Empirical Best Predictor for Nested Error Regression Small Area Models

Dian Handayani¹, Khairil Anwar Notodiputro¹*, Asep Saefuddin¹, I Wayan Mangku², Anang Kurnia¹

¹Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia
²Department of Mathematics, Bogor Agricultural University, Bogor, 16680, Indonesia
³Department of Statistics, State University of Jakarta, Jakarta, 13220, Indonesia

*Email: khairil@apps.ipb.ac.id

Abstract. Nested error regression models have played an important role in small area estimation (SAE) especially for deriving indirect or model-based estimates of small area parameters. The models are valid to be employed whenever the auxiliary information is available at level units, as well as the random effect is independent from the sampling error. Furthermore, the models also assume normality the random effects and sampling errors. The standard SAE method, specifically the Empirical Best Linear Unbiased Predictor (EBLUP), which derives the estimates for small area parameters under the nested error regression models, certainly have to satisfy the strictly assumptions. In this paper, we study the Empirical Best Predictor (EBP) which can be utilized for deriving estimates of small area parameters whenever the variable of interest has skewed distributions. We apply the EBP to estimate the poverty incidence and poverty gap for the regions (‘kabupaten’ and ‘kota’) in West Java Province – Indonesia.

Keywords: Empirical Best Linear Unbiased Predictor (EBLUP), log-normal distribution, poverty measures Small Area Estimation (SAE), skewed data.

1. Introduction
Nowadays, the demand of statistics for a certain sub population is increasing both from government and private sectors. The statistics is important because it can support the central government for funding allocation to regional government. Furthermore, the statistics from sub population also can help private sectors for making business decision especially for which depends on local socio economic condition. The source of information to produce the sub population statistics is often obtained from survey sample. Because the survey sample is usually designed to provide statistics for the large (whole) population, it can cause the sample size from some sub populations is too small (or it can be zero) for producing the accurate direct estimates. The sub population which the sample size from it is not large enough to produce the accurate direct estimates is called by small area. The statistical methods which deal with the inference for small area is called Small Area Estimation (SAE).

To increase the quality of estimation and to obtain more accurate small area estimates, we can increase the “effective” sample size from small area by borrowing strength the information from other
sources. For example, we can utilize the observation of variable interest from other related areas and we can also use some auxiliary information from census, administrative records or other survey. Then, instead direct estimates, we can apply indirect estimation by employing statistical models to link the variable of interest with some auxiliary information from other sources. The estimates that are derived utilizing statistical models are also called by model-based estimates.

The SAE model can be classified by SAE area level models and SAE unit level models. The SAE area level models are employed whenever the auxiliary information is available for area level. On the other hand, if the auxiliary information is available for unit level, then the SAE unit level models should be employed.

The standard SAE method estimates small area parameter under the SAE model which assumes the normality for sampling error and random area effect. On the other hand, these assumptions are often not valid to be applied in practice. Specifically, in business or socio economic data, the variable of interest often follows skewed distribution and certainly it could yield that the normality assumption for sampling error can not be satisfied.

In this paper, we review the Empirical Based Predictor (EBP) to estimate small area linear parameter as well as nonlinear parameter whenever variable interest has skewed distribution. We assume that the variable of interest would follow log normal distribution after taking logarithm transformation. Then, the EBP would be derived under SAE unit level model (nested error regression model) using this transformed of variable of interest.

The structure of the paper is as follows. Section 2 describes the Empirical Best Linear Unbiased Prediction (EBLUP). Section 3 and section 4 describe the Empirical Best Prediction (EBP) for linear parameter and nonlinear parameter respectively. The application is provided in section 5 and concluding remarks is in section 6.

2. Empirical Best Linear Unbiased Prediction (EBLUP)

The standard SAE method that estimates small area parameter under SAE unit level model or nested error regression model is originally developed by Battese et al (1988). The nested error regression model assumes that random area effect and random sampling error follow normal distribution. Moreover, it is also assumed that there are no correlation among the random area effect, no correlation among random sampling errors and no correlation between random area effect and random sampling errors.

The nested error regression model relates the specific values of variable of interest \( y_{ij} \) to \( p \) specific values of auxiliary information \( x_{ij} \), random area effect \( v_i \) and random sampling error \( e_{ij} \) as follows:

\[
y_{ij} = x_{ij}'\beta + v_i + e_{ij} \quad j = 1,2 \ldots N_i; \quad i = 1,2 \ldots M
\]

where \( N_i \) is the number of units in \( i^{th} \) small area, the random area effect \( v_i \) is assumed to be \( v_i \sim iid N(0,\sigma^2_v) \) and it is independence with random sampling error \( e_{ij} \); \( e_{ij} \sim iid N(0,\sigma^2_e) \).

Model (1) is hold for population and it is actually a special case of the general linear mixed model which also could be written as follows:

\[
y = X\beta + Zv + e
\]

where \( y = (y_1, \ldots, y_M)' \) is \((N \times 1)\) vector of observation of variable of interest \( Y \) where \( y_i = (y_{i1}, y_{i12}, \ldots, y_{iN_i})' \), \( i = 1,2 \ldots M, \quad N = \sum_{i=1}^{M} N_i, \quad X = (x_{1i}, \ldots, x_{Mi})' \) is \((N \times p)\) matrix of auxiliary information, where \( x_{ij} = (x_{1ij}, \ldots, x_{p(ij)})' \), \( i = 1,2 \ldots M, \quad j = 1,2 \ldots N_i \), \( \beta = (\beta_1, \ldots, \beta_p)' \) is \((p \times 1)\) vector of regression coefficient, \( Z = diag(Z_i = 1_{N_i}) \), \( i = 1,2 \ldots M \) is \((N \times M)\) matrix, \( v = (v_{i1}, \ldots, v_{iM})' \) is \((M \times 1)\) vector random area effect, \( e = (e_{i1}, \ldots, e_{iM})' \) is \((M \times 1)\) vector sampling error, \( e_{i} = (e_{i1}, e_{i12}, \ldots, e_{iN_i})' \), \( i = 1,2 \ldots M \). The assumption for \( v \) and \( e \) are \( v \sim N(0,\mathbf{G}) \) and \( e \sim N(0,\mathbf{R}) \) respectively. Therefore, the distribution for \( y \) is \( y \sim N(X\beta, V) \) where \( V = ZGZ' + R \).

In this paper, we will assume that all of \( M \) small areas are selected in the sample. It means that we assume that there is no sample selection bias of areas. Furthermore, we also assume that the sample
selection bias within areas is absent. As a result, the model (1) that is valid for population will be also valid for sample units. In the model for sample, the notation \( N \) (number of population units) will be replaced with \( n \) (number of sample units). Similarly, \( N_i \) (number of population units in \( i^{th} \) small area) is replace with \( n_i \) (number of sample units in \( i^{th} \) small area).

Suppose that the parameter of interest is the mean for \( i^{th} \) small area (denoted by \( \mu_i \)):

\[
\mu_i = \frac{1}{n_i} \sum_{i=1}^{N_i} y_{ij} = \frac{1}{n_i} \left[ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} y_{ij} \right]
\]

where : \( s_i \) denotes the set of \( j \) in the sample from \( i^{th} \) small area and \( r_i \) denotes the set of \( j \) in the non-sample for area \( i \).

The Best Linear Unbiased Predictor (BLUP) for \( \mu_i \) under nested error regression model (1) is given by :

\[
\hat{\mu}_i^{BLUP} = \frac{1}{n_i} \sum_{i=1}^{N_i} y_{ij} = \frac{1}{n_i} \left[ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{BLUP} \right]
\]

where : \( \hat{y}_{ij}^{BLUP} \) is given by :

\[
\hat{y}_{ij}^{BLUP} = E \left[ y_{ij} | v_i, \bar{x}_i \right] = x_i^{\top} \hat{\beta} + \hat{\upsilon}_i
\]

The best linear unbiased estimate (BLUE) for \( \beta \) is given by :

\[
\hat{\beta} = \left( \sum_{i=1}^{N} x_i^{\top} v_i^{-1} x_i \right)^{-1} \left( \sum_{i=1}^{N} x_i^{\top} v_i^{-1} y_i \right)
\]

On the other hand, the best linear unbiased predictor (BLUP) for \( v_i \) is given by :

\[
\hat{\upsilon}_i = y_i - \bar{y}_i - \bar{x}_i^{\top} \hat{\beta} = \frac{\sigma^2}{\sigma^2 + \frac{\sigma^4}{n_i}} \left( y_i - \bar{y}_i - \bar{x}_i^{\top} \hat{\beta} \right)
\]

where: \( \bar{y}_i = \frac{1}{n_i} \sum_{j \in s_i} y_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \); \( \bar{x}_i = \frac{1}{n_i} \sum_{j \in s_i} x_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \).

The BLUP for \( \mu_i \) depends on the parameter variance \( \theta = (\sigma^2, \sigma^4) \) which is usually unknown in practice. The parameter variance can be estimated by maximum likelihood (ML) method, restricted maximum likelihood (REML) or method of moments. If the parameter variance is replaced with their estimates then we will obtain the Empirical Best Linear Unbiased Prediction (EBLP) for \( \mu_i \) (denoted by \( \hat{\mu}_i^{EBLP} \)). The \( MSE(\hat{\mu}_i^{EBLP}) \) and the approximately unbiased estimator for \( MSE(\hat{\mu}_i^{EBLP}) \) can be seen further in Rao and Molina (2015, chapter 7).

Prasad and Rao (1990) proposed an approximation for the \( MSE(\hat{\mu}_i^{EBLP}) \) as follows :

\[
MSE(\hat{\mu}_i^{EBLP}) \approx g_{1i}(\sigma^2, \sigma^4) + g_{2i}(\sigma^4, \sigma^2) + g_{3i}(\sigma^4, \sigma^4)
\]

where :

\[
g_{1i}(\sigma^2, \sigma^4) = \gamma(\sigma^2, \sigma^4/n_i) \\
g_{2i}(\sigma^2, \sigma^4) = (\bar{X}_i - y_i \bar{x}_i)^T (X^T V^{-1} X)^{-1} (\bar{X}_i - y_i \bar{x}_i) \\
g_{3i}(\sigma^4, \sigma^4) = n_i^{-2} (\sigma^2 + \sigma^4/n_i)^{-3} [\sigma^2 var(\sigma^2) + \sigma^4 var(\sigma^2) - 2 \sigma^2 \sigma^4 cov(\sigma^2, \sigma^4) ] , \text{ var}(\sigma^2) = 2(n - M - p + \lambda)^{-1} \sigma^4 ; \lambda = 0 \text{ if model (1) has no intercept and } \lambda = 1 \text{ otherwise.}
\]

Moreover, Prasad and Rao (1990) also proposed an approximately unbiased estimator for the \( MSE(\hat{\mu}_i^{EBLP}) \) as follows :

\[
\text{mse}(\hat{\mu}_i^{EBLP}) = g_{1i}(\hat{\sigma}^2, \hat{\sigma}^4) + g_{2i}(\hat{\sigma}^4, \hat{\sigma}^2) + 2 g_{3i}(\hat{\sigma}^2, \hat{\sigma}^4)
\]

where \( g_{1i}(\hat{\sigma}^2, \hat{\sigma}^4), g_{2i}(\hat{\sigma}^4, \hat{\sigma}^2) \) and \( g_{3i}(\hat{\sigma}^2, \hat{\sigma}^4) \) are obtained from \( g_{1i}(\sigma^2, \sigma^4), g_{2i}(\sigma^4, \sigma^2) \) and \( g_{3i}(\sigma^2, \sigma^4) \) respectively, replacing \( \theta = (\sigma^2, \sigma^4) \) by \( \hat{\theta} = (\hat{\sigma}^2, \hat{\sigma}^4) \).

3. Empirical Best Prediction (EBP) for Linear Parameter

Parameter in small area \( i^{th} \), \( \tau_i \), is be called by linear parameter if \( \tau_i = h(y_i) \), where \( h \) is a linear function of \( y_i \). Berg and Chandra (2014) have developed the Empirical Best Prediction/EBP to estimate linear parameter, specifically for the mean of variable of interest which has skewed distribution, i.e log normal distribution. The EBP is derived under unit level log linear model as follows :
\[
\log(y_{ij}) = y_{ij}^* = x_{ij}^T \beta + v_i + e_{ij}
\]
where \(v_i \sim iid N(0, \sigma_v^2), e_{ij} \sim iid N(0, \sigma_e^2), y_{ij}^* \sim N(x_{ij}^T \beta, \sigma_v^2 + \sigma_e^2)\).

Suppose that the linear parameter of interest that will be estimated is small area mean as given by formula (3). The Best Predictor (BP) for \(\mu_i\) is given by:

\[
\hat{\mu}_i^{BP} = \frac{1}{N_i} \left[ \sum_{j \in s_i} y_{ij} + \sum_{r \in \bar{s}_i} \hat{y}_{ij}^{BP} \right]
\]
where:

\[
\hat{y}_{ij}^{BP} = E[y_{ij}|(y, x)] = \exp\left(x_{ij}^T \hat{\beta} + y_i(\bar{y}_is - \bar{x}_is \hat{\beta}) + 0.5 \sigma_e^2 (y_i n_i^{-1} + 1)\right) = \sigma_v^2 (\sigma_v^2 + n_i^{-1} \sigma_e^2)^{-1},
\]
where \((y, x) = \{y_{ij}; i = 1, 2, \ldots, M, j \in s_i \cup \{x_{ij}; i = 1, 2, \ldots, M, j \in U_i\}, U_i\) denotes the set of \(N_i\) indexes in the population for area \(i\), \(\bar{y}_is = \frac{1}{N_i} \sum_{j \in s_i} y_{ij}, \bar{x}_is = \frac{1}{n_i} \sum_{j \in s_i} x_{ij}\).

In practice, it is seldom to obtain \(\hat{\mu}_i^{BP}\) because the predictor is a function of the parameter \(\theta = (\sigma_v^2, \sigma_e^2)\) which are unknown. If \(\theta\) is replaced with its consistent estimates, \(\hat{\theta} = (\hat{\sigma}_v^2, \hat{\sigma}_e^2)\), then we obtain the Empirical Best Predictor (EBP) for \(\mu_i\) (denoted by \(\hat{\mu}_i^{EBP}\)). The details for \(\text{MSE}(\hat{\mu}_i^{EBP})\) and the estimate of \(\text{MSE}(\hat{\mu}_i^{EBP})\) could be seen further in Berg and Chandra (2014).

The mean square error of the EBP of \(\mu_i\), denoted by \(\text{MSE}(\hat{\mu}_i^{EBP})\), is given by:

\[
\text{MSE}(\hat{\mu}_i^{EBP}) = M_{s1}(\theta) + M_{s2}(\theta)
\]
where:

\[
M_{s1}(\theta) = E \left[ (\hat{\mu}_i^{EBP} - \mu_i)^2 \right] \quad \text{and} \quad M_{s2}(\theta) = E \left[ (\hat{\mu}_i^{EBP} - \hat{\mu}_i^{BP}(\theta))^2 \right].
\]

Berg and Chandra (2014) give a closed form expression for \(M_{s1}(\theta)\) and a linear approximation for \(M_{s2}(\theta)\). The MSE of \(\hat{\mu}_i^{BP}\) is given by:

\[
M_{s1}(\theta) = E \left[ (\hat{\mu}_i^{BP}(\theta) - \mu_i)^2 \right] = \frac{N_i}{N_i} \left( \sum_{j \in e_i} \exp(x_{ij}^T \beta) \right)^2 \xi_i + \sum_{j \in e_i} \exp(2x_{ij}^T \beta) \psi_i
\]
where:

\[
\xi_i = \exp(y_i n_i^{-1} \sigma_v^2) - 1, \quad \psi_i = \exp(y_i n_i^{-1} \sigma_v^2 + \sigma_e^2) - \exp(y_i n_i^{-1} \sigma_v^2) \quad \text{and} \quad \kappa_i = \exp(2 \beta_0 + 2 \gamma_i \sigma_v^2 + n_i^{-1} \sigma_v^2) + y_i n_i^{-1} \sigma_v^2 + \sigma_v^2).
\]

An approximation for \(M_{s2}(\theta)\) is:

\[
M_{s2}(\theta) = N_i^{-1} \sum_{j \in e_i} \exp(\mu_{ij} + \gamma_i (\sigma_v^2 + n_i^{-1} \sigma_v^2) + 0.5 \delta_{ij} [i \neq k] + 2 \gamma_i [i = j]) - 2 \sum_{j \in e_i} \exp(\mu_{ij} + \mu_{ik} + 2 \gamma_i (\sigma_v^2 + n_i^{-1} \sigma_v^2) + 0.5 \gamma_i) + \sum_{j \in e_i} \exp(\mu_{ij} + \mu_{jk} + 2 \gamma_i (\sigma_v^2 + n_i^{-1} \sigma_v^2) + 0.5 \gamma_i) = M_{s2}(\theta)
\]
where:

\[
\delta_{ijk} = (b_{ij} + b_{ik})^T V[\hat{\beta}] (b_{ij} + b_{ik}) + 4 tr (E[f_i f_i^T] V[\hat{\sigma}^2])
\]
\[
\nu_{ij} = b_{ij}^T V[\hat{\beta}] b_{ij} + tr (E[f_i f_i^T] V[\hat{\sigma}^2])
\]
\[
\hat{\beta} = (\beta_{00} \hat{B}_1)^T ; \quad \hat{\sigma} = (\hat{\sigma}_v^2 \hat{\sigma}_e^2)^T ; \quad f_i = (f_{i1}, f_{i2})^T
\]
\[
b_{ij}^* = (1 - y_i, x_{ij} - y_i \bar{x}_is)^T
\]
\[
f_{i1} = \frac{1}{1 - y_i} (\sigma_v^2 + n_i^{-1} \sigma_v^2)^{-1} (-1 - y_i) (\bar{y}_is - \beta_0 - \bar{x}_is \hat{\beta}_1) + 0.5 (1 - y_i)^2
\]
\[
f_{i2} = -\frac{1}{1 - y_i} (\sigma_v^2 + n_i^{-1} \sigma_v^2)^{-1} n_i^{-1} (\bar{y}_is - \beta_0 - \bar{x}_is \hat{\beta}_1) + 0.5 y_i n_i^{-1} + 0.5
\]

The \(V[\hat{\beta}]\) and \(V[\hat{\sigma}^2]\) are the asymptotic covariance matrix of the \(\hat{\beta}\) and \(\hat{\sigma}\) respectively. If the estimator of \(\hat{\sigma}\) are obtained by REML or ML, then the asymptotic variances are obtained from the inverse of the Fisher information matrix.

The elements of the second term of the formula (13) are:

\[
E[f_{i1}^2] = (\sigma_v^2 + n_i^{-1} \sigma_v^2)^{-1} (1 - y_i)^2 + 0.5 (1 - y_i)^2
\]
\[
E[f_{i1} f_{i2}] = -n_i^{-1} y_i (\sigma_v^2 + n_i^{-1} \sigma_v^2)^{-1} + 0.5 (1 - y_i)^2 \{0.5 (y_i n_i^{-1} + 1)\}
\]
\[
E[f_{i2}^2] = n_i^{-1} y_i^2 (\sigma_v^2 + n_i^{-1} \sigma_v^2)^{-1} + 0.5 (y_i n_i^{-1} + 1)^2.
\]
A naïve estimator for $MSE(\hat{\rho}^{\text{EBP}})$ is given by

$$MSE_{1i} = M_{3i}(\hat{\theta}) + \hat{M}_{2i}(\hat{\theta})$$

where $M_{3i}(\hat{\theta})$ and $\hat{M}_{2i}(\hat{\theta})$ are expressions in (13) and (14) respectively which is evaluated at $\hat{\theta}$. Because $M_{3i}(\hat{\theta})$ is a nonlinear function of $\theta$, the naïve estimator of $M_{3i}(\theta)$ in (22) is biased.

Berg and Chandra (2014) correct the bias and propose the MSE estimator as follows:

$$MSE_{2i} = M_{3i}(\hat{\theta}) + \hat{M}_{2i}(\hat{\theta})$$

where:

$$\hat{\beta}_{0i} = \beta_0 - 0.5 \left( \frac{\partial M_{3i}(\theta)}{\partial \hat{\theta}_0} \right)^{-1} M_{3i}(\hat{\theta}),$$

$$M_{3i}(\hat{\theta}) = \text{tr} \left( \frac{\partial^2 M_{3i}(\theta)}{\partial \hat{\theta}_i \partial \hat{\theta}_0} \right) V\{\beta\} + \frac{\partial^2 M_{3i}(\theta)}{\partial \sigma^2 \partial (\sigma^2)^T} V\{\hat{\sigma}_{ij}^2\}$$

$$\frac{\partial^2 M_{3i}(\theta)}{\partial \hat{\theta}_i \partial \hat{\theta}_0}$$

and

$$\frac{\partial^2 M_{3i}(\theta)}{\partial \sigma^2 \partial (\sigma^2)^T}$$

are the matrices of second derivatives of $M_{3i}(\theta)$ respect to $\beta$ dan $\sigma^2$ respectively.

4. Empirical Best Prediction (EBP) for Nonlinear Parameter

Parameter in small area $i^{th}$, $\tau_i$, is called by nonlinear parameter if $\tau_i = h(y_i)$, where $h$ is a nonlinear function of $y_i$. Poverty measures such as poverty incidence, poverty gap and poverty severity are the examples of nonlinear parameter. Molina and Rao (2010) developed the EBP method to estimate the poverty measures FGT (Foster, Greer & Thorbecke, 1984) under SAE unit level model. The EBP method is developed by assuming that the result of transformation of variable of interest follow normal distribution. The poverty measure FGT for small area $i$ (denoted by $P_{ai,i}$) is given by:

$$P_{ai} = \frac{1}{N_i} \sum_{j=1}^{N_i} P_{ai,ij} \; ; \; i = 1,2 \ldots M, \; \alpha = 0,1,2$$

where $P_{ai,ij}$ is given by:

$$P_{ai,ij} = \left( \frac{z-y_{ij}}{z} \right)^{\alpha} I(y_{ij} < z); \; j = 1,2 \ldots N_i; \; \alpha = 0,1,2$$

where $y_{ij}$ is welfare measure such as income or expenditure, $z$ is poverty line, $I(y_{ij} < z) = 1$ if $y_{ij} < z$ (person under poverty) and $I(y_{ij} < z) = 0$ if $y_{ij} \geq z$ (person not under poverty). For $\alpha = 0$, the poverty measure $P_{ai}$ is also called by poverty incidence (Head Count Ratio). It represents about the proportion of person under poverty. For $\alpha = 1$, the poverty measure $P_{ai}$ is also called poverty gap which represents the area mean of relative distance to the poverty line of each individual. For $\alpha = 2$, the poverty measure $P_{ai}$ is also called by poverty severity which represents the average squares of the poverty gaps relative to the poverty line.

The FGT poverty measures for small area $i$ can also be written by:

$$P_{ai} = \frac{1}{N_i} \left[ \sum_{j \in s_i} P_{ai,ij} + \sum_{j \in r_i} P_{ai,ij} \right]$$

where $\sum_{j \in s_i} P_{ai,ij}$ is the sum of $P_{ai,ij}$ for sample values and $\sum_{j \in r_i} P_{ai,ij}$ is the sum of $P_{ai,ij}$ for non-sampled values.

The EBP of $P_{ai}$ is given by:

$$\beta_{ai}^{\text{EBP}} = \frac{1}{N_i} \left[ \sum_{j \in s_i} P_{ai,ij} + \sum_{j \in r_i} \beta_{ai,ij}^{\text{EBP}} \right]$$

$$\beta_{ai,ij}^{\text{EBP}} = E[P_{ai,ij}|y_s]$$

is obtained by Monte Carlo approximation; $y_s$ is the values of $y$ that is selected as sample.

Molina and Rao (2010) evaluate the $\beta_{ai}^{\text{EBP}}$ under nested error regression model (1). Subject to the model (1), distribution for variable of interest $y$ in $i^{th}$ small area is $y_i \sim N(\mu_i, V_i)$, $\mu_i = X_i \beta$ and $V_i = \sigma^2 I_{N_i}$. The EBP of $P_{ai}$ is given by:

$$\beta_{ai}^{\text{EBP}} = \frac{1}{N_i} \left[ \sum_{j \in s_i} P_{ai,ij} + \sum_{j \in r_i} \beta_{ai,ij}^{\text{EBP}} \right]$$

$$\beta_{ai,ij}^{\text{EBP}} = E[P_{ai,ij}|y_s]$$

is obtained by Monte Carlo approximation; $y_s$ is the values of $y$ that is selected as sample.
Consider that the values of $y$ in $i^{th}$ small area is decomposed into sample and non-sampled values, $y_i = (y_{is}, y_{ir})^T$. Then, the distribution of $y_{ir} | y_{is}$ is 

$$y_{ir} | y_{is} \sim N(\mu_{ir|s}, V_{ir|s})$$

(28)

where:

$$\mu_{ir|s} = X_{ir}\beta + \sigma_{v}^{2}1_{N_{i}-n_{i}}1_{n_{i}}^TV_{is}^{-1}(y_{is} - X_{is}\beta)$$ \hspace{1cm} (29)

$$V_{ir|s} = \sigma_{v}^{2}(1 - \gamma_{i})1_{N_{i}-n_{i}}1_{n_{i}} + \sigma_{e}^{2}\Sigma_{N_{i}-n_{i}}$$ \hspace{1cm} (30)

where: $V_{is} = \sigma_{v}^{2}1_{n_{i}}1_{n_{i}}^T + \sigma_{e}^{2}I_{n_{i}}$ and $y_{i} = \sigma_{v}^{2}(\sigma_{v}^{2} + n_{i}^{-1}\sigma_{e}^{2})^{-1}$.

Generating $y_{ir}$ from (28) by Monte Carlo approximation will involve $M$ multivariate normal vector $y_{ir}$ of size $(N_{i} - n_{i})$, $i = 1, ..., M$. Moreover, this process must be repeated $L$ times. This computation would be very intensive and would be unfeasible for large $N_{i}$.

This process can be avoided by noting that $V_{ir|s}$ given by (30) corresponds to the covariance matrix of $y_{ir}$ that is generated from:

$$y_{ir} = \mu_{ir|s} + w_{i}1_{N_{i}-n_{i}} + \varepsilon_{ir}$$ \hspace{1cm} (31)

with new random effect $w_{i}$ and error $\varepsilon_{ir}$ are independence and satisfy $w_{i} \sim N[0, \sigma_{v}^{2}(1 - \gamma_{i})]$, $\varepsilon_{ir} \sim N[0_{N_{i}-n_{i}}, \sigma_{e}^{2}I_{n_{i}-n_{i}}]$ independently for $j \in r_{i}$.

Based on model (31), it is not necessary to generate multivariate normal vector $y_{ir}$ of size $(N_{i} - n_{i})$. We only need to generate $(1 + N_{i} - n_{i})$ independent univariate normal variable $w_{i} \sim N[0, \sigma_{v}^{2}(1 - \gamma_{i})]$ and $\varepsilon_{ir} \sim N[0_{N_{i}-n_{i}}, \sigma_{e}^{2}I_{n_{i}-n_{i}}]$. Then, the generated values of $w_{i}$ and $\varepsilon_{ir}$ is utilized to obtain the corresponding elements $y_{ij}$, $j \in r_{i}$ from (31) with $\mu_{ir|s}$ is given by (29). In practice, the parameter $\theta = (\beta, \sigma_{v}^{2}, \sigma_{e}^{2})$ are replaced with $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_{v}^{2}, \hat{\sigma}_{e}^{2})$. In this situation, the $y_{ij}$ are generated from the corresponding estimated normal distribution.

Molina and Rao (2010) develop bootstrap parametric to approximate mean square error of the EBP of the FGT poverty measures. The steps of procedures are:

1. Based on sample data $(y_{s}, X_{s})$ and model (1), obtain $\hat{\beta}$, $\hat{\sigma}_{v}^{2}$, $\hat{\sigma}_{e}^{2}$.
2. Generate $v_{i}^{*} \sim iidN(0, \hat{\sigma}_{v}^{2})$, $i = 1, ..., M$ and independently generate $\varepsilon_{ij}^{*} \sim iidN(0, \hat{\sigma}_{e}^{2})$, $j = 1, ..., N_{i}$, $i = 1, ..., M$.
3. By using $\hat{\beta}$, $V_{*i}$, $\varepsilon_{ij}^{*}$, $x_{ij}$, $i = 1, ..., M$, $j = 1, ..., N_{i}$ construct bootstrap population model:

$$y_{ij}^{*} = x_{ij}^{*}\hat{\beta} + v_{i}^{*} + \varepsilon_{ij}^{*}, j = 1, ..., N_{i}, i = 1, ..., M.$$ \hspace{1cm} (32)

4. Based on bootstrap population model (32), generate a large number $B$ of bootstrap population $\{y_{ij}^{(b)}\}; j = 1, ..., N_{i}, i = 1, ..., M$ which is independent and identically distributed. Then, calculate bootstrap population parameter $p^{(b)}_{ai} = N_{i}^{-1}\sum_{j=1}^{N_{i}}p_{aij}^{*}$, where $p_{aij}^{*} = h_{a}(y_{ij}^{*})^{(b)}$, $b = 1, ..., B$.
5. Select sample from each population bootstrap that is generated from step (4) and calculate the bootstrap EBP of $p_{ai}$ (denoted by $\hat{p}_{ai}^{EBP(b)}$) using the bootstrap sample data $y_{s}^{*}$ and known population values $x_{ij}$.
6. Calculate the estimates for MSE($\hat{p}_{ai}^{EBP}$) as follows:

$$mse_{i}(\hat{p}_{ai}^{EBP}) = \frac{1}{B} \sum_{b=1}^{B} \left( \hat{p}_{ai}^{EBP(b)} - p_{ai}^{*} \right)^{2}$$ \hspace{1cm} (33)

5. Application

In our study, we apply the EBP method to estimate the average of monthly expenditure per capita for each ‘kabupaten’ and ‘kota’ in West Java Province. For this, we use data from Socio Economic
Survey, specifically September 2015, to obtain the outcome variable. We also use data from Potensi Desa 2014 to obtain some auxiliary information.

The outcome variable is monthly expenditure per capita of household. On the other hand, for auxiliary variables, we have considered the number of minimarket, the number of groceries, the number of traditional market, and the status of village where the selected household is located (the status is “urban” or “rural”). All of the auxiliary variables are available in village level.

![Histogram of kapita](image1)

![Histogram of log_kapita](image2)

**Figure 1.** (a) Distribution of monthly average per capita expenditure of household in West Java Province (b) Distribution of logarithmic transformation of monthly average per capita expenditure of household in West Java Province

We assume that the logarithmic transformation of outcome variable would follow normal distribution. Moreover, the transformed variable and some auxiliary information would also satisfy the nested error regression model (10). Figure 1(a) shows that the distribution of monthly average per capita expenditure of household in West Java Province has positively skewed distribution. On the other hand, we could see from Figure 1(b) that the distribution of logarithmic transformation of the capita follow normal distribution approximately.

The values of direct estimates (DIRECT), EBLUP and the Empirical Best Predictor (EBP) of the average of monthly expenditure per capita for each ‘kabupaten’ and ‘kota’ in West Java Province are listed in Table 1.

Based on Table 1, we can see that the direct estimates of monthly expenditure for each ‘kabupaten’ and ‘kota’ in West Java which is produced by DIRECT estimator is the largest, compared to EBLUP as well as EBP. It is happened because the variable of interest has positively skewness distribution which has some large (extreme) positive values. Moreover, if sample size is too small, the direct estimates would be very unstable (the variability would be very large). Refer to Appendix 1 for the square root of the mean square error.

The other estimators, EBLUP and EBP are model-based estimator. From Table 1, we can see that the EBLUP of the average of monthly expenditure per capita for each ‘kabupaten’ and ‘kota’ in West Java is approximately similar to the direct estimates. It can happen because the EBLUP is calculated under normal distribution that will also tend to be influenced by some large (extreme) values. On the other hand, the EBP which is calculated under log-normal nested error regression model can reduce the influence of some large (extreme) values.
Table 1. Population size, sample size, Direct, EBLUP and EBP of the average of monthly expenditure per capita for each ‘kabupaten’ and ‘kota’ in West Java.

| No | Area         | Population Size | Sample Size | DIRECT  | EBLUP  | EBP    |
|----|--------------|-----------------|-------------|---------|--------|--------|
| 1  | Bogor        | 1,265,161       | 288         | 1,370,685 | 1,354,377 | 1,015,718 |
| 2  | Sukabumi     | 715,810         | 225         | 862,500 | 872,935 | 762,541 |
| 3  | Cianjur      | 673,184         | 225         | 672,413 | 682,699 | 615,982 |
| 4  | Bandung      | 976,061         | 253         | 874,523 | 887,592 | 776,427 |
| 5  | Garut        | 717,408         | 251         | 666,484 | 682,868 | 556,964 |
| 6  | Tasikmalaya  | 533,685         | 215         | 531,203 | 548,419 | 516,237 |
| 7  | Ciamis       | 375,499         | 228         | 694,950 | 701,825 | 621,525 |
| 8  | Kuningan     | 325,347         | 183         | 906,854 | 906,192 | 754,216 |
| 9  | Cirebon      | 647,238         | 244         | 662,164 | 675,229 | 623,822 |
| 10 | Majalengka   | 416,801         | 208         | 836,048 | 842,156 | 696,241 |
| 11 | Sumedang     | 357,878         | 202         | 1,042,595 | 1,045,522 | 877,500 |
| 12 | Indramayu    | 573,119         | 228         | 749,385 | 758,675 | 736,620 |
| 13 | Subang       | 445,617         | 219         | 855,405 | 853,299 | 799,407 |
| 14 | Purwakarta   | 290,533         | 183         | 954,318 | 954,254 | 883,012 |
| 15 | Karawang     | 649,442         | 241         | 928,455 | 925,195 | 872,016 |
| 16 | Bekasi       | 808,658         | 253         | 1,490,900 | 1,459,189 | 1,253,626 |
| 17 | Bandung Barat| 482,193         | 233         | 617,310 | 633,514 | 583,321 |
| 18 | Pangandaran  | 141,023         | 169         | 764,882 | 774,698 | 681,792 |
| 19 | Kota Bogor   | 238,787         | 178         | 1,879,861 | 1,849,907 | 1,493,311 |
| 20 | Kota Sukabumi| 88,704          | 132         | 1,122,818 | 1,116,579 | 1,021,905 |
| 21 | Kota Bandung | 635,058         | 239         | 1,854,295 | 1,823,912 | 1,435,998 |
| 22 | Kota Cirebon | 82,209          | 136         | 1,082,500 | 1,081,442 | 884,283 |
| 23 | Kota Bekasi  | 683,588         | 231         | 1,650,234 | 1,623,641 | 1,374,449 |
| 24 | Kota Depok   | 486,876         | 227         | 2,344,213 | 2,298,002 | 1,770,883 |
| 25 | Kota Cimahi  | 178,404         | 175         | 1,332,124 | 1,314,305 | 1,102,590 |
| 26 | Kota Tasikmalaya | 200,858 | 166         | 980,844 | 982,014 | 801,014 |
| 27 | Kota Banjar  | 65,355          | 127         | 958,447 | 956,361 | 785,612 |

6. Concluding Remarks

This article attempts to study the Empirical Best Predictor (EBP) for linear parameter as well as nonlinear parameter when the log transformed of variable of interest follows a normal distribution, in particular, it follows a nested error regression model. The EBP of linear parameter as well as nonlinear parameter is derived so that the mean square error of the EBP minimum. It is given by the conditional expectation. For linear parameter, the explicit expression for the conditional expectation can be obtained. However, for nonlinear parameter, the conditional expectation can be approximated by Monte Carlo simulation. An application to real data, especially for the estimation of the average of monthly expenditure per capita for each ‘kabupaten’ and ‘kota’ in West Java Province - Indonesia demonstrates the satisfactory performance of the EBP.
References

[1] Battese GE, Harter RM and Fuller 1988 An error components model for prediction of county crops areas using survey and satellite data Journal of the American Statistical Association, 80 28-36

[2] Berg E and Chandra H 2014 Small Area Prediction for a Unit-Level Lognormal Computational Statistics and Data Analysis 78 159-175

[3] Foster J, Greer J and Thorbecke E 1984 A class of decomposable poverty measures Econometrica 52 761–766

[4] Molina I and Rao JNK 2010 Small Area Estimation of Poverty Indicators The Canadian Journal of Statistics 38(3) 369-385

[5] Molina I and Rao JNK 2014 Small Area Estimation of General Parameters With Application to Poverty Indicators : Hierarchical Bayes Approach The Annals of Applied Statistics 8(2) 852-885

[6] Prasad NGN and Rao JNK 1990 The Estimation of the Mean Squared Error of Small-Area Estimators Journal of the American Statistical Association 85(409) 163-171

[7] Rao  JNK and Molina I 2015 Small Area Estimation New York : John Wiley and Sons.