

Pienalue-estimointi (78189)

Kevät 2009

Risto Lehtonen

6.4.2009

OSA 4

Laajennettu GREG-estimaattoreiden perhe

Avustavat mallit

Yleistetty lineaarinen malli

Lineaarinen sekamalli

Yleistetty lineaarinen sekamalli

GREG-estimaattoreiden teoreettisten ominaisuuksien
(harha, MSE) tutkiminen empiirisesti

Monte Carlo -simulointi

LAAJENNETTU GREG-ESTIMAATTOREIDEN PERHE

GREG-tyyppiset estimaattorit joissa avustava malli on yleistettyjen lineaaristen mallien (GLMM, *Generalized linear mixed models*) perheen jäsen

Avustavat mallit

(1) Yleistetty lineaarinen (kiinteiden tekijöiden) malli

$$E_m(Y_k) = f(\mathbf{x}_k; \boldsymbol{\beta}) \quad (51)$$

missä $f(\cdot; \boldsymbol{\beta})$ on annettu funktio (lineaarinen funktio, logistinen funktio), $\boldsymbol{\beta}$ on estimoitava parametrivektori ja E_m viittaa odotusarvoon mallin suhteen

Malli sovitetaan otosdatalle $\{(y_k, \mathbf{x}_k); k \in s\}$

Saadaan parametrin \mathbf{B} estimaatti $\hat{\mathbf{B}}$, missä \mathbf{B} on mallin parametrin $\boldsymbol{\beta}$ äärellisen perusjoukon vastine

Sovitteet $\hat{y}_k = f(\mathbf{x}_k; \hat{\mathbf{B}})$ lasketaan jokaiselle $k \in U$ käyttämällä vektoria $\hat{\mathbf{B}}$ ja apumuuttujavektoria \mathbf{x}_k

Logistisen mallin avustama GREG-estimaattori LGREG

Lehtonen, R. and A. Veijanen (1998). Logistic generalized regression estimators. *Survey Methodology* **24**, 51-55.

(2) Yleistetty lineaarinen sekamalli

$$E_m(Y_k | \mathbf{u}_d) = f(\mathbf{x}'_k (\boldsymbol{\beta} + \mathbf{u}_d)) \quad (52)$$

missä \mathbf{u}_d on domain-spesifien satunnaistermien vektori

Sovitteet $\hat{y}_k = f(\mathbf{x}'_k (\hat{\mathbf{B}} + \hat{\mathbf{u}}_d))$ lasketaan jokaiselle $k \in U$ käyttämällä estimaattivektoreita $\hat{\mathbf{B}}$, $\hat{\mathbf{u}}_d$ ja apumuuttujavektoria \mathbf{x}_k

(a) Lineaarinen sekamalli

$$\begin{aligned} E_m(Y_k | \mathbf{u}_d) &= \mathbf{x}'_k (\boldsymbol{\beta} + \mathbf{u}_d) \\ &= (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + \dots + (\beta_J + u_{Jd})x_{Jk} \end{aligned} \quad (53)$$

missä $\mathbf{u}_d = (u_{0d}, u_{1d}, \dots, u_{Jd})'$ on domain-spesifien satunnaistermien vektori

Käytännössä usein vain osa u -termeistä on mallissa:

$$E_m(Y_k | \mathbf{u}_d) = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + \beta_2 x_{2k} \quad (54)$$

Vastaava kiinteiden tekijöiden malli

$$E_m(Y_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} \quad (55)$$

Lehtonen, R. and A. Veijanen (1999). Domain estimation with logistic generalized regression and related estimators. *Proceedings, IASS Satellite Conference on Small Area Estimation*, Riga, August 1999. Riga: Latvian Council of Science, 121-128.

(b) Logistinen sekamalli

Binominen logistinen sekamalli on muotoa

$$E_m(y_k | \mathbf{u}_d) = P\{y_k = 1 | \mathbf{u}_d\} = \frac{\exp(\mathbf{x}'_k (\boldsymbol{\beta} + \mathbf{u}_d))}{1 + \exp(\mathbf{x}'_k (\boldsymbol{\beta} + \mathbf{u}_d))} \quad (56)$$

missä tulosmuuttuja y on **binäärinen**

Esim: 0: Työllinen

1: Työtön

Tulosmuuttuja voi olla myös **moniluokkainen**

Multinomiaalinen logistinen sekamalli

Esim: 1: Työllinen

2: Työtön

3: Ei kuulu työvoimaan

Lehtonen, R., C.-E. Särndal, and A. Veijanen (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33-44.

Lehtonen, R., C.-E. Särndal, and A. Veijanen (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* **7**, 649-673.

HUOM: Mallia (56) vastaava kiinteiden tekijöiden logitmalli on

$$E_m(y_k) = P\{y_k = 1\} = \frac{\exp(\mathbf{x}'_k \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_k \boldsymbol{\beta})} \quad (57)$$

ESIMERKKI

Tutkitaan osajoukkototaalien GREG-estimaattoreiden teoreettisia ominaisuuksia empiirisesti simulointikokeiden avulla

Parametrit $t_d = \sum_{k \in U_d} y_k, d = 1, \dots, D$

Kiinnostuksen kohteena estimaattorin \hat{t}_d harha ja MSE

$$\text{Bias}(\hat{t}_d) = E(\hat{t}_d) - t_d$$

$$\text{MSE}(\hat{t}_d) = E(\hat{t}_d - t_d)^2$$

Tutkimusmenetelmä: Monte Carlo -kokeet

Otokset $s_v; v = 1, 2, \dots, K$

Kullekin osajoukolle U_d lasketaan otosten perusteella:

Absoluuttinen suhteellinen harha

Absolute relative bias ARB

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

Suhteellinen RMSE (Root MSE)

Relative root mean squared error RRMSE

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

Simuloinneissa poimitaan generoitavasta perusjoukosta $K = 1000$ riippumatonta otosta

Keinotekoisen perusjoukon generointi

Perusjoukon koko $N = 1,000,000$ alkiota

Osajoukot: $D = 100$

Osajoukon U_d koko N_d on suhteellinen lukuun $\exp(q_d)$ missä q_d generoidaan tasajakaumasta $U(0,2.9)$

Pienimmässä osajoukossa $N_d = 1721$

Suurimmassa osajoukossa $N_d = 28614$

Muuttuja x_1 generoidaan tasajakaumasta $U(1,11)$

Muuttuja x_2 generoidaan tasajakaumasta $U(-5,5)$

Domain-kohtaiset satunnaistermit u_d ja v_{id} , $i=1,2$ generoidaan multinormaalijakaumasta

Varianssit $Var(u_d) = 1$

$$Var(v_{id}) = 0.125$$

Korrelaatiot $Corr(u_d, v_{id}) = -0.5$

$$Corr(v_{1d}, v_{2d}) = 0$$

Jäännöstermi ε generoidaan jakaumasta $N(0,100)$

Tulosmuuttujan y arvot generoidaan mallilla

$$y_k = (1 + u_d) + (1 + v_{1d})x_{1k} + (1 + v_{2d})x_{2k} + \varepsilon_k$$

missä u_d satunnaiset vakiotermit (*intercept*)

v_{1d} ja v_{2d} satunnaiset kulmakertoimet (*slope*)

HUOM: Mallin kiinteät parametrit

$$\beta_0 = \beta_1 = \beta_2 = 1$$

Populaatiokorrelaatiot:

$$\text{corr}(y, x_1) = 0.44$$

$$\text{corr}(y, x_2) = 0.45$$

$$\text{corr}(x_1, x_2) \approx 0$$

Tulosmuuttujan domain-kohtaiset keskiarvot olivat likimain yhtäsuuria

Kokonaismäärät poikkesivat toisistaan paljon:

Osajoukon koko	Keskimääräinen totaali perusjoukossa
Pieni	50,977
Keskisuuri	131,776
Suuri	263,979

Otanta-asetelma

Ei-suunnitellut (*unplanned*) osajoukot

Systemaattinen PPS-otanta (*Sampling with probabilities proportional to size*)

PPS-otannan kokomuuttuja x_1

Alkion k sisältymistodennäköisyys

$$\Pr\{k \in s\} = \pi_k = \frac{nx_{1k}}{\sum_{k \in U} x_{1k}}$$

Otoskoko $n = 10,000$

Asetelmapainot $a_k = 1/\pi_k$
vaihteluväli 54.5 - 599.8

Osajoukkojen kokoluokittelu

Osajoukko	Otoskoko	Osajoukkoja
Pieni	< 70	47
Keskisuuri	70–119	19
Suuri	>119	34
Yht.		100

Domain-totaalien estimaattorit

HUOM: Yksikkötason lisäinfo x_1 ja x_2 käytettävissä kaikista perusjoukon alkioista estimointia varten

GREG-estimaattorit tavanomaista muotoa:

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

missä sovitteet \hat{y}_k määräytyvät valitun mallin mukaisesti

Avustavat regressiomallit

(1) Kiinteiden vaikutusten D-mallit
(Esimerkiksi malli D1)

$$Y_k = \mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k, \quad k \in U_d$$

missä $\mathbf{x}_k = (1, 1, \dots, 1, x_{1k}, x_{2k})'$ ja $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \dots, \beta_{0D}, \beta_1, \beta_2)'$

Mallien parametrien estimointi: WLS

(2) Lineaariset sekamallit
(Esimerkiksi malli B2)

$$Y_k = \mathbf{x}'_k \boldsymbol{\beta} + u_d + \varepsilon_k, \quad k \in U$$

missä $\mathbf{x}_k = (1, x_{1k}, x_{2k})'$ ja $\boldsymbol{\beta} = (\beta_{01}, \beta_1, \beta_2)'$

Mallien parametrien estimointi: GWLS ja REML

Estimaattorit ja avustavat mallit

Estimaattori	Malli
GREG-A1	$Y_k = \beta_{0d} + \varepsilon_k, k \in U_d$
MGREG-A2	$Y_k = \beta_0 + u_d + \varepsilon_k, k \in U$
GREG-B1	$Y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k, k \in U$
MGREG-B2	$Y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k, k \in U$
GREG-C1	$Y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k, k \in U$
MGREG-C2	$Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k, k \in U$
GREG-D1	$Y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k, k \in U$
MGREG-D2	$Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k, k \in U$

GREG, avustavana mallina lineaarinen **kiinteiden tekijöiden** regressiomalli muotoa (55)

MGREG: Avustavana mallina lineaarinen **sekamalli** (*Mixed model*) muotoa (54)

HUOM:

Kaikki mallit A-D ovat “väärin” spesifioituja
Miksi?

A- ja B-mallit: Otanta-asetelma on **informatiivinen** (*informative sampling*) koska y-arvot riippuvat PPS-otannan kokomuuttujasta x_1 mutta muuttuja ei ole mukana malleissa

C- ja D-mallit: PPS-otannan kokomuuttuja x_1 on mukana

“Double-use” of the auxiliary information
(Särndal 1996)

Osajoukkojen erojen huomioon ottaminen

Mallit A1, B1, C1 ja D1

Kiinteät vakiotermit β_{0d}

Mallit A2, B2, C2 ja D2

Satunnaiset vakiotermit $\beta_0 + u_d$

Kumpi tapa on “parempi”? Miksi?

Taulukko 4. GREG-estimaattoreiden keskimääräinen absoluuttinen suhteellinen harha (*Absolute relative bias ARB %*) ja keskimääräinen suhteellinen RMSE (*Relative root mean squared error RRMSE %*) simulointikokeissa.

Avustava malli ja estimaattori	Keskimääräinen ARB (%)			Keskimääräinen RRMSE (%)		
	Otoksen kokoluokka			Otoksen kokoluokka		
	Pieni (20-69)	Keski- suuri (70-119)	Suuri (120+)	Pieni (20-69)	Keski- suuri (70-119)	Suuri (120+)
Malli A1 $Y_k = \beta_{0d} + \varepsilon_k$						
GREG-A1	1.2	0.7	0.3	20.2	11.9	8.5
Malli A2 $Y_k = \beta_0 + u_d + \varepsilon_k$						
MGREG-A2	0.5	0.5	0.3	19.9	11.8	8.5
Malli B1 $Y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-B1	1.2	0.6	0.3	18.3	10.7	7.7
Malli B2 $Y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-B2	0.5	0.4	0.2	18.0	10.6	7.7
Malli C1 $Y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k$						
GREG-C1	0.4	0.3	0.2	17.5	10.3	7.5
Malli C2 $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
MGREG-C2	0.3	0.3	0.2	17.3	10.2	7.5
Malli D1 $Y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-D1	0.4	0.3	0.2	15.3	8.8	6.5
Malli D2 $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-D2	0.3	0.3	0.2	15.1	8.7	6.5