HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi

Kevät 2012

TEEMA 5:

Tilastollinen mallinnus

Asetelmaperusteisten ja malliperusteisten menetelmien vertailu, ohjelmasovelluksia, PISA-esimerkki

Risto Lehtonen
risto.lehtonen@helsinki.fi

Korreloituneiden havaintojen analyysi

- Lineaariset mallit *Linear models*
  - Estimointi: LS, WLS, GEE
- Yleistetyt lineaariset mallit *Generalized linear models*
  - Estimointi: ML, PML, GEE
- Yleistetyt lineaariset sekamallit *Generalized linear mixed models GLMM*
  - Monitasomallit - *Multilevel models*
  - Hierarkkiset mallit - *Hierarchical models*
  - Estimointi: GLS ja ML, REML,…
- YHTEENVETOTAULUKKO

Yleistetty lineaarinen sekamalli GLMM

Malli:

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

missä $f(.)$ linkkifunktio, esimerkiksi

- lineaarinen sekamalli
- logistinen sekamalli

$\mathbf{x}_k = (1, x_{1k}, ..., x_{pk})'$ selittävien muuttujien vektori

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ kiinteät (fixed) parametrit

$\mathbf{u}_d = (u_{0d}, ..., u_{pd})'$ satunnaistermit (random effects)

Estimointi: SAS GLIMMIX

Lineaarinen kiinteiden tekijöiden malli

Malli

$$E_m(y_k) = \mathbf{x}_k'\boldsymbol{\beta}$$

missä

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$

β_j mallin kiinteät parametrit, $j = 0, \ldots, p$

Esim: $y_k = \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k$

Estimointi: SAS SURVEYREG (WLS)

Lineaarinen sekamalli

Malli

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

missä

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ kiinteät parametrit

$\mathbf{u}_d = (u_{0d}, \ldots, u_{pd})'$ satunnaistermit (random effects)

Esim: $y_k = \beta_0 + u_{0d} + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k$

Estimointi: SAS MIXED (GLS ja ML tai REML)

Logistinen kiinteiden tekijöiden malli

Malli

$$E_m(y_k) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta})}$$

missä y_k on binäärinen

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$

β_j mallin kiinteät parametrit, $j = 0, \ldots, p$

Estimointi: SAS SURVEYLOGISTIC (PML)

Logistinen sekamalli

Malli

$$E_m(y_k \,|\, \mathbf{u}_d) = \frac{\exp(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d))}{1 + \exp(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d))}$$

missä

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ kiinteät parametrit

$\mathbf{u}_d = (u_{0d}, \ldots, u_{pd})'$ satunnaistermit (random effects)

Estimointi: SAS GLIMMIX (ML)

SAS - Lineaariset ja yleistetyt lineaariset mallit ja sekamallit

- **Asetelmaperusteiset proseduurit**
  - Otanta-asetelma otetaan huomioon
  - Ositus, ryvästyminen (korreloituneet havainnot), painot
  - SURVEYREG
  - SURVEYLOGISTIC
  - SURVEYPHREG
- **Malliperusteiset proseduurit**
  - Korreloituneet havainnot, ryväsotanta
    - GENMOD
    - MIXED
    - GLIMMIX
  - Korreloimattomat havainnot, SRS-oletus
    - REG
    - LOGISTIC

SAS-sovellukset – Korreloituneiden havaintojen analyysimenetelmät 1a

- Asetelmaperusteiset proseduurit
- PROC SURVEYREG
  - **Lineaariset kiinteiden tekijöiden mallit**
  - Jatkuva tulosmuuttuja
  - Regressiomalli, ANOVA, ANCOVA
  - Reagoidaan otanta-asetelman ominaisuuksiin
    - Ositus: STRATA-lause
    - Ryvästyminen (sisäkorrelaatio): CLUSTER-lause
    - Analyysipainot: WEIGHT-lause

- PROC SURVEYREG

(1) Asetelmaperusteinen analyysi

PROC SURVEYREG

Lineaarinen ANCOVA-malli

Päävaikutusmalli

Jatkuvat selittäjät: age, phys, chron

Diskreetti selittäjä: sex

```
proc surveyreg data=ohc;
   class sex;
   model psych=sex age phys chron
   / deff solution;
   strata osite;
   cluster ryvas;
```

PROC SURVEYREG
Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Design Effect
Intercept	-0.1300951	0.05265252	-2.47	0.0142	1.51
SEX 1	-0.2816240	0.02924999	-9.63	<.0001	1.61
SEX 2	0.0000000	0.00000000	.	.	.
AGE	0.0031016	0.00130925	2.37	0.0186	1.49
PHYS	0.1730267	0.02898974	5.97	<.0001	1.45
CHRON	0.3925469	0.02939601	13.35	<.0001	1.36

NOTE: The denominator degrees of freedom for the t tests is 245.

SAS-sovellukset – Korreloituneiden havaintojen analyysimenetelmät 1b

- Malliperusteiset proseduurit
- PROC MIXED
  - **Lineaariset sekamallit**
  - Jatkuva tulosmuuttuja
  - Regressiomalli, ANOVA, ANCOVA
  - Reagoidaan otanta-asetelman ominaisuuksiin
    - Ryvästyminen (sisäkorrelaatio): RANDOM-lause tai REPEATED-lause
    - Analyysipainot: WEIGHT-lause

- PROC MIXED

MIXED: Kiinteiden vaikutusten kovarianssimatriisin estimointi

EMPIRICAL

computes the estimated variance-covariance matrix of the fixed-effects parameters by using the asymptotically consistent estimator described in Huber (1967), White (1980), Liang and Zeger (1986), and Diggle, Liang, and Zeger (1994). This estimator is commonly referred to as the "sandwich" estimator, and it is computed as follows:

$$(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-}\left(\sum_{i=1}^{S}\mathbf{X}_i'\widehat{\mathbf{V}}_i^{-1}\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i'\widehat{\mathbf{V}}_i^{-1}\mathbf{X}_i\right)(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-}$$

Here, $\widehat{\boldsymbol{\epsilon}}_i = y_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}$, S is the number of subjects, and matrices with an i subscript are those for the ith subject. You must include the SUBJECT= option in either a RANDOM or REPEATED statement for this option to take effect

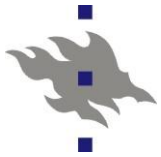
(2a) Malliperusteinen analyysi

PROC MIXED, RANDOM-lause

Lineaarinen ANCOVA-malli

Päävaikutusmalli

```
proc mixed data=ohc empirical
   method=reml;
   class sex ryvas;
   model psych=sex age phys chron
      / solution;
   random intercept / subject=ryvas
      type=vc;
*vc: variance components model;
```



(2b) Malliperusteinen analyysi

PROC MIXED, REPEATED-lause

Lineaarinen ANCOVA-malli

Päävaikutusmalli

```
proc mixed data=ohc empirical
   method=reml;
   class sex ryvas;
   model psych=sex age phys chron
      / solution;
   repeated / subject=ryvas
   type=vc;
```

PROC MIXED

Effect	Gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		-0.1301	0.05245	249	-2.48	0.0138
SEX	1	-0.2816	0.02923	235	-9.63	<.0001
SEX	2	0	.	.	.	.
AGE		0.003102	0.001306	7587	2.38	0.0176
PHYS		0.1730	0.02922	7587	5.92	<.0001
CHRON		0.3925	0.02920	7587	13.44	<.000

SAS-sovellukset – Korreloituneiden havaintojen analyysimenetelmät 2a

- Asetelmaperusteiset proseduurit
- PROC SURVEYLOGISTIC
  - **Logistiset kiinteiden tekijöiden mallit**
  - Binäärinen tai moniluokkainen tulosmuuttuja
  - Logistinen regressiomalli, ANOVA, ANCOVA
  - Reagoidaan otanta-asetelman ominaisuuksiin
    - Ositus: STRATA-lause
    - Ryvästyminen (sisäkorrelaatio): CLUSTER-lause
    - Analyysipainot: WEIGHT-lause

- PROC SURVEYLOGISTIC

PROC SURVEYLOGISTIC < options >**;**

BY variables **;**

CLASS variable <(v-options)> ... >**;**

CLUSTER variables **;**

CONTRAST 'label' effect values <,... /options >**;**

FREQ variable **;**

MODEL events/trials = < effects > < / options >**;**

MODEL variable < (variable_options) > = < effects > < / options >**;**

STRATA variables < / options > **;** < *label:* >

TEST equation1 < , ... , < equationk >> < /option >**;**

UNITS independent1 = list1 < ... /option > **;**

WEIGHT variable </ option >**;**

(1)Asetelmaperusteinen analyysi

PROC SURVEYLOGISTIC

Logistinen ANCOVA-malli

Päävaikutusmalli

```
proc surveylogistic data=ohc;
   class sex(ref=first);
   model psych2(ref=last)=
   sex age phys chron / link=logit;
   strata osite;
   cluster ryvas;
```

PROC SURVEYLOGISTIC

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.3230	0.1012	10.1921	0.0014
SEX	2	1	0.2494	0.0294	72.0593	<.0001
AGE		1	0.00268	0.00260	1.0612	0.3029
PHYS		1	0.2665	0.0591	20.3611	<.0001
CHRON		1	0.5652	0.0574	96.8233	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
SEX 2 vs 1	1.647	1.467	1.848
AGE	1.003	0.998	1.008
PHYS	1.305	1.163	1.466
CHRON	1.760	1.572	1.970

SAS-sovellukset – Korreloituneiden havaintojen analyysimenetelmät 2b

- Malliperusteiset proseduurit
- PROC GENMOD
  - Yleistetyt lineaariset mallit
  - **Logistiset kiinteiden tekijöiden mallit**
  - **Yleistetyt estimointiyhtälöt**
    - GEE - Generalized estimating equations
    - WGEE – Weighted GEE
    - TYPE=IND tai TYPE=EXCH (exchangeable)
  - Reagoidaan otanta-asetelman ominaisuuksiin
    - Ryvästyminen (sisäkorrelaatio): REPEATED-lause
    - Analyysipainot: WEIGHT-lause
- PROC GENMOD

(2)Malliperusteinen analyysi

PROC GENMOD

Logistinen ANCOVA-malli

Yhdysvaikutusmalli

```
proc genmod data=ohc descending;
  class sex(ref=first) ryvas;
  model psych2=sex age phys chron
    sex*age /
    dist=bin link=logit;
  repeated subject=ryvas /
    type=exch;
* exch = exchangeable correlation
  structure;
```

PROC GENMOD

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		0.2258	0.1522	-0.0724	0.5240	1.48	0.1378
SEX	1	-1.0252	0.1993	-1.4159	-0.6345	-5.14	<.0001
SEX	2	0.0000	0.0000	0.0000	0.0000	.	.
AGE		-0.0055	0.0039	-0.0132	0.0021	-1.41	0.1579
PHYS		0.2983	0.0593	0.1820	0.4145	5.03	<.0001
CHRON		0.5575	0.0568	0.4461	0.6688	9.81	<.0001
AGE*SEX	1	0.0142	0.0050	0.0045	0.0239	2.86	0.0043
AGE*SEX	2	0.0000	0.0000	0.0000	0.0000	.	.

Exchangeable Working Correlation
Correlation 0.0156016243

SAS-sovellukset – Korreloituneiden havaintojen analyysimenetelmät 2c

- Malliperusteiset proseduurit
- PROC GLIMMIX
  - Yleistetyt lineaariset sekamallit GLMM
  - **Logistiset sekamallit**
  - Reagoidaan otanta-asetelman ominaisuuksiin
    - Ryvästyminen (sisäkorrelaatio): RANDOM-lause
    - Analyysipainot: WEIGHT-lause

- PROC GLIMMIX

(3)Malliperusteinen analyysi

PROC GLIMMIX

Logistinen ANCOVA-malli

Yhdysvaikutusmalli

```
proc glimmix data=ohc empirical;
  model psych2=sex age phys chron
    sex*age  / dist=bin link=logit
    solution;
  random int / subject=ryvas
    type=vc;
* empirical: cvastaava kuin MIXED;
```

PROC GLIMMIX

Effect	Gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		0.2292	0.1531	249	1.50	0.1355
SEX	1	-1.0334	0.2007	7586	-5.15	<.0001
SEX	2	0	.	.	.	.
AGE		-0.00565	0.003946	7586	-1.43	0.1521
PHYS		0.3025	0.05966	7586	5.07	<.0001
CHRON		0.5609	0.05717	7586	9.81	<.0001
AGE*SEX	1	0.01437	0.005002	7586	2.87	0.0041
AGE*SEX	2	0	.	.	.	.

Case: PISA 2000

- Lehtonen R. and Pahkinen E. Practical Methods for Design and Analysis of Complex Surveys. Second Edition. Chichester: John Wiley & Sons
  - **Section 9.4.**
  - **MULTILEVEL MODELLING IN EDUCATIONAL SURVEY**
  - **Koulusaavutusaineiston monitasomallinnus**

PISA 2000

Programme for International Student Assessment

- Tiedonkeruu vuonna 2000
  - 32 maata
- Aihepiirit
  - **Lukeminen (reading literacy)**
  - Matematiikka
  - Luonnontieteet
- **Maat tässä esimerkissä**
  - Brazil, Finland, Germany, Hungary, Republic of Korea, United Kingdom, and United States

- Aineiston hierarkinen rakenne maittain
  - Taso 1: Oppilas
  - Taso 2: Koulu
- Tyypillinen otanta-asetelma
  - Ositettu kaksiasteinen ryväsotanta
  - Rypäänä koulu
  - Koulujen poiminta
    - Systemaattinen PPS-otanta (*Sampling with probabilities proportional to size*)

PISA – Analyysistrategia

Malliperusteinen analyysi *(Model-based)*

- Aineiston hierarkkisen rakenteen mallintaminen
- **Sekamallit** (*mixed models*)
- **Monitasomallit** (*multilevel models*)
- Painotus (*analysis weights*)
- Ryvästyminen, rypäiden sisäkorrelaatio (*clustering effect*)

- **Laskentatyökalut**
  - **SAS-proseduurit**
    - MIXED
    - GLIMMIX
    - NLMIXED

- **MLwiN** (Harvey Goldstein)

http://www.mlwin.com/

- **HML** Hierarchical Linear and Nonlinear Modeling

http://www.ssicentral.com/hlm/

Miksi monitasomalli (sekamalli)?

- Perusjoukko on hierarkkisesti rakentunut
  - Koulutaso
  - Oppilastaso koulujen sisällä
- Otanta-asetelmana ryväsotanta
  - Poimitaan ensin otos kouluista
  - Otoskouluista poimitaan oppilasotokset
- Ryvästymisen aiheuttaman sisäkorrelaation hallinta sekamallin avulla
- Vaihtoehto: Asetelmaperusteinen analyysi
  - Tulokset halutaan yleistää kaikkiin kouluihin
  - Kiinteiden vaikutusten malli (jossa kouluefekti on ”fixed effect”) olisi perusteltu, jos yleistettäisiin vain otoskouluihin!

PISA 2000 – Painotus

- **Painotus**
  - Alkiotason asetelma-painon konstruointi
    - Koulun sisältymistn
    - Oppilaan sisältymistn
    - Vastauskadon adjustointi
    - Maakohtaiset erityispiirteet
- Indeksointi
  - Koulu *i*
  - Oppilas *k*

- **Painojen uudelleenskaalaus maittain**
  - Analyysipaino
  - Painojen summa = *n* (aineiston maakohtainen koko)
  - Painojen keskiarvo = 1
  - Yksityiskohdat: OECD (2002b)

Weighting procedure (design weight, asetelmapaino)

Weight w_{ik} for student k in school i:

$$w_{ik} = w_{1i} \times w_{2ik} \times f_i, \quad i = 1,...,m \text{ and } k = 1,...,n_i,$$

where

$w_{1i} = 1/(\pi_i \hat{\theta}_i)$ is the reciprocal of the product of the inclusion probability π_i and the estimated participation probability $\hat{\theta}_i$ of school i;

$w_{2ik} = 1/(\pi_{k|i} \hat{\theta}_{k|i})$ is the reciprocal of the product of the conditional inclusion probability $\pi_{k|i}$ and estimated conditional response probability $\hat{\theta}_{k|i}$ of student k from within the selected school i;

f_i is an adjustment factor for school i to compensate any country-specific refinements in the survey design, and

m is the number of sample schools in a given country and n_i is the number of sample students in school i.

PISA 2000 – Tulosmuuttuja

- Tulosmuuttuja *y*
  - *Student's combined reading literacy score*
  - Oppilaan lukemisen osaamista kuvaava kokonaispistemäärä
  - Yhdistelmämuuttuja
    - Konstruoitu viiden lukemisen osaamista kuvaavan muuttujan avulla
- Tulosmuuttujan skaalaus:
  - Keskiarvo yli osallistuneiden OECD maiden = 500
  - Keskihajonta = 100
- Minimi 402 (Brazil)
- Maksimi 550 (Finland)

PISA 2000 – Kuvailua

Table 9.8 Descriptive statistics for combined reading literacy score in the PISA 2000 Survey by country (in alphabetical order).

Country	Combined reading literacy score						
	Mean	Standard error	Overall design effect (1)	Design effect accounting for stratification and clustering (2)	Effective sample size of students	Number of observations in data set	
						Students	Schools
Brazil	402.9	3.82	8.33	5.17	476	3961	290
Finland	550.7	2.15	2.79	2.74	1600	4465	147
Germany	497.4	5.68	13.47	11.68	305	4108	183
Hungary	485.7	6.02	20.00	16.20	231	4613	184
Republic of Korea	526.6	3.66	12.99	11.67	351	4564	144
United Kingdom	531.4	4.08	14.08	7.16	564	7935	328
United States	517.0	5.16	6.93	5.46	354	2455	112

Data source: OECD PISA database, 2001.

PISA 2000 – Asetelmakertoimet

- *Overall design effect (1)*
  - Mittaa
    - Osituksen
    - Ryvästymisen
    - Painotuksen

    vaikutusta keskiarvon varianssiestimaattiin
    - SRS-varianssiestimaatti lasketaan **painottamattomalle** keskiarvolle

- *Deff accounting for stratification and clustering (2)*
  - Mittaa osituksen ja ryvästymisen vaikutusta keskiarvon varianssiestimaattiin
  - Painotuksen vaikutus on puhdistettu
    - SRS-varianssiestimaatti lasketaan **painotetulle** keskiarvolle

Asetelmakerroin *Deff*

Asetelmakerroin (Design effect, deff, Kish 1965) mittaa otanta-asetelman ryvästymisen vaikutusta estimaattorin keskivirheeseen

Keskiarvon **estimoitu asetelmakerroin** (1) (*overall deff*) on muotoa:

$$deff(\bar{y}^*) = \frac{\hat{v}(\bar{y}^*)}{\hat{v}_{srs}(\bar{y})}$$

missä
$\bar{y}^*$ on painotettu keskiarvo ja $\bar{y}$ on vastaava painottamaton keskiarvo
Osoittajassa oleva keskiarvon varianssiestimaattori on käytetyn otanta-asetelman mukainen (ositettu ryväsotanta)
Nimittäjässä on SRS-perusteinen varianssilauseke

Asetelmakerroin (2) on

$$deff(\bar{y}^*) = \frac{\hat{v}(\bar{y}^*)}{\hat{v}_{srs}(\bar{y}^*)}$$

PISA 2000 – Tehokas otoskoko

- *Effective sample size*
  Tehokas otoskoko
- Tehokas otoskoko = Alkuperäinen oppilastason otoskoko jaettuna asetelmakertoimella
- Tehokas otoskoko ilmaisee SRS-otoskoon, jolla saadaan sama estimointitarkkuus (keskivirhe) kuin käytetyn ryväsotanta-asetelman mukaisella oppilastason otoskoolla

- Esim: Hungary

$$n_{eff} = \frac{n}{deff} = \frac{4613}{20.00} = 231$$

- Voimakas sisäkorreloituneisuus pienentää paljon tehokasta otoskokoa!

PISA – Kaksitasoinen hierarkinen lineaarinen malli

Fitting a Two-Level Hierarchical Linear Model

- Tulosmuuttuja *y*: Combined scaled reading literacy score
- Selittäjät
  - Koulutaso
    - School size (SSIZE)
    - Teacher autonomy (AUTONOMY)
  - Standardointi

    Keskiarvo (yli maiden) = 0

    Varianssi = 1
- Oppilastaso
  - FEMALE (1 is for females and 0 is for males)
  - Socio-economic background (SEB)
  - Engagement in reading (ENGAGEMENT)
  - Achievement press (ACHPRESS)
- Standardointi

  Keskiarvo (yli maiden) = 0

  Varianssi = 1

.

PISA – Lineaarinen kaksitasomalli

$$y_{ik} = \text{INTERCEPT} + \gamma_1 \times \text{SSIZE}_i + \gamma_2 \times \text{AUTONOMY}_i + \beta_1 \times \text{FEMALE}_{ik} + \beta_2 \times \text{SEB}_{ik} + \beta_3 \times \text{ENGAGEMENT}_{ik} + \beta_4 \times \text{ACHPRESS}_{ik} + u_i + e_{ik}$$

Indeksi k : Tason 1 alkiot (oppilaat)
Indeksi i: Tason 2 alkiot (koulut)

Kiinteät vaikutukset γ ja β:
Regressiokertoimet koulu- ja oppilastasolla

Satunnaistermit:
u_i : Koulutason satunnaistermi (*random intercept*)
Jakaumaoletus: Normaalijakauma, keskiarvo 0 ja varianssi σ_u^2

e_{ik} : Oppilastason satunnaistermi
Jakaumaoletus: Normaalijakauma, keskiarvo 0 ja varianssi σ_e^2

Satunnaistermit u_i ja e_{ik} oletetaan riippumattomiksi
Analyysissa käytetään oppilastason painoja w_{ik}

PISA – Sisäkorrelaatio

Sisäkorrelaatio (*intra-cluster correlation*)

Skinner et al. (1989), Goldstein (2003), Snijders & Bosker (2002)

$$\hat{\rho}_{\text{int}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}^2}$$

Estimoitu tulosmuuttujan kokonaisvarianssi $\hat{\sigma}^2$ on jaettu kahteen komponenttiin:

Koulujen välinen (*between-school*) varianssi $\hat{\sigma}_u^2$

Koulujen sisäinen (*within-school*) varianssi $\hat{\sigma}_e^2$

Sisäkorrelaatio mittaa pareittaista korrelaatiota samaan rypääseen (kouluun) kuuluvien oppilaiden välillä

PISA – Lineaarinen kaksitasomalli

Nollamalli (a) Taulukko 9.9

$$y_{ik} = \text{INTERCEPT} + u_i + e_{ik}$$

Selittäviä muuttujia sisältävä malli (b) Taulukko 9.10

$$\begin{aligned} y_{ik} = {} & \text{INTERCEPT} + \gamma_1 \times \text{SSIZE}_i + \gamma_2 \times \text{AUTONOMY}_i \\ & + \beta_1 \times \text{FEMALE}_{ik} + \beta_2 \times \text{SEB}_{ik} + \beta_3 \times \text{ENGAGEMENT}_{ik} \\ & + \beta_4 \times \text{ACHPRESS}_{ik} + u_i + e_{ik} \end{aligned}$$

Indeksi k : Tason 1 alkiot (oppilaat)
Indeksi i: Tason 2 alkiot (koulut)

PISA – Sisäkorrelaatio mallille (a)

Esimerkki: Sisäkorrelaatio

(a) Nollamallista (*multilevel model with only intercept and residuals at both levels*) estimoitu sisäkorrelaatio (Hungary in Table 9.9)

$$\hat{\rho}_{\text{int}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}^2} = \frac{6093.7}{6093.7+3148.3} = 0.659$$

PISA – Sisäkorrelaatio mallille (b)

(b) Selittäviä muuttuja sisältävästä mallista estimoitu sisäkorrelaatio
Residual intra-school correlation coefficient
(Hungary in Table 9.10)

$$\hat{\rho}_{\text{int}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}^2} = \frac{4744.2}{4744.2+2897.4} = 0.621$$

Table 9.9 Estimates of two-level variance component models (null models) for combined reading literacy score in the PISA 2000 Survey by country (ordered by the size of the estimated intra-school correlation coefficient). MALLI (a)

Country	Intra-school correlation coefficient	Variance components		Intercept	Standard error
		School level	Student level		
Hungary	0.659	6093.7	3148.3	464.1	5.84
Germany	0.553	5572.2	4507.8	496.1	5.61
Brazil	0.428	3146.9	4201.4	387.9	3.61
Republic of Korea	0.375	1828.6	3043.0	520.9	3.74
United States	0.241	2318.2	7315.5	503.3	4.97
United Kingdom	0.212	1917.5	7126.5	529.0	2.88
Finland	0.063	470.7	6960.9	550.6	2.18

Data source: OECD PISA database, 2001.

Table 9.10 Estimates of two-level models for combined reading literacy score in the PISA 2000 Survey by country. MALLI (b)

		Hungary	Germany	Brazil	Republic of Korea	United States	United Kingdom	Finland
Fixed effects: Coefficient								
Intercept	γ_0	471.2	496.4	382.0	506.8	496.6	524.9	531.6
	s.e.	6.36	4.58	4.56	6.29	6.05	3.38	4.91
	t-ratio	74.14	108.37	83.75	80.53	82.12	155.06	108.27
	p-value	.000	.000	.000	.000	.000	.000	.000
School level variables:								
School size	γ_1	30.6	27.4	2.4	7.1	1.0	3.8	5.9
	s.e.	9.00	9.22	1.47	3.44	2.54	3.14	7.35
	t-ratio	3.41	2.97	1.64	2.07	0.38	1.20	0.80
	p-value	.001	.003	.100	.039	.705	.232	.426
Teacher autonomy	γ_2	4.8	-7.1	-3.1	2.5	4.1	-2.3	2.8
	s.e.	5.62	5.22	4.24	5.39	3.63	2.61	2.68
	t-ratio	0.86	-1.37	-0.74	0.47	1.14	-0.89	1.06
	p-value	.392	.171	.459	.641	.256	.374	.291
Student level variables:								
Female	β_1	6.4	3.6	3.1	15.9	14.9	9.8	19.6
	s.e.	2.22	2.41	2.54	2.49	3.71	2.64	2.43
	t-ratio	2.89	1.50	1.21	6.38	4.00	3.71	8.09
	p-value	.004	.133	.228	.000	.000	.000	.000
Socio-economic background	β_2	6.0	11.5	9.9	2.2	16.7	23.3	15.8
	s.e.	1.09	1.53	1.35	0.92	2.22	1.32	1.34
	t-ratio	5.56	7.50	7.34	2.40	7.51	17.70	11.78
	p-value	.000	.000	.000	.016	.000	.000	.000
Engagement in reading	β_3	19.5	19.0	19.5	16.6	28.9	31.5	33.9
	s.e.	1.04	0.98	1.51	1.04	1.99	1.40	1.26
	t-ratio	18.68	19.36	12.87	15.94	14.49	22.59	27.05
	p-value	.000	.000	.000	.000	.000	.000	.000
Achievement press	β_4	0.9	-1.6	3.4	3.4	-3.3	-7.2	-3.7
	s.e.	0.93	1.16	1.44	0.89	2.04	1.59	1.40
	t-ratio	0.92	-1.35	2.36	3.85	-1.62	-4.52	-2.65
	p-value	.356	.176	.018	.000	.106	.000	.008

Data source: OECD PISA database, 2001.

- Lehtonen-Pahkinen (2004)
- Section 9.4:
- **MULTI-LEVEL MODELLING IN AN EDUCATIONAL SURVEY**

- Table 9.10

	Hungary	Germany	Brazil	Republic of Korea	United States	United Kingdom	Finland
Random effects: Variance component							
School level	4744.2	3501.6	2730.5	1387.3	1770.6	999.6	394.8
Student level	2897.4	3981.9	3830.6	2809.6	6094.1	5779.0	4984.3
Residual intra-school correlation coefficient	0.621	0.468	0.416	0.331	0.225	0.147	0.073
Proportional reduction in variance components, compared to null model (%)							
School level	22.1	37.2	13.2	24.1	23.6	47.9	16.1
Student level	8.0	11.7	8.8	7.7	16.7	18.9	28.4
Total	17.3	25.8	10.7	13.8	18.4	25.0	27.6

PISA – Vertailu

- Vertailu: Painotettu SRS-analyysi
  - *Weighted SRS analysis option*
  - Oletetaan (virheellisesti), että aineisto on poimittu SRS-otannalla suoraan oppilastason perusjoukosta
    - Oletetaan, että havainnot ovat riippumattomia
    - Toisin sanoen, jätetään huomioimatta ryvästymisen aiheuttama havaintojen korreloituneisuus
  - Käytetään painotettuja estimaatteja

PISA – Vertailtavat mallit

Sekamalli (two-level model; ryväsotantaan perustuva kaksitasomalli):

$$y_{ik} = \text{INTERCEPT} + \gamma_1 \times \text{SSIZE}_i + \gamma_2 \times \text{AUTONOMY}_i + \beta_1 \times \text{FEMALE}_{ik} + \beta_2 \times \text{SEB}_{ik} + \beta_3 \times \text{ENGAGEMENT}_{ik} + \beta_4 \times \text{ACHPRESS}_{ik} + u_i + e_{ik}$$

Kiiteiden vaikutusten malli (Weighted SRS option):

$$y_{ik} = \text{INTERCEPT} + \gamma_1 \times \text{SSIZE}_i + \gamma_2 \times \text{AUTONOMY}_i + \beta_1 \times \text{FEMALE}_{ik} + \beta_2 \times \text{SEB}_{ik} + \beta_3 \times \text{ENGAGEMENT}_{ik} + \beta_4 \times \text{ACHPRESS}_{ik} + e_{ik}$$

Indeksi k : Tason 1 alkiot (oppilaat)
Indeksi i: Tason 2 alkiot (koulut)

Table 9.11 Comparison of estimated coefficients of a two-level model for combined reading literacy score and a fixed-effects model fitted under the weighted SRS analysis option (the German data are used as an example).

Coefficient		Two-level model	Weighted SRS option
Intercept	γ_0	496.4	497.5
	s.e.	4.58	1.93
	t-ratio	108.37	258.08
	p-value	.000	.000
School size	γ_1	27.4	20.1
	s.e.	9.22	1.74
	t-ratio	2.97	11.52
	p-value	.003	.000
Teacher autonomy	γ_2	-7.1	-7.3
	s.e.	5.22	1.38
	t-ratio	-1.37	-5.26
	p-value	.171	.000
Female	β_1	3.6	3.3
	s.e.	2.41	2.74
	t-ratio	1.50	1.20
	p-value	.133	.229
Socio-economic background	β_2	11.5	31.5
	s.e.	1.53	1.38
	t-ratio	7.50	22.9
	p-value	.000	.000
Engagement in reading	β_3	19.0	28.9
	s.e.	0.98	1.17
	t-ratio	19.36	24.6
	p-value	.000	.000
Achievement press	β_4	-1.6	-4.7
	s.e.	1.16	1.31
	t-ratio	-1.35	-3.64
	p-value	.176	.000

Data source: OECD PISA database, 2001.

Tilastollinen ohjelmisto

- SAS
- SPSS
- Stata
- Lisrel
- Mplus

Tilastolliset ohjelmistot: Korreloituneiden aineistojen analyysi

- Hierarkkisesti rakentunut aineisto
  - Ryväsrakenne
  - Ositerakenne
- Asetelmaperusteinen analyysi
  - Painomuuttuja
  - Ositusmuuttuja
  - Ryväsmuuttuja
- Malliperusteinen analyysi
  - Painomuuttuja
  - Ryväsmuuttuja

Tilastollinen ohjelmisto: SAS

- Asetelmaperusteinen analyysi
- SURVEY-proseduurit (SAS versio 9)
- SURVEYMEANS
  - Keskiarvot
- SURVEYFREQ
  - Ristiintaulukointi
  - Asetelmaperusteiset testit
- SURVEYREG
  - Lineaarinen regressioanalyysi, ANOVA, ANCOVA
- SURVEYLOGISTIC
  - Logistiset mallit

Tilastollinen ohjelmisto: SAS

- Malliperusteinen analyysi
- Monitasomallien (sekamallien) sovittaminen
- MIXED
  - Lineaariset sekamallit
- GLIMMIX
  - Yleistetyt lineaariset sekamallit
- NLMIXED
  - Epälineaariset sekamallit

Tilastollinen ohjelmisto: SPSS

- Complex samples (SPSS versio 16)
- Hierarkkinen data
  - Ositettu ryväsotanta
- Asetelmaperusteinen analyysi
  - Asetelmapainot tai analyysipainot
  - Ositusmuuttuja
  - Ryväsmuuttuja
- Modulit
  - CSPLAN ja CSSELECT Otoksen poiminta
  - CSDESCRIPTIVES Kuvailevat tunnusluvut
  - CSTABULATE Ristiintaulukointi ja testit
  - CSGLM, CSLOGISTIC Lineaariset ja logistiset mallit

Tilastollinen ohjelmisto: STATA

- STATA (versio 10)
- Hierarkkinen data
  - Ositettu ryväsotanta
- Asetelmaperusteinen analyysi
  - Analyysipainot
  - Ositusmuuttuja
  - Ryväsmuuttuja
- SVY-optiot (SurVeY data)
  - Kuvailevat tunnusluvut ja testisuureet
  - Yleistetyt lineaariset mallit
  - Biometrian menetelmiä ja malleja
  - Ekonometrian menetelmiä ja malleja

Tilastollinen ohjelmisto: LISREL

- LISREL 8.7 Win
- Hierarkkinen data
  - Ositettu ryväsotanta
- Asetelmaperusteinen analyysi
  - Analyysipainot
  - Ositusmuuttuja
  - Ryväsmuuttuja
- Menetelmät, esimerkiksi:
  - Yleistetyt lineaariset mallit
  - Lineaariset sekamallit

Tilastollinen ohjelmisto: Mplus

- Mplus
- Hierarkkinen data
  - Ositettu ryväsotanta
- Asetelmaperusteinen analyysi
  - Analyysipainot
  - Ositusmuuttuja
  - Ryväsmuuttuja
- Menetelmät, esimerkiksi:
  - Yleistetyt lineaariset mallit
  - Yleistetyt lineaariset sekamallit

Kirjallisuutta

- Chambers R.L. and Skinner C.J. (Eds.) (2004). *Analysis of Survey Data*. Chichester: Wiley.
- Demidenko E. (2004). *Mixed Models. Theory and Applications*. New York: Wiley.
- Diggle P. J., Liang K.-Y. & Zeger S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.
- Gelman A. and Hill J. (2009). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.
- Goldstein H. (2003). *Multilevel Statistical Models. 3rd Edition*. London: Edward Arnold. http://www.cmm.bristol.ac.uk/MLwiN/index.shtml
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Second Edition. Chichester: Wiley.
- OECD (2002a). PISA 2000 Technical Report. Paris: OECD. http://www.pisa.oecd.org/
- Snijders T. and Bosker R. (2002). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage Publications.