

SMALL AREA ESTIMATION UNDER RANDOM
COEFFICIENT REGRESSION MODELS
WITH MEASUREMENT ERROR

Parimal Mukhopadhyay
University of Qatar
Doha

ABSTRACT

In this note we consider simultaneous estimation of small area totals of a finite population assuming multiple regression equation of the study variable on a set of p auxiliary variables for each small area, when the regression coefficients are independent samples from a p -variate distribution and when the true value of the variables cannot be measured but only values mixed with measurement errors.

KEY WORDS

Small Area, Random Coefficient Regression, Measurement Error.

1. INTRODUCTION

Let \mathcal{P} be a finite population consisting of A mutually exclusive and exhaustive small areas or domains \mathcal{P}_a of size N_a , $\mathcal{P} = \cup_a \mathcal{P}_a$, $N = \sum_a N_a$ ($a = 1, \dots, A$). A sample s of size n is drawn from \mathcal{P} using any sampling design, s_a of size n_a being the part of s which falls in \mathcal{P}_a , $s = \cup_a s_a$, $n = \sum_a n_a$. Assume that $n_a > 0 \forall a$. Let ' y ' be a study variable taking value y_{ai} on unit i belonging to area a ($i = 1, \dots, N_a; a = 1, \dots, A$). We assume that the true value y_{ai} cannot be observed but a different value Y_{ai} mixed with measurement errors. Also, assume that associated with (ai) there are fixed (non-stochastic) true values x_{aij} of p auxiliary variables x_j ($j = 1, \dots, p$) each of which is closely related to the main variable y of interest. However, like y, x_{aij} 's also

cannot be measured correctly, but some other values X_{aij} mixed with measurement errors. Our aim is to estimate simultaneously the area totals $T_a = \sum_{i=1}^{N_a} y_{ai}$ on the basis of the data $(Y_{ai}, X_{aij}, a = 1, \dots, A, i \in s_a, j = 1, \dots, p)$ and an assumed relationship between y and x_j 's.

2. THE MODEL

We assume that the true value y_{ai} is actually a realization of a random variable \mathcal{Y}_{ai} , the vector $\mathcal{Y} = (\mathcal{Y}_{11}, \dots, \mathcal{Y}_{1N_1}, \dots, \mathcal{Y}_{A1}, \dots, \mathcal{Y}_{1N_A})'$ having a joint distribution ξ_N . However, since both y_{ai} and \mathcal{Y}_{ai} are unknown, we make no notational distinction between them. We also assume that the sampling design is such that no selection bias is present i.e. the vector $y_s = (y'_{1s}, y'_{2s}, \dots, y'_{As})'$ where $y_{as} = (y_{a1}, y_{a2}, \dots, y_{an_a})'$ ($a = 1, \dots, A$) obeys ξ_n . Let

$$Y_{as} = y_{as} + u_{as}$$

with

$$E(u_{as}) = 0, E(u_{as}u'_{as}) = \sigma_{uu}I_{n_a} \tag{2.1}$$

$$E(u_{as}u'_{bs}) = 0 \quad (a \neq b = 1, \dots, A)$$

Here u_{ai} is a measurement error corresponding to observation Y_{ai} . Further, assume that

$$y_{as} = X_{as}\beta_a + \alpha_a 1_{n_a} + e_{as}$$

with

$$E(e_{as}) = 0, E(e_{as}e'_{as}) = \sigma_{ee}I_{n_a} \tag{2.2}$$

$$E(e_{as}e'_{bs}) = 0, (a \neq b = 1, \dots, A)$$

e_{as} being distributed independently of u_{as} . Here, $1_q = (1, \dots, 1)_{q \times 1}$, $X_{as} = ((X_{aij}, i \in s_a, j = 1, \dots, p))_{n_a \times p}$, $e_{as} = (e_{ai}, i \in s_a)_{n_a \times 1}$, e_{ai} being an error in equation corresponding to true value y_{ai} , $\beta_a = (\beta_{a1}, \dots, \beta_{ap})'$, a vector of regression coefficients, and α_a is a general effect due to area a . Assume $\text{rank}(X_{as}) = p \ \forall a$ and $\min_a n_a > p + 1$.

Also assume that

$$X_{as} = x_{as} + v_{as} \tag{2.3}$$

where $v_{as} = ((v_{aij}, i \in s_a; j = 1, \dots, p))_{n_a \times p}$, v_{aij} being the error of measurement corresponding to the true value x_{aij} . Assume that the errors $v_{ai} = (v_{ai1}, \dots, v_{aip})'$ are distributed independently of $v_{bk} [(ai) \neq (bk)]$, each having mean 0 and dispersion matrix $\mathcal{D}(v_{ai}) = \text{Diag}(\sigma_{v11}, \dots, \sigma_{vpp})$. Also, assume that v_{ai} is distributed independently of u_{ai} and e_{ai} .

Again, assume that the regression coefficient β_a are independent samples from a distribution with mean $\bar{\beta}$ and dispersion matrix D , i.e.

$$\beta_a = \bar{\beta} + \eta_a \tag{2.4}$$

with

$$\begin{aligned} E(\eta_a) &= 0, \mathcal{D}(\eta_a) = D \\ E(\eta_a \eta_b') &= 0 \quad (a \neq b = 1, \dots, A) \end{aligned}$$

We assume that η_a is distributed independently of u_{ai}, e_{ai} and v_{ai} . We shall assume that $\sigma_{uu}, \sigma_{vjj} (j = 1, \dots, p)$ are known and $\bar{\beta}, \alpha_a, \sigma_{eea} (a = 1, \dots, A)$ and D are unknown and require to be estimated. The general theory of random coefficient regression models have been considered by Swamy (1970), Dempster, Rubin and Tsutakawa (1981) and its applications in small area estimation by Prasad and Rao (1990), Lahiri and Rao (1995), among others.

3. BEST LINEAR UNBIASED ESTIMATION

We have

$$\begin{aligned} Y_{as} &= y_{as} + u_{as} \\ &= X_{as}\beta_a + \alpha_a 1 + e_{as} + u_{as} \\ &= X_{as}\bar{\beta} + X_{as}\eta_a + \alpha_a 1 + \phi_{as} \end{aligned} \tag{3.1}$$

i.e.

$$Y_s = \tilde{X}_s \gamma + Z_s \eta + \phi_s \tag{3.2}$$

where

$$\begin{aligned} Y_s &= (Y'_{1s}, \dots, Y'_{As})'_{n \times 1} \\ \gamma &= (\bar{\beta}', \alpha')'_{(p+A) \times 1}, \quad \alpha = (\alpha_1, \dots, \alpha_A)' \\ \eta &= (\eta'_1, \dots, \eta'_A)'_{(Ap) \times 1}, \quad \phi_s = (\phi'_{1s}, \dots, \phi'_{As})'_{n \times 1} \\ \phi_{as} &= (\phi_{a1}, \dots, \phi_{an_a})', \quad \phi_{ai} = e_{ai} + u_{ai} \\ \tilde{X}_s &= \begin{bmatrix} X_{1s} & 1_{n_1} & 0 & \dots & 0 \\ X_{2s} & 0 & 1_{n_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ X_{As} & 0 & 0 & \dots & 1_{n_A} \end{bmatrix}_{n \times (p+A)}, \quad Z_s = \bigotimes_{a=1}^A X_{as} \end{aligned} \tag{3.3}$$

where $\bigotimes_{t=1}^m (A_t)$ denotes the block diagonal matrix $\text{Diag} (A_1, \dots, A_m)$. Hence

$$\begin{aligned} \mathcal{D}(\phi) &= \sigma_{uu} I_n + \bigotimes_{a=1}^A (I_{n_a} \sigma_{eea}) \\ &= R \text{ (say)} \end{aligned} \tag{3.4}$$

We are required to predict the domain total $T_a = \sum_{i=1}^{N_a} y_{ai}$, given the data $\{Y_{as}, X_{as}, a = 1, \dots, A\}$. Using Royall's (1970) approach, a predictor of T_a is, therefore,

$$\hat{T}_a = \sum_{i \in s_a} Y_{ai} + \hat{\mu}_a \tag{3.5}$$

where $\hat{\mu}_a$ is a model-unbiased estimator of

$$\begin{aligned} \mu_a &= \sum_{i \in (\mathcal{P}_a - s_a)} E(y_{ai}) \\ &= l'_a \gamma + m'_a \eta \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} l'_a &= 1'_{N_a - n_a} [X_{a\bar{s}} \ 0 \ 0 \ \dots \ 0 \ 1 \ 0 \ \dots \ 0]_{(N_a - n_a) \times (p+A)} \\ m'_a &= 1'_{N_a - n_a} [0 \ \dots \ 0 \ X_{a\bar{s}} \ 0 \ \dots \ 0]_{(N_a - n_a) \times pA} \end{aligned} \tag{3.7}$$

$$X_{a\bar{s}} = ((x_{aij}, i \in \mathcal{P}_a - s_a, j = 1, \dots, p))_{(N_a - n_a) \times p}$$

Here \hat{T}_a is a predictor of a realized value of T_a .

Let θ be the $\nu \times 1$ vector containing the distinct elements of D and $\sigma_{ee1}, \dots, \sigma_{eeA}$, where $\nu = [\frac{p(p+1)}{2} + A]$. Assume that θ is unknown. It is then known due to Henderson (1975) that the best linear unbiased predictor (BLUP) of μ_a is

$$\hat{\mu}_a(\theta) = l'_a \hat{\gamma}_s + m'_a G Z'_s V_s^{-1} (Y_s - \tilde{X}'_s \hat{\gamma}_s) \tag{3.8a}$$

where

$$G = \text{Diag}(D)_{Ap \times Ap} \tag{3.8b}$$

$$V_s = R + Z_s G Z'_s \tag{3.8c}$$

is the variance-covariance matrix of Y_s and

$$\hat{\gamma}_s = (\tilde{X}'_s V_s^{-1} \tilde{X}_s)^{-1} \tilde{X}'_s V_s^{-1} Y_s \tag{3.8d}$$

is the generalized least squares estimator of γ . It is assumed that $V_s(\theta)$ is non-singular for all $\theta \in I$, which is an interval in ν -dim. space containing the true value of θ as an interior point. Now

$$V_s = \bigotimes_{a=1}^A \Phi_{as} \tag{3.9}$$

where

$$\Phi_{as} = \sigma_{uu} I_{n_a} + \sigma_{eea} I_{n_a} + X_{as} D X'_{as}$$

Therefore,

$$\hat{\gamma}_s = \left[\begin{array}{cccc} \sum_{a=1}^A X'_{as} \Phi_{as}^{-1} X_{as} & X'_{1s} \Phi_{1s}^{-1} 1_{n_1} & X'_{2s} \Phi_{2s}^{-1} 1_{n_2} & \dots & X'_{As} \Phi_{As}^{-1} 1_{n_A} \\ 1'_{n_1} \Phi_{1s}^{-1} X_{1s} & 1'_{n_1} \Phi_{1s}^{-1} 1_{n_1} & 0 & \dots & 0 \\ 1'_{n_2} \Phi_{2s}^{-1} X_{2s} & 0 & 1'_{n_2} \Phi_{2s}^{-1} 1_{n_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1'_{n_A} \Phi_{As}^{-1} X_{As} & 0 & 0 & \dots & 1'_{n_A} \Phi_{As}^{-1} 1_{n_A} \end{array} \right]_{(p+A) \times (p+A)}^{-1}$$

$$\begin{bmatrix} \sum_{a=1}^A X'_{as} \Phi_{as}^{-1} Y_{as} \\ 1_{n_1} \Phi_{1s}^{-1} Y_{1s} \\ \dots \\ 1'_{n_A} \Phi_{As}^{-1} Y_{As} \end{bmatrix}_{(p+A) \times 1} \tag{3.10}$$

Here

$$GZ'_s = \otimes(DX_{as}) \tag{3.11}$$

Using (3.5)-(3.11), \hat{T}_a can be found out. Now, estimate of $\bar{\beta}$ is obtained from the first p components of $\hat{\gamma}_s$. Clearly, $\hat{\beta}, \hat{\alpha}$ depend on the data from all the small areas $a(= 1, \dots, A)$. Thus, the estimate of the small area total \hat{T}_a borrows strength from all the other areas, $b(\neq a) = 1, \dots, A$.

Now we consider estimation of parameters in θ . Let us denote $V(\phi_{ai}) = \sigma_{\phi\phi a} = \sigma_{uu} + \sigma_{eea}$. We have

$$\hat{\sigma}_{\phi\phi a} = \frac{1}{n_a - p - 1} \Phi'_{as} H_a \Phi_{as} = s_{aa} \text{ (say)} \tag{3.12a}$$

where

$$\Phi_{as} = X_{as} \eta_a + e_{as} + u_{as} \tag{3.12b}$$

$$H_a = I_{n_a} - W_{as} (W'_{as} W_{as})^{-1} W'_{as} \tag{3.12c}$$

$$W_{as} = [X_{as} 1_{n_a}] \tag{3.12d}$$

Since σ_{uu} is assumed to be known, $\hat{\sigma}_{eea} = s_{aa} - \sigma_{uu}$.

Let b_a be the ordinary least squares estimator of β_a based on the sampled units in the small area a . To estimate D we treat the quantities b_a as a random sample of size A .

$$b_a = (X'_{as} M_a X_{as})^{-1} X'_{as} M_a Y_{as}, \tag{3.13a}$$

where

$$M_a = \begin{bmatrix} 1 - \frac{1}{n_a} & -\frac{1}{n_a} & \dots & -\frac{1}{n_a} \\ -\frac{1}{n_a} & 1 - \frac{1}{n_a} & \dots & -\frac{1}{n_a} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n_a} & -\frac{1}{n_a} & \dots & 1 - \frac{1}{n_a} \end{bmatrix}_{n_a \times n_a} \tag{3.13b}$$

Let

$$s_b = \sum_{a=1}^A b'_a b_a - \frac{1}{A} \sum_{a=1}^A b_a \sum_{a=1}^A b'_a, \tag{3.14}$$

$s_b/(A - 1)$ is the sample variance-covariance matrix of b_a 's.

Writing

$$b_a = \beta_a + (X'_{as} M_a X_{as})^{-1} X'_{as} M_a (e_{as} + u_{as}) \tag{3.15}$$

and taking expectation of both sides of (3.15),

$$E(s_b) = (A - 1)D + \frac{A - 1}{A} \sum_{a=1}^A \sigma_{\phi\phi a} (X'_{as} M_a X_{as})^{-1} \tag{3.16}$$

Hence

$$\hat{D} = \frac{s_b}{A-1} - \frac{1}{A} \sum_{a=1}^A s_{aa} (X'_{as} M_a X_{as})^{-1} \tag{3.17}$$

The two-stage estimator or empirical best linear unbiased prediction (EBLUP) estimator of T_a is given by

$$\hat{T}_a(\hat{\theta}) = \sum_{i \in s_a} Y_{is} + \hat{\mu}_a(\hat{\theta})$$

Kackar and Harville (1984) showed that under some generally satisfiable conditions both the estimators $\hat{T}_a(\theta)$ and $\hat{T}_a(\hat{\theta})$ have the same expected value. In large samples, $\hat{\gamma}_s(\theta)$ has the same asymptotic properties as $\hat{\gamma}_s(\hat{\theta})$. Under general regularity conditions both the estimators $\hat{T}_a(\theta)$, $\hat{T}_a(\hat{\theta})$ are consistent.

4. ESTIMATION OF MSE ($\hat{T}_a(\hat{\theta})$)

It is known from Kackar and Harville (1984) that

$$\begin{aligned} MSE(\hat{T}_a(\hat{\theta})) &\approx MSE(\hat{T}_a(\theta)) + trace[A(\theta)B(\theta)] \\ &= \tilde{M}(\theta) + trace[A(\theta)B(\theta)] \text{ (say)} \\ &= M_{KH}(\theta) \text{ (say)} \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} A(\theta) &= Var[d(Y_s; \theta)] \\ B(\theta) &= E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] \\ d(Y_s; \theta) &= \left. \frac{\partial \hat{T}_a(Y_s; \theta)}{\partial \theta} \right]_{\theta = \hat{\theta}} \end{aligned}$$

Prasad and Rao (1990) approximated $trace[A(\theta)B(\theta)]$ by

$$\begin{aligned} trace[(\nabla c'_a) V_s (\nabla c'_a)' E\{(\theta - \hat{\theta})(\theta - \hat{\theta})'\}] \\ = g(\theta) \text{ (say)} \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} \nabla c'_a &= \frac{\partial c'_a}{\partial \theta} \\ c'_a &= m'_a G Z'_s V_s^{-1} \end{aligned}$$

Hence, an approximation of $MSE(\hat{T}_a(\hat{\theta}))$ is

$$\begin{aligned} \tilde{M}(\theta) + g(\theta) \\ = M_{PR}(\theta) \text{ (say)} \end{aligned} \tag{4.3}$$

It is customary to ignore the term $g(\theta)$ and use $\tilde{M}(\hat{\theta})$ as an estimator of $MSE(\hat{T}_a(\theta))$, but this approximation may lead to serious under-estimation. An approximately unbiased estimator of $MSE(\hat{T}_a(\hat{\theta}))$ is

$$\begin{aligned} & \tilde{M}(\hat{\theta}) + 2g(\hat{\theta}) \\ & = mse(\hat{T}_a(\hat{\theta})) \text{ (say)} \end{aligned} \quad (4.4)$$

For the present case,

$$\begin{aligned} \tilde{M}(\theta) &= l'_a (\tilde{X}'_s V_s^{-1} \tilde{X}_s)^{-1} l_a \\ &+ m'_a G Z'_s [I - V_s^{-1} \tilde{X}_s (\tilde{X}'_s V_s^{-1} \tilde{X}_s)^{-1} \tilde{X}'_s] V_s^{-1} Z_s G' m_a \end{aligned} \quad (4.5)$$

For the case of one regressor variable x ,

$$\begin{aligned} c'_a &= [0 \ 0 \ \dots 0 \ D(\sum_{k \in \bar{s}_a} x_{ak}) x'_{as} \phi_{as}^{-1} \ 0 \ 0 \dots 0]_{1 \times n} \\ D(\sum_{k \in \bar{s}} x_{ak}) x'_{as} \phi_{as}^{-1} &= F(\sum_{k \in \bar{s}} x_{ak}) x'_{as} [I - \frac{F x_{as} x'_{as}}{1 + F \sum_{k \in s_a} x_{ak}^2}] \end{aligned} \quad (4.6)$$

where

$$\sigma_{kka} = \sigma_{uu} + \sigma_{eea}, \quad F = \frac{D}{\sigma_{kka}}$$

For further details on results of estimation of mean square error of estimators of small area parameters, the reader may refer to the reader may refer to Prasad and Rao (1990), Lahiri and Rao (1995), Mukhopadhyay (1998), among many others.

ACKNOWLEDGEMENT

The author is thankful to the referee for his valuable suggestions which also led to the improvement of the paper.

REFERENCES

- (1) Dempster, A.P., Rubin, D.B. and Tsutakawa, R.A.K. (1981). Estimation in covariance component models, *Jour. Am. Stat. Assoc.* 76, 341-353.
- (2) Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics*, 31, 423-447.
- (3) Kacker, R.N. and Harville, D.A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models, *Comm. Stat. T & M.*, 10, 1249 - 1261.
- (4) Kacker, R.N. and Harville, D.A. (1984). Approximation for standard error of estimators for fixed and random effects in mixed models, *J. Amer. Stat. Assoc.*, 79, 85- 862.

- (5) Lahiri, P. and Rao, J.N.K.(1995). Robust estimation of mean squared error of small area estimators, *Jour. Am. Stat. Assoc.* 90, 758-766.
- (6) Mukhopadhyay, P. (1998). *Small Area Estimation in Survey Sampling*, Narosa Publishing House, New Delhi, India.
- (7) Prasada, N.G.N. and Rao, J.K.K.(1990). The estimation of mean squared errors of small area estimators. *Jour. Am. Stat. Assoc.*, 85, 163-171.
- (8) Royall, R.M.(1970).On finite population sampling theory under certain linear regression models, *Biometrika*, 57, 377-387.
- (9) Swamy, P.A.V.B.(1970). Efficient estimation in a random regression coefficient model,*Econometrika*, 38(2), 311-323.