



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

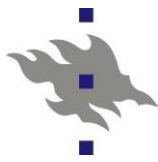
Topics in Survey Methodology and Survey Analysis

PART 3

The analysis of complex survey data

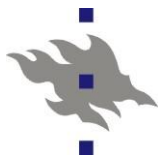
Risto Lehtonen
University of Helsinki

Lecture notes, 26 and 29 September 2011



Course description: PART 3

- Lectures (SSKH IT-sal) (6 hours)
 - Monday 26 Sept. at 14-18
 - Thursday 29 Sept. at 14-16
- PC training (3 hours)
 - Analysis of OHC Survey data
 - Thursday 29 Sept. at 16-19
- Homework, options:
 - Analysis of OHC Survey data (for all)
 - Analysis of own data set (personal)



PART 3: Topics and materials

- Hierarchical or multilevel data, analysis of multilevel or cluster correlated data, basic multilevel modelling, linear and logistic mixed models, software.
- Lehtonen R. and Pahkinen E. (2004).
Practical Methods for Design and Analysis of Complex Surveys. Second Edition.
Chichester: John Wiley & Sons

[VLISS](#)-virtual laboratory in survey sampling



Survey methodology and analysis

Social statistics

- Social statistics focuses on statistical methods for describing and analyzing social phenomena and change
 - Welfare, Living conditions
 - Poverty, Social exclusion
 - Labour market
 - ...
- The methods of social statistics are widely used in empirical research in many fields, including social and behavioral sciences and economics and official statistics production



Sub-areas of social statistics 1

- Survey sampling
- Survey methodology
- Survey analysis

- Survey sampling
 - Sampling techniques
 - Estimation methods

- Survey methodology
 - Data collection methods
 - Nonresponse treatment
 - Measurement



Sub-areas of social statistics 2

- **Survey analysis**
- **Descriptive methods**
 - Means, proportions
 - Standard errors
 - Confidence intervals
 - Frequency tables
 - Simple test statistics
- **Analytic methods**
 - Linear regression analysis
 - Logistic regression analysis
 - Multilevel modelling



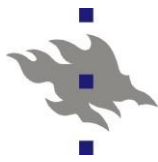
Survey analysis

- Sampling design
 - Stratification
 - Clustering
 - Weighting
- Study design
 - Cross-sectional
 - Longitudinal
- Analysis design
 - Analysis methods that account for complex sampling and/or hierarchical structure of data



Complex survey data

- Hierarchically structured data
- Clustered data, Multilevel data
- Common in quantitative research in sociology, psychology and educational sciences
- Typical hierarchical structures (clusters)
 - Schools - Students
 - Work places - Staff members
 - Health centers – Patients
 - Household - Hh members
- Multilevel structure



Examples of hierarchical structure

- Multi-stage sampling design with clustering of population elements
- Occupational Health Care Survey (OHC)
 - Workplaces as clusters
- PISA Survey
 - Schools as clusters
- Health 2000
 - Health center districts as clusters



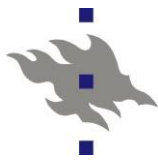
Analysis of complex survey data 1

- The hierarchical structure of the data involves correlations between observations
- The correlations must be accounted for to obtain proper statistical inference
- Two main approaches
 - Design-based methods
 - SAS / SURVEY procedures
 - SPSS / COMPLEX SAMPLES module
 - Mplus COMPLEX type analysis
 - Hierarchical / Multilevel / Mixed models
 - SAS / Procedure GENMOD, MIXED and GLIMMIX
 - SPSS: Similar options
 - Mplus, MLwiN: Generalized linear mixed models



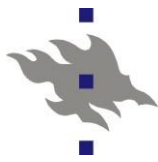
Hierarchical or clustered structure and sources of correlation of observations

Levels of hierarchy	Research design	
	a. Cross-sectional	b. Longitudinal (Panel design)
1. Single-level data (no clustering)	1a. No correlation between observations	1b. Positive autocorrelation between observations
2. Two or more levels (clustered data)	2a. Positive intra-class correlation between observations	2b. Complex covariance structures



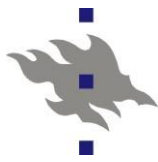
Analysis of complex survey data 2

- Terminology
 - Fixed-effects models
 - Mixed models, multilevel models, hierarchical models
- Linear fixed-effects models
- Linear mixed models
 - Continuous response variable
- Logistic fixed-effects models
- Logistic mixed models
 - Binary response
 - Polytomous response
 - Nominal level measurement
 - Ordinal level measurement



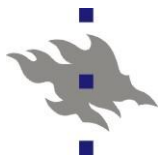
Analysis of complex survey data 3

- Modelling framework:
Generalized linear mixed models GLMM
 - Continuous response
 - Linear regression/ANOVA/ANCOVA model
 - Binary response
 - Logistic regression/ANOVA/ANCOVA model
 - Polytomous response
 - Multinomial logistic regression/ANOVA/ANCOVA model
 - Count response
 - Poisson regression/ANOVA/ANCOVA model



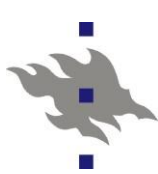
Design-based analysis - SAS

- 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
 - More...



Design-based analysis - PASW (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



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.



Model-based analysis - SAS

- Model-based procedures for hierarchical (multilevel) or clustered data
 - Means, proportions: MEANS
 - Two-way tables: FREQ
 - Linear regression: MIXED
 - Logistic regression: GENMOD
 - Generalized linear mixed models: GLIMMIX



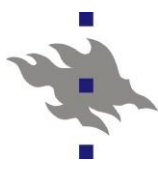
Design-based and model-based analysis - Mplus

- Two approaches for clustered / multilevel data
- Design-based analysis for complex survey data
 - Clustering, Stratification, Weighting
 - Linear, log-linear and logistic models
- Model-based analysis of multilevel data
 - Two-level data
 - Clustering, Stratification, Weighting
 - Linear and logistic mixed models
- Structural Equation Models (SEM) for multilevel data (not discussed here)
- [More...](#) [Mplus Language Summary](#)



Model-based analysis - MLwiN

- Multilevel modelling - MLwiN
- Goldstein H. (2003). *Multilevel Statistical Models*, 3rd Ed. London: Arnold.
- [MLwiN](http://www.cmm.bristol.ac.uk/MLwiN/)
www.cmm.bristol.ac.uk/MLwiN/
- [LEMMA](http://www.cmm.bristol.ac.uk/learning-training/index.shtml) Learning Environment for Multilevel Methods and Applications
www.cmm.bristol.ac.uk/learning-training/index.shtml



Software for multilevel modeling

- MLWIN
 - Multilevel (generalized linear mixed) modeling
- HLM
 - Hierarchical (linear mixed) modeling
- MIXED (SAS)
 - Linear mixed modeling
- GLIMMIX (SAS)
 - Generalized linear mixed modeling
- GLLAMM (Stata)
 - Generalized linear latent and mixed modeling
- LISREL
 - Structural equation modeling (SEM)
- MPLUS
 - Structural equation modeling (SEM)



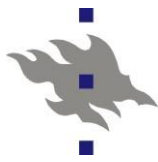
Virtual training materials

- Web extension of Lehtonen and Pahkinen (2004)

VLISS-Virtual Laboratory in Survey Sampling

<http://mathstat.helsinki.fi/VLISS/>

- Analysis of a complex survey data set involving stratification and clustering



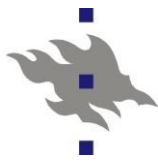
Basic sampling methods

- Lehtonen R. and Djerf K. (2008). *Survey sampling reference guidelines*. Luxembourg: Eurostat Methodologies and Working papers.

- [PDF](#)

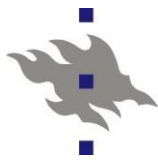
- Free download at:

http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF



Occupational Health Care Survey OHC

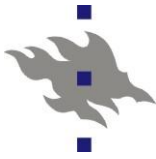
- Study design: Cross-sectional
- Sampling design
 - Stratified one-stage and two-stage cluster sampling with workplaces as clusters
- Stratification by cluster size and type of industry
 - Small workplaces: One-stage cluster sampling
 - Large workplaces: Two-stage sampling
- Positive intra-cluster correlation of observations within clusters



OHC-data

- Demonstration data: SAS data OHC
 - Real data cleaned for pedagogical purposes
 - Workplaces with more than 10 workers
 - $H = 5$ strata
 - $m = 250$ workplaces
 - Primary Sampling Units, PSU (clusters)
 - $n = 7841$ persons
 - 10 variables
 - Varying number of elements per cluster

- VLISS [Section 5.6](#)

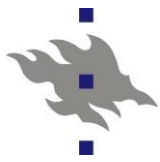


OHC Data

The CONTENTS Procedure

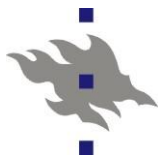
Variables in Creation Order

#	Variable	Type	Len	Label
1	ID	Num	8	Element identifier
2	STRATUM	Num	8	Stratum identifier
3	SEX	Num	8	Gender
4	AGE	Num	8	Age in years
5	AGE2	Num	8	Age under/over 45
6	PHYS	Num	8	Physical health hazards of work
7	CHRON	Num	8	Chronic morbidity
8	PSYCH	Num	8	Psychic strain - 1st princomp
9	PSYCH2	Num	8	Psychic strain - dichotomy
10	PSU	Num	8	Primary sampling unit (Cluster)
11	wd	Num	8	Design weight
12	wa	Num	8	Rescaled weight (Analysis weight)



Analysis of OHC data

- Hierarchical (clustered) structure
 - Workplaces as clusters
- Positive intra-cluster correlation of observations within clusters
- Measures of correlation
 - Design effect (deff)
 - Intra-cluster correlation ICC



Design effect deff 1

■ Overall design effect (1)

- Measures the effect of:
 - Stratification
 - Clustering
 - Weighting

on variance estimate of the mean estimate

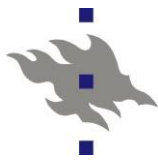
- SRS variance estimate is for **unweighted** mean estimate

■ Deff accounting for stratification and clustering (2)

- Measures the effect of:
 - Stratification
 - Clustering

on variance estimate of the mean estimate

- SRS variance estimate is for **weighted** mean estimate



Design effect deff 2

Design effect, deff (Kish 1965) measures the magnitude of the clustering effect to variance (standard error) estimate

Estimated overall deff (1):

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

where

\bar{y}^* is weighted mean estimate and \bar{y} is the corresponding unweighted mean estimate

$\hat{v}(\bar{y}^*)$ is based on the actual sampling design

$\hat{v}_{srs}(\bar{y})$ is the SRS-based variance estimate

Deff (2):

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



Deff for proportion estimate

Example: Deff for proportion estimate \hat{p}

$$deff(\hat{p}) = \frac{v_{clu}(\hat{p})}{v_{srs}(\hat{p})} = \frac{v_{clu}(\hat{p})}{\hat{p}(1-\hat{p})/n}$$

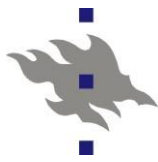
where

\hat{p} is the estimated proportion

v_{clu} is the variance estimate of \hat{p} based on the actual cluster sampling design

v_{srs} is the variance estimate of \hat{p} based on an assumption of simple random sampling (here: binomial variance formula)

n is the actual sample size



Interpretation of d_{eff}

- $d_{eff} < 1$
 - The actual sampling design is **more efficient** than SRS

- $d_{eff} = 1$
 - The actual sampling design is **equally efficient** as SRS

- $d_{eff} > 1$
 - The actual sampling design is **less efficient** than SRS
 - Typical case for clustered data
 - OHC, PISA, Health2000



OHC data: *Deff estimates* (Lehtonen&Pahkinen 2004)

Table 5.8

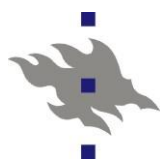
Averages of design-effect estimates of proportion estimates of selected groups of binary response variables in the OHC Survey data set (number of variables in parentheses).

Study variable	Mean deff
Physical working conditions (12)	6.5
Psycho-social working conditions (11)	3.3
Psychosomatic symptoms (8)	2.0
Psychic symptoms (9)	1.8



• The effect of positive intra-cluster correlation on statistical analysis

- When compared with an element-level simple random sample (SRS) of the same element sample size n :
 - Decreasing effective sample sizes
 - Increasing standard errors
 - Larger confidence intervals
 - Weaker statistical significance of statistical tests



Deff, ICC and effective sample size

Deff and ICC

$$\hat{\rho}_{ICC} = \frac{deff(\hat{p}) - 1}{\bar{n} - 1}$$

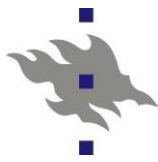
Effective sample size

$$n_{eff} = \frac{n}{deff(\hat{p})} = \frac{n}{1 + (\bar{n} - 1)\hat{\rho}_{ICC}}$$

where

n is element sample size

\bar{n} is average cluster sample size



OHC example: Effective sample size

■ Physical working conditions

- Design effect $deff = 6.5$
- Intra-cluster correlation
ICC = 0.181
- Element sample size
 $n = 7841$ persons
- Effective sample size
 $n(\text{eff}) = 7841/6.5$
 $= 1206$ persons

■ Psychic symptoms

- Design effect $deff = 1.8$
- Intra-cluster correlation
ICC = 0.026
- Element sample size
 $n = 7841$ persons
- Effective sample size
 $n(\text{eff}) = 7841/1.8$
 $= 4356$ persons



PISA example: Effective sample size

Table 2. Descriptive statistics for combined reading literacy score in the PISA 2000 Survey by country (in alphabetical order).

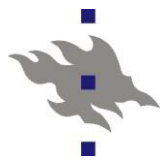
				Effective sample size of students	Number of observations in data set	
Country	Mean	Standard error	Design effect		Students	Schools
Brazil	402.9	3.82	8.33	476	3961	290
Finland	550.7	2.15	2.79	1600	4465	147
Germany	497.4	5.68	13.47	305	4108	183
Hungary	485.7	6.02	20.00	231	4613	184
Republic of Korea	526.6	3.66	12.99	351	4564	144
United Kingdom	531.4	4.08	14.08	564	7935	328
United States	517.0	5.16	6.93	354	2455	112
All	500.0			3881	32101	1388

Data source: OECD PISA database, 2001.



Accounting for sampling design complexities in the analysis phase

- The sample data set prepared for the analysis should include the following technical variables:
 - Stratum indicator
 - Cluster indicator
 - Weight variables
 - Design weight
 - Analysis weight
 - Indicators for imputed variable values



Design weight

Design weight: $w_k = 1/\pi_k$ for element k ,
 $k = 1, \dots, n$, where π_k is inclusion probability for
element k and n is the size of sample data set

$$\sum_{k=1}^n w_k = N,$$

where N is the population size

Design weights are needed when estimating
population totals



Analysis weight

Analysis weight: Rescaled design weight

$$w_k^* = (n/N)w_k, \quad k = 1, \dots, n,$$

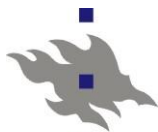
where n is sample size and N is population size

$$\sum_{k=1}^n w_k^* = n \text{ (sample size)}$$

and thus the average analysis weight = 1

Analysis weights are used in statistical analysis

NOTE: For SRS sample analysis weight = 1



EXAMPLE of complex weighting procedure in PISA

Design weight w_{ik} for student k in school i :

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

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 ;

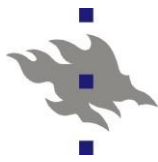
f_i is an adjustment factor for school i to compensate any country-specific refinements in the survey design, and m is the number of sample schools in a given country and n_i is the number of sample students in school i .



OHC Survey – Statistical analysis 1

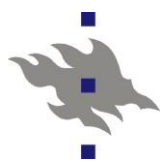
- Alternative statistical methods for proper statistical analysis of OHC Survey data set?

- Recall:
 - Stratified cluster sampling design
 - Weighting (simple here)
 - Analysis weights = 1
 - Stratification by STRATUM
 - Clustering by PSU



OHC Survey – Statistical analysis 2

- Design-based methods
 - Linear fixed-effects models for continuous response
 - Logistic fixed-effects models for binary response
- Model-based methods
 - Linear mixed models for continuous response
 - Logistic mixed models for binary response
- Generalized linear mixed models GLMM
 - Continuous response
 - Binary response
 - Polytomous response
 - Nominal or ordinal level of measurement
 - Count response



Generalized linear mixed model GLMM

Model:

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

where $f(\cdot)$ refers to the link function, e.g.

- linear model
- logistic model

$\mathbf{x}_k = (1, x_{1k}, \dots, x_{pk})'$ vector of explanatory variable values for element k

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

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

Special case 1

Linear fixed-effects model

Model:

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

where

$\mathbf{x}_k = (1, x_{1k}, \dots, x_{pk})'$ vector of explanatory variable values for element k

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

E.g. $y_k = \beta_0 + \beta_1 x_{1k} + \dots + \beta_p x_{pk} + \varepsilon_k$

Special case 2

Linear mixed model

Model:

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

where

$\mathbf{x}_k = (1, x_{1k}, \dots, x_{pk})'$ vector of explanatory variable values for element k

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

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

E.g. $y_k = \beta_0 + u_{0d} + \beta_1 x_{1k} + \dots + \beta_p x_{pk} + \varepsilon_k$

Special case 3

Logistic fixed-effects model

Model

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

where

$\mathbf{x}_k = (1, x_{1k}, \dots, x_{pk})'$ vector of explanatory variable

values for element k

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

Special case 4

Logistic mixed model

Model

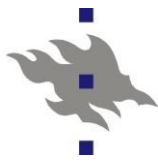
$$E_m(y_k | \mathbf{u}_d) = \frac{\exp(\mathbf{x}'_k \boldsymbol{\beta} + \mathbf{u}_d)}{1 + \exp(\mathbf{x}'_k \boldsymbol{\beta} + \mathbf{u}_d)}$$

where

$\mathbf{x}_k = (1, x_{1k}, \dots, x_{pk})'$ vector of explanatory variable values for element k

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

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



Computation tools

■ Descriptive methods

■ SAS procedures for means and proportions

■ MEANS

- SRS assumption

■ SURVEYMEANS

- General sampling design

■ SAS procedures for frequency tables

■ FREQ

- SRS assumption

■ SURVEYFREQ

- General (complex) sampling design



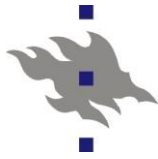
PROC SURVEYMEANS - OHC data

(1) Valid analysis by accounting for stratification and clustering

```
proc surveymeans data=ohc mean;  
  var psych2 phys chron age sex;  
  strata stratum;  
  cluster PSU; * Primary Sampling Unit;
```

(2) Invalid analysis assuming SRS

```
proc surveymeans data=ohc mean;  
  var psych2 phys chron age sex;
```

(1) Valid analysis (actual cluster sampling design)

Variable	Label	Mean	Std error of Mean
PSYCH2	Psychic strain - dichotomy	0.499426	0.007336
PHYS	Physical health hazards of work	0.345747	0.014385
CHRON	Chronic morbidity	0.292437	0.006808
AGE	Age in years	37.581941	0.251905
SEX	Gender	1.428007	0.01851

(2) Invalid analysis (SRS assumption)

Variable	Label	Mean	Std error of Mean
PSYCH2	Psychic strain - dichotomy	0.499426	0.005647
PHYS	Physical health hazards of work	0.345747	0.005371
CHRON	Chronic morbidity	0.292437	0.005137
AGE	Age in years	37.581941	0.120721
SEX	Gender	1.428007	0.005588



PROC SURVEYFREQ - OHC data

(1) Valid analysis by accounting for stratification and clustering

```
proc surveyfreq data=ohc;  
  tables phys*psych2 / chisq;  
  strata stratum;  
  cluster PSU;
```

(2) Invalid analysis assuming SRS

```
proc surveyfreq data=ohc;  
  tables phys*psych2 / chisq;
```

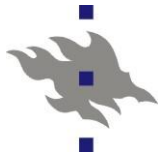


(1) Valid analysis (actual cluster sampling design)

PHYS	PSYCH2	Frequency	Percent	Std err of Percent
0	0	2629	33.5289	0.8321
	1	2501	31.8964	0.9890
	Total	5130	65.4253	1.4385

1	0	1296	16.5285	0.8304
	1	1415	18.0462	0.8266
	Total	2711	34.5747	1.4385

Total	0	3925	50.0574	0.7336
	1	3916	49.9426	0.7336
	Total	7841	100.000	

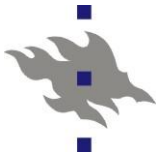


(2) Invalid analysis (SRS assumption)

PHYS	PSYCH2	Frequency	Percent	Std err of Percent
0	0	2629	33.5289	0.5332
	1	2501	31.8964	0.5264
	Total	5130	65.4253	0.5371

1	0	1296	16.5285	0.4195
	1	1415	18.0462	0.4343
	Total	2711	34.5747	0.5371

Total	0	3925	50.0574	0.5647
	1	3916	49.9426	0.5647
	Total	7841	100.00	



**(1) Valid design-based
statistical test**

Rao-Scott Chi-Square Test

Pearson Chi-Square	8.4070
Design Correction	1.4032

Rao-Scott Chi-Square	5.9913
DF	1
Pr > ChiSq	0.0144

F Value	5.9913
Num DF	1
Den DF	245
Pr > F	0.0151

Sample Size = 7841

(2) Invalid test (SRS-based)

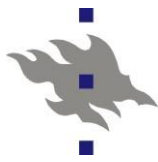
Rao-Scott Chi-Square Test

Pearson Chi-Square	8.4070
Design Correction	1.0000

Rao-Scott Chi-Square	8.4070
DF	1
Pr > ChiSq	0.0037

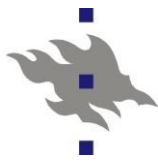
F Value	8.4070
Num DF	1
Den DF	7840
Pr > F	0.0037

Sample Size = 7841



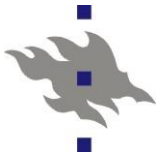
Computation tools for linear models

- **Analytical methods**
- SAS procedures for **linear** models
- Continuous response variable y
- Procedure REG
 - SRS assumption
- Procedure SURVEYREG
 - General sampling design
 - e.g. Stratified cluster sampling



Computation tools for logistic models

- **Analytical methods**
- SAS procedures for **logistic** models
- Binary or polytomous response
- Procedure LOGISTIC
 - SRS assumption
- Procedure SURVEYLOGISTIC
 - General sampling design
- Summary table



PROC SURVEYLOGISTIC < options >;

BY variables ;

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

CLUSTER variables ;

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

FREQ variable ;

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

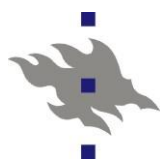
MODEL variable < (variable_options) > = < effects
> < / options >;

STRATA variables < / options > ; < label: >

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

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

WEIGHT variable </ option >;



Two simple logistic models

Logistic fixed-effects model

One x-variable

$$\text{logit}(y_k) = \log\left(\frac{y_k}{1-y_k}\right) = \mathbf{x}'_k \boldsymbol{\beta} = \beta_0 + \beta_1 \mathbf{x}_{1k}$$

where β_0 is the *fixed intercept effect*

β_1 is the *fixed slope effect*

Logistic multilevel model (mixed model)

$$\text{logit}(y_k | u_d) = \log\left(\frac{y_k}{1-y_k}\right) = \beta_0 + u_{0d} + \beta_1 \mathbf{x}_{1k}$$

where u_{0d} are *cluster-specific random intercepts*



Estimation of parameters of logistic model 1

- GWLS method (simplest method)
 - *Generalized weighted least squares*
 - Non-iterative method
 - Generalized linear model (not applied in SAS)

- PML method
 - *Pseudo maximum likelihood*
 - Iterative method
 - Logistic model: SAS/SURVEYLOGISTIC
 - Generalized linear model: SAS/ GENMOD



Estimation of parameters of logistic model 2

■ GEE method

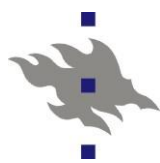
- *Generalized estimating equations*
- SAS/GENMOD
 - Generalized linear model
 - Exchangeable correlation structure

■ More complex numerical approximation methods

- SAS/GLIMMIX
 - Generalized linear mixed model

Estimation of parameters of logistic model 3

SUMMARY: Accounting for intra-class correlation in estimation of the model		
	Accounting for clustering...	
	In estimation of model parameters	In estimation of standard errors
Least squares		
LS	No	No
GWLS	Yes	Yes
Maximum likelihood		
ML	No	No
PML	No	Yes
GEE-IND	No	Yes
GEE-EXCH	Yes	Yes
LS Ordinary least squares GWLS Generalized Weighted Least Squares ML Standard Maximum Likelihood PML Pseudo Maximum Likelihood	GEE Generalized Estimating Equations GEE-IND = PML GEE-EXCH Exchangeable correlation structure Multivariate quasilielihood	



Design-based Wald statistic

Asymptotically χ^2 distributed test statistic
with $df=1$

$$X^2_{des}(\beta_j) = \frac{\hat{\beta}_j^2}{v_{des}(\hat{\beta}_j)}, \quad j = 1, \dots, p+1$$

where

$\hat{\beta}_j$ is estimated logistic regression coefficient (esim. PML)

$v_{des}(\hat{\beta}_j)$ design-based variance estimate of $\hat{\beta}_j$

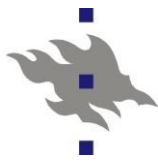
The corresponding t test statistic is $t_{des}(\beta_j) = \frac{\hat{\beta}_j}{\text{s.e.}_{des}(\hat{\beta}_j)}$

(signed square root of Wald statistic)



- **EXAMPLE**
- **Lehtonen&Pahkinen (2004) Example 8.1**

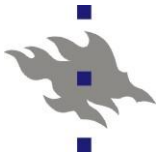
- **The analysis of frequency data**
- Design based logistic ANOVA
- Multidimensional frequency table
- One discrete response variable
 - Binary (0 / 1)
 - Polytomous (>2 classes)
 - Several discrete predictors
- Modelling of the relationship between response variable and predictors with a logistic ANOVA model



Design-based logistic modelling

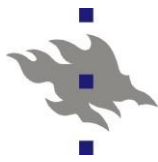
- SAS Procedure SURVEYLOGISTIC
 - Binary response
 - Polytomous response
 - Nominal level (A / B / C /...)
 - Ordinal level (1 / 2 / 3 /...)

- Properties of sampling design must be accounted for
 - Stratification (STRATA statement)
 - Clustering (CLUSTER statement)
 - Weighting (WEIGHT statement)



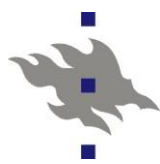
Logit ANOVA model 1

- Simplest case
 - Binary (0/1) response
- OHC data
 - Response variable y : PSYCH2
 - 1 - More severe psychic strain
 - 0 - Less severe psychic strain
- Dichotomized by the median of the continuous measurement PSYCH
 - PSYCH = Standardized first principal component of nine measures of psychic strain



Logit ANOVA model 2

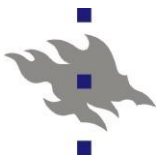
- Discrete predictors (x -variables):
 - SEX (M/F)
 - AGE2 (-44/45-)
 - Physical health hazards of work PHYS (0/1)
- **Table 8.2** Lehtonen&Pahkinen (2004)
 - PHYS2 proportion estimated for eight subgroups (classes)
- Statistical inference: To identify statistically significant sources of variation of class proportions of PSYCH2 according to the three predictors



OHC-survey: Frequency table (Lehtonen&Pahkinen 2004) Logit-ANOVA

Table 8.2 Proportion \hat{p}_j of persons in the upper psychic strain group, with standard error estimates $s.e_j$ and design-effect estimates \hat{d}_j of the proportions, and domain sample sizes \hat{n}_j and the number of sample clusters m_j (the OHC Survey).

Domain j	SEX	AGE	PHYS	\hat{p}_j	s.e _{j}	\hat{d}_j	\hat{n}_j	m_j
1	Males	-44	0	0.419	0.0128	1.16	1734	230
2			1	0.472	0.0145	1.33	1578	198
3	Females	45-	0	0.461	0.0178	0.88	690	186
4			1	0.520	0.0247	1.18	483	138
5			0	0.541	0.0125	1.23	1966	240
6			1	0.620	0.0270	1.38	447	152
7		45-	0	0.532	0.0236	1.65	740	185
8			1	0.700	0.0391	1.48	203	101
All					0.500	0.0073	1.69	7841



Saturated logistic model

■ Logit ANOVA model

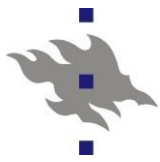
$$\begin{aligned} \text{logit}(P) = & \text{INTERCEPT} + \text{SEX} + \text{AGE2} + \text{PHYS} \\ & + \text{SEX} * \text{AGE2} + \text{SEX} * \text{PHYS} + \text{AGE2} * \text{PHYS} \\ & + \text{SEX} * \text{AGE2} * \text{PHYS} \end{aligned}$$

where

$$P = \text{Prob}(\text{Psych2} = 1 \mid X)$$

Unknown proportion parameter

Probability of belonging to the **more severe** psychic strain class



Reduced logit ANOVA model

- Main effects model

$$\text{logit}(P) = \text{INTERCEPT} + \text{SEX} + \text{AGE2} + \text{PHYS}$$

- NOTE: None of the interaction terms appear statistically significant
- **Table 8.4** Lehtonen and Pahkinen (2004)



Table 8.4 Estimates from design-based logit ANOVA on overall psychic strain (model fitting by the GWLS method).

Model term	Beta coefficient	Design effect	Standard error	<i>t</i> -test	<i>p</i> -value	Odds ratio	95% confidence interval for OR	
							Lower	Upper
Intercept	-0.3282	1.32	0.0635	-7.02	0.0000	0.72	0.66	0.79
Sex								
Males*	0	n.a.	0	n.a.	n.a.	1	1	1
Females	0.4663	1.44	0.0579	8.06	0.0000	1.59	1.42	1.79
Age								
-44*	0	n.a.	0	n.a.	n.a.	1	1	1
45-	0.1385	1.23	0.0570	2.43	0.0159	1.15	1.03	1.28
Physical health hazards								
No*	0	n.a.	0	n.a.	n.a.	1	1	1
Yes	0.2568	1.30	0.0574	4.48	0.0000	1.29	1.16	1.45

* Reference class; parameter value set to zero.

n.a. not available.



Odds Ratio (OR)

Odds Ratio estimation

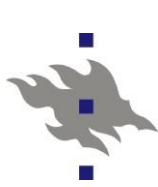
Sex-age adjusted OR for PHYS

$$\text{OR}(\hat{\beta}_3) = \exp(\hat{\beta}_3) = \exp(0.2568) = 1.29$$

where

$\hat{\beta}_1$ is the estimated regression coefficient
for variable PHYS

Interpretation: The probability to belong to the more severe PSYCH2 class is 1.29 times larger for persons who experience physical health hazards of work than for persons who do not experience such hazards



VLISS

Virtual Laboratory in Survey Sampling

- *Practical Methods for Design and Analysis of Complex Surveys.*

Risto Lehtonen and Erkki Pahkinen

- **TRAINING KEY 277: Logit ANOVA**

- In **Training Key 277**, design-based logit ANOVA modelling is examined reproducing the results of Example 8.1. A step-wise ANOVA model building procedure is demonstrated. A program for generalized weighted least squares (GWLS) estimation is examined in detail. The Occupational Health Care Survey data set is used.



Logit ANOVA: technical summary

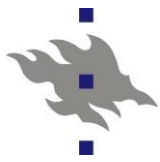
- Lehtonen&Pahkinen (2004)
- 8.3 ANALYSIS OF CATEGORICAL DATA
 - Design-based GWLS Estimation
 - Goodness of Fit and Related Tests
 - Unstable Situations
 - Residual Analysis
 - Design Effect Estimation
- Example 8.1



EXAMPLE

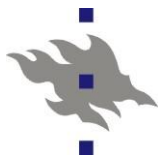
Lehtonen&Pahkinen (2004) Example 8.2

- **Design-based logistic ANCOVA**
- 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)
- Discrete predictors
 - SEX (M/F)
- Continuous predictor
 - AGE (in years)
- Binary predictors
 - Physical health hazards of work PHYS (0/1)
 - Chronic morbidity CHRON (0/1)



Logistic model 1

■ Logit ANCOVA model

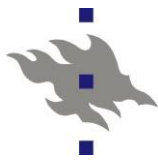
$$\text{logit}(P) = \text{INTERCEPT} + \text{SEX} + \text{AGE} + \text{PHYS} \\ + \text{CHRON} + \text{SEX*AGE} + \text{SEX*PHYS} + \text{SEX*CHRON}$$

where

$$P = \text{Prob}(\text{Psych2} = 1 \mid X)$$

Unknown proportion parameter

Probability of belonging to the **more severe** psychic strain class



Logistic model 2

- Estimation of model parameters
 - PML method (Pseudolikelihood)
 - Accounting for stratification and clustering)

- SAS/SURVEYLOGISTIC

- Final (reduced) model

$$\text{logit}(P) = \text{INTERCEPT} + \text{SEX} + \text{AGE} \\ + \text{PHYS} + \text{CHRON} + \text{SEX} * \text{AGE}$$



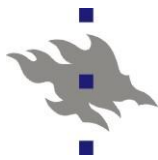
Lehtonen & Pahkinen (2004) Table 8.8

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

Model term	Beta coefficient	Design effect	Standard error	<i>t</i> -test	<i>p</i> -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.



Odds Ratio OR

- Sex-age adjusted Odds Ratio OR
(design-based 95% confidence interval):

$$\text{OR(PHYS)} = 1.32 (1.17, 1.48)$$

$$\text{OR(CHRON)} = 1.76 (1.57, 1.97)$$

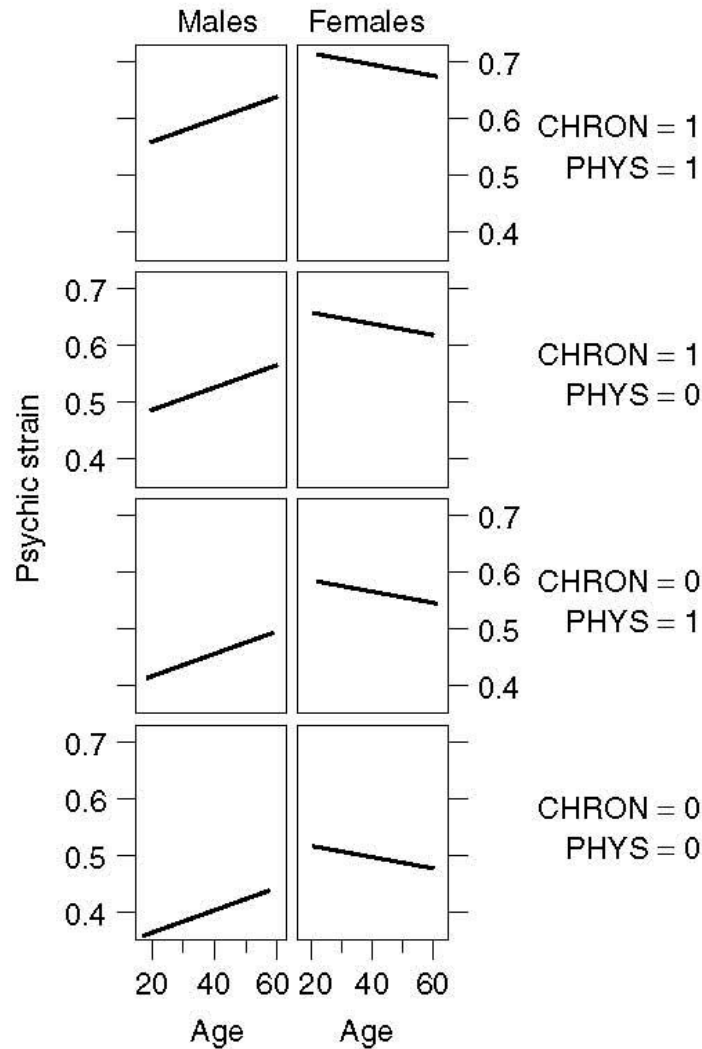


Figure 8.2 Fitted proportions of falling into the high psychic strain group for the final logistic ANCOVA model.



VLISS

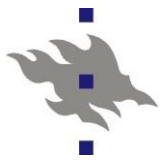
Virtual Laboratory in Survey Sampling

- *Practical Methods for Design and Analysis of Complex Surveys.*

Risto Lehtonen and Erkki Pahkinen

- **TRAINING KEY 288: Logistic ANCOVA**

- In **Training Key 288**, logistic analysis of covariance (ANCOVA) is demonstrated for a binary response variable and the results of Example 8.2 are reproduced. Pseudolikelihood (PML) estimation is used for the OHC Survey data set, accounting for the sampling complexities. An option is provided for a detailed examination of the role of interaction effects in a logistic ANCOVA model.



Logit ANCOVA: technical summary

- Lehtonen&Pahkinen (2004)

- 8.4 LOGISTIC AND LINEAR REGRESSION

- Design-based and Binomial PML Methods
- Logistic Regression
- Example 8.2
- SAS application



SAS Procedure SURVEYLOGISTIC: Program code

Logistic ANCOVA model
Reduced (final) model

```
proc surveylogistic data=ohc;  
title1 "Design-based analysis";  
strata stratum; * Stratification;  
cluster PSU; * Clustering;  
class sex / param=ref;  
model psych2(event=last)=sex age phys  
chron sex*age / link=logit rsquare;  
run;
```



SPSS: Analyze / Complex samples / Prepare for analysis: CSPLAN file creation

```
GET
```

```
FILE='F:\Root\USB\HY\Social Statistics Course 2010 and 2011\  
Course 2011\SPSS\ohc.sav'.
```

```
DATASET NAME DataSet1 WINDOW=FRONT.
```

```
* Analysis Preparation Wizard.
```

```
CSPLAN ANALYSIS
```

```
/PLAN FILE='F:\Root\USB\HY\Social Statistics Course 2010 and 2011\  
Course 2011\SPSS\OHC.csaplan'
```

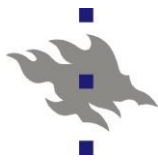
```
/PLANVARS ANALYSISWEIGHT=wd
```

```
/SRSESTIMATOR TYPE=WOR
```

```
/PRINT PLAN
```

```
/DESIGN STRATA=STRATUM CLUSTER=PSU
```

```
/ESTIMATOR TYPE=WR.
```



SPSS: Analyze / Complex samples / Logistic regression: Program code for logistic ANCOVA

```
* Complex Samples Logistic Regression.
CSLOGISTIC PSYCH2(HIGH) BY SEX WITH AGE PHYS CHRON
  /PLAN FILE='F:\Root\USB\HY\Social Statistics Course 2010 and 2011\
Course 2011\SPSS\OHC.csaplan'
  /MODEL SEX AGE PHYS CHRON AGE*SEX
  /INTERCEPT INCLUDE=YES SHOW=YES
  /STATISTICS PARAMETER EXP SE CINTERVAL TTEST DEFF
  /TEST TYPE=F PADJUST=LSD
  /MISSING CLASSMISSING=EXCLUDE
  /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-006 RELATIVE]
LCONVERGE=[0] CHKSEP=20 CILEVEL=95
  /PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```



Mplus: Program code for logistic regression

TITLE:

Mplus Logistic Regression for OHC Survey data;
COMPLEX type analysis for cluster correlated data;

DATA:

FILE IS

"H:\Root\USB\HY\Social Statistics Course 2010 and 2011\Course 2011\Data\Mplus\OHC.inp";

TYPE IS INDIVIDUAL;

VARIABLE:

NAMES ARE ID STRATUM SEX AGE AGE2 PHYS CHRON PSYCH2 PSU;

USEVARIABLES ARE STRATUM PSU PSYCH2 AGE PHYS CHRON SEX01 INT;

CATEGORICAL IS PSYCH2;

STRATIFICATION IS STRATUM;

CLUSTER IS PSU;

DEFINE:

SEX01=SEX-1;

INT=SEX01*AGE;

ANALYSIS:

TYPE IS COMPLEX;

ESTIMATOR IS MLR;

LINK IS LOGIT;

ITERATIONS = 1000;

CONVERGENCE = 0.00005;

MODEL:

PSYCH2 ON SEX01 AGE PHYS CHRON INT;

OUTPUT: SAMPSTAT CINTERVAL;



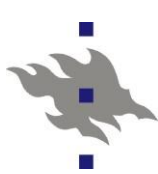
Logit ANCOVA: Computation

- SAS Procedure SURVEYLOGISTIC
 - Run SAS code for SURVEYLOGISTIC
 - [SAS output](#)
- SPSS Complex Samples module
 - Specify sampling design CSPLAN
 - Run logistic regression
 - [SPSS output](#)
- Mplus Logistic regression analysis
 - Specify and run COMPLEX type analysis
 - [Mplus output](#)
- SAS/SURVEYLOGISTIC, SPSS/Complex Samples and Mplus/COMPLEX
 - Similar numerical results!



Comparative analysis with model-based methods

- Generalized linear models
 - SAS Procedure GENMOD
- Generalized linear mixed models
 - SAS Procedure GLIMMIX
 - Logistic mixed models



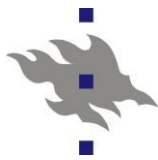
Model-based analysis: GENMOD

■ SAS Procedure GENMOD

- Generalized linear models
- Accounting for clustering effect with the GEE method

- GENMOD fits generalized linear models, as defined by Nelder and Wedderburn ([1972](#)). The class of generalized linear models is an extension of traditional linear models that allows the mean of a population to depend on a **linear predictor** through a nonlinear **link function** and allows the response probability distribution to be any member of an exponential family of distributions.

■ PROC GENMOD



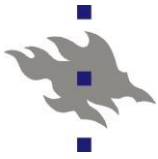
Model-based analysis

PROC GENMOD

Logistic ANCOVA model

Reduced (final) model

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

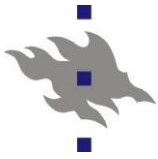


PROC GENMOD

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



Model-based analysis: GLIMMIX

- SAS Procedure [GLIMMIX](#)
 - Logistic mixed model
- Accounting for clustering effect
 - Mixed model formulation with cluster-specific random intercepts
 - Logistic variance components (vc) model



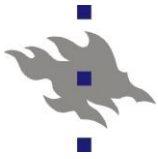
Model-based analysis

PROC GLIMMIX

Logistic mixed ANCOVA model

Reduced (final) model

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



PROC GLIMMIX

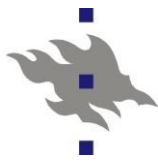
Effect	Gender	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		0.2292	0.1531	249	1.50	0.1355
SEX	1	-1.0334	0.2007	7586	-5.15	<.0001
SEX	2	0
AGE		-0.00565	0.003946	7586	-1.43	0.1521
PHYS		0.3025	0.05966	7586	5.07	<.0001
CHRON		0.5609	0.05717	7586	9.81	<.0001
AGE*SEX	1	0.01437	0.005002	7586	2.87	0.0041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



Comparison of results

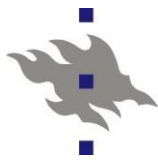
- Interaction term AGE*SEX

- SAS Procedures
 - SURVEYLOGISTIC
 - design-based
 - GENMOD
 - model-based with GEE estimation
 - GLIMMIX
 - model-based with mixed model specification



Comparison of results

	Model term	Beta coefficient	Standard error	Test statistic	p-value
Analysis accounting for clustering					
SURVEYLOGISTIC	AGE*SEX	0.0131	0.0051	2.56	0.0111
GENMOD	AGE*SEX	0.0142	0.0050	2.86	0.0043
GLIMMIX	AGE*SEX	0.0144	0.0050	2.87	0.0041
Analysis ignoring clustering (SRS based)					
SRS based analysis	AGE*SEX	0.0131	0.0043	9.2507	0.0024

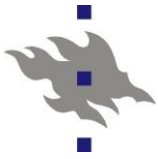


Conclusion

- Design-based analysis SURVEYLOGISTIC
 - Accounting for stratification and clustering effect
 - Most conservative (largest p-value)

- Model-based methods GENMOD, GLIMMIX
 - Accounting for clustering effect
 - Similar results in both cases

- SRS-based analysis
 - Overly liberal
 - SRS assumption obviously wrong in this case

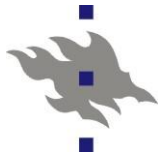


Capabilities of software: Aspects

- Coverage of model types
 - MLM - Multilevel modelling (Mixed models)
 - SEM analysis - Structural Equation Models

- Coverage of members of GLMM's
 - Continuous responses - Linear models
 - Binary responses - Binomial logistic models
 - Polytomous responses - Multinomial logistic models
 - Count data - Poisson regression models

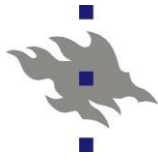
- Accounting for research design complexities
 - Stratification
 - Clustering
 - Weighting



Capabilities of selected software 1

(adjusted from Chantala et al. 2005)

	SEM Analysis	MLM Analysis	Adjust for Clustering	Adjust for Stratification
MPLUS	Yes	Yes	Yes	Yes
LISREL	Yes	Yes	Yes	Yes
GLLAMM (Stata)	Yes	Yes	Yes	
MLWIN		Yes	Yes	
HLM		Yes	Yes	
MIXED (SAS)		Yes	Yes	
GLIMMIX (SAS)		Yes	Yes	



Capabilities of selected software 2

(adjusted from Chantala et al. 2005)

	Normal	Binary	Poisson	Multinomial Categorical	Ordered Categorical
MPLUS	Yes	Yes	Yes		
LISREL	Yes				
GLLAMM (Stata)	Yes	Yes	Yes	Yes	Yes
MLWIN	Yes	Yes	Yes	Yes	Yes
HLM	Yes	Yes	Yes	Yes	Yes
MIXED (SAS)	Yes				
GLIMMIX (SAS)	Yes	Yes	Yes	Yes	Yes



Literature

- 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.
- Goldstein H. (2003). *Multilevel Statistical Models*. 3rd edition. London: Arnold; New York: John Wiley & Sons.
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Second Edition. Chichester: Wiley. Chapters 5, 7-8