Helsingin yliopisto
Matematiikan ja tilastotieteen laitos

Pienalue-estimointi (78189)

Kevät 2009
Risto Lehtonen
30.3.2009

OSA 3

GREG-estimaattori

Yleinen tilanne (*unequal probability sampling*)

Komposiittiestimaattorit (*Composite estimators*)

Estimointi SAS-proseduureilla SURVEYREG ja SURVEYMEANS

Estimointi SAS-makrolla EBLUPGREG

Esimerkkejä

HUOM: Tässä osassa merkinnät kuten Lehtonen & Veijanen (2009)

Taulukko 1. Estimaattorin tyypin ja osajoukon tyypin tavallisimmat yhdistelmät käytännön kannalta

		Estimaattorin tyyppi	
		Suora *Direct*	**Epäsuora** *Indirect*
Osajoukon tyyppi	**Suunniteltu** *Planned* Ositettu otanta	HT Kiinteiden tekijöiden D-malli D-tyypin GREG	 Mahdollinen mutta ei kovin yleinen käytännössä
	Ei-suunniteltu *Unplanned*	Mahdollinen mutta ei kovin yleinen käytännössä	Kiinteiden tekijöiden P-malli P-tyypin GREG Sekamallin tyyppinen D-malli D-tyypin GREG

GREG-ESTIMAATTORI: YLEINEN TILANNE
(*Unequal probability sampling*)

(1) Suunniteltujen osajoukkojen tilanne
(*Planned domains*)

GREG-estimaattori:

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k \qquad (30)$$

Avustava regressiomalli: D-malli:

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta}_d + \varepsilon_k \qquad (31)$$

$k \in U_d$, $\mathbf{x}_k = (1, x_{1k}, ..., x_{Jk})'$ ja $Var(\varepsilon_k) = \sigma_k^2$

Oletetaan vakiovarianssi $\sigma_k^2 = \sigma^2$

Osajoukon U_d parametrivektorin $\mathbf{B}_d$ WLS-estimaattori:

$$\hat{\mathbf{B}}_d = \left(\sum_{k \in s_d} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s_d} a_k \mathbf{x}_k y_k \qquad (32)$$

Sovitteet: $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}_d$

Jäännökset: $e_k = y_k - \hat{y}_k$

HUOM: $a_k = 1/\pi_k$

GREG-estimaattorin vaihtoehtoiset muodot

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{B}}_d \qquad (33)$$

missä

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

$$\mathbf{t}_{dx} = \sum_{k \in U_d} \mathbf{x}_k = \left(N_d, \sum_{k \in U_d} \mathbf{x}_{1k}, \ldots, \sum_{k \in U_d} \mathbf{x}_{Jk}\right)'$$

$$\hat{\mathbf{t}}_{dx} = \sum_{k \in s_d} a_k \mathbf{x}_k$$

Kalibrointiestimaattori

$$\hat{t}_{dGREG} = \sum_{k \in s_d} a_k g_{dk} y_k \qquad (34)$$

missä

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

$$\mathbf{t}_{dx} = (t_{dx_1}, \ldots, t_{dx_1})' \text{ ja } \hat{\mathbf{t}}_{dx} = (\hat{t}_{dx_1}, \ldots, \hat{t}_{dx_1})'$$

$$t_{dx_j} = \sum_{k \in U_d} x_{jk} \text{ ja } \hat{t}_{dx_j} = \sum_{k \in s_d} a_k x_{jk}$$

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

$$I_{dk} = I\{k \in U_d\} \quad \text{(domain-indikaattorit)}$$

Varinssiestimaattorit

$$\hat{V}_1\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) e_k e_l. \qquad (35)$$

$$\hat{V}_2\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \qquad (36)$$

missä a_{kl} on 2. kertaluvun sisältymistodennäköisyys

(2) Ei-suunniteltujen osajoukkojen tilanne (*Unplanned domains*)

GREG-estimaattori:

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k \qquad (37)$$

Avustava regressiomalli: P-malli:

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k \qquad (38)$$

$k \in U$ ja $Var(\varepsilon_k) = \sigma_k^2$

Oletetaan vakiovarianssi $\sigma_k^2 = \sigma^2$

Parametrivektorin $\mathbf{B}$ WLS-estimaattori:

$$\hat{\mathbf{B}} = \left(\sum_{k \in s} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} a_k \mathbf{x}_k y_k \qquad (39)$$

Sovitteet: $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}$

Jäännökset: $e_k = y_k - \hat{y}_k$

HUOM: Vertaa kaavaan (32)

GREG-estimaattorin vaihtoehtoiset muodot

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{B}} \qquad (40)$$

missä

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

$$\mathbf{t}_{dx} = \sum_{k \in U_d} \mathbf{x}_k = \left(N_d, \sum_{k \in U_d} \mathbf{x}_{1k}, \ldots, \sum_{k \in U_d} \mathbf{x}_{Jk} \right)'$$

$$\hat{\mathbf{t}}_{dx} = \sum_{k \in s_d} a_k \mathbf{x}_k$$

Kalibrointiestimaattori

$$\hat{t}_{dGREG} = \sum_{k \in s} a_k g_{dk} y_k \qquad (41)$$

missä

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

$$\mathbf{t}_{dx} = (t_{dx_1}, \ldots, t_{dx_1})' \quad \text{ja} \quad \hat{\mathbf{t}}_{dx} = (\hat{t}_{dx_1}, \ldots, \hat{t}_{dx_1})'$$

$$t_{dx_j} = \sum_{k \in U_d} x_{jk} \quad \text{ja} \quad \hat{t}_{dx_j} = \sum_{k \in s_d} a_k x_{jk}$$

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

$$I_{dk} = I\{k \in U_d\} \quad \text{(domain-indikaattorit)}$$

HUOM. Vertaa estimaattoriin (34)

Varinssiestimaattorit

$$\hat{v}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s}(a_k a_l - a_{kl})g_{dk}e_k g_{dl}e_l \qquad (42)$$

missä a_{kl} on 2. kertaluvun sisältymistn

HUOM: Varianssiestimaattorissa (42) kaksoissumma on yli koko otoksen *s*

Vaihtoehtoja:

(1) Summataan yli domain-otoksen s_d

Hidiroglou, M. A. and Z. Patak (2004). Domain estimation using linear regression. *Survey Methodology* 30, 67-78.

(2) Käytetään domain-kohtaisia tulosmuuttujia $y_{dk} = I\{k \in U_d\}y_k$ mallin sovittamisessa

Estevao, V. M., M. A. Hidiroglou, and C.-E. Särndal (1995). Methodological principles for a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics* 11, 181-204.

(3) Sovitetaan malli alkuperäisille arvoille y_k ja korvataan varianssiestimaattorissa jäännökset e_k domain-kohtaisilla jäännöksillä $e_{dk} = I\{k \in U_d\}y_k - \hat{y}_k$

Lehtonen, R. and E. Pahkinen (2004). *Practical methods for design and analysis of complex surveys.* Second Edition. John Wiley & Sons, Chichester, p. 39.

Särndal, C.-E. (2001). Design-based methodologies for domain estimation. In: R. Lehtonen and K. Djerf, eds., *Proceedings of the Symposium on Advances in Domain Estimation*. Statistics Finland, Reviews 2001/5, p. 202.

(4) Kuten (3) mutta $e_{dk} = 0$ kun $k \in s$ ja $k \notin U_d$

Sas-yhdistelmä SURVEYREG ja SURVEYMEANS/DOMAIN-lause

KAKSI ESIMERKKIÄ

Lehtonen and Veijanen (2009)

Estimointitilanne: Käytettävissä olevien tulojen kokonaismäärän alueittainen estimointi Länsi-Suomen $D = 12$ seutukunnassa

Estimoitavat parametrit:

Käytettävissä olevien tulojen kokonaismäärä

$$t_d = \sum_{k \in U_d} y_k$$

Osajoukot (domains) U_d, $d = 1,...,12$

Populaatiodata: $N = 431{,}000$ kotitaloutta

Lisäinformaatio tilastorekistereistä:

EDUC: Kotitalouden jäsenten lkm joilla on korkea-asteen koulutus

EMP: Kotitalouden jäsenten yhteenlaskettu työllisyyskuukausien määrä edellisenä vuonna

Lisäksi (tässä pedagogisessa tilanteessa) tiedetään tulosmuuttujan arvot y_k kaikilta perusjoukon alkioilta

Esimerkki 1

Suora estimointi (*direct estimation*)
Suunnitellut osajoukot (*planned domains*)

Kotitalousotos: Ositettu πPS (WOR- tyyppinen PPS)
Kokomuuttuja: Kotitalouden jäsenten lukumäärä
Ositteet: Seutukunnat (domains)

HUOM: Ositteiden otoskoot on kiinnitetty otanta-asetelmassa (suhteellinen kiintiöinti)

Estimaattorit: HT, kaava (21)

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

$$\hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d - 1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2$$

Suora GREG, kaavat (30) ja (36)

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

$$\hat{V}_2\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$$

GREG-estimaattorin avustavat D-mallit:

$$Y_k = \beta_{0d} + \beta_{1d} \mathrm{EMP}_k + \varepsilon_k \quad \text{(sarake 2)}$$

$$Y_k = \beta_{0d} + \beta_{1d} \mathrm{EMP}_k + \beta_{2d} \mathrm{EDUC}_k + \varepsilon_k \quad \text{(sarake 3)}$$

Taulukko 2. Suorien HT- ja GREG-estimaattoreiden keskimääräinen absoluuttinen suhteellinen virhe (Mean absolute relative error MARE %) ja keskimääräinen variaatiokerroin (mean coefficient of variation MCV %) pienissä, keskisuurissa ja suurissa domaineissa:
Suunniteltujen domainien tilanne

Auxiliary information	HT		GREG			
	1 None		2 Domain sizes and domain totals of EMP		3 Domain sizes and domain totals of EMP and EDUC	
Domain sample size class	MARE %	MCV %	MARE %	MCV %	MARE %	MCV %
Minor $8 \le n_d \le 33$	11.5	11.9	5.8	7.7	6.4	6.8
Medium $34 \le n_d \le 45$	7.6	9.0	3.7	8.0	3.6	8.1
Major $46 \le n_d \le 277$	12.5	5.2	4.3	4.7	5.2	3.7

HUOM:

$$\mathrm{ARE}(\hat{t}_d) = \left|\hat{t}_d - t_d\right| / t_d, \quad d = 1,...,12$$

$$\mathrm{CV}(\hat{t}_d) = \mathrm{s.e}(\hat{t}_d) / \hat{t}_d, \quad d = 1,...,12$$

MARE ja MCV ovat vastaavia keskiarvoja kussakin domainien kokoluokassa

Esimerkki 2

HT: Suora estimointi
GREG: Epäsuora estimointi
Ei-suunnitellut osajoukot (*unplanned domains*)

Kotitalousotos: πPS (WOR- tyyppinen PPS)
Kokomuuttuja: Kotitalouden jäsenten lukumäärä

HUOM: Domainien otoskokoja ei ole kiinnitetty otanta-asetelmassa

Estimaattorit: HT, kaava (21)

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

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

GREG, kaavat (30) ja (42)

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

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

GREG-estimaattorin avustava P-malli:

$$Y_k = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k \quad \text{(sarake 2)}$$

Taulukko 3. Suoran HT-estimaattorin ja epäsuoran GREG-estimaattoreiden keskimääräinen absoluuttinen suhteellinen virhe (Mean absolute relative error MARE %) ja keskimääräinen variaatiokerroin (mean coefficient of variation MCV %) pienissä, keskisuurissa ja suurissa domaineissa:
Ei-suunniteltujen domainien tilanne

Auxiliary information	HT		GREG	
	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

HUOM:

$$\mathrm{ARE}(\hat{t}_d) = \left|\hat{t}_d - t_d\right| / t_d, \quad d = 1,...,12$$

$$\mathrm{CV}(\hat{t}_d) = \mathrm{s.e}(\hat{t}_d) / \hat{t}_d, \quad d = 1,...,12$$

MARE ja MCV ovat vastaavia keskiarvoja kussakin domainien kokoluokassa

KOMPOSIITTITYYPPISET ESTIMAATTORIT
(*Composite estimators*)

Yhdistelmäestimaattori on muotoa

$$\hat{t}_{dCOMB} = \lambda_d \hat{t}_{dGREG} + (1-\lambda_d)\hat{t}_{dSYN} \qquad (43)$$

joka on muodostettu asetelmaperusteisen GREG-estimaattorin

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \qquad (44)$$

ja malliperusteisen synteettisen estimaattorin

$$\hat{t}_{dSYN} = \sum\nolimits_{k \in U_d} \hat{y}_k = \sum\nolimits_{k \in U_d} \mathbf{x}_k' \hat{\mathbf{B}} \qquad (45)$$

painotettuna summana

Domainkohtaiset painot λ_d $(0 \le \lambda_d \le 1)$ valitaan niin, että λ_d on suurille domaineille (”suuri” n_d) lähellä ykköstä ja lähestyy nollaa kun n_d on ”pieni”

Pienille domaineille $\hat{t}_{dCOMB}$ on lähellä SYN-estimaattoria $\hat{t}_{dSYN}$

Suurille domaineille $\hat{t}_{dCOMB}$ on lähellä GREG-estimaattoria $\hat{t}_{dGREG}$

Estimaattori (45) voidaan kirjoittaa muotoon

$$\hat{t}_{dCOMB} = \hat{t}_{dSYN} + \lambda_d \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \qquad (46)$$

HUOM:

GREG-estimaattori kun $\lambda_d = 1$
SYN-estimaattori kun $\lambda_d = 0$

Esimerkki 1.

$$\hat{t}_{dGREG(N)} = \sum_{k \in U_d} \hat{y}_k + \left(N_d / \hat{N}_d \right) \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \qquad (47)$$

missä $\hat{N}_d = \sum_{k \in s_d} a_k$

Esimerkki 2.

Dampened regression estimator (Särndal and Hidiroglou 1989)

$$\hat{t}_{dDRE} = \sum_{k \in U_d} \hat{y}_k + \left(\hat{N}_d / N_d \right)^{c-1} \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \qquad (48)$$

missä

$c = 0$ kun $\hat{N}_d \geq N_d$
$c = 2$ kun $\hat{N}_d < N_d$

GREG-ESTIMOINTI SAS-PROSEDUUREILLA SURVEYREG JA SURVEYMEANS

Ei-suunniteltujen domainien tilanne

Metodi:

(1) Kiinteiden vaikutusten regressiomalli (P-malli) sovitetaan proseduurilla SURVEYREG

(2) Lasketaan sovitteet $\hat{y}_k$ ja jäännökset $e_k = y_k - \hat{y}_k$

(3) Lasketaan GREG-estimaatit $\hat{t}_{dGREG}$

(4) Estimoidaan GREG-estimaattorin varianssi proseduurilla SURVEYMEANS käyttämällä y-muuttujana jäännösvektoria ja varianssiestimoinnissa DOMAIN-lausetta samaan tapaan kuin HT-estimoinnissa tulosmuuttujalle *y*

Varianssiestimaattori on muotoa (SRSWOR-tilanne)

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2(1-\frac{n}{N})(\frac{1}{n})\sum_{k\in s}\frac{(e_{dk}-\bar{e})^2}{n-1}$$

missä

$e_{dk} = e_k$ kun $k \in U_d$ ja $k \in s$

$e_{dk} = 0$ kun $k \notin U_d$ ja $k \in s$

$\bar{e} = \sum_{k\in s} e_k / n$ jäännösten keskiarvo koko otoksessa

HUOM: Vertaa estimaattoriin (28)

ESIMERKKI

Vaihe (1)

```
proc surveyreg data=omaotos total=966;
model y=x;
weight samplingweight;
ods output
ParameterEstimates=beta(keep=estimate);
run;
```

Vaiheet (2) ja (4)

```
proc transpose data=beta
   out=beta2(drop=_name_
   rename=(col1=b0 col2=b1));
run;

data pj;
if _n_=1 then set beta2;
set pj;
yhat=b0+b1*x;
ehat=y-yhat;
run;

proc surveymeans data=pj nobs sum
total=966;
where ind=1;
weight SamplingWeight;
var ehat;
domain domain;
run;
```

GREG-ESTIMOINTI SAS-MAKROLLA EBLUPGREG

GREG-estimaattori

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \qquad (49)$$

Perustuu **kiinteiden tekijöiden lineaariseen malliin**

Malli on P-malli:

$$Y_k = \beta_0 + \beta_1 x_{1k} + ... + \beta_J x_{Jk} + \varepsilon_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k \qquad (50)$$

$k \in U$, missä $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$

Estimointi: WLS, mukana painot $a_k = 1/\pi_k$

Prediktiot: $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ kaikille $k \in U$

Varianssiestimaattori domainissa *d*:

Kaava (42)

ESIMERKKI

Ohjelmakutsu:

```
%eblupgreg
    (sample=omaotos,
    population=pj,
    y=y,
    xlist=x,
    regionIdentifier=domain,
    test=1,
    estimateMeans=0,
    weights=samplingweight,
    convergenceCrit=1e-8,
    maxiterations=200,
    initialSigma2=1,
    modules=modules.eurarea,
    parametersEstimatedBy='REML',
    eblup=0,
    greg=1,
    synthetic=0,
    stratified=0,
    output=greg
    );
```