one estimates the ’individual-level’ parameters and then, at the second-stage, combines them in some manner to obtain the ’population-level’ parameter estimates. However despite of its simplicity, the main challenge to such a two-stage estimation approach is in obtaining the sampling distributions and related properties (accuracy, precision, consitency, etc..) of the final estimators, either in finite or in large sample settings. For most part, the performance of these two-stage estimation methods have been evaluated primarily via Monte-Carlo simulations– see related references including: Sheiner and Beal (1981, 1982, 1983), Steimer et al. (1984), and Davidian and Giltinan (1995, 2003). Hence, an alternate and a more data oriented evaluation methodology should be considered in assesing this type of hierarchical models. Using a variant of the random weighting technique, Bar-Lev and Boukai (2015) proposed a re-sampling scheme, which is termed herein *recycling*, as a valuable and valid alternative methodology for evaluation and comparison of the estimation procedure. Zhang and Boukai (2019b) studied the validity and established the asymptotic consistency and asymptotic normality of the *recycled* estimates in a one-layered nonlinear regression model.

In the present paper we extend Bar-Lev and Boukai (2015) approach to include general random weights in the case of hierarchical nonlinear regression models with minimal moments assumptions on the random error-terms/effects. In Sect. 2, we present the basic framework for the hierarchical nonlinear regression models with fixed and random effects. In Sect. 3, we describe the *Standard Two-Stage* (STS) estimation procedure for the population parameters appropriate in this hierarchical nonlinear regression settings. Along these lines, we introduce a corresponding re-sampling scheme especially devised, based on general random weights, to obtain the *recycled* version of the STS estimators. In Sect. 4, we establish the asymptotic consistency and asymptotic normality of the STS estimators in such general settings. As we mentioned before, our rigorous results do not depend on specifying the distribution(s) of the the random component terms in the model (both errors and effects), but rather, are obtained largely based on minimal moments assumptions. As far as we know, these are the first provably valid asymptotic results concerning the sampling distribution and implied sampling properties of the estimators obtained using the STS procedure in the context of hierarchical nonlinear regression. In addition, we demonstrate the applicability via the asymptotic consistency and normality of the *recycled* version of the STS estimators. These results enable us to use the sampling distribution of the *recycled* version of the STS estimators to approximate the unknown sampling distribution of the actual STS estimators in a general 'distribution-free' framework. The results of extensive simulation studies and a detailed application to the *Theophylline data* are provided in Sect. 5. The proofs of our main results along with many other technical details are provided in the "Appendix".

2 The basic hierarchical (population) model

Consider a study involving a random sample of N individuals, where the nonlinear regression model (as in Zhang and Boukai (2019b)) is assumed to hold for each of the i-th individuals. That is, for each i, $(i = 1, 2, \ldots, N)$, we have available the n_i (repeated) observations on the response variable in the form of $\mathbf{y}_i := (y_{i1}, y_{i2}, \ldots, y_{in_i})^{\mathbf{t}}$, where

$$y_{ij} = f(\mathbf{x}_{ij};\, \boldsymbol{\theta}_i) + \epsilon_{ij}, \qquad j = 1, \ldots, n_i, \tag{1}$$

and $\mathbf{x}_{ij}$ is the j-th covariate for the i-th individual, which gives rise to the response, y_{ij}, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. Here, $f(\cdot)$ is a given nonlinear function and ϵ_{ij} denote some *i.i.d.* $(0, \sigma^2)$ error-terms. That is, if we set $\boldsymbol{\epsilon}_i := (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{in_i})^{\mathbf{t}}$, then $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon}_i) \equiv Cov(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^{\mathbf{t}}) = \sigma^2 \mathbf{I}_{n_i}$. In the current context of hierarchical modeling, the parameter vector $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})^{\mathbf{t}} \in \Theta \subset I\!R^p$, (with $p < n_i$), can vary from individual to individual, so that $\boldsymbol{\theta}_i$ is seen as the individual-specific realization of $\boldsymbol{\theta}$. More specifically, it is assumed that, independent of the error terms, $\boldsymbol{\epsilon}_i$,

$$\boldsymbol{\theta}_i := \boldsymbol{\theta}_0 + \mathbf{b}_i, \tag{2}$$

where $\boldsymbol{\theta}_0 := (\theta_{01}, \theta_{02}, \ldots, \theta_{0p})^{\mathbf{t}}$, is a fixed population parameter, though unknown, and $\mathbf{b}_i = (b_{i1}, b_{i2}, \ldots, b_{ip})^{\mathbf{t}}$ is a $p \times 1$ vector representing the random effects associated with i-th individual. It is assumed that the random effects, $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ are independent and identically distributed random vectors satisfying,

$$E(\mathbf{b}_i) = \mathbf{0} \quad \text{and} \quad Var(\mathbf{b}_i) \equiv Cov(\mathbf{b}_i, \mathbf{b}_i^t) = \mathbf{D}.$$

Here σ^2 represents the *within individual* variability and $\mathbf{D}$ describes the *between individuals* variability. Thus, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ are *i.i.d.* random vectors with

$$E(\boldsymbol{\theta}_i) = \mathbf{0} \quad \text{and} \quad Var(\boldsymbol{\theta}_i) = \mathbf{D}.$$

In the simple (i.e.: standard) hierarchical modeling it is often assumed that $\mathbf{D}$ is some diagonal matrix of the form $\mathbf{D} = Diag(\lambda_1^2, \lambda_2^2, \ldots, \lambda_p^2)$ or even simpler, as $\mathbf{D} = \lambda^2 \mathbf{I}_p$ for some $\lambda > 0$, and that $Var(\boldsymbol{\epsilon}_i) = \sigma^2 \mathbf{I}_{n_i}$ for each $i = 1, \ldots, N$ for some $\sigma > 0$.

In the more complex hierarchical modeling, more general structures of the *within individual* variability $Var(\boldsymbol{\epsilon}_i) = \Gamma_i$ (for some Γ_i) and of the *between individuals* variability, $\mathbf{D}$, are possible. However, even in the simplest structure, the available estimation methods for these model's parameters, $\boldsymbol{\theta}_0, \sigma^2$ and $\mathbf{D}$ are typically highly iterative in their nature and are based on the variations of the least squares estimation. Similarly, even when considered under some specific distributional assumptions, such as that both, the error terms $\boldsymbol{\epsilon}_i$, and the random effects $\mathbf{b}_i$ are normally distributed, so that, $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$ and $\mathbf{b}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{D})$, for each $= i = 1, \ldots, N$. In fact, many of the available results in the literature hinge on the specific normality assumption and on the ability to effectively 'linearize' the regression function $f(\cdot)$ (see for example Bates and Watts (2007)) on order to obtain some assessment of the resulting sampling distributions fo the parameters' estimates. We point out that here we require no specific distributional assumptions (such as normality, or otherwise) on either the intra-individual and the inter-individual error terms, $\boldsymbol{\epsilon}_i$ nor $\mathbf{b}_i$, respectively.

3 The two-stage estimation procedure

For each $i = 1, \ldots, N$, let $\mathbf{f}_i(\boldsymbol{\theta})$ denote the $n_i \times 1$ vectors whose elements are $f(\mathbf{x}_{ij}, \boldsymbol{\theta}), j = 1, \ldots, n_i$ then model (1) can be written more succinctly as

$$\mathbf{y}_i = \mathbf{f}_i(\boldsymbol{\theta}_i) + \boldsymbol{\epsilon}_i \tag{3}$$

Accordingly, the STS estimation procedure can be described as follows:

On Stage I: For each $i = 1, \ldots, N$ obtain $\hat{\boldsymbol{\theta}}_{ni}$ as the minimizer of

$$Q_i(\boldsymbol{\theta}) := (\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\theta}))(\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\theta}))^{\mathbf{t}} \equiv \sum_{j=1}^{n_i}(y_{ij} - f(\mathbf{x}_{ij}, \boldsymbol{\theta}))^2, \quad (4)$$

so as to form $\hat{\boldsymbol{\theta}}_{n1}, \hat{\boldsymbol{\theta}}_{n2}, \ldots, \hat{\boldsymbol{\theta}}_{nN}$, based on all the $M := \sum_i^N n_i$ available observations. Next, estimate the *within-individual* variability component, σ^2, by

$$\hat{\sigma}^2_M := \frac{1}{M - pN} \sum_{i=1}^{N} Q_i(\hat{\boldsymbol{\theta}}_{ni}).$$

On Stage II: Estimate the 'population' parameter $\boldsymbol{\theta}_0$ by the average

$$\hat{\boldsymbol{\theta}} - STS := \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_{ni}. \quad (5)$$

Next, estimate $Var(\hat{\boldsymbol{\theta}} - STS)$ by $\mathbf{S}^2 - STS/N$, where

$$\mathbf{S}^2 - STS := \sum_{i=1}^{N} (\hat{\boldsymbol{\theta}}_{ni} - \hat{\boldsymbol{\theta}} - STS)(\hat{\boldsymbol{\theta}}_{ni} - \hat{\boldsymbol{\theta}} - STS)^{\mathbf{t}}.$$

Finally estimate the *between-individual* variability component, $\mathbf{D}$, by

$$\hat{\boldsymbol{D}} = \mathbf{S}^2 - STS - \min(\hat{v}, \hat{\sigma}^2_M)\, \hat{\boldsymbol{\Sigma}}_N, \quad (6)$$

where $\hat{\boldsymbol{\Sigma}}_N := \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\Sigma}_{n_i}(\hat{\boldsymbol{\theta}}_{ni})$, with $\boldsymbol{\Sigma}^{-1}_{n_i}$ defined as,

$$\boldsymbol{\Sigma}^{-1}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\boldsymbol{\theta}) \nabla f_i(\boldsymbol{\theta})^{\mathbf{t}}, \quad (7)$$

and where $\hat{v}$ is the smallest root of the equation $|\mathbf{S}^2 - STS - v\hat{\boldsymbol{\Sigma}}_N| = 0$, see Davidian and Giltinan (2003) for details.

Bar-Lev and Boukai (2015) provided a numerical study of this two-stage estimation procedure in the context of a (hierarchical) pharmacokinetics modeling under the normality assumption. They also proposed a corresponding two-stage resampling scheme based on specific $\mathcal{D}irichlet(\boldsymbol{1})$ random weights. However, in this paper we consider a more general framework for the random weights to be used.

As in Zhang and Boukai (2019b), we let for each $n \geq 1$, the random weights, $\mathbf{w}_n = (w_{1:n}, w_{2:n}, \ldots, w_{n:n})^{\mathbf{t}}$, be a vector of exchangeable nonnegative random variables with $E(w_{i:n}) = 1$ and $Var(w_{i:n}) := \tau_n^2$, and let $W_i \equiv W_{1:n} = (w_{i:n} - 1)/\tau_n$ be the standardized version of $w_{i:n}$, $i = 1, \ldots, n$. In addition we also assume:

Assumption W The underlying distribution of the random weights $\mathbf{w}_n$ satisfies

1. For all $n \geq 1$, the random weights $\mathbf{w}_n$ are independent of $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^{\mathbf{t}}$;

2. $\tau_n^2 = o(n)$, $E(W_i W_j) = O(n^{-1})$ and $E(W_i^2 W_j^2) \to 1$ for all $i \neq j$, $E(W_i^4) < \infty$ for all i.

Some examples of random weights, $\mathbf{w}_n$ that satisfy the above conditions in Assumption W are: the Multinomial weights, $\mathbf{w}_n \sim \mathcal{M}ultinomial(n, 1/n, 1/n, \ldots, 1/n)$, which correspond to the classical bootstrap of Efron (1979) and the Dirichlet weights, $\mathbf{w}_n \equiv n \times \mathbf{z}_n$ where $\mathbf{z}_n \sim \mathcal{D}irichlet(\alpha, \alpha, \ldots, \alpha)$, with $\alpha > 0$ which often refer to as the Bayesian bootstrap (see Rubin (1981), and its variants as in Zheng and Tu (1988) and Lo (1991)).

We will assume throughout this paper that all the random weights we use in the sequel do satisfy Assumption W. With such random weights $\mathbf{w}_n$ at hand, we define in similarity to (3), the *recycled* version $\hat{\boldsymbol{\theta}}_n^*$ of $\hat{\boldsymbol{\theta}}_n$ as the minimizer of the *randomly weighted* least squares criterion. With such general random weights, the *recycled* version of the STS estimation procedure described in (4-7) above is:

On Stage I *: For each $i = 1, \ldots, N$, independently generate random weights, $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in_i})^{\mathbf{t}}$ that satisfy Assumption W with $Var(w_{ij}) = \tau_{n_i}^2$ and obtain $\hat{\boldsymbol{\theta}}_{ni}^*$ as the minimizer of

$$Q_i^*(\boldsymbol{\theta}) := \sum_{j=1}^{n_i} w_{ij}(y_{ij} - f(\mathbf{x}_{ij}, \boldsymbol{\theta}))^2, \tag{8}$$

so as to form $\hat{\boldsymbol{\theta}}_{n1}^*, \hat{\boldsymbol{\theta}}_{n2}^*, \ldots, \hat{\boldsymbol{\theta}}_{nN}^*$.

On Stage II *: Independent of **Stage I** *, generate random weights, $\mathbf{u} = (u_1, u_2, \ldots, u_N)^{\mathbf{t}}$ that satisfy Assumption W with $Var(u_i) = \tau_N^2$, and obtained the *recycled* version of $\hat{\boldsymbol{\theta}} - STS$ as:

$$\hat{\boldsymbol{\theta}} - RTS^* := \frac{1}{N}\sum_{i=1}^{N} u_i \hat{\boldsymbol{\theta}}_{ni}^* \tag{9}$$

The *recycled* version $\mathbf{D}^*$ of $\mathbf{D}$ can be subsequently obtained as described in **Stage II** above.

4 Consistency of the STS and the recycled estimation procedures

In this section we present some asymptotic results that establish and validate the consistency and asymptotic normality of the STS estimator, $\hat{\boldsymbol{\theta}} - STS$ (Theorems 1 and 2) and of its *recycled* version $\hat{\boldsymbol{\theta}} - RTS^*$ (Theorems 3 and 4), obtained using the general random weights satisfying the premises of Assumption W. We establish these results without the 'typical' normality assumption on the *within-individual* error terms, ϵ_{ij}, nor on the *between-individual* random effects $\mathbf{b}_i$. However, for simplicity of the exposition, we state these results in the case of $p = 1$, so that $\Theta \in I\!R$. With that in mind, we denote for each $i = 1, \ldots, N$,

$$f_{ij}(\theta) \equiv f(x_{ij}, \theta), \quad \text{for} \quad j = 1, \ldots, n_i.$$

Accordingly, the least squares criterion in (1), becomes

$$Q_{ni}(\theta) := \sum_{j=1}^{n_i} (y_{ij} - f_{ij}(\theta))^2,$$

and the LS estimator $\hat{\theta}_{ni}$ is readily seen as the solution of

$$Q'_{ni}(\theta) := 2 \sum_{j=1}^{n_i} \phi_{ij}(\theta) = 0 \tag{10}$$

where,

$$\phi_{ij}(\theta) := -(y_{ij} - f_{ij}(\theta)) f'_{ij}(\theta), \tag{11}$$

with $f'_{ij}(\theta) := df_{ij}(\theta)/d\theta$, for $j = 1\ldots, n_i$ and for each $i = 1\ldots, N$. We write $f''_{ij}(\theta) := df'_{ij}(\theta)/d\theta$ and $\phi'_{ij}(\theta) := d\phi_{ij}(\theta)/d\theta$, etc. As in Zhang and Boukai (2019b), we also assume that $f'_{ij}(\theta)$ and $f''_{ij}(\theta)$ exist for all θ near θ_0. However, to account for the inclusion of the $(0, \lambda^2)$ random effect term, b_i, in the model, we also assume that,

Assumption A For each $i = 1, \ldots, N$

1. $a_{n_i}^2 := \sigma^2 \sum_{j=1}^{n_i} E(f'^2_{ij}(\theta_0 + b_i)) \to \infty \quad as \quad n_i \to \infty,$;
2. $\limsup_{n_i \to \infty} \; a_{n_i}^{-2} \sum_{j=1}^{n_i} \sup_{|\theta - \theta_0 - b_i| \le \delta} f''^2_{ij}(\theta) < \infty$
3. $a_{n_i}^{-2} \sum_{j=1}^{n_i} f'^2_{ij}(\theta) \to \frac{1}{\sigma^2}$ uniformly in $|\theta - \theta_0 - b_i| \le \delta$.

In the following two Theorems we establish, under the conditions of Assumption A, the asymptotic consistency and normality of $\hat{\theta}_{STS}$. Their proofs and some related technical results are given in Sect. 7.1.

Theorem 1 *Suppose that Assumption A holds, then there exists a sequence* $\hat{\theta}_{ni}$ *of solutions of* (10) *such that* $\hat{\theta}_{ni} = \theta_0 + b_i + a_{ni}^{-1} T_{ni}$, *where* $|T_{ni}| < K$ *in probability, for each* $i = 1, 2, \ldots, N$. *Further, there exists a sequence* $\hat{\theta}_{STS}$ *as expressed in* (5) *such that* $\hat{\theta}_{STS} - \theta_0 \xrightarrow{p} 0$, *as* $n_i \to \infty$, *for* $i = 1, 2, \ldots, N$, *and as* $N \to \infty$.

Theorem 2 *Suppose that Assumption A holds. If* $\lim_{N, ni \to \infty} N/a_{ni}^2 < \infty$, *for all* $i = 1, 2, \ldots, N$, *then there exists a sequence* $\hat{\theta}_{STS}$ *as expressed in* (5) *such that* $\hat{\theta} - STS - \theta_0 = \frac{1}{N} \sum_{i=1}^{N} b_i - \psi - N, n_i$, *where* $\sqrt{N}\psi - N, n_i \xrightarrow{p} 0$. *Further,*

$$\mathcal{R}_{\mathcal{N}} := \frac{\sqrt{N}}{\lambda}(\hat{\theta} - STS - \theta_0) \Rightarrow \mathcal{N}(0,1)$$

as $n_i \to \infty$, for $i = 1, 2, \ldots, N$, and as $N \to \infty$.

For the *recycled* STS estimation procedure as described in Sect. 3, the *recycled* version $\hat{\theta}^*_{ni}$ of $\hat{\theta}_{ni}$ is the minimizer of (8), or alternatively, the direct solution of

$$Q_i^{*\prime}(\theta) := 2\sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\theta) = 0, \tag{12}$$

where $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in_i})^{\mathbf{t}}$ are the randomly drawn weights (satisfying Assumption W), for the ith individual, $i = 1, 2, \ldots, N$. For establishing comparable results to those given in Theorems 1 and 2 for the *recycled* version, $\hat{\theta}^* - RTS = \sum_{i=1}^N u_i\hat{\theta}^*_{ni}/N$ of $\hat{\theta} - STS = \sum_{i=1}^N \hat{\theta}_{ni}/N$, with the random weights $\mathbf{u} = (u_1, u_2, \ldots, u_N)^{\mathbf{t}}$ as in **Stage II** *, we need the following additional assumptions.

Assumption B In addition to Assumption A, we assume that $E(\epsilon_{ij}^4) < \infty$ and that for each $i = 1, 2, \ldots, N$,

1. $\limsup_{n_i\to\infty} a_{n_i}^{-2}\sum_{j=1}^{n_i} \sup_{|\theta-\theta_0-b_i|\le\delta} f_{ij}^{\prime 4}(\theta) < \infty$,
2. $\limsup_{n_i\to\infty} a_{n_i}^{-2}\sum_{j=1}^{n_i} \sup_{|\theta-\theta_0-b_i|\le\delta} f_{ij}^{\prime\prime 4}(\theta) < \infty$,
3. As $n_i \to \infty, n_i a_{n_i}^{-2} \to c_i \ge 0$.

In Theorems 3 and 4 below we establish, under the conditions of Assumptions A and B, the asymptotic consistency and normality of the *recycled* estimator $\hat{\theta}^* - RTS$. Their proofs and some related technical results are given in Sect. 7.2.

Theorem 3 *Suppose that Assumptions A and B hold. Then there exists a sequence* $\hat{\theta}^*_{ni}$ *as the solution of* (12) *such that* $\hat{\theta}^*_{ni} = \hat{\theta}_{ni} + a_{ni}^{-1}T^*_{ni}$, *where* $|T^*_{ni}| < K\tau_{n_i}$ *in probability, for* $i = 1, \ldots, N$. *Further for any* $\epsilon > 0$, *we have* $P^*(|\hat{\theta}^* - RTS - \theta_0| > \epsilon) = o_p(1)$, *as* $n_i \to \infty$, *for* $i = 1, 2, \ldots, N$, *and as* $N \to \infty$.

Theorem 4 *Suppose that Assumptions A and B hold. If for each* $i = 1, 2, \ldots, N$, $\frac{\tau_{n_i}}{\tau - N} = o(\sqrt{n_i})$, *then we have* $\hat{\theta}^*_{RTS} - \hat{\theta}_{STS} = \frac{1}{N}\sum_{i=1}^N (u_i - 1)\hat{\theta}_{ni} - \psi^*_{N,n_i}$, *where* $\frac{\sqrt{N}}{\tau_N}\psi^*_{N,n_i} \xrightarrow{p^*} 0$ *as* $N, n_i \to \infty$. *Additionally,*

$$\mathcal{R}^*_{\mathcal{N}} := \frac{\sqrt{N}}{\lambda\tau_N}(\hat{\theta}^*_{RTS} - \hat{\theta}_{STS}) \Rightarrow \mathcal{N}(0,1),$$

as $n_i \to \infty$, for $i = 1, 2, \ldots, N$, and as $N \to \infty$.

The proofs of Theorems 3 and 4 and some related technical results are given in Sect. 7.1. The following Corollary is an immediate consequence of the above results. It suggest that the sampling distribution of $\hat{\theta}_{STS}$ can be well approximated by that of the *recycled* or re-sampled version of it, $\hat{\theta}^*_{RTS}$.

Corollary 1 *For all $t \in I\!R$, let $\mathcal{H}_N(t) = P(\mathcal{R}_{\mathcal{N}} \le t)$, and $\mathcal{H}^*_N(t) = P^*(\mathcal{R}^*_{\mathcal{N}} \le t)$, denote the corresponding c.d.f of $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{R}^*_{\mathcal{N}}$, respectively. Then by Theorems 2 and 4,*

$$\sup_t |\mathcal{H}^*_n(t) - \mathcal{H}_n(t)| \to 0 \quad in \;\; probability.$$

5 Implementation and numerical results

5.1 Illustrating the STS estimation procedure

To illustrate the main results of Sect. 4 for the hierarchical nonlinear regression model and the corresponding STS estimation procedure as described in (4-7) above, we consider a typical compartmental modeling from pharmacokinetics. In the standard two-compartment model, the relationship between the measure drug concentration and the post-dosage time t, (following an intravenous administration), can be described through the nonlinear function of the form:

$$f(t; \boldsymbol{\eta}) = \eta_1 e^{-\eta_2 t} + \eta_3 e^{-\eta_4 t}, \tag{13}$$

with $\boldsymbol{\eta} := (\eta_1, \eta_2, \eta_3, \eta_4)'$ being a parameter representing the various kinetics rate constants, such as the rate of elimination, rate of absorption, clearance, volume, etc. Since these constants (i.e. parameters) must be positive, we re-parametrize the model with $\boldsymbol{\theta} \equiv \log(\boldsymbol{\eta})$ (with $\theta_k = log(\eta_k)$, $k = 1, 2, 3, 4$), so that with $t > 0$,

$$f(t; \boldsymbol{\theta}) = exp(\theta_1) exp\{-exp(\theta_2)t\} + exp(\theta_3) exp\{-exp(\theta_4)t\}, \tag{14}$$

with $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^{\mathbf{t}} \in I\!R^4$. For the simulation study we consider a situation in which the (plasma) drug concentrations $\{y_{ij}\}$ of N individuals were measure at post-dose times t_{ij} and are related as in model (1) via the nonlinear regression model,

$$y_{ij} = f(t_{ij}; \boldsymbol{\theta}_i) + \epsilon_{ij},$$

for $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. Here, as in Sect. 4, ϵ_{ij} are standard *i.i.d.* $(0, \sigma^2)$ random error terms and $\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \mathbf{b}_i$, where $\mathbf{b}_i$ are independent identically distributed random effects terms, with mean $\mathbf{0}$ and unknown variance $\lambda^2 \boldsymbol{I}_{4\times 4}$. Accordingly, we have in all a total of 6 unknown parameters, namely, $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40})^{\mathbf{t}}$, σ and λ.

Since σ and λ represent variation within and between individuals (respectively), different setting for these two lead to very different situations. For instance, Fig. 2a below, depicts the situation for $N = 5$ and $n_i \equiv n = 15$, each, when $\sigma = 0.1$ and $\lambda = 0.1$, so that the variation between individuals are similar to variation within individuals. Figure 2b depicts the situation with $\sigma = 0.05, \lambda = 1$, so that the variation between individuals is much larger than variation within individuals.

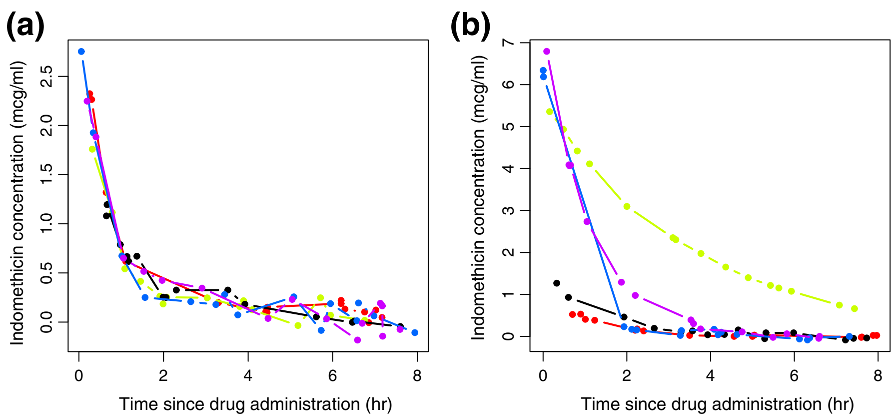

Fig. 2 Illustrating drug plasma concentration vs time for 5 individuals (colored) for the cases: **a** $\sigma = 0.1, \lambda = 0.1$ and **b** $\sigma = 0.05, \lambda = 1$

For the simulation, we set $\boldsymbol{\theta}_0 = (1, 0.8, -0.5, -1)^{\mathbf{t}}$, and for each i, the times $t_{ij}, j = 1, \ldots, n$ were generated uniformly from [0, 8] interval. To allow for different 'distributions', the error terms, ϵ_{ij}, as well as the random effect terms, $\mathbf{b}_i$, were generated either from the **(a)** *Truncated Normal*, **(b)** *Normal* and **(c)** *Laplace* distributions – all in consideration of Assumption A in our main results.

For each simulation run, with the *Truncated Normal* distribution for the error-terms and the random effects terms, we calculated the value of $\hat{\boldsymbol{\theta}}^k{}_{STS}$ as an estimator of $\boldsymbol{\theta}_0$ and repeated this procedure $M = 1000$ times to calculate the corresponding Mean Square Error (MSE) as followed,

$$MSE = \frac{1}{M} \sum_{k=1}^{M} ||\hat{\boldsymbol{\theta}}^k_{STS} - \boldsymbol{\theta}_0||^2$$

The corresponding simulation results obtained for various values of N and n, are presented in Table 1 for $\sigma = 0.1, \lambda = 0.1$ and $\sigma = 0.05, \lambda = 1$.

Table 1 The MSE of STS estimates for *truncated Normal* error-terms/effects with $\sigma = 0.1, \lambda = 0.1$ and $\sigma = 0.05, \lambda = 1.0$

n	$\sigma = 0.1$ $\lambda = 0.1$					$\sigma = 0.05$ $\lambda = 1.0$				
N	15	30	50	100	200	15	30	50	100	200
15	0.8662	0.2289	0.0465	0.0114	0.0063	1.0001	0.6383	0.5688	0.4730	0.4602
30	0.5767	0.1071	0.0244	0.0057	0.0033	0.6997	0.3950	0.3315	0.3523	0.3263
50	0.4584	0.0893	0.0210	0.0038	0.0020	0.5568	0.2944	0.2594	0.2500	0.2347
100	0.3785	0.0692	0.0125	0.0022	0.0010	0.3982	0.2245	0.2021	0.1973	0.2200
200	0.3506	0.0590	0.0089	0.0014	0.0006	0.3492	0.1945	0.1748	0.1882	0.1958

From Table 1, we see that with n and N both increasing, the MSE is decreasing, as expected. However, when $\sigma = 0.05, \lambda = 1$, n increasing for a fixed N, doesn't contribute to smaller MSE, which is consistent with our main result Theorem 1, the STS estimate is not consistent with only $n_i \to \infty$, (this effect is more obvious in the case λ is relatively large).

For simulating the results of Theorem 2, we choose θ_2 to be the unknown parameter, and use the main result to construct 95% Confidence Interval as

$$(\hat{\theta}_{STS} - 1.96\frac{\hat{\lambda}}{\sqrt{N}},\ \hat{\theta}_{STS} + 1.96\frac{\hat{\lambda}}{\sqrt{N}}),$$

where

$$\hat{\lambda}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\theta}_{ni} - \hat{\theta}_{STS})^2.$$

The estimate for $\hat{\lambda}$ used here is the simple STS estimate, not the corrected one as in (6). M=1,000 replications of such simulations were executed to determine the percentage of times the true value of the parameter estimates was contained in the interval. We use $\sigma = 0.5, \lambda = 0.5$ and observed Coverage Percentages are provided in Table 2.

From these results we can observe that with n and N both increase, the Coverage Percentage approximate to 0.95. While when n is small (15), with N increase, the Coverage Percentage is drifting farther away from the desired level of 0.95. This finding is consistent with our main result, the convergence require the condition $\lim_{N,ni\to\infty} N/a_{ni}^2 < \infty$, which in this case becomes $\lim_{n\to\infty} \frac{1}{n} a_n^2/\sigma^2 < \infty$, that is $\lim_{N,n\to\infty} N/n < \infty$ is required. Hence, when N is much large than n, this condition does not hold. Although for this model, error terms that follow the normal distribution do not satisfy Assumption A, we used normal error terms in the simulations, and reported the resulting MSE and Coverage Percentage for 95% confidence interval in Tables 2 and 3. From the results we can observe that with

Table 2 Coverage Percentage of the CI for the *truncated Normal* and *Normal* error-terms/effects with $\sigma = 0.5, \lambda = 0.5$

n	truncated Normal					Normal				
N	15	30	50	100	200	15	30	50	100	200
15	0.903	0.934	0.933	0.931	0.931	0.918	0.927	0.939	0.951	0.922
30	0.896	0.940	0.940	0.943	0.944	0.901	0.939	0.944	0.931	0.932
50	0.883	0.941	0.959	0.944	0.944	0.871	0.947	0.949	0.950	0.944
100	0.828	0.948	0.946	0.941	0.944	0.851	0.950	0.934	0.949	0.948
200	0.759	0.943	0.932	0.935	0.949	0.740	0.949	0.944	0.951	0.945

Table 3 The MSE of STS estimates for *Normal* error-terms/effects with $\sigma = 0.1, \lambda = 0.1$

	n				
N	15	30	50	100	200
15	0.7718	0.1746	0.0788	0.0112	0.0062
30	0.5548	0.1185	0.0297	0.0061	0.0032
50	0.4772	0.0928	0.0216	0.0044	0.0020
100	0.3828	0.0742	0.0122	0.0023	0.0010
200	0.3384	0.0563	0.0089	0.0014	0.0006

n and N increasing, the MSE are smaller and Coverage Percentage are closer to 0.95.

We further considered simulations using the Laplace distributions for the error terms and random effects terms. The complete results are reported in Zhang and Boukai (2019b) and indicate of similar conclusions.

5.2 Illustrating the *recycled* STS estimation procedure

Here we provide the results of the simulation studies corresponding to Theorem 3 and 4 concerning the *recycled* STS estimator, $\hat{\theta}^*_{RTS}$. We considered the same nonlinear (compartmental) model as given in the previous subsection, however again with $p = 1$. Accordingly, we choose θ_2 to represent the model's unknown parameter and set, for the simulations, $\theta_0 = 0.8$, for each i. As before, we generated the values of $\{t_{ij}, j = 1, \ldots, n\}$ uniformly from the [0, 8] interval, and drew the error terms, ϵ_{ij} and the random effects terms, b_i, from the *truncated Normal* distribution.

For each simulation run, we calculated the value of $\hat{\theta}_{STS}$ as in Sect. 3, then with $B = 1,000$, we generated $B \times N$ independent replications of the random weights $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in})$ and $B = 1,000$ independent replications of the random weight $\mathbf{u} = (u_1, u_2, \ldots, u_N)$, to obtain $\hat{\theta}^{*1}_{STS}$, $\hat{\theta}^{*2}_{STS}$, …, $\hat{\theta}^{*B}_{STS}$. The correspond 95% Confidence Intervals were formed. With $\sigma = 1$, $\lambda = 1$ a total of $M = 2000$ replications of such simulations were executed to determine the percentage of times the true value of the parameter estimates was contained in the interval and average confidence interval length was calculated. The Coverage Percentages with average confidence interval lengths are reported in Tables 4 and 5.

Table 4 demonstrates the results of the asymptotic results of Sect. 4. Table 5 provide Coverage Percentages with average confidence interval lengths, with random weights set to be *Multinomial, Dirichlet* or *Exponential* distributed . From these results we can see with N and n both increase, the Coverage Percentages converges to 0.95 as expected (see Corollary 5). Also notice that Coverage Percentages derived from the *recycled* STS are more accurate (closer to 0.95) than the asymptotic result, especially when n and N are small.

We further consider the case when n is even smaller. Table 6 provides Coverage Percentages and the average confidence interval length when $n = 10$ for the case of the Multinomial, Dirichlet or Exponential distributed random weights. As can be seen, in these cases, our procedure produces reasonable results. However, we must

Table 4 Simulated Coverage Percentage of the CI for the *truncated Normal* error-terms/effects with $\sigma = 1, \lambda = 1$

	n			
N	15	30	50	100
15	0.755	0.880	0.905	0.920
	0.999	1.004	1.009	1.038
30	0.590	0.860	0.930	0.955
	0.730	0.722	0.729	0.740
50	0.48	0.815	0.885	0.955
	0.566	0.576	0.568	0.573
100	0.170	0.680	0.895	0.935
	0.397	0.403	0.410	0.406

Table 5 Coverage Percentage of the CI for the *truncated Normal* error-terms/effects with $\sigma = 1, \lambda = 1$ and with *Multinomial* random weights

n	*Multinomial*				*Dirichlet*				*Exponential*			
N	15	30	50	100	15	30	50	100	15	30	50	100
15	0.860	0.910	0.930	0.940	0.810	0.905	0.930	0.950	0.810	0.895	0.920	0.945
	1.222	1.191	1.179	1.170	1.303	1.362	1.364	1.407	1.296	1.351	1.347	1.397
30	0.780	0.915	0.955	0.960	0.695	0.900	0.955	0.965	0.680	0.890	0.960	0.965
	0.881	0.855	0.851	0.832	0.936	0.965	0.993	1.001	0.935	0.965	0.990	0.999
50	0.760	0.890	0.940	0.940	0.605	0.870	0.930	0.965	0.590	0.855	0.930	0.940
	0.787	0.683	0.660	0.648	0.725	0.761	0.766	0.773	0.729	0.765	0.765	0.771
100	0.500	0.850	0.935	0.945	0.305	0.795	0.935	0.950	0.300	0.805	0.935	0.950
	0.478	0.473	0.471	0.458	0.509	0.534	0.550	0.546	0.507	0.532	0.550	0.546

Table 6 Coverage Percentage of the CI for the *truncated Normal* error-terms/effects with $\sigma = 0.05, \lambda = 1$ and with different choices of random weights

	n=10		
N	Multinomial	Dirichlet	Exponential
15	0.880	0.875	0.865
	0.899	1.114	1.121
100	0.705	0.705	0.685
	0.361	0.438	0.435

point out that the effects of a small sample size on our procedure depend also on the dimensionality, p, of the parameter $\boldsymbol{\theta}$, on the "nature"of the non-linear regression function $\mathbf{f}_i(\cdot)$ and its gradients, and on the particular minimization (optimization) algorithm used on $Q_i(\boldsymbol{\theta})$ in (4) and on $Q_i^*(\boldsymbol{\theta})$ in (8). Clearly, further numerical experimentation could be instructive in these regards.

5.3 An example–the*theophylline* data

We illustrate our proposed *recycled* two-stage estimation procedure with the *Theophylline* data set (which is widely available as `Theoph` under the R package, R Core Team (2020)). This well-known data set provides the concentration-time profiles (see Fig. 1) as were obtained in the pharmacokinetic study of the anti-asthmatic agent *Theophylline*, and reported by Boeckmann et al. (1994) and subsequently analyzed by Davidian and Giltinan (1995) (NLME), Kurada and Chen (2018) (NLMIXED) as well as in Adeniyi et al. (2018). In this experiment, the drug was administered orally to $N = 12$ subjects, and serum concentrations were measured at 11 time points per subject over the subsequent 25 hours. However, as in Davidian and Giltinan (1995), we also excluded here the zero time point from the analysis to simplify the modeling of the within-subject mean variance relationship.

For the analysis, a one-compartment version of the model in (13) in which $\eta_1 = -\eta_3$ was fitted to the data. The resulting pharmacokineticmodel is described by the three parameters $\boldsymbol{\theta}_i = (K_{a_i}, K_{e_i}, Cl)'$, (with $K_{a_i} > K_{e_i}$), representing the absorption rate (1/hr), the elimination rate (1/hr) and fundamental clearance (L/hr), per each of the N individual under study. Often however, the model parametrization is given in term of the compartmental volume, V, where $V = Cl/K_e$ (L). Thus, the mean concentration at time t_{ij}, $(i = 1, \ldots, N, j = 1, \ldots, n)$, following a single dose of size d_0 administrated at time t_{i1}, by the i-th individual, $(i = 1, \ldots, N)$, is,

$$f(t_{ij};\, \boldsymbol{\theta}_i) \equiv \frac{d_0 K_{a_i} K_{ei}}{Cl_i(K_{a_i} - K_{e_i})} (\exp(-K_{e_i} t_{ij}) - \exp(-K_{a_i} t_{ij})). \tag{15}$$

The statistical model accounts for the errors intervening between true and the observed drug concentrations, and with the inter-individual variability in the model's parameters. To deal with the first, it is assumed that for each $i = 1, \ldots, N$,

$$y_{ij} = f(t_{ij};\, \boldsymbol{\theta}_i) + \epsilon_{ij}$$

where y_{ij} is the observed j^{th} drug concentration of the i^{th} individual, obtained at time t_{ij}, and where ϵ_{ij} are some i.i.d random error terms with mean 0 and variance σ^2_ϵ. Here σ^2_ϵ is assumed to be the only intra-individual random effect parameter of concern. Similarly, for modeling the inter-individual variability in the parameters, we assume that $\boldsymbol{\beta_i} := log(\boldsymbol{\theta_i}) \equiv (lK_a, lK_e, lCl)'$, represent some random effects with $E(\boldsymbol{\beta_i}) = \boldsymbol{\beta_0} \equiv (lK_{a0}, lK_{e0}, lCl_0)'$ and where $Var(\boldsymbol{\beta_i}) = \mathbf{D} \equiv diag(\sigma^2_{lK_a}, \sigma^2_{lK_e}, \sigma^2_{lCl})$, for each $i = 1, \ldots, N$. Accordingly, $\boldsymbol{\theta}_0 \equiv exp(\boldsymbol{\beta_0}) = (K_{e0}, K_{a0}, Cl_0)'$ represents the fixed-effect population parameter. In all, there are seven population parameters, namely: K_{a0}, K_{e0}, Cl_0 and $\sigma^2_{K_a}, \sigma^2_{K_e}, \sigma^2_{Cl}$ and σ_ϵ. Because of the logarithmic scale, all these standard deviations are dimensionless quantities and they may be regarded as approximate coefficients of variation. We emphasize that, unlike the other cited approaches (namely NLME and NLMIXED), our modeling here does not depend on any specific distributional assumption (i.e. normality) for the random effects, $\boldsymbol{\beta_i}$ nor for the error terms $\boldsymbol{\epsilon_i}$.

In fact, after standardization to a unit dose, (so that $d_0 = 1$ in (15)) the data on each individual may be viewed as consisting of 10 observations. Using the *Dirichlet*($\mathbf{1}$) weights in $B = 1000$ iterations, we obtained the *recycled* two-stage estimates $\hat{\boldsymbol{\theta}} - RTS^*$ of $\boldsymbol{\theta}_0 = E(\boldsymbol{\theta}_i)$ as well as the STS estimates, $\hat{\boldsymbol{\theta}} - STS$ for these data. The results are presented in Table 7, which also provides the estimates for the variance components in $\mathbf{D}$ and in $\mathbf{D}^*$, as well as the 95% confidence intervals for the fixed parameters K_{e0}, K_{a0} and Cl_0 as where obtained directly from the corresponding *recycled* sampling distributions. For sake of comparison, we also provide in Table 8 the results of NLME estimation procedure as in Lindstrom and Bates (1990) (using `nlme` R package) and those obtained from the NLMIXED estimation procedure as as were reported by Kurada and Chen (2018). We point out again that, while the results in Tables 7 and 8 are largely similar, the estimation procedures utilized in Table 8 (NLME and NLMIXED) hinge on the normality assumptions for the random effect terms (both within and between). In contrast, the result presented in Table 7 using our *recycled* two-stage estimation procedure are entirely free of such specific distributional assumptions.

6 Summary and discussion

We considered the general random weights approach as a viable re-sampling technique in the case of hierarchical nonlinear regression models involving fixed and random effects. We revisit the Standard Two-Stage (STS) estimation procedure for the population parameters, say $\boldsymbol{\theta}_0$, appropriate in this hierarchical nonlinear regression settings. While intuitively appealing, this STS approach was studied in the literature primarily via simulations and with an underlying normality assumption. Here, we establish at first the asymptotic consistency and the asymptotic normality of the STS estimator, $\hat{\boldsymbol{\theta}} - STS$, in the more general context. Our rigorous results, as stated in Theorems 1 and 2, do not hinge on any specific distributional assumptions (e.g., normality) on the random component terms in the model (both errors-terms and random effects), but rather, they are obtained largely

Table 7 The Recycled STS estimation from the PK-data on Theophylline

	Parameter			
	K_a (hr^{-1})	K_e (hr^{-1})	Cl (L/hr)	σ^2_ϵ
$\hat{\boldsymbol{\theta}} - STS$	1.610	0.088	0.040	0.022
$\hat{\boldsymbol{\theta}} - RTS^*$	1.639	0.089	0.040	0.017
D_{ii}	0.588	0.039	0.082	
D^*_{ii}	0.649	0.051	0.085	
95%CI	(1.070, 2.513)	(0.078,0.099)	(0.034, 0.046)	

Table 8 Estimation Result of the PK-data on Theophylline using NLME and NLMIXED (published) procedures

	Parameter		
	K_a (hr^{-1})	K_e (hr^{-1})	Cl (L/hr)
NLMIXED est.	1.617	0.086	0.040
95%CI	(1.038, 2.519)	(0.076,0.096)	(0.035, 0.045)
NLME est.	1.571	0.088	0.040
95%CI	(1.069,2.306)	(0.078,0.099)	(0.034,0.047)

based on minimal moments assumptions. Next, we presented the *recycled* (or the re-sampled) version, $\hat{\boldsymbol{\theta}} - RTS^{*}$, of the STS estimates, $\hat{\boldsymbol{\theta}} - STS$, in this hierarchical nonlinear regression context and established its applicability under general random weighting scheme (Assumption W). In Theorems 3 and 4 we established, the consistency and asymptotic normality of the corresponding re-sampled estimator, $\hat{\boldsymbol{\theta}} - RTS^{*}$. These results enable us to use the *recycled* sampling distribution $\hat{\boldsymbol{\theta}} - RTS^{*}$, as is generated by the re-sampling procedure using the random weights technique to approximate the actual, though unknown, sampling distribution of the STS estimator, $\hat{\boldsymbol{\theta}} - STS$ (see Corollary 5). Thereby allowing us to validly assess the sampling properties of $\hat{\boldsymbol{\theta}} - STS$, such as precision and coverage probabilities based on the re-sampled (via the random weights) data. Toward that end, we augmented our rigorous theoretical results with a detailed simulation study (covering various sample sizes) illustrating the properties of the estimators, $\hat{\boldsymbol{\theta}} - STS$ and $\hat{\boldsymbol{\theta}} - RTS^{*}$ under various scenarios involving normal as well as non-normal error terms and utilizing different choices of random weights (*Multinomial, Dirichlet and Exponential*). Clearly, the effects of the choice of random weights on the numeral minimization (optimization) procedures used by various software, will depend also on the non-linear regression function, its curvatures and the number of data points used. However, this choice could be instructed by experimentation. Additionally, we provided a detailed application of our two-stage *recycled* estimation procedure to the data of the *Theophylline* study, and provided a comparison with the (normality-based) estimation procedures, NLME and NLMIXED. This real-data example, with $N = 12$ and $n = 10$, also illustrates the applicability of our approach even to data involving small sample sizes. In any case, we believe that the gamut of results presented here, both theoretical and numerical, are indicative of the potential and promise of the random weighting *recycled* (re-sampled) STS estimation procedure method to other more complex hierarchical non-linear regression models involving more structured mixed-effects parameters. For instance, extension to cases in which (2) is generalized to $\boldsymbol{\theta}_i = \mathbf{A}_i\boldsymbol{\theta}_0 + \mathbf{B}_i\mathbf{b}_i$, where: $\mathbf{A}_i$, $\mathbf{B}_i$ are some design matrices. However, for sake of scope and space, this and other related issues will have to be pursued elsewhere.

7. Appendix7.1. Technical details and proofs, the STS estimation case

In this section of the "Appendix" we provide the technical results needed for the proofs of Theorems 1 and 2 on the STS estimator $\hat{\theta}_{STS}$ in the hierarchical nonlinear regression model. In the sequel, we let $\phi_{1ij}(\theta) := \phi_{ij}^{'}(\theta)$ (see (11)), and set K to denote a *generic* constant. Recall that (see Assumption A(1)),

$$a_{n_i}^2 := \sigma^2 \sum_{j=1}^{n_i} E(f_{ij}^{'2}(\theta_0 + b_i)) \rightarrow \infty \quad as \quad n_i \rightarrow \infty.$$

Lemma 1 *Under the conditions of Assumption A, for some* $K > 0$

$$a_{n_i}^{-2} \sup_{|t| \leq K} \sum_{j=1}^{n_i} \phi_{1ij}(b_{i1}) - \frac{1}{\sigma^2} \rightarrow 0 \quad a.s.,$$

where $b_{i1} := b_{1n_i}(t)$ is a sequence such that $\sup_{|t| \leq K} |b_{i1} - b_i - \theta_0| \rightarrow 0, \quad a.s.,$ as $n_i \rightarrow \infty$.

Proof of Lemma 1 Since $\phi_{1ij}(\theta) := \phi_{ij}^{'}(\theta)$, we have

$$\phi_{1ij}(\theta) \equiv f_{ij}^{'2}(\theta) - \epsilon_{ij} f_{ij}^{''}(\theta) - (f_{ij}(\theta_0 + b_i) - f_{ij}(\theta)) f_{ij}^{''}(\theta).$$

Accordingly, we first note that,

$$\begin{aligned} \left| a_{n_i}^{-2} \sup_{|t| \leq K} \sum_{j=1}^{n_i} \phi_{1ij}(b_{i1}) - \frac{1}{\sigma^2} \right| \leq & \left| a_{n_i}^{-2} \sup_{|t| \leq K} \sum_{j=1}^{n_i} f_{ij}^{'2}(b_{i1}) - \frac{1}{\sigma^2} \right| \\ & + a_{n_i}^{-2} \sup_{|t| \leq K} \left| \sum_{j=1}^{n_i} \epsilon_{ij} f_{ij}^{''}(b_{i1}) \right| \\ & + a_{n_i}^{-2} \sup_{|t| \leq K} \left| \sum_{j=1}^{n_i} (f_{ij}(\theta_0 + b_i) - f_{ij}(b_{i1})) f_{ij}^{''}(b_{i1}) \right|. \end{aligned}$$

By Assumption A (3), we have $a_{n_i}^{-2} \sup_{|t| \leq K} \sum_{j=1}^{n_i} f_{ij}^{'2}(b_{i1}) - \frac{1}{\sigma^2} \rightarrow 0 \quad a.s.$, and by Assumption A (2) and Corollary A in Wu (1981), we also have,

$$a_{n_i}^{-2} \sup_{|t| \leq K} \left| \sum_{j=1}^{n_i} \epsilon_{ij} f_{ij}^{''}(b_{i1}) \right| \rightarrow 0 \quad a.s..$$

Finally, the last term converge to 0 a.s. by Assumption A, an application of Cauchy-Schwarz inequality and Corollary A in Wu (1981). Thus we have

$$a_{n_i}^{-2}\sup_{|t|\leq K}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1})-\frac{1}{\sigma^2}\to 0 \quad a.s..$$

. □

Lemma 2 *Let X_i be a sequence of random variables bounded in probability and let Y_i be a sequence of random variables which satisfies $\frac{1}{n}\sum_{i=1}^n|Y_i|\to 0$ in probability. Then $\frac{1}{n}\sum_{i=1}^n X_iY_i\xrightarrow{p}0$.*

Proof of Lemma 2 Since X_i is bounded in probability, for any $\epsilon>0$, there is K_ϵ such that with sufficient large i, $P(|X_i|>K_\epsilon)<\epsilon$. Then

$$\begin{aligned}\lim_{n\to\infty}P(|\frac{1}{n}\sum_{i=1}^n X_iY_i|>\epsilon)=&\lim_{n\to\infty}\left[P(|\frac{1}{n}\sum_{i=1}^n X_iY_i|>\epsilon,|X_i|<K_\epsilon)\right]\\&+\lim_{n\to\infty}\left[P(|\frac{1}{n}\sum_{i=1}^n X_iY_i|>\epsilon,|X_i|>K_\epsilon)\right]\\&\leq\lim_{n\to\infty}P(\frac{1}{n}\sum_{i=1}^n|\frac{X_i}{K_\epsilon}Y_i|>\frac{\epsilon}{K_\epsilon},|X_i|<K_\epsilon)+\epsilon\\&\leq\lim_{n\to\infty}P(\frac{1}{n}\sum_{i=1}^n|Y_i|>\frac{\epsilon}{K_\epsilon},|X_i|<K_\epsilon)+\epsilon=\epsilon,\end{aligned}$$

from which the desired result follows. □

Lemma 3 *There exists a $K>0$ such that for any $\epsilon>0$, for any i,*

$$P\left[\left|a_{n_i}^{-1}\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right|>K\right]<\frac{\epsilon}{2}.$$

Proof of Lemma 3 Since ϵ_{ij} and b_i are independent, for each $i=1,\ldots,N$, we have that for any $j_1\neq j_2$,

$$\begin{aligned}E(\phi_{ij_1}(\theta_0+b_i)\phi_{ij_2}(\theta_0+b_i))=&E[E(\phi_{ij_1}(\theta_0+b_i)\phi_{ij_2}(\theta_0+b_i)|b_i)]\\=&E[E(\epsilon_{ij_1}\epsilon_{ij_2}f'_{ij_1}(\theta_0+b_i)f'_{ij_2}(\theta_0+b_i)|b_i)]\\=&E[E(\epsilon_{ij_1})E(\epsilon_{ij_2})f'_{ij_1}(\theta_0+b_i)f'_{ij_2}(\theta_0+b_i)]\\=&0.\end{aligned}$$

Similarly,

$$E(\phi_{ij_1}(\theta_0+b_i))=E[E(\epsilon_{ij_1}f'_{ij_1}(\theta_0+b_i)|b_i)]=E[E(\epsilon_{ij_1})f'_{ij_1}(\theta_0+b_i)]=0.$$

Hence, we have, $E(\phi_{ij_1}(\theta_0+b_i)\phi_{ij_2}(\theta_0+b_i))=E(\phi_{ij_1}(\theta_0+b_i))E(\phi_{ij_2}(\theta_0+b_i))$. To conclude that,

$$
\begin{aligned}
Var\left(\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right) &= \sum_{j=1}^{n_i} Var(\phi_{ij}(\theta_0+b_i)) \\
&= \sum_{j=1}^{n_i} Var(\epsilon_{ij} f_{ij}^{'}(\theta_0+b_i)) \\
&= \sum_{j=1}^{n_i} E(\epsilon_{ij}^2)E(f_{ij}^{'2}(\theta_0+b_i)) \\
&= \sigma^2 \sum_{j=1}^{n_i} E(f_{ij}^{'2}(\theta_0+b_i)) \equiv a_{n_i}^2.
\end{aligned}
$$

Accordingly, there exists a $K>0$ such that for any $\epsilon>0$, for any i,

$$
P\left[\left|a_{n_i}^{-1}\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right|>K\right]<\frac{\epsilon}{2}.
$$

□

We are now ready to prove Theorem 1

Proof of Theorem 1 Let

$$
S_{n_i}(t) := a_{n_i}^{-1}\sum_{j=1}^{n_i}\left[\phi_{ij}(\theta_0+b_i+a_{n_i}^{-1}t)-\phi_{ij}(\theta_0+b_i)\right]-\frac{t}{\sigma^2}. \tag{16}
$$

Next we will show for any given constant K,

$$
\sup_{|t|\le K}|S_{n_i}(t)|\to 0 \quad a.s. \tag{17}
$$

By a Taylor expansion, $\phi_{ij}(\theta_0+b_i+a_{n_i}^{-1}t)=\phi_{ij}(\theta_0+b_i)+\phi_{1ij}(b_{i1})a_{n_i}^{-1}t$, where $b_{i1}=\theta_0+b_i+ca_{n_i}^{-1}t$ for some $0<c<1$. Accordingly we obtain that,

$$
\begin{aligned}
\sup_{|t|\le K}|S_{n_i}(t)| &= \sup_{|t|\le K}\left|a_{n_i}^{-1}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1})a_{n_i}^{-1}t-\frac{t}{\sigma^2}\right| \\
&= K\left|a_{n_i}^{-2}\sup_{|t|\le K}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1})-\frac{1}{\sigma^2}\right|.
\end{aligned}
$$

By Lemma 1, $a_{n_i}^{-2}\sup_{|t|\le K}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1})-\frac{1}{\sigma^2}\to 0 \;\; a.s.$ Thus, we have proved (17). Next, by (16),

$$
A_{n_i}(t) := a_{n_i}^{-1}t\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i+a_{n_i}^{-1}t) = tS_{n_i}(t)+a_{n_i}^{-1}t\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)+\frac{t^2}{\sigma^2}.
$$

Thus,

$$\inf_{|t|=K} A_{n_i}(t) \geq -K\sup_{|t|=K}|S_{n_i}(t)| - Ka_{n_i}^{-1}\left|\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right| + \frac{K^2}{\sigma^2}.$$

By lemma 3 there exists a $K > 0$ such that for any $\epsilon > 0$, for any i,

$$P\left[\left|a_{n_i}^{-1}\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right| > K\right] < \frac{\epsilon}{2}. \tag{18}$$

So that by (18) and (17) we may choose K large enough such that for sufficiently large n_i,

$$\begin{aligned}
P(\inf_{|t|=K} A_{n_i}(t) \geq 0) \geq & P(\sup_{|t|=K}|S_{n_i}(t)| + a_{n_i}^{-1}\left|\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right| \leq \frac{K}{\sigma^2}) \\
=& 1 - P\left(\sup_{|t|=K}|S_{n_i}(t)| + a_{n_i}^{-1}\left|\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right| > \frac{K}{\sigma^2}\right) \\
\geq & 1 - P\left(\sup_{|t|=K}|S_{n_i}(t)| > \frac{K}{4\sigma^2}) - P(a_{n_i}^{-1}\left|\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i)\right| > \frac{K}{4\sigma^2}\right) \\
\geq & 1 - \epsilon.
\end{aligned}$$

By the continuity of $\sum_{j=1}^{n_i}\phi_{ij}(\theta)$ in θ, we have, for sufficiently large n_i, that there exists a constant K such that the equation

$$\sum_{j=1}^{n_i}\phi_{ij}(\theta_0 + b_i + a_{n_i}^{-1}t) = 0,$$

has a root $t = T_{ni}$ in $|t| \leq K$ with probability larger than $1-\epsilon$. That is, we have $\hat{\theta}_{ni} = \theta_0 + b_i + a_{ni}^{-1}T_{ni}$, where $|T_{ni}| < K$ in probability. Thus, by Lemma 2,

$$\hat{\theta}_{STS} - \theta_0 = \frac{1}{N}\sum_{i=1}^{N} b_i + \frac{1}{N}\sum_{i=1}^{N} a_{ni}^{-1}T_{ni} \xrightarrow{p} 0.$$

□

For establishing the asymptotic normality result as stated in Theorem 2, we need the following Lemma.

Lemma 4 *Under the conditions of Assumptions A,*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_{n_i}^{-2}\sum_{j=1}^{n_i}\phi_{ij}(\theta_0+b_i) \xrightarrow{p} 0.$$

Proof of Lemma 4 Let $X_{ni} := a_{n_i}^{-1} \sum_{j=1}^{n_i} \phi_{ij}(\theta_0 + b_i)$, where, by proof of Theorem 1 we have $E(X_{ni}) = 0$ and $Var(X_{ni}) = 1$. Thus,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-2} \sum_{j=1}^{n_i} \phi_{ij}(\theta_0 + b_i) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-1} X_{ni}.$$

Now, for any $\epsilon > 0$,

$$P\left(\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-1} X_{ni} \right| > \epsilon\right) \leq \frac{\sum_{i=1}^{N} \frac{1}{a_{ni}^2}}{N\epsilon^2} \to 0.$$

Accordingly, we have $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-2} \sum_{j=1}^{n_i} \phi_{ij}(\theta_0 + b_i) \xrightarrow{p} 0$, as required. □

Proof of Theorem 2 We first note that by Lemma 1 and (16),

$$\hat{\theta}_{ni} - \theta_0 - b_i = -a_{ni}^{-2} \sigma^2 \sum_{j=1}^{n_i} \phi_{ij}(\theta_0 + b_i) - a_{n_i}^{-1} \sigma^2 S_{n_i}(T_{ni}).$$

Thus,

$$\hat{\theta}_{STS} - \theta_0 = \frac{1}{N} \sum_{i=1}^{N} b_i - \frac{\sigma^2}{N} \sum_{i=1}^{N} a_{n_i}^{-2} \sum_{j=1}^{n_i} \phi_{ij}(\theta_0 + b_i) - \frac{\sigma^2}{N} \sum_{i=1}^{N} a_{n_i}^{-1} S_{n_i}(T_{ni}).$$

Recall that $\sum_{i=1}^{N} b_i / N \to E(b_1) \equiv 0$. In view of (17) and since, $\lim_{N, ni \to \infty} N / a_{ni}^2 < \infty$, we have

$$\frac{\sigma^2}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-1} S_{n_i}(T_{ni}) \to 0 \quad a.s..$$

Finally, from Lemma 4,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-2} \sum_{j=1}^{n_i} \phi_{ij}(\theta_0 + b_i) \xrightarrow{p} 0.$$

Thus, it follows that $\lambda^{-1} \sqrt{N}(\hat{\theta}_{STS} - \theta_0) \Rightarrow \mathcal{N}(0, 1)$. □

7.2. Technical details and proofs, the recycled STS estimation case

In this section of the "Appendix" we provide the technical results needed for the proofs of Theorems 3 and 4 on the *recycled* STS estimator, $\hat{\theta}^*_{RTS}$, in the hierarchical nonlinear regression model. We begin with a re-statement of Lemma 2 from Boukai and Zhang (2018) which is concerned with the general random weights under Assumption W.

Lemma 5 *Let* $\mathbf{w}_n = (w_{1:n}, w_{1:n}, \ldots, w_{n:n})^{\mathbf{t}}$ *be random weights that satisfy the conditions of Assumption W. Then With* $W_i = (w_{i:n} - 1)/\tau_n,\ i = 1\ldots, n$ *and* $\bar{W}_n := \frac{1}{n}\sum_{i=1}^{n} W_i$ *we have, as* $n \to \infty$*, that (i)* $\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{p^*} 0$ *(ii)* $\frac{1}{n}\sum_{i=1}^{n} W_i^2 \xrightarrow{p^*} 1$ *and hence (iii)* $\frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W}_n)^2 \xrightarrow{p^*} 1$.

Lemma 6 *Under the conditions of Assumption W,* $\frac{1}{n}\sum_{i=1}^{n} w_{i:n} - 1 \xrightarrow{p^*} 0$*, Further, let* $\mathbf{u}_n = (u_1, u_2, \ldots, u_n)^{\mathbf{t}}$ *denote a vector of n i.i.d random variables that is independent of* $\mathbf{w}_n$ *with* $E(u_i) = 0$, $E(u_i^2) < \infty$*. Then, conditional on the given value of the* $\mathbf{u}_n$*, we have* $\frac{1}{n}\sum_{i=1}^{n} u_i w_{i:n} \xrightarrow{p^*} 0$*, as* $n \to \infty$.

Proof of Lemma 6 We first note that

$$\begin{aligned} E^*\left(\frac{1}{n}\sum_{i=1}^{n}(w_{i:n} - 1)\right)^2 =& E^*\left(\frac{\tau_n}{n}\sum_{i=1}^{n} W_i\right)^2 \\ =& \frac{\tau_n^2}{n^2}\sum_{i=1}^{n} E^*(W_i^2) + \frac{\tau_n^2}{n^2}\sum_{i_1 \neq i_2} E^*(W_{i_1}W_{i_2}) \\ =& \frac{\tau_n^2}{n} + \frac{\tau_n^2}{n^2} n(n-1) O\left(\frac{1}{n}\right) \to 0, \quad \text{as} \quad n \to \infty. \end{aligned}$$

To conclude that, $\frac{1}{n}\sum_{i=1}^{n} w_i - 1 \xrightarrow{p^*} 0$, as $n \to \infty$. As for the second assertion, we note that since

$$\frac{1}{n}\sum_{i=1}^{n} u_i w_{i:n} = \frac{\tau_n}{n}\sum_{i=1}^{n} u_i W_i + \frac{1}{n}\sum_{i=1}^{n} u_i,$$

and since $\sum_{i=1}^{n} u_i/n \to 0$, as $n \to \infty$, we may only consider the first term. To that end, we note that

$$\begin{aligned} E^*\left(\frac{\tau_n}{n}\sum_{i=1}^{n} u_i W_i\right)^2 =& \frac{\tau_n^2}{n^2}\sum_{i=1}^{n} E^*(u_i^2 W_i^2) + \frac{\tau_n^2}{n^2}\sum_{i_1 \neq i_2} E^*(W_{i_1}W_{i_2}u_{i_1}u_{i_2}) \\ \leq& \left[1 + (n-1)O\left(\frac{1}{n}\right)\right]\frac{\tau_n^2}{n^2}\sum_{i=1}^{n} u_i^2 \to 0, \end{aligned}$$

as $n \to \infty$. We therefore conclude that $\frac{1}{n}\sum_{i=1}^{n} u_i w_{i:n} \xrightarrow{p^*} 0$, as required. □

Lemma 7 *Under the conditions of Assumptions Aand B, we have that* $a_{n_i}^{-2}\sum_{j=1}^{n_i}\phi_{ij}^2(\hat{\theta}_{ni}) \xrightarrow{p} 1$*, for all* $i = 1, 2, \ldots, N$.

Proof of Lemma 7: Since $\hat{\theta}_{ni} \xrightarrow{p} \theta_0$, we have

$$\begin{aligned} a_{n_i}^{-2}\sum_{j=1}^{n_i}\phi_{ij}^2(\hat{\theta}_{ni}) =& a_{n_i}^{-2}\sum_{j=1}^{n_i}(y_{ij}-f_{ij}(\hat{\theta}_{ni}))^2 f_{ij}'^2(\hat{\theta}_{ni}) \\ =& a_{n_i}^{-2}\sum_{j=1}^{n_i}\epsilon_{ij}^2 f_{ij}'^2(\hat{\theta}_{ni}) + a_{n_i}^{-2}\sum_{j=1}^{n_i}(f_{ij}(\theta_0+b_i)-f_{ij}(\hat{\theta}_{ni}))^2 f_{ij}'^2(\hat{\theta}_{ni}) \\ &+2a_{n_i}^{-2}\sum_{j=1}^{n_i}\epsilon_{ij}(f_{ij}(\theta_0+b_i)-f_{ij}(\hat{\theta}_{ni}))f_{ij}'^2(\hat{\theta}_{ni}) \\ \equiv& B_1+B_2+B_3. \end{aligned}$$

Write,

$$B_1 = a_{n_i}^{-2}\sum_{j=1}^{n_i}(\epsilon_{ij}^2-\sigma^2)f_{ij}'^2(\hat{\theta}_{ni}) + a_{n_i}^{-2}\sigma^2\sum_{j=1}^{n_i}f_{ij}'^2(\hat{\theta}_{ni}).$$

The first term in B_1 converges to 0 by Assumption A (3), and Corollary A of Wu (1981) while the second term in B_1 converges to 1 by Assumption A(3). Hence $B_1 \xrightarrow{p} 1$. As for the second and third terms, B_2 and B_3, it follows by a direct application of the Cauchy-Schwarz inequality ogether with Assumption B (1), that $B_2 \xrightarrow{p} 0$ and $B_3 \xrightarrow{p} 0$. Accordingly, it follows that $a_{n_i}^{-2}\sum_{j=1}^{n_i}\phi_{ij}^2(\hat{\theta}_{ni}) \xrightarrow{p} 1$, as required. □

Lemma 8 *Under the conditions of Assumptions A and B, for all i,*

$$E^*\left[\tau_{n_i}a_{n_i}^{-2}\sup_{|t|\le K\tau_{n_i}}\sum_{j=1}^{n_i}W_{ij}\phi_{1ij}(b_{i1}^*)\right]^2 \to 0$$

where $b_{i1}^* = \hat{\theta}_{ni} + ca_{n_i}^{-1}t$ for some $0<c<1$, as $n_i\to\infty$.

Proof of Lemma 8 We first note that since by Theorem 1, we have $\hat{\theta}_{ni}-b_i-\theta_0 \xrightarrow{p} 0$, and since

$$\begin{aligned} |b_{i1}^*-b_i-\theta_0| =& |\hat{\theta}_{ni}-b_i-\theta_0+ca_{n_i}^{-1}t| \\ &\le |\hat{\theta}_{ni}-b_i-\theta_0| + \frac{c\tau_{n_i}}{\sqrt{n_i}}\frac{\sqrt{n_i}}{a_{n_i}}\frac{|t|}{\tau_{n_i}}, \end{aligned}$$

it follows under Assumption B (3) that with $|t|\le K\tau_{n_i}$, we have $b_{i1}^*-b_i-\theta_0 \xrightarrow{p} 0$. Thus,

$$
\begin{aligned}
& E^*[\tau_{n_i} a_{n_i}^{-2} \sup_{|t| \le K\tau_{n_i}} \sum_{j=1}^{n_i} W_{ij}\phi_{1ij}(b_{i1}^*)]^2 \\
\le & \tau_{n_i}^2 a_{n_i}^{-4} \sup_{|t| \le K\tau_{n_i}} [\sum_{j=1}^{n_i} \phi_{1ij}^2(b_{i1}^*) + O(\frac{1}{n_i}) \sum_{j_1 \ne j_2} \phi_{1ij_1}(b_{i1}^*)\phi_{1ij_2}(b_{i1}^*)] \\
\le & \tau_{n_i}^2 a_{n_i}^{-4} \sup_{|t| \le K\tau_{n_i}} [\sum_{j=1}^{n_i} \phi_{1ij}^2(b_{i1}^*) + O(\frac{1}{n_i})(n_i - 1) \sum_{j=1}^{n_i} \phi_{1ij}^2(b_{i1}^*)] \\
= & \tau_{n_i}^2 a_{n_i}^{-4} [O(\frac{1}{n_i})(n_i - 1) + 1] \sup_{|t| \le K\tau_{n_i}} \sum_{j=1}^{n_i} \phi_{1ij}^2(b_{i1}^*).
\end{aligned}
$$

In light of Assumption B (2–3) , and that $\tau_{n_i}^2/n_i \to 0$, we only need to show, in order to complete the proof of Lemma 8, that

$$
\lim_{n_i \to \infty} \quad a_{n_i}^{-2} \sup_{|t| \le K\tau_{n_i}} \sum_{j=1}^{n_i} \phi_{1ij}^2(b_{i1}^*) < \infty.
$$

Toward that end, we note that,

$$
\begin{aligned}
& a_{n_i}^{-2} \sup_{|t| \le K\tau_{n_i}} \sum_{j=1}^{n_i} \phi_{1ij}^2(b_{i1}^*) \\
= & a_{n_i}^{-2} \sup_{|t| \le K\tau_{n_i}} \sum_{j=1}^{n_i} (f_{ij}^{\prime 2}(b_{i1}^*) - (y_{ij} - f_{ij}(b_{i1}^*))f_{ij}^{\prime\prime}(b_{i1}^*))^2 \\
\le & a_{n_i}^{-2} \sup_{|t| \le K\tau_{n_i}} \sum_{j=1}^{n_i} f_{ij}^{\prime 4}(b_{i1}^*) + a_{n_i}^{-2} \sup_{|t| \le K\tau_{n_i}} \sum_{j=1}^{n_i} (y_{ij} - f_{ij}(b_{i1}^*))^2 f_{ij}^{\prime\prime 2}(b_{i1}^*) \\
& + 2a_{n_i}^{-2} \sup_{|t| \le K\tau_{n_i}} \left| \sum_{j=1}^{n_i} f_{ij}^{\prime 2}(b_{i1}^*)(y_{ij} - f_{ij}(b_{i1}^*))f_{ij}^{\prime\prime}(b_{i1}^*) \right| \\
\equiv & I_1 + I_2 + I_3.
\end{aligned}
$$

It is straight forward to see that by Assumption B (1), $\lim_{n_i \to \infty} I_1 < \infty$, and that by Cauchy-Schwarz inequality $\lim_{n_i \to \infty} I_3 < \infty$. Finally we write

$$\begin{aligned} I_2 =& a_{n_i}^{-2} \sup_{|t|\le K\tau_{n_i}} \sum_{j=1}^{n_i} (\epsilon_{ij}^2 - \sigma^2) f_{ij}''^2(b_{i1}^*) + a_{n_i}^{-2} \sup_{|t|\le K\tau_{n_i}} \sum_{j=1}^{n_i} \sigma^2 f_{ij}''^2(b_{i1}^*) \\ &+ a_{n_i}^{-2} \sup_{|t|\le K\tau_{n_i}} \sum_{j=1}^{n_i} (f_{ij}(\theta_0 + b_i) - f_{ij}(b_{i1}^*))^2 f_{ij}''^2(b_{i1}^*) \\ &+ 2a_{n_i}^{-2} \sup_{|t|\le K\tau_{n_i}} \left| \sum_{j=1}^{n_i} \epsilon_{ij}(f_{ij}(\theta_0 + b_i) - f_{ij}(b_{i1}^*)) f_{ij}''^2(b_{i1}^*) \right|. \end{aligned}$$

The first term converges to 0 in probability by Assumption B (2) and Corollary A of Wu (1981). Then, according to Assumption A (2),

$$\lim_{n_i\to\infty} a_{n_i}^{-2} \sup_{|t|\le K\tau_{n_i}} \sum_{j=1}^{n_i} \sigma^2 f_{ij}''^2(b_{i1}^*) < \infty.$$

The third term in I_2 converges to 0 in probability by an application of the Cauchy-Schwarz inequality combined with Assumption B (1) and (2). Finally, the fourth term in I_2, converges to 0 in probability again, by an application of the Cauchy-Schwarz inequality. Thus we have $\lim_{n_i\to\infty} I_2 < \infty$. Accordingly, we have established that as $n_i \to \infty$,

$$E^* \left[\tau_{n_i} a_{n_i}^{-2} \sup_{|t|\le K\tau_{n_i}} \sum_{j=1}^{n_i} W_{ij} \phi_{1ij}(b_{i1}^*) \right]^2 \to 0.$$

□

Lemma 9 *Under the conditions of Assumptions A and B, there exists a* $K > 0$ *such that for any* $\epsilon > 0$,

$$P^* \left[\left| a_{n_i}^{-1} \sum_{j=1}^{n_i} W_{ij} \phi_{ij}(\hat{\theta}_{ni}) \right| > K \right] < \frac{\epsilon}{2}.$$

Proof of Lemma 9 By Lemma 7,

$$
\begin{aligned}
&V^*(a_{ni}^{-1}\sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni}))\\
=&a_{ni}^{-2}\sum_{j=1}^{n_i}\phi_{ij}^2(\hat{\theta}_{ni}) + a_{ni}^{-2}O(\frac{1}{n_i})\sum_{j_1\neq j_2}\phi_{ij_1}(\hat{\theta}_{ni})\phi_{ij_2}(\hat{\theta}_{ni})\\
=&a_{ni}^{-2}\sum_{j=1}^{n_i}\phi_{ij}^2(\hat{\theta}_{ni}) + a_{ni}^{-2}O(\frac{1}{n_i})(\sum_{j=1}^{n_i}\phi_{ij}(\hat{\theta}_{ni}))^2 - a_{ni}^{-2}O(\frac{1}{n_i})\sum_{j=1}^{n_i}\phi_{ij}^2(\hat{\theta}_{ni})\\
\leq& a_{ni}^{-2}(1-O(\frac{1}{n_i}))\sum_{j=1}^{n_i}\phi_{ij}^2(\hat{\theta}_{ni})\xrightarrow{p}1.
\end{aligned}
$$

Hence we obtain,

$$
P^*(\left|a_{ni}^{-1}\sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni})\right| > K) \leq \frac{V^*(a_{ni}^{-1}\sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni}))}{K^2}\xrightarrow{p}\frac{1}{K^2}.
$$

Accordingly, there exists a $K > 0$ such that for any $\epsilon > 0$,

$$
P^*\left[\left|a_{n_i}^{-1}\sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni})\right| > K\right] < \frac{\epsilon}{2}.
$$

□

Proof of Theorem 3 Let

$$
S_{n_i}^*(t) := a_{n_i}^{-1}\sum_{j=1}^{n_i} w_{ij}\left[\phi_{ij}(\hat{\theta}_{ni} + a_{n_i}^{-1}t) - \phi_{ij}(\hat{\theta}_{ni})\right] - \frac{t}{\sigma^2}. \tag{19}
$$

First, we will show that for any given $K > 0$,

$$
E^*\left[\tau_{n_i}^{-1}\sup_{|t|\leq K\tau_{n_i}}|S_{n_i}^*(t)|\right]^2 \xrightarrow{p^*} 0. \tag{20}
$$

By a Taylor expansion we have that $\phi_{ij}(\hat{\theta}_{ni} + a_{n_i}^{-1}t) = \phi_{ij}(\hat{\theta}_{ni}) + \phi_{1ij}(b_{i1}^*)a_{n_i}^{-1}t$, where as before, $b_{i1}^* = \hat{\theta}_{ni} + ca_{n_i}^{-1}t$ for some $0<c<1$. Accordingly we obtain,

$$
\begin{aligned}
\tau_{n_i}^{-1}\sup_{|t|\leq K\tau_{n_i}}|S_{n_i}^*(t)| =& \tau_{n_i}^{-1}\sup_{|t|\leq K\tau_{n_i}}\left|a_{n_i}^{-1}\sum_{j=1}^{n_i} w_{ij}\phi_{1ij}(b_{i1}^*)a_{n_i}^{-1}t - \frac{t}{\sigma^2}\right|\\
=& K\left|a_{n_i}^{-2}\sup_{|t|\leq K\tau_{n_i}}\sum_{j=1}^{n_i} w_{ij}\phi_{1ij}(b_{i1}^*) - \frac{1}{\sigma^2}\right|\\
\leq& K\left|\tau_{n_i}a_{n_i}^{-2}\sup_{|t|\leq K\tau_{n_i}}\sum_{j=1}^{n_i} W_{ij}\phi_{1ij}(b_{i1}^*)\right| + K\left|a_{n_i}^{-2}\sup_{|t|\leq K\tau_{n_i}}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1}^*) - \frac{1}{\sigma^2}\right|.
\end{aligned}
$$

Further,

$$E^*\left[\tau_{n_i}^{-1}\sup_{|t|\le K\tau_{n_i}}|S_{n_i}^*(t)|\right]^2 \le K^2E^*\left|\tau_{n_i}a_{n_i}^{-2}\sup_{|t|\le K\tau_{n_i}}\sum_{j=1}^{n_i}W_{ij}\phi_{1ij}(b_{i1}^*)\right|^2$$
$$+K^2E^*\left|a_{n_i}^{-2}\sup_{|t|\le K\tau_{n_i}}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1}^*)-\frac{1}{\sigma^2}\right|^2$$
$$+K^2E^*\left|\tau_{n_i}a_{n_i}^{-2}\sup_{|t|\le K\tau_{n_i}}\sum_{j=1}^{n_i}W_{ij}\phi_{1ij}(b_{i1}^*)\right|\left|a_{n_i}^{-2}\sup_{|t|\le K\tau_{n_i}}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1}^*)-\frac{1}{\sigma^2}\right|.$$

By Lemma 8 and Lemma 1, we have

$$E^*\left|\tau_{n_i}a_{n_i}^{-2}\sup_{|t|\le K\tau_{n_i}}\sum_{j=1}^{n_i}W_{ij}\phi_{1ij}(b_{i1}^*)\right|^2\to 0,$$

and

$$E^*\left|a_{n_i}^{-2}\sup_{|t|\le K\tau_{n_i}}\sum_{j=1}^{n_i}\phi_{1ij}(b_{i1}^*)-\frac{1}{\sigma^2}\right|^2\to 0.$$

Thus, by an application of the Cauchy-Schwarz inequality we have proved (20). Next, in light of (19) we define

$$A_{n_i}^*(t) := a_{n_i}^{-1}t\sum_{j=1}^{n_i}w_{ij}\phi_{ij}(\hat{\theta}_{ni}+a_{n_i}^{-1}t) = tS_{n_i}^*(t)+a_{n_i}^{-1}t\sum_{j=1}^{n_i}w_{ij}\phi_{ij}(\hat{\theta}_{ni})+\frac{t^2}{\sigma^2}.$$

Accordingly,

$$\inf_{|t|=K\tau_{n_i}}A_{n_i}^*(t)\ge -K\tau_{n_i}\sup_{|t|=K\tau_{n_i}}|S_{n_i}^*(t)|-K\tau_{n_i}a_{n_i}^{-1}\left|\sum_{j=1}^{n_i}w_{ij}\phi_{ij}(\hat{\theta}_{ni})\right|+\frac{K^2\tau_{n_i}^2}{\sigma^2}.$$

Recall that by Lemma 9, there exists a $K>0$ such that for any $\epsilon>0$,

$$P^*\left[\left|a_{n_i}^{-1}\sum_{j=1}^{n_i}W_{ij}\phi_{ij}(\hat{\theta}_{ni})\right|>K\right]<\frac{\epsilon}{2}. \tag{21}$$

Accordingly, by (21) and (20) we may choose large enough K such that for sufficiently large n_i,

$$
\begin{aligned}
P^*\left(\inf_{|t|=K\tau_{n_i}} A_{n_i}(t) \geq 0\right) &\geq P^*\left[\sup_{|t|=K\tau_{n_i}} |S^*_{n_i}(t)| + a_{n_i}^{-1}\left|\sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni})\right| \leq \frac{K\tau_{n_i}}{\sigma^2}\right] \\
&= P^*\left[\sup_{|t|=K\tau_{n_i}} |S^*_{n_i}(t)| + a_{n_i}^{-1}\tau_{n_i}\left|\sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni})\right| \leq \frac{K\tau_{n_i}}{\sigma^2}\right] \\
&= 1 - P^*\left[\sup_{|t|=K\tau_{n_i}} |S^*_{n_i}(t)| + a_{n_i}^{-1}\tau_{n_i}\left|\sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni})\right| > \frac{K\tau_{n_i}}{\sigma^2}\right] \\
&\geq 1 - P^*\left[\tau_{n_i}^{-1}\sup_{|t|=K\tau_{n_i}} |S^*_{n_i}(t)| > \frac{K}{4\sigma^2}\right] - P^*\left[a_{n_i}^{-1}\left|\sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni})\right| > \frac{K}{4\sigma^2}\right] \\
&\geq 1 - \epsilon.
\end{aligned}
$$

From the continuity of $\sum_{j=1}^{n_i}\phi_{ij}(\theta)$ in θ, we have for sufficiently large n_i, that there exists a K such that the equation $\sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni} + a_{n_i}^{-1}t) = 0$, has a root, $t = T^*_{ni}$ in $|t| \leq K\tau_{n_i}$, with a probability larger than $1-\epsilon$. That is, we have $\hat{\theta}^*_{ni} = \hat{\theta}_{ni} + a_{ni}^{-1}T^*_{ni}$, where $|\tau_{n_i}^{-1}T^*_{ni}| < K$ in probability. Accordingly we may rewrite $\hat{\theta}^*_{RTS}$ as,

$$
\begin{aligned}
\hat{\theta}^*_{RTS} &= \frac{1}{N}\sum_{i=1}^{N} u_i\hat{\theta}_{ni} + \frac{1}{N}\sum_{i=1}^{N} u_i a_{ni}^{-1}T^*_{ni} \\
&= \frac{1}{N}\sum_{i=1}^{N} u_i(\theta_0 + b_i + a_{ni}^{-1}T_{ni}) + \frac{1}{N}\sum_{i=1}^{N} u_i a_{ni}^{-1}T^*_{ni} \\
&= \frac{1}{N}\sum_{i=1}^{N} u_i\theta_0 + \frac{1}{N}\sum_{i=1}^{N} u_i b_i + \frac{1}{N}\sum_{i=1}^{N} u_i a_{ni}^{-1}T_{ni} + \frac{1}{N}\sum_{i=1}^{N} u_i a_{ni}^{-1}T^*_{ni}.
\end{aligned}
$$

That is,

$$
\hat{\theta}^*_{RTS} - \theta_0 = \frac{1}{N}\sum_{i=1}^{N}(u_i - 1)\theta_0 + \frac{1}{N}\sum_{i=1}^{N} u_i b_i + \frac{1}{N}\sum_{i=1}^{N} u_i a_{ni}^{-1}T_{ni} + \frac{1}{N}\sum_{i=1}^{N} u_i a_{ni}^{-1}T^*_{ni}.
$$

Additionally, by Lemma 6, we have $\frac{1}{N}\sum_{i=1}^{N}(u_i - 1)\xrightarrow{p^*}0$, as well as, $\frac{1}{N}\sum_{i=1}^{N} u_i b_i \xrightarrow{p^*} 0$. Further, we also have that

$$
\frac{1}{N}\sum_{i=1}^{N} u_i a_{ni}^{-1}T_{ni} = \frac{1}{N}\sum_{i=1}^{N}(u_i - 1)a_{ni}^{-1}T_{ni} + \frac{1}{N}\sum_{i=1}^{N} a_{ni}^{-1}T_{ni}.
$$

Now by Lemma 2 and the fact $T_{ni} = O_p(1)$, we obtain, with $U_i := (u_i - 1)/\tau_N$, that

$$E^*(\frac{1}{N}\sum_{i=1}^{N}(u_i-1)a_{ni}^{-1}T_{ni})^2 = E^*(\frac{\tau_N}{N}\sum_{i=1}^{N}U_i a_{ni}^{-1}T_{ni})^2$$
$$\leq \frac{\tau_N^2}{N^2}\sum_{i=1}^{N}a_{ni}^{-2}T_{ni}^2 + (N-1)O(\frac{1}{N})\frac{\tau_N^2}{N^2}\sum_{i=1}^{N}a_{ni}^{-2}T_{ni}^2 \xrightarrow{p} 0,$$

as well as, $\frac{1}{N}\sum_{i=1}^{N}a_{ni}^{-1}T_{ni}\xrightarrow{p}0$. That is, we have established that, $E^*(\frac{1}{N}\sum_{i=1}^{N}u_i a_{ni}^{-1}T_{ni})^2\xrightarrow{p}0$. Accordingly we conclude, $P^*(|\frac{1}{N}\sum_{i=1}^{N}u_i a_{ni}^{-1}T_{ni}|>\epsilon) = o_p(1)$. Similarly,

$$\frac{1}{N}\sum_{i=1}^{N}u_i a_{ni}^{-1}T_{ni}^* = \frac{1}{N}\sum_{i=1}^{N}(u_i-1)a_{ni}^{-1}T_{ni}^* + \frac{1}{N}\sum_{i=1}^{N}a_{ni}^{-1}T_{ni}^*,$$

where by Lemma 2, Assumption B (3) and the fact $\tau_{n_i}^{-1}T_{ni}^* = O_{p^*}(1)$, we obtain,

$$E^*(\frac{1}{N}\sum_{i=1}^{N}(u_i-1)a_{ni}^{-1}T_{ni}^*)^2 = E^*(\frac{\tau_N}{N}\sum_{i=1}^{N}U_i a_{ni}^{-1}T_{ni}^*)^2$$
$$\leq \frac{\tau_N^2}{N^2}\sum_{i=1}^{N}a_{ni}^{-2}T_{ni}^{*2} + (N-1)O(\frac{1}{N})\frac{\tau_N^2}{N^2}\sum_{i=1}^{N}a_{ni}^{-2}T_{ni}^{*2}$$
$$= (1+(N-1)O(\frac{1}{N}))\frac{\tau_N^2}{N^2}\sum_{i=1}^{N}\frac{\tau_{n_i}^2}{a_{ni}^2}\tau_{n_i}^{-2}T_{ni}^{*2}\xrightarrow{p}0.$$

Finally, by Lemma 2,

$$\frac{1}{N}\sum_{i=1}^{N}a_{ni}^{-1}T_{ni}^* = \frac{1}{N}\sum_{i=1}^{N}\frac{\tau_{n_i}}{a_{ni}}\tau_{n_i}^{-1}T_{ni}^* \to 0.$$

Accordingly we also conclude that, $P^*(|\frac{1}{N}\sum_{i=1}^{N}u_i a_{ni}^{-1}T_{ni}^*|>\epsilon) = o_p(1)$. Hence, we have proved that $P^*(|\hat{\theta}_{RTS}^* - \theta_0|>\epsilon) = o_p(1)$. □

For the related asymptotic normality results as stated in Theorem 4, we need the following two Lemmas.

Lemma 10 *Suppose that the conditions of Assumptions A and B hold. If* $\frac{\tau_{n_i}}{\tau_N} = o(\sqrt{n_i})$ *then as* $n_i\to\infty$ *and* $N\to\infty$,

$$\frac{\tau_N^{-1}}{\sqrt{N}}\sum_{i=1}^{N}u_i a_{n_i}^{-2}\sum_{j=1}^{n_i}w_{ij}\phi_{ij}(\hat{\theta}_{ni})\xrightarrow{p^*}0.$$

Proof of Lemma 10 Let

$$X^*_{ni} := \tau_{n_i}^{-1} a_{n_i}^{-1} \sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni}) = a_{n_i}^{-1} \sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni}).$$

Clearly $E^*(X^*_{ni}) = 0$, and $X^*_{n_i}$ are independent for i in $1, 2, \ldots, N$. Further, by Lemma 7 we have, as $n_i \to \infty$, that

$$\begin{aligned}
E^*(X^{*2}_{ni}) =& E^*(a_{n_i}^{-1} \sum_{j=1}^{n_i} W_{ij}\phi_{ij}(\hat{\theta}_{ni}))^2 \\
=& a_{n_i}^{-2}\left[\sum_{j=1}^{n_i} \phi_{ij}^2(\hat{\theta}_{ni}) + O(\frac{1}{n_i})\sum_{j_1 \neq j_2} \phi_{ij_1}(\hat{\theta}_{ni})\phi_{ij_2}(\hat{\theta}_{ni})\right] \\
=& a_{n_i}^{-2}\left[\sum_{j=1}^{n_i} \phi_{ij}^2(\hat{\theta}_{ni}) + O(\frac{1}{n_i})\left(\sum_{j=1}^{n_i} \phi_{ij}(\hat{\theta}_{ni})\right)^2 - O(\frac{1}{n_i})\sum_{j=1}^{n_i} \phi_{ij}^2(\hat{\theta}_{ni})\right] \\
=& (1 - O(\frac{1}{n_i}))a_{n_i}^{-2} \sum_{j=1}^{n_i} \phi_{ij}^2(\hat{\theta}_{ni}) \to 1.
\end{aligned}$$

Thus, with $U_i = (u_i - 1)/\sqrt{\tau_N}$,

$$\begin{aligned}
\frac{\tau_N^{-1}}{\sqrt{N}} \sum_{i=1}^{N} u_i a_{n_i}^{-2} \sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni}) =& \frac{\tau_N^{-1}}{\sqrt{N}} \sum_{i=1}^{N} u_i a_{n_i}^{-1} \tau_{n_i} X^*_{ni} \\
=& \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i a_{n_i}^{-1} \tau_{n_i} X^*_{ni} + \frac{\tau_N^{-1}}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-1} \tau_{n_i} X^*_{ni}.
\end{aligned}$$

Since U_i and X^*_{ni} are independent, we obtain,

$$\begin{aligned}
E^*(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i a_{n_i}^{-1} \tau_{n_i} X^*_{ni})^2 =& \frac{1}{N} \sum_{i=1}^{N} E^*(U_i^2 a_{n_i}^{-2} \tau_{n_i}^2 X^{*2}_{ni}) \\
&+ \sum_{i_1 \neq i_2} E^*(U_{i_1} U_{i_2} a_{n_{i_1}}^{-1} a_{n_{i_2}}^{-1} \tau_{n_{i_1}} \tau_{n_{i_2}} X^*_{ni_1} X^*_{ni_2}) \\
=& \frac{1}{N} \sum_{i=1}^{N} a_{n_i}^{-2} \tau_{n_i}^2 E^*(X^{*2}_{ni}) \to 0.
\end{aligned}$$

Finally, since $\frac{\tau_{n_i}}{\tau_N} = o(\sqrt{n_i})$, we also have,

$$\begin{aligned}
E^*(\frac{\tau_N^{-1}}{\sqrt{N}} \sum_{i=1}^{N} a_{n_i}^{-1} \tau_{n_i} X^*_{ni})^2 =& \frac{\tau_N^{-2}}{N} \sum_{i=1}^{N} a_{n_i}^{-2} \tau_{n_i}^2 E^*(X^{*2}_{ni}) \\
=& \frac{1}{N} \sum_{i=1}^{N} \frac{\tau_{n_i}^2}{\tau_N^2} a_{n_i}^{-2} E^*(X^{*2}_{ni}) \to 0.
\end{aligned}$$

Accordingly we obtain that,

$$\frac{\tau_N^{-1}}{\sqrt{N}}\sum_{i=1}^{N} u_i a_{n_i}^{-2}\sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni})\xrightarrow{p^*}0.$$

□

Lemma 11 *Suppose that the conditions of Assumptions A and B hold. If* $\frac{\tau_{n_i}}{\tau_N} = o(\sqrt{n_i})$ *then as* $n_i \to \infty$ *and* $N \to \infty$,

$$\frac{\lambda^{-1}\tau_N^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} u_i a_{n_i}^{-1} S_{n_i}(T_{ni}^*)\xrightarrow{p^*}0.$$

Proof of Lemma 11 We first write

$$\frac{\lambda^{-1}\tau_N^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} u_i a_{n_i}^{-1} S_{n_i}(T_{ni}^*) = \frac{\lambda^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} U_i a_{n_i}^{-1} S_{n_i}(T_{ni}^*) + \frac{\lambda^{-1}\tau_N^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} a_{n_i}^{-1} S_{n_i}(T_{ni}^*).$$

By Lemma 2, Assumption B (3) and the fact $\tau_N^{-1} S_{n_i}(T_{ni}^*)\xrightarrow{p^*}0$,

$$\frac{\lambda^{-1}\tau_N^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} a_{n_i}^{-1} S_{n_i}(T_{ni}^*)\xrightarrow{p^*}0.$$

Further, it can be seen that,

$$E^*(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_i a_{ni}^{-1} S_{n_i}(T_{ni}^*))^2 \le \frac{1}{N}\left[1 + (N-1)O(\frac{1}{N})\right]\sum_{i=1}^{N} a_{ni}^{-2} E^*(S_{n_i}^2(T_{ni}^*)) \to 0.$$

Thus we have,

$$\frac{\lambda^{-1}\tau_N^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} u_i a_{n_i}^{-1} S_{n_i}(T_{ni}^*)\xrightarrow{p^*}0.$$

□

We conclude the "Appendix" with a proof of Theorem 4.

Proof of Theorem 4 By Theorem 3 and (19) we express,

$$\hat{\theta}_{ni}^* - \hat{\theta}_{ni} = -a_{ni}^{-2}\sigma^2\sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni}) - a_{n_i}^{-1}\sigma^2 S_{n_i}(T_{ni}^*).$$

Accordingly we have,

$$\hat{\theta}^*_{SRS} - \hat{\theta}_{STS} = \frac{1}{N}\sum_{i=1}^{N}(u_i - 1)\hat{\theta}_{ni} - \frac{\sigma^2}{N}\sum_{i=1}^{N} u_i a_{n_i}^{-2}\sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni}) - \frac{\sigma^2}{N}\sum_{i=1}^{N} u_i a_{n_i}^{-1} S_{n_i}(T^*_{ni}),$$

where $|T^*_{ni}| < K\tau_{n_i}$ in probability. Further,

$$\begin{aligned}\lambda^{-1}\tau_N^{-1}\sqrt{N}(\hat{\theta}^*_{RTS} - \hat{\theta}_{STS}) = &\frac{\lambda^{-1}\tau_N^{-1}}{\sqrt{N}}\sum_{i=1}^{N}(u_i - 1)\hat{\theta}_{ni} \\ &- \frac{\lambda^{-1}\tau_N^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} u_i a_{n_i}^{-2}\sum_{j=1}^{n_i} w_{ij}\phi_{ij}(\hat{\theta}_{ni}) \\ &- \frac{\lambda^{-1}\tau_N^{-1}\sigma^2}{\sqrt{N}}\sum_{i=1}^{N} u_i a_{n_i}^{-1} S_{n_i}(T^*_{ni}) \\ \equiv & I_1 + I_2 + I_3.\end{aligned}$$

By Lemma 10, $I_2 \xrightarrow{p^*} 0$, and by Lemma 11, $I_3 \xrightarrow{p^*} 0$, and therefore it remains only to consider I_1. Now, observe that,

$$I_1 := \frac{\lambda^{-1}\tau_N^{-1}}{\sqrt{N}}\sum_{i=1}^{N}(u_i - 1)\hat{\theta}_{ni} = \frac{\lambda^{-1}}{\sqrt{N}}\sum_{i=1}^{N} U_i(b_i + \theta_0) + \frac{\lambda^{-1}}{\sqrt{N}}\sum_{i=1}^{N} U_i a_{n_i}^{-1} T_{ni}.$$

By Lemma 2,

$$E^*\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_i a_{ni}^{-1} T_{ni}\right)^2 \le \frac{1}{N}\sum_{i=1}^{N} a_{ni}^{-2} T_{ni}^2 + (N-1)O\left(\frac{1}{N}\right)\frac{1}{N}\sum_{i=1}^{N} a_{ni}^{-2} T_{ni}^2 \xrightarrow{p} 0.$$

Further by Lemma 5,

$$\bar{U}_N := \frac{1}{N}\sum_{i=1}^{N} U_i \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{u_i - 1}{\tau_N} \xrightarrow{p^*} 0,$$

and clearly, $\sqrt{N}(\bar{b} + \theta_0) \Rightarrow \mathcal{N}(\theta_0, \lambda^2)$. Accordingly we have, $\frac{\lambda^{-1}}{N}\sum_{i=1}^{N}(b_i - \bar{b})^2 \to 1$ $a.s.$ as well as $\sqrt{N}\bar{U}(\bar{b} + \theta_0) \xrightarrow{p^*} 0$. Further, by Lemma 4.6 of Praestgaard and Wellner (1993), we have that

$$\frac{\lambda^{-1}}{\sqrt{N}}\sum_{i=1}^{N} U_i(b_i + \theta_0) \Rightarrow \mathcal{N}(0, 1).$$

Thus we have

$$\frac{\lambda^{-1}\tau_N^{-1}}{\sqrt{N}}\sum_{i=1}^{N}(u_i - 1)\hat{\theta}_{ni} \Rightarrow \mathcal{N}(0, 1).$$

Finally we conclude that as $n_i \to \infty$ and $N \to \infty$,

$$\lambda^{-1}\tau_N^{-1}\sqrt{N}(\hat{\theta}^*_{RTS} - \hat{\theta}_{STS}) \Rightarrow \mathcal{N}(0,1).$$

□

Acknowledgements We would like to thank the Editor and the anonymous referees for their careful review, useful comments and helpful suggestions that have led to much improvements of this paper.

Funding This research was not supported by any funding.

Availability of data and material Derived data supporting the findings of this study are available as `Theoph` under the R package.

Declarations

Conflict of interest No conflicts of interest.

Code availability The code supporting this study are available from the corresponding author upon reasonable request.

References

Adeniyi I, Yahya WB, Ezenweke C (2018) A note on pharmacokinetics modelling of theophylline concentration data on patients with respiratory diseases. Turk Klin J Biostat. 10(1):27–45

Bar-Lev SK, Boukai B (2015) Recycled estimation of population pharmacokinetics models. Adv Appl Stat 47:247–263

Bates DM, Watts DG (2007) Nonlinear regression analysis and its applications. Wiley, New York

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9:1196–1217

Boeckmann AJ, Sheiner LB, Beal SL (1994) NONMEM users guide: part V, NONMEM project group. University of California, San Francisco

Boukai B, Zhang Y (2018) Recycled least squares estimation in nonlinear regression. *arXiv Preprint (2018)*ArXiv:1812.06167 [stat.ME]

Chatterjee S, Bose A (2005) Generalized bootstrap for estimating equations. Ann Stat 33:414–436

Davidian M, Gallant AR (1993a) The non-linear mixed effects model with a smooth random effects density. Biometrika 80:475–488

Davidian M, Giltinan DM (1993b) Some simple methods for estimating intra-individual variability in non-linear mixed effects models. Biometrics 49:59–73

Davidian M, Giltinan DM (1995) Monographs on statistics and applied probability. Nonlinear models for repeated measurements data. Chapman and Hall, London

Davidian M, Giltinan DM (2003) Nonlinear models for repeated measurement data: an overview and update. J Agric Biol Environ Stat 8:387–419

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Efron B (1979) Bootstrap methods: another Look at the Jackknife. Ann Stat 7:1–26

Efron B, Tibshirani R (1994) An introduction to the bootstrap. Chapman and Hall, New York

Eicker F (1963) Asymptotic normality and consistency of the least squares estimators for families of linear regressions. Ann Math Stat 34:447–456

Flachaire E (2005) Bootstrapping heteroskedastic regression models: wild bootstrap vs pairs bootstrap. CSDA 49:361–476

Fan J, Mei C (1991) The convergence rate of randomly weighted approximation for errors of estimated parameters of AR(I) models. Xian Jiaotong DaXue Xuebao 25:1–6

Freedman DA (1981) Bootstrapping regression models. Ann Stat 9:1218–1228

Hartigan JA (1969) Using subs ample values as typical value. J Am Stat Assoc 64:1303–1317

Ito K, Nisio M (1968) On the convergence of sums of independent Banach space valued random variables. Osaka J Math 5:33–48

Jennrich IR (1969) Asymptotic properties of non-linear least squares estimatiors. Ann Stat 40:633–643
Kurada RR, Chen F (2018) Fitting compartment models using PROC NLMIXED. *Paper SAS1883-2018*, https://bit.ly/2Cds4cT
Lo AY (1987) A large sample study of the Bayesian bootstrap. Ann Stat 15:360–375
Lo AY (1991) Bayesian bootstrap clones and a biometry function. Sankhya A 53:320–333
Lindstrom MJ, Bates DM (1990) Non-linear mixed effects models for repeated measures data. Biometrics 46:673–687
Mallet A (1986) A maximum likelihood estimation method for random coefficient regression models. Biometrika 73:645–656
Mammen E (1989) Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann Stat 17:382–400
Mason DM, Newton MA (1992) A rank statistics approach to the consistency of a general bootstrap. Ann Stat 20:1611–1624
Newton MA, Raftery AE (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). J R Stat Soc Ser B 56:3–48
Praestgaard J, Wellner JA (1993) Exchangeably weighted bootstraps of the general empirical process. Ann Probab 21:2053–2086
Quenouille M (1949) Approximate tests of correlation in time-series. Math Proc Cambridge Philos Soc 45(03):483
R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 3-900051-07-0 (http://www.R-project.org/)
Rao CR, Zhao L (1992) Approximation to the distribution of Mestimates in linear models by randomly weighted bootstrap. Sankhya A 54:323–331
Rubin DB (1981) The Bayesian bootstrap. Ann Stat 9:130–134
Shao J, Tu DS (1995) The Jackknife and Bootstrap. Springer-Verlag, New York. Singh K (1981) On the asymptotic accuracy of Efron's bootstrap. Ann Stat 9:1187–1195
Sheiner LB, Rosenberg B, Melmon KL (1972) Modelling of individual pharmacokinetics for computer-aided drug dosage. Comput Biomed Res 5:441–459
Sheiner LB, Beal SL (1981) Evaluation of methods for estimating population pharmacokinetic parameters. II. Bioexponential model: routine clinical pharmacokinetic data. J Pharmacokinet Biopharm 9:635–651
Sheiner LB, Beal SL (1982) Bayesian individualization of pharmacokinetics: simple implementation and comparison with non-Bayesian methods. J Pharm Sci 71(12):1344–1348
Sheiner LB, Beal SL (1983) Evaluation of methods for estimating population pharmacokinetic parameters. III. Monoexponential model: routine clinical pharmacokinetic data. J Pharmacokinet Biopharm 11:303–319
Singh K (1981) On the asymptotic accuracy of Efron's bootstrap. Ann Stat 9:1187–1195
Steimer JL, Mallet A, Golmard JL, Boisvieux JF (1984) Alternative approaches to estimation of population pharmacokinetic parameters: comparison with the non-linear mixed effect model. Drug Metab Rev 15:265–292
Vonesh EF, Carter RL (1992) Mixed effects non-linear regression for unbalanced repeated measures. Biometrics 48:1–17
Weng CS (1989) On a second order property of the Bayesian bootstrap. Ann Stat 17:705–710
Wu CF (1981) Asymptotic theory of nonlinear least squares estimation. Ann Stat 9:501–513
Wu CFJ (1986) Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). Ann Stat 14:1261–1350
Yu K (1988) The random weighting approximation of sample variance estimates with applications to sampling survey. Chin J Appl Prob Stat 3:340–347
Zhang Y, Boukai B (2019) Recycled estimates for nonlinear regression models. Stat 8:e230
Zhang Y, Boukai B (2019) Recycled two-stage estimation in nonlinear mixed effects regression models, arXiv:1902.00917
Zheng Z (1987) Random weighting methods. Acta Math Appl Sinica 10:247–253
Zheng Z, Tu D (1988) Random weighting method in regression models. Sci China Ser A-Math Phys Astron Technol Sci 31(12):1442–1459

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.