Small area estimation to quantify discontinuities in sample surveys

Jan A. van den Brakel [1 2] Bart Buelens [1]
Harm-Jan Boonstra [1]

First Asian ISI Satellite Meeting on Small Area Estimation,
Bangkok, Thailand,
1-4 September 2013.

[1] Statistics Netherlands, Department of Statistical Methods
[2] Maastricht University, Department of Quantitative Economics

Outline

Introduction

- Survey $\rightarrow$ measurement error: $y_{k,i} = u_{k,i} + b_i + e_{k,i}$
- Survey redesign $\rightarrow$ affects measurement error: b_i
- Discontinuities: $\Delta_i = y_i^{(a)} - y_i^{(r)}$
- Quantification through a parallel run:
  - Regular survey full sample size: direct estimators $\hat{y}_i^{(r)}$
  - Alternative sample reduced sample size: small area estimation for $\tilde{y}_i^{(a)}$
  - Additional auxiliary information: direct estimates regular survey $\hat{y}_i^{(r)}$

Introduction

This paper:

- Direct estimates regular survey $\hat{y}_i^{(r)}$ as additional information in models for small area estimators
- Variance estimation discontinuities:
  $var(\hat{\Delta}_i) = var(\hat{y}_i^{(r)}) + var(\tilde{y}_i^{(a)}) - 2cov(\hat{y}_i^{(r)}, \tilde{y}_i^{(a)})$

Introduction

Redesign Crime Victimization Survey (CVS) in 2008:

- Regular (new) survey design (ISM):
  - Stratified simple random sampling, with 25 police regions as strata
  - Sample size: 19000 responses (about 750 per domain)
  - GREG estimator domains: $\hat{y}_i^{(r)}$
- Alternative (old) survey (NSM):
  - Stratified simple random sampling, with 25 police regions as strata
  - Sample size: 6000 responses (proportional allocation)
  - SAE domains: $\tilde{y}_i^{(a)}$

Small area estimators

Auxiliary information

- Municipal Basic Administration (gender, age, household size, nationality, urbanization, municipality, province, etc.)
- Police Register of Reported Offences
- Direct estimates target variable and related variables from the regular survey
- Direct estimates preceding editions of the survey

Area level model (Fay and Herriot, 1979):

$$\begin{aligned} \hat{y}_i^{(a)} &= y_i^{(a)} + e_i = z_i^t \beta + v_i + e_i, \\ & v_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2), \quad e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_i) \end{aligned}$$

Small area estimators

1. EBLUP for auxiliary variables with error (Ybarra and Lohr, 2008)

$$\begin{aligned}\tilde{y}_i^{(a)} &= \hat{\gamma}_i \hat{y}_i^{(a)} + (1-\hat{\gamma}_i)\hat{z}_i^t\hat{\beta}, \\ \hat{\gamma}_i &= \frac{\hat{\sigma}_v^2 + \hat{\beta}^t \widehat{cov}(\hat{z}_i)\hat{\beta}}{\hat{\sigma}_v^2 + \hat{\beta}^t \widehat{cov}(\hat{z}_i)\hat{\beta} + \psi_i},\end{aligned}$$

2. Standard EBLUP (Rao, 2003)

$$\begin{aligned}\tilde{y}_i^{(a)} &= \hat{\gamma}_i \hat{y}_i^{(a)} + (1-\hat{\gamma}_i)z_i^t\hat{\beta}, \\ \hat{\gamma}_i &= \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i},\end{aligned}$$

3. Hierarchical Bayesian approach (Rao, 2003, Section 10.3). Posterior mean and variance for the area level model with a flat prior on β and σ_v^2.

Model selection

Procedure:

- Step forward variable selection
- Criterion: conditional AIC
- Penalty: trace of the "hat" matrix $\hat{y} = Hy$

Percentage improvement in coefficient of variation of the HB estimates compared to the direct estimates for optimal models based on different sets of covariates.

variable	admin	+ Police register	+ ISM
`offtot`	47%	49%	56%
`unsafe`	24%	29%	37%
`nuisance`	29%	35%	51%
`satispol`	50%	50%	55%
`propvict`	49%	49%	51%

Model selection

Optimal models

variable	cAIC-based model
offtot	REG_victim
unsafe	REG_nuisance, ADM_benefit, PR_propcrim, PR_drugs
nuisance	REG_nuisance, ADM_old
satispol	REG_funcpol
propvict	PR_propcrim, ADM_old

All models also include an intercept (not shown).

- REG_*: direct estimate regular survey
- PR_*: Police Register of Reported Offences
- ADM_*: Municipal Basic Administration

Analyzing discontinuities

Discontinuity: $\hat{\Delta}_i = \hat{y}_i^{(r)} - \tilde{y}_i^{(a)}$

Variance $var(\hat{\Delta}_i) = var(\hat{y}_i^{(r)}) + MSE(\tilde{y}_i^{(a)}) - 2cov(\hat{y}_i^{(r)}, \tilde{y}_i^{(a)})$.

Problem:

- $\tilde{y}_i^{(a)} = \hat{\gamma}_i \hat{y}_i^{(a)} + (1 - \hat{\gamma}_i)\hat{z}_i^t \hat{\beta}$
- $\hat{z}_i$ and $\hat{\beta}$ contain survey estimates from the regular survey (same target variable or related variables):
  - Design correlation between $\hat{y}_i^{(r)}$ and $\tilde{y}_i^{(a)}$
  - Nonlinear term: $\hat{z}_i^t \hat{\beta}$

Analyzing discontinuities

Covariance $cov(\hat{y}_i^{(r)}, \tilde{y}_i^{(a)})$:

- First order Taylor approximation for $\hat{z}_i^t \hat{\beta}$ around z_i and $y_i^{(a)}$.
- Approximation for $cov(\hat{y}_i^{(r)}, \tilde{y}_i^{(a)})$:

$$(1-\hat{\gamma}_i)[(1-\hat{\gamma}_i \hat{z}_i^t \hat{T}^{-1} \hat{z}_i)\hat{\beta}^t + \hat{\gamma}_i(\hat{\theta}_i - \hat{\beta}^t \hat{z}_i)\hat{z}_i^t \hat{T}^{-1}]\widehat{cov}(\hat{y}_i^{(r)}, \hat{z}_i),$$

with:

- $\hat{\beta} = \hat{T}^{-1}\hat{t}, \quad \hat{T} = \sum_{i=1}^m \hat{\gamma}_i \hat{z}_i \hat{z}_i^t, \quad \hat{t} = \sum_{i=1}^m \hat{\gamma}_i \hat{z}_i \hat{\theta}_i$
- $\widehat{cov}(\hat{y}_i^{(r)}, \hat{z}_i)$: vector with design covariances between $\hat{y}_i^{(r)}$ and $\hat{z}_i$

Analyzing discontinuities

$var(\hat{\Delta}_i) = var(\hat{y}_i^{(r)}) + MSE(\tilde{y}_i^{(a)}) - 2cov(\hat{y}_i^{(r)}, \tilde{y}_i^{(a)})$.

$MSE(\tilde{y}_i^{(a)})$:

1. Posterior variance of the HB estimator
2. Design-based approximation:
   - Taylor approximation for $\tilde{y}_i^{(a)}$ around z_i and $y_i^{(a)}$
   - Approximation for $MSE(\tilde{y}_i^{(a)})$:

$$\hat{\gamma}_i^2 \widehat{var}(\hat{y}_i^{(a)}) + (1-\hat{\gamma}_i)^2 \left[\sum_{j=1}^{m} \hat{B}_{i,j} \widehat{cov}(\hat{z}_j) \hat{B}_{i,j}^t + \sum_{j=1}^{m} \hat{C}_{i,j}^2 \widehat{var}(\hat{y}_j^{(a)}) \right] + 2\hat{\gamma}_i(1-\hat{\gamma}_i)\hat{C}_{i,i} \widehat{var}(\hat{y}_i^{(a)}),$$

with

$$\begin{aligned} \hat{B}_{i,j} &= (\delta_{i,j} - \hat{\gamma}_j \hat{z}_i^t \hat{T}^{-1} \hat{z}_j)\hat{\beta}^t + \hat{\gamma}_j(\hat{y}_j^{(a)} - \hat{z}_j^t \hat{\beta})\hat{z}_i^t \hat{T}^{-1}, \\ \hat{C}_{i,j} &= \hat{z}_i^t \hat{T}^{-1} \hat{z}_j \hat{\gamma}_j. \end{aligned}$$

Analyzing discontinuities

Three estimators for
$var(\hat{\Delta}_i) = var(\hat{y}_i^{(r)}) + MSE(\tilde{y}_i^{(a)}) - 2cov(\hat{y}_i^{(r)}, \tilde{y}_i^{(a)})$:

1. Posterior variance of the HB estimator for $MSE(\tilde{y}_i^{(a)})$
2. Design-based approximation for $MSE(\tilde{y}_i^{(a)})$
3. Bootstrap approximation
   - Draw repeatedly bootstrap samples from the original sample (regular and alternative sample)
   - Calculate $\hat{\Delta}_{i,b} = \hat{y}_{i,b}^{(r)} - \tilde{y}_{i,b}^{(a)}$, $b = 1, \ldots, B$
   - $\widehat{var}(\hat{\Delta}_i) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\Delta}_{i,b} - \bar{\hat{\Delta}}_i)^2$

Results discontinuities

Comparison HB point and SE estimates with bootstrap results averaged over districts.

variable	Analytic			Bootstrap	
	HB est.	SE(1)	SE(2)	HB est.	SE
offtot	33.21	2.43	2.90	33.29	3.13
unsafe	19.83	1.76	1.64	19.84	1.92
nuisance	1.29	0.06	0.08	1.28	0.08
satispol	55.09	3.00	2.54	55.29	3.58
propvict	9.85	1.09	0.84	9.88	1.12

Results discontinuities

Analysis results discontinuities averaged over districts.

variable	Analytic			Bootstrap		GREG	
	Disc.	SE(1)	SE(2)	Disc.	SE	Disc.	SE
offtot	9.08	3.54	3.92	9.02	4.92	9.01	7.69
unsafe	4.55	2.54	2.46	4.54	2.69	4.52	3.57
nuisance	0.33	0.05*	0.07	0.33	0.11	0.33	0.17
satispol	5.52	4.98	4.72	5.33	5.43	5.04	8.21
propvict	2.70	1.95	1.84	2.70	1.97	2.78	2.77

*: For nuisance 2 districts with negative variance estimates for the estimated discontinuity are truncated at zero.

Conclusions

- Additional information regular survey useful for SAE models (substantial reduction standard errors)
- Variance approximations:
  - Design-based covariance approximation
  - Design-based approximation MSE of SAE predictions
  - Avoids negative variance estimates
- Alternative (further research): bivariate area level model