Bootstrap Confidence Interval of Prediction for Small Area Estimation Based on Linear Mixed Model

F Novkaniza\textsuperscript{1,2}, K A Notodiputro\textsuperscript{2,*}, B Sartono\textsuperscript{2}

\textsuperscript{1}Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor 16680, Indonesia
\textsuperscript{2}Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Indonesia, Depok 16424, Indonesia
\*email: khairlnotodiputro@gmail.com

Abstract. Linear Mixed Model (LMM) analyzes the relationship between Gaussian response and predictors with either fixed and random effects. Procedures based on LMM have been used to construct estimates of the means of small areas, by exploiting auxiliary information. In this article, we show how to resample fixed effects coefficient estimates via bootstrapping and we construct nonparametric and parametric bootstrap confidence interval of predictions for small area estimation, based on mixed-effects linear models. Examples of computation for bootstrap confidence intervals of prediction are given for Battese, Harter and Fuller Data (1988).

Keywords: mixed model, prediction interval, parametric bootstrap, nonparametric bootstrap

1. Introduction
In Generalized Linear Model (GLM), we can use the predictor function to get the standard error for the predicted values on either the observed data or on new data [1]. It is then used to construct confidence or prediction intervals around the fitted regression curve. The confidence intervals focus on the regression curve and suppose for a 95% significant level can be interpreted as follows: “If we would repeat our sampling n times, the regression curve would fall between this interval 95% of the time.” On the other hand, a prediction interval focuses on single data points and could be interpreted as: “If we would sample n times at these values for the explanatory variables, the response value would fall between this interval 95% of the time”. There are two types of confidence intervals used for predictions in regression and other linear models i.e prediction interval and confidence interval of the prediction. Prediction interval represents a range that a single new observation is likely to fall given specified settings of the predictors. Confidence interval of the prediction represents a range that the mean response is likely to fall given specified settings of the predictors.

Statistical models containing fixed effects and random effects are called mixed models. Mixed-effects models, like many other types of statistical models, describe a relationship between a response variable and some of the covariates that have been measured or observed along with the response. In mixed-effects models at least one of the covariates is a categorical covariate representing experimental or observational “units” in the dataset. For Generalized Linear Mixed Model (GLMM), the predict function does not allow one to derive standard error, the reason being that there is no option for computing standard errors of predictions because it is difficult to define an efficient method that
incorporates uncertainty in the variance parameters. This means there is no way to include in the computation of the standard error for predicted values, the fact that the fitted random effect standard deviation is just estimated and may be well estimated. However, we can still derive confidence or prediction intervals that we might underestimate the uncertainty around the estimates.

Most researchers are interested in constructing reliable estimates for areas with small sample sizes, where areas often refer to geographic areas and demographic groups. The estimation for such areas is known as small area estimation (SAE). SAE concept was developed by statisticians as a methodology related to estimation for the survey area or domain that has small samples or even no sample. Data obtained through appropriate survey techniques will be highly effective and have the reliability properties to predict the total or mean of variables. The property of such estimators can be achieved if the sample data from the survey includes large areas or domains. A small area estimation is a method for predicting parameters in a subpopulation with small sample sizes. The method developed in the estimation of small areas is an indirect method of estimation by utilizing the strength of the surrounding area and data sources outside the area.

In the small area models, the area specific random effects explain the between area variations in the data which is not explained by the fixed effects part of the model. One of the most developed methods for estimating small areas is a model based on GLMM. Mixed models are suitable for small area estimation because they combine different sources of information and contain different sources of error. Prediction methods for the small area mean and confidence intervals (CI’s) of prediction for the small area means are presented for the case when the response variable is normal. SAE has two basic model types: basic area-level model and basic unit-level model [9]. The fundamental difference between the two models is on the use of available support data. In SAE area-level models, supporting data are available only for certain area levels. This model connects the direct estimator with the corresponding variable from the other domain for each area, whereas the unit-level model assumes that the available common variables correspond individually with the response variable. Battese, Harter, and Fuller (1988) [3] use a linear mixed model to predict the area planted with corn and soybeans in Iowa counties.

Bootstrap is a method based on data simulations for statistical inference purposes. Bootstrap is used to enable inference on the statistic of interest when the true distribution of this statistic is unknown. For example, in a linear model, the parameter of interest has a known distribution from which standard errors and formal tests can be performed. On the other hand, for some statistics, if we do not want to spend time writing down equations, bootstrapping might be a great approach to get standard errors and confidence intervals from the bootstrapped distribution. Bootstrap is one of the most famous resampling technique and is very useful to get confidence intervals in situations where classical approach would fail. Bootstrap methods were introduced to construct confidence intervals in an algorithmic fashion, using fewer assumptions than those based on the normal approximation. Also, bootstrap confidence intervals can be constructed for complicated models and data structures.

Efron [6] introduced the bootstrap as a nonparametric tool for estimating standard errors and biases. Confidence intervals require more effort than parametric estimation. Methods of improvements have been developed, such as the bootstrap accelerated method, bootstrap-t, iterated bootstrap, and calibration. Most small area studies focus on constructing predictors for the area means and on estimating the variance of the prediction errors. Most studies that report confidence intervals (CIs) for the small area means, Dass et al [5], Diao et al [6] and Yoshimori and Lahiri [10]. In this article, we use nonparametric bootstrap method and parametric bootstrap method for constructing confidence intervals of prediction for the small area estimation. The most common CI is based on the estimated prediction Mean Square Error (MSE) and approximates the distribution of parameter estimates with a normal distribution. The coverage error for such an interval can be large when the distribution of the parameter estimate is skewed and when the standard error is poorly estimated. We present two-sided bootstrap CIs for prediction in small area estimation for a normal response variable with covariates.
2. Parametric and Nonparametric Bootstrap Confidence Interval of Prediction

Bootstrapping is a way of estimating statistical parameters from the sample by means of resampling with replacement. Like other non-parametric approaches, bootstrapping does not make any assumptions about the distribution of the sample. The major assumption behind bootstrapping is that the sample distribution is a good approximation to the population distribution, i.e. that the sample is representative of the population. Bootstrap is a nonparametric resampling technique aimed at determining estimated standard errors and confidence intervals from population parameters such as mean, ratio, median, proportion, correlation coefficient or regression coefficient without using the distribution assumption. The bootstrap method is used to find the sampling distribution of an estimator through a resampling procedure with the return of the original data and the same size as the original sample size. In the Bootstrap method, the original sample position is viewed as a population.

Suppose that \( X_1, X_2, ..., X_n \) is a random sample from a distribution \( F \) and that \( \theta \) is the parameter of interest. Let \( \hat{\theta} \) be a sample estimator of \( \theta \). The idea of bootstrap is to treat the sample \( X_1, X_2, ..., X_n \) as the population and to draw samples of size \( n \), with replacement, from \( X_1, X_2, ..., X_n \) denoted \( X'_1, X'_2, ..., X'_n \). The bootstrap estimate of \( \theta \) is \( \hat{\theta}^* \), a function of the bootstrap sample denoted \( X'_1, X'_2, ..., X'_n \). The procedure of drawing a sample of size \( n \) from the original sample that \( X_1, X_2, ..., X_n \), treated as the population, is called the nonparametric bootstrap, proposed by Efron [6].

In nonparametric bootstrapping for LMM, we match \( y_i \) and \( X_i \) to form the pairs of data structures \( (y_{i_1}, X_{i_1}), i = 1, ..., n \) and then draw a sample size of \( n \) with replacement from the \( n \) pairs. We denote the pairs in the bootstrap sample \( (y^*_i, X^*_i), ..., (y^n_i, X^n_i) \). Therefore, a bootstrap sample is comprised of a set of \( n \) pairs \( (y^*_i, X^*_i), i = 1, ..., n \), randomly drawn from the pairs \( (y_i, X_i), i = 1, ..., n \). In LMM, the predictor \( \eta_i \) of an individual \( i \) is defined as comprising the \( c \) associated fixed effects \( \beta \). The idea of a mixed model approach is to extend the predictor by so-called random effects. A common notation for the model equation of LMM is

\[
y_{ij} = x_{ij}^T \beta + z_{ij} \gamma + \epsilon_{ij}
\]

Suppose that the random vector \( u \) comes from a distribution in the parametric family \( \{F_\lambda: \lambda \in \Lambda\} \), where \( \lambda \) is unknown. Let \( \hat{\lambda} \) be an estimate of \( \lambda \) by \( \mathbf{u} \). Suppose we want to suspect a function \( t(F_\lambda) \) of an unknown distribution like a raw error of a particular component of \( \hat{\lambda} \), then the bootstrap estimator of \( t(F_\lambda) \) is \( t(F^*_\lambda) \). In LMM, estimation of standard errors for the fitted linear predictor \( \hat{\eta} = x^T \hat{\beta} + z^T \hat{\gamma} \) at a set of covariate values \( x \) and \( z \). Standard error to mean the square root of the mean squared error of prediction:

\[
u^2 = E[[\hat{\eta} - \eta]^2] = E \left[ (x^T (\hat{\alpha} - \alpha) + z^T (\hat{\beta} - \beta))^2 \right]
\]

The bootstrap estimate of \( \nu^2 \) is:

\[
\hat{\nu}^2 = E^*((\hat{\eta}^* - \eta^*)^2) = E^* \left[ (x^T (\hat{\alpha}^* - \alpha) + z^T (\hat{\beta}^* - \beta))^2 \right]
\]

where the “*” notation indicates that the expected value is taken with respect to the fitted distribution with parameter \( \hat{\lambda} = (\hat{\alpha}, \hat{\phi}, \hat{\gamma}) \) i.e. \( (\hat{\gamma}^*, \hat{\beta}^*) \sim F^*_\lambda \). Let \( (y^*_{i_1}, b^*_{i_1}), ..., (y^*_{i_R}, b^*_{i_R}) \) denote identically independent vectors generated from the fitted model \( F^*_\lambda \). That is, for each \( r = 1, ..., R \), \( b^*_{i_r} \sim N(0, D(\hat{\gamma})) \) and \( y^*_{i_r} \) is a “resample” consisting of \( n \) independent data values with \( \phi, \alpha \) dan \( \beta \) replaced by \( \hat{\phi}, \hat{\alpha} \) dan \( \hat{\beta} \).

3. Bootstrap Confidence Interval of Prediction For Small Area Estimation

Based on linear mixed model with a gaussian response, covariates, and random area effects, we consider bootstrap procedures to estimate the small area means and to construct confidence intervals for prediction. Suppose that the response values come from \( k \) different small areas or groups. Let \( y_{ij}, j = 1, ..., n_i \) denote the \( j \)th response in the \( i \)th small area and let \( x_{ij} \ (p \times 1) \) be a vector of known
covariates associated with the \((i,j)\)th response. The value of \(n = \sum_{i=1}^{k} n_i\) is often chosen so that characteristics of the overall population can be estimated accurately. If \(k\) is large, the number of responses per small area may be very small, even zero in some cases if the sampling scheme allows it. Thus, direct estimates of individual small area characteristics are likely to be extremely unreliable. One way of dealing with this problem is to “borrow strength” from other small areas using a LMM with linear predictor:

\[
\eta_{ij} = x_{ij}' \alpha + b_i
\]

for the \((i,j)\)th response, where the \(b_i\) is a random effect associated with the small area \(i\). It is usually assumed that the \(b_i\)'s are i.i.d normal with mean zero and variance \(\sigma_b^2\) [4].

Maximum likelihood or restricted maximum likelihood (REML) estimates of the parameters in linear mixed-effects models can be determined using the \texttt{lmer} function in the \texttt{lme4} package for R. As for most model fitting functions in R, the model is described in an \texttt{lmer} call by a formula, in this case including both fixed and random effects terms. The formula and data together determine a numerical representation of the model from which the profiled deviance or the profiled REML criterion can be evaluated as a function of some of the model parameters. The appropriate criterion is optimized, using one of the constrained optimization functions in R, to provide the parameter estimates. To generate a proper prediction interval, a prediction must account for three sources of uncertainty in mixed models: the residual (observation-level) variance, the uncertainty in the fixed coefficients, and the uncertainty in the variance parameters for the grouping factors.

We estimate confidence intervals by parametric bootstrapping, that is, by simulating data from the fitted model, refitting the model, and extracting the new estimated parameters. This task is quite straightforward since there is already a simulate method, and a refit function which re-estimates the REML parameters for new data, starting from the previous REML estimates and re-using the previously computed model structures for efficiency.

4. Application to BHF Data

BHF [3] use a LMM to estimate the area used for growing corn in 12 Iowa counties. Their response values are precise areas based on surveying a small fraction of "segments" in each county. The sample sizes, \(n_i\), range from 1 to 5 with the total sample size being \(n = 36\). In addition to the survey measurements, information on crop areas for the entire region is available from satellite images. Number of pixels of sample segments for corn and soybean are used as covariates. Thus, the complete model considered by Battese et al [3] is:

\[
y_{ij} = x_i^T \alpha + b_i + \epsilon_{ij}
\]

where:

- \(y_{ij}\) = area of growing corn in the sample \(j\)th segment in county \(i\) (hectares)
- \(x_{ij} = (x_{1ij}, x_{2ij})\) = number of pixels corn and soybean based on satellite photos
- \(\epsilon_{ij}\) are i.i.d normal errors with variance \(\sigma^2\)

The format data is represented in Table 1.
Table 1. Survey and Satellite Data for Corn and Soybean in 12 Iowa Counties

| County      | No. of Segments | Reported No. of Pixels | Reported Hectares | No. of Pixels in Sample Segments |
|-------------|-----------------|------------------------|-------------------|---------------------------------|
| Cerro Gordo | 1               | 545                    | 165.76            | 374, 55                         |
| Hamilton    | 1               | 566                    | 96.32             | 209, 218                        |
| Worth       | 1               | 394                    | 76.08             | 253, 250                        |
| Humbolt     | 2               | 424                    | 185.35            | 432, 96                         |
|             |                 |                        | 116.43            | 367, 178                        |
| Franklin    | 3               | 564                    | 162.08            | 361, 137                        |
|             |                 |                        | 152.04            | 288, 165                        |
|             |                 |                        | 161.75            | 369, 165                        |
| Pocahontas  | 3               | 570                    | 92.88             | 206, 218                        |

Using nonparametric bootstrap, we drew randomly from the 36 pairs of observations, 10 new bootstrap samples with replacement. We denote the pairs in the bootstrap sample \((y_{11}^*, x_{11}^*), \ldots, (y_{125}^*, x_{125}^*)\). For each bootstrap sample, we fitted LMM with corn pixels and soybean pixels are used as covariates and county as random effect. We get the following summary statistics for LMM and Bootstrap confidence interval of model parameter using 10 Bootstrap samples:

![Figure 1. Nonparametric Bootstrap CI of Prediction](image)

In order to construct a parametric bootstrap confidence interval of predictions for BHF data, the steps are as follows:

1. Fit the mixed model (1) to the BHF data. The model has five parameters: three regression coefficients (including the intercept) and two variance components.
2. Using the estimated coefficients, we simulated \( B = 500, 1000 \) and \( 5000 \) new data sets (resamples) with the same structure as the BHF data (i.e. 12 counties, 36 observations, etc.). Simulated each data set involves simulating a random county specific effect.

3. Fit model to each simulated dataset and calculate the small area estimates for model parameter using "lme4" and "boot" packages from R, we got parameter estimation for LMM and 95\% parametric bootstrap CI’s of prediction for each county in the following results:

![Linear mixed model fit by REML](image)

**Table 2.** Bootstrap CI of Parameter Estimation (Bootstrap=500)

| Parameter       | Confidence Interval |
|-----------------|---------------------|
| \( \sigma_0 \)  | 2.442367 19.28502   |
| \( \sigma \)     | 8.648499 15.48753   |
| Intercept       | 3.662079 99.44178   |
| CornPix         | 0.225309 0.42978    |
| SoyBeansPix     | -0.24834 -0.02751   |

![Bootstrap Confidence Interval of Prediction](image)

**Figure 2.** Parameter Estimation of LMM for BHF Data

**Figure 3.** Parametric Bootstrap CI of Prediction for 1000 Bootstrap samples
5. Discussions
In this paper we discuss two sided parametric and nonparametric Bootstrap confidence interval of prediction for small area means based on linear mixed model with a gaussian response. Prediction methods for the small area mean, estimation of the prediction mean squared error (MSE) and confidence intervals (CIs) for the small area means are important in constructing reliable estimates for areas with small sample sizes. It is shown that parametric Bootstrap CIs for the small area mean is narrower than nonparametric Bootstrap CIs. In further research, we will study method for constructing confidence interval using different type of Bootstrap with respect to the coverage errors and can improve the coverage accuracy for general level.

Acknowledgments
The authors also want to thank LPDP (Lembaga Pengelola Dana Pendidikan) and Kemenristek DIKTI for supporting BUDI (Beasiswa Unggulan Dosen Indonesia) Scholarship for the first author.

References
[1] A Agresti 2015 Foundation of Linear and Generalized Linear Models John Wiley & Sons Inc Hoboken New Jersey
[2] Bates D, Martin Mächler, B M Bolker, S C Walker 2015 Fitting Linear Mixed-Effects Models Using lme4 Journal of Statistical Software 67 1
[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 1995 Bootstrap Methods for Generalized Linear Mixed Models with Applications to Small Area Estimation, Proceedings of the 10th International Workshop on Statistical Modelling (Innsbruck, Austria, 10-14 July 1995)
[5] Dass S, Maiti T, Ren H, Sinha S 2012 Confidence interval estimation of small area parameters shrinking both means and variances Survey Methodology 38(2) 173-187
[6] Diao L, Smith D, Datta G S, Maiti T, Opsomer J 2014 Accurate Confidence Interval Estimation of Small Area Parameters Under the Fay-Herriot Model Scandinavian Journal of Statistics 41 497-515
[7] Efron B and Tibshirani R 1993 An Introduction to the Bootstrap *Chapman and Hall* New York London
[8] Sánchez-Espigares J A Ocaña J 2009 An R Implementation of Bootstrap Procedures for Mixed Models *Conference presentation*
[9] Rao JNK and Isabel Molina 2015 Small Area Estimation 2nd Edition *Wiley Series in Survey Methodology*
[10] Yoshimori M, Lahiri P, 2014 A Second-Order Efficient Empirical Bayes Confidence Interval The Annals of Statistics *42*(4) 1233–1261