Bias in ML Estimation of Multilevel Models: Check the Algebra Before You Gamble in Monte Carlo

Martin Elff and Susumu Shikano*

Prepared for delivery at the 4th Annual General Conference of the European Political Science Association, Royal Society of Edinburgh and the Royal College of Physicians of Edinburgh, Edinburgh, UK, 19–21 June 2014

Abstract

In a recent article published in the American Journal of Political Science Stegmüller claims that ML estimation of multilevel models leads to bias and confidence interval undercoverage of coefficient estimates if the number of higher-level units is small and that this bias and undercoverage can be avoided by using methods of Bayesian inference instead.

In our paper we examine the sources of bias in ML estimation of multilevel models with a small number of higher-level units. It can be shown that, for given variance components, ML estimation of multilevel model coefficients is a particular case of Generalised Least Squares, hence must be unbiased. However, ML estimation of variance parameters tends to be biased if the number of higher-level units is small. We discuss a modification of ML – restricted maximum likelihood (REML) – which reduces this bias considerably. Further, we show that the coverage error of interval estimates of coefficients found by Stegmüller can be addressed, first, by using unbiased estimates of variance parameters and, second, by using an appropriately selected t-distribution for the construction of confidence intervals. In addition to discussing the relevant literature establishing these adjustment, we replicate Stegmüller's Monte Carlo study and extend it, by taking into account REML and improved interval estimate construction. We conclude that Stegmüller claim that frequentist estimators of multilevel models are flawed is misleading and that the claim that Bayesian estimators are consequentially superior is premature.

*University of Konstanz, Department of Politics and Public Administration, martin.elff@uni-konstanz.de

1 Introduction

Multilevel modelling has emerged as a widely used tool for the comparative analysis of political attitudes and behaviour, especially for the cross-national analysis of these phenomena. A problem that may cause some concern is that often the number of higher-level units of comparison, i.e. countries in the case of cross-national studies, is small due to limitations of available data. While one may hear about various rules of thumb regarding the minimum number of countries required for reliable results, it is not easy to find justifications for them. In a recent article in the *American Journal of Political Science* Stegmüller (2013) examined the consequences of a small number of higher-level units or countries for estimation and inference with multilevel models. In this article Stegmüller compares frequentist and Bayesian approaches at point and interval estimation of coefficients of these models and finds that "that maximum likelihood estimates and confidence intervals can be severely biased, especially in models including cross-level interactions. In contrast, the Bayesian approach proves to be far more robust and yields considerably more conservative tests" (Stegmüller 2013).

In the present paper we question the stark claim about the superiority of the Bayesian approach over frequentist estimation for several reasons. Firstly, while Stegmüller claims to have established – by a Monte Carlo study – that coefficient estimates may be biased, and even severely so, statistical theory implies that no such bias exists in common frequentist estimators of multilevel model coefficients. In fact even OLS estimates of coefficients are unbiased, yet not efficient. Secondly, Stegmüller only considers only maximum likelihood (ML) estimators and not a more appropriate variant for estimation and inference in multilevel models with small numbers of higher-level units, namely restricted maximum likelihood (REML) (Patterson and Thompson 1971). Second, when analysing the coverage performance of interval estimates, Stegmüller only considers as frequentist confidence intervals only those based on asymptotic normality, and does not consider improved confidence intervals, as suggested by Kenward and Roger (1997) that explicitely take into account the failure of asymptotic normality of estimates for coefficients of group-level covariates. Third, in his Monte Carlo studies, Stegmüller does not take into account the inevitable variance of simulated averages and thus is not able to distinguish between true simulated bias and the mere random departures created by Monte Carlo sampling error. We find that, if these limitations are addressed, frequentist estimators are not as flawed as apparent from Stegmüllers simulations and that Bayesian inference is not the panacea that he suggests.

The paper is organised as follows: In the next section we introduce the necessary notation of multilevel models. In particular, we show how these models can be written in matrix form that allows the formulate the relevant techniques of estimation and inference pertinent to multilevel analysis. In the ensuing section we use this notation to show why coefficient estimates of normal linear multilevel models are unbiased when estimated using maximum likelihood techniques, and even using OLS, notwithstanding the number of higher-level units. We further discuss the bias in variance parameters, that was not much discussed by Stegmüller (2013) and the potential that alternatives to ML have to address this bias. We further discuss in that section how these arguments can be extended to generalised linear mixed models, of which mixed probit, analysed by Stegmüller, is a special case. And finally we discuss in this section the potential of Bayesian inference to address such biases. This theoretical section is followed by a report on a Monte Carlo study of three frequentist estimators of parameters of multilevel models. Beside ML, we also consider OLS as an estimator of coefficient and REML as an estimator of coefficient and variance parameters. We also investigate the performance of variously constructed interval estimators. The Monte Carlo study of frequent estimators is followed in another section by a Monte Carlo Study that compares frequentist and Bayesian techniques and investigates the sensitivity of the Bayesian approach to the choice of prior distributions. The conclusion of the paper summarises its results and highlights its most important implications.

2 Linear and Generalised Linear Multilevel Models

Multilevel models are commonly used in the social sciences to simultaneously analyse the effects of both individual and context-specific factors and covariates on individual behaviour as well as the interaction of these two types of effects. In comparative analysis of political behaviour, these contexts typically are countries. Often such multilevel models are specified in a hierarchical form. While this hierarchical form is often more easy to understand from a substantial perspective and therefore serves as the framework of the discussion of the performance of estimators for the parameters of such model in Stegmüller (2013), the expanded form is better suited to explain estimation and inference for such models. This is why prefer to discuss multilevel model in this expanded form.

To understand the relation between the hierarchical form and the expanded form of multilevel models consider the case where there are individual observations, e.g. citizens, nested

in contexts, e.g. countries, thereby also considered as members in "groups", "clusters", or "upper-level units". Suppose one observes for each individual i, which is member in group j, a value y_{ij} of a variable of interest, e.g. support for the European Union, which has metric level of measurement. Suppose further that one is interested in the effects of an individual-level covariate with values x_{1ij}, e.g. the amount of information about the EU, which may vary across gross groups j. In addition, suppose that the group-level average of the response variable y_{ij} is influenced by a group-level covariate x_{2j} and that the effect of the individual-level covariate is also conceived as influenced by another group-level covariate x_{3j}. This situation can be expressed by the following set of equations:

$$y_{ij} = a_j + b_j x_{1ij} + \epsilon_{ij} \tag{1}$$

$$a_j = \beta_{00} + \beta_{01} x_{2j} + u_{1j} \tag{2}$$

$$b_j = \beta_{10} + \beta_{11} x_{3j} + u_{2j} \tag{3}$$

where β_{00}, β_{01}, β_{10}, and β_{11} are fixed, but unknown coefficients (which one wants to estimate), and ϵ_{ij}, u_{1j}, and u_{2j} are individual level and group-level disturbances or error terms (which e.g. may represent either pure randomness or the effects unmeasured covariates). Substituting equations (2) and (3) into equation (1) leads to the expanded form of the model:

$$y_{ij} = \beta_{00} + \beta_{10} x_{1ij} + \beta_{01} x_{2j} + \beta_{11} x_{3j} x_{1ij} + u_{1j} + x_{1ij} u_{2j} + \epsilon_{ij}. \tag{4}$$

In the language of multilevel modelling one usually would call β_{00}, β_{10}, β_{01}, and β_{11} "fixed-effects coefficients" of the covariates x_{1ij}, x_{2j}, and $x_{3j}x_{1ij}$, and in particular γ_{11} also as a "cross-level interaction" of x_{1ij} and x_{2j} and the product $x_{3j}x_{1ij}$ a cross-level interaction term. Further u_{1j} is typically referred to as a random intercept and u_{2j} as a random slope.

Equation (4) can also be written in a form involving vectors:

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u} + \epsilon_{ij} \tag{5}$$

where $\boldsymbol{\beta}$ is the vector with elements β_{00}, β_{10}, β_{01}, and β_{11}; $\boldsymbol{u}$ is the vector with elements $u_{11}, u_{21}, \ldots, u_{1m}, u_{2m}$ (with m as the number of groups) $\boldsymbol{x}_{ij}$ is the vector with elements 1, x_{1ij}, x_{2ij}, and $x_{3j}x_{1ij}$ (and $\boldsymbol{x}'_{ij}$ its transpose); and finally $\boldsymbol{z}_{ij}$ a vector with all elements equal to zero except for those elements equal to 1 and x_{1ij}, respectively, at the appropriate places such that $\boldsymbol{z}'_{ij}\boldsymbol{u} = u_{1j} + x_{1ij}u_{2j}$. If we collect the response observations y_{ij} into the vector $\boldsymbol{y}$ and the errors ϵ_{ij} into the vector $\boldsymbol{\epsilon}$, if we further construct the matrix $\boldsymbol{X}$ with the vectors $\boldsymbol{x}_{ij}$ as rows and

the matrix $\boldsymbol{Z}$ with the vectors $\boldsymbol{z}_{ij}$ then we arrive at a special case of the more general matrix form of multilevel models given by the following equation (6).

In general a *linear mixed model* (another term for the type of models just discussed) with response vector $\boldsymbol{y}$, random effects vector $\boldsymbol{u}$, and disturbance vector $\boldsymbol{\epsilon}$ can be written in the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \quad \text{with} \quad \operatorname{Var}(\boldsymbol{u}) = \boldsymbol{\Phi} \quad \text{and} \quad \operatorname{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}, \tag{6}$$

where $\boldsymbol{X}$ is a regressor matrix that contains the values of the independent variables, $\boldsymbol{\beta}$ is a vector of "fixed effects", and $\boldsymbol{Z}$ is a matrix appropriately constructed to reflect random intercepts random slopes, and the grouping structure of the random effects. Further, $\boldsymbol{\Phi}$ is a symmetric positive definite matrix, and $\boldsymbol{I}$ is a identity matrix of appropriate size. Usually, it is assumed that the elements of $\boldsymbol{\epsilon}$ are assumed to be normal i.i.d. with zero expectation and common variance σ^2 and $\boldsymbol{u}$ can be split into independent multivariate normal distributed sub-vectors with zero expectation, such that $\boldsymbol{\Phi}$ is block-diagonal.

In case of a two-level random-intercept model with n observations, m groups at the second level, and one independent variable with values $x_1, \ldots, x_n$, then $\boldsymbol{X}$ is a $n \times 2$ matrix, $\boldsymbol{Z}$ is a $n \times m$, $\boldsymbol{\Phi}$ is a $m \times m$ diagonal matrix and $\sigma^2\boldsymbol{I}$ is a $n \times n$ diagonal matrix with

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} 1 & & 0 \\ \vdots & & \vdots \\ 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ \vdots & & \vdots \\ 0 & & 1 \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} \theta & & \\ & \ddots & \\ & & \theta \end{bmatrix}, \quad \text{and} \quad \sigma^2\boldsymbol{I} = \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix}.$$

In case of a two-level model with random intercepts and random slopes of the independent

variable $\boldsymbol{Z}$ is a $n \times 2m$ matrix and $\boldsymbol{\Phi}$ is a $2m \times 2m$ matrix with

$$\boldsymbol{Z} = \begin{bmatrix} 1 & x_1 & & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{i_1} & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & x_{i_m+1} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 1 & x_n \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} \theta_{11} & \theta_{12} & & & \\ \theta_{12} & \theta_{22} & & & \\ & & \ddots & & \\ & & & \theta_{11} & \theta_{12} \\ & & & \theta_{12} & \theta_{22} \end{bmatrix}$$

Generalised linear mixed models extend linear mixed models much in the same way as generalised linear models (McCullagh and Nelder 1989) extend linear regression models. A generalised linear mixed effects model involves response vector $\boldsymbol{y}$ with expectation $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu}$ extends a generalised linear model by the inclusion of a random-effects vector $\boldsymbol{u}$ and thus takes the form

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}, \quad \text{or in matrix form} \quad g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tag{7}$$

again, with $\mathrm{Var}(\boldsymbol{u}) = \boldsymbol{\Phi}$. Here, η_i refers to the *linear component* of the model. In most applications the elements of the response vectors are assumed to come from a distribution that is a member of a exponential family which may involve further parameters, some of them related to the dispersion or variance of the response. The class of generalised linear mixed-effects models includes normal-linear mixed-effects models, where $g(\boldsymbol{\mu}) = \boldsymbol{\mu}$ and $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$. It also includes logistic mixed-effects models for binary responses, where $g(\mu_i) = \ln[\mu_i/(1-\mu_i)]$.

The matrix notation of multilevel models — as linear mixed models and generalised linear mixed models — allows to formulate the estimators for the parameters of these models in a compact form and to elucidate their theoretical properties. This is the topic of the following section.

3 Estimation and Inference

In his AJPS article Stegmüller (2013) claims to have found that the estimates of covariate effects (or fixed-effects coefficients, the elements of the vector $\boldsymbol{\beta}$ introduced in the previous section) obtained by maximum likelihood (ML) methods may be severely biased. By discussing elementary implications of the structure of maximum likelihood estimators of fixed-effects coefficients, we are able to show that point estimators of these are generally unbiased in linear mixed models. Only estimates of the variances of random intercepts and random slopes may be biased when estimated by maximum likelihood. We further discuss an already well-known modification of ML — restricted maximum likelihood (REML) (Patterson and Thompson 1971) — that at least leads to a bias reduction in the estimation of variance parameters, if it does not even eliminate this bias. We also discuss the construction of interval estimates and point out, referring to appropriate literature, that the usual assumption of asymptotic normality may indeed lead to biased interval estimates (which tend to be too short and lead to undercoverage) if the number of upper-level units of multi-level models is small, but that there are ways to take this into account in such a way that one may obtain interval estimates that are, if perhaps not completely unbiased, much less biased than normality-based ones. We further discuss whether these arguments can be extended to the case of generalised linear models.

The big challenge posed by linear and generalised linear models is that the random intercepts and random slopes (which are the elements of the random vector $\boldsymbol{u}$) are neither observed, nor estimable model parameters. Any technique of estimating the parameters of these models therefore cannot depend on the values of these random components. If in case of a linear model with normal disturbance and an observed random vector $\boldsymbol{u}$ the "complete-data" log-likelihood function would take the form

$$\ell_{\text{cpl}}(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}) = c - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}) - \frac{1}{2}\boldsymbol{u}'\boldsymbol{\Phi}^{-1}\boldsymbol{u} \quad (8)$$

so that ML estimation of the model parameters would be relatively straightforward. Yet since $\boldsymbol{u}$ is unobserved, the actual log-likelihood used for estimating the model parameters cannot depend on it.

The usual technique to eliminate unobserved data from a log-likehood function is to integrate them out. In case of a linear mixed model with normal distribution of disturbances and

random effects, this leads to:

$$\begin{aligned}
\ell(\boldsymbol{\theta};\boldsymbol{y}) &= c - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) \\
&\quad + \ln\int \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})'(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}) - \frac{1}{2}\boldsymbol{u}'\boldsymbol{\Phi}^{-1}\boldsymbol{u}\right]\mathrm{d}\,\boldsymbol{u} \\
&= c^\dagger - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) - \frac{1}{2}\ln\det\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Phi}^{-1}\right) \\
&\quad - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{Z}\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Phi}^{-1}\right)^{-1}\boldsymbol{Z}'(\boldsymbol{y}-\boldsymbol{X\beta}) \\
&= c^\dagger - \frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})
\end{aligned} \tag{9}$$

where $\boldsymbol{V} = \sigma^2\boldsymbol{I} + \boldsymbol{Z\Phi Z}'$ and c and $c^\dagger$ are normalising constants independent from the data and the parameters (see Jiang 2007; the integration rules relevant here can be found in Harville 1997).

If the variance parameters in σ^2 and $\boldsymbol{\Phi}$ are given, obtaining maximum likelihood likelihood estimates for the fixed-effects vector $\boldsymbol{\beta}$ is relatively straightforward. One only needs to solve the likelihood equation

$$\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) = 0 \tag{10}$$

which leads to the GLS estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{11}$$

One important implication of this is that for any choice of σ^2 and $\boldsymbol{\Phi}$ the ML estimator for $\boldsymbol{\beta}$ is unbiased: If $\boldsymbol{\beta}_0$ is the true value of the fixed-effects vector $\boldsymbol{\beta}$ and if $\boldsymbol{V}$ is given by predetermined values of σ^2 and $\boldsymbol{\Phi}$ then:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\,\mathrm{E}(\boldsymbol{y}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X\beta}_0 = \boldsymbol{\beta}_0. \tag{12}$$

It should be noted that even OLS estimators exhibit this unbiasedness, because

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{E}(\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X\beta}_0 = \boldsymbol{\beta}_0. \tag{13}$$

That notwithstanding, OLS estimators are less efficient than GLS estimators in the presence of group-level variance components.

The main problem here is that no such simple solution exists for the maximum likelihood

estimates of the variance parameters, that is, the functionally independent elements of σ^2 and $\boldsymbol{\Phi}$, which in the following are assumed to be collected in the vector $\boldsymbol{\theta}$. Since any ML estimator $\hat{\boldsymbol{\theta}}$ therefore is not a linear function of any sufficient statistics computed from $\boldsymbol{y}$, it cannot be unbiased (see e.g. the literature cited in Elff 2014 [forthcoming]). That ML estimates are often biased is however quite common and well-known. For example, let $\boldsymbol{x}$ be a vector of n normal i.i.d. random variables with mean μ and variance σ^2 then ML estimator for σ^2 is the uncorrected sample variance

$$\hat{\sigma}^2_{ML} = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

which has expectation and bias (if the true value is σ^2_0)

$$\mathrm{E}(\hat{\sigma}^2_{ML}) = \frac{n-1}{n}\sigma^2_0 \quad \text{and} \quad \mathrm{Bias}(\hat{\sigma}^2_{ML}) = -\frac{1}{n}\sigma^2_0.$$

Of course, the bias becomes negligible if n grows to infinity, but in small samples it may be substantial. In case of variance parameters of a multilevel model, the situation may be analogous, yet for the amount of bias of variance parameters other than σ^2 it is not the overall sample size that is relevant, but the number of groups the variance between of is expressed by $\boldsymbol{\Phi}$ as shown by the simulation studies reported in Stegmüller (2013).

The bias in the ML estimator of $\boldsymbol{\theta}$ and hence $\boldsymbol{\Phi}$ fortunately does not carry over to the ML estimator of $\boldsymbol{\beta}$: From equation (12) it follows that $\mathrm{E}(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\Phi}}) = \boldsymbol{\beta}_0$ (where $\boldsymbol{\beta}_0$ is the true value of the fixed-effects coefficient vector) so that the law of iterated expectations leads to

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}_{\hat{\boldsymbol{\Phi}}}[\mathrm{E}(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\Phi}})] = \mathrm{E}_{\hat{\boldsymbol{\Phi}}}(\boldsymbol{\beta}_0) = \boldsymbol{\beta}_0 \tag{14}$$

(a more rigorous proof is given by Harville 1976, see also Jiang 1999). On the other hand, the bias in the ML estimator of $\boldsymbol{\theta}$ is likely to affect inferences about $\boldsymbol{\beta}$: For given $\boldsymbol{\theta}$ (and thus given $\boldsymbol{V}$) the variance of the GLS-estimator of $\boldsymbol{\beta}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \left(\boldsymbol{X}'\boldsymbol{Z}\left[\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Phi}^{-1}\right]^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \tag{15}$$

which increases with the variance parameters $\boldsymbol{\theta}$. Thus if $\hat{\boldsymbol{\theta}}$ has a downward bias, so will $\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ with $\hat{\boldsymbol{V}}$ computed from the ML estimator $\hat{\boldsymbol{\theta}}$. As a consequence, significance tests and Wald tests of hypotheses about $\boldsymbol{\beta}$ will be anti-conservative and confidence

intervals will be too short in length.

That ML estimators of error variances are biased even in the case of linear regression with normal distributed errors is a fact well-known enough to statisticians so that attempts were made early on to correct the bias of variance parameters of linear mixed models. To this purpose, a modified version of the ML estimator of variance parameters was developed by Patterson and Thompson (1971), the restricted maximum likelihood estimator or residual maximum likelihood estimator (both abreviated as "REML"). This estimator for variance parameters of mixed-effects models has been compared to the role that the corrected sample variance

$$s^2_{\text{corr}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

plays as an unbiased estimator for a population variance. In the following we show that REML is a special case of an estimator that maximizes the *modified profile likelihood* suggested by Cox and Reid (1987) (see also McCullagh and Tibshirani 1990a) of which the usual unbiased estimator of the error variance in linear regression is another special case.

To understand maximum modified profile likelihood estimators, one needs of course to understand profile likelihood functions. Suppose $\boldsymbol{\psi}$ is the parameter vector of a statistical model and suppose further that it can be split in to parts $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ such that the log-likelihood function $\ell(\boldsymbol{\psi};\boldsymbol{y}) = \ell(\boldsymbol{\lambda},\boldsymbol{\theta};\boldsymbol{y})$ can easily maximized for $\boldsymbol{\lambda}$ with $\boldsymbol{\theta}$ held fixed (as in the case of regression coefficients in a normal linear model or the fixed-effects coefficients in a normal linear mixed effects model). If $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}}$ denotes the value of $\boldsymbol{\lambda}$ that maximizes $\ell(\boldsymbol{\lambda},\boldsymbol{\theta};\boldsymbol{y})$ for $\boldsymbol{\theta}$ held fixed then the profile log-likelihood function is a function of $\boldsymbol{\theta}$ alone defined as:

$$\ell_{\text{p}}(\boldsymbol{\theta};\boldsymbol{y}) = \ell(\hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}},\boldsymbol{\theta};\boldsymbol{y})$$

In case of linear regression with normal i.i.d. disturbances the profile log-likehood function with respect to the disturbance variance σ^2 is

$$\ell_{\text{p}}(\sigma^2;\boldsymbol{y}) = c - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}) = c - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{y}$$

where n is the number of observations and $\boldsymbol{P}_X = \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'$ (Harville 1997, 166ff). In case of the linear normal mixed-effects model the profile log-likelihood function with respect to

the variance parameters in $\boldsymbol{\theta}$ is

$$\ell_{\mathrm{p}}(\boldsymbol{\theta};\boldsymbol{y}) = c^{\dagger} - \frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}) = c^{\dagger} - \frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}\boldsymbol{y}'\boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{V}^{-1}})\boldsymbol{y}$$

where $\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{V}^{-1}} = \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$ (Harville 1997, 260ff).

The *modified* profile likelihood approach suggested by Cox and Reid (1987) consists in maximizing, instead of the profile log-likelihood just discussed, a modified version thereof, namely

$$\ell_{\mathrm{mp}}(\boldsymbol{\theta};\boldsymbol{y}) = \ell(\hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}},\boldsymbol{\theta};\boldsymbol{y}) + \frac{1}{2}\ln\det\left(-\frac{\partial^2\ell(\hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}};\boldsymbol{\theta})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right). \tag{16}$$

In case of linear regression with normal i.i.d. disturbances, k independent variables and an intercept, the modified log-likelihood function is

$$\begin{aligned}\ell_{\mathrm{mp}}(\sigma^2;\boldsymbol{y}) &= c - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) + \frac{1}{2}\ln\det\left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}\right)\\ &= c - \frac{n-k-1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) + \frac{1}{2}\ln\det\left(\boldsymbol{X}'\boldsymbol{X}\right)\end{aligned}$$

since $\det\left(\sigma^{-2}\boldsymbol{X}'\boldsymbol{X}\right) = (\sigma^{-2})^{k+1}\det\left(\boldsymbol{X}'\boldsymbol{X}\right)$, because $\boldsymbol{X}'\boldsymbol{X}$ is a $(k+1)\times(k+1)$ matrix. Now it is easy to see that setting $\partial\ell_{\mathrm{mp}}(\sigma^2)/\partial\sigma^2$ to zero leads to the well-known unbiased estimator of σ^2,

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})}{n-k-1}.$$

In case of normal-linear mixed models, the modified log-likelihood leads to the REML estimator proposed by Patterson and Thompson (1971):

$$\ell_{\mathrm{mp}}(\boldsymbol{\theta};\boldsymbol{y}) = \ell_{\mathrm{REML}}(\boldsymbol{\theta};\boldsymbol{y}) = c^{\dagger} - \frac{1}{2}\ln\det(\boldsymbol{V}) + \frac{1}{2}\ln\det\left(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}\right) - \frac{1}{2}\boldsymbol{y}'\boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{V}^{-1}})\boldsymbol{y}. \tag{17}$$

In contrast to linear regression with normal i.i.d. disturbances, there does not exist a simple solution formula for the variance parameters, as already the case with respect to maximum likelihood. Nevertheless, REML estimators differ from ML estimators only in the way estimates of variance parameters are computed, while the way fixed effects coefficients are treated is, for given values of the variance parameters essentially the same. Most importantly, in so far as REML estimators can be viewed as a generalisation to multilevel models of unbiased estimators of the error variance in linear regression. Thus, it seems reasonable to expect that if variance parameters of multilevel models are estimated by REML instead of ML then the

coverage error of interval estimates of fixed-effects coefficients found by Stegmüller (2013) can be considerably reduced if not eliminated.

The correctness of confidence intervals, i.e. what Stegmüller's terminology would be the unbiasedness of interval estimates, hinges not only on unbiased estimates of estimated standard errors (which in turn depends on the unbiasedness of estimators of variance parameters), but usually also on the correctness of the distributional assumptions about the sampling distribution of point estimators. Standard statistical theory states that ML estimators are, under suitable regularity conditions, consistent and asymptotically normal. Roughly speaking this means that, first, whatever bias an ML estimator may have in small samples, this bias will become negligible if the samples size approaches infinity and second, that the sampling distribution of the estimator will become close to a (multivariate) normal distribution centred on the true parameter value with a known variance matrix. Further, not only are estimators of parameters consistent under the suitable conditions, but also standard estimators of this variance matrix of the sampling distribution of the parameters, the inverse of the information matrix evaluated at the parameter estimates. Relying on this theory, most statistical software uses a normal distribution for single-parameter Wald-tests and confidence intervals and a χ^2 distribution for likelihood-ratio tests and multi-parameter Wald-tests. The usual confidence interval for coefficient β_k then can be constructed by

$$\hat{\beta}_{k,\text{lower}} = \hat{\beta}_k + z_{0.025}\widehat{\text{SE}}\left(\hat{\beta}_k\right) \quad \text{and} \quad \hat{\beta}_{k,\text{upper}} = \hat{\beta}_k + z_{0.975}\widehat{\text{SE}}\left(\hat{\beta}_k\right) \tag{18}$$

For the asymptotic theory behind such interval estimators to be valid, the number of *independent* observations has to increase without bounds. Yet this does not hold for multi-level mixed-effects models if the number of observations increases only *within* groups and the number of groups themselves stays constant, because of the interdependence of the observations within the groups. If asymptotic normality cannot be guaranteed to hold multi-level mixed-effects models with a small number of groups, normality-based single-parameter tests and confidence intervals, as well as χ^2-based multi-parameter tests would be misleading, *even if* the variance of the sampling distribution of parameter estimates were estimated without bias.

Of course, the failure of asymptotic theory to account for the small-sample behaviour of ML and related estimators has already been noted by theoretical statisticians. Indeed, it has been shown that GLS-based standard errors (as in equation (15)) are only asymptotically valid variance parameters need to be estimated and are biased if the sample size is small (Kenward

and Roger 1997). Kenward and Roger (1997) develop a formula to adjust for this bias in small samples. The same authors also suggest a degrees of freedom adjustment for multi-parameter Wald tests that goes back to Satterthwaite (1941). Such degree-of-freedom adjustments can of course also be used for single-coefficient t-tests and for confidence intervals based on a t-distribution (see also Manor and Zucker 2004). Such a confidence interval for a coefficient β_k could then be constructed by

$$\hat{\beta}_{k,\text{lower}} = \hat{\beta}_k + t_{d,0.025}\widehat{\text{SE}}\left(\hat{\beta}_k\right) \quad \text{and} \quad \hat{\beta}_{k,\text{upper}} = \hat{\beta}_k + t_{d,0.975}\widehat{\text{SE}}\left(\hat{\beta}_k\right) \tag{19}$$

where $t_{d,0.025}$ and $t_{d,0.975}$ are the 2.5 and 97.5 percentiles of the t-distribution with d degrees of freedom. Using such confidence intervals based on a t-distribution instead of a normal distribution could reduce the coverage error found by Stegmüller (2013) even further, if it does not eliminate it altogether.

Generalised linear mixed models beyond the normal-linear type pose additional challenges. First, they lead to likelihood functions that involve (sometimes high-dimensional) integrals that do not have a closed-form solution. Second, due to the non-linearity in the link between coefficients and the conditional expectation of the response variable, coefficient estimates inevitably are biased in small samples McCullagh and Nelder 1989 and it may be difficult to establish how quickly this bias vanishes as the sample size increases (for bias correction in generalised linear models, see Firth 1993). So the relatively reassuring result about the unbiasedness of estimators for parameters in normal-linear mixed model, summarized by equation (14), does not necessary carry over to generalised linear mixed models. In the following we first present the structure of generalised linear mixed models and discuss how ML and REML estimation would work in this type of models.

If $f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\mu},\tau)$ is the density or probability mass function of the conditional distribution of the response for given values of the random effects vector, then the log-likelihood function for a generalised linear mixed model takes the form of the integral

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = c - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) + \ln\int f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\mu},\tau)\exp\left[-\frac{1}{2}\boldsymbol{u}'\boldsymbol{\Phi}^{-1}\boldsymbol{u}\right]\mathrm{d}\,\boldsymbol{u} \tag{20}$$

for which a solution formula exists only if the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ is normal. In the absence of a solution formula the integral involved in the log-likelihood function of generalised linear mixed effecs models cannot be computed exactly, but only be approximated. The chief analytical approximation in use is the Laplace approximation (Breslow and

Clayton 1993), whereas the most widely used numeric approximations are Gauss-Hermite quadrature and Monte Carlo integration (McCulloch 1997; Booth and Hobert 1999; Caffo et al. 2005).

The crucial advantage of the Laplace approximation, introduced by Breslow and Clayton (1993) as penalised quasi-likelihood (PQL), is that it makes it easy to translate the concept of restricted maximimum likelihood to generalised linear mixed models beyond the normal-linear case. The Laplace approximation rests on the second-order Taylor expansion of the exponent integrand in (20) around its maximum and integration of the this quadratic approximation of the integrand, which leads, with

$$\frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\mu},\tau)}{\partial \boldsymbol{u}}(\tilde{\boldsymbol{u}}) - \boldsymbol{\Phi}^{-1}\tilde{\boldsymbol{u}} = 0$$

to

$$\begin{aligned}\ell(\boldsymbol{\theta};\boldsymbol{y}) &\approx c - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) + \ln\int \exp\left[\ln f(\boldsymbol{y}|\tilde{\boldsymbol{u}};\boldsymbol{\mu},\tau) + \frac{1}{2}(\boldsymbol{u}-\tilde{\boldsymbol{u}})'\tilde{\boldsymbol{K}}(\boldsymbol{u}-\tilde{\boldsymbol{u}}) - \frac{1}{2}\boldsymbol{u}'\boldsymbol{\Phi}^{-1}\boldsymbol{u}\right]\mathrm{d}\,\boldsymbol{u} \\ &= c^{\dagger} - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) + \ln f(\boldsymbol{y}|\tilde{\boldsymbol{u}};\boldsymbol{\mu},\tau) - \frac{1}{2}\ln\det\left(\tilde{\boldsymbol{K}} + \boldsymbol{\Phi}^{-1}\right) - \frac{1}{2}\tilde{\boldsymbol{u}}'\boldsymbol{\Phi}^{-1}\tilde{\boldsymbol{u}}\end{aligned} \tag{21}$$

Breslow and Clayton point out that, if the conditional distribution of the response is in a exponential family, finding the PQL estimates for the fixed-effects coefficients for given $\boldsymbol{\Phi}$ leads to the GLS-like estimation equation

$$\boldsymbol{X}'\tilde{\boldsymbol{V}}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\tilde{\boldsymbol{V}}^{-1}\boldsymbol{y}^{*} \tag{22}$$

where $\boldsymbol{y}^{*}$ is the usual "working response" known from the GLM literature (McCullagh and Nelder 1989) with components

$$y_i^{*} = \tilde{\eta}_i + (y_i - \tilde{\mu}_i)\left(\frac{\partial\mu}{\partial\eta}\right)^{-1}$$

and

$$\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{W}}^{-} + \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}'.$$

$\tilde{\boldsymbol{W}}^{-}$ is a generalised inverse of a diagonal matrix with diagonal elements

$$\tilde{w}_{ii} = \frac{\partial^2\mu_i}{(\partial\eta_i)^2}(\tilde{\eta}_i)$$

where the tilde (e.g. $\tilde{\mu}_i$) indicates these quantities are evaluated at $\boldsymbol{u} = \tilde{\boldsymbol{u}}$. Note that in this situation we also have

$$\tilde{\boldsymbol{K}} = \boldsymbol{Z}'\tilde{\boldsymbol{W}}\boldsymbol{Z}$$

By substituting the estimation for $\tilde{\boldsymbol{u}}$

$$(\boldsymbol{Z}'\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{\Phi}^{-1})\tilde{\boldsymbol{u}} = \boldsymbol{Z}'\tilde{\boldsymbol{W}}(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta})$$

and using a quadratic expansion for $\ln f(\boldsymbol{y}|\tilde{\boldsymbol{u}}; \boldsymbol{\mu}, \tau)$ (22) can further approximated by

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) \approx c^\dagger - \frac{1}{2}\det(\tilde{\boldsymbol{V}}) - \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta})\tilde{\boldsymbol{V}}^{-1}(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta}). \tag{23}$$

which is similar to the log-likelihood in the normal-linear case. It should be noted that the derivatives of (22) and (23) for the parameters involved in $\boldsymbol{\Phi}$ are virtually identical if the dependence of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$ on $\boldsymbol{\Phi}$ is taken into account. As pointed out by Breslow and Clayton (1993), equation (23) can be used to introduce a "REML-like" variant of PQL:

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) \approx c^\dagger - \frac{1}{2}\det(\tilde{\boldsymbol{V}}) - \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta})\tilde{\boldsymbol{V}}^{-1}(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\det\left(\boldsymbol{X}\tilde{\boldsymbol{V}}^{-1}\boldsymbol{X}\right). \tag{24}$$

It should be noted that the accuracy of the Laplace approximation depends on the group size (and but not on the number of groups). With smaller group sizes, the Laplace approximation may lead to bias (usually a downward bias of the variance parameters). For dealing with such situations, bias-corrections based on a higher-order Laplace approximation have been proposed (Breslow and Lin 1995; Lin and Breslow 1996) as well as Monte-Carlo integration approaches, that allow to increase the accuracy of the appoximation of the integrals involved in the likelihood to any desired degree by increasing the number of Monte Carlo replicates (algorithms for automatically increasing the Monte Carlo sample sizes have been proposed by Booth and Hobert 1999 and Caffo et al. 2005). How the "logic" of REML can be applied to these setups is much less straightforward, but see McCullagh and Tibshirani (1990b).

The implications of this discussion can be summarised as follows: (1) ML estimates of fixed-effects coefficients in linear mixed models are generally unbiased but (2) ML estimates of variance parameters in these models are biased, especially if the number of groups is small. (3) A downward bias of ML estimates of variance parameters leads to a downward bias of estimated standard errors and to interval estimates with that cover the true parameter value with lower probability than the nominal confidence level. (4) REML, a modification of ML,

has the potential to reduce the bias of variance parameter estimates and the undercoverage of interval estimates of coefficient estimates. (5) If the number of groups is small, the bias correction provided by REML may not be sufficient to guarentee a satisfactory coverage performance of interval estimates if these are based on the assumption of asymptotic normality. (6) The failure of asymptotic normality can at least approximately corrected by basing interval estimates on an appropriate t-distribution. (7) Since coefficient estimates in generalised linear models are generally biased in small samples, they are no less likely to be biased in generalised linear mixed models. (8) It remains an open question whether REML estimation of variance paremeters in generalised linear mixed models leads to a bias reduction in the same way as it does for normal linear mixed models and whether the use of a t-distribution for interval estimates of coefficients in generalised linear mixed models leads to the same amount of reduction in coverage error as in the case of normal linear mixed models.

The central claim made in Stegmüller's (2013) article is that frequentist point and interval estimators of coefficients in multilevel models are seriously biased if the number of higher-level units is small and that Bayesian techniques are superior in so far as they exhibit much less bias. The previous discussion should make clear that it would be premature to conclude from a bias in ML estimators that frequentist techniques in general are flawed, since ML is not the only estimator relevant for multilevel models and asymptotic normality is not the only possible assumption on which to base the construction of interval estimates. The question thus arises what Bayesian techniques can contribute to the improvement of ML estimation. We therefore discuss how ML and Bayesian techniques differ.

The fundamental difference between frequentist and Bayesian inference is that, while in the frequentist framework model parameters are treated as fixed but unknown constants, Bayesian inference treats them random variables. Here one is interested in the conditional distribution of the model parameters given the data and a pre-determined prior distribution, a probability distribution supposed to reflect knowledge about the parameters before the data are acquired and analysed. In essence, Bayesian inference rests on applying Bayes' theorem to model parameters. To illustrate, let $\boldsymbol{\theta}$ be the vector of all parameters in a multilevel model and let $p(\boldsymbol{\theta})$ the density function of the prior distribution of the model parameters (in short, the prior density), let $p(\boldsymbol{y}|\boldsymbol{\theta})$ be the conditional distribution – the *likelihood* – of the observed data (of the response variable in the model) given the values of the parameter values $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|\boldsymbol{y})$ the density function – the *posterior density* – of the conditional distribution of the parameter vector given the observed data – the *posterior distribution*. Then this posterior

density is given by Bayes' theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\,\boldsymbol{\theta}}. \tag{25}$$

The principal relation between Bayesian inference and maximum likelihood estimation rests on the fact that the likelihood function in both approaches is identical, that is, $p(\boldsymbol{y}|\boldsymbol{\theta}) = \exp(\ell(\boldsymbol{\theta}; \boldsymbol{y}))$.

Once the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$ is derived, it is possible to construct "Bayesian point estimates" of the parameters, which come in two principal variants. One variant of Bayesian point estimates is the posterior mean

$$\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{y}) = \int \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\,\boldsymbol{\theta}$$

and the other variant is the posterior mode

$$\mathrm{mode}(\boldsymbol{\theta}|\boldsymbol{y}) = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta}|\boldsymbol{y}) = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))$$

where the latter equality holds because the normalising constant $p(\boldsymbol{y})$ of the posterior density does not depend on $\boldsymbol{\theta}$. It is easy to see that if the posterior expectation exists and if the posterior density is symmetric, then these two variants give identical results.

Bayesian inference is in general confronted with two principal challenges: (1) a prior distribution has to be chosen – and this decision may influence its results – for which many options are available; and (2) the functional form of the posterior is often unknown because the integral in the denominator does not have a closed-form solution. The only setting where the functional form of the posterior is known arises when the prior distribution, identified by $p(\boldsymbol{\theta})$ is conjugate to the conditional distribution of the data, identified by $p(\boldsymbol{y}|\boldsymbol{\theta})$. The first problem is minor if the amount of data is large enough so that the shape of the posterior density is dominated by the likelihood. But in this case, ML and Bayesian inference are unlikely to yield different results. The second problem is usually addressed by a Monte Carlo approximation to the posterior distribution, the most popular approximation being the Markov Chain Monte Carlo (MCMC) method using a Gibbs sampler.

If inferences about random intercepts and slopes are made, they are already Bayesian even if the variance parameters are estimated from the data. In this context, the distribution of the

random intercepts and slopes, which usually is assumed to be (multivariate) normal, can be considered as a prior distribution, and if ML (or REML) estimates of the model parameters are computed using an Expectation-Maximization algorithm the posterior means of the random intercepts and slopes emerge as a "by-product" of the computations (Dempster et al. 1977; McLachlan and Krishnan 2008). In this case one can talk of thus predicted values of the random intercepts and slopes as *empirical Bayes* predictions (Carlin and Louis 1996; Casella 1985). A "fully Bayesian" approch in contrast also assumes that the fixed-effects coefficients and the variance parameters have a probabity distribution.

In a typical Bayesian approach to inference about linear mixed model of the form given by equation (6) one uses a multivariate normal distribution as prior for the coefficent vector $\boldsymbol{\beta}$, an inverse-Gamma distribution as prior for the individual level variance parameter σ^2, and an inverse-Wishart distribution as prior for the variance matrix $\boldsymbol{\Phi}$. While it is in general difficult to desribe the resulting posterior distribution, a reason why also in this setting MCMC simulation with Gibbs-sampling is used, some statements can be made about posteriors in a simple version of this model.

Consider a simple two-level model with only a random intercept (and no random slope). Such model can be written as

$$\begin{aligned} y_{ij} &\sim N(a_j + bx_{1ij}, \sigma^2) \\ a_j &\sim N(\beta_{00} + \beta_{01}x_{2j}, \phi) \end{aligned}$$

for which we derive two conditional posteriors for β_{01} and ϕ. This is because both parameters are discussed by Stegmüller (2013).[1] As stated above, the data are assumed to be normally distributed, which leads to the following fully conditional distribution of the observation y_{ij}:

$$p(y_{ij}|a_j, b, \sigma^2) \propto \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{(y_{ij} - a_j - \beta_{10}x_{1i})^2}{2\sigma^2}\right\}$$

The conditional density of the random intercept a_j is

$$p(a_j|\beta_{00}, \beta_{01}, \phi) \propto \frac{1}{\sqrt{\phi}} \exp\left\{-\frac{(a_j - \beta_{00} - \beta_{01}x_{2j})^2}{2\phi}\right\}$$

[1] For the posteriors of a model including random slope see e.g. Seltzer et al. (1996).

By assuming a normal prior for β_{01}:

$$\beta_{01} \sim N(\mu_{01}, \tau_{01}) \tag{26}$$

β_{01}'s conditional posteriors can be derived as follows:

$$\begin{aligned}
p(\beta_{01}|.) &\propto p(a_j|\beta_{00}, \beta_{01}, \phi)p(\beta_{01}) \\
&\propto \left(\prod_j \frac{1}{\sqrt{\phi}} \exp\left\{-\frac{(a_j - \beta_{00} - \beta_{01}x_{2j})^2}{2\phi}\right\}\right) \frac{1}{\sqrt{\tau_{01}}} \exp\left\{-\frac{(\beta_{01} - \mu_{01})^2}{2\tau_{01}}\right\} \\
&\propto \exp\left\{\left(-\frac{1}{2}\right)\left(\frac{\Phi(\beta_{01} - \mu_{01})^2 + \tau_{01}\sum_j (a_j - \beta_{00} - \beta_{01}x_{2j})^2}{\phi\tau_{01}}\right)\right\} \\
&\propto \exp\left\{\left(-\frac{1}{2}\right)\left(\frac{(\phi + \tau_{01}\sum_j x_{2j}^2)\beta_{01}^2 - 2(\phi\mu_{01} + \tau_{01}\sum_j x_{2j}(a_j - \beta_{00}))\beta_{01}}{\phi\tau_{01}}\right)\right\}
\end{aligned}$$

Therefore:

$$\beta_{01}|. \sim N\left(\frac{\phi\mu_{01} + \tau_{01}\sum_j x_{2j}(a_j - \beta_{00})}{\phi + \tau_{01}\sum_j x_{2j}^2}; \frac{\phi\tau_{01}}{\phi + \tau_{01}\sum_j x_{2j}^2}\right)$$

Obviously the prior is an important source for the difference in the ML and Bayesian estimates. That is, by setting a smaller value in τ_{01} researchers can set up certain informative priors which significantly affect the posterior. In this case, Bayesian estimation can lead to point and interval estimates which are different from ML. In contrast: If τ_{01} approaches to ∞, that is, in case of very flat prior the posterior is reduced to:

$$\lim_{\tau_{01}\to\infty} \beta_{01}|. \sim N\left(\frac{\sum_j x_{2j}(a_j - \beta_{00})}{\sum_j x_{2j}^2}; \frac{\phi}{\sum_j x_{2j}^2}\right) \tag{27}$$

Note that the posterior's expected value is only conditioned by a_j and β_{00} and corresponds to the maximum likelihood point estimate for given a_j and β_{00}. Analogous conditional posterior can be also derived for a_j and β_{00} which also corresponds to the maximum likelihood point estimate. Therefore, the maximum likelihood estimation and Bayesian estimation should deliver a same point estimate independently from ϕ in case of uninformative priors.

Now we move on to ϕ's conditional posterior. By assuming inverse gamma prior:

$$\phi \sim IG(a, b)$$

we can derive ϕ's conditional posterior as follows:

$$p(\phi|.) \propto p(a_j|\beta_{00},\beta_{01},\phi)p(\phi) \propto \left(\prod_j \frac{1}{\sqrt{\phi}} \exp\left\{-\frac{(a_j-\beta_{00}-\beta_{01}x_{2j})^2}{2\phi}\right\}\right)\frac{1}{\phi^{a+1}}\exp\left(-\frac{b}{\phi}\right)$$

After some algebra we obtain:

$$p(\phi|.) \sim IG\left(\frac{m}{2}+a+1, \frac{\sum_j(a_j-\beta_{00}-\beta_{01}x_{2j})^2+2b}{2}\right)$$

Here again, by setting a significantly large number in a and b prior information can add additional information to the likelihood. If we, however, specify a very small value for a and b the ML and Bayesian point estimates can still differ since the inverse gamma distribution is asymmetric and right-skewed. That is, the maximum of likelihood is smaller than the expected value of posterior. Correspondingly ML estimates of ϕ is smaller than the corresponding posterior mean (Rubin 1981).

The issue of skewed posterior distribution of ϕ is in particular relevant if we have a small number of groups e.g. m=5 (Rubin 1981; Draper 1995). In this case, not only the difference between ML and Bayesian estimates, but also the issue about the choice of the prior is important. It is well known that the prior distribution has a larger impact if the amount of observed information pertinent to the parameter, in the present case the number of gruops, is small. In the case of inverse gamma priors one can see both parameters of the posterior distribution in equation 3 strongly depend on m. And if one has smaller value for m different values in a and b lead to a significant difference in the resulting posterior distribution of ϕ (Gelman and others 2006).

As an alternative to such skewed prior distributions uniform priors on the the variance parameters have been suggested in the literature. Yet the choice of such a uniform prior is not straightforward, rather there are different options: (1) a uniform distribution on $\log\phi$, (2) a uniform distribution on $\sqrt{\phi}$, and (3) a uniform distribution on ϕ. It is well known that the latter two priors leads to larger posteriors of ϕ. In particular the uniform distribution on ϕ (variance) strongly boosts the posterior distribution (Gelman and others 2006). Additionally Gelman and others (2006) suggests a half t-family priors which can also deal with a totally skewed and censored posterior whose maximum corresponds to zero.

4 The Bias of ML and REML Point and Interval Estimators – A Monte Carlo Study

In the present section we present a Monte Carlo study on the bias of point estimates of fixed-effects coefficients and variance parameters of linear mixed models and generalised linear mixed models and on the coverage error of interval estimates of fixed-effects coefficients in these models. This Monte Carlo study serves several purposes. Firstly, we aim to examine how it is possible for Stegmüller (2013) to find a bias in fixed-effect coefficient estimates even if the theory discussed in a previous section indicates that such a bias does not exist in theory. Secondly, we aim to examine the degree to which REML and a t-distribution assumption leads to interval estimates of fixed-effect coefficients that achieve a coverage of true parameter values close to the nominal level of confidence. Third, we aim to examine the degree to which arguments pertaining normal linear mixed models also work for generalised linear mixed models at least in situations with large group sizes typical for cross-national survey research on political attitudes and behaviour.

To these purposes we replicate a part of Stegmüller's Monte Carlo study, but we also extend it by considering OLS and REML estimators for fixed-effects coefficients and REML estimators for variance parameters in linear mixed models. We further extend Stegmüller's study by considering REML-type estimators for parameters in mixed probit models. Like Stegmüller, we vary the number of groups in the Monte Carlo simulations of the linear mixed and mixed probit models as we vary the true values of the variance parameters, but in contrast to Stegmüller, we use a larger number of Monte Carlo replications (2,000 instead of 500) and present the result in somewhat more detail. Not only do we show the results for different true values of the variance parameters – to elucidate the effect of group-level variance on the performance of the various estimators – but, more importantly, we make explicit the consequences of Monte Carlo error.

It is of paramount importance in the interpretation of the results of Monte Carlo studies to keep in mind that they cannot be exact. To clarify, consider a typical Monte Carlo study of the bias of an estimator of a parameter θ. Here one generates R data sets from a probability distribution with true parameter value θ_0, applies the estimator to each of data sets to obtain

parameter estimates $\hat{\theta}^{(r)}$ (with $r = 1, \ldots, R$) and estimates the bias of the estimator by

$$\text{bias}_R^*(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}^{(r)} - \theta_0.$$

By the law of large numbers we have that

$$\text{plim}_{R\to\infty} \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}^{(r)} = \text{E}(\hat{\theta}) \quad \text{hence} \quad \text{plim}_{R\to\infty} \text{bias}_R^*(\hat{\theta}) = \text{bias}_R(\hat{\theta}),$$

but if the parameter space of θ is continous, $\widehat{\text{bias}}_R(\hat{\theta}) \neq 0$ almost surely, even if $\text{bias}_R(\hat{\theta}) = 0$, as long as R is finite. In other words, as long as the Monte Carlo study has a finite number of replications, a simulated bias in a Monte Carlo study will randomly differ from zero even if the true bias is zero. In order not to be mislead by random departures that occur almost surely (i.e. with probability one) Monte Carlo estimates of bias should always be accompanied by a measure of Monte Carlo error, for example by confidence intervals. Since such a simulated bias is a arithmetic mean, one can rely on the central limit theorem and construct a confidence interval for $\text{bias}_R(\hat{\theta})$ based on normal distribution quantiles and the Monte Carlo standard error

$$\text{SE}\left(\frac{1}{R} \sum_{r=1}^{R} \hat{\theta}^{(r)}\right) = \sqrt{\frac{1}{R}\left(\frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)})^2 - \left[\frac{1}{R} \sum_{r=1}^{R} \hat{\theta}^{(r)}\right]^2\right)}.$$

In the presentation of our Monte Carlo study we not only report the simulated biases, but also 95 percent confidence intervals for various settings of the number of groups and the size of variance parameters. Such confidence intervals will contain the true bias with 95 percent probability. However, if we simulate a bias under various settings, where the number of settings may be for example $3 \cdot 5 \cdot 6 = 90$ — for three estimators (OLS, ML, and REML), five settings of the variance parameter and six settings for the number of groups – then the probability that *all* confidence intervals of these biases cover the true biases is only $0.95^{90} = 0.00988$, i.e. less than one percent. For this reason we not only report conventional confidence intervals but also Šidák-corrected confidence intervals, which are wide enough so that the true bias is contained with 95 percent probability by all confidence intervals (Šidák 1967). The Monte Carlo study was conducted with the statistical software *R* (R Core Team 2013).

In the first experiment, we examine the performance of OLS, ML, and REML estimators for the coefficient of a two-level random-intercept model with two independent variables $\boldsymbol{x}_1$ and

$\boldsymbol{x}_2$ and a dependent variable $\boldsymbol{y}$. The model is formulated as

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j + \epsilon_{ij}$$

where i denotes the running number of the individual observation and j denotes the running number of the group which the individual observation is a member of, y_{ij}, x_{ij}, and x_{2j} are elements of $\boldsymbol{y}$, $\boldsymbol{x}_1$, and $\boldsymbol{x}_2$, respectively, u_j are group-level elements of the random effects vector $\boldsymbol{u}$ (each with a normal distribution with zero mean and variance θ), ϵ_{ij} are the elements of an individual-level disturbance vector $\boldsymbol{\epsilon}$ (each with a normal distribution with zero mean and variance σ^2). That is, the elements of the vector $\boldsymbol{x}_1$ vary across individuals and groups, whereas the values of the vector $\boldsymbol{x}_2$ vary only across groups, but are constant within groups.

In the experiment, we vary the number of groups m within the set $\{5, 10, 15, 20, 25, 30\}$ as well as the intra-class correlation $\rho = \theta/(\sigma^2 + \theta)$ within the set $\{0, 0.05, 0.1, 0.15, 0.3\}$. Througout the settings, the values of the coefficients are fixed at $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 1$. Further, the individual level disturbance variance is held fixed at $\sigma^2 = 1$, so that the variance of the random effects variance is determined by σ^2 and ρ as $\theta = \sigma^2\rho/(1 - \rho)$. Like in Stegmüller (2013), the number of observations within each group is 500. For each combination of settings of intra-class correlation and number of groups 2000 times data were generated accoring to the two-level random intercept model (the values of the independent variables each having a standard normal distribution). We use the R-package `nlme` (Pinheiro et al. 2013) for obtaining point and interval estimates of fixed-effects coefficients and variance parameters.

Figure 1 shows the simulated bias of the OLS, ML, and REML estimators of β_1, the coefficient of the predictor $\boldsymbol{x}_1$ that varies both between individuals and groups. The dots connected by lines in the diagram show the average differences between estimates and the true value of the coefficient in terms of the percentage of the true value. The dark-gray areas represent conventional 95 percent confidence intervals of the simulated relative bias, the light-gray areas represent Šidák-corrected confidence intervals. These gray areas represent the Monte Carlo error created by the fact that the number of replications is finite. In the same vein Figure 2 shows the simulated bias of the tree estimators of β_2, the coefficient of the predictor $\boldsymbol{x}_2$ that is constant within and varies only between groups. Finally, Figure 3 shows the simulated bias of the three estimators of θ, the variance of the group-level random slopes.

The two figures provide Monte Carlo evidence for the claim that there is no systematic bias

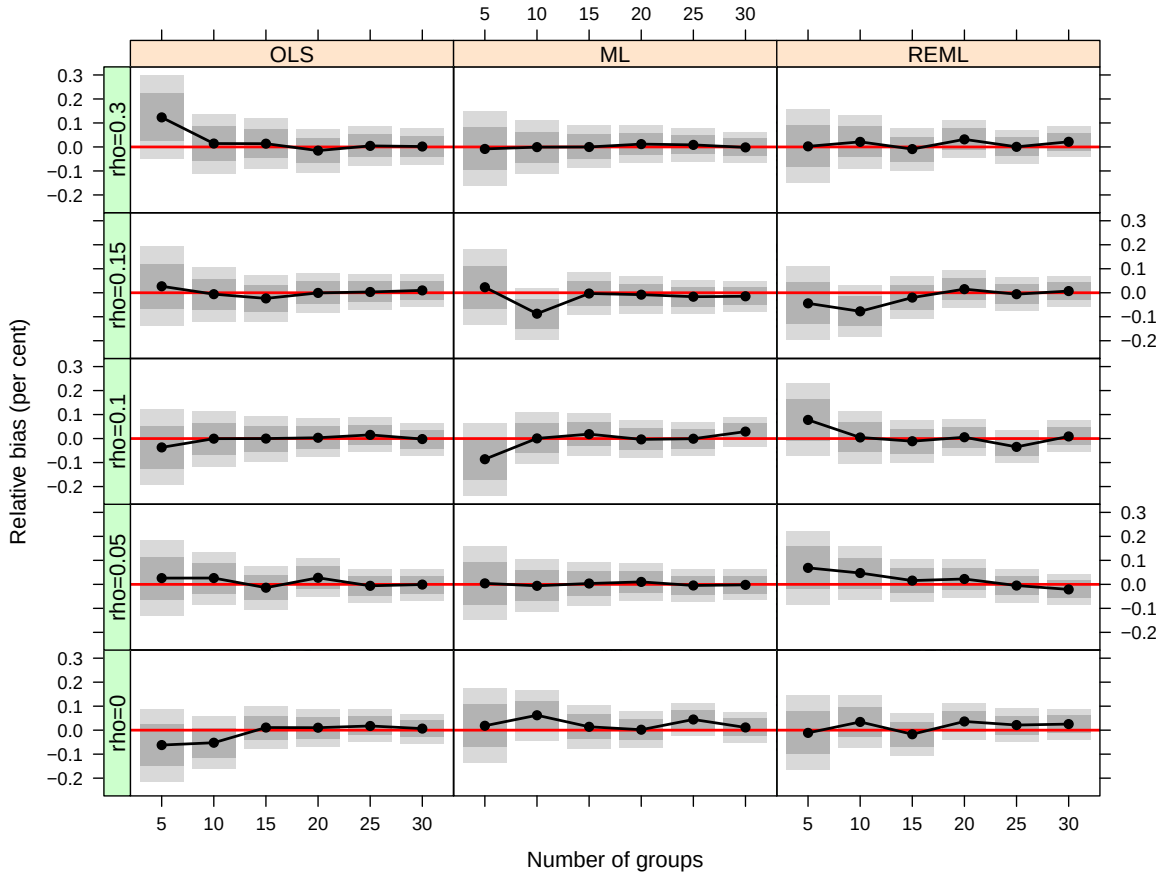

Figure 1: Relative simulated bias (in per cent) of the estimated (fixed effect) coefficient β_1 of the predictor $\boldsymbol{x}_1$ that varies across individuals and across croups.

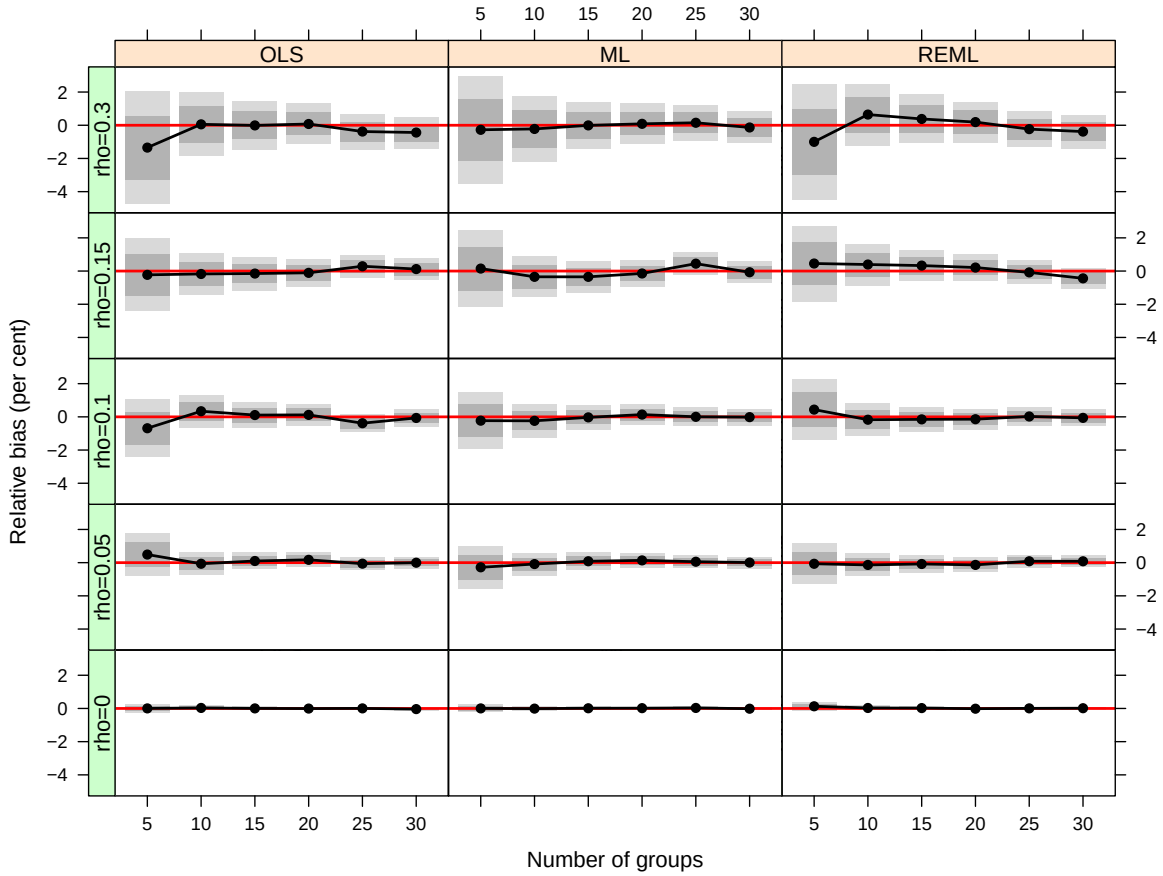

Figure 2: Relative simulated bias (in per cent) of the estimated (fixed effect) coefficient β_2 of the predictor $\boldsymbol{x}_2$ that varies only across croups and is constant within groups.

in the estimates produced by any of the estimators whatever the number of the groups or the amount of intra-class correlation. The average simulated bias departs from zero considerably in some of the settings, especially in Figure 2, but this should not be mistaken for a substantial bias. Instead, these departures stem from the fact that estimates vary and their averages therefore also vary. That is, the departures from zero that average simulated bias exhibits in the simulation study is merely the consequence from the fact that the number of replications (simulation runs) is finite. In most instances their 95 per cent confidence intervals envelope the zero line and the corrected confidence do so in all settings. That is, if the number of simulation runs were increased, these confidence intervals would get shorter, but also the departures of the simulated averages would get closer to zero. (The fact that the diagrams in Stegmüller (2013) do not show such simulation confidence intervals may however convey the misleading impression that there is a substantial bias in the coefficient estimates.) What the unsystematic departures of the average simulated bias of the estimates of β_2 in Figure 2 mostly show however is how much their variability increases with the intraclass correlation. Yet the choice of the estimator seems to have little impact of the amount of variability. Of course, all this should not surprise, since it can be mathematically proved that no such bias exists.

While the theoretical discussion above indicated that no bias exists in ML estimates of fixed-effects coefficients, it also suggested that ML estimates of variance parameters may be biased, especially if the number of groups is small, even though it was not possible to derive the size of this bias (as opposed to the bias of ML estimates of error variances in normal linear regression). The dicussion of REML estimators also suggested that they may lead to a reduced bias, if not its elimination. Figure 3 allows to appreachiate the bias of ML and REML estimates of a variance parameter, it shows the relative bias of the group-level random intercept variance θ when estimated by these methods (since OLS does not provide such estimates OLS, results are not shown in the figure). ML leads to a clear downward bias of the variance parameter estimates, which amounts to minus 40 percent when there are only five groups, but which gets smaller in size as the number of groups increases. Yet still if there are 30 groups, the bias is at least minus five percent. In contrast, the REML estimator does not exhibit any such bias. Instead, the average simulated bias stays closely to the zero line and its confidence intervals encompass it in all settings of number of groups and intraclass correlation. That is, even if the number of groups is only five, the bias in the estimated random-effects variance is negligible.

In the second experiment, model of the first experiment is extended by the inclusion of ran-

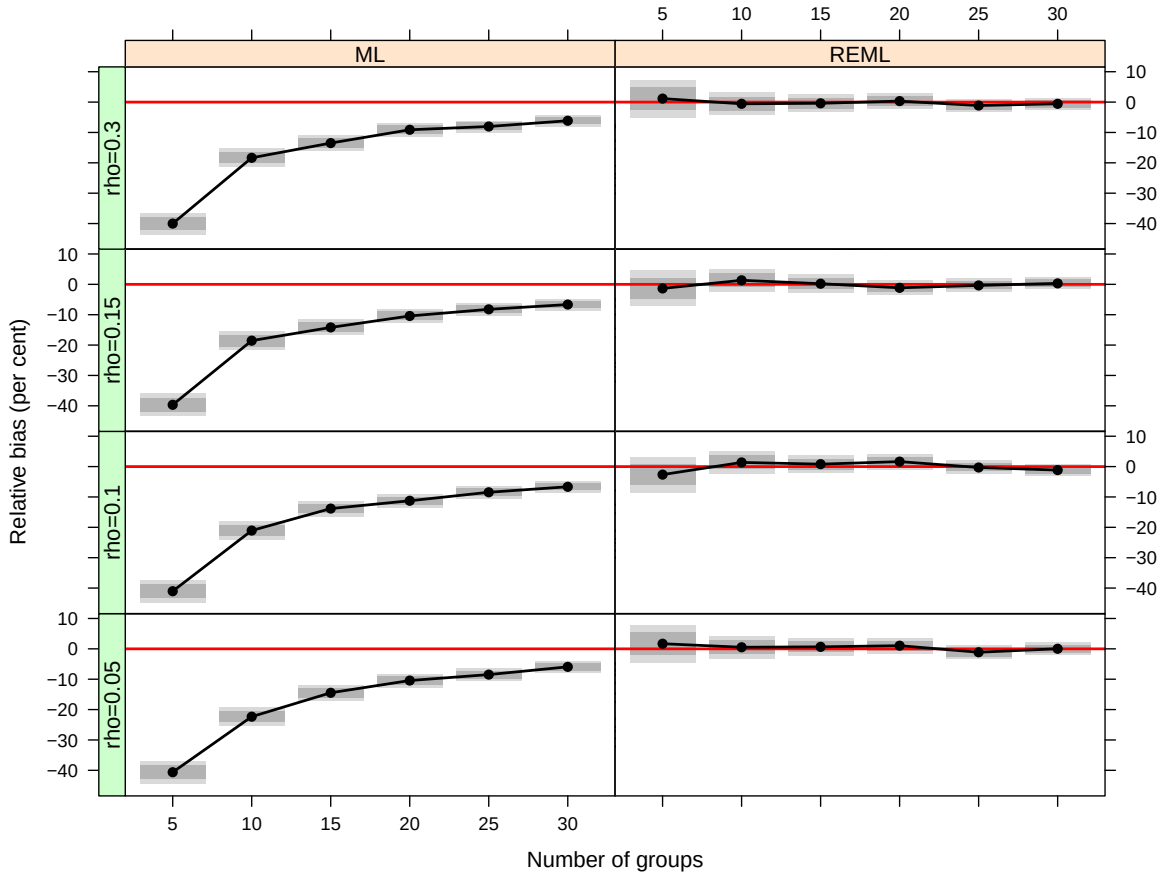

Figure 3: Relative simulated bias (in per cent) of the estimated variance of group-level random effects.

dom slopes of the predictor $\boldsymbol{x}_1$, so that it takes the form

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_{0j} + u_{1j} x_{1ij} + \epsilon_{ij}.$$

In this simulation experiment we vary the number of groups and the proportion of the random-effects variances relative to the residual variance (designated again as ρ, where $\theta_{11} = \theta_{22} = \rho\sigma^2$) and the number of groups. Yet we hold the fixed-effects coefficients constant at $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 1$. Further, we fix the correlation between the random slopes and random intercepts at $\theta_{12}/\sqrt{\theta_{11}\theta_{22}} = 0.05$. In a third experiment we add to the random-slope model just described a cross-level interaction, so that the resulting model takes the form

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{1ij} x_{2j} + u_{0j} + u_{1j} x_{1ij} + \epsilon_{ij}.$$

We vary the number of groups and the random-effects variances in exactly the same way as in the second simulation experiment and also fix the correlation between random slopes and random intercept at $\theta_{12}/\sqrt{\theta_{11}\theta_{22}} = 0.05$. The coefficients of the fixed effects coefficients are set to $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$, and $\beta_3 = 0.3$.

With respect to the estimates of the coefficients, our second and third simulation experiments do not lead to any different conclusions in terms of bias: any departure from zero by the average simulated bias stays within the bound of sampling error. Therefore, to save space and to avoid repetition, we do no show the the results from the second and third simulation experiments with respect to the simulated bias of ML and REML estimators of the coefficients of the individual-level predictor $\boldsymbol{x}_1$ and the group-level predictor $\boldsymbol{x}_2$. But since Stegmüller in particular claims that ML estimators "can be severely biased . . . especially in models including cross-level interactions" we do show in Figure 4 the simulation results with respect to the ML and REML estimators of the coefficient β_3 of the cross-level interaction term $\boldsymbol{x}_1 * \boldsymbol{x}_2$. As becomes obvious from this figure, not even the estimates of the coefficient representing cross-level interaction effects shows any systematic bias or any departures from the true value of the coefficients that cannot be attributed to the inevitable Monte Carlo error.

That notwithstanding, REML estimates for variance parameters when only 5 groups are present can be highly unstable. First, we incur requent non-convergence, and second, when the REML algorithm does converge the resulting estimates are highly volatile. This becomes apparent in Figure 5 that shows the simulated bias of the estimates of the variance of the random intercept in the second simulation experiment: The estimates now appear to be up-

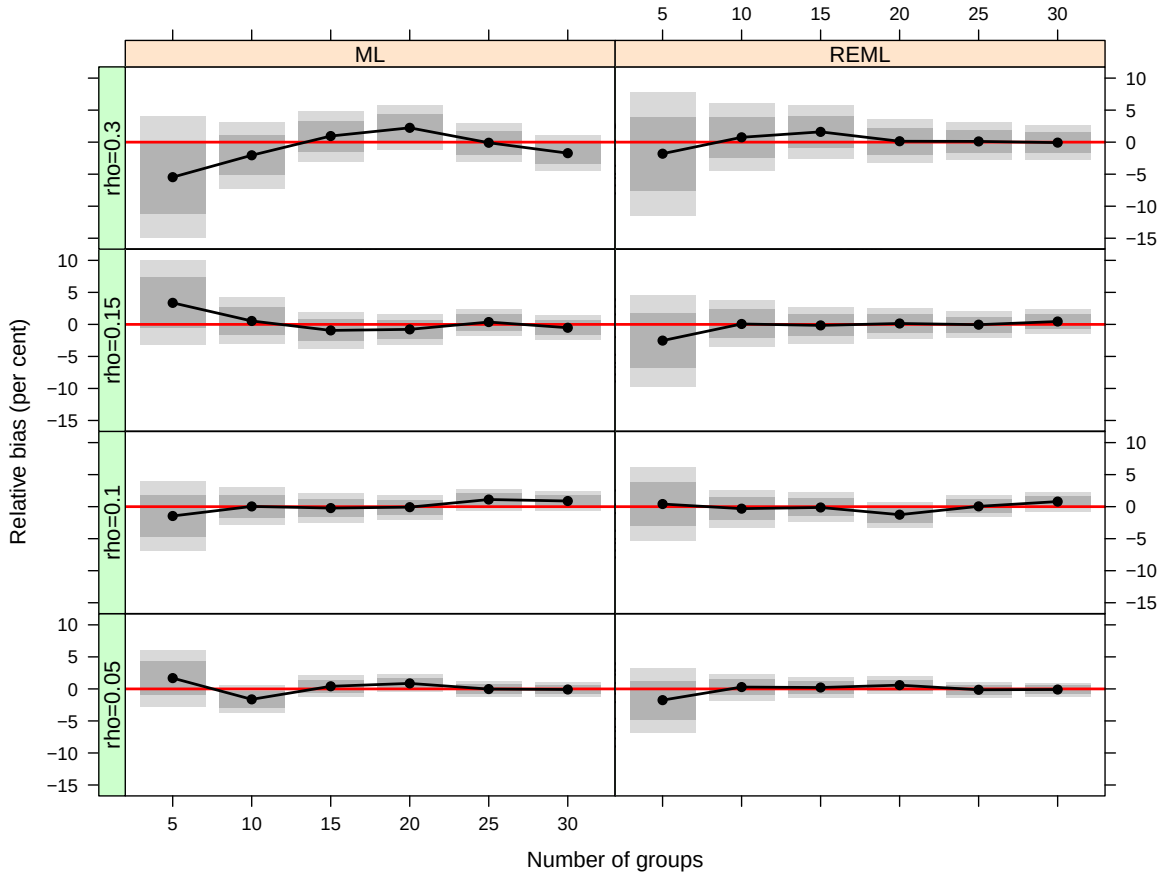

Figure 4: Relative simulated bias (in per cent) of the estimated (fixed effect) coefficient β_3 of the cross-level interaction term $\boldsymbol{x}_1 * \boldsymbol{x}_2$

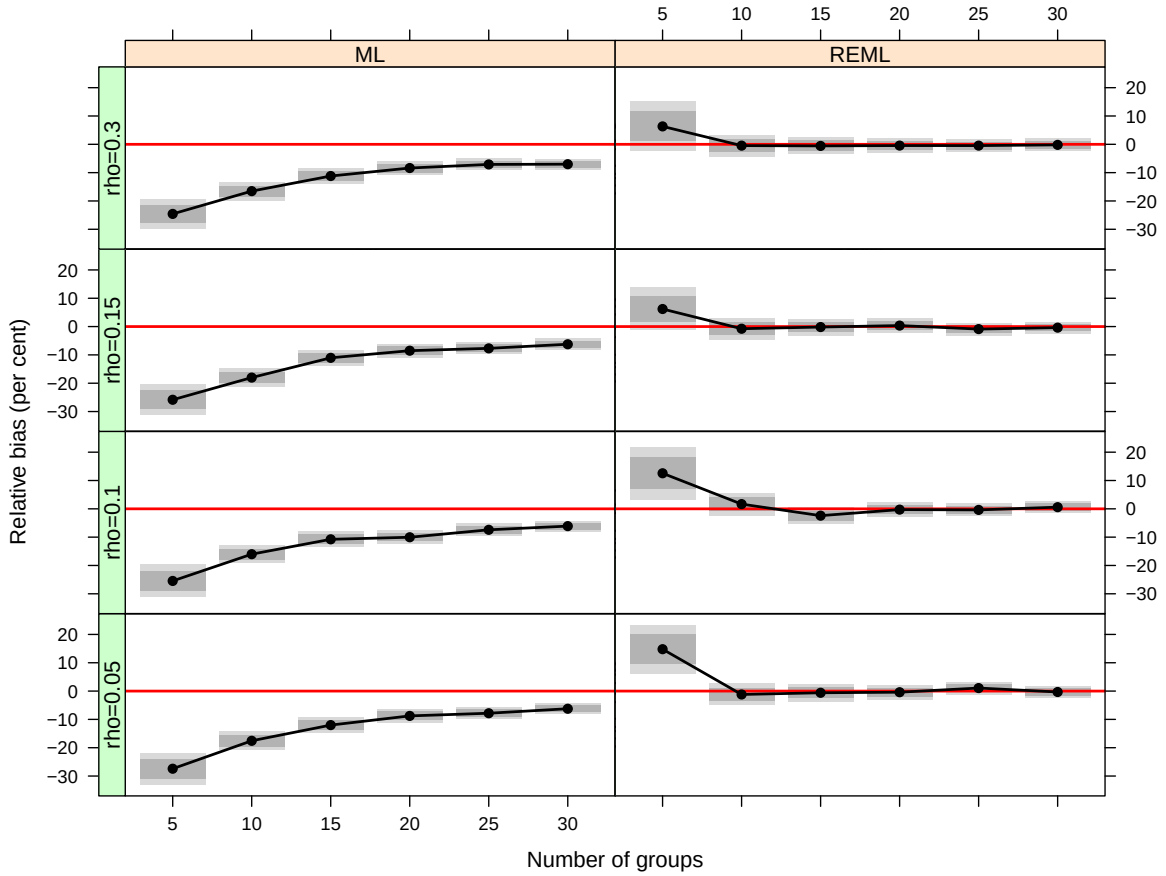

Figure 5: Relative simulated bias (in per cent) of the estimated variance of group-level random effects in a random-slope model.

wardly biased. However, this seems to be rather a consequence of a high volatility of the estimates due to non-convergence or near non-convergence. That REML struggles with a random slope model for 5 groups should not surprise, however. In this case there are effectively 5 degrees of freedom available for estimating three parameters in case of ML. REML now further corrects this by taking into account the loss of degrees of freedom incurred by the presence of fixed-effects predictors in the model. In effect this reduces the number of available degrees of freedom to zero.

In the theoretical discussion of interval estimates of fixed-effects coefficients, we indicated two sources of potential undercoverage of true coefficient values: First, estimated standard errors may be too small because variance parameter estimates are biased downward and, second, asymptotic normality may fail to apply to the sampling distribution of the parameters, thus invalidating the usual technique of constructing confidence intervals. In the following we present further results of our Monte Carlo study. Instead of discussing the simulated bias we discuss the simulated coverage performance of various interval estimates: normality-based confidence intervals with variance parameters estimated by ML and REML, confidence intervals based on a t-distribution with degrees of freedom obtained with the heuristic method implemented in the R package `nlme`, and confidence intervals based on a t-distribution with degrees of freedom obtained by the method of Kenward and Roger (1997).

Figure 6 shows the simulated coverage error of normality-based confidence intervals from ML and REML estimates of the coefficient β_1 of a predictor variable $\boldsymbol{x}_1$ that varies mainly between individuals. Quite obviously, normality-based confidence intervals show a satisfactory coverage performance. Irrespective of group size, intraclass correlation, or estimator, the true parameter value is covered in 95 per cent of the replications, barring sampling error. That is, the average coverage percentages do depart from 95 percent but these departures stay within the limits of random variation due to the finiteness of the simulated samples (again, the 95 per cent of the confidence intervals of the coverage errors include zero).

With regards to the coverage of the coefficient β_2 of a predictor $\boldsymbol{x}_2$ that is constant within groups and variant only between groups the finding is different, the finding is different. ML estimation and normality-based interval estimators show a serious amount of undercoverage, especially if the number of groups is small. With only 5 groups, the 95 percent interval estimate covers the true parameter values in just about 80 percent of the replications. The coverage performance increases as the number of groups increases, but persists even with

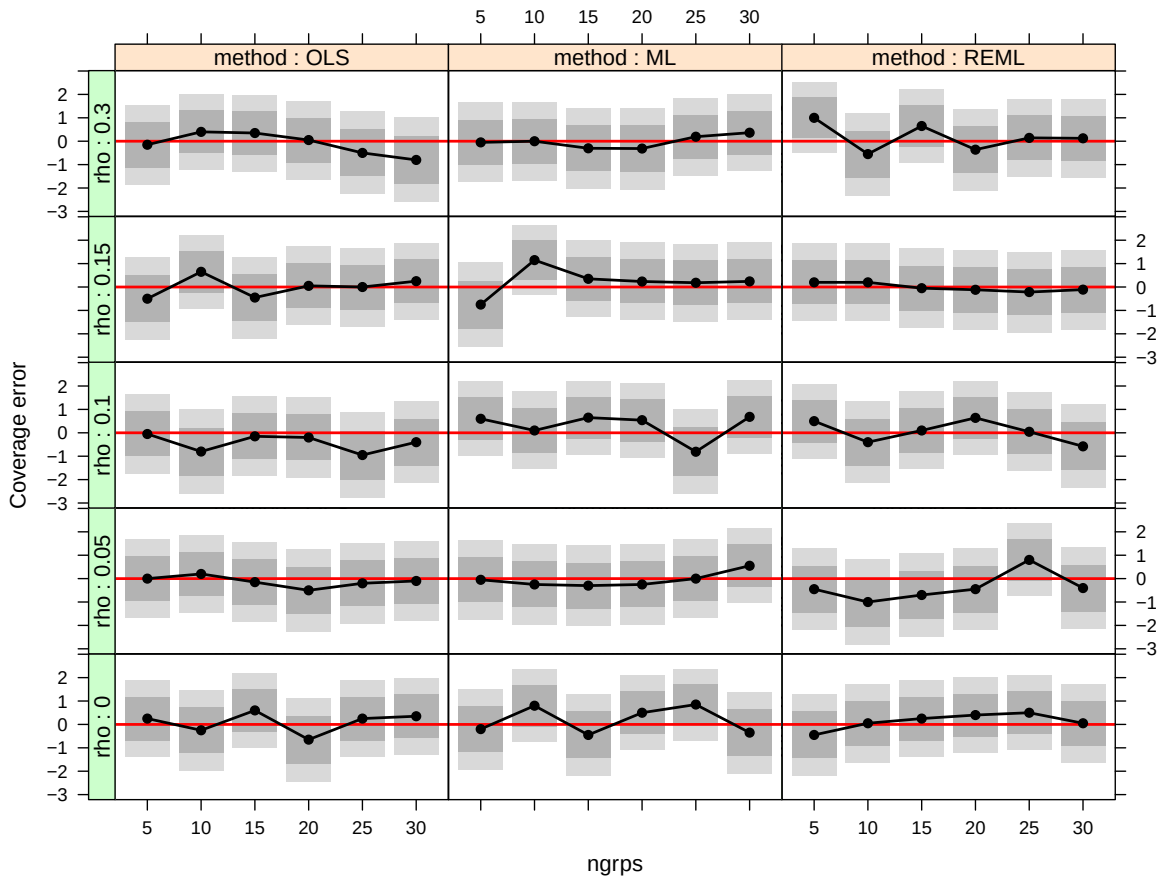

Figure 6: Simulated coverage error of normality-based nominal 95 per cent confidence intervals for coefficient β_1 of the predictor $\boldsymbol{x}_1$ that varies across individuals and across croups.

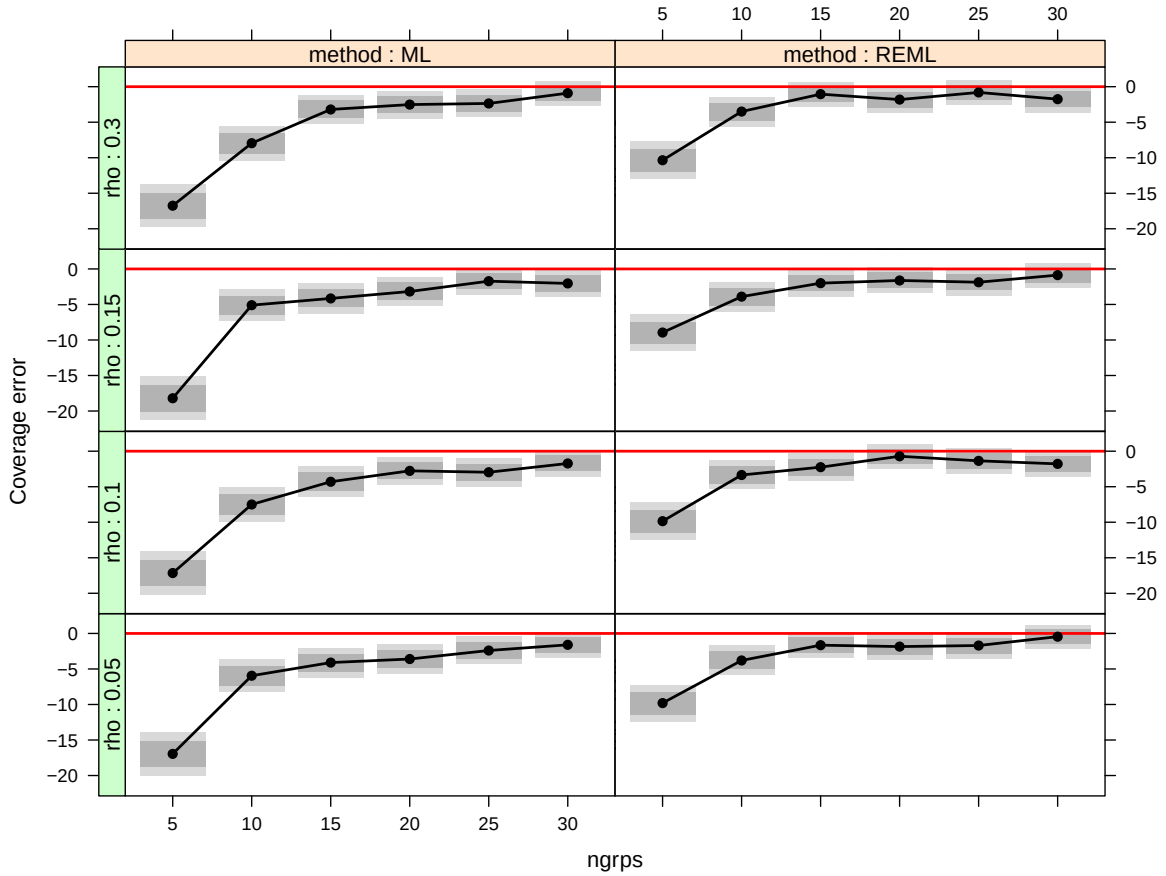

Figure 7: Simulated coverage error of normality-based nominal 95 per cent confidence intervals for coefficient β_2 of the predictor $\boldsymbol{x}_2$ that varies only across croups.

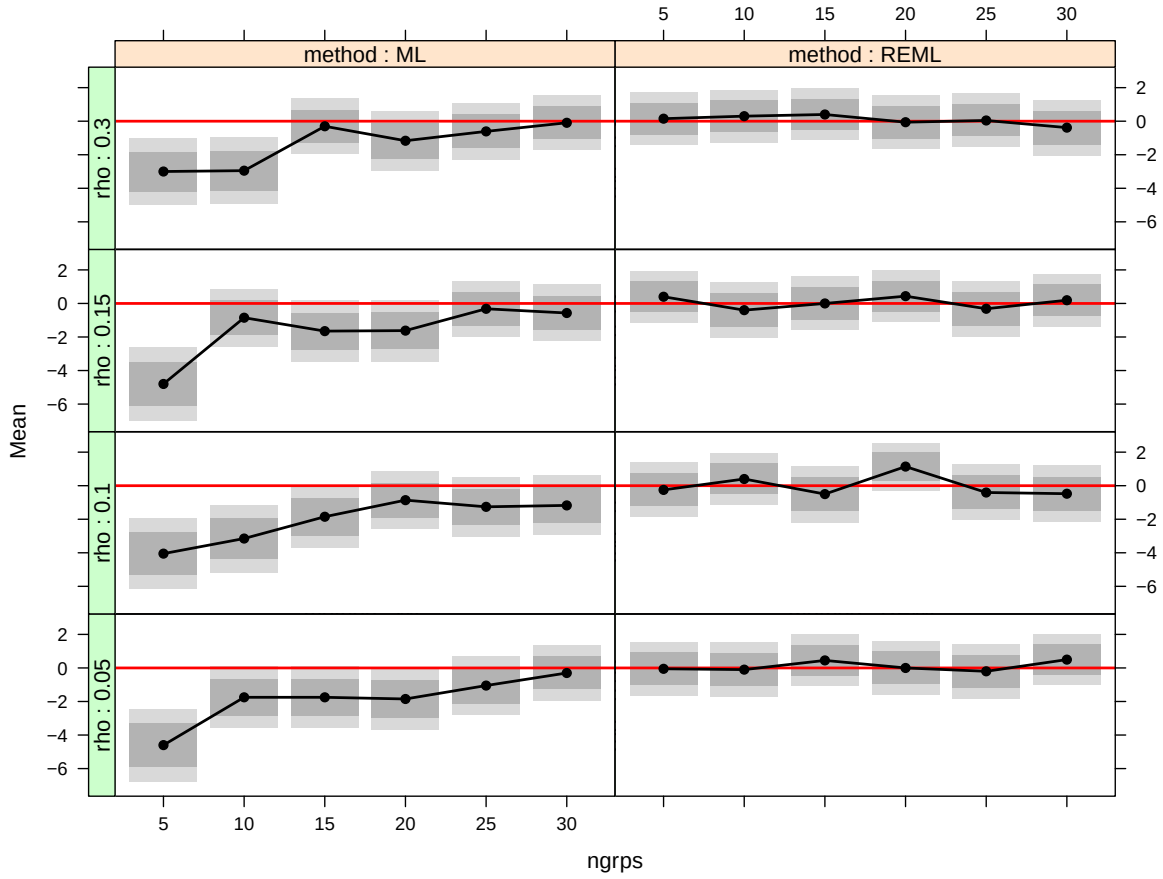

Figure 8: Simulated coverage error of nominal 95 per cent confidence intervals based on a Student's t-distribution for coefficient β_2 of the predictor $\boldsymbol{x}_2$ that varies only across croups, where the degrees of freedom are determined by the `nlme` heuristic.

30 groups. When REML estimation is used, the coverage error is only slighlty better, in the worst case of only 5 groups the undercoverage is about 10 percent instead of 15. That is, the abysmal coverage performance of the interval estimates of the coefficient of a group-level variable cannot be solely attributed to the downward bias of the estimate of the group-level variance.

Figure 8 shows the coverage performance of interval estimates based on a t-distribution with degrees of freedom determined using the heuristic method implemented in the R-package `nlme` (Pinheiro et al. 2013). This method proceeds as follows: if the covariate is constant within groups, then the number of groups minus the number of coefficients is chosen as the degrees of freedom, otherwise the total number of individual observation minus the number of coefficients is chosen as the degrees of freedom (just as usual for test statistics and

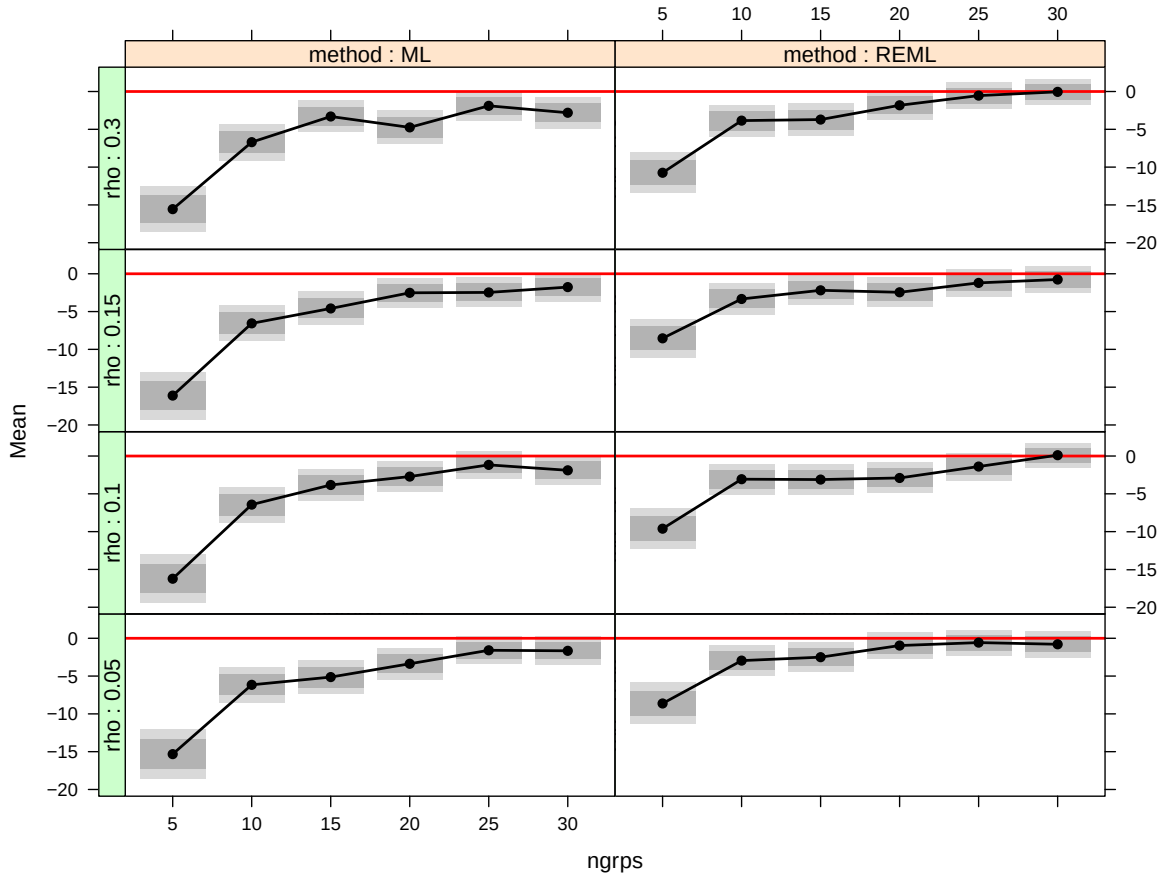

Figure 9: Simulated coverage error of nominal 95 per cent confidence intervals based on a Student's t-distribution for coefficient β_3 of the cross-level interaction term $\boldsymbol{x}_1 * \boldsymbol{x}_2$, where the degrees of freedom are determined by the `nlme` heuristic.

confidence intervals in linear regression estimated by OLS). As the left-hand panels in the diagram show, moving from a normal to a t-distribution does not lead to a satisfactory coverage of interval estimates based on ML point estimates if the number of groups is lower than 25. But the right-hand side panels indicate that nominally 95 percent interval estimates obtained from REML and a t-distributions attain, within the limits of inevitable Monte Carlo error, the correct coverage of 95 percent of the replications.

Unfortunately, the relative simple heuristic to determine approximately the correct degrees of freedom does not seem to work for coefficients of cross-level interaction terms. As Figure 9 shows, the interval estimates computed based on a t-distribution with degrees of freedom determined by the heuristic method show a considerable amount of undercoverage, even when they are based on REML. Since cross-level interaction terms are products of individual-

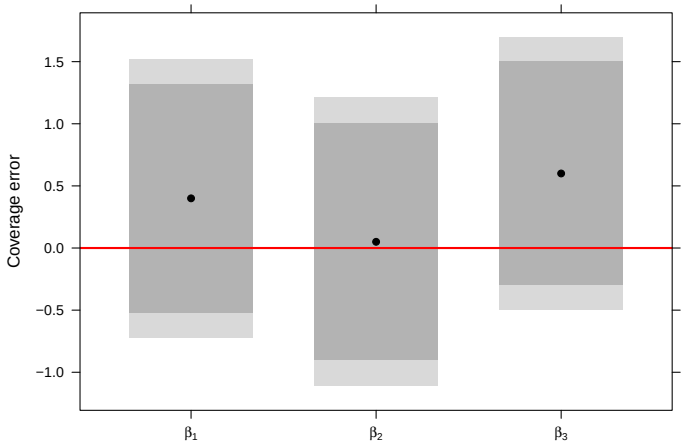

Figure 10: Simulated coverage error of nominal 95 per cent confidence intervals based on a Student's t-distribution for coefficient β_3 of the cross-level interaction term $\boldsymbol{x}_1 * \boldsymbol{x}_2$.

level covariates and group-level covariates, they vary within groups, so that the heuristic method of degree-of-freedom assignment uses the number of observations rather than the number of groups as point of departure.

Because the assignment of degrees of freedom based on the heuristic method does not seem to work, we repeated our simulation experiment, with estimating the models using the R-package `lme4` (Bates et al. 2014) instead of `nlme` and with interval estimates obtained from the package `pbkrtest` (Halekoh and Højsgaard 2013). Because replications with this setup were quite time-consuming[2], we restricted the simulation experiment to a single setting, with 10 groups and the ratio of random-intercept/random-slope variance and individual-level variance set to 0.3. The results of this reduced simulation study are shown in Figure 10.

Figure 10 depicts the coverage error of interval estimates for the three fixed-effects coefficients of a two-level model with a random-intercept, a random slope, and a cross-level interaction. While we found for the cross-level interaction coefficient a considerable degree of undercoverage if the heuristic method of determining the degrees of freedom of the t-distribution is used, we do not find such an undercoverage if the more rigorous Kenward-Rodger method is used. Instead of an undercoverage, the degree of coverage error is rather in

[2] 2000 replication with a single setting of the simulation study required several hours, even when run in 80 parallel tasks on the recently installed `bwUniCluster`.

the positive direction, the Kenward-Rogers interval estimates are very lightly conservative. However, this overcoverage is not statistically significant, the 95 percent confidence intervals of the coverage errors of all three coefficients contain zero.

To summarise the results of our Monte Carlo study so far: The theoretical argument implying the unbiasedness of fixed-effects coeffient estimates is borne out by simulation. There is hardly any coverage error in conventional interval estimates of fixed-effects coefficients of covariate that vary within groups, while interval estimates based on a t-distribution with the appropriately determined degrees of freedom also attain their nominal level of confidence.

In the following, we present a Monte Carlo study to examine the degree of bias of PQL and PQL/REML estimators in the case of large group sizes and small numbers of groups. Since our simulation results for the normal-linear case were essentially the same for a model with random intercepts only, for random-intercepts and random-slopes, and for random-intercepts, random-slopes, and cross-level interactions, we conduct our simulation study of a generalised linear mixed-effects model with random-intecepts only. If the adjustments discussed previously break down with the simplest variant of a generalised linear mixed-effects model, then they should break down also for more complex specifications. However, if they do work, we do not expect them to perform differently in more complicated setups.

Following Stegmüller (2013), we consider a two-level probit model with a predictor variable $\boldsymbol{x}_1$ that varies within and between groups, a predictor variable $\boldsymbol{x}_2$ that is constant within but varies between groups, and group-level random intercepts. The model is formulated as

$$\ln \frac{\mu_{ij}}{1-\mu_{ij}} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j \quad \text{where} \quad \mu_{ij} = \mathrm{E}(Y_{ij}) = \Pr(Y_{ij} = 1),$$

with $\beta_0 = \beta_1 = \beta_2 = 1$. Again we vary in our Monte Carlo study the number of groups between 5 and 25, and also the variance of the group-level random intercepts between 0.1 and 0.3. In each Monte Carlo replication, data were generated according to the random-intercept two-level probit model and fitted with a modified version of Venables and Ripley's function `glmmPQL` (Venables and Ripley 2002). The original version of this function does not support the REML modification of the PQL estimator discussed above, so we needed to slightly modify this function to allow for this. The code for this modified function, which we call `glmmPQL1` can be obtained from the first author of this paper.

Figure 11 shows the simulated bias of the fixed-effects coefficient of the individual-level co-

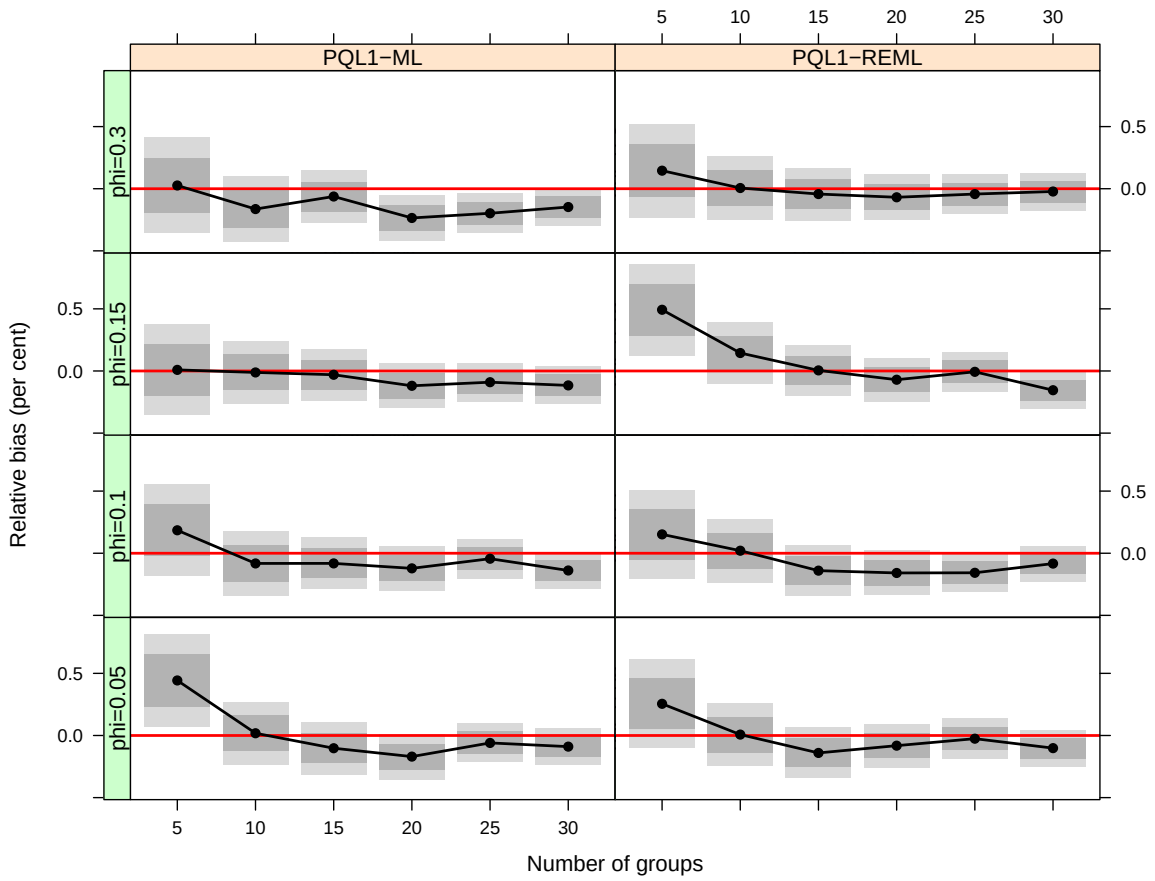

Figure 11: Relative simulated bias (in per cent) of the estimated (fixed effect) coefficient β_1 of the predictor $\boldsymbol{x}_1$ that varies across individuals and across croups; two-level random-intercept probit model.

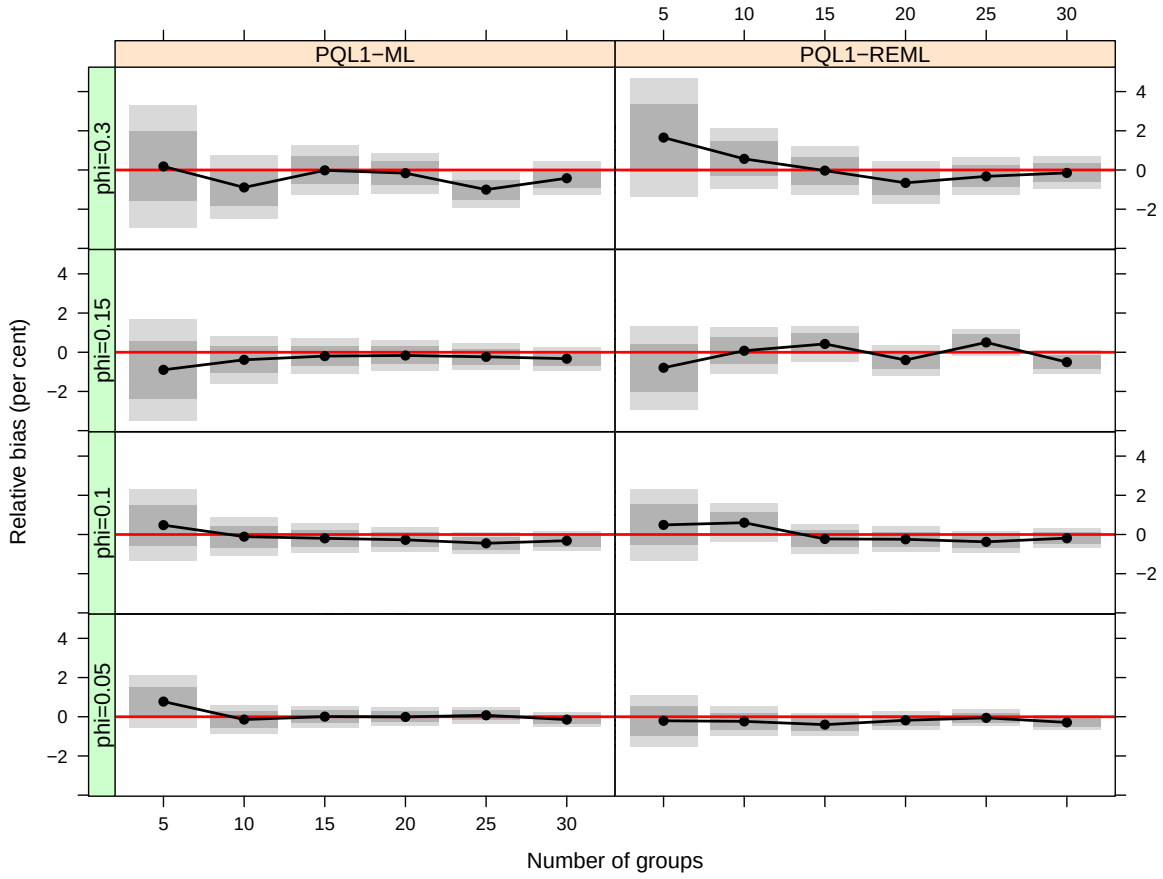

Figure 12: Relative simulated bias (in per cent) of the estimated (fixed effect) coefficient β_2 of the predictor $\boldsymbol{x}_2$ that varies only across croups and is constant within groups; two-level random-intercept probit model.

variate $\boldsymbol{x}_1$ with various group sizes and values of the random-intercept variance. Apart from a group size of 5, there does not seem to be any apparent bias that cannot be attributed to Monte Carlo sampling error, since all confidence intervals of the simulated bias include zero. For a group size of 5, the confidence intervals of the simulated bias does not include zero in two instances, yet even for a 95 percent confidence interval this could be expected to happen occasionally. Further even if we had to consider these departures as significant, they are hardly substantial in their relative size: They all seem to stay below one half of a percent.

As can be seen in Figure 12, the results with regards to the bias of the fixed-effects coefficient of the group-level covariate $\boldsymbol{x}_2$ in the two-level probit model are hardly different from those obtained for the two-level normal-linear model: Coefficient estimates show a substantial dispersion if the group sizes are small, yet on average they do not exhibit a bias, since the

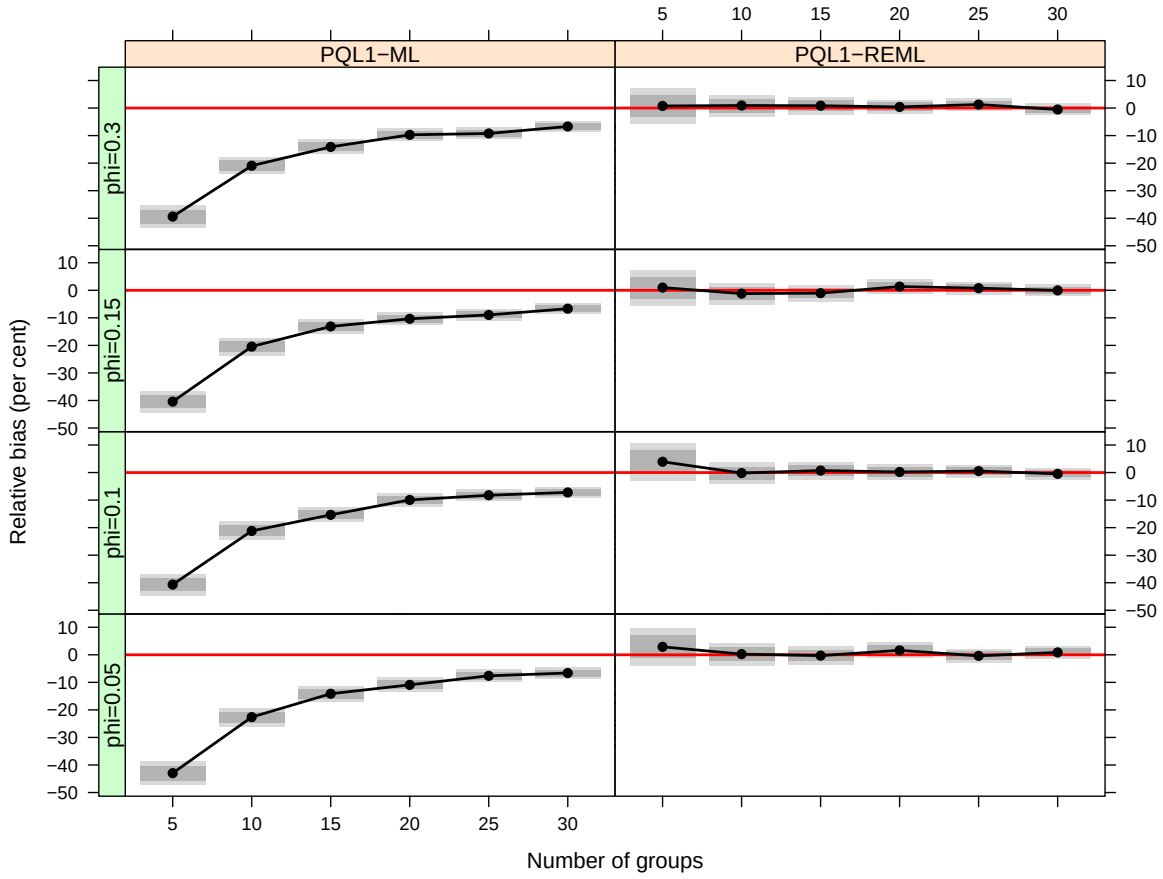

Figure 13: Relative simulated bias (in per cent) of the estimated variance of group-level random effects; two-level random-intercept probit model.

bias confidence intervals include zero throughout.

Also with regards to a bias of the estimates of the random-intercept variance we arrive at similar results to those obtained for the normal-linear case. If estimated by original PQL, the variance estimates show a clear downward bias of up to -40 percent. But if the REML variant is used instead of unmodified PQL, this bias seems to disappear, or rather, its 95 percent confidence interval includes zero in all settings of the number of group and size of the true random-intercept variance.

With regards to the interval estimates of fixed-effects coefficients, our Monte Carlo experiment does not lead to any substantially different results as with regards to the normal-linear mixed-effects model. Since it were the interval estimates of fixed-effects coefficients of group-level covariate of fixed-effects coefficients of group-level covariates that were particularly

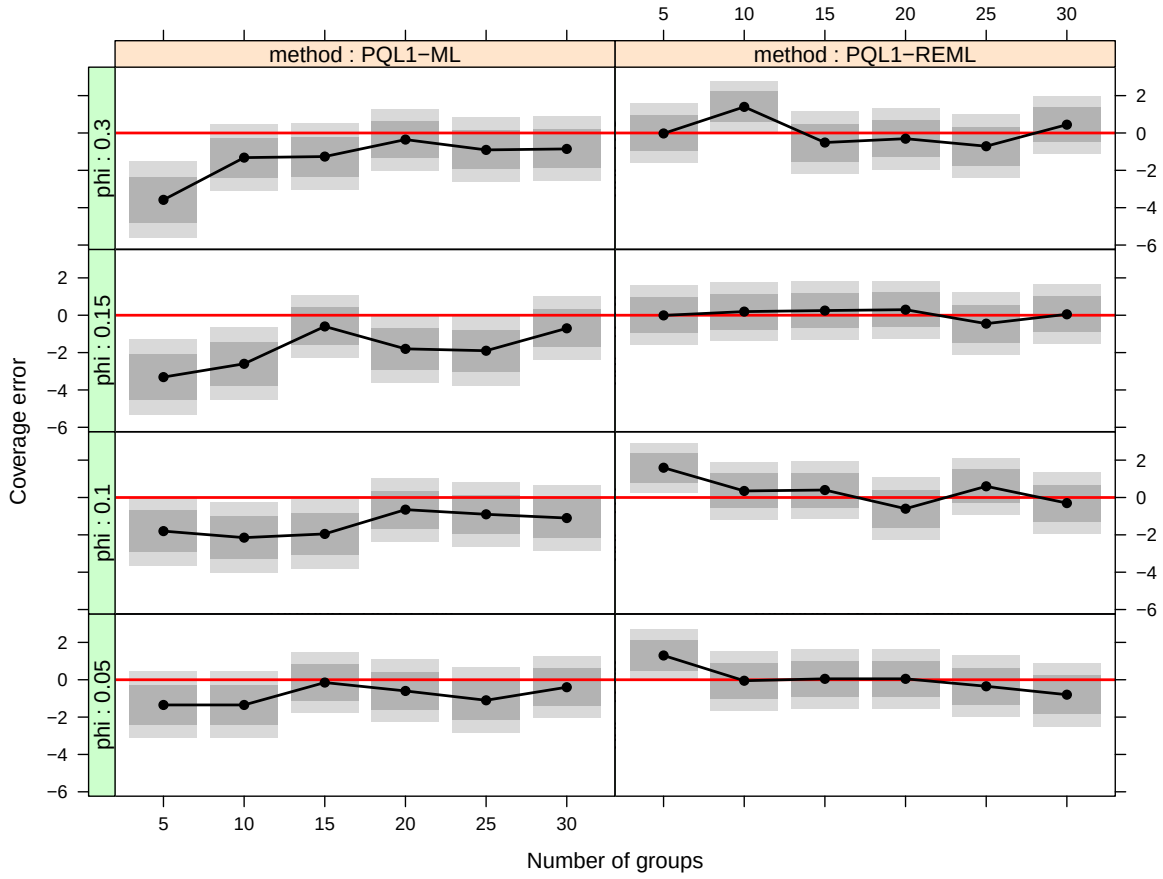

Figure 14: Simulated coverage error of nominal 95 per cent confidence intervals based on a Student's t-distribution for coefficient β_2 of the predictor $\boldsymbol{x}_2$ that varies only across croups, where the degrees of freedom are determined by the `nlme` heuristic; two-level random-intercept probit model.

affected by a bias in point estimates of random-intercept variances, we discuss only these, to save place.

We already saw in the previous section that interval estimates constructed on the assumption of normality of the sampling distribution of the estimates tend to be too short to attain their nominal coverage of the true parameters, whereas if interval estimates based on a t-distribution with the appropriate number of degrees of freedom cover the true parameter value roughly in at their nominal level if parameters are estimated via REML rather than ML. Figure 14 indicate that the same seems to apply to the case of two-level probit models with random intercepts, at least if the group sizes are large enough. As can be seen in the panels on the left half of the diagram, interval estimates based on unmodified PQL exhibit

a clear under-coverage of the true parameter value if the group sizes are small, a bias that tends to get smaller as the number of groups increases. This bias in coverage is substantial, the actual coverage of true parameter values occurs between 2 and 4 percent less often than to be expected from a nominal 95 confidence interval. But in case of PQL-REML no such undercoverage seems to occur. The simulated coverage error stays close to zero and the sole instances where confidence intervals of the coverage error do not include zero indicate overcoverage rather than undercoverage.

Our Monte Carlo study concerning mixed probit models essentially do not lead to substantially different results than the study of normal linear mixed models: At least if the group size is large enough (i.e. 500) there is no substantial bias in the estimates of the fixed-effects coefficients and the bias in the estimates of variance parameters can (almost) eliminated by estimating them with ML instead of REML. Further, if variance parameters are estimated without substantial bias and if a t-distribution with the appropriate number of degrees of is used to construct them, interval estimates attain (approximately) their nominal confidence level.

5 Differences between Bayesian and ML estimates

As discussed above ML and Bayesian estimates should differ only in their variance estimates which also affect the interval estimates of fixed effect parameters. However, their point estimates should not be affected. To show this we focus on the random intercept model above with $m = 5$ and $\rho = 0.1$. As in the Monte Carlo simulations in the previous section we generate 2000 random data sets from which we calculated ML estimates as well as Bayes posterior means based on an MCMC algorithm with Gibbs samplers, using the following priors:[3]

- $\phi \sim IG(0.0001, 0.0001)$
- $\phi \sim IG(0.1, 0.1)$
- $\sqrt{\phi} \sim Unif(0, 1000)$
- $\phi \sim Unif(0, 1000)$

[3]For the Gibbs samplers we used JAGS (Plummer 2013). For each simulated data three chains with different initial parameter values were run. In each chain posterior information was collected in 2000 iterations after 2000 "burn-in" iterations. Convergence checks indicated no evidence of non-convergence of three chains.

	$\phi \sim IG(0.001)$	$\phi \sim IG(0.1)$	$\sqrt{\phi} \sim Unif$	$\phi \sim Unif$	ML	REML
β_1	0.950	0.999	0.983	0.999	0.955	0.955
β_2	0.947	0.997	0.983	1.000	0.896	0.950
ϕ	0.946	0.957	0.939	0.831	0.714	0.913

Table 1: Coverage rate of 95% confidence/credible intervals of different estimates.

Figure 15 presents the direct comparison of the estimated (fixed effect) coefficient β_2 of the predictor $\boldsymbol{x}_2$. The left-side panels compare naive ML estimates and Bayes estimates with different priors. The right-side panels compare REML estimates and Bayes estimates. If we first look at the point estimates (dots in Figure) they are most identical among ML/REML and Bayesian estimates. That is, if ML/REML is biased, Bayesian estimates should be also biased.[4] Only the Bayesian estimates with the uniform prior on ϕ are instable in comparison with the other estimates (both panes at the bottom). This is clearly because the boosted posterior of ϕ by the prior specification. This is confirmed by the corresponding interval estimates which show much longer credible intervals (vertical lines in the panels at the bottom).

If one checks the coverage rate of the parameter values it is clear to see that REML estimates and Bayesian ones with inverse distribution with 0.001 outperforms the other estimates. Both estimates have coverage rates very close to 95%. Only at the estimate of ϕ the REML shows a undercoverage. In contrast, the other estimates show either clear overcoverage (the other Bayesian estimates) or undercoverage (naive ML).[5]

From these simulation results, we confirmed the following: First, the point estimates are identical among the ML and Bayesian estimation so long uninformed priors are specified. Second, the REML and Bayesian estimation with uninformative prior perform well to a similar degree. Third, a Bayesian posterior can be quite sensitive to prior choice. In particular, the uniform prior for the variance parameter perform much worse than naive ML even though the prior looks uninformative. One might object that this result is based on a random intercept model with specific parameter value ($m = 5$ and $\rho = 0.1$). Using the same model specification, Stegmüller (2013) found partly large differences between ML and Bayesian es-

[4]To be precise, the term "bias" is in the Bayesian context not correct since it assumes no true parameter value from which estimates are "biased". Gelman and others (2006), e.g., uses the term miscalibration as difference between posterior mean and the true parameter mean.

[5]One might wonder that the uniform prior for ϕ has overcoverage for both estimated β's and undercoverage for estimated ϕ. This is because the posterior of ϕ has too large expected value therefore its credible interval is often larger than the true value (undercoverage). It leads to the larger dispersions of the posterior of β's which in turn result in overcoverage.

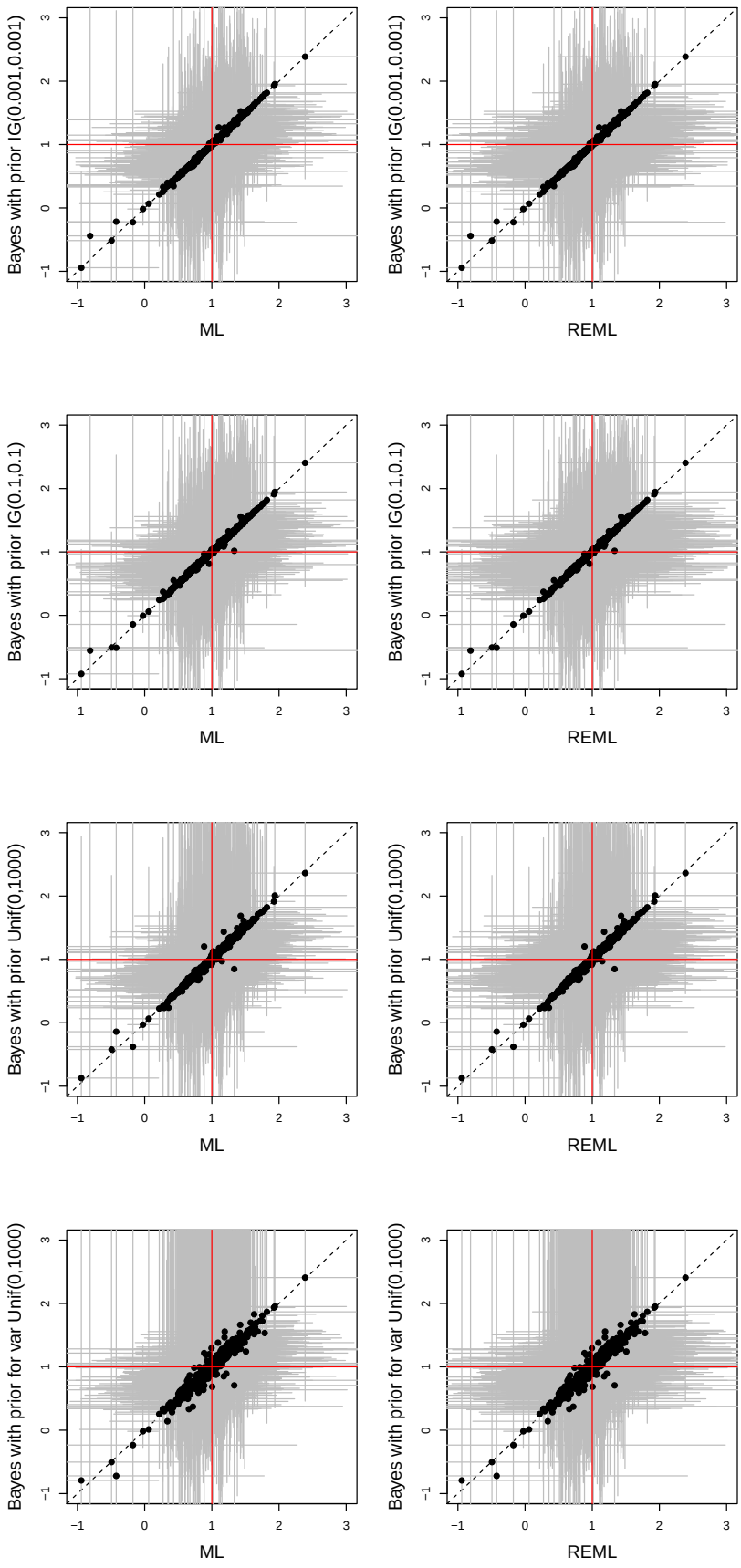

Figure 15: Direct comparison of estimated (fixed effect) coefficient β_2 of the predictor x_2 based on ML and Bayesian; two-level random intercept linear model. The dots are points estimate. The grey lines are 95% confidence/credible interval. The red lines are drawn on the true parameter value.

timates, which however could not be confirmed by the simulation study here.

6 Conclusion

In the present paper we have shown, both theoretically and by way of a Monte Carlo study, that frequentist estimators of coefficients in normal linear multilevel models are unbiased irrespective of the number of higher-level units (e.g. countries in case of cross-national comparative studies), and irrespective of whether maximum likelihood, restricted maximum likelihood and even (in the present context) crude OLS estimators are used. Secondly, we have shown that confidence intervals, if appropriately constructed, do not exhibit any serious undercoverage. Thirdly, we have found that the result to multilevel probit models are hardly different to normal linear multilevel models even if a relatively simple Laplace approximation is used. Fourth we have found that Bayesian point estimates of coefficient estimates in the Monte Carlo study hardly differ from ML estimates, while we do find that the choice of the prior distribution of a variance parameters has consequences for the coverage of true coefficient values by Bayesian credibility intervals and that these consequences are not always beneficial. That is, "the Bayesian approach" is not "far more robust" (Stegmüller 2013) but rather quite sensitive, namely to the choice of the prior distribution.

The question now arises why Stegmüller (2013) finds a bias in the point estimates for coefficients in multilevel models while we do not. The major reason for this is that he does not account for Monte Carlo sampling error. Whereas we do find (mostly minor) departures from zero of the simulated bias, we are able to attribute these departures to random fluctuations, as they are inevitable in any Monte Carlo study. Therefore we conclude that reporting of Monte Carlo results without taking into account their variance is misleading. The second question is why Stegmüller finds a serious undercoverage of frequentist confidence intervals while we do not. Again part of the question is the failure to account for Monte Carlo variance. Yet most importantly, Stegmüller's study ignores REML as a frequentist estimator and the more improved confidence intervals based on t-distributions as proposed by Kenward and Roger (1997). Thus when Stegmüller (2013) conveys the impression that frequentist estimation and inference in multilevel models is flawed, we have to disagree. And if the other impression is that the Bayesian approach is superior, we also have to disagree. Rather we would argue that one does not need to become a Bayesian to address the limits of maximum likelihood.

We should not forget to point to a limitation of our Monte Carlo study as well as Stegmüller's. Throughout the discussion the size of the higher-level units (e.g. country samples) is relatively large. While smaller group sizes are unlikely to lead to a bias in coefficient estimates of normal linear multilevel models, a small group size may have consequences for the performance of PQL and PQL-REML estimates. But for this problems there may again be remedies within the frequentist framework. One is Monte Carlo integration as an approximation of the log-likelihood function (McCulloch 1997; Booth and Hobert 1999; Caffo et al. 2005), another one is using a higher-order Laplace approximation (Breslow and Lin 1995; Lin and Breslow 1996). Finally, one can use parametric bootstrap or iterated parametric bootstrap bias correction methods (e.g. Kuk 1995).

References

Bates, Douglas, Martin Maechler, Ben Bolker and Steven Walker, 2014. *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.0-6.
URL `http://CRAN.R-project.org/package=lme4`

Booth, James G and James P Hobert, 1999. "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61(1): 265–285.

Breslow, N. E. and D. G. Clayton, 1993. "Approximate Inference in Generalized Linear Mixed Models". *Journal of the American Statistical Association* 88(421): 9–25.

Breslow, Norman E. and Xihong Lin, 1995. "Bias correction in generalised linear mixed models with a single component of dispersion". *Biometrika* 82(1): 81–91.

Caffo, Brian S., Wolfgang Jank and Galin L. Jones, 2005. "Ascent-based Monte Carlo expectation-maximization". *Journal of the Royal Statistical Society. Series B (Methodological)* 67: 235–251.

Carlin, Bradley P. and Thomas A. Louis, 1996. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, London.

Casella, George, 1985. "An Introduction to Empirical Bayes Data Analysis". *American Statistician* 39(2): 83–87.

Cox, D. R. and N. Reid, 1987. "Parameter Orthogonality and Approximate Conditional Inference". *Journal of the Royal Statistical Society. Series B (Methodological)* 49(1): 1–39.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1): 1–38.

Draper, David, 1995. "Inference and hierarchical modeling in the social sciences". *Journal of Educational and Behavioral Statistics* 20(2): 115–47.

Elff, Martin, 2014 [forthcoming]. "Estimation Techniques: OLS and MLE". In Best, Henning and Christoph Wolf (Eds.), "Regression Analysis and Causal Inference", London: Sage, pp. 9–32.

Firth, David, 1993. "Bias reduction of maximum likelihood estimates". *Biometrika* 80(1): 27–38.

Gelman, Andrew and others, 2006. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". *Bayesian analysis* 1(3): 515–534.

Halekoh, Ulrich and Søren Højsgaard, 2013. *pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison.* R package version 0.3-8.
URL `http://CRAN.R-project.org/package=pbkrtest`

Harville, David, 1976. "Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects". *The Annals of Statistics* 4(2): 384–395.

Harville, David A, 1997. *Matrix Algebra From a Statistician's Perspective.* New York: Springer.

Jiang, Jiming, 1999. "On unbiasedness of the empirical BLUE and BLUP". *Statistics & Probability Letters* 41(1): 19–24.

Jiang, Jiming, 2007. *Linear and Generalized Linear Mixed Models and Their Applications.* New York: Springer.

Kenward, Michael G. and James H. Roger, 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood". *Biometrics* 53(3): 983–997.

Kuk, Anthony Y. C., 1995. "Asymptotically Unbiased Estimation in Generalized Linear Models with Random Effects". *Journal of the Royal Statistical Society. Series B (Methodological)* 57(2): 395–407.

Lin, Xihong and Norman E. Breslow, 1996. "Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion". *Journal of the American Statistical Association* 91(435): 1007–1016.

Manor, Orly and David M Zucker, 2004. "Small sample inference for the fixed effects in the mixed linear model". *Computational Statistics & Data Analysis* 46(4): 801–817.

McCullagh, P. and J. A. Nelder, 1989. *Generalized Linear Models. Second Edition*. London, New York: Chapman and Hall.

McCullagh, Peter and Robert Tibshirani, 1990a. "A Simple Method for the Adjustment of Profile Likelihoods". *Journal of the Royal Statistical Society. Series B (Methodological)* 52(2): 325–344.

McCullagh, Peter and Robert Tibshirani, 1990b. "A Simple Method for the Adjustment of Profile Likelihoods". *Journal of the Royal Statistical Society. Series B (Methodological)* 52(2): 325–344.

McCulloch, Charles E., 1997. "Maximum Likelihood Algortithms for General Linear Mixed Models". *Journal of the American Statistical Association* 92(437): 162–170.

McLachlan, Geoffrey and Thriyambakam Krishnan, 2008. *The EM Algorithm and Extensions*. Hoboken, NJ: John Wiley & Sons, 2 ed..

Patterson, H. D. and R. Thompson, 1971. "Recovery of inter-block information when block sizes are unequal". *Biometrika* 58(3): 545–554.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar and R Core Team, 2013. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-111.

Plummer, Martyn, 2013. *JAGS: Just Another Gibbs Sampler*. Version 3.4.0.
URL `http://mcmc-jags.sourceforge.net/`

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org/`

Rubin, Donald B., 1981. "Estimation in parallel randomiozed experiments". *Journal of Educational Statistics* 6(4): 377–400.

Satterthwaite, Franklin E., 1941. "Synthesis of variance". *Psychometrika* 6(5): 309–316.

Seltzer, Michael H., WIng Hung Wong and Anthony S. Bryk, 1996. "Bayesian analysis in applications of hierarchical models: issues and methods". *Journal of Educational and Behavioral Statistics* 21(2): 131–67.

Stegmüller, Daniel, 2013. "How Many Countries for Multilevel Modeling? A Comparison of

Frequentist and Bayesian Approaches". *American Journal of Political Science* 57(3): 748–761.

Venables, W. N. and B. D. Ripley, 2002. *Modern Applied Statistics with S.* New York: Springer, fourth ed.. ISBN 0-387-95457-0.
URL `http://www.stats.ox.ac.uk/pub/MASS4`

Šidák, Zbyněk, 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". *Journal of the American Statistical Association* 62(318): 626–633.