



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi

Kevät 2012

TEEMA 4: Asetelmaperusteinen monimuuttuja-analyysi

Risto Lehtonen

risto.lehtonen@helsinki.fi



Analyysimenetelmiä ja työkaluja

- Lineaariset mallit
 - Regressioanalyysi
 - Varianssianalyysi ANOVA (Analysis of Variance)
 - Kovarianssianalyysi ANCOVA

- Yleistetyt lineaariset mallit
 - Logistiset mallit
 - Poisson-mallit

- YHTEENVETOTAULUKKO



ESIMERKKI

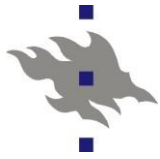
Moniulotteisten frekvenssitaulujen analyysi

- **Asetelmaperusteinen logistinen ANOVA**
- **Moniulotteinen frekvenssitaulu**
 - Usean muuttujan avulla muodostettu moniulotteinen frekvenssitaulu:
- **Epäsymmetrinen tilanne**
 - Yksi diskreetti tulosmuuttuja
 - Binäärinen (0 / 1)
 - Moniluokkainen (>2 luokkaa)
 - Useita diskreettejä selittäjiä
- Tulosmuuttujan ja selittäjien riippuvuusrakenteen mallintaminen logististen mallien avulla



Asetelmaperusteinen analyysi logistisilla malleilla

- **SAS-proseduuri SURVEYLOGISTIC**
- Logistinen malli: Yleistettyjen lineaaristen mallien perheen jäsen
 - *Generalized linear models*
 - Binäärinen (0 / 1) tulosmuuttuja
 - Moniluokkainen tulosmuuttuja
 - Nominaalinen (laatueroasteikko) (A / B / C /...)
 - Ordinaalinen (järjestysasteikko) (1 / 2 / 3 /...)
- Otanta-asetelman ominaisuudet
 - Ositus STRATA-lause
 - Ryvästys CLUSTER-lause
 - Painotus WEIGHT-lause

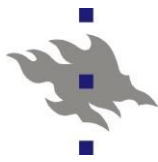


Logistinen ANOVA

- **Logistiset ANOVA-mallit (logit-ANOVA)**
- Yksinkertainen tilanne
 - Binäärinen (0/1) tulosmuuttuja

- **ESIMERKKI: OHC-aineisto**

- Tulosmuuttuja y : PSYCH2
 - 1 - keskimääräistä vakavampi psyykinen rasittuneisuus
 - 0 - keskimääräistä lievempi psyykinen rasittuneisuus



Logistinen ANOVA

- Diskreetit selittäjät (x-muuttujat):
 - Sukupuoli SEX (M/F)
 - Ikä AGE2 (-44/45-)
 - Työn fysikaaliset haitat PHYS (0/1)

- **Table 8.2** Lehtonen&Pahkinen (2004)
 - Taulukossa on 8 osajoukkoa
 - Tavoite: Tutkitaan, missä määrin ja miten tulosmuuttujan PSYCH2 osuudet vaihtelevat selittäjämuuttujien mukaan

- **Table 8.4** Lehtonen and Pahkinen (2004)
 - Beta-kertoimet, keskivirheet, t-testit
 - Ristitulosuhteet (Odds ratio, OR) ja 95 % lv

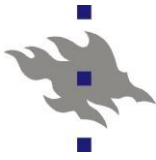


Table 8.2 Proportion \hat{p}_j of persons in the upper psychic strain group, with standard error estimates $s.e_j$ and design-effect estimates \hat{d}_j of the proportions, and domain sample sizes \hat{n}_j and the number of sample clusters m_j (the OHC Survey).

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		-44	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	250



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.



Tilastollinen malli: Logistinen kiinteiden tekijöiden malli

Tulosmuuttuja y alkion y_k : $y_k = 1$ jos tutkittava ilmiö tapahtuu
 $y_k = 0$ muulloin

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

missä $\mathbf{x}_k = (1, x_{1k}, \dots, x_{pk})'$ on selittävien muuttujien arvojen vektori alkion k

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ on estimoitavien parametrien vektori

Estimointi: PML (*pseudolikelihood*)

SAS/SURVEYLOGISTIC



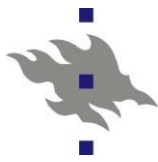
Logistinen kiinteiden tekijöiden malli

Logistisen mallin vaihtoehtoinen muoto

Yksinkertainen tilanne: Yksi selittävä x-muuttuja

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

missä β_0 on mallin kiinteä (*fixed*) vakiotermi (*intercept*)
 β_1 on “kulmakerroin” (*slope*)



Logistinen sekamalli

Logistinen sekamalli (*mixed model*) (monitasomalli)

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

missä u_{0d} on satunnainen vakiotermi (*random intercept*)

Sisäkorreloituneisuus hallitaan satunnaistermillä u_{0d}

u_{0d} : Oletetaan normaalijakauma $N(0, \sigma_u^2)$

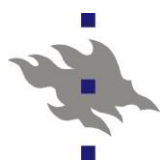
Estimointi: ML

SAS/GLIMMIX



Logistisen kiinteiden tekijöiden mallin parametrien asetelmaperusteinen estimointi

- GWLS-estimointi – ei-iteratiivinen menetelmä
 - Painotettu PNS – *Generalized weighted least squares*
 - Soveltuu logit-ANOVA-malleille
 - (Ei enää käytössä SAS-ohjelmistossa)
- PML-estimointi – yleisimmin käytetty menetelmä
 - Pseudo-uskottavuus – *Pseudo maximum likelihood*
 - Iteratiivinen menetelmä
 - Analyysipainot tuodaan mukaan estimointiproseduriin
 - SAS/SURVEYLOGISTIC, GENMOD, ym.
- GEE-estimointi – vaihtoehto PML-menetelmälle
 - Yleistetyt estimointiyhtälöt –
Generalized estimating equations, GEE
 - Vaihtoehtona proseduurissa SAS/GENMOD
 - Alunperin paneeliaineistoja varten (autokorrelaatio)



ESIM: Logistisen mallin parametrivektorin GWLS-estimaattori

Design-based GWLS Estimation

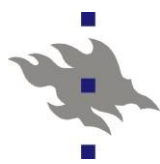
Under the design-based option, a consistent *GWLS estimator* $\hat{\mathbf{b}}_{des}$, denoted $\hat{\mathbf{b}}$ for short in this section, of the $s \times 1$ model coefficient vector \mathbf{b} for a model $F(\mathbf{p}) = \mathbf{X}\mathbf{b}$ is given by

$$\hat{\mathbf{b}} = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}), \quad (8.5)$$

where $\hat{\mathbf{V}}_{des}$ is a consistent estimator of the covariance matrix of the consistent domain proportion estimator vector $\hat{\mathbf{p}}$, and $\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H}$ is a covariance-matrix estimator of the function vector $F(\hat{\mathbf{p}})$. An estimate $\hat{\mathbf{V}}_{des}$ is obtained using, for example, the linearization method as described in Chapter 5. The GWLS estimating

```
SAS/IML-toteutus
```

```
b=ginv(x`*ginv(h*vdesp*h)*x)*x`*ginv(h*vdesp*h)*logitp;
```



Mallin parametreja koskevien hypoteesien testaus: Asetelmaperusteinen Waldin testi

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

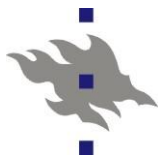
joka on asympotoottisesti χ^2 -jakautunut vapausastein $df=1$

Termi $\hat{\beta}_j$ on estimoitu logit-regressiokerroin (esim. PML)

Termi $v_{des}(\hat{\beta}_j)$ on **asetelmaperusteisesti** estimoitu varianssi (esim. linearisointimenetelmä, jackknife, bootstrap)

$$\text{Vastaava t-testisuure } t_{des}(\beta_j) = \frac{\hat{\beta}_j}{\text{s.e}_{des}(\hat{\beta}_j)}$$

on Waldin testisuureen merkinen neliöjuuri



Logistinen ANOVA: Tekninen tarkastelu

- Logistisen mallin parametrien estimointimenetelmät
 - GWLS
 - PML
 - GEE
- Laskentatyökalut
 - SAS / IML
 - SAS / SURVEYLOGISTIC
- EXAMPLE 8.1 (Lehtonen-Pahkinen 2004)
- [Diat 4b](#)



Logistinen ANOVA, tilastometodinen kuvaus

- 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



Logistinen ANCOVA, tilastometodinen kuvaus

- Lehtonen&Pahkinen (2004)

- 8.4 LOGISTIC AND LINEAR REGRESSION

- Design-based and Binomial PML Methods
- Logistic Regression

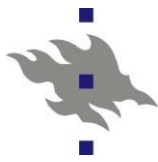
- Example 8.2



ESIMERKKI

Lehtonen&Pahkinen (2004) Example 8.2

- Asetelmaperusteinen logistinen ANCOVA
- OHC Survey
- Ositettu ryväsotanta-asetelma
 $H = 5$ ositetta
 $m = 250$ toimipaikkaa (otosryvästä)
 $n = 7841$ otoshenkilöä



Asetelmaperusteinen logistinen ANCOVA

- Binäärinen tulosmuuttuja:
PSYCH2 Psykkinen rasittuneisuus
 - 0: Lievä (alle mediaanin)
 - 1: Vakava (yli mediaanin)
- Diskreetti selittäjä
 - Sukupuoli SEX (M/F)
- Jatkuva selittäjä
 - Ikä AGE (vuosina)
- Binääriset selittäjät
 - Työn fysikaaliset haitat: PHYS (0/1)
 - Pitkäaikaissairastavuus: CHRON (0/1)



Tilastollinen malli

- Logistinen ANCOVA-malli

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

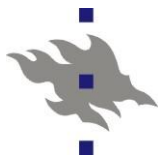
missä $\text{logit}(P) = \log(P/(1-P))$ on “vedonlyöntisuhteen” (Odds Ratio) logaritmi

OR: “ristitulosuhte”, “suhteellinen riski”

Mallinnetaan todennäköisyyttä:

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

Todennäköisyys kuulua vakavamman psyykkisen rasittuneisuuden luokkaan



Lopullinen (redusoitu) malli

- Eksploratiivinen analyysi tuotti lopulliseksi malliksi:

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

missä SEX, AGE, PHYS ja CHRON
ovat mallin **päävaikutustermejä**

SEX*AGE on **yhdysvaikutustermi**

- Mallin parametrivektorin estimointi
 - PML-estimointi (Pseudolikelihood)
 - SAS/SURVEYLOGISTIC



SAS Procedure SURVEYLOGISTIC

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



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.



Suhteellinen riski Odds Ratio OR

- Sukupuoli-ikävakioitu suhteellinen riski
Odds Ratio, OR
(asetelmaperusteinen 95% luottamusväli):

$$\text{OR(PHYS)} = \exp(0.2765) = 1.32 (1.17, 1.48)$$

$$\text{OR(CHRON)} = \exp(0.5641) = 1.76 (1.57, 1.97)$$

- Tulkinta: Henkilöillä, joilla on pitkäaikainen sairaus, on 1.76ertainen riski kuulua vakavamman psyykkisten rasittuneisuuden luokkaan verrattuna henkilöihin joilla sairautta ei ole

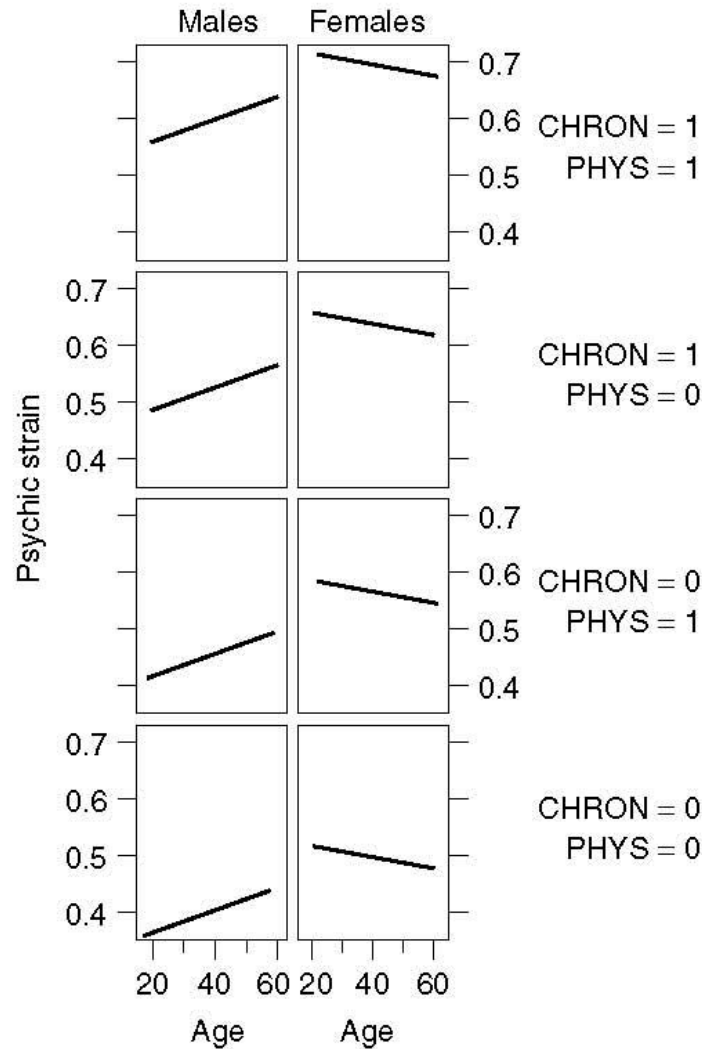
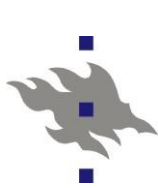


Figure 8.2 Fitted proportions of falling into the high psychic strain group for the final logistic ANCOVA model.



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.