Helsingin yliopisto

Otanta-aineistojen analyysi

Kevät 2010 Periodi III

Risto Lehtonen

Teema 2
Estimaattoreiden varianssien estimointi

Survey-analyysin lähestymistavat

Kuvaileva survey
Descriptive survey

Analyyttinen survey
Analytical survey

Asetelmaperusteinen survey-analyysi
Design-based survey analysis

Malliperusteinen / mallista riippuva survey-analyysi
Model-based survey analysis
Model-dependent survey analysis

Estimaattoreiden varianssin estimointi linearisointimenetelmällä ja pseudotoisto-otannan menetelmillä

Linearisointimenetelmä (TAYLOR)

Linearization method
Taylor series expansion

Pseudotoisto-otantaan perustuvat menetelmät

Pseudoreplication, Sample re-use

Jackknife-menetelmä (JACKKNIFE)
Balanced Repeated Replications (BRR)
Bootstrap-menetelmä (BOOT)

Ohjelmasovellukset

SAS: SURVEYMEANS, SURVEYREG, SURVEYFREQ, SURVEYLOGISTIC

SAS Vers. 9.1.3
Linearisointimenetelmä (TAYLOR)

SAS Vers. 9.2 (uusi versio)

TAYLOR
JACKKNIFE
BRR

LINEARISOINTIMENETELMÄ

Survey-analyysin ohjelmistot:

SAS
SURVEYMEANS, SURVEYREG,
SURVEYFREQ, SURVEYLOGISTIC

Epälineaariset estimaattorit

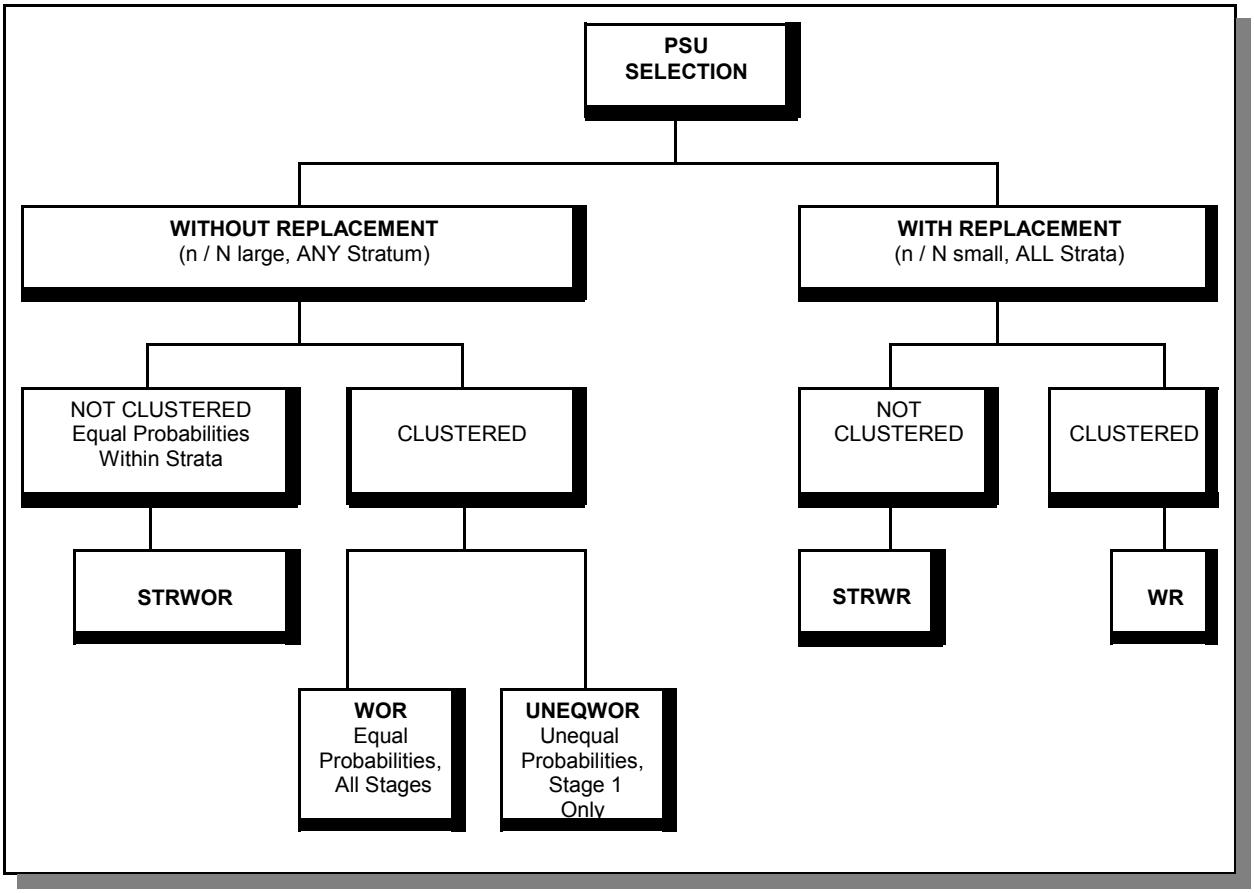
Osajoukon koko satunnaismuuttuja
- Osajoukkojen osuusestimaattorit
- Osajoukkojen keskiarvoestimaattorit

Regressiokertoimien estimaattorit
Logitmallin kerroinestimaattorit

HUOM:
Ohjelmasovelluksissa (SAS, SPSS, Stata) estimaattoreiden varianssien estimointi perustuu otosrypäiden välisen varianssin estimointiin ositteittain

Poikkeus: SUDAAN-ohjelmisto

Exhibit 3-1. Choosing the Taylor Series Design Option



* PERUSJOUKON OSAJOUKKOJA KOSKEVIEN OSUUKSIEN JA KESKIARVOJEN ESTIMOINTI

Perusjoukko jaettu D **osajoukkoon** $U_1,...,U_D$

Binäärinen (0/1) indikaattorimuuttuja δ

$\delta_{jhik}=1$ jos ositteen h rypään i alkio $k \in U_j$
$= 0$ muulloin

Binäärinen (0/1) tulosmuuttuja y

$y_{hik}=1$ jos ositteen h rypään i alkiolla k on tutkittava ominaisuus
$= 0$ muulloin

Estimoitavana osuusparametri

$$p_j = \frac{\sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{k=1}^{N_{hi}} \delta_{jhik} y_{hik}}{N_j} = \frac{T_j}{N_j} \qquad (j=1,...,D)$$

missä

H ositteiden lkm
M_h perusjoukon rypäiden lkm ositteessa h
N_{hi} perusjoukon alkioiden lkm ositteen h rypäässä i
T_j tulosmuuttujan y totaali osajoukossa j
N_j osajoukon alkioiden lkm

* SUHTEEN JA OSUUDEN ESTIMAATTORI
Combined ratio estimator

Osuusestimaattori $\hat{p}_j$, j=1,...,D (D osajoukkoa)

$$\hat{p}_j = \frac{y_j}{x_j} = \frac{\hat{t}_j}{\hat{N}_j} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h} y_{jhi}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h} x_{jhi}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} \delta_{jhik} w_{hik} y_{hik}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} w_{hik}}$$

missä $\hat{t}_j$ Tulosmuuttujan y totaaliestimaattori osajoukossa j

$\hat{N}_j$ Osajoukon koon estimaattori

$y_j=(n/\hat{N})\hat{t}_j$ ja $x_j=(n/\hat{N})\hat{N}_j$ vastaavat skaalatut luvut

w_{hik} Painomuuttuja

HUOM: Analyyttisissa tutkimuksissa painot w skaalataan usein niin, että niiden keskiarvo koko aineistossa = 1

Suhteen estimaattori on yksinkertaisin esimerkki epälineaarisesta estimaattorista

HUOM: Merkintätapa x_j (eikä n_j) korostaa sitä, että myös nimittäjä on satunnaismuuttuja

* LINEARISOINTIMENETELMÄ

Suhteen estimaattorissa sekä osoittaja y_j että nimittäjä x_j ovat satunnaismuuttujia

Tästä syystä asetelmaperusteisen varianssiestimaattorin tulee käsittää:

- osoittajan varianssi *v(y)*
- nimittäjän varianssi *v(x)*
- osoittajan ja nimittäjän kovarianssi *cov(y,x)*

Osajoukon osuustunnusluvun $\hat{p}_j$ linearisointimenetelmään perustuva varianssiestimaattori on:

$$\hat{v}_{des}(\hat{p}_j)=\hat{p}_j^2(y_j^{-2}\hat{v}(y_j)+x_j^{-2}\hat{v}(x_j)-2(y_jx_j)^{-1}\hat{cov}(y_j,x_j))$$

HUOM: Vastaava malliperusteinen (binominen, SRS-perusteinen) varianssiestimaattori:

$$\hat{v}_{bin}(\hat{p}_j)=\hat{p}_j(1-\hat{p}_j)/\hat{n}_j$$

ESIMERKKI. Osuusestimaattorin varianssin approksimointi linearisointimenetelmällä
Lehtonen&Pahkinen 2004, Example 5.5

OHC Survey demodata

Ositettu ryväsotanta-asetelma

H= 5 ositetta
m= 250 toimipaikkaa (otosryvästä)
n = 7841 henkilöä

Binäärinen tulosmuuttuja
PHYS Työn fysikaaliset terveyshaitat
0 = Ei ole
1 = On

Estimointi:

Työn fysikaalisista haitoista kärsivien miesten osuus

Osuuden estimaatti:

$$\hat{p}_1 = \frac{y_1}{x_1} = \frac{2061}{4485} = 0.4595$$

Varianssiapproksimaatio:

SAS / SURVEYMEANS

Osuusestimaattorin asetelmaperusteinen varianssiestimaatti linearisointimenetelmän avulla:

$$\hat{v}_{des}(\hat{p}_1)=\hat{p}_1^2(y_1^{-2}\hat{v}(y_1)+x_1^{-2}\hat{v}(x_1)-2(y_1x_1)^{-1}\hat{cov}(y_1,x_1))=0.2775\times10^{-3}$$

SRS-perusteinen (binomimalliin perustuva) varianssiestimaatti:

$$\hat{v}_{bin}(\hat{p}_1)=\hat{p}_1(1-\hat{p}_1)/\hat{n}_1=0.4595(1-0.4595)/4485=0.554\times10^{-4}$$

Estimoitu asetelmakerroin:

$$deff(\hat{p}_1) = 0.0002775/0.0000554 = 5.01$$

Suuri deff-estimaatti viittaa tulosmuuttujan PHYS voimakkaaseen positiiviseen sisäkorrelaatioon rypäissä

Binominen varianssiestimaatti aliestimoi selvästi todellista varianssia