Statistica Sinica **12**(2002), 537-553

DISTRIBUTION-FREE PREDICTION INTERVALS IN MIXED LINEAR MODELS

Jiming Jiang and Weihong Zhang

Case Western Reserve University and Cleveland Clinic Foundation

Abstract: This paper considers prediction intervals for a future observation in the context of mixed linear models. For such prediction problems, it is reasonable to assume that the future observation is independent of the current ones. Our approach is distribution-free, that is, we do not assume that the distributions of the random effects and errors are normal or specified up to a finite number of parameters. We show that for standard mixed linear models, a simple method based on the (regression) residuals works well for constructing prediction intervals. For nonstandard mixed linear models, however, a more complicated method may have to be used, based on estimation of the distribution of the random effects. Simulation studies compare prediction intervals based on the ordinary least squares estimators and those based on the empirical best linear unbiased estimators. We apply the method to a data set regarding lead contamination of soil.

Key words and phrases: Asymptotic coverage probability, consistent estimator, empirical distribution, semiparametric inference.

1. Introduction

A prediction interval for a single future observation is an interval that will, with a specified coverage probability, contain a future observation from a population. In model-based statistical inference, it is assumed that the future observation has a certain distribution. Sometimes, the distribution is specified up to a finite number of unknown parameters, e.g., those of the normal distribution. Then, a prediction interval may be obtained, if the parameters are adequately estimated, and the uncertainty in the parameter estimation suitably assessed. Clearly, such a procedure is dependent on the underlying distribution in that, if the distributional assumption fails, the prediction interval may be seriously off, i.e., it either is wider than necessary, or does not have the claimed coverage probability. An alternative to the parametric method is a distribution-free approach, in which we do not assume the form of the distribution is known.

The problem of prediction intervals is, of course, an old one. One of the earliest work in this field is Baker (1935). Patel (1989) provides a review of the literature on prediction intervals when the future observation is independent of the observed sample, including results based on parametric distributions and on distribution-free methods. Hahn and Meeker (1991) review three types of statistical intervals that are used most frequently in practice: the confidence interval, the prediction interval, and the tolerance interval. For a more recent overview, and developments on nonparametric prediction intervals, see Zhou (1997). Although many results on prediction intervals are for the i.i.d. case, the problem is also well-studied in some non-i.i.d. cases, such as linear regression.

In this paper, we are interested in prediction intervals in mixed linear models. Let y be a $N \times 1$ vector of observations. The model is

$$y = X\beta + Z_1\alpha_1 + \cdots + Z_s\alpha_s + \epsilon \ , \tag{1}$$

where X is an $N \times p$ known matrix of full rank p (p a fixed integer), β is a $p \times 1$ vector of unknown constants (the fixed effects), Z_r is an $N \times m_r$ known matrix, α_r is an $m_r \times 1$ vector of i.i.d. random variables with mean 0 and unknown distribution F_r, $r = 1, \cdots, s$ (the random effects), ϵ is an $N \times 1$ vector of i.i.d. random variables with mean 0 and unknown distribution F_0 (the errors), and $\alpha_1, \ldots, \alpha_s, \epsilon$ are independent. We say (1) is *standard* if each Z_r consists only of 0's and 1's, and there is exactly one 1 in each row and at least one 1 in each column. Thus standard mixed linear model simply has it that the random effects appear in the model in the form of an analysis of variance, but there is no restriction on the fixed effects (see examples in Section 2). For such a reason, a standard mixed linear model is also known as a mixed model of the analysis of variance (e.g., Miller (1977)). Note that it is not assumed that the random effects and errors are normally distributed. However, it is crucial to assume that the F_r, $0 \leq r \leq s$, have finite variances.

1.1. Two types of prediction problems

In linear regression, observations are assumed to be independent. Therefore, it is a natural assumption that any future observation is independent of the current ones. In mixed linear models, however, this is not necessarily true and one distinguishes two types of prediction problems.

The first type of problem arises when one wishes to predict a future observation from a cluster or unit not previously observed, and it is reasonable to assume that it is independent of the current ones. The second type of problem occurs when one wishes to predict another observation from a cluster or unit from which samples have already been collected, and it is unrealistic to make such an assumption here. We offer some examples.

Example 1. In longitudinal studies, one may be interested in prediction, based on repeated measurements from the observed individuals, of a future observation from an individual not previously observed. It is of less interest to predict another observation from an observed individual, since longitudinal studies often aim at applications to a larger population (e.g., drugs going to the market after clinical trials).

Example 2. In sample surveys, responses may be collected in two steps: in the first step, a number of families are randomly selected; in the second step, some family members (e.g., all family members) are interviewed for each of the selected families. Again, one may be more interested in predicting what happens to a family not selected, because one already knows enough about selected families (especially when all family members in the selected families are interviewed).

Example 3. In small area estimation (e.g., Ghosh and Rao (1994)), small-area specific effects are often treated as random effects. In such cases, one may be interested in predicting the outcome from any small area, whether samples have been collected from it or not. This is related to prediction of random effects. Jeske and Harville (1988) consider prediction intervals for mixed effects, assuming that the joint distribution of α and $y - E(y)$ is known up to a vector of unknown parameters. Thus, their approach is not distribution-free.

1.2. The scope of this paper

We focus on problems of the first type, and assume a future observation, y_*, is independent of the current ones. Then $E(y_*|y) = E(y_*) = x_*^t\beta$, so the best predictor is $x_*^t\beta$, if β is known. Even if β is unknown, it is still fairly easy to obtain a prediction interval for y_* $\underline{\text{if}}$ one is willing to make the assumption that the distributions of the random effects and errors are known up to a vector of parameters (e.g., variance components). To see this, consider a simple case: $y_{ij} = x_{ij}^t\beta + \alpha_i + \epsilon_{ij}$, where the random effect α_i and error ϵ_{ij} are independent such that $\alpha_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, \tau^2)$. It follows that the distribution of y_{ij} is $N(x_{ij}^t\beta, \sigma^2 + \tau^2)$. In mixed linear models, methods are well-developed for estimating fixed parameters such as β, σ^2, and τ^2. Once the parameters are consistently estimated, a prediction interval with asymptotic coverage probability $1-\alpha$ is easy to obtain. However, it is much more difficult if one does not know the forms of the distributions of the random effects and errors, and this is the case we work on. In other words, our approach is distribution-free. Still to consistently estimate the fixed effects and variance components in a mixed linear model, one need not assume that the random effects and errors are normally distributed (Jiang (1996, 1998)).

1.3. Outline of the main results

Our approaches are quite different for standard and non-standard mixed linear models. For standard mixed linear models, the method is surprisingly simple, and can be described as follows: first, one throws away the middle terms in (1) that involve the random effects, and pretends that it is a linear regression model with i.i.d. errors: $y = X\beta + \epsilon$. Next, one computes the least squares (LS) estimator $\hat{\beta} = (X^t X)^{-1} X^t y$ and the residuals $\hat{\epsilon} = y - X\hat{\beta}$. Let $\hat{a}$ and $\hat{b}$ be the $\alpha/2$ and $1 - \alpha/2$ quantiles of the residuals. Then a prediction interval for y_* with asymptotic coverage probability $1 - \alpha$ is $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$, where $\hat{y}_* = x_*^t \hat{\beta}$ (see the discussion in the first paragraph of Section 1.2). Note that, although the method sounds almost the same as the residual method in linear regression, its justification is not so obvious because, unlike linear regression, the observations in a (standard) mixed linear model are not independent. The method may be improved if one uses more efficient estimators, such as the empirical best linear unbiased estimator (EBLUE, e.g., Jiang (1998)), instead of the LS estimator. For nonstandard mixed linear models, the method of obtaining prediction intervals involves estimation of the distributions of the random effects and errors.

The rest of the paper is organized as follows: In Section 2 we consider standard mixed linear models. Section 3 deals with the nonstandard case. Section 4 contains some simulation results. In Section 5, we apply our method to a data set regarding lead contamination of soil. Finally, in Section 6 we have a few concluding remarks. Proofs are given in the Appendix.

2. Standard Mixed Linear Models

In this section, we consider standard mixed linear models. Note that one can express the mixed model (1) as

$$y_i = x_i^t \beta + z_{i1}^t \alpha_1 + \cdots + z_{is}^t \alpha_s + \epsilon_i \ , \tag{2}$$

$i = 1, \ldots, N$, where x_i^t is the ith row of X, and z_{ir}^t the ith row of Z_r, $1 \le r \le s$. If we let $\alpha_r = (\alpha_{r,u})_{1 \le u \le m_r}$, $1 \le r \le s$, then, in the standard case, we have the equivalent expression

$$y_i = x_i^t \beta + \alpha_{1,u(i,1)} + \cdots + \alpha_{s,u(i,s)} + \epsilon_i \ , \tag{3}$$

where $u(i,r)$, $1 \le r \le s$, are some indices such that $1 \le u(i,r) \le m_r$. Note that $\alpha_{r,u(i,r)} \sim F_r$, $1 \le r \le s$, $\epsilon_i \sim F_0$, and $\alpha_{1,u(i,1)}, \ldots, \alpha_{s,u(i,s)}, \epsilon_i$ are independent. Let $\delta_i = y_i - x_i^t \beta$. Then, $\delta_1, \ldots, \delta_N$ are identically distributed, but not independent. Let F be the common distribution of the δ_i's.

For linear regression, which may be regarded as (2) without the terms,

$$\xi_i = z_{i1}^t \alpha_1 + \cdots + z_{is}^t \alpha_s \ , \tag{4}$$

Lai and Wei (1982) showed that under suitable conditions the LS estimator of β is consistent. The result does not require normality of the data.

In mixed linear models, the following theorem shows that, under suitable conditions, the LS estimator

$$\hat{\beta}_{OLS} = (X^t X)^{-1} X^t y \tag{5}$$

is still consistent, where OLS stands for ordinary least squares. The result does not require normality, nor the standardness of the mixed model. Let λ_{max} (λ_{min}) represent the largest (smallest) eigenvalue.

Theorem 1. *For the mixed model* (1) (*not necessarily standard*)*, if the variances of the random effects and errors are finite, and* $\lambda_{\text{min}}(X^t X) \to \infty$,

$$\frac{\lambda_{\text{max}}(Z_r^t Z_r)}{\lambda_{\text{min}}(X^t X)} \longrightarrow 0 \ , \quad 1 \le r \le s \ , \tag{6}$$

then $\hat{\beta}_{OLS} \xrightarrow{L^2} \beta$, *hence* $\hat{\beta}_{OLS} \xrightarrow{P} \beta$.

The proof is given in the Appendix. To see what (6) means, first note that for standard mixed linear models, $\lambda_{\text{max}}(Z_r^t Z_r) = \max_{1 \le u \le m_r} n_{ru}$, where n_{ru} is the number of 1's in the uth column of Z_r, $1 \le r \le s$, or equivalently, the number of times $\alpha_{r,u}$ appears in the model. Thus, if the number of times each individual random effect appears in the model is bounded, which implicitly assumes that m_r increases with N (which is typical in situations where mixed effects models are used, e.g., in small-area estimation, Ghosh and Rao (1994)), then condition (6) reduces to $\lambda_{\text{min}}(X^t X) \to \infty$. (See Example 4 below.) Note that $\lambda_{\text{min}}(X^t X) \to \infty$ is the standard requirement for consistency of LS estimator in linear regression (Lai and Wei (1982)).

Alternatively, Jiang (1998) considered the EBLUE of β,

$$\hat{\beta}_{EBLUE} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} y \ , \tag{7}$$

where $V = \text{Var}(y) = \sigma_0^2 I + \sum_{r=1}^s \sigma_r^2 Z_r Z_r^t$ (I represents the identity matrix), σ_0^2 is the variance of the components of ϵ, σ_r^2 is the variance of the components of α_r, $1 \le r \le s$; and $\hat{V}$ is V with the variance components σ_r^2, $0 \le r \le s$, replaced by their estimators. He showed that if the restricted maximum likelihood (REML) estimators $\hat{\sigma}_r^2$, $0 \le r \le s$, are used, $\hat{\beta}_{EBLUE}$ is consistent, and $\hat{\beta}_{EBLUE}$ and $\hat{\sigma}_r^2$, $0 \le r \le s$, are jointly asymptotically normal. Again, the result does not require normality or the standardness of the mixed model, and the REML estimators in non-normal cases are defined as M-estimators, i.e., they save the REML equations derived under normality. Similar results hold when the REML estimators are replaced by the maximum likelihood (ML) estimators. See Jiang (1996) for asymptotic properties of the REML and ML estimators in the non-normal cases.

Let y_* be a future observation which we wish to predict. Suppose that y_* satisfies a standard mixed linear model with the same structure as (2), but with possibly different covariates. Thus, by (3), y_* can be expressed as

$$y_* = x_*^t\beta + \alpha_{*1} + \cdots + \alpha_{*s} + \epsilon_* \ ,$$

where x_* is a known vector of covariates (not necessarily present with the data), α_{*r}'s are random effects, and ϵ_* is an error, such that $\alpha_{*r} \sim F_r$, $1 \le r \le s$, $\epsilon_* \sim F_0$, and $\alpha_{*1}, \ldots, \alpha_{*s}, \epsilon_*$ are independent. According to the discussion in Section 1.1, we assume that the future observations are independent of the current ones, thus y_* is independent of $y = (y_j)_{1\le j\le N}$. It follows that the best (point) predictor of y_*, when β is known, is $E(y_*|y) = E(y_*) = x_*^t\beta$. Because β is unknown, it is naturally replaced by $\hat\beta$, resulting in the empirical best predictor:

$$\hat y_* = x_*^t\hat\beta \ . \tag{8}$$

Let $\hat\delta_i = y_i - x_i^t\hat\beta$. Define

$$\hat F(x) = \frac{\#\{1 \le i \le N : \hat\delta_i \le x\}}{N} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} 1_{(\hat\delta_i \le x)} \ . \tag{9}$$

Note that, although (9) resembles the empirical distribution, it is not one in the classic sense because the $\hat\delta_i$'s are not independent (the y_i's are dependent, and $\hat\beta$ depends on all the data). Let $\hat a < \hat b$ be any numbers satisfying

$$\hat F(\hat b) - \hat F(\hat a) = 1 - \alpha \ . \tag{10}$$

Then, a prediction interval for y_* with asymptotic coverage probability $1-\alpha$ is given by

$$[\hat y_* + \hat a,\ \hat y_* + \hat b] \ . \tag{11}$$

Note 1. A "typical" choice has

$$\hat F(\hat a) = \frac{\alpha}{2}\ , \quad \hat F(\hat b) = 1 - \frac{\alpha}{2} \ . \tag{12}$$

Another choice would be to select $\hat a$ and $\hat b$ to minimize $\hat b - \hat a$, the length of the prediction interval. Usually, $\hat a$, $\hat b$ are selected such that the former is negative and the latter positive, so that $\hat y_*$ is contained in the interval.

Note 2. If one considers linear regression as a special case of the mixed linear model, in which the middle terms ξ_i (see (4)) disappear, then $\hat\delta_i$ is the same as $\hat\epsilon_i$, the *residual*, if $\hat\beta$ is the LS estimator. In this case, $\hat F$ is the empirical distribution of the residuals, and the prediction interval (11) corresponds to that obtained by the *bootstrap* method (Efron (1979)). The difference is that our prediction interval (11) is obtained in closed form rather than by a Monte Carlo method. For more discussion on bootstrap prediction intervals, see Shao and Tu (1995), §7.3.

Although (11) is defined as a prediction interval, one has to demonstrate that its asymptotic coverage probability is, indeed, $1 - \alpha$. This is shown by the following theorem. Let $z_i = (z_{i1}^t \cdots z_{is}^t)^t$, and $S_N = \{(i,j) : 1 \le i,j \le N, z_i^t z_j > 0\}$. Note that S_N is the set of all pairs (i,j) such that $z_{ir} = z_{jr}$ for some $1 \le r \le s$. Let $|S|$ represent the cardinality of a set S.

Theorem 2. *Suppose that* (i) *the mixed model is standard;* (ii) F *is continuous;* (iii) $|x_i|$, $1 \le i \le N$, *are bounded; and* (iv) $|S_N|/N^2 \to 0$ *as* $N \to \infty$. *Then, for any consistent estimator* $\hat{\beta}$,

$$P(\hat{y}_* + \hat{a} \le y_* \le \hat{y}_* + \hat{b}) \longrightarrow 1 - \alpha\ , \quad \text{as } N \to \infty\ . \tag{13}$$

The proof is given in the Appendix.

Note that F is continuous if one of the distributions $F_0, F_1, \ldots, F_s$ is (this follows from the convolution formula, e.g., Billingsley (1986), page 272).

We use examples to illustrate condition (iv) of Theorem 2.

Example 4. Consider the mixed linear model $y_{ij} = x_{ij}^t\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, where the α_i's are i.i.d. random effects with mean 0 and distribution F_1, ϵ_{ij}'s are i.i.d. errors with mean 0 and distribution F_0, and α_i's and ϵ_{ij}'s are independent. It is clear that the model is standard with $Z_1 = I_m \otimes 1_n$, where I_m and 1_n represent the m-dimensional identity matrix and the n-dimensional vector of 1's, respectively ($\otimes$ means Kronecker product). Note that here the index i is replaced by the multiple index (i,j), and $N = mn$. It follows that $S_N = \{((i,j),(i,j')) : 1 \le i \le m, 1 \le j, j' \le n\}$. Thus, $|S_N|/N^2 = 1/m \to 0$ if and only if $m \to \infty$. Note that the result holds regardless of n. In particular, it holds when $n = 1$. It is interesting to note that in the latter case F_0 and F_1 are not identifiable, but one could still construct a prediction interval with an asymptotically correct coverage probability.

Example 5. Consider the mixed linear model $y_{ij} = x_{ij}^t\beta + u_i + v_j + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, where the u_i's are i.i.d. random effects with mean 0 and distribution F_1, v_j's are i.i.d. random effects with mean 0 and distribution F_2, e_{ij}'s are i.i.d. errors with mean 0 and distribution F_0, and u, v, and e are independent. Again, this is a standard mixed model with $Z_1 = I_m \otimes 1_n$ and $Z_2 = 1_m \otimes I_n$. Furthermore, $S_N = \{((i,j),(i',j')) : 1 \le i,i' \le m, 1 \le j,j' \le n, i = i' \text{ or } j = j'\}$. Thus, $|S_N|/N^2 \le (mn^2 + m^2n)/(mn)^2 = 1/m + 1/n \to 0$ if and only if $m, n \to \infty$.

The finite-sample performance of the prediction intervals will be investigated in Section 4. In particular, comparison between the prediction interval based on $\hat{\beta}_{OLS}$ and that based on $\hat{\beta}_{EBLUE}$ will be considered.

3. Nonstandard Cases

Although most mixed linear models used in practice are standard, nonstandard mixed models are also used. In this section, we consider how to construct prediction intervals in such cases.

First, the method developed in the previous section may be applied to some of the nonstandard cases. To illustrate this, consider the following example.

Example 6. Suppose that the data is divided into two parts. For the first part, we have $y_{ij} = x_{ij}^t\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \dots, m$, $j = 1, \dots, n_i$, where $\alpha_1, \dots, \alpha_m$ are i.i.d. random effects with mean 0 and distribution F_1; ϵ_{ij}'s are i.i.d. errors with mean 0 and distribution F_0, and the α's and ϵ's are independent. For the second part of the data, we have $y_k = x_k^t\beta + \epsilon_k$, $k = N+1, \dots, N+K$, where $N = \sum_{i=1}^m n_i$, and the ϵ_k's are i.i.d. errors with mean 0 and distribution F_0. Note that the random effects only appear in the first part of the data (and hence there is no need to use a double index for the second part).

For the first part, let the distribution of $\delta_{ij} = y_{ij} - x_{ij}^t\beta$ be F $(= F_0 * F_1)$. For the second part, let $\delta_k = y_k - x_k^t\beta$. If β were known, the δ_{ij}'s (δ_k's) would be sufficient statistics for F (F_0). Therefore it suffices to consider an estimator of F (F_0) based on the δ_{ij}'s (δ_k's). Note that the prediction interval for any future observation is determined either by F or by F_0, depending on which part the observation corresponds to. Now, since β is unknown, it is customary to replace it by $\hat\beta$. Thus, a prediction interval for y_*, a future observation corresponding to the first part, is

$$[\hat y_* + \hat a, \hat y_* + \hat b]\ , \tag{14}$$

where $\hat y_* = x_*^t\hat\beta$, $\hat a$, $\hat b$ are determined by (10) with

$$\hat F(x) = \frac{1}{N}\#\{(i,j) : 1 \le i \le m, 1 \le j \le n_i, \hat\delta_{ij} \le x\} \tag{15}$$

and $\hat\delta_{ij} = y_{ij} - x_{ij}^t\hat\beta$. Similarly, a prediction interval for y_*, a future observation corresponding to the second part, is

$$[\hat y_* + \hat a, \hat y_* + \hat b]\ , \tag{16}$$

where $\hat y_* = x_*^t\hat\beta$, $\hat a$, $\hat b$ are determined by (10) with $\hat F$ replaced by

$$\hat F_0(x) = \frac{1}{K}\#\{k : N+1 \le k \le N+K, \hat\delta_k \le x\} \tag{17}$$

and $\hat\delta_k = y_k - x_k^t\hat\beta$. It can be argued, as before, that the prediction interval (14) ((16)) has asymptotic coverage probability $1-\alpha$, as $N \to \infty$ ($K \to \infty$).

If we look more carefully, we see that the model in Example 6 can be divided into two standard submodels, so that the method of Section 2 is applied to each submodel, and, because of the sufficient statistics, there is not much loss of efficiency in doing so. Of course, not every nonstandard mixed linear model can be divided into standard submodels. For example, some of the z_{ir}'s in (2) may involve covariates other than 0 and 1, such as in a random slope model. For such nonstandard models we consider a different approach, described as follows.

Jiang (1998) has considered estimation of the distributions of the random effects and errors. His approach is the following. Consider the empirical best linear unbiased predictors (EBLUP) of the random effects:

$$\hat{\alpha}_r = \hat{\sigma}_r^2 Z_r^t \hat{V}^{-1}(y - X\hat{\beta}) \ , \quad 1 \leq r \leq s \ , \tag{18}$$

where $\hat{\beta}$ is the EBLUE (see (7), and the definition of $\hat{V}$ below it), and the "EBLUP" for the errors

$$\hat{\epsilon} = y - X\hat{\beta} - \sum_{r=1}^{s} Z_r \hat{\alpha}_r \ . \tag{19}$$

Jiang (1998) showed that, if the REML or ML estimators of the variance components are used, then, under suitable conditions,

$$\hat{F}_r(x) = \frac{1}{m_r} \sum_{u=1}^{m_r} 1_{(\hat{\alpha}_{r,u} \leq x)} \xrightarrow{P} F_r(x) \ , \quad x \in CF_r \ , \tag{20}$$

where $\hat{\alpha}_{r,u}$ is the uth component of $\hat{\alpha}_r$, $1 \leq r \leq s$, and

$$\hat{F}_0(x) = \frac{1}{N} \sum_{i=1}^{N} 1_{(\hat{\epsilon}_i \leq x)} \xrightarrow{P} F_0(x) \ , \quad x \in CF_0 \ , \tag{21}$$

where $\hat{\epsilon}_i$ is the ith component of $\hat{\epsilon}$. Here CF_r represents the set of all continuity points of F_r, $0 \leq r \leq s$.

For simplicity, in the following we assume that all the distributions $F_0, \ldots, F_s$ are continuous. Let y_* be a future observation we would like to predict. As before, we assume that y_* is independent of y and satisfies a mixed linear model with the same structure as (2), but with possibly different covariates. The latter means that y_* can be expressed as

$$y_* = x_*^t \beta + \sum_{j=1}^{l} w_j \gamma_j + \epsilon_* \ , \tag{22}$$

where x_* is a known vector of covariates (not necessarily present with the data); w_j's are known nonzero constants; γ_j's are unobservable random effects; ϵ_* is an error. In addition, there is a partition of the indices $\{1, \ldots, l\} = \cup_{k=1}^{q} I_k$, such that $\gamma_j \sim F_{r(k)}$ if $j \in I_k$, where $r(1), \ldots, r(q)$ are distinct integers between 1 and s (so $q \leq s$); $\epsilon_* \sim F_0$; $\gamma_1, \ldots, \gamma_l, \epsilon_*$ are independent. Define

$$\hat{F}^{(j)}(x) = m_{r(k)}^{-1} \sum_{u=1}^{m_{r(k)}} 1_{(w_j \hat{\alpha}_{r(k),u} \leq x)} \ , \quad \text{if } j \in I_k \tag{23}$$

for $1 \leq k \leq q$. Let

$$\begin{aligned} \hat{F}(x) &= (\hat{F}^{(1)} * \cdots * \hat{F}^{(l)} * \hat{F}_0)(x) \\ &= \frac{\#\{(u_1, \ldots, u_l, i) : \sum_{k=1}^q \sum_{j \in I_k} w_j \hat{\alpha}_{r(k),u_j} + \hat{\epsilon}_i \leq x\}}{\left(\prod_{k=1}^q m_{r(k)}^{|I_k|}\right) N} \ , \end{aligned} \tag{24}$$

where $*$ represents convolution; $1 \leq u_j \leq m_{r(k)}$ if $j \in I_k$, $1 \leq k \leq q$; $1 \leq i \leq N$. Then it is easy to show that (20) and (21) imply

$$\sup_x |\hat{F}(x) - F(x)| \xrightarrow{P} 0 \ , \tag{25}$$

where $F = F^{(1)} * \cdots * F^{(l)} * F_0$, and $F^{(j)}$ is the distribution of $w_j \gamma_j$, $1 \leq j \leq l$. Note that F is the distribution of $y_* - x_*^t \beta$. Let $\hat{y}_*$ be defined at (8) with $\hat{\beta}$ a consistent estimator, and $\hat{a}$, $\hat{b}$ by (10), where $\hat{F}$ is given by (24). Then, by the same argument as the first part of the proof of Theorem 2, it can be shown that the prediction interval

$$[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}] \tag{26}$$

has asymptotic coverage probability $1 - \alpha$.

4. A Simulated Example

In this section we consider the mixed linear model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i + \epsilon_{ij} \ , \tag{27}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where the α_i's are i.i.d. random effects with mean 0 and distribution F_1, and ϵ_{ij}'s are i.i.d. errors with mean 0 and distribution F_0. This model may be considered as associated with a sample survey, where α_i is a random effect related to the ith family in the sample, and n_i is the sample size for the family (e.g., the family size, if all family members are to be surveyed). The x_{ij}'s are covariates associated with the individuals sampled from the family and, in this case, correspond to people's ages. The ages are categorized by the following groups: $0-4$, $5-9$, $\ldots$, $55-59$, so that $x_{ij} = k$ if the person's age falls into the kth category (people whose ages are 60 or over are not included in the survey). The true parameters for β_0 and β_1 are 2.0 and 0.2, respectively.

In the following simulations, four combinations of the distributions F_0, F_1 are considered. These are Case I: $F_0 = F_1 = N(0,1)$; Case II: $F_0 = F_1 = t_3$; Case III: $F_0 =$ logistic (the distribution of $\log\{U/(1-U)\}$, where $U \sim \text{Uniform}(0,1)$), $F_1 =$ centralized lognormal (the distribution of $e^X - \sqrt{e}$, where $X \sim N(0,1)$); Case IV: $F_0 =$ double exponential (the distribution of $X_1 - X_2$, where X_1, X_2 are independent $\sim$ exponential(1)), $F_1 =$ a mixture of $N(-4,1)$ and $N(4,1)$ with equal probability. Note that Case II - IV are related to the following types of departure from normality: heavy-tail, asymmetry, and bimodal. In each case, the following sample size configuration is considered: $m = 100$; $n_1 = \cdots = n_{m/2} = 2$, and $n_{m/2+1} = \cdots = n_m = 6$. Finally, for each of the above cases, three prediction intervals are considered. The first is the prediction interval based on the OLS estimator of β; the second is that based on the EBLUE of β, where the variance components are estimated by REML (see discussion in Section 2); the third is the linear regression (LR) prediction interval (e.g., Casella and Berger (1990), page 576-577), which assumes that the observations are independent and normally distributed. The third one is considered here for comparison.

For each of the four cases, 1000 data sets are generated. First, the following are independently generated: (i) x_{ij}, $1 \le i \le m$, $1 \le j \le n_i$, uniformly from the integers $1, \ldots, 12$ (twelve age categories); (ii) α_i, $1 \le i \le m$, from F_1; (iii) ϵ_{ij}, $1 \le i \le m$, $1 \le j \le n_i$, from F_0. Then y_{ij} is obtained from (27) with β_0, β_1 being the true parameters. Because of the way that the data is generated, conditional on the x_{ij}'s, the y_{ij}'s satisfy (27) with its distributional assumptions. For each data set generated, and for each of the twelve age categories, three prediction intervals are obtained according to (12), where $\alpha = 0.10$ (nominal level 90%): OLS, EBLUE, and LR; then one additional observation is generated, which corresponds to a future observation in that category. The percentages of coverage and average lengths of the intervals over the 1000 data sets are reported.

The results are given in Table 1, in which the letters O, E, and L stand for OLS, EBLUE, and LR, respectively. The numbers shown in the table are coverage probabilities based on the simulations, in terms of percentages, and average lengths of the prediction intervals. Note that for OLS and EBLUE the length of the prediction intervals do not depend on the covariates (if they are all determined by (12)), while for LR the length of the prediction interval depends on the covariate, but will be almost constant if the sample size is large. This, of course, follows from the definition of the prediction intervals, but there is also an intuitive interpretation. Consider, for example, the normal case. The distribution of a future observation y_* corresponding to a covariate x_* is $N(\beta_0 + \beta_1 x_*, \sigma^2)$, where $\sigma^2 = \text{var}(\alpha_i) + \text{var}(\epsilon_{ij})$ is a constant. So, if the β's were known the length of any prediction interval for y_* would not depend on x_*. If the β's are unknown but replaced by consistent estimators, then if the sample size is large, one also expects the length of the prediction interval to be almost constant (not dependent on x_*). For such a reason, there is no need to exhibit the lengths of the prediction intervals for different categories, and we only give the averages over all categories.

Table 1.

Coverage Probability (%)												
	Case I			Case II			Case III			Case IV		
x	O	E	L	O	E	L	O	E	L	O	E	L
1	90	90	90	89	89	92	90	91	93	90	90	94
2	90	90	90	89	89	91	91	91	93	89	90	96
3	88	88	88	91	91	93	90	89	92	88	89	96
4	90	90	89	91	91	93	89	89	91	89	89	97
5	89	89	89	89	89	92	90	90	92	90	90	96
6	89	89	90	89	89	92	91	91	93	90	90	97
7	89	88	89	90	90	92	90	90	93	88	89	96
8	90	90	90	90	90	92	89	89	91	90	90	97
9	90	90	91	89	89	92	89	89	91	89	89	96
10	89	89	90	91	90	93	89	89	93	88	88	95
11	90	90	90	89	89	93	89	89	92	89	89	97
12	89	89	89	89	89	92	91	91	93	89	89	96
Average Length												
	4.6	4.6	4.7	7.0	7.0	7.9	8.1	8.1	9.0	12.1	12.1	14.3

It is seen that in the normal case there is not much difference among all three methods. This is not surprising. The difference appears in the non-normal cases. First, the LR prediction intervals are wider than the OLS and EBLUE ones. Second, as a consequence, the coverage probabilities for the LR prediction intervals seem to be higher than 90%. Overall, the OLS and EBLUE perform better than LR in the non-normal cases. This is not surprising, because the OLS and EBLUE prediction intervals are distribution-free. The EBLUE does not seem to do better than the OLS. This was a bit unexpected. On the other hand, it shows that at least in this special case the OLS, although much simpler than the EBLUE in that one does not need to estimate the variance components, can do just as well as more sophisticated methods such as the EBLUE.

5. An Application: Lead Contamination of Soil

Childhood lead poisoning has been declared by the U.S. Public Health Service to be “the most common and societally devastating environmental disease of young children”. Lead poisoning’s role in reduced intelligence, poor school performance, and impaired social functioning has been well established. It is believed that lead poisoning is largely due to exposure to lead hazards from deteriorated lead paint and lead-contaminated bare soil in substandard housing. Children living in well-maintained housing can be poisoned by lead-contaminated soil and by lead dust created by unsafe repainting and renovation practices that disturb lead-based paint.

The data for this example came from a survey which was part of a joint project by the Ohio Air Quality Development Authority and Case Western Reserve University. The overall goal for the project was to assess the sources, nature, and extent of lead contamination of residential soils in Cleveland. Sixty-nine houses were randomly selected from the area. For each selected house, thirteen samples of soil were taken: one from the foundation, twelve from the backyard of the house. Overall, a total of 726 samples were collected. Concentration of lead in those samples were then measured in a laboratory. It was found that there was a tremendous amount of between-house variation as well as within-house variation in the lead concentration. To model such variations, the following nested error regression model, a special case of the mixed linear model (1), was proposed:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 p_i + \beta_3 x_{ij} \cdot p_i + \beta_4 w_i + \alpha_i + \epsilon_{ij} \ , \tag{28}$$

$i = 1, \dots, 77$, $j = 1, \dots, n_i$, where y_{ij} is the measure of lead concentration (in unit of parts per million, or PPM) for the jth sample from the ith house; the β's are unknown regression coefficients; $x_{ij} = 1$ if the measure was taken from the foundation, and 0 otherwise; $p_i = 0$, 1, 2, or 3, if the house is not painted, painted in good condition, painted in fair condition, or painted in poor condition; $w_i = -1$ or 1 if the house was built before 1950 or after 1950, and $w_i = 0$ if that information is not available; α_i represents a house-specific random effect, and ϵ_{ij} is a random error which corresponds to within-house variation. Note that an interaction between x_{ij} and p_i is included. It is assumed that the random effects are independent with an unknown distribution F, the errors are independent with an unknown distribution G, and the random effects and errors are independent. The sample sizes n_i were supposed to be 13 but, in reality, many were less.

One problem of practical interest is to predict the level of lead concentration at a randomly selected house given the information of age of the house and paint condition. It would also be interesting to see if there is a difference in lead concentration between the foundation and the backyard, and whether that interacts with the paint condition. Since the mixed linear model (28) is standard, we may apply the method developed in Section 2 to obtain prediction intervals. Two methods are used in estimating the β's: OLS and EBLUE, where for EBLUE the REML estimators of the variance components are used. Table 2 gives the estimates of parameters, where the numbers in parentheses are the corresponding p-values (here 10^{-4} means that the p-value is less than 10^{-4}).

Note. Although the OLS estimator (5), as a point estimator, is the same as the LS estimator, its standard errors are different from those obtained in linear regression. One has

$$\mathrm{Var}(\hat{\beta}_{OLS}) = (X^t X)^{-1} X^t V X (X^t X)^{-1} , \tag{29}$$

where V is given below (7) (e.g., Diggle, Liang, and Zeger (1996), §4.3).

Table 2. Estimates of Parameters

	β_0	β_1	β_2	β_3	β_4	σ_α	σ_ϵ
OLS	228.8	68.9	77.8	524.1	-145.0		
	(.0002)	(.29)	(.09)	(10^{-4})	(.01)		
EBLUE	230.9 (.0002)	48.5 (.37)	54.4 (.21)	570.9 (10^{-4})	-137.6 (.01)	306.0 (10^{-4})	258.1 (10^{-4})

It is seen that the OLS and EBLUE estimates of the regression coefficients that are significant (at 5 percent level) are quite close. It is also notable that, although the effects of location of the sample (foundation/backyard) and paint conditions are found individually insignificant, their interaction is highly significant with a large (in absolute value) coefficient, β_3. This suggests that the location of the sample only matters for painted houses, indicating that the paint is a main source of lead in soil. Based on the above estimates, 90% prediction intervals are obtained for each combination of levels of the covariates. Table 3 shows the results for houses that were built prior to 1950.

Table 3. Selected Prediction Intervals (Pre '50)

Covariate			Prediction Interval	
x	p	w	OLS	EBLUE
1	0	-1	[53, 1164]	[49, 1161]
0	0	-1	[0, 1095]†	[1, 1113]
1	1	-1	[655, 1766]	[674, 1787]
0	1	-1	[62, 1173]	[55, 1167]
1	2	-1	[1257, 2367]	[1300, 2412]
0	2	-1	[140, 1250]	[109, 1222]
1	3	-1	[1859, 2969]	[1925, 3037]
0	3	-1	[218, 1328]	[164, 1276]

† Lower end is truncated because it is negative ($= -15$).

The prediction intervals confirm an earlier speculation that there is a large difference in lead concentration between foundation and backyard for painted houses; however, such a difference is not significant for houses that are not painted. Note that, unless the house is either not painted or painted but in good condition, the prediction intervals for foundation and for backyard do not overlap.

6. Concluding Remarks

Distribution-free methods play, perhaps, a much more significant role in prediction intervals than they do in confidence intervals, especially in large samples. Central limit theorems are often useful in constructing confidence intervals, but they may not help in obtaining prediction intervals. For example, when predicting a single future observation, as is in this paper, the prediction interval for a future observation is governed by the actual distribution of the observation, not by an approximate normal distribution, unless the observation itself is normal; and this is true whether the sample size is large or small.

Despite the well-known fact that methods developed in regression analysis do not necessarily apply to mixed linear models, a simple method based on the residuals prevails in obtaining prediction intervals for standard mixed linear models. The latter, in fact, are credited with the majority of mixed linear model applications. Although the EBLUE method has the potential of making an improvement, this is not seen in our simulation study.

Acknowledgement

The research of the first author was supported, in part, by National Security Agency Grant MDA904-98-1-0038.

A. Appendix: Proofs

Proof of Theorem 1. Write $\hat{\beta} = \hat{\beta}_{OLS}$. Because $E(\hat{\beta}) = \beta$, we have

$$\begin{aligned} E|\hat{\beta} - \beta|^2 &= E\text{tr}((\hat{\beta} - \beta)(\hat{\beta} - \beta)^t) \\ &= \text{tr}(\text{Var}(\hat{\beta})) \\ &= \text{tr}((X^tX)^{-1}X^tVX(X^tX)^{-1}) \ , \end{aligned}$$

where V is given below (7). It follows that $V \le \lambda I$, where $\lambda = \sigma_0^2 + \sum_{r=1}^s \sigma_r^2 \lambda\text{max}(Z_r Z_r^t) = \sigma_0^2 + \sum_{r=1}^s \sigma_r^2 \lambda\text{max}(Z_r^t Z_r)$. Thus,

$$\begin{aligned} \text{tr}((X^tX)^{-1}X^tVX(X^tX)^{-1}) &\le \lambda \text{tr}((X^tX)^{-1}) \\ &\le p\left[\frac{\sigma_0^2}{\lambda_{\text{min}}(X^tX)} + \sum_{r=1}^s \sigma_r^2 \frac{\lambda\text{max}(Z_r^t Z_r)}{\lambda_{\text{min}}(X^tX)}\right] \ . \end{aligned}$$

The result follows.

Proof of Theorem 2. By standard arguments, it suffices to show $\hat{F}(x) \to F(x)$ in probability for each x. For any $\eta > 0$, we have

$$\begin{aligned} \hat{F}(x) &= \frac{1}{N}\sum_{i=1}^N \left[1_{(\hat{\delta}_i \le x)} - 1_{(\delta_i \le x+\eta)}\right] + \frac{1}{N}\sum_{i=1}^N \left[1_{(\delta_i \le x+\eta)} - F(x+\eta)\right] + F(x+\eta) \\ &= I_1 + I_2 + F(x+\eta) \ . \end{aligned} \tag{A.1}$$

If $1_{(\hat{\delta}_i \le x)} > 1_{(\delta_i \le x+\eta)}$, then $\hat{\delta}_i \le x$ and $\delta_i > x+\eta$. Thus, $x_i^t(\hat{\beta}-\beta) = \delta_i - \hat{\delta}_i > \eta$. Therefore, assuming $|x_i| \le B$ where $B > 0$, we have $B|\hat{\beta}-\beta| \ge |x_i^t(\hat{\beta}-\beta)| > \eta$, or $|\hat{\beta}-\beta| > \eta/B$. It follows that

$$\begin{aligned} P(I_1 > 0) &\le P\left(1_{(\hat{\delta}_i \le x)} > 1_{(\delta_i \le x+\eta)} \text{ for some } i\right) \\ &\le P(|\hat{\beta}-\beta| > \eta/B) \quad \longrightarrow \quad 0 \, . \end{aligned} \tag{A.2}$$

On the other hand, let $\alpha = (\alpha_1^t \cdots \alpha_s^t)^t$. We have

$$\begin{aligned} I_2 &= \frac{1}{N}\sum_{i=1}^N \left[1_{(\delta_i \le x+\eta)} - P(\delta_i \le x+\eta|\alpha)\right] + \frac{1}{N}\sum_{i=1}^N [P(\delta_i \le x+\eta|\alpha) - F(x+\eta)] \\ &= I_{21} + I_{22} \, . \end{aligned} \tag{A.3}$$

By conditional independence,

$$\begin{aligned} E(I_{21}^2) &= \frac{1}{N^2} E\Big\{E\Big[\Big(\sum_{i=1}^N (\cdots)\Big)^2 \,\Big|\, \alpha \,\Big]\Big\} \\ &= \frac{1}{N^2} E\Big\{\sum_{i=1}^N \text{var}\left(1_{(\delta_i \le x+\eta)}|\alpha\right)\Big\} \\ &\le \frac{1}{4N} \quad \longrightarrow \quad 0 \, . \end{aligned} \tag{A.4}$$

Also, it is easy to show that $P(\delta_i \le x+\eta|\alpha) = F_0(x+\eta-\xi_i)$, where ξ_i is defined by (4), and $F(x+\eta) = EF_0(x+\eta-\xi_i)$. Note that, if $z_{ir}^t z_{jr} = 0$ (i.e., z_{ir} and z_{jr} have 1 in different places), $1 \le r \le s$, or, equivalently, if $z_i^t z_j = 0$, then ξ_i and ξ_j are independent. It follows that

$$\begin{aligned} E(I_{22}^2) &= \frac{1}{N^2}\sum_{i,j=1}^N \text{cov}(F_0(x+\eta-\xi_i), F_0(x+\eta-\xi_j)) \\ &= \frac{1}{N^2}\sum_{z_i^t z_j > 0} \text{cov}(F_0(x+\eta-\xi_i), F_0(x+\eta-\xi_j)) \\ &\le \frac{|S_N|}{4N^2} \quad \longrightarrow \quad 0 \, . \end{aligned} \tag{A.5}$$

Combining (A.1) — (A.5), we have $\hat{F}(x) \le F(x+\eta) + o_P(1)$, where $o_P(1)$ represents a term that $\to 0$ in probability. Similarly, we have $\hat{F}(x) \ge F(x-\eta) + o_P(1)$. It follows, by the arbitrariness of η and continuity of F, that

$$\hat{F}(x) - F(x) \xrightarrow{P} 0 \, . \tag{A.6}$$

It then follows (e.g., Chow and Teicher (1997), page 283) that (A.6) remains true with the left side replaced by $\sup_x |\hat{F}(x) - F(x)|$. The rest of the proof is easy to complete.

References

Baker, G. A. (1935). The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample. *Ann. Math. Statist.* **6**, 197-201.

Billingsley, P. (1986). *Probability and Measure*, 2nd Edition. John Wiley, New York.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Wadsworth/Brooks Cole, Pacific Grove, CA.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory*, 3rd Edition. Springer-Verlag, New York.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1996). *Analysis of Longitudinal Data.* Oxford Univ. Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statist. Sci.* **9**, 55-93.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals - A Guide for Practitioners.* John Wiley, New York.

Jeske, D. R. and Harville, D. A. (1988). Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models. *Commun. Statist. - Theory Meth.* **17**, 1053-1087.

Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *Ann. Statist.* **24**, 255-286.

Jiang, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statist. Sinica* **8**, 861-885.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10**, 154-166.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann. Statist.* **5**, 746-762.

Patel, J. K. (1989). Prediction intervals - A review. *Commun. Statist. - Theory Meth.* **18**, 2393-2465.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer, New York.

Zhou, L. (1997). Nonparametric prediction intervals. Ph. D. dissertation, Univ. of Calif. at Berkeley, Berkeley, CA.

Department of Statistics, University of California, One Shields Avenue, Davis, CA95616, U.S.A.

E-mail: jiang@wald.ucdavis.edu

Cancer Detection Section, California Department of Health Services, 601 North 7th street, MS428, P.O. Box 942732, Sacramento, CA94234-7320, U.S.A.

E-mail: wzhang@dns.ca.gov

(Received June 2000; accepted November 2001)