Nested Error Non-parametric Unit Level Model performance in the context of empirical Bayes (EB) approach

Patrick Munyangabo[1], **Anthony Waititu** [2] **and Anthony Kibira Wanjoya**[3]

Abstract

In this paper, we assess the performance of the proposed model compared with the standard unit-level model, the performance of both models were evaluated in the case of skewed data. The empirical Bayes (EB) estimates obtained under both models were compared through different criteria proposed by the panel on small area estimates of population and income set up by the United States Committee on National Statistics in 1978 and the proposed model was found to fit better than the standard one. The MSE for both models were checked using the bootstrap technique and the proposed model win over the standard model at each bootstrap iterations.

Mathematics Subject Classification : xxxxx
Keywords: xxxxx, xxxxx, Small area estimation,

[1] Department of Mathematics, Pan African University, Institute of Basic Sciences, Technology, and Innovation, Kenya, e-mail: munyangabo.patrick@students.jkuat.ac.ke
[2]Affiliation of the first author, e-mail: awaititu@jkuat.ac.ke
[3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Kenya, e-mail: awanjoya@fsc.jkuat.ac.ke

Article Info: *Revised* : ..... *Published online* :

1 Introduction

In statistical inference, the empirical Bayes methods are procedures used when the prior distribution is not known but estimated from data. In small area estimation, the empirical Bayes methods have been proven to be efficient data-analysis tools when data violent some normality assumptions. There are used in the estimation of the nonlinear parameter under the basic unit model and quite number of application to poverty mapping[6]. The empirical Bayes model is much richer than either the classical or the ordinary Bayes model and often provides superior estimates of parameters[8]. In this paper, we assess the performance of the proposed model [5] comparing with the basic unit-level model proposed by Fuller, Battese and Harter,1988)[7]. To achieve our goal, we examined the the empircal Bayes estimates for both models through out two general standpoints: the accuracy of point estimates and Mean Square Error(MSE) of the estimates . The former was checked through the relative errors and absolute relative errors of the empirical Bayes estimates, the later was estimated by bootstrap procedure similar to the bootstrap algorithm proposed by[3].

2 Empirical Bayes (EB) approaches

2.1 Basic Unit level Model

In small area estimation, a basic unit-level model(FBH) is based on unit level auxiliary variables i.e $X_{ij} = (x_{ij1}, \cdots, x_{ijp})^T$ are available for population element j in each small area i.These variables are related to unit level values of response y_{ij} through a nested error linear regression model[6].

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij}. \tag{1}$$

where y_{ij} is the response of unit $j, j = 1, 2, \cdots, N_i$ in area $i, i = 1, 2, \cdots, m$. x_{ij} is the vector of auxiliary variables, β is the vector of the regression parameters, the random effect v_i represents the effect of area characteristics are

not accounted for by auxiliary variables X_{ij} and e_{ij} the individual unit error term.

The area effects v_i are assumed independent with mean zero and variance σ_u^2, The errors e_{ij} are independent with mean zero and variance σ_e^2. In addition, the v_i's and e_{ij}'s are assumed to be independent.

The small area parameter of interest, θ_i, may be approximated by

$$\theta_i = \bar{X}_i^T \beta + v_i, \quad i = 1, \cdots, m. \tag{2}$$

assuming that N_i is large, where $\bar{X}_i$ is the average vector of population of the x_{ij} for the ith area, that is, $\bar{X}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$. The sample data $\{y_{ij}, x_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, m\}$ are assumed to obey model 1, i.e.,

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij}, j = 1, \cdots, n_i, \quad i = 1, \cdots, m. \tag{3}$$

where n_i is the sample size in the ith small area. This implies that the selection bias is absent. For a proof of this absence bias, see (Rao, 2003). Assuming that $v_i \sim \mathcal{N}(0, \sigma_u^2), i = 1, \cdots, m$ and expressing $y_{ij}|\beta, v_i, \sigma_e^2 \sim \mathcal{N}(x_{ij}^T \beta + v_i, \sigma_e^2), j = 1, \cdots, n_i, i = 1, \cdots, m$, we have the distribution of v_i from the unit-level model 3 conditional on $y_{ij}, \beta, \sigma_u^2$ and σ_e^2 as

$$v_i|y_{ij}, \beta, \sigma_u^2, \sigma_e^2 \sim \mathcal{N}\left(\Omega_i(\bar{y}_i - \bar{x}_i^T \beta), \Omega_i \frac{\sigma_e^2}{n_i}\right)$$

where $\Omega_i = \sigma_u^2/(\sigma_u^2 + n_i^{-1}\sigma_e^2)$, $(\bar{y}_i, \bar{x}_i)$ are the sample means for the i^th area and the regression coefficient are obtained by using the relation
$\beta = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y$. Therefore,

$$\mathbb{E}[y_{ij}, \beta, \sigma_u^2, \sigma_e^2] = \Omega_i(\bar{y}_i - \bar{x}_i^T \beta) \tag{4}$$

Using equation 2 and 4, we get the Bayes predictor of θ_i, θ_i^B as

$$\begin{aligned} \theta_i^B &= \mathbb{E}[\theta_i|y_{ij}, \beta, \sigma_u^2, \sigma_e^2] \\ &= (\bar{X}_i^T \beta + \Omega_i(\bar{y}_i - \bar{x}_i^T \beta) \\ &= (1 - \Omega_i)\bar{X}_i^T \beta + \Omega_i[\bar{y}_i + (\bar{X}_i - \bar{x}_i)^T \beta] \end{aligned} \tag{5}$$

Replacing β, σ_u^2 and σ_e^2 by $\hat{\beta}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ respectively we obtain the Empirical Bayes predictor of θ_i, θ_i^{EB}

$$\theta_i^{EB} = (1 - \hat{\Omega}_i)\bar{X}_i^T\hat{\beta} + \hat{\Omega}_i[\bar{y}_i + (\bar{X}_i - \bar{x}_i)^T\hat{\beta}] \tag{6}$$

where $\hat{\Omega}_i = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + n_i^{-1}\hat{\sigma}_e^2)$ and $\hat{\beta} = \left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}y$

2.2 Proposed unit level Model

In the proposed model, we assume that unit specific auxiliary data $X_{jd} = (1, x_1, jd, \cdots, x_{p-1}, jd)$ are available for each population element d in each small area j. Further, variable of interest y_{jd}, is assumed to follow log-normal distribution and to be related to X_{jd} through a nested error non-parametric unit level model[5].

$$log(y_{jd}) := l_{jd} = m(x_{jd}) + u_j + e_{jd} \tag{7}$$

where the function $m(\cdot)$ is unknown, but estimated by locally weighted regression (loess or lowess) and $log(y_{jd})$ is the response of unit d, $\quad d = 1, 2, \cdots, N_j$ in area j, $\quad j = 1, 2, \cdots, p$, x_{jd} is the vector of auxiliary variables and The u_d and e_{jd} are mutually independent with zero mean and variance σ_u^2 and σ_e^2 respectively.

The small area parameter of interest, θ_j, may be approximated by

$$\theta_j = m(\bar{X}_j^T) + v_j, \quad j = 1, \cdots, p. \tag{8}$$

assuming that N_j is large, where $\bar{X}_j$ is the average vector of population of the x_{jd} for the jth area, that is, $\bar{X}_j = \sum_{d=1}^{N_j} x_{jd}/N_j$. The sample data $\{l_{jd}, x_{jd}, d = 1, \cdots, n_j, j = 1, \cdots, p\}$ are assumed to obey model 7, i.e.,

$$l_{jd} = m(x_{jd}^T) + v_j + e_{jd}, d = 1, \cdots, n_j, \quad j = 1, \cdots, p. \tag{9}$$

where n_j is the sample size in the jth small area. This implies that the selection bias is absent. Assuming that $v_j \sim \mathcal{N}(0, \sigma_u^2), j = 1, \cdots, p$ and

expressing $l_{jd}|m(\cdot), v_j, \sigma_e^2 \sim \mathcal{N}(m(x_{jd}^T) + v_j, \sigma_e^2), d = 1, \cdots, n_j, j = 1, \cdots, p$, we have the distribution of v_j from the model 9 conditional on $l_{jd}, m(\cdot), \sigma_u^2$ and σ_e^2 as

$$v_j|l_{jd}, m(\cdot), \sigma_u^2, \sigma_e^2 \sim \mathcal{N}\left(B_j(\bar{l}_j - m(\bar{x}_j^T)), B_j\frac{\sigma_e^2}{n_j}\right)$$

where $B_j = \sigma_u^2/(\sigma_u^2 + n_j^{-1}\sigma_e^2)$, $(\bar{l}_j, \bar{x}_j)$ are the sample means for the j^th area and the mean function is given by $m(\cdot)$.

Therefore,

$$\mathbb{E}[l_{jd}, m(\cdot), \sigma_u^2, \sigma_e^2] = B_j(\bar{l}_j - m(\bar{x}_j^T)) \tag{10}$$

Using equation 8 and 10, we get the Bayes predictor of θ_j, θ_j^B as

$$\begin{aligned}\theta_j^B &= \mathbb{E}[\theta_j|l_{jd}, m(\cdot), \sigma_u^2, \sigma_e^2] \\ &= (m(\bar{x}_j^T) + B_j(\bar{l}_j - m(\bar{x}_j^T)) \\ &= (1 - B_j)m(\bar{x}_j^T) + B_j[\bar{l}_j + m(\bar{X}_i - \bar{x}_i)^T]\end{aligned} \tag{11}$$

Replacing $m(\bar{x}_j^T), \sigma_u^2$ and σ_e^2 by $\widehat{m}(\bar{x}_j^T)$,$\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ respectively we obtain the Empirical Bayes predictor of θ_j, θ_j^{EB}

$$\theta_j^{EB} = (1 - \hat{B}_j)\widehat{m}(\bar{x}_j^T) + \hat{B}_j[\bar{l}_j + \widehat{m}(\bar{X}_j - \bar{x}_j)^T] \tag{12}$$

where $\hat{B}_j = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + n_j^{-1}\hat{\sigma}_e^2)$ and $\widehat{m}(\cdot)$ is a mean function estimated using loess.

3 Simulation Study

In our simulation study, the performance of the basic unit-level(FBH) and proposed model were checked using left skewed simulated data. The variable of interest y_{ij} or y_{jd} were randomly selected from Burr Distribution with parameters $(c = 1, k = 4, rate = 1)$ and the covariate values X_{ij} or X_{jd} were generated from Poisson distribution of $\lambda = (5)$, for 30 small areas of sizes $(n = 1000, N = 3000)$ to satisfy the skewness condition for both models. The

random errors and the random area effects are mutually independent with zero mean and variance σ_u^2 and σ_e^2 respectively for both models. The plots for both models using simulated data are shown in figure 2 and 3.

4 Performance Criteria

For the performance of Empirical Bayes estimators were examined from two general standpoints: the accuracy of point estimates and Mean Square Error of the estimates. The former was checked through the relative errors and absolute relative errors of the Empirical Bayes estimates, the latter was estimated by similar Bootstrap Technique proposed by [3]. For the empirical comparison of relative errors and absolute relative errors of the EB estimates, the following four different criteria recommended by the panel on small area estimates of population and income set up by the United States Committee on National Statistics in 1978,[4].

4.1 Empirical Comparison of Empirical Bayes Estimates

Suppose e_{iTR} denotes the true income for the i^{th} area, and e_i is any estimate of $e_{iTR}; i = 1; \cdots ; m$. Then

1 Average relative bias : $(ARB) = \frac{1}{m} \sum_{i=1}^{m} |\frac{e_i - e_{iTR}}{e_{iTR}}|$,

2 Average squared relative bias : $(ASRB) = \frac{1}{m} \sum_{i=1}^{m} (\frac{e_i - e_{iTR}}{e_{iTR}})^2$,

3 Average absolute bias : $(AAB) = \frac{1}{m} \sum_{i=1}^{m} |e_i - e_{iTR}|$,

4 Average squared deviation : $(ASD) = \frac{1}{m} \sum_{i=1}^{m} (e_i - e_{iTR})^2$.

4.2 Estimation of MSE Using Bootstrap Algorithm for the Basic Unit level Model

After getting the parameter $\hat{\beta}$ for the Basic unit-level model from the simulated data, we need to gauge the performance of the Basic unit-level model by estimating the mean-squared error of small area parameter θ_i^{EB}. We have achieved this by using the bootstrap technique of [3]. This technique enables us to obtain a bias-corrected, mean-squared error estimator of θ_i^{EB}. We generated 100 bootstrap samples at each level of the bootstrap as described in the algorithm below:

$$1.\text{Do this} \quad b_1 = 1; \cdots ; B_1 \text{times}$$

$$\text{generate} \quad x_{b_1}^* \quad \text{from} \quad \mathcal{P}(n = 1000, \lambda = 5)$$

$$\text{generate} \quad y_{b_1}^* \quad \text{from} \quad rburr(n = 1000, 1, 4, 1)$$

$$\text{get}\hat{\beta}_{b_1}^* \quad \text{from the bootstrap data} \quad (y_{b_1}^*; x_{b_1}^*)$$

$$\text{calculate} \quad v_{b1}^* = \bar{y}_{b1}^* - \bar{X}_{b1}^*$$

$$\text{calculate} \quad e_{b1}^* = y_{b1}^* - x_{b1}^* \beta_{b1}^* - v_{b1}^*$$

$$\text{calculate} \quad \theta_{b_1}^*(\hat{\beta}_{b1}^*) = \bar{X}_{b1}^* \hat{\beta}_{b1}^* + v_{b_1}^*$$

$$\text{calculate} \quad \hat{\theta}_{b_1}^{*EBP} = \bar{X}_{b_1}^* \hat{\beta}_{b_1}^* + \hat{\Omega}[\bar{y}_{b_1}^* + (\bar{X}_{b_1}^* - \bar{x}_{b_1}^*)^T \hat{\beta}_{b_1}^*]$$

Finally, the Mean Sum square is obtained as:

$$\widehat{(MSE)} = \frac{1}{B_1} \sum_{b_1=1}^{B_1} (\hat{\theta}_{b_1}^{*EBP} - \theta_{b_1}^*(\hat{\beta}_{b_1}^*))^2$$

4.3 Estimation of MSE Using Bootstrap Technique for the Proposed Model

After generating skewed data, the proposed model was used. We considered the span/bandwidth of 0.5 to tackle the issue of over and underestimation, more details [5]. The loess method was used to estimate the mean function $m(\cdot)$ of the proposed model. The performance of the proposed model was measured by estimating the mean squared error of small area θ_j^{EB}. We have achieved this by using the bootstrap technique of [3]. This technique enables us to obtain a bias-corrected, mean-squared error estimator of θ_j^{EB}. We generated 100 bootstrap samples at each level of the bootstrap as described in the algorithm below:

$$\text{1.Do this} \quad b_1 = 1; \cdots ; B_1 \text{times}$$

$$\text{generate} \quad x^*_{b_1} \text{from} \quad \mathcal{P}(n = 1000, \lambda = 5)$$

$$\text{generate} \quad y^*_{b_1} \quad \text{from} \quad rburr(n = 1000, 1, 4, 1)$$

$$\text{calculate} \quad v^*_{b1} = \bar{y}^*_{b1} - m(\bar{X}^*_{b1})$$

$$\text{calculate} \quad e^*_{b1} = l^*_{b1} - m(x_{b1})^* - v^*_{b1}; \quad l^*_{b1} = log(y^*_{b1})$$

$$\text{calculate} \quad \theta^*_{b_1} = m(\bar{X}^*_{b1}) + v^*_{b_1}$$

$$\text{calculate} \quad \hat{\theta}^{*EBP}_{b_1} = m(\bar{X}^*_{b_1}) + \hat{B}[\bar{l}^*_{b_1} + m(\bar{X}^*_{b_1} - \bar{x}^*_{b_1})^T]$$

Finally, the Mean Sum square is obtained as:

$$\widehat{(MSE)} = \frac{1}{B_1} \sum_{b_1=1}^{B_1} (\hat{\theta}^{*EBP}_{b_1} - \theta^*_{b_1})^2$$

5 Simulation Results

In this section, we presented the results from the simulation study and their corresponding discussions.

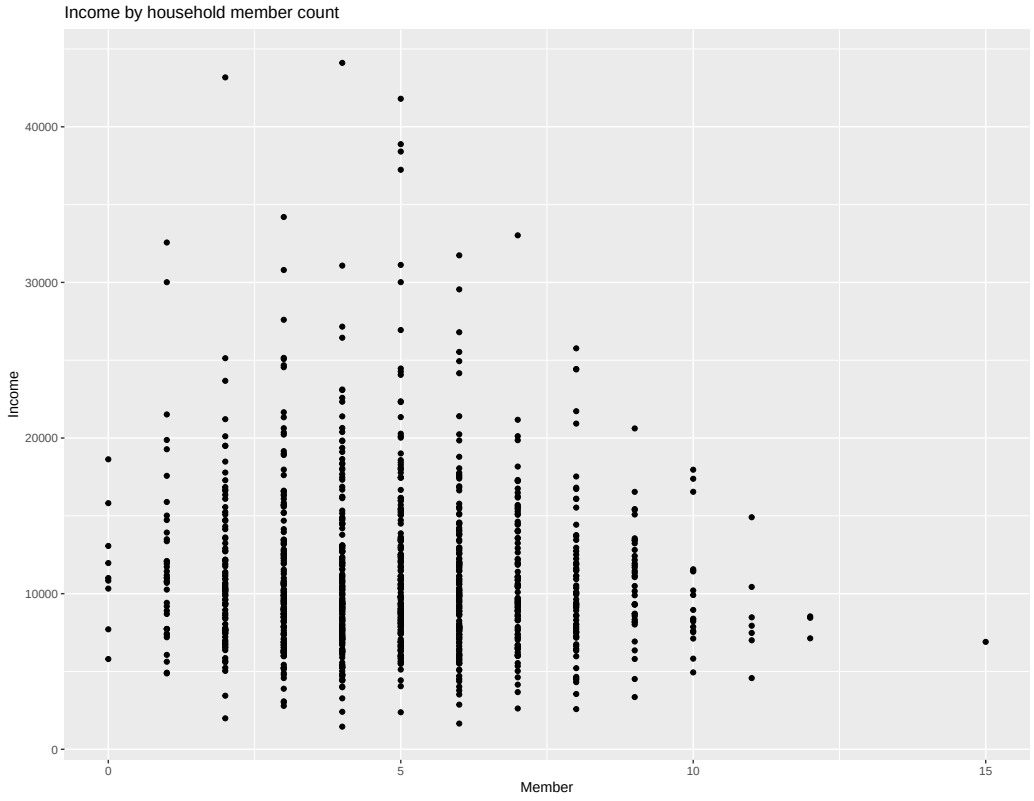

Figure 1: Plot of the simulated data for i=1.

The figure 1 presents the simulated data for $i = 1$ small area, it is left skewed. The maximun income is 40000 and the maximun number of members in household is 15 for $i = 1$. We have simulated data for $i = 2, \cdots, 30$ in all small areas but we won't plot them here.

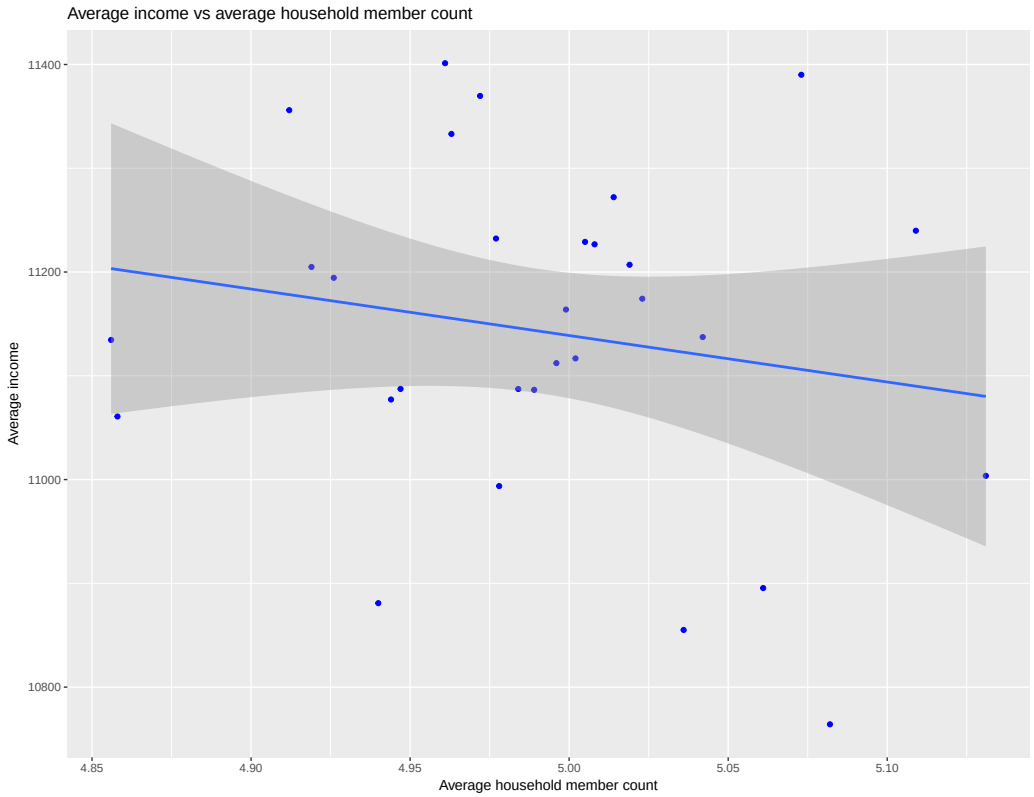

Figure 2: Fitted of Simulated data using basic unit-level model(FBH)

The figure 2 describes the fitted data using the basic unit-level model(FBH). The model uses the linear method to fit the simulated data for $i = 1, \cdots, 30$, it is uses the average income and average household members in for $i = 1, \cdots, 30$ in all small areas.

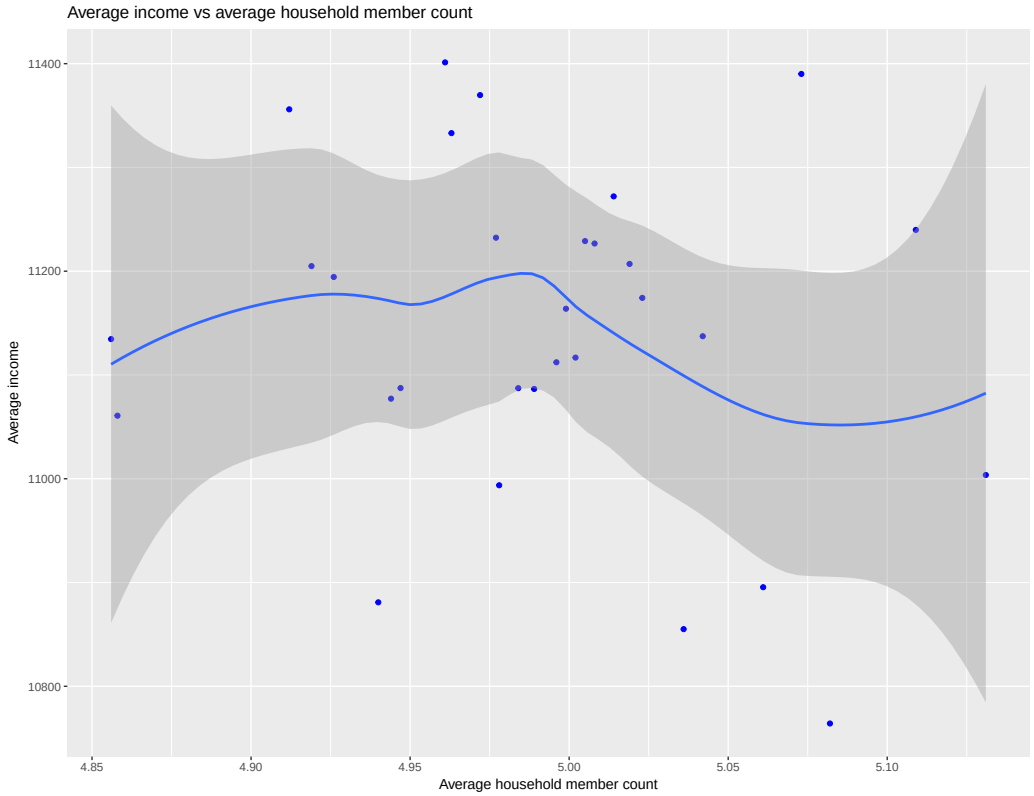

Figure 3: Fitted Simulated data using proposed model

The figure 3 describes the fitted data using the proposed model. The model uses the loess method to fit the simulated data for $j = 1, \cdots, 30$,by averaging income and average household members in for $j = 1, \cdots, 30$ in all small areas.

Table 1: Empirical Comparison of EB Estimates under Standard unit level Model(FBH) and Proposed Model (PM) at span = 0.5 for skewed data using Four Different Criteria.

Models	**Average Absolute relative bias**	**Average squared relative bias**	**Average Absolute Bias**	**Average Squared Deviation**
PM	1.228593e-03	2.633468e-06	1.131202e-02	2.229792e-04
FBHM	1.206998e-02	2.286032e-04	1.344029e+02	2.832072e+04

The simulation results obtained for both models using the skewed data is presented in Table 1. The study was to investigated how the small area parameters θ would be affected under simulated data using the two models. We have compared the results of Basic unit level model and proposed model on the basis of average absolute relative bias, average squared relative bias, average absolute bias and average squared deviation under the simulation setting where data simulated is left skewed data for both models. Table 1 reports the values of these measures of empirical comparision. The results appear to suggest a better fit for the proposed method compared to the standard unit level model(FBH).

Table 2: Plot of empirical values of MSE using bootstrap approach: Comparison of the basic unit-level model(FBH) and the proposed model at span = 0.5

Bootstrap iterations	**FBH model**	**Proposed model**
10	31380.69	4e-04
50	34580.97	0.00041
80	31455.222	0.00038
100	33739.27	0.00039

Table 2 provides a clear comparison of the basic unit-level model(FBH) and the proposed model in terms of empirical MSE obtained via the bootstrap approach. The empirical value of MSE over 100 bootstrap iterations were obtained for $i = 1, \cdots, 30$ small area. We see that the proposed model is markedly better than the basic unit-level model(FBH) at span=0.5 in each bootstrap iterations.

6 Conclusion

Assuming the normally assumptions are not satisfied within the the small area unit, the proposed model at span=0.5 appears to performance well comparing to the standard unit-level model. However, more research is needed to integrate the non-parametric methods into small area estimation. We are currently working on the performance the proposed model under the hierarchical Bayes approach.

References

[1] Maiti, G. D. P. L. T. (2002). Empirical Bayes estimation of the median income of four-person families by state using time series and cross-sectional data. Journal of Statistical Planning and Inference, 102

[2] Pratesi, M., Giusti, C., and Marchetti, S. (2013). Small area estimation of poverty indicators. In Survey data collection and integration, pages 89101. Springer.

[3] Erciulescu, A. L. and Fuller, W. A. (2014). Parametric bootstrap procedures for small area prediction variance. In Proceedings of the Survey Research Methods Section. American Statistical Association Washington, DC.

[4] Fan, J., and Gijbels, I. (1996). Local polynomial modeling and its applications: monographs on statistics and applied probability 66, volume 66. CRC Press.

[5] Munyangabo, P., Waititu, A., Wanjoya, A. K., et al. (2019). Estimation of nested error non-parametric unit-level model. Journal of Statistical and Econometric Methods, 8(1):13

[6] Rao, J. N. K. (2003). Small Area Estimation. Wiley series in survey methodology, (May): xxiii, 313 p.

[7] 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

[8] Statistical Association, 83(401):28-36. Casella, G. (1985). An introduction to empirical Bayes data analysis. American Statistician - AMER STATIST, 39:83-87