HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi
(78136 , 78405)
Kevät 2010
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 logit-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 logitmalleilla

- **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

Logit-ANOVA-mallit

- **Logit-ANOVA-mallit**
- Yksinkertaisin tilanne
  - Binäärinen (0/1) tulosmuuttuja
- **ESIMERKKI**: OHC-aineisto
- Tulosmuuttuja *y*: PSYCH2
  - 1 - keskimääräistä vakavampi psyykkinen rasittuneisuus
  - 0 - keskimääräistä lievempi psyykkinen rasittuneisuus

Logit-ANOVA-mallit

- 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)
  - Tulokset

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

Table 8.2
Proportion *p* of persons in the upper psychic strain group, with standard error estimates s.e and design-effect estimates *deff* of the proportions, and domain sample sizes *n* and the number of sample clusters *m* (the OHC Survey).

Domain	SEX	AGE	PHYS	*p*	s.e	*deff*	*n*	*m*
1	Males	–44	0	0.419	0.0128	1.16	1734	230
2			1	0.472	0.0145	1.33	1578	198
3		45–	0	0.461	0.0178	0.88	690	186
4			1	0.520	0.0247	1.18	483	138
5	Females	–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	*t*-test	*p*-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
Logitmalli (logistinen malli)

Tulosmuuttuja *y* alkiolle 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 $\mathbf{x}_k = (1, x_{1k}, ..., x_{pk})'$ on selittävien muuttujien arvojen vektori alkiolle *k*

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

Tilastollinen malli
Logitmalli (logistinen malli)

Logitmalli

Vaihtoehtoinen muoto
Yksinkertainen tilanne: Yksi selittävä muuttuja *x*

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

missä β_0 on mallin vakiotermi (*intercept*)
β_1 on "kulmakerroin" (*slope*)

ESIMERKKI

Kiinteiden tekijöiden logitmalli

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

missä β_0 on mallin kiinteä vakiotermi (*intercept*)
β_1 on "kulmakerroin" (*slope*)

Monitasomalli (sekamalli)

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

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

Logitmallin parametrien asetelmaperusteinen estimointi

- GWLS-estimointi – ei-iteratiivinen menetelmä
  - Painotettu PNS – *Generalized weighted least squares*
- PML-estimointi – yleisimmin käytetty menetelmä
  - Pseudo-uskottavuus – *Pseudo maximum likelihood*
  - Iteratiivinen menetelmä
  - SAS/SURVEYLOGISTIC, GENMOD, ym.

Logitmallin parametrien asetelmaperusteinen estimointi

- GEE-estimointi – vaihtoehto PML-menetelmälle
  - Yleistetyt estimointiyhtälöt – *Generalized estimating equations*
  - SAS/GENMOD (malliperusteinen)

Asetelmaperusteinen Waldin testisuure

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

joka on asymptoottisesti 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)

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

on Waldin testisuureen merkkinen neliöjuuri

Logit ANOVA: Tekninen tarkastelu

- Logitmallin parametrien estimointimenetelmät
  - GWLS
  - PML
  - GEE
- Laskentatyökalut
  - SAS / IML
  - SAS / SURVEYLOGISTIC
- EXAMPLE 8.1 (Lehtonen-Pahkinen 2004)
- Diat 5b

Logit 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

Logit 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 Psyykkinen 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

- Logit-ANCOVA-malli

  logit(P) = INTERCEPT + SEX + AGE + PHYS + CHRON + SEX*AGE + SEX*PHYS + SEX*CHRON

missä

  P = Prob(Psych2 = 1 | X)
  Tuntematon osuusparametri

  Todennäköisyys kuulua vakavamman psyykkisen rasittuneisuuden luokkaan

Tilastollinen malli

- Mallin parametrivektorin estimointi
  - PML-estimointi
    Pseudolikelihood
- SAS/SURVEYLOGISTIC
- Lopullinen redusoitu malli:

  logit(P) = INTERCEPT + SEX + AGE
  + PHYS + CHRON + SEX*AGE

SAS Procedure SURVEYLOGISTIC

```
proc surveylogistic data=ohc;
strata stratum;
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	*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.

23

Suhteellinen riski
Odds Ratio OR

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

  OR(PHYS) = 1.32 (1.17, 1.48)

  OR(CHRON) = 1.76 (1.57, 1.97)

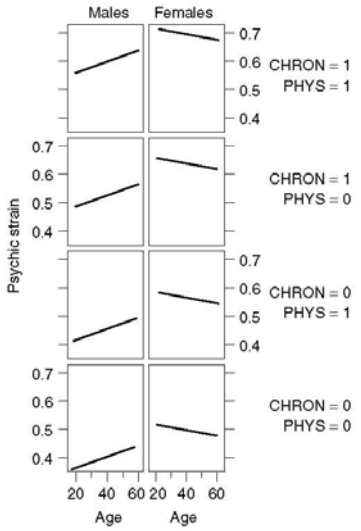

Figure 8.2 Fitted proportions of falling into the high psychic strain group for the final logistic ANCOVA model.

25

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.

26 21.10.2008 Risto