Bias-Corrected Least Squares (BCLS) Method for Estimating Parameters in Nonlinear Ordinary Differential Equations Models with Mixed-Effects

Yalan Xing

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2018

Committee:
Sarah Holte, Chair
Ali Shojaie

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Bias-Corrected Least Squares (BCLS) Method for Estimating Parameters in Nonlinear Ordinary Differential Equations Models with Mixed-Effects

Yalan Xing

Chair of the Supervisory Committee:
Professor Sarah Holte
Public Health Sciences

Ordinary Differential Equation (ODE) is a popular framework for modeling temporal dynamics in a variety of scientific settings. When repeated measurements are available for the dynamic process, it is of great interest to estimate random effects in the ODE model for the process. Current algorithms for parameter estimation in the nonlinear mixed-effects model are often challenging due to intensive computation and slow convergence; while alternative methods mostly rely on smoothing spline functions and therefore a careful choice of smoothing parameters. We propose an extension of the original Bias-Corrected Least Squares (BCLS) method, referred to as BCLS-LME, which reduces the nonlinear mixed-effects model to a linear mixed-effects model that does not require any smoothing. We demonstrate that, using both simulated data and an actual birth cohort data and with proper sampling schemes, the BCLS-LME method achieves comparable accuracy with the NLME method for estimating both ODE parameters and random effects; however, it dramatically improves the computational efficiency by reducing computation time and alleviating convergence difficulties. Given proper sampling schemes and further fine-tuning, the BCLS-LME method is likely to provide a powerful tool for parameter estimation in complex ODE mixed-effects models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first extend my deepest gratitude to my thesis advisor, Dr. Sarah Holte at the Fred Hutchinson Cancer Research Center. She has been my strong support through research, and inspirational role model in life. I learned from her academically in research, and professionally at work in the CFAR Biometrics Core. Her motto, "It is nice to be important, but more important to be nice", has shed light on me and will go a long way in my future career and life.

I would also like to thank my thesis committee member, Dr. Ali Shojaie, for his suggestions and constructive questions during our regular meetings. All those discussions and inputs from Ali have fundamentally shaped my thesis to the final form.

I must acknowledge my peer students, Lei Wang, Yichen Zhang, Xiaoyue Wang, Kendrick Li, Adam Elder and many more, with whom I have gone through stressful days and nights studying helping each other. The unforgettable experience will lead to life-long friendships.

Last but not the least, I want to give my thanks to our program director, Gitana Garofalo, for her devotion to supporting all the Biostatistics students and organizing department affairs. I deeply appreciate all the talks we had, all the positive energy from her; and I will miss her notes and sweets during the final weeks.

DEDICATIONS

This thesis is dedicated to my precious sons, Luca and Oliver. They have motivated me, made me stronger, and driven me to challenge myself constantly and become the better me. I also dedicate my thesis to my parents and parents-in-law for their unconditional love and support for me pursuing my dreams through years. Last, this thesis is dedicated to my husband, Pin Lu, who has been a great friend, a reliable teammate, and the love of my life.

Chapter 1. Introduction

Mathematical modeling has been extensively utilized in a wide variety of areas such as engineering, medicine, economics and social sciences. Differential equations are one of the essential and widely applied techniques in mathematical modeling. In this chapter, we will introduce Ordinary Differential Equations (ODEs) models, and discuss scientific fields where ODEs are most applicable. We will also review current methods for estimating parameters in simple ODE models and ODE models with mixed-effects, and the greatest challenges encountered, which motivated the bias-corrected least squares method (BCLS). As the original BCLS method was developed for longitudinal data from a single individual, the central goal of this thesis lies in developing the BCLS-LME method, which extends the BCLS method to the mixed-effects models depicting population longitudinal data, with the ultimate aim of reducing the computational burden in the nonlinear mixed-effects (NLME) model.

1.1 Ordinary Differential Equations (ODEs)

In mathematics, an ordinary differential equation (ODE) is an equation containing one or more functions of one independent variable and its derivatives. It can be written as:

$$\frac{dX(t)}{dt} = F\{X(t), \beta\} \tag{1.1}$$

where $X(t) = \{X_1(t), \dots, X_k(t)\}^T$ is a vector of observations depending on time, $\beta = (\beta_1, \dots, \beta_m)^T$ is a vector of unknown parameters, and $F(.)$ can be a linear or nonlinear function.

ODEs have been used extensively in a wide variety of scientific areas to describe time-varying phenomena. The most historical application was utilizing ODEs to understand the motion of objects in physics. Many fundamental laws of physics and chemistry can be formulated as differential equations. In ecology, ODEs have been utilized to understand the dynamics of animal and plant populations (Freedman 1980); meteorologists have used them to understand and predict patterns in the weather and atmosphere (Lorenz 1963); economists have applied them to understanding and predicting patterns in financial markets (Zhang 2005). One of the rising trends in biological research and health care research is utilizing ODEs to make sense of observed nonlinear behaviors. Particularly, the studies of pharmacokinetics (PK) and pharmacodynamics (PD), areas where evaluation of drug metabolism is essentially connected with design and development of biomedical compounds, have extensively integrated ODEs with data analysis (Csajka and Verotta 2006, Dartois, Brendel et al. 2007, Danhof, de Lange et al. 2008). In some clinical research, ODE models have been used as a tool in exploring the etiology of HIV and Hepatitis B and C infections as well as the effects of therapy on those diseases. Multiple studies have implemented ODEs to analyze the temporal dynamics of HIV viral load measurements in AIDS patients (Ho, Neumann et al. 1995, Wei, Ghosh et al. 1995, Perelson, Neumann et al. 1996, Perelson, Essunger et al. 1997), and the results had a tremendous scientific impact, revealing that the HIV virus replicates rapidly and continuously, in spite of the prolonged interval between infection and development of AIDS.

1.2 Current Methods for Parameter Estimation in ODE Models

ODEs are more and more commonly used in combination with observed data for parameter estimation and statistical inference. Parameter estimates are often obtained using the standard nonlinear least squares (NLS) estimation techniques. One drawback of NLS estimation using conventional gradient-based optimization methods such as Gauss-Newton or Levenberg-Marquart

is the instability of the procedure since the estimates are obtained by an iterative numerical procedure (Press 1986). NLS procedures may fail to converge to global minima depending on the choice of starting values for iteration, particularly if the least squares solution has multiple local minima or saddle points. Certain likelihood surfaces or cost functions may be exceptionally complex with ridges that appear near bifurcation values of the parameters in the ODE model. Furthermore, it is well known that nonlinear regression estimates may be biased when small samples are used for estimation (Cook, Tsai et al. 1986).

In contrast, the non-iterative ("direct") methods, which can be based on derivatives or integrals, are generally combined with a linear least squares estimation step. Therefore, when the parameters in an ODE model are linear in the vector field, it is plausible to estimate parameters either iteratively (e.g. the NLS method) or non-iteratively (the direct methods). The direct methods relying on derivatives were originally described by (Bard 1974, Swartz 1975, Hosten 1979, Varah 1982). Those methods involve constructing a cubic spline using the observed data for each dependent variable, and estimating unknown parameters using linear least squares. The direct methods relying on integrals were originally described by (Himmelblau, Jones et al. 1967) and further developed in (Foss 1971, Jacquez 1972, Bard 1974, Hosten 1979). In integral-based methods, the ODEs are transformed into integral equations; the integrals are approximated via methods of quadrature yielding algebraic equations, which can be solved for the approximate parameters using a linear least squares approach. Both approaches lead to linear least squares problems which can be solved without iteration, and provide similar estimates given negligible measurement errors and evenly spaced values of dependent variable. In the scenario where the assumptions are not met, the approaches which rely on integrals are preferred to those rely on derivatives (Wikstrom 1997). Methods that rely on derivatives require differencing of observed data, an approach known to inflate errors; while methods that rely on integrals involve sums of observed data, a smoothing approach

that can reduce errors. Methods that rely on integrals also have the advantage of capability to estimate the initial states of the system, if these are unknown.

The parameter estimates from the original direct methods are not optimal, in the sense that they are not unbiased, efficient, or consistent. A recent study (Liang and Wu 2008) described a direct method based on differentiation, local smoothing of the data and linear least squares regression of correlated quantities, which is referred to as the pseudo-least squares (PsLS) method. The PsLS estimates of ODE model parameters were proven to be consistent with an asymptotically normal distribution, under reasonable general conditions. However, this method requires a critical choice of bandwidth for the kernel smoothing, which may demand additional research to obtain the most accurate and efficient results. Other investigators have used functional estimation or principle differential analysis (PDA) approaches to replace the state variables with smoothed estimates, and proceed with estimation of parameters using estimated values of the solutions to ODEs (Ramsay and Silverman 2005, Ramsay, Hooker et al. 2007). These methods, like the PsLS method, use derivatives rather than integrals to form the least squares regression models. Dattner and Klaassen developed an integral based two-step estimation approach which is based on first rewriting the ODE model as an integral equation, next smoothing the data, and finally using a linear regression step with the smoothed observations as described above to obtain estimated parameter values (Dattner and Klaassen 2015). Importantly, all studies described above have used a pre- or simultaneous smoothing of the observed data, which often results in additional rather than less bias if the smoothing parameter(s) are not correctly chosen (personal communications with Dr. Sarah Holte). In fact, the choice of smoothing parameter(s) may result in unbiased estimates for some of the parameters in the system but introduce bias in estimation of other parameters in the system. Thus it's likely that pre- or simultaneously smoothed data will result in bias in some of the estimated parameters.

In contrast, we are utilizing a direct method relying on integrals, the bias-corrected least squares method (BCLS) (Holte 2016). By modifying a direct least squares method similar to the direct methods developed by Himmelblau, Jones and Bischoff (Himmelblau, Jones et al. 1967), the BCLS method incorporates a bias correction factor to obtain consistent estimates, and does not require any smoothing or functional estimation. It has been proven in (Holte, 2016) to produce asymptotically unbiased estimates for parameters in nonlinear ODE problems, compared with NLS, PsLS, and Dattner & Klaassen's methods.

1.3 ODE Models with Mixed-Effects and Current Methods for Parameter Estimation

Longitudinal studies play a key role in epidemiology, clinical research and general therapeutic evaluations. The key feature of longitudinal studies is that measurements of the same individual are taken repeatedly through time, thereby allowing the direct study of change over time. The primary goal of a longitudinal study is to characterize the change in response over time and the factors that influence the change (Fitzmaurice, Laird et al. 2011). As longitudinal studies generally yield repeated measurements on each subject, they involve correlated data within subjects, thus require statistical methods to properly account for the intra-subject correlation of response measurements (Van Belle and Fisher 2004). A mixed-effects model is a powerful tool for longitudinal data, which incorporates both fixed effects and random effects for each individual or cluster to account for correlations within individuals in addition to variations between individuals. Since each individual or cluster shares the same random effects, the measurements within the individual or cluster are correlated; and the random effects require individual-specific inference.

All the methods described in section 1.2 were developed for longitudinal data from a single individual. To obtain accurate estimates for growth rates and individual random effects, given

population nonlinear growth data over time, we need mixed-effects estimation and inference methods to account for the inter-class correlation in an aggregate population with time-varying mean described by an ODE model. This approach flexibly represents the correlation structure, thus an efficient method for estimation. When nonlinear ODEs have analytic solutions, the "Gold Standard" method for parameters estimation is the nonlinear mixed-effects model (NLME), commonly tackled by available packages such as the nlme( ) function in R. As most ODEs have no analytic solutions, methods have been developed to estimate mixed-effects ODE models by solving them numerically. Software package NONMEM has been developed and widely used to estimate nonlinear mixed-effects models in the pharmacometrics community (Wang 2007); EM-type algorithms, such as stochastic approximation EM (SAEM), have been developed to tackle computational challenges from the maximum likelihood method (Kuhn 2005). Those methods and software programs all require precise initial values for the ODE model parameters in order to obtain numerical solutions, which are often difficult to obtain. Wang and Cao et al. proposed a semi-parametric method to estimate mixed-effects ODE model parameters without a requirement of analytic solutions (Wang 2014). Rather than using the ODE numerical solution directly, which requires providing initial conditions, their method estimates a spline function to approximate the dynamic process using smoothing splines. However, current methods face common challenges from the nature of population longitudinal data: it is impractical to estimate spline functions using smoothing for each individual, who has his/her own growth rate and random effects.

Based on the statistical properties of the BCLS method in estimating parameters for ODE models, we propose to extend the BCLS method for a single individual to a linear mixed-effects model for a population of individuals, referred to as BCLS-LME. We are particularly interested in evaluating the estimation accuracy and computational efficiency of the BCLS-LME method, compared to the NLME method. The novel finding from this thesis is demonstrating the

comparable accuracy, precision and great advantage in computational efficiency of the BCLS-LME method, which will motivate further efforts to prove the method's consistency in mixed-effects ODE models.

Chapter 2. Methods

2.1 Background for the BCLS Method

2.1.1 BCLS model assumptions

As described in (Holte 2016), the bias-corrected least squares (BCLS) method applies to data, $y(t)$, on observations with time-varying expectation given by the solutions of system of $q = 1, \dots, s$ ODEs, observed at $t = (t_0, \dots, t_n)$, n observation times. The statistical and mathematical model assumptions for the data $y(t)$ with distribution $Y(t)$ are summarized as follows:

(A.1): Specification of the mean structure via a mathematical model

- There exists a vector of functions $\boldsymbol{X}(t) = \{X_1(t), \dots, X_s(t)\}$ which satisfies the system of differential equations:

$$\frac{dX_q}{dt} = \sum_{k=1}^{m_q} \beta_{q,k} h_{q,k}\left(X_1, \dots, X_s\right), \quad X_q(t_0) = X_{q,0}, q = 1, \dots, s \tag{2.1}$$

such that the expectation, $E\{\boldsymbol{Y}(t)\} = \boldsymbol{X}(t)$.

(A.2): Specification of variance/covariance structure

- $Y_q(\boldsymbol{t}) = X_q(\boldsymbol{t}) + \boldsymbol{\epsilon}_q, \quad \boldsymbol{\epsilon}_q \sim (\mathbf{0}, \Sigma_q), \ q = 1, \dots, s.$

- Σ_q is a diagonal matrix with constant diagonal entries $\sigma_q, \ q = 1, \dots, s.$

- $\boldsymbol{\epsilon}_r \ and \ \boldsymbol{\epsilon}_q$ are independent for all $1 \leq r < q \leq s$.

- $var(Y_i)$ and $var\{h_{q,k}(Y_i)\}$ are bounded by a common bound B for all i, q, and k.

(A.3): Data sampling requirements

- To obtain asymptotic properties, the maximum interval length defined by sampling times is $O(n^{-1})$.

- At least one of the compartments is not in equilibrium throughout the entire time course of data collection.

Condition (A.1) provides the relationship between the time-varying expectation of the data and a system of ODEs. Condition (A.2) requires that conditional on the expected value, observations from different compartments of the system of ODE's are independent. This requirement is not overly restrictive since the use of ODEs is intended to capture the relationships (correlations) between the observed compartments. Condition (A.2) also specifies that the variance of observations from different compartments can differ. The second part of condition (A.3) is included to prevent parameter non-identifiability but does not ensure identifiability of the parameters of interest.

The key component of BCLS method is the use of bias correction functions as weights in the least squares estimation, to compensate for bias introduced during transformation of random variables. These functions can be defined in the following ways to satisfy $E[h^*\{\boldsymbol{Y}(t)\}] = h\{\boldsymbol{X}(t)\}$:

1) For a function $h = h_{q,k}$, identify the difference from expectation, $E[h\{\boldsymbol{Y}(t)\}] - h\{E[\boldsymbol{Y}(t)]\}$ and subtract it from h to define h^* as $h^*\{\boldsymbol{Y}(t)\} = h\{\boldsymbol{Y}(t)\} - (E[h\{\boldsymbol{Y}(t)\}] - h\{E[\boldsymbol{Y}(t)]\})$;

2) Alternatively, h^* can be defined by multiplying h with the ratio of $\frac{h\{E[\boldsymbol{Y}(t)]\}}{E[h\{\boldsymbol{Y}(t)\}]}$, as $h^*\{\boldsymbol{Y}(t)\} = h\{\boldsymbol{Y}(t)\}\frac{h\{E[\boldsymbol{Y}(t)]\}}{E[h\{\boldsymbol{Y}(t)\}]}$.

2.1.2 Estimation algorithm

The BCLS method is designed to simplify a nonlinear regression problem by reducing it to a linear regression problem (Holte 2016). This is achieved by 1) transforming the system of differential equations into a system of integral equations; 2) treating the transformed system as a model for linear regressions involving covariates that are approximates of the integrals.

Using children's growth data as an example, we assume that children's weight/height, $\boldsymbol{y(t)}$, follows a log-normal distribution, with

$$\log\{Y(\boldsymbol{t})\} = \log\{X(\boldsymbol{t})\} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim i.i.d\ N(0, \sigma^2) \tag{2.2}$$

where $X(t)$ satisfies the standard logistic growth curve described by the nonlinear differential equation:

$$\frac{dX}{dt} = X(a - bX), \qquad \boldsymbol{X}(0) = y_0 \tag{2.3}$$

Parameters a and b define the growth rate per capita, a/b is the carrying capacity of the growth, and y_0 is the initial weight/height. To estimate a and b, the following algorithm is executed:

Step 0: Log-transform differential equation (2.3) to obtain normally distributed observations:

$$\frac{d\{\log(X)\}}{dt} = a - bX,\ \ \log\{X(0)\} = \log(y_0). \tag{2.4}$$

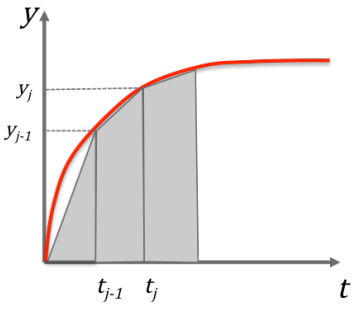
Step 1: Transform differential equation (2.4) into an integral equation:

$$\log\{X(t)\} - \log(y_0) = \ a\int_0^t ds - b\int_0^t h\{X(s)\}ds \tag{2.5}$$

where $h\{X\} = X$.

Step 2: Determine the bias correction function. As $y(t)$ follows a log-normal distribution, we can calculate the expectation of h : $E[h\{Y(t_i)\}] = E[Y(t_i)] = X(t_i)e^{\frac{\sigma^2}{2}}$. To make sure $E[h^*\{Y(t_i)\}] = X(t_i)$, we can obtain the correct form of bias correction function h^*: $h^*(X) = Xe^{\frac{-\sigma^2}{2}}$, and use it for weighting.

Step 3: Approximate the integrals in equation (2.5). The first covariate $\int_0^t ds$ is simply t, thus no weighting is necessary; the second covariate integral $\int_0^t h\{X(s)\}ds$ can be approximated using methods such as the trapezoid rule, with consideration of the bias correction function. Thus the second covariate can be approximated by defining covariates $z(t) = \{z(t_0), z(t_1), \dots, z(t_n)\}$ in the form of: $z(t_0) = 0$ and for $i = 1, \dots, n$, where $e^{\frac{-\sigma^2}{2}}$ is the necessary weight to correct for bias



$$z(t_i) = e^{\frac{-\sigma^2}{2}}\left\{\sum_{j=1}^{i-1}\frac{(y_j + y_{j-1}) * (t_j - t_{j-1})}{2}\right\} \quad (2.6)$$

Step 4: Fit linear regression model. With the known initial value, y_0, we use $y'(\boldsymbol{t}) = log\{y(\boldsymbol{t})\} - log\{y_0\}$ as the response for the regression model in the BCLS method, and obtain the estimates for $\boldsymbol{a}$ and $\boldsymbol{b}$ using standard linear regression with the following model without intercept:

$$y'(\boldsymbol{t}) \sim a\,\boldsymbol{t} + b\,z(\boldsymbol{t}) + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim i.i.d.N(0, \sigma^2) \quad (2.7)$$

If the initial value, y_0, is unknown, it can be treated as an additional parameter and estimated by including an intercept in the regression model (2.7), using $log\{y(\boldsymbol{t})\}$ as the response.

2.1.3 Simulation of growth curve data for an individual

The growth curve data was simulated based on the statistical model

$$\log\{Y(\boldsymbol{t})\} = \log\{X(\boldsymbol{t})\} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim i.i.d\ N(0, \sigma^2)$$

where $X(\boldsymbol{t})$ is the solution of equation (2.3). The parameters for simulation were set as follows:

- $y_0 = 2$;
- Means of parameters: $a = 0.8, b = 0.0015$;
- Measurement error: $\epsilon \sim i.i.d.N(0, \sigma^2)$, $\sigma = (0.1, 0.2, 0.3, 0.4)$.

Four sets of observations times were evaluated, each consisted of 21 or 201 total observations, evenly spaced between $t = 0$ and $t = 20$ months, or $t = 0$ and $t = 60$ months. We simulated an individual dataset for each of the four sets of observation times as the sampling scheme and each of the values of $\sigma = (0.1, 0.2, 0.3, 0.4)$. For each combination of sampling schemes and values of σ, the data were simulated 1000 times. We then followed steps described in section 2.1.2 to approximate the integrals and fit the linear regression model to obtain the BCLS estimates, in comparison with LS (without weights to adjust for bias) and NLS estimates. Relative biases in the parameter estimates and computational efficiency of each method were investigated and compared.

2.2 Mixed-Effects Models

The mixed-effects models are commonly used in longitudinal studies, where the repeated measurements are clustered by subject. This section for the first time describes the BCLS-LME method extended from the BCLS method, applying to mixed-effects ODE models.

2.2.1 Simulation of growth curve data for an aggregate population of individuals

In our mixed-effects model, we use the same growth curve ODE as described above, but take into consideration the random effects from both a and b, to account for individual-specific growth rates. We first simulated parameters a and b for each individual, following normal distributions

$$a \sim i.i.d.\ N(\mu_a, \sigma_a^2),\ \ b \sim i.i.d.\ N(\mu_b, \sigma_b^2)$$

with random effects. A range of values for random effects, σ_a and σ_b, were evaluated. $\sigma_a = (0.1, 0.2, 0.4), \sigma_b = (0.0001, 0.0005)$. The growth curve data set for each individual out of a population of 100 subjects was then simulated based on the same statistical model as equation (2.2), using similar strategies as described in section 2.1.3. The parameters for simulation were set as follows:

- $y_0 = 2$;
- Means of parameters: $\mu_a = 0.8,\ \mu_b = 0.0015$;
- Measurement error: $\epsilon \sim i.i.d.N(0, \sigma^2)\ , \sigma\ =\ (0.1, 0.2, 0.3, 0.4)$.

Instead of evenly spaced samplings, we designed a sampling scheme mimicking the sampling scenario of longitudinal studies in reality, where the repeated measurements from follow-up visits are often sampled at random time points, and tend to become more sparse towards the end of study. The observation times of each individual consisted of 21 or 201 random observations between $t = 0$ and $t = 20$ months, or $t = 0$ and $t = 60$ months, following an exponential distribution $t \sim Exp(\lambda), \lambda = 0.05$. For each combination of sampling schemes and values of σ, the data were simulated for 100 subjects. We then followed steps described in section 2.1.2 to approximate the integrals, fit the linear mixed-effect model and nonlinear mixed-effect model respectively to obtain the BCLS-LME and NLME estimates. Relative biases in the estimates, convergence rate, and computational efficiency of each method were investigated in details in Chapter 3.

2.2.2 BCLS-LME model fit to simulated data sets

Instead of the original BCLS method using linear least squares, we estimated parameters using the simulated data sets with extended method, BCLS-LME, which uses a linear mixed-effects model. Simulated data sets were fit with BCLS-LME and NLME, the nonlinear mixed-effects model, respectively, using lme( ) and nlme( ) functions from the nlme package of R (Pinheiro J 2017). In the BCLS-LME model, we defined the linear formula as $y'(t) \sim t + z(t) - 1$ to describe the fixed-effects parts of the model, specifying no intercept to be estimated; we also specified the random-effects formula as $\sim t + z(t) - 1 \mid ID$, grouped by patient ID, allowing different random effects for each grouping level. In the NLME model, we defined the same fixed-effects and random-effects formula and grouping structure as the BCLS-LME model. We specified $a = 0.8$ and $b = 0.0015$ as the starting values for the NLME algorithm. The accuracy of each model was evaluated by the relative biases of estimates for a and b. The computational efficiency was evaluated by both the successful convergence rate out of 100 iterations and the average computation time of 1000 iterations.

2.3 The Observational Malaria Cohort (OMC) Data Set

To demonstrate the capacity of BCLS method in an actual data set, we utilized a population data set from a longitudinal birth cohort in Muheza district of Tanzania, an area of intensive malaria transmission rate (Goncalves, Huang et al. 2014). We named the data set observational malaria cohort (OMC) for simplicity. A total of 882 newborns were enrolled between September 2002 and November 2005, followed for an average of 2 years and for as long as 4 years. Of the 882 children in the study, 457 (51.8%) were males, 201 (11.6%) had severe malaria, 663 (75.1%) had follow-up measurements for over one year. The enrolled children were examined once every two weeks during

infancy, once every month after infancy, and during any illness. Baseline characteristics, including weight, height and head circumference, living/nutritional conditions and physical vitals including body temperature, parasite burden, and malaria symptom diagnosis results were collected at each visit.

We utilized the repeated measurements of children's weights as the main outcome of interest, as body weight is a direct indicator of growth rate and easily influenced by major health events such as malaria episodes. The weight data generally follows a logistic growth curve but has not fully reached the steady state by the end of the study. From the original OMC data set, we first excluded two subjects who had only baseline weight measurement. We defined the initial value W_0 as the mean of initial weights from all children, the measurement error σ as the standard deviation of initial weights from all children. We then defined the response $W'(t)$ by subtracting the log of the initial weight from the log of weight $W'(t) = log\,(W(t)) - log\,(W_0)$, and calculated $Z(t)$ following equation (2.6). After pre-processing, the BCLS-LME and NLME models were fit to the data set respectively, as described in section 2.2, to estimate the fixed effects and random effects of parameters a and b. The starting values for NLME model were set as $a = 0.8, b = 0.008$. We used the NLME estimates as the "unbiased" standards and evaluated the BCLS-LME method by comparing the Monte Carlo means and standard deviations of the parameters estimates.

The effects from other covariates, Qs, on growth rate can also be evaluated by estimating the coefficient γ for Q in the alternative form of ODE

$$\frac{dW}{dt} = W\big((a^* + \gamma * Q) - bW\big). \tag{2.8}$$

We investigated the effects on children's weight growth from multiple variables, including gender (*gender*), positive parasite counts in blood smear samples (*blood*), ever diagnosis of malaria (*malever*),

and the number of malaria episodes (*nmal*). The estimated coefficients and clinical inference will be further discussed in Chapter 3.

Chapter 3. Results

To illustrate the application of the BCLS and BCLS-LME methods using a model defined by the nonlinear differential equation (2.3), we employed data on the growth of children, from simulations as well as the OMC birth cohort data set. In this chapter, we first implemented the original BCLS method to fit simulated longitudinal data from an individual, and compared it to the NLS method in accuracy and computational efficiency. We then implemented the extended method, BCLS-LME, to mixed-effects models and compared the accuracy and computational efficiency against the NLME method, using simulated growth data for an aggregate population of individuals. Furthermore, we applied the BCLS-LME method to estimate the growth rate of the population from the OMC data set, and provided insights on the effects of parasite burden and malaria episodes on children's growth rate.

3.1 The BCLS Method Compared to NLS Method in Parameter Estimation

3.1.1 Data simulation

We simulated data for a single individual, following the statistical model in equation (2.2) in section 2.1.2

$$\log\{Y(t)\} = \log\{X(t)\} + \epsilon, \;\; \epsilon \sim i.i.d\; N(0, \sigma^2)$$

with the parameters set as: $a = 0.8, b = 0.0015, y_0 = 2.$ As described in section 2.1.3, a range of values for measurement error, $\sigma = (0.1, 0.2, 0.3, 0.4)$, was evaluated to demonstrate the increase in bias in the estimate of the parameter b using ordinary least squares (LS) without the bias-correction adjustment. Data sets were simulated with multiple sampling schemes as mentioned in section 2.1.3.

The transformed data from one simulation and the fitted trajectory by the NLS method are demonstrated in Figure 1. All data points are from a simulation with 201 evenly spaced observations between t = 0 and t = 20 months, measurement error $\sigma = 0.4$. For each combination of observation times and values of σ, data was simulated 1000 times and the LS, BCLS and NLS models were fit to estimate model parameters a and b. For the NLS model, the initial values were specified as $a = 0.8, b = 0.0015$. Among all methods, the relative biases and computational efficiency are compared.

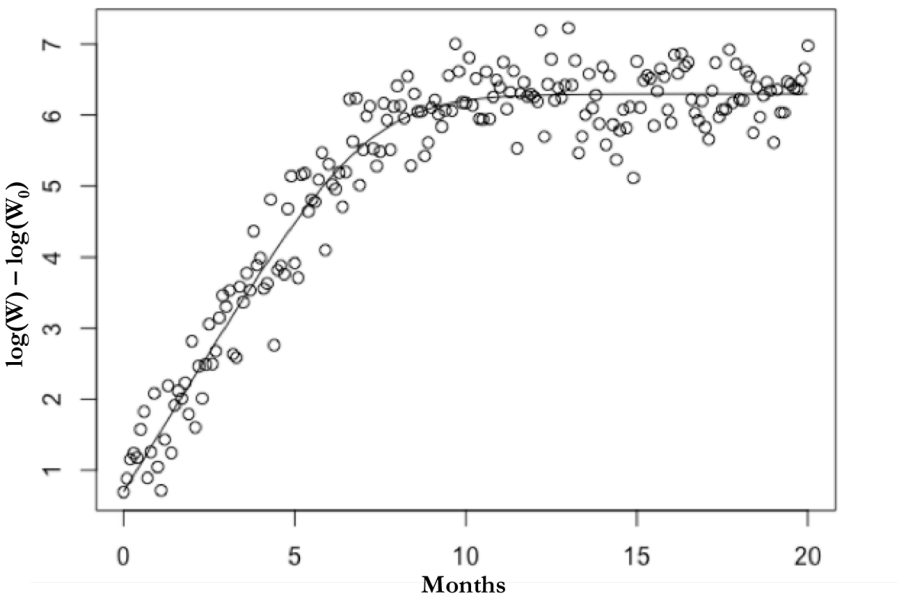


Figure 1. Simulated weight data based on 201 evenly spaced observations between t = 0 and t = 20 months, measurement error $\sigma = 0.4$, and associated NLS trajectory. X-axis: time in months; Y-axis: absolute log-transformed weight.

3.1.2 Comparison of the BCLS and NLS models in accuracy

Figure 2 summarizes the relative biases from estimates of parameters a and b as the measurement error σ increases, using the LS method (dotted lines), the BCLS method (solid line), and the NLS

method (red dotted line). In our model, as the first covariate $\int_0^t ds$ is simply t, we do not expect any bias in estimate for a. Therefore, the BCLS and LS estimates of a are always identical. Panel A of Figure 2 shows how relative biases in the parameter estimates vary with measurement errors in the simulated data when 21 evenly spaced observations are sampled between t = 0 and t = 20 months. Panel B depicts the relative biases in the parameter estimates when the number of observations is increased to 201, within the same time span. Note that in both scenarios, the biases of the LS estimates are dramatically corrected by the BCLS method. When observation points are not densely sampled, we lose some accuracy with the BCLS method, due to integral approximation. The BCLS estimates from densely sampled data sets are more comparable to the NLS estimates. The Monte Carlo means and standard errors, as well as relative biases of parameter estimates from each method, are summarized in Table 1. The efficiency of BCLS method can be evaluated by the Monte Carlo standard errors, which indicate that the variability of the estimation procedure by BCLS method is comparable to NLS, with the NLS method providing slightly more precise estimates (lower standard error). This is expected, as NLS is the maximum likelihood method for parameter estimation in this statistical model.

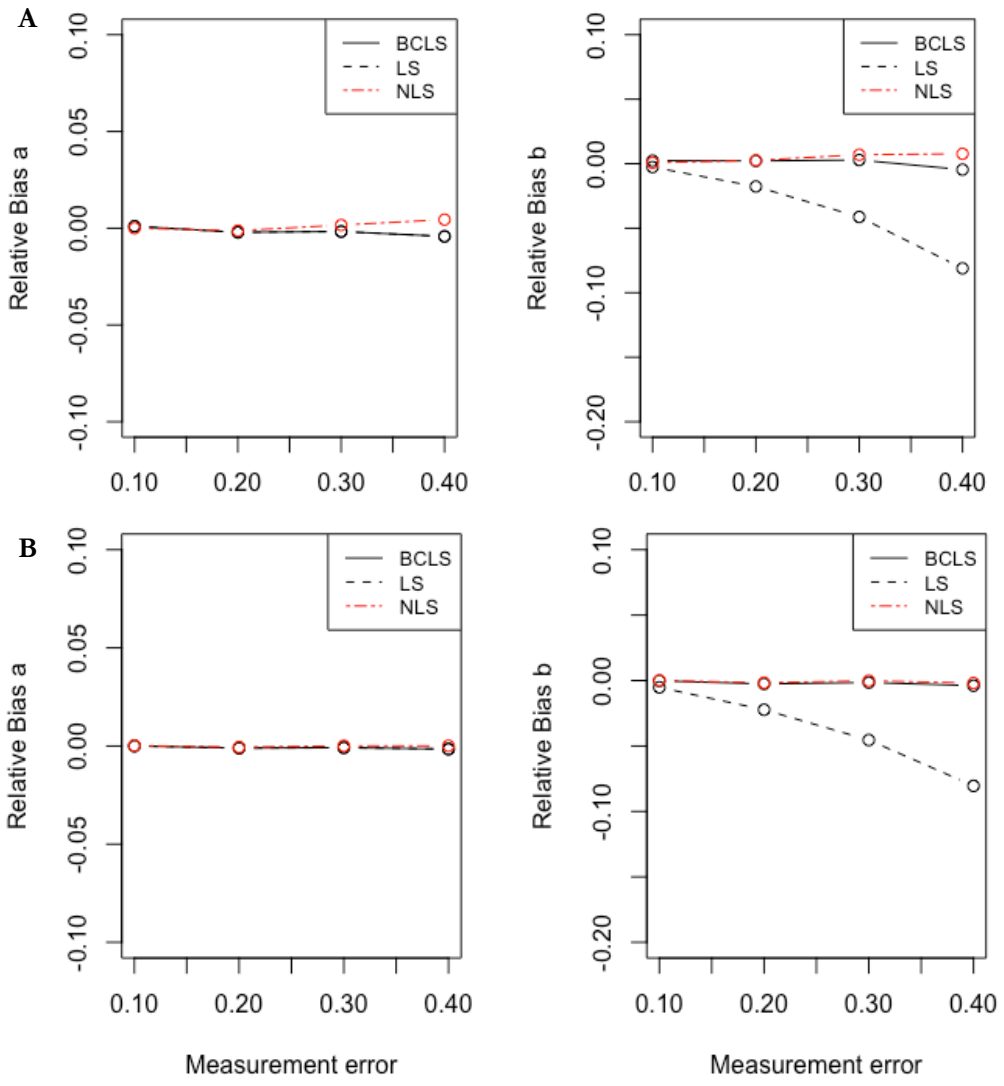


Figure 2. Relative bias of LS (ordinary least squares), BCLS, and NLS estimates using data sets simulated with a range of measurement errors. (A) Relative bias of estimates from data simulated from 21 evenly spaced observations between [0, 20] months. (B) Relative bias of estimates from data simulated from 201 evenly spaced observations between [0, 20] months.

Table 1. Monte Carlo means, standard errors and relative biases in parameter estimates using LS, BCLS and NLS methods.

Number of Observations = 21							
Methods	**Measurement Error**	**a.mean**	**a.sd**	**b.mean**	**b.sd**	**a.relative bias**	**b.relative bias**
LS	0.1	0.8008	0.0148	0.001496	0.000060	0.000987	-0.002645
	0.2	0.7983	0.0301	0.001473	0.000117	-0.002134	-0.017679
	0.3	0.7986	0.0435	0.001438	0.000174	-0.001722	-0.041271
	0.4	0.7966	0.0571	0.001378	0.000215	-0.004245	-0.081128
BCLS	0.1	0.8008	0.0148	0.001504	0.000061	0.000987	0.002354
	0.2	0.7983	0.0301	0.001503	0.000119	-0.002134	0.002166
	0.3	0.7986	0.0435	0.001504	0.000182	-0.001722	0.002857
	0.4	0.7966	0.0571	0.001493	0.000233	-0.004245	-0.004598
NLS	0.1	0.8000	0.0120	0.001501	0.000057	0.000013	0.000765
	0.2	0.7990	0.0232	0.001504	0.000109	-0.001273	0.002597
	0.3	0.8013	0.0348	0.001510	0.000169	0.001624	0.006990
	0.4	0.8036	0.0462	0.001512	0.000212	0.004442	0.007707
Number of Observations = 201							
Methods	**Measurement Error**	**a.mean**	**a.sd**	**b.mean**	**b.sd**	**a.relative bias**	**b.relative bias**
LS	0.1	0.8000	0.0047	0.001492	0.000020	0.000023	-0.005215
	0.2	0.7991	0.0090	0.001467	0.000037	-0.001080	-0.022299
	0.3	0.7993	0.0139	0.001432	0.000056	-0.000850	-0.045434
	0.4	0.7987	0.0192	0.001379	0.000073	-0.001569	-0.080469
BCLS	0.1	0.8000	0.0047	0.001500	0.000020	0.000023	-0.000229
	0.2	0.7991	0.0090	0.001496	0.000037	-0.001080	-0.002548
	0.3	0.7993	0.0139	0.001498	0.000059	-0.000850	-0.001497
	0.4	0.7987	0.0192	0.001494	0.000079	-0.001569	-0.003884
NLS	0.1	0.8002	0.0038	0.001500	0.000019	0.000204	0.000048
	0.2	0.7996	0.0070	0.001497	0.000035	-0.000562	-0.001673
	0.3	0.8000	0.0109	0.001500	0.000053	0.000056	-0.000139
	0.4	0.8001	0.0146	0.001497	0.000069	0.000078	-0.001902

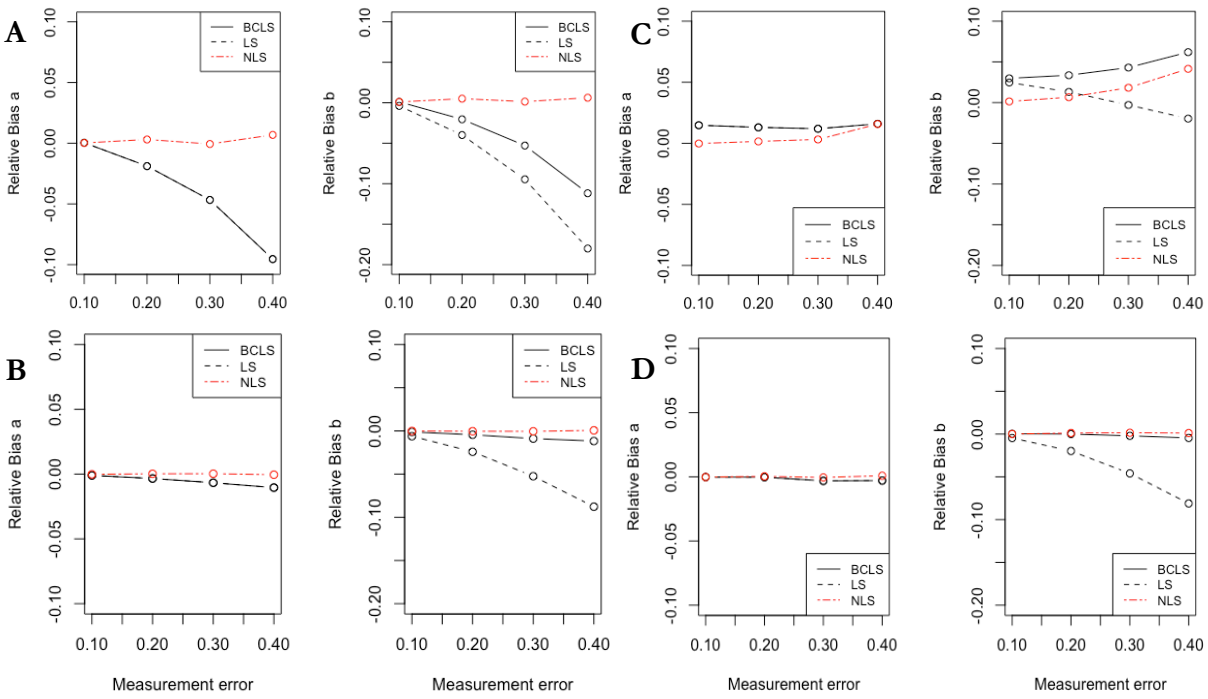


Figure 3. Relative bias associated with BCLS, LS, and NLS estimates based on 1000 iterations with measurement errors varying between 0.1 and 0.4. (A) Relative bias of estimates from data simulated from 21 evenly spaced observations between [0, 60] months. (B) Relative bias of estimates from data simulated from 201 evenly spaced observations between [0, 60] months. (C) Relative bias of estimates from truncated data, the first 7 out of 21 evenly spaced observations between [0, 60] months, within a range of [0, 20] months. (D) Relative bias of estimates from truncated data, the first 67 out of 201 evenly spaced observations between [0, 60] months, within a range of [0, 20] months.

As we expanded the time span for sampling from between t = 0 and t = 20 months to between t = 0 and t = 60 months and sampled 21 evenly spaced observations within the time span, we observed unexpected biases in BCLS estimates for both a and b (Figure 3A). The increased biases of BCLS estimates are likely to originate from the small sample properties due to sparse samplings, as well as increased sampling ratio from the steady state, which may violate the

assumption A.3 for BCLS method described in section 2.1.1 (Holte 2016). To compensate for the small sample properties, we increased the sampling density to 201 evenly spaced observations between t = 0 and t = 60 months, which significantly corrected the biases with large measurement errors, validating the expected asymptotic unbiasedness of estimate for b (Figure 3B).

To investigate whether the biases observed can be attributed to samplings through the steady state, we simulated data from 21 or 201 evenly spaced observations between t = 0 and t = 60 months, but fit the LS, BCLS, and NLS models using truncated data from observations spanning only the dynamic range between t = 0 and t = 20 months (totally 7 or 67 observations). The relative biases of parameter estimates using truncated data sets are summarized in Figure 3C and D. In contrast with the estimates from complete data, other than slight biases in the initial estimates, the BCLS estimates from the dynamic range of the data are generally unbiased, comparable to NLS estimates. With more details discussed in Chapter 4, the fundamental mechanism of the biases needs further investigation and proof.

3.1.3 Comparison of the BCLS and NLS models in computational efficiency

In the simple least squares models, both the BCLS and NLS methods successfully converged under a range of measurement errors (σ up to 0.4). To compare the BCLS and NLS methods in computational efficiency, we simulated data sets with 201 evenly spaced observations between t = 0 and t = 20 months as described in section 3.1.1, with parameters defined as $a = 0.8$, $b = 0.0015$, $\sigma = 0.1$ .Each model was fit to the same data set and the computation time was recorded through iterations. The mean computation time and standard errors from ten 100-iteration trials (1000 iterations in total) using either the BCLS method or NLS method are summarized in Table 2. From 1000 iterations, the BCLS method reduced the computation time by 24%, suggesting superior computational efficiency.

Table 2. Comparison of BCLS and NLS methods in computation time using simulated data.

Methods	**Mean Computation Time**	**Standard Errors**
BCLS	15.680	0.907
NLS	20.622	1.196

3.2 The BCLS Method in Mixed-Effects Models Parameter Estimation

As the BCLS method has been demonstrated to be superior to NLS in computational efficiency, we further explore the method's validity and performance in mixed-effects ODE models. In realistic longitudinal data, such as the OMC data set described in section 2.3, repeated measurements are taken for an aggregate population of individuals, thus requiring consideration of the intra-subject correlation. By considering the random effects for measurements clustered by individuals, the mixed-effects ODE models can flexibly represent the correlation structure.

Although the asymptotic results of the BCLS method in mixed-effects models are yet to be proven, we first applied the BCLS-LME to simulated data with random effects and evaluated the estimates for parameters a and b. The major goals of the analyses include 1) To compare the accuracy and precision of the BCLS-LME and NLME methods; 2) To identify the range of scenarios where the BCLS-LME and/or NLME methods can be applied; 3) To compare the computational efficiency of the two methods.

3.2.1 Simulation of growth curve data with random effects for an aggregate population of individuals

We simulated data representing growth curves of a children's population, taking into account the correlations within repeated measurements from each child, which can be reflected by the random effects from a and b. For each child out of a population of 100, we simulated growth data from 21 or 201 randomly sampled observations between t = 0 and t = 20 months following the same strategy described in section 2.2.1 and fit the data with linear mixed-effects model adopting the BCLS method (BCLS-LME) and common nonlinear mixed-effects model (NLME) respectively. The accuracy of parameter estimates and computational efficiencies were evaluated and compared.

3.2.2 Comparison of the BCLS-LME and NLME models in accuracy

To avoid potential convergence difficulties in estimation using the NLME method, we deliberately used small random effects for a and b in data simulation ($\sigma_a = 0.1, \sigma_b = 0.0001$). The relative biases of parameter estimates from the BCLS-LME and NLME models, given the same range of measurement errors, are plotted in Figure 4; the Monte Carlo means and standard errors for the parameters estimates from both methods are summarized in Table 3. Small biases were observed in the BCLS-LME estimates when measurements were collected from 21 observation times, compared to the NLME estimates. Increasing the sampling density to 201 observations within the same time range successfully corrected this bias (Figure 4B), demonstrating the expected asymptotic unbiasedness of BCLS estimates in the mixed-effects models. In addition, the Monte Carlo standard errors of the parameter estimates by BCLS-LME method are reasonably close to the standard errors of the NLME estimates, especially with 201 observations sampled (Table 3). This suggests that in the mixed-effects models, the BCLS-LME method is also comparable to the NLME method in efficiency.

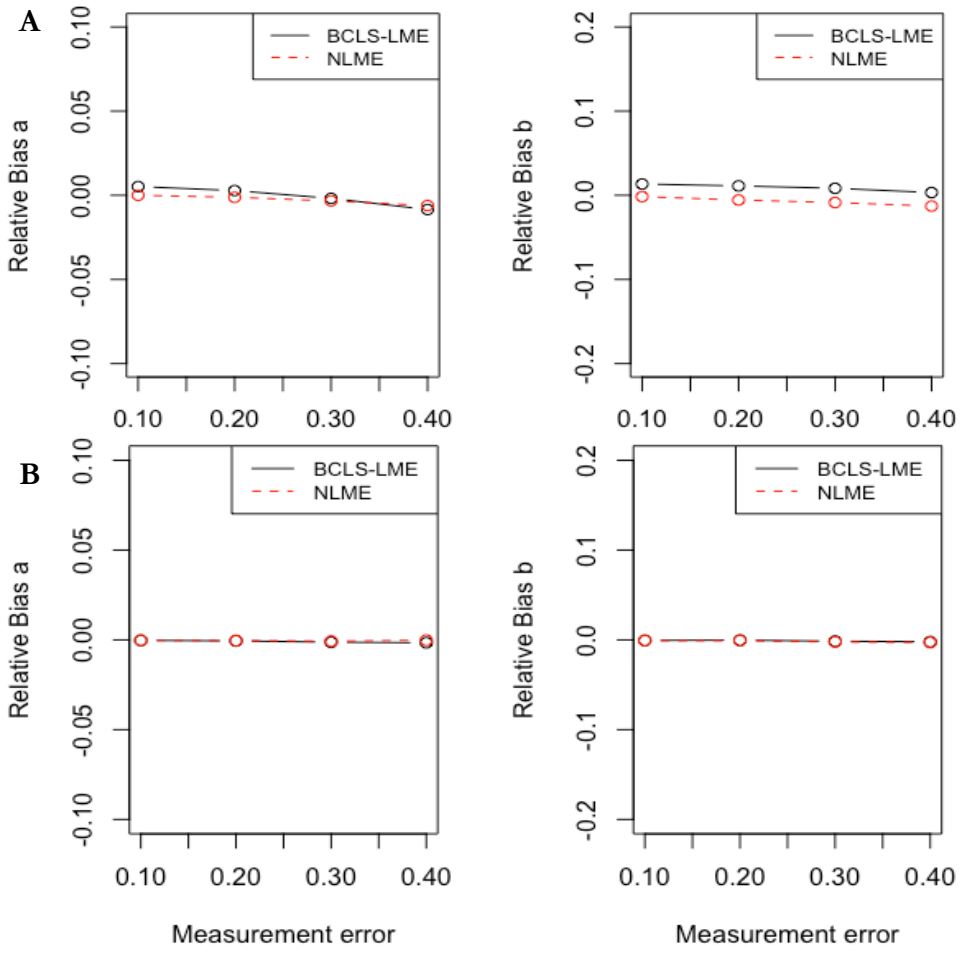


Figure 4. Relative bias in parameter estimates using the BCLS-LME or NLME methods with measurement error varying between 0.1 and 0.4. (A) Relative bias of estimates from data simulated with 21 randomly sampled observations between [0, 20] months. (B) Relative bias of estimates from data simulated with 201 randomly sampled observations between [0, 20] months.

Table 3. Monte Carlo means, standard errors and relative bias of parameter estimates using BCLS-LME or NLME methods.

Number of observations = 21							
Method	**Measurement Error**	**a.mean**	**a.sd**	**b.mean**	**b.sd**	**a.relative bias**	**b.relative bias**
BCLS-LME	0.1	0.8041	0.01432	0.001520	0.000019	0.005170	0.013367
	0.2	0.8022	0.01562	0.001517	0.000028	0.002795	0.011113
	0.3	0.7985	0.01552	0.001512	0.000034	-0.001836	0.008313
	0.4	0.7932	0.01654	0.001505	0.000047	-0.008491	0.003233
NLME	0.1	0.8000	0.01423	0.001498	0.000017	0.000004	-0.001615
	0.2	0.7990	0.01520	0.001492	0.000024	-0.001223	-0.005644
	0.3	0.7973	0.01479	0.001487	0.000028	-0.003334	-0.008644
	0.4	0.7951	0.01561	0.001481	0.000037	-0.006112	-0.012860
Number of observations = 201							
Method	**Measurement Error**	**a.mean**	**a.sd**	**b.mean**	**b.sd**	**a.relative bias**	**b.relative bias**
BCLS-LME	0.1	0.7998	0.01465	0.001499	0.000015	-0.000279	-0.000609
	0.2	0.7995	0.01413	0.001499	0.000016	-0.000576	-0.000351
	0.3	0.7990	0.01419	0.001498	0.000017	-0.001304	-0.001484
	0.4	0.7987	0.01449	0.001497	0.000019	-0.001607	-0.002146
NLME	0.1	0.7998	0.01464	0.001499	0.000015	-0.000279	-0.000872
	0.2	0.7997	0.01411	0.001499	0.000016	-0.000394	-0.000873
	0.3	0.7994	0.01425	0.001497	0.000016	-0.000722	-0.002283
	0.4	0.7998	0.01445	0.001496	0.000018	-0.000205	-0.002916

Similar to what we have observed in section 3.1.2, increased biases also arise in the estimates for both a and b, when the measurements are sampled from 21 observation times within an expanded time range, between t = 0 and t = 60 months (data not shown). We increased the

sampling density to 201 or 601 observations within the same time range and observed parameters estimates from the BCLS-LME method became comparably unbiased, suggesting the expected asymptotic unbiasedness of the BCLS-LME estimates. The potential causes of the observed biases will be further discussed in Chapter 4; and the solutions are yet to be explored.

3.2.3 Comparison of the BCLS-LME and NLME models in computational efficiency

We evaluated the efficiency of the BCLS-LME and NLME methods through comparison in convergence rates and computation time on simulated data sets. To compare the two methods in convergence rate, we first set the measurement error as $\sigma = 0.1$, and simulated data sets with combinations of various random effects from a and b : $\sigma_a = (0.1, 0.2, 0.4)$, $\sigma_b = (0.0001, 0.0005)$. Then the BCLS-LME and NLME models are respectively fit to the data sets and evaluated for successfully converged iterations out of 100 total iterations. To ensure fair comparisons, we set the maximum number of iterations for the optimization algorithm as 200 for both methods (default 50). As shown in Table 4, the NLME method encountered convergence difficulties as random effects increased in a, b, or both, while models fitted with the BCLS-LME method successfully converged at all trials, suggesting a superior convergence efficiency of the BCLS-LME method.

Table 4. Convergence rates for the NLME and BCLS-LME methods.

σ_a	σ_b	Conv. Rate-NLME	Conv.Rate- BCLS-LME
0.1	0.0001	100/100	100/100
0.1	0.0005	39/100	100/100
0.2	0.0001	100/100	100/100
0.2	0.0005	51/100	100/100
0.4	0.0001	99/100	100/100
0.4	0.0005	67/100	100/100

We then evaluated the computation time for each method, using data sets simulated with 21 or 201 observations. We set the measurement error and the random effects at permissive values, $\sigma = 0.1, \sigma_a = 0.1, \sigma_b = 0.0001,$ to avoid the convergence difficulties that the NLME method may encounter (described above) and ensure fair comparisons. The mean computation time from ten 100-iteration trials (a total of 1000 iterations) of each method is summarized in Table 5. In all sampling schemes, the BCLS-LME method made dramatic improvement compared to the NLME method, by shortening the computation time by 2.5-4.1 folds.

Table 5. Computation time for the NLME and BCLS-LME methods under differential sampling schemes.

Method	**Observations Sampled**	**Mean Computation Time**
NLME	21	12.894
	201	75.514
BCLS-LME	21	5.169
	201	18.284

3.3 Children's Growth Data from the Observational Malaria Cohort (OMC)

One of the well-known applications of ODE models with mixed-effects is to estimate and interpret the growth rates in population growth data. We utilized a population data set from a longitudinal birth cohort in Tanzania (Goncalves, Huang et al. 2014), the observational malaria cohort (OMC) data set, to test the BCLS-LME and NLME methods in parameter estimation with mixed-effects taken into consideration.

The OMC data set and the characteristics of the study population were previously described in section 2.3. The longitudinal data of log-transformed weight for the first 33 subjects are plotted in Figure 5, suggesting a logistic growth curve in general. It is noticeable that although the growth

curves had not reached the fully steady state by the end of the study, the general growth rate significantly slowed down after the first year of dynamic range. Using the OMC data set pre-processed as described in section 2.3, we also applied the BCLS-LME and NLME methods to investigate the effects of health events, such as parasite burden and malaria episodes, on children's growth rate.

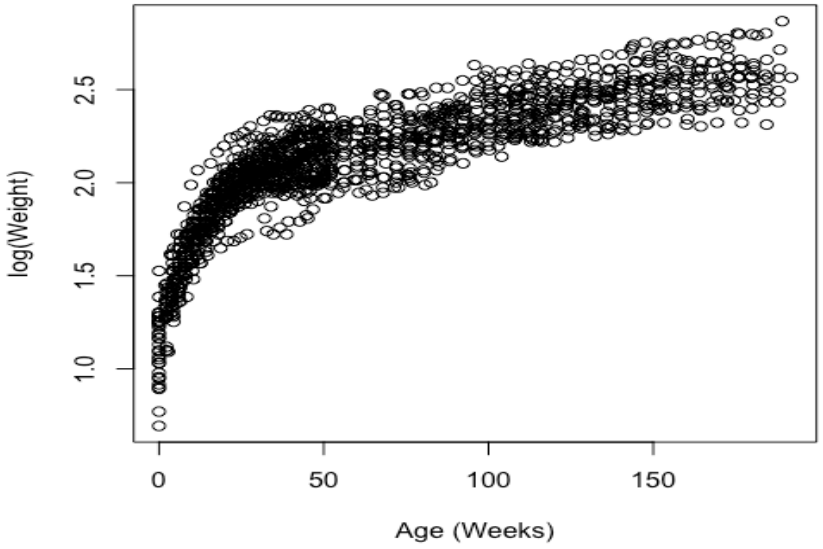


Figure 5. A plot of log-transformed weight data from the first 33 subjects of the OMC data set.

3.3.1 Estimate the growth rate using the BCLS-LME and NLME models

Using the BCLS-LME and NLME methods described in section 3.2, we estimated the parameters a and b, and the standard errors by fitting models to the cleaned OMC data set (Table 6). The BCLS-LME method yielded comparable estimates for b with estimates by the NLME method, while slight biases existed in estimates for a. The standard errors of parameters estimates suggest similar efficiencies of estimation using either the BCLS-LME method or the NLME method. The observed biases are likely attributed to small sample properties and samplings during the steady state, as we discussed above. As it is non-practical to increase the sampling density in a completed study, we

focus on testing the second possibility. We truncated the data set by removing all measurements collected after a year since enrolled in the study (365 days, 52 weeks) when the growth curve is reaching a relatively steady state. We also excluded any subject who had less than two measurements in the truncated data set. Using the truncated data set, we again fit the BCLS-LME and NLME models respectively to estimate the parameters. As shown in Table 6, the estimates for a and b by BCLS-LME model (BCLS-LME.365) are approaching the estimates by the NLME model (NLME.365).

Table 6. Estimates of means and random effects for growth rates using the OMC data set.

Model	**a.mean**	**a.sd**	**b.mean**	**b.sd**	γ
NLME	0.084911	0.026327	0.008917	0.003461	/
BCLS-LME	0.078105	0.026798	0.008840	0.003740	/
NLME + r_{gender}	0.082755	0.026165	0.008924	0.003466	0.004237
BCLS-LME+r_{gender}	0.076202	0.026659	0.008844	0.003744	0.003694
NLME + r_{blood}	0.084833	0.026143	0.008893	0.003435	-0.000608
BCLS-LME+r_{blood}	0.077962	0.026645	0.008811	0.003713	-0.000259
NLME + $r_{malever}$	**FTC***				
BCLS-LME+ $r_{malever}$	0.079496	0.026546	0.008819	0.003727	-0.001805
NLME + r_{nmal}	**FTC**				
BCLS-LME + r_{nmal}	0.078659	0.026484	0.008831	0.003741	-0.000092
NLME.365	0.110537	0.026117	0.013025	0.003757	/
BCLS-LME.365	0.110019	0.026695	0.014203	0.004196	/
NLME.365 + r_{gender}	**FTC**				
BCLS-LME.365 + r_{gender}	0.107070	0.026622	0.014207	0.004203	0.005717
NLME.365 + r_{blood}	0.110490	0.025937	0.013021	0.003730	-0.000171
BCLS-LME.365 + r_{blood}	0.109993	0.026579	0.014201	0.004178	-0.000057
NLME.365 + $r_{malever}$	**FTC**				
BCLS-LME.365+ $r_{malever}$	0.111400	0.026622	0.014192	0.004195	-0.001749
NLME.365 + r_{nmal}	**FTC**				
BCLS-LME.365+ r_{nmal}	0.110960	0.026624	0.014193	0.004196	-0.000153

***** FTC: Failed to converge.

3.3.2 Estimate the effects from other covariates on children's growth rate

One of our major goals analyzing the OMC data set is to investigate the effects of health events on children's growth curve. We are interested in evaluating the effects from multiple covariates on children's growth rate, including gender (*gender*), positive parasite counts in blood smear samples (*blood*), ever diagnosis of malaria (*malever*), and the number of malaria episodes (*nmal*). As an example, we included gender as a covariate in the alternative form of ODE:

$$\frac{dW}{dt} = W\big((a^* + \gamma * gender) - bW\big)$$

Using the same procedure described in section 3.2, we fit the BCLS-LME and NLME models to estimate the new growth rate, a^* and b, and the coefficient for gender, γ. Both the BCLS-LME and NLME methods yielded similar conclusions based on the estimates of the coefficients (Table 6): the study subjects tend to have slower growth rate if they were female, with positive parasite counts in the blood smear samples, or any previously diagnosed malaria episodes, although the effects are subtle for all. Taking the standard errors into consideration, the BCLS-LME method yielded estimates of the same direction with the NLME estimates, with slightly different altitudes.

As increasing the sampling density is nonrealistic with a completed study, we again truncated the data set to measurements limited to the first year of study and fit both models, with the purpose of evaluating the effects only during the dynamic state. The BCLS-LME and NLME estimates for a^* are now indistinguishable, confirming our previous hypothesis that additional samplings in steady state lead to further biases. Interestingly, the NLME method failed to converge under a number of model designs, while the BCLS-LME method succeeded in all, suggesting the superiority of the BCLS-LME method in computational efficiency, especially given extremely sparse observations.

Chapter 4. Discussion

ODEs are popular modeling tools to describe dynamic systems in biology, engineering, health care and many other fields. However, estimating parameters from observational data has always been a challenging statistical problem, as most ODEs have no analytic solutions, and solving ODEs numerically is computationally intensive. Among various parameter estimation algorithms for nonlinear ODE models, the numerical method, nonlinear least squares (NLS), has been regarded as the "Gold Standard" with ideal asymptotic properties (Bates and Watts 1988). The NLS method faces many challenges in ODE parameter estimation. One of those originates from the nature of NLS estimation using conventional gradient-based optimization methods such as Gauss-Newton or Levenberg-Marquart, that the iterative numerical procedure to obtain the estimates leads to instability of the procedure (Press 1986). In addition, by the nature that the NLS estimators are computed using iterative procedures, it requires a guess of the starting values for a minimizer. Another major challenge is that the NLS procedures may fail to converge to global minima depending on the choice of starting values for iteration, particularly if the least squares solution has multiple local minima or saddle points. Alternative methods have been proposed for parameter estimation in nonlinear ODE models. For example, the PsLS method described as a two-step direct method relying on differentiation (Liang and Wu 2008), and the method by Ramsay et al., which introduced a parameter cascading procedure and a smoothing-involved approximation of ODE solutions (Ramsay, Hooker et al. 2007).

Estimating ODE models with mixed-effects is generally even more challenging. Incorporating random effects in the classic nonlinear mixed-effects models are already known to be difficult statistical problems (Pinheiro and Bates 2000). When the ODEs have analytic solutions, they are commonly solved as a nonlinear mixed-effects model (NLME) with available packages like

nlme( ) (Pinheiro J 2017). Given the great burden of random effects in ODE models, estimation using the NLME method is often computationally intensive, and commonly faces convergence difficulties. When the analytic solutions are not available, the ODEs can only be solved by numerical methods. Software packages and methods have been developed to solve the ODEs numerically and estimate the mixed-effects ODE models. NONMEM is widely used in the pharmacometrics industry to estimate nonlinear mixed-effect models (Wang 2007); NONMEM and another software, MONOLIX, have also implemented EM-type algorithms such as stochastic approximation EM (SAEM) to estimate nonlinear mixed-effects models (Kuhn 2005). Both software programs require pre-specified initial conditions for ODE models to obtain numerical solutions, which is often hard to determine. Wang et, al. proposed a semi-parametric approach, which estimates a spline function that approximates the ODE solution using smoothing splines (Wang 2014). Similar to the other direct methods, Wang's method also involves arbitrarily specifying the number and location of knots for spline functions, especially when the observations are sparse.

The bias-corrected least squares (BCLS) method (Holte 2016) is a computationally simple, non-iterative, and easily implemented method for parameters estimation in ODE models. It retains the simplicity of early direct methods and has most of the desirable statistical properties of more recently developed direct methods. A major advantage of the BCLS method (and other direct methods) is that it does not require starting values for the estimation algorithm. Furthermore, it does not require the choice of smoothing bandwidth of functional data analysis methods. Since it is based on integration rather than differentiation, the BCLS method can be used to estimate the initial ODE model states in addition to the ODE model parameters, which is superior to other direct methods that rely on differentiation, such as the PsLS method (Liang and Wu 2008).

In this thesis, we have simulated longitudinal data sets to represent the growth curve of a single individual and demonstrated the accuracy and computational efficiency of the BCLS method

in parameter estimation in contrast with the LS and NLS methods (section 3.1). With simulated longitudinal data involving only fixed effects, and a proper sampling scheme with sufficient observations, the BCLS method provides unbiased estimates of ODE parameters and Monte Carlo standard errors comparable to those of the NLS method. It surpasses the NLS method with remarkably reduced computation time, averagely 24% through 1000 iterations. We demonstrated strong results to show that the common method for estimating nonlinear mixed-effects model, NLME, encountered convergence difficulties as increased random effects were introduced to a, b, or both; while the BCLS-LME method successfully converged in all trials. Furthermore, due to excessive estimation for large random effects, the computation time gets extremely long when fitting the NLME model to longitudinal data from an aggregate of population. By reducing the nonlinear model to a linear model, the BCLS-LME method was able to reduce the computation time by 2.5-4.1 folds. With dramatically improved computational efficiency, the BCLS-LME method still managed to obtain generally unbiased estimates under a wide range of random effects and measurement errors, given properly sampled observations. With the asymptotic unbiasedness yet to be proven in ODE models with mixed-effects, we are impressed by the performance of the BCLS-LME method in parameter and random effects estimation, particularly in computational efficiency.

Through the tests we performed with differential sampling schemes, we noticed biases arising when we applied the BCLS method to data sets simulated with sparse observations in a longer time span. These unexpected biases were first observed in the BCLS estimates for both a and b when data were simulated with 21 evenly spaced observations between t = 0 and t = 60 months, and the measurement errors being as large as 0.4 (Figure 3). The observed biases could be attributed to the following causes: 1) When the observations are not densely sampled, the integral approximations by the BCLS method may lead to biases in estimation. 2) Referring to the model assumption A.3 in Chapter 2, our data-sampling scheme in BCLS-LME may have violated the

assumption that not all data points are from the equilibrium state (steady state). We tested the first potential cause by increasing the sampling density from 21 to 201 and 601 evenly spaced observations within the same time span (Figure 3 and data not shown). As the observations become denser, the biases from the BCLS estimates for a and b shrunk dramatically, suggesting the asymptotic unbiasedness of the BCLS estimators. To test the second cause, we sampled 21 or 201 evenly spaced observations between t = 0 and t = 60 months in simulations, but fit the BCLS model using only data points from observations spanning the dynamic state, between t = 0 and t =20 months (totally 7 or 67 observations). In contrast with the estimates obtained using the full-range data, other than slight bias in the initial estimates, the BCLS estimates from the dynamic state are generally unbiased, comparable to the NLS estimates (Figure 3C, D). These tests suggest a mysterious phenomenon, in which additional samplings from the steady state of a logistic growth curve actually lead to increased biases in parameter estimates. The underlying mechanism requires more careful thinking and further efforts to be discovered.

An important application of the BCLS-LME method in this thesis is to estimate the growth rates and effects from covariates of interest using the OMC data set on children's growth. Taking the standard errors into consideration, the BCLS-LME method yielded comparable estimates for growth rate to the NLME estimates. When fitting both models to evaluate the effects of covariates, such as gender, parasite count, and malaria episodes, the coefficients estimated by BCLS-LME and NLME methods were of the same effect direction. As we truncated the data set to measurements only within the first year of study, eliminating samplings from the steady state, the BCLS-LME and NLME estimates for a became indistinguishable, confirming our previous hypothesis that additional samplings in the steady state may lead to further biases. Interestingly, the NLME method failed to converge under a number of model designs to estimate covariates effects, while the BCLS-LME method succeeded in all, suggesting its notable superiority in computational efficiency.

In conclusion, this thesis has implemented the BCLS method in nonlinear ODE models, with fixed effects only and with random effects taken into consideration. With decent accuracy in parameter estimation compared to the traditional NLS method, we demonstrate the great superiority of the BCLS method in convergence rate and computation time, especially in the mixed-effects models, given properly sampled observations. Further investigations are required to prove the asymptotic unbiasedness of the BCLS-LME estimates in mixed-effects model, and to explain the mysterious biases caused by additional samplings in the steady state. With those questions answered, the BCLS-LME method is likely to provide a powerful tool for parameter estimation in complex ODE mixed-effects model.

Bibliography

1. Bard, Y. (1974). Nonlinear parameter estimation. New York, Academic Press.

2. Bates, D. M. and D. G. Watts (1988). Nonlinear regression analysis and its applications. New York, Wiley.

3. Cook, R. D., C. L. Tsai and B. C. Wei (1986). "Bias in Nonlinear-Regression." Biometrika **73**(3): 615-623.

4. Csajka, C. and D. Verotta (2006). "Pharmacokinetic-pharmacodynamic modelling: history and perspectives." J Pharmacokinet Pharmacodyn **33**(3): 227-279.

5. Danhof, M., E. C. de Lange, O. E. Della Pasqua, B. A. Ploeger and R. A. Voskuyl (2008). "Mechanism-based pharmacokinetic-pharmacodynamic (PK-PD) modeling in translational drug research." Trends Pharmacol Sci **29**(4): 186-191.

6. Dartois, C., K. Brendel, E. Comets, C. M. Laffont, C. Laveille, B. Tranchand, F. Mentre, A. Lemenuel-Diot and P. Girard (2007). "Overview of model-building strategies in population PK/PD analyses: 2002-2004 literature survey." Br J Clin Pharmacol **64**(5): 603-612.

7. Dattner, I. and C. Klaassen (2015). "Optimal Rate of Direct Estimators in Systems of Ordinary Differential Equations Linear in Functions of the Parameters." Electronic Journal of Statistics.

8. Fitzmaurice, G. M., N. M. Laird and J. H. Ware (2011). Applied longitudinal analysis. Hoboken, N.J., Wiley.

9. Foss, S. D. (1971). "Estimates of Chemical Kinetic Rate Constants by Numerical Integration." Chemical Engineering Science **26**(3): 485-&.

10. Freedman, H. I. (1980). Deterministic mathematical models in population ecology. New York ; Basel, Dekker.

11. Goncalves, B. P., C. Y. Huang, R. Morrison, S. Holte, E. Kabyemela, D. R. Prevots, M. Fried and P. E. Duffy (2014). "Parasite burden and severity of malaria in Tanzanian children." N Engl J Med **370**(19): 1799-1808.

12. Himmelblau, D. M., C. R. Jones and K. B. Bischoff (1967). "Determination of Rate Constants for Complex Kinetics Models." Industrial & Engineering Chemistry Fundamentals **6**(4): 539-+.

13. Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard and M. Markowitz (1995). "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection." Nature **373**(6510): 123-126.

14. Holte, S. (2016). "A consistent direct method for estimating parameters in models defined by ordinary differential equation. ." https://arxiv.org/abs/1601.04736.

15. Hosten, L. H. (1979). "A Comparative-Study of Short Cut Procedures for Parameter-Estimation in Differential-Equations." Computers & Chemical Engineering **3**(1-4): 117-126.

16. Jacquez, J. A. (1972). Compartmental analysis in biology and medicine. Kinetics of distribution of tracer-labeled materials. Amsterdam, New York,, Elsevier Pub. Co.

17. Kuhn, E., Lavielle, M. (2005). "Maximum likelihood estimation in nonlinear mixed effects models." Comput. Stat. Data Anal.(49): 1020–1038.

18. Liang, H. and H. Wu (2008). "Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models." J Am Stat Assoc **103**(484): 1570-1583.

19. Lorenz, E. N. (1963). "Deterministic Nonperiodic Flow." Journal of Atmospheric Sciences **26**: 130-141.

20. Perelson, A. S., P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz and D. D. Ho (1997). "Decay characteristics of HIV-1-infected compartments during combination therapy." Nature **387**(6629): 188-191.

21. Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard and D. D. Ho (1996). "HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time." Science **271**(5255): 1582-1586.

22. Pinheiro J, B. D., DebRoy S, Sarkar D and R Core Team (2017). "nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131."

23. Pinheiro, J. C. and D. M. Bates (2000). Mixed effects models in S and S-PLUS. New York, Springer.

24. Press, W. H. (1986). Numerical recipes in FORTRAN : the art of scientific computing. Cambridge England ; New York, NY, USA, Cambridge University Press.

25. Ramsay, J. O., G. Hooker, D. Campbell and J. Cao (2007). "Parameter estimation for differential equations: a generalized smoothing approach." Journal of the Royal Statistical Society Series B-Statistical Methodology **69**: 741-770.

26. Ramsay, J. O. and B. W. Silverman (2005). Functional data analysis. New York, Springer.

27. Swartz, J. B., H. (1975). "Discussion of parameter estimation in biological modelling: Algorithms for estimation and evaluation of the estimates." Journal of Mathematical Biology **1**: 241-257.

28. Van Belle, G. and L. Fisher (2004). Biostatistics : a methodology for the health sciences. Hoboken, NJ, John Wiley & Sons.

29. Varah, J. M. (1982). "A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations." SIAM, Journal of Scientific and Statistical Computation(3): 28-46.

30. Wang, L., Cao, J., Ramsay, J.O. (2014). "Estimating mixed-effects differential equation models." Stat Comput **24:111–121**.

31. Wang, Y. (2007). "Derivation of various NONMEM estimation methods." J Pharmacokinet Pharmacodyn **34**(5): 575-593.

32. Wei, X., S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn and et al. (1995). "Viral dynamics in human immunodeficiency virus type 1 infection." Nature **373**(6510): 117-122.

33. Wikstrom, G. (1997). "Computation of parameters occurring linearly in systems of ordinary differential equations, Part i." Technical Report, Department of Computing Science, Umea University.

34. Zhang, W.-B. (2005). Differential equations, bifurcations, and chaos in economics. Hackensack, N.J., World Scientific.

APPENDIX

APPENDIX.1 R script to simulate data and fit LS/BCLS/NLS models.

```
############################################################
################## BCLS vs. LS vs. NLS ####################
############################################################
library(deSolve)
library(nlme)
library(minpack.lm)

# Analytical solution function
spanal<-function(t,p0,a,b) {
  k<-p0/(a-b*p0)
  a*k*exp(a*t)/(1+b*k*exp(a*t))
}

# Initial conditions
y0 = 2
t = c(0:200)/10
n = length(t)
a = 0.8
b = 0.0015
parms<-c(a,b)
reps = 100
mp<-spanal(t, y0, a, b)
merr<-c(0.1,0.2,0.3,0.4)

# Simulate data and fit models
ests = NULL

for (sigma in merr) {
  for (i in c(1:reps)) {
    outTable1 = NULL
    e  = rnorm(n, 0, sigma)
    logy = log(mp) + e
    logy[1] = log(y0)
    newy = exp(logy)
    resnewy = logy - log(y0)

    zi.uncor = c(rep(0,n))
    zi = c(rep(0,n))
    ztotal.uncor = c(rep(0,n))
    ztotal = c(rep(0,n))

    for (p in c(2:n)) {
      zi.uncor[p] = 0.5 * (t[p] - t[p-1]) * (newy[p] + newy[p-1])
      zi[p] = 0.5 * (t[p] - t[p-1]) * (newy[p] + newy[p-1]) * exp(-
0.5*sigma^2)
      ztotal.uncor[p] = sum(zi.uncor[1:p])
      ztotal[p] = sum(zi[1:p])
    }
    outTable1 = data.frame(cbind(t, logy, resnewy, ztotal, ztotal.uncor))

    lm.fit.uncor = lm(resnewy~ t + ztotal.uncor -1, data = outTable1)
```

```
    lm.fit = lm(resnewy~ t + ztotal -1, data = outTable1)
    nls.fit = nlsLM(logy ~ log(spanal(t, y0, a, b)),
                    data = outTable1,
                    start = c(a = 0.8, b = 0.0015),
                    trace = FALSE)
    ests=rbind(ests,c(sigma,lm.fit.uncor$coef,lm.fit$coef,
summary(nls.fit)$coefficient[,1]))
  }
}

ests=data.frame(ests)
names(ests)=c("sigma","a.uncorr","b.uncorr","a.corr","b.corr", "a.nls",
"b.nls")

res=NULL
for (sigma in unique(ests$sigma) ) {
  each = ests[ests$sigma==sigma, ]
  res = rbind(res,c(sigma,round(mean(each$a.uncorr),dig=5),sd(each$a.uncorr),
                    -round(mean(each$b.uncorr),dig=6),sd(each$b.uncorr),
                    (mean(each$a.uncorr)-a)/a,(-mean(each$b.uncorr)-b)/b,
                    round(mean(each$a.corr),dig=5),sd(each$a.corr),
                    -round(mean(each$b.corr),dig=6),sd(each$b.corr),
                    (mean(each$a.corr)-a)/a,(-mean(each$b.corr)-b)/b,
                    round(mean(each$a.nls),dig=5), sd(each$a.nls),
                    round(mean(each$b.nls), dig = 6),sd(each$b.nls),
                    (mean(each$a.nls)-a)/a, (mean(each$b.nls) -b)/b
  ) )
}

res<-data.frame(res)
names(res)<-c("sigma","a.LS", "a.LS.sd",
              "b.LS","b.LS.sd",
              "biasA.LS","biasB.LS",
              "a.BCLS","a.BCLS.sd",
              "b.BCLS","b.BCLS.sd",
              "biasA.BCLS","biasB.BCLS",
              "a.NLS", "a.NLS.sd",
              "b.NLS", "b.NLS.sd",
              "biasA.NLS", "biasB.NLS")

# Plot estimates
par(mfrow=c(1,2))
plot(res$sigma,res$biasA.BCLS, xlab="Measurement error",ylab="Relative Bias
a",type="b", ylim = c(-0.1, 0.1))
points(res$sigma,res$biasA.LS,lty=2,type="b")
points(res$sigma, res$biasA.NLS,lty=6, type = "b", col = "red")
legend("bottomright", cex=0.8,
       legend=c("BCLS", "LS", "NLS"),
       lty=c(1, 2, 6),
       col = c("black", "black", "red"))

plot(res$sigma,res$biasB.BCLS, xlab="Measurement error",ylab="Relative Bias
b",type="b", ylim = c(-0.2, 0.1))
points(res$sigma,res$biasB.LS,lty=2,type="b")
points(res$sigma, res$biasB.NLS,lty=6, type = "b", col = "red")
legend("bottomright", cex=0.8,
       legend=c("BCLS", "LS", "NLS"),
       lty=c(1, 2, 6),
```

```
       col = c("black", "black", "red"))
```

APPENDIX.2 R script to simulate data and fit BCLS-LME/NLME models.

```
############################################################
################ BCLS-LME vs. NLME   ######################
############################################################
library(deSolve)
library(nlme)
library(minpack.lm)

# Analytical solution function
spanal<-function(t,p0,a,b) {
  k<-p0/(a-b*p0)
  a*k*exp(a*t)/(1+b*k*exp(a*t))
}

# Initialize parameters
y0 = 2
reps = 50
a.mean = 0.8
b.mean = 0.0015
r = 0.05
sigmaA = 0.1
sigmaB = 0.0001
iter = 1000
merr = c(0.1, 0.2, 0.3, 0.4)
num = 200

# Function to simulate data
simuData = function(outTable, sigma){
  for (i in c(1:reps)) {
    min = 0; max = 20
    r = 0.05
    u = runif(num, min = exp(-r*max), max = exp(-r*min))
    newt = c(0,sort((1/r)*log(1/u)))
    n = length(newt)
    id = rep(i, n)
    a = rnorm(1, a.mean, sigmaA)
    b = rnorm(1, b.mean, sigmaB)
    solx = spanal(newt,y0,a,b)

    e = rnorm(n, 0, sigma)
    logy = log(solx)+e
    logy[1] = log(y0)
    newy = exp(logy)
    resy = logy-log(y0)
    zi = c(rep(0,n))
    ztotal = c(rep(0,n))

    for (p in c(2:n)) {
      zi[p] = 0.5 * (newt[p] - newt[p-1]) * (newy[p] + newy[p-1]) * exp(-
0.5*sigma^2)
      ztotal[p] = sum(zi[1:p])
    }
```

```
    each = cbind(id, t = newt, y.noIntercept = resy, y.intercept = logy, z =
ztotal,
                 ranA = rep(a, n), ranB = rep(b, n), y.init = rep(y0,n))
    outTable = data.frame(rbind(outTable, each))
  }
  outTable
}

outTable.MC = NULL
nlmeEsts.MC = NULL
lmeEsts.MC = NULL

for (sigma in merr) {
  for (j in 1:iter){
    outTable.MC = NULL
    outTable.MC = simuData(outTable.MC, sigma)

    lmeFit = NULL
    lmeFit = try(lme(y.noIntercept ~ t + z -1,
                     method = "ML",
                     random = ~ t + z -1| id,
                     control = lmeControl(opt='optim'),
                     data = outTable.MC))

    if(length(lmeFit) > 1) {
      lmeEsts.MC = data.frame(rbind(lmeEsts.MC, c(sigma,
round(lmeFit$coefficients$fixed, 6))))
    }

    nlmeFit = NULL
    f1 = y.intercept ~ log(spanal(t, y.init, ranA, ranB))
    nlmeFit = try(nlme(f1,
                       data = outTable.MC,
                       fixed = ranA + ranB ~ 1,
                       random = ranA + ranB ~ 1,
                       group = ~ id,
                       start = c(ranA = 0.8, ranB = 0.0015)))
    #control = nlmeControl(maxIter = 200,pnlsTol=0.01, pnlsMaxIter = 20)))
    if(length(nlmeFit) > 1) {
      nlmeEsts.MC = data.frame(rbind(nlmeEsts.MC, c(sigma,
round(nlmeFit$coefficients$fixed, 6))))
    }
  }
}

names(lmeEsts.MC) = c("sigma", "a", "b")
names(nlmeEsts.MC) = c("sigma", "a", "b")

res.MC.lme=NULL
for (sigma in unique(lmeEsts.MC$sigma) ) {
  each = lmeEsts.MC[lmeEsts.MC$sigma==sigma, ]
  res.MC.lme = rbind(res.MC.lme,c(sigma,round(mean(each$a),dig=5), #monte
carlo mean of a
                                  sd(each$a),  #monte carlo sd of a
                                  -round(mean(each$b),dig=6), #monte carlo
mean of b
                                  sd(each$b), #monte carlo sd of b
```

```
                                    (mean(each$a)-a.mean)/a.mean, # monte carlo
mean of relative bias from a
                                    (-mean(each$b)-b.mean)/b.mean)) # monte
carlo mean of relative bias from b
}
res.MC.lme = data.frame(res.MC.lme)
names(res.MC.lme) = c("sigma", "meanA", "sdA","meanB", "sdB", "biasA",
"biasB")
res.MC.lme

res.MC.nlme = NULL
for (sigma in unique(nlmeEsts.MC$sigma) ) {
  each = nlmeEsts.MC[nlmeEsts.MC$sigma==sigma, ]
  res.MC.nlme = rbind(res.MC.nlme,c(sigma,round(mean(each$a),dig=5),
                                    sd(each$a),
                                    round(mean(each$b),dig=6),
                                    sd(each$b),
                                    (mean(each$a)-a.mean)/a.mean,
                                    (mean(each$b)-b.mean)/b.mean))
}
res.MC.nlme = data.frame(res.MC.nlme)
names(res.MC.nlme) = c("sigma", "meanA", "sdA","meanB", "sdB", "biasA",
"biasB")
res.MC.nlme

write.csv(res.MC.nlme, file = "Table3_1000iter_201timepoints_0to20_NLME.csv")
write.csv(res.MC.lme, file =
"Table3_1000iter_201timepoints_0to20_BCLS_LME.csv")

BiasEst = cbind(res.MC.lme[,c(1,6,7)], res.MC.nlme[,6:7])
names(BiasEst) = c("sigma", "biasA.lme", "biasB.lme", "biasA.nlme",
"biasB.nlme")
BiasEst

## Compare relative bias from both methods
par(mfrow=c(1,2))
plot(BiasEst$sigma,BiasEst$biasA.lme, xlab="Measurement error",ylab="Relative
Bias a",type="b", ylim = c(-0.1, 0.1))
points(BiasEst$sigma,BiasEst$biasA.nlme,lty=2,type="b", col = "red")
legend("topright", cex=0.8,
       legend=c("BCLS-LME", "NLME"),
       lty=c(1, 2),
       col = c("black","red"))

plot(BiasEst$sigma,BiasEst$biasB.lme, xlab="Measurement error",ylab="Relative
Bias b",type="b", ylim = c(-0.2, 0.2))
points(BiasEst$sigma,BiasEst$biasB.nlme,lty=2,type="b", col = "red")
legend("topright", cex=0.8,
       legend=c("BCLS-LME", "NLME"),
       lty=c(1, 2),
       col = c("black","red"))
```

APPENDIX.3 R script to process OMC data set and fit BCLS-LME/NLME models for parameter estimation.

```
#####################################################################
################## BCLS-LME vs. NLME in OMC data set ##############
#####################################################################
library(nlme)
library(deSolve)
library(tidyr)
library(dplyr)
library(minpack.lm)
library(xtable)

# differential equation
dx<-function(t,x,parms) {
  list(c(x[1]*(parms[1]-parms[2]*x[1])))
}

# analytic solution
spanal <- function(t,x0,a,b) {
  k = x0/(a-b*x0)
  a*k*exp(a*t)/(1+b*k*exp(a*t))
}

# Set working directory
setwd("/Users/yalanxing/Desktop/Thesis project/Dataset")

################################ Pre-processing data sets
#####################################
grdat1 = read.csv(file="growthmeasures.csv", header=TRUE)
grdat2 = read.csv(file="grdat_5.10csv.csv", header=TRUE)
head(grdat1, 20)
head(grdat2, 20)

myvar1 = c("ID", "age","VisitDate","weight", "height", "Gender", "vml",
"Bloodsmear")
myvar2 =  c("id", "age","weight", "height", "vml", "malever", "nmal")
shortdata1 = grdat1[myvar1]
names(shortdata1) = c("id", "age","VisitDate","weight", "height", "gender",
"vml", "bloodsmear")

shortdata2 = grdat2[myvar2]
shortdataAll = merge(shortdata1, shortdata2, by = c("id", "age",
"weight","height", "vml"))

# Create new variables subtracting the initial non-empty height or weight
newdata = shortdataAll %>%
  group_by(id) %>%
  mutate("H0" = height[which.min(is.na(height))]) %>%   # Initial height for
each id
  mutate("W0" = weight[which.min(is.na(weight))]) %>%   # Initial weight for
each id
  mutate("absWeight" = weight - W0) %>%   # Absolute weight by subtracting
initial weight
  mutate("absLogWeight" = log(weight) - log(W0))

newweight = newdata %>%
```

```
  drop_na(weight)                      # Drop entries with NA in weight
newweight = data.frame(newweight)
newweight$ageWeek = newweight$age / 7.0   # age in weeks (continuous)
newweight$sigma = sd(newweight$W0)        # Measurement error calculated as
SD of all initial weight values
weight0 = mean(newweight$W0)              #initial value of weight calculated
as mean of all initial weight
newweight$Gender[newweight$gender == "Male"] <- 1
newweight$Gender[newweight$gender =="Female"] <- 0

# Remove id who had only one valid entry for weight data
length(unique(newweight$id))
for (id in unique(newweight$id)) {
  if (nrow(newweight[newweight$id == id,]) < 2) {
    newweight = newweight[!(newweight$id == id), ]
  }
}
length(unique(newweight$id))

# Calculate z
Z = NULL
for (p in unique(newweight$id)) {
  subset = newweight[newweight$id == p,]
  n = nrow(subset)
  zi = c(rep(0,n))
  ztotal = c(rep(0,n))

  for (i in c(2:n)){
    zi[i] = 0.5 * (subset$ageWeek[i] - subset$ageWeek[i-1]) *
(subset$weight[i] + subset$weight[i-1]) * exp(-0.5*subset$sigma[i]^2)
    ztotal[i] = sum(zi[1:i])
  }
  Z = append(Z, ztotal)
}

newweight$z = Z
newweight$resWeight = log(newweight$weight)-log(newweight$W0)
newweight$logWeight = log(newweight$weight)

write.csv(newweight,file = "SBRI_weight_5.15.csv")

################################### Analyasis
#########################################
shortdataAll = read.csv(file="SBRI_weight_5.15.csv", header=TRUE)

# Plot weight curve for ID < 50 = 33id
plot(shortdataAll[shortdataAll$id<50,]$ageWeek,
shortdataAll[shortdataAll$id<50,]$logWeight, xlab = "Age (Weeks)", ylab =
"log(Weight)")

# Fit NLME to full data set
nlmeEsts2 = NULL
nlme.f1 = logWeight ~ log(spanal(ageWeek, W0, a, b))
nlmeFit.weight2 = try(nlme(nlme.f1,
                           data = shortdataAll,
                           fixed = a + b ~ 1,
                           random = a + b ~ 1,
                           group = ~ id,
```

```
                          start = c(a = 0.08, b = 0.008),
                          control = nlmeControl(maxIter = 200,pnlsTol=0.01,
pnlsMaxIter = 20),
                          verbose = TRUE),
                      silent = FALSE)

if (length(nlmeFit.weight2) > 1) {
  nlmeEsts2 = signif(nlmeFit.weight2$coefficients$fixed, 5)
}
nlmeFit.weight2
nlmeEsts2

# Fit BCLS-LME to full data set
lmeFit2 = NULL
lmeEsts2 = NULL
lmeFit2 = try(lme(resWeight ~ ageWeek +  z -1,           # -1 to force no
intercept in the model
                  method = "ML",
                  random = ~ ageWeek + z -1| id, # random effect from a and
b, -1 to force no intercept in the model
                  data = shortdataAll))
if (length(lmeFit2) > 1) {
  est = lmeFit2$coefficients$fixed
  lmeEsts2 = c(signif(est[1], 3), -signif(est[2],5))
}
lmeFit2
lmeEsts2

# Effect from covariates including: gender, blood, malever, nmal
# NLME + malever  -- failed to converge
nlmeEsts.malever = NULL
nlmeFit.weight.malever = NULL

nlmeFit.weight.malever = try(nlme(nlme.f1,
                                  data = shortdataAll,
                                  fixed = list(a ~ malever,b~1),
                                  random = a + b ~ 1,
                                  group = ~ id,
                                  start = c(W0 = weight0, a = 0.08, b =
0.008),
                                  control = nlmeControl(maxIter =
200,pnlsTol=0.01, pnlsMaxIter = 20)),
                             silent = FALSE)
if (length(nlmeFit.weight.malever) > 1) {
  nlmeEsts.malever = rbind(nlmeEsts.malever,
signif(nlmeFit.weight.malever$coefficients$fixed, 5))
}
nlmeFit.weight.malever
nlmeEsts.malever

# BCLS-LME + malever -- converged
lmeFit.malever = NULL
lmeEsts.malever = NULL
lmeFit.malever = try(lme(resWeight ~ ageWeek + malEverTime + z -1,
# -1 to force no intercept in the model
                         method = "ML",
                         random = ~ ageWeek + z -1| id, # random effect from
a and b, -1 to force no intercept in the model
```

```
                        control = ctrl,
                        data = shortdataAll))

if (length(lmeFit.malever) > 1) {
  est.arb = lmeFit.malever$coefficients$fixed
  lmeEsts.malever = c(signif(est.arb[1], 5), signif(est.arb[2],5), -
signif(est.arb[3],5))
}
lmeFit.malever
lmeEsts.malever

##################### Truncated data for 1st year
################################
equailDays = 365
weight.1y = subset(shortdataAll, age <= equailDays )
summary(weight.1y$age)

## Effect from gender in 1st year data
# NLME.365 + gender -- failed to converge
nlmeEsts.gender.1y= NULL
nlme.f2 = logWeight ~ log(spanal(ageWeek, W0, a, b))
nlmeFit.gender.1y = try(nlme(nlme.f2,
                             data = weight.1y,
                             fixed = list(a ~ gender,b~1),
                             random = a + b ~ 1,
                             group = ~ id,
                             start = c(W0 = weight0, a = 0.08, b = 0.008),
                             control = nlmeControl(maxIter =
200,pnlsTol=0.01, pnlsMaxIter = 20)),
                        silent = FALSE)
if (length(nlmeFit.gender.1y) > 1) {
  nlmeEsts.gender.1y = signif(nlmeFit.gender.1y$coefficients$fixed, 5)
}
nlmeFit.gender.1y
nlmeEsts.gender.1y

# BCLS-LME.365 + gender --converged
lmeFit.gender.1y = NULL
lmeEsts.gender.1y = NULL
lmeFit.gender.1y = try(lme(resWeight ~ ageWeek + gentime + z -1,           #
-1 to force no intercept in the model
                           method = "ML",
                           random = ~ ageWeek + z -1| id, # random effect
from a and b, -1 to force no intercept in the model
                           control = ctrl,
                           data = weight.1y))

if (length(lmeFit.gender.1y) > 1) {
  est.arb = lmeFit.gender.1y$coefficients$fixed
  lmeEsts.gender.1y = c(signif(est.arb[1], 5), signif(est.arb[2],5), -
signif(est.arb[3],5))
}
lmeFit.gender.1y
lmeEsts.gender.1y
```