HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi

Kevät 2014

TEEMA 5: Korreloituneiden havaintojen tilastollinen mallinnus

Asetelmaperusteisten ja malliperusteisten menetelmien vertailu, ohjelmasovelluksia

Englanninkielinen terminologia

Risto Lehtonen
risto.lehtonen@helsinki.fi

Korreloituneiden havaintojen analyysi

- Havaintojen korrelaation lähteet
  - Ryväsotanta: Rypäänsisäinen korrelaatio (*intra-cluster correlation*)
  - Paneeliasetelma: Havaintojen autokorrelaatio
- 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

SAS - Lineaariset ja yleistetyt lineaariset mallit ja sekamallit

- **Asetelmaperusteiset proseduurit**
  - Otanta-asetelman kaikki ominaisuudet otetaan huomioon asetelmaperusteisilla menetelmillä (TEEMA 4)
    - Ositus
    - Ryvästyminen (havaintojen rypäänsisäinen korrelaatio)
    - Painot
  - SURVEYREG, SURVEYLOGISTIC, SURVEYPHREG
- **Malliperusteiset proseduurit**
  - Pyritään reagoimaan havaintojen korreloituneisuuteen malliperusteisilla menetelmillä
    - GEE-menetelmä: GENMOD
    - Sekamallit: MIXED, GLIMMIX
  - Korreloimattomat havainnot, SRS-oletus, iid-tilanne
    - REG, LOGISTIC

Asetelmaperusteinen analyysi
Lineaariset kiinteiden tekijöiden mallit

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

- PROC SURVEYREG

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)

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.

HUOM: f=m-H =250-5=245 df

Malliperusteinen analyysi
Lineaariset sekamallit

- Malliperusteiset proseduurit
- PROC MIXED
  - **Lineaariset sekamallit**
  - Jatkuva tulosmuuttuja *y*
  - Regressiomalli, ANOVA, ANCOVA
  - Reagoidaan otanta-asetelman ominaisuuksiin
    - Ryvästyminen (sisäkorrelaatio): RANDOM-lause tai REPEATED-lause
    - Analyysipainot: WEIGHT-lause tai painomuuttuja malliin kiinteänä vaikutuksena (oma beta-parametri)
- PROC MIXED

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)

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

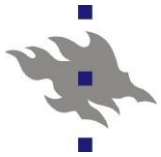
Malliperusteinen analyysi

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;
```



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

Asetelmaperusteinen analyysi
Logistiset kiinteiden tekijöiden mallit

- Asetelmaperusteiset proseduurit (TEEMA 4)
- 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

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)

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

Yhdysvaikutusmalli

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

Lehtonen & Pahkinen (2004) Table 8.8

Table 8.8 Design-based logistic ANCOVA on overall psychic strain with the PML method.

Model term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value	Odds ratio	95% confidence interval for OR	
							Lower	Upper
Intercept	0.1964	1.56	0.1572	1.25	0.2127	1.22	0.89	1.66
Sex								
Males	−0.9926	1.43	0.2033	−4.88	0.0000	0.37	0.25	0.55
Females*	0	n.a.	0	n.a.	n.a.	1	1	1
Age	−0.0046	1.55	0.0041	−1.12	0.2624	1.00	0.99	1.00
Physical health hazards	0.2765	1.39	0.0596	4.64	0.0000	1.32	1.17	1.48
Chronic morbidity	0.5641	1.17	0.0575	9.82	0.0000	1.76	1.57	1.97
Sex, Age								
Males	0.0131	1.41	0.0051	2.56	0.0111	1.01	1.00	1.02
Females*	0	n.a.	0	n.a.	n.a.	1	1	1

* Reference class; parameter value set to zero.
n.a. not available.

Malliperusteinen analyysi
Model-based analysis 1

- **Model-based methods**
  - ”Mainstream statistics” framework
  - Correlation structure is incorporated in the statistical model by including **random effects** in addition to the fixed effects
- **Mixed / Hierarchical / Multilevel models**
  - Huge literature!
  - Established statistical software
  - Popular approach in many scientific disciplines

Malliperusteinen analyysi
Model-based analysis 2

- Typical modelling framework
  - **Generalized linear mixed models GLMM**
- Complexities of the survey data are incorporated in the model by including:
  - Random effects to account for clustering
  - Random effects to account for stratification
  - Fixed effects to account for weighting or WEIGHT statement
- Software
  - SAS Procedures GENMOD, MIXED and GLIMMIX
  - SPSS, Stata: Similar options
  - Mplus, MLwiN: Generalized linear mixed models
  - R packages, e.g. **`nlme, lme4`**

Malliperusteinen analyysi
Model-based analysis 3

- **Some open questions**
- Accounting for stratification
  - Not straightforward
  - Often specified as strata-level random effects
  - In fact, there is no randomness when stratifying the whole population into strata!
- Accounting for unequal probability sampling
  - Not any consensus within statistics community
    e.g. Pfeffermann D., Skinner C.J., Holmes D.J., Goldstein H. and Rasbash, J. (1998)
  - An option: Include weight variable as covariate in the model (perhaps better than WEIGHT statement)

Malliperusteinen analyysi
Logistiset kiinteiden tekijöiden mallit

- 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 tai painomuuttuja malliin!
- PROC GENMOD

Malliperusteinen analyysi
Logistiset sekamallit

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

- PROC GLIMMIX

Technical annex: Linear mixed model

Linear mixed model

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

where

X is design matrix for fixed effects

Z is design matrix for random effects

β is vector of fixed effects

u is vector of random effects

ε is the residual vector

Key assumption: **u** and **ε** are normally distributed with a certain type of covariance structure

Technical annex: Linear mixed model

Assumptions

$$E\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad \text{cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

The variance of **y** then is

$$V(\mathbf{y})=\mathbf{ZGZ}'+\mathbf{R}$$

G: Covariance structure for random effects

R: Covariance structure for residuals

Modelling: Set up the random effects design matrix **Z** and specify covariance structures to **G** and/or **R**

Technical annex: GLMM

Generalized linear mixed model

$$E(y \mid \mathbf{u}) = g^{-1}(\mathbf{X\beta} + \mathbf{Zu})$$

where

$g(.)$ is the link function (linear, logistic,...)

X is design matrix for fixed effects

Z is design matrix for random effects

β is vector of fixed effects

u is vector of random effects

Assumption: **u** is normally distributed with covariance matrix **G**

Modelling: Set up the random effects design matrix **Z**

and specify covariance structures to **G**

Technical annex: GLMM specification

Model:

$$E(y_k \,|\, \mathbf{u}_d) = g^{-1}(\mathbf{x}_k'\boldsymbol{\beta} + \mathbf{z}_k'\mathbf{u}_d))$$

where d refers to cluster and k refers to element in a cluster

$g(.)$ refers to link function:

- linear link function
- logistic link function

$\mathbf{x}_k = (1, x_{1k}, ..., x_{pk})'$ design vector values for fixed effects

$\mathbf{z}_k = (1, z_{1k}, ..., z_{qk})'$ design vector values for random effects

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

$\mathbf{u}_d = (u_{0d}, ..., u_{qd})'$ cluster-specific random effects

(intercepts and slopes) assumed $N(\mathbf{0}, \mathbf{G})$

Special case 1: Linear mixed model with random intercepts and slopes

Model:

$$E(y_k \,|\, \mathbf{u}_d) = \mathbf{x}_k'\boldsymbol{\beta} + \mathbf{z}_k'\mathbf{u}_d, \text{ where}$$

$\mathbf{x}_k = (1, x_{1k}, x_{2k})'$ design vector for fixed effects

$\mathbf{z}_k = (1, x_{1k})'$ design vector for random effects

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ fixed effects

$\mathbf{u}_d = (u_{0d}, u_{1d})'$ cluster-specific random effects

$$y_k = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + \beta_2 x_{2k} + \varepsilon_k$$

where $\mathbf{u}_d = (u_{0d}, u_{1d})'$ is $N(\mathbf{0}, \mathbf{G})$ with $\mathbf{G} = \begin{pmatrix} \sigma^2_{u0} & \sigma^2_{u10} \\ \sigma^2_{u01} & \sigma^2_{u1} \end{pmatrix}$

Special case 2: Logistic mixed model with random intercept

Model:

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

where

$\mathbf{x}_k = (1, x_{1k}, x_{2k})'$ design vector for fixed effects

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ fixed effects

u_d is cluster-specific random intercept and is assumed $N(0, \sigma_u^2)$

Model-based estimation of parameters of generalized linear mixed model GLMM

- Clustering is accounted for by specifying covariance structures to the random effects (covariance matrix **G)**
  - Complex numerical approximation methods
- Parameter estimation (linear mixed model)
  - REML – restricted ML for random effects
  - GLS – generalized least squares for fixed effects
  - Standard errors: ”*Sandwich form*” variances
- SAS/PROC GLIMMIX
- Demidenko E. (2004)
- Goldstein H. (2011)

Model-based estimation of parameters of logistic model with GEE method

- GEE method: *Generalized estimating equations*
  - Originally developed for longitudinal surveys
  - Diggle, P. J., Liang, K.-Y. & Zeger, S. L. (1994
- Clustering is accounted for by specifying covariance structures to the multivariate responses
  - Independent correlation structure (= PML method)
  - Exchangeable correlation structure (common intra-cluster correlation assumed)
- Standard errors: ”*Sandwich form*”
  - “Empirical”, “Robust” covariance matrix
- SAS PROC GENMOD

Estimation of parameters of logistic model 3

SUMMARY: SAS modelling procedures Accounting for intra-class correlation in fitting **logistic model**		
Method	Accounting for **clustering**…	
	In estimation of model parameters	In estimation of standard errors
LOGISTIC	No	No
SURVEYLOGISTIC	No	Yes
GENMOD(GEE-IND)	No	Yes
GENMOD(GEE-EXCH)	Yes	Yes
GLIMIMX (GLMM)	Yes	Yes

LOGISTIC: Standard ML
SURVEYLOGISTIC: PML with "sandwich form" covariance matrix
GENMOD(GEE-IND): Generalized Estimating Equations with independent cov. structure
GENMOD(GEE-EXCH): Generalized Estimating Equations with exchangeable cov. structure
GLIMMIX(GLMM): Logistic mixed model with cluster-specific random terms

EXAMPLE: Design-based logistic ANCOVA - Revisited

- OHC Survey

- Stratified cluster sampling
  H= 5 strata
  m= 250 sample clusters (workplaces)
  n = 7841 sample persons

Design-based logistic ANCOVA

- Binary response
  PSYCH2 Psychic strain
  - 0: Less severe (equal or less than median)
  - 1: More severe (greater than median)

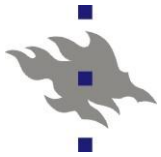
- Categorical explanatory variable
  - SEX (M/F)
- Continuous explanatory variable
  - AGE (in years)
- Binary explanatory variables
  - Physical health hazards of work PHYS (0/1)
  - Chronic morbidity CHRON (0/1)

Final reduced logistic fixed-effects model (GEE)

$$\text{logit}(y_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k} + \beta_7 x_{7k}$$

where

- β_0 is for intercept
- β_1 and β_2 are for SEX ($\beta_2 = 0$)
- β_3 is for AGE
- β_4 is for PHYS
- β_5 is for CHRON
- β_6 and β_7 are for SEX*AGE interaction ($\beta_7 = 0$)



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	.	.

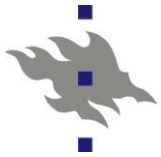
Exchangeable Working Correlation
Correlation 0.0156016243

Final reduced logistic mixed model (GLMM)

$$\text{logit}(y_k) = (\beta_0 + u_d) + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k} + \beta_7 x_{7k}$$

where

- β_0 is for intercept
- β_1 and β_2 are for SEX ($\beta_2 = 0$)
- β_3 is for AGE
- β_4 is for PHYS
- β_5 is for CHRON
- β_6 and β_7 are for SEX*AGE interaction ($\beta_7 = 0$)



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: vastaava 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	.	.	.	.

Comparison of results

- Interaction term AGE*SEX
- SAS Procedures
  - SURVEYLOGISTIC
    - design-based
  - GENMOD
    - model-based with GEE estimation
  - GLIMMIX
    - model-based with mixed model specification

Comparison of test results for interaction term AGE*SEX in OHC

	Model term	Beta coefficient	Standard error	t-test statistic	p-value
Analysis accounting for clustering					
Design-based Fixed-effects model PML method	SEX*AGE	0.0131	0.0051	2.56	0.0111
Model-based Fixed-effects model GEE method	SEX*AGE	0.0142	0.0050	2.86	0.0046
Model-based Mixed model, REML method	SEX*AGE	0.0144	0.0050	2.87	0.0045
Analysis ignoring clustering (SRS based)					
SRS based Fixed-effects model ML method	SEX*AGE	0.0131	0.0043	3.042	0.0026

Conclusions for accounting for clustering and stratification in logistic ANCOVA

- **Design-based analysis** SURVEYLOGISTIC
  - Accounting for stratification and clustering effect
  - Most conservative (largest p-value)
- **Model-based methods** GENMOD, GLIMMIX
  - Accounting for clustering effect with GEE and GLMM methods
  - Similar results in both cases
- **SRS-based analysis** *(iid assumption)*
  - Stratification and clustering ignored
  - Overly liberal test results
  - SRS assumption obviously wrong in this case

SUMMARY
Software for design-based modelling

- SAS / SURVEY procedures
- SPSS / COMPLEX SAMPLES module
- Stata/svy-procedures
- Sudaan software
- R SURVEY package (Lumley)
- Mplus COMPLEX type analysis

Design-based analysis – SAS

- SAS - Design-based analysis procedures for cluster correlated data
  - Sample selection: SURVEYSELECT
  - Means, proportions: SURVEYMEANS
  - Two-way tables: SURVEYFREQ
  - Linear regression: SURVEYREG
  - Logistic regression: SURVEYLOGISTIC
  - Cox proportional hazards model: SURVEYPHREG
- Current version: SAS 9.3

Design-based analysis – IBM SPSS

- SPSS Complex Samples package
  - CSPLAN Complex samples plan
  - CSGLM Numerical outcome prediction through the Complex Samples General Linear Model
  - CSORDINAL Ordinal outcome prediction through Complex Samples Ordinal Regression
  - CSLOGISTIC Categorical outcome prediction through Complex Samples Logistic Regression
  - CSCOXREG Time to an event prediction through Complex Samples Cox Regression
- Current version: IBM SPSS ver. 21

Design-based analysis - Mplus

- Mplus Software: MODELING WITH COMPLEX SURVEY DATA
- Design-based analysis
  - Standard errors and a chi-square test of model fit are computed taking into account stratification, non-independence of observations due to cluster sampling, and/or unequal probability of selection.
- Multilevel analysis
  - A second approach is to specify a model for each level of the multilevel data thereby modeling the non-independence of observations due to cluster sampling.
  - In both approaches, observed outcome variables can be continuous, censored, binary, ordered categorical (ordinal), unordered categorical (nominal), counts, or combinations of these variable types.

Kirjallisuutta

- 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
- Heeringa S.G., West B.T. and Berglund P.A. (2010). Applied Survey Data Analysis. Chapman and Hall/CRC.
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Second Edition. Chichester: Wiley.
- Lumley T. (2010) *Complex Surveys: A Guide to Analysis Using R.* Wiley.
- Vonesh E.F. (2012). Generalized Linear and Nonlinear Models for Correlated Data. Theory and Applications Using SAS. SAS Institute.