Helsingin yliopisto
Sosiaalitieteiden laitos

Pienalue-estimointi (78189)

Kevät 2011

Risto Lehtonen

OSA 5

Malliperusteinen SAE

- Synteettinen estimaattori
- EBLUP-estimaattori

Yhteenvetoa

Malliperusteiset pienalue-estimaattorit:
Ominaispiirre on pyrkimys ”voiman lainaamiseen”
Borrow strength

MALLIPERUSTEINEN ESTIMOINTI
Model-based estimation

Domain-totalien $t_d = \sum_{k \in U_d} y_k$ malliperusteiset estimaattorit:

Synteettinen estimaattori
EBLUP-estimaattori

Oletetaan että käytettävissä on alkiotasoinen perusjoukkodata (kehikkoperusjoukko), joka sisältää lisäinformaatiomuuttujat $x_1, x_2, ..., x_j, ..., x_J$

Vektorin $\mathbf{x}_k$ arvot x_{jk} tunnettu kaikille $k \in U$

”Perinteinen” synteettinen estimaattori SYN:

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k \quad , \quad d = 1, 2, ..., D \qquad (58)$$

missä sovitteet (prediktiot)

$$\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}, \quad k \in U \qquad (59)$$

saadaan kiinteiden tekijöiden regressiomallista

$$E_m(Y_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_J x_{Jk}$$

Regressiokertoimet estimoidaan PNS-menetelmällä (ilman asetelmapainoja)

SYN (58) on epäsuora estimaattori. Miksi?

SYN on malliperusteinen epäsuora estimaattori
Voiman lainaaminen – *Borrow strength*

Parametrien t_d SYN-estimaattorit $\hat{t}_{dSYN}$ ovat (määritelmän mukaan) harhaisia asetelman suhteen

Harhan $\text{Bias}(\hat{t}_{dSYN})$ suuruus domainissa *d* riippuu siitä, miten hyvin malli sopii kyseisessä domainissa

Harhan suuruutta ei voida tietää yhden poimitun otoksen perusteella

SYN on siten herkkä mallin valinnalle!

Vaihtoehtoinen malli

ESIM: Kiinteiden tekijöiden malli, jossa mukana domain-spesifit vakiotermit (huom: Tässä $\beta_0 = 0$)

$$E_m(Y_k) = \mathbf{x}_k'\boldsymbol{\beta}$$
$$= \beta_{01} + ... + \beta_{0D} + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_J x_{Jk}$$

missä

$$\mathbf{x}_k = (\delta_{1k}, ..., \delta_{Dk}, x_{1k}, ..., x_{Jk})'$$

ja domain-indikaattorit ovat:

$$\delta_{dk} = 1 \text{ kun } k \in U_d$$
$$\delta_{dk} = 0 \text{ kun } k \notin U_d$$

Muita mallivaihtoehtoja, jotka tuottavat epäsuoran SYN-estimaattorin?

SYN-estimaattorin varianssin estimointi

$$\hat{v}\left(\hat{t}_{dSYN}\right) = \sum_{k \in U_d} \mathbf{x}_k' \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \mathbf{x}_k \,, \quad d = 1, 2, ..., D$$

tai

$$\hat{v}\left(\hat{\mathbf{t}}_{SYN}\right) = \mathbf{t}_x \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \mathbf{t}_x'$$

missä

$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ on estimoidun regressiokerroinvektorin $\hat{\boldsymbol{\beta}}$ kovarianssimatriisin estimaatti

$\mathbf{t}_x$ on apumuuttujien domain-totaalivektori pj:ssa

Keskivirheiden estimointi:

$$\mathrm{s.e}\left(\hat{t}_{dSYN}\right) = \sqrt{\hat{v}\left(\hat{t}_{dSYN}\right)}$$

HUOM: Synteettisen estimaattorin käyttöä ei yleensä suositella

Miksi?

- SYN on ”yliherkkä” mallin spesifioinnille
- Varianssiestimaatit ja keskivirhe-estimaatit yleensä epärealistisen pieniä

SYN laskenta

ESIM: SAS macro EBLUPGREG

Aseta makroparametri SYN=1 (Synthetic estimator)

Synteettinen estimaattori

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k$$

perustuu sekamallin

$$E_m(Y_k \mid u_{0d}) = (\beta_0 + u_{0d}) + \beta_1 x_{1k} + ... + \beta_J x_{Jk} \quad (60)$$

kiinteään osaan, missä malli on sovitettu ilman asetelmapainoja

Sovitteet $\hat{y}_k$ on laskettu mallin (60) perusteella mutta ilman estimoitujen satunnaiskomponenttien $\hat{u}_{0d}$ kontribuutiota:

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + ... + \hat{\beta}_J x_{Jk}$$

Termit $\hat{u}_{0d}$ tulevat mukaan laskentaan EBLUP-estimaattoreissa

Empirical Best Linear Unbiased Predictor EBLUP

EBLUP-estimaattorin yleinen muoto:

$$\hat{t}_{dEBLUP} = \sum_{k \in s_d} y_k + \sum_{k \in U_d - s_d} \hat{y}_k \,, \quad d = 1,2,...,D \qquad (61)$$

missä sovitteet (prediktiot)

$$\hat{y}_k = \mathbf{x}_k'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d), \quad k \in U_d - s_d \qquad (62)$$

lasketaan lineaarisesta sekamallista, johon sisältyy domain-kohtaisia satunnaistermejä.

HUOM: Vertaa (61) SYN-estimaattoriin (58)
HUOM: EBLUP on epäsuora estimaattori. Miksi?

ESIM: Lineaarinen sekamalli:

$$\begin{aligned} &E_m(Y_k \mid \mathbf{u}_d) = \mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d) \\ &= (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})\, x_{1k} + ... + (\beta_J + u_{Jd})\, x_{Jk} \end{aligned} \qquad (63)$$

missä $\mathbf{u}_d = (u_{0d}, u_{1d}, ..., u_{Jd})'$ on domain-kohtaisten satunnaistermien vektori, $d = 1,2,...,D$

Käytännössä sovelletaan usein mallia, jossa satunnaistermeinä ovat domain-spesifit vakiotermit:

$$E_m(Y_k \mid u_{0d}) = (\beta_0 + u_{0d}) + \beta_1\, x_{1k} + ... + \beta_J\, x_{Jk} \qquad (64)$$

HUOM: Vertaa SYN-estimaattorin vaihtoehtoiseen malliin!

HUOM: Perinteisessä EBLUP-estimaattorissa prediktiot $\hat{y}_k$ lasketaan vain joukolle $U_d - s_d$

HUOM: Osajoukossa *d* EBLUP-estimaattori $\hat{t}_{dEBLUP}$ on lähellä SYN-estimaattoria $\hat{t}_{dSYN}$ kun osajoukon otoskoko n_{s_d} on pieni ja SYN-estimaattorin malli sopii hyvin osajoukossa *d*

Vaihtoehtoinen EBLUP-estimaattori on muotoa

$$\hat{t}_{dEBLUP} = \sum_{k \in U_d} \hat{y}_k , \quad d = 1, 2, ..., D \qquad (65)$$

HUOM: Vertaa estimaattoriin (61)

Sekamallin parametrien estimointi

Macro EBLUPGREG:
Yhdistelmä GLS (Generalized least squares) ja REML (Restricted ML) tai ML (Maximum likelihood)
EBLUP-estimaattori on muotoa (65)

Ohjelma Domest: Lisäksi painotetut versiot GWLS ja REML tai ML
EBLUP-estimaattorit (61) ja (65)

Domest ja EBLUPGREG: Malli on muotoa (64)

EBLUP:n (65) **keskineliövirheen** (MSE) estimointi

Macro EBLUPGREG ja ohjelma Domest

Lasketaan MCPE = *Mean Cross Product Error matrix* (Saei and Chambers, 2004, Chapter. 3.3)

$$\text{MCPE} = g_1 + g_2 + 2g_3 + g_4$$

MCPE:n komponentit:
g_1 (general estimate of variation),
g_2 (uncertainty of estimating the beta coefficients),
g_3 (uncertainty of estimating variance components)
g_4 (uncertainty of estimating the model).

MSE-estimaatit ovat estimoidun MCPE-matriisin diagonaalialkioita

Domest ja EBLUPGREG laskevat kaikki neljä komponenttia ja tulostavat MSE-estimaatin ja komponentit g_1, g_2, g_3 ja g_4

EBLUPGREG tulostaa lisäksi MSE-estimaatin josta komponentti g_3 on poistettu

Joissain tilanteissa g_3 on epästabiili ja voi tuottaa yllättäviä tuloksia (malli ehkä spesifioitu väärin)

ESIMERKKI (eri paperi)

YHTEENVETOA

Perusjoukon osajoukkoja koskeva estimointi
Estimation for domains

Keskeiset näkökulmat ja valinnat

A. Osajoukkorakenne (*Domains*)
- A.1 Suunniteltu (*Planned*)
- A.2 Ei-suunniteltu (*Unplanned*)

B. Tilastollinen malli

B.1 Parametrisointi
- B.1.1 Kiinteiden tekijöiden (*fixed effects*) malli
- B.1.2 Sekamalli (*mixed model*)

B.2 Funktionaalinen muoto
- B.2.1 Lineaarinen malli
- B.2.2 Yleistetty lineaarinen malli (GLMM)

C. Osajoukkoparametrien estimaattori

C.1 Estimaattorin tyyppi
- C.1.1 Asetelmaperusteinen (*design-based*)
- C.1.2 Malliperusteinen (*model-based*)

C.2 Suora vai epäsuora?
- C.2.1 Suora (*direct*) estimaattori
- C.2.2 Epäsuora (*indirect*) estimaattori

C.1 Estimaattorin tyyppi

C.1.1 Asetelmaperusteinen (*design-based*)

a) Estimaattorit, joissa ei käytetä lisäinformaatiota
HT-estimaattori (suora)
Hájek-estimaattori (suora)

b) Malliavusteiset estimaattorit
Model assisted estimators

Suoria tai epäsuoria estimaattoreita
Tilastollinen malli: B1.1, B.1.2, B.2.1, B2.2

Yleistetyt regressioestimaattorit
Generalized regression (GREG)

Mallikalibrointiestimaattorit (MC)
Model calibration estimators

c) Kalibrointiestimaattorit
Model-free calibration estimators

C.1.2 Malliperusteinen (***model-based***)

Suoria tai epäsuoria estimaattoreita
Tilastollinen malli: B1.1, B.1.2, B.2.1, B2.2

a) Synteettiset (SYN) estimaattorit
Synthetic estimators – ei suositella

b) EBLUP-estimaattorit
Empirical best linear unbiased predictor

Table 1. Malliavusteisten ja malliperusteisten estimaattoreiden ominaisuuksia
(Lehtonen and Veijanen 2009)

	Asetelmaperusteiset HT, Hájek GREG	**Malliperusteiset** Syntettiset SYN EBLUP
Harha *Bias*	Harhaton (ainakin likimain)	Harhainen Harha ei välttämättä lähene nollaa osajoukon otoskoon kasvaessa
Tarkkuus *Precision* (Varianssi)	Varianssi voi olla suuri pienissä osajoukoissa Varianssi pienenee osajoukon otoskoon kasvaessa	Varianssi voi olla pieni myös pienissä osajoukoissa Varianssi pienenee osajoukon otoskoon kasvaessa
Täsmällisyys *Accuracy* (MSE)	MSE = Variance (likimain)	MSE = Variance + squared Bias Täsmällisyys voi olla huono jos harha on suuri
Luottamus-välit *Confidence intervals*	Asetelmaperusteiset luottamusvälit OK	Asetelmaperusteiset luottamusvälit ei välttämättä OK