Helsingin yliopisto
Sosiaalitieteiden laitos

Pienalue-estimointi (78189)

Kevät 2013

Risto Lehtonen

OSA 6

JATKOESIMERKKEJÄ:

EBLUPGREG-makro

GREG-estimointi: Mallin valinta
EBLUP-estimointi

SAS PROC MIXED

Sekamallin estimointi
EBLUP-estimointi

Tulosten vertailu

ESIMERKKI: SAS-makro EBLUPGREG
GREG-estimointi: Mallin valinta

GREG-estimaattori

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

GREG-estimaattorin avustavat mallit

(a) Kiinteiden tekijöiden P-malli:

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k$$

missä $\mathbf{x}_k = (1, x_{1k}, ..., x_{Jk})'$,
$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$

(b) Kiinteiden tekijöiden D-malli:

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k ,$$

missä $\mathbf{x}_k = (\delta_{1k}, \delta_{2k}, ..., \delta_{Dk}, x_{1k}, x_{2k})'$,
$\delta_{dk} = 1$ kun $k \in U_d$, nolla muulloin
$\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, ..., \beta_{0D}, \beta_1, \beta_2)'$

Mallien parametrien estimointi: WLS

SAS-toteutus (SAE_demot_5a.sas)

(1) Luodaan perusjoukkodataan domain-indikaattormuuttujat (dummyt d1,...,d10)

```
data pj(drop=d);
set a.pj;
array c(10) d1-d10;
   do d=1 to 10;
   c(d)=0;
   if domain=d then c(d)=1;
   end;
run;
```

(2) Sovitetaan mallit poimitulle SRSWOR-otokselle

Otoskoko n=100
Käytetään SURVEYREG-proseduuria

Malli (a) P-malli

```
proc surveyreg data=omaotos;
model y=x / adjrsq solution;
run;
```

Malli (b) D-malli

```
proc surveyreg data=omaotos;
model y=d1-d9 x / adjrsq solution;
run;
```

Totaalin GREG-estimointi / SURVEYREG Procedure

(2a) GREG: Kiinteiden tekijöiden P-malli

The SURVEYREG Procedure

Regression Analysis for Dependent Variable y

Data Summary

Number of Observations	100
Mean of y	21.35558
Sum of y	2135.6

Fit Statistics

R-square	0.8097
Adjusted R-square	0.8078
Root MSE	1.4873
Denominator DF	99

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	9.50785465	0.58225585	16.33	<.0001
x	0.53190892	0.02562875	20.75	<.0001

(2b) GREG: Kiinteiden tekijöiden D-malli

Fit Statistics

R-square	0.8503
Adjusted R-square	0.8334
Root MSE	1.3846
Denominator DF	99

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	10.0462261	0.81933710	12.26	<.0001
d1	-0.8614483	0.58625075	-1.47	0.1449
d2	-0.0317282	0.55009855	-0.06	0.9541
d3	-0.6423523	0.60087142	-1.07	0.2877
d4	-0.1996628	0.62685368	-0.32	0.7508
d5	-0.6330242	0.89125208	-0.71	0.4792
d6	0.7373336	0.48007503	1.54	0.1278
d7	0.2117662	0.89650212	0.24	0.8138
d8	-1.3761111	0.55944698	-2.46	0.0156
d9	-1.7609842	0.58639783	-3.00	0.0034
x	0.5139214	0.03213692	15.99	<.0001

3) EBLUPGREG-ajo ja GREG-tulostukset

Malli (a) P-malli

```
%eblupgreg
     (sample=omaotos,
     population=pj,
     y=y,
     xlist=x,
     regionIdentifier=domain,
     test=0,
     estimateMeans=0,
     weights=samplingweight,
     convergenceCrit=1e-8,
     maxiterations=200,
     initialSigma2=1,
     modules=modules7.eurarea7,
     parametersEstimatedBy='REML',
     eblup=0,
     greg=1,
     synthetic=0,
     stratified=0,
     output=out1
     );
```

SAS Macro EBLUPGREG / GREG-estimointi
(2a) GREG: Kiinteiden tekijöiden P-malli

	parameters
parameter	in GREG
intercept	9.5078546479
X	0.5319089174

Malli (b) D-malli

```
%eblupgreg
     (sample=omaotos,
     population=pj,
     y=y,
     xlist=d1 d2 d3 d4 d5 d6 d7 d8 d9 x,
     regionIdentifier=domain,
     test=0,
     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=out2
     );
```

(2b) GREG: Kiinteiden tekijöiden D-malli

	parameters
parameter	in GREG
intercept	10.046226141
D1	-0.861448313
D2	-0.031728206
D3	-0.642352308
D4	-0.199662814
D5	-0.633024214
D6	0.7373336423
D7	0.2117662202
D8	-1.376111136
D9	-1.760984199
X	0.5139213998

GREG-estimointi: Tulosten vertailu

GREG ja stdGREG: Kiinteiden tekijöiden P-malli
GREG2 ja stdGREG2: Kiinteiden tekijöiden D-malli

region	GREG	std GREG	GREG2	std GREG2	Parametrity
1	1262.9	31.1	1256.8	24.5	1299.3
2	2568.1	38.9	2564.3	39.9	2532.8
3	1849.9	38.0	1871.4	36.2	1839.1
4	1870.5	37.5	1866.6	36.5	1864.6
5	1720.5	62.2	1706.2	60.9	1737.9
6	4779.9	51.7	4746.0	45.2	4662.6
7	850.5	31.6	852.3	30.6	835.2
8	985.4	29.4	989.2	15.5	1022.1
9	854.6	34.1	851.2	15.5	884.2
10	3638.3	64.7	3648.0	63.3	3593.9

Huomataan, että D-tyypin avustavalla mallilla on taipumus tuottaa GREG-estimaattorille pienempi keskivirhe verrattuna P-tyyppiseen malliin

Miksi?

ESIMERKKI: SAS-makro EBLUPGREG EBLUP-estimaattori

EBLUP-estimaattori:

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

Malli: Lineaarinen sekamalli

Domain-kohtaiset satunnaiset vakiotermit

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

Parametrien estimointi GLS ja ML tai REML

MCPE-estimaattori:

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

MSE-estimaatit ovat estimoidun MCPE-matriisin diagonaalialkioita

Ks: Luentomateriaalien Osa 5
EBLUPGREG Manual, ss. 18–20

Makrokutsu:

```
%eblupgreg
     (sample=omaotos,
     population=pj,
     y=y,
     xlist=x,
     regionIdentifier=domain,
     test=0,
     estimateMeans=0,
     weights=samplingweight,
     convergenceCrit=1e-8,
     maxiterations=200,
     initialSigma2=1,
     modules=modules7.eurarea7,
     parametersEstimatedBy='REML',
     eblup=1,
     greg=0,
     synthetic=0,
     stratified=0,
     output=out3
     );
```

SAS Macro EBLUPGREG / EBLUP-estimointi

(3a) EBLUP: Lineaarinen sekamalli

	parameters	
parameter	in mixed model	std. error
intercept	9.7244765393	0.6155013902
X	0.5127605458	0.0263034313
phi(area)	0.1680298307	
area effect variance	0.3207946654	
sigma2	1.9091530595	

SAS Macro EBLUPGREG / GREG-estimointi
(3a) EBLUP: Lineaarinen sekamalli

region	EBLUP	area Effect	effect StdErr	sqrtMSE	sqrt MSENoG3
1	1273.15	-0.26015	0.41896	32.5638	28.7714
2	2555.43	0.23010	0.34187	38.1675	34.4088
3	1878.54	-0.20764	0.35122	31.4690	28.3228
4	1861.06	0.07915	0.40482	38.6951	34.0882
5	1717.56	-0.14526	0.41852	40.4884	35.5560
6	4710.18	0.88593	0.32248	52.2145	48.0131
7	838.33	0.22084	0.45594	24.0584	21.4341
8	1012.72	-0.46729	0.43545	22.9981	20.4191
9	881.53	-0.56746	0.44878	20.9591	18.7503
10	3629.38	0.23178	0.36252	64.5895	57.1321

Tarkastellaan esimerkinomaisesti sekamallin estimointia SAS-proseduurilla MIXED

SAS PROC MIXED

Lineaarinen sekamalli:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

missä

$\mathbf{X}$ on kiinteiden tekijöiden $\boldsymbol{\beta}$ mallimatriisi

$\mathbf{Z}$ on satunnaistermien $\mathbf{u}$ mallimatriisi

Perusoletukset:

Satunnaismuuttujat $\mathbf{u}$ ja $\boldsymbol{\varepsilon}$ ovat normaalijakautuneita ja korreloimattomia:

$\mathbf{u}$ on $N(\mathbf{0},\mathbf{G})$ ja $\boldsymbol{\varepsilon}$ on $N(\mathbf{0},\mathbf{R})$

$$E\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \quad \text{ja} \quad Var\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

missä

$\mathbf{G}$ on satunnaistermien $\mathbf{u}$ kovarianssimatriisi

$\mathbf{R}$ on jäännösten $\boldsymbol{\varepsilon}$ kovarianssimatriisi

jolloin $\mathbf{V}=\mathbf{ZGZ}'+\mathbf{R}$

on vektorin $\mathbf{y}$ kovarianssimatriisi

Kovarianssimatriisin $\mathbf{V}$ mallinnus:

Kiinnitä mallimatriisi $\mathbf{Z}$ ja spesifioi kovarianssirakenteet $\mathbf{G}$ ja $\mathbf{R}$

HUOM: Jos $\mathbf{R}=\sigma^2\mathbf{I}$ ja $\mathbf{Z}=0$ saadaan kiinteiden tekijöiden lineaarinen malli $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

SAS PROC MIXED

Paljon erilaisia kovarianssirakenteita, esim:

`type=vc` (*variance components model*)

Mallin parametrien estimointi

Beta-vektorin ja satunnaistermien estimointi:
GLS (*Generalized least squares*) ja ML tai REML (*Restricted ML*)

Mixed model equations (Henderson 1984)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \quad \text{(EBLUE)}$$

$$\hat{\mathbf{u}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad \text{(EBLUP)}$$

Iteratiivinen laskenta (Newton-Raphson-algoritmi)
Prediktiot: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}$

Perusteellisempi kuvaus: Veijanen-Lehtonen (2010):
PROGRAM *DOMEST* FOR ESTIMATION FOR DOMAINS AND SMALL AREAS. PART 1: Technical documentation

Yksinkertainen malli (*Nested error regression model*)

Malli: Satunnainen domain-tasoinen vakiotermi u_d (*random intercept*) ja yksi selittävä x-muuttuja

$$y_k = \beta_0 + \beta_1 x_k + u_d + \varepsilon_k$$

missä u_d on $N(0, \sigma_u^2)$ ja ε_k on $N(0, \sigma^2)$

```
proc mixed data=omaotos asycov;
class domain;
model y=x / solution;
random intercept/solution type=vc subject=domain g;
run;
```

SAS PROC MIXED

(3a) Sekamallin estimointi

The Mixed Procedure

Estimated G Matrix

Row	Effect	domain	Col1
1	Intercept	1	0.3208

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	domain	0.3208
Residual		1.9092

Asymptotic Covariance Matrix of Estimates

Row	Cov Parm	CovP1	CovP2
1	Intercept	0.06269	-0.00708
2	Residual	-0.00708	0.08071

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	9.7245	0.6155	9	15.80	<.0001
x	0.5128	0.02630	89	19.49	<.0001

Solution for Random Effects

Effect	domain	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	-0.2601	0.4189	89	-0.62	0.5362
Intercept	2	0.2301	0.3419	89	0.67	0.5027
Intercept	3	-0.2076	0.3512	89	-0.59	0.5559
Intercept	4	0.07914	0.4048	89	0.20	0.8454
Intercept	5	-0.1453	0.4185	89	-0.35	0.7294
Intercept	6	0.8859	0.3225	89	2.75	0.0073
Intercept	7	0.2208	0.4559	89	0.48	0.6293
Intercept	8	-0.4673	0.4354	89	-1.07	0.2861
Intercept	9	-0.5674	0.4488	89	-1.26	0.2094
Intercept	10	0.2318	0.3625	89	0.64	0.5242

ESIMERKKI
EBLUP-estimointi proseduurilla MIXED

SAS-koodi:

```
data omaotos;
set omaotos;
otos=1;run;

data pj2;
merge pj omaotos;
by id;
if otos=. then y=.;
run;

proc mixed data=pj2 asycov ratio;
class domain;
model y=x /
    solution
    covb
    outpred=pj2;
random intercept/
    solution
    type=vc
    subject=domain g;
run;

proc summary data=pj2 nway;
class domain;
var pred;
output out=EBLUP
    (drop=_type_
    rename=(_freq_=n))
    sum(pred)=EBLUP
    sum(StdErrPred)=RootMSE;
run;

proc print data=eblup noobs;
var domain n EBLUP rootMSE;run;
```

Tulostus (olennaiset osat) SAS PROC MIXED

(3b) Prediktioiden laskenta

```
                  Model Information

Data Set                     WORK.PJ2
Dependent Variable           y
Covariance Structure         Variance Components
Subject Effect               domain
Estimation Method            REML
Residual Variance Method     Profile
Fixed Effects SE Method      Model-Based
Degrees of Freedom Method    Containment


             Class Level Information

Class     Levels    Values

domain        10    1 2 3 4 5 6 7 8 9 10


            Dimensions

Covariance Parameters           2
Columns in X                    2
Columns in Z Per Subject        1
Subjects                       10
Max Obs Per Subject           204


          Number of Observations

Number of Observations Read            966
Number of Observations Used            100
Number of Observations Not Used        866


           Estimated G Matrix

 Row    Effect       domain        Col1

   1    Intercept     1          0.3208


       Covariance Parameter Estimates

Cov Parm      Subject       Ratio    Estimate

Intercept     domain       0.1680      0.3208
Residual                   1.0000      1.9092
```

Asymptotic Covariance Matrix of Estimates

Row	Cov Parm	CovP1	CovP2
1	Intercept	0.06269	-0.00708
2	Residual	-0.00708	0.08071

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	9.7245	0.6155	9	15.80	<.0001
x	0.5128	0.02630	89	19.49	<.0001

Solution for Random Effects

Effect	domain	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	-0.2601	0.4189	89	-0.62	0.5362
Intercept	2	0.2301	0.3419	89	0.67	0.5027
Intercept	3	-0.2076	0.3512	89	-0.59	0.5559
Intercept	4	0.07914	0.4048	89	0.20	0.8454
Intercept	5	-0.1453	0.4185	89	-0.35	0.7294
Intercept	6	0.8859	0.3225	89	2.75	0.0073
Intercept	7	0.2208	0.4559	89	0.48	0.6293
Intercept	8	-0.4673	0.4354	89	-1.07	0.2861
Intercept	9	-0.5674	0.4488	89	-1.26	0.2094
Intercept	10	0.2318	0.3625	89	0.64	0.5242

(3c) EBLUP-tulokset

domain	n	EBLUP	RootMSE
1	69	1274.70	29.9112
2	120	2554.06	39.8338
3	94	1879.78	32.2174
4	86	1860.59	36.1126
5	86	1718.43	37.8369
6	204	4704.91	56.8927
7	46	837.02	22.1081
8	47	1015.51	21.0579
9	40	884.90	19.1317
10	174	3628.00	61.8845

HUOM: Vertaa SAS-makron EBLUPGREG tuloksiin!