Helsingin yliopisto
Sosiaalitieteiden laitos

Pienalue-estimointi (78189)

Kevät 2013

Risto Lehtonen

OSA 5

Malliperusteinen SAE

- Synteettinen estimaattori
- EBLUP-estimaattori

ESIMERKKI: EBLUPGREG-makro

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 $\text{Bias}(\hat{t}_{dSYN}) = E(\hat{t}_{dSYN}) - t_d$
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 D-malli, jossa mukana domain-spesifit vakiotermit (huom: Tässä $\beta_0 = 0$)

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

missä
$\mathbf{x}_k = (\delta_{1k},...,\delta_{Dk}, x_{1k},...,x_{Jk})'$, $\boldsymbol{\beta} = (\beta_{01},...,\beta_{0D},\beta_1,...,\beta_J)'$

ja domain-indikaattorit ovat:

$\delta_{dk} = 1$ kun $k \in U_d$
$\delta_{dk} = 0$ 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
- Muista:
  ”*all models are wrong (but some are useful*”)
- Varianssiestimaatit ja keskivirhe-estimaatit yleensä epärealistisen pieniä

SYN laskenta

ESIM: SAS makro 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 (61) 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)
HUOM: (65) yleisesti käytetty

Sekamallin parametrien estimointi

Macro EBLUPGREG:
Yhdistelmä GLS (Generalized least squares) ja REML (Restricted ML) tai ML (Maximum likelihood)
EBLUP-estimaattori on muotoa (65)
Ks. EBLUPGREG Manual ss. 18–20

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)

$$\mathrm{MCPE} = g_1 + g_2 + 2g_3 + g_4 \qquad (66)$$

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
SAS-MAKRO EBLUPGREG

Vertailu: GREG, SYN ja EBLUP

GREG: Katso osan 4 esimerkki

SYN-estimaattori: Kaava (58)
Malli: Lineaarinen sekamalli (kiinteä osa), kaava (59)

EBLUP-estimaattori: Kaava (65)
MCPE-estimaattori: Kaava (66)
Malli: Lineaarinen sekamalli (kiinteä osa ja satunnaisosa) (64)

Makrokutsu:

```
%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=1,
      greg=1,
      synthetic=1,
      stratified=0,
      output=out1
      );
```

TULOSTUS

(1) SAS Macro EBLUPGREG / GREG- ja SYN-estimointi

SAS Macro EBLUPGREG / GREG ja SYN

domain	n	**GREG**	synthetic	sqrt MSEsyn	**stdGREG**	true Value
1	8	**1292.93**	1302.88	16.3148	**24.8442**	1299.27
2	13	**2472.54**	2560.54	24.7205	**42.2263**	2532.79
3	14	**1816.11**	1921.33	19.4337	**32.3286**	1839.14
4	5	**1913.76**	1880.74	18.1961	**41.4054**	1864.56
5	7	**1747.49**	1750.69	17.8414	**53.4452**	1737.94
6	19	**4745.43**	4595.54	45.7247	**78.6909**	4662.57
7	8	**874.72**	831.64	12.0799	**31.9516**	835.20
8	6	**1026.49**	1053.62	10.4176	**31.4930**	1022.06
9	6	**939.52**	922.79	9.5419	**32.7146**	884.18
10	14	**3630.78**	3632.68	35.6288	**48.2893**	3593.91

sqrtMSEsyn	SYN-estimaattorin Root MSE (keskineliövirheen neliöjuuri)
stdGREG	GREG-estimaattorin keskivirhe s.e

HUOM:
GREG on asetelmaperusteinen malliavusteinen estimaattori

SYN on malliperusteinen estimaattori

(2) SAS Macro EBLUPGREG / EBLUP-estimointi

SAS Macro EBLUPGREG / EBLUP

domain	n	EBLUP	sqrtMSE	sqrt MSENoG3	sqrtg1	sqrtg2	sqrtg3	sqrtg4	true Value
1	8	1299.00	30.0123	24.9866	19.8890	9.3051	11.7560	11.9236	1299.27
2	13	2515.35	45.9894	36.9976	31.4816	11.3290	19.3158	15.7919	2532.79
3	14	1879.40	34.5820	28.1648	23.1123	8.5218	14.1891	13.6549	1839.14
4	5	1898.53	40.4164	34.0563	28.4264	12.7666	15.3892	13.7400	1864.56
5	7	1749.16	38.3467	31.6833	26.3662	11.1594	15.2748	13.5693	1737.94
6	19	4697.90	70.5633	56.5358	49.2219	18.5020	29.8570	20.7648	4662.57
7	8	843.73	19.6283	16.6735	12.3899	5.9942	7.3234	9.4110	835.20
8	6	1045.11	21.3437	18.1871	14.0229	6.2106	7.8988	9.7754	1022.06
9	6	927.46	18.4963	16.0095	11.6288	6.4678	6.5502	8.9019	884.18
10	14	3631.15	66.2352	52.6922	46.2247	16.3351	28.3781	19.3109	3593.91

`sqrtMSE` EBLUP-estimaattorin Root MSE (keskineliövirheen neliöjuuri)

`MSENoG3` MSE ilman komponenttia g3

EBLUP-estimaattorin MSE:n komponentit g1, g2, g3 ja g4
`sqrtg1` `sqrtg2` `sqrtg3` `sqrtg4`

Katso EBLUPGREG-makron manuaali s. 7

HUOM:
EBLUP on malliperusteinen estimaattori

YHTEENVETOA

Perusjoukon osajoukkoja koskeva estimointi
Estimation for domains (small OR large)

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 - REVISITED (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