HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small Area Estimation

Spring 2015

Topic 4: GREG and calibration estimators
PART II: Indirect GREG estimators

Risto Lehtonen, University of Helsinki

Topic 4 Part I

- **GREG and calibration estimators**
  PART II: Indirect GREG estimators
  - Indirect linear GREG estimator for domain totals
  - Variance estimators
  - Example

Indirect estimators

- **Recall definition**
- Indirect estimator uses y-values not only from the domain of interest itself but also outside the domain or from earlier time points
- “Borrowing strength” from other domains or in a temporal dimension
- Borrowing strength can be exercised both in design-based SAE and model-based SAE

Linear GREG estimator

- GREG estimator assisted by a linear fixed-effects model (Särndal, Swensson and Wretman, 1992)

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

Assisting models, examples:

Common model for all domains

$$y_k = \beta_0 + \beta_1 x_k + \ldots + \beta_J x_{Jk} + \varepsilon_k$$

Domain-specific fixed intercepts and common slopes

$$y_k = \beta_{01} I_{1k} + \beta_{02} I_{2k} + \ldots + \beta_{0D} I_{Dk} + \beta_1 x_k + \ldots + \beta_J x_{Jk} + \varepsilon_k$$

where $I_{dk} = I\{k \in U_d\}$ (domain membership indicator)

Indirect GREG estimator for domain total

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k), \text{ where}$$

$\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ are predictions from the model, $k \in U$

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{Jk})'$, known for all $k \in U_d$

$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_J)'$ is the vector of estimated regression coefficients common for all domains

We fit the model by WLS: $\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in s} a_k \mathbf{x}_k y_k \right)$

This GREG is an indirect estimator, since all *y*-values in the sample contribute

Indirect GREG estimator – another form

Since assisting model is linear, GREG estimation does not require unit-level information on $\mathbf{x}_k$

It is enough to have access to the vector $\mathbf{t}_{dx} = \sum_{k \in U_d} \mathbf{x}_k$ of domain totals of auxiliary x-variables in the population and the corresponding HT estimates $\hat{\mathbf{t}}_{dx} = \sum_{k \in s_d} \mathbf{x}_k$ in the sample

Standard textbook form:

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\boldsymbol{\beta}}, \text{ where } \hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$$

Details

$\mathbf{t}_{dx} = (t_{dx_0},\ldots,t_{dx_J})'$ known domain totals of auxiliary x-variables in population, $d = 1,\ldots,D$

$$t_{dx_j} = \sum_{k \in U_d} x_{jk}\ , \quad j = 0,\ldots,J$$

$\hat{\mathbf{t}}_{dx} = (\hat{t}_{dx_0},\ldots,\hat{t}_{dx_J})'$ HT estimators of domain totals

$$\hat{t}_{dx_j} = \sum_{k \in s_d} a_k x_{jk}\ , \quad j = 0,\ldots,J$$

NOTE: $x_{0k} = 1$ for all $k \in U$

Practical variance estimator for indirect GREG for unplanned domains

Approximate variance estimator of GREG by using *extended residuals*:

$$\hat{V}_U\left(\hat{t}_{dGREG}\right) = \frac{n}{n-1}\sum_{k\in s}\left(a_k e_{dk} - \hat{t}_{dHTe} / n\right)^2, \qquad (15)$$

where

n is the total sample size and $a_k = 1/\pi_k$ (design weights)

$e_{dk} = I\{k \in U_d\}e_k$ are extended residuals, where $e_k = y_k - \hat{y}_k$

NOTE: $e_{dk} = e_k$ if $k \in s_d$ and $e_{dk} = 0$ if $k \notin s_d$

$\hat{t}_{dHTe} = \sum_{k\in s_d} a_k e_k$ is HT estimator of residual total in domain d

NOTE: Similarity of (15) with HT variance estimator (5) for unplanned domains (both (5) and (15) are used in RDomest software)

Indirect GREG estimator expressed as calibration estimator

$$\hat{t}_{dGREG} = \sum_{k \in s} a_k g_{dk} y_k$$

where

$g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k$ are extended g-weights

$I_{dk} = 1$ if $k \in U_d$, 0 otherwise (domain membership indicator)

$\hat{\mathbf{M}} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i'$ NOTE: Extends over whole sample

NOTE: Calibration property holds for the auxiliary x-variables

Variance estimator

Variance estimator for unplanned domains

$$\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \qquad (16)$$

where

$e_k = y_k - \hat{y}_k$ are residuals

$$g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k$$

$$\hat{\mathbf{M}} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i'$$

NOTE: Extended g-weights are used

Indirect design-based model-assisted GREG estimators

- **SUMMARY page for:**
  - Direct GREG estimator for planned domains under SRS
  - Indirect GREG for unplanned domains under SRS
  - Assisting model: linear fixed-effects model of common model type:

  - Example: Comparison of results for HT and GREG under more complex unequal probability sampling

$$y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \ldots + \beta_J x_{Jk} + \varepsilon_k$$

  - See separate sheet for Topic 4, Part 2, available at course website

EXAMPLE: HT and indirect GREG for unplanned domains

- Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.

- Section 4.2. Computational example with direct and indirect estimation under an unplanned domain structure

Sampling design

- Population: N = 431,000 households
- Household sampling: πPS (PPS-WOR)
- Size variable in PPS-WOR: Number of household members
- Domains: D = 12 NUTS4 regions (domains)
    - Domain sample sizes are assumed random
- Sample size: n = 1000 households

Variables

- **Study variable** ***y***
  - Disposable household income
- **Auxiliary x-variable (known for all HHs)**
  - EMP: the number of months in total the household members were employed during last year
  - Variable is derived from administrative registers
  - Domain sizes in population and domain totals of EMP are assumed known

- NOTE: Also here we have access to unit-level population values of our study variable y and auxiliary x-variable
- This gives option to compare results with true values

Estimators of domain totals

- HT estimator with variance estimator (5)
- Linear GREG estimator with variance estimator (15)

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$$

$$\hat{V}_U\left(\hat{t}_{dHT}\right) = \frac{n}{n-1} \sum_{k \in s} \left(a_k y_{dk} - \hat{t}_{dHT} / n\right)^2$$

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\boldsymbol{\beta}}$$

$$\hat{V}_U\left(\hat{t}_{dGREG}\right) = \frac{n}{n-1} \sum_{k \in s} \left(a_k e_{dk} - \hat{t}_{dHTe} / n\right)^2$$

Assisting model in GREG

GREG estimator is assisted by a linear fixed-effects model

$$y_k = \beta_0 + \beta_1 \text{EMP}_k + \varepsilon_k$$

fitted to the whole sample

NOTE: Common intercept and slope for all domains - therefore, this GREG is indirect

Quality measures of estimators

ARE Absolute relative error of an estimator in domain *d*

$\text{ARE}(\hat{t}_d) = |\hat{t}_d - t_d| / t_d$, $d = 1,\ldots,D$

MARE in domain group:
The mean of absolute relative errors over domains in the group

MCV The mean coefficient of variation of the estimate over domain group

The coefficient of variation is calculated as $s.e(\hat{t}_d) / \hat{t}_d$
where s.e refers to the estimated standard error of an estimator

Table 4. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of HT and indirect GREG estimators of totals for minor, medium-sized and major domains for **unplanned domains**.

	HT		**GREG**	
	Auxiliary information			
	1 None		2 Domain sizes and domain totals of EMP	
Domain sample size class	MARE %	**MCV %**	MARE %	**MCV %**
Minor $8 \le n_d \le 33$	11.5	**28.3**	7.6	**9.0**
Medium $34 \le n_d \le 45$	7.6	**20.3**	3.8	**8.1**
Major $46 \le n_d \le 277$	12.5	**9.6**	4.1	**5.0**

Lessons learned from the two examples

- **Planned domains, direct estimators**
  - GREG better than HT in terms of accuracy
- **Unplanned domains, indirect estimators**
  - GREG again better than HT in terms of accuracy
- **Use of auxiliary data makes sense!**
- **Planned vs. unplanned case**
  - Accuracy tends to be better in planned domains case
- **Stratification for important domains of interest makes sense!**
  - An issue of the survey planning stage!