HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small Area Estimation

Spring 2015

CASE STUDY 2
Pseudo EBLUP estimation

Risto Lehtonen, University of Helsinki

Model-based estimation under unequal probability sampling

- How to account for unequal probability sampling in model-based EBLUP estimation?
  - Stratified sampling with non-proportional allocation
  - PPS type sampling designs
- The role of design weights?
- The role of design variables in the model?

Options considered

- PPS-WOR sampling design
- Continuous study variable y
- Linear mixed model with random intercepts

- **Model-based EBLUP**
  - Inclusion of PPS size variable in the model

- **Pseudo model-based EBLUPW**
  - Incorporation of design weights in the estimation procedure of the model

Simulation experiments - 1

Population N = 1 million elements

D = 100 domains

Size of domain U_d is proportional to $\exp(q_d)$
where q_d is simulated from Uniform(0,2.9)

47 minor domains (-69 elements)
19 medium-sized domains (70-119)
34 major domains (120-)

Simulation experiments - 2

PPS size variable x_1: Uniform(1,11)

Variable x_2 (unrelated to the sampling design): Uniform(-5,5)

Random intercept u_{0d} and random slopes u_{1d} and u_{2d}: Multinormal distribution

$Var(u_{0d}) = 1,\ Var(u_{1d}) = Var(u_{2d}) = 0.125$

$Corr(u_{0d}, u_{1d}) = Corr(u_{0d}, u_{2d}) = -0.5,\ Corr(u_{1d}, u_{2d}) = 0$

Residual ε followed $N(0,100)$

Simulation experiments - 3

Values of the *y*-variable were simulated as

$$y_k = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + (\beta_2 + u_{2d})x_{2k} + \varepsilon_k$$
$$\beta_0 = \beta_1 = \beta_2 = 1$$

Correlations of the variables in the population

$$corr(y, x_1) = 0.441$$
$$corr(y, x_2) = 0.446$$

Simulation experiments - 4

Population N = 1,000,000
Sample n = 10,000

Monte Carlo experiments

K = 1000 independent PPS-WOR samples

Inclusion probabilities: $\pi_k = nx_{1k} / \sum_{k \in U} x_{1k}$

Weights $a_k = 1/\pi_k$ varied between 54.5 and 599.8

Models and estimators

EBLUP estimator of domain totals - basic form

$$\hat{t}_{dEBLUP} = \sum_{k \in U_d} \hat{y}_k, \quad d = 1,\ldots,D$$

Fitted models:

Special cases of linear mixed models with random intercepts: $y_k = \beta_0 + u_{0d} + \beta_1 x_k + \varepsilon_k$

Models fitted by REML or pseudo REML (REML-W)

Predicted values: $\hat{y}_k = \hat{\beta}_0 + \hat{u}_{0d} + \hat{\beta}_1 x_k,$

$k \in U_d, \ d = 1,\ldots,D$

Pseudo EBLUP

Linear mixed model $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$

Pseudo EBLUP (EBLUPW) estimators are derived by incorporating design weights a_k in ML-W and REML-W estimation procedures of model parameters by using HT estimators for certain matrix products (Domest and RDomest programs of Ari Veijanen)

Modification of matrix products of $\mathbf{X}$, $\mathbf{y}$, $\mathbf{Z}$ matrix (whose columns are domain indicators), and $\mathbf{e}$ (the vector of residuals):
Matrix product $\mathbf{A}'\mathbf{B}$ ($\mathbf{A},\mathbf{B} = \mathbf{X},\mathbf{Z},\mathbf{y},\mathbf{e}$) was replaced by $\mathbf{A}'\mathbf{W}\mathbf{B}$, where $\mathbf{W}$ is the diagonal matrix of design weights a_k

Quality measures

Absolute relative bias (ARB)

$$\mathrm{ARB}(\hat{t}_d) = \left| \frac{1}{K} \sum_{v=1}^{K} \hat{t}_d(s_v) - t_d \right| / t_d$$

Relative root mean squared error (RRMSE)

$$\mathrm{RRMSE}(\hat{t}_d) = \sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{t}_d(s_v) - t_d)^2} / t_d$$

where K is the number of simulated samples

Table 1. Average ARB (%) and average RRMSE (%) of EBLUP estimators.

Model and estimator	**Average ARB** (%)			**Average RRMSE** (%)		
	Domain size class			Domain size class		
	Minor (20-69)	Medium (70-119)	Major (120+)	Minor (20-69)	Medium (70-119)	Major (120+)
Model 1 $y_k = \beta_0 + u_d + \varepsilon_k$						
EBLUP	19.7	19.5	20.3	19.9	19.8	20.6
EBLUPW	3.7	3.1	2.1	6.8	6.8	6.1
Model 2 $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
EBLUP	4.0	3.6	2.3	5.4	5.2	4.5
EBLUPW	3.6	3.0	1.9	6.3	6.1	5.5
Model 3 $y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
EBLUP	19.6	19.6	20.2	19.9	19.9	20.5
EBLUPW	3.4	2.9	1.9	6.5	6.4	5.7
NOTE: Variable x_1 is the PPS size variable						

Lessons learned

- Bias can be large for a misspecified model
- Unequal probability sampling can be successfully accounted for with two options
  - Inclusion of the size variable into the model for model-based EBLUP
  - Use of pseudo EBLUP by incorporating design weights in the estimation procedure of the model
- Squared bias component can still dominate MSE
  - Can be difficult to obtain proper confidence intervals