Reduced-bias estimation and inference for mixed-effects models

Sophia Kyriakou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

at

University College London.

Department of Statistical Science

University College London

In loving memory of
thios Athinagoras

Contents

Acknowledgements

First and foremost, I would like to thank my supervisor Dr Ioannis Kosmidis for all the advice and guidance he gave me during my PhD studies and for helping develop myself as a researcher. I especially thank him for carefully reading this thesis, providing valuable feedback that improved the presentation of my work.

Appreciation is expressed to the department of Statistical Science of the University College of London for awarding me a studentship for postgraduate work in the department for 3.5 years. I also thank the Erasmus+ programme of the European Union which funded me for a 2-month traineeship at the University of Padova, Italy.

I am thankful to Professor Nicola Sartori who very kindly offered his knowledge and help whenever it was needed during and after my time in Italy. It has been a pleasure working with my co-authors Dr Ioannis Kosmidis and Professor Nicola Sartori who provided insight and expertise that greatly assisted the research on "Median bias reduction in random-effects meta-analysis and meta-regression" (to be appeared in Statistical Methods for Medical Research).

I would like to thank (in random order) Ioannis Kosmidis this time not for his excellent supervisory skills but for his many efforts with me to keep me motivated and positive throughout these years; Tom Fearn and Christian Hennig for the useful feedback during my upgrade examination; Yannis Kasparis for the constructive comments on parts of my work and for inviting me to present my work at the University of Cyprus; Panagiotis for the understanding and for all the moments together; Stelianos, Antigoni, and Katerina for their youthful spirit and for reminding me that noone can crush mine; Matina, Sofia-Maria, Nikos, and Menelaos for the hours of fun and joy; Mariangela, Carlo, and Caterina for the hospitality and the beautiful moments in Padova; Mariam for making the house in London feel like home; Simon for being caring and patient

during the writing up of my thesis; Antonis, Elpidios, Odysseas, Simoni, and Sophia for being constant pillars of strength and support throughout the years.

A heartfelt thanks goes to my family for their constant support and non-judgemental ears. Christina, you have supported me no matter what and you are the person I can always turn to. Dad, my voice of reason, you might not know much about statistics but you are always there for me whenever I need to talk about my work. Last, but most certainly not least, mum, without your love and encouragement I would not have made it this far. The three of you have been the rocks in my life.

This thesis is dedicated to my beloved uncle Athinagoras who has lost the battle with cancer in April 2016.

Declaration

I, Sophia Kyriakou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

A popular method for reducing the mean and median bias of the maximum likelihood estimator in regular parametric models is through the additive adjustment of the score equation (Firth, 1993; Kenne Pagui et al., 2017). The current work focuses on mean and median bias-reducing adjusted score equations in models with latent variables. First, we give estimating equations based on a mean bias-reducing adjustment of the score function for mean bias reduction in linear mixed models. Second, we propose an extension of the adjusted score equation approach (Firth, 1993) to obtain bias-reduced estimates for models with either computationally infeasible adjusted score equations and/or intractable likelihood. The proposed bias-reduced estimator is obtained by solving an approximate adjusted score equation, which uses an approximation of the log-likelihood to obtain tractable derivatives, and Monte Carlo approximation of the bias function to get feasible expressions. Under certain general conditions, we prove that the feasible and tractable bias-reduced estimator is consistent and asymptotically normally distributed. The "iterated bootstrap with likelihood adjustment" algorithm is presented that can compute the solution of the new bias-reducing adjusted score equation. The effectiveness of the proposed method is demonstrated via simulation studies and real data examples in the case of generalised linear models and generalised linear mixed models. Finally, we derive the median bias-reducing adjusted scores for linear mixed models and random-effects meta-analysis and meta-regression models.

Impact Statement

The current thesis explores solutions to the important problem of reducing the bias in the estimation of mixed models. This problem is a common concern of practitioners and statisticians, because the magnitude of bias can affect the performance of standard procedures for hypothesis testing and construction of confidence intervals. For instance, the underestimation of standard errors may lead to shorter than expected confidence intervals, which in turn result in spuriously strong conclusions.

More specifically, we extend existing work for mean and median bias-reduction using adjusted score equations in linear mixed models and random-effects meta-analysis and meta-regression. We also develop a new bias-reducing methodology for models that have intractable likelihoods and infeasible bias functions. This is a major development because this class of models includes many complex and widely used models, such as generalised linear mixed models.

Generalised linear mixed models are broadly used in various fields for modeling dependence within clustered data. For instance, in medical science mixed models become fruitful in analysing data from longitudinal studies that compare a new drug with a standard one for treating patients suffering from an illness. In social sciences mixed models can be used for the estimation of county-specific characteristics, such as the unemployment rate. Geneticists and evolutionary biologists use mixed modeling when they are interested in quantifying the magnitude of variation among genotypes.

Generalised linear mixed models are generally challenging to fit and standard estimation methods tend to underestimate the variance components. Use of the adjusted score equations in mixed modeling yields variance component estimates with smaller bias, which in turn improves inference on the fixed effects. Much work remains to reveal the full power of the bias-reducing adjusted score equations approach to these modern statistical models, but we strongly believe that it will offer practitioners a formal and flexible statistical framework for bias reduction, that will make an impact in many application areas where bias reduction is beneficial.

Abbreviations

ML	maximum likelihood
REML	restricted maximum likelihood
IBLA	iterated bootstrap with likelihood adjustment
BOOT	parametric bootstrap
mean BR	mean bias reduction
median BR	median bias reduction
DL	DerSimonian & Laird (1986) method
KR	Kenward & Roger (1997) statistic
LR	likelihood ratio
mean BRPL ratio	mean bias reducing penalised likelihood ratio
median BRPL ratio	median bias reducing penalised likelihood ratio
MSE	mean squared error
PU	percentage of underestimation
La-ML	Laplace-based maximum likelihood
PQL	penalised quasi-likelihood
CPQL	corrected penalised quasi-likelihood
La-BOOT	Laplace-based parametric bootstrap
La-IBLA	Laplace-based iterated bootstrap with likelihood adjustment

Notation

$\Re$	the set of real numbers
$\Re^p$	the p-dimensional Euclidean space
$\xrightarrow{p}$	converges in probability
$\xrightarrow{d}$	converges in distribution
$\|\|x\|\|$	the norm of x in the domain of x
$E(X)$, $\mathrm{Var}(X)$, $\mathrm{Cov}(X)$	expected value, variance, covariance of X
$A \circ B$	Hadamard (elementwise) product of matrices A and B
A^{T}	the transpose of a matrix A
A^{-1}	the inverse of a square matrix A
$\|A\|$	the determinant of a square matrix A
$\mathrm{tr}(A)$	the trace of a square matrix A
$\dim(A)$	the dimension of a matrix A
0_p	a $p \times 1$ vector of zeros
$0_{n \times p}$	a $n \times p$ matrix of zeros
I_n	the $n \times n$ identity matrix
$\nabla f(x)$, $\nabla\nabla^{\mathrm{T}} f(x)$	the gradient, the Hessian of a function f with respect to x

$\hat{\theta}$	the ML estimator, root of the score function $s(\theta)$
$\hat{\theta}^{\ddagger}$	the REML estimator, root of the REML score function $s^{\ddagger}(\theta)$
$\hat{\theta}^{*}$	the mean BR estimator, root of the mean BR adjusted score function $s^{*}(\theta)$
$\hat{\theta}^{\dagger}$	the median BR estimator, root of the median BR adjusted score function $s^{\dagger}(\theta)$
$\hat{\theta}^{*}_{n,R}$	the IBLA estimator, root of the simulation-based adjusted score function $s^{*}_{n,R}(\theta)$
$\tilde{\theta}$	the maximum approximate likelihood estimator, root of the approximate score function $\tilde{s}(\theta)$
$\tilde{\theta}^{*}_{n,R}$	the IBLA estimator, root of the simulation-based approximate adjusted score function $\tilde{s}^{*}_{n,R}(\theta)$

$B_n(\theta)$	the bias function of $\hat{\theta}$, $B_n(\theta) = E_{\theta}(\hat{\theta} - \theta)$
$\hat{B}_{n,R}(\theta)$	the simulation-based estimate of $B_n(\theta)$
$\tilde{B}_n(\theta)$	the bias function of $\tilde{\theta}$, $\tilde{B}_n(\theta) = E_{\theta}(\tilde{\theta} - \theta)$
$\tilde{B}_{n,R}(\theta)$	the simulation-based estimate of $\tilde{B}_n(\theta)$

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Bias in estimation

The bias of an estimator $\hat{\theta}$ of a model parameter θ is the difference between the expected value of $\hat{\theta}$ with respect to the model and θ. An estimator whose bias is equal to zero is called unbiased and satisfies $E_\theta(\hat{\theta}) = \theta$, for all θ.

The current work focuses on the bias of maximum likelihood (ML) estimators. The essence of ML estimation is to view the likelihood as a function of the parameter θ, and to derive the ML estimate as the value of θ that maximises the likelihood of the observed data within the parameter space. The ML estimate is formally defined as $\hat{\theta}_n = \arg\max_\theta l(\theta; y)$, where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are the observations of n independent random variables with density functions or probability distributions $f_i(y;\theta)$ and $l(\theta;y) = \log \prod_{i=1}^n f_i(y_i;\theta)$ is the log-likelihood function.

Maximum likelihood is a widely used method of estimation in regular parametric models, and its popularity is partly because, under standard regularity conditions, the ML estimator has asymptotically desirable behaviour (Cox & Hinkley, 1979, Section 9.1). It can be shown, for example, that the ML estimator is consistent, asymptotically normally distributed, and asymptotically unbiased.

However, the finite-sample bias of the ML estimator is a common concern for statisticians, because the magnitude of bias can affect the performance of standard procedures for hypothesis testing and construction of confidence intervals. For instance, the underestimation of standard errors may lead to shorter than expected confidence intervals, which in turn result in spuriously strong conclusions.

The poor coverage properties that confidence intervals can have due to bias in the ML estimator are illustrated via the following motivating example.

Consider the one-way random effects model (Jiang, 2007, Example 1.1) with observations y_{ij}, $(i = 1,\dots,m$ and $j = 1,\dots,k_i)$ with $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. The parameter μ is the only fixed effect and is an unknown mean, $\alpha_1,\dots,\alpha_m$ are random effects that are independent and normally distributed with mean zero and unknown variance σ^2_α, $\varepsilon_{11},\dots,\varepsilon_{1k_1},\dots,\varepsilon_{m1},\dots,\varepsilon_{mk_m}$ are independent and normally distributed errors with mean zero and unknown variance σ^2_ε, and the random effects are independent of the errors. The log-likelihood function of the one-way random effects model is

$$\begin{aligned} l(\mu,\sigma^2_\alpha,\sigma^2_\varepsilon) &= c - \frac{1}{2}(n-m)\log(\sigma^2_\varepsilon) - \frac{1}{2}\sum_{i=1}^{m}\log(\sigma^2_\varepsilon + k_i\sigma^2_\alpha) \\ &\quad - \frac{1}{2\sigma^2_\varepsilon}\sum_{i=1}^{m}\sum_{j=1}^{k_i}(y_{ij}-\mu)^2 + \frac{\sigma^2_\alpha}{2\sigma^2_\varepsilon}\sum_{i=1}^{m}\frac{k_i^2}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}(\bar{y}_{i\cdot}-\mu)^2 , \end{aligned}$$

where c is a constant, $n = \sum_{i=1}^{m} k_i$, and $\bar{y}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} y_{ij}$. The ML estimator of μ is the solution to

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{m}\frac{k_i}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}(\bar{y}_{i\cdot}-\mu) = 0 ,$$

and it is equal to

$$\hat{\mu} = \sum_{i=1}^{m}\frac{\frac{k_i}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}}{\sum_{i=1}^{m}\frac{k_i}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}}\bar{y}_{i\cdot} .$$

It is easy to show that $\hat{\mu}$ is an unbiased estimator of μ, because

$$E(\hat{\mu}) = \sum_{i=1}^{m}\frac{\frac{k_i}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}}{\sum_{i=1}^{m}\frac{k_i}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}}E(\bar{y}_{i\cdot}) = \sum_{i=1}^{m}\frac{\frac{k_i}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}}{\sum_{i=1}^{m}\frac{k_i}{\sigma^2_\varepsilon + k_i\sigma^2_\alpha}}\mu = \mu .$$

In the current example we simulated 10 000 independent datasets from the one-way random effects model with parameter values $\theta = (\mu,\sigma^2_\alpha,\sigma^2_\varepsilon)^{\mathrm{T}} = (0,0.5,0.5)^{\mathrm{T}}$. We chose three values of m, specifically 5, 10, 15, and we set $k_i = 5$ for all $i \in \{1,\dots,m\}$. Table 1.1 gives the empirical bias of the ML estimates of θ and the empirical p-value distribution for the Wald statistic under the null hypothesis $\mu = 0$. For a given parameter θ and corresponding estimates $\hat{\theta} = (\hat{\theta}_1,\dots,\hat{\theta}_S)^{\mathrm{T}}$ from S independent simulated

Table 1.1: Empirical mean bias of the ML estimates for the parameters $(\mu, \sigma_\alpha^2, \sigma_\varepsilon^2)^\top$ of the one-way random effects model, and empirical p-value distribution (%) for the two-sided Wald test that $\mu = 0$.

		Bias			Empirical p-value distribution (%) for the Wald test										
m	$\hat{\mu}$	$\hat{\sigma}_\alpha^2$	$\hat{\sigma}_\varepsilon^2$	$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
5	-0.001	-0.125	-0.007		7.3	10.6	14.3	20.5	35.1	57.7	79.1	91.8	96.0	98.1	99.4
10	0.006	-0.069	-0.003		3.6	6.1	9.0	15.4	30.6	54.3	77.8	91.0	95.4	98.0	99.2
15	0.004	-0.042	0.001		2.1	4.2	7.5	13.0	28.4	52.5	76.3	90.3	95.2	97.6	99.1

Notes: $\hat{\mu}$, ML estimator of μ; $\hat{\sigma}_\alpha^2$, ML estimator of σ_α^2; $\hat{\sigma}_\varepsilon^2$, ML estimator of σ_ε^2. The empirical p-value distribution (%) represents the coverage probability of $(1-\alpha)\%$ confidence intervals based on the Wald statistic.

samples, the empirical bias is defined as $\text{Bias}(\theta) = (1/S)\sum_{s=1}^{S}(\hat{\theta}_s - \theta)$. The empirical p-value distribution represents the coverage probability of $(1-\alpha)\%$ confidence intervals based on the Wald statistic, where the term coverage probability refers to the estimated probability that the Wald-type confidence intervals at a specific nominal level $(1-\alpha)\%$ include the true parameter value. The results in Table 1.1 illustrate that the unbiasedness of $\hat{\mu}$ does not guarantee good coverage properties for confidence intervals for μ, because ML underestimates the variance of the random effects σ_α^2. As the sample size increases the bias of $\hat{\sigma}_\alpha^2$ decreases, and the empirical p-value distribution for the Wald statistic gets closer to uniformity.

1.2 Methods for bias reduction

Numerous methods have been proposed to correct the finite-sample bias of the ML estimator. Kosmidis (2014a) classifies the bias reduction methods into two large groups, the explicit and implicit methods.

Let $B_n(\theta) = E_\theta(\hat{\theta}_n - \theta)$ be the bias function of the ML estimator $\hat{\theta}_n$. Explicit methods estimate $B_n(\theta)$ by an estimator $\hat{B}_n$, and compute a bias-reduced estimator as $\hat{\theta}_n - \hat{B}_n$. The most popular explicit bias-reduction methods are jackknife (Quenouille, 1956), asymptotic bias corrections (Efron, 1975), and bootstrap (Efron, 1979). The main advantage of the explicit methods is that they are simple to implement; once we estimate the bias, we obtain the bias-reduced estimator by simply doing a subtraction. However, explicit methods inherit any instabilities the ML estimator may have. These instabilities involve infinite ML estimates which occur with positive probability when fitting models with categorical responses (Albert & Anderson, 1984), or more generally ML estimates at the boundary of the parameter space.

Implicit methods replace $B_n(\theta)$ with an estimate of the whole bias function $\hat{B}_n(\theta)$, and the bias-reduced estimator is derived by solving the implicit equation $\hat{\theta}_n - \theta = \hat{B}_n(\theta)$ with respect to θ. The most popular implicit bias-reduction methods are indirect inference (Gourieroux et al., 1993) and adjusted score equation (Firth, 1993). The main disadvantage of implicit methods is that the solution of the implicit equation often requires numerical optimisation. Also, similar to explicit methods, indirect inference depends on the ML estimator and therefore inherits any potential instabilities that it might have. The adjusted score equation approach does not depend on the ML estimator, but it is applicable only when the score function, the expected information matrix, and the first-order bias term of the ML estimator are available in closed form.

The methods proposed in the current work are extensions of the adjusted score equation approach, and they allow the adjustment of the score function in cases where the direct use of the vanilla method is computationally infeasible or intractable.

1.3 Adjusted score equation for bias reduction

Rather than correcting the ML estimator itself, Firth (1993) systematically corrects the mechanism that produces the estimator. Specifically, Firth (1993) showed that a bias-reduced estimate $\hat{\theta}^*$ is obtained by solving an adjusted score equation of the form

$$s^*(\theta) = s(\theta) + A(\theta) = 0_p\,, \tag{1.1}$$

where $s(\theta) = \nabla_\theta l(\theta)$ is the score function and $A(\theta)$ is the bias-reducing adjustment of order $O_p(1)$ as $n \to \infty$. Firth (1993) gives two suitable candidates for $A(\theta)$, both of which can be used for removal of the $O(n^{-1})$ bias of the ML estimator. These are

$$A^{(O)}(\theta) = j(\theta)\{i(\theta)\}^{-1}A^{(E)}(\theta)\,, \tag{1.2}$$

and $A^{(E)}(\theta)$ with components

$$A_t^{(E)}(\theta) = \frac{1}{2}\operatorname{tr}[\{i(\theta)\}^{-1}\{P_t(\theta) + Q_t(\theta)\}]\,, \tag{1.3}$$

where $P_t(\theta) = E_\theta\{s(\theta)s^{\mathrm{T}}(\theta)s_t(\theta)\}$, $Q_t(\theta) = -E_\theta\{j(\theta)s_t(\theta)\}$, $i(\theta) = E_\theta(j(\theta))$ is the expected information matrix, $j(\theta) = -\nabla_\theta \nabla_\theta^{\mathrm{T}} l(\theta)$ is the observed information matrix, and $s_t(\theta)$ is the tth component of the score vector $s(\theta)$.

Kosmidis & Firth (2009) give a general family of candidates for a bias-reducing choice of $A(\theta)$. The general adjustment is

$$A(\theta) = -\{G(\theta) + R(\theta)\}b(\theta), \tag{1.4}$$

where $G(\theta)$ is either $j(\theta)$ or $i(\theta)$ or some other matrix with expectation $i(\theta)$, $R(\theta)$ is any matrix with expectation of order $O(n^{1/2})$, and $b(\theta) = n^{-1}b_1(\theta) = -\{i(\theta)\}^{-1}A^{(E)}(\theta)$ is the first-order bias term in the expansion of the asymptotic bias of the ML estimator $B(\theta) = n^{-1}b_1(\theta) + n^{-2}b_2(\theta) + \ldots$, where the functions $b_1(\theta), b_2(\theta), \ldots$, are of order $O(1)$ (see, for example, McCullagh, 1987). For example, if we let $G(\theta) = i(\theta)$ with $R(\theta) = 0$ then

$$s^*(\theta) = s(\theta) + A^{(E)}(\theta) = s(\theta) - i(\theta)b(\theta), \tag{1.5}$$

and if we let $G(\theta) = j(\theta)$ with $R(\theta) = 0$ then

$$s^*(\theta) = s(\theta) + A^{(O)}(\theta) = s(\theta) - j(\theta)b(\theta). \tag{1.6}$$

Any suitable bias-reducing adjustment $A(\theta)$ gives the same asymptotic results and removes the $O(n^{-1})$ bias of the ML estimator (Kosmidis & Firth, 2009). For this reason, the choice of $A(\theta)$ in the following chapters is based solely on how easy the derivation of the quantities involved in the adjustment is.

The current work is intended to extend the adjusted score equation method in two ways. First, we show how an adjustment of the score function can be used for obtaining bias-reduced estimators in models with tractable likelihood and infeasible bias function. We define a function as infeasible when it is not possible to calculate it easily. Second, we propose an adjusted score equation which can be used as a bias reduction method in models with intractable likelihood. We define a likelihood as intractable when it cannot be evaluated analytically, and the integrals involved in the function require approximations. The intractability of the likelihood in such models prevents the direct use of the approach in Firth (1993), because all quantities involved cannot generally be written in closed form.

1.4 Thesis outline

A considerable part of the present work is devoted to mixed-effects models and how bias in their estimation can be reduced. Mixed modelling has become a major area of statistical research during the last decades because it provides a flexible approach to clustered data. The parameters in a mixed-effects model are classified into fixed effects and variance components. Fixed effects are associated with the average effect of predictors on the response, and variance components are associated with the variance-covariance structure of the random effects. In this thesis we will restrict ourselves to models in which the random effects are normally distributed.

In Chapter 2 we study mean bias reduction for linear mixed models (Longford, 1993). Linear mixed models are models in which both the fixed and the random effects contribute linearly to the response. The two most popular estimation methods in linear mixed models are ML and restricted or residual maximum likelihood (REML). REML improves ML estimation by effectively adjusting for degrees of freedom lost in estimation, delivering estimators with less bias. In this chapter we derive the mean bias-reducing adjusted score equation and link the estimating equation with REML score equation. Simulation studies and real data applications are used to assess the performance of estimation and inference based on the mean bias-reducing adjusted score equation and compare it to ML and REML under various parameterisations. Chapter 2 also includes a special case of linear mixed models, random effects meta-analysis and meta-regression models. These models are used for synthesising the results from independent studies investigating a common effect of interest. Kosmidis et al. (2017) derived the adjusted score equation for mean bias reduction in random effects meta-analysis and meta-regression models and proposed a likelihood-based test for conducting inference. In this chapter, we complement Kosmidis et al. (2017) with new results on computational efficiency, estimation, and inference.

In Chapter 3 we consider variations of the mean bias-reducing adjusted score equation for models with tractable likelihood. Firth (1993) uses the bias function to obtain the adjusted score equation. We show that solving an adjusted score equation where the bias function is replaced by its simulation-based estimate, also leads to estimators with $o(n^{-1})$ bias, and the estimators are consistent and asymptotically normally distributed. The "iterated bootstrap with likelihood adjustment" algorithm (IBLA) is presented that can compute the solution of the new bias-reducing adjusted score equation. The simulation-based adjusted score equation approach is applied and evaluated in generalised linear models (McCullagh & Nelder, 1989). These models extend standard linear regression models to encompass non-normally distributed data and possibly nonlinear functions of the mean. Implementing the proposed simulation-based bias reduction method on generalised linear models allows the evaluation of its performance on estimation and inference compared to the traditional adjusted score equation approach (Firth, 1993).

Chapter 4 extends the use of the simulation-based adjusted score equation approach derived in Chapter 3 as a mean bias reduction method in the case of models with intractable likelihood. We give conditions under which an approximation of the likelihood function may be used in order to derive mean bias-reduced estimates. In this chapter we also prove the asymptotic properties of the proposed estimators and modify the IBLA algorithm such that it can be used for the calculation of the mean bias-reduced estimates.

Chapter 5 evaluates the performance of the mean bias-reduction method proposed in Chapter 4 on generalised linear mixed models (McCulloch et al., 2008). These models are an extension of linear mixed models that can handle non-normally distributed clustered data. They can also be seen as an extension of generalised linear models that include random effects in addition to the usual fixed effects. We evaluate the performance of IBLA against the most popular existing methods used in fitting generalised linear mixed models.

In Chapter 6 we deviate from mean bias reduction and turn to median bias reduction, while staying in the familiar framework proposed in Firth (1993). Kenne Pagui et al. (2017) consider the median as a centering index for the score, and an adjusted score function for median bias reduction then results by subtracting from the score its approximate median. In this chapter we first derive the adjusted score equation for median bias reduction in linear mixed models, and compare it to the relative equation for mean bias reduction derived in Chapter 2. Second, we derive the adjusted score equation based on a median adjustment of the score function for median unbiased estimation for random effects meta-analysis and meta-regression models, and compare it to the relative equation for mean bias reduction proposed in Kosmidis et al. (2017).

A summary of the main results is given in Chapter 7, where we also indicate some related open topics for further work in the area.

Computing environment and typeset

For the computational requirements of the thesis, the R language (R Core Team, 2017) was used and all the figures were created using the `ggplot2` R package (Wickham, 2009). The simulation results were computed using a workstation with 24 cores at 2.90GHz and 80GB memory running under the CentOS 7 operating system, using one core per data set.

Chapter 2

Mean bias reduction in linear mixed models

2.1 Introduction

Linear mixed models are widely used for analysing clustered data, that is data in which the observations are grouped into disjoint classes (clusters) according to some classification criterion (Longford, 1993). Mixed models are also suitable for analysing longitudinal data collected from studies designed to investigate changes over time about a characteristic which is measured repeatedly for each individual (Laird & Ware, 1982), as well as repeated measurements from experimental designs where several individuals participate and multiple measurements are taken on each individual (Lindstrom & Bates, 1988).

Typically the linear mixed model parameters, which consist of the fixed effects and the variance components, are estimated by ML. The ML estimators of the fixed effects are unbiased, but the ML estimators of the variance components are negatively biased, because they do not take into account the loss in degrees of freedom resulting from the estimation of the fixed effects (see, for example, Harville, 1977; Kackar & Harville, 1984; Lindstrom & Bates, 1988).

To reduce the bias in the variance component estimators Patterson & Thompson (1971) and Harville (1974) suggest modifying the log-likelihood using generalised least squares residuals. This modification leads to a likelihood-based function which differs from the log-likelihood (ignoring any quantities that are constant at the parameters) by an extra additive term, and the estimation method is referred to as the residual or restricted maximum likelihood (REML). Harville (1977) applied REML to the linear mixed model as given in Section 2.2. REML takes into account the degrees of freedom for the fixed effects in the linear mixed model, and consequently the estimation of the variance components is unbiased. However, standard inferential procedures, such as the Wald test, tend to be anti-conservative when using the REML estimates and the sample size is small (Gumedze & Dunne, 2011).

A prominent special case of linear mixed models is random effects meta-analysis and meta-regression, a core tool for synthesising the results from independent studies investigating a common effect of interest. Introduced in DerSimonian & Laird (1986), the model expresses the heterogeneity between studies in terms of a variance component that can be estimated through standard estimation techniques. Contrary to linear mixed models, the error variances in the random effects meta-analysis and meta-regression are assumed to be known, and hence are not being estimated.

The random effects meta-analysis and meta-regression is an interesting special case of linear mixed models, because frequentist inference is not performing well, especially when the number of studies is small or moderate. Specifically, the estimation of the heterogeneity parameter can be highly imprecise, which in turn results in misleading conclusions (Guolo & Varin, 2017; Kosmidis et al., 2017). Examples of recently proposed methods that attempt to improve inference are the resampling (Jackson & Bowden, 2009) and double resampling (Zeng & Lin, 2015) approaches, and the mean bias-reducing penalised likelihood (mean BRPL) approach (Kosmidis et al., 2017). Specifically, Kosmidis et al. (2017) show that maximisation of the mean BRPL results in an estimator of the heterogeneity parameter that has notably smaller bias than ML with small loss in efficiency, and illustrate that inference based on the mean BRPL outperforms its competitors in terms of inferential performance.

In this chapter we will restrict ourselves to the framework of linear mixed models in which the errors and the random effects are normally distributed. First, we use the adjusted score equation approach (Firth, 1993) to derive the mean bias-reducing adjusted score equation for linear mixed models. The derived bias-reducing score equations can be used for model estimation under any parameterisation of the variance-covariance matrix of the random effects. We show that the mean bias-reduced (mean BR) estimates of the variance components are identical to the REML estimates under certain parameterisations. The performance of the mean bias-reducing adjusted score equations is investigated through simulation studies and a real-data example from Potthoff & Roy (1964). Our results illustrate that confidence intervals based on the Wald statistic and the mean BRPL estimates outperform confidence intervals based on the ordinary Wald or likelihood ratio statistics, in terms of coverage. Next, we focus on random effects meta-analysis and meta-regression and we use simulation studies and real data applications to complement Kosmidis et al. (2017) with new results on the computational efficiency of mean BRPL estimation and distribution of p-values.

2.2 Linear mixed model

In a linear mixed model the observations $y_1,\ldots,y_n$ are assumed to be realisations of the random variables $Y_1,\ldots,Y_n$, respectively, and $Y_1,\ldots,Y_n$ are independent conditionally on random effects. The model can be expressed in matrix form as

$$Y = X\beta + Z\alpha + \varepsilon\,, \tag{2.1}$$

where $Y = (Y_1,\ldots,Y_n)^{\mathrm{T}}$, X is the $n\times p$ matrix of known covariates, β is a p-dimensional vector of the fixed effects, Z is a $n\times q$ known design matrix, α is the q-dimensional vector of random effects, and ε is the vector of errors. Both α and ε are unobservable. Typically, we assume the random effects and errors to be independent and normally distributed with

$$\begin{pmatrix}\alpha\\ \varepsilon\end{pmatrix} \sim N\left(\begin{pmatrix}0_q\\ 0_n\end{pmatrix}, \begin{pmatrix}\Sigma(\sigma^2) & 0_{q\times n}\\ 0_{n\times q} & \sigma_\varepsilon^2 I_n\end{pmatrix}\right), \tag{2.2}$$

where σ^2 represents the vector of all the unknown dispersion parameters, σ_ε^2 is the error variance, 0_q denotes a q-dimensional vector of zeros, $0_{q\times n}$ denotes a $q\times n$ matrix of zeros, and I_n is the $n\times n$ identity matrix.

Let $\psi = (\sigma^{2^{\mathrm{T}}}, \sigma_\varepsilon^2)^{\mathrm{T}}$ be the m-dimensional vector of all unknown variance components. The marginal distribution of Y is multivariate normal with mean $X\beta$ and variance-covariance matrix $V(\psi) = Z\Sigma(\sigma^2)Z^{\mathrm{T}} + \sigma_\varepsilon^2 I_n$. The elements of $V(\psi)$ are assumed to be differentiable up to second order with respect to the elements of ψ.

2.3 Methods of estimation

The most common likelihood-based estimation methods adopted in linear mixed models are ML and REML (Longford, 1993). The log-likelihood function for the parameter $\theta = (\beta^{\mathrm{T}}, \psi^{\mathrm{T}})^{\mathrm{T}}$ in model (2.1) is up to an additive constant given by

$$l(\theta) = -\frac{1}{2}\left[\log|V(\psi)| + R(\beta)^{\mathrm{T}} V(\psi)^{-1} R(\beta)\right], \tag{2.3}$$

where $|V(\psi)|$ denotes the determinant of $V(\psi)$, and $R(\beta) = y - X\beta$, $y = (y_1, \ldots, y_n)^{\mathrm{T}}$. By differentiating $l(\theta)$ with respect to the model parameters we obtain the score function $s(\theta)$ with components $s_\beta(\theta) = X^{\mathrm{T}} V(\psi)^{-1} R(\beta)$ and

$$s_{\psi_r}(\theta) = \frac{1}{2}\left\{R(\beta)^{\mathrm{T}} V(\psi)^{-1} \frac{\partial V(\psi)}{\partial \psi_r} V(\psi)^{-1} R(\beta) - \mathrm{tr}\left(V(\psi)^{-1} \frac{\partial V(\psi)}{\partial \psi_r}\right)\right\},$$

where ψ_r is the rth component of ψ, $r \in \{1, \ldots, m\}$.

The expected information matrix $i(\theta)$ is a block-diagonal matrix with diagonal blocks $i_{\beta\beta} = E_\theta(j_{\beta\beta})$ and $i_{\psi\psi} = E_\theta(j_{\psi\psi})$, where $j_{\beta\beta}$ and $j_{\psi\psi}$ are the diagonal blocks of the observed information matrix $j(\theta) = \nabla\nabla^{\mathrm{T}} l(\theta)$, whose expression is given in Appendix A. Specifically, $i_{\beta\beta} = X^{\mathrm{T}} V(\psi)^{-1} X$ and $i_{\psi\psi}$ is a $m \times m$ matrix with (r,s)th element

$$i_{\psi_r\psi_s} = \frac{1}{2}\mathrm{tr}\left(V(\psi)^{-1} \frac{\partial V(\psi)}{\partial \psi_r} V(\psi)^{-1} \frac{\partial V(\psi)}{\partial \psi_s}\right).$$

The ML estimator of θ, $\hat{\theta} = (\hat{\beta}^{\mathrm{T}}, \hat{\psi}^{\mathrm{T}})^{\mathrm{T}}$, results from solving the equations $s_\beta(\theta) = 0_p$ and $s_{\psi_r}(\theta) = 0$ for all $r \in \{1, \ldots, m\}$. Once $\hat{\psi}$ is found, $\hat{\beta}$ can be calculated by the closed-form expression $\hat{\beta} = \{X^{\mathrm{T}} V(\hat{\psi})^{-1} X\}^{-1} X^{\mathrm{T}} V(\hat{\psi})^{-1} y$.

Following Jiang (2007, Chapter 1) we assume that rank$(X) = p$ and that G is an $n \times (n-p)$ matrix such that rank$(G) = n - p$ and $G^{\mathrm{T}} X = 0_{(n-p)\times p}$. We can find up to $n - p$ linearly independent combinations $G^{\mathrm{T}} y$ whose distribution does not depend on β, for example, any $n - p$ of the least-squares residuals of the regression of y on X. In REML we treat $G^{\mathrm{T}} y$ as the data and use ML estimation for the variance components.

The REML estimates of the variance components do not depend on the choice of G and maximise the restricted log-likelihood which can be written up to a constant as

$$l^{\ddagger}(\theta) = l(\theta) - \frac{1}{2}\log|X^{\mathrm{T}}V(\psi)^{-1}X|\,. \tag{2.4}$$

The derivative of $l^{\ddagger}(\theta)$ with respect to the rth variance component is

$$s^{\ddagger}_{\psi_r}(\theta) = s_{\psi_r}(\theta) + \frac{1}{2}\operatorname{tr}\left(V(\psi)^{-1}H(\psi)\frac{\partial V(\psi)}{\partial\psi_r}\right)\,, \tag{2.5}$$

with $H(\psi) = X(X^{\mathrm{T}}V(\psi)^{-1}X)^{-1}X^{\mathrm{T}}V(\psi)^{-1}$.

Let $\hat{\psi}^{\ddagger}$ be the REML estimator of ψ that solves $s^{\ddagger}_{\psi}(\theta) = 0_m$. The REML estimates for the fixed effects are usually obtained using the generalised least squares estimator $\hat{\beta}^{\ddagger} = \Phi(\hat{\psi}^{\ddagger})X^{\mathrm{T}}V(\hat{\psi}^{\ddagger})^{-1}y$, where $\hat{\beta}^{\ddagger}$ is an unbiased estimator of β (Kackar & Harville, 1984). The matrix $\Phi(\psi) = (X^{\mathrm{T}}V(\psi)^{-1}X)^{-1}$ is the variance-covariance matrix of the asymptotic limiting distribution of $\hat{\beta}^{\ddagger}$ as the number of clusters goes to infinity, and $\hat{\Phi}^{\ddagger} = \Phi(\hat{\psi}^{\ddagger})$ can be used as an approximation to the variance-covariance matrix of $\hat{\beta}^{\ddagger}$ (Pinheiro, 1994, Chapter 3). However, $\hat{\Phi}^{\ddagger}$ is biased when the sample size is small which can seriously overestimate β (Kackar & Harville, 1984). Ignoring any possible bias in $\hat{\psi}^{\ddagger}$, Kenward & Roger (1997) propose a better approximation to the small sample variance-covariance matrix of $\hat{\beta}^{\ddagger}$ through an adjusted estimator of the variance-covariance matrix of the fixed effects, which can be used to form a scaled Wald-type statistic that can result in better performance when conducting small sample inference for fixed effects.

2.4 Mean bias reduction

We recall from Section 1.3 that a suitable bias-reducing adjustment $A(\theta)$ to the score vector has components

$$A_t(\theta) = \frac{1}{2}\operatorname{tr}\left[\{i(\theta)\}^{-1}\{P_t(\theta) + Q_t(\theta)\}\right]\,, \tag{2.6}$$

for $t \in \{1,\ldots,p+m\}$, where $P_t(\theta) = E_\theta[s(\theta)s(\theta)^{\mathrm{T}}s_t(\theta)]$ and $Q_t(\theta) = E_\theta[-j(\theta)s_t(\theta)]$. Let $t \in \{1,\ldots,p\}$ correspond to an element of parameter β and $t \in \{p+1,\ldots,p+m\}$

correspond to an element of parameter ψ. We find that $A_t(\theta) = 0$ for $t \in \{1, \ldots, p\}$ and

$$A_t(\theta) = \frac{1}{2}\operatorname{tr}\left[V(\psi)^{-1}H(\psi)\frac{\partial V(\psi)}{\partial \psi_{t-p}}\right] + \frac{1}{2}\operatorname{tr}\left[\{i_{\psi\psi}\}^{-1}P_{4t}(\psi)\right]$$

for $t \in \{p+1, \ldots, p+m\}$, where $P_{4t}(\psi)$ is a $m \times m$ matrix with (r,s)th element

$$(P_{4t})_{r,s} = \frac{1}{2}\operatorname{tr}\left(V(\psi)^{-1}\frac{\partial^2 V(\psi)}{\partial \psi_r \partial \psi_s}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial \psi_{t-p}}\right).$$

The detailed calculation of the above results is given in Appendix A.

The mean BR adjusted score function for the fixed effects and the variance components of linear mixed models is $s^*_\beta(\theta) = s_\beta(\theta)$ and

$$s^*_{\psi_r}(\theta) = s^\ddagger_{\psi_r}(\theta) + \frac{1}{2}\operatorname{tr}\left[\{i_{\psi\psi}\}^{-1}P_{4\psi_r}(\psi)\right], \quad r \in \{1, \ldots, m\}$$

respectively, and the mean BR estimates $\hat{\theta}^* = (\hat{\beta}^{*\mathrm{T}}, \hat{\psi}^{*\mathrm{T}})^{\mathrm{T}}$ are the roots of the $p+m$ equations $s^*_\beta(\theta) = 0_p$ and $s^*_\psi(\theta) = 0_m$.

Theorem 1. *For covariance structures where*

$$\frac{\partial^2 V(\psi)}{\partial \psi_r \partial \psi_s} = 0 \tag{2.7}$$

for all pairs (r,s), $r,s \in \{1, \ldots, m\}$, the mean bias-reducing adjusted score function $s^*_{\psi_r}(\theta)$ coincides with the derivative of the restricted log-likelihood function $s^\ddagger_{\psi_r}(\theta)$. Then $s^*_\beta(\theta)$ and $s^*_\psi(\theta)$ are the derivatives of the mean BRPL function $l^*(\theta) = l(\theta) - \frac{1}{2}\log|X^{\mathrm{T}}V(\psi)^{-1}X|$.

Proof of Theorem 1: The condition in (2.7) implies that $P_{4t}(\psi) = 0_{m\times m}$ for all $t \in \{p+1, \ldots, p+m\}$, and therefore, $s^*_\psi(\theta) = s^\ddagger_\psi(\theta)$. □

A class of covariance structures for linear mixed models for which the condition in Theorem 1 holds is defined by $\Sigma(\sigma^2) = \sum_{i=1}^{m-1}\sigma_i^2 G_i$, where σ_i^2 are the elements of the variance-covariance matrix $\Sigma(\sigma^2)$ and G_i are known matrices. It is worth noting that the natural parameterisation for the unique elements in the variance-covariance matrix of the random effects satisfies the condition in Theorem 1.

2.5 Parameter estimation

Calculating the parameter estimates can be challenging especially when the random-effect structure is complex and/or the number of subjects or the number of observations per subject is small. In such cases any estimation algorithm might converge to parameter estimates that correspond to degenerate or singular variance-covariance matrices (Bates et al., 2015, Section 3.1). For example, if we are trying to fit a linear mixed model with a random intercept we might get zero random-effects variance, or in a linear mixed model with correlated random intercepts and slopes we might get a boundary correlation estimate of -1 or 1. Moreover, the variance components are constrained in complicated ways because the variance-covariance matrix of the unconditional distribution of the random effects has to be positive-definite.

For this reason we recommend using a suitable transformation of the parameters in the variance-covariance matrix of the random effects such that the resulting mean BR estimates of the variance-covariance matrix are positive-definite. Lindstrom & Bates (1988) describe the use of Cholesky factors for implementing unconstrained ML and REML estimation of the variance components in linear mixed models. The idea in Lindstrom & Bates (1988) is to replace $\Sigma(\sigma^2)$ in (2.2) by LL^{T}, where L is the Cholesky factor of $\Sigma(\sigma^2)$, whose unique elements form an unconstrained parameter vector. Then instead of estimating the natural random-effect parameters, we estimate the parameters on and below the diagonal of the lower triangular Cholesky factor. This ensures positive-definiteness and, hence, the invertibility of the variance-covariance matrix when evaluated at the parameter estimates. The disadvantage of this reparameterisation is that the elements in the Cholesky factor lack direct interpretation in terms of the original variances and covariances.

Let $\hat{\theta}^* = (\hat{\beta}^{*\mathrm{T}}, \hat{\lambda}^{*\mathrm{T}}, \hat{\sigma}_\varepsilon^{2*})^{\mathrm{T}}$ be the mean BR estimates, where $\hat{\lambda}^*$ are the estimates of the lower triangular elements of L. The Cholesky factor of $\Sigma(\sigma^2)$ does not satisfy the condition in Theorem 1, and therefore $\hat{\theta}^*$ is different than the REML estimator $\hat{\theta}^\ddagger$. We obtain $\hat{\theta}^*$ with the `nleqslv` *R* function (Hasselman, 2017) which numerically solves the system of adjusted score equations applying the Newton method with a numerical Jacobian matrix. The algorithm we implemented for computing the mean BR estimates may fail in estimating the variance components, especially for models with complex random-effect structure, unless good starting values are available. For this reason, we suggest using the REML estimates as starting values in the algorithm. We declare the algorithm has converged when the components of the mean bias-reducing adjusted score function are all smaller than $\varepsilon = 10^{-6}$ in absolute value when evaluated at $\hat{\theta}^*$.

2.6 Statistical inference

When the variance components are estimated using ML one can use the Wald test statistic to test the null hypothesis $\beta = \beta_0$ against the alternative $\beta \neq \beta_0$. Other appropriate inferential procedures are the likelihood ratio (LR) and the score tests. The three tests are asymptotically equivalent with a χ^2_p asymptotic null distribution (see, for example, Pace & Salvan, 1997, Section 5.9).

When the variance components are estimated using REML the Wald statistic has been found to be anti-conservative especially for small sample sizes, i.e. the test indicates that an effect may be important more often than expected under the null hypothesis of no effect (Gumedze & Dunne, 2011). The LR and score tests are also not reliable when REML estimation has been used. Kenward & Roger (1997) proposed a scaled Wald statistic, based on the adjusted estimator of the variance-covariance matrix of the REML fixed-effect estimates $\hat{\Phi}^{\ddagger}_A$, which accounts for the extra variability introduced by estimating the variance components by REML. Specifically, the bias-corrected variance-covariance matrix of the REML fixed effects is

$$\hat{\Phi}^{\ddagger}_A = \hat{\Phi}^{\ddagger} + 2\hat{\Phi}^{\ddagger}\left\{\sum_{r=1}^{m}\sum_{s=1}^{m} W_{rs}\left(Q_{rs} - P_r\hat{\Phi}^{\ddagger}P_s - \frac{1}{4}R_{rs}\right)\right\}\hat{\Phi}^{\ddagger}, \tag{2.8}$$

where W_{rs} is the (r,s)th element of the inverse of the expected information matrix $i^{\ddagger}_{\psi\psi}$ evaluated at $\hat{\psi}^{\ddagger}$, and

$$\begin{aligned} Q_{rs} &= X^{\mathrm{T}}V(\hat{\psi}^{\ddagger})^{-1}\frac{\partial V(\hat{\psi}^{\ddagger})}{\partial \psi_r}V(\hat{\psi}^{\ddagger})^{-1}\frac{\partial V(\hat{\psi}^{\ddagger})}{\partial \psi_s}V(\hat{\psi}^{\ddagger})^{-1}X, \\ P_r &= -X^{\mathrm{T}}V(\hat{\psi}^{\ddagger})^{-1}\frac{\partial V(\hat{\psi}^{\ddagger})}{\partial \psi_r}V(\hat{\psi}^{\ddagger})^{-1}X, \\ R_{rs} &= X^{\mathrm{T}}V(\hat{\psi}^{\ddagger})^{-1}\frac{\partial^2 V(\hat{\psi}^{\ddagger})}{\partial \psi_r \partial \psi_s}V(\hat{\psi}^{\ddagger})^{-1}X. \end{aligned}$$

The expected information matrix $i^{\ddagger}_{\psi\psi}$ is a $m \times m$ matrix with (r,s)th element

$$\begin{aligned} i^{\ddagger}_{\psi_r\psi_s} &= i_{\psi_r\psi_s} - \operatorname{tr}\left(V(\psi)^{-1}H(\psi)\frac{\partial V(\psi)}{\partial \psi_r}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial \psi_s}\right) \\ &\quad + \frac{1}{2}\operatorname{tr}\left(V(\psi)^{-1}H(\psi)\frac{\partial V(\psi)}{\partial \psi_r}V(\psi)^{-1}H(\psi)\frac{\partial V(\psi)}{\partial \psi_s}\right) \\ &\quad + \frac{1}{2}\operatorname{tr}\left(V(\psi)^{-1}H(\psi)\frac{\partial^2 V(\psi)}{\partial \psi_r \partial \psi_s}\right) . \end{aligned}$$

In the current work we focus on Wald-type inference. Specifically, we compare the performance of the Wald test using the mean BR estimates with the Wald test using the ML and REML estimates, as well as the Kenward & Roger (1997) scaled Wald test (KR) and the LR test.

2.7 Dental data

In this section we use the dental dataset (Potthoff & Roy, 1964) to illustrate the problematic behaviour of ML in terms of estimation and inference due to bias in the variance component estimates. The dataset consists of 108 measurements on the distance (mm) from the center of the pituitary to the pterygomaxillary fissure collected from 27 children (11 girls and 16 boys) at ages 8, 10, 12, 14 years. The objective of this orthodontic study was to determine whether distances are on average larger for boys than for girls over time.

Following Edwards et al. (2008) we fit a linear mixed model to this data with three different fixed-effect structures and two different random-effect structures. Specifically we fit the following six models written on the observational level as

$$\begin{aligned} \text{Model I:}\quad & Y_{ij} = \beta_0 + \beta_1 a_{ij} + u_{i0} + \varepsilon_{ij} \\ \text{Model II:}\quad & Y_{ij} = \beta_0 + \beta_1 a_{ij} + \beta_2 g_i + u_{i0} + \varepsilon_{ij} \\ \text{Model III:}\quad & Y_{ij} = \beta_0 + \beta_1 a_{ij} + \beta_2 g_i + \beta_3 a_{ij} g_i + u_{i0} + \varepsilon_{ij} \\ \text{Model IV:}\quad & Y_{ij} = \beta_0 + \beta_1 a_{ij} + u_{i0} + a_{ij} u_{i1} + \varepsilon_{ij} \\ \text{Model V:}\quad & Y_{ij} = \beta_0 + \beta_1 a_{ij} + \beta_2 g_i + u_{i0} + a_{ij} u_{i1} + \varepsilon_{ij} \\ \text{Model VI:}\quad & Y_{ij} = \beta_0 + \beta_1 a_{ij} + \beta_2 g_i + \beta_3 a_{ij} g_i + u_{i0} + a_{ij} u_{i1} + \varepsilon_{ij} . \end{aligned}$$

In the above models Y_{ij} denotes the jth measurement ($j = 1, \ldots, 4$) of the ith child ($i = 1, \ldots, 27$), a_{ij} denotes the age of the ith child when the jth measurement was made, g_i denotes the gender of the ith child (0 for male, 1 for female), u_{i0} is a random intercept that takes into account heterogeneity between children, and u_{i1} is a random slope of age within children correlated with the random intercept. For models I-III we assume that u_{i0} are independent and normally distributed with $N(0, \sigma^2_{u_0})$. For models IV-VI we assume that the random vectors $(u_{i0}, u_{i1})^{\mathrm{T}}$ are independent and identically distributed bivariate normal with mean zero and variance-covariance matrix Σ. The matrix Σ and its Cholesky factor L are given by

$$\Sigma = \begin{pmatrix} \sigma^2_{u_0} & \rho\sigma_{u_0}\sigma_{u_1} \\ \rho\sigma_{u_0}\sigma_{u_1} & \sigma^2_{u_1} \end{pmatrix} \quad \text{and} \quad L = \begin{pmatrix} \lambda_1 & 0 \\ \lambda_2 & \lambda_3 \end{pmatrix}.$$

Tables 2.1 and 2.2 give the ML, REML, and mean BR estimates of the parameters involved in models I-VI where the reported variance component estimates are the estimates of $\psi = (\sigma^2_{u_0}, \sigma^2_{u_1}, \rho, \sigma^2_{\varepsilon})^{\mathrm{T}}$ and $\psi = (\lambda_1, \lambda_2, \lambda_3, \sigma^2_{\varepsilon})^{\mathrm{T}}$, respectively. The parameterisation $\psi = (\sigma^2_{u_0}, \sigma^2_{\varepsilon})^{\mathrm{T}}$ of models I-III in Table 2.1 satisfies the condition in (2.7), and hence the REML and mean BR results are identical. Tables 2.1 and 2.2 illustrate that the fixed-effect estimates are similar between the three methods but their standard errors generally differ. Inclusion of random slopes in models IV-VI reduces error variance. The results also show that adding a fixed effect for any of the covariance structures impacts the estimates of the random effects variances, with the impact being more evident on the variance of the random intercepts. For example, comparing the estimates of the parameters in models IV and V given in Table 2.1, we see that the inclusion of gender as a fixed effect inflates the estimates of $\sigma^2_{u_0}$ from around 5 to nearly 8. This also affects the correlation between random intercept and random slope, increasing the estimate in absolute value. As a general conclusion, we argue that under various parameterisations and for various estimation methods, adding a fixed effect in the model can result in marked changes to the estimates of the variance components. Also, the bias reduced estimates of the variance components are larger than the ML estimates. This is typical, because the ML estimates do not account for the degrees of freedom used to estimate the fixed effects and they tend to be negatively biased.

Table 2.1: ML, REML, and mean BR estimates of the parameters in linear mixed models I-VI for the dental data using the parameterisation $\psi = (\sigma^2_{u_0}, \sigma^2_{u_1}, \rho, \sigma^2_{\varepsilon})^{\mathrm{T}}$. Estimated standard errors are reported in parentheses.

		Fixed effects				Variance components			
Model	Method	β_0	β_1	β_2	β_3	$\sigma^2_{u_0}$	$\sigma^2_{u_1}$	ρ	σ^2_{ε}
I	ML	16.76 (0.79)	0.66 (0.06)	-	-	4.29	-	-	2.02
	REML/Mean BR	16.76 (0.80)	0.66 (0.06)	-	-	4.47	-	-	2.05
II	ML	17.71 (0.82)	0.66 (0.06)	-2.32 (0.73)	-	2.99	-	-	2.02
	REML/Mean BR	17.71 (0.83)	0.66 (0.06)	-2.32 (0.76)	-	3.27	-	-	2.05
III	ML	16.34 (0.96)	0.78 (0.08)	1.03 (1.51)	-0.30 (0.12)	3.03	-	-	1.87
	REML/Mean BR	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.12)	3.30	-	-	1.92
IV	ML	16.76 (0.76)	0.66 (0.07)	-	-	4.81	0.05	-0.58	1.72
	REML	16.76 (0.78)	0.66 (0.07)	-	-	5.42	0.05	-0.61	1.72
	Mean BR	16.76 (0.78)	0.66 (0.07)	-	-	5.42	0.05	-0.75	1.72
V	ML	17.64 (0.86)	0.66 (0.07)	-2.15 (0.73)	-	6.99	0.05	-0.76	1.72
	REML	17.64 (0.89)	0.66 (0.07)	-2.15 (0.76)	-	7.82	0.05	-0.77	1.72
	Mean BR	17.62 (0.88)	0.66 (0.07)	-2.12 (0.66)	-	7.97	0.05	-0.84	1.72
VI	ML	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.13)	4.56	0.02	-0.60	1.72
	REML	16.34 (1.02)	0.78 (0.09)	1.03 (1.60)	-0.30 (0.13)	5.79	0.03	-0.67	1.72
	Mean BR	16.34 (1.02)	0.78 (0.09)	1.03 (1.60)	-0.30 (0.13)	5.79	0.03	-0.81	1.72

Table 2.2: ML, REML, and mean BR estimates of the parameters in linear mixed models I-VI for the dental data using the Cholesky parameterisation. Estimated standard errors are reported in parentheses.

		Fixed effects				Variance components			
Model	Method	β_0	β_1	β_2	β_3	λ_1	λ_2	λ_3	σ^2_{ε}
I	ML	16.76 (0.79)	0.66 (0.06)	-	-	2.07	-	-	2.02
	REML	16.76 (0.80)	0.66 (0.06)	-	-	2.11	-	-	2.05
	Mean BR	16.76 (0.81)	0.66 (0.06)	-	-	2.17	-	-	2.05
II	ML	17.71 (0.82)	0.66 (0.06)	-2.32 (0.73)	-	1.73	-	-	2.02
	REML	17.71 (0.83)	0.66 (0.06)	-2.32 (0.76)	-	1.81	-	-	2.05
	Mean BR	17.71 (0.84)	0.66 (0.06)	-2.32 (0.78)	-	1.86	-	-	2.05
III	ML	16.34 (0.96)	0.78 (0.08)	1.03 (1.51)	-0.30 (0.12)	1.74	-	-	1.87
	REML	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.12)	1.82	-	-	1.92
	Mean BR	16.34 (0.99)	0.78 (0.08)	1.03 (1.55)	-0.30 (0.12)	1.87	-	-	1.92
IV	ML	16.76 (0.76)	0.66 (0.07)	-	-	2.19	-0.12	0.17	1.72
	REML	16.76 (0.78)	0.66 (0.07)	-	-	2.33	-0.14	0.18	1.72
	Mean BR	16.76 (0.80)	0.66 (0.08)	-	-	2.54	-0.17	0.20	1.72
V	ML	17.64 (0.86)	0.66 (0.07)	-2.15 (0.73)	-	2.64	-0.16	0.14	1.72
	REML	17.64 (0.89)	0.66 (0.07)	-2.15 (0.76)	-	2.80	-0.17	0.15	1.72
	Mean BR	17.60 (0.91)	0.66 (0.07)	-2.06 (0.80)	-	2.95	-0.19	0.16	1.72
VI	ML	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.13)	2.13	-0.09	0.12	1.72
	REML	16.34 (1.02)	0.78 (0.09)	1.03 (1.60)	-0.30 (0.13)	2.41	-0.12	0.13	1.72
	Mean BR	16.34 (1.05)	0.78 (0.09)	1.03 (1.65)	-0.30 (0.14)	2.62	-0.15	0.16	1.72

In order to further investigate the performance of the ML, REML, and mean BR methods we performed a simulation study where we considered only linear mixed models I-III. From each of the models I-III we simulated 10 000 independent samples with true parameter values equal to the ML estimates shown in Table 2.1.

Table 2.3 gives the estimated mean bias of the estimates under the $\psi = (\sigma^2_{u_0}, \sigma^2_{\varepsilon})^{\mathrm{T}}$ parameterisation, the percentage of underestimation, the mean squared error, and the estimated relative increase in the mean squared error from its absolute minimum (the variance) due to bias (Kosmidis, 2014b, Table 5). The latter is calculated as the square of the bias divided by the variance. Table 2.3 illustrates the underestimation of the variance components by ML. The REML/mean BR methods correct for this underestimation. Comparing the values in the last column of Table 2.3 we can see the significance of the effect of estimation bias, especially in models II and III. The mean squared errors of the ML estimates of the variance components are inflated by as much as 7.5% due to bias from their minimum values (the variances). The corresponding inflation factors for the REML/mean BR estimators are almost zero.

The simulated samples were also used to calculate the empirical p-value distribution for the two-sided tests that each parameter is equal to the true values based on the LR and the Wald-type statistics. Table 2.4 shows that the empirical p-value distribution for the KR and the Wald statistic using the mean BR estimates are closest to uniformity.

Next, we repeated the simulation study where instead of estimating the variance components under the $\psi = (\sigma^2_{u_0}, \sigma^2_{\varepsilon})^{\mathrm{T}}$ parameterisation, we estimated $\psi = (\sigma_{u_0}, \sigma^2_{\varepsilon})^{\mathrm{T}}$. The results from this simulation setting are reported in Tables 2.5 and 2.6. The magnitude of mean bias of the ML estimates is smaller under this parameterisation, but ML still underestimates the variance components. The REML and mean BR methods reduce this bias and they also perform better in terms of percentage of underestimation. The mean squared error is similar across all methods. The mean squared errors of the ML estimates of the variance components are inflated by as much as 12.3% due to bias from their minimum values, while the corresponding inflation factors for the REML and mean BR estimators are significantly smaller and do not exceed 1%.

Similar to the previous parameterisation, we used the simulated samples to calcu-

Table 2.3: Mean bias, percentage of underestimation (PU), and mean squared error (MSE) of the variance component estimates for the linear mixed models I-III using the dental data setting and the $\psi = (\sigma^2_{u_0}, \sigma^2_\varepsilon)^\top$ parameterisation.

Model	Parameter	Method	Bias	PU	MSE	Bias^2/Variance (%)
I	$\sigma^2_{u_0}$	ML	-0.166	58.3	1.645	1.707
		REML/Mean BR	0.006	53.0	1.743	0.002
	σ^2_ε	ML	-0.019	54.0	0.100	0.362
		REML/Mean BR	0.006	51.2	0.102	0.036
II	$\sigma^2_{u_0}$	ML	-0.250	63.7	0.897	7.471
		REML/Mean BR	0.004	53.3	0.972	0.001
	σ^2_ε	ML	-0.019	54.0	0.100	0.362
		REML/Mean BR	0.006	51.2	0.102	0.036
III	$\sigma^2_{u_0}$	ML	-0.244	63.6	0.892	7.152
		REML/Mean BR	0.004	53.4	0.971	0.002
	σ^2_ε	ML	-0.041	57.2	0.087	1.939
		REML/Mean BR	0.006	51.4	0.090	0.037

Table 2.4: Empirical p-value distribution (%) for the likelihood ratio test and the tests based on the Wald statistic using the dental data setting and the $\psi = (\sigma^2_{u_0}, \sigma^2_\varepsilon)^\top$ parameterisation.

Model	$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
I	Likelihood ratio	1.1	2.7	4.9	9.4	24.7	49.9	74.8	89.7	94.7	97.4	99.0
	Wald using ML	1.2	2.8	5.3	9.7	24.9	50.0	74.8	89.7	94.7	97.4	99.0
	Kenward-Roger	1.0	2.5	4.7	9.1	24.4	49.6	74.6	89.5	94.6	97.4	99.0
	Wald using mean BR	1.2	2.7	5.2	9.4	24.6	49.7	74.6	89.6	94.7	97.4	99.0
II	Likelihood ratio	1.6	3.2	6.0	12.0	27.4	51.5	76.2	90.8	95.4	97.6	99.3
	Wald using ML	2.2	4.1	6.7	12.8	27.9	51.6	76.2	90.8	95.4	97.6	99.3
	Kenward-Roger	1.2	2.6	4.9	10.1	25.4	49.3	74.9	90.3	95.3	97.5	99.2
	Wald using mean BR	1.8	3.3	5.9	11.6	26.3	50.1	75.3	90.3	95.3	97.5	99.2
III	Likelihood ratio	1.0	2.7	5.6	10.4	26.0	50.1	74.7	89.9	95.4	97.7	99.0
	Wald using ML	1.2	3.1	6.0	10.9	26.1	50.1	74.7	89.9	95.4	97.7	99.0
	Kenward-Roger	0.9	2.5	5.4	10.0	25.2	49.5	74.5	89.8	95.3	97.7	99.0
	Wald using mean BR	1.0	2.8	5.6	10.4	25.5	49.6	74.5	89.8	95.3	97.7	99.0

late the empirical p-value distribution for the two-sided tests that each parameter is equal to the true values based on the LR and the Wald-type statistics. Table 2.6 shows that the empirical p-value distribution for the KR and the Wald statistic using the mean BR estimates are closest to uniformity.

In conclusion, the REML and mean BR estimators have smaller estimated mean bias and percentage of underestimation than the ML estimator. Also, the simulation studies suggest that the Wald-type statistics using REML or mean BR estimates and the KR statistic are more appropriate for statistical inference than LR and Wald statistics, and result in confidence intervals with good coverage properties.

Table 2.5: Mean bias, percentage of underestimation (PU), and mean squared error (MSE) of the Cholesky parameter estimates for the linear mixed models I-III using the dental data setting and the $\psi = (\sigma_{u_0}, \sigma_\varepsilon^2)^\mathrm{T}$ parameterisation.

Model	Parameter	Method	Bias	PU	MSE	Bias^2/Variance (%)
I	σ_{u_0}	ML	-0.065	58.3	0.102	4.279
		REML	-0.023	53.0	0.102	0.532
		Mean BR	0.028	47.1	0.105	0.766
	σ_ε^2	ML	-0.019	54.0	0.100	0.362
		REML	0.006	51.2	0.102	0.036
		Mean BR	0.006	51.2	0.102	0.036
II	σ_{u_0}	ML	-0.097	63.7	0.086	12.267
		REML	-0.023	53.3	0.082	0.633
		Mean BR	0.027	46.5	0.084	0.851
	σ_ε^2	ML	-0.019	54.0	0.100	0.362
		REML	0.006	51.2	0.102	0.036
		Mean BR	0.006	51.2	0.102	0.036
III	σ_{u_0}	ML	-0.094	63.6	0.084	11.802
		REML	-0.022	53.4	0.081	0.609
		Mean BR	0.026	46.8	0.083	0.835
	σ_ε^2	ML	-0.041	57.2	0.087	1.939
		REML	0.006	51.4	0.090	0.037
		Mean BR	0.006	51.4	0.090	0.037

Table 2.6: Empirical p-value distribution (%) for the tests based on the Wald statistic using the dental data setting and the $\psi = (\sigma_{u_0}, \sigma_\varepsilon^2)^\mathrm{T}$ parameterisation.

Model	$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
I	Likelihood ratio	1.1	2.7	4.9	9.4	24.7	49.9	74.8	89.7	94.7	97.4	99.0
	Wald using ML	1.2	2.8	5.3	9.7	24.9	50.0	74.8	89.7	94.7	97.4	99.0
	Wald using REML	1.2	2.7	5.2	9.4	24.6	49.7	74.6	89.6	94.7	97.4	99.0
	Kenward-Roger	1.0	2.5	4.7	9.1	24.4	49.6	74.6	89.5	94.6	97.4	99.0
	Wald using mean BR	1.2	2.7	5.2	9.4	24.6	49.7	74.6	89.6	94.7	97.4	99.0
II	Likelihood ratio	1.6	3.2	6.0	12.0	27.4	51.5	76.2	90.8	95.4	97.6	99.3
	Wald using ML	2.2	4.1	6.7	12.8	27.9	51.6	76.2	90.8	95.4	97.6	99.3
	Wald using REML	1.8	3.3	5.9	11.6	26.3	50.1	75.3	90.3	95.3	97.5	99.2
	Kenward-Roger	1.3	2.8	5.2	10.7	26.0	50.0	75.3	90.4	95.3	97.5	99.2
	Wald using mean BR	1.5	3.0	5.4	10.5	25.4	48.9	74.6	90.2	95.2	97.5	99.2
III	Likelihood ratio	1.0	2.7	5.6	10.4	26.0	50.1	74.7	89.9	95.4	97.7	99.0
	Wald using ML	1.2	3.1	6.0	10.9	26.1	50.1	74.7	89.9	95.4	97.7	99.0
	Wald using REML	1.0	2.8	5.6	10.4	25.5	49.6	74.5	89.8	95.3	97.7	99.0
	Kenward-Roger	0.9	2.5	5.4	10.0	25.2	49.5	74.5	89.8	95.3	97.7	99.0
	Wald using mean BR	1.0	2.8	5.6	10.4	25.5	49.6	74.5	89.8	95.3	97.7	99.0

2.8 Simulation study

In this section we present simulation results to study the behaviour of the mean BR method under small and moderate sample sizes when fitting a linear mixed model with a random intercept and a correlated random slope.

We generated data from the model used by Vaida & Blanchard (2005), defined as

$$y_{ij} = (\beta_0 + \beta_1 t_j) + (\alpha_i + t_j b_i) + \varepsilon_{ij} \quad (i = 1, \ldots, 10, \;\; j = 1, \ldots, n_i) \tag{2.9}$$

where $\beta_0 = -2.78$, $\beta_1 = -0.186$, $t_j = 5j$, $(\alpha_i, b_i)^{\mathrm{T}}$ is normally distributed with mean zero and variance-covariance matrix

$$\begin{pmatrix} 3.67 & -0.126 \\ -0.126 & 0.279 \end{pmatrix},$$

and ε_{ij} are independent and identically distributed with $N(0, \sigma_\varepsilon^2)$. Following Vaida & Blanchard (2005) and Liang et al. (2008) we let the error variance be equal to $\sigma_\varepsilon^2 = 0.0705^2$, 0.141^2 and 0.282^2. Also, we let j take the values (i) $j \in \{0, 1, \ldots, 5\}$ and (ii) $j \in \{0, 1, \ldots, 25\}$, giving cluster sizes $n_i = 6$ and 26, respectively. For each of these six scenarios we simulated 10 000 samples and we estimated the mean bias, the percentage of underestimation, the mean squared error, and the variance of the parameter estimates under the ML, REML, and mean BR fit of model (2.9).

Tables 2.7 and 2.8 summarise the results of the simulation study. The REML and mean BR methods reduce the bias of the ML estimates of the variance components, especially of the Cholesky parameter λ_1, with the mean BR yielding the smallest bias. The mean squared error is in all scenarios similar across the three estimation methods. The mean squared errors of the ML and REML estimates are inflated by as much as 35% and 15% due to bias, respectively. On the other hand, the corresponding inflation factor for the mean BR estimates is very close to zero and does not exceed 0.4%. Table 2.8 illustrates once again that the empirical p-value distributions for the KR and the Wald statistic using the mean BR estimates are closest to uniformity.

Lastly, the results suggest that as the cluster size increases, the differences between the ML, REML and mean BR estimators do not decrease, and the performance of the LR and traditional Wald tests does not improve. Also, we should note that the KR statistic is a strong competitor of the Wald statistic using the mean BR estimates when conducting inference, but the latter has the advantage of being computationally less expensive and easier to implement.

Table 2.7: Mean bias, percentage of underestimation (PU), and mean squared error (MSE) of the Cholesky parameter estimates under the linear mixed model (2.9) with cluster size n_i and variance error σ_ε^2.

	Bias				PU				MSE				Bias^2/Variance (%)			
Method	λ_1	λ_2	λ_3	σ_ε^2	λ_1	λ_2	λ_3	σ_ε^2	λ_1	λ_2	λ_3	σ_ε^2	λ_1	λ_2	λ_3	σ_ε^2
							$n_i = 6, \sigma_\varepsilon^2 = 0.0705^2$									
ML	-0.16	0.01	-0.07	0.00	65.8	48.7	72.5	54.5	0.20	0.03	0.02	0.00	13.53	0.14	35.01	0.11
REML	-0.06	0.00	-0.04	0.00	56.9	49.4	64.9	54.5	0.20	0.03	0.02	0.00	1.83	0.03	12.82	0.11
Mean BR	-0.01	0.00	0.00	0.00	52.0	49.7	50.2	54.5	0.21	0.03	0.02	0.00	0.02	0.00	0.03	0.11
							$n_i = 6, \sigma_\varepsilon^2 = 0.141^2$									
ML	-0.16	0.01	-0.07	0.00	66.0	48.5	72.5	54.5	0.20	0.03	0.02	0.00	13.54	0.15	35.10	0.11
REML	-0.06	0.00	-0.04	0.00	56.8	49.4	64.9	54.5	0.20	0.03	0.02	0.00	1.84	0.03	12.87	0.11
Mean BR	-0.01	0.00	0.00	0.00	51.9	49.6	50.1	54.5	0.21	0.03	0.02	0.00	0.02	0.01	0.03	0.11
							$n_i = 6, \sigma_\varepsilon^2 = 0.282^2$									
ML	-0.16	0.01	-0.07	0.00	65.8	48.4	72.7	54.5	0.21	0.03	0.02	0.00	13.58	0.16	35.45	0.11
REML	-0.06	0.00	-0.04	0.00	56.9	49.2	65.3	54.5	0.21	0.03	0.02	0.00	1.86	0.04	13.08	0.11
Mean BR	-0.01	0.00	0.00	0.00	51.8	49.6	50.0	54.5	0.22	0.03	0.02	0.00	0.02	0.01	0.04	0.11
							$n_i = 26, \sigma_\varepsilon^2 = 0.0705^2$									
ML	-0.15	0.00	-0.07	0.00	64.9	49.6	73.4	52.0	0.20	0.03	0.02	0.00	12.24	0.07	35.56	0.04
REML	-0.06	0.00	-0.05	0.00	56.5	50.2	65.8	52.0	0.21	0.03	0.02	0.00	1.77	0.00	15.14	0.04
Mean BR	0.00	0.00	0.00	0.00	51.5	50.7	51.8	51.8	0.21	0.03	0.02	0.00	0.01	0.01	0.01	0.02
							$n_i = 26, \sigma_\varepsilon^2 = 0.141^2$									
ML	-0.15	0.00	-0.07	0.00	64.8	49.5	73.5	52.0	0.20	0.03	0.02	0.00	12.29	0.07	35.59	0.04
REML	-0.05	0.00	-0.04	0.00	56.1	50.3	65.2	52.0	0.21	0.03	0.02	0.00	1.44	0.00	13.11	0.04
Mean BR	0.00	0.00	0.00	0.00	51.6	50.6	51.2	52.0	0.21	0.03	0.02	0.00	0.00	0.00	0.03	0.04
							$n_i = 26, \sigma_\varepsilon^2 = 0.282^2$									
ML	-0.15	0.00	-0.07	0.00	64.7	49.3	73.3	52.0	0.21	0.03	0.02	0.00	12.38	0.07	35.69	0.04
REML	-0.05	0.00	-0.04	0.00	56.2	50.2	65.3	52.0	0.21	0.03	0.02	0.00	1.47	0.00	13.15	0.04
Mean BR	0.00	0.00	0.00	0.00	51.7	50.7	51.0	52.0	0.22	0.03	0.02	0.00	0.00	0.00	0.03	0.04

Table 2.8: Empirical p-value distribution (%) for the tests based on the Wald statistic using the Cholesky parameter estimates under the linear mixed model (2.9) with cluster size n_i and variance error σ_ε^2.

n_i	σ_ε^2	$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
6	0.0705^2	Likelihood ratio	2.3	4.6	8.1	14.5	31.2	56.0	77.1	90.8	95.6	97.8	99.1
		Wald using ML	3.8	6.4	9.9	15.6	30.8	54.6	76.1	90.3	95.5	97.6	99.1
		Wald using REML	3.1	5.2	8.2	13.7	28.5	52.7	75.1	89.8	95.2	97.5	99.0
		Kenward-Roger	1.7	3.7	6.9	12.9	28.9	54.2	76.0	90.2	95.4	97.6	99.1
		Wald using mean BR	2.3	3.8	6.3	11.1	24.7	48.7	73.3	88.8	94.8	97.2	99.0
6	0.141^2	Likelihood ratio	2.3	4.6	8.0	14.5	31.3	56.0	77.0	90.7	95.6	97.7	99.1
		Wald using ML	3.9	6.4	9.9	15.5	30.8	54.7	76.1	90.2	95.5	97.6	99.1
		Wald using REML	3.2	5.3	8.2	13.6	28.5	52.7	75.1	89.7	95.3	97.4	99.0
		Kenward-Roger	1.7	3.8	6.9	12.8	29.0	54.1	76.1	90.2	95.4	97.6	99.1
		Wald using mean BR	2.3	3.9	6.3	11.1	24.6	48.8	73.4	88.9	94.8	97.2	98.9
6	0.282^2	Likelihood ratio	2.2	4.6	8.1	14.4	31.3	56.0	77.1	90.6	95.7	97.7	99.1
		Wald using ML	3.8	6.4	10.0	15.5	30.7	54.6	76.4	90.2	95.6	97.6	99.1
		Wald using REML	3.2	5.3	8.3	13.6	28.5	52.6	75.1	89.6	95.3	97.5	99.0
		Kenward-Roger	1.8	3.7	6.9	12.9	29.0	53.9	76.2	90.1	95.6	97.6	99.1
		Wald using mean BR	2.2	3.9	6.2	11.1	24.7	48.7	73.2	88.9	94.8	97.3	98.9
26	0.0705^2	Likelihood ratio	2.2	4.8	8.6	14.9	32.0	55.9	78.3	91.4	95.4	97.7	99.1
		Wald using ML	3.9	6.7	10.5	16.1	31.8	54.6	77.4	91.0	95.1	97.5	99.0
		Wald using REML	3.1	5.7	8.8	14.2	29.5	52.8	76.5	90.3	94.9	97.4	99.0
		Kenward-Roger	1.7	3.7	7.0	13.1	29.7	53.9	77.2	91.0	95.1	97.4	99.1
		Wald using mean BR	2.3	4.0	6.6	11.4	25.5	49.1	74.2	89.5	94.5	97.2	98.9
26	0.141^2	Likelihood ratio	2.2	4.8	8.6	15.0	32.0	55.9	78.3	91.4	95.4	97.6	99.1
		Wald using ML	3.9	6.6	10.3	15.9	31.5	54.4	77.3	91.0	95.2	97.5	99.1
		Wald using REML	3.1	5.6	8.7	13.8	29.1	52.5	76.3	90.3	94.9	97.4	99.0
		Kenward-Roger	1.7	3.7	7.0	13.1	29.7	53.9	77.2	91.0	95.1	97.5	99.1
		Wald using mean BR	2.3	3.9	6.5	11.3	25.2	48.8	74.1	89.5	94.4	97.2	99.0
26	0.282^2	Likelihood ratio	2.2	4.8	8.6	14.9	32.0	55.9	78.3	91.4	95.4	97.7	99.1
		Wald using ML	3.9	6.5	10.4	15.9	31.5	54.4	77.3	90.9	95.2	97.5	99.1
		Wald using REML	3.1	5.6	8.7	13.9	29.1	52.5	76.2	90.3	94.9	97.4	99.0
		Kenward-Roger	1.7	3.7	7.0	13.1	29.7	53.9	77.2	90.9	95.2	97.4	99.1
		Wald using mean BR	2.3	3.9	6.5	11.3	25.2	48.8	74.1	89.6	94.4	97.2	98.9

2.9 Mean bias reduction in random effects meta-analysis and meta-regression

Linear mixed models is a wide class of parametric models including as special cases the random effects meta-analysis and meta-regression models considered in this section.

2.9.1 Random effects meta-regression model

Let y_i and $\hat{\sigma}_i^2$ denote the estimate of the effect from the ith study $(i = 1, \ldots, K)$ and the associated within-study variance, respectively, and $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ denote a p-dimensional vector of study-specific covariates that can be used to account for the heterogeneity across studies.

The within-study variances $\hat{\sigma}_i^2$ are usually assumed to be estimated well-enough to be considered as known and equal to the values reported in each study. Then the observations $y_1, \ldots, y_K$ are assumed to be realisations of the random variables $Y_1, \ldots, Y_K$, which are independent conditionally on independent random effects $U_1, \ldots, U_K$. The conditional distribution of Y_i given $U_i = u_i$ is $N(u_i + x_i^{\mathrm{T}}\beta, \hat{\sigma}_i^2)$, where β is an unknown p-dimensional vector of fixed effects. The random effects $U_1, \ldots, U_K$ are typically assumed to be independent with U_i having a $N(0, \psi)$ distribution, where ψ is a parameter that attempts to capture the unexplained between-study heterogeneity. In matrix notation, the random effects meta-regression model has

$$Y = X\beta + U + \varepsilon\,, \tag{2.10}$$

where $Y = (Y_1, \ldots, Y_K)^{\mathrm{T}}$, X is the $K \times p$ model matrix with x_i^{T} in its ith row, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_K)^{\mathrm{T}}$ is a vector of independent errors each with a $N(0, \hat{\sigma}_i^2)$ distribution and independent of $U = (U_1, \ldots, U_K)^{\mathrm{T}}$. Under this specification, the marginal distribution of Y is multivariate normal with mean $X\beta$ and variance-covariance matrix $\hat{\Sigma} + \psi I_K$, where I_K is the $K \times K$ identity matrix and $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2)$. The random effects meta-analysis results as a special case of meta-regression, by setting X to be a column of ones.

2.9.2 Methods of estimation

Under the random effects meta-regression model there are a number of options for estimating the parameters. A traditional method is ML, which maximises the log-likelihood function for $\theta = (\beta^{\mathrm{T}}, \psi)^{\mathrm{T}}$ given by

$$l(\theta) = \frac{1}{2}\{\log|W(\psi)| - R(\beta)^{\mathrm{T}} W(\psi) R(\beta)\},$$

where $|W(\psi)|$ denotes the determinant of $W(\psi) = (\hat{\Sigma} + \psi I_K)^{-1}$ and $R(\beta) = y - X\beta$. The score function is

$$s(\theta) = \begin{pmatrix} X^{\mathrm{T}} W(\psi) R(\beta) \\ \frac{1}{2}\{R(\beta)^{\mathrm{T}} W(\psi)^2 R(\beta) - \mathrm{tr}[W(\psi)]\} \end{pmatrix} \tag{2.11}$$

and the ML estimator $\hat{\theta} = (\hat{\beta}^{\mathrm{T}}, \hat{\psi})^{\mathrm{T}}$ is obtained as the solution of $s(\theta) = 0_{p+1}$. Even though ML is attractive for its asymptotic properties, the resulting between-study heterogeneity estimate $\hat{\psi}$ is negatively biased if the number of studies is small (Schwarzer et al., 2015, Section 2.3).

Another traditional approach for fitting random effects meta-regression models is the DerSimonian & Laird (1986) procedure (DL). The DL estimator of β is the weighted average $\hat{\beta}_{DL} = \hat{\beta}(\hat{\psi}_{DL}) = (X^{\mathrm{T}} W(\hat{\psi}_{DL}) X)^{-1} X^{\mathrm{T}} W(\hat{\psi}_{DL}) y$, where $\hat{\psi}_{DL}$ is the DL estimator of ψ calculated as

$$\hat{\psi}_{DL} = \max\left\{0, \frac{Q - (K-p)}{\mathrm{tr}(\hat{\Sigma}^{-1}) - \mathrm{tr}\{(X^{\mathrm{T}}\hat{\Sigma}^{-1}X)^{-1} X^{\mathrm{T}} \hat{\Sigma}^{-2} X\}}\right\}. \tag{2.12}$$

The quantity Q involved in (2.12) is the observed value of Cochran's statistic defined as $(y - X\hat{\beta}(0))^{\mathrm{T}} \hat{\Sigma}^{-1} (y - X\hat{\beta}(0))$ (Cochran, 1937). Even though the DL method is simple to implement, it can lead to unreliable inferential conclusions. Specifically, confidence intervals for the fixed effects are generally narrower than they should be, because the variability associated to the estimation of the between-study heterogeneity is not taken into account (Brockwell & Gordon, 2001; Guolo & Varin, 2017).

Several alternatives to the ML and DL methods have been proposed in the literature to account for the uncertainty in estimating the between-study heterogeneity, especially when the number of studies is small. For example, Knapp & Hartung (2003) introduced a modified limiting distribution of test statistics based on an improved estimator of the variance of the parameter estimates, Zeng & Lin (2015) suggested a double resampling approach which accounts for the variation in the estimation of ψ, and recently Kosmidis et al. (2017) suggested maximising a mean BRPL to get reduced-bias estimates of ψ. The aforementioned proposals have been shown to improve coverage accuracy by yielding wider confidence intervals for β than those obtained from the ML or DL approaches.

In this chapter we focus only on the mean BRPL method. The maximum mean BRPL estimator $\hat{\theta}^* = (\hat{\beta}^{*\mathrm{T}}, \hat{\psi}^*)^{\mathrm{T}}$ solves the mean bias-reducing adjusted score equations for β and ψ, specifically $s^*_\beta(\theta) = s_\beta(\theta)$ and $s^*_\psi(\theta) = s_\psi(\theta) + \mathrm{tr}[W(\psi)H(\psi)]/2$, respectively (Kosmidis et al., 2017). A direct approach for computing $\hat{\theta}^*$ is through the following two-step iterative process (Kosmidis et al., 2017). At the jth iteration $(j = 1, 2, \ldots)$

1. calculate $\beta^{(j)}$ by weighted least squares as

$$\beta^{(j)} = (X^{\mathrm{T}} W(\psi^{(j-1)})X)^{-1} X^{\mathrm{T}} W(\psi^{(j-1)})y$$

2. solve $s^*_\psi(\theta^{(j)}(\psi)) = 0$ with respect to ψ, where $\theta^{(j)}(\psi) = (\beta^{(j)\mathrm{T}}, \psi)^{\mathrm{T}}$.

In the above steps, $\beta^{(j)}$ is the candidate value for $\hat{\beta}^*$ at the jth iteration and $\psi^{(j-1)}$ is the candidate value for $\hat{\psi}^*$ at the $(j-1)$th iteration. The equation in step 2 is solved numerically, by searching for the root of the function $s^*_\psi(\beta^{(j)}, \psi)$ in a predefined positive interval. For the computations in this chapter we use the DL estimate of ψ as starting value $\psi^{(0)}$. The iterative process is then repeated until the components of the score function $s^*(\theta)$ are all less than $\varepsilon = 10^{-6}$ in absolute value at the current estimates.

The remainder of this section uses the real data applications used in Kosmidis et al. (2017) to compare the performance in estimation, inference, and computational speed of mean BRPL against ML, a method which has not been taken into consideration in Kosmidis et al. (2017). The first application uses the cocoa data (Taubert et al., 2007) to study the performance of the methods in a random effects meta-analysis setting, and the second application uses the meat consumption data (Larsson & Orsini,

2014) to study the performance of the methods in a random effects meta-regression setting. The results in Kosmidis et al. (2017) demonstrate the superior performance in estimation and inference against other existing estimation methods for random effects meta-analysis and meta-regression models, and for this reason we did not include any of these methods in our study.

2.9.3 Cocoa intake and blood pressure reduction data

Consider the setting in Bellio & Guolo (2016) who carry out a meta-analysis of five randomised controlled trials from Taubert et al. (2007) on the efficacy of two weeks of cocoa consumption on lowering diastolic blood pressure. The top panel in Figure 2.1 is a forest plot with the estimated mean difference in diastolic blood pressure before and after cocoa intake from each study, and the associated 95% Wald-type confidence intervals. Four out of the five studies reported a reduction of diastolic blood pressure from cocoa intake.

The random effects meta-analysis model is used to synthesise the evidence from the five studies. In particular, let Y_i be a random variable representing the mean difference in the diastolic blood pressure after two weeks of cocoa intake in the ith study. We assume that $Y_1,\ldots,Y_5$ are independent random variables where Y_i has a Normal distribution with mean the overall effect β and variance $\hat{\sigma}_i^2+\psi$.

The bottom panel in Figure 2.1 depicts nominally 95% confidence intervals for β using various alternative methods. As is apparent, the conclusions when testing the hypothesis $\beta=0$ can vary depending on the method used. More specifically, the Wald test using the ML estimates, the DL method, double resampling, and the LR test give evidence that there is a relationship between cocoa consumption and diastolic blood pressure, with p-values 0.005, 0.006, 0.016, 0.030, respectively. On the other hand, Knapp & Hartung (2003) method, the mean BRPL ratio, the Bartlett-corrected LR (Huizenga et al., 2011), and Skovgaard's test, suggest that the evidence that cocoa consumption affects diastolic blood pressure is weaker, with p-values 0.050, 0.053, 0.058, 0.067, 0.077, respectively. The median BRPL ratio that is also reported in Figure 2.1 is a newly proposed statistic derived in Chapter 6 that can also be used for carrying out hypothesis tests and constructing confidence intervals for the random effects meta-regression model parameters.

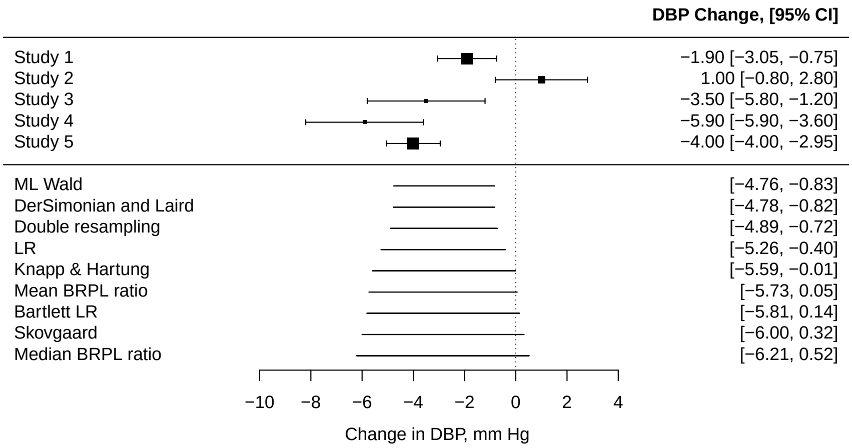

Figure 2.1: Forest plot of cocoa data. The outcomes from the five studies are reported in terms of the diastolic blood pressure (DBP) difference after two weeks of cocoa consumption. A negative change in DBP indicates favourable hypotensive cocoa actions. Squares represent the mean effect estimate for each study; the size of the square reflects the weight that the corresponding study exerts in the meta-analysis calculated as the within-study's inverse variance. Horizontal line segments represent 95% Wald-type confidence intervals for the effect estimate of individual studies. In the bottom panel of the plot horizontal line segments represent the corresponding 95% confidence interval as computed based on various statistics (for details, see text). The confidence intervals are ordered according to their length.

The ML and the mean BRPL estimates of the heterogeneity parameter in the meta-analysis model are $\hat{\psi} = 4.199$ and $\hat{\psi}^* = 5.546$. The estimates of the common effect are $\hat{\beta} = -2.799$, and $\hat{\beta}^* = -2.811$, with estimated standard errors 1.002 and 1.129, respectively. The bias-reduced estimate of ψ and, as a consequence, the corresponding estimated standard error for β are larger than their ML counterparts, which is typical in random effects meta-analysis. The iterative process used for computing the ML and maximum mean BRPL estimates converged in 4 and 5 iterations, respectively. The computational run-time for the two-step iterative process which computes the ML and maximum mean BRPL estimates is 1.1×10^{-2} and 1.8×10^{-2} seconds, respectively.

The 95% confidence intervals for β are $(-5.26, -0.40)$ and $(-5.73, 0.05)$ for the LR statistic and the mean BRPL ratio statistic, respectively. The corresponding 95% confidence intervals for ψ are $(1.1, 23.5)$ and $(1.4, 58.0)$, respectively.

In order to further investigate the performance of the two approaches to estimation and inference, we performed a simulation study where we simulated 10 000 independent samples from the random effects meta-analysis model with parameter values set to the ML estimates reported earlier, i.e. $\beta_0 = -2.799$ and $\psi_0 = 4.199$. The esti-

Table 2.9: Empirical p-value distribution (%) for the tests based on the LR statistic and the mean BRPL ratio statistic in the cocoa data setting.

$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
LR	5.9	8.4	11.7	18.2	34.5	57.8	79.1	91.7	96.0	98.0	99.2
Mean BRPL ratio	1.6	3.7	6.7	12.1	28.3	52.8	76.6	90.9	95.5	97.9	99.1

Notes: Each column gives the coverage probability of $(1-\alpha)\%$ confidence intervals based on the LR or the mean BRPL ratio statistics.

mated mean bias of the ML estimator of ψ is -0.990 and the bias corresponding to the maximum mean BRPL estimator is considerably smaller and equal to 0.028. Let $\psi_0 = 4.199$. The simulation-based estimates of the probabilities of underestimation for ψ, $P_{\psi_0}(\hat{\psi} \leq \psi_0)$ and $P_{\psi_0}(\hat{\psi}^* \leq \psi_0)$ are 0.708 and 0.591 for the ML and maximum mean BRPL, respectively.

The simulated samples were also used to calculate the empirical p-value distribution for the two-sided tests that each parameter is equal to the true values based on the LR statistic and the mean BRPL ratio statistic. Table 2.9 shows that the empirical p-value distribution for the mean BRPL ratio statistic is closest to uniformity. The coverage probability of the 95% confidence intervals for β based on the mean BRPL ratio is notably closer to the nominal level than the one based on the LR. Specifically, the coverage probabilities for β are 88% and 93% for LR and mean BRPL ratio, respectively, and the corresponding coverage probabilities for ψ are 88% and 94%, respectively.

Overall, the results indicate that mean BRPL is superior in estimation against ML with a small additional computational run-time, and mean BRPL ratio outperforms LR resulting in confidence intervals with better coverage properties.

2.9.4 Meat consumption data

A well used example in the random effects meta-regression literature is the meat consumption data (Larsson & Orsini, 2014) used for investigating the association between meat consumption and relative risk of all-cause mortality. The data consists of 16 prospective studies, eight of which are about unprocessed red meat consumption and eight about processed meat consumption. Figure 2.2 displays the information provided by each study in the meta-analysis. The results from the studies point towards the conclusion that high consumption of red meat, in particular processed red meat, is associated with higher all-cause mortality.

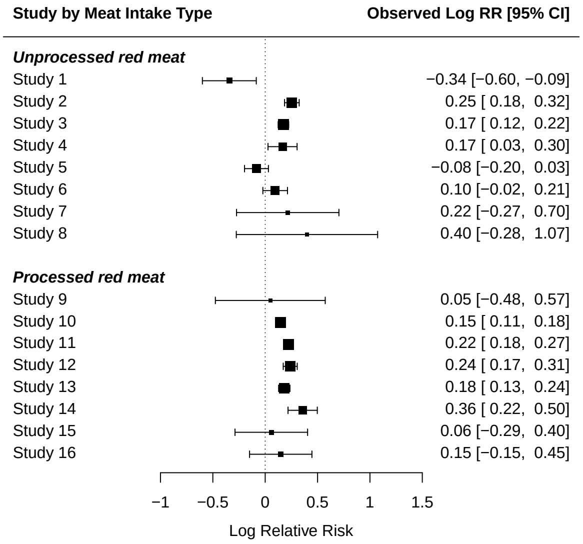

Figure 2.2: The meat consumption data. Outcomes from 16 studies are reported in terms of the logarithm of the relative risk (Log RR) of all-cause mortality for the highest versus lowest category of unprocessed red meat, and processed meat consumption. Squares represent the mean effect estimate for each study; the size of the square reflects the weight that the corresponding study exerts in the meta-analysis. Horizontal lines represent 95% Wald-type confidence intervals for the effect estimate of individual studies.

We consider the random effects meta-regression model assuming that Y_i has a $N(\beta_0 + \beta_1 x_i, \hat{\sigma}_i^2 + \psi)$, where Y_i is the random variable representing the logarithm of the relative risk reported in the ith study, and x_i takes value 1 if the consumption in the ith study is about processed red meat and 0 if it is about unprocessed meat $(i = 1, \ldots, 16)$.

Table 2.10 gives the ML estimates and the maximum mean BRPL estimates of the fixed effects and the heterogeneity parameter, along with the corresponding estimated standard errors and the 95% confidence intervals. The results show that the ML estimate of ψ, as well as the estimated standard errors for the fixed effects have the smallest values. The LR test indicates some evidence for a higher risk associated to the consumption of red processed meat with a p-value of 0.047. On the other hand, the mean BRPL ratio test suggests that there is weaker evidence for higher risk with p-value of 0.066. The iterative process used for computing the ML and maximum mean BRPL estimates converged in 8 and 9 iterations, respectively. The computational runtime for the two-step iterative process which computes the ML and maximum mean BRPL estimates is 1.2×10^{-2} and 2.4×10^{-2} seconds, respectively.

Similar to Section 2.9.3, we performed a simulation study in order to further in-

Table 2.10: ML and mean BRPL estimates of the model parameters for the meat consumption data. Estimated standard errors are reported in parentheses. The 95% confidence intervals based on the LR and mean BRPL ratio are reported in squared brackets.

Method	β_0	β_1	ψ
ML	0.099 (0.044)	0.106 (0.061)	0.009
	[-0.004,0.189]	[-0.022,0.244]	[0.003,0.030]
Mean BRPL	0.095 (0.050)	0.110 (0.069)	0.012
	[-0.020,0.199]	[-0.040,0.264]	[0.003,0.042]

Table 2.11: Empirical p-value distribution (%) for the tests based on the LR statistic and the mean BRPL ratio statistic using the meat consumption data.

$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
LR	2.2	4.5	7.7	13.1	28.0	50.0	71.7	86.6	92.1	95.3	97.7
Mean BRPL ratio	1.3	3.0	5.6	11.1	25.9	49.8	73.8	89.0	94.2	96.9	98.6

Notes: Each column gives the coverage probability of $(1-\alpha)\%$ confidence intervals based on the LR or the mean BRPL ratio statistics.

vestigate the performance of the two methods in a meta-regression context. We simulated 10 000 independent samples from the meta-regression model at the ML estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\psi})^{\mathrm{T}}$ reported in Table 2.10. ML slightly underestimates the heterogeneity parameter with mean bias -0.002, while mean BRPL almost fully compensate for the negative bias of ML estimate, with mean bias 3×10^{-5}. The percentages of underestimation are 72.6% and 56.6% for the ML and maximum mean BRPL estimators, respectively.

The simulated samples were also used to calculate the empirical p-value distribution for the tests based on the LR and mean BRPL ratio statistics. Table 2.11 shows that the empirical p-value distribution for the mean BRPL ratio statistic is closer to uniformity.

Similar to the cocoa consumption study, the results indicate that in random effects meta-regression mean BRPL is superior in estimation against ML with a small additional computational run-time, and mean BRPL ratio outperforms LR resulting in confidence intervals with better coverage properties.

2.10 Concluding remarks

In this chapter we derive the adjusted score equations for mean bias reduction of the ML estimator for linear mixed models under any parameterisation of the variance components. We show that under certain parameterisations the solution of the mean bias-reducing adjusted score equations is identical to the REML estimates, and we give a sufficient condition for the equality of mean BR with REML estimates. Our simulation studies indicate that adding a predictor in the fixed effects can increase the estimated variance of the random effects. The results also illustrate a significant improvement of the random-effect parameter estimation compared to ML and REML.

This chapter also highlights the need for reducing estimation bias in linear mixed models. Our results provide evidence that the bias of ML estimates can affect Wald-type inference. On the other hand, the proposed mean bias reduction method corrects the anti-conservativeness of the traditional Wald test. The good performance of the Wald statistic using the mean BR estimates is comparable with the KR statistic (Kenward & Roger, 1997). A disadvantage of the well-established KR approximation is its complexity in calculating the approximate variance-covariance matrix as well as the denominator degrees of freedom of its small sample F-distribution, which make the hypothesis testing computationally more expensive, than simply computing the Wald statistic using the mean BR estimates. The results were qualitatively similar across all parameterisations considered.

Lastly, using two simulation settings we were able to retrieve enough information on the performance of the maximum mean BRPL estimators proposed in Kosmidis et al. (2017) for mean bias reduction of the ML estimator for random effects meta-analysis and meta-regression models. All the results illustrate that use of the mean BRPL succeeds in achieving bias reduction in estimation, which leads to confidence intervals with good coverage properties. The computation of the maximum mean BRPL estimates is not expensive, as illustrated by the computational run-times and number of iterations reported.

In Chapter 6 we derive the adjusted score equations for median bias reduction of the ML estimator for linear mixed models and random effects meta-analysis and meta-regression, and include more simulation studies to assess the performance of estimation and inference based on the proposed median bias reduction method, mean BRPL, and ML.

Chapter 3

Mean bias reduction through simulation-based adjusted score equations

3.1 Introduction

In this chapter we consider variants of the adjusted score function proposed in Firth (1993) to reduce mean bias of the ML estimator regardless of the feasibility of the bias function, where the term "feasibility" refers to the possibility of a function to be calculated. Specifically, we show that solving an adjusted score equation where the bias function is replaced by its simulation-based estimate, also leads to estimators with $o(n^{-1})$ bias. Moreover, we introduce the "iterated bootstrap with likelihood adjustment" (IBLA) algorithm, which can be used for the computation of the bias-reduced estimates.

Additionally, this chapter provides the implementation of IBLA algorithm in the context of generalised linear models (McCullagh & Nelder, 1989). We choose generalised linear models to evaluate the performance of IBLA because they comprise an important class of statistical models; they extend linear models to encompass non-normal response distributions and modelling functions of the mean. Moreover, the likelihood function of generalised linear models can be written in closed form, which allows the easy comparison of IBLA to the traditional adjusted score equations method (Firth, 1993). IBLA is also compared with ML and parametric bootstrap (Efron & Tibshirani, 1993, Chapter 10) methods. Under the parametric bootstrap framework R bootstrap samples are generated from the model using the estimated parameter values. For each of the R bootstrap samples the parameter θ is estimated. The bias of an estimator $\hat{\theta}_n$ is then estimated as $B^{(\mathrm{boot})} = \bar{\theta} - \hat{\theta}_n$, where $\bar{\theta}$ is the average of the estimates based on each of the R bootstrap samples. The parametric bootstrap bias-reduced estimate is calculated as $\hat{\theta}_{\mathrm{boot}} = \hat{\theta}_n - B^{(\mathrm{boot})} = 2\hat{\theta}_n - \bar{\theta}$.

The evaluation of ML, adjusted score equations (Firth, 1993), parametric bootstrap, and IBLA is performed via simulation studies and real data examples. The first simulation study considers data from a continuous probability distribution, which satisfies the continuity condition assumed in Section 3.3, whereas the second simulation study considers data from a discrete probability distribution. Specifically, in the former simulation study we use a generalised linear model with the log link function and gamma distributed data and in the latter study we use a generalised linear model with the logistic link function and binary data. Both simulation studies are designed to compare the finite sample properties of the IBLA estimates and the associated inferential procedures.

Finally, we apply IBLA on a real-data example. The endometrial cancer grade study (Heinze & Schemper, 2002) illustrates a binary-response logistic regression analysis for which one parameter has an infinite ML estimate. The adjusted score equations do not depend on the finiteness of the ML estimates and yield finite estimates (Firth, 1993). On the contrary, the parametric bootstrap estimates are by definition undefined when the ML estimates are infinite. In this chapter we demonstrate how adjusting suitably the simulated samples involved in the simulation-based adjusted score equations results in finite IBLA estimates even when the ML estimates are infinite with positive probability.

3.2 Simulation-based adjusted score function

Firth (1993) shows that bias reduced estimates can be obtained by solving $s_n(\theta;y) - j_n(\theta;y)b_n(\theta) = 0$. In the latter equation, $s_n(\theta;y)$ is the score function, $j_n(\theta;y)$ is the observed information matrix, and $b_n(\theta)$ is the first-order term in the expansion of the bias of the ML estimator $\hat{\theta}_n$, as defined in Section 1.3, where the use of the subindex highlights the dependence of these quantities on n.

More generally, the theory in Firth (1993) and Kosmidis & Firth (2009) guarantees that estimators with $o(n^{-1})$ bias result by the solution of the equation

$$s_n^*(\theta;y) = s_n(\theta;y) - j_n(\theta;y)B_n(\theta) + v_n(\theta;y) = 0\,, \tag{3.1}$$

where $B_n(\theta) = E_\theta(\hat{\theta}_n - \theta)$ is the bias of $\hat{\theta}_n$ and $v_n(\theta;y) = O_p(n^{-1/2})$. There are cases where it is difficult to analytically evaluate $B_n(\theta)$ and hence equation (3.1) is difficult to be solved. For this reason, we propose to replace $B_n(\theta)$ in (3.1) by its simulation-based estimate, which makes the adjusted score equation feasible.

Definition 1. Let $\hat{B}_{n,R}(\theta) = (1/R)\sum_{r=1}^{R} \hat{\theta}_n(Z_r) - \theta$, where $\hat{\theta}_n(Z_r)$ is the solution of $s_n(\theta;Z_r) = 0$, and $Z_r = z(\theta;\Xi_r)$ is a sample of responses simulated from the model at θ, based on $\Xi_1,\ldots,\Xi_R$ independent copies of a random variable Ξ that does not depend on θ. Suppressing the dependence of $s_{n,R}^*(\theta;y)$, $s_n(\theta;y)$ and $j_n(\theta;y)$ on y, the simulation-based adjusted score function is expressed as

$$s_{n,R}^*(\theta) = s_n(\theta) - j_n(\theta)\hat{B}_{n,R}(\theta)\,. \tag{3.2}$$

3.3 Asymptotic properties

In this section, we consider the asymptotic properties of the estimator obtained from the equation $s_{n,R}^*(\theta) = 0_p$, and we also give some guidance of what values of R guarantee a reduction of bias in terms of n. The consistency and asymptotic normality of a sequence $\hat{\theta}_{n,R}^*$ of roots of $s_{n,R}^*(\theta)$ is shown using the results listed in Appendix E and under the following conditions.

Condition 1. The parameter space Θ is a compact subset of $\Re^p$.

Condition 2. $s_n(\theta)$ and $j_n(\theta)$ are continuous functions of θ.

Condition 3. $s_n^*(\theta)$ has a unique zero at $\hat{\theta}_n^* \in \Theta$.

Condition 4. $s_{n,R}^*(\theta)$ is continuously differentiable for all θ in a neighbourhood of the true unknown θ_0, and the matrix $H_{n,R}^*(\theta)$ with rows $\partial s_{n,R}^*(\theta)/\partial\theta_j$, $j \in \{1,\ldots,p\}$ is nonsingular.

Condition 5. For all $\theta \in \Theta$, $i \in \{1,\ldots,n\}$ and $\{j,k\} \in \{1,\ldots,p\}$,

$$E\left(\frac{\partial^2 \log f_i(y_i;\theta)}{\partial\theta_j\partial\theta_k}\right) \quad \text{and} \quad \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E\left(\frac{\partial^2 \log f_i(y_i;\theta)}{\partial\theta_j\partial\theta_k}\right)$$

exist and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_i(y_i;\theta)}{\partial\theta_j\partial\theta_k} \xrightarrow{p} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E\left(\frac{\partial^2 \log f_i(y_i;\theta)}{\partial\theta_j\partial\theta_k}\right).$$

Condition 6. The matrix $\bar{F}(\theta_0) = \left(\bar{F}_{jk}(\theta_0)\right)_{1\le j,k\le p}$, where

$$\bar{F}_{jk}(\theta_0) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E\left(-\frac{\partial^2 \log f_i(y_i;\theta)}{\partial\theta_j\partial\theta_k}\bigg|_{\theta_0}\right),$$

is positive definite.

Condition 7. As n and R go to infinity, $n^{-1}[-H^*_{n,R}(\theta)] \xrightarrow{p} \bar{F}(\theta)$ for θ in a neighbourhood of θ_0.

Conditions 1-3 are used to show the consistency of $\hat{\theta}^*_{n,R}$ and the extra conditions 4-7 are useful to derive the asymptotic distribution of $\hat{\theta}^*_{n,R}$. The weak law of large numbers (Davison, 2003, p. 28) gives sufficient conditions for the convergence in probability in condition 5. Condition 5 also ensures that $\bar{F}(\theta_0)$ in condition 6 exists. Condition 7 is justified if $[H^*_n(\theta) - H^*_{n,R}(\theta)]/n$ converges to zero, where $H^*_n(\theta)$ is the derivative of $s^*_n(\theta)$ with respect to θ. The columns of $[H^*_n(\theta) - H^*_{n,R}(\theta)]/n$ are

$$\frac{\partial}{\partial\theta_j}\left\{\frac{j_n(\theta)[\hat{B}_{n,R}(\theta) - B_n(\theta)]}{n}\right\}.$$

Given $n^{-1}j_n(\theta)[\hat{B}_{n,R}(\theta) - B_n(\theta)]$ is $O_p(n^{-1}R^{-1/2})$ and assuming the higher order derivatives of an $O_p(1)$ term are also $O_p(1)$, we have $n^{-1}[H^*_n(\theta) - H^*_{n,R}(\theta)] = O_p(n^{-1}R^{-1/2}) = o_P(1)$. Hence $n^{-1}[H^*_n(\theta) - H^*_{n,R}(\theta)] \xrightarrow{p} 0$ as $R \to \infty$.

Theorem 2 shows that with probability one the simulation-based adjusted score function $s^*_{n,R}(\theta)$ converges to the adjusted score function $s^*_n(\theta)$ uniformly in θ, as $R \to \infty$, and Corollary 1 shows that uniform convergence to the adjusted score function implies convergence of the simulation-based bias reduced estimator $\hat{\theta}^*_{n,R}$ to the bias reduced estimator $\hat{\theta}_n^*$ proposed in Firth (1993).

Theorem 2. *If conditions 1 and 2 are satisfied, then $s_n^*(\theta)$ and $s_{n,R}^*(\theta)$ are such that $\sup_{\theta\in\Theta}\|s_{n,R}^*(\theta)-s_n^*(\theta)\| \xrightarrow{p} 0$ as $R\to\infty$.*

Proof of Theorem 2: Let $s_{n,R}^*(\theta)$ be written as the average of R functions, such that $s_{n,R}^*(\theta) = (1/R)\sum_{r=1}^{R}\{s_n(\theta) - j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)\}$. Van der Vaart (2000, Theorem 5.9) gives a set of sufficient conditions for uniform convergence of functions that can be written in the form of an average, according to which we need Θ to be a compact space, $s_n(\theta) - j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)$ to be continuous for every θ, and $s_n(\theta) - j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)$ to be dominated by an integrable function. Condition 1 covers for compactness, and $s_n(\theta) - j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)$ is continuous as the sum of continuous functions. From the triangle inequality, $s_n(\theta) - j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)$ is bounded on Θ because there exists a positive number $K_n(\theta) = \|s_n(\theta)\| + \|j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)\|$ such that $\|s_n(\theta) - j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)\| \leq K_n(\theta)$. In order to show that $K_n(\theta)$ is integrable we need to show that it is continuous on a rectangle in $\Re^p$ (Trench, 2003, Theorem 7.1.13). The space Θ is compact, which is equivalent by the Heine-Borel theorem (Rudin, 1976, pp. 39-40) to Θ being closed and bounded. Then Θ is a closed subset of a rectangle that is a product of bounded intervals (Lavrent'ev & Savel'ev, 2006, p. 165). Also, the function $K_n(\theta)$ is continuous as it is the sum of vector norms. Thus $K_n(\theta)$ is integrable. □

Remark 1. Theorem 2 does not cover the case of discrete-response models. This is because for these models the continuity condition for $s_n(\theta) - j_n(\theta)(\hat{\theta}_n(Z_r) - \theta)$, which is one of the sufficient conditions for uniform convergence (see Van der Vaart, 2000, p. 46, and proof of Theorem 2), is not valid. The continuity condition is not valid, because the sample of responses $Z_1, \ldots, Z_R$ simulated from the model at θ are not continuous in terms of θ. However, even though formally our theory does not cover this case, simulation studies in Section 3.6 demonstrate that the simulation-based adjusted score equation approach behaves well.

Corollary 1. *If conditions 1-3 are satisfied and $s_{n,R}^*(\theta)$ converges uniformly to $s_n^*(\theta)$ as $R\to\infty$, then any $\hat{\theta}_{n,R}^* \in \Theta$ such that $s_{n,R}^*(\hat{\theta}_{n,R}^*) = 0$ converges in probability to $\hat{\theta}_n^*$ as $R\to\infty$.*

Proof of Corollary 1: In Theorem 2 we established uniform convergence of $s^*_{n,R}(\theta)$ to $s^*_n(\theta)$ as $R \to \infty$. Then for every $\varepsilon > 0$, there exists $M > 0$ such that for $R > M$, $\varepsilon > \sup_{\theta\in\Theta} \|s^*_{n,R}(\theta) - s^*_n(\theta)\| \geq \|s^*_{n,R}(\hat{\theta}^*_{n,R}) - s^*_n(\hat{\theta}^*_{n,R})\| = \|s^*_n(\hat{\theta}^*_{n,R})\|$. So the sequence $\{\hat{\theta}^*_{n,R}\}$ will converge to the unique $\hat{\theta}^*_n$. □

Having established the consistency of $\hat{\theta}^*_{n,R}$ as an estimator of $\hat{\theta}^*_n$ as $R \to \infty$, we can proceed, under some additional conditions, to prove asymptotic normality for $n^{1/2}(\hat{\theta}^*_{n,R} - \theta_0)$.

Theorem 3. *If conditions 1-7 are satisfied, the observations are independent and identically distributed, and the number of Monte Carlo samples R is fixed with n* $\to \infty$*, then* $n^{1/2}(\hat{\theta}^*_{n,R} - \theta_0)$ *is asymptotically normally distributed with zero mean and variance-covariance matrix* $(1+R^{-1})\{E[j_i(\theta_0)]\}^{-1}$*, where* $j_i(\theta)$ *is the observed information matrix for the ith observation.*

Proof of Theorem 3: Because $\hat{\theta}^*_{n,R}$ is a consistent estimator of the true parameter θ_0 as n and R go to infinity, it makes sense to expand $s^*_{n,R}(\theta)$ in a Taylor series around θ_0. Application of Taylor's theorem to $s^*_{n,R}(\theta)$ about its solution $\hat{\theta}^*_{n,R}$ gives $0 = s^*_{n,R}(\theta_0) + \nabla s^*_{n,R}(\breve{\theta})(\hat{\theta}^*_{n,R} - \theta_0)$, where $\breve{\theta} = \theta_0 + t(\hat{\theta}^*_{n,R} - \theta_0)$, with $t \in (0,1)$. Thus

$$n^{1/2}(\hat{\theta}^*_{n,R} - \theta_0) \quad = \quad \left\{-\frac{\nabla s^*_{n,R}(\breve{\theta})}{n}\right\}^{-1} \frac{s^*_{n,R}(\theta_0)}{n^{1/2}}.$$

By the central limit theorem (Van der Vaart, 2000, Proposition 2.17) $n^{-1/2}s_n(\theta_0) \xrightarrow{d} N(0_p, E[j_i(\theta_0)])$ as $n \to \infty$. Again by the central limit theorem and for all $r \in \{1,\dots,R\}$ $n^{1/2}(\hat{\theta}_{n,r} - \theta_0) \xrightarrow{d} N(0_p, \{E[j_{i,r}(\theta_0)]\}^{-1}) = N(0_p, \{E[j_i(\theta_0)]\}^{-1})$ as $n \to \infty$ and R is fixed. Then because $\{n^{1/2}(\hat{\theta}_{n,r} - \theta_0)\}_{r=1}^R$ are independent we have the joint limit

$$\begin{bmatrix} n^{1/2}\left(\hat{\theta}_{n,1} - \theta_0\right) \\ n^{1/2}\left(\hat{\theta}_{n,2} - \theta_0\right) \\ \vdots \\ n^{1/2}\left(\hat{\theta}_{n,R} - \theta_0\right) \end{bmatrix} \xrightarrow{d} N(0_{pR}, D)$$

where D is a block diagonal matrix with main diagonal blocks the matrices

$\{E[j_i(\theta_0)]\}^{-1}$. In view of the joint convergence in distribution (joint for all elements of the vector above) the continuous mapping theorem (Van der Vaart, 2000, Theorem 2.3) gives

$$n^{1/2}\hat{B}_{n,R}(\theta_0) = \frac{1}{R}\sum_{r=1}^{R} n^{1/2}(\hat{\theta}_{n,r} - \theta_0) \xrightarrow{d} N\left(0_p, \frac{1}{R}\{E[j_i(\theta_0)]\}^{-1}\right).$$

Further, because $n^{-1/2}\sum_{i=1}^{n} s_i(\theta_0)$ and $(1/R)\sum_{r=1}^{R} n^{1/2}(\hat{\theta}_{n,r} - \theta_0)$ are independent we have the joint limit as $n \to \infty$

$$\begin{bmatrix} n^{-1/2}s_n(\theta_0) \\ n^{1/2}\hat{B}_{n,R}(\theta_0) \end{bmatrix} \xrightarrow{d} N\left(0_{2p}, \begin{matrix} E[j_i(\theta_0)] & 0_{p\times p} \\ 0_{p\times p} & \frac{1}{R}\{E[j_i(\theta_0)]\}^{-1} \end{matrix}\right).$$

In view of the above and the fact that by the weak law of large numbers (Davison, 2003, p. 28) $n^{-1}j_n(\theta_0) \xrightarrow{p} E[j_i(\theta_0)]$ as $n \to \infty$, we have that

$$n^{-1/2}s^*_{n,R}(\theta_0) = \frac{s_n(\theta_0)}{n^{1/2}} - \frac{j_n(\theta_0)}{n} n^{1/2}\hat{B}_{n,R}(\theta_0) \xrightarrow{d} N\left(0_p, \left(1+R^{-1}\right)E[j_i(\theta_0)]\right).$$

Under the assumption of independent and identically distributed observations the matrix $\bar{F}(\theta)$ in Condition 7 is $E[j_i(\theta)]$. Using this result, the consistency of $\breve{\theta}$, and Slutsky's Lemma (Van der Vaart, 2000, Lemma 2.8) we have $n^{1/2}(\hat{\theta}^*_{n,R} - \theta_0) \xrightarrow{d} N\left(0_p, \left(1+R^{-1}\right)\{E[j_i(\theta_0)]\}^{-1}\right)$. □

Lastly, Theorem 4 shows that, as long as the number of Monte Carlo samples R is $O(n^a)$ with $a \geq 1$, the estimator $\hat{\theta}^*_{n,R}$ has smaller bias than the ML estimator whose bias is of order $O(n^{-1})$.

Theorem 4. *Let* $R = O(n^a)$, $a \geq 1$. *Then* $E_\theta(\hat{\theta}^*_{n,R} - \theta_0) = O(n^{-3/2})$.

The proof of Theorem 4 is included in Appendix B.

3.4 Some examples of simulation-based bias-reduced estimators

Before giving the algorithm that can be used for computing the simulation-based bias reduced estimator $\hat{\theta}^*_{n,R}$, we give two examples that illustrate how the simulation-based adjusted score function $s^*_{n,R}(\theta)$ is constructed. In the first example we estimate the exponential distribution parameter, and in the second example we estimate the normal distribution parameters.

Example 1. (Exponential distribution parameter). Let the observations $y_1,\dots,y_n$ be realisations of the random variables $Y_1,\dots,Y_n$ that are assumed to be independent and exponentially distributed with mean $1/\lambda$. The score function for the parameter λ is $s_n(\lambda;y) = n/\lambda - n\bar{y}$, where $y = (y_1,\dots,y_n)^{\mathrm{T}}$, $\bar{y} = (1/n)\sum_{i=1}^n y_i$, and the ML estimator of λ is $\hat{\lambda}_n = 1/\bar{y}$. The bias function $B_n(\lambda) = E(\hat{\lambda}_n) - \lambda$ is calculated easily if we consider that $\sum_{i=1}^n y_i$ is a Gamma random variable with shape parameter n and rate parameter λ. Then $B_n(\lambda) = \lambda/(n-1)$ and thus $s^*_n(\lambda;y) = n(n-2)/[(n-1)\lambda] - n\bar{y}$. The bias reduced estimator that solves $s^*_n(\lambda;y) = 0$ is $\hat{\lambda}^*_n = [(n-2)/(n-1)](1/\bar{y})$. Now consider independent random variables Ξ_{ir} from the Uniform$(0,1)$ distribution, with $i \in \{1,\dots,n\}$, $r \in \{1,\dots,R\}$. The Monte Carlo estimate of the bias function is

$$\hat{B}_{n,R}(\lambda) = -\lambda\left(\frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1} + 1\right)$$

and the simulation-based adjusted score function is

$$s^*_{n,R}(\lambda) = \frac{2n}{\lambda} - n\bar{y} + \frac{n}{\lambda R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1}.$$

The simulation-based bias reduced estimator that solves $s^*_{n,R}(\lambda) = 0$ is

$$\hat{\lambda}^*_{n,R} = \frac{1}{\bar{y}}\left(\frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1} + 2\right).$$

The bias, variance, and mean squared error of the ML estimator, the bias reduced estimator obtained by solving the adjusted score function proposed in Firth (1993), and the simulation-based bias reduced estimator are given in Table 3.1. The two bias reduced estimators have the same bias of order $O(n^{-2})$. The variance of $\hat{\lambda}^*_{n,R}$ converges to the variance of $\hat{\lambda}^*_n$ as the Monte Carlo size $R \to \infty$. Note that in order to keep the mean squared error of $\hat{\lambda}^*_{n,R}$ smaller than the mean squared error of the ML estimator,

Table 3.1: Bias, variance, and mean squared error (MSE) of the three estimators of λ calculated in Example 1.

	$\hat{\lambda}_n$	$\hat{\lambda}_n^*$	$\hat{\lambda}_{n,R}^*$
Estimator	$\frac{1}{\bar{y}}$	$\frac{n-2}{n-1}\frac{1}{\bar{y}}$	$\left(\frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1}+2\right)\frac{1}{\bar{y}}$
Bias	$\frac{\lambda}{n-1}$	$-\frac{\lambda}{(n-1)^2}$	$-\frac{\lambda}{(n-1)^2}$
Variance	$\frac{n^2\lambda^2}{(n-1)^2(n-2)}$	$\frac{n^2(n-2)\lambda^2}{(n-1)^4}$	$\frac{n^2(n-2)\lambda^2}{(n-1)^4}+\frac{1}{R}\frac{n^4\lambda^2}{(n-1)^3(n-2)^2}$
MSE	$\frac{(n+2)\lambda^2}{(n-1)(n-2)}$	$\frac{(n^2-n-1)\lambda^2}{(n-1)^3}$	$\frac{(n^2-n-1)\lambda^2}{(n-1)^3}+\frac{1}{R}\frac{n^4\lambda^2}{(n-1)^3(n-2)^2}$

Notes: $\hat{\lambda}_n$, maximum likelihood estimator; $\hat{\lambda}_n^*$, bias reduced estimator that results from solving $s_n^*(\lambda)=0$; $\hat{\lambda}_{n,R}^*$, bias reduced estimator that results from solving $s_{n,R}^*(\lambda)=0$.

R should be greater than $n^3/[(3n-4)(n-2)] \approx n/3$. In this example, the parametric bootstrap estimate of λ is identical to $\hat{\lambda}_{n,R}^*$.

Let $\Xi \sim \text{Uniform}(0,1)$. Then $Z = -\sum_{i=1}^{n}\log\Xi_{ir} \sim \text{Gamma}(n,1)$ and

$$n\,E(1/Z) = n\int_0^\infty \frac{1}{z}\frac{z^{n-1}e^{-z}}{\Gamma(n)}dz = \frac{n}{n-1}\int_0^\infty \frac{z^{n-2}e^{-z}}{\Gamma(n-1)}dz = \frac{n}{n-1}\,.$$

By the weak law of large numbers (Davison, 2003, p. 28) we have that

$$\frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1} \xrightarrow{p} E\left(\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1}\right) = -\frac{n}{n-1}$$

as $R\to\infty$, and as a result $\hat{\lambda}_{n,R}^* \xrightarrow{p} \hat{\lambda}_n^*$ as $R\to\infty$. If we fix R and let $n\to\infty$ then $\hat{\lambda}_{n,R}^* \xrightarrow{p} \lambda$. This can be shown using the results

$$\frac{1}{\bar{y}} \xrightarrow{p} \lambda \quad \text{and} \quad \frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1} \xrightarrow{p} \frac{1}{R}\sum_{r=1}^{R}\left\{E\log\Xi_{ir}\right\}^{-1} = -1$$

as $n\to\infty$. Given $\hat{\lambda}_{n,R}^* \xrightarrow{p} \lambda$ as $n\to\infty$ and R is fixed, we can use Taylor's theorem to express $\sqrt{n}(\hat{\lambda}_{n,R}^*-\lambda_0)$ as

$$\begin{aligned}\sqrt{n}(\hat{\lambda}_{n,R}^*-\lambda_0) &= \left\{-\frac{H_{n,R}^*(\breve{\lambda})}{n}\right\}^{-1}\frac{s_{n,R}^*(\lambda_0)}{\sqrt{n}}, \ \text{where } \breve{\lambda} = \hat{\lambda}_{n,R}^*+t(\hat{\lambda}_{n,R}^*-\lambda_0)\,,\ t\in(0,1)\\ &= \left\{-\frac{H_{n,R}^*(\breve{\lambda})}{n}\right\}^{-1}\left\{\frac{s_n(\lambda_0)}{\sqrt{n}}-\frac{j_n(\lambda_0)}{n}\sqrt{n}\hat{B}_{n,R}(\lambda_0)\right\}.\end{aligned}$$

By Slutsky's Lemma (Van der Vaart, 2000, Lemma 2.8) and the following asymptotic results:

$$\begin{aligned}
-\frac{H^*_{n,R}(\breve{\lambda})}{n} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{2}{\breve{\lambda}^2}+\frac{1}{\breve{\lambda}^2}\frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1}\right]\xrightarrow{p}\frac{1}{\lambda_0^2}\\
\frac{j_n(\lambda_0)}{n} &= \frac{1}{\lambda_0^2}\rightarrow\frac{1}{\lambda_0^2}\\
\frac{s_n(\lambda_0)}{\sqrt{n}} &= \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\lambda_0}-y_i\right]\right\}\xrightarrow{d}N\left(0,\frac{1}{\lambda_0^2}\right)\\
\sqrt{n}\hat{B}_{n,R}(\lambda_0) &= -\sqrt{n}\lambda\left(\frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1}+1\right)\xrightarrow{d}N\left(0,\frac{\lambda_0^2}{R}\right)
\end{aligned}$$

we get that $\sqrt{n}(\hat{\lambda}^*_{n,R}-\lambda_0)\xrightarrow{d}N(0,(1+R^{-1})\lambda_0^2)$ as $n\rightarrow\infty$. The first result is obtained by the law of large numbers (Davison, 2003, p. 28) and the consistency of $\breve{\lambda}$, and the last two results are obtained by the central limit theorem (Van der Vaart, 2000, Proposition 2.17). In the above calculations we used

$$\begin{aligned}
E\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1}\right] &= -\frac{n}{n-1};\\
\mathrm{Var}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\log\Xi_{ir}\right\}^{-1}\right] &= \frac{n^2}{(n-1)^2(n-2)}.
\end{aligned}$$

Example 2. (Normal distribution parameters). Let the observations $y_1,\ldots,y_n$ be realisations of the random variables $Y_1,\ldots,Y_n$ that are assumed to be independent and normally distributed with unknown mean μ and unknown variance σ^2. The score function for the parameter $\theta=(\mu,\sigma^2)^{\mathrm{T}}$ is

$$s_n(\theta;y)=\begin{bmatrix}\frac{1}{\sigma^2}\sum\limits_{i=1}^{n}(y_i-\mu)\\ \frac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(y_i-\mu)^2-\frac{n}{2\sigma^2}\end{bmatrix}.$$

and the ML estimator of θ is $\hat{\theta}_n=\left(\bar{y},\ (1/n)\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{\mathrm{T}}$, where $\bar{y}=(1/n)\sum_{i=1}^{n}y_i$. The ML estimator $\hat{\mu}_n=\bar{y}$ is an unbiased estimator of μ because $E(\bar{y})=E(y_i)=\mu$. On the other hand, the ML estimator $\hat{\sigma}_n^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$ of σ^2 is biased because $E(\hat{\sigma}_n^2)=E(y_i^2)-E(\bar{y}^2)=(\sigma^2+\mu^2)-(\sigma^2/n+\mu^2)=\sigma^2(1-1/n)$. Then, the bias function is $B_n(\theta) = (0, -\sigma^2/n)^{\mathrm{T}}$, and thus

$$s_n^*(\theta, y) = s_n(\theta, y) - i_n(\theta, y) B_n(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} \sum\limits_{i=1}^{n} (y_i - \mu) \\ \frac{1}{2\sigma^4} \sum\limits_{i=1}^{n} (y_i - \mu)^2 - \frac{n}{2\sigma^2} + \frac{1}{2\sigma^2} \end{bmatrix},$$

where $i_n(\theta; y)$ is the expected information matrix. The bias reduced estimator proposed in Firth (1993) solves $s_n^*(\theta; y) = 0$ and is equal to $\hat{\theta}_n^* = \left(\bar{y},\ (1/(n-1)) \sum_{i=1}^n (y_i - \bar{y})^2\right)^{\mathrm{T}}$.

Now let Ξ_{ir} be independent random variables from the Uniform(0,1) distribution for $i \in \{1, \ldots, n\}$, $r \in \{1, \ldots, R\}$. Then $U_{ir} = \Phi^{-1}(\Xi_{ir})$ are independent random variables from the standard normal distribution, and $Z_{ir} = \mu + \sigma U_{ir}$ are normally distributed with $N(\mu, \sigma^2)$. The Monte Carlo estimate of the bias function is

$$\hat{B}_{n,R}(\theta) = \begin{bmatrix} \frac{1}{R} \sum\limits_{r=1}^{R} \left(\frac{1}{n} \sum\limits_{i=1}^{n} Z_{ir} \right) - \mu \\ \frac{1}{R} \sum\limits_{r=1}^{R} \left(\frac{1}{n} \sum\limits_{i=1}^{n} (Z_{ir} - \bar{Z}_r)^2 \right) - \sigma^2 \end{bmatrix}$$

and the simulation-based adjusted score function is

$$s_{n,R}^*(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} \sum\limits_{i=1}^{n} y_i - \frac{n\mu}{\sigma^2} - \frac{1}{\sigma R} \sum\limits_{r=1}^{R} \sum\limits_{i=1}^{n} U_{ir} \\ \frac{1}{2\sigma^4} \sum\limits_{i=1}^{n} (y_i - \mu)^2 - \frac{1}{2\sigma^2 R} \sum\limits_{r=1}^{R} \sum\limits_{i=1}^{n} (U_{ir} - \bar{U}_r)^2 \end{bmatrix}.$$

Solving $s_{n,R}^*(\theta) = 0$ gives

$$\hat{\theta}_{n,R}^* = \begin{bmatrix} \hat{\mu}_{n,R}^* \\ \hat{\sigma}_{n,R}^{2*} \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{\sqrt{\hat{\sigma}_{n,R}^{2*}}}{Rn} \sum\limits_{r=1}^{R} \sum\limits_{i=1}^{n} U_{ir} \\ \frac{1}{n} \sum\limits_{i=1}^{n} (y_i - \hat{\mu}_{n,R}^*)^2 \left\{ \frac{1}{Rn} \sum\limits_{r=1}^{R} \sum\limits_{i=1}^{n} (U_{ir} - \bar{U}_r)^2 \right\}^{-1} \end{bmatrix}.$$

A simpler form of the estimator $\hat{\sigma}_{n,R}^{2*}$ is obtained if we replace $\hat{\mu}_{n,R}^*$ by the unbiased estimator $\bar{y}$.

The bias, variance, and mean squared error of the ML estimator, the parametric bootstrap estimator, the bias reduced estimator proposed in Firth (1993), and the simulation-based bias reduced estimator are given in Table 3.2. The bias of $\hat{\sigma}_{n,R}^{2*}$ is smaller than the bias of the bootstrap estimator for all $R > 2(n^2+1)/(n-1) \approx 2(n+1)$.

Table 3.2: Bias, variance, and mean squared error (MSE) of the four estimators of σ^2 calculated in Example 2.

	$\hat{\sigma}^2_n$	$\hat{\sigma}^2_{\text{boot}}$
Estimator	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\left\{2-\frac{1}{Rn}\sum_{r=1}^{R}\sum_{i=1}^{n}(U_{ir}-\bar{U}_r)^2\right\}$
Bias	$-\frac{\sigma^2}{n}$	$-\frac{\sigma^2}{n^2}$
Variance	$2\sigma^4\left(\frac{n-1}{n^2}\right)$	$\frac{2(n-1)(n+1)\sigma^4}{n^4}\left\{n+1+\frac{(2n-1)(n-1)}{Rn^2}\right\}$
MSE	$\sigma^4\left(\frac{2n-1}{n^2}\right)$	$\frac{2(n-1)(n+1)\sigma^4}{n^4}\left\{n+1+\frac{(2n-1)(n-1)}{Rn^2}\right\}+\frac{\sigma^4}{n^4}$

	$\hat{\sigma}^{2*}_n$	$\hat{\sigma}^{2*}_{n,R}$
Estimator	$\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\left\{\frac{1}{Rn}\sum_{r=1}^{R}\sum_{i=1}^{n}(U_{ir}-\bar{U}_r)^2\right\}^{-1}$
Bias	0	$\frac{2\sigma^2}{R(n-1)-2}$
Variance	$2\sigma^4\left(\frac{1}{n-1}\right)$	$2\sigma^4\left(\frac{R^2(n-1)[R(n-1)+n-3]}{[R(n-1)-2]^2[R(n-1)-4]}\right)$
MSE	$\sigma^4\left(\frac{2}{n-1}\right)$	$\sigma^4\left\{\frac{2}{[R(n-1)-2]^2}\left[2+\frac{R^2(n-1)[R(n-1)+n-3]}{R(n-1)-4}\right]\right\}$

Notes: $\hat{\sigma}^2_n$, maximum likelihood estimator; $\hat{\sigma}^2_{\text{boot}}$, parametric bootstrap estimator; $\hat{\sigma}^{2*}_n$, bias reduced estimator that results from solving $s^*_n(\theta)=0$; $\hat{\sigma}^{2*}_{n,R}$, bias reduced estimator that results from solving $s^*_{n,R}(\theta)=0$.

From the results in Table 3.2 we also notice that in order for $\hat{\sigma}^{2*}_{n,R}$ to be $o(n^{-1})$, R must be $O(n)$. The $\hat{\sigma}^{2*}_{n,R}$ estimator converges to $\hat{\sigma}^{2*}_n$ in probability for fixed n as $R\to\infty$. This can be shown using the weak law of large numbers (Davison, 2003, p. 28) and the fact that $E[(U_{ir}-\bar{U}_r)^2]=1-n^{-1}$.

If we fix R and let $n\to\infty$ then we can show that $\hat{\sigma}^{2*}_{n,R}\to\sigma^2$ by using the asymptotic results $n^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\xrightarrow{p}\sigma^2$ and $n^{-1}\sum_{i=1}^{n}\left\{R^{-1}\sum_{r=1}^{R}(U_{ir}-\bar{U}_r)^2\right\}\xrightarrow{p}1$ which are based on $E[(y_i-\bar{y})^2]=\sigma^2-\sigma^2/n\to\sigma^2$ and $E[(U_{ir}-\bar{U}_r)^2]=1-1/n\to1$. Given $\hat{\sigma}^{2*}_{n,R}\to\sigma^2$ as $n\to\infty$ and R is fixed, we use Taylor's theorem to express $\hat{\theta}^*_{n,R}-\theta_0$ as

$$\sqrt{n}(\hat{\theta}^*_{n,R}-\theta_0) \quad = \quad \left\{-\frac{H^*_{n,R}(\breve{\theta})}{n}\right\}^{-1}\frac{s^*_{n,R}(\theta_0)}{\sqrt{n}},$$

where $\breve{\theta}=\hat{\theta}^*_{n,R}+t(\hat{\theta}^*_{n,R}-\theta_0)$, $t\in(0,1)$. By the weak law of large numbers (Davison, 2003, p. 28)

$$-\frac{H^*_{n,R}(\breve{\theta})}{n}=\frac{1}{n}\begin{bmatrix}\sum_{i=1}^{n}\frac{1}{\sigma^2} & \sum_{i=1}^{n}\left\{\frac{y_i-\mu}{\sigma^4}-\frac{1}{2R\sigma^{3/2}}\sum_{r=1}^{R}U_{ir}\right\}\\ \sum_{i=1}^{n}\frac{y_i-\mu}{\sigma^4} & \sum_{i=1}^{n}\left\{\frac{(y_i-\mu)^2}{\sigma^6}-\frac{1}{2R\sigma^4}\sum_{r=1}^{R}(U_{ir}-\bar{U}_r)^2\right\}\end{bmatrix}\xrightarrow{p}\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}^{-1} \tag{3.3}$$

and by the central limit theorem (Van der Vaart, 2000, Proposition 2.17)

$$\begin{aligned}\frac{s^*_{n,R}(\theta)}{\sqrt{n}} &= \sqrt{n}\left\{\frac{1}{n}\begin{bmatrix}\sum_{i=1}^{n}\left\{\frac{y_i-\mu}{\sigma^2}-\frac{1}{\sigma R}\sum_{r=1}^{R}U_{ir}\right\}\\ \sum_{i=1}^{n}\left\{\frac{(y_i-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2 R}\sum_{r=1}^{R}(U_{ir}-\bar{U}_r)^2\right\}\end{bmatrix}\right\}\\ &\xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\frac{1}{\sigma^2}\left(1+\frac{1}{R}\right) & 0\\ 0 & \frac{1}{2\sigma^4}\left(1+\frac{1}{R}\right)\end{bmatrix}\right) \qquad (3.4)\end{aligned}$$

as $n \to \infty$. Then by Slutsky's Lemma (Van der Vaart, 2000, Lemma 2.8)

$$\sqrt{n}(\hat{\theta}^*_{n,R}-\theta_0)\xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix},\left(1+\frac{1}{R}\right)\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}\right).$$

In order to get (3.3) and (3.4) we used the following results

$$\begin{aligned}E\left[\frac{y_i-\mu}{\sigma^2}-\frac{1}{\sigma R}\sum_{r=1}^{R}U_{ir}\right] &= 0,\\ E\left[\frac{(y_i-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2 R}\sum_{r=1}^{R}(U_{ir}-\bar{U}_r)^2\right] &= \frac{1}{2\sigma^2}-\frac{1}{2\sigma^2}\left(1-\frac{1}{n}\right)=\frac{1}{2\sigma^2 n}\to 0,\\ \mathrm{Var}\left[\frac{y_i-\mu}{\sigma^2}-\frac{1}{\sigma R}\sum_{r=1}^{R}U_{ir}\right] &= \frac{1}{\sigma^4}\mathrm{Var}(y_i)+\frac{1}{\sigma^2 R}\mathrm{Var}(U_{ir})=\frac{1}{\sigma^2}\left(1+\frac{1}{R}\right),\\ \mathrm{Var}\left[\frac{(y_i-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2 R}\sum_{r=1}^{R}(U_{ir}-\bar{U}_r)^2\right] &= \frac{1}{2\sigma^4}\left(1+\frac{1}{R}\right),\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Cov}\left(\frac{y_i-\mu}{\sigma^2}-\frac{1}{\sigma R}\sum_{r=1}^{R}U_{ir},\frac{(y_i-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2 R}\sum_{r=1}^{R}(U_{ir}-\bar{U}_r)^2\right)=\\ &= \frac{1}{2\sigma^6}\mathrm{Cov}\left(y_i,(y_i-\mu)^2\right)+\frac{1}{2\sigma^3 R^2}\mathrm{Cov}\left(\sum_{r=1}^{R}U_{ir},\sum_{r=1}^{R}(U_{ir}-\bar{U}_r)^2\right)\\ &= \frac{1}{2\sigma^6}\mathrm{Cov}\left(y_i,y_i^2\right)-\frac{\mu}{\sigma^6}\mathrm{Var}(y_i)+\frac{1}{2\sigma^3 R}\mathrm{Cov}\left(U_{ir},(U_{ir}-\bar{U}_r)^2\right)\\ &= \frac{1}{2\sigma^6}2\mu\sigma^2-\frac{\mu}{\sigma^6}\sigma^2+\frac{1}{2\sigma^3 R}\mathrm{Cov}\left(U_{ir},(U_{ir}-\bar{U}_r)^2\right)\\ &= \frac{1}{2\sigma^3 R}\mathrm{Cov}\left(U_{ir},(U_{ir}-\bar{U}_r)^2\right)\\ &\to \frac{1}{2\sigma^3 R}\mathrm{Cov}\left(U_{ir},U_{ir}^2\right)=0\quad\text{as } n\to\infty.\end{aligned}$$

As a general observation from Examples 1 and 2 we notice that both examples suggest R should be $O(n)$ in order to achieve bias reduction when solving the equation $s^*_{n,R}(\theta) = 0_p$.

3.5 Iterated bootstrap with likelihood adjustment

A direct approach for computing the simulation-based bias-reduced estimator $\hat{\theta}^*_{n,R}$, the solution of $s^*_{n,R}(\theta) = s_n(\theta) - j_n(\theta)\hat{B}_{n,R}(\theta) = 0$, is through a quasi Newton-Raphson iteration. Specifically, $\hat{\theta}^*_{n,R}$ is obtained through an iteration of the form

$$\theta_n^{(j+1)} = (2\theta_n^{(j)} - \bar{\theta}_{n,R}^{(j)}) + \{j_n(\theta_n^{(j)})\}^{-1} s_n(\theta_n^{(j)}), \tag{3.5}$$

where $\theta_n^{(j)}$ is the candidate value for $\hat{\theta}^*_{n,R}$ at the jth iteration, and $\bar{\theta}_{n,R}^{(j)}$ is the average of the ML estimates calculated for each of R simulated samples from the model at $\theta_n^{(j)}$. Starting from the ML estimate, a single iteration gives the parametric bootstrap corrected estimate, so iteration (3.5) can be seen as a generalisation of the bootstrap for bias correction (Efron & Tibshirani, 1993, Chapter 10). The extra term $\{j_n(\theta_n^{(j)})\}^{-1} s_n(\theta_n^{(j)})$ is the reason we refer to the proposed bias reduction method as *iterated bootstrap with likelihood adjustment* (IBLA).

A stopping criterion for the iterations is the absolute difference of two consecutive candidate values for $\hat{\theta}^*_{n,R}$ evaluated at each of the parameters is less than some prespecified $\varepsilon > 0$. Based on practical experimentation, IBLA reaches the neighbourhood of the solution of the simulation-based adjusted score equation quickly, and then varies in that neighbourhood. In all our simulations we assume the algorithm converges when $|\theta_n^{(j+1)} - \theta_n^{(j)}| < 10^{-6}$. We recommend using the same initial state for the random number generator in each iteration in order to achieve a smooth estimator of the bias function.

3.6 Bias reduction in generalised linear models

3.6.1 Generalised linear model

In a generalised linear model the observations $y_1, \ldots, y_n$ are assumed to be realisations of the independent random variables $Y_1, \ldots, Y_n$, respectively, with Y_i having a distribution in the exponential family (McCullagh & Nelder, 1989). The exponential family has probability density function or mass function of the form

$$f_{Y_i}(y_i;\gamma_i,\phi) = \exp[\phi^{-1}\{y_i\gamma_i - b(\gamma_i)\} + c(y_i,\phi)], \qquad (3.6)$$

where $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions, and γ_i and $\phi > 0$ are possibly unknown parameters. In this setting, we have $E(Y_i) = \mu_i = \partial b(\gamma_i)/\partial\gamma_i$ and $\text{Var}(Y_i) = \phi v_i$, where ϕ is the dispersion parameter and $v_i = v(\mu_i) = \partial\mu_i/\partial\gamma_i$ is the variance function. The natural parameter $\gamma = (\gamma_1,\ldots,\gamma_n)^{\text{T}}$ is a strictly monotonic function of $\mu = (\mu_1,\ldots,\mu_n)^{\text{T}}$.

A generalised linear model is defined by (3.6) and by a one-to-one twice differentiable link function $g(\mu) = \eta$, relating the mean μ of the probability distribution of $Y = (Y_1,\ldots,Y_n)^{\text{T}}$ to the linear predictor of the model $\eta = (\eta_1,\ldots,\eta_n)^{\text{T}}$. The linear predictor is expressed as $\eta = X\beta$, where X is the $n \times p$ known design matrix of rank p, and $\beta = (\beta_1,\ldots,\beta_p)^{\text{T}}$ is the vector of unknown parameters to be estimated.

In general, the link function allows μ to be non-linearly related to the predictors, and is used to map μ onto the real line allowing the parameters to take any value on the Euclidean space without violating the possibly bounded range of μ implied by the model. This enables the modelling of non-normally distributed responses, such as categorical data and count data, without having to transform the data. For example, log-linear models use the log link function $g(\mu) = \log\mu$ which is appropriate when μ cannot be negative, such as with count data. Another important generalised linear model is the logistic regression model, which models the log of an odds via the logit link function $g(\mu) = \log[\mu/(1-\mu)]$. The logit link function is appropriate when μ is between 0 and 1, such as a probability.

3.6.2 Gamma-response log-linear model

In this simulation study we evaluate the ML, adjusted score equations (Firth, 1993), parametric bootstrap, and IBLA methods using data from the Gamma distribution, a probability distribution which satisfies the continuity condition assumed in Section 3.3.

Specifically, we simulated 10 000 samples from a log-linear model with linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i \in \{1,\ldots,n\}$, and the observations are assumed to be realisations of independent Gamma distributed random variables with shape parameter $1/\phi$

Table 3.3: Bias, mean squared error (MSE), and empirical coverage probability of Wald-type confidence intervals for the parameters of the gamma-response log-linear model with $n = 10$ and true parameter values $\beta_0 = 0.5$, $\beta_1 = 1.2$, $\phi = 0.25$.

			β_0			β_1			ϕ	
Method	R	Bias	MSE	Coverage	Bias	MSE	Coverage	Bias	MSE	Coverage
ML	-	-0.025	0.028	0.880	0.008	0.086	0.864	-0.051	0.012	0.721
mean BR	-	-0.001	0.027	0.912	-0.008	0.086	0.903	-0.003	0.014	0.826
BOOT	50	0.024	0.028	0.897	-0.002	0.086	0.890	-0.031	0.012	0.808
	100	0.018	0.028	0.896	0.012	0.086	0.886	-0.033	0.012	0.786
	500	0.002	0.028	0.908	-0.007	0.086	0.897	-0.010	0.013	0.817
IBLA	50	0.039	0.029	0.907	-0.009	0.086	0.900	-0.013	0.013	0.822
	100	0.030	0.029	0.905	0.005	0.086	0.896	-0.017	0.013	0.829
	500	0.009	0.028	0.912	-0.009	0.086	0.905	0.001	0.015	0.837

Notes: The coverage probabilities correspond to nominally 95% confidence intervals.

and scale parameter $\phi\mu$. The true parameter values were set to $\beta_0 = 0.5$, $\beta_1 = 1.2$, $\phi = 0.25$, and the covariate was simulated from a uniform distribution $U[-1,1]$. The sample size was set to $n = 10$. For each simulated sample we estimated the parameter $\theta = (\beta_0, \beta_1, \phi)^{\mathrm{T}}$.

A summary of the simulations is given in Table 3.3. The results indicate that in order for the parametric bootstrap and IBLA methods to reduce the bias of the ML estimators, R must be large compared to the sample size. The empirical coverage of the confidence intervals is smaller than the nominal 95% level for all statistics, but the confidence intervals obtained using the mean BR estimates obtained from solving the traditional adjusted score equations, the parametric bootstrap, and IBLA estimates have empirical coverage that is relatively closer to the nominal level.

Figure 3.1 shows how the Monte Carlo estimate of the bias function of the dispersion parameter behaves for each value of R used in the simulation setting. As R gets larger, the Monte Carlo estimate of the bias function becomes smoother and approaches the first-order term of the bias of the ML estimator used in the adjusted score equation proposed in Firth (1993).

3.6.3 Logistic regression

The generalised linear model that uses the logit link function, i.e.

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n$$

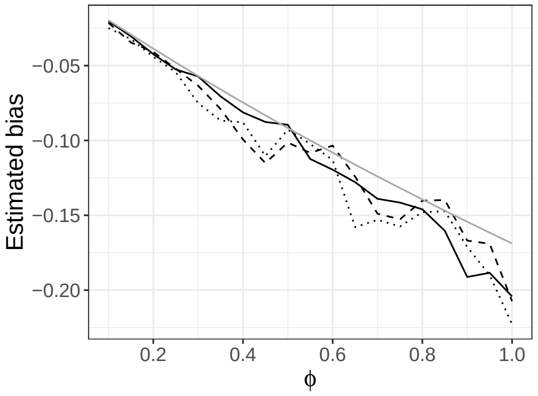

Figure 3.1: Monte Carlo estimate of bias of the ML estimator of ϕ in a gamma-response log-linear model with $n = 10$ and $\beta_0 = 0.5$, $\beta_1 = 1.2$. The curves correspond to Monte Carlo size R equal to 50 (dotted), 100 (dashed), and 500 (solid). The grey line is the first-order term of the bias of the ML estimator of ϕ.

is referred to as logistic regression. This is the most important model for binary-response data, where the observations $y_1, \ldots, y_n$ are assumed to be realisations of independent Bernoulli distributed random variables $Y_1, \ldots, Y_n$ with probability $\pi_1, \ldots, \pi_n$.

In binary-response logistic regression models the ML estimates of the parameters can be infinite, which happens when a hyperplane separates the set of explanatory variable values having $y = 0$ from the set having $y = 1$ (Albert & Anderson, 1984). In this case the space of explanatory variable values is said to have complete separation (Agresti, 2015, Chapter 5.4.2). The bias-reduced estimates obtained from the adjusted score function approach do not depend upon the finiteness of the ML estimates (Firth, 1993). Other bias reduced estimates, such as the parametric bootstrap estimates, are undefined when the ML estimates are infinite.

By definition, the IBLA estimates also depend on the finiteness of the ML estimates through the simulation-based estimator of the bias function. The use of the trimmed mean or the median instead of the mean of the R parameter estimates when estimating the bias of $\hat{\theta}_n$ can reduce, but not eliminate, the possibility of infinite IBLA estimates. This is because there are cases in which more than half bootstrap samples give infinite ML estimates. For this reason, we propose modifying at each iteration the simulated binary responses y to $y^c = c + y(1 - 2c)$, where c is a small positive constant. By this simple modification of the algorithm we eliminate the possibility of infinite IBLA estimates. In our simulations we set $c = 10^{-8}$.

3.6.3.1 Simulation study

In this simulation study we simulated 10 000 samples from a logistic regression model with linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i \in \{1,\ldots,n\}$, in order to evaluate the performance of ML, adjusted score equations (Firth, 1993), parametric bootstrap, and IBLA methods using data that do not satisfy the continuity condition assumed in Section 3.3. The parameters of interest $\theta = (\beta_0, \beta_1)$ were set to $\beta_0 = 1$, $\beta_1 = 1.5$, and the covariate was simulated from a standard normal distribution. The parameter values and the model matrix were obtained from Siino et al. (2016). The sample size was set to $n = 60, 120$ and 240.

The ML estimator $\hat{\theta}_n$ of the binary logistic model solves $s_n(\theta) = \sum_{i=1}^{n} (y_i - \pi_i) x_{ij} = 0$ and the mean bias-reducing estimator $\hat{\theta}_n^*$ proposed in Firth (1993) solves $s_n^*(\theta) = \sum_{i=1}^{n} (y_i + h_i/2 - h_i \pi_i - \pi_i) x_{ij} = 0$, where $j \in \{1,\ldots,p\}$, and h_i is the leverage for the ith observation (Firth, 1992). The leverage is defined as the ith diagonal element of the "hat" matrix $WX(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}$, with $W = \phi^{-1}\mathrm{diag}\{\kappa_{2i}\}$ and κ_{2i} being the variance of the ith observation. The Bernoulli distribution has $\phi = 1$ and $\kappa_{2i} = \pi_i(1-\pi_i)$.

The number of Monte Carlo samples R used to calculate the simulation-based estimate of the bias function in $s_{n,R}^*(\theta)$ was set to be proportional to n. We denote by $\hat{\theta}_{n,R_1}^*$ the IBLA estimates obtained when the Monte Carlo size is $R_1 = n$. Similarly, $\hat{\theta}_{n,R_2}^*$, $\hat{\theta}_{n,R_3}^*$, and $\hat{\theta}_{n,R_4}^*$ correspond to the IBLA estimates obtained when the Monte Carlo size is $R_2 = 2n$, $R_3 = 3n$, and $R_4 = 4n$, respectively. The IBLA estimates were calculated by using the adjusted simulated samples y^c in each iteration of the algorithm.

The results of the simulation study are summarised in Table 3.4. We observe that for all n ML yields the largest bias and mean squared error. The IBLA approach is in all cases among the best two methods in terms of bias and mean squared error. Regarding coverage properties, all fitting methods perform equally well for moderate and large sample sizes, giving coverage probabilities close to the nominal 95% level. When $n = 60$ the estimated coverage probabilities are slightly larger than the nominal level. In general, from this simulation study, it seems evident that IBLA is a good alternative for improving the estimation of regular statistical models, especially in terms of mean squared error.

Figure 3.2 shows how the simulation-based estimate of the bias function behaves

Table 3.4: Bias, mean squared error (MSE), and empirical coverage probability of Wald-type confidence intervals for the parameters of the binary logistic model with true values $\beta_0 = 1$, $\beta_1 = 1.5$.

			$n = 60$			$n = 120$			$n = 240$		
	Method	R	Bias	MSE	Coverage	Bias	MSE	Coverage	Bias	MSE	Coverage
β_0	ML	-	0.064	0.152	0.962	0.035	0.067	0.958	0.019	0.032	0.958
	mean BR	-	0.002	0.124	0.963	0.006	0.061	0.957	0.004	0.030	0.958
	BOOT	n	-0.028	0.113	0.962	-0.034	0.056	0.953	-0.017	0.029	0.956
		$2n$	-0.060	0.088	0.971	-0.032	0.056	0.955	-0.003	0.029	0.959
		$3n$	-0.055	0.090	0.970	-0.022	0.056	0.958	0.001	0.030	0.959
		$4n$	-0.054	0.094	0.967	-0.021	0.057	0.957	0.001	0.030	0.958
	IBLA	n	-0.017	0.118	0.964	-0.030	0.056	0.954	-0.014	0.029	0.958
		$2n$	-0.050	0.095	0.971	-0.026	0.056	0.958	-0.001	0.030	0.960
		$3n$	-0.051	0.093	0.969	-0.017	0.056	0.959	0.003	0.030	0.959
		$4n$	-0.051	0.095	0.969	-0.017	0.057	0.958	0.002	0.030	0.959
β_1	ML	-	0.144	0.336	0.960	0.064	0.134	0.955	0.027	0.058	0.954
	mean BR	-	0.011	0.245	0.954	0.002	0.117	0.952	-0.003	0.055	0.951
	BOOT	n	0.025	0.224	0.960	-0.060	0.109	0.943	-0.030	0.054	0.945
		$2n$	-0.083	0.185	0.953	-0.048	0.110	0.947	-0.017	0.053	0.948
		$3n$	-0.116	0.189	0.944	-0.038	0.107	0.950	-0.014	0.054	0.949
		$4n$	-0.093	0.192	0.948	-0.036	0.110	0.949	-0.011	0.054	0.949
	IBLA	n	0.040	0.241	0.961	-0.050	0.109	0.947	-0.025	0.053	0.946
		$2n$	-0.057	0.197	0.956	-0.039	0.109	0.949	-0.012	0.053	0.950
		$3n$	-0.092	0.197	0.948	-0.029	0.107	0.953	-0.009	0.054	0.951
		$4n$	-0.072	0.198	0.953	-0.028	0.109	0.951	-0.007	0.054	0.951

Notes: The coverage probabilities correspond to nominally 95% confidence intervals.

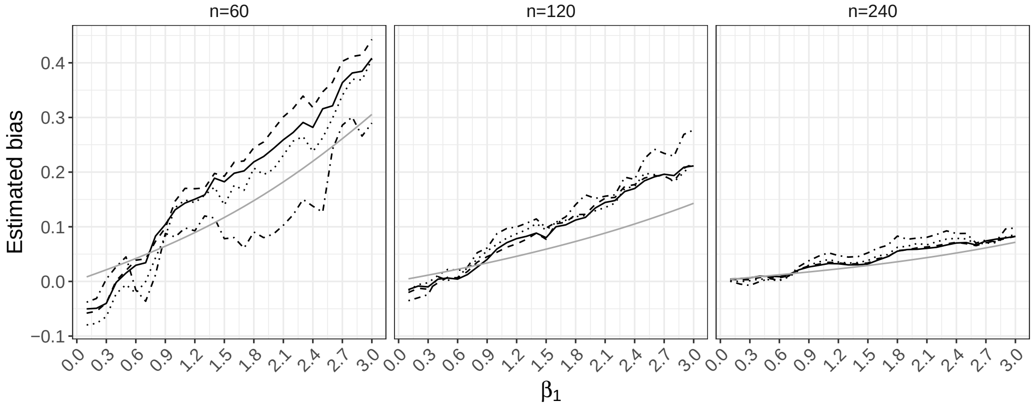

Figure 3.2: Monte Carlo estimate of bias of the ML estimator of β_1 in a binary logistic model with $\beta_0 = 0.5$. The curves correspond to Monte Carlo size R equal to n (dotdashed), $2n$ (dotted), $3n$ (dashed), and $4n$ (solid). The grey line is the first-order term of the bias of $\hat{\beta}_1$.

for values of $R = \{n, 2n, 3n, 4n\}$ for each of the three values of sample size n considered. As R gets larger, the Monte Carlo estimate of the bias function becomes a smoother and better approximation of the first-order term in the bias expansion.

3.6.3.2 Endometrial cancer grade study

As a real-data example we consider the endometrial cancer grade dataset analysed in Heinze & Schemper (2002) and in Agresti (2015, Chapter 5.7.1). The goal of the study was to evaluate the relationship between the histology of the endometrium of 79 patients ($0 =$ low grade for 30 patients, $1 =$ high grade for 49 patients) and three risk factors: neovasculation, pulsatility index of arteria uterina, and endometrium height. A logistic regression model has been fitted with parameter $\theta = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, where β_0 is an intercept and the remaining parameters correspond to neovasculation, pulsatility index of arteria uterina, and endometrium height, respectively.

Table 3.5 shows the ML, the mean BR estimates, the parametric bootstrap estimates, and the IBLA estimates of the binary-response logistic model parameters. For the parametric bootstrap and IBLA, the estimates were calculated by sampling $R = \{n, 2n, 3n, 4n\}$ bootstrap samples from the fitted model at the candidate value. In this dataset 13 patients have neovasculation and they all are with high grade histologic type of the endometrium. This leads to infinite ML estimate of β_1 due to the quasi-complete separation problem (Heinze & Schemper, 2002). The other ML estimates are not affected by the quasi-complete separation. The parametric bootstrap estimate of β_1 is also infinite, because by definition it depends on the ML estimate. On the contrary, the mean BR method does not depend on the finiteness of the ML estimate and yields a finite estimate of β_1. In order for IBLA to yield a finite estimate of β_1 we set the quasi Newton-Raphson algorithm start from the mean BR instead of the ML estimates, and we also used the adjusted simulated responses y^c in each iteration.

Figure 3.3 shows the iterations for IBLA for the endometrial cancer grade data for the four values of R considered. Comparing the mean BR to the IBLA estimates of β_1 reported in Table 3.5 we notice that the mean BR estimate is more than twice the relative IBLA estimate.

In order to further investigate the performance of IBLA, we performed a simulation study where we considered a binary-response logistic model with true parameter values equal to $\theta = (1.5, 2, 0, -2)^{\mathrm{T}}$, and used the neovasculation, pulsatility index of arteria uterina, and endometrium height risk factors from the endometrial cancer grade data as the covariates. Based upon 10 000 replications of simulated data we present in

Table 3.5: ML, mean BR, parametric bootstrap (BOOT), and IBLA estimates of the model parameters for the endometrial cancer study. The estimated standard errors are reported in parentheses.

Method	R	β_0	β_1	β_2	β_3
ML	-	4.305 (1.637)	$+\infty$ ($+\infty$)	-0.042 (0.044)	-2.903 (0.846)
mean BR	-	3.775 (1.489)	2.929 (1.551)	-0.035 (0.040)	-2.604 (0.776)
BOOT	n	3.839 (1.534)	$+\infty$ ($+\infty$)	-0.035 (0.042)	-2.632 (0.787)
	$2n$	3.687 (1.503)	$+\infty$ ($+\infty$)	-0.032 (0.041)	-2.561 (0.772)
	$3n$	3.604 (1.491)	$+\infty$ ($+\infty$)	-0.033 (0.041)	-2.493 (0.961)
	$4n$	3.704 (1.508)	$+\infty$ ($+\infty$)	-0.035 (0.042)	-2.532 (0.767)
IBLA	n	3.489 (1.372)	1.220 (0.914)	-0.028 (0.034)	-2.465 (0.736)
	$2n$	3.402 (1.349)	1.072 (0.882)	-0.025 (0.033)	-2.442 (0.729)
	$3n$	3.366 (1.343)	1.112 (0.883)	-0.027 (0.033)	-2.400 (0.721)
	$4n$	3.472 (1.357)	1.093 (0.883)	-0.030 (0.033)	-2.445 (0.729)

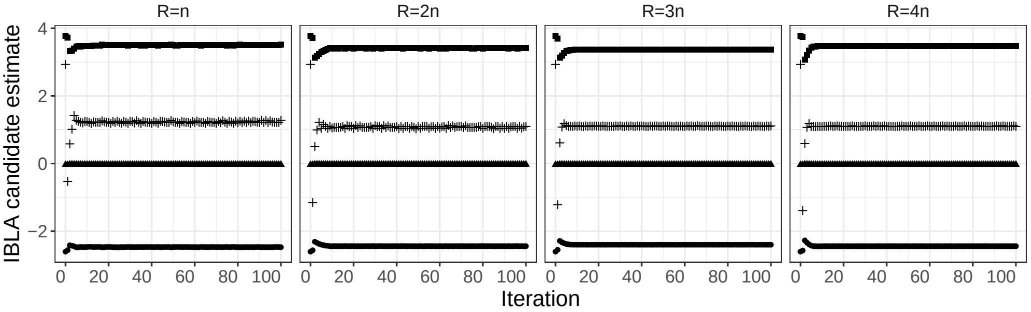

Figure 3.3: Plot of the candidate values for the IBLA estimates of the binary logistic model parameters β_0 (square), β_1 (cross), β_2 (triangle), and β_3 (circle). The starting values of the algorithm are set to the mean BR estimates shown in Table 3.5.

Table 3.6 the bias and mean squared error of the estimates from the four fitting methods (ML, mean BR, BOOT, IBLA). We also give the estimated coverage probability of the individual Wald-type confidence intervals at levels 90, 95, and 99%. The mean BR and IBLA methods reduce the bias and mean squared error of the ML estimates, with the former being best in terms of bias and the latter being best in terms of mean squared error. Parametric bootstrap does not always achieve an improvement in the estimation. Finally, comparing the coverage probabilities obtained at the three nominal levels, it is evident that the Wald-type confidence intervals calculated based on any of the methods are close to the nominal level, except for the confidence intervals calculated based on the parametric bootstrap estimates, which significantly undercover the true parameter value across parameter β_1.

Table 3.6: Bias, mean squared error (MSE), and empirical coverage probability of Wald-type confidence intervals for β_i ($i \in \{0,1,2,3\}$) in the endometrial cancer grade setting.

		Nominal level (%)	ML	mean BR	BOOT			IBLA		
R			-	-	n	$2n$	$3n$	n	$2n$	$3n$
β_0	Bias	-	0.287	-0.010	-0.381	-0.496	-0.522	-0.229	-0.213	-0.127
	MSE	-	2.595	1.900	2.945	3.955	5.125	1.710	1.567	1.454
	Coverage	90	0.903	0.916	0.895	0.872	0.882	0.913	0.922	0.928
		95	0.956	0.961	0.952	0.927	0.930	0.961	0.963	0.969
		99	0.996	0.995	0.991	0.970	0. 971	0.995	0.996	0.997
β_1	Bias	-	1.007	-0.013	-1.595	-2.188	-2.342	-0.359	-0.499	-0.536
	MSE	-	1.425	0.754	2.555	2.852	2.907	0.548	0.591	0.536
	Coverage	90	0.938	0.937	0.536	0.368	0.331	0.910	0.912	0.910
		95	0.972	0.971	0.626	0.487	0.440	0.954	0.958	0.957
		99	0.995	0.995	0.753	0.690	0.657	0.994	0.995	0.994
β_2	Bias	-	-0.002	-0.001	0.001	0.001	0.002	0.001	0.001	-0.001
	MSE	-	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Coverage	90	0.888	0.910	0.897	0.917	0.923	0.928	0.929	0.934
		95	0.948	0.960	0.954	0.963	0.966	0.966	0.971	0.973
		99	0.993	0.996	0.993	0.994	0.996	0.998	0.997	0.998
β_3	Bias	-	-0.251	0.011	0.329	0.474	0.520	0.172	0.183	0.161
	MSE	-	1.018	0.659	2.076	2.522	3.593	0.566	0.513	0.460
	Coverage	90	0.906	0.902	0.869	0.816	0.839	0.893	0.894	0.911
		95	0.961	0.950	0.923	0.876	0.884	0.945	0.943	0.952
		99	0.993	0.989	0.974	0.926	0.926	0.988	0.986	0.988

3.7 Concluding remarks

In this chapter we propose IBLA, a computational method for the reduction of bias of the ML estimator that is applicable regardless of the infeasibility of the bias function. Our method extends the framework in Firth (1993) and Kosmidis & Firth (2009) and systematically corrects the mechanism that produces the ML estimates by introducing a small bias in the score function. This extension relies on the use of the Monte Carlo approximation of the bias function instead of the first-order bias term. Under suitable conditions we show the consistency and asymptotic normality of the IBLA estimator as the Monte Carlo size R goes to infinity. We also show that bias reduction is achieved for $R = O(n^{\alpha})$, $\alpha \geq 1$. A formal proof for the asymptotic normality of the IBLA estimator when the observations are independent but non identically distributed remains to be formulated, and further work is required in this direction.

The Monte Carlo approximation of the bias function depends upon the existence of the ML estimates and so IBLA may fail, for example, in situations like logistic regression where the ML estimates are infinite with positive probability. We propose adjusting suitably the bootstrap samples generated during the computation of the approximated bias function so that all ML estimates obtained are finite, thus producing finite IBLA estimates.

Finally, we evaluate the performance of IBLA in terms of bias, mean squared error, and coverage probability in the framework of generalised linear models, including models for which the simulation-based adjusted score function is discontinuous in terms of θ. The simulation results suggest that IBLA does not outperform the traditional adjusted score function approach (Firth, 1993) in terms of bias, but it yields the smallest mean squared error. Also, IBLA is better than ML and parametric bootstrap in terms of bias, mean squared error, and coverage probabilities.

Summing up, we conclude that IBLA can be regarded as an overall improvement over ML and parametric bootstrap, even for discrete-response models. Another advantage of IBLA estimates is that they do not depend upon the finiteness of the ML estimates, whereas by definition, the parametric bootstrap estimates are infinite when the ML estimates are infinite. Lastly, the implementation of IBLA is rather attractive because it does not require an analytic expression for the first-order term of the bias, like the traditional adjusted score equation (Firth, 1993) does. IBLA allows practitioners to obtain bias reduced estimates through the solution of a feasible equation, just by having the score function, the observed information matrix, and the ability to simulate samples from the model.

Chapter 4

Mean bias reduction for models with intractable likelihood

4.1 Introduction

In this chapter we consider variations of the adjusted score functions proposed in Firth (1993) to reduce mean bias of the ML estimator, that apply regardless of the feasibility of the bias function and the tractability of the likelihood function. A likelihood is defined as intractable when it cannot be evaluated analytically, and the integrals involved in it require approximations, or its evaluation is prohibitively expensive for practical purposes.

In Chapter 3 we show that solving an adjusted score equation where the bias function is replaced by its simulation-based estimate, also leads to estimators with $o(n^{-1})$ bias. In this chapter we further extend the typical framework of adjusting the score function and show that a suitable approximation of the likelihood can be used in order to obtain an approximate adjusted score equation. The solution to this equation yields estimators with smaller bias than the maximum approximate likelihood estimator that maximises the approximate log-likelihood. We give the conditions under which an approximation of the likelihood may be used in order to derive bias-reduced estimates, and we show that the Laplace approximation (Tierney & Kadane, 1986) satisfies them. However, the tractable approximate adjusted score equation is infeasible due to the infeasibility of the approximate bias function. For this reason, we replace the approximate bias function by its simulation-based estimate, which leads to the development of a feasible and tractable adjusted score equation method for removing the first-order term in the asymptotic expansion of the bias of the maximum approximate likelihood estimator.

4.2 Tractable simulation-based adjusted score function

Definition 2. Let $\tilde{s}_n(\theta)$ and $\tilde{j}_n(\theta)$ be the gradient and negative Hessian matrix of an approximation of the log-likelihood, and $\tilde{B}_n(\theta) = E_\theta(\tilde{\theta}_n - \theta)$, with $\tilde{\theta}_n$ being a maximum approximate likelihood estimator such that $\tilde{s}_n(\tilde{\theta}_n) = 0$. Also consider the tractable but infeasible approximate adjusted score function

$$\tilde{s}^*_n(\theta) = \tilde{s}_n(\theta) - \tilde{j}_n(\theta)\tilde{B}_n(\theta)\,. \tag{4.1}$$

Let $\tilde{\theta}^*_n$ be the solution of $\tilde{s}^*_n(\theta) = 0$. Theorem 5 gives the conditions that an approximation method needs to satisfy in terms of n, in order for the estimator $\tilde{\theta}^*_n$ to have smaller bias than $\tilde{\theta}_n$.

Theorem 5. *If $\tilde{s}_n(\theta) - s_n(\theta) = O(n^a)$, $\tilde{j}_n(\theta) - j_n(\theta) = O(n^b)$, and $\tilde{B}_n(\theta) - B_n(\theta) = O(n^c)$ with $\max\{a, b-1, c+1\} \leq -1/2$ then $\tilde{\theta}^*_n$ has $o(n^{-1})$ bias.*

Proof of Theorem 5: Using the expressions $\tilde{s}_n(\theta) - s_n(\theta) = O(n^a)$, $\tilde{j}_n(\theta) - j_n(\theta) = O(n^b)$, $\tilde{B}_n(\theta) - B_n(\theta) = O(n^c)$ and given also that $B_n(\theta) = O(n^{-1})$, we can write the mean bias reducing adjusted score function based on the observed information matrix (Firth, 1993) in the form

$$\begin{aligned}
s^*_n(\theta) &= s_n(\theta) - j_n(\theta)B_n(\theta) + v(\theta) \\
&= [\tilde{s}_n(\theta) - O(n^a)] - [\tilde{j}_n(\theta) - O(n^b)]B_n(\theta) + v(\theta) \\
&= \tilde{s}_n(\theta) - O(n^a) - \tilde{j}_n(\theta)B_n(\theta) + O(n^{b-1}) + v(\theta) \\
&= \tilde{s}_n(\theta) - O(n^a) - \tilde{j}_n(\theta)[\tilde{B}_n(\theta) - O(n^c)] + O(n^{b-1}) + v(\theta) \\
&= \tilde{s}_n(\theta) - \tilde{j}_n(\theta)\tilde{B}_n(\theta) - O(n^a) + [j_n(\theta) + O(n^b)]O(n^c) + O(n^{b-1}) + v(\theta) \\
&= \tilde{s}^*_n(\theta) + O(n^{\max\{a,b-1,c+1\}}) + v(\theta)\,.
\end{aligned}$$

Hence, from (3.1), the solution of (4.1) has $o(n^{-1})$ bias if $\max\{a, b-1, c+1\} \leq -1/2$.

□

The adjusted score function in (4.1) is generally infeasible because the bias function $\tilde{B}_n(\theta) = E_\theta(\tilde{\theta}_n - \theta)$ cannot be computed. A natural way to tackle the infeasibility of (4.1) is to replace the bias function $\tilde{B}_n(\theta)$ by its simulation-based estimate.

Definition 3. The tractable and feasible estimating equation is

$$\tilde{s}^*_{n,R}(\theta) = \frac{1}{R}\sum_{r=1}^{R} \tilde{s}^*_n(\theta; Z_r) = 0\,, \tag{4.2}$$

where $\tilde{s}^*_n(\theta;Z_r) = \tilde{s}_n(\theta) - \tilde{j}_n(\theta)(\tilde{\theta}_n(Z_r) - \theta)$. In the above expressions, $\tilde{\theta}_n(Z_r)$ is the solution of $\tilde{s}_n(\theta;Z_r) = 0$, and $Z_r = z(\theta;\Xi_r)$ is a sample of responses simulated from the model at θ, based on $\Xi_1,\ldots,\Xi_R$ independent copies of a random variable Ξ that does not depend on θ. In this way, $E[\tilde{s}^*_n(\theta;Z)] = \tilde{s}^*_n(\theta)$ for $Z = z(\theta,\Xi)$ and any $\theta \in \Theta$.

Definition 4. Let the Monte Carlo estimate of the bias function $\tilde{B}_n(\theta)$ be $\tilde{B}_{n,R}(\theta) = \bar{\theta}_{n,R} - \theta$ with $\bar{\theta}_{n,R} = (1/R)\sum_{r=1}^{R} \tilde{\theta}_n(Z_r)$. The tractable simulation-based adjusted score function in (4.2) can be expressed as $\tilde{s}^*_{n,R}(\theta) = \tilde{s}_n(\theta) - \tilde{j}_n(\theta)\tilde{B}_{n,R}(\theta)$.

4.3 Asymptotic properties

The following conditions result in the consistency and asymptotic normality of the root $\tilde{\theta}^*_{n,R}$ of the tractable simulation-based adjusted score function $\tilde{s}^*_{n,R}(\theta)$. Note that we still assume compactness of the parameter space as stated in condition 1.

Condition 8. $\tilde{s}_n(\theta)$ and $\tilde{j}_n(\theta)$ are continuous for all $\theta \in \Theta$.

Condition 9. $\tilde{s}^*_n(\theta)$ has a unique zero at $\tilde{\theta}^*_n \in \Theta$.

Condition 10. $\tilde{s}^*_{n,R}(\theta)$ is continuously differentiable for all θ in a neighbourhood of the true unknown θ_0, and the matrix $\tilde{H}^*_{n,R}(\theta)$ with elements $\partial\tilde{s}^*_{n,R}(\theta)/\partial\theta_j$, $j \in \{1,\ldots,p\}$ is nonsingular.

Condition 11. For all $\theta \in \Theta$, $i \in \{1,\ldots,n\}$ and $\{j,k\} \in \{1,\ldots,p\}$,

$$E\left(\frac{\partial^2 \log \tilde{f}_i(y_i;\theta)}{\partial\theta_j\partial\theta_k}\right) \quad \text{and} \quad \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{\partial^2 \log \tilde{f}_i(y_i;\theta)}{\partial\theta_j\partial\theta_k}\right)$$

exist and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log \tilde{f}_i(y_i;\theta)}{\partial\theta_j\partial\theta_k} \xrightarrow{p} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{\partial^2 \log \tilde{f}_i(y_i;\theta)}{\partial\theta_j\partial\theta_k}\right).$$

Condition 12. The matrix $\bar{\bar{F}}(\theta_0) = \left(\bar{\bar{F}}_{jk}(\theta_0)\right)_{1 \leq j,k \leq p}$, where

$$\bar{\bar{F}}_{jk}(\theta_0) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\left(- \frac{\partial^2 \log \tilde{f}_i(y_i;\theta)}{\partial\theta_j \partial\theta_k}\bigg|_{\theta_0} \right),$$

is positive definite.

Condition 13. As n and R go to infinity, $n^{-1}[-\tilde{H}^*_{n,R}(\theta)] \xrightarrow{p} \bar{\bar{F}}(\theta)$ for θ in a neighbourhood of θ_0.

Conditions 8-9 are for the consistency of $\tilde{\theta}^*_{n,R}$ and conditions 10-13 are additionally necessary for the asymptotic normality of $\tilde{\theta}^*_{n,R}$. Using Lemma 1, where we prove the continuity of $\tilde{s}^*_{n,R}(\theta)$ and $\tilde{s}^*_n(\theta)$, we show in Theorem 6 that with probability one the tractable simulation-based adjusted score function $\tilde{s}^*_{n,R}(\theta)$ converges to the tractable but infeasible adjusted score function $\tilde{s}^*_n(\theta)$ in (4.1) uniformly in θ, as $R \to \infty$, which in turn results to $\tilde{\theta}^*_{n,R}$ being a consistent estimator of θ_0.

Lemma 1. *If conditions 1 and 8 are satisfied, then (i) the feasible and tractable adjusted score function $\tilde{s}^*_{n,R}(\theta)$, and (ii) the infeasible but tractable adjusted score function $\tilde{s}^*_n(\theta)$, are continuous.*

Proof of Lemma 1: Condition 8 implies result (i) because $\tilde{s}^*_{n,R}(\theta)$ is the sum of continuous functions (Rudin, 1976, Theorem 4.9). In order to prove continuity of $\tilde{s}^*_n(\theta)$, consider some $\theta \in \Theta$ and a sequence $\{\theta_j\}$ such that $\theta_j \to \theta$. Then, $\lim_{j\to\infty} \tilde{s}^*_n(\theta_j) = \lim_{j\to\infty} E[\tilde{s}^*_n(\theta_j;Z)]$. The function $\tilde{s}^*_n(\theta;Z)$ is continuous and bounded by its supremum for all $\theta \in \Theta$, which is finite as it is the supremum of a continuous function over a compact set (Rudin, 1976, Theorem 4.16). Applying the bounded convergence theorem (Feller, 2008, p. 111), $\lim_{j\to\infty} E[\tilde{s}^*_n(\theta_j;Z)] = E[\lim_{j\to\infty} \tilde{s}^*_n(\theta_j;Z)]$. Thus, $\lim_{j\to\infty} \tilde{s}^*_n(\theta_j) = \tilde{s}^*_n(\theta)$ and result (ii) is obtained. □

The law of large numbers (Van der Vaart, 2000, Proposition 2.16) implies that the Monte Carlo estimate of the bias function $\tilde{B}_{n,R}(\theta)$ converges in probability to $B_n(\theta)$ for every $\theta \in \Theta$ as $R \to \infty$, which implies that $\tilde{s}^*_{n,R}(\theta)$ converges to $\tilde{s}^*_n(\theta)$. This pointwise convergence is, though, not strong enough to guarantee convergence of the solutions of equation (4.2) to the infeasible $\tilde{\theta}_n^*$, which according to Theorem 5 has $o(n^{-1})$ bias. According to Van der Vaart (2000, Section 5.2), a sufficient condition for the convergence of the solutions is the uniform convergence of $\tilde{s}_{n,R}^*(\theta)$ to $\tilde{s}_n^*(\theta)$. Similar to Theorem 2, Theorem 6 outlines sufficient conditions for the uniform convergence and is inspired by Van der Vaart (2000, Theorem 5.9) and the subsequent discussion therein.

Theorem 6. *If conditions 1 and 8 are satisfied, then $\tilde{s}_n^*(\theta)$ and $\tilde{s}_{n,R}^*(\theta)$ are such that $\sup_{\theta\in\Theta} \|\tilde{s}_{n,R}^*(\theta) - \tilde{s}_n^*(\theta)\| \xrightarrow{p} 0$ as $R \to \infty$.*

Proof of Theorem 6: A set of sufficient conditions for uniform convergence of functions that can be written as an average is that Θ is compact, that the functions $\tilde{s}_n^*(\theta;Z)$ are continuous for every θ, and that they are dominated by an integrable function (Van der Vaart, 2000, Theorem 5.9). Condition 1 and Lemma 1 cover for compactness and continuity. From the triangle inequality the function $\tilde{s}_n^*(\theta;Z)$ is bounded on Θ because there exists a positive number $K_n(\theta) = \|\tilde{s}_n(\theta)\| + \|\tilde{j}_n(\theta)(\tilde{\theta}_n(Z) - \theta)\|$ such that $\tilde{s}_n^*(\theta;Z) \leq K_n(\theta)$. In order to show that $K_n(\theta)$ is integrable we need to show that it is continuous on a rectangle in $\Re^p$ (see, Trench, 2003, Theorem 7.1.13). The space Θ is compact, which is equivalent by the Heine-Borel theorem (Rudin, 1976, pp. 39-40) to Θ being closed and bounded. Then Θ is a closed subset of a rectangle that is product of bounded intervals (Lavrent'ev & Savel'ev, 2006, p. 165). Also, the function $K(\theta)$ is continuous as it is the sum of vector norms. Thus it is integrable. □

The following corollary establishes that the tractable and feasible estimator $\tilde{\theta}_{n,R}^*$ converges in probability to the reduced-bias tractable but infeasible estimator $\tilde{\theta}_n^*$, and thus $\tilde{\theta}_{n,R}^*$ is a consistent estimator of $\tilde{\theta}_n^*$. Consistency of $\tilde{\theta}_{n,R}^*$ requires that the number of Monte Carlo samples R in the construction of $\tilde{B}_{n,R}(\theta)$ goes to infinity as $n \to \infty$.

Corollary 2. *If conditions 1, 8, and 9 are satisfied and $\tilde{s}_{n,R}^*(\theta)$ converges uniformly to $\tilde{s}_n^*(\theta)$, then any $\tilde{\theta}_{n,R}^* \in \Theta$ such that $\tilde{s}_{n,R}^*(\tilde{\theta}_{n,R}^*) = 0$ converges in probability to $\tilde{\theta}_n^*$ as $R \to \infty$. Also, $\tilde{\theta}_{n,R}^*$ is a consistent estimator of θ_0.*

Proof of Corollary 2: In Theorem 6 we established uniform convergence of $\tilde{s}_{n,R}^*(\theta)$ to $\tilde{s}_n^*(\theta)$, as $R \to \infty$. Then for every $\varepsilon > 0$, there exists $M > 0$ such that for $R > M$,

$\varepsilon > \sup_{\theta\in\Theta} \|\tilde{s}^*_{n,R}(\theta) - \tilde{s}^*_n(\theta)\| \geq \|\tilde{s}^*_{n,R}(\tilde{\theta}^*_{n,R}) - \tilde{s}^*_n(\tilde{\theta}^*_{n,R})\| = \|\tilde{s}^*_n(\tilde{\theta}^*_{n,R})\|$. So the sequence $\{\tilde{\theta}^*_{n,R}\}$ will converge to the unique $\tilde{\theta}^*_n$. Also, from Theorem 5 we have that $\tilde{\theta}^*_n \xrightarrow{p} \theta_0$ as $n \to \infty$. Thus $\tilde{\theta}^*_{n,R} \xrightarrow{p} \theta_0$. □

Having established the consistency of the tractable and feasible estimator as an estimator of the unknown θ_0 as R and n go to infinity, we proceed under some additional conditions to prove the asymptotic normality of $n^{1/2}(\tilde{\theta}^*_{n,R} - \theta_0)$.

Theorem 7. *If conditions 1 and 8-13 are satisfied, the observations are independent and identically distributed, and the number of Monte Carlo samples R is fixed with* $n \to \infty$, *then* $n^{1/2}(\tilde{\theta}^*_{n,R} - \theta_0)$ *is asymptotically normally distributed with zero mean and covariance matrix* $(1 + R^{-1})\{E[\tilde{j}_i(\theta_0)]\}^{-1}$.

The proof of Theorem 7 is similar to the proof of Theorem 3, and is given below.

Proof of Theorem 7: Because $\tilde{\theta}^*_{n,R}$ is a consistent estimator of the true parameter θ_0 as n and R go to infinity, it makes sense to expand $\tilde{s}^*_{n,R}(\theta)$ in a Taylor series around θ_0. Application of Taylor's theorem to $\tilde{s}^*_{n,R}(\theta)$ about its solution $\tilde{\theta}^*_{n,R}$ gives $0 = \tilde{s}^*_{n,R}(\theta_0) + \nabla\tilde{s}^*_{n,R}(\breve{\theta})(\tilde{\theta}^*_{n,R} - \theta_0)$, where $\breve{\theta} = \theta_0 + t(\tilde{\theta}^*_{n,R} - \theta_0)$, with $t \in (0,1)$. Thus

$$n^{1/2}(\tilde{\theta}^*_{n,R} - \theta_0) = \left\{ -\frac{\nabla\tilde{s}^*_{n,R}(\breve{\theta})}{n} \right\}^{-1} \frac{\tilde{s}^*_{n,R}(\theta_0)}{n^{1/2}}.$$

By the central limit theorem (Van der Vaart, 2000, Proposition 2.17) $n^{-1/2}\tilde{s}_n(\theta_0) \xrightarrow{d} N(0_p, E[\tilde{j}_i(\theta_0)])$ as $n \to \infty$. Again by the central limit theorem and for all $r \in \{1,\ldots,R\}$ $n^{1/2}(\tilde{\theta}_{n,r} - \theta_0) \xrightarrow{d} N(0_p, \{E[\tilde{j}_{i,r}(\theta_0)]\}^{-1}) = N(0_p, \{E[\tilde{j}_i(\theta_0)]\}^{-1})$ as $n \to \infty$ and R is fixed. Then because $\{n^{1/2}(\tilde{\theta}_{n,r} - \theta_0)\}_{r=1}^R$ are independent we have the joint limit

$$\begin{bmatrix} n^{1/2}\left(\tilde{\theta}_{n,1} - \theta_0\right) \\ n^{1/2}\left(\tilde{\theta}_{n,2} - \theta_0\right) \\ \vdots \\ n^{1/2}\left(\tilde{\theta}_{n,R} - \theta_0\right) \end{bmatrix} \xrightarrow{d} N(0_{pR}, D)$$

where D is a block diagonal matrix with main diagonal blocks the matrices $\{E[\tilde{j}_i(\theta_0)]\}^{-1}$. In view of the joint convergence in distribution (joint for all elements of the vector above) the continuous mapping theorem (Van der Vaart, 2000, Theorem 2.3) gives

$$n^{1/2}\tilde{B}_{n,R}(\theta_0) = \frac{1}{R}\sum_{r=1}^{R} n^{1/2}(\tilde{\theta}_{n,r} - \theta_0) \xrightarrow{d} N\left(0_p, \frac{1}{R}\{E[\tilde{j}_i(\theta_0)]\}^{-1}\right).$$

Further, because $n^{-1/2}\sum_{i=1}^{n}\tilde{s}_i(\theta_0)$ and $(1/R)\sum_{r=1}^{R}n^{1/2}(\tilde{\theta}_{n,r} - \theta_0)$ are independent we have the joint limit as $n \to \infty$

$$\begin{bmatrix} n^{-1/2}\tilde{s}_n(\theta_0) \\ n^{1/2}\tilde{B}_{n,R}(\theta_0) \end{bmatrix} \xrightarrow{d} N\begin{pmatrix} 0_{2p}, & E[\tilde{j}_i(\theta_0)] & 0_{p\times p} \\ & 0_{p\times p} & \frac{1}{R}\{E[\tilde{j}_i(\theta_0)]\}^{-1} \end{pmatrix}.$$

In view of the above and the fact that by the weak law of large numbers (Davison, 2003, p. 28) $n^{-1}\tilde{j}_n(\theta_0) \xrightarrow{p} E[\tilde{j}_i(\theta_0)]$ as $n \to \infty$, we have that

$$n^{-1/2}\tilde{s}^*_{n,R}(\theta_0) = \frac{\tilde{s}_n(\theta_0)}{n^{1/2}} - \frac{\tilde{j}_n(\theta_0)}{n}n^{1/2}\tilde{B}_{n,R}(\theta_0) \xrightarrow{d} N\left(0_p, \left(1+R^{-1}\right)E[\tilde{j}_i(\theta_0)]\right).$$

Under the assumption of independent and identically distributed observations the matrix $\bar{\bar{F}}(\theta)$ in Condition 13 is $E[\tilde{j}_i(\theta)]$. Using this result, the consistency of $\breve{\theta}$, and Slutsky's Lemma (Van der Vaart, 2000, Lemma 2.8) we have $n^{1/2}(\tilde{\theta}^*_{n,R} - \theta_0) \xrightarrow{d} N\left(0_p, \left(1+R^{-1}\right)\{E[\tilde{j}_i(\theta_0)]\}^{-1}\right)$. □

4.4 Iterated bootstrap with likelihood adjustment

A direct approach for computing the tractable simulation-based bias-reduced estimator that solves $\tilde{s}^*_{n,R}(\tilde{\theta}^*_{n,R}) = 0$ with $\tilde{s}^*_{n,R}(\theta) = \tilde{s}_n(\theta) - \tilde{j}_n(\theta)\tilde{B}_{n,R}(\theta)$ is through IBLA algorithm that was introduced in Section 3.5. Specifically, $\tilde{\theta}^*_{n,R}$ is obtained through a similar iteration with (3.5), where the derivatives of the likelihood are replaced by the derivatives of the approximated likelihood, i.e.

$$\theta_n^{(j+1)} = (2\theta_n^{(j)} - \bar{\theta}_{n,R}^{(j)}) + \{\tilde{j}_n(\theta_n^{(j)})\}^{-1}\tilde{s}_n(\theta_n^{(j)}). \tag{4.3}$$

In the above iterations, $\theta_n^{(j)}$ is the candidate value for $\tilde{\theta}^*_{n,R}$ at the jth iteration, and $\bar{\theta}_{n,R}^{(j)}$ is the average of the maximum approximate likelihood estimators calculated for each of R simulated samples from the model at $\theta_n^{(j)}$.

Starting from the maximum approximate likelihood estimate, a single iteration gives the approximate parametric bootstrap corrected estimate. A stopping criterion for the iterations is $|\theta_n^{(j+1)} - \theta_n^{(j)}| < \varepsilon$, for some prespecified $\varepsilon > 0$. We recommend using the same initial state for the random number generator in each iteration in order to achieve a smooth estimator of the bias function.

4.5 Adjusted score functions with Laplace approximation

In this section we show that Laplace approximation (Tierney & Kadane, 1986) satisfies the conditions in Theorem 5, and therefore it can be used to approximate the likelihood and yield a tractable adjusted score equation whose solution has smaller mean bias.

Suppose the marginal likelihood $L_n(\theta; y)$ of a model involves integrals of the form $\int_{\Re^d} \exp\{nq(\alpha|y,\theta)\}d\alpha$. For each fixed θ, the Laplace approximation relies on an approximation to $q(\alpha|y,\theta)$, using its second-order Taylor series expansion. This is given by $q(\alpha|y,\theta) \approx q(\alpha_{\max}|y,\theta) + \frac{1}{2}(\alpha - \alpha_{\max})^{\mathrm{T}}\Sigma_{\alpha_{\max}}(\theta)(\alpha - \alpha_{\max})$, where $\alpha_{\max}$ is the maximum of $q(\alpha|y,\theta)$, and $\Sigma_{\alpha_{\max}}(\theta)$ is the Hessian matrix of $q(\alpha|y,\theta)$ evaluated at $\alpha_{\max}$. When we integrate the approximation of $\exp\{nq(\alpha|y,\theta)\}$ over α, we have

$$\int_{\Re^d} e^{nq(\alpha|y,\theta)}d\alpha \approx (2\pi/n)^{d/2}\sigma_{\alpha_{\max}} e^{nq(\alpha_{\max}|y,\theta)}, \tag{4.4}$$

where $\sigma_{\alpha_{\max}} = |-\Sigma_{\alpha_{\max}}(\theta)|^{-1/2}$.

The expansion in (4.4) is accurate to order $O(n^{-1})$ since we only consider the first-order terms of Laplace approximation (Tierney & Kadane, 1986), i.e.

$$\int_{\Re^d} e^{nq(\alpha|y,\theta)}d\alpha = (2\pi/n)^{d/2}\sigma_{\alpha_{\max}} e^{nq(\alpha_{\max}|y,\theta)} \left(1 + O(n^{-1})\right).$$

A more refined result is given by

$$\int_{\Re^d} e^{nq(\alpha|y,\theta)}d\alpha = (2\pi/n)^{d/2}\sigma_{\alpha_{\max}} e^{nq(\alpha_{\max}|y,\theta)} \left(1 + \frac{c_1(y,\theta)}{n} + \frac{c_2(y,\theta)}{n^2} + O(n^{-3})\right),$$

where $c_1(y,\theta)$ and $c_2(y,\theta)$ are assumed to be $O(1)$ (Tierney & Kadane, 1986). Let $q_k = q_k(\alpha_{\max}|y,\theta)$ where $q_k(\alpha|y,\theta)$ is the kth derivative of $q(\alpha|y,\theta)$ with respect to α. The constants $c_1(y,\theta)$ and $c_2(y,\theta)$ are given by

$$\begin{aligned} c_1(y,\theta) &= \frac{1}{8}\sigma^4_{\alpha_{\max}} q_4 + \frac{5}{24}\sigma^6_{\alpha_{\max}} q_3^2 \\ c_2(y,\theta) &= \frac{1}{48}\sigma^6_{\alpha_{\max}} q_6 + \frac{35}{384}\sigma^8_{\alpha_{\max}} q_4^2 + \frac{7}{48}\sigma^8_{\alpha_{\max}} q_3 q_5 + \frac{35}{64}\sigma^{10}_{\alpha_{\max}} q_3^2 q_4 + \frac{385}{1152}\sigma^{12}_{\alpha_{\max}} q_3^4 . \end{aligned}$$

Let $\tilde{L}_n(\theta;y)$ be the Laplace approximation of $L_n(\theta;y)$. Then the log-likelihood $l_n(\theta;y)$ is

$$\begin{aligned} l_n(\theta;y) &= \log L_n(\theta;y) \\ &= \log\left(\tilde{L}_n(\theta;y)\left(1 + \frac{c_1(y,\theta)}{n} + \frac{c_2(y,\theta)}{n^2} + O(n^{-3})\right)\right) \\ &= \log \tilde{L}_n(\theta;y) + \log\left(1 + \frac{c_1(y,\theta)}{n} + \frac{c_2(y,\theta)}{n^2} + O(n^{-3})\right) \\ &= \tilde{l}_n(\theta;y) + \log\left(1 + O(n^{-1})\right) \\ &= \tilde{l}_n(\theta;y) + O(n^{-1}) . \end{aligned} \tag{4.5}$$

The first derivative of (4.5) yields

$$\begin{aligned} s_n(\theta;y) &= \tilde{s}_n(\theta;y) + \nabla_\theta \log\left(1 + \frac{c_1(y,\theta)}{n} + \frac{c_2(y,\theta)}{n^2} + O(n^{-3})\right) \\ &= \tilde{s}_n(\theta;y) \\ &\quad + \left(\frac{\nabla_\theta c_1(y,\theta)}{n} + \frac{\nabla_\theta c_2(y,\theta)}{n^2} + O(n^{-3})\right) \circ \left(1 + \frac{c_1(y,\theta)}{n} + \frac{c_2(y,\theta)}{n^2} + O(n^{-3})\right)^{-1} \\ &= \tilde{s}_n(\theta;y) + \frac{\nabla_\theta c_1(y,\theta)}{n} + \frac{\nabla_\theta c_2(y,\theta) - \nabla_\theta c_1(y,\theta) c_1(y,\theta)}{n^2} + O(n^{-3}) \\ &= \tilde{s}_n(\theta;y) + O(n^{-1}) . \end{aligned} \tag{4.6}$$

The second derivative of (4.5) is

$$\begin{aligned} H_n(\theta;y) &= \tilde{H}_n(\theta;y) + \nabla_\theta \left(\frac{\nabla_\theta c_1(y,\theta)}{n} + \frac{\nabla_\theta c_2(y,\theta) - \nabla_\theta c_1(y,\theta) c_1(y,\theta)}{n^2} + O(n^{-3})\right) \\ &= \tilde{H}_n(\theta;y) + \frac{\nabla\nabla_\theta^{\mathrm{T}} c_1(y_i,\theta)}{n} + O(n^{-2}) \\ &= \tilde{H}_n(\theta;y) + O(n^{-1}) . \end{aligned} \tag{4.7}$$

Results 4.6 and 4.7, show that the first two conditions in Theorem 5 are satisfied, where in this example $\tilde{s}_n(\theta)$ and $\tilde{j}_n(\theta)$ denote the gradient and negative Hessian matrix of the Laplace approximation of the log-likelihood.

The last condition we need to check is if $\tilde{B}_n(\theta) - B_n(\theta) = E_\theta(\tilde{\theta}_n - \hat{\theta}_n) = O(n^c)$ with $c \leq -3/2$. In order to find the order of $\tilde{B}_n(\theta) - B_n(\theta)$ we need the asymptotic expansion of $\tilde{\theta}_n - \hat{\theta}_n$, where $\tilde{\theta}_n$ maximises $\tilde{l}_n(\theta;y)$ and $\hat{\theta}_n$ maximises $l_n(\theta;y)$. Appendix C contains all the details on the asymptotic expansion of $\tilde{\theta}_n - \hat{\theta}_n$, which is calculated to be $\tilde{\theta}_n - \hat{\theta}_n = n^{-1}\{E[j_n(\theta_0;y)]\}^{-1}\nabla_\theta b(y,\theta_0) + O_p(n^{-5/2})$. Taking expectations on both sides we have

$$E_\theta(\tilde{\theta}_n - \hat{\theta}_n) = n^{-1}\{E[j_n(\theta_0;y)]\}^{-1}E\left(\nabla_\theta b(y,\theta_0)\right) + O_p(n^{-5/2}),$$

which results to $E_\theta(\tilde{\theta}_n - \hat{\theta}_n) = O(n^{-2})$.

To sum up, we showed that $\tilde{s}_n(\theta) - s_n(\theta) = O(n^{-1})$, $\tilde{j}_n(\theta) - j_n(\theta) = O(n^{-1})$, and $\tilde{B}_n(\theta) - B_n(\theta) = O(n^{-2})$, and the conditions in Theorem 5 hold. Therefore, the Laplace-based mean bias reduced estimator $\tilde{\theta}_n^*$ has $o(n^{-1})$ bias.

4.6 Concluding remarks

The tractable simulation-based adjusted score function proposed in Section 4.2 allows the calculation of bias-reduced estimates in models with intractable likelihood. We give the three conditions that need to be satisfied so that an approximation of the likelihood is suitable for bias reduction. We established that Laplace approximation matches the conditions for the applicability of the tractable simulation-based adjusted score function method. Also, we established the consistency and asymptotic normality of the bias-reduced estimates under suitable conditions. A formal proof for the asymptotic normality of the tractable simulation-based bias-reduced estimator when the observations are independent but non-identically distributed remains to be formulated, and further work is required in this direction.

The performance of the tractable simulation-based adjusted score equation approach is evaluated in the framework of generalised linear mixed models (McCulloch et al., 2008) in Chapter 5.

Chapter 5

Bias reduction in generalised linear mixed models

5.1 Introduction

Generalised linear mixed models are widely used for analysing non-normally distributed clustered data. The key characteristic of such models is the use of random effects to capture the between-cluster heterogeneity. The mixed model assumes that the responses are conditionally independent given a random effect and that the conditional means depend on random effects, fixed effects and covariates according to a generalised linear model specification. McCulloch et al. (2008, Chapter 7) provide a thorough overview of the models.

The integrals involved in the likelihood function of a generalised linear mixed model have, generally, no closed-form, and, hence, the likelihood function is intractable. There have been many proposals for estimating the parameters of the model, including approximating the likelihood by numerical integration or using a penalised quasi-likelihood (see, for example, Pinheiro & Chao, 2012; Breslow & Clayton, 1993). However, the estimators that are obtained from maximising an approximated likelihood have usually poor frequentist properties resulting in problems in inference, specifically in hypothesis testing and confidence intervals (McCulloch et al., 2008, Chapter 7).

In this chapter we focus on remedying such phenomena by producing variants of maximum approximate likelihood for generalised linear mixed models based on the methodology and the associated computational procedures developed in Chapter 4.

Specifically, we employ the bias reduction method that operates via the adjustment of the derivative of the approximate likelihood and the approximation of the bias function using Monte Carlo, and use IBLA and Laplace approximation for the computation of the tractable simulation-based bias reduced estimates.

First, we study a simple generalised linear mixed model, a binomial-response model with a fixed intercept and a random intercept only. For this model we analytically derive the adjusted score function (Firth, 1993). The purpose of this exercise is to demonstrate the challenges that one has to face when implementing the traditional adjusted score function approach for generalised linear mixed models. We also highlight the necessity for an extension of the method, such as IBLA, which can handle more complex and realistic models with covariates in the linear predictor or with complex random effect specifications.

Second, we use real data sets and conduct simulation studies to evaluate the performance of IBLA against some of the existing estimation methods used in the literature.

5.2 Generalised linear mixed model

A generalised linear mixed model is specified by the linear predictor, the link function, the conditional distribution for the response variable given the random effects, and the random effects distribution. The linear predictor is $X\beta + Z\alpha$, where X is the $n \times k$ design matrix of fixed-effects terms associated with the k regressors, β is the corresponding $k \times 1$ vector of the fixed-effects regression coefficients, Z is the $n \times q$ design matrix for the q random effects and α is the $q \times 1$ vector of the random effects. The conditional mean μ_i of the response is modelled as $g(\mu_i) = X_i^{\mathrm{T}}\beta + Z_i^{\mathrm{T}}\alpha$, where $g(\cdot)$ is the link function.

The observations $y_1, \ldots, y_n$ are assumed to be realisations of the random variables $Y_1, \ldots, Y_n$ from the exponential family of distributions (3.6), which are independent conditionally on the unobserved random effects. To complete the specification of the model we assign a distribution to the random effects, which are commonly assumed to follow a multivariate normal distribution with zero mean.

The likelihood function of a generalised linear mixed model can be written as

$$L(\theta) = \int \prod_{i} f_{Y_i|\alpha}(y_i|\alpha) f_{\alpha}(\alpha) d\alpha, \tag{5.1}$$

where the integration is over the q-dimensional distribution of α, $f_{Y_i|\alpha}(y_i|\alpha)$ is the conditional probability function of Y_i given α, and $f_{\alpha}(\alpha)$ is the density function of the random effects. In general, (5.1) cannot be evaluated in closed form. In simple models, e.g. generalised linear mixed models with a random intercept, the log-likelihood is the sum of independent contributions from each cluster, each of which involves just an one-dimensional integral, which can be readily and accurately evaluated using numerical integration techniques (McCulloch et al., 2008, Chapter 7).

5.3 Estimation in generalised linear mixed model

This section presents some of the most common methods that have been proposed for estimating generalised linear mixed models, and variants of those that have been proposed in an attempt to improve estimation quality. Some of the methods described here are used in the simulation studies presented later in this chapter when evaluating the performance of IBLA in terms of estimation and inference.

5.3.1 Standard estimation methods

(i) Maximum approximate likelihood

Fitting generalised linear mixed models via maximum likelihood involves integrating over the random effects. In general, these integrals are intractable, and numerical integration techniques can be used to evaluate them. Maximising the approximated likelihood then yields the maximum approximate likelihood estimates. Laplace approximation of the log-likelihood (Tierney & Kadane, 1986; Pinheiro & Chao, 2012) is the computationally least intensive compared to other Gaussian quadrature rules (Liu & Pierce, 1994) or Monte Carlo integration techniques (McCulloch, 1997) that can also be used to compute maximum approximate likelihood estimates.

Maximum approximate likelihood estimators have the desirable behaviour of being asymptotically unbiased and consistent under increasing cluster size and the number of clusters (Jiang et al., 2013). However, the underestimation of the variance components for finite sample sizes is a common concern for statisticians (see, for example, Kuk, 1995; Raudenbush et al., 2000). This underestimation can result in severe problems in inference, because bias in the variance components estimates leads to the underestimation of the standard errors for the fixed effects, which in turn result in shorter confidence intervals and smaller p-values.

(ii) Penalised quasi-likelihood

Another popular method for fitting generalised linear mixed models is penalised quasi-likelihood (PQL) proposed in Breslow & Clayton (1993), who linearise the conditional mean of the model and then repeatedly apply linear mixed model techniques to the approximated model. The linearisation is achieved by expanding the logarithm of the integrand of the likelihood defined in (5.1) as a quadratic Taylor expansion about its maximum and then applying Laplace approximation to the integrals over the random effects.

Let the log-likelihood of the observations in the ith cluster be written as $l_i = c\int e^{-k(\alpha_i)}d\alpha_i$. Breslow & Clayton (1993) proposed a Laplace-based penalised quasi-likelihood defined as

$$l_P(\beta,\sigma^2) = \sum_{i=1}^{q}\left(\tilde{l}_i - \frac{\{\alpha_i^{(\max)}\}^2}{2\sigma^2}\right),$$

where $\alpha_i^{(\max)}$ denotes the solution to $\partial k(\alpha_i)/\partial\alpha_i = 0$, $\tilde{l}_i = l_i(\beta,\alpha_i^{(\max)})$ is the log-conditional density at the maximising value $\alpha_i^{(\max)}$, and $\sum_{i=1}^{q}\{\alpha_i^{(\max)}\}^2/(2\sigma^2)$ is a penalty term. The PQL estimators $(\hat{\beta}_{\mathrm{PQL}},\hat{\sigma}^2_{\mathrm{PQL}})^{\mathrm{T}}$ simultaneously solve the mean and variance score equations

$$\begin{aligned}\frac{\partial l_P(\beta,\sigma^2)}{\partial\beta} &= 0;\\ \frac{1}{2}\sum_{i=1}^{q}\left(\tilde{l}_i^{(1)2} + \frac{\tilde{l}_i^{(2)}}{1-\sigma^2\tilde{l}_i^{(2)}}\right)\Bigg|_{\beta=\hat{\beta}_{\mathrm{PQL}}} &= 0,\end{aligned}$$

where $\tilde{l}_i^{(k)} = l_i^{(k)}(\beta,\alpha_i^{(\max)})$ and $l_i^{(k)}(\beta,\alpha_i) = \partial^k l_i/\partial\alpha_i^k$. Breslow & Clayton (1993) showed that this approximation leads eventually to estimating equations based on PQL for the mean parameters and pseudo-likelihood for the variances.

Even though this estimation procedure is relatively simple and fast to implement it is known to produce biased estimates, especially in generalised linear mixed models with a small number of observations for each random effect (Breslow & Lin, 1995; Lin & Breslow, 1996). Also, because we compute a quasi-likelihood rather than a true likelihood, this makes PQL not directly useful for the definition of pivotal quantities for hypothesis testing and confidence intervals.

5.3.2 Improved estimation methods

The problem of inaccurate estimation has led to the development of several proposals for reducing the bias.

(i) Approximate parametric bootstrap

A popular method for correcting bias is parametric bootstrap (Efron & Tibshirani, 1993, Chapter 10). In the framework of generalised linear mixed models the maximum approximate likelihood estimates can be used in the computation of the approximate parametric bootstrap estimates. The bias of the maximum approximate likelihood estimator $\tilde{\theta}$ is estimated as $\tilde{B}^{(\text{boot})} = \bar{\theta} - \tilde{\theta}$, where $\bar{\theta}$ is the average of the maximum approximate likelihood estimates based on each of the R bootstrap samples. The approximate parametric bootstrap estimate is calculated as $\tilde{\theta}_{\text{boot}} = \tilde{\theta} - \tilde{B}^{(\text{boot})} = 2\tilde{\theta} - \bar{\theta}$.

(ii) Corrected penalised quasi-likelihood

Breslow & Lin (1995) and Lin & Breslow (1996) studied the bias of penalised quasi-likelihood estimators. They showed that the size of the asymptotic bias can be serious when the random effects have large variance and the cluster size is small, and developed bias correction methods for the regression parameters and the variance components. The simulation studies in Lin & Breslow (1996) demonstrate that correction to the PQL regression coefficient estimates fails to reduce bias unless the amount of dispersion is small. When the magnitude of dispersion is moderate or large, Lin & Breslow (1996) recommend only correction of the variance components and recalculation of the regression coefficients using the corrected PQL variance components.

(iii) Iterative bootstrap

Kuk (1995) proposed a method of adjusting initial estimates to yield consistent estimates, via a computationally intensive, iterated version of bootstrap which gives asymptotically consistent and unbiased estimates. To obtain an initial estimate $\theta^{(0)}$

Kuk (1995) suggest maximising the log-likelihood function $l(\theta; y, \alpha^{(0)})$ with respect to θ, where $\alpha^{(0)}$ is some imputed value of the unobserved α that can depend on both y and θ. The iterative bootstrap estimate results from iterative bias correction for $\theta^{(0)}$, where in each iteration an updated estimate of bias of $\theta^{(0)}$ leads to an updated bias-corrected estimate of θ. Let $b^{(k)}$ be the updated estimate of bias at the kth iteration and $h(\theta)$ be the asymptotic mean of $\theta^{(0)}$. The iterative procedure is

$$\begin{aligned} b^{(k+1)} &= h(\theta^{(k)}) - \theta^{(k)}; \\ \theta^{(k+1)} &= \theta^{(0)} - b^{(k+1)}, \end{aligned}$$

where $b^{(0)}$ is set to 0 so that the initial candidate estimate for the iterative bootstrap estimate is $\theta^{(0)}$. The function $h(\theta)$ can be approximated by $\hat{h}_R(\theta)$, the average of the $\theta^{(0)}$ values calculated for each of R simulated samples. In each iterative step a new set of R bootstrap samples is generated from the model at $\theta^{(k)}$ and then we subtract $\theta^{(k)}$ from the new mean bootstrap parameter estimates to obtain updated bias estimates. These bias estimates are then subtracted from the initial estimates $\theta^{(0)}$ to obtain a new set of bias-corrected estimates. The cycle is continued until some appropriate convergence criterion is satisfied.

5.4 Logistic random intercept model

In this section we apply the adjusted score function approach (Firth, 1993) to the binomial-response generalised linear model with logistic link and a random intercept. Implementing the adjusted score function approach on this simple model identifies the prohibitive challenges involved with directly implementing the vanilla method to more complex mixed models. We also conduct a simulation study to compare the performance of the adjusted score function approach with IBLA.

5.4.1 The traditional adjusted score equations

The binomial-response generalised linear mixed model with logistic link and a random intercept is defined as

$$\operatorname{logit}(\pi_i) = \beta + \alpha_i, \tag{5.2}$$

where β is the intercept of the model and α_i is the effect for the ith subject, $i \in \{1,\ldots,q\}$, assumed to be normally distributed with mean 0 and variance σ^2. The random variable $Y = (Y_1,\ldots,Y_q)^{\mathrm{T}}$ is assumed to consist of conditionally independent elements, each following a binomial distribution with m_i number of trials and success probability π_i in each trial. Let $\theta = (\beta,\sigma^2)^{\mathrm{T}}$. The log-likelihood of the binomial-response logistic random intercept model is

$$l(\theta) = \sum_{i=1}^{q} \log\left(\int \binom{m_i}{y_i}\left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)^{y_i}\left(\frac{1}{1+e^{\alpha_i+\beta}}\right)^{m_i-y_i}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}d\alpha_i\right). \quad (5.3)$$

We recall from (1.5) that the adjusted score function (Firth, 1993) can be expressed as

$$s_t^*(\theta) = s_t(\theta) + \frac{1}{2}\mathrm{tr}[\{i(\theta)\}^{-1}\{P_t(\theta)+Q_t(\theta)\}],$$

where $t=1$ corresponds to parameter β and $t=2$ corresponds to parameter σ^2. The score function for parameter θ is derived analytically in Appendix D.1 and has elements $s_1(\theta) = \sum_{i=1}^{q}(y_i - m_i E(\pi_i|y_i))$ and $s_2(\theta) = \sum_{i=1}^{q}\left(-(2\sigma^2)^{-1} + (2\sigma^4)^{-1}E(\alpha_i^2|y_i)\right)$, where

$$\begin{aligned} E(\pi_i|y_i) &= \int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i)d\alpha_i\,, \\ E(\alpha_i^2|y_i) &= \int \alpha_i^2 P(\alpha_i|y_i)d\alpha_i\,. \end{aligned}$$

In Appendix D.1 we also derive the adjusted score functions under the assumption of balanced data (i.e. $m_i = m$ for all i). These can be expressed as

$$\begin{aligned} s_1^*(\theta) &= s_1(\theta) + \frac{1}{2}\left(\frac{Q_1(\theta)}{|i(\theta)|}+1\right) \\ &= \sum_{i=1}^{q}\left[y_i + \frac{1}{2q} - mE(\pi_i|y_i) + \frac{1}{2q}\frac{Q_1(\theta)}{|i(\theta)|}\right] \qquad (5.4) \\ s_2^*(\theta) &= s_2(\theta) + \frac{1}{2}\left(\frac{Q_2(\theta)}{|i(\theta)|} - \frac{1}{2\sigma^2}\right) \\ &= \sum_{i=1}^{q}\left[-\frac{1}{2\sigma^2} - \frac{1}{4\sigma^2 q} + \frac{1}{2\sigma^4}E(\alpha_i^2|y_i) + \frac{1}{2q}\frac{Q_2(\theta)}{|i(\theta)|}\right]. \qquad (5.5) \end{aligned}$$

The quantities $Q_1(\theta)/|i(\theta)|$ and $Q_2(\theta)/|i(\theta)|$ do not depend on q and have

$$\begin{aligned}
Q_1(\theta) &= \frac{q^2 m}{4\sigma^8}\Big\{ m E^2_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] - m E^2_{Y_i}[E(\alpha_i^2\pi_i|y_i)] + \Big(E_{Y_i}[E^2(\alpha_i^2|y_i)] - \sigma^4\Big) \\
&\times \Big(-2E_{Y_i}[E(\pi_i^2|y_i)] + 2mE_{Y_i}[y_i E^2(\pi_i|y_i)] - m(m+1)E_{Y_i}[E(\pi_i|y_i)E(\pi_i^2|y_i)] \\
&- (m+1)(m-2)E_{Y_i}[E(\pi_i^3|y_i)]\Big) + 2m\Big(E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]\Big) \\
&\times \Big(E_{Y_i}[E(\alpha_i^2\pi_i^2|y_i)] - mE_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2\pi_i|y_i)] + E_{Y_i}[y_i E(\pi_i|y_i)E(\alpha_i^2|y_i)]\Big) \\
&- 4\sigma^2 m\Big(E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]\Big) \\
&\times \Big(E_{Y_i}[E(\pi_i|y_i)] - (m+1)E_{Y_i}[E(\pi_i^2|y_i)] + mE_{Y_i}[E^2(\pi_i|y_i)]\Big) \\
&- m\Big(E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^4|y_i)] - E_{Y_i}[E(\alpha_i^4\pi_i|y_i)]\Big) \\
&\times \Big(E_{Y_i}[E(\pi_i|y_i)] - (m+1)E_{Y_i}[E(\pi_i^2|y_i)] + mE_{Y_i}[E^2(\pi_i|y_i)]\Big)\Big\}; \\
Q_2(\theta) &= \frac{q^2 m}{8\sigma^{12}}\Big\{ m\sigma^2\Big(E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]\Big)^2 \\
&+ \Big(E_{Y_i}[E^2(\alpha_i^2|y_i)] - \sigma^4\Big)\Big(E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] \\
&- 2E_{Y_i}[y_i E(\pi_i|y_i)E(\alpha_i^2|y_i)] + (m+1)E_{Y_i}[E(\pi_i^2|y_i)E(\alpha_i^2|y_i)] + (m-1)E_{Y_i}[E(\alpha_i^2\pi_i^2|y_i)]\Big) \\
&- 2\Big(E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i))\Big)\Big(E_{Y_i}[y_i E^2(\alpha_i^2|y_i)] \\
&- mE_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^2\pi_i|y_i)]\Big) + \Big(E_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^4|y_i)] - 5\sigma^2 E_{Y_i}[E^2(\alpha_i^2|y_i)] + 2\sigma^6\Big) \\
&\times \Big(E_{Y_i}[E(\pi_i|y_i)] - (m+1)E_{Y_i}[E(\pi_i^2|y_i)] + mE_{Y_i}[E^2(\pi_i|y_i)]\Big)\Big\},
\end{aligned}$$

and the determinant of the expected information matrix is

$$\begin{aligned}
|i(\theta)| &= \frac{q^2 m}{4\sigma^8}\Big\{ -m\left(E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]\right)^2 \\
&+ \left(E_{Y_i}[E^2(\alpha_i^2|y_i)] - \sigma^4\right)\left(E_{Y_i}[E(\pi_i|y_i)] - (m+1)E_{Y_i}[E(\pi_i^2|y_i)] + mE_{Y_i}[E^2(\pi_i|y_i)]\right)\Big\}.
\end{aligned}$$

Based on the theory in Kosmidis & Firth (2009), we expect the quantity $Q_1(\theta)/|i(\theta)|$ to approach $-2\pi(\beta)$, where $\pi(\beta) = e^\beta/(1+e^\beta)$, as we let σ^2 go to zero, i.e. we are under the framework of generalised linear models. Kosmidis & Firth (2009) expressed the adjusted score function for the regression coefficients of a binomial-response generalised linear model via a pseudo-data representation as $s^*(\beta) = \sum_{i=1}^{q}\left(y_i + h_i(\beta)/2 - \left(m_i + h_i(\beta)\right)\pi_i(\beta)\right)x_{ir}$, where h_i is the ith diagonal element of the hat matrix $H = X(X^TWX)^{-1}X^TW$ and $W = \text{diag}\{\kappa_{2i}\}$, κ_{2i} being the variance of y_i. For the simple case of a generalised linear model with $\text{logit}(\pi_i) = \beta$, $s^*(\beta)$ takes the form

$$s^*(\beta) = \sum_{i=1}^{q} \left[y_i + \frac{1}{2q} - \left(m_i + \frac{1}{q} \right) \pi_i \right]. \tag{5.6}$$

Comparing the functions in (5.4) and (5.6) it is clear that if the derivations above are correct, $Q_1(\theta)/|i(\theta)|$ should approach $-2\pi(\beta)$ as σ^2 approaches zero.

When the variance of a normal distribution tends to zero, the probability density $f_X(x)$, with $X \sim N(\mu, \sigma^2)$, eventually tends to zero at any $x \neq \mu$, but grows without limit if $x = \mu$, while its integral remains equal to 1. The Dirac delta function can be viewed as a limit of the Gaussian distribution, with

$$\delta(x-\mu) = \lim_{\sigma^2 \to 0} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \begin{cases} 0, & \text{if } x \neq \mu \\ \infty, & \text{if } x = \mu \end{cases} \tag{5.7}$$

and is constrained to satisfy the identity $\int_{-\infty}^{\infty} \delta(x-\mu) d(x-\mu) = 1$. Using (5.7), basic properties of limits, and applying Lebesgue's dominated convergence theorem (Rudin, 1976, p. 318), we obtain Results 1-5, which are proved in Appendix D.2:

Result 1. $\lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] = \lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\alpha_i^2 \pi_i|y_i)]$.

Result 2. $\lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^4|y_i)] = \lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\alpha_i^4 \pi_i|y_i)]$.

Result 3. $\lim\limits_{\sigma^2 \to 0} E_{Y_i}[E^2(\pi_i|y_i)] = \lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i^2|y_i)]$.

Result 4. $\lim\limits_{\sigma^2 \to 0} E_{Y_i}[y_i E^2(\pi_i|y_i)] = m_i \lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i^3|y_i)]$.

Result 5. $\lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i|y_i)E(\pi_i^2|y_i)] = \lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i^3|y_i)]$.

As expected, using Results 1-5 we obtain

$$\lim_{\sigma^2 \to 0} \frac{Q_1(\theta)}{|i(\theta)|} = \lim_{\sigma^2 \to 0} \frac{-2E_{\alpha_i}(\pi_i^2) + 2E_{\alpha_i}(\pi_i^3)}{E_{\alpha_i}(\pi_i) - E_{\alpha_i}(\pi_i^2)} = -2\left(\frac{\pi(\beta)^3 - \pi(\beta)^2}{\pi(\beta)^2 - \pi(\beta)} \right) = -2\pi(\beta).$$

5.4.2 Approximating the intractable integrals

The intractable integrals involved in the traditional adjusted score equations can be approximated by numerical integration. Numerical integration techniques include the Laplace approximation (Tierney & Kadane, 1986), the Gauss-Hermite quadrature approximation (Abramowitz & Stegun, 1965), and the adaptive Gauss-Hermite quadrature approximation (Liu & Pierce, 1994).

A quadrature rule is an approximation of an integral, usually stated as a weighted sum of function values at K specified points (nodes or abscissae) within the domain of integration. The Gauss-Hermite quadrature rule can be employed for approximating the value of integrals of the form $\int_{-\infty}^{\infty} e^{-x^2} f(x)\, dx$, as

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)\, dx \approx \sum_{k=1}^{K} \omega_k f(u_k), \tag{5.8}$$

where u_k are the roots of the Hermite polynomial $H_K(u)$, and ω_k are the weights with

$$H_K(u) = (-1)^K e^{u^2} \frac{d^K}{du^K} e^{-u^2}; \; \omega_k = \frac{2^{K-1} K! \sqrt{\pi}}{K^2 [H_{K-1}(u_k)]^2}\,.$$

The weights and nodes used in Gaussian quadrature rules can be obtained from Abramowitz & Stegun (1965, Table 25.8). Generally, using large values of K increases the accuracy but also the computational run-time of the approximation. Although Gauss-Hermite quadrature is easy to implement, it can completely miss the maximum of the integrand and, as a consequence, lead to biased estimators (Huber et al., 2004). The bias with Gauss-Hermite quadrature approximation is explained by the fact that it is based on the summation over prespecified and fixed quadrature points irrespective of the range where the function is concentrated.

An improvement of Gauss-Hermite quadrature and Laplace approximations is provided by an adaptive Gaussian quadrature which searches for the maximum of the integrand, and approximates adaptively the function in the neighbourhood of the maximum. Consequently, adaptive Gauss-Hermite quadrature improves efficiency by dramatically reducing the number of quadrature points needed to effectively approximate the integral than ordinary Gauss-Hermite quadrature. This technique centres and rescales the quadrature nodes and approximates $\int_{-\infty}^{\infty} g(x)dx$, where $g(x) = e^{q(x)}$, as

$$\int_{-\infty}^{\infty} g(x)dx \approx \sqrt{2}\hat{\sigma} \sum_{k=1}^{K} \omega_k^* g(\hat{\mu} + \sqrt{2}\hat{\sigma} u_k), \tag{5.9}$$

where $\omega_k^* = \omega_k e^{u_k^2}$ and the quadrature points are centred at the mode of the Laplace approximation, $\hat{\mu} = \max_x q(x)$, with spread determined by $\hat{\sigma} = \{-q''(\hat{\mu})\}^{-1/2}$ (Liu & Pierce, 1994). Adaptive Gauss-Hermite quadrature with one node becomes the Laplace approximation. This result follows immediately from (4.4) and (5.9) for one quadrature point given that the node and weight are equal to $u_1 = 0$ and $w_1 = \sqrt{\pi}$. Thus adaptive Gauss-Hermite quadrature can be thought of as a higher-order Laplace approximation.

An implicit assumption of the Laplace and adaptive Gauss-Hermite quadrature approximations is that the function $q(x)$ is unimodal. Some of the integrals in the adjusted score function derived in Section 5.4.1 are bimodal, because α_i^k, k being an even number, is involved in the integrand. For example,

$$\begin{aligned} E_{\alpha_i}(\alpha_i^2 \pi_i) &= \int \alpha_i^2 \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} f_{\alpha_i}(\alpha_i)\, d\alpha_i, \\ E_{\alpha_i}(\alpha_i^4 \pi_i) &= \int \alpha_i^4 \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} f_{\alpha_i}(\alpha_i)\, d\alpha_i, \end{aligned}$$

where $f_{\alpha_i}(\alpha_i)$ is the density function of α_i, are bimodal. As a result, their approximations may be poor whichever maximum point is taken.

Demidenko (2013, Chapter 7) proposed improving the Laplace approximation by splitting the domain of integration $(-\infty, \infty)$ into $(-\infty, c)$ and (c, ∞), where c is any point between the two maxima. The improved Laplace approximation is then expressed as

$$\int_{-\infty}^{\infty} e^{q(x)}dx \approx \sqrt{2\pi}\left\{\hat{\sigma}_1 e^{q(\hat{\mu}_1)}\Phi\left(\frac{c-\hat{\mu}_1}{\hat{\sigma}_1}\right) + \hat{\sigma}_2 e^{q(\hat{\mu}_2)}\left[1-\Phi\left(\frac{c-\hat{\mu}_2}{\hat{\sigma}_2}\right)\right]\right\}, \tag{5.10}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\hat{\mu}_1$, $\hat{\mu}_2$ are the two maxima points of $q(x)$ and $\hat{\sigma}_i = \{-q''(\hat{\mu}_i)\}^{-1/2}$, $i \in \{1, 2\}$.

Below we introduce a new approximation method for numerical integration which is an extension of the Liu & Pierce (1994) approximation and can be used for approximating integrals with bimodal integrands. For this reason, the method can be considered as an "improved adaptive Gauss-Hermite quadrature" approximation. Similarly to improved Laplace approximation, when the function $q(x)$ is bimodal we can split the integral into two integrals around each maximum.

Proposition 1. *The Kth-order improved adaptive Gauss-Hermite quadrature approximation of $\int_{-\infty}^{\infty} g(x)dx$ is*

$$\int_{-\infty}^{\infty} g(x)dx \approx \sqrt{2}\left[\hat{\sigma}_1 \sum_{k=1}^{K} w_k^* g(-\sqrt{2}\hat{\sigma}_1 u_k + c) + \hat{\sigma}_2 \sum_{k=1}^{K} w_k^* g(\sqrt{2}\hat{\sigma}_2 u_k + c)\right], \quad (5.11)$$

where $\hat{\sigma}_i$, $i \in \{1,2\}$, is the scale of the Laplace approximation calculated at the ith maximum point of $q(x)$ with $g(x) = e^{q(x)}$.

Proof of Proposition 1: Steen et al. (1969) and Galant (1969) developed Gaussian quadratures for the semi-infinite integral $\int_0^\infty e^{-x^2} f(x)dx \approx \sum_{k=1}^K w_k f(u_k)$, where w_k and u_k are the weights and abscissae given in Steen et al. (1969, Table II) for $K \in \{2,\dots,15\}$. Based on this result and using integration by substitution ($t = x - c$) we obtain the general form of the semi-infinite integral for any threshold c,

$$\begin{aligned}
\int_c^\infty e^{-x^2} f(x)dx &\approx \sum_{k=1}^{K} w_k f(u_k + c) e^{-(c^2+2cu_k)};\\
\int_{-\infty}^c e^{-x^2} f(x)dx &\approx \sum_{k=1}^{K} w_k f(-u_k + c) e^{-(c^2-2cu_k)}.
\end{aligned}$$

Following the Liu & Pierce (1994) methodology and using the above results we obtain

$$\begin{aligned}
\int_c^\infty g(x)dx &= \int_c^\infty \frac{g(x)}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx\\
&\overset{t=\frac{x-\mu}{\sqrt{2}\sigma}}{=} \int_{\frac{c-\mu}{\sqrt{2}\sigma}}^\infty \sqrt{2}\sigma e^{t^2} g(\mu + \sqrt{2}\sigma t) e^{-t^2} dt\\
&\approx \sqrt{2}\hat{\sigma} \sum_{k=1}^{K} w_k e^{\left(u_k + \frac{c-\hat{\mu}}{\sqrt{2}\hat{\sigma}}\right)^2} g(\hat{\mu} + \sqrt{2}\hat{\sigma} u_k + (c - \hat{\mu})) e^{-\left(\left(\frac{c-\hat{\mu}}{\sqrt{2}\hat{\sigma}}\right)^2 + 2\left(\frac{c-\hat{\mu}}{\sqrt{2}\hat{\sigma}}\right) u_k\right)}\\
&= \sqrt{2}\hat{\sigma} \sum_{k=1}^{K} w_k e^{u_k^2} g(\sqrt{2}\hat{\sigma} u_k + c), \qquad (5.12)
\end{aligned}$$

and, similarly,

$$\int_{-\infty}^{c} g(x)dx \approx \sqrt{2}\hat{\sigma}\sum_{k=1}^{K} w_k^* g(-\sqrt{2}\hat{\sigma}u_k + c). \tag{5.13}$$

Summing (5.12) and (5.13) yields (5.11). □

It is worth noting that eventually improved adaptive Gauss-Hermite quadrature approximation depends on the maxima points of $q(x)$ only through $\hat{\sigma}_i$. We can also derive an expression for improved adaptive Gauss-Hermite quadrature in the interval $[c_1, c_2]$, which can be used when the domain $(-\infty, \infty)$ needs to be split into more than two intervals.

Proposition 2. *The Kth-order improved adaptive Gauss-Hermite quadrature approximation of $\int_{c_1}^{c_2} g(x)dx$ is*

$$\int_{c_1}^{c_2} g(x)dx \approx (c_2 - c_1)\sum_{k=1}^{K} w_k^* g\left((c_2 - c_1)u_k + c_1\right). \tag{5.14}$$

Proof of Proposition 2: Using the transformation $x = (c_2 - c_1)t + c_1$ we obtain the standard Gauss-Hermite quadrature approximation

$$\int_{c_1}^{c_2} e^{-x^2} f(x)dx \approx (c_2 - c_1)\sum_{k=1}^{K} w_k e^{u_k^2} f((c_2 - c_1)u_k + c_1)e^{-((c_2-c_1)u_k+c_1)^2},$$

where w_k and u_k are the weights and abscissae given in Steen et al. (1969, Table III) for $K \in \{2, \dots, 10\}$. Using the same line of argument as in the proof of Proposition 1, we obtain the result in (5.14). □

To verify whether the first-order improved adaptive Gauss-Hermite quadrature and improved Laplace approximations are identical, we need to calculate the weight and abscissa of the former approximation when $K = 1$, because these are not reported in Steen et al. (1969). The abscissa $u_1 = 1/\sqrt{\pi}$ solves the first-order polynomial, $p_1(u) = u - (1 - e^{-b^2})/(\sqrt{\pi}\mathrm{erf}(b))$ when $b \to \infty$, where $\mathrm{erf}(b)$ is the error function and $\sqrt{\pi}\mathrm{erf}(b) = \int_{-b}^{b} e^{-t^2}dt$. The weight $w_1 = \sqrt{\pi/2}$ is calculated from the weight expression $w_1 = \gamma_0/[p_1'(u_1)p_0(u_1)]$, where $p_0(u) = 1$ and $\gamma_0 = \int_0^\infty e^{-u^2}du$. Using u_1

and w_1 in (5.11) we calculate the first-order improved adaptive quadrature approximation to be $\sqrt{\pi}e^{1/\pi}\left[\hat{\sigma}_1 e^{q(-\sqrt{2}\hat{\sigma}_1/\sqrt{\pi}+c)} + \hat{\sigma}_2 e^{q(\sqrt{2}\hat{\sigma}_2/\sqrt{\pi}+c)}\right]$. This result is different from the improved Laplace approximation in (5.10), and therefore, the improved adaptive Gauss-Hermite quadrature cannot be seen as a higher-order improved Laplace approximation.

Summing up, the most accurate numerical integration technique between the Laplace approximation, the Gauss-Hermite quadrature and the adaptive Gauss-Hermite quadrature approximations, is the latter, whereas the fastest technique is Laplace approximation (Liu & Pierce, 1994; Demidenko, 2013, Chapter 7).

Example 3. (Illustration of a bimodal integral) We consider the example in Demidenko (2013, p. 343) to illustrate the performance of improved Laplace and improved adaptive Gauss-Hermite quadrature approximations by approximating the integral

$$\int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}(x^2+x)} dx$$

for which $q(x) = \frac{1}{2}(x^2+x) - 2\log|x|$. The function $-q(x)$ has two local maxima points, $\hat{\mu}_1 = -1.686$ and $\hat{\mu}_2 = 1.186$ (Figure 5.1). For this integral it is natural to set the threshold c to zero. Table 5.1 shows the values of the integral for the approximation techniques under consideration using various quadrature points when applicable. The results show that integration must be carried out with care when a function is bimodal. Standard Laplace and adaptive Gauss-Hermite quadrature fail to estimate the integral unless a large number of quadrature points is used. On the other hand, improved Laplace yields an approximation of 3.62 which is close to the true value of 3.55. Improved adaptive Gauss-Hermite quadrature yields even better results approaching the true value of the integral with the use of just 5 quadrature points and reaching the exact value in 3 decimal places for 10 quadrature points. Lastly, as expected, we observe that the values calculated using adaptive Gauss-Hermite quadrature with one node are identical to the ones obtained using Laplace approximation, but this does not extend to the improved versions of the two approximations.

Figure 5.1: An example with bimodal function, $-q(x) = 2\log|x| - \frac{1}{2}(x^2 + x)$.

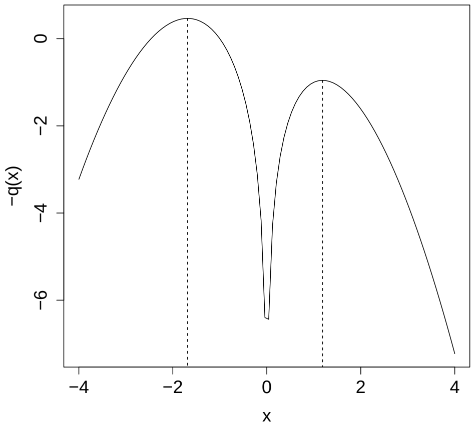

Table 5.1: Approximations of $\int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}(x^2+x)} dx$

Approximation Method	Details	Result
Laplace	using 1st maximum	3.062
Laplace	using 2nd maximum	0.620
improved Laplace		3.619
adaptive GHQ	1 quadrature point using 1st maximum	3.062
adaptive GHQ	1 quadrature point using 2nd maximum	0.620
improved adaptive GHQ	1 quadrature point	1.065
adaptive GHQ	2 quadrature points using 1st maximum	2.995
adaptive GHQ	2 quadrature points using 2nd maximum	0.592
improved adaptive GHQ	2 quadrature points	2.388
adaptive GHQ	5 quadrature points using 1st maximum	3.187
adaptive GHQ	5 quadrature points using 2nd maximum	1.097
improved adaptive GHQ	5 quadrature points	3.519
adaptive GHQ	10 quadrature points using 1st maximum	3.540
adaptive GHQ	10 quadrature points using 2nd maximum	3.003
improved adaptive GHQ	10 quadrature points	3.550
adaptive GHQ	15 quadrature points using 1st maximum	3.550
adaptive GHQ	15 quadrature points using 2nd maximum	3.493
improved adaptive GHQ	15 quadrature points	3.550

Notes: The exact value of the integral is 3.550. The Laplace and adaptive Gauss-Hermite quadrature (adaptive GHQ) approximations at the first or second minimum point of $q(x)$ are calculated using (4.4) and (5.9), respectively. The improved versions of these approximations are given in (5.10) for the Laplace and (5.11) for the improved adaptive Gauss-Hermite quadrature.

5.4.3 Simulation study

This section compares via simulations the performance of the adjusted score equation approach (Firth, 1993) with IBLA and maximum approximate likelihood.

The maximum approximate likelihood estimates are calculated by maximising the Laplace approximation of the log-likelihood. The IBLA estimates are calculated by solving the tractable simulation-based adjusted score equations in (4.2) using the algorithm in Section 4.4, where the first and second derivatives of the Laplace approximation of the log-likelihood are used, the maximum approximate likelihood estimates are set to be the starting values, and R is set to 50. Recall from Section 4.5 that Laplace approximation is suitable for bias reduction through the adjustment of the approximate score function. The unimodal integrals involved in the adjusted score equations derived in Section 5.4.1 are approximated using the adaptive Gauss-Hermite quadrature (Liu & Pierce, 1994) with $K = 20$, and the bimodal integrals are approximated using the improved adaptive Gauss-Hermite quadrature with $K = 15$.

In this simulation study we simulated the data from the binomial-response generalised linear mixed model with logistic link and a random intercept with true fixed-effect parameter $\beta = 0.5$. We assume that the data is balanced, with binomial denominator $m = 10$, and the number of clusters is set to $q = 20$. Five values of the variance component σ^2 are chosen, specifically $\sigma^2 \in \{0.25, 0.5, 1.0, 1.5, 2.0\}$. For each value of σ^2 considered, we simulated 10 000 data sets initialising the random number generator at a common state. The maximum approximate likelihood, adjusted score equation approach, and IBLA are evaluated in terms of mean bias, mean squared error, and coverage probability. The results of the simulation study are summarised in Table 5.2.

The results illustrate the underestimation of the variance component by Laplace-based maximum likelihood, which increases as random effects become more heterogeneous. Laplace-based IBLA and the adjusted score equation approach based on adaptive quadrature approximation reduce the bias of the maximum approximate likelihood estimates, with the latter yielding the smallest mean bias. Our explanation for the better improvement of the estimation of variance components by the traditional adjusted score equations is that it relies on finding the roots of an adjusted version of the scores based on the theory in Firth (1993) rather than finding the roots of an objective function. Also, a more precise approximation technique has been used in the adjusted score equations compared to the Laplace approximation used in IBLA. The bias of the fixed effect estimates is close to zero for all methods. Comparing the mean squared errors, we observe that the three methods yield similar mean squared errors for the fixed effect, whereas for the variance component maximum approximate likelihood has the

Table 5.2: Mean bias and mean squared error (MSE) for the parameters of the binomial-response generalised linear mixed model with logistic link and a random intercept, and empirical coverage probability of 90%, 95% and 99% confidence intervals for β based on the Wald statistic.

Parameter	True value	Method	Bias	MSE	Coverage probability (%)		
					90%	95%	99%
β	0.5	La-ML	0.003	0.036	88.6	93.8	98.5
		adGHQ-AS	0.002	0.036	89.4	94.3	98.6
		La-IBLA	-0.002	0.037	89.0	93.9	98.6
σ^2	0.25	La-ML	-0.018	0.050	-	-	-
		adGHQ-AS	0.005	0.056	-	-	-
		La-IBLA	-0.007	0.058	-	-	-
β	0.5	La-ML	0.002	0.048	89.0	94.0	98.4
		adGHQ-AS	0.001	0.048	89.7	94.5	98.5
		La-IBLA	-0.003	0.047	89.4	94.2	98.5
σ^2	0.5	La-ML	-0.038	0.114	-	-	-
		adGHQ-AS	-0.003	0.121	-	-	-
		La-IBLA	-0.010	0.127	-	-	-
β	0.5	La-ML	0.006	0.076	88.9	94.1	98.3
		adGHQ-AS	0.005	0.075	89.5	94.5	98.3
		La-IBLA	-0.004	0.075	89.5	94.7	98.4
σ^2	1.0	La-ML	-0.048	0.338	-	-	-
		adGHQ-AS	-0.001	0.346	-	-	-
		La-IBLA	0.015	0.355	-	-	-
β	0.5	La-ML	0.007	0.102	89.1	94.1	98.5
		adGHQ-AS	0.004	0.101	89.8	94.5	98.7
		La-IBLA	-0.005	0.102	90.2	95.2	98.8
σ^2	1.5	La-ML	-0.060	0.668	-	-	-
		adGHQ-AS	-0.006	0.675	-	-	-
		La-IBLA	0.024	0.732	-	-	-
β	0.5	La-ML	0.008	0.132	89.1	94.0	98.5
		adGHQ-AS	0.004	0.130	89.5	94.4	98.7
		La-IBLA	-0.007	0.131	89.8	94.8	98.8
σ^2	2.0	La-ML	-0.063	1.108	-	-	-
		adGHQ-AS	-0.006	1.113	-	-	-
		La-IBLA	0.021	1.185	-	-	-

Notes: La-ML, Laplace-based maximum likelihood; adGHQ-AS, adjusted score equation approach (Firth, 1993) using (improved) adaptive Gauss-Hermite quadrature; La-IBLA, Laplace-based iterated bootstrap with likelihood adjustment with $R = 50$.

smallest mean squared error, followed closely by the adjusted score equation approach and IBLA. Lastly, the empirical coverage probabilities of the 90%, 95% and 99% confidence intervals for β based on the Wald statistic and the mean bias-reduced estimates are closer to the nominal level than those based on the Wald statistic and the maximum approximate likelihood estimates. The conservativeness of the latter inferential procedure is a result of the underestimation of the variance components by maximum approximate likelihood.

To sum up, IBLA is found to be a good alternative for reducing the bias of maximum approximate likelihood estimates, having the additional advantage of being easier to apply in more complex mixed models.

5.4.4 Challenges with adjusted score functions

In Section 5.4.1 we have seen how the approximation of the first-order bias can be achieved via the direct approximation of a number of expectations (see Appendix D.1), because the model under consideration is relatively simple. However, in general, the requirement for joint null moments of the log-likelihood derivatives makes the calculation of the adjusted score functions proposed in Firth (1993) challenging.

Consider, for example, adding a single continuous covariate to the binomial-response generalised linear mixed model with logistic link and a random intercept, i.e.

$$\operatorname{logit}(\pi_{ij}) = \beta x_{ij} + \alpha_i\,, \tag{5.15}$$

with $y_{ij}|\alpha_i \sim \text{Binomial}(m_i, \pi_{ij})$ and $\alpha_i \sim N(0, \sigma^2)$, for $i \in \{1, \ldots, q\}$, $j \in \{1, \ldots, n_i\}$. The log-likelihood for this model is

$$l(\beta, \sigma^2) = \sum_{i=1}^{q} \log\left(\int \prod_{j=1}^{n_i} \binom{m_i}{y_{ij}} \frac{e^{y_{ij}(\alpha_i + \beta x_{ij})}}{(1 + e^{\alpha_i + \beta x_{ij}})^{m_i}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\alpha_i^2}{2\sigma^2}} d\alpha_i\right). \tag{5.16}$$

The score functions are

$$\begin{aligned} s_1(\theta) &= \sum_{i=1}^{q}\sum_{j=1}^{n_i} \left[x_{ij}y_{ij} - m_i x_{ij} E(\pi_{ij}|y_{ij})\right]\,, \\ s_2(\theta) &= \sum_{i=1}^{q}\sum_{j=1}^{n_i} \left[\frac{1}{2\sigma^4} E(\alpha_i^2|y_{ij}) - \frac{1}{2\sigma^2}\right]\,, \end{aligned}$$

and the observed information matrix elements are

$$\begin{aligned} j_{11}(\theta) &= \sum_{i=1}^{q}\sum_{j=1}^{n_i} \left[m_i x_{ij}^2 E(\pi_{ij}|y_{ij}) - m_i x_{ij}(m_i + x_{ij}) E(\pi_{ij}^2|y_{ij}) + m_i^2 x_{ij} \{E(\pi_{ij}|y_{ij})\}^2\right]\,, \\ j_{12}(\theta) &= \sum_{i=1}^{q}\sum_{j=1}^{n_i} \left[\frac{m_i}{2\sigma^4} E(\pi_{ij}|y_{ij}) E(\alpha_i^2|y_{ij}) - \frac{m_i}{2\sigma^4} E(\alpha_i^2 \pi_{ij}|y_{ij})\right]\,, \\ j_{22}(\theta) &= \sum_{i=1}^{q}\sum_{j=1}^{n_i} \left[-\frac{1}{2\sigma^4} + \frac{1}{\sigma^6} E(\alpha_i^2|y_{ij}) - \frac{1}{4\sigma^8} E(\alpha_i^4|y_{ij}) + \frac{1}{4\sigma^8}\{E(\alpha_i^2|y_{ij})\}^2\right]. \end{aligned}$$

Obtaining the expectations involved in the expected information matrix and the adjusted score function requires computationally more expensive calculations. In Section 5.4.1 we studied the simple generalised linear mixed model with binomial responses, a random intercept and a fixed intercept on the logit scale. For that model we had to approximate one integral for each expectation. Adding covariates in the linear predictor of the generalised linear mixed model requires the approximation of n_i integrals for each expectation. For example, one of the integrals we had to approximate for model (5.2) was

$$E(\pi_i|y_i) = \int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} P(\alpha_i|y_i)d\alpha_i \,.$$

The corresponding integral for model (5.15) is

$$E(\pi_{ij}|y_{ij}) = \int \frac{e^{\alpha_i+\beta x_{ij}}}{1+e^{\alpha_i+\beta x_{ij}}} P(\alpha_i|y_i)d\alpha_i \,,$$

which needs to be approximated n_i times, one for each $j \in \{1,\ldots,n_i\}$. Also, for more complicated structures (e.g. crossed random effects) the integrals involved in the adjusted score function are no longer one-dimensional and numerical integration becomes more difficult (McCulloch et al., 2008, Section 7.4).

This study illustrates the challenges in evaluating the traditional adjusted score function for generalised linear mixed models and highlights the extent of the need for an alternative way of approximating the bias function, which would enable the implementation of an adjusted score function approach in models with more complicated linear predictors. The simulations in Section 5.4.3 indicate that a prominent extension of the traditional adjusted score function for models with intractable likelihood is the tractable simulation-based adjusted score function introduced in Chapter 4, using IBLA for the computation of the bias-reduced estimates.

5.5 Real-data examples

This section considers two real-data examples that are used to evaluate the performance of IBLA in estimation and inference against the maximum approximate likelihood, PQL, corrected PQL, and approximate parametric bootstrap methods. The data sets used are the multicenter clinical trial (Beitler & Landis, 1985) and the Culcita sea stars (McKeon et al., 2012). The responses of the first data set follow a binomial distribution and the responses of the second data set are binary. We fit a logistic linear model with a random intercept to both data sets.

Both examples are challenging in terms of model fitting. The multicenter clinical trial (Beitler & Landis, 1985) example is a small dataset with only 16 observations, and the Culcita sea stars (McKeon et al., 2012) example is challenging because the variance of the random effects is large ($\hat{\sigma}^2 = 11.8$). Given the challenging format of the data sets, there is a large possibility of convergence issues for all the methods under consideration. Convergence issues are generated, for example, when the gradient at the estimated values is not equal to zero and the Hessian matrix is not positive definite. One could argue that convergence failures are considered to be a minor issue and choose to discard any samples that give convergence issues. However, it is generally bad practice to discard samples as this can skew and bias the results when estimating the distribution of the estimators. Our strategy for avoiding any convergence issues when fitting the model to the multicenter clinical trial (Beitler & Landis, 1985) and the Culcita sea stars (McKeon et al., 2012) data sets is to estimate the standard deviation, instead of the variance, of the random effects.

5.5.1 Multicenter clinical trial

Beitler & Landis (1985) consider a multicenter clinical trial which investigates the success of two topical cream treatments (active drug and control) in curing an infection. The number of trials and favorable cures were recorded for each treatment for a total of 8 clinics. Table 5.3 shows the data of favorable response to active drug and control treatment from the multicenter randomised clinical trial. The clinics are ordered arbitrarily according to decreasing sample sizes. In all clinics, except for the eighth (which has the smallest number of patients), the drug produced a higher proportion of favorable responses, albeit only slightly higher than the control therapy in some clinics.

For this data we consider the generalised linear mixed model with linear predictor $\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + \alpha_i$, where β_0 is the fixed intercept, β_1 is the fixed effect associated with the treatment indicator variable x_{ij} that takes value 0 for control and 1 for the active drug, and α_i is the normally distributed random intercept with zero mean and variance σ^2 that takes into account heterogeneity between clinics. The subindex

Table 5.3: Data for the multicenter randomised clinical trial example

		Response			
Clinic	Treatment	Favourable	Unfavourable	Total	Proportion favorable
1	Drug	11	25	36	0.306
	Control	10	27	37	0.270
2	Drug	16	4	20	0.800
	Control	22	10	32	0.688
3	Drug	14	5	19	0.737
	Control	7	12	19	0.368
4	Drug	2	14	16	0.125
	Control	1	16	17	0.059
5	Drug	6	11	17	0.353
	Control	0	12	12	0.000
6	Drug	1	10	11	0.091
	Control	0	10	10	0.000
7	Drug	1	4	5	0.200
	Control	1	8	9	0.111
8	Drug	4	2	6	0.667
	Control	6	1	7	0.857

$i \in \{1, \ldots, 8\}$ indicates the clinic and the subindex $j \in \{1,2\}$ indicates the treatment. The observations are assumed to be realisations of random variables which are independent conditionally on the random effects, following a binomial distribution with m_{ij} number of trials and success probability π_{ij} in each trial. The binomial denominators m_{ij} range between 5 and 37.

Table 5.4 gives the parameter estimates of β_0, β_1, and σ, when the model is fitted by maximum approximate likelihood, PQL, corrected PQL, approximate parametric bootstrap, and IBLA. The maximum approximate likelihood, approximate parametric bootstrap and IBLA methods use the Laplace approximation. The maximum approximate likelihood yields the smallest estimate for σ. This might be an indication of the underperformance of the method which can be explained by the small sample size of the dataset. According to Breslow & Lin (1995), maximum approximate likelihood is expected to perform better than PQL when the outcomes are binomials having moderately large denominators, and corrected PQL is expected to behave similarly to PQL. The approximate parametric bootstrap and IBLA also yield larger estimates of σ than maximum approximate likelihood. The standard errors of the estimated parameters are similar for all methods except for the approximate parametric bootstrap which yields smaller standard errors for $R = 64$. The computational run-times for the methods under consideration are, in increasing order, $0.04, 0.30, 1.42, 4.68, 7.81, 8.31, 11.89$ seconds

Table 5.4: Estimates of the model parameters for the multicenter clinical trial data. The estimated standard errors are reported in parentheses.

	La-ML	PQL	CPQL	La-BOOT		La-IBLA	
R	-	-	-	64	128	64	128
β_0	-1.197	-1.147	-1.148	-1.215	-1.205	-1.221	-1.218
	(0.553)	(0.559)	(0.564)	(0.502)	(0.538)	(0.580)	(0.599)
β_1	0.739	0.726	0.727	0.732	0.719	0.715	0.728
	(0.300)	(0.296)	(0.296)	(0.257)	(0.305)	(0.303)	(0.303)
σ	1.390	1.426	1.442	1.560	1.530	1.450	1.524

Notes: La-ML, Laplace-based maximum likelihood; PQL, penalised quasi-likelihood; CPQL, corrected penalised quasi-likelihood; La-BOOT, Laplace-based parametric bootstrap; La-IBLA, Laplace-based iterated bootstrap with likelihood adjustment.

Table 5.5: Mean bias and mean squared error (MSE) for the parameters of the logistic mixed model for the multicenter clinical trial setting, and empirical coverage probability of 90%, 95% and 99% confidence intervals for β_0 and β_1 based on the Wald statistic.

				La-ML	PQL	CPQL	La-BOOT		La-IBLA	
		$(1-\alpha)\%$	R	-	-	-	64	128	64	128
β_0	Bias	-		-0.012	0.039	0.037	-0.031	-0.013	-0.023	-0.010
	MSE	-		0.328	0.286	0.287	0.284	0.315	0.315	0.307
	Coverage	90		0.858	0.864	0.868	0.826	0.836	0.880	0.887
		95		0.911	0.914	0.916	0.888	0.899	0.932	0.937
		99		0.968	0.965	0.966	0.958	0.962	0.975	0.977
β_1	Bias	-		0.006	-0.010	-0.009	-0.012	-0.017	-0.020	-0.013
	MSE	-		0.100	0.096	0.096	0.098	0.100	0.098	0.100
	Coverage	90		0.900	0.894	0.894	0.884	0.886	0.904	0.899
		95		0.953	0.945	0.945	0.942	0.946	0.955	0.951
		99		0.989	0.985	0.985	0.986	0.987	0.992	0.991
σ	Bias	-		-0.123	-0.091	-0.076	0.068	0.017	-0.062	-0.006
	MSE	-		0.229	0.193	0.195	0.217	0.220	0.218	0.212

Notes: La-ML, Laplace-based maximum likelihood; PQL, penalised quasi-likelihood; CPQL, corrected penalised quasi-likelihood; La-BOOT, Laplace-based parametric bootstrap; La-IBLA, Laplace-based iterated bootstrap with likelihood adjustment

for the PQL, corrected PQL, maximum approximate likelihood, approximate parametric bootstrap with $R = 64$, IBLA with $R = 64$, approximate parametric bootstrap with $R = 128$, and IBLA with $R = 128$, respectively. Also, we note that the number of iterations taken per fit for the iterative process of Section 4.4 to converge was 7 and 4 iterations when R was set to 64 and 128, respectively. The starting values used for the iterative process were the maximum approximate likelihood estimates.

In order to further investigate the performance of the fitting methods, we performed a simulation study where we simulated 10 000 independent samples from the generalised linear mixed model with parameter values set to the maximum approximate likelihood estimates reported in Table 5.4.

The results in Table 5.5 illustrate the underestimation of σ by maximum approximate likelihood using Laplace approximation, which yields the largest bias in absolute value across all methods considered (PQL, corrected PQL, approximate parametric bootstrap and IBLA using Laplace approximation). PQL and CPQL perform better than maximum approximate likelihood, but they also underestimate σ. The bias of these three methods is explained by the small sample size of the data set. Laplace-based parametric bootstrap and IBLA are the best methods in terms of improving the estimation of σ, with the latter being the best method when $R = 128$. The bias of the fixed effect estimates is close to zero for all methods. Comparing the mean squared errors, we observe that maximum approximate likelihood has the largest mean squared error. Lastly, the estimated coverage probabilities of the 90%, 95% and 99% confidence intervals for β_0 based on the Wald statistic and any of the estimates under consideration are smaller than the nominal level, but IBLA gives the best results. The corresponding estimated coverage probabilities for β_1 are all notably closer to the nominal level. Overall, Laplace-based IBLA is the best method in terms of reducing the bias of Laplace-based ML estimates. It also reduces the mean squared error and the confidence intervals based on the IBLA estimates have generally better coverage properties.

The average computational run-times for the methods under consideration are, in increasing order, $0.04, 0.05, 0.72, 6.69, 12.65, 13.74, 16.44$ seconds for the PQL, corrected PQL, maximum approximate likelihood, approximate parametric bootstrap with $R = 64$, IBLA with $R = 64$, approximate parametric bootstrap with $R = 128$, and IBLA with $R = 128$, respectively. Also, we note that the average number of iterations taken per fit for the iterative process of Section 4.4 to converge was 7.4 and 3.1 iterations when R was set to 64 and 128, respectively.

5.5.2 Culcita sea stars

This example represents a small-scale ecological field experiment. The data are from McKeon et al. (2012), and represent trials of coral-eating sea stars Culcita novaeguineae (hereafter Culcita) attacking coral that harbour differing combinations of protective symbionts (crabs and shrimp). The design is a randomised complete block design with two replications per treatment per block, four treatments (no symbionts,

Table 5.6: Estimates of the model parameters for the Culcita sea stars data. The estimated standard errors are reported in parentheses.

	La-ML	PQL	CPQL	La-BOOT		La-IBLA	
R	-	-	-	50	100	50	100
β_0	5.096	3.847	3.938	4.741	4.508	5.238	4.912
	(1.813)	(1.288)	(1.342)	(1.554)	(1.624)	(1.814)	(1.721)
β_1	-3.842	-2.978	-3.036	-3.205	-3.146	-3.991	-3.701
	(1.465)	(1.165)	(1.184)	(1.516)	(1.682)	(1.493)	(1.411)
β_2	-4.431	-3.483	-3.562	-4.318	-4.006	-4.643	-4.285
	(1.552)	(1.207)	(1.231)	(1.628)	(1.780)	(1.590)	(1.493)
β_3	-5.599	-4.445	-4.561	-5.327	-5.076	-5.392	-5.102
	(1.725)	(1.264)	(1.294)	(1.831)	(1.897)	(1.704)	(1.612)
σ	3.437	2.716	2.910	3.293	3.137	3.034	2.952

Notes: La-ML, Laplace-based maximum likelihood; PQL, penalised quasi-likelihood; CPQL, corrected penalised quasi-likelihood; La-BOOT, Laplace-based parametric bootstrap; La-IBLA, Laplace-based iterated bootstrap with likelihood adjustment

crabs only, shrimp only, both crabs and shrimp), with each of these units of eight repeated in ten blocks, giving a total of 80 observations. A natural way to model this data is by means of a generalised linear mixed model that takes into account heterogeneity between blocks. Predation is the binary response, and the linear predictor is

$$\operatorname{logit}[P(y_{ij}=1|\alpha_i)] = \beta_0 + \beta_1 c_{ij} + \beta_2 s_{ij} + \beta_3 b_{ij} + \alpha_i\,, \tag{5.17}$$

where c_{ij}, s_{ij} and b_{ij} are dummy variables denoting the treatment used in the ith block at the jth repetition, namely crabs only, shrimp only, and both crabs and shrimp, respectively.

Table 5.6 gives the estimates of model (5.17) for the methods under consideration. PQL and corrected PQL give the smallest estimates of σ, which based on Breslow & Lin (1995) and Lin & Breslow (1996) can be an indication of underperformance of the methods due to the large variance of the random effects and the small cluster size. Also, the PQL and corrected PQL estimates of the fixed effects are the smallest in absolute value. The computational run-times for the methods under consideration are, in increasing order, $0.04, 0.05, 0.37, 19.68, 23.38, 39.34, 45.77$ seconds for the PQL, corrected PQL, maximum approximate likelihood, approximate parametric bootstrap with $R = 50$, IBLA with $R = 50$, approximate parametric bootstrap with $R = 100$, and IBLA with $R = 100$, respectively. The iterative process of Section 4.4 converged in 11 and 10 iterations when R was set to 50 and 100, respectively.

Table 5.7: Mean bias and mean squared error (MSE) for the parameters of the logistic mixed model for the Culcita sea stars setting, and empirical coverage probability of 90%, 95% and 99% confidence intervals for β_i ($i \in \{0,1,2,3\}$) based on the Wald statistic.

				La-ML	PQL	CPQL	La-BOOT		La-IBLA	
		$(1-\alpha)\%$	R	-	-	-	50	100	50	100
β_0	Bias	-		3.707	1.830	1.415	0.878	0.756	0.842	0.711
	MSE	-		80.372	71.565	47.683	32.352	27.011	31.144	27.275
	Coverage	90		0.747	0.779	0.810	0.843	0.862	0.856	0.882
		95		0.815	0.845	0.865	0.917	0.932	0.928	0.941
		99		0.904	0.906	0.917	0.922	0.943	0.935	0.956
β_1	Bias	-		-2.920	-2.165	-1.726	-0.902	-0.726	-0.825	-0.713
	MSE	-		55.693	68.754	45.301	23.563	19.745	24.012	19.847
	Coverage	90		0.856	0.838	0.847	0.873	0.884	0.880	0.891
		95		0.908	0.885	0.892	0.920	0.932	0.928	0.939
		99		0.957	0.926	0.927	0.961	0.970	0.972	0.981
β_2	Bias	-		-3.068	-2.116	-1.619	-0.803	-0.697	-0.788	-0.676
	MSE	-		58.448	69.089	45.580	24.231	20.338	24.958	20.504
	Coverage	90		0.838	0.824	0.839	0.859	0.865	0.871	0.882
		95		0.893	0.875	0.886	0.918	0.925	0.931	0.942
		99		0.950	0.924	0.927	0.968	0.979	0.970	0.979
β_3	Bias	-		-3.332	-2.005	-1.619	-0.799	-0.697	-0.775	-0.676
	MSE	-		63.178	68.273	45.161	25.323	22.568	25.452	22.638
	Coverage	90		0.830	0.796	0.818	0.864	0.875	0.872	0.886
		95		0.882	0.856	0.870	0.909	0.916	0.923	0.934
		99		0.945	0.916	0.921	0.968	0.976	0.976	0.983
σ	Bias	-		1.338	-0.738	-0.543	0.563	0.417	0.356	0.284
	MSE	-		27.331	1.100	0.935	9.154	8.973	9.232	9.568

Notes: La-ML, Laplace-based maximum likelihood; PQL, penalised quasi-likelihood; CPQL, corrected penalised quasi-likelihood; La-BOOT, Laplace-based parametric bootstrap; La-IBLA, Laplace-based iterated bootstrap with likelihood adjustment

Similar to the multicenter clinical trial example, we performed a simulation study in order to further investigate the performance of the five methods in a generalised linear mixed model context with a large random-effect variance. We simulated 10000 independent samples from model (5.17) at the maximum approximate likelihood estimates reported in Table 5.6.

Table 5.7 presents the mean bias, mean squared error and empirical coverage probabilities of two-sided confidence intervals based on the Wald statistic for the fixed effects. The results for the fixed effects are similar across the five methods in broad outline; the methods overestimate the fixed intercept β_0 and underestimate the fixed effects associated with the treatment β_i, $i = \{1,2,3\}$. As expected, the standard estimation methods (Laplace-based maximum likelihood and PQL) yield the largest bias, and corrected PQL reduces the bias of PQL estimates. Laplace-based parametric bootstrap and IBLA yield the best results in terms of bias, with the latter giving slightly better results. Mean squared errors tend to be smaller under the Laplace-based parametric bootstrap and IBLA estimation, with some of them being less than half the mean squared errors under the maximum approximate likelihood or PQL estimation.

The good performance of Laplace-based IBLA is also evident in the estimation of the standard deviation of the random effects σ, yielding the smallest bias across the five methods. The PQL and corrected PQL methods underestimate σ, contrary to the rest of the methods that yield positive bias. Laplace-based parametric bootstrap and IBLA reduce the mean squared error of Laplace-based maximum likelihood, but PQL and corrected PQL perform best in terms of mean squared error even though they suffer from larger bias of the σ estimates. Callens & Croux (2005) in their study compared PQL with adaptive Gaussian quadrature and ordinary Gaussian quadrature in estimating parameters for logistic regression mixed models and found that in terms of mean squared error, the quadrature methods perform relatively poorly in comparison with PQL, which concurs with our findings. An explanation given in Callens & Croux (2005) is that the number of quadrature points used in the numerical integration techniques is not adequate to outperform the mean squared error of PQL.

Comparing the empirical coverage probabilities, the results in Table 5.7 indicate the conservativeness of the Wald statistic based on the maximum approximate likelihood estimates. On the contrary, the coverage probabilities of the 90, 95 and 99% confidence intervals for any of the fixed effects based on the Wald statistic and the approximate parametric bootstrap or IBLA estimates are notably closer to the nominal level, with the former being slightly more conservative than the latter.

The average computational run-times for the methods under consideration are, in increasing order, $0.12, 0.14, 1.63, 9.38, 11.62, 13.93, 18.07$ seconds for the PQL, corrected PQL, maximum approximate likelihood, approximate parametric bootstrap with $R = 50$, IBLA with $R = 50$, approximate parametric bootstrap with $R = 100$, and IBLA with $R = 100$, respectively. On average, the iterative process of Section 4.4 converged in 8.4 and 5.1 iterations when R was set to 50 and 100, respectively.

In summary, Laplace-based IBLA seems to do well in estimating logistic mixed models producing mean bias-reduced estimates, even in the case of a large random effects dispersion and a small cluster size. The method also appears to perform well when the resulting estimates are used to construct Wald-type inferences.

5.6 Concluding remarks

In this chapter we tested the performance of IBLA of Section 4.4 in the context of generalised linear mixed models. First, we compared the performance of IBLA in estimation and inference with the traditional adjusted score equation (Firth, 1993). In order to do that, we used a simple logistic mixed model with a fixed intercept and a random intercept only. The results indicate that the traditional adjusted score equation method is better in reducing the bias in the estimated variance component parameters, but IBLA also improves estimation. As expected, the Monte Carlo size R governs the accuracy of IBLA. Accuracy increases with increasing R, but at the cost of longer computational run-time. However, the calculations required for obtaining the traditional adjusted score equations and the large number of intractable integrals that are involved in the equations, make their derivation a challenging task, especially for generalised linear mixed models with more complex random and fixed effects structures. On the other hand, IBLA depends only on the first two derivatives of a suitably approximated log-likelihood and hence it can be implemented virtually for any model.

Second, we tested IBLA through two real-data applications to generalised linear mixed models with a random intercept. These examples illustrate that when the cluster size is small or the dispersion parameter is large, standard estimation methods such as the maximum approximate likelihood and the PQL tend to be inaccurate. The large bias in the estimated dispersion parameter affects inference on regression coefficients, and the Wald test is conservative. Corrected PQL attempts to improve PQL estimation but the promised improvement depends on the cluster size and the variance structure. Approximate parametric bootstrap reduces the bias in the maximum approximate likelihood estimates and also improves inference. Even though IBLA is the computationally most expensive method, in most cases it was found to be the best method in terms of mean bias, mean squared error and empirical coverage probabilities.

Overall, IBLA appears to be a promising algorithm for computing bias-reduced estimates in the framework of generalised linear mixed models. More numerical studies are needed to evaluate its behaviour in other generalised linear mixed models.

Chapter 6

Median bias reduction in linear mixed models

6.1 Introduction

Chapter 2 dealt with mean bias reduction in linear mixed models. In this chapter, we consider a different type of bias, namely, median bias reduction in linear mixed models.

Section 6.2 presents the median bias reducing method proposed in Kenne Pagui et al. (2017). Kenne Pagui et al. (2017) show that under suitable conditions third-order median unbiased estimators can be obtained by the solution of a suitably adjusted score equation. Such median bias-reduced (median BR) estimators have component-wise the same probability of over and under-estimating the true parameter value. A key property of these estimators, not shared with the mean bias-reduced ones, is that any monotone component-wise transformation of the estimators results automatically in median bias-reduced estimators of the transformed parameters (Kenne Pagui et al., 2017). Such equivariance property can be useful e.g. in the context of random effects meta-analysis we considered in Section 2.9 where the Fisher information and, hence, the asymptotic variances of various likelihood-based estimators depend only on the heterogeneity parameter.

In Section 6.3 we derive the adjusted score equation for median bias reduction in linear mixed models defined in Chapter 2, and compare it to the corresponding equation for mean bias reduction derived in Section 2.4. The two equations differ by one extra term. Sections 6.4 and 6.5 use the same dataset and simulation study used in Sections 2.7 and 2.8 to compare the performance of the mean and median BR methods in reducing the bias of ML estimates. The results provide evidence on the effectiveness of median bias reduction in improving estimation and Wald-type inference.

In Section 6.6 we derive the median bias-reducing adjusted score functions for random effects meta-analysis and meta-regression. The adjusted score functions are found to correspond to a median bias reducing penalised likelihood (median BRPL), whose logarithm differs from the logarithm of the mean BRPL in Kosmidis et al. (2017) by a simple additive term that depends on the heterogeneity parameter. Since the adjustments to the score function for mean and median bias reduction are both of order $O(1)$, the same arguments as in Kosmidis et al. (2017) are used to obtain a median BRPL ratio statistic with known asymptotic null distribution that can be used for carrying out hypothesis tests and constructing confidence regions or intervals for either the fixed effect or the heterogeneity parameter. The simulation studies and real data applications used in Kosmidis et al. (2017) are used to assess the performance of estimation based on the median BRPL, and compare it to ML and the mean BRPL. We also compare the performance of median BRPL ratio statistic with LR and mean BRPL ratio statistics. Comparison with other methods is not done, because the mean BRPL ratio statistic is already a strong competitor against them in terms of inferential performance as is illustrated in the comparisons in Kosmidis et al. (2017).

6.2 Adjusted score equation for median bias reduction

Kenne Pagui et al. (2017) propose an adjusted score equation approach which can be used to obtain median bias-reduced estimators. Specifically, under the model, the new estimator has a distribution with component-wise medians closer to the “true” parameter values than the ML estimator. Kenne Pagui et al. (2017) consider the median as a centring index for the score, and the adjusted score function for median bias reduction then results by subtracting from the score its approximate median.

Let $j(\theta)$ be the observed information matrix and $i(\theta)$ be the expected information matrix with tth column $i_t(\theta)$. Let also $i^t(\theta)$ and $i^{tt}(\theta)$ be the tth column and the tth diagonal element of $\{i(\theta)\}^{-1}$, with $t \in \{1,\ldots,p+m\}$. Kenne Pagui et al. (2017) show that a median bias-reduced estimator $\hat{\theta}^{\dagger}$ can be obtained by solving an adjusted score equation of the form $s^{\dagger}(\theta) = s(\theta) + A^{\dagger}(\theta) = 0$, where the extra additive term $A^{\dagger}(\theta)$ is of order $O(1)$, with tth element

$$A_t^{\dagger}(\theta) = \frac{1}{2}\operatorname{tr}\left[\{i(\theta)\}^{-1}(P_t(\theta) + Q_t(\theta))\right] - \{i_t(\theta)\}^{\mathrm{T}} K^{\dagger}(\theta)\,. \tag{6.1}$$

The quantities $P_t(\theta) = E_{\theta}[s(\theta)s^{\mathrm{T}}(\theta)s_t(\theta)]$ and $Q_t(\theta) = E_{\theta}[-j(\theta)s_t(\theta)]$ in (6.1) are those introduced by Kosmidis & Firth (2009) for mean bias-reduction, and $K^{\dagger}(\theta)$ is a $(p+m)$-vector with tth element $K_t^{\dagger}(\theta) = \{i^t(\theta)\}^{\mathrm{T}} K_t(\theta)$ where $K_t(\theta)$ is another $(p+m)$-vector with uth element

$$K_{tu}(\theta) = \operatorname{tr}\left[\frac{i^t(\theta)\{i^t(\theta)\}^{\mathrm{T}}}{i^{tt}(\theta)}\left(\frac{1}{3}P_u(\theta) + \frac{1}{2}Q_u(\theta)\right)\right].$$

Given that $A^{\dagger}(\theta)$ is of order $O(1)$, $\hat{\theta}^{\dagger}$ has the same asymptotic distribution as $\hat{\theta}$ (Kenne Pagui et al., 2017), i.e. multivariate normal with mean θ and variance-covariance matrix $\{i(\theta)\}^{-1}$, which can be consistently estimated as $\{i(\hat{\theta}^{\dagger})\}^{-1}$.

6.3 Median bias reducing adjusted score equation for linear mixed models

In the context of linear mixed models values of t and u in $\{1,\ldots,p\}$ correspond to the elements of parameter β, and values of t and u in $\{p+1,\ldots,p+m\}$ correspond to the elements of parameter ψ. The quantity $K_t(\theta)$ has the form

$$K_t(\theta) = \left[\begin{array}{c} 0_p \\ \hline \kappa_{1t} \end{array}\right] \text{ for } t \in \{1,\ldots,p\} \text{ or } K_t(\theta) = \left[\begin{array}{c} 0_p \\ \hline \kappa_{2t} \end{array}\right] \text{ for } t \in \{p+1,\ldots,p+m\} \tag{6.2}$$

where κ_{1t} and κ_{2t} are column vectors with uth elements

$$\begin{aligned}
\kappa_{1tu} &= \frac{1}{3}\operatorname{tr}\left(\frac{\{i_{\beta\beta}^{-1}\}_t\{i_{\beta\beta}^{-1}\}_t^{\mathrm{T}}}{i^{tt}}P_{2u}(\psi)\right), \\
\kappa_{2tu} &= -\frac{1}{6}\operatorname{tr}\left(\frac{\{i_{\psi\psi}^{-1}\}_{t-p}\{i_{\psi\psi}^{-1}\}_{t-p}^{\mathrm{T}}}{i^{tt}}P_{3u}(\psi)\right) + \frac{1}{2}\operatorname{tr}\left(\frac{\{i_{\psi\psi}^{-1}\}_{t-p}\{i_{\psi\psi}^{-1}\}_{t-p}^{\mathrm{T}}}{i^{tt}}P_{4u}(\psi)\right)
\end{aligned}$$

for $u \in \{1, \ldots, m\}$, where the matrices $P_{2u}(\psi)$, $P_{3u}(\psi)$, and $P_{4u}(\psi)$ are defined in Appendix A. Using the result in (6.2) we have $A_t^\dagger(\theta) = 0$ for $t \in \{1, \ldots, p\}$ (regression parameters) and $A_t^\dagger(\theta) = A_t(\theta) - \{i_{\psi\psi}\}_{t-p}^{\mathrm{T}} \kappa^\dagger$ for $t \in \{p+1, \ldots, p+m\}$ (variance components), where $\kappa^\dagger$ is a column vector with rth element $\kappa_r^\dagger = \{i_{\psi\psi}^{-1}\}_r^{\mathrm{T}} \kappa_{2,p+r}$, for $r \in \{1, \ldots, m\}$.

The median BR adjusted score functions for the fixed- and random-effect parameters of linear mixed models are $s_\beta^\dagger(\theta) = s_\beta(\theta)$ and $s_{\psi_r}^\dagger(\theta) = s_{\psi_r}^*(\theta) - \{i_{\psi\psi}\}_r^{\mathrm{T}} \kappa^\dagger$, respectively. The median BR estimates $\hat{\theta}^\dagger = (\hat{\beta}^{\dagger\mathrm{T}}, \hat{\psi}^{\dagger\mathrm{T}})^{\mathrm{T}}$ solve the equations $s_\beta^\dagger(\theta) = 0_p$ and $s_\psi^\dagger(\theta) = 0_m$.

Let $\hat{\theta}^\dagger = (\hat{\beta}^{\dagger\mathrm{T}}, \hat{\lambda}^{\dagger\mathrm{T}}, \hat{\sigma}_\varepsilon^{2\dagger})^{\mathrm{T}}$ be the median BR estimate of θ, where $\hat{\lambda}^\dagger$ is the estimate of the lower triangular elements of the Cholesky factor L. We compute $\hat{\theta}^\dagger$ using the `nleqslv` R package (Hasselman, 2017), and we use the mean BR estimates as starting values in the algorithm that calculates the median BR estimates.

6.4 Dental data

In this section, we revisit the example of dental data (Potthoff & Roy, 1964) discussed in Section 2.7, where the ML, REML, and mean BR methods were evaluated under six different model structures and two different parameterisations. All the tables in this section are the same tables given in Section 2.7 where we further add the results obtained from the median BR method.

The median BR estimates of the linear mixed model parameters in models I-VI are added in Table 6.1. We observe that the median BR estimates of the fixed effects are similar to the ML, REML, and mean BR estimates. The estimates of the variance components differ. Specifically, the median BR estimates of $\sigma_{u_0}^2$ are the largest among the four methods, and the median BR estimates of ρ seem to be closer to the REML estimates. The corresponding results based on the Cholesky parameterisation are given in Table 6.2. For models I-III, which are the models with only a random intercept, the median BR estimates of the variance components are similar to the mean BR estimates. For models IV-VI, which are the models with a random slope correlated with a random intercept, the mean BR estimates of λ_1 are the largest, followed by the median BR estimates.

Table 6.1: ML, REML, mean BR, and median BR estimates of the linear mixed model parameters in models I-VI for the dental data (Potthoff & Roy, 1964) using the parameterisation $\psi = (\sigma^2_{u_0}, \sigma^2_{u_1}, \rho, \sigma^2_{\varepsilon})^{\mathrm{T}}$. Estimated standard errors are reported in parentheses.

		Fixed effects				Variance components			
Model	Method	β_0	β_1	β_2	β_3	$\sigma^2_{u_0}$	$\sigma^2_{u_1}$	ρ	σ^2_{ε}
I	ML	16.76 (0.79)	0.66 (0.06)	-	-	4.29	-	-	2.02
	REML/Mean BR	16.76 (0.80)	0.66 (0.06)	-	-	4.47	-	-	2.05
	Median BR	16.76 (0.81)	0.66 (0.06)	-	-	4.60	-	-	2.07
II	ML	17.71 (0.82)	0.66 (0.06)	-2.32 (0.73)	-	2.99	-	-	2.02
	REML/Mean BR	17.71 (0.83)	0.66 (0.06)	-2.32 (0.76)	-	3.27	-	-	2.05
	Median BR	17.71 (0.84)	0.66 (0.06)	-2.32 (0.77)	-	3.37	-	-	2.07
III	ML	16.34 (0.96)	0.78 (0.08)	1.03 (1.51)	-0.30 (0.12)	3.03	-	-	1.87
	REML/Mean BR	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.12)	3.30	-	-	1.92
	Median BR	16.34 (0.99)	0.78 (0.08)	1.03 (1.55)	-0.30 (0.12)	3.40	-	-	1.94
IV	ML	16.76 (0.76)	0.66 (0.07)	-	-	4.81	0.05	-0.58	1.72
	REML	16.76 (0.78)	0.66 (0.07)	-	-	5.42	0.05	-0.61	1.72
	Mean BR	16.76 (0.78)	0.66 (0.07)	-	-	5.42	0.05	-0.75	1.72
	Median BR	16.76 (0.79)	0.66 (0.07)	-	-	5.75	0.06	-0.60	1.74
V	ML	17.64 (0.86)	0.66 (0.07)	-2.15 (0.73)	-	6.99	0.05	-0.76	1.72
	REML	17.64 (0.89)	0.66 (0.07)	-2.15 (0.76)	-	7.82	0.05	-0.77	1.72
	Mean BR	17.62 (0.88)	0.66 (0.07)	-2.12 (0.66)	-	7.97	0.05	-0.84	1.72
	Median BR	17.62 (0.90)	0.66 (0.07)	-2.12 (0.79)	-	8.20	0.05	-0.75	1.74
VI	ML	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.13)	4.56	0.02	-0.60	1.72
	REML	16.34 (1.02)	0.78 (0.09)	1.03 (1.60)	-0.30 (0.13)	5.79	0.03	-0.67	1.72
	Mean BR	16.34 (1.02)	0.78 (0.09)	1.03 (1.60)	-0.30 (0.13)	5.79	0.03	-0.81	1.72
	Median BR	16.34 (1.03)	0.78 (0.09)	1.03 (1.62)	-0.30 (0.14)	6.15	0.03	-0.65	1.74

Next, we used the same independent samples generated in the simulation study in Section 2.7, and computed the estimated mean bias of the median BR estimates under the $\psi = (\sigma^2_{u_0}, \sigma^2_{\varepsilon})^{\mathrm{T}}$ parameterisation, the percentage of underestimation, the mean squared error, and the estimated relative increase in the mean squared error from its absolute minimum (the variance) due to bias (Kosmidis, 2014b, Table 5). We see in Table 6.3 that median BR reduces the mean bias of the ML estimates, but it does not perform as well as the mean BR does in this respect. The mean BR approach is also better than the median BR in terms of mean squared error, but the median BR is the best method in terms of percentage of underestimation.

The simulated samples were also used to calculate the empirical p-value distribution for the two-sided tests that each parameter is equal to the true values based on the LR and the Wald-type statistics. The results in Table 6.4 suggest that the empirical p-value distribution for the Wald statistic using the median BR estimates is similar to

Table 6.2: ML, REML, mean BR, and median BR estimates of the linear mixed model parameters in models I-VI for the dental data (Potthoff & Roy, 1964) using the Cholesky parameterisation. Estimated standard errors are reported in parentheses.

		Fixed effects				Variance components			
Model	Method	β_0	β_1	β_2	β_3	λ_1	λ_2	λ_3	σ_ε^2
I	ML	16.76 (0.79)	0.66 (0.06)	-	-	2.07	-	-	2.02
	REML	16.76 (0.80)	0.66 (0.06)	-	-	2.11	-	-	2.05
	Mean BR	16.76 (0.81)	0.66 (0.06)	-	-	2.17	-	-	2.05
	Median BR	16.76 (0.81)	0.66 (0.06)	-	-	2.14	-	-	2.07
II	ML	17.71 (0.82)	0.66 (0.06)	-2.32 (0.73)	-	1.73	-	-	2.02
	REML	17.71 (0.83)	0.66 (0.06)	-2.32 (0.76)	-	1.81	-	-	2.05
	Mean BR	17.71 (0.84)	0.66 (0.06)	-2.32 (0.78)	-	1.86	-	-	2.05
	Median BR	17.71 (0.84)	0.66 (0.06)	-2.32 (0.78)	-	1.84	-	-	2.07
III	ML	16.34 (0.96)	0.78 (0.08)	1.03 (1.51)	-0.30 (0.12)	1.74	-	-	1.87
	REML	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.12)	1.82	-	-	1.92
	Mean BR	16.34 (0.99)	0.78 (0.08)	1.03 (1.55)	-0.30 (0.12)	1.87	-	-	1.92
	Median BR	16.34 (0.99)	0.78 (0.08)	1.03 (1.55)	-0.30 (0.12)	1.84	-	-	1.94
IV	ML	16.76 (0.76)	0.66 (0.07)	-	-	2.19	-0.12	0.17	1.72
	REML	16.76 (0.78)	0.66 (0.07)	-	-	2.33	-0.14	0.18	1.72
	Mean BR	16.76 (0.80)	0.66 (0.08)	-	-	2.54	-0.17	0.20	1.72
	Median BR	16.76 (0.79)	0.66 (0.07)	-	-	2.40	-0.14	0.20	1.74
V	ML	17.64 (0.86)	0.66 (0.07)	-2.15 (0.73)	-	2.64	-0.16	0.14	1.72
	REML	17.64 (0.89)	0.66 (0.07)	-2.15 (0.76)	-	2.80	-0.17	0.15	1.72
	Mean BR	17.60 (0.91)	0.66 (0.07)	-2.06 (0.80)	-	2.95	-0.19	0.16	1.72
	Median BR	17.60 (0.90)	0.66 (0.07)	-2.07 (0.80)	-	2.85	-0.18	0.16	1.74
VI	ML	16.34 (0.98)	0.78 (0.08)	1.03 (1.54)	-0.30 (0.13)	2.13	-0.09	0.12	1.72
	REML	16.34 (1.02)	0.78 (0.09)	1.03 (1.60)	-0.30 (0.13)	2.41	-0.12	0.13	1.72
	Mean BR	16.34 (1.05)	0.78 (0.09)	1.03 (1.65)	-0.30 (0.14)	2.62	-0.15	0.16	1.72
	Median BR	16.34 (1.03)	0.78 (0.09)	1.03 (1.62)	-0.30 (0.14)	2.48	-0.13	0.15	1.74

that of the Wald statistic using the mean BR estimates, and close to uniformity.

Lastly, we obtained the corresponding results under the $\psi = (\sigma_{u_0}, \sigma_\varepsilon^2)^\top$ parameterisation. We did not repeat the simulation study in order to obtain the estimates of ψ. Instead, because of the equivariance of the median BRPL estimators under monotone component-wise parameter transformations, the median BR estimates of σ_{u_0} were calculated by taking the square root of the median BR estimates of $\sigma_{u_0}^2$ obtained in the last simulation study. The results for the $\psi = (\sigma_{u_0}, \sigma_\varepsilon^2)^\top$ setting are reported in Tables 6.5 and 6.6. Median BR reduces the bias of the ML estimates and it also performs well in terms of percentage of underestimation. The mean squared error is similar across all methods. The mean squared errors of the median BR estimates of the variance components are inflated by as much as 1% due to bias from their minimum values.

Table 6.3: Mean bias, percentage of underestimation (PU), and mean squared error (MSE) of the variance component estimates for the linear mixed models I-III using the dental data setting and the $\psi = (\sigma_{u_0}^2, \sigma_\varepsilon^2)^\top$ parameterisation.

Model	Parameter	Method	Bias	PU	MSE	Bias^2/Variance (%)
I	$\sigma_{u_0}^2$	ML	-0.166	58.3	1.645	1.707
		REML/Mean BR	0.006	53.0	1.743	0.002
		Median BR	0.132	49.7	1.854	0.946
	σ_ε^2	ML	-0.019	54.0	0.100	0.362
		REML/Mean BR	0.006	51.2	0.102	0.036
		Median BR	0.023	49.2	0.105	0.513
II	$\sigma_{u_0}^2$	ML	-0.250	63.7	0.897	7.471
		REML/Mean BR	0.004	53.3	0.972	0.001
		Median BR	0.099	49.5	1.036	0.959
	σ_ε^2	ML	-0.019	54.0	0.100	0.362
		REML/Mean BR	0.006	51.2	0.102	0.036
		Median BR	0.018	49.2	0.105	0.513
III	$\sigma_{u_0}^2$	ML	-0.244	63.6	0.892	7.152
		REML/Mean BR	0.004	53.4	0.971	0.002
		Median BR	0.100	49.5	1.034	0.968
	σ_ε^2	ML	-0.041	57.2	0.087	1.939
		REML/Mean BR	0.006	51.4	0.090	0.037
		Median BR	0.022	49.3	0.092	0.520

Table 6.4: Empirical p-value distribution (%) for the likelihood ratio test and the tests based on the Wald statistic using the dental data setting and the $\psi = (\sigma_{u_0}^2, \sigma_\varepsilon^2)^\top$ parameterisation.

Model	$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
I	Likelihood ratio	1.1	2.7	4.9	9.4	24.7	49.9	74.8	89.7	94.7	97.4	99.0
	Wald using ML	1.2	2.8	5.3	9.7	24.9	50.0	74.8	89.7	94.7	97.4	99.0
	Kenward-Roger	1.0	2.5	4.7	9.1	24.4	49.6	74.6	89.5	94.6	97.4	99.0
	Wald using mean BR	1.2	2.7	5.2	9.4	24.6	49.7	74.6	89.6	94.7	97.4	99.0
	Wald using median BR	1.1	2.7	4.9	9.3	24.6	49.6	74.6	89.5	94.6	97.4	99.0
II	Likelihood ratio	1.6	3.2	6.0	12.0	27.4	51.5	76.2	90.8	95.4	97.6	99.3
	Wald using ML	2.2	4.1	6.7	12.8	27.9	51.6	76.2	90.8	95.4	97.6	99.3
	Kenward-Roger	1.2	2.6	4.9	10.1	25.4	49.3	74.9	90.3	95.3	97.5	99.2
	Wald using mean BR	1.8	3.3	5.9	11.6	26.3	50.1	75.3	90.3	95.3	97.5	99.2
	Wald using median BR	1.6	3.2	5.6	11.2	25.7	49.4	74.8	90.2	95.3	97.5	99.2
III	Likelihood ratio	1.0	2.7	5.6	10.4	26.0	50.1	74.7	89.9	95.4	97.7	99.0
	Wald using ML	1.2	3.1	6.0	10.9	26.1	50.1	74.7	89.9	95.4	97.7	99.0
	Kenward-Roger	0.9	2.5	5.4	10.0	25.2	49.5	74.5	89.8	95.3	97.7	99.0
	Wald using mean BR	1.0	2.8	5.6	10.4	25.5	49.6	74.5	89.8	95.3	97.7	99.0
	Wald using median BR	1.0	2.7	5.5	10.3	25.3	49.5	74.5	89.8	95.3	97.7	99.0

Similar to the previous parameterisation, we used the simulated samples to calculate the empirical p-value distribution for the two-sided tests that each parameter is equal to the true values based on the LR and the Wald statistic. The results are reported in Table 6.6 and are qualitatively similar to the ones obtained using the first parameterisation in Table 6.4.

Table 6.5: Mean bias, percentage of underestimation (PU), and mean squared error (MSE) of the Cholesky parameter estimates for the linear mixed models I-III using the dental data setting and the $\psi = (\sigma_{u_0}, \sigma_\varepsilon^2)^\mathrm{T}$ parameterisation.

Model	Parameter	Method	Bias	PU	MSE	Bias^2/Variance (%)
I	σ_{u_0}	ML	-0.065	58.3	0.102	4.279
		REML	-0.023	53.0	0.102	0.532
		Mean BR	0.028	47.1	0.105	0.766
		Median BR	0.033	46.5	0.106	1.044
	σ_ε^2	ML	-0.019	54.0	0.100	0.362
		REML	0.006	51.2	0.102	0.036
		Mean BR	0.006	51.2	0.102	0.036
		Median BR	0.023	49.2	0.105	0.513
II	σ_{u_0}	ML	-0.097	63.7	0.086	12.267
		REML	-0.023	53.3	0.082	0.633
		Mean BR	0.027	46.5	0.084	0.851
		Median BR	0.030	46.1	0.085	1.050
	σ_ε^2	ML	-0.019	54.0	0.100	0.362
		REML	0.006	51.2	0.102	0.036
		Mean BR	0.006	51.2	0.102	0.036
		Median BR	0.018	49.2	0.105	0.513
III	σ_{u_0}	ML	-0.094	63.6	0.084	11.802
		REML	-0.022	53.4	0.081	0.609
		Mean BR	0.026	46.8	0.083	0.835
		Median BR	0.030	46.1	0.084	1.060
	σ_ε^2	ML	-0.041	57.2	0.087	1.939
		REML	0.006	51.4	0.090	0.037
		Mean BR	0.006	51.4	0.090	0.037
		Median BR	0.022	49.3	0.092	0.520

Table 6.6: Empirical p-value distribution (%) for the tests based on the Wald statistic using the dental data setting and the $\psi = (\sigma_{u_0}, \sigma_\varepsilon^2)^\mathrm{T}$ parameterisation.

Model	$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
I	Likelihood ratio	1.1	2.7	4.9	9.4	24.7	49.9	74.8	89.7	94.7	97.4	99.0
	Wald using ML	1.2	2.8	5.3	9.7	24.9	50.0	74.8	89.7	94.7	97.4	99.0
	Wald using REML	1.2	2.7	5.2	9.4	24.6	49.7	74.6	89.6	94.7	97.4	99.0
	Kenward-Roger	1.0	2.5	4.7	9.1	24.4	49.6	74.6	89.5	94.6	97.4	99.0
	Wald using mean BR	1.2	2.7	5.2	9.4	24.6	49.7	74.6	89.6	94.7	97.4	99.0
	Wald using median BR	1.1	2.7	4.9	9.3	24.6	49.6	74.6	89.5	94.6	97.4	99.0
II	Likelihood ratio	1.6	3.2	6.0	12.0	27.4	51.5	76.2	90.8	95.4	97.6	99.3
	Wald using ML	2.2	4.1	6.7	12.8	27.9	51.6	76.2	90.8	95.4	97.6	99.3
	Wald using REML	1.8	3.3	5.9	11.6	26.3	50.1	75.3	90.3	95.3	97.5	99.2
	Kenward-Roger	1.3	2.8	5.2	10.7	26.0	50.0	75.3	90.4	95.3	97.5	99.2
	Wald using mean BR	1.5	3.0	5.4	10.5	25.4	48.9	74.6	90.2	95.2	97.5	99.2
	Wald using median BR	1.5	2.9	5.3	10.4	25.2	48.8	74.6	90.1	95.2	97.5	99.2
III	Likelihood ratio	1.0	2.7	5.6	10.4	26.0	50.1	74.7	89.9	95.4	97.7	99.0
	Wald using ML	1.2	3.1	6.0	10.9	26.1	50.1	74.7	89.9	95.4	97.7	99.0
	Wald using REML	1.0	2.8	5.6	10.4	25.5	49.6	74.5	89.8	95.3	97.7	99.0
	Kenward-Roger	0.9	2.5	5.4	10.0	25.2	49.5	74.5	89.8	95.3	97.7	99.0
	Wald using mean BR	1.0	2.8	5.6	10.4	25.5	49.6	74.5	89.8	95.3	97.7	99.0
	Wald using median BR	1.0	2.7	5.5	10.3	25.3	49.5	74.5	89.8	95.3	97.7	99.0

In a nutshell, the results in this section suggest that median BR reduces the bias and mean squared error of ML estimates and is a good competitor to mean BR.

6.5 Simulation study

In this section, we revisit the simulation study performed in Section 2.8, where we studied the behaviour of the ML, REML, and mean BR methods under small and moderate sample sizes when fitting a linear mixed model with a random intercept and a correlated random slope. Similar to the previous section, we generated the same samples and added the results obtained from the median BR method in Tables 6.7 and 6.8.

Table 6.7 indicates that the mean BR and median BR methods reduce the bias of the ML estimates of the variance components, especially λ_1, yielding smaller bias than REML. The results on the percentage of underestimation for the Cholesky parameters and the variance error indicate that median bias reduction is achieved by solving the median BR adjusted score equations. The mean squared error is in all scenarios similar across the four estimation methods. The mean squared errors of the ML and REML estimates are inflated by as much as 35% and 15% due to bias from their variance, respectively. On the other hand, the corresponding inflation factors for the mean BR and median BR estimators are very close to zero and do not exceed 0.4%. Table 6.8 indicates once again that the empirical p-value distribution for the KR and the Wald statistic using the mean BR and median BR estimates are closest to uniformity.

Table 6.7: Mean bias, percentage of underestimation (PU), and mean squared error (MSE) of the Cholesky parameter estimates under the linear mixed model (2.9) with cluster size n_i and variance error σ_ε^2.

	Bias				PU				MSE				Bias^2/Variance (%)			
Method	λ_1	λ_2	λ_3	σ_ε^2	λ_1	λ_2	λ_3	σ_ε^2	λ_1	λ_2	λ_3	σ_ε^2	λ_1	λ_2	λ_3	σ_ε^2
$n_i = 6, \sigma_\varepsilon^2 = 0.0705^2$																
ML	-0.16	0.01	-0.07	0.00	65.8	48.7	72.5	54.5	0.20	0.03	0.02	0.00	13.53	0.14	35.01	0.11
REML	-0.06	0.00	-0.04	0.00	56.9	49.4	64.9	54.5	0.20	0.03	0.02	0.00	1.83	0.03	12.82	0.11
Mean BR	-0.01	0.00	0.00	0.00	52.0	49.7	50.2	54.5	0.21	0.03	0.02	0.00	0.02	0.00	0.03	0.11
Median BR	0.01	0.00	0.01	0.00	50.3	49.7	48.5	51.5	0.22	0.03	0.02	0.00	0.07	0.00	0.39	0.17
$n_i = 6, \sigma_\varepsilon^2 = 0.141^2$																
ML	-0.16	0.01	-0.07	0.00	66.0	48.5	72.5	54.5	0.20	0.03	0.02	0.00	13.54	0.15	35.10	0.11
REML	-0.06	0.00	-0.04	0.00	56.8	49.4	64.9	54.5	0.20	0.03	0.02	0.00	1.84	0.03	12.87	0.11
Mean BR	-0.01	0.00	0.00	0.00	51.9	49.6	50.1	54.5	0.21	0.03	0.02	0.00	0.02	0.01	0.03	0.11
Median BR	0.01	0.00	0.01	0.00	50.3	49.6	48.4	51.5	0.22	0.03	0.02	0.00	0.07	0.00	0.39	0.17
$n_i = 6, \sigma_\varepsilon^2 = 0.282^2$																
ML	-0.16	0.01	-0.07	0.00	65.8	48.4	72.7	54.5	0.21	0.03	0.02	0.00	13.58	0.16	35.45	0.11
REML	-0.06	0.00	-0.04	0.00	56.9	49.2	65.3	54.5	0.21	0.03	0.02	0.00	1.86	0.04	13.08	0.11
Mean BR	-0.01	0.00	0.00	0.00	51.8	49.6	50.0	54.5	0.22	0.03	0.02	0.00	0.02	0.01	0.04	0.11
Median BR	0.01	0.00	0.01	0.00	50.3	49.6	48.5	51.5	0.22	0.03	0.02	0.00	0.06	0.01	0.39	0.16
$n_i = 26, \sigma_\varepsilon^2 = 0.0705^2$																
ML	-0.15	0.00	-0.07	0.00	64.9	49.6	73.4	52.0	0.20	0.03	0.02	0.00	12.24	0.07	35.56	0.04
REML	-0.06	0.00	-0.05	0.00	56.5	50.2	65.8	52.0	0.21	0.03	0.02	0.00	1.77	0.00	15.14	0.04
Mean BR	0.00	0.00	0.00	0.00	51.5	50.7	51.8	51.8	0.21	0.03	0.02	0.00	0.01	0.01	0.01	0.02
Median BR	0.02	0.00	0.00	0.00	49.8	50.7	50.3	50.4	0.22	0.03	0.02	0.00	0.26	0.01	0.09	0.03
$n_i = 26, \sigma_\varepsilon^2 = 0.141^2$																
ML	-0.15	0.00	-0.07	0.00	64.8	49.5	73.5	52.0	0.20	0.03	0.02	0.00	12.29	0.07	35.59	0.04
REML	-0.05	0.00	-0.04	0.00	56.1	50.3	65.2	52.0	0.21	0.03	0.02	0.00	1.44	0.00	13.11	0.04
Mean BR	0.00	0.00	0.00	0.00	51.6	50.6	51.2	52.0	0.21	0.03	0.02	0.00	0.00	0.00	0.03	0.04
Median BR	0.02	0.00	0.01	0.00	50.2	50.6	49.3	50.6	0.22	0.03	0.02	0.00	0.17	0.00	0.37	0.01
$n_i = 26, \sigma_\varepsilon^2 = 0.282^2$																
ML	-0.15	0.00	-0.07	0.00	64.7	49.3	73.3	52.0	0.21	0.03	0.02	0.00	12.38	0.07	35.69	0.04
REML	-0.05	0.00	-0.04	0.00	56.2	50.2	65.3	52.0	0.21	0.03	0.02	0.00	1.47	0.00	13.15	0.04
Mean BR	0.00	0.00	0.00	0.00	51.7	50.7	51.0	52.0	0.22	0.03	0.02	0.00	0.00	0.00	0.03	0.04
Median BR	0.02	0.00	0.01	0.00	50.2	50.7	49.5	50.6	0.22	0.03	0.02	0.00	0.16	0.00	0.37	0.01

Table 6.8: Empirical p-value distribution (%) for the tests based on the Wald statistic using the Cholesky parameter estimates under the linear mixed model (2.9) with cluster size n_i and variance error σ_ε^2.

n_i	σ_ε^2	$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
6	0.0705^2	Likelihood ratio	2.3	4.6	8.1	14.5	31.2	56.0	77.1	90.8	95.6	97.8	99.1
		Wald using ML	3.8	6.4	9.9	15.6	30.8	54.6	76.1	90.3	95.5	97.6	99.1
		Wald using REML	3.1	5.2	8.2	13.7	28.5	52.7	75.1	89.8	95.2	97.5	99.0
		Kenward-Roger	1.7	3.7	6.9	12.9	28.9	54.2	76.0	90.2	95.4	97.6	99.1
		Wald using mean BR	2.3	3.8	6.3	11.1	24.7	48.7	73.3	88.8	94.8	97.2	99.0
		Wald using median BR	2.2	3.7	6.2	10.9	24.3	48.4	73.1	88.6	94.8	97.2	99.0
6	0.141^2	Likelihood ratio	2.3	4.6	8.0	14.5	31.3	56.0	77.0	90.7	95.6	97.7	99.1
		Wald using ML	3.9	6.4	9.9	15.5	30.8	54.7	76.1	90.2	95.5	97.6	99.1
		Wald using REML	3.2	5.3	8.2	13.6	28.5	52.7	75.1	89.7	95.3	97.4	99.0
		Kenward-Roger	1.7	3.8	6.9	12.8	29.0	54.1	76.1	90.2	95.4	97.6	99.1
		Wald using mean BR	2.3	3.9	6.3	11.1	24.6	48.8	73.4	88.9	94.8	97.2	98.9
		Wald using median BR	2.2	3.7	6.1	10.8	24.2	48.3	73.1	88.7	94.7	97.2	98.9
6	0.282^2	Likelihood ratio	2.2	4.6	8.1	14.4	31.3	56.0	77.1	90.6	95.7	97.7	99.1
		Wald using ML	3.8	6.4	10.0	15.5	30.7	54.6	76.4	90.2	95.6	97.6	99.1
		Wald using REML	3.2	5.3	8.3	13.6	28.5	52.6	75.1	89.6	95.3	97.5	99.0
		Kenward-Roger	1.8	3.7	6.9	12.9	29.0	53.9	76.2	90.1	95.6	97.6	99.1
		Wald using mean BR	2.2	3.9	6.2	11.1	24.7	48.7	73.2	88.9	94.8	97.3	98.9
		Wald using median BR	2.1	3.8	6.1	10.7	24.3	48.3	73.1	88.9	94.8	97.2	98.9
26	0.0705^2	Likelihood ratio	2.2	4.8	8.6	14.9	32.0	55.9	78.3	91.4	95.4	97.7	99.1
		Wald using ML	3.9	6.7	10.5	16.1	31.8	54.6	77.4	91.0	95.1	97.5	99.0
		Wald using REML	3.1	5.7	8.8	14.2	29.5	52.8	76.5	90.3	94.9	97.4	99.0
		Kenward-Roger	1.7	3.7	7.0	13.1	29.7	53.9	77.2	91.0	95.1	97.4	99.1
		Wald using mean BR	2.3	4.0	6.6	11.4	25.5	49.1	74.2	89.5	94.5	97.2	98.9
		Wald using median BR	2.3	3.8	6.4	11.2	25.0	48.7	74.0	89.3	94.4	97.1	98.9
26	0.141^2	Likelihood ratio	2.2	4.8	8.6	15.0	32.0	55.9	78.3	91.4	95.4	97.6	99.1
		Wald using ML	3.9	6.6	10.3	15.9	31.5	54.4	77.3	91.0	95.2	97.5	99.1
		Wald using REML	3.1	5.6	8.7	13.8	29.1	52.5	76.3	90.3	94.9	97.4	99.0
		Kenward-Roger	1.7	3.7	7.0	13.1	29.7	53.9	77.2	91.0	95.1	97.5	99.1
		Wald using mean BR	2.3	3.9	6.5	11.3	25.2	48.8	74.1	89.5	94.4	97.2	99.0
		Wald using median BR	2.2	3.8	6.3	11.1	24.7	48.4	73.9	89.3	94.4	97.2	99.0
26	0.282^2	Likelihood ratio	2.2	4.8	8.6	14.9	32.0	55.9	78.3	91.4	95.4	97.7	99.1
		Wald using ML	3.9	6.5	10.4	15.9	31.5	54.4	77.3	90.9	95.2	97.5	99.1
		Wald using REML	3.1	5.6	8.7	13.9	29.1	52.5	76.2	90.3	94.9	97.4	99.0
		Kenward-Roger	1.7	3.7	7.0	13.1	29.7	53.9	77.2	90.9	95.2	97.4	99.1
		Wald using mean BR	2.3	3.9	6.5	11.3	25.2	48.8	74.1	89.6	94.4	97.2	98.9
		Wald using median BR	2.2	3.7	6.3	11.1	24.7	48.4	73.9	89.4	94.4	97.2	98.9

6.6 Median bias reduction in random effects meta-analysis and meta-regression

In the context of random effects meta-regression values of t and u in $\{1,\ldots,p\}$ correspond to the elements of parameter β, and $t,u = p+1$ correspond to parameter ψ.

The observed information matrix $j(\theta)$ for the random effects meta-regression model (2.10) is

$$j(\theta) = \begin{pmatrix} X^{\mathrm{T}}W(\psi)X & X^{\mathrm{T}}W(\psi)^2R(\beta) \\ X^{\mathrm{T}}W(\psi)^2R(\beta) & R(\beta)^{\mathrm{T}}W(\psi)^3R(\beta) - \frac{1}{2}\operatorname{tr}[W(\psi)^2] \end{pmatrix}$$

and the expected information matrix $i(\theta)$ is

$$i(\theta) = \begin{pmatrix} X^{\mathrm{T}}W(\psi)X & 0_p \\ 0_p^{\mathrm{T}} & \frac{1}{2}\operatorname{tr}[W(\psi)^2] \end{pmatrix}. \tag{6.3}$$

For this model

$$P_t(\theta) = -Q_t(\theta) = \begin{pmatrix} 0_{p\times p} & X^{\mathrm{T}}W(\psi)^2X_t \\ X^{\mathrm{T}}W(\psi)^2X_t & 0 \end{pmatrix} \quad (t = 1, \ldots, p),$$

and

$$P_{p+1}(\theta) = \begin{pmatrix} X^{\mathrm{T}}W(\psi)^2X & 0_p \\ 0_p^{\mathrm{T}} & \operatorname{tr}(W(\psi)^3) \end{pmatrix} \quad \text{and} \quad Q_{p+1}(\theta) = \begin{pmatrix} 0_{p\times p} & 0_p \\ 0_p^{\mathrm{T}} & -\operatorname{tr}(W(\psi)^3) \end{pmatrix},$$

where X_t is the tth column of X.

The median bias-reducing adjustment for the random effects meta-analysis and meta-regression models is obtained by plugging the above expressions into (6.1) and has the form

$$A^{\dagger}(\theta) = \begin{pmatrix} 0_p \\ \frac{1}{2}\operatorname{tr}[W(\psi)H(\psi)] + \frac{1}{3}\frac{\operatorname{tr}[W(\psi)^3]}{\operatorname{tr}[W(\psi)^2]} \end{pmatrix}, \tag{6.4}$$

where $H(\psi) = X(X^{\mathrm{T}}W(\psi)X)^{-1}X^{\mathrm{T}}W(\psi)$. Substituting (6.4) in the expression for $s^{\dagger}(\theta)$ gives that the median bias-reducing adjusted score functions for β and ψ are $s^{\dagger}_{\beta}(\theta) = s_{\beta}(\theta)$ and

$$s^{\dagger}_{\psi}(\theta) = s_{\psi}(\theta) + \frac{1}{2}\operatorname{tr}[W(\psi)H(\psi)] + \frac{1}{3}\frac{\operatorname{tr}[W(\psi)^3]}{\operatorname{tr}[W(\psi)^2]},$$

respectively.

The median BR adjusted score function for the variance component ψ differs from the corresponding mean BR adjusted score function by one extra additive term, whereas the two adjusted score functions for the regression coefficients β are identical.

6.7 Computation of median bias-reduced estimator

A direct approach for computing the estimator $\hat{\theta}^{\dagger} = (\hat{\beta}^{\dagger\mathrm{T}}, \hat{\psi}^{\dagger})^{\mathrm{T}}$ is through a modification of the two-step iterative process in Kosmidis et al. (2017). At the jth iteration $(j = 1, 2, \ldots)$

1. calculate $\beta^{(j)}$ by weighted least squares as

$$\beta^{(j)} = (X^{\mathrm{T}}W(\psi^{(j-1)})X)^{-1}X^{\mathrm{T}}W(\psi^{(j-1)})y$$

2. solve $s^{\dagger}_{\psi}(\theta^{(j)}(\psi)) = 0$ with respect to ψ, where $\theta^{(j)}(\psi) = (\beta^{(j)\mathrm{T}}, \psi)^{\mathrm{T}}$.

In the above steps, $\beta^{(j)}$ is the candidate value for $\hat{\beta}^{\dagger}$ at the jth iteration and $\psi^{(j-1)}$ is the candidate value for $\hat{\psi}^{\dagger}$ at the $(j-1)$th iteration. The equation in step 2 is solved numerically, by searching for the root of the function $s^{\dagger}_{\psi}(\beta^{(j)}, \psi)$ in a predefined positive interval. For the computations in this chapter we use the DL estimate of ψ as starting value $\psi^{(0)}$. The iterative process is then repeated until the components of the score function $s^{\dagger}(\theta)$ are all less than $\varepsilon = 10^{-6}$ in absolute value at the current estimates.

6.8 Median bias-reducing penalised likelihood

Although it is not generally true that the median BR adjusted scores are the gradients of a suitable penalised log-likelihood, in this case $s^{\dagger}(\theta)$ is the gradient of the median BRPL

$$l^{\dagger}(\theta) = l(\theta) - \frac{1}{2}\log|X^{\mathrm{T}}W(\psi)X| - \frac{1}{6}\log[\mathrm{tr}(W(\psi)^2)], \tag{6.5}$$

where the expression for the differential of the log-determinant has been used in the derivation of (6.5). Hence, $\hat{\theta}^{\dagger}$ is also the maximum median BRPL estimator. The median BRPL in (6.5) differs from the mean BRPL derived in Kosmidis et al. (2017) by the term $-\log[\mathrm{tr}(W(\psi)^2)]/6$.

An advantage of the median BRPL estimators over mean BRPL ones is that the former are equivariant under monotone component-wise parameter transformations (Kenne Pagui et al., 2017). In the context of random effects meta-analysis and meta-regression, this equivariance implies that not only we get a median bias-reduced estimator of ψ, but we also get median bias-reduced estimates of the standard errors for β by calculating the square roots of the diagonal elements of $\{i(\theta)\}^{-1}$ in (6.3) at $\psi^{\dagger}$. This is because $i(\theta)$ is a function of ψ only, and moreover the square roots of the diagonal elements of $\{i(\theta)\}^{-1}$ are monotone functions of ψ. The monotonicity of the standard errors for β can be shown by showing that the diagonal elements of the first derivative with respect to ψ of $(X^{\mathrm{T}}W(\psi)X)^{-1}$ are positive. The above holds because $d(X^{\mathrm{T}}W(\psi)X)^{-1}/d\psi = (X^{\mathrm{T}}W(\psi)X)^{-1}(X^{\mathrm{T}}W(\psi)^2X)(X^{\mathrm{T}}W(\psi)X)^{-1}$ is a product of positive definite matrices, which in turn results to a positive definite matrix.

6.9 Penalised likelihood-based inference

For inference about either the components of the fixed effects β or the between-study heterogeneity ψ we propose the use of the median BRPL ratio. If $\theta = (\tau^{\mathrm{T}}, \lambda^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\lambda}^{\dagger}_{\tau}$ is the maximiser of $l^{\dagger}(\theta)$ for fixed τ, then the same arguments as in Kosmidis et al. (2017) can be used to show that the logarithm of the median BRPL ratio statistic

$$2\{l^{\dagger}(\hat{\tau}^{\dagger}, \hat{\lambda}^{\dagger}) - l^{\dagger}(\tau, \hat{\lambda}^{\dagger}_{\tau})\} \tag{6.6}$$

has a $\chi^2_{\dim(\tau)}$ asymptotic distribution, as K goes to infinity. Specifically, the adjustment to the score function is additive and of order $O(1)$. As a result, the extra terms in the asymptotic expansion of the logarithm of the median BRPL that depend on the penalty and its derivatives disappear as information increases, and the expansion has the same leading term as that of the log-likelihood (see, for example, Pace & Salvan, 1997, Section 9.4, for an asymptotic expansion of the log-likelihood).

6.10 Cocoa intake and blood pressure reduction data

In this section, we revisit the example of the cocoa data analysed in Section 2.9.3, where the performance of estimation and inference based on the ML and mean BR was

Table 6.9: ML, mean BRPL, and median BRPL estimates of the model parameters for the cocoa data. Estimated standard errors are reported in parentheses. The 95% confidence intervals based on the LR, mean BRPL ratio and median BRPL ratio are reported in squared brackets.

	Parameter		Iterations until	Computational
Method	β	ψ	convergence	run-time (sec)
ML	-2.799 (1.002)	4.199	4	1.1×10^{-2}
	[-5.26, -0.40]	[1.10, 23.5]		
Mean BRPL	-2.811 (1.121)	5.546	6	1.8×10^{-2}
	[-5.73, 0.05]	[1.00, 38.5]		
Median BRPL	-2.818 (1.244)	6.897	11	1.5×10^{-2}
	[-6.21, 0.52]	[1.40, 58.0]		

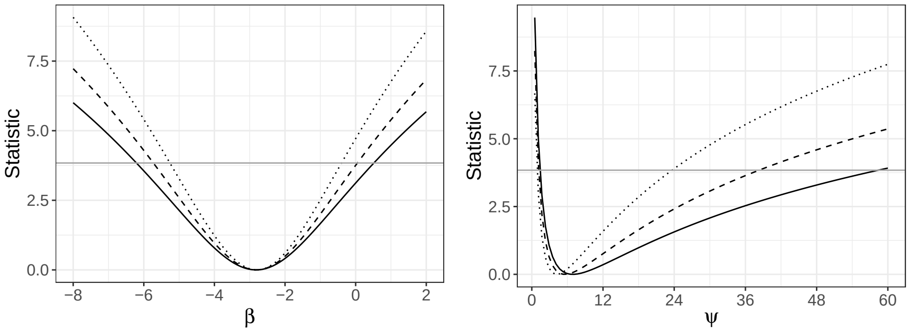

Figure 6.1: Plot of LR (dotted), mean BRPL (dashed) and median BRPL (solid) ratio statistic in (6.6) when τ is β (left) and ψ (right). The horizontal line is the 95% quantile of the limiting χ_1^2 distribution, and its intersection with the values of the statistics results in the endpoints of the corresponding 95% confidence intervals.

evaluated.

The ML, the maximum mean BRPL and the maximum median BRPL estimates of the meta-analysis model parameters are given in Table 6.9. We observe that the bias-reduced estimates of ψ and, as a consequence, the corresponding estimated standard errors for β are larger than their ML counterparts, which is typical in random effects meta-analysis. Also, for both β and ψ, the confidence intervals based on the LR statistic are the narrowest and the confidence intervals based on the median BRPL ratio statistic are the widest. The confidence intervals are also illustrated in Figure 6.1, which gives the value of LR, mean BRPL and median BRPL ratio statistic in (6.6) for a range of values of τ, when τ is either β or ψ. The horizontal line in Figure 6.1 is the 95% quantile of the limiting χ_1^2 distribution, and its intersection with the values of the statistics results in the endpoints of the corresponding 95% confidence intervals.

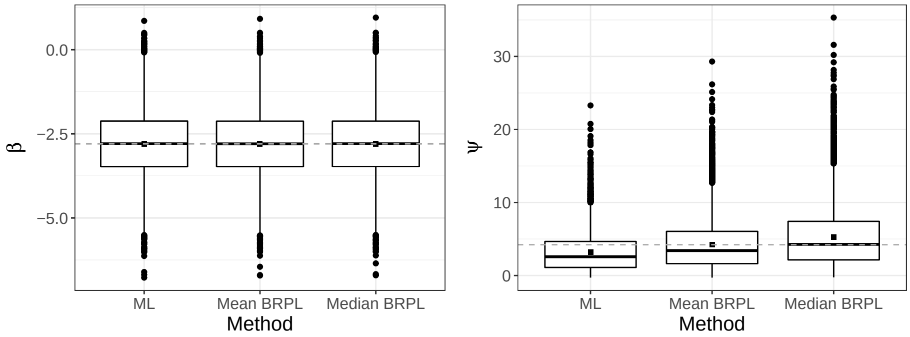

Figure 6.2: Boxplots for the ML, the maximum mean BRPL, and the maximum median BRPL estimates of β and ψ as calculated from 10 000 simulated samples under the ML fit using the cocoa data. The square point is the empirical mean of the estimates. The dashed grey horizontal line is at the parameter value used to generate the data.

Table 6.10: Empirical p-value distribution (%) for the tests based on the LR statistic, the mean BRPL ratio statistic, and the median BRPL ratio statistic in the cocoa data setting.

$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
LR	5.9	8.4	11.7	18.2	34.5	57.8	79.1	91.7	96.0	98.0	99.2
Mean BRPL ratio	1.6	3.7	6.7	12.1	28.3	52.8	76.6	90.9	95.5	97.9	99.1
Median BRPL ratio	0.6	1.8	4.1	8.6	23.1	48.5	74.2	89.9	94.9	97.5	99.1

Notes: Each column gives the coverage probability of $(1-\alpha)\%$ confidence intervals based on the three statistics.

In order to further investigate the performance of the three approaches to estimation and inference, we performed a simulation study where we simulated 10 000 independent samples from the random effects meta-analysis model with parameter values set to the ML estimates reported in Table 6.9, i.e. $\beta_0 = -2.799$ and $\psi_0 = 4.199$. Figure 6.2 shows boxplots of the estimates of β and ψ calculated from each of the 10 000 simulated samples. The distributions of the three alternative estimators for β are similar. On the other hand, the ML estimator of ψ has a large negative mean bias, maximum median BRPL tends to over-correct for that bias, while maximum mean BRPL almost fully corrects for the bias of ML estimator. The distribution of the median BRPL estimates has the heaviest right tail. The simulation-based estimates of the probabilities of underestimation for ψ, $P_{\psi_0}(\hat{\psi} \leq \psi_0)$, $P_{\psi_0}(\hat{\psi}^* \leq \psi_0)$ and $P_{\psi_0}(\hat{\psi}^\dagger \leq \psi_0)$ are 0.708, 0.591 and 0.493 for the ML, maximum mean BRPL, and maximum median BRPL, respectively, illustrating how effective maximising the median BRPL in (6.5) is in reducing the median bias of the ML estimator of ψ.

The simulated samples were also used to calculate the empirical p-value distribution for the two-sided tests that each parameter is equal to the true values based on the

LR statistic, the mean BRPL ratio statistic, and the median BRPL ratio statistic. Table 6.10 shows that the empirical p-value distribution for the mean and median BRPL ratio statistics are closest to uniformity, with the latter being slightly more conservative than the former. The coverage probability of the 95% confidence intervals for β based on the mean BRPL ratio and the median BRPL ratio are notably closer to the nominal level than those based on the LR. Specifically, the coverage probabilities for β are 88%, 93%, and 96% for LR, mean BRPL ratio, and median BRPL ratio respectively, and the corresponding coverage probabilities for ψ are 88%, 94%, and 96%, respectively.

6.11 Simulation study

More extensive simulations under the random effects meta-analysis model (2.10) are performed here using the design in Brockwell & Gordon (2001). Specifically, the data y_i, $i \in \{1,\ldots,K\}$, are simulated from model (2.10) with true fixed-effect parameter $\beta = 0.5$. The within-study variances $\hat{\sigma}_i^2$ are independently generated from a χ_1^2 distribution and are multiplied by 0.25 before restricted to the interval $(0.009, 0.6)$. Eleven values of the between-study variance ψ ranging from 0 to 0.1 are chosen, and the number of studies K ranges from 5 to 200. For each combination of ψ and K considered, we simulated 10 000 data sets initialising the random number generator at a common state. The within-study variances where generated only once and kept fixed while generating the samples.

Zeng & Lin (2015) compared the performance of their proposed double resampling method with the DerSimonian & Laird (1986) method, the profile likelihood method in Hardy & Thompson (1996), and the resampling method in Jackson & Bowden (2009) and showed that the double resampling method improves the accuracy of statistical inference. Based on these results Kosmidis et al. (2017) compared the performance of their mean BRPL approach with the double resampling method and illustrated that the former results in confidence intervals with coverage probabilities closer to the nominal level that the alternative methods.

We take advantage of the results reported in Zeng & Lin (2015) and Kosmidis et al. (2017) and evaluate the performance of estimation and inference based only on the median BRPL with that based on the ML and the mean BRPL. The estimators of the fixed and random-effect parameters obtained from the three methods are calculated using variants of the two-step algorithm described in Section 6.7. In the second step of the algorithm the candidate values for the ML, and maximum mean and median BRPL estimators of the between-study variance ψ are calculated by searching for the root of the partial derivatives of $l(\theta)$, $l^*(\theta)$, and $l^\dagger(\theta)$ with respect to ψ, in the interval $(0,3)$.

First, we compare the performance of the ML, maximum mean BRPL and maximum median BRPL estimators in terms of percentage of underestimation. Figure 6.3 shows that the median bias-reducing adjustment is the most effective in reducing median bias even for small values of K. As expected, the ML and maximum mean BRPL estimators also approach the limit of 50% underestimation as K grows, with the latter being closer to 50% than the former. Figure 6.4 shows that maximum median BRPL is also effective in reducing the mean bias of the ML estimator of ψ but only for moderate to large values of K, while maximum mean BRPL results in estimators with the smallest bias.

Figures 6.5 and 6.6 show the estimated coverage probability for the one-sided and two-sided confidence intervals for β based on the LR, mean BRPL ratio and median BRPL ratio statistics at the 95% nominal level. Figure 6.7 shows the estimated coverage probability for the two-sided confidence intervals for ψ based on the LR, mean BRPL ratio and median BRPL ratio statistics at the 95% nominal level. For small values of ψ or small and moderate number of studies K the empirical coverage of the intervals is larger than the nominal 95% level. In general, the confidence intervals based on mean and median BRPL ratio have empirical coverage that is closer to the nominal level with the latter having generally better coverage. The differences between the three methods diminish as the number of studies K increases.

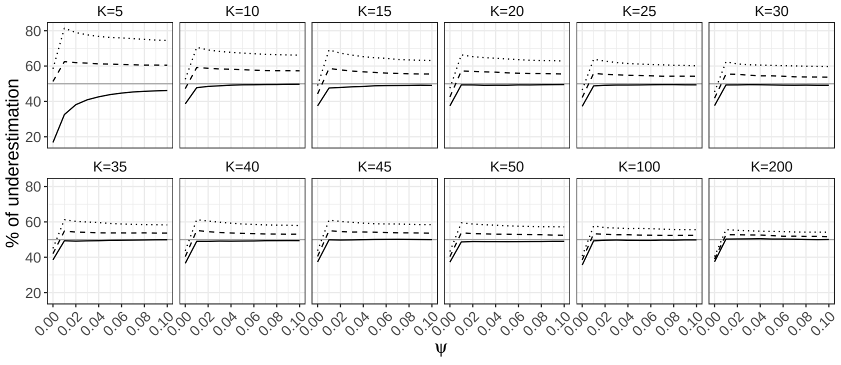

Figure 6.3: Empirical percentage of underestimation for ψ for random effects meta-analysis. The percentage of underestimation is calculated for increasing values of ψ in the interval $[0, 0.1]$ and with K ranging from 5 to 200. The curves correspond to the maximum median BRPL (solid), maximum mean BRPL (dashed), and ML (dotted) estimators. The grey horizontal line is at the target of 50% underestimation.

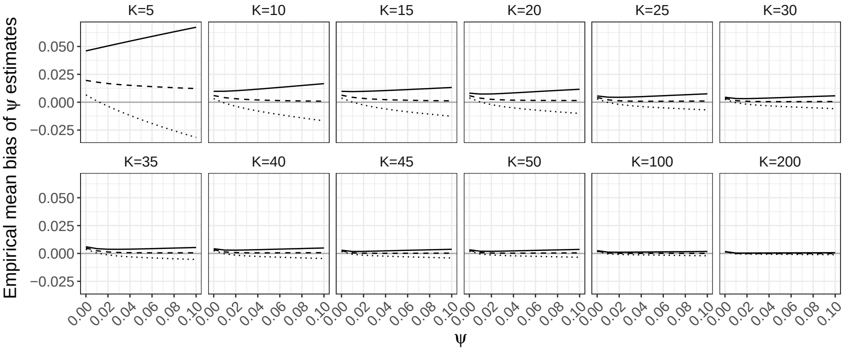

Figure 6.4: Empirical mean bias of ψ estimates for random effects meta-analysis. The mean bias is calculated for increasing values of ψ in the interval $[0, 0.1]$ and with K ranging from 5 to 200. The curves correspond to the maximum median BRPL (solid), maximum mean BRPL (dashed), and ML (dotted) estimators. The grey horizontal line is at zero.

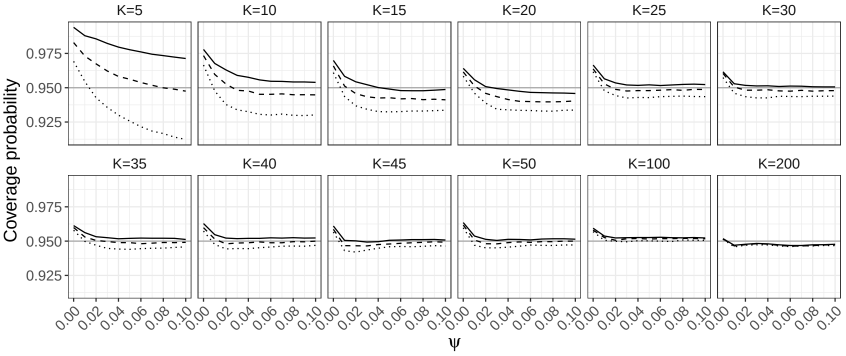

Figure 6.5: Empirical coverage probabilities of one-sided (right) confidence intervals for β for random effects meta-analysis. The curves correspond to nominally 95% confidence intervals based on the median BRPL ratio (solid), the mean BRPL ratio (dashed), and the LR (dotted). The grey horizontal line is at the 95% nominal level.

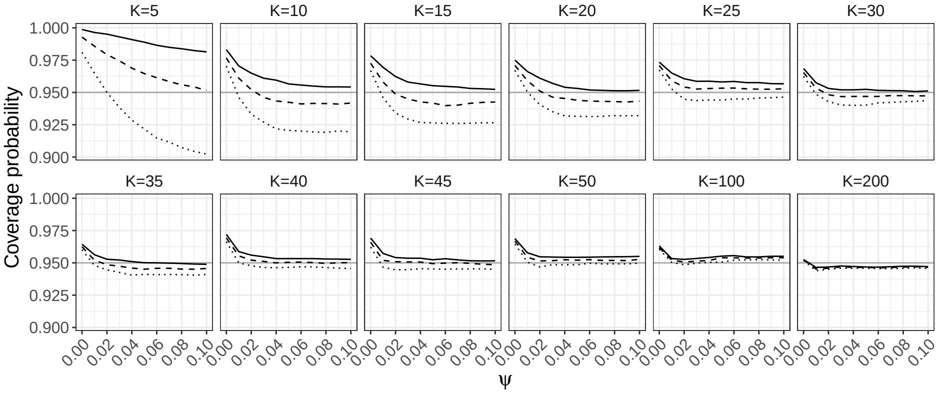

Figure 6.6: Empirical coverage probabilities of two-sided confidence intervals for β for random effects meta-analysis. The curves correspond to nominally 95% confidence intervals based on the median BRPL ratio (solid), the mean BRPL ratio (dashed), and the LR (dotted). The grey horizontal line is at the 95% nominal level.

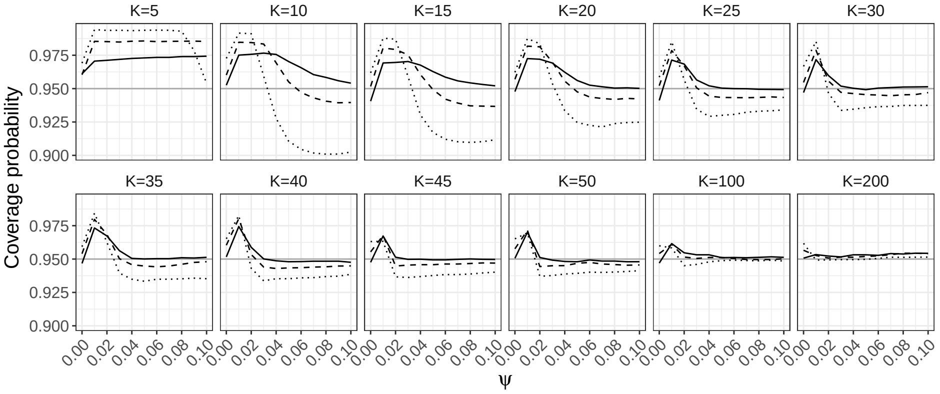

Figure 6.7: Empirical coverage probabilities of two-sided confidence intervals for ψ for random effects meta-analysis. The empirical coverage is calculated with $\beta = 0.5$. The curves correspond to nominally 95% confidence intervals based on the median BRPL ratio (solid), the mean BRPL ratio (dashed), and the LR (dotted). The grey horizontal line is at the 95% nominal level.

Figures 6.8 and 6.9 give the power of the LR, the mean BRPL ratio, and the median BRPL ratio tests for testing the null hypothesis $\beta = 0.5$ against various alternatives. Specifically, we simulated 10 000 data sets under the alternative hypothesis that parameter β is equal to $b = 0.5 + \delta K^{-1/2}$, where δ ranges from 0 to 2.25. In Figure 6.8 the power is calculated using critical values of the the asymptotic null χ^2_1 distribution of the statistics. In Figure 6.9 the power is calculated using critical values based on the exact null distribution of each statistic, obtained by simulation under the null hypothesis. In this way, the three tests are calibrated to have size 5%.

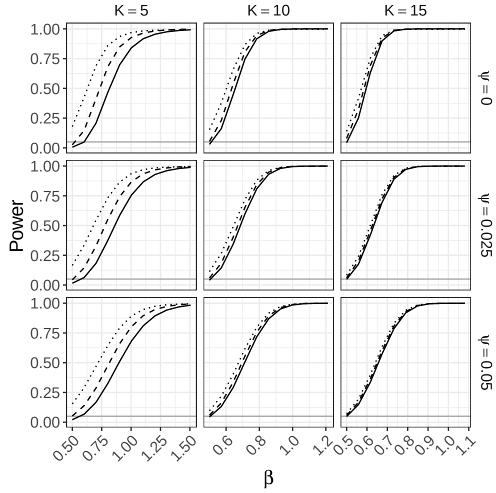

Figure 6.8: Empirical power of the likelihood-based tests of asymptotic level 0.05 for random effects meta-analysis for testing $\beta = 0.5$. The empirical power is calculated for increasing values of β, for $K \in \{5,10,15\}$ and $\psi \in \{0,0.025,0.05\}$. The curves correspond to median BRPL ratio (solid), mean BRPL ratio (dashed), and LR (dotted) tests. The grey horizontal line is at the 5% nominal size.

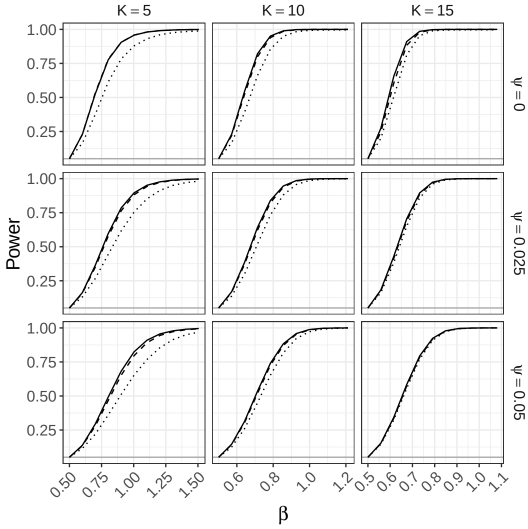

Figure 6.9: Empirical power of the likelihood-based tests of exact level 0.05 for random effects meta-analysis for testing $\beta = 0.5$. The empirical power is calculated for increasing values of β, for $K \in \{5,10,15\}$ and $\psi \in \{0,0.025,0.05\}$. The curves correspond to median BRPL ratio (solid), mean BRPL ratio (dashed), and LR (dotted) tests. The grey horizontal line is at the 5% nominal size.

Figure 6.8 shows that the three tests have monotone power and for small values of K the LR test yields the largest power. This is because the LR test is oversized, while the mean and median BRPL ratio tests are slightly more conservative and this conservativeness comes at the cost of lower power. As the number of studies K increases the three tests approach the nominal size and provide similar power.

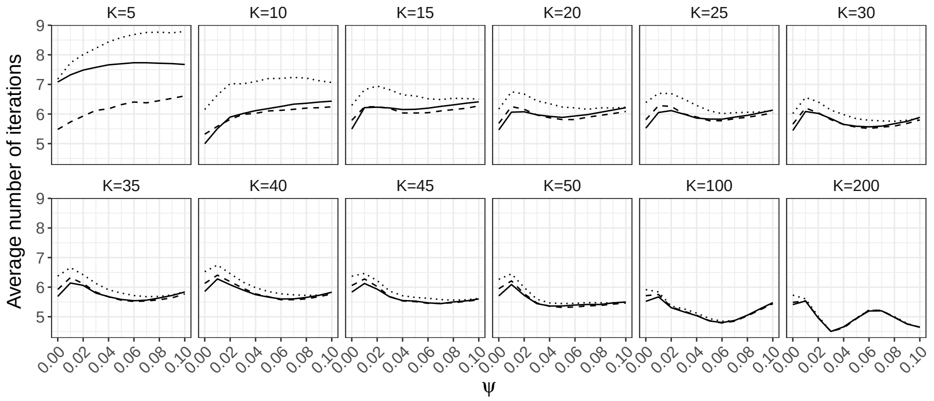

Figure 6.10: Average number of iterations until the two-step iterative process converges for random effects meta-analysis for increasing values of ψ in the interval $[0, 0.1]$ and with K ranging from 5 to 200. The curves correspond to the maximum median BRPL (solid), maximum mean BRPL (dashed), and ML (dotted) estimators.

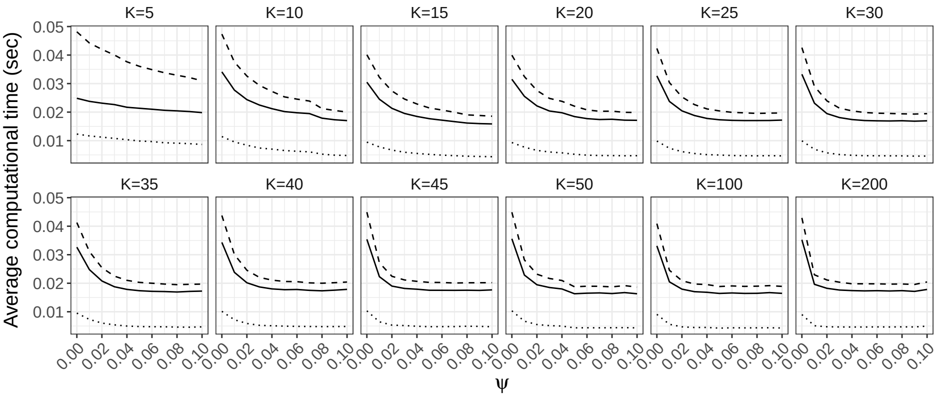

Figure 6.11: Average computational run-time per fit for random effects meta-analysis for increasing values of ψ in the interval $[0, 0.1]$ and with K ranging from 5 to 200. The curves correspond to the maximum median BRPL (solid), maximum mean BRPL (dashed), and ML (dotted) estimators.

The use of the exact critical values in Figure 6.9 allows us to compare the performance of the tests without letting the oversizing or the conservativeness of a test skew the power results. Figure 6.9 shows that the power of the median BRPL ratio test is almost identical to that of the mean BRPL ratio test, and both tests have larger power than the LR test. Again, inference based on either of the two penalised likelihoods becomes indistinguishable from classical likelihood inference as the number of studies increases.

Lastly, it is worth noting that across all ψ and K values considered, the average number of iterations taken per fit for the two-step iterative process to converge is six iterations for every method. The average computational run-time for the algorithm to run is $0.008, 0.022$, and 0.017 seconds for the ML, mean BRPL, and median BRPL methods, respectively. Figures 6.10 and 6.11 show the average number of iterations and the average computational run-time taken per fit for the two-step iterative process to converge for each value of K and ψ used in the simulation study. The results show that in all cases convergence of the algorithm is achieved after a small number of iterations for all three methods, and the difference in computational run-time from ML for the two bias reducing methods is small.

6.12 Meat consumption data

In this section, we revisit the meat consumption data (Larsson & Orsini, 2014) used in Section 2.9.4 as an example of random effects meta-regression.

Table 6.11 gives the ML estimates, the maximum mean BRPL estimates, and the maximum median BRPL estimates of the fixed effects and the heterogeneity parameter, along with the corresponding estimated standard errors and the 95% confidence intervals. The ML estimates of ψ and the estimated standard errors for the fixed effects have the smallest values. The LR test indicates some evidence for a higher risk associated to the consumption of red processed meat with a p-value of 0.047. On the other hand, the mean BRPL ratio test suggests that there is weaker evidence for higher risk with a p-value of 0.066. The median BRPL ratio test also gives weak evidence for higher risk with a p-value of 0.074. The estimation algorithm used for computing the ML, maximum mean BRPL, and maximum median BRPL estimates converged in 8, 9, and 12 iterations, respectively. The computational run-time for the two-step iterative process which computes the ML, maximum mean BRPL, and maximum median BRPL estimates is 1.2×10^{-2}, 2.4×10^{-2}, and 1.5×10^{-2} seconds, respectively.

Next, we performed a simulation study in order to further investigate the performance of the three methods in a meta-regression context. We simulated 10000 independent samples from the meta-regression model at the ML estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\psi})^{\mathrm{T}}$ reported in Table 6.11. Figure 6.12 shows boxplots of the estimates of ψ. ML underestimates the heterogeneity parameter, while mean BRPL and median BRPL almost fully compensate for the negative bias of ML estimates, with the latter having slightly

Table 6.11: ML, mean BRPL, and median BRPL estimates of the model parameters for the meat consumption data. Estimated standard errors are reported in parentheses. The 95% confidence intervals based on the LR, mean BRPL ratio and median BRPL ratio are reported in squared brackets.

Method	β_0	β_1	ψ
ML	0.099 (0.044)	0.106 (0.061)	0.009
	[-0.004,0.189]	[-0.022,0.244]	[0.003,0.030]
Mean BRPL	0.095 (0.050)	0.110 (0.069)	0.012
	[-0.020,0.199]	[-0.040,0.264]	[0.003,0.042]
Median BRPL	0.093 (0.052)	0.111 (0.072)	0.013
	[-0.027,0.203]	[-0.048,0.271]	[0.004,0.048]

Notes: ML, Maximum likelihood; Mean BRPL, Mean bias-reducing penalised likelihood; Median BRPL, Median bias-reducing penalised likelihood.

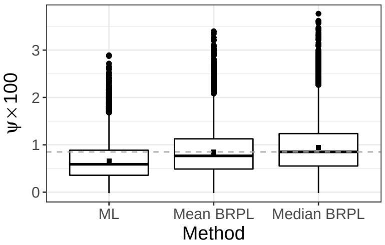

Figure 6.12: Boxplots for the ML, mean BRPL, and median BRPL estimates of ψ as calculated from 10000 simulated samples under the ML fit using the meat consumption data. The square point is the mean of the estimates obtained from each method. The dashed grey horizontal line is at the parameter value used to generate the data.

Table 6.12: Empirical p-value distribution (%) for the tests based on the LR statistic, the mean BRPL ratio statistic, and the median BRPL ratio ratio statistic using the meat consumption data.

$\alpha \times 100$	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
LR	2.2	4.5	7.7	13.1	28.0	50.0	71.7	86.6	92.1	95.3	97.7
Mean BRPL ratio	1.3	3.0	5.6	11.1	25.9	49.8	73.8	89.0	94.2	96.9	98.6
Median BRPL ratio	1.0	2.5	4.9	9.9	25.1	49.7	74.7	89.8	94.8	97.5	98.9

Notes: LR, Likelihood ratio statistic; Mean BRPL ratio, Mean BRPL ratio statistic; Median BRPL ratio, Median BRPL ratio statistic. Each column gives the coverage probability of $(1-\alpha)$% confidence intervals based on the three statistics.

heavier tails.

The percentages of underestimation when estimating ψ are 72.6%, 56.6%, and 49.9% for the ML, maximum mean BRPL, and maximum median BRPL estimators, respectively. The results indicate that median BRPL performs best in terms of percentage of underestimation, followed by the mean BRPL.

The simulated samples were also used to calculate the empirical p-value distribution for the tests based on the likelihood, mean BRPL and median BRPL ratio statistics. Table 6.12 shows that the empirical p-value distributions for the mean and the median BRPL ratio statistics are the ones closest to uniformity.

6.13 Concluding remarks

In this chapter we derive the adjusted score equations for the median bias reduction of the ML estimator for linear mixed models under any parameterisation of the variance components. We show that the median bias-reducing adjusted score function of the variance components differs from the relative mean bias reducing function by one extra additive term.

We also derive the median bias-reducing adjusted score equations for random effects meta-analysis and meta-regression models. In the random-effects meta-regression context, we show that the solution of the median bias-reducing adjusted score equations is equivalent to maximising a penalised log-likelihood. The logarithm of that penalised likelihood differs from the logarithm of the mean BRPL in Kosmidis et al. (2017) by a simple additive term. The main advantage of the maximum median BRPL estimators from the maximum mean BRPL ones is their equivariance under monotone component-wise parameter transformations, which leads to median bias-reduced standard errors.

Using various settings we were able to retrieve enough information on the performance of the maximum median BRPL estimators. All the simulation studies illustrate that use of the median bias-reducing adjusted scores succeeds in achieving median centring in estimation, and results in confidence intervals with good coverage properties. This chapter also provides evidence that median bias reduction corrects the anti-conservativeness of the traditional Wald test in the linear mixed models framework, and the good performance of the Wald statistic using the median BR estimates is comparable with the KR statistic (Kenward & Roger, 1997) and the Wald statistic using the mean BR estimates. Furthermore, in the random effects meta-regression framework, while tests based on the LR suffer from size distortions, the median BRPL ratio statistic results in tests with size and power properties, sometimes better than those of the mean BRPL ratio statistic in Kosmidis et al. (2017).

Chapter 7

Final Remarks

7.1 Summary of the thesis

The current thesis explores solutions to the important problem of reducing the bias in the estimation of mixed models, which are widely used for modelling dependence within clustered data. Reducing the bias is important because bias affects the performance of standard inferential procedures, such as the Wald and likelihood ratio tests.

A popular method for reducing the mean and median bias of the ML estimator in regular parametric models is the adjusted score equations (Firth, 1993; Kenne Pagui et al., 2017). The superior properties of the mean and median BR estimators over the ML estimator (see, for example, Kosmidis & Firth, 2009; Kenne Pagui et al., 2017), motivated us to further enhance the impact of the method in improving estimation and inference for mixed models. Specifically, we aimed:

(i) To derive the mean BR adjusted score equations (Firth, 1993) in the case of linear mixed models and simple generalised linear mixed models, and to assess the performance of the bias-reduction method in estimation and inference.

(ii) To widen the applicability of the mean BR adjusted score equations method (Firth, 1993) in models where it cannot be directly implemented, e.g. models with infeasible bias function or intractable likelihood, and to demonstrate the effectiveness of the extended mean BR method, IBLA, in removing the first-order bias from the (approximate) ML estimators.

(iii) To derive the median BR adjusted score equations (Kenne Pagui et al., 2017) in the case of linear mixed models and random effects meta-analysis and meta-regression, and to assess the performance of the bias-reduction method in estimation and inference.

In what follows we describe the main findings from our research related to each of the aforementioned aims.

(i) Derive the mean BR adjusted score equations (Firth, 1993) in the case of linear mixed models and simple generalised linear mixed models, and assess the performance of the bias-reduction method in estimation and inference.

Chapter 2 derives the adjusted score equations in the case of linear mixed models, and Chapter 5 derives the adjusted score equations in the case of logistic linear models with a fixed intercept and a random intercept only. For linear mixed models we showed that, for general parameterisations of the variance-covariance matrix of the distribution of the random effects, the adjusted score function of the variance components differs from the REML score function by an extra additive term. We also showed that for covariance structures with $\partial^2 V(\psi)/\partial\psi_r\partial\psi_s = 0$ for all pairs (r,s), $r,s \in \{1,\dots,m\}$, where $V(\psi)$ is the variance-covariance matrix of the marginal distribution of the responses and ψ is the vector of variance components, the mean BR adjusted score function is the derivative of a mean BRPL which coincides with the REML likelihood. Hence, maximising the mean BRPL is equivalent to calculating the REML estimator for ψ. Furthermore, Chapter 2 demonstrates that the bias of the ML estimates affects Wald-type inference. Numerical studies provide evidence that the adjusted score equations method corrects the anti-conservativeness of the Wald test. The Wald-type confidence intervals for the mean BR estimates have good coverage properties. The corresponding coverage probabilities obtained from the Kenward & Roger (1997) method are also close to the nominal level, but the method has the disadvantage of being computationally more expensive. Finally, complementing the work in Kosmidis et al. (2017), we demonstrate the successful mean bias reduction in estimation and the good coverage properties of the LR confidence intervals under the framework of the random-effects meta-analysis and meta-regression models.

Chapter 5 demonstrates that the adjusted score equations method preserves the good estimation and coverage properties in the framework of logistic linear models with a fixed intercept and a random intercept only. However, the implementation of the method in more complex generalised linear mixed models can be challenging, because it requires the approximation of a large number of intractable integrals, and a high number of quadrature points for their accurate approximation.

(ii) Widen the applicability of the mean BR adjusted score equations method (Firth, 1993) in models where it cannot be directly implemented, e.g. models with infeasible bias function or intractable likelihood, and demonstrate the effectiveness of the extended mean BR method, IBLA, in removing the first-order bias from the (approximate) ML estimators.

Chapter 3 and Chapter 4 present extended versions of the traditional adjusted score equations for mean bias reduction, and introduce the IBLA algorithm for the computation of the bias-reduced estimates. The asymptotic properties of the resulting bias-reduced estimators are also derived. Specifically, in Chapter 3 we propose an adjusted score equation that can be used to derive mean bias-reduced estimates for models with infeasible bias function, and in Chapter 4 we propose an adjusted score equation that can be used for bias reduction in models with intractable likelihood. There are two main differences between the proposed adjusted score equations and the traditional adjusted score equation (Firth, 1993). These are (i) the use of a Monte Carlo approximation of the bias function, instead of using the bias function, in order to achieve feasibility of the equations, and (ii) the use of the derivatives of a suitably approximated log-likelihood in order to achieve tractability, instead of obtaining explicit expressions for the adjusted score functions and then approximating them. We show that the IBLA estimators obtained for models with infeasible bias function or intractable likelihood are consistent and asymptotically normally distributed. Two additional important findings are that the Monte Carlo size used for approximating the bias function must be of order $O(n^{\alpha})$, with $\alpha \geq 1$, in order to achieve mean bias reduction, and that Laplace approximation is suitable for use in the simulation-based approximate adjusted score function. Even though our theoretical findings do not cover the case of models where the bias function is not continuous (e.g. in generalised linear models with discrete responses), numerical studies in Chapter 3 demonstrate that the simulation-based adjusted score equation method performs well in terms of estimation and inference.

The IBLA algorithm is a quasi Newton-Raphson iteration that can compute the solution of the new bias-reducing adjusted score equations. Starting from the ML estimate (or the maximum approximate likelihood estimate in the case of models with intractable likelihood), a single iteration gives the (approximate) parametric bootstrap estimates. The advantages of IBLA against parametric bootstrap are: (i) based on numerical studies we have evidence that even though parametric bootstrap performs well in reducing the bias of ML estimates, IBLA performs better both in terms of bias and mean squared error, (ii) numerical studies also indicate that Wald-type confidence intervals have better coverage properties when using IBLA instead of parametric bootstrap estimates, (iii) IBLA gives finite estimates even when the ML estimates are infinite with positive probability, while parametric bootstrap estimates by definition depend on the finiteness of the ML estimates. The finiteness of IBLA estimates is achieved by suitably modifying in each iteration the simulated responses y to $y^c = c + y(1-2c)$, where c is a small positive constant. The only advantage of parametric bootstrap against IBLA is that it is computationally less expensive.

Chapter 5 examines the performance of IBLA in the framework of generalised linear mixed models and compares it with standard estimation methods (ML using numerical integration, PQL) and bias reduction methods (corrected PQL, approximate parametric bootstrap) that have been developed in the literature. Even though IBLA is the computationally most expensive method, it is the most accurate in fitting generalised linear mixed models with a random intercept, even under challenging settings with a small cluster size or with a large random-effects variance.

(iii) Derive the median BR adjusted score equations (Kenne Pagui et al., 2017) in the case of linear mixed models and random effects meta-analysis and meta-regression, and assess the performance of the bias-reduction method in estimation and inference.

Lastly, we derived the adjusted score equations for median instead of mean bias reduction in the framework of linear mixed models and their special case, random-effects meta-analysis and meta-regression. We showed that the median BR adjusted score function of the variance components differs from the mean BR function by one extra additive term. In the random-effects meta-regression context, we showed that the solution of the median BR adjusted score equations is equivalent to maximising a penalised log-likelihood. The main advantage of the maximum median BRPL estimators from the maximum mean BRPL ones is their equivariance under monotone component-wise parameter transformations, which, in the case of random effects meta-regression, leads to median bias-reduced standard errors.

All in all, achieving our aims offers a large flexibility in the applicability of the adjusted score equations method (Firth, 1993) in linear mixed models. Moreover, there is strong indication that IBLA is superior than other popular estimation methods in generalised linear mixed models in that it yields variance component estimates with smaller bias, which in turn improve inference on the fixed effects. Despite the theoretical and computational challenges of IBLA, we believe that it will offer practitioners a formal and flexible statistical framework for bias reduction for models with intractable likelihood, that will make an impact in many application areas where bias reduction is beneficial.

7.2 Further work on IBLA

This section lists some topics for future research that were not covered in this thesis. Some of the items in the list have already been mentioned in the concluding remarks of the chapters and others relate to implementing IBLA on various application areas.

1. Reduce the computational run-time of IBLA.

   The computational efficiency of the algorithm mainly depends on the Monte Carlo size R used for the calculation of the simulation-based bias function. The simulation cost of this procedure could be reduced by the use of variance-reduction techniques (see, for example, Fieller & Hartley, 1954; Davidson & MacKinnon, 1992), which would increase the precision of the estimated bias function by decreasing the variability of the Monte Carlo simulation output. Hence, a more accurate estimate of the bias function could be obtained with a smaller value of R.

2. Develop a rule for the selection of an optimal Monte Carlo size R.

   Theorem 4 shows that in order to achieve bias reduction, R must be of order $O(n^{\alpha})$, with $\alpha \geq 1$. We aim to research if we can develop a more precise selection rule for R that governs the accuracy of the simulation-based adjusted score function and, as a result, the accuracy of the IBLA estimates.

3. Assess the adequacy of numerical integration for bias reduction using IBLA.

   In Section 4.5 we showed that Laplace approximation (Tierney & Kadane, 1986) satisfies the conditions in Theorem 5, and therefore it can be used to approximate the likelihood and yield a tractable adjusted score equation whose solution has smaller mean bias than the maximum approximate likelihood estimate. We aim to investigate if other approximation techniques, such as the Gauss-Hermite quadrature (Abramowitz & Stegun, 1965), the adaptive Gauss-Hermite quadrature (Liu & Pierce, 1994) or the sequential reduction method (Ogden, 2015), also satisfy the conditions in Theorem 5, and can therefore be used for bias reduction via IBLA.

4. Explore further the asymptotic properties of the IBLA estimates.

   A formal proof for the asymptotic normality of the IBLA estimator when the observations are independent but non-identically distributed remains to be formulated. The proof is going to make use of the Lindeberg central limit theorem (Van der Vaart, 2000, Proposition 2.27).

5. Study the performance of IBLA using generalised linear mixed models with more complex structures and other response distributions.

   The numerical studies in Chapter 5 evaluate the performance of IBLA in estimation and inference for the logistic linear models with a random intercept and responses following a binomial distribution. It would be interesting to see if IBLA performs well in the estimation of more complex models, e.g. models with a random intercept and a random slope, or models with crossed random effects. Moreover, we plan to explore the performance of IBLA in generalised linear mixed models with other response distributions, such as the Poisson distribution.

6. Derive which parameterisation for variance-covariance matrix is best for IBLA.

   In the current thesis we only considered one parameterisation for unstructured variance-covariance matrices (the elements of the Cholesky factor of the variance-covariance matrix) and one parameterisation for structured (the variance of the random effects and their correlation). We plan to investigate how

IBLA behaves under other parameterisations for variance-covariance matrices that allow unconstrained estimation of the associated parameters. For example, it may be that for some of the parameterisations the algorithm convergences faster to a solution of the adjusted score equations, and this could be used as a criterion for choosing among them.

7. Development of statistical software for the public use of IBLA.

   We plan to create an R package (R Core Team, 2017) for bias reduction via adjusted score equations in mixed models, in order to make it easy for practitioners to use either IBLA or the traditional adjusted score equations method (where applicable).

8. Implement IBLA in Item Response Theory (IRT) models.

   Our future research agenda also includes studying the performance of IBLA in IRT models, a special modelling setting which has multiple uses in educational testing, psychometrics and other disciplines (Baker & Kim, 2004). The label "item response theory" reflects the dependence of the theory upon an examinee's responses to items. Consider realisations y_{is} of independent random variables Y_{is}, $i = 1,\ldots,N$, $s = 1,\ldots,n$. A general model for the probability of correct response for the sth examinee in the ith item given the ability level γ_s is given by $P(y_{is} = 1|\gamma_s) = c_i + (1-c_i)g\{d_i(\gamma_s + \beta_i)\}$, where $g(\cdot)$ is a link function, y_{is} is the dichotomous response, γ_s is the sth examinee's level on the latent scale, c_i is the guessing parameter, d_i is the discrimination parameter, and β_i is the easiness parameter for the ith item. The guessing parameter expresses the probability that an examinee with very low ability responds correctly to an item by chance, the discrimination parameter quantifies how well the item distinguishes between subjects with low/high standing in the latent scale, and the easiness parameter expresses the difficulty level of the item (Rizopoulos, 2006).

   The IRT models fit within the framework of generalised linear mixed models, because they introduce latent variables to account for the heterogeneity across participants. For example, the one-parameter logistic model assumes that the link function is the logistic function, that there is no guessing parameter ($c_i = 0$)

and the discrimination parameter d_i is fixed. A special case of the one-parameter logistic model is the Rasch model (Rasch, 1960) for which $d_i = 1$, for all i and is defined as $P(y_{is} = 1|\gamma_s) = g(\gamma_s + \beta_i)$. The Rasch model can be seen as a generalised linear mixed model if β_i (items) are considered as fixed effects and γ_s (abilities) are considered as random effects.

Despite the simplicity of the models getting good parameter estimates is challenging. These challenges involve dealing with biased ML variance components estimates (Agresti, 2002, Section 12.1.5) and infinite estimates, for example, when the persons are perfectly separated with respect to a specific item. As such, they provide a natural testing ground for the methodology that has been developed in this thesis.

9. Implement IBLA in logistic regression models in meta-analysis.

   The logistic regression models in meta-analysis are used for synthesizing information from different studies reporting dichotomous data, such as death or occurrence of a disease (Simmonds & Higgins, 2016). When raw individual participant data of dichotomous outcomes are available from each study a one-stage mixed-effects logistic regression model may be used to estimate an overall summary effect in a single analysis (Turner et al., 2000), instead of using the traditional two-stage approach which first estimates the effect in each study and then combines the effects in a meta-analysis.

   For analyses of controlled trials comparing an experimental against a control treatment Simmonds & Higgins (2016) define the one-stage mixed-effects logistic regression model as $g(p_{ij}) = (t + v_i)x_{ij} + \phi_i$, where p_{ij} is the probability of an event in treatment arm j (1 for experimental, 0 for control) of trial i, t is the average treatment effect, v_i represents the deviation of each trial's true treatment effect (log-odds ratio) from the average, $x_{ij} = 0/1$ indicates the control/treatment group for the jth individual in the ith trial, and ϕ_i is the baseline risk of the event in the ith trial. The link function $g(\cdot)$ is typically the logit link, in which case t is a log risk ratio. It is assumed that v_i is normally distributed with mean zero and variance τ^2, i.e. τ^2 is the heterogeneity in treatment effects across trials. It is also assumed that the parameters ϕ_i are fixed and unrelated across studies.

ML does not take into account the use of the same data in the estimation of the fixed effects, and therefore the estimate of τ^2 is generally negatively biased (Turner et al., 2000). The results in Turner et al. (2000) also provide evidence that the bias in the ML estimates affects the Wald confidence intervals, which are narrower than expected. As such, it would be interesting to derive the adjusted score equations for the logistic regression models in meta-analysis, and study the estimation and inferential properties of the IBLA bias-reduced estimates.

Bibliography

ABRAMOWITZ, M. & STEGUN, I. A. (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical table*, vol. 2172. Dover New York.

AGRESTI, A. (2002). *Categorical Data Analysis*. Wiley, New Jersey, 2nd edition.

AGRESTI, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

ALBERT, A. & ANDERSON, J. (1984). On the existence of maximum likelihood estimates in logistic models. *Biometrika* **71**, 1–10.

BAKER, F. B. & KIM, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

BARTLETT, M. (1953). Approximate confidence intervals. *Biometrika* **40**, 12–19.

BATES, D., KLIEGL, R., VASISHTH, S. & BAAYEN, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967* .

BEITLER, P. J. & LANDIS, J. R. (1985). A mixed-effects model for categorical data. *Biometrics* **41**, 991–1000.

BELLIO, R. & GUOLO, A. (2016). Integrated likelihood inference in small sample meta-analysis for continuous outcomes. *Scandinavian Journal of Statistics* **43**, 191–201.

BENEDETTO, J. J. & CZAJA, W. (2010). *Integration and modern analysis*. Springer Science & Business Media.

BRESLOW, N. E. & CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

BRESLOW, N. E. & LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

BROCKWELL, S. E. & GORDON, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20**, 825–840.

CALLENS, M. & CROUX, C. (2005). Performance of likelihood-based estimation methods for multilevel binary regression models. *Journal of Statistical Computation and Simulation* **75**, 1003–1017.

COCHRAN, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society* **4**, 102–118.

COX, D. R. & HINKLEY, D. V. (1979). *Theoretical statistics*. London: Chapman & Hall Ltd.

DAVIDSON, R. & MACKINNON, J. G. (1992). Regression-based methods for using control variates in monte carlo experiments. *Journal of Econometrics* **54**, 203–222.

DAVISON, A. C. (2003). *Statistical models*, vol. 11. Cambridge University Press.

DEMIDENKO, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

DERSIMONIAN, R. & LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

EDWARDS, L. J., MULLER, K. E., WOLFINGER, R. D., QAQISH, B. F. & SCHABENBERGER, O. (2008). An r^2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine* **27**, 6137–6157.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics* **3**, 1189–1242.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

EFRON, B. & TIBSHIRANI, R. J. (1993). An introduction to the bootstrap, monographs on statistics and applied probability, vol. 57. *New York and London: Chapman and Hall* .

FELLER, W. (2008). *An introduction to probability theory and its applications*, vol. 2. John Wiley & Sons.

FIELLER, E. & HARTLEY, H. (1954). Sampling with control variables. *Biometrika* **41**, 494–501.

FIRTH, D. (1992). Bias reduction, the jeffreys prior and glim. In *Advances in GLIM and Statistical Modelling*. Springer, pp. 91–100.

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

GALANT, D. (1969). Gauss quadrature rules for the evaluation of $2\pi^{-1/2}\int_0^\infty \exp(-x^2)f(x)dx$. *Mathematics of Computation* **23**, 674–s39.

GOURIEROUX, C., MONFORT, A. & RENAULT, E. (1993). Indirect inference. *Journal of Applied Econometrics* **8**, 85–118.

GUMEDZE, F. & DUNNE, T. (2011). Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications* **435**, 1920–1944.

GUOLO, A. & VARIN, C. (2017). Random-effects meta-analysis: the number of studies matters. *Statistical methods in medical research* **26**, 1500–1518.

HARDY, R. J. & THOMPSON, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* **15**, 619–629.

HARVILLE, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

HASSELMAN, B. (2017). *nleqslv: Solve Systems of Nonlinear Equations*. R package version 3.3.1.

HEINZE, G. & SCHEMPER, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

HUBER, P., RONCHETTI, E. & VICTORIA-FESER, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 893–908.

HUIZENGA, H. M., VISSER, I. & DOLAN, C. V. (2011). Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology* **64**, 1–19.

JACKSON, D. & BOWDEN, J. (2009). A re-evaluation of the "quantile approximation method" for random effects meta-analysis. *Statistics in Medicine* **28**, 338–348.

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

JIANG, J. et al. (2013). The subset argument and consistency of mle in glmm: Answer to an open problem and beyond. *The Annals of Statistics* **41**, 177–195.

KACKAR, R. N. & HARVILLE, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853–862.

KENNE PAGUI, E. C., SALVAN, A. & SARTORI, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* **104**, 923–938.

KENWARD, M. G. & ROGER, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997.

KNAPP, G. & HARTUNG, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* **22**, 2693–2710.

KOSMIDIS, I. (2014a). Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics* **6**, 185–196.

KOSMIDIS, I. (2014b). Improved estimation in cumulative link models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 169–196.

KOSMIDIS, I. & FIRTH, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96**, 793–804.

KOSMIDIS, I., GUOLO, A. & VARIN, C. (2017). Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika* **104**, 489–496.

KUK, A. Y. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 395–407.

LAIRD, N. M. & WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

LARSSON, S. C. & ORSINI, N. (2014). Red meat and processed meat consumption and all-cause mortality: a meta-analysis. *American Journal of Epidemiology* **179**, 282–289.

LAVRENT'EV, M. M. & SAVEL'EV, L. J. (2006). *Operator theory and ill-posed problems*, vol. 50. Walter de Gruyter.

LIANG, H., WU, H. & ZOU, G. (2008). A note on conditional aic for linear mixed-effects models. *Biometrika* **95**, 773–778.

LIN, X. & BRESLOW, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, 1007–1016.

LINDSTROM, M. J. & BATES, D. M. (1988). Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**, 1014–1022.

LIU, Q. & PIERCE, D. A. (1994). A note on gausshermite quadrature. *Biometrika* **81**, 624–629.

LONGFORD, N. (1993). *Random Coefficient Models*. Oxford University Press.

MCCULLAGH, P. (1987). *Tensor methods in statistics*, vol. 161. Chapman and Hall London.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized linear models*, vol. 37. CRC press.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.

MCCULLOCH, C. E., SEARLE, S. R. & NEUHAUS, J. M. (2008). *Generalized linear mixed models*. Wiley Online Library.

MCKEON, C. S., STIER, A. C., MCILROY, S. E. & BOLKER, B. M. (2012). Multiple defender effects: synergistic coral defense by mutualist crustaceans. *Oecologia* **169**, 1095–1103.

OGDEN, H. E. (2015). A sequential reduction method for inference in generalized linear mixed models. *Electronic Journal of Statistics* **9**, 135–152.

PACE, L. & SALVAN, A. (1997). *Principles of statistical inference from a neo-Fisherian perspective*, vol. 4. World Scientific Publishing Co Inc.

PATTERSON, H. D. & THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

PINHEIRO, J. C. (1994). *Topics in mixed effects models*. Ph.D. thesis, University of Wisconsin–Madison.

PINHEIRO, J. C. & CHAO, E. C. (2012). Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics* **15**, 58–81.

POTTHOFF, R. F. & ROY, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313–326.

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360.

R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RASCH, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. .

RAUDENBUSH, S. W., YANG, M.-L. & YOSEF, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141–157.

RIZOPOULOS, D. (2006). ltm: An r package for latent variable modeling and item response theory analyses. *Journal of statistical software* **17**, 1–25.

RUDIN, W. (1976). *Principles of Mathematical Analysis (International Series in Pure & Applied Mathematics)*. McGraw-Hill Publishing Co.

SCHWARZER, G., CARPENTER, J. R. & RÜCKER, G. (2015). *Meta-analysis with R*. Springer.

SIINO, M., FASOLA, S. & MUGGEO, V. M. (2016). Inferential tools in penalized logistic regression for small and sparse data: A comparative study. *Statistical Methods in Medical Research* **27**, 1365–1375.

SIMMONDS, M. C. & HIGGINS, J. P. (2016). A general framework for the use of logistic regression models in meta-analysis. *Statistical Methods in Medical Research* **25**, 2858–2877.

STEEN, N., BYRNE, G. & GELBARD, E. (1969). Gaussian quadratures for the integrals $\int_0^\infty \exp(-x^2) f(x) dx$ and $\int_0^b \exp(-x^2) f(x) dx$. *Mathematics of Computation* **23**, 661–671.

TAUBERT, D., ROESEN, R. & SCHÖMIG, E. (2007). Effect of cocoa and tea intake on blood pressure: a meta-analysis. *Archives of Internal Medicine* **167**, 626–634.

TIERNEY, L. & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

TRENCH, W. F. (2003). *Introduction to real analysis*. Pearson education.

TURNER, R. M., OMAR, R. Z., YANG, M., GOLDSTEIN, H. & THOMPSON, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **19**, 3417–3432.

VAIDA, F. & BLANCHARD, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika* **92**, 351–370.

VAN DER VAART, A. W. (2000). *Asymptotic statistics*, vol. 3. Cambridge university press.

WAND, M. (2007). Fisher information for generalised linear mixed models. *Journal of Multivariate Analysis* **98**, 1412–1416.

WICKHAM, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

ZENG, D. & LIN, D. Y. (2015). On random-effects meta-analysis. *Biometrika* **102**, 281–294.

Appendix A

Derivation of mean bias-reducing adjusted scores in linear mixed models

The elements of the observed information matrix $j(\theta)$ for the linear mixed model in (2.1) are

$$\begin{aligned}
j_{\beta\beta} &= X^{\mathrm{T}}V(\psi)^{-1}X \\
j_{\beta\psi_r} &= X^{\mathrm{T}}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_r}V(\psi)^{-1}R(\beta) \\
j_{\psi_r\psi_s} &= \frac{1}{2}R(\beta)^{\mathrm{T}}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_r}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_s}V(\psi)^{-1}R(\beta) \\
&\quad +\frac{1}{2}R(\beta)^{\mathrm{T}}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_s}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_r}V(\psi)^{-1}R(\beta) \\
&\quad -\frac{1}{2}R(\beta)^{\mathrm{T}}V(\psi)^{-1}\frac{\partial^2 V(\psi)}{\partial\psi_r\partial\psi_s}V(\psi)^{-1}R(\beta) \\
&\quad -\frac{1}{2}\operatorname{tr}\left(V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_r}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_s}\right)+\frac{1}{2}\operatorname{tr}\left(V(\psi)^{-1}\frac{\partial^2 V(\psi)}{\partial\psi_r\partial\psi_s}\right)
\end{aligned}$$

for $r,s \in \{1,\ldots,m\}$.

Let $t \in \{1,\ldots,p\}$ correspond to an element of parameter β. Then

$$P_t(\theta) = -Q_t(\theta) = \left[\begin{array}{c|c} 0_{p\times p} & P_{1t}(\psi) \\ \hline P_{1t}(\psi)^{\mathrm{T}} & 0_{m\times m}\end{array}\right], \tag{A.1}$$

where $P_{1t}(\psi)$ is a $p\times m$ matrix with (r,s)th element

$$(P_{1t})_{r,s} = X_r^{\mathrm{T}}V(\psi)^{-1}\frac{\partial V(\psi)}{\partial\psi_s}V(\psi)^{-1}X_t^{\mathrm{T}}, \ \ r\in\{1,\ldots,p\}, s\in\{1,\ldots,m\}.$$

Let $t \in \{p+1,\dots,p+m\}$ correspond to an element of parameter ψ. Then

$$P_t(\theta) = \left[\begin{array}{c|c} P_{2t}(\psi) & 0_{p\times m} \\ \hline 0_{m\times p} & P_{3t}(\psi) \end{array}\right], \quad Q_t(\theta) = \left[\begin{array}{c|c} 0_{p\times p} & 0_{p\times m} \\ \hline 0_{m\times p} & -P_{3t}(\psi)+P_{4t}(\psi) \end{array}\right], \qquad \text{(A.2)}$$

where $P_{2t}(\psi)$ is a $p\times p$ matrix with (r,s)th element

$$(P_{2t})_{r,s} = X_r^{\mathrm{T}} V(\psi)^{-1} \frac{\partial V(\psi)}{\partial \psi_{t-p}} V(\psi)^{-1} X_s^{\mathrm{T}}, \ \ r,s \in \{1,\dots,p\}$$

$P_{3t}(\psi)$ is a $m\times m$ matrix with (r,s)th element

$$(P_{3t})_{r,s} = \operatorname{tr}\left(V(\psi)^{-1}\frac{\partial V(\psi)}{\partial \psi_r} V(\psi)^{-1}\frac{\partial V(\psi)}{\partial \psi_s} V(\psi)^{-1}\frac{\partial V(\psi)}{\partial \psi_{t-p}}\right), \ \ r,s \in \{1,\dots,m\}$$

and $P_{4t}(\psi)$ is a $m\times m$ matrix with (r,s)th element

$$(P_{4t})_{r,s} = \frac{1}{2}\operatorname{tr}\left(V(\psi)^{-1}\frac{\partial^2 V(\psi)}{\partial \psi_r \partial \psi_s} V(\psi)^{-1}\frac{\partial V(\psi)}{\partial \psi_{t-p}}\right), \ \ r,s \in \{1,\dots,m\}.$$

The mean bias-reducing adjustment for θ is obtained by plugging the above expressions into (2.6).

Appendix B

Proof of Theorem 4

Proof of Theorem 4: Let $\theta = (\theta_1, \dots, \theta_p)^{\mathrm{T}}$, $s^*_{n,R}(\theta) = (s^*_1(\theta), \dots, s^*_p(\theta))^{\mathrm{T}}$, and $\delta = \hat{\theta}^*_{n,R} - \theta_0$, where $\delta = O_p(n^{-1/2})$ by Theorem 3. Using Einstein summation convention we write the expansion of $s^*_j(\theta)$ around θ_0 as

$$0 = s^*_j(\hat{\theta}^*_{n,R}) = s^*_j + \delta^s s^*_{js} + \frac{1}{2}\delta^{st} s^*_{jst} + \frac{1}{6}\delta^{stu} s^*_{jstu} + O_p(n^{-1}), \tag{B.1}$$

where s^*_{js}, s^*_{jst}, s^*_{jstu} denote the partial derivatives of s^*_j with respect to θ_j ($j \in \{1, \dots, p\}$) evaluated at θ_0, $\delta^{st} = \delta^s \delta^t$, $\delta^{stu} = \delta^s \delta^t \delta^u$, and s, t, u take values in the index set $\{1, \dots, p\}$.

Let $s^*_j = s_j + \hat{A}_j$, where $\hat{A}_j$ is the jth element of $-j_n(\theta_0)\hat{B}_{n,R}(\theta_0)$. We express $\hat{A}_j$ as $\hat{A}_j = A_j + C_j$, where A_j is the jth element of $-j_n(\theta_0)B_n(\theta_0)$ and C_j is the jth element of $-j_n(\theta_0)[\hat{B}_{n,R}(\theta_0) - B_n(\theta_0)]$. The term A_j is $O_p(1)$ and C_j is $O_p(R^{-1/2})O_p(1)$. Let s_{js}, s_{jst}, s_{jstu} denote the partial derivatives of s_j evaluated at θ_0, and $\hat{A}_{js}$, $\hat{A}_{jst}$, $\hat{A}_{jstu}$ denote the partial derivatives of $\hat{A}_j$ evaluated at θ_0. Assume the higher order derivatives of the $O_p(1)$ terms are also $O_p(1)$. Then

$$\begin{aligned}
0 &= s_j + \hat{A}_j + \delta^s s_{js} + \delta^s \hat{A}_{js} + \frac{1}{2}\delta^{st} s_{jst} + \frac{1}{2}\delta^{st}\hat{A}_{jst} + \frac{1}{6}\delta^{stu} s_{jstu} + \frac{1}{6}\delta^{stu}\hat{A}_{jstu} + O_p(n^{-1}) \\
&= s_j + A_j + C_j + \delta^s s_{js} + \delta^s A_{js} + \delta^s C_{js} + \frac{1}{2}\delta^{st} s_{jst} + \frac{1}{2}\delta^{st} A_{jst} + \frac{1}{2}\delta^{st} C_{jst} \\
&\quad + \frac{1}{6}\delta^{stu} s_{jstu} + \frac{1}{6}\delta^{stu} A_{jstu} + \frac{1}{6}\delta^{stu} C_{jstu} + O_p(n^{-1}) \\
&= s_j + A_j + C_j + \delta^s s_{js} + \delta^s A_{js} + \delta^s C_{js} + \frac{1}{2}\delta^{st} s_{jst} + \frac{1}{6}\delta^{stu} s_{jstu} + O_p(n^{-1}),
\end{aligned}$$

where the terms $\delta^{st} A_{jst} = O_p(n^{-1})$, $\delta^{st} C_{jst} = O_p(n^{-(2+a)/2})$, $\delta^{stu} A_{jstu} = O_p(n^{-3/2})$,

$\delta^{stu}C_{jstu} = O_p(n^{-(3+a)/2})$ are incorporated in the $O_p(n^{-1})$ remainder, and $\delta^s C_{js} = O_p(n^{-(1+a)/2})$. Re-expressing in terms of the centred log-likelihood derivatives we have

$$\begin{aligned} 0 &= s_j + A_j + C_j + \delta^s(s_{js} - \mu_{js}) + \delta^s\mu_{js} + \delta^s A_{js} + \delta^s C_{js} + \frac{1}{2}\delta^{st}(s_{jst} - \mu_{jst}) \\ &\quad + \frac{1}{2}\delta^{st}\mu_{jst} + \frac{1}{6}\delta^{stu}(s_{jstu} - \mu_{jstu}) + \frac{1}{6}\delta^{stu}\mu_{jstu} + O_p(n^{-1}) \,. \end{aligned}$$

For simplicity denote $H_{js} = s_{js} - \mu_{js}$, $H_{jst} = s_{jst} - \mu_{jst}$, and $H_{jstu} = s_{jstu} - \mu_{jstu}$, where all are $O_p(n^{1/2})$. Then

$$\begin{aligned} 0 &= s_j + A_j + C_j + H_{js}\delta^s + \mu_{js}\delta^s + A_{js}\delta^s + C_{js}\delta^s + \frac{1}{2}H_{jst}\delta^{st} + \frac{1}{2}\mu_{jst}\delta^{st} \\ &\quad + \frac{1}{6}H_{jstu}\delta^{stu} + \frac{1}{6}\mu_{jstu}\delta^{stu} + O_p(n^{-1}) \\ &= s_j + A_j + C_j + H_{js}\delta^s + \mu_{js}\delta^s + A_{js}\delta^s + C_{js}\delta^s + \frac{1}{2}H_{jst}\delta^{st} + \frac{1}{2}\mu_{jst}\delta^{st} \\ &\quad + \frac{1}{6}\mu_{jstu}\delta^{stu} + O_p(n^{-1}) \,, \end{aligned}$$

where $H_{jstu}\delta^{stu} = O_p(n^{-1})$ and is incorporated in the $O_p(n^{-1})$ remainder.

The above can be expressed as

$$\begin{aligned} -\mu_{js}\delta^s &= s_j + A_j + C_j + H_{js}\delta^s + A_{js}\delta^s + C_{js}\delta^s + \frac{1}{2}H_{jst}\delta^{st} + \frac{1}{2}\mu_{jst}\delta^{st} \\ &\quad + \frac{1}{6}\mu_{jstu}\delta^{stu} + O_p(n^{-1}) \,. \end{aligned}$$

By the second Bartlett identity (Bartlett, 1953) $\mu_{j,s} = -\mu_{js}$ and so, for the matrix inverse of the Fisher information we have $\mu^{j,s} = -\mu^{js}$. Using the latter and solving with respect to δ^r we have

$$\begin{aligned} \delta^r &= \mu^{r,j}s_j + \mu^{r,j}A_j + \mu^{r,j}C_j + \mu^{r,j}H_{js}\delta^s + \mu^{r,j}A_{js}\delta^s + \mu^{r,j}C_{js}\delta^s + \frac{1}{2}\mu^{r,j}H_{jst}\delta^{st} \\ &\quad + \frac{1}{2}\mu^{r,j}\mu_{jst}\delta^{st} + \frac{1}{6}\mu^{r,j}\mu_{jstu}\delta^{stu} + O_p(n^{-2}) \,. \end{aligned}$$

Again, for simplicity denote $s^r = \mu^{r,j}s_j$, $A^r = \mu^{r,j}A_j$, $H^r_s = \mu^{r,j}H_{js}$, $\mu^r_{st} = \mu^{r,j}\mu_{jst}$ and so on, and express the preceding expression as

$$\delta^r = s^r + A^r + C^r + H^r_s\delta^s + A^r_s\delta^s + C^r_s\delta^s + \frac{1}{2}H^r_{st}\delta^{st} + \frac{1}{2}\mu^r_{st}\delta^{st} + \frac{1}{6}\mu^r_{stu}\delta^{stu} + O_p(n^{-2}),$$

where

$$\begin{aligned}
s^r &= O_p(n^{-1/2}) \\
A^r &= O_p(n^{-1}) \\
C^r &= O_p(n^{-(2+a)/2}) \\
H^r_s\delta^s &= O_p(n^{-1/2})O_p(n^{-1/2}) = O_p(n^{-1}) \\
A^r_s\delta^s &= O_p(n^{-1})O_p(n^{-1/2}) = O_p(n^{-3/2}) \\
C^r_s\delta^s &= \mu^{r,j}C_{js}\delta^s = O_p(n^{-1})O_p(n^{-a/2})O_p(n^{-1/2}) = O_p(n^{-(3+a)/2}) \\
H^r_{st}\delta^{st} &= O_p(n^{-1/2})O_p(n^{-1}) = O_p(n^{-3/2}) \\
\mu^r_{st}\delta^{st} &= O_p(1)O_p(n^{-1}) = O_p(n^{-1}) \\
\mu^r_{stu}\delta^{stu} &= O_p(1)O_p(n^{-3/2}) = O_p(n^{-3/2}).
\end{aligned}$$

Reordering the terms in decreasing order we get:

$$\delta^r = \begin{cases} s^r + A^r + H^r_s\delta^s + \frac{1}{2}\mu^r_{st}\delta^{st} + C^r + A^r_s\delta^s + \frac{1}{2}H^r_{st}\delta^{st} + & \\ \frac{1}{6}\mu^r_{stu}\delta^{stu} + C^r_s\delta^s + O_p(n^{-2}) & \text{if } 0 < a < 1 \\ s^r + A^r + H^r_s\delta^s + \frac{1}{2}\mu^r_{st}\delta^{st} + A^r_s\delta^s + \frac{1}{2}H^r_{st}\delta^{st} + & \\ \frac{1}{6}\mu^r_{stu}\delta^{stu} + C^r + O_p(n^{-2}) & \text{if } 1 \le a < 2 \\ s^r + A^r + H^r_s\delta^s + \frac{1}{2}\mu^r_{st}\delta^{st} + A^r_s\delta^s + \frac{1}{2}H^r_{st}\delta^{st} + & \\ \frac{1}{6}\mu^r_{stu}\delta^{stu} + O_p(n^{-2}) & \text{if } a \ge 2 \end{cases}$$

Using the iterative substitution method with

$$\delta^r = \begin{cases} s^r + A^r + H^r_s\delta^s + \frac{1}{2}\mu^r_{st}\delta^{st} + C^r + O_p(n^{-3/2}) & \text{if } 0 < a < 1 \\ s^r + A^r + H^r_s\delta^s + \frac{1}{2}\mu^r_{st}\delta^{st} + O_p(n^{-3/2}) & \text{if } a \ge 1 \end{cases}$$

we get

$$\delta^r = \begin{cases} s^r + A^r + H^r_s s^s + \frac{1}{2}\mu^r_{st}s^{st} + C^r + C^r_s s^s + H^r_s A^s + \mu^r_{st}s^s A^t \\ +A^r_s s^s + H^r_s H^s_t s^t + \mu^r_{st}H^s_u s^{tu} + \frac{1}{2}H^r_s \mu^s_{tu}s^{tu} + \frac{1}{2}\mu^r_{st}\mu^s_{uv}s^{tuv} \\ +\frac{1}{2}H^r_{st}s^{st} + \frac{1}{6}\mu^r_{stu}s^{stu} + H^r_s C^s + \mu^r_{st}s^s C^t + C^r_s A^s \\ +C^r_s H^s_t s^t + \frac{1}{2}C^r_s \mu^s_{tu}s^{tu} + C^r_s C^s + O_p(n^{-2}) & \text{if } 0 < a < \frac{1}{2} \\ \\ s^r + A^r + H^r_s s^s + \frac{1}{2}\mu^r_{st}s^{st} + C^r + C^r_s s^s + H^r_s A^s + \mu^r_{st}s^s A^t \\ +A^r_s s^s + H^r_s H^s_t s^t + \mu^r_{st}H^s_u s^{tu} + \frac{1}{2}H^r_s \mu^s_{tu}s^{tu} + \frac{1}{2}\mu^r_{st}\mu^s_{uv}s^{tuv} \\ +\frac{1}{2}H^r_{st}s^{st} + \frac{1}{6}\mu^r_{stu}s^{stu} + H^r_s C^s + \mu^r_{st}s^s C^t + C^r_s A^s \\ +C^r_s H^s_t s^t + \frac{1}{2}C^r_s \mu^s_{tu}s^{tu} + O_p(n^{-2}) & \text{if } \frac{1}{2} \leq a < 1 \\ \\ s^r + A^r + H^r_s s^s + \frac{1}{2}\mu^r_{st}s^{st} + H^r_s A^s + \mu^r_{st}s^s A^t + A^r_s s^s \\ +H^r_s H^s_t s^t + \mu^r_{st}H^s_u s^{tu} + \frac{1}{2}H^r_s \mu^s_{tu}s^{tu} + \frac{1}{2}\mu^r_{st}\mu^s_{uv}s^{tuv} \\ +\frac{1}{2}H^r_{st}s^{st} + \frac{1}{6}\mu^r_{stu}s^{stu} + C^r + O_p(n^{-2}) & \text{if } 1 \leq a < 2 \\ \\ s^r + A^r + H^r_s s^s + \frac{1}{2}\mu^r_{st}s^{st} + H^r_s A^s + \mu^r_{st}s^s A^t + A^r_s s^s \\ +H^r_s H^s_t s^t + \mu^r_{st}H^s_u s^{tu} + \frac{1}{2}H^r_s \mu^s_{tu}s^{tu} + \frac{1}{2}\mu^r_{st}\mu^s_{uv}s^{tuv} \\ +\frac{1}{2}H^r_{st}s^{st} + \frac{1}{6}\mu^r_{stu}s^{stu} + O_p(n^{-2}) & \text{if } a \geq 2 \end{cases}$$

The asymptotic bias of $\hat{\theta}^*_{n,R}$ is obtained by taking expectations in both sides, where we make use of the following rule:

$$E(H_d) = \begin{cases} O(n^{d/2}) & \text{if } d \text{ is even} \\ O(n^{(d-1)/2}) & \text{if } d \text{ is odd} \end{cases}$$

with H_d being the product of d centered log-likelihood derivatives.

Thus for $a \geq 1$ we have $E(\delta^r) = E\left[s^r + A^r + H^r_s s^s + \frac{1}{2}\mu^r_{st}s^{st}\right] + O_p(n^{-3/2})$, where

$$\begin{aligned}
E(s^r) &= E(\mu^{r,j}s_j) = \mu^{r,j}E(s_j) = 0\\
E(A^r) &= E(\mu^{r,j}A_j) = \mu^{r,j}E(A_j) = \mu^{r,s}E(A_s)\\
E(H^r_s s^s) &= E(\mu^{r,j}H_{js}\mu^{s,k}s_k) = \mu^{r,j}\mu^{s,k}E(H_{js}s_k) = \mu^{r,j}\mu^{s,k}E[(s_{js}-\mu_{js})s_k]\\
&= \mu^{r,j}\mu^{s,k}E(s_{js}s_k) = \mu^{r,j}\mu^{s,k}\mu_{js,k} = \mu^{r,s}\mu^{t,u}\mu_{st,u}\\
E(\mu^r_{st}s^{st}) &= \mu^{r,j}\mu_{jst}E(\mu^{s,j}s_j\mu^{t,k}s_k) = \mu^{r,j}\mu^{s,j}\mu^{t,k}\mu_{jst}E(s_j s_k)\\
&= \mu^{r,j}\mu^{s,j}\mu^{t,k}\mu_{jst}\mu_{j,k} = \mu^{r,j}\mu^{t,k}\mu_{jst}\mu^{s,j}\mu_{j,k}\\
&= \mu^{r,j}\mu^{t,k}\mu_{jst}\delta^s_k, \text{ where } \delta^s_k = 1 \text{ if } s = k \text{ and } 0 \text{ otherwise}\\
&= \mu^{r,j}\mu^{t,s}\mu_{jst} = \mu^{r,s}\mu^{t,u}\mu_{stu}\\
E(C^r) &= E(\mu^{r,j}C_j) = \mu^{r,j}E(C_j)\\
&= \mu^{r,j}E(\{-j_n(\theta_0)[\hat{B}_{n,R}(\theta_0) - B_n(\theta_0)]\}_j)\\
&= 0.
\end{aligned}$$

The latter expectation is equal to zero because $j_n(\theta_0)$ and $[\hat{B}_{n,R}(\theta_0) - B_n(\theta_0)]$ are independent, and $E[\hat{B}_{n,R}(\theta_0) - B_n(\theta_0)] = 0$.

The final expression of $E(\delta^r)$ for $\alpha \geq 1$ is

$$E(\delta^r) = \mu^{r,s}E(A_s) + \frac{1}{2}\mu^{r,s}\mu^{t,u}(2\mu_{st,u} + \mu_{stu}) + O(n^{-3/2}).$$

□

Appendix C

Asymptotic expansion of $\tilde{\theta}_n - \hat{\theta}_n$ in Section 4.5

Let $\theta \in \Re^p$, $\tilde{s}_j(\theta)$ be the jth element of $\tilde{s}_n(\theta)$ ($j \in \{1,\ldots,p\}$), and $s_j(\theta)$ be the jth element of $s_n(\theta)$. Also, let $\delta = \tilde{\theta}_n - \hat{\theta}_n$, $\tilde{\delta} = \tilde{\theta}_n - \theta_0$ and $\hat{\delta} = \hat{\theta}_n - \theta_0$, such that $\delta = \tilde{\delta} - \hat{\delta}$ with $\tilde{\delta} = O_p(n^{-1/2})$ by Theorem 7, and $\hat{\delta} = O_p(n^{-1/2})$. Consider the expansion of $\tilde{s}_j(\theta)$ around θ_0

$$0 = \tilde{s}_j(\tilde{\theta}_n) = \tilde{s}_j + \tilde{\delta}^s \tilde{s}_{js} + \frac{1}{2}\tilde{\delta}^{st}\tilde{s}_{jst} + \frac{1}{6}\tilde{\delta}^{stu}\tilde{s}_{jstu} + \frac{1}{24}\tilde{\delta}^{stuv}\tilde{s}_{jstuv} + \frac{1}{120}\tilde{\delta}^{stuvx}\tilde{s}_{jstuvx} + O_p(n^{-2}),$$

where $\tilde{s}_{js}$, $\tilde{s}_{jst}$, $\tilde{s}_{jstu}$, $\tilde{s}_{jstuv}$, $\tilde{s}_{jstuvx}$ denote the partial derivatives of $\tilde{s}_j$ evaluated at θ_0, $\tilde{\delta}^{st} = \tilde{\delta}^s\tilde{\delta}^t$, $\tilde{\delta}^{stu} = \tilde{\delta}^s\tilde{\delta}^t\tilde{\delta}^u$, $\tilde{\delta}^{stuv} = \tilde{\delta}^s\tilde{\delta}^t\tilde{\delta}^u\tilde{\delta}^v$, $\tilde{\delta}^{stuvx} = \tilde{\delta}^s\tilde{\delta}^t\tilde{\delta}^u\tilde{\delta}^v\tilde{\delta}^x$, and s,t,u,v,x take values in the index set $\{1,\ldots,p\}$. Given the Laplace approximation of the derivatives of the likelihood function has an error of order $O(n^{-1})$, then we can express the above expansion as

$$\begin{aligned}
0 &= s_j + O(n^{-1}) + \tilde{\delta}^s s_{js} + \tilde{\delta}^s O(n^{-1}) + \frac{1}{2}\tilde{\delta}^{st}s_{jst} + \frac{1}{2}\tilde{\delta}^{st}O(n^{-1}) + \frac{1}{6}\tilde{\delta}^{stu}s_{jstu} + \frac{1}{6}\tilde{\delta}^{stu}O(n^{-1}) \\
&\quad + \frac{1}{24}\tilde{\delta}^{stuv}s_{jstuv} + \frac{1}{24}\tilde{\delta}^{stuv}O(n^{-1}) + \frac{1}{120}\tilde{\delta}^{stuvz}s_{jstuvz} + \frac{1}{120}\tilde{\delta}^{stuvz}O(n^{-1}) + O_p(n^{-2}) \\
&= s_j + O(n^{-1}) + \tilde{\delta}^s s_{js} + \tilde{\delta}^s O(n^{-1}) + \frac{1}{2}\tilde{\delta}^{st}s_{jst} + \frac{1}{6}\tilde{\delta}^{stu}s_{jstu} + \frac{1}{24}\tilde{\delta}^{stuv}s_{jstuv} \\
&\quad + \frac{1}{120}\tilde{\delta}^{stuvx}s_{jstuvx} + O_p(n^{-2})
\end{aligned}$$

where all the terms that are of equal or smaller order than $O_p(n^{-2})$ are incorporated in the remainder term. Re-expressing in terms of the centered log-likelihood derivatives we have

$$\begin{aligned} 0 &= s_j + O(n^{-1}) + \tilde{\delta}^s(s_{js} - \mu_{js}) + \tilde{\delta}^s\mu_{js} + \tilde{\delta}^s O(n^{-1}) + \frac{1}{2}\tilde{\delta}^{st}(s_{jst} - \mu_{jst}) + \frac{1}{2}\tilde{\delta}^{st}\mu_{jst} \\ &\quad + \frac{1}{6}\tilde{\delta}^{stu}(s_{jstu} - \mu_{jstu}) + \frac{1}{6}\tilde{\delta}^{stu}\mu_{jstu} + \frac{1}{24}\tilde{\delta}^{stuv}(s_{jstuv} - \mu_{jstuv}) + \frac{1}{24}\tilde{\delta}^{stuv}\mu_{jstuv} \\ &\quad + \frac{1}{120}\tilde{\delta}^{stuvx}(s_{jstuvx} - \mu_{jstuvx}) + \frac{1}{120}\tilde{\delta}^{stuvx}\mu_{jstuvx} + O_p(n^{-2}) . \end{aligned}$$

For simplicity denote the centered log-likelihood derivatives by H_{js}, H_{jst}, H_{jstu}, $\ldots$, where all are $O_p(n^{1/2})$. Then

$$\begin{aligned} 0 &= s_j + O(n^{-1}) + \tilde{\delta}^s H_{js} + \tilde{\delta}^s\mu_{js} + \tilde{\delta}^s O(n^{-1}) + \frac{1}{2}\tilde{\delta}^{st}H_{jst} + \frac{1}{2}\tilde{\delta}^{st}\mu_{jst} + \frac{1}{6}\tilde{\delta}^{stu}H_{jstu} \\ &\quad + \frac{1}{6}\tilde{\delta}^{stu}\mu_{jstu} + \frac{1}{24}\tilde{\delta}^{stuv}H_{jstuv} + \frac{1}{24}\tilde{\delta}^{stuv}\mu_{jstuv} + \frac{1}{120}\tilde{\delta}^{stuvx}\mu_{jstuvx} + O_p(n^{-2}) , \end{aligned}$$

where $H_{jstuvx}\tilde{\delta}^{stuvx}$ is $O_p(n^{-2})$ and is incorporated in the remainder term. The above can be expressed as

$$\begin{aligned} -\mu_{js}\tilde{\delta}^s &= s_j + O(n^{-1}) + H_{js}\tilde{\delta}^s + O(n^{-1})\tilde{\delta}^s + \frac{1}{2}H_{jst}\tilde{\delta}^{st} + \frac{1}{2}\mu_{jst}\tilde{\delta}^{st} + \frac{1}{6}H_{jstu}\tilde{\delta}^{stu} \\ &\quad + \frac{1}{6}\mu_{jstu}\tilde{\delta}^{stu} + \frac{1}{24}H_{jstuv}\tilde{\delta}^{stuv} + \frac{1}{24}\mu_{jstuv}\tilde{\delta}^{stuv} + \frac{1}{120}\mu_{jstuvx}\tilde{\delta}^{stuvx} + O_p(n^{-2}) . \end{aligned}$$

By the second Bartlett identity (Bartlett, 1953) $\mu_{j,s} = -\mu_{js}$ and so, for the matrix inverse of the Fisher information we have $\mu^{j,s} = -\mu^{js}$. Using the latter and solving with respect to $\tilde{\delta}^r$ we have

$$\begin{aligned} \tilde{\delta}^r &= \mu^{r,j}s_j + \mu^{r,j}O(n^{-1}) + \mu^{r,j}H_{js}\tilde{\delta}^s + O(n^{-1})\mu^{r,j}\tilde{\delta}^s + \frac{1}{2}\mu^{r,j}H_{jst}\tilde{\delta}^{st} + \frac{1}{2}\mu^{r,j}\mu_{jst}\tilde{\delta}^{st} \\ &\quad + \frac{1}{6}\mu^{r,j}H_{jstu}\tilde{\delta}^{stu} + \frac{1}{6}\mu^{r,j}\mu_{jstu}\tilde{\delta}^{stu} + \frac{1}{24}\mu^{r,j}H_{jstuv}\tilde{\delta}^{stuv} + \frac{1}{24}\mu^{r,j}\mu_{jstuv}\tilde{\delta}^{stuv} \\ &\quad + \frac{1}{120}\mu^{r,j}\mu_{jstuvx}\tilde{\delta}^{stuvx} + O_p(n^{-3}) . \end{aligned}$$

Again, for simplicity denote $s^r = \mu^{r,j}s_j$, $H^r_s = \mu^{r,j}H_{js}$, $\mu^r_{st} = \mu^{r,j}\mu_{jst}$ and so on, and express the preceding expression as

$$\begin{aligned}\tilde{\delta}^r &= s^r + \mu^{r,j}O(n^{-1}) + H^r_s\tilde{\delta}^s + O(n^{-1})\mu^{r,j}\tilde{\delta}^s + \frac{1}{2}H^r_{st}\tilde{\delta}^{st} + \frac{1}{2}\mu^r_{st}\tilde{\delta}^{st} + \frac{1}{6}H^r_{stu}\tilde{\delta}^{stu} \\ &\quad + \frac{1}{6}\mu^r_{stu}\tilde{\delta}^{stu} + \frac{1}{24}H^r_{stuv}\tilde{\delta}^{stuv} + \frac{1}{24}\mu^r_{stuv}\tilde{\delta}^{stuv} + \frac{1}{120}\mu^r_{stuvx}\tilde{\delta}^{stuvx} + O_p(n^{-3}),\end{aligned}$$

where

$$\begin{aligned} s^r &= O_p(n^{-1/2}) \\ H^r_s\tilde{\delta}^s &= O_p(n^{-1/2})O_p(n^{-1/2}) = O_p(n^{-1}) \\ H^r_{st}\tilde{\delta}^{st} &= O_p(n^{-1/2})O_p(n^{-1}) = O_p(n^{-3/2}) \\ \mu^r_{st}\tilde{\delta}^{st} &= O_p(1)O_p(n^{-1}) = O_p(n^{-1}) \\ H^r_{stu}\tilde{\delta}^{stu} &= O_p(n^{-1/2})O_p(n^{-3/2}) = O_p(n^{-2}) \\ \mu^r_{stu}\tilde{\delta}^{stu} &= O_p(1)O_p(n^{-3/2}) = O_p(n^{-3/2}) \\ H^r_{stuv}\tilde{\delta}^{stuv} &= O_p(n^{-1/2})O_p(n^{-2}) = O_p(n^{-5/2}) \\ \mu^r_{stuv}\tilde{\delta}^{stuv} &= O_p(1)O_p(n^{-2}) = O_p(n^{-2}) \\ \mu^r_{stuvx}\tilde{\delta}^{stuvx} &= O_p(1)O_p(n^{-5/2}) = O_p(n^{-5/2}).\end{aligned}$$

Reordering the terms in decreasing order we get:

$$\begin{aligned}\tilde{\delta}^r &= s^r + H^r_s\tilde{\delta}^s + \frac{1}{2}\mu^r_{st}\tilde{\delta}^{st} + \frac{1}{2}H^r_{st}\tilde{\delta}^{st} + \frac{1}{6}\mu^r_{stu}\tilde{\delta}^{stu} + \frac{1}{6}H^r_{stu}\tilde{\delta}^{stu} + \frac{1}{24}\mu^r_{stuv}\tilde{\delta}^{stuv} \\ &\quad + \mu^{r,j}O(n^{-1}) + \frac{1}{24}H^r_{stuv}\tilde{\delta}^{stuv} + \frac{1}{120}\mu^r_{stuvx}\tilde{\delta}^{stuvx} + O(n^{-1})\mu^{r,j}\tilde{\delta}^s + O_p(n^{-3}).\end{aligned}$$

Similarly,

$$\begin{aligned}\hat{\delta}^r &= s^r + H^r_s\hat{\delta}^s + \frac{1}{2}\mu^r_{st}\hat{\delta}^{st} + \frac{1}{2}H^r_{st}\hat{\delta}^{st} + \frac{1}{6}\mu^r_{stu}\hat{\delta}^{stu} + \frac{1}{6}H^r_{stu}\hat{\delta}^{stu} + \frac{1}{24}\mu^r_{stuv}\hat{\delta}^{stuv} \\ &\quad + \frac{1}{24}H^r_{stuv}\hat{\delta}^{stuv} + \frac{1}{120}\mu^r_{stuvx}\hat{\delta}^{stuvx} + O_p(n^{-3}).\end{aligned}$$

Using the iterative substitution method we get expressions for $\tilde{\delta}^r$ and $\hat{\delta}^r$, which when subtracted lead to an expression for $\delta^r = \tilde{\delta}^r - \hat{\delta}^r$. This is given by

$$\begin{aligned}
\delta^r &= \mu^{r,j}O(n^{-1}) + H^r_s \mu^{s,j}O(n^{-1}) + \frac{1}{2}\mu^r_{st}s^s\mu^{t,j}O(n^{-1}) + \frac{1}{2}\mu^r_{st}\mu^{s,j}O(n^{-1})s^t \\
&\quad + O(n^{-1})\mu^{r,j}s^s + O_p(n^{-3}) \\
&= \mu^{r,j}O(n^{-1}) + O_p(n^{-5/2}) \\
&= n^{-1}\mu^{r,j}G(\theta_0; y) + O_p(n^{-5/2}),
\end{aligned}$$

where $G(\theta_0; y)$ is the $O(1)$ quantity in $\tilde{s}_n(\theta_0) = s_n(\theta_0) + n^{-1}G(\theta_0; y)$.

Appendix D

Results for the logistic mixed model with a random intercept

D.1 Derivation of the adjusted score function

The first derivatives of the log-likelihood in (5.3) are calculated using the Bayes' theorem, stated as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(A|B)$ is the conditional probability of A given B. If we let $P(\theta)$ be the prior distribution of θ, then because $f(y,\theta) = f(y|\theta)P(\theta) = P(\theta|y)f(y)$, we have

$$P(\theta|y) = \frac{f(y|\theta)P(\theta)}{f(y)} = \frac{f(y|\theta)P(\theta)}{\int f(y|\theta)P(\theta)d\theta},$$

where $P(\theta|y)$ is called the posterior density of θ. Under the typical framework of generalised linear mixed models, $P(\theta)$ is assumed to be a normal probability density function and in the current example $f(y|\theta)$ is a binomial probability mass function.

Let

$$I_{0i} = \int \frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\alpha_i^2}{2\sigma^2}} d\alpha_i$$

and

$$P(\alpha_i|y_i) = I_{0i}^{-1} \left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\alpha_i^2}{2\sigma^2}} \right).$$

Before we calculate the first and second derivatives of the log-likelihood we first calculate the derivatives of $P(\alpha_i|y_i)$. These are

$$\begin{aligned}
\frac{\partial P(\alpha_i|y_i)}{\partial \beta} &= I_{0i}^{-2}\left[I_{0i}\frac{\partial}{\partial \beta}\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right)-\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\frac{\partial I_{0i}}{\partial \beta}\right] \\
&= P(\alpha_i|y_i)\frac{\partial}{\partial \beta}\log\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right) \\
&- P(\alpha_i|y_i)\int P(\alpha_i|y_i)\frac{\partial}{\partial \beta}\log\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right)d\alpha_i \\
&= P(\alpha_i|y_i)\left[\left(y_i-m_i\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)-\int P(\alpha_i|y_i)\left(y_i-m_i\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)d\alpha_i\right] \\
&= P(\alpha_i|y_i)\left[y_i-m_i\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}-y_i+m_i\int\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i)d\alpha_i\right] \\
&= m_iP(\alpha_i|y_i)\left[\int\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i)d\alpha_i-\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right]
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial P(\alpha_i|y_i)}{\partial \sigma^2} &= I_{0i}^{-2}\left[I_{0i}\frac{\partial}{\partial \sigma^2}\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right)-\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\frac{\partial I_{0i}}{\partial \sigma^2}\right] \\
&= P(\alpha_i|y_i)\frac{\partial}{\partial \sigma^2}\log\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right) \\
&- P(\alpha_i|y_i)\int P(\alpha_i|y_i)\frac{\partial}{\partial \sigma^2}\log\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right)d\alpha_i \\
&= P(\alpha_i|y_i)\left[\left(\frac{\alpha_i^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right)-\int P(\alpha_i|y_i)\left(\frac{\alpha_i^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right)d\alpha_i\right] \\
&= P(\alpha_i|y_i)\left(\frac{\alpha_i^2}{2\sigma^4}-\frac{1}{2\sigma^4}\int\alpha_i^2P(\alpha_i|y_i)d\alpha_i\right).
\end{aligned}$$

The first derivatives of the log-likelihood are

$$
\begin{aligned}
\frac{\partial l}{\partial \beta} &= \frac{\partial}{\partial \beta}\sum_{i=1}^{q}\log\left(\int \frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}d\alpha_i\right)\\
&= \sum_{i=1}^{q}\frac{\partial}{\partial \beta}\log\left(\int \frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}d\alpha_i\right)\\
&= \sum_{i=1}^{q}\left(\int \frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}d\alpha_i\right)^{-1}\frac{\partial}{\partial \beta}\left(\int \frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}d\alpha_i\right)\\
&= \sum_{i=1}^{q}\left(\int \frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}d\alpha_i\right)^{-1}\int \frac{\partial}{\partial \beta}\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right)d\alpha_i\\
&= \sum_{i=1}^{q}\int P(\alpha_i|y_i)\,\frac{\partial}{\partial \beta}\log\left(\frac{e^{y_i(\alpha_i+\beta)}}{(1+e^{\alpha_i+\beta})^{m_i}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha_i^2}{2\sigma^2}}\right)d\alpha_i\\
&= \sum_{i=1}^{q}\int P(\alpha_i|y_i)\,\left(y_i-m_i\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)d\alpha_i\\
&= \sum_{i=1}^{q}\left\{y_i-m_i\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i)d\alpha_i\right\}\\
&= \sum_{i=1}^{q}[y_i-m_iE(\pi_i|y_i)];
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma^2} &= \sum_{i=1}^{q}\int P(\alpha_i|y_i)\,\frac{\partial}{\partial \sigma^2}\left(y_i(\alpha_i+\beta)-\frac{\alpha_i^2}{2\sigma^2}-m_i\log(1+e^{\alpha_i+\beta})-\frac{1}{2}\log(2\pi\sigma^2)\right)d\alpha_i\\
&= \sum_{i=1}^{q}\int P(\alpha_i|y_i)\,\left(\frac{\alpha_i^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right)d\alpha_i\\
&= \sum_{i=1}^{q}\left(-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\int \alpha_i^2P(\alpha_i|y_i)d\alpha_i\right)\\
&= \sum_{i=1}^{q}\left(-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}E(\alpha_i^2|y_i)\right).
\end{aligned}
$$

The second derivatives of the log-likelihood are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta^2} &= \frac{\partial}{\partial \beta}\sum_{i=1}^{q}\left\{y_i - m_i \int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} P(\alpha_i|y_i)\,d\alpha_i\right\} \\
&= -\sum_{i=1}^{q} m_i \int \frac{\partial}{\partial \beta}\left[P(\alpha_i|y_i)\left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)\right] d\alpha_i \\
&= -\sum_{i=1}^{q} m_i \int \left[\frac{\partial P(\alpha_i|y_i)}{\partial \beta}\left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right) + \frac{\partial}{\partial \beta}\left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right) P(\alpha_i|y_i)\right] d\alpha_i \\
&= -\sum_{i=1}^{q} m_i \left\{ \int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} m_i P(\alpha_i|y_i)\left(\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} P(\alpha_i|y_i)\,d\alpha_i\right) d\alpha_i \right. \\
&- \left. \int \left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)^2 m_i P(\alpha_i|y_i)\,d\alpha_i + \int \frac{e^{\alpha_i+\beta}}{(1+e^{\alpha_i+\beta})^2} P(\alpha_i|y_i)\,d\alpha_i \right\} \\
&= -\sum_{i=1}^{q}\left[m_i^2 E^2(\pi_i|y_i) - m_i(1+m_i)E(\pi_i^2|y_i) + m_i E(\pi_i|y_i)\right];
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \sigma^4} &= \frac{\partial}{\partial \sigma^2}\sum_{i=1}^{q}\left(-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\int \alpha_i^2 P(\alpha_i|y_i)\,d\alpha_i\right) \\
&= \sum_{i=1}^{q}\left[\frac{1}{2\sigma^4} + \frac{\partial}{\partial \sigma^2}\left(\frac{1}{2\sigma^4}\right)\int \alpha_i^2 P(\alpha_i|y_i)\,d\alpha_i + \frac{1}{2\sigma^4}\frac{\partial}{\partial \sigma^2}\left(\int \alpha_i^2 P(\alpha_i|y_i)\,d\alpha_i\right)\right] \\
&= \sum_{i=1}^{q}\left[\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\int \alpha_i^2 P(\alpha_i|y_i)\,d\alpha_i + \frac{1}{2\sigma^4}\int \alpha_i^2 \frac{\partial P(\alpha_i|y_i)}{\partial \sigma^2}\,d\alpha_i\right] \\
&= -\sum_{i=1}^{q}\left[-\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}E(\alpha_i^2|y_i) - \frac{1}{4\sigma^8}E(\alpha_i^4|y_i) + \frac{1}{4\sigma^8}E^2(\alpha_i^2|y_i)\right];
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \sigma^2 \partial \beta} &= \frac{\partial}{\partial \beta}\left(-\frac{q}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{q} E(\alpha_i^2|y_i)\right) \\
&= \frac{1}{2\sigma^4}\sum_{i=1}^{q}\int \alpha_i^2 \frac{\partial P(\alpha_i|y_i)}{\partial \beta}\,d\alpha_i \\
&= \frac{1}{2\sigma^4}\sum_{i=1}^{q}\int \alpha_i^2 m_i P(\alpha_i|y_i)\left[\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} P(\alpha_i|y_i)\,d\alpha_i - \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right] d\alpha_i \\
&= -\sum_{i=1}^{q}\left[\frac{m_i}{2\sigma^4}E(\alpha_i^2\pi_i|y_i) - \frac{m_i}{2\sigma^4}E(\pi_i|y_i)E(\alpha_i^2|y_i)\right].
\end{aligned}$$

We now have all the quantities required to derive the score function (first derivative of the log-likelihood), the observed information matrix $j(\theta)$ (minus the second derivative of the log-likelihood) and the expected information matrix $i(\theta) = E_\theta\{j(\theta)\}$. Let

$$s(\theta) = \begin{pmatrix} s_1(\theta) \\ s_2(\theta) \end{pmatrix}; \; j(\theta) = \begin{pmatrix} j_{11}(\theta) & j_{12}(\theta) \\ j_{21}(\theta) & j_{22}(\theta) \end{pmatrix}; \; i(\theta) = \begin{pmatrix} i_{11}(\theta) & i_{12}(\theta) \\ i_{21}(\theta) & i_{22}(\theta) \end{pmatrix}.$$

Then

$$\begin{aligned}
s_1(\theta) &= \sum_{i=1}^{q}(y_i - m_i E(\pi_i|y_i)); \\
s_2(\theta) &= \sum_{i=1}^{q}\left(-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}E(\alpha_i^2|y_i)\right); \\
j_{11}(\theta) &= \sum_{i=1}^{q}\left(m_i^2 E^2(\pi_i|y_i) - m_i(1+m_i)E(\pi_i^2|y_i) + m_i E(\pi_i|y_i)\right); \\
j_{12}(\theta) &= \sum_{i=1}^{q}\left(\frac{m_i}{2\sigma^4}E(\alpha_i^2\pi_i|y_i) - \frac{m_i}{2\sigma^4}E(\pi_i|Y_i)E(\alpha_i^2|y_i)\right); \\
j_{22}(\theta) &= \sum_{i=1}^{q}\left(-\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}E(\alpha_i^2|y_i) - \frac{1}{4\sigma^8}E(\alpha_i^4|y_i) + \frac{1}{4\sigma^8}E^2(\alpha_i^2|y_i)\right); \\
i_{11}(\theta) &= \sum_{i=1}^{q}\left\{m_i^2 E_{Y_i}[E^2(\pi_i|y_i)] - m_i(1+m_i)E_{Y_i}[E(\pi_i^2|y_i)] + m_i E_{Y_i}[E(\pi_i|y_i)]\right\}; \\
i_{12}(\theta) &= \sum_{i=1}^{q}\left\{\frac{m_i}{2\sigma^4}E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - \frac{m_i}{2\sigma^4}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]\right\}; \\
i_{22}(\theta) &= \sum_{i=1}^{q}\left\{-\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}E_{Y_i}[E(\alpha_i^2|y_i)] - \frac{1}{4\sigma^8}E_{Y_i}[E(\alpha_i^4|y_i)] + \frac{1}{4\sigma^8}E_{Y_i}[E^2(\alpha_i^2|y_i)]\right\} \\
&= \sum_{i=1}^{q}\left\{-\frac{1}{4\sigma^4} + \frac{1}{4\sigma^8}E_{Y_i}[E^2(\alpha_i^2|y_i)]\right\}.
\end{aligned}$$

The above expressions for the elements of the expected information matrix are identical to the expressions for the expected information matrix derived in Wand (2007) for the binomial-response generalised linear mixed model in (5.2).

The adjusted score functions proposed in Firth (1993) are

$$s_1^*(\theta) = s_1(\theta) + \frac{1}{2|i(\theta)|}\left(i_{22}E_{Y_i}[s_1^3 - j_{11}s_1] - 2i_{12}E_{Y_i}[s_1^2 s_2 - j_{12}s_1] + i_{11}E_{Y_i}[s_1 s_2^2 - j_{22}s_1]\right) \quad \text{(D.1)}$$

$$s_2^*(\theta) = s_2(\theta) + \frac{1}{2|i(\theta)|}\left(i_{22}E_{Y_i}[s_1^2 s_2 - j_{11}s_2] - 2i_{12}E_{Y_i}[s_1 s_2^2 - j_{12}s_2] + i_{11}E_{Y_i}[s_2^3 - j_{22}s_2]\right) \quad \text{(D.2)}$$

where

$$\begin{aligned}
E_{Y_i}[s_1^3 - j_{11}s_1] &= \sum_{i=1}^{q}\Bigg\{E_{Y_i}[y_i] - 3m_iE_{Y_i}[y_i^2E(\pi_i|y_i)] + 2m_i^2E_{Y_i}[y_iE^2(\pi_i|y_i)] \\
&- m_iE_{Y_i}[y_iE(\pi_i|y_i)] + m_i(1+m_i)E_{Y_i}[y_iE(\pi_i^2|y_i)] \\
&- m_i^2(1+m_i)E_{Y_i}[E(\pi_i|y_i)E(\pi_i^2|y_i)] + m_i^2E_{Y_i}[E^2(\pi_i|y_i)]\Bigg\}; \\
E_{Y_i}[s_1^2s_2 - j_{12}s_1] &= \sum_{i=1}^{q}\Bigg\{-\frac{1}{2\sigma^2}E_{Y_i}[y_i^2] + \frac{m_i}{\sigma^2}E_{Y_i}[y_iE(\pi_i|y_i)] - \frac{m_i^2}{2\sigma^2}E_{Y_i}[E^2(\pi_i|y_i)] \\
&+ \frac{1}{2\sigma^4}E_{Y_i}[y_i^2E(\alpha_i^2|y_i)] - \frac{m_i}{2\sigma^4}E_{Y_i}[y_iE(\alpha_i^2\pi_i|y_i)] \\
&+ \frac{m_i^2}{2\sigma^4}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2\pi_i|y_i)] - \frac{m_i}{2\sigma^4}E_{Y_i}[y_iE(\pi_i|y_i)E(\alpha_i^2|y_i)]\Bigg\}; \\
E_{Y_i}[S_1S_2^2 - I_{22}S_1] &= \sum_{i=1}^{q}\Bigg\{-\frac{3}{2\sigma^6}E_{Y_i}[y_iE(\alpha_i^2|y_i)] + \frac{3m_i}{2\sigma^6}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] \\
&+ \frac{1}{4\sigma^8}E_{Y_i}[y_iE(\alpha_i^4|y_i)] - \frac{m_i}{4\sigma^8}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^4|y_i)]\Bigg\}; \\
E_{Y_i}[s_1^2s_2 - j_{11}s_2] &= \sum_{i=1}^{q}\Bigg\{-\frac{1}{2\sigma^2}E_{Y_i}[y_i^2] + \frac{m_i}{\sigma^2}E_{Y_i}[y_iE(\pi_i|y_i)] + \frac{1}{2\sigma^4}E_{Y_i}[y_i^2E(\alpha_i^2|y_i)] \\
&- \frac{m_i(1+m_i)}{2\sigma^2}E_{Y_i}[E(\pi_i^2|y_i)] + \frac{m_i}{2\sigma^2}E_{Y_i}[E(\pi_i|y_i)] - \frac{m_i}{\sigma^4}E_{Y_i}[y_iE(\pi_i|y_i)E(\alpha_i^2|y_i)] \\
&+ \frac{m_i(1+m_i)}{2\sigma^4}E_{Y_i}[E(\pi_i^2|y_i)E(\alpha_i^2|y_i)] - \frac{m_i}{2\sigma^4}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]\Bigg\}; \\
E_{Y_i}[s_1s_2^2 - j_{12}s_2] &= \sum_{i=1}^{q}\Bigg\{-\frac{1}{2\sigma^6}E_{Y_i}[y_iE(\alpha_i^2|y_i)] + \frac{1}{4\sigma^8}E_{Y_i}[y_iE^2(\alpha_i^2|y_i)] + \frac{m_i}{4\sigma^6}E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] \\
&+ \frac{m_i}{4\sigma^6}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] - \frac{m_i}{4\sigma^8}E_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^2\pi_i|y_i)]\Bigg\}; \\
E_{Y_i}[s_2^3 - j_{22}s_2] &= \sum_{i=1}^{q}\Bigg\{\frac{3}{8\sigma^6} - \frac{3}{4\sigma^{10}}E_{Y_i}[E^2(\alpha_i^2|y_i)] + \frac{1}{8\sigma^{12}}E_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^4|y_i)]\Bigg\}.
\end{aligned}$$

Below we give a list of all the expectations that appear in $s_1^*(\theta)$ and $s_2^*(\theta)$:

1. $E_{Y_i}[E(\pi_i|y_i)] = E_{\alpha_i}(\pi_i)$
2. $E_{Y_i}[E(\pi_i^2|y_i)] = E_{\alpha_i}(\pi_i^2)$
3. $E_{Y_i}[E^2(\pi_i|y_i)]$
4. $E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] = E_{\alpha_i}(\alpha_i^2\pi_i)$
5. $E_{Y_i}[E^2(\alpha_i^2|y_i)]$
6. $E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]$
7. $E_{Y_i}[y_i^3] = m_iE_{\alpha_i}(\pi_i) + 3m_i(m_i-1)E_{\alpha_i}(\pi_i^2) + m_i(m_i-1)(m_i-2)E_{\alpha_i}(\pi_i^3)$
8. $E_{Y_i}[y_i^2E(\pi_i|y_i)] = m_iE_{\alpha_i}(\pi_i^2) + m_i(m_i-1)E_{\alpha_i}(\pi_i^3)$
9. $E_{Y_i}[y_iE^2(\pi_i|y_i)]$
10. $E_{Y_i}[y_iE(\pi_i|y_i)] = m_iE_{\alpha_i}(\pi_i^2)$
11. $E_{Y_i}[y_iE(\pi_i^2|y_i)] = m_iE_{\alpha_i}(\pi_i^3)$
12. $E_{Y_i}[E(\pi_i|y_i)E(\pi_i^2|y_i)]$
13. $E_{Y_i}[y_i^2] = m_iE_{\alpha_i}(\pi_i) + m_i(m_i-1)E_{\alpha_i}(\pi_i^2)$
14. $E_{Y_i}[y_i^2E(\alpha_i^2|y_i)] = m_iE_{\alpha_i}(\alpha_i^2\pi_i) + m_i(m_i-1)E_{\alpha_i}(\alpha_i^2\pi_i^2)$
15. $E_{Y_i}[y_iE(\alpha_i^2\pi_i|y_i)] = m_iE_{\alpha_i}(\alpha_i^2\pi_i^2)$
16. $E_{Y_i}[y_iE(\pi_i|y_i)E(\alpha_i^2|y_i)]$
17. $E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2\pi_i|y_i)]$
18. $E_{Y_i}[y_iE(\alpha_i^2|y_i)] = m_iE_{\alpha_i}(\alpha_i^2\pi_i)$
19. $E_{Y_i}[y_iE(\alpha_i^4|y_i)] = m_iE_{\alpha_i}(\alpha_i^4\pi_i)$
20. $E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^4|y_i)]$
21. $E_{Y_i}[E(\pi_i^2|y_i)E(\alpha_i^2|y_i)]$
22. $E_{Y_i}[y_iE^2(\alpha_i^2|y_i)]$
23. $E_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^2\pi_i|y_i)]$
24. $E_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^4|y_i)]$

Notice that some of the above expectations are expressed in a simpler form, and this has been achieved by using the Bartlett identities (Bartlett, 1953) and the identities on conditional expectations.

For example, we show that $E_{Y_i}[y_i E(\pi_i|y_i)] = m_i E_{\alpha_i}(\pi_i^2)$, using the identity of conditional expectations $f(y_i)E(\pi_i|y_i) = E(f(y_i)\pi_i|y_i)$, which gives $E_{Y_i}[y_i E(\pi_i|y_i)] = E_{Y_i}[E(y_i\pi_i|y_i)] = E_{\alpha_i}[E_{Y_i|\alpha_i}(y_i\pi_i|y_i)] = E_{\alpha_i}[\pi_i E_{Y_i|\alpha_i}(y_i|y_i)] = E_{\alpha_i}[\pi_i(m_i\pi_i)] = m_i E_{\alpha_i}(\pi_i^2)$. Similarly, $E_{Y_i}[y_i^2 E(\pi_i|y_i)] = E_{\alpha_i}[\pi_i E_{Y_i|\alpha_i}(y_i^2|y_i)] = E_{\alpha_i}[\pi_i(m_i\pi_i(1-\pi_i)+(m_i\pi_i)^2)]$, and thus $E_{Y_i}[y_i^2 E(\pi_i|y_i)] = m_i E_{\alpha_i}(\pi_i^2) + m_i(m_i-1)E_{\alpha_i}(\pi_i^3)$.

Using the above simplifications we have

$$\begin{aligned}
E_{Y_i}[s_1^3 - j_{11}s_1] &= \sum_{i=1}^{q}\Big\{ m_i E_{Y_i}[E(\pi_i|y_i)] - m_i(m_i+3)E_{Y_i}[E(\pi_i^2|y_i)] \\
&\quad - m_i(m_i+1)(m_i-2)E_{Y_i}[E(\pi_i^3|y_i)] + m_i^2 E_{Y_i}[E^2(\pi_i|y_i)] \\
&\quad + 2m^2 E_{Y_i}[y_i E^2(\pi_i|y_i)] - m_i^2(m_i+1)E_{Y_i}[E(\pi_i|y_i)E(\pi_i^2|y_i)] \Big\}; \\
E_{Y_i}[s_1^2 s_2 - j_{12}s_1] &= \sum_{i=1}^{q}\Big\{ -\frac{m_i}{2\sigma^2}E_{Y_i}[E(\pi_i|y_i)] + \frac{m_i(m_i+1)}{2\sigma^2}E_{Y_i}[E(\pi_i^2|y_i)] \\
&\quad - \frac{m_i^2}{2\sigma^2}E_{Y_i}[E^2(\pi_i|y_i)] + \frac{m_i}{2\sigma^4}E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - \frac{m_i}{2\sigma^4}E_{Y_i}[E(\alpha_i^2\pi_i^2|y_i)] \\
&\quad + \frac{m_i^2}{2\sigma^4}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2\pi_i|y_i)] - \frac{m_i}{2\sigma^4}E_{Y_i}[y_i E(\pi_i|y_i)E(\alpha_i^2|y_i)] \Big\}; \\
E_{Y_i}[s_1 s_2^2 - j_{22}s_1] &= \sum_{i=1}^{q}\Big\{ -\frac{3m_i}{2\sigma^6}E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] + \frac{3m_i}{2\sigma^6}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] \\
&\quad + \frac{m_i}{4\sigma^8}E_{Y_i}[E(\alpha_i^4\pi_i|y_i)] - \frac{m_i}{4\sigma^8}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^4|y_i)] \Big\}; \\
E_{Y_i}[s_1^2 s_2 - j_{11}s_2] &= \sum_{i=1}^{q}\Big\{ \frac{m_i}{2\sigma^4}E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] - \frac{m_i}{2\sigma^4}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] \\
&\quad - \frac{m_i}{\sigma^4}E_{Y_i}[y_i E(\pi_i|y_i)E(\alpha_i^2|y_i)] + \frac{m_i(m_i+1)}{2\sigma^4}E_{Y_i}[E(\pi_i^2|y_i)E(\alpha_i^2|y_i)] \\
&\quad + \frac{m_i(m_i-1)}{2\sigma^4}E_{Y_i}[E(\alpha_i^2\pi_i^2|y_i)] \Big\}; \\
E_{Y_i}[s_1 s_2^2 - j_{12}s_2] &= \sum_{i=1}^{q}\Big\{ -\frac{m_i}{4\sigma^6}E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] + \frac{m_i}{4\sigma^6}E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)] \\
&\quad + \frac{1}{4\sigma^8}E_{Y_i}[y_i E^2(\alpha_i^2|y_i)] - \frac{m_i}{4\sigma^8}E_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^2\pi_i|y_i)] \Big\}; \\
E_{Y_i}[s_2^3 - j_{22}s_2] &= \sum_{i=1}^{q}\Big\{ \frac{3}{8\sigma^6} - \frac{3}{4\sigma^{10}}E_{Y_i}[E^2(\alpha_i^2|y_i)] + \frac{1}{8\sigma^{12}}E_{Y_i}[E(\alpha_i^2|y_i)E(\alpha_i^4|y_i)] \Big\}.
\end{aligned}$$

We substitute these quantities into (D.1) and (D.2) to get the final form of the adjusted score functions.

D.2 Proof of Results 1-5

Proof of Result 1: Let $L_1 = \lim\limits_{\sigma^2\to 0} E_{Y_i}[E(\pi_i|y_i)E(\alpha_i^2|y_i)]$ and $L_2 = \lim\limits_{\sigma^2\to 0} E_{Y_i}[E(\alpha_i^2\pi_i|y_i)]$.

$$\begin{aligned}
L_1 &= \lim_{\sigma^2\to 0}\int \sum_{k=1}^{m} E(\pi_i|y_i=k)E(\alpha_i^2|y_i=k)P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i \\
&= \lim_{\sigma^2\to 0}\int \sum_{k=1}^{m}\left(\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i=k)d\alpha_i\right)\left(\int \alpha_i^2 P(\alpha_i|y_i=k)d\alpha_i\right)P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i \\
&= \sum_{k=1}^{m}\lim_{\sigma^2\to 0}\int \left(\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i=k)d\alpha_i\right)\left(\int \alpha_i^2 P(\alpha_i|y_i=k)d\alpha_i\right)P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i \\
&= \sum_{k=1}^{m}\lim_{\sigma^2\to 0}\left(\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i=k)d\alpha_i\int \alpha_i^2 P(\alpha_i|y_i=k)d\alpha_i\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right) \\
&= \sum_{k=1}^{m}\left[\lim_{\sigma^2\to 0}\left(\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i=k)d\alpha_i\right)\times\lim_{\sigma^2\to 0}\left(\int \alpha_i^2 P(\alpha_i|y_i=k)d\alpha_i\right)\right. \\
&\qquad \left.\times\lim_{\sigma^2\to 0}\left(\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)\right] \\
&= \sum_{k=1}^{m}\left[\lim_{\sigma^2\to 0}\left(\int \frac{\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(y_i=k|\alpha_i)f(\alpha_i)}{\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}d\alpha_i\right)\times\lim_{\sigma^2\to 0}\left(\int \frac{\alpha_i^2 P(y_i=k|\alpha_i)f(\alpha_i)}{\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}d\alpha_i\right)\right. \\
&\qquad \left.\times\lim_{\sigma^2\to 0}\left(\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)\right] \\
&= \sum_{k=1}^{m}\left[\frac{\lim\limits_{\sigma^2\to 0}\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}{\lim\limits_{\sigma^2\to 0}\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}\times\frac{\lim\limits_{\sigma^2\to 0}\int \alpha_i^2 P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}{\lim\limits_{\sigma^2\to 0}\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}\right. \\
&\qquad \left.\times\lim_{\sigma^2\to 0}\left(\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)\right] \\
&= \sum_{k=1}^{m}\left[\frac{\lim\limits_{\sigma^2\to 0}\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}{\lim\limits_{\sigma^2\to 0}\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}\times\lim_{\sigma^2\to 0}\int \alpha_i^2 P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right]. \qquad \text{(D.3)}
\end{aligned}$$

Here it is tempting to move the limit inside the integrals, but this is not always valid. Lebesgue integration theory has a powerful criterion called Lebesgue's dominated convergence theorem (Rudin, 1976, p. 318). This theorem tells us that if the limit of the integrand exists for almost all x, and there is a function $H(x) \geq 0$, $\int_{-\infty}^{\infty} H(x)dx < \infty$, such that $|f(x)| < H(x)$ then the interchange of limit and integration is valid. In other words,

$$\lim_{t\to t_0}\int_X f(x,t)dx = \int_X f(x,t_0)dx$$

is justified when $|f(x,t)|$ is bounded (Benedetto & Czaja, 2010, Theorem 3.6.1). In (D.3) we can interchange limits and integration only in the first part of L_1,

$$\frac{\lim_{\sigma^2\to 0}\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i}{\lim_{\sigma^2\to 0}\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i},$$

where both integrands are bounded in $[0,(2\pi\sigma^2)^{-1/2}]$. We then have

$$L_1 = \sum_{k=1}^{m}\left[\frac{\int \frac{e^{\beta}}{1+e^{\beta}} P(y_i=k|\alpha_i)\delta(\alpha_i)d\alpha_i}{\int P(y_i=k|\alpha_i)\delta(\alpha_i)d\alpha_i}\lim_{\sigma^2\to 0}\int \alpha_i^2 P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right].$$

The quantities $e^{\beta}/(1+e^{\beta})$ and $P(y_i=k|\alpha_i)$ are independent of α_i, because when σ^2 approaches zero, α_i also approaches zero. They can then be taken out of the integration as constants, and the remaining $\int \delta(\alpha_i)d\alpha_i$ is equal to unity by construction. The limit L_1 is therefore further simplified to

$$\begin{aligned}
L_1 &= \sum_{k=1}^{m}\left[\frac{e^{\beta}}{1+e^{\beta}}\lim_{\sigma^2\to 0}\int \alpha_i^2 P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right] \\
&= \lim_{\sigma^2\to 0}\int \sum_{k=1}^{m}\alpha_i^2\frac{e^{\beta}}{1+e^{\beta}}P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i \\
&= \lim_{\sigma^2\to 0}\int \alpha_i^2\frac{e^{\beta}}{1+e^{\beta}}\left(\sum_{k=1}^{m}P(y_i=k|\alpha_i)\right)f(\alpha_i)d\alpha_i \\
&= \lim_{\sigma^2\to 0}\int \alpha_i^2\frac{e^{\beta}}{1+e^{\beta}}f(\alpha_i)d\alpha_i \\
&= \lim_{\sigma^2\to 0}E_{\alpha_i}(\alpha_i^2\pi_i) \\
&= \lim_{\sigma^2\to 0}E_{Y_i}[E(\alpha_i^2\pi_i|y_i)] \\
&= L_2.
\end{aligned}$$

□

Similarly, we can prove Result 2.

Proof of Result 3: Let $L_3 = \lim_{\sigma^2\to 0} E_{Y_i}[E^2(\pi_i|y_i)]$ and $L_4 = \lim_{\sigma^2\to 0} E_{Y_i}[E(\pi_i^2|y_i)]$. Using the same properties of limits and Lebesgue dominated convergence theorem we have:

$$\begin{aligned}
L_3 &= \lim_{\sigma^2\to 0}\int\sum_{k=1}^{m} E(\pi_i|y_i=k)^2 P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i \\
&= \lim_{\sigma^2\to 0}\int\sum_{k=1}^{m}\left(\int\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i=k)d\alpha_i\right)^2 P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i \\
&= \sum_{k=1}^{m}\lim_{\sigma^2\to 0}\left(\int\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i=k)d\alpha_i\right)^2\left(\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right) \\
&= \sum_{k=1}^{m}\left[\lim_{\sigma^2\to 0}\left(\int\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(\alpha_i|y_i=k)d\alpha_i\right)^2\times\lim_{\sigma^2\to 0}\left(\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)\right] \\
&= \sum_{k=1}^{m}\left[\frac{\lim_{\sigma^2\to 0}\left(\int\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)^2}{\lim_{\sigma^2\to 0}\left(\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)^2}\times\lim_{\sigma^2\to 0}\left(\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)\right] \\
&= \sum_{k=1}^{m}\frac{\left(\lim_{\sigma^2\to 0}\int\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i\right)^2}{\lim_{\sigma^2\to 0}\int P(y_i=k|\alpha_i)f(\alpha_i)d\alpha_i} \\
&= \sum_{k=1}^{m}\frac{\left(\int\frac{e^{\beta}}{1+e^{\beta}}P(y_i=k|\alpha_i)\delta(\alpha_i)d\alpha_i\right)^2}{\int P(y_i=k|\alpha_i)\delta(\alpha_i)d\alpha_i} \\
&= \sum_{k=1}^{m}\left(\frac{e^{\beta}}{1+e^{\beta}}\right)^2 P(y_i=k|\alpha_i) = \left(\frac{e^{\beta}}{1+e^{\beta}}\right)^2;
\end{aligned}$$

$$\begin{aligned}
L_4 &= \lim_{\sigma^2\to 0} E_{Y_i}[E(\pi_i^2|y_i)] = \lim_{\sigma^2\to 0} E_{\alpha_i}(\pi_i^2) \\
&= \lim_{\sigma^2\to 0}\int\left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)^2 f(\alpha_i)d\alpha_i \\
&= \int\left(\frac{e^{\beta}}{1+e^{\beta}}\right)^2\delta(\alpha_i)d\alpha_i = \left(\frac{e^{\beta}}{1+e^{\beta}}\right)^2.
\end{aligned}$$

Then $L_3 = L_4$. □

Proof of Result 4: Let $L_5 = \lim\limits_{\sigma^2 \to 0} E_{Y_i}[y_i E^2(\pi_i|y_i)]$ and $L_6 = \lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i^3|y_i)]$.

$$\begin{aligned}
L_5 &= \lim_{\sigma^2 \to 0} \int \sum_{k=1}^{m} Y_k E(\pi_i|y_i = k)^2 P(y_i = k|\alpha_i) f(\alpha_i) d\alpha_i \\
&= \lim_{\sigma^2 \to 0} \int \sum_{k=1}^{m} Y_k \left(\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} P(\alpha_i|y_i = k) d\alpha_i \right)^2 P(y_i = k|\alpha_i) f(\alpha_i) d\alpha_i \\
&= \sum_{k=1}^{m} Y_k \lim_{\sigma^2 \to 0} \left(\int \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} P(\alpha_i|y_i = k) d\alpha_i \right)^2 \left(\int P(y_i = k|\alpha_i) f(\alpha_i) d\alpha_i \right) \\
&= \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^2 \sum_{k=1}^{m} Y_k P(y_i = k|\alpha_i) \\
&= \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^2 E(y_i|\alpha_i) = m_i \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^3;
\end{aligned}$$

$$\begin{aligned}
L_6 &= \lim_{\sigma^2 \to 0} E_{\alpha_i}(\pi_i^3) \\
&= \lim_{\sigma^2 \to 0} \int \left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \right)^3 f(\alpha_i) d\alpha_i \\
&= \int \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^3 \delta(\alpha_i) d\alpha_i = \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^3.
\end{aligned}$$

Then $L_5 = m_i L_6$. □

Proof of Result 5: Let $L_7 = \lim\limits_{\sigma^2 \to 0} E_{Y_i}[E(\pi_i|y_i)E(\pi_i^2|y_i)]$.

$$\begin{aligned}
L_7 &= \lim_{\sigma^2 \to 0} \int \sum_{k=1}^{m} E(\pi_i|y_i = k) E(\pi_i^2|y_i = k) P(y_i = k|\alpha_i) f(\alpha_i) d\alpha_i \\
&= \sum_{k=1}^{m} \left[\frac{\lim_{\sigma^2 \to 0} \int \pi_i P(y_i = k|\alpha_i) f(\alpha_i) d\alpha_i}{\lim_{\sigma^2 \to 0} \int P(y_i = k|\alpha_i) f(\alpha_i) d\alpha_i} \times \lim_{\sigma^2 \to 0} \int \pi_i^2 P(y_i = k|\alpha_i) f(\alpha_i) d\alpha_i \right] \\
&= \sum_{k=1}^{m} \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^3 P(y_i = k|\alpha_i) = \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^3.
\end{aligned}$$

Then $L_6 = L_7$. □

Appendix E

Key results used to support proofs of theorems in Chapter 3 and Chapter 4

This appendix contains some key results used to support proofs of theorems and derivation of results in Chapter 3 and Chapter 4. In order to prove the consistency and asymptotic normality of the mean bias-reduced IBLA estimators we use the following main results:

Lemma E.1. *Slutsky lemma (Van der Vaart, 2000, Lemma 2.8)*
Let X_n, X and Y_n be random vectors or variables. If $X_n \rightarrow X$ and $Y_n \rightarrow c$ for a constant c, then (i) $X_n + Y_n \rightarrow X + c$, (ii) $Y_n X_n \rightarrow cX$, (iii) $Y_n^{-1} X_n \rightarrow c^{-1} X$ provided $c \neq 0$.

Theorem E.1. *Weak law of large numbers (Davison, 2003, p. 28).*
If $Y_1, Y_2, \ldots$ is a sequence of independent identically distributed random variables each with finite mean μ, and if $\bar{Y} = n^{-1}(Y_1 + \ldots + Y_n)$ is the average of $Y_1, \ldots, Y_n$, then $\bar{Y} \xrightarrow{p} \mu$.

Theorem E.2. *Van der Vaart (2000, Theorem 5.9) Let Ψ_n be random vector-valued functions and let Ψ be a fixed vector-valued function of θ such that for every $\varepsilon > 0$*

$$\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{p} 0,$$

$$\inf_{\theta : d(\theta, \theta_0) \geq \varepsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|.$$

Then any sequence of estimators $\hat{\theta}_n$ such that $\Psi_n(\hat{\theta}_n) = o_p(1)$ converges in probability to θ_0.

Theorem E.3. *Central limit theorem (Van der Vaart, 2000, Proposition 2.17)*

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $EY_i = 0$ and $EY_i^2 = 1$. Then the sequence $\sqrt{n}\bar{Y}_n$ converges in distribution to the standard normal distribution.

Theorem E.4. *Continuous mapping theorem (Van der Vaart, 2000, Theorem 2.3)*

Let $g : \Re^k \mapsto \Re^m$ be continuous at every point of a set C such that $P(X \in C) = 2$.

(i) If $X_n \to X$, then $g(X_n) \to g(X)$;

(ii) If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$;

(iii) If $X_n \xrightarrow{as} X$, then $g(X_n) \xrightarrow{as} g(X)$.

Theorem E.5. *(Trench, 2003, Theorem 7.1.13)*

If f is continuous on a rectangle R in $\Re^n$, then f is integrable on R.

Theorem E.6. *Lebesgue's theorem (Lavrent'ev & Savel'ev, 2006, p. 165)*

A subset of $\Re^m$ is compact if and only if it is bounded and closed. If $X \subset \Re^m$ is bounded and closed, then X is a closed subset of a rectangle that is a product of intervals.

The equivalence of (a) and (b) in Theorem E.7 is known as the Heine-Borel theorem.

Theorem E.7. *(Rudin, 1976, Theorem 2.41)*

If a set E in $\Re^k$ has one of the following two properties, then it has the other two:

(a) E is closed and bounded.

(b) E is compact.

(c) Every infinite subset of E has a limit point in E.

Theorem E.8. *(Rudin, 1976, Theorem 4.9)*

Let f and g be complex continuous functions on a metric space X. Then $f+g, fg$ and f/g are continuous on X.

Theorem E.9. *(Rudin, 1976, Theorem 4.16)*

Suppose f is a continuous real function on a compact metric space X, and $M = \sup_{p \in X} f(p)$, $m = \inf_{p \in X} f(p)$. Then there exist points $p, q \in X$ such that $f(p) = M$ and $f(q) = m$.

Theorem E.10. *Bounded convergence theorem (Feller, 2008, p. 111)*

Let u_n be integrable and $u_n \to u$ pointwise. If there exists an integrable U such that $|u_n| \leq U$ for all n, then u is integrable and $E(u_n) \to E(u)$.

The stochastic order symbols O_p and o_p (see, for example, Van der Vaart, 2000, Section 2.2) are used for describing the asymptotic order of random quantities and are defined as follows:

Definition E.1. Consider a sequence of random variables $\{X_n\}$ and a sequence of constants $\{a_n\}$. We write $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$.

Definition E.2. Consider a sequence of random variables $\{X_n\}$ and a sequence of constants $\{a_n\}$. We write $X_n = O_p(a_n)$ if for every $\varepsilon > 0$ there exists $K(\varepsilon) > 0$ and $n_0(\varepsilon)$ such that, for all $n > n_0(\varepsilon)$,

$$P\left(\left|\frac{X_n}{a_n}\right| \leq K(\varepsilon)\right) > 1 - \varepsilon .$$

The statement $X_n = O_p(1)$ is equivalent to saying that $\{X_n\}$ is bounded in probability.

We also make use of the Landau symbols $o(\cdot)$ and $O(\cdot)$.

Definition E.3. Consider two sequences of real constants $\{a_n\}$ and $\{b_n\}$. We write $b_n = o(a_n)$ if $\lim_{n\to\infty} |b_n/a_n| = 0$.

Definition E.4. Consider two sequences of real constants $\{a_n\}$ and $\{b_n\}$. We write $b_n = O(a_n)$ if there exists $\varepsilon > 0$ and positive integer $N(\varepsilon)$ such that if $n \geq N(\varepsilon)$ then $\limsup |b_n|/|a_n| < \infty$.