Improved random effects prediction *

Ruggero Bellio and Paolo Vidoni

Department of Economics and Statistics, University of Udine

via Tomadini 30/a, I-33100 Udine, Italy.

ruggero.bellio@uniud.it paolo.vidoni@uniud.it

*Running title: Improved random effects prediction

Abstract

This paper focuses on the prediction of parametric functions of random effects in generalized linear mixed models, for settings with independent groups. Borrowing from recent results on frequentist prediction, a methodology to obtain accurate prediction is introduced. The proposal is defined by conditioning on the observed value of the response for the relevant group. The resulting procedure has a simple form, and can be applied both for obtaining accurate prediction limits as well as the entire predictive distribution. Prediction intervals are provided for commonly used models, such as linear mixed models and logistic regression with random intercepts. Analytical results as well as simulation results support the good properties of the methodology, which is also illustrated by some numerical examples.

Keywords: Generalized linear mixed model; Parametric bootstrap; Prediction interval; Predictive distribution; Random effect.

1 Introduction

Mixed effect models are a widely used class of statistical models, and prediction of random effects is one of the most fundamental usage of such models; see, for example, Jiang (2007, §2.3 and §3.6) and McCulloch et al. (2008). Indeed, early applications of mixed models readily attempted to address the issue, as vividly illustrated in the discussion paper by Robinson (1991).

Here we briefly summarize the main results about the literature on random effects prediction. For linear mixed models, and under the Gaussian assumption for both the random effects and the residual error, the frequentist approach to random effects prediction employs the empirical BLUP. Namely, the conditional mean of the random effects given the observed data is computed, with model parameters replaced by corresponding estimates. When the assumptions of normality or linearity do not hold, the tendency is to mirror what done in the linear normal case. This amounts to employing as point predictor a summary of the conditional distribution of the random effects given the data, such as the mean or the mode,

as described by Skrondaal and Rabe-Hesketh (2009). Prediction intervals almost always consist of Wald-type intervals, given by the point predictor plus or minus the standard error of prediction multiplied by a suitable constant. The standard error of prediction is given by the square root of some estimate of the mean squared error of prediction. As reported in Skrondaal and Rabe-Hesketh (2009), a common choice in mixed models is given by the variance of the conditional distribution of the random effects given the data, with model parameters replaced by estimated values. For the special case of normal linear mixed models it is relatively straightforward to incorporate the variability of fixed effects into the standard error of prediction. Even in such favourable setting, however, accounting for the variability of variance parameters in a simple fashion is virtually impossible. In short, two are the main pitfalls of the aforementioned frequentist techniques for random effects prediction: (i) Symmetric prediction intervals are often employed, even if the conditional distribution of the random effects given the observed data need not be symmetric; (ii) Variability of model parameters is typically neglected, even if in some cases it might be substantial, especially in the case of variance parameters.

The paper by Booth and Hobert (1998) set the tone for much of the recent research on the topic. Though they did not departed from symmetric prediction intervals, they stressed the importance on basing prediction standard errors computation on *Conditional Mean Squared Error of Prediction* (CMSEP), as opposed to *Unconditional Mean Squared Error of Prediction* (UMSEP), which is sometimes used for normal linear mixed models. At the same time, they emphasized the need to adjust for both bias and variability of naive point predictions, employing a suitable bootstrap methodology. Their results were influential, with echoes also in the literature on small-area estimation (Lohr and Rao, 2009; Datta and Gosh, 2012). Booth and Hobert (1998), Subsection 1.5, also included a reference to the research on predictive likelihoods, stressing its importance. Quoting their writing "These predictive likelihood formulas allow for non-Bayesian inference about the entire distribution of an unobserved random variable". This is of course a commendable objective, but they went on claiming "(...) we are not aware of any simple, non-Bayesian methodologies that are directly applicable to the general problem addressed in this article", thus acknowledging that such objective was not achieved at the time of writing.

The aim of this paper is to illustrate how recent advances in techniques for frequentist prediction make possible to obtain accurate interval prediction of scalar functions of parameter models and random effects in mixed models. Results are provided for a general class of latent variable models (e.g. Skrondal and Rabe-Hesketh, 2004) for independent groups, with special emphasis on Generalized Linear Mixed Models (GLMMs). We endorse a conditional approach, along the lines of Booth and Hobert (1998). The method presented here provides a predictive distribution, thus allowing for non-Bayesian inference about the entire distribution of the target random variable. This fulfils exactly the task ruled out by Booth and Hobert in the latter of the above quotations, but with the noteworthy distinction that the methodology employed in the current paper is not based on predictive likelihood, but rather on high-order prediction based on asymptotic methods and the bootstrap, as proposed in Vidoni (1998), Ueki and Fueda (2007) and Fonseca et al. (2012).

The plan of the paper is as follows. Section 2 provides some background results on mixed models and modern prediction methods. Section 3 focuses on random effects prediction, studying the order of error terms arising in estimative approaches and quantifying the improvement achieved by higher-order methods. Sections 4 and 5 apply the methodology to some commonly used models, providing numerical support to the theory of the previous sections. Section 6 concludes the paper with a discussion, whereas some technical results are reported in Appendices A and B. Some computational details are given in Appendix C.

2 Background

2.1 Model and notation

The general setting of interest here is as follows. Assume that the available response data are arranged in $k \geq 1$ groups, related to specific clusters or subjects, and that the random variable Y_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, describes the response of unit j in the i-th group, having dimension $n_i \geq 1$. The m-dimensional continuous random vector $U_i = (U_{i1}, \ldots, U_{im})^T$, $i = 1, \ldots, k$, specifies the unobservable random effects associated with the i-th group. Suppose that $U_1, \ldots, U_k$ are independent, identically distributed, with marginal density $q(u_i; \gamma)$,

$i = 1, \ldots, k$, being $\gamma \in \mathcal{G} \subseteq \mathbf{R}^p$, $p \geq 1$, an unknown p-dimensional parameter, and that the pairs (Y_i, U_i), $i = 1, \ldots, k$, with $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$, are independent. Moreover, conditional on $U_i = u_i$, the responses in the i-th group are independent, having conditional density, with respect to a suitable dominating measure, $p(y_{ij}|u_i; \delta)$, $j = 1, \ldots, n_i$, with $\delta \in \mathcal{D} \subseteq \mathbf{R}^{q+1}$, $q \geq 0$, an unknown $(q+1)$-dimensional parameter.

According to this model, prediction of random effects U_i, related to the i-th group, $i = 1, \ldots, k$, should be based on the conditional density of U_i given the data $y = (y_1, \ldots, y_k)^T$. Due to the independence assumption, this reduces to

$$f(u_i|y_i; \omega) = \frac{q(u_i; \gamma) \prod_{j=1}^{n_i} p(y_{ij}|u_i; \delta)}{L_i(\omega; y_i)}, \tag{1}$$

where $\omega = (\gamma^T, \delta^T)^T$ is the model parameter and

$$L_i(\omega; y_i) = \int_{\mathcal{U}} q(u_i; \gamma) \prod_{j=1}^{n_i} p(y_{ij}|u_i; \delta) du_i \tag{2}$$

is the i-th likelihood component, being $\mathcal{U}$ the support of U_i. The quantity $L_i(\omega; y_i)$ is a normalizing term and with the exception of some special cases, mainly related to the normal distribution, it is not known explicitly. Thus, the calculation of $L_i(\omega; y_i)$ requires suitable numeric or approximation-based techniques. Since the groups are independent, the likelihood function is given by $L(\omega; y) = \prod_{i=1}^{k} L_i(\omega; y_i)$.

Predicted random effects are usually considered for inference concerning suitable one-dimensional transformations $R = R(U_i, \omega)$ of vector U_i, with regard to the i-th group, $i = 1, \ldots, k$, such as the conditional mean of the response Y_{ij} given $U_i = u_i$.

Let us consider the one-to-one transformation $H : \mathbf{R}^m \to \mathbf{R}^m$, defined as $H(u_i) = (R(u_i, \omega), u_{i2}, \ldots, u_{im}) = (h_1, h_2, \ldots, h_m) = h$. If the components are continuous and differentiable functions, using basic probability we get from (1) the conditional density of R given $Y_i = y_i$, namely

$$f(r|y_i; \omega) = \frac{\int_{\mathcal{H}} q(H^{-1}(h); \gamma) \prod_{j=1}^{n_i} p(y_{ij}|H^{-1}(h); \delta)|J(h)| dh_2 \cdots dh_m}{L_i(\omega; y_i)}, \tag{3}$$

with $|J(h)| = |\sum_{k=1}^{m} \partial R^{-1}(h)/\partial h_k|$, $\mathcal{H}$ the support of the transformed random vector $H(U_i)$ and $H^{-1}(\cdot)$ the inverse of function $H(\cdot)$, with first component $R^{-1}(\cdot)$. Whenever the random effect U_i follows a normal distribution and R is a suitable function of a linear transformation of U_i, the numerator in (3) takes a simple explicit form, as shown in Section 4.

In this paper we discuss prediction of one-dimensional reductions R, such as the expected responses. Our approach involves, as predictive distribution, a suitable estimator of the conditional distribution of R given the observation y_i. The target are computationally tractable prediction intervals for R, with both conditional and unconditional coverage probability close to the nominal value. We note in passing that the prediction of R is logically distinct from the problem of predicting a new observation of the response variable for a subject, which was considered in Vidoni (2006). Although there are some common aspects, the prediction problem for a new response is simpler, and the results in Vidoni (2006) suggest that in such case the need to adjust standard estimative procedures is less pronounced.

2.2 Review of improved predictive procedures

Suppose that the observable random vector $Y = (Y_1, \ldots, Y_k)^T$ and the future or unobservable random variable R follow a joint distribution depending on an unknown parameter $\omega = (\omega_1, \ldots, \omega_d)^T$, $d \geq 1$. In the following, $\hat{\omega} = \hat{\omega}(Y)$ denotes the maximum likelihood estimator for ω, or an asymptotically equivalent alternative estimator. Whenever Y and R are conditionally independent given a transitive statistics $T = T(Y)$ (see Barndorff-Nielsen and Cox, 1996), we consider for prediction purposes the conditional distribution of R given $T = t$, with density and distribution functions given by $f(r|t;\omega)$ and $F(r|t;\omega)$, respectively.

An α-prediction interval for R or, in particular, an (upper) α-prediction limit $l_\alpha(y)$ is such that, exactly or approximately,

$$P_{Y,R|T}\{R \leq l_\alpha(Y)|T = t\} = \alpha, \tag{4}$$

for all ω, where the target value $\alpha \in (0, 1)$ is fixed. The above probability is called conditional coverage probability and it refers to the conditional distribution of (Y, R) given $T = t$. Thus, prediction evaluation is done conditionally on the observed value of T. Furthermore, it is important to stress that a conditional solution to (4) has an unconditional coverage probability $P_{Y,R}\{R \leq l_\alpha(Y)\}$ equal to α as well.

Since exact solutions to (4) can be found only in special cases, we shall consider, as a simple approximate solution, the estimative or plug-in prediction limit. That is, if $r_\alpha(\omega, t)$ denotes the α-quantile of the conditional distribution of R given $T = t$, namely $r_\alpha(\omega, t) =$

$F^{-1}(\alpha|t;\omega)$, where $F^{-1}(\cdot|t;\omega)$ is the inverse function of $F(\cdot|t;\omega)$, the estimative prediction limit is given by $\hat{r}_\alpha = r_\alpha(\hat{\omega}, t)$. It is well-known that estimative prediction limits are usually imprecise, since the associated coverage error can be substantial. To reduce the asymptotic order of the coverage error term, suitable modifications of $\hat{r}_\alpha$ have been proposed by using asymptotic calculations (Barndorff-Nielsen and Cox, 1996; Vidoni, 1998), simulation-based calibration arguments (Beran, 1990; Hall et al. 1999) and approximate pivotal quantities (Lawless and Fredette, 2005).

For random effects prediction, since some of the above mentioned modifications usually require complicated asymptotic expansions, it may be convenient to consider the procedure by Ueki and Fueda (2007), that gives asymptotically equivalent improved prediction limits by means of a simple simulation-based technique. This solution is now briefly reviewed.

Under regularity assumptions, by means of suitable asymptotic expansions, we have that the conditional coverage probability of $\hat{r}_\alpha$ is such that

$$\begin{aligned} \hat{\alpha}(\omega, t) &= P_{Y,R|T}\left\{R \leq r_\alpha(\hat{\omega}, T)|T = t\right\} = E_{Y|T}\left[F\{r_\alpha(\hat{\omega}, T)|T; \omega\}|T = t\right] \\ &= \alpha + c(\alpha, \omega, t) + o(M^{-1}), \end{aligned}$$

where the expectation is with respect to the conditional distribution of Y given $T = t$, and M is the asymptotic index, typically related to the sample size. The coverage error term $c(\alpha, \omega, t)$ has an asymptotic order $O(M^{-1})$ which depends on the consistency features of $\hat{\omega}$. Indeed,

$$\begin{aligned} c(\alpha, \omega, t) &= -\sum_{s=1}^{d} b_s(\omega, t) F_s(r_\alpha|t; \omega) \\ &- \frac{1}{2}\sum_{s,v=1}^{d} i^{sv}(\omega, t)\left\{F_{sv}(r_\alpha|t; \omega) - 2\, F_s(r_\alpha|t; \omega)\ell_v(\omega; r_\alpha, t)\right\}. \end{aligned} \tag{5}$$

Here, $r_\alpha = r_\alpha(\omega, t)$, $b_s(\omega, t)$ is the first-order conditional (on $T = t$) bias term of the s-th component of the maximum likelihood estimator $\hat{\omega}$ and $i^{sv}(\omega, t)$ is the (s, v)-element of the inverse of the conditional (on $T = t$) expected information matrix. Moreover, $F_s(r|t; \omega)$ and $F_{sv}(r|t; \omega)$ are the first and the second partial derivatives of $F(r|t; \omega)$ with respect to the corresponding components of vector ω and $\ell_v(\omega; r, t)$ is $\partial\ell(\omega; r, t)/\partial\omega_v$, where $\ell(\omega; r, t) = \log f(r|t; \omega)$. Usually, the conditional expected information matrix may be substituted with the unconditional one, maintaining the same approximation order.

It is possible to verify (Barndorff-Nielsen and Cox, 1996; Vidoni, 1998) that the modified estimative prediction limit

$$r_\alpha(\hat{\omega}, t) + a(\alpha, \hat{\omega}, t),$$

where $a(\alpha, \omega, t) = -c(\alpha, \omega, t)/f(r_\alpha|t; \omega)$, reduces the coverage error to order $o(M^{-1})$. However, since the computation of the modifying term can be troublesome, Ueki and Fueda (2007) obtained the following asymptotic equivalent expression

$$a(\alpha, \omega, t) = -c(\alpha, \omega, t)/f(r_\alpha|t; \omega) = r_\alpha(\omega, t) - r_{\hat{\alpha}(\omega,t)}(\omega, t) + o(M^{-1}),$$

which gives the improved prediction limit

$$\tilde{r}_\alpha(\hat{\omega}, t) = r_\alpha(\hat{\omega}, t) + r_\alpha(\hat{\omega}, t) - r_{\hat{\alpha}(\hat{\omega},t)}(\hat{\omega}, t) = 2r_\alpha(\hat{\omega}, t) - r_{\hat{\alpha}(\hat{\omega},t)}(\hat{\omega}, t)\,.$$

Here the notation $\hat{\alpha}(\hat{\omega}, t)$ means

$$\hat{\alpha}(\hat{\omega}, t) = E_{Y|T}\left[F\{r_\alpha(\hat{\omega}, T)|T; \omega\}|T = t\right]|_{\omega=\hat{\omega}}\,,$$

i.e. the evaluation at $\hat{\omega}$ takes place after computing the expected value. This task can be performed by means of a suitable parametric bootstrap procedure conditional on $T = t$.

Fonseca et al. (2012) complete the Ueki and Fueda's procedure by defining the predictive distribution function which gives the improved prediction limit $\tilde{r}_\alpha(\hat{\omega}, t)$ as its α-quantile, for all $\alpha \in (0, 1)$. Neglecting terms of order $o(M^{-1})$, it corresponds to

$$\tilde{F}(r|t; Y) = F(r|t; \hat{\omega}) + f(r|t; \hat{\omega})\left[F^{-1}\{\hat{\alpha}(\hat{\omega}, t)|t; \hat{\omega}\}|_{\alpha=F(r|t;\hat{\omega})} - r\right], \tag{6}$$

where $F^{-1}\{\hat{\alpha}(\hat{\omega}, t)|t; \hat{\omega}\}$ depends on α through $\hat{\alpha}(\hat{\omega}, t)$.

3 Prediction of random effects

As emphasized in Section 2.1, our objective is to provide a relatively simple procedure for predicting random effects, and in particular some associated one-dimensional transformations, by means of prediction intervals with good coverage properties.

Starting from the conditional density $f(r|y_i; \omega)$ for the interest quantity R, as defined by (3), an obvious choice is to take as the transitive statistic of Section 2.2 the observed

value of $T = Y_i$. Let us assume that the numerator of (3) may be specified explicitly, while the denominator $L_i(\omega; y_i)$ has to be computed using numerical or analytical approximation techniques. In particular, a simple application of the Laplace formula (see, for example, Barndorff-Nielsen and Cox, 1989, Chapter 6) gives the following approximation $L_i(\omega; y_i) = \bar{L}_i(\omega; y_i)\{1 + e(\omega, y_i)\}$, with

$$\bar{L}_i(\omega; y_i) = \frac{(2\pi)^{m/2} q(\bar{u}_i; \gamma) \prod_{j=1}^{n_i} p(y_{ij}|\bar{u}_i; \delta)}{\det\{I(\bar{u}_i; \omega)\}}. \tag{7}$$

Here, $\bar{u}_i$ is the (unique) minimum of function $-\log \prod_{j=1}^{n_i} p(y_{ij}|u_i; \delta)$ with respect of u_i in the interior of $\mathcal{U}$ and $I(\bar{u}_i; \omega)$ is a matrix with (s, v)-element $-\partial^2 \log \prod_{j=1}^{n_i} p(y_{ij}|u_i; \delta)/\partial u_{is} \partial u_{iv}$, $s, v = 1, \ldots, m$, evaluated at $u_i = \bar{u}_i$. Notice that, since $-\log \prod_{j=1}^{n_i} p(y_{ij}|u_i; \delta) = O(n_i)$, the error term $e(\omega, y_i)$ is of order $O(n_i^{-1})$, as $n_i \to \infty$. An alternative application of the Laplace formula is also possible, if we consider the integrand in (2) written in a fully exponential form and we specify the (unique) minimum of function $-\log \prod_{j=1}^{n_i} p(y_{ij}|u_i; \delta) - \log q(u_i; \gamma)$, instead. The approximation is asymptotically equivalent to the previous one, but it lacks invariance with respect to transformations of u_i.

Although the Laplace approximation is often very accurate, there is a relative error term which may not be negligible, if n_i is small. On the other hand, it is usually possible to approximate the i-th likelihood component $L_i(\omega; y_i)$ also by means of accurate numerical techniques, as done for the logistic model presented in Section 5.2. In the following calculations, we shall consider explicitly the two components of the coverage error term, namely that one related to the approximate evaluation of $L_i(\omega; y_i)$ and that one induced by the estimative procedure. As in Booth and Hobert (1998), we assume that the dimension n_i, $i = 1, \ldots, k$, of the k groups is bounded, so that the asymptotic expansions are computed assuming $k \to \infty$. In other words, the number of groups k plays the role of the asymptotic index M in the results of Section 2.2.

We shall consider prediction limits obtained from the approximate density $\bar{f}(r|y_i; \omega)$, specified from (3) by substituting $L_i(\omega; y_i)$ with $\bar{L}_i(\omega; y_i)$. Since $L_i(\omega; y_i) = \bar{L}_i(\omega; y_i)\{1 + e(\omega, y_i)\}$, it is easy to see that $\bar{f}(r|y_i; \omega) = f(r|y_i; \omega)\{1 + e(\omega, y_i)\}$ and that the associated distribution function corresponds to $\bar{F}(r|y_i; \omega) = F(r|y_i; \omega)\{1 + e(\omega, y_i)\}$. The (approximate) estimative prediction limit $\hat{r}_\alpha = r_\alpha(\hat{\omega}, y_i)$ is obtained from $r_\alpha(\omega, y_i) = \bar{F}^{-1}(\alpha|y_i; \omega)$

by substituting ω with $\hat{\omega}$. Following the procedure outlined in Section 2.2, we find that the corresponding conditional coverage probability is

$$\begin{aligned}\hat{\alpha}(\omega, y_i) &= E_{Y|Y_i}\left[F\{\hat{r}_\alpha|Y_i;\omega\}|Y_i = y_i\right] = E_{Y|Y_i}\left[\bar{F}\{\hat{r}_\alpha|Y_i;\omega\}|Y_i = y_i\right]\{1 + e(\omega, y_i)\}^{-1} \\ &= \alpha + \bar{c}(\alpha, \omega, y_i) + o(k^{-1}), \end{aligned} \tag{8}$$

where

$$\bar{c}(\alpha, \omega, y_i) = \frac{c(\alpha, \omega, y_i) - \alpha\, e(\omega, y_i)}{1 + e(\omega, y_i)}, \tag{9}$$

with $c(\alpha, \omega, t)$ given by (5) with $t = y_i$ and $\bar{F}(\cdot)$ instead of $F(\cdot)$. Using a result outlined in Vidoni (2006, Appendix B), we find that, since n_i is bounded, $c(\alpha, \omega, y_i) = O(k^{-1})$ and the remainder term in (8) is of order $o(k^{-1})$. Notice that, in this case, the first-order coverage error term depends on both the error $c(\alpha, \omega, y_i)$, due to the estimative procedure, and the error $e(\omega, y_i)$ related to the approximation mentioned before.

Whenever the normalizing constant $L_i(\omega; y_i)$ is explicitly known, or accurately estimated with numerical procedures that lead to negligible error, we have that $e(\omega, y_i) \doteq 0$ and $\hat{r}_\alpha$ is the α-quantile of the true estimative distribution function $F(r|y_i; \hat{\omega})$. As a matter of fact, in both cases, the conditional coverage probability of the (approximate) estimative prediction limit differs from the target nominal value α by an error term which can be substantial for small values of k, and a suitable correction is usually needed for making accurate prediction statements. Under this respect, we prove that the Ueki and Fueda's procedure improves the estimative one and it gives a simple solution for improved random effects prediction.

As stated in Section 2.2, the improved prediction limit may be expressed as

$$\tilde{r}_\alpha(\hat{\omega}, y_i) = 2r_\alpha(\hat{\omega}, y_i) - r_{\hat{\alpha}(\hat{\omega}, y_i)}(\hat{\omega}, y_i), \tag{10}$$

and it corresponds to the α-quantile of the predictive distribution function (6) with $t = y_i$ and, eventually, $\bar{F}(\cdot)$ substituted for $F(\cdot)$. Since the estimative coverage probability $\hat{\alpha}(\omega, y_i)$ is usually unknown, we may consider an estimate based on a suitable bootstrap parametric technique conditional on $Y_i = y_i$.

We prove that the coverage error of the modified prediction limit $\tilde{r}_\alpha = \tilde{r}_\alpha(\hat{\omega}, y_i)$ is reduced with respect to that of the estimative solution. Using relation (8) and the fact that

$$\tilde{r}_\alpha = \hat{r}_\alpha - \frac{\bar{c}(\alpha, \hat{\omega}, y_i)}{\bar{f}(\hat{r}_\alpha|y_i; \hat{\omega})} + o(k^{-1}),$$

we obtain, by means of a simple stochastic expansions around $\tilde{r}_\alpha = \hat{r}_\alpha$, that the conditional coverage probability of $\tilde{r}_\alpha$ is

$$\begin{aligned} E_{Y|Y_i}\left[F\{\tilde{r}_\alpha|Y_i;\omega\}|Y_i = y_i\right] &= \hat{\alpha}(\omega, y_i) - \bar{c}(\alpha, \omega, y_i)\frac{f(r_\alpha|y_i;\omega)}{\bar{f}(r_\alpha|y_i;\omega)} + o(k^{-1}) \\ &= \alpha + \bar{c}(\alpha, \omega, y_i)\frac{\bar{f}(r_\alpha|y_i;\omega) - f(r_\alpha|y_i;\omega)}{\bar{f}(r_\alpha|y_i;\omega)} + o(k^{-1}) \\ &= \alpha + \bar{c}(\alpha, \omega, y_i)\frac{e(\omega, y_i)}{1 + e(\omega, y_i)} + o(k^{-1}). \end{aligned} \tag{11}$$

Comparing (11) with (8), we conclude that the coverage error term of the modified prediction limit $\tilde{r}_\alpha$ is still of order $O(k^{-1})$, but it is uniformly lower than that one of the estimative prediction limit $\hat{r}_\alpha$. Moreover, when $L_i(\omega; y_i)$ is explicitly known, or an accurate numerical estimate is available, the term $e(\omega, y_i)$ is null or close to zero and the $O(k^{-1})$ order term in (11) vanishes. Thus, in this case, the asymptotic order of the coverage error of the modified prediction limit $\tilde{r}_\alpha$ is $o(k^{-1})$, lower than that of the estimative solution.

4 Application to generalized linear mixed models

This section applies the general theory introduced above for improved random effects prediction in GLMMs. We also cover the important issue of preventing degenerate estimation of variance components.

4.1 Preliminaries

An interesting application of the theoretical findings presented in the previous section concerns the case of GLMMs. These models represent an extension of generalized linear models, including random effects into the linear predictor; McCulloch at al. (2008) provide an overview. In this framework, the responses Y_{ij}, $j = 1, \ldots, n_i$, in the i-th group, $i = 1, \ldots, k$, have conditional density functions of the form

$$p_{ij}(y_{ij}|u_i; \beta, \lambda) = c(\lambda, y_{ij}) \exp[\lambda\{y_{ij}\theta_{ij} - K(\theta_{ij})\}], \quad y_{ij} \in \mathcal{Y} \subseteq \mathbf{R}, \tag{12}$$

where $\theta_{ij} = x_{ij}^T\beta + z_{ij}^T u_i$ is the linear predictor, with $x_{ij} = (x_{ij1}, \ldots, x_{ijq})^T$ and $z_{ij} = (z_{ij1}, \ldots, z_{ijm})^T$ known covariate values, and $\beta = (\beta_1, \ldots, \beta_q)^T$ an unknown q-dimensional

parameter. Here $\lambda \in \Lambda \subseteq \mathbf{R}^+$ is the index parameter, whereas $\sigma^2 = 1/\lambda$ is the dispersion parameter. According to the notation introduced in Section 2.1, $\delta = (\beta^T, \sigma^2)^T$.

The model function (12) corresponds to a reproductive exponential dispersion model (see Jørgensen, 1997, Chapter 3) with mean $\mu_{ij} = \mu(\theta_{ij}) = dK(\theta_{ij})/d\theta_{ij}$ and variance $\sigma^2 V(\mu_{ij})$, where the variance function $V(\mu_{ij}) = d^2K(\theta_{ij})/d\theta_{ij}^2|_{\theta_{ij}=\theta(\mu_{ij})}$, with $\theta(\cdot)$ the inverse of $\mu(\cdot)$. The class of GLMMs is obtained by considering a monotonic differentiable link function $g(\cdot)$ such that $g(\mu_{ij}) = x_{ij}^T\beta + z_{ij}^T u_i$. Moreover, $\mu_{ij} = g^{-1}(x_{ij}^T\beta + z_{ij}^T u_i)$ and $\theta_{ij} = \theta\{g^{-1}(x_{ij}^T\beta + z_{ij}^T u_i)\}$, with $g^{-1}(\cdot)$ the inverse of $g(\cdot)$. If the canonical link function $g(\cdot) = \theta(\cdot)$ is considered, we obtain $\theta_{ij} = x_{ij}^T\beta + z_{ij}^T u_i$.

With regard to the random effects, we assume that $U_i = (U_{i1}, \ldots, U_{im})^T$, $i = 1, \ldots, k$, follows a m-dimensional Gaussian distribution with null mean vector and Σ as variance matrix. For the notable case where the components U_{is}, $s = 1, \ldots, m$ are independent Gaussian random variables with mean 0 and variance σ_s^2, according to the notation introduced previously $\gamma = (\sigma_1^2, \ldots, \sigma_m^2)^T$ denotes the vector of the variance components, so that the model parameter corresponds to $\omega = (\delta^T, \gamma^T)^T$.

As already emphasized, this article concerns prediction of one-dimensional transformations $R = R(U_i, \omega)$ of the random effects U_i and here we consider, in particular, a linear combination of the form $x_{ij}^T\beta + z_{ij}^T U_i$, or the corresponding mean response $R = g^{-1}(x_{ij}^T\beta + z_{ij}^T U_i)$. In this particular situation, assuming Gaussian random effects, the conditional density of R given $Y_i = y_i$, specified by (3), simplifies to

$$f(r|y_i, \omega) = \frac{f(r; \omega) \prod_{j=1}^{n_i} c(\lambda, y_{ij}) \exp[\lambda\{y_{ij} g(r) - K(g(r))\}]}{L_i(\omega; y_i)}. \tag{13}$$

The function $f(r; \omega)$ is the marginal density of R, which may be computed explicitly since the linear combination $x_{ij}^T\beta + z_{ij}^T U_i$ follows a Gaussian distribution with mean $x_{ij}^T\beta$ and variance equal to $\sum_{s=1}^m z_{ijs}^2 \sigma_s^2$, whenever the components $U_{is}, s = 1, \ldots, m$, are independent.

4.2 Prediction based on the CMSEP

The CMSEP, proposed by Booth and Hobert (1998) for $R = R(u_i, \omega) = x_i^T\beta + z_i^T u_i$, has the following form

$$\text{CMSEP}(\omega; y_i) = E_{Y|Y_i}\left[\{R(\hat{u}_i, \hat{\omega}) - R(u_i, \omega)\}^2 | Y_i = y_i\right] .$$

Booth and Hobert showed that it can be written as

$$\text{CMSEP}(\omega; y_i) = \text{Var}_{Y|Y_i}(R|Y_i = y_i) + E_{Y|Y_i}\left[\{R(\hat{u}_i, \hat{\omega}) - R(\hat{u}_i(\omega), \omega)\}^2 | Y_i = y_i\right] ,$$

where the first term on the right hand side has order $O(1)$ and the second has order $O(k^{-1})$. The latter, denoted by $v(\omega, y_i)$, accounts for the estimation of ω, and it can be approximated by a first-order Taylor expansion, so that

$$v(\omega, y_i) = \left\{\frac{\partial R(\hat{u}_i(\omega), \omega)}{\partial \omega}\right\}^T j(\omega; y)^{-1} \left\{\frac{\partial R(\hat{u}_i(\omega), \omega)}{\partial \omega}\right\} , \tag{14}$$

with $j(\omega; y)$ denoting the observed Fisher information matrix at ω. Finally, the CMSEP has to be evaluated at $\omega = \hat{\omega}$, and this introduces a bias of order $O(k^{-1})$ in $\widehat{\text{Var}}_{Y|Y_i}(R|Y_i = y_i)$. The main order of such bias can be removed by a conditional parametric bootstrap, that is the same resampling scheme required by the improved prediction limits (10).

The α-level prediction limit based on the CMSEP method has the simple form

$$r_\alpha^{(C)}(\hat{\omega}, y_i) = \hat{u}_i + q_\alpha \sqrt{\text{CMSEP}(\hat{\omega}; y_i)} , \tag{15}$$

with q_α the α-quantile of the standard normal distribution. In the following, the method employing formula (15) will be refereed to as the Booth and Hobert CMSEP (BHC) method.

4.3 Avoiding boundary estimates of variance components

A serious problem with random effects prediction is that maximum likelihood estimation may give zero estimates for variance components, leading to total shrinkage for the estimative and improved prediction intervals. The literature presents some proposals addressing this issue, such as the Adjustment for Density Maximization (ADM) method (Morris, 2006; Li and Lahiri, 2010; Morris and Tang, 2011) or the penalized likelihood approach proposed by

Chung et al. (2013) for random-intercept linear mixed models, extended to general linear mixed models in Chung et al. (2015). The ADM and the penalized likelihood approach are strictly connected, as pointed out in Chung et al. (2013), and actually the extension of the ADM approach to random effects survival models proposed in Ha et al. (2013) could be considered as an extension of the penalized likelihood approach.

The penalized likelihood approach suits well the proposal of this paper, and it is worth considering. The idea is to penalize the likelihood function for the data at hand with a suitable density for the variance components, obtaining the penalized log likelihood

$$\ell_p(\delta, \gamma) = \sum_{i=1}^{k} \log L_i(\delta, \gamma; y_i) + \log p(\gamma)\,. \tag{16}$$

For choosing the penalty term $p(\gamma)$, the two papers by Chung and co-workers provide some guidelines, leading to estimated variance components that are always finite, but nonetheless close to the maximum likelihood estimate. For example, for random intercepts linear mixed models, a suitable penalty term for the random effects variance σ_1^2 can be obtained by taking $p(\sigma_1^2)$ as the gamma density $G_a(\alpha, \lambda)$, with $\alpha = 1.5$ and $\lambda \to 0$. Chung et al. (2013) showed that such choice keeps the estimated value of σ_1 within one estimated standard error from the maximum likelihood estimate. This choice has some further properties, such as an interesting connection with REML estimation, but for the aims of this paper the important point is that the $O(1)$ penalty term introduced in (16) does not affect the asymptotic properties of the estimator; see Chung et al. (2013, §4.2). The implication is that the results on Section 3 remain valid when the maximizer of $\ell_p(\delta, \gamma)$ is used in place of the maximum likelihood estimate, and the form of the prediction limit (10) is exactly the same.

5 Results for Gaussian and logistic mixed models

5.1 Gaussian models

We consider an important special case of linear mixed models. In this context it is possible to highlight the nature of the correction introduced by the prediction limit (10), in order to account for the additional variability introduced by the plug-in procedure and to improve

the coverage accuracy. Moreover, we compare our correction to the BHC method. Although the calculations concern the simple case of a normal linear model with random intercepts, we conjecture that the conclusions may have, to some extent, a general validity.

According to the notation of Section 4.1, let us assume that the random variables Y_{ij} given $U_i = u_i$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, are mutually independent normally distributed with mean $\mu_{ij} \in \mathbf{R}$ and variance $\sigma^2 > 0$, and we take $R = x_{ij}^T \beta + z_{ij}^T U_i$. Using standard properties of the normal distribution, it readily follows that the conditional distribution of the mean response R given $Y_i = y_i$ is Gaussian with mean $\mu_{R|Y_i} = x_{ij}^T \beta + z_{ij}^T \, \Sigma \, Z_i^T \, \Delta^{-1} (y_i - X_i \, \beta)$ and variance $\sigma^2_{R|Y_i} = z_{ij}^T (\Sigma - \Sigma \, Z_i^T \, \Delta^{-1} \, Z_i \, \Sigma) z_{ij}$. Here, $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$, $X_i = (x_{i1}, \ldots, x_{in_i})^T$, $Z_i = (z_{i1}, \ldots, z_{in_i})^T$ and $\Delta = Z_i \, \Sigma \, Z_i^T + \mathrm{diag}(\sigma^2, \ldots, \sigma^2)$.

Let us consider, as a special case, the normal linear model with a random intercept. In this case, $m = 1$, $z_{ij} \equiv 1$ and $\Sigma = \sigma_1^2$. It follows that

$$\mu_{R|Y_i} = x_{ij}^T \beta + \frac{\gamma_i}{n_i} \sum_{j=1}^{n_i} (y_{ij} - x_{ij}^T \beta), \quad \sigma^2_{R|Y_i} = \gamma_i \, \sigma^2 / n_i, \tag{17}$$

with $\gamma_i = \sigma_1^2 / (\sigma_1^2 + \sigma^2 / n_i)$. Then, the distribution and the density functions of R given $Y_i = y_i$ are, respectively, $F(r|y_i; \omega) = \Phi\{(r - \mu_{R|Y_i}) / \sigma^2_{R|Y_i}\}$ and $f(r|y_i; \omega) = \sigma^{-1}_{R|Y_i} \phi\{(r - \mu_{R|Y_i}) / \sigma_{R|Y_i}\}$, with $\Phi(\cdot)$ and $\phi(\cdot)$ denoting the standard normal distribution and density functions. Furthermore, the estimative α-prediction limit is $\hat{r}_\alpha = r_\alpha(\hat{\omega}, y_i) = \hat{\mu}_{R|Y_i} + q_\alpha \, \hat{\sigma}_{R|Y_i}$, where the hat denotes, as usual, evaluation at $\omega = \hat{\omega}$. As stated previously, $\hat{\omega}$ is the maximum likelihood estimate or its counterpart obtained from the penalized log likelihood (16).

Although it is computationally convenient to calculate the improved prediction limit using formula (10), in order to investigate the nature of the modifying term, we consider the following asymptotic equivalent expression

$$\tilde{r}_\alpha(\hat{\omega}, y_i) = \hat{r}_\alpha - \frac{c(\alpha, \hat{\omega}, y_i)}{\hat{\sigma}^{-1}_{R|Y_i} \phi\{(\hat{r}_\alpha - \hat{\mu}_{R|Y_i}) / \hat{\sigma}_{R|Y_i}\}}, \tag{18}$$

where $c(\alpha, \hat{\omega}, y_i)$ is given by (5) with $\hat{\omega}$ and y_i substituted for ω and t. As in Booth and Hobert (1998), we assume that n_i, $i = 1, \ldots, k$, is bounded, so that $c(\alpha, \omega, y_i) = O(k^{-1})$. With simple algebra (details in Appendix A) we find that, neglecting terms of order $o(k^{-1})$,

$$\tilde{r}_\alpha(\hat{\omega}, y_i) \doteq \hat{\mu}_{R|Y_i} - \eta_1(\hat{\omega}, y_i) + q_\alpha \sqrt{\hat{\sigma}^2_{R|Y_i} - \eta_2(\hat{\omega}, y_i) + \eta_3(\hat{\omega}, y_i)} + \varepsilon(\alpha, \hat{\omega}, y_i) \, . \tag{19}$$

Here $\eta_1(\omega, y_i) \doteq E_{Y|Y_i}(\hat{\mu}_{R|Y_i} - \mu_{R|Y_i}|Y_i = y_i)$ and $\eta_2(\omega, y_i) \doteq E_{Y|Y_i}(\hat{\sigma}^2_{R|Y_i} - \sigma^2_{R|Y_i}|Y_i = y_i)$ are the first-order conditional bias terms of $\hat{\mu}_{R|Y_i}$ and $\hat{\sigma}^2_{R|Y_i}$, as plug-in estimators for $\mu_{R|Y_i}$ and $\sigma^2_{R|Y_i}$, respectively. Namely, $\eta_3(\omega, y_i) = v(\omega, y_i)$ in (14), and $\eta_2(\omega, y_i)$ is the further additional correction for the conditional bias of the plug-in estimator of $\sigma^2_{R|Y_i}$. Finally, the last quantity $\varepsilon(\alpha, \hat{\omega}, y_i)$ is of order $O(k^{-1})$ and it involves the quantile q_α to a power greater than one.

This result, though related to a simple linear Gaussian mixed model, is very useful since it helps to clarify some general issues concerning random effects prediction by means of prediction intervals. First, even for Gaussian models, a prediction limit involving the bias correction for $\hat{\mu}_{R|Y_i}$ and both the bias and the variance corrections for $\hat{\sigma}^2_{R|Y_i}$, as defined by Booth and Hobert (1998), does not improve properly over the estimative solution. In particular, the conditional coverage error is still of order $O(k^{-1})$ and it could be not negligible. Furthermore, since the improved prediction limit $\tilde{r}_\alpha(\hat{\omega}, y_i)$ differs from a Gaussian quantile, the improved predictive distribution may have a density function quite different from a Gaussian one, e.g. a density with fatter tails, as in this case.

A further interesting point concerns the relation between the dimension n_i of the i-th group and the coverage error term $c(\alpha, \omega, y_i)$ of the estimative prediction limit. To study this relation we determine the order of $c(\alpha, \omega, y_i)$, assuming a more general asymptotic regime where $k \to \infty$, $n_i \to \infty$, $i = 1, \ldots, k$, with $n_i = o(\sum_{l=1}^k n_l)$. Since $c(\alpha, \omega, y_i)$ is given by (5) with t replaced by y_i, using the results presented in Appendix B, and recalling that the conditional expected information matrix may be substituted by the unconditional one, we find that the order is $O(n_i^{3/2}/\sum_{l=1}^k n_l)$. In the particular case of a balanced design where $n_i = \bar{n}$, $i = 1, \ldots, k$, the order corresponds to $O(\bar{n}^{1/2}/k)$. Namely, the coverage error term of the estimative prediction limit does not decrease with the group size, contrary to what the intuition may suggest. Thus, when there is a small number k of groups and the dimension n_i of the i-th group is comparatively large, random effects prediction based on the estimative procedure may have coverage probability far from the nominal value. In this case, an alternative, improved prediction limit is recommendable.

Example 1: Linear model with random intercepts.

We illustrate some further points by means of a numerical example. In particular, we consider the survey and satellite data given in full in Battese et al. (1988), already employed by many

authors, including Booth and Hobert (1998). This is a small-area dataset, including data on corn and soybean production for 12 Iowa counties, each comprising observations from various segments. A random intercept model is assumed for the hectares of corn y_{ij} in segment j of county i

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i + \varepsilon_{ij} ,$$

where x_1 and x_2 are the two covariates obtained from satellite data. Here $i = 1, \ldots, 12$ and $j = 1, \ldots, n_i$, with $1 \leq n_i \leq 5$. The goal is to predict

$$R_i = \beta_0 + \beta_1 \overline{x}_{1i(p)} + \beta_2 \overline{x}_{2i(p)} + U_i , \quad i = 1, \ldots, 12 , \tag{20}$$

where the mean values $\overline{x}_{1i(p)}$ and $\overline{x}_{2i(p)}$ for the i-th county are the population values of the satellite covariates. Figure 1 reports the result of three different predictive intervals, obtained by the estimative method, the BHC method and the improved method based on (10), using 5,000 conditional bootstrap simulations for the latter two methods. The three sets of prediction intervals have been computed using the penalized estimate defined by (16) rather than the maximum likelihood estimate, as for the setting of interest there is a non-negligible probability of zero estimate of σ_1^2 in the conditional bootstrap resampling. To be more precise, the average percentage of null estimates of σ_1^2 varies between 0.1% and 6.2% across the 12 different conditional bootstrap samples, with an average of 3.8%.

[Figure 1 about here.]

There is notable adjustment made by both the BHC method and the improved, with an average length of prediction intervals about 15.6% and 17.5% larger in the two cases.

Simulation study 1: Linear model with random intercepts.

In order to assess the properties of the various methods, a small-scale simulation study was performed. In particular, the same setting of the *Corn and soybean data* analysed above was considered, and the actual coverages of the various methods were estimated via simulation. Taking two different counties, a conditional simulation was performed, mirroring the same kind of computation done for the improved prediction. In particular, 10,000 data sets were simulated with true parameter value fixed at the observed estimate and with the data of a given county held fixed at the observed value; the study was repeated for the two counties

Cerro Gordo (with $n_i = 1$) and Hardin ($n_i = 5$). The estimated coverages of prediction intervals for the three methods used for Figure 1 are reported in Table 1, obtained with 2,000 conditional simulations for the bootstrap-based adjustment. All the methods were computed by employing again the penalized likelihood estimator defined by (16). Some further computational details are reported in Appendix C.

[Table 1 about here.]

The results show the under-coverage of the estimative procedure, which is to some extent corrected by both the BHC and the improved methods. In some cases, with small estimated random effects variance, the CMSEP may be not computable, as either the $v(\hat{\omega}, y_i)$ term in (14) or the bias-correction term for $\widehat{\mathrm{Var}}_{Y|Y_i}(R|Y_i = y_i)$ may render the whole expression negative. In such cases, which occur rarely (less than 20 data sets in either study), the CMSEP was computed without the higher-order adjustment. We also notice that the coverage levels of prediction intervals for either the BHC and the improved method are below the nominal level. This is mainly to be ascribed to the small number of groups, as confirmed by a further simulation study performed in the case of $n_i = 1$ with a very similar setting with enlarged data, i.e. 24 groups of size between 1 and 5. See again Table 1.

Finally, we conclude this section by noticing that a similar study based on the ordinary maximum likelihood estimator, discarding the data sets with the estimate of σ_1^2 close to zero in both the main simulation and the bootstrap computations, gave results similar to that of Table 1. However, the solution based on the penalized estimator is more appealing, as it provides a methodology which can be applied to any dataset, in an automatic fashion.

5.2 Logistic models

Logistic regression with random effects are an important instance of GLMMs, for which random effects prediction has been considered by several authors; see, among others, Jiang (2007, §3.6) and Skrondal and Rabe-Hesketh (2009). The problem can be challenging, as documented by the simulation studies reported in Ten Have and Localio (1999). Differently from the linear case, no analytic expression exists even for the simplest cases, and the relevant theory is that reported in Section 3.

Example 1: Logistic regression with random intercepts.

As an illustrative example, we consider the logistic-normal example already analysed by Booth and Hobert (1998), who made use of the multicenter clinical trial data from Beitler and Landis (1985). We adopt the same logistic model with random clinic effects for the binary response, with linear predictor

$$\eta_{ij} = \beta_0 + \beta_1 \, x_{ij} + u_i \, .$$

Here x_{ij} is the binary treatment indicator, and $U_i \sim N(0, \sigma_1^2)$ the random intercept, $i = 1, \ldots, 8$, $j = 1, \ldots, n_i$ ($13 \leq n_i \leq 73$). Like in Booth and Hobert (1998), the interest is on the conditional linear predictors for the treated, given for each clinic by $R_i = \beta_0 + \beta_1 + U_i$.

The methods of the previous sections have been applied, approximating the integrals required for the likelihood function and the predictive distribution $f(u_i|y_i;\omega)$ by adaptive Gaussian quadrature, thus (essentially) removing the error component $e(\omega, y_i)$ in (9) when computing the prediction limits with the improved method. As the percentage of data sets with estimated random effects variance equal to zero was negligible (i.e. average percentage equal to 0.34% across the 8 clinics), we did not make any adjustment to the ordinary maximum likelihood estimates. At any rate, as mentioned above the penalized estimation method (16) could be adopted also for this setting. Figure 2 reports the three predictive densities corresponding to the three methods under study, for two different clinics with different sample sizes. As reminded in Section 4.2, the predictive distribution based on the BHC method is just a Gaussian density, with mean equal to the BLUP estimate of R_i and variance equal to the CMSEP, in accordance with (15). The adjusted methods were obtained with 5,000 conditional bootstrap simulations; some computational details are reported in Appendix C.

[Figure 2 about here.]

The adjustments made by both the BHC method and the improved method are apparent. Although both the two adjusted methods produce larger prediction quantiles, as they adjust for the sample variability of the parameter estimates, it is clear that the two adjustments are of different nature. The improved distribution (6), in particular, may deviate from symmetry substantially, differently from the BHC method.

Simulation study 2: Logistic regression with random intercepts.

The properties of the various methods were assessed also in this case by simulation, designing a study based on the observed setting for the multicenter clinical trial data used above, with true parameter set equal to the maximum likelihood estimate. Like for the observed data, ordinary maximum likelihood estimation was used, discarding the data sets with zero estimated variance. The empirical coverages were computed on the remaining data sets, treating all the methods on an equal footing. The simulation study was repeated with different group sizes, using the same sizes employed for Figure 2. Like in the previous study, conditional simulations were performed, generating data sets with the data of the i-th group (with $i = 1, 8$) held fixed at the observed value. However, also some unconditional simulations were performed, with data generating from the assumed model for all the groups. In the latter case, however, the sample-by-sample improved prediction limits were still obtained conditionally. In either case, 2,000 bootstrap samples were used for the adjusted methods. Also here, the CMSEP for the BHC method was computed without the high-order adjustment in the problematic cases, occurring in moderate number for all the settings.

[Table 2 about here.]

The simulation results show that both the BHC method and the improved one provide a useful adjustment to the estimative procedure, removing most of its inaccuracy. The two adjusted methods have similar coverages, though the improved method gives slightly more symmetric coverages. As stated in Section 2.2, the unconditional coverages of both the improved procedure and the BHC method are also quite satisfactory.

6 Discussion

This paper has proposed a novel method for frequentist prediction of parametric functions of random effects, which combines simplicity of use with good accuracy. The only price is merely computational, but with nowadays computing power this is not an issue in many cases. A remarkably property of the methodology is the possibility to obtain a bonafide predictive distribution, whose quantiles supply the prediction interval at any given level.

In all the examples and simulation studies, a good agreement was found between the BHC method and the proposal of this paper. To some extent, this can be explained, as both methods improve over estimative prediction following a conditional approach, with pronounced agreement for linear forms of R; see equation (19). However, it can be said that the improved prediction method offers some theoretical advantages, such as the possibility to obtain asymmetric prediction distributions (and prediction intervals), possible extension to nonlinear specifications of R, and invariance to model parameterization. The latter property is not satisfied by the CMSEP adjustment, as suggested by formula (14). A further factor is the higher accuracy of the improved procedure, as displayed in formula (11). Finally, it is should be noted that the computational cost of the two procedures is similar, as parametric bootstrap is required for both methods. The only difference may lie in the fact that the bias correction for the BHC method would require a smaller number of bootstrap trials, but this seems a minor issue.

The methodology of this paper has been developed for the case of independent groups. Extension to other designs, including crossed designs or the mixed models approach to smoothing, would be of considerable appeal, as testified by current research (e.g. Marra and Wood, 2012). For such complex settings, analytical solutions may play an important role, despite their awkward nature. At any rate, the conditional approach employed here appears of difficult generalization, and therefore unconditional solutions may be the only possibility. Some further research on these topics is called for.

Appendix A: Formulas for the Gaussian case

We consider the normal linear model with random intercept specified in Section 4.2 and we aim at computing the improved prediction limit $\tilde{r}_\alpha(\hat{\omega}, y_i)$ for the expected response $R = x_{ij}^T\beta + U_i$ as given by (18). We need an explicit expression for the quantity $c(\alpha, \omega, y_i)$, defined by (5). This is obtained with a simple algebra by calculating the first and the second partial derivatives of $\Phi\{(r - \mu_{R|Y_i})/\sigma^2_{R|Y_i}\}$ and the first partial derivatives of $\log[\sigma^{-1}_{R|Y_i}\phi\{(r - \mu_{R|Y_i})/\sigma_{R|Y_i}\}]$, with respect to the components of the parameter vector $\omega = (\beta_1, \ldots, \beta_q, \sigma^2, \sigma_1^2)$. In the computation we use the fact that the conditional expected

information matrix may be substituted with the unconditional one, maintaining the same approximation order, and that, in this case, the expected information matrix is block diagonal. Indeed, for the asymptotic calculations, we assume that n_i, $i = 1, \ldots, k$, is bounded, while $k \to \infty$. The final expression is rather long and, neglecting terms of order $o(k^{-1})$, it can be summarized as

$$\begin{aligned}\tilde{r}_\alpha(\hat{\omega}, y_i) &\doteq \hat{\mu}_{R|Y_i} - \eta_1(\hat{\omega}, y_i) + q_\alpha \left\{ \hat{\sigma}_{R|Y_i} - \frac{\eta_2(\hat{\omega}, y_i)}{2\hat{\sigma}_{R|Y_i}} + \frac{\eta_3(\hat{\omega}, y_i)}{2\hat{\sigma}_{R|Y_i}} \right\} + \varepsilon(\alpha, \hat{\omega}, y_i) \\ &\doteq \hat{\mu}_{R|Y_i} - \eta_1(\hat{\omega}, y_i) + q_\alpha \sqrt{\hat{\sigma}^2_{R|Y_i} - \eta_2(\hat{\omega}, y_i) + \eta_3(\hat{\omega}, y_i)} + \varepsilon(\alpha, \hat{\omega}, y_i),\end{aligned}$$

which is exactly the result presented in Section 4.2. Quantities $\eta_1(\omega, y_i) \doteq E_{Y|Y_i}(\hat{\mu}_{R|Y_i} - \mu_{R|Y_i}|Y_i = y_i)$ and $\eta_2(\omega, y_i) \doteq E_{Y|Y_i}(\hat{\sigma}^2_{R|Y_i} - \sigma^2_{R|Y_i}|Y_i = y_i)$ are the first-order conditional bias terms of the plug-in estimators $\hat{\mu}_{R|Y_i}$ and $\hat{\sigma}^2_{R|Y_i}$ and correspond to

$$\eta_1(\omega, y_i) = \sum_{s=1}^{q+2} \frac{\partial \mu_{R|Y_i}}{\partial \omega_s} b_s(\omega, y_i) + \frac{1}{2} \sum_{s,v=1}^{q+2} \frac{\partial^2 \mu_{R|Y_i}}{\partial \omega_s \partial \omega_v} i^{sv}(\omega),$$

$$\eta_2(\omega, y_i) = \sum_{s=1}^{q+2} \frac{\partial \sigma^2_{R|Y_i}}{\partial \omega_s} b_s(\omega, y_i) + \frac{1}{2} \sum_{s,v=1}^{q+2} \frac{\partial^2 \sigma^2_{R|Y_i}}{\partial \omega_s \partial \omega_v} i^{sv}(\omega).$$

Here, $b_s(\omega, y_i)$ is the first-order conditional bias term of s-th component of the maximum likelihood estimator $\hat{\omega}$ and $i^{sv}(\omega)$ is the (s, v)-element of the inverse of the (unconditional) expected information matrix. Furthermore,

$$\begin{aligned}\eta_3(\omega, y_i) &= \sum_{s,v=1}^{q} i^{sv}(\omega)(x_{ijs} - \gamma_i \bar{x}_{ijs})(x_{ijv} - \gamma_i \bar{x}_{ijv}) \\ &+ \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - x_{ij}^T \beta) \right\}^2 \left\{ i^{q+1q+1}(\omega) \frac{\gamma_i^2 (1 - \gamma_i)^2}{\sigma^4} \right. \\ &- \left. 2\, i^{q+1q+2}(\omega) \frac{\gamma_i^3 (1 - \gamma_i)}{n_i \sigma_1^4} + i^{q+2q+2}(\omega) \frac{\gamma_i^4 \sigma^4}{n_i^2 \sigma_1^8} \right\},\end{aligned}$$

with $\bar{x}_{ijr} = n_i^{-1} \sum_{j=1}^{n_i} x_{ijr}$, $r = 1, \ldots, q$, and it coincides with the variance correction term calculated by Booth and Hobert (1998, Equation (24)). Finally, the further quantity $\varepsilon(\alpha, \hat{\omega}, y_i)$ is not negligible, since it is of order $O(k^{-1})$, and it involves the standard normal α-quantile q_α to a power greater than one.

Appendix B: Some results for two-index asymptotics

We consider the normal linear model with random intercept specified in Section 4.2 and we aim at specifying the asymptotic order of the quantities required for the calculation of the coverage error term $c(\alpha, \omega, y_i)$ of the estimative prediction limit, given by (5) with t substituted by y_i. Unlike Appendix A, we assume that $k \to \infty$, $n_i \to \infty$, $i = 1, \ldots, k$, with $n_i = o(\sum_{l=1}^{k} n_l)$.

Let us consider $\omega = (\omega_1, \ldots, \omega_q, \omega_{q+1}, \omega_{q+2}) = (\beta_1, \ldots, \beta_q, \sigma^2, \sigma_1^2)$; recalling the results on the asymptotic conditional bias and variance of the maximum likelihood estimator presented in Vidoni (2006, Appendix B), we state that

$$b_s(\omega, y_i) = \begin{cases} O(1/\sum_{l=1}^{k} n_l) & s = 1, \ldots, q \\ O(k^{-1}) & s = q+1, q+2, \end{cases}$$

$$i^{sv}(\omega, y_i) = \begin{cases} O(k^{-1}) & s, v = q+1, q+2 \\ O(1/\sum_{l=1}^{k} n_l) & \text{otherwise.} \end{cases}$$

Indeed, by considering the first and the second partial derivatives of the conditional distribution function $\Phi\{(r - \mu_{R|Y_i})/\sigma^2_{R|Y_i}\}$ and the first partial derivatives of function $\log[\sigma^{-1}_{R|Y_i}\phi\{(r - \mu_{R|Y_i})/\sigma_{R|Y_i}\}]$, with respect to the components of the parameter ω, and recalling that $\mu_{R|Y_i}$ and $\sigma^2_{R|Y_i}$ are given by (17), we have that

$$F_s(r|y_i; \omega) = \begin{cases} O(n_i^{1/2}) & s = 1, \ldots, q+1 \\ O(n_i^{-1/2}) & s = q+2, \end{cases}$$

$$F_{sv}(r|y_i; \omega) - 2\, F_s(r|y_i; \omega)\ell_v(\omega; r, y_i) = \begin{cases} O(n_i^{3/2}) & s, v = 1, \ldots, q+1 \\ O(n_i^{1/2}) & s, v = q+1, q+2 \\ O(n_i^{-1/2}) & s, v = q+2. \end{cases}$$

Appendix C: Computational details

The computation to obtain the improved prediction limits are in principle straightforward, but, as the method is based on parametric bootstrap, some care in its implementation is in order. For the linear model case, we relied on the `R` (R Core Team, 2015) package `blme`

(Dorie, 2014), which implements the penalized estimate defined by (16) for several GLMMs. For the logistic regression case, which is more demanding, we wrote our own code based on adaptive quadrature. A further complication for the logistic case is given by the fact that the quantiles of the conditional distribution $F^{-1}_{R|Y_i}(\alpha|y_i;\omega)$ are not available in close form, but they have to be obtained by solving a nonlinear equation. To this end, the `R` package `nleqslv` (Hasselman, 2015) was employed. A key factor for performing the simulation studies for the improved prediction limits was given by the parallel capabilities of `R`, provided by the `parallel` package, which is part of the standard `R` distribution. The predictive density represented in Figure (2) has been obtained by computing the prediction limits over a grid of values for α, and then proceeding to a suitable interpolation to obtain the distribution function (6). The predictive density was then obtained by numerical differentiation of the predictive distribution function.

Acknowledgement

A note based on a preliminary version of this work was presented at the 28th International Workshop on Statistical Modelling, held in Palermo in July 2013.

References

[1] Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for use in Statistics*. Chapman and Hall, London.

[2] Barndorff-Nielsen, O.E. and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli* **2**, 319-40.

[3] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28-36.

[4] Booth, J.G. and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 262-72.

[5] Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J. and Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics* **40**, 136–157.

[6] Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* **78**, 685–709.

[7] Datta, G. and Ghosh, M. (2012). Small area shrinkage estimation. *Statistical Science* **27**, 95-114.

[8] Dorie, V. (2014). *blme: Bayesian linear mixed-effects models*. R package version 1.0-2.

[9] Fonseca, G., Giummolè, F. and Vidoni, P. (2012). A note about calibrated prediction regions and distributions. *Journal of Statistical Planning and Inference* **142**, 2726-34.

[10] Ha, I.D., Vaida, F. and Lee, Y. (2013). Interval estimation of random effects in proportional hazards models with frailties. *Statistical Methods in Medical Research*, to appear.

[11] Hall, P., Peng, L. and Tajvidi, N. (1999). On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika* **86**, 871-80.

[12] Hasselman, B. (2015). *nleqslv: Solve systems of nonlinear equations*. R package version 2.8.

[13] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.

[14] Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92**, 529-42.

[15] Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis* **101**, 882-892.

[16] Lohr, S. L. and Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika* **96**, 457-68.

[17] Marra, G. and Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics* **39**, 53–74.

[18] McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd ed. Wiley, Hoboken, NJ.

[19] Morris, C. (2006). Mixed model prediction and small area estimation (with discussions). *Test* **15**, 72–76.

[20] Morris, C. and Tang, R. (2011). Estimating random effects via adjustment for density maximization. *Statistical Science* **26**, 271–287.

[21] R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

[22] Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science* **6**, 15-51.

[23] Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC Press, Boca Raton.

[24] Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series A* **172**, 659-87.

[25] Ten Have, T. R. and Localio, A. R. (1999). Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics* **55**, 1022-29.

[26] Ueki, M. and Fueda, K. (2007). Adjusting estimative prediction limits. *Biometrika* **94**, 509-11.

[27] Vidoni, P. (1998). A note on modified estimative prediction limits and distributions. *Biometrika* **85**, 949-53.

[28] Vidoni, P. (2006). Response prediction in mixed effects models. *Journal of Statistical Planning and Inference* **136**, 3948-66.

Table 1: Estimated coverages of 95% prediction intervals, with left- and right-tail 2.5% errors in brackets. Based on 10,000 conditional simulations.

	County	
	Cerro Gordo ($n_i = 1$)	Hardin ($n_i = 5$)
	Design based on original data (12 counties)	
Method		
Estimative	0.881 (0.060, 0.059)	0.891 (0.046, 0.063)
BHC	0.921 (0.040, 0.039)	0.939 (0.027, 0.034)
Improved	0.933 (0.034, 0.033)	0.927 (0.034, 0.040)
	Design based on enlarged data (24 counties)	
Method		
Estimative	0.916 (0.044, 0.040)	0.912 (0.037, 0.050)
BHC	0.934 (0.035, 0.031)	0.934 (0.028, 0.038)
Improved	0.945 (0.028, 0.027)	0.937 (0.030, 0.033)

Table 2: Estimated coverages of 95% prediction intervals, with left- and right-tail 2.5% errors in brackets. Based on 10,000 simulations for each entry.

	Clinic	
	1 ($n_i = 73$)	8 ($n_i = 13$)
	Conditional simulations[a]	
Method		
Estimative	0.898 (0.023, 0.078)	0.912 (0.028, 0.059)
BHC	0.954 (0.012, 0.035)	0.949 (0.026, 0.039)
Improved	0.953 (0.016, 0.031)	0.945 (0.025, 0.030)
	Unconditional simulations[b]	
Method		
Estimative	0.908 (0.044, 0.048)	0.903 (0.051, 0.046)
BHC	0.952 (0.025, 0.022)	0.943 (0.034, 0.023)
Improved	0.945 (0.027, 0.028)	0.954 (0.025, 0.021)

[a] The data sets with $\hat{\sigma}_1^2 \doteq 0$ were 75 for the study based on Clinic 1 and 6 for the study based on Clinic 8

[b] The data sets with $\hat{\sigma}_1^2 \doteq 0$ were 37

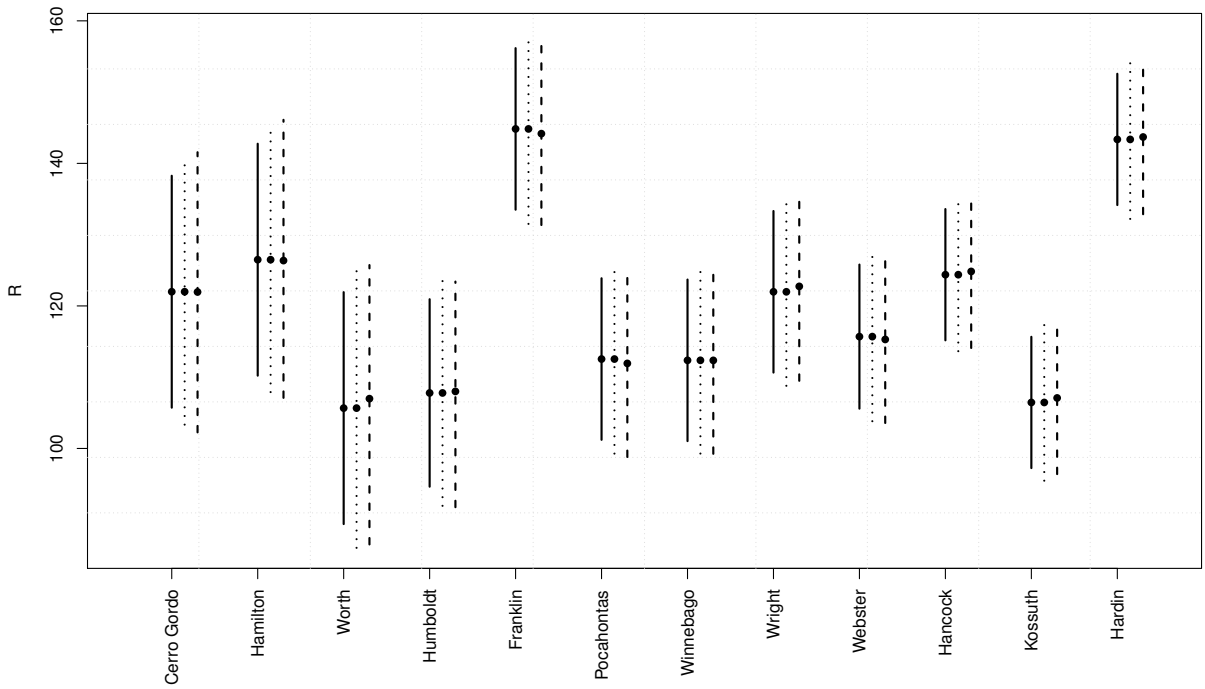

Figure 1: *Corn and soybean data*: 95% prediction intervals for R_i in (20) for 12 Iowa counties, based on the estimative method (solid), BHC method (dotted) and improved method (dashed). Dots denote point predictors.

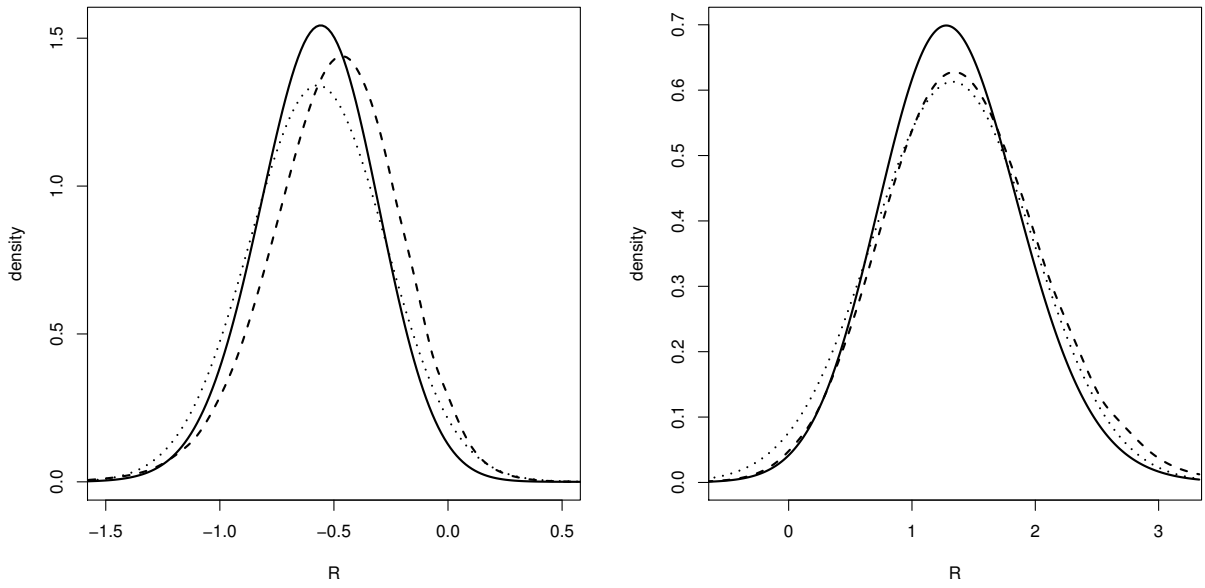

Figure 2: *Multicenter clinical trial data*: Predictive densities for Clinic 1 with $n_1 = 73$ (left panel) and Clinic 8 with $n_8 = 13$ (right panel), based on the estimative method (solid), BHC method (dotted) and improved method (dashed).