

Introduktion til Logistisk Regression

Morten Frydenberg

© 2006 Afdeling for Biostatistik, Århus Universitet

MPH 1. studieår specialmodul 4

Cand. San. uddannelsen 1. studieår

1. gang:

Hvorfor logistisk regression

Præsentation af Framingham databasen.

Repetition af Odds Ratio beregning i 2x2 tabeller.

Repetition af stratificeret (Mantel-Haenszel) analyse.

Logistic regression med en og to binære forklarende variable.

1

Probability of Developing Coronary Heart Disease in 6 years.

Women (Aged 35-70)

Systolic Blood Pressure

Left Ventricular Hypertrophy

Glucose Intolerance

Age

Serum Cholesterol

160

No

Yes

70

270

Points

012345678910

SBP

100110120130140150160170180190200

6

LVH

NoYes

0

GLU

NoYes

3

Note: No points added for Smoking

2

Age

Chol

363840424446485055606570

165

024579101215182021

180

134689111216182022

195

2357910121316192122

210

34681011131417192122

225

45791012131517202123

240

468101113141518202223

255

579111213151619212223

270

6810111314161719212324

285

7911121415161720222324

300

81012131516171821222424

315

91113141517181921232425

9

4

24

Interaktion/effekt modifikation !!

3

Calculation of Probability

Enter Points (in red) for

6

0

3

24

33

Systolic Blood Pressure

Left Ventricular Hypertrophy

Glucose Intolerance

Age/Serum Cholesterol

Total Points

TP

Prob

TP

Prob

TP

Prob

5

0.002

17

0.014

29

0.081

6

0.003

18

0.016

30

0.093

7

0.003

19

0.019

31

0.110

8

0.004

20

0.022

32

0.120

9

0.004

21

0.025

33

0.140

10

0.004

22

0.029

34

0.160

11

0.006

23

0.034

35

0.180

12

0.007

24

0.040

36

0.200

13

0.008

25

0.046

37

0.230

14

0.009

26

0.053

38

0.260

15

0.010

27

0.061

39

0.290

16

0.012

28

0.071

40

0.320

6

FRAMINGHAM HEART STUDY

Follow-up undersøgelse påbegyndt i 1948.

Planlagt som et 20 års studie.

Forekomst af cardio-vaskulære sygdomme.

Klinisk undersøgelse hvert andet år.

5209 personer (inkl. 740 frivillige).

Vi/I har :

Oplysninger fra de 10 første gen-undersøgelser (18 års follow-up) for 1406 personer.

Udeladt: Personer under 45 år ved 1. us..

Udeladt: Serum kolesterol værdi mangler ved 1. us..

5

Vi/I har 13 ud af de oprindelige 394 variable.

OBS, SEX, CHD, AGE, SBP, SBP10, DBP, CHOL, FRW, CIG, YRS_CHD, YRS_DTH, DEATH, CAUSE.

SEX

1,2

1 for mand, 2 for kvinde

CHD

0,...10

0: ingen CHD diag. ved de 10 første us.
1: CHD diag. ved 1. us (prævalent case)
2-10: us.-nummer ved første CHD diag.

AGE

45-62

Alder i år ved første us.

SBP

90-300

Systolisk blodtryk ved første us.

CIG

0-60

0: Ikke ryger
1-60: Antal cigaretter pr. dag

6

Hvad vil vi se på?

Analysere den **kumulerede incidens** af hjertesygdom.

Tidsperiode: Opfølgningsperioden efter 1 us.!

Risikopopulation: **1363** personer uden CHD ved 1. us.
43 prævalente person udelades.

Risikofaktorer/indikatorer:

Systolisk **blodtryk**: over/under 160 mmHg.

Rygning: ja/nej.

Alder: 45-48, 49-52, 53-56 og over 56 år.

Køn.

7

STATA Syntax:

```
use fram, clear

generate chdever=chd>1                                if( chd!=.)
generate age4=(age>48)+(age>52)+(age>56)             if( age!=.)
generate sys160=sbp>160                               if( sbp!=.)
generate smoke=cig>0                                  if( cig!=.)

drop if chd==1

tabodds chdever sex, base(2) w or
tabodds chdever sys160, base(1) w or
tabodds chdever smoke if smoke!=., base(1) w or
tabodds chdever age4, base(1) w or
```

"if (cig!=.)" sikrer at missing **cig** giver missing i **smoke**
Check **altid** hvordan 'dit' program reagerer på missing!!!

8

Odds ratio beregning:

SygRaskTotal

+Exp

a

b

-Exp

c

d

Total

a+c

b+c

a+b+c+d

$$OR = \frac{a \cdot d}{b \cdot c}$$

SygRask

Mand

164

479

Kvinde

104

616

$$OR = \frac{164 \cdot 616}{479 \cdot 104} = 2.03$$

RaskSyg

Mand

479

164

Kvinde

616

104

$$OR = \frac{479 \cdot 104}{164 \cdot 616} = 0.493 = \frac{1}{2.03}$$

9

CHDEVER

yes (1)no (0)Total

kum.incidentsORln(OR)

SEXmale (1)

164

479

643

0.2552.030.71

female (2)

104

616

720

0.144ref1.000.00

Total

268

1095

1363

2.03 = 1/0.493

CHDEVER

yes (1)no (0)Total

kum.incidentsORln(OR)

SYS160>160 (1)

101

215

316

0.3202.480.91

=<160 (0)

167

880

1047

0.160ref1.000.00

Total

268

1095

1363

CHDEVER

yes (1)no (0)Total

kum.incidentsORln(OR)

SMOKEyes (1)

134

480

614

0.2181.280.25

no (0)

134

614

748

0.179ref1.000.00

Total

268

1094

1362

CHDEVER

yes (1)no (0)Total

kum.incidentsORln(OR)

AGE445-48 (0)

51

308

359

0.142ref1.000.00

49-52 (1)

61

298

359

0.1701.240.21

53-56 (2)

64

254

318

0.2011.520.42

57-62 (3)

92

235

327

0.2812.360.86

Total

268

1095

1363

10

Et eksempel på beregning i STATA.

```
cs chdever sys160, w or
```

	sys160		
	Exposed	Unexposed	Total
Cases	101	167	268
Noncases	215	880	1095
Total	316	1047	1363
Risk	.3196203	.1595033	.1966251
	Point estimate	[95% Conf. Interval]	
Risk difference	.1601169	.1041216	.2161122
Risk ratio	2.003847	1.620021	2.478612
Attr. frac. ex.	.5009598	.3827239	.5965483
Attr. frac. pop	.1887946		
Odds ratio	2.475421	1.854928	3.303476 (Woolf)

chi2 (1) = 39.40 Pr>chi2 = 0.0000

11

Tilbage til "vores" problem:

Køn, blodtryk og hjertesygdom.

- Forhøjet risiko for **mænd** estimeret **OR=2.03**.
- Forhøjet risiko hvis **SBP>160** estimeret **OR=2.48**.
- Begge odds ratioer er **ukorrigerede** !

SYS160

>160<=160Total

OR

SEXmale

112

531

643

0.53

female

204

516

720

Total

316

1047

1363

- Mænd ikke så ofte har SBP>160 .
- Dermed undervurderes betydningen af SBP>160 !
- Kønnet er en mulig **konfunder** for SBP>160→CHDEVER.

12


```
logit chdever sys160
```

chdever	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
sys160	.9064	.14723	6.16	0.000	.617846 1.194975
_cons	-1.6612	.08441	-19.69	0.000	-1.82736 -1.496495

β_1 β_0 $se(\hat{\beta}_1) = se(\ln(\widehat{OR}))$

Sikkerhedsinterval for β_1

P-værdi for hypotesen: $\beta_1 = 0$ dvs. $OR=1$.

$\widehat{OR} = \exp(\hat{\beta}_1)$

```
logit chdever sys160 , or
```

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
sys160	2.475421	.0644551	6.16	0.000	1.854928 3.303476

Sikkerhedsinterval for OR

Intet nyt: Det ku' vi godt i forvejen !

Men med logistik regression kan vi mere!!

19

Analyse af en 2x2x2 tabel vha. logistisk regression.

Køn og forhøjet blodtryk : **SEX** og **SYS160**

Mantel-Haenszel analyse kan give:

- OR for **SYS160** korrigeret for **SEX** (har vi lavet)
- OR for **SEX** korrigeret for **SYS160** (har vi ikke lavet)

To analyser for at få de **to** korrigerede estimater.

En logistisk regression vil give **to** korrigerede estimater!

20

SEX	SBP	CHD	CHD	Odds	ln(odds)	OR	ln(OR)
		1	0				
Mand	>160	47	65	0.723	-0.324	6.74	1.908
Mand	≤160	117	414	0.283	-1.264	2.63	0.968
Kvinde	>160	54	150	0.360	-1.022	3.36	1.211
Kvinde	≤160	50	466	0.107	-2.232	1.00	0.00

Referenceværdi $odds(0,2)$, dvs. kvinde med 'lavt' blodtryk

Ny variabel: $MALE = 2 - SEX$ 1 hvis mand og 0 hvis kvinde

Logistisk regression med **SEX** og **SYS160**
som forklarende variable:

$$\ln(odds(SYS160, SEX)) = \beta_0 + \beta_1 \cdot SYS160 + \beta_2 \cdot MALE$$

21

$$\ln(odds(SYS160, SEX)) = \beta_0 + \beta_1 \cdot SYS160 + \beta_2 \cdot MALE$$
$$= \beta_0 + \beta_1 \cdot SYS160 + \beta_2 \cdot (2 - SEX)$$

Hvad er β_0 , β_1 og β_2 ?

$$\ln(odds(0,2)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot (2 - 2) = \beta_0$$
$$\ln(odds(1,2)) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot (2 - 2) = \beta_0 + \beta_1$$
$$\ln(odds(0,1)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot (2 - 1) = \beta_0 + \beta_2$$
$$\ln(odds(1,1)) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot (2 - 1) = \beta_0 + \beta_1 + \beta_2$$

22

$\ln(odds(0,2)) = \beta_0 \Rightarrow \beta_0 = \ln(odds(0,2))$ $\ln(odds(1,2)) = \beta_0 + \beta_1$ $\ln(odds(0,1)) = \beta_0 + \beta_2$ $\ln(odds(1,1)) = \beta_0 + \beta_1 + \beta_2$

$\beta_1 = \ln(odds(1,2)) - \ln(odds(0,2)) = \ln(OR_{SYS160 \text{ for } KVINDER})$ $\beta_1 = \ln(odds(1,1)) - \ln(odds(0,1)) = \ln(OR_{SYS160 \text{ for } MÆND})$

Effekten af forhøjet blodtryk antages er være uafhængig af køn !!

Når den beskrives ved odds ratio !

Ingen effektmodifikation!

Præcis som den stratificerede analyse, Mantel-Haenszel !

23

$\ln(odds(0,2)) = \beta_0$ $\ln(odds(1,2)) = \beta_0 + \beta_1$ $\ln(odds(0,1)) = \beta_0 + \beta_2$ $\ln(odds(1,1)) = \beta_0 + \beta_1 + \beta_2$

$\beta_2 = \ln(odds(0,1)) - \ln(odds(0,2)) = \ln(OR_{SEX \text{ for } SBP < 160})$ $\beta_2 = \ln(odds(1,1)) - \ln(odds(1,2)) = \ln(OR_{SEX \text{ for } SBP > 160})$

Effekten af køn antages er være uafhængig af forhøjet blodtryk !!

Når den beskrives ved odds ratio !

Ingen effektmodifikation!

Præcis som den stratificerede analyse, Mantel-Haenszel !

24

$$\ln(\text{odds}(\text{SYS160}, \text{SEX})) = \beta_0 + \beta_1 \cdot \text{SYS160} + \beta_2 \cdot \text{MALE}$$

$\ln(\text{odds}(\text{SBP} \leq 160, \text{kvinde}))$ $\ln(OR_{\text{SYS160}})$ $\ln(OR_{\text{SEX}})$

Tre ukendte størrelser, der skal findes !?
Det klarer statistikprogrammet for os !

`char sex[omit] 2``char sys160[omit] 0`

`xi: logit chdever i.sys160 i.sex``xi: logit chdever i.sys160 i.sex, or`

`xi` og `i`-erne betyder at disse variable er kategoriske.

25

`xi:logit chdever i.sys160 i.sex`

chdever	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_Isys160_1	1.07529	.1536891	7.00	0.000	.7740655 1.376516
_Issex_1	.874965	.1457309	6.00	0.000	.5893373 1.160592
_cons	-2.1703	.1279629	-16.96	0.000	-2.421124 -1.919519

$\hat{\beta}_1 = \ln(\widehat{OR}_{\text{SYS160}})$ $\hat{\beta}_2 = \ln(\widehat{OR}_{\text{SEX}})$

`xi:logit chdever i.sys160 i.sex, or`

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Isys160_1	2.93084	.4504388	7.00	0.000	2.168565 3.961076
_Issex_1	2.39879	.3495779	6.00	0.000	1.802793 3.191822

$\widehat{OR}_{\text{SYS160}} = 2.931$ $\widehat{OR}_{\text{SEX}} = 2.932$

Korrektion for køn vha. logistiske regression : $\widehat{OR}_{\text{SYS160}} = 2.931$
Den stratificerede (Mantel-Haenszel) analysen gav 2.932
En lille og ubetydelig forskel.

26