

# **Pienalue-estimointi (78189)**

Kevät 2013

Risto Lehtonen

## **OSA 1: TEKNINEN MATERIAALI**

### **Käsiteltävät teemat:**

Estimaattorin tyyppi

Mallin valinta

Asetelmaperusteinen estimointi  
Horvitz-Thompson (HT)

Malliavusteinen estimointi  
Yleistetyt regressioestimaattorit  
Generalized Regression Estimator (GREG)  
Kalibrointiestimaattorit

Malliperusteinen estimointi  
Synteettiset estimaattorit (SYN)

Esimerkkejä

-----

Lehtonen-Pahkinen (2004) Practical Methods for Design and Analysis of Complex Surveys. 2nd Edition. Chichester: Wiley.

# ESTIMAATTORIN TYYPPI

## Päätyypit

### 1. Asetelmaperusteiset estimaattorit

*Desig-based estimators*

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

b) Malliavusteiset estimaattorit

*Model assisted estimators*

Yleistetyt regressioestimaattorit

*Generalized regression (GREG)*

Mallikalibrointiestimaattorit (MC)

*Model calibration estimators*

c) Kalibrointiestimaattorit

*Model-free calibration estimators*

### 2. Malliperusteiset estimaattorit

*Model based estimators*

a) Synteettiset (SYN) estimaattorit

*Synthetic estimators*

b) EBLUP- ja EBP-estimaattorit

*Empirical best linear unbiased predictor*

*Empirical Best Predictor*

**Table 1.** Malliavusteisten ja malliperusteisten estimaattoreiden ominaisuuksia (Lehtonen and Pahkinen 2004)

	<b>Asetelmaperusteiset</b> HT , GREG, MC	<b>Malliperusteiset</b> Syntettiset SYN, EBLUP, EBP
<b>Harha</b> <i>Bias</i>	Harhaton (ainakin likimain)	Harhainen  Harha ei välttämättä lähene nollaa osajoukon otoskoon kasvaessa
<b>Tarkkuus</b> <i>Precision</i> (Varianssi)	Varianssi voi olla suuri pienissä osajoukoissa  Varianssi pienenee osajoukon otoskoon kasvaessa	Varianssi voi olla pieni myös pienissä osajoukoissa  Varianssi pienenee osajoukon otoskoon kasvaessa
<b>Täsmällisyys</b> <i>Accuracy</i> (MSE)	MSE = Variance (likimain)	MSE = Variance + squared Bias  Täsmällisyys voi olla huono jos harha on suuri
<b>Luottamusvälit</b> <i>Confidence intervals</i>	Asetelmaperusteiset luottamusvälit OK	Asetelmaperusteiset luottamusvälit ei välttämättä OK

## Estimaattoreiden teoreettisia ominaisuuksia voidaan tutkia empiirisesti simulointikokeilla

Simulointikokeessa poimitaan  $N$  alkion perusjoukosta  $U$   $J$  riippumatonta  $n$  alkion otosta otanta-asetelmalla  $p(s)$ .

**Parametri (totaali)**  $T = \sum_{k=1}^N Y_k$

**Estimaattori (esim. HT)**  $\hat{t} = \sum_{k=1}^n y_k / \pi_k$

**Harha**  $Bias(\hat{t}) = E(\hat{t}) - T \approx \sum_{j=1}^J \hat{t}_j / J - T$

missä  $\hat{t}_j$  on totaaliestimaatti simulaatiosta  $j$

Absolute relative bias  $ARB(\hat{t}) = | \sum_{j=1}^J \hat{t}_j / J - T | / T$

**Täsmällisyys Precision**

$$\begin{aligned} Var(\hat{t}) &= E(\hat{t} - E(\hat{t}))^2 \\ &\approx \sum_{j=1}^J (\hat{t}_j - \sum_{j=1}^J \hat{t}_j / J)^2 / (J - 1) \end{aligned}$$

**Tarkkuus Accuracy**

$$\begin{aligned} MSE(\hat{t}) &= E(\hat{t} - T)^2 = Var(\hat{t}) + Bias^2(\hat{t}) \\ &\approx \sum_{j=1}^J (\hat{t}_j - T)^2 / J \end{aligned}$$

# MALLIN VALINTA

## Kaksi näkökulmaa

Mallin matemaattinen muoto

Mallin parametrisointi

## ESIMERKKI: Matemaattinen muoto

Jatkuva tulosmuuttuja

Lineaarinen malli

Binäärinen tulosmuuttuja

Binominen logistinen malli

Moniluokkainen tulosmuuttuja

Multinomiaalinen logistinen malli

Lukumäärämuuttuja

Poisson-regressiomalli

## HUOM:

Mallit ovat yleistettyjen lineaaristen sekamallien (*Generalized Linear Mixed Models* GLMM) erikoistapauksia (McCulloch and Searle 2001)

## Taulukko 2. Tilastolliset mallit matemaattisen muodon, tulosmuuttujan tyypin ja selittäjien tyypin mukaan

	Mallin tyyppi		
	Lineaariset mallit	Logistiset mallit	Logaritmiset (Poisson) mallit
Selittäjämuuttujat	Tulosmuuttuja jatkuva	Tulosmuuttuja binäärinen tai moniluokkainen	Tulosmuuttuja lukumäärämuuttuja
Diskreettejä	<i>Lineaarinen ANOVA</i>	<i>Logit-ANOVA</i>	<i>Logaritminen (Poisson) ANOVA</i>
Jatkuvia	<i>Lineaarinen regressio</i>	<i>Logit-regressio</i>	<i>Logaritminen (Poisson) regressio</i>
Diskreettejä ja jatkuvia	<i>Lineaarinen ANCOVA</i>	<i>Logit-ANCOVA</i>	<i>Logaritminen (Poisson) ANCOVA</i>

# MALLIN PARAMETRISOINTI

Populaation  $U$  alkio  $k \in U$

Osajoukko (*domain, small area*)  $U_d \subset U$ ,  $d = 1, \dots, D$

**Kaksi perustyyppiä:**

**Kiinteiden tekijöiden malli**

*Fixed-effects model formulation*

Esimerkiksi: Lineaarinen malli

$$y_k = \beta_0 + \beta_1 z_{1k} + \varepsilon_k, k \in U$$

Kiinteät vaikutukset  $\beta_0$  ja  $\beta_1$

**Sekamalli / Hierarkkinen malli / Monitasomalli**

*Mixed model / Hierarchical model / Multilevel model formulation*

Esimerkiksi: Lineaarinen sekamalli

$$y_k = \beta_0 + u_{0d} + \beta_1 z_{1k} + \varepsilon_k, k \in U_d$$

Domain-kohtaiset satunnaistermit  $u_{0d}$

**HUOM:** Kutakin mallia vastaava malliavusteinen (GREG; MC) ja malliperusteinen (SYN, EBLUP, EBP) estimaattori voidaan konstruoida

# ESIMERKKI

(Lehtonen, Särndal and Veijanen 2003)

**Table 3.** Estimaattoreiden luokittelu mallin valinnan ja estimaattorin tyypin mukaan

<i>MALLIN VALINTA</i>			<i>ESTIMAATTORIN TYYPPI</i>	
Mallin parametrisointi	Aggregoinnin taso	Matemaattinen muoto	Malli-perusteinen	Asetelma-perusteinen malli-avusteinen
<b>Kiinteiden tekijöiden mallit</b>	<b>Population models (P-models)</b>	<b>1. Lineaarinen</b>	SYN-P	GREG-P
		<b>2. Logistinen</b>	LSYN-P	LGREG-P
	<b>Domain models (D-models)</b>	<b>3. Lineaarinen</b>	SYN-D	GREG-D
		<b>4. Logistinen</b>	LSYN-D	LGREG-D
<b>Sekamallit</b>	<b>Domain models (D-models)</b>	<b>5. Lineaarinen</b>	MSYN-D	MGREG-D
		<b>6. Logistinen</b>	MLSYN-D	MLGREG-D

**P-mallit (Perusjoukon tasoinen):** Kiinteiden tekijöiden mallit, parametrisointi populaatiotasoisena

**D-mallit (Domain-tasoinen):** Mallissa domain-kohtaisia parametreja (kiinteitä tai satunnaisia)



# TARKASTELUKEHIKKO JA PERUSTEITA

## Notaatio

Äärellinen perusjoukko  $U = \{1, 2, \dots, k, \dots, N\}$

Toisensa poissulkevat perusjoukon osajoukot  
(domains)  $U_1, \dots, U_d, \dots, U_D$

Oletetaan ensin että **alkiotasoinen** (*unit-level*)  
**perusjoukko**  $U$  on käytettävissä  
kehikkoperusjoukon muodossa (esim. SAS-tiedosto,  
relaatiotietokanta tms.)

Tilastorekisteri

Väestörekisteri

Yritysrekisteri

Oletetaan että  $U$  sisältää jokaiselle alkiolle  $k \in U$   
muuttujat:

Identifikaatiomuuttuja (ID)

Osajoukkoon kuulumisindikaattorit

Osateindikaattorit

Ryväsindikaattorit

Apumuuttujatiedot (z-muuttujat)

**Tulosmuuttuja:**  $y$ 

$Y_k$  Tulosmuuttujan (tuntematon) arvo alkiolle  $k$

**Kohdeparametrit:** Osajoukkototaalit  
(*Domain totals*)

$$T_d = \sum_{k \in U_d} Y_k, \quad d = 1, \dots, D$$

**Apumuuttujat:**

$$\mathbf{z}_k = (z_{1k}, \dots, z_{jk}, \dots, z_{Jk})'$$

dimensio  $J \geq 1$

**Domain-indikaattorivektori:**

$\delta_k = (\delta_{1k}, \dots, \delta_{dk}, \dots, \delta_{Dk})'$ :  $\delta_{dk} = 1$  kun  $k \in U_d$ , nolla muulloin

**Ositeindikaattorivektori:**  $\tau_k$ :

$\tau_{hk} = 1$  kun  $k \in U_h$ ,  $h = 1, \dots, H$ , nolla muulloin, missä  $U_h$  viittaa ositteeseen  $h$  ja  $H$  on ositteiden lukumäärä.

**HUOM:** Vektori  $\mathbf{z}_k$  oletetaan tunnetuksi kaikille alkioille  $k \in U$

## ESIMERKKI

Henkilötutkimus: Vektori  $\mathbf{z}_k$  sisältää muuttujat ikä, sukupuoli, verotustiedot, koulutustiedot, työllisyystiedot ym. jatkuvia ja diskreettejä muuttujia henkilölle  $k$

Yritystutkimus: Vektori  $\mathbf{z}_k$  sisältää muuttujat liikevaihto ja henkilöstön lukumäärä yritykselle  $k$

**Miksi apumuuttujavektori  $\mathbf{z}_k$  oletetaan tunnetuksi?**

**Joustavuusperiaate.** Data voidaan tarvittaessa aggregoida osajoukko- tai ositetasolle.

Parhaat mallit saadaan alkiotasoisina.

**HUOM:** Yksinkertaisimmissa tapauksissa riittää että tunnetaan aggregaatteja, kuten osajoukkojen totaalit  $T_{dz_1}, \dots, T_{dz_j}$  apumuuttujille  $z_j$

Mallinnusvaiheessa tavallisesti oletetaan että vakio 1 on vektorin  $\mathbf{z}_k$  ensimmäinen alkio

## Otanta ja tiedonkeruu

**Satunnaisotos**  $s$  kokoa  $n$  poimitaan perusjoukosta  $U$  käyttämällä otanta-asetelmaa  $p(s)$  jossa sisällysmistodennäköisyys  $\pi_k$  ( $0 < \pi_k \leq 1$ ) kiinnitetään alkiolle  $k \in U$

**Asetelmapaino:**  $w_k = 1/\pi_k$

**Tulosmuuttujan arvot**  $y_k$  mitataan otosalkioilta  $k \in s$

### **Mahdolliset lisätoimenpiteet:**

Vastauskadon adjustointi

Yksikkökato (*unit nonresponse*):

Uudelleenpainotus tarvittaessa

Eräkato (*item nonresponse*):

Imputointi tarvittaessa

## KAKSI VAIHTOEHTOISTA DOMAIN- RAKENNETTA

### **Ei-suunniteltu** (*unplanned*) **domain-rakenne:**

Osajoukkojen otokset:  $s_d = U_d \cap s$ ,  $d = 1, \dots, D$

Osajoukkojen  $d$  otoskokoja  $n_{s_d}$  ei ole kiinnitetty otanta-asetelmassa

Otoskoot  $n_{s_d}$  ovat satunnaismuuttujia

### **Suunniteltu** (*planned*) **domain-rakenne:**

Osajoukkojen otokset:  $s_d \subset U_d$ ,  $d = 1, \dots, D$

Osajoukkojen  $d$  otoskoot  $n_d$  on kiinnitetty otanta-asetelmassa (ositettu otanta)

Osajoukkojen otoskoot  $n_d$  ovat kiinteitä

### **Ositettu otanta ja sopiva kiintiöintimenetelmä**

Optimaalinen (Neyman) -kiintiöinti

Bankier-kiintiöinti

Tasakiintiöinti

**Table 4.** Planned and unplanned domain structures in a stratified sample of  $n$  elements, Lehtonen and Pahkinen (2004)

<b>Unplanned domains</b>	<i>Strata (planned domains)</i>						Sum
	1	2	...	$h$	...	$H$	
1	$n_{s_{11}}$	$n_{s_{12}}$	...	$n_{s_{1h}}$	...	$n_{s_{1H}}$	$n_{s_1}$
2	$n_{s_{21}}$	$n_{s_{22}}$	...	$n_{s_{2h}}$	...	$n_{s_{2H}}$	$n_{s_2}$
•	...	...	...	...	...	...	...
•							
•							
$d$	$n_{s_{d1}}$	$n_{s_{d2}}$	...	$n_{s_{dh}}$	...	$n_{s_{dH}}$	$n_{s_d}$
•	...	...	...	...	...	...	...
•							
•							
$D$	$n_{s_{D1}}$	$n_{s_{D2}}$	...	$n_{s_{Dh}}$	...	$n_{s_{DH}}$	$n_{s_D}$
Sum	$n_1$	$n_2$	...	$n_h$	...	$n_H$	$n$

Stratum sample sizes  $n_h$ ,  $h = 1, \dots, H$ , are fixed in the sampling design. Thus, the strata are defined as *planned* domains.

Sample sizes  $n_{s_d}$ ,  $d = 1, \dots, D$ , for *unplanned* domains are not fixed in advance and thus are random variables.

Cell sample sizes  $n_{s_{dh}}$  are random variables in both cases.

## ESIMERKKI

**Ei-suunniteltu rakenne:** Odotettu otoskoko osajoukossa  $d$ , otanta-asetelmana SRSWOR:

$$E(n_{s_d}) = n \times (N_d / N)$$

**Suunniteltu rakenne:** Osajoukot on määritelty ositteiksi

Oletetaan että tulosmuuttujan  $y$  variaatiokertoimet  $C.V_{dy} = S_{dy} / \bar{Y}_d$  tunnetaan kaikissa osajoukoissa, missä  $S_{dy}$  ja  $\bar{Y}_d$  ovat perusjoukon keskihajonta ja keskiarvo domainissa  $d$

**Bankier-kiintiöinti:** Domain-otoskoot ovat

$$n_{d,pow} = n \times \frac{T_{dz}^a \times C.V_{dy}}{\sum_{d=1}^D T_{dz}^a \times C.V_{dy}}$$

Vakio  $a = 0$  tässä tapauksessa.

**Perusjoukko:** Occupational Health Care Survey (OHC),  $N = 7841$  henkilöä

**Parametrit:** Domain-totaalit

$$T_d = \sum_{k \in U_d} Y_k, \quad d = 1, \dots, D$$

Pitkäaikaisesti sairaiden lukumäärä osajoukoissa  $D = 30$  osajoukkoa

**Otos:** SRSWOR, otoskoko  $n = 392$

**Horvitz-Thompson-estimaattori:**

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k, \quad d = 1, \dots, D$$

missä  $w_k = 1 / \pi_k$

**Laatuindikaattori:**

Estimaattorin variaatiokerroin  
*coefficient of variation*

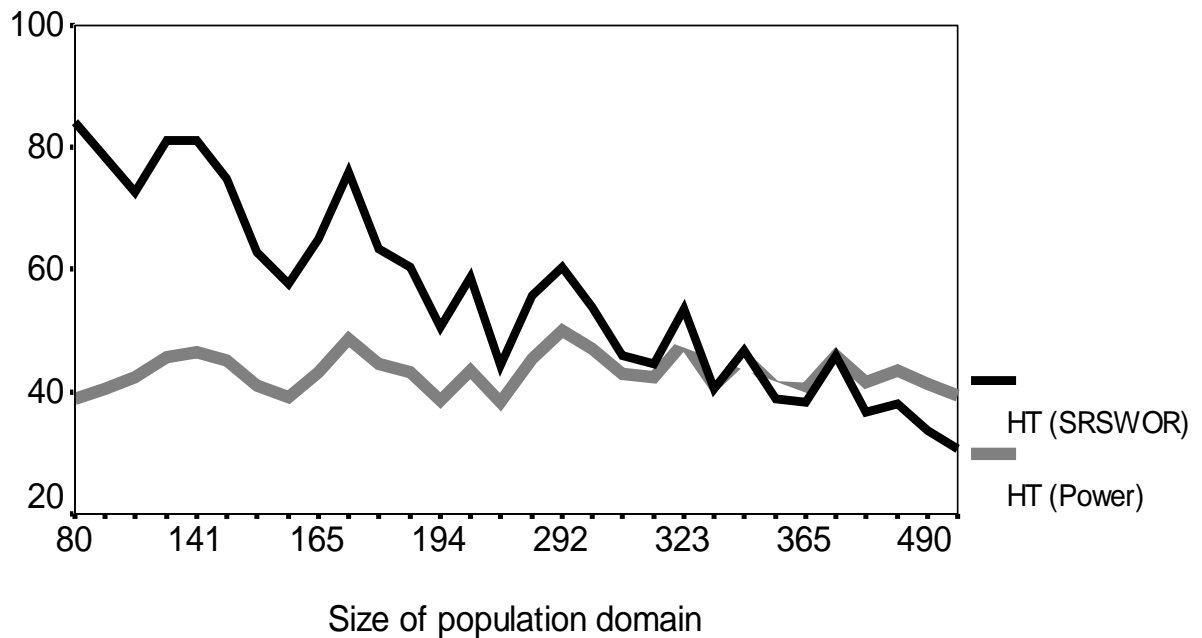
$$C.V(\hat{t}_{dHT}) = S.E(\hat{t}_{dHT}) / T_d$$

missä  $S.E(\hat{t}_{dHT})$  on HT-estimaattorin keskivirhe



**Table 5.** HT-estimaattoreiden CV (%) ei-suunnitellussa ja suunnitellussa domain-rakenteessa (Lehtonen and Pahkinen 2004).

Domain		Domain-otoskoko		HT-estimaattoreiden C.V (%)	
		<b>Ei-suunniteltu rakenne</b> SRSWOR $E(n_{s_d})$	<b>Suunniteltu rakenne</b> Bankier-kiintiöinti $n_d$	<b>Ei-suunniteltu rakenne</b> SRSWOR $C.V(\hat{i}_{dHT})$	<b>Suunniteltu rakenne</b> Bankier-kiintiöinti $C.V(\hat{i}_{dHT})$
$D$	$N_d$				
10	81	4	11	84.10	38.88
20	101	5	12	78.41	40.54
18	129	6	13	72.69	42.38
3	133	7	15	81.04	45.63
8	141	7	16	81.03	46.54
30	146	7	15	74.80	45.03
21	153	8	12	62.87	41.15
23	156	8	11	57.65	39.05
16	165	8	13	64.94	43.19
1	181	9	17	75.90	48.78
11	187	9	14	63.52	44.52
6	188	9	13	60.37	43.22
28	194	10	10	50.52	38.69
24	200	10	13	58.68	43.39
22	242	12	10	44.27	38.30
15	252	13	14	55.68	45.50
7	292	15	17	60.34	50.06
4	295	15	15	53.92	47.04
13	305	15	13	46.00	43.04
12	311	16	12	44.50	42.38
5	323	16	16	53.50	48.23
25	339	17	11	40.57	41.03
2	352	18	14	46.80	45.74
26	364	18	11	38.87	40.88
29	365	18	11	38.25	40.45
9	366	18	14	45.99	45.85
17	426	21	12	36.67	41.62
14	447	22	13	37.95	43.37
19	490	24	11	33.60	41.22
27	517	26	10	30.68	39.34
<b>Sum</b>	<b>7841</b>	<b>392</b>	<b>392</b>		



**Figure 1.** (Lehtonen and Pahkinen 2004)

Horvitz-Thompson-estimaattorin variaatiokerroin (%) SRSWOR-otannan tilanteessa (vastaa unplanned-rakennetta) ja ositetun SRSWOR-otannan tilanteessa (Bankier-kiintiöinti,  $\alpha = 0$ ) (vastaa planned-rakennetta).

[VLISS-demonstraatio](#)

## **BOX 1. Estimointiproseduurin operationaaliset vaiheet**

*Vaihe 1: Kehikkoperusjoukon konstruointi.* Muodostetaan  $N$  alkion perusjoukko  $U$ , joka sisältää seuraavat muuttujat:

ID-tieto, domain-indikaattorit, ositeindikaattorit, sisältymistodennäköisyydet  $n$  alkion otosta varten asetelmalla  $p(s)$ , ja apumuuttujavektorit kaikille alkioille  $k \in U$ .

*Vaihe 2: Otanta ja mittaus.* Poimitaan otos asetelmalla  $p(s)$  ja kerätään tiedot tulosmuuttujasta  $y$ . Muodostetaan otostiedosto  $s(y)$ , joka sisältää seuraavat muuttujat: ID-tieto, havaittu  $y$ -muuttujan arvo ja asetelmapainot kaikille alkioille  $k \in s$ .

*Vaihe 3: Yhdistetään  $U$  ja  $s(y)$ .* Muodostetaan yhdistetty tiedosto mikrolinkkaamalla (merge) avaimen ID avulla kehikkopj  $U$  ja otosaineisto  $s(y)$ .

*Vaihe 4: Mallin valinta ja mallin sovitus.* Mallin matemaattisen muodon valinta, parametrisointi ja sovittaminen otosaineistolle. Mallin diagnostiikka. Lasketaan sovitetun mallin avulla tulosmuuttujan  $y$  sovitteet kaikille  $p_j$ :n alkioille  $k \in U$  sekä residuaalit kaikille otosalkioille  $k \in s$ .

*Vaihe 5. Domain-estimaattoreiden valinta ja estimointi.* Käyttämällä sovitteita, residuaaleja ja asetelmapainoja lasketaan estimaatit kullekin osajoukolle  $d$ .

*Vaihe 6: Estimaattoreiden laatuindikaattorit.* Domain-estimaattoreiden varianssien, keskivirheiden ja variaatiokertoimien estimointi.

(Lehtonen and Pahkinen 2004)

**Table 6.** Vaiheiden 1, 3 ja 4 havainnollistaminen.

<b>Vaihe 1:</b> Kehikkoperusjoukon $U$ konstruointi					<b>Vaihe 3:</b> Yhdistetään $U$ ja $s(y)$			<b>Vaihe 4:</b> Lasketaan sovitteet ja residuaalit	
Alkio ID	Domain $\delta'_k$	Osite $\tau'_k$	$\pi_k$	Apu- muuttujat $\mathbf{z}'_k$	Asetelma- painot $w_k$	Otos- Indik. $I_k$	Tulos- muuttuja $y_k$	Sovitteet $\hat{y}_k$	Resi- duaalit $\hat{e}_k$
1	$\delta'_1$	$\tau'_1$	$\pi_1$	$\mathbf{z}'_1$	0	0	...	$\hat{y}_1$	...
2	$\delta'_2$	$\tau'_2$	$\pi_2$	$\mathbf{z}'_2$	0	0	...	$\hat{y}_2$	...
3	$\delta'_3$	$\tau'_3$	$\pi_3$	$\mathbf{z}'_3$	$w_3$	1	$y_3$	$\hat{y}_3$	$\hat{e}_3$
4	$\delta'_4$	$\tau'_4$	$\pi_4$	$\mathbf{z}'_4$	0	0	...	$\hat{y}_4$	...
5	$\delta'_5$	$\tau'_5$	$\pi_5$	$\mathbf{z}'_5$	$w_5$	1	$y_5$	$\hat{y}_5$	$\hat{e}_5$
⋮									
$k$	$\delta'_k$	$\tau'_k$	$\pi_k$	$\mathbf{z}'_k$	$w_k$	1	$y_k$	$\hat{y}_k$	$\hat{e}_k$
⋮									
$N$	$\delta'_N$	$\tau'_N$	$\pi_N$	$\mathbf{z}'_N$	0	0	...	$\hat{y}_N$	...
... Non-sampled element									

## HUOM:

Apumuuttujavektorit  $\mathbf{z}_k = (z_{1k}, \dots, z_{Jk})'$  oletetaan tunnetuksi kaikille  $p_j$ :n alkioille

Tällöin apumuuttujien totaalien vektori

$\mathbf{T}_z = (T_{z_1}, \dots, T_{z_J})'$  missä  $T_{z_j} = \sum_{k \in U} z_{jk}$ ,  $j = 1, \dots, J$ , on tunnettu

Koska domain-indikaattorit  $\delta_{dk}$  tunnetaan, voidaan laskea apumuuttujien domain-totaalit

$$T_{dz_j} = \sum_{k \in U_d} z_{jk}, \quad d = 1, \dots, D \text{ ja } j = 1, \dots, J$$

Mallin sovitusvaiheessa lasketaan sovitteet  $\hat{y}_k$  kaikille  $N$  alkioille  $k \in U$

Residuaalit  $\hat{e}_k = y_k - \hat{y}_k$  voidaan laskea vain otoshavainnoille  $k \in s$

Sovitteet  $\hat{y}_k$ ,  $k \in U$  vaihtelevat spesifioidusta mallista riippuen.

## DOMAIN-TOTAALIEN ESTIMAATTORIT

Osajoukkototaalien  $T_d = \sum_{k \in U_d} y_k$  estimaattoreiden päätyypit:

### Horvitz-Thompson –estimaattori HT

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k = \sum_{k \in s_d} y_k / \pi_k$$

### Synteettinen estimaattori SYN

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k \quad (1)$$

### Yleistetty regressioestimaattori GREG (*Generalized regression estimator*)

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k) \quad (2)$$

### Yhdistelmäestimaattori (*Composite estimator*)

$$\hat{t}_{dCOMP} = \sum_{U_d} \hat{y}_k + \hat{\gamma}_d \sum_{s_d} a_k (y_k - \hat{y}_k)$$

missä  $w_k = 1 / \pi_k$ ,  $s_d = s \cap U_d$  ja  $d = 1, \dots, D$

COMP-estimaattorissa  $\hat{\gamma}_d$  on domain-spesifi paino,  $0 \leq \hat{\gamma}_d \leq 1$ , jota tarvitaan erityisesti EBLUP-estimaattorin yhteydessä

# ESTIMAATTOREIDEN KONSTRUOINTI JA MALLIN SPESIFIOINTI

Työvaiheet:

- (1) Estimoidaan valitun mallin parametrit käyttämällä otosaineistoa  $s(y) = \{(y_k, \mathbf{z}_k); k \in s\}$ .
- (2) Mallin parametriestimaattien ja apumuuttujavektoreiden  $\mathbf{z}_k$  avulla lasketaan sovitteet  $\hat{y}_k$  kaikille perusjoukon alkioille  $k$  (otosalkiot ja otoksen ulkopuoliset alkiot)
- (3) Domain-totaalin  $T_d$  estimaattia  $\hat{t}_d$  varten domainissa  $d$  sijoitetaan sovitteet  $\{\hat{y}_k; k \in U\}$  ja otoshavainnot  $\{y_k; k \in s\}$  vastaaviin estimaattorikaavoihin (GREG, SYN, COMP tai EBLUP).

[GREG-estimaattorin](#) teoreettinen motivaatio:

*Population fit regression estimator*

(Lehtonen and Veijanen 2009)

## ESIMERKKI

### a) Kiinteiden tekijöiden lineaarinen malli:

$$y_k = \mathbf{z}'_k \boldsymbol{\beta} + \varepsilon_k$$

missä  $\boldsymbol{\beta}$  on mallin tuntematon parametrivektori ja residuaalit ovat  $\varepsilon_k$

Sovitetaan malli, saadaan estimaatti  $\hat{\boldsymbol{\beta}}$

Lasketaan sovitteet  $\hat{y}_k = \mathbf{z}'_k \hat{\boldsymbol{\beta}}$  kaikille  $k \in U$

### b) Lineaarinen sekamalli:

$$y_k = \mathbf{z}'_k (\boldsymbol{\beta} + \mathbf{u}_d) + \varepsilon_k$$

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

Estimoidaan mallin parametrit ja lasketaan sovitteet  $\hat{y}_k = \mathbf{z}'_k (\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d)$  kaikille  $k \in U$



## MALLIN SPESIFIOINTI

Olkoon  $(J+1)$ -dimensioinen apumuuttujavektori

$$\mathbf{z}_k = (1, z_{1k}, \dots, z_{jk}, \dots, z_{Jk})', \quad j = 1, \dots, J$$

Vektoria tarvitaan sovitteiden  $\hat{y}_k$ ,  $k \in U$  laskentaa varten

### (1) Kiinteiden tekijöiden P-mallit

Estimaattorit SYN-P ja GREG-P perustuvat **lineaariseen malliin**

$$y_k = \beta_0 + \beta_1 z_{1k} + \dots + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k \boldsymbol{\beta} + \varepsilon_k \quad (3)$$

$k \in U$ , missä  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_J)'$  on kiinteiden tekijöiden vektori joka on määritelty **koko populaatiolle**

Malli (3) on **kiinteiden tekijöiden P-malli**

## Mallin parametrien estimointi

Perusjoukon tasolla:

Vektorin  $\beta$  PNS-estimaattori:

$$\mathbf{B} = \left( \sum_{k \in U} \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in U} \mathbf{z}_k y_k \quad (4)$$

Käytettävissä otosaineisto:

Painotettu PNS (Weighted least-squares, WLS) – estimaattori parametrille (4) lasketaan käyttämällä otoshavaintoja:

$$\hat{\mathbf{b}} = \left( \sum_{k \in S} w_k \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in S} w_k \mathbf{z}_k y_k \quad (5)$$

missä  $w_k = 1/\pi_k$  on alkion  $k$  asetelmapaino

Sovitteet ovat:

$$\hat{y}_k = \mathbf{z}'_k \hat{\mathbf{b}}, \quad k \in U \quad (6)$$

## **HUOM: Epäsuora domain-estimaattori**

Kun käytetään P-mallia ositteelle  $d$ , myös muiden osajoukkojen  $y$ -arvot vaikuttavat osajoukon  $d$  totaaliestimaattoreihin SYN-P ja GREG-P sijoitettaviin sovitteisiin  $\hat{y}_k$

Tästä syystä kiinteiden tekijöiden P-malliin perustuvia estimaattoreita  $\hat{t}_{dSYN-P}$  ja  $\hat{t}_{dGREG-P}$  kutsutaan **epäsuoriksi** (*indirect*)

**(2) Kiinteiden tekijöiden D-mallit. Estimaattorit**  
 SYN-D ja GREG-D perustuvat samaan apumuuttujavektoriin  $\mathbf{z}_k$ , mutta malli määrittellään domainkohtaiseksi:

$$y_k = \mathbf{z}'_k \boldsymbol{\beta}_d + \varepsilon_k \quad (7)$$

$k \in U_d, d = 1, \dots, D$ , tai

$$y_k = \sum_{d=1}^D \delta_{dk} \mathbf{z}'_k \boldsymbol{\beta}_d + \varepsilon_k \quad (8)$$

$k \in U$ , missä  $\delta_{dk}$  on alkion  $k$  domain-indikaattori:  
 $\delta_{dk} = 1$  kun  $k \in U_d$ , nolla muulloin,  $d = 1, \dots, D$ , ja  
 $\boldsymbol{\beta}_d$  on domain-kohtainen parametrivektori

Malli (7) on **kiinteiden tekijöiden D-malli**

**PNS-estimaattori** parametrille  $\beta_d$ :

$$\mathbf{B}_d = \left( \sum_{k \in U_d} \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in U_d} \mathbf{z}_k y_k \quad (9)$$

$d = 1, \dots, D$

## Otosdataan perustuva WLS estimaattori:

$$\hat{\mathbf{b}}_d = \left( \sum_{k \in S_d} w_k \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in S_d} w_k \mathbf{z}_k y_k \quad (10)$$

$$d = 1, \dots, D$$

Sovitteet ovat:

$$\hat{y}_k = \mathbf{z}'_k \hat{\mathbf{b}}_d \quad (11)$$

$$k \in U_d; d = 1, \dots, D$$

Sijoittamalla sovitteet  $\hat{y}_k$  kaavoihin (1) ja (2) saadaan vastaavat estimaattorit SYN-D ja GREG-D

### **HUOM: Suora domain-estimaattori**

D-mallien sovituksessa kussakin domainissa käytetään vain kyseisen domainin y-arvoja

Vastaavia estimaattoreita  $\hat{t}_{dSYN-D}$  ja  $\hat{t}_{dGREG-D}$  kutsutaan **suoriksi** (*direct*)

## HUOM:

Estimaattorin (9) täydellisempi muoto on GLS-estimaattori (Generalized least squares)

$$\mathbf{B}_d = \left( \sum_{k \in U_d} \mathbf{z}_k \mathbf{z}'_k / c_k \right)^{-1} \sum_{k \in U_d} \mathbf{z}_k y_k / c_k$$

missä  $c_k$  on muotoa  $c_k = \boldsymbol{\lambda}' \mathbf{z}_k$  alkiolle  $k \in U$  ja  $(J+1)$ -vektori  $\boldsymbol{\lambda}$  ei riipu arvosta  $k$ .

Käytännössä asetetaan usein  $c_k = 1$  kaikille  $k$

Koska nyt  $c_k = \boldsymbol{\lambda}' \mathbf{z}_k = 1$ , seuraa siitä että GREG-estimaattorin jäännöstotaalin HT-estimaatti

$$\sum_{k \in s_d} w_k (y_k - \hat{y}_k) = 0$$

Tästä seuraa että SYN-D ja GREG-D ovat identtiset, eli

$$\hat{t}_{dSYN-D} = \hat{t}_{dGREG-D}$$

jokaiselle otokselle  $s$ , kun käytetään kiinteiden tekijöiden D-mallia

**(3) Sekamallit.** Estimaattorit MSYN-D ja MGREG-D perustuvat lineaariseen kaksitasomalliin (sekamalliin), jota kutsumme **lineaariseksi D-tyypin sekamalliksi**

Mallissa on kiinteitä tekijöitä ja domainkohtaisia satunnaisia tekijöitä:

$$\begin{aligned} y_k &= \beta_0 + u_{0d} + (\beta_1 + u_{1d})z_{1k} + \dots + (\beta_J + u_{Jd})z_{Jk} + \varepsilon_k \\ &= \mathbf{z}'_k (\boldsymbol{\beta} + \mathbf{u}_d) + \varepsilon_k \end{aligned} \quad (12)$$

$$k \in U_d, \quad d = 1, \dots, D$$

Kukin mallin termi voidaan ajatella populaatiotasoisena kiinteänä tekijänä ja domainkohtaisen satunnaistekijän summaksi:

$$\begin{aligned} &\beta_0 + u_{0d} \text{ vakiotermille (intercept)} \\ &\beta_j + u_{jd}, j = 1, \dots, J \text{ "kulmakertoimille" (slopes)} \end{aligned}$$

Termit  $\mathbf{u}_d = (u_{0d}, u_{1d}, \dots, u_{Jd})'$  edustavat poikkeamia mallin **kiinteän osan** parametreista

$$y_k = \beta_0 + \beta_1 z_{1k} + \dots + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k \boldsymbol{\beta} + \varepsilon_k \quad (13)$$

## HUOM:

Käytännössä vai osa termeistä määritellään satunnaisiksi, jolloin joillekin  $j$ ,  $u_{jd} = 0$  kaikissa domaineissa  $d$

Erikoistapaus, jota käytetään paljon käytännön sovelluksissa, on malli jossa on vain **domain-kohtaiset satunnaiset vakiotermit**  $u_{0d}$ :

$$y_k = (\beta_0 + u_{0d}) + \beta_1 z_{1k} + \dots + \beta_J z_{Jk} + \varepsilon_k$$

Sovitteet lasketaan kaavalla

$$\hat{y}_k = \mathbf{z}'_k (\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d) \quad (14)$$

Saadaan estimaattorit MSYN-D ja MGREG-D (Lehtonen and Veijanen 1999)

D-malli (12) voidaan sovittaa esimerkiksi estimoimalla varianssikomponentit suurimman uskottavuuden (ML) tai rajoitetulla suurimman uskottavuuden (*restricted maximum likelihood* REML) menetelmällä ja kiinteät tekijät GLS-menetelmällä ehdolla varianssikomponentit (esim. Goldstein 2003 tai McCulloch and Searle 2001).



**Yleistettyjen lineaaristen sekamallien GLMM**  
kehikossa voidaan kirjoittaa malli:

$$E_m(y_k | \mathbf{u}_d) = g(\mathbf{z}'_k (\boldsymbol{\beta} + \mathbf{u}_d))$$

**Erikoistapauksia:**

**Lineaarinen malli** (jatkuva tulosmuuttuja):

$$E_m(y_k | \mathbf{u}_d) = \mathbf{z}'_k (\boldsymbol{\beta} + \mathbf{u}_d)$$

**Multinomiaalinen logistinen sekamalli**  
(moniluokkainen tulosmuuttuja):

$$E_m(y_{ik} | \mathbf{u}_d) = \frac{\exp(\mathbf{z}'_k (\boldsymbol{\beta}_i + \mathbf{u}_{id}))}{1 + \sum_{r=2}^m \exp(\mathbf{z}'_k (\boldsymbol{\beta}_r + \mathbf{u}_{rd}))}$$

(Lehtonen, Särndal and Veijanen 2003,  
Lehtonen and Veijanen 2009)

## ESIMERKKI

Jatkuvatyypinen  $y$ , jonka totaali  $T_d$  estimoidaan  
domaineille  $U_d$ ,  $d = 1, \dots, D$

Oletetaan yksi jatkuvatyypinen apumuuttuja  $z$

Avustavat mallit:

### (1) Kiinteiden tekijöiden P-mallit

$y_k$ ,  $k \in U$ :

$$(1a) \quad y_k = \beta_0 + \varepsilon_k$$

$$(1b) \quad y_k = \beta_1 z_k + \varepsilon_k$$

$$(1c) \quad y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$$

### (2) Kiinteiden tekijöiden D-mallit

$y_k$ ,  $k \in U_d$ ,  $d = 1, \dots, D$ :

$$(2a) \quad y_k = \beta_{0d} + \varepsilon_k$$

$$(2b) \quad y_k = \beta_{1d} z_k + \varepsilon_k$$

$$(2c) \quad y_k = \beta_{0d} + \beta_{1d} z_k + \varepsilon_k$$

### (3) Sekamallit

$y_k$ ,  $k \in U_d$ ,  $d = 1, \dots, D$ :

$$(3a) \quad y_k = \beta_{0d} + \varepsilon_k = \beta_0 + u_{0d} + \varepsilon_k$$

$$(3b) \quad y_k = \beta_{0d} + \beta_{1d} z_k + \varepsilon_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$$

**HUOM:**

Mallit (1b) ja (2b):

**Suhdetehosteinen estimointi** (*Ratio estimation*)  
osajoukoille  $d$

Mallit (1c) ja (2c):

**Regressioestimointi** osajoukoille  $d$

**HUOM:**

Mallit (1) ja (3):

**Epäsuorat** (*Indirect*) estimaattorit SYN ja  
GREG

Malli (2):

**Suorat** (*Direct*) estimaattorit SYN ja GREG

## ESIMERKKI

P-malli (1b)

SYN-estimaattori (1) totaaleille  $T_d$ :

$$\begin{aligned}\hat{t}_{dSYN-P} &= \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \hat{b}_1 z_k \\ &= T_{dz} \hat{b}_1 = T_{dz} \times \hat{t}_{HT} / \hat{t}_{zHT}\end{aligned}\quad (18)$$

$d = 1, \dots, D$

Parametrin (*slope*)  $B_1$  estimaattori on:

$$\hat{b}_1 = \frac{\sum_{k \in S} w_k y_k}{\sum_{k \in S} w_k z_k} = \frac{\hat{t}_{HT}}{\hat{t}_{zHT}}$$

Onko tämä estimaattori suora (*direct*) vai epäsuora (*indirect*)?

Estimaattori (18) on **epäsuora**. Miksi?

Estimaattori  $\hat{t}_{dSYN-P}$  osajoukolle  $d$  käyttää y-muuttujan arvoja koko otoksesta ja pyrkii siten **lainaamaan voimaa** (*borrowing strength*) myös muista domaineista

**HUOM:****SYN-estimaattorin (18) harha**

Estimaattorin  $\hat{t}_{dSYN-P}$  harhaa approksimoidaan kaavalla

$$\text{BIAS}(\hat{t}_{dSYN-P}) = E(\hat{t}_{dSYN-P}) - T_d \doteq -T_{dz}(B_{1d} - B_1)$$

missä

$B_{1d} = \sum_{k \in U_d} y_k / \sum_{k \in U_d} z_k$  on domain-kohtainen parametri (*slope*),  $d = 1, \dots, D$

$B_1 = \sum_{k \in U} y_k / \sum_{k \in U} z_k$  on perusjoukkotasoinen parametri

Domainille  $d$  harha on pieni, jos perusjoukkotasoinen parametri  $B_1$  approksimoi hyvin osajoukkokohtaista parametria  $B_{1d}$

**Merkittävä harha seuraa jos ehto ei ole voimassa.**

Vastaava epäsuora **GREG**-estimaattori (2)  
domain-totaaleille  $T_d$ :

$$\begin{aligned}
 \hat{t}_{dGREG-P} &= \sum_{k \in U_d} \hat{y}_k + \sum_{k \in S_d} w_k (y_k - \hat{y}_k) \\
 &= \hat{t}_{dSYN-P} + \sum_{k \in S_d} w_k (y_k - \hat{b}_1 z_k) \\
 &= \hat{t}_{dHT} + \frac{\hat{t}_{HT}}{\hat{t}_{zHT}} (T_{dz} - \hat{t}_{dzHT}) \tag{19}
 \end{aligned}$$

**HUOM:**

Yritys “lainata voimaa” pätee myös tälle  
estimaattorille

**Suorat estimaattorit SYN ja GREG** tyyppiä (2b) käyttävät  $y$ -arvoja vain kyseisestä domainista

Korvataan  $\hat{b}_1$  kaavassa (18) domain-kohtaisella estimaattorilla  $\hat{b}_{1d}$ :

$$\hat{b}_{1d} = \frac{\sum_{k \in s_d} w_k y_k}{\sum_{k \in s_d} w_k z_k} = \frac{\hat{t}_{dHT}}{\hat{t}_{dzHT}}, \quad d = 1, \dots, D,$$

missä  $\hat{t}_{dHT}$  ja  $\hat{t}_{dzHT}$  ovat totaalien  $T_d$  ja  $T_{dz}$  domain-kohtaisia HT-estimaattoreita

**Suora estimaattori SYN**  $\hat{t}_{dSYN-D}$

$$\hat{t}_{dSYN-D} = \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \hat{b}_{1d} z_k = T_{dz} \times \hat{t}_{dHT} / \hat{t}_{dzHT},$$

$$d = 1, \dots, D. \tag{20}$$

Tässä tapauksessa suora GREG-estimaattori  $\hat{t}_{dGREG-D}$  on identtinen suoran SYN-estimaattorin kanssa, koska GREG-estimaattorin harhankorjaustermi on tällöin nolla.