The extended nested-error regression model with penalized spline function for estimation under informative sampling

A N R Chytrasari\textsuperscript{1,2*}, Sri Haryatmi Kartiko\textsuperscript{1}, Danardono\textsuperscript{1}

\textsuperscript{1} Department of Mathematics, Universitas Gadjah Mada, Jl. Sekip Utara Yogyakarta, Indonesia
\textsuperscript{2} Department of Mathematics Education, Universitas Pancasakti Tegal, Jl. Halmahera Km 1 Kota Tegal, Indonesia

* angela.nina.r@mail.ugm.ac.id

Abstract. Model-based estimates have been developed in statistics. Often, the statisticians use the data that obtained through complex sampling designs for estimation. If the probability of inclusion in the sampling design is informative, then the informativeness of sampling must be taken into account in the estimation process. We extended the nested error regression model at the unit level by adding the probability of inclusion as a covariate in the model but with an unknown functional form to reduce the informativeness effect. This extended part in the model is then approached using a penalized spline function. In the mixed model framework, we derived the EBLUP estimator for the mean areas of population. A simulation is given to applying this approach by using the first order of the p-spline function. The RMSPE value and the average absolute bias value obtained through the use of the bootstrap method then compared with that results from the approach using the nested error regression model.

1. Introduction

Estimates in statistics require data. Often, the data used obtained from informative sampling where the probability of inclusion correlates with the observed variables. Some researchers have taken into account the effects of informativeness sampling in the estimation process as an effort to reduce the resulting bias. One attempt is the use of weights in the prediction approach based on the model. \cite{1} has used design weights in the estimation of statistical models in the Bayesian context, while \cite{2} has built a pseudo-posterior distribution that uses sampling weights based on the probability of marginal inclusion. By \cite{3}, the inversion weight was multiplied by the covariate function and then smoothed it with the covariate value and the result variable. Several researchers offer different approaches in reducing the informativeness effect of the sample. \cite{4} extends the nested-error regression model by adding a chosen function of inclusion probability as an auxiliary variable in the sample model. \cite{5} extends the best predictor empirically under the normal-log model and uses a weighted-survey estimate for the parameter model using the same design weights as the inverse chance of inclusion. Although \cite{4} has provided an approach that does not require a new method; it will be difficult to determine the appropriate form of the probability inclusion function when the plot of the nested model of each of the four inclusion probability function options does not show a linear relationship.
This paper uses ideas [4], but the penalized spline (p-spline) function approach is used to approach the probability inclusion function added as a covariate to the model. We propose this approach for cases, where the probability inclusion function cannot be specified beforehand. The use of p-spline in mixed models is not new, see [7], [8], and [9]. But its use as an approach added to the model to take into account the effect of sample informativeness is the new approach proposed in this paper. The coefficient in the model is estimated using the BLUE Method. The estimated model is then used to predict the average area in the population.

2. Methodology

2.1. The extended nested-error regression model

Let a population U consists of units partitioned in m areas. The area \( i \) \( (i = 1, \ldots, m) \) contains \( N_i \) population units. In each \( i \), we take a sample of \( n_i \) units with probability \( \pi_i \). Battese, Harter, and Fuller in [8], have stated the error-nested regression model at unit level as

\[
y_{ij} = x_{ij} \beta + w_i + e_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n_i, \tag{1}
\]

where \( x_{ij} \) an auxiliary variable for unit \( j \) in area \( i \), \( w_i \) is the random effect of area \( i \), and \( e_{ij} \) is a unit \( j \) error in area \( i \). In the literature, (1) widely used in estimating small area problems. In this paper, \( \pi_i \) is an inclusion probability of unit \( j \) in area \( i \) resulting from the specified informative sampling design.

Equation (1.1) extended by adding a function of the probability of inclusion, \( m(\pi_{ij}) \), which is then approached by the penalized-spline function, i.e.

\[
y_{ij} = x_{ij} \beta + m(\pi_{ij}) + w_i + e_{ij}
\]

\[
= x_{ij} \beta + \sum_{k=1}^{p} \tau_{ij} t_k + \sum_{k=1}^{s} \kappa_{ij} (\pi_{ij} - Q_k)^p + w_i + e_{ij}
\tag{2}
\]

for \( i = 1, \ldots, m; \quad j = 1, \ldots, n_i \), where \( x_{ij} = [1 \ x_{ij}] \), \( p \) is order of spline, \( t_k \) and \( \kappa_k \) respectively are vector coefficients of parametric and non-parametric parts, \( Q_k \) is the \( k \)-th knot in \( \pi \), and \( s \) \( (\pi_{ij} - Q_k)^p \) is \( \max(0,(\pi_{ij} - Q_k)) \). Using all sample data, mixed-model (2) can be restated in matrix form as

\[
y = X \beta + \pi \tau + Z_1 s + Z_2 w + e
\tag{3}
\]

Estimates for 1 and 2 and predictors for 3 and four can be obtained simultaneously through the Henderson mixed model equation. This equation is generated from maximizing the function of combined density log-likelihood

2.2. Estimation for coefficients of fixed and random effect

Estimators for \( \beta \) and \( \tau \) and predictors for \( s \) and \( w \) can be obtained simultaneously through the Henderson mixed model equation. This equation is generated from maximizing the log-likelihood of the joint density function of \( y, s, \) and \( w \). We assume \( s \sim \text{iid } N(0, V_1), w \sim \text{iid } N(0, V_2), e \sim N(0, V_3) \) where \( V_1 = \sigma_1^2 I_p, V_2 = \sigma_2^2 I_k, V_3 = \sigma_3^2 I_n \) with \( n = \sum_{i=1}^{m} n_i \). Also, we assume the distribution of conditional density \( y \) given \( s \) and \( w \) as \( (y | s, w) \sim N(x \beta + \pi \tau + z_1 s + z_2 w, V_3) \). Furthermore, using the similarity
\( f(y, s, w) = f(y|s, w).f(s).f(w) \), we can state the function of the logarithm of the likelihood function \( s \) and \( w \), i.e.

\[
\log L(s, w) = -\frac{1}{2} \left\{ (y - xb - \pi t - z_i s - z_2 t)^T V^{-1}_3 (y - xb - \pi t - z_i s - z_2 t) + s^T V^{-1}_2 s + w^T V^{-1}_3 w \right\}
\]

Equation (4) is then maximized against \( b, t, s, \) and \( w \) to obtain Henderson's mixed model equation,

\[
\begin{bmatrix}
x V^{-1}_2 x & x V^{-1}_2 z_2 & x V^{-1}_2 z_2 & x V^{-1}_2 z_2 & b \\
p V^{-1}_2 x & p V^{-1}_2 z_2 & p V^{-1}_2 z_2 & p V^{-1}_2 z_2 & t \\
z_i V^{-1}_2 x & z_i V^{-1}_2 z_2 & (z_i V^{-1}_2 z_i + V^{-1}_i) & z_i V^{-1}_2 z_2 & s \\
z_2 V^{-1}_2 x & z_2 V^{-1}_2 z_2 & z_2 V^{-1}_2 z_2 & (z_2 V^{-1}_2 z_2 + V^{-1}_i) & w \\
\end{bmatrix}
\begin{bmatrix}
x V^{-1}_2 y \\
p V^{-1}_2 y \\
z_i V^{-1}_2 y \\
z_2 V^{-1}_2 y \\
\end{bmatrix}
= \begin{bmatrix}
\end{bmatrix}
\]

If the variance \( V_1, V_2, V_3 \) is known, then the solution of (5) is the estimators \( \hat{b}, \hat{t}, \hat{s}, \) and \( \hat{w} \). In line with [9], it can be proved that estimators and predictors are unbiased. Also, \( \hat{s} \) and \( \hat{w} \) are the best predictors for \( s \) and \( w \).

2.3. REML estimation for variance components

If the variance in the model is unknown, it must be estimated. In this paper, the variance component is estimated using the restricted maximum likelihood (REML) method. REML estimates based on a linear combination of element \( y \), namely \( K'y \), which chosen so that \( K'y \) does not contain a fixed effect, i.e \( K'x = 0 \) and \( K'\pi = 0 \). In this case, the \( K' \) matrix that meets is \( K' = c'M \) where \( c' \) any vector of constant constants (1xn) and \( M = I - R(R'R)^{-1}R' \) with partition matrix \( R = [x \ t] \).

Assume \( y \sim N(xb + \pi t, V) \) where \( V = \text{var}(y) = \sum_{i=1}^{3} z_i V_i z_i' \) so \( K'y \sim N(0, K'VK) \). Furthermore, maximizing the likelihood function for \( K'y \), we can obtain variance component estimators in this case, namely

\[
\hat{\sigma}^2_s = \frac{y'(y - \hat{x}\hat{b} - \pi\hat{t} - z_i \hat{s} - z_2 \hat{w})}{n-p-2}
\]

\[
\hat{\sigma}^2_s = \frac{s's}{p-tr(T_{11})}
\]

\[
\hat{\sigma}^2_w = \frac{w'w}{m-tr(T_{22})}
\]

Where \( T_{i} \) is the \( i \)-th diagonal matrix of \( T = (I + z'HzD)^{-1} \), with \( z = [z_1, z_2] \), \( H = V^{-1}_3 - V^{-1}_2 R (R'V^{-1}_3 R)^{-1} R'V^{-1}_2 \), and \( D = I_d V_{i+j}^2 \). Equation (8) – (10) is then used to estimate \( V_1, V_2, \) and \( V_3 \) iteratively with the following steps: (a) determine the initial value for the component of variance \( \gamma^{(0)} = (\sigma^2_s, \sigma^2_w) \); (b) determine \( \hat{b}^{(0)}, \hat{t}^{(0)}, \hat{s}^{(0)}, \) and \( \hat{w}^{(0)} \) by substituting results of (a) to (5) then completing it; (c) calculate ; (d) calculate \( \gamma^{(i)} \) using results (c) on (6)–(8); (e) repeat step (d) to get a convergent \( \gamma \) value; (f) the results (e) obtained are then substituted to the solution of (5) to obtain empirical estimates for \( b, t, s, \) and \( w \).

2.4. Mean area prediction and performance

Predictions for the mean areas in a population are predicted using formulas
\[
\hat{\mu}_i = \frac{1}{N_i} \left\{ \sum_{s_{ij}} y_{ij} + \sum_{s_{ij}} \hat{y}_{ij} \right\}
\] (9)

where \( s_i \) and \( \bar{s}_i \) successively states the set of sample units and the set of non-sample units in area-\( i \), while \( \hat{y}_{ij} = x_{ij} \hat{b} + \pi_{ij} \hat{t} + z_{ij} \hat{h}_1 + \ldots + \hat{w}_i \) is the predictor for \( y_{ij} \); \( j \in \bar{s}_i \) using the proposed approach. This prediction performance measured through root mean squared prediction error (RMSPE) which is calculated using the bootstrap approach with repetition of \( B = 1000 \) times using the formula

\[
\text{RMSPE}_{\text{boot}}(\hat{\mu}_i) = \left( \frac{1}{B} \sum_{b=1}^{B} (\hat{\mu}_i^*(b) - \bar{y}_i^*(b))^2 \right)^{1/2}
\] (10)

where \( \hat{\mu}_i^*(b) \) and \( \bar{y}_i^*(b) \) successively are values \( \hat{\mu}_i^* \) and \( \bar{y}_i^* \) for the \( b \) bootstrap repetition of the sample and population. We also calculate the average of RMSPE for every sample measure and determine absolute bias with

\[
\text{ABS}(\hat{\mu}_i) = \left| \frac{1}{B} \sum_{b=1}^{B} (\hat{\mu}_i^*(b) - \bar{y}_i^*(b)) \right|
\] (11)

3. Results And Discussion

The following simulation applies the proposed approach. We built 10,000 population units which included \( x \sim N(0,1) \), \( w \sim N(0,1) \), \( e \sim N(0,0.5)^2 \). The values \( y \) are built using (1). Partitioned population in 20 areas, each area contains 500 units. In each area, units sample taken with size variations of 5, 10, 25, 50, and 100 unit, respectively with probability \( \pi_{ij} = n_i b_{ij} / \sum_{k=1}^{20} b_{kj} \) where \( b_{ij} = \exp(-e_{ij} + \nu_i + \delta_{ij}) \), with \( \delta \sim N(0,1) \). This simulation is limited to the use of the first order of \( p \)-spline with knots determined using the quantile formula refer to [6], i.e

\[
Q_k = \text{quartile number} \left( \frac{k + 1}{K + 2} \right) \text{ of unique value of } \pi_{ij},
\]

for \( k = 1, \ldots, K \) with \( K = \min \left( \frac{1}{4} \times \text{amount of unique } \pi_{ij}, 35 \right) \). We determine the number of unique \( \pi_{ij} \) from the population. Our model approach is

\[
M1: \quad y_{ij} = x_{ij} b + \pi_{ij}^* t_i + \sum_{k=1}^{K} s_k (\pi_{ij} - Q_k) + w_i + e_{ij}, \quad i = 1, \ldots, 20; \quad j = 1, \ldots, n_i
\]

The model coefficient is estimated by completing (5), and the corresponding variance component is estimated using (6), (7), and (8). Furthermore, the estimated model is used to calculate the twenty average areas. Model performance is then calculated using (10) and (11). We compare the results of this approach with the results obtained using the nested error regression model (1) without taking into account informativeness.

The M1 model produces RMSPE values that are smaller than those produced by the nested-error regression model with an average difference of 1.0764 for the five sample size variants. The larger the sample size does not guarantee the smaller RMSPE values and the absolute bias produced. It was also seen that the distribution of RMSPE values generated by the nesting error regression model tends to be more spread than the RMSPE value produced by M1. This condition shows that the M1 model
approach tends to provide more accurate results than the nested-error regression model. As shown in the boxplot graph in Figure 1 below.

![Figure 1](image)

**Figure 1.** Distribution of RMSPE values for extended model approaches using p-spline first order is better than that for nested-error models.

From table 1 it can be seen that the average RMSPE value and the average absolute bias of the M1 model are smaller than those generated through the nested-error regression model with the mean differences of 1.074951 and 0.03404 respectively. However, the addition of sample size does not seem to guarantee a decrease in the RMSPE value or an average absolute bias in the proposed approach.

| $n_i$ | M1 model | Nested error regression model |
|------|----------|-------------------------------|
|      | RMSPE    | $AB$                          | RMSPE           | $AB$                          |
| 5    | 0.048274706 | 0.001526580                  | 1.124.252.035   | 0.035551971                   |
| 10   | 0.049441526 | 0.001563478                  | 1.126.788.126   | 0.035632169                   |
| 25   | 0.053823390 | 0.001702045                  | 1.123.651.658   | 0.035532985                   |
| 50   | 0.040645328 | 0.001285318                  | 1.122.838.008   | 0.035507255                   |
| 100  | 0.041805233 | 0.001321998                  | 1.118.456.404   | 0.035368697                   |

**Table 1.** The RMSPE average and the mean absolute bias values generated by the mixed model are extended with the first-order p-spline approach better than those provided by error-nest regression models for each sample size.

4. Conclusion
From the simulation results above, it can be concluded that the proposed approach can be used as an alternative predictive approach under informative sampling. This approach is an initial idea that can still be developed to overcome more complex problems. As in [4], our method requires information about the probability of inclusion values for all population units. However, this approach does not need graph plot analysis to determine the form of the inclusion probability function chosen to add to the model.
References

[1] Little R J 2012 Calibrated Bayes, an alternative inferential paradigm for official statistics Journal of Official Statistics 28 309–34
[2] Savitsky T D and Toth D 2015 Bayesian estimation under informative sampling arXiv:1507.07050 [math, stat]
[3] Kim J K and Skinner C J 2013 Weighting in survey analysis under informative sampling Biometrika 100 385–98
[4] Verret F 2015 Model-based small area estimation under informative sampling Survey Methodology 41 333–47
[5] Zimmermann T and Münnich R T 2018 Small area estimation with a lognormal mixed model under informative sampling Journal of Official Statistics 34 523–42
[6] Chen H and Wang Y 2011 A penalized spline approach to functional mixed effects model analysis Biometrics 67 861–70
[7] Opsomer J D, Claeskens G, Ranalli M G, Kauermann G and Breidt F J 2008 Non-Parametric Small Area Estimation Using Penalized Spline Regression Journal of the Royal Statistical Society. Series B (Statistical Methodology) 70 265–86
[8] Rao J N K, Sinha S K and Dumitrescu L 2014 Robust small area estimation under semiparametric mixed models The Canadian Journal of Statistics / La Revue Canadienne de Statistique 42 126–41
[9] Ruppert D, Wand M P and Carroll R J 2003 Semiparametric Regression (Cambridge: Cambridge University Press 404) p 126