

1

Introduktion til Logistisk Regression

Morten Frydenberg, Inst. f. Biostatistik

© 2006 Institut for Biostatistik, Århus Universitet

MPH 1. studieår Specialmodul 4

Cand. San. uddannelsen 1. studieår

2. gang:

- Test og sikkerhedsinterval for en parameter.
- Kategoriske forklarende variable med mere to kategorier.
- Logistisk regression med en kontinuert forklarende variabel.

2

RESUME:

Hvad vil vi se på?

Analysere den **kumulerede incidens** af hjertesygdom.

Tidsperiode: Opfølgningsperioden efter 1 us.!

Risikopopulation: 1363 personer uden CHD ved 1. us.

Risikofaktorer/indikatorer:

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

Rygning: ja/nej.

Alder: I år og AGE4: 45-48, 49-52, 53-56 over 56 år.

Køn.

3

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

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

$$\text{odds}(\text{SYS160}, \text{SEX}) = \text{odds}(0, 2) \cdot (OR_{\text{SYS160}})^{\text{SYS160}} \cdot (OR_{\text{SEX}})^{\text{MALE}}$$

Model antagelse:

- Ingen effektmodifikation/interaktion mellem køn og blodtryk.

```
char sex[omit] 2
char sys160[omit] 0
xi: logit chdever i.sys160 i.sex
```

4

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

chdever	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_sys160_1	1.07529	.1536891	7.00	0.000	.7740655 1.376516
_isex_1	.874965	.1457309	6.00	0.000	.5893373 1.160592
_cons	-2.1703	.1279629	-16.96	0.000	-2.421124 -1.919519

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

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

$\ln(\widehat{\text{odds}}(\text{SBP} \leq 160, \text{kvinde}))$

Hypotesen : $OR_{\text{SYS160}} = 1$ forkastes!

Hypotesen : $OR_{\text{SEX}} = 1$ forkastes!

- Hvordan passer modellen til data ?
- Hvad er de fittede kumulerede incidenser ?

5

På basis af :

$$\ln(\text{odds}(\text{SYS160}, \text{SEX})) = \beta_0 + \beta_1 \cdot \text{SYS160} + \beta_2 \cdot \text{MALE}$$
$$\ln(\widehat{\text{odds}}(\text{SYS160}, \text{SEX})) = -2.1703 + 1.0753 \cdot \text{SYS160} + 0.8750 \cdot \text{MALE}$$

kan man beregne:

$$\text{odds} = \exp(\ln(\text{odds})) \quad p = \frac{\text{odds}}{1 + \text{odds}}$$

SEX	SYS160	Fra den logistisk regressionsanalyse ln(odds)	odds	kum.inc.	Data kum.inc.
male	>160	-2.1703+1.0753+ 0.8750	0.803	0.445	0.420
male	<=160	-2.1703+ 0.8750	0.274	0.215	0.220
female	>160	-2.1703+1.0753	0.335	0.251	0.265
female	<=160	-2.1703	0.114	0.102	0.097

Rimelig god overensstemmelse !

Modellen fitter godt !

Antagelsen om ingen effektmodifikation synes ok !

Næste gang vil vi se på et test for ingen effektmodifikation !

6

I STATA en variable med de fittede værdier laves således

```
predict fitted if e(sample)
```

En tabel over observerede således:

```
table sys160 sex if e(sample) ///
, c(sum fitted sum chdever)
```

sys160	sex	
	male	female
0	114.1371	52.86289
	117	50
1	49.86289	51.13711
	47	54

Sikkerhedsintervaller (CI) (Wolf/Wald- methods)
Ingen estimater uden sikkerhedsintervaller!
(Approximativt) 95% CI: $\text{Estimat} \pm 1.96 \times \text{s.e.}$

SygRask

+Exp

-Exp

a

b

c

d

$$\widehat{OR} = \frac{a \cdot d}{b \cdot c}$$
$$\text{s.e.}(\ln(\widehat{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

95% CI for OR:
Først 95% CI for $\ln(OR)$:
$$\ln(\widehat{OR}) \pm 1.96 \times \text{s.e.}(\ln(\widehat{OR}))$$

Dernæst tag eksponentialfunktionen til disse grænser.

Beregning af sikkerhedsintervaller "i hånden".

Parameter	β	s.e.	95% grænser		OR	95% grænser	
			nedre	øvre		nedre	øvre
SYS160	1.0753	0.1537	0.774	1.377	2.93	2.17	3.96
SEX	0.8750	0.1457	0.589	1.161	2.40	1.80	3.19
Constant	-2.1703	0.1280	-2.421	-1.919	0.11	0.09	0.15

Fra outputEstimat $\pm$ 1.96 $\times$ s.e.Exp

Disse sikkerhedsintervaller er i præcist de, der er angivet i STATA outputtet.

Alder og risiko for hjertesygdom:
AGE4 inddeler i **fire** grupper: 45-48, 49-52, 53-56 over 56 år.

	CHDEVER		Total	kum.incidens	ln(odds)	OR	ln(OR)	
	yes (1)	no (0)						
AGE4 45-48 (0)	51	308	359	0.142	ref	-1.80	1.00	0.00
49-52 (1)	61	298	359	0.170		-1.59	1.24	0.21
53-56 (2)	64	254	318	0.201		-1.38	1.52	0.42
57-62 (3)	92	235	327	0.281		-0.94	2.36	0.86
Total	268	1095	1363					

Fire parametre skal bestemmes.
Fire mulige valg af reference gruppe.
Vi vælger **45-48** som reference gruppe.
De **fire** parameter er så:
• $\ln(\text{odds})$ i reference gruppen.
• $\ln(\text{OR})$ relativt til referencegruppen

Der er brug for **tre** dummy variable:
 $AKAT1 = 1$ hvis $AGE4 = 1$ og 0 ellers
 $AKAT2 = 1$ hvis $AGE4 = 2$ og 0 ellers
 $AKAT3 = 1$ hvis $AGE4 = 3$ og 0 ellers

Modellen :
$$\ln(\text{odds}(AGE4)) = \beta_0 + \beta_1 \cdot AKAT1 + \beta_2 \cdot AKAT2 + \beta_3 \cdot AKAT3$$

Giver følgende:
$$\ln(\text{odds}(AGE4 = 0)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 = \beta_0$$
$$\ln(\text{odds}(AGE4 = 1)) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 0 = \beta_0 + \beta_1$$
$$\ln(\text{odds}(AGE4 = 2)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 + \beta_3 \cdot 0 = \beta_0 + \beta_2$$
$$\ln(\text{odds}(AGE4 = 3)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 1 = \beta_0 + \beta_3$$

$$\ln(\text{odds}(AGE4)) = \beta_0 + \beta_1 \cdot AKAT1 + \beta_2 \cdot AKAT2 + \beta_3 \cdot AKAT3$$
$$\rightarrow \ln(\text{odds}(AGE4 = 0)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 = \beta_0$$
$$\rightarrow \ln(\text{odds}(AGE4 = 1)) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 0 = \beta_0 + \beta_1$$
$$\rightarrow \ln(\text{odds}(AGE4 = 2)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 + \beta_3 \cdot 0 = \beta_0 + \beta_2$$
$$\rightarrow \ln(\text{odds}(AGE4 = 3)) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 1 = \beta_0 + \beta_3$$
$$\beta_1 = \ln(\text{odds}(AGE4 = 1)) - \ln(\text{odds}(AGE4 = 0)) = \ln(OR_{1 \text{ mod } 0})$$
$$\beta_2 = \ln(\text{odds}(AGE4 = 2)) - \ln(\text{odds}(AGE4 = 0)) = \ln(OR_{2 \text{ mod } 0})$$
$$\beta_3 = \ln(\text{odds}(AGE4 = 3)) - \ln(\text{odds}(AGE4 = 0)) = \ln(OR_{3 \text{ mod } 0})$$

I praksis

STATA klarer det for os !
`char age4[omit] 0`
`xi: logit chdever i.age4`

STATA laver 3 nye variable svarende til $AKAT1$, $AKAT2$ og $AKAT3$

I output:
`i.age4 _Iage4_0-3` (naturally coded; `_Iage4_0` omitted)

Age4 har værdier fra 0 til 3

Age4 = 0 er valgt som reference

Estimaterne i STATA:
`xi: logit chdever i.age4`

chdever	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_Iage4_1	.212055	.2064067	1.03	0.304	-.1924952 .6166043
_Iage4_2	.419823	.2059529	2.04	0.042	-.0161627 .8234833
_Iage4_3	.860477	.1948832	4.42	0.000	.4785131 1.242441
_cons	-1.79827	.1511774	-11.90	0.000	-2.09458 -1.501972

Som ved direkte beregning.

P- værdi for hypotese "samme risiko blandt 53-56 og 45-48 årige".
`xi: logit chdever i.age4, or`

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Iage4_1	1.236215	.2551632	1.03	0.304	.824898 1.852626
_Iage4_2	1.521692	.3133969	2.04	0.042	1.016294 2.278422
_Iage4_3	2.364289	.4607602	4.42	0.000	1.613673 3.46406

OR relativt til AGE4=0, dvs. de yngste

I STATA outputtet var der test for

$OR=1$ eller ækvivalent $\ln(OR)=0$

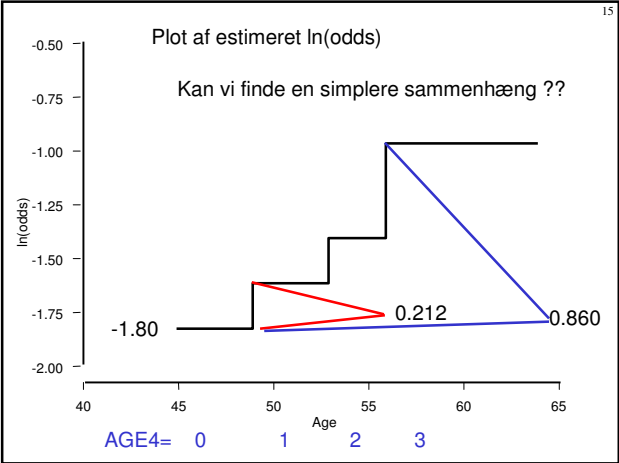
ved sammenligning med **referencegruppen**.

Men hvad med hypotesen: "Alder betyder ikke noget", dvs alle tre $\ln(OR)$ er lig 0?

Dette test kan testes vha. et "Wald-test", som i STATA beregnes således:

```
testparm _Iage4*  
( 1) _Iage4_1 = 0  
( 2) _Iage4_2 = 0  
( 3) _Iage4_3 = 0  
      chi2( 3) = 22.74  
      Prob > chi2 = 0.0000
```

Hypotesen forkastes klart - alder betyder noget for risikoen.



En simplifikation:

- Alle trinene lige store
- Dvs. lineær sammenhæng med AGE4:

$$\ln(\text{odds}(\text{AGE4})) = \beta_0 + \beta_1 \cdot \text{AGE4}$$
$$\beta_1 = \ln(\text{odds}(\text{AGE4}=1)) - \ln(\text{odds}(\text{AGE4}=0)) = \ln(OR_{1 \text{ mod } 0})$$
$$\beta_1 = \ln(\text{odds}(\text{AGE4}=2)) - \ln(\text{odds}(\text{AGE4}=1)) = \ln(OR_{2 \text{ mod } 1})$$
$$\beta_1 = \ln(\text{odds}(\text{AGE4}=3)) - \ln(\text{odds}(\text{AGE4}=2)) = \ln(OR_{3 \text{ mod } 2})$$

Dvs $\beta_1 = \ln(OR_{\text{forskøl p\aa 1}})$

I STATA

```
xi: logit chdever age4  
eller  
xi: logit chdever age4, or
```

Obs intet i. før age4!

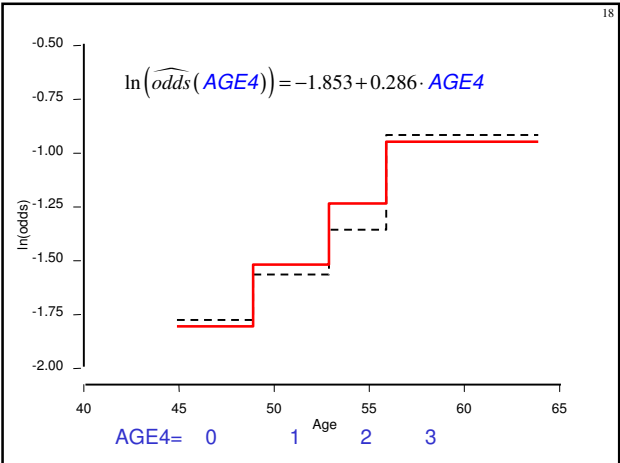
chdever	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age4	.2860376	.061997	4.61	0.000	.1645258 .407549
_cons	-1.853176	.1240054	-14.94	0.000	-2.096222 -1.61013

$\ln(\widehat{\text{odds}}(\text{AGE4}=0))$ $\beta_1 = \ln(\widehat{OR}_{\text{forskøl p\aa 1}})$

$$\ln(\widehat{\text{odds}}(\text{AGE4})) = -1.853 + 0.286 \cdot \text{AGE4}$$
$$\widehat{\text{odds}}(\text{AGE4}) = 0.157 \cdot 1.331^{\text{AGE4}}$$

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age4	1.331143	.0825268	4.61	0.000	1.178834 1.50313

$\widehat{OR}_{\text{forskøl p\aa 1}}$



Det så da meget pænt ud !

Kan denne forsimpning testet ?

Ja! V.h.a. et **likelihood ratio test** !!

I STATA gøres det ved kommandoerne:

```
quietly xi: logit chdever i.age4
estimates store model1
```

Den første model fittes uden output vises og den gemmes

```
quietly logit chdever age4
estimates store model2
```

Den anden model fittes uden output vises og den gemmes

```
lrtest model1 model2
```

De sammenlignes.

Output

likelihood-ratio test	LR chi2(2) =	0.86
(Assumption: model2 nested in model1)	Prob > chi2 =	0.6513

Der er ikke statistisk signifikant forskel på de to modeller.

En **anden** simplifikation af aldersafhængigheden:

$$\ln(\text{odds}(\text{AGE})) = \beta_0 + \beta_1 \cdot \text{AGE}$$

Bemærk: **AGE** og ikke **AGE4** !!

Lineær afhængighed af (den ikke-grupperede) alder.

$$\beta_1 = \ln(\text{OR}_{\text{Forskel på 1 år}})$$
$$\beta_0 = \ln(\text{odds}(\text{AGE} = 0))$$

Nonsens ! Det ser vi på om lidt !

I STATA

```
xi: logit chdever age
```

eller

```
xi: logit chdever age , or
```

Obs **age** ikke **age4**!

```
logit chdever age
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.0658098	.0144649	4.55	0.000	.0374591 .0941605
_cons	-4.884311	.7737227	-6.31	0.000	-6.40078 -3.367842

Reference alder=0 år!

```
logit chdever age, or
```

	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age	1.068024	.0154488	4.55	0.000	1.03817 1.098736

$\widehat{OR}_{\text{Forskel på 1 år}}$

$\widehat{OR}_{\text{Forskel på 4 år}} = \exp(\hat{\beta}_1 \times 4) = \exp(0.0658 \times 4) = 1.3011$

Aldersgrupper i **AGE4** spænder over ca. 4 år

$\widehat{OR}_{\text{Forskel 1 i gruppe}} = 1.3309$

En mere fornuftig "_cons"; brug: **AGEC50=AGE-50**

$$\ln(\text{odds}(\text{AGE})) = \beta_0 + \beta_1 \cdot (\text{AGE} - 50)$$
$$\beta_1 = \ln(\text{OR}_{\text{Forskel på 1 år}})$$
$$\beta_0 = \ln(\text{odds}(\text{AGE} = 50))$$

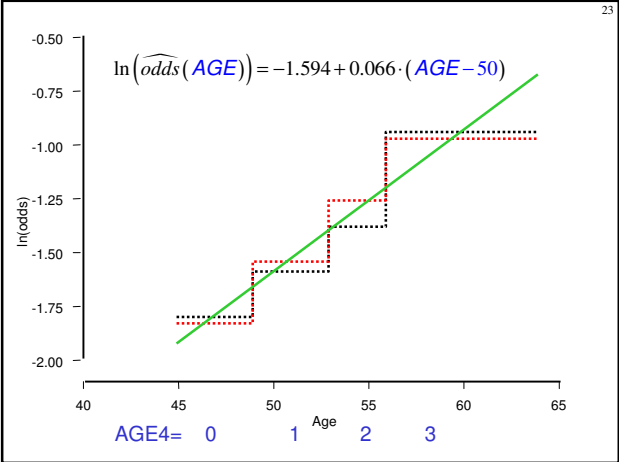
Det giver mening !

Uændret

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
agec50	.0658098	.0144649	4.55	0.000	.0374591 .0941605
_cons	-1.593821	.0834682	-19.09	0.000	-1.757416 -1.430226

$\ln(\widehat{\text{odds}}(\text{AGE} = 50))$ $\widehat{\text{odds}}(\text{AGE} = 50) = \exp(-1.594) = 0.203$

$$\ln(\widehat{\text{odds}}(\text{AGE})) = -1.594 + 0.066 \cdot (\text{AGE} - 50)$$
$$\widehat{\text{odds}}(\text{AGE}) = 0.203 \cdot 1.068^{(\text{AGE} - 50)}$$



Valg af en anden reference gruppe

```
char age4[omit] 0
xi: logit chdever i.age4, or
```

Aldersgruppe kodet 0 er reference

i.age4 _Iage4_0-3 (naturally coded; _Iage4_0 omitted)

Iteration 0: log likelihood = -675.62283

Iteration 1: log likelihood = -664.52107

Iteration 2: log likelihood = -664.32526

Iteration 3: log likelihood = -664.32522

Logit estimates

	Number of obs	=	1363
	LR chi2(3)	=	22.60
	Prob > chi2	=	0.0000
	Pseudo R2	=	0.0167

Log likelihood = -664.32522

	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Iage4_1	1.236215	.2551632	1.03	0.304	.8248982 1.852626
_Iage4_2	1.521692	.3133969	2.04	0.042	1.016294 2.278422
_Iage4_3	2.364289	.4607602	4.42	0.000	1.613673 3.46406

Hvad hvis vi hellere vil sammenligne med de 49-52 årige ?

Reference **AGE4=1**.

Valg af en anden reference gruppe

```
char age4[omit] 1
xi: logit chdever i.age4, or
```

Aldersgruppe kodet 1 er reference

```
i.age4      _Iage4_0-3      (naturally coded; _Iage4_1 omitted)
Iteration 0:  log likelihood = -675.62283
Iteration 1:  log likelihood = -664.52107
Iteration 2:  log likelihood = -664.32526
Iteration 3:  log likelihood = -664.32522
Logit estimates
                Number of obs   =    1363
                LR chi2(3)       =    22.60
                Prob > chi2      =    0.0000
                Pseudo R2        =    0.0167

Log likelihood = -664.32522
```

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Iage4_0	.8089206	.1669667	-1.03	0.304	.5397742 1.212271
_Iage4_2	1.230928	.2440566	1.05	0.295	.8345735 1.815519
_Iage4_3	1.912522	.3571556	3.47	0.001	1.326319 2.757813

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Iage4_0	.8089206	.1669667	-1.03	0.304	.5397742 1.212271
_Iage4_2	1.230928	.2440566	1.05	0.295	.8345735 1.815519
_Iage4_3	1.912522	.3571556	3.47	0.001	1.326319 2.757813

$\widehat{OR}_{45-48 \text{ mod } 49-52}$ $\widehat{OR}_{53-56 \text{ mod } 49-52}$

Test for: 'samme risiko for de 53-56 som de 49-52 årige'

Sammenligning med før:

$$\widehat{OR}_{45-48 \text{ mod } 49-52} = 0.8089 = \frac{1}{\widehat{OR}_{49-52 \text{ mod } 45-48}} = \frac{1}{1.2362}$$
$$\widehat{OR}_{53-56 \text{ mod } 49-52} = 1.2309 = \frac{\widehat{OR}_{53-56 \text{ mod } 45-48}}{\widehat{OR}_{49-52 \text{ mod } 45-48}} = \frac{1.5217}{1.2362}$$

SYS160 korrigeret for køn og alder : 3 modeller

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Isys160_1	2.770273	.4322908	6.53	0.000	2.040308 3.761399
_Issex_1	2.398956	.3527809	5.95	0.000	1.798241 3.200343
_Iage4_1	1.21038	.2566128	0.90	0.368	.798837 1.83394
_Iage4_2	1.354989	.2887571	1.43	0.154	.892357 2.057465
_Iage4_3	2.133587	.428766	3.77	0.000	1.438974 3.163501

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Isys160_1	2.732757	.4246137	6.47	0.000	2.01531 3.705613
_Issex_1	2.408494	.3535652	5.99	0.000	1.8063 3.211451
age4	1.280259	.0819389	3.86	0.000	1.129326 1.451364

chdever	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Isys160_1	2.75052	.426968	6.52	0.000	2.028997 3.728622
_Issex_1	2.411299	.3539757	6.00	0.000	1.808405 3.215188
agec50	1.059597	.0158635	3.87	0.000	1.028957 1.091149