Helsingin yliopisto
Matematiikan ja tilastotieteen laitos

Pienalue-estimointi (78189)

Kevät 2009
Risto Lehtonen
7.4.2009

OSA 5

SAE-OHJELMA DOMEST
SAS Makro EBLUPGREG

Malliperusteinen SAE
- Synteettinen estimaattori
- EBLUP-estimaattori

Estimaattoreiden ja ohjelmien vertailu

MALLIPERUSTEINEN ESTIMOINTI
Model-based estimation

Estimaattorit

Synteettinen estimaattori
EBLUP-estimaattori

Oletetaan edelleen että käytettävissä on alkiotasoinen perusjoukkodata

Vektori $\mathbf{x}_k$ 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

ESIM: Kiinteiden tekijöiden malli

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

SYN on malliperusteinen estimaattori

Parametrin $t_d = \sum_{k \in U_d} Y_k$ SYN-estimaattori $\hat{t}_{dSYN}$ on (määritelmän mukaan) harhainen asetelman suhteen

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

Harhan suuruutta ei voida tietää yhden poimitun otoksen perusteella

SYN on siten herkkä mallin valinnalle!

SYN-estimaattorin varianssin estimointi

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

tai (ks. IML-muoto)

$$\hat{v}\left(\hat{\mathbf{t}}_{SYN}\right) = \mathbf{t}_x \text{Cov}(\hat{\mathbf{B}}) \mathbf{t}_x'$$

missä
$\text{Cov}(\hat{\mathbf{B}})$ on estimoidun regressiokerroinvektorin kovarianssimatriisin estimaatti
$\mathbf{t}_x$ on apumuuttujien domain-totaalit pj:ssa

Keskivirheen estimointi apumuuttujien pj:n

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

Laskenta SAS-proseduurilla IML

```
******************;
* Sovitetaan otosdatalle D-tyypin
regressiomalli;
* Tulosmuuttuja y
* Selittäjä (apumuuttuja) x;

ods trace on;

proc surveyreg data=omaotos total=966;
model y=x / covb;
weight samplingweight;
ods output CovB=covb(drop=parameter);
run;

* Muokataan perusjoukkodata havaintojen
lukumäärän ja x-muuttujan domain-kohtaisten
totaalien laskentaa varten;

data x;
retain x0 0;
set pj(keep=domain x);
x0=1;
run;

* Lasketaan totaalit tiedostoon xs;

proc summary data=x nway;
class domain;
output out=xs(drop=_TYPE_ _FREQ_ domain)
sum(x0)=x0 sum(x)=x1;
run;
```

```
* SAS-proseduuri IML ;
* IML = Interactive Matrix Language;

proc iml;
* Luetaan data xs IML-matriisiksi xs;
use xs;
read all into xs;
print xs;

* Luetaan data covb IML-matriisiksi covb;
use covb;
read all into covb;
print covb;

* Lasketaan kovarianssimatriisi var;
var=xs*covb*xs`;
print var;

* Lasketaan keskivirheet ja laitetaan ne
sarakevektoriksi se;
se=vecdiag(sqrt(var));
print se;
quit;
```

```
x-muuttujein domain-totaalit
          XS
       69 1212.3723
      120 2651.3597
       94 1921.3543
       86 1984.3166
       86 1744.7016
      204 4954.3262
       46 740.17422
       47 1131.9455
       40 1011.4321
      174 3696.883
```

Mallin betavektorin kov.matriisi

	COVB
0.3849613	-0.016761
-0.016761	0.0007749

Domain-totaalien SYN-estimaattien kov.matriisi

	VAR				
	COL1	COL2	COL3	COL4	COL5
ROW1	167.46901	173.39201	169.59177	105.9901	158.00958
ROW2	173.39201	324.95169	247.84097	236.51031	226.18204
ROW3	169.59177	247.84097	207.63723	170.31623	191.10542
ROW4	105.9901	236.51031	170.31623	177.58962	154.55893
ROW5	158.00958	226.18204	191.10542	154.55893	176.01725
ROW6	197.72129	571.68582	382.5463	445.03251	344.47782
ROW7	126.42461	112.66044	118.96733	64.117108	111.43597
ROW8	47.614205	131.30277	88.959239	101.61963	80.215068
ROW9	30.085986	113.81922	71.539027	91.105472	63.962931
ROW10	283.45929	464.81835	372.1743	328.75444	341.18098

	VAR				
	COL6	COL7	COL8	COL9	COL10
ROW1	197.72129	126.42461	47.614205	30.085986	283.45929
ROW2	571.68582	112.66044	131.30277	113.81922	464.81835
ROW3	382.5463	118.96733	88.959239	71.539027	372.1743
ROW4	445.03251	64.117108	101.61963	91.105472	328.75444
ROW5	344.47782	111.43597	80.215068	63.962931	341.18098
ROW6	1159.4561	103.24686	263.14607	244.12853	767.00881
ROW7	103.24686	97.726244	25.661821	12.346197	192.50321
ROW8	263.14607	25.661821	59.778842	55.170038	177.2051
ROW9	244.12853	12.346197	55.170038	52.398869	148.31969
ROW10	767.00881	192.50321	177.2051	148.31969	681.64061

Domain-totaalien SYN-estimaattien keskivirheet

SE
12.940982
18.026416
14.409623
13.326275
13.267149
34.050786
9.8856585
7.7316778
7.2387063
26.108248

Empirical Best Linear Unbiased Predictor EBLUP

EBLUP-estimaattorin yleinen muoto:

$$\hat{t}_{dEBLUP} = \sum_{k \in s_d} y_k + \sum_{k \in U_d - s_d} \hat{y}_k \,, \quad d = 1,2,...,D \qquad (60)$$

missä sovitteet (prediktiot)

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

saadaan **lineaarisesta sekamallista**

ESIM: Lineaarinen sekamalli (53):

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

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

HUOM: EBLUP-estimaattorissa prediktiot $\hat{y}_k$ lasketaan vain otoksen kuulumattomaan pj:n osaan

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*

SAS-MAKRO EBLUPGREG

EURAREA-hankkeen yhteydessä laadittu SAS-makro pienalue-estimointia varten

PIENALUE-ESTIMOINNIN OHJELMA DOMEST

Java-kielinen erillinen ohjelma pienalue-estimointia varten

Ari Veijanen, Tilastokeskus

Katso erillinen paperi