

Basal Statistik

Logistisk regression mm. i R

Lene Theil Skovgaard

28. oktober 2019

Logistisk regression mm.

- ▶ Modeller for binært outcome
 - ▶ Repetition vedrørende tabeller
 - ▶ **Logistisk regression**
 - ▶ Modelkontrol
 - ▶ Alternative links
- ▶ Ordinal regression
 - og hvis vi når det...:
- ▶ Poisson regression

Hjemmesider:

http://publicifsv.sund.ku.dk/~lts/basal19_2

E-mail: ltsk@sund.ku.dk

1 / 114



2 / 114



Typer af outcome

- ▶ Kvantitative data
 - Den generelle lineære model
- ▶ **Binære data** 0/1-data
 - Logistisk regression
- ▶ **Ordinale** data
 - Proportional odds regression, Ordinal regression
- ▶ **Tælledata**
 - Poisson regression
- ▶ Censurerede data (overlevelsesdata)
 - Cox regression

Eksempel fra tidligere: Farveblindhed og køn

	Farveblindhed?		Total
	nej	ja	
Piger	119	1	120
Drenge	144	6	150
Total	263	7	270

Outcome Y: Farveblindhed
dikotom, 0/1, nej/ja

Kovariat: Køn:
dikotom, 0/1, pige/dreng

Er farveblindhed lige hyppigt blandt drenge og piger?, dvs.
Er farveblindhed **uafhængig** af køn?

p_d Sandsynlighed for farveblindhed blandt drenge

p_p Sandsynlighed for farveblindhed blandt piger

Hypotese: $p_d = p_p$

3 / 114



4 / 114



Mål for kønseffekt på farveblindhed

f.eks. drenge vs. piger:

Differens: $p_d - p_p$

Risk Ratio: $RR = \frac{p_d}{p_p}$

Odds Ratio: $OR = \frac{p_d/(1-p_d)}{p_p/(1-p_p)}$

5 / 114



Test for uafhængighed

Hypotese: $p_d = p_p$ (RR=OR=1)

testes med

- ▶ Chi-i-anden test (χ^2 -test),
med mindre tabellerne er ret tynde, dvs.
hvis der er forventede værdier under 5...

Så bruges i stedet

- ▶ Fishers eksakte test
(kan altid anvendes)

som altså er test for uafhængighed,
mellem kovariat (køn) og outcome (farveblindhed)

6 / 114



Opsætning af tabellen

Se samlet kode s. 95

```
fb = matrix(c(1,6,119,144), nrow=2)
rownames(fb) = c("Piger","Drenge")
colnames(fb) = c("ja","nej")
```

```
> addmargins(fb)
      ja nej Sum
Piger  1 119 120
Drenge  6 144 150
Sum     7 263 270
```

```
> prop.table(fb,1)*100
      ja      nej
Piger 0.8333333 99.16667
Drenge 4.0000000 96.00000
```

```
> chisq.test(fb)$expected
      ja      nej
Piger  3.111111 116.8889
Drenge  3.888889 146.1111
Warning message:
In chisq.test(fb) : Chi-squared approximation may be incorrect
```

7 / 114



Test af hypotesen om uafhængighed

```
> chisq.test(fb)
Pearson's Chi-squared test with Yates' continuity correction
```

```
data: fb
X-squared = 1.5418, df = 1, p-value = 0.2144
```

```
Warning message:
In chisq.test(fb) : Chi-squared approximation may be incorrect
```

```
> fisher.test(fb)
Fisher's Exact Test for Count Data
```

```
data: fb
p-value = 0.1364
alternative hypothesis: true odds ratio is not equal to 1
```

Vi kan altså *ikke afvise*, at forekomsten af farveblindhed er ens for
drenge og piger (P=0.14)

8 / 114



Differens mellem sandsynlighederne

Vi må selv udregne differensen:

$$\hat{p}_p - \hat{p}_d = -0.0317, CI = (-0.0745, 0.0112):$$

```
> prop.test(fb)

2-sample test for equality of proportions with continuity correction

data:  fb
X-squared = 1.5418, df = 1, p-value = 0.2144
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07449312  0.01115979
sample estimates:
 prop 1      prop 2 
0.008333333 0.040000000

Warning message:
In prop.test(fb) : Chi-squared approximation may be incorrect
```

Der er altså ca. **3% flere drenge end piger**, der er farveblinde.

9 / 114



Relativ risiko og odds ratio

Risk ratio = relativ risiko (RR=4.80),

```
install.packages("epitools")
library("epitools")

> epitab(fb,method="riskratio",rev="columns")$tab
      nej      p0 ja      p1 riskratio lower upper p.value
Piger 119 0.9916667 1 0.008333333      1.0      NA      NA      NA
Drenge 144 0.9600000 6 0.040000000      4.8 0.5858252 39.32914 0.1363846
```

Odds ratio (OR=4.96)

```
> epitab(fb,method="oddsratio",rev="rows")$tab
      ja      p0 nej      p1 oddsratio lower upper p.value
Drenge 6 0.8571429 144 0.5475285 1.000000      NA      NA      NA
Piger 1 0.1428571 119 0.4524715 4.958333 0.5887152 41.76055 0.1363846
```

10 / 114



Nyt eksempel: Postoperative sårinfektioner

194 patienter opereret på Frederiksberg Hospital.

For hver patient i registreres:

- ▶ x_{i1} =optid, operationens varighed i minutter
- ▶ x_{i2} =alder, i år
- ▶ y_i =infektion=

$$\begin{cases} 1 : & \text{postoperativ sårinfektion (n=23)} \\ 0 : & \text{ingen postoperativ sårinfektion (n=171)} \end{cases}$$

```
> summary(inf[,2:4])
      infektion      optid      alder
Min.   :0.0000   Min.   :  5.00   Min.   :  7.00
1st Qu.:0.0000   1st Qu.: 50.00   1st Qu.:30.00
Median :0.0000   Median : 75.00   Median :64.50
Mean   :0.1186   Mean   : 93.94   Mean   :55.97
3rd Qu.:0.0000   3rd Qu.:120.00   3rd Qu.:77.75
Max.   :1.0000   Max.   :390.00   Max.   :90.00
```

11 / 114



Problemstilling

Hvordan afhænger sandsynligheden for at få postoperativ sårinfektion af alder og operationstid?

Outcome: Sårinfektion ja/nej,
et binært/dikotomt outcome (0-1 variabel)

Kovariater vælges her som:

- ▶ Patientens alder, som lineær effekt
(evt. som alder pr. 10-år ved at skalere til
 $\text{alder}_{10} = \text{alder}/10$)
- ▶ Indikator for operationsvarighed over 2 timer:

$$\text{over2timer} = \begin{cases} 1 : & \text{hvis optid} > 120 \\ 0 : & \text{ellers} \end{cases}$$

Model for p_i 'erne? Først plejer vi at tegne...

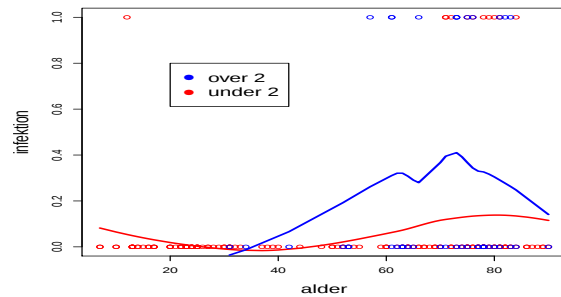
12 / 114



Figurer

Plot af infektion vs. alder, med loess-smoother (kode s. 96)

Blå: over 2 timer, Rød: under 2 timer



Figureerne er rigtigt grimme,
når outcome kun antager 2 forskellige værdier

13 / 114

Model for p_i 'erne

Adskillige ting *dur ikke* i denne situation med binært outcome og kvantitativ kovariat (her alder):

- ▶ Figurer er ikke særlig kønne, fordi outcome er binært
- ▶ Tabeller dur ikke, fordi alder har for mange værdier – her 68 forskellige aldre
- ▶ Den sædvanlige linearitetsantagelse (for effekten af kovariaten alder) er ikke god for sandsynligheder, fordi en linie ikke er begrænset, hverken nedadtil eller opadtil, så vi kommer let udenfor området mellem 0 og 1
 - og det er mildest talt ikke så rart, når man har med sandsynligheder at gøre

14 / 114

Model for p_i 'erne, fortsat

Vi har brug for at **transformere** p_i 'erne, før vi kan antage linearitet

- ▶ Derfor bruger man en funktion $g(p_i)$, kaldet **link**-funktionen, og antager linearitet på denne skala.

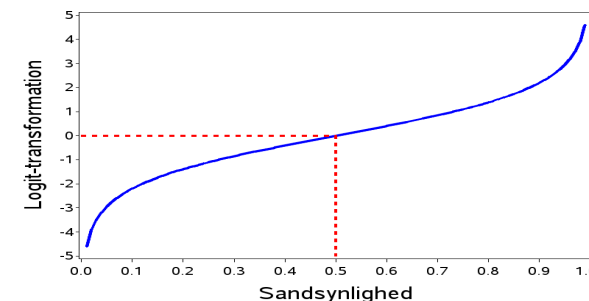
$$\eta_i = g(p_i) = \beta_0 + \beta_1 \text{alder}_i + \beta_2 \text{over2timer}_i$$

Den hyppigst anvendte funktion er **logit**-funktionen, og analysen kaldes derfor **logistisk regression**:

$$\eta_i = g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

15 / 114

Logit funktionen



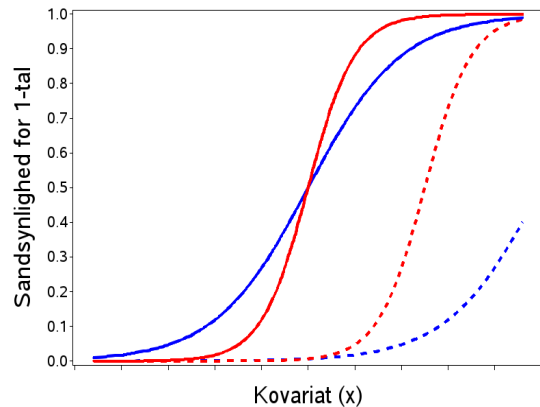
Sandsynligheder tæt på 0 giver store negative logit-værdier,
Sandsynligheder tæt på 1 giver store positive logit-værdier

Så på denne skala giver det mening at bygge lineære modeller

16 / 114

Logistiske kurver – for en enkelt kovariat, x

$$p(x) = g^{-1}(\eta) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$



Parameterværdier:

- ▶ β_0 : -5, 0
- ▶ β_1 : 1, 2

17 / 114



Fortolkning af parametre - for en enkelt kovariat

Model: $Y_i \sim \text{Binomial}(1, p_i)$, hvor $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$

Sammenlign to personer, A og B, med alder $x_A = a$ hhv. $x_B = a + 10$, altså med en aldersforskel på 10 år:

Person B: $\text{logit}(p_B) = \beta_0 + \beta_1 \times (a + 10)$

Person A: $\text{logit}(p_A) = \beta_0 + \beta_1 \times a$

Forskel: $\text{logit}(p_B) - \text{logit}(p_A) = \beta_1 \times 10$

Men $\text{logit}(p_B) - \text{logit}(p_A) = \log(OR)$

og derfor er **Odds Ratio** for infektion for person B vs. person A netop

$OR = \exp(\beta_1 \times 10)$, fordi der var 10 års forskel på de to personer.

18 / 114



Fortolkning af parametre - for to kovariater

Nu er

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Sammenlign igen to personer, A og B, med alder $x_{1A} = a$ hhv. $x_{1B} = a + 10$, altså med en aldersforskel på 10 år, men **samme operationsvarighed(sgruppe)** $x_{2A} = x_{2B} = b$:

Person B: $\text{logit}(p_B) = \beta_0 + \beta_1 \times (a + 10) + \beta_2 \times b$

Person A: $\text{logit}(p_A) = \beta_0 + \beta_1 \times a + \beta_2 \times b$

Forskel: $\text{logit}(p_B) - \text{logit}(p_A) = \beta_1 \times 10$

altså **samme svar** som ved en enkelt kovariat, ganske som i almindelig regression

19 / 114



Fortolkning af parametre, fortsat

Fortolkning af interceptet, β_0

- ▶ har (som altid) noget at gøre med en person, hvor alle kovariater er 0, nemlig $\log(\frac{p}{1-p})$ for en sådan person.
- ▶ dvs. her en nyfødt (alder=0), der har en operationstid på under 2 timer (over2timer=0)

Dette er irrelevant!

For at få et fortolkeligt intercept, kan vi i stedet **centrere** alderen ved f.eks. 50 år, altså benytte kovariaten

$\text{alder_minus50} = \text{alder} - 50$

eller prediktere for denne udvalgte alder

20 / 114



Software til logistisk regression

- ▶ SAS
 - ▶ LOGISTIC: let, kan anvendes til ordinale data
<http://www.ats.ucla.edu/stat/sas/dae/logit.htm>
 - ▶ GENMOD: minder om GLM, kan anvende andre links og andre fordelinger, men kræver lidt mere kodning
<http://support.sas.com/kb/42/728.html>
- ▶ SPSS: Analyze/Regression/Binary Logistic
<https://www.youtube.com/watch?v=zj15KUXtC7M>
<http://www.ats.ucla.edu/stat/spss/topics>
- ▶ R: glm
<http://www.statmethods.net/advstats/glm.html>
- ▶ STATA: logit
<http://www.ats.ucla.edu/stat/stata/dae/logit.htm>

21 / 114



Software til logistisk regression - HUSK

Hvis man ikke passer på, kommer man til at benytte en sædvanlig normalfordelingsmodel

– ikke så smart, når outcome kun kan være 0 eller 1

Sørg for at få specificeret:

- ▶ Fordelingen: Binomial eller Bernoulli (Binomial med N=1) (family=binomial)
- ▶ Link-funktionen, der vælger hvilken skala, vi benytter til den lineære struktur, sædvanligvis *logit* (link="logit"), men evt. en anden — hvis man har styr på, hvad man gør
- ▶ hvad "event" er, altså om man vil modellere sandsynlighed for et 1-tal eller et 0 (R vælger at modellere sandsynligheden for et 1-tal)

22 / 114



Logistisk regression i praksis

Vi lægger modelresultatet i model1:

```
model1 = glm(formula=infektion ~
  factor(over2timer)+alder,
  family=binomial(link="logit"), data=inf)
```

hvorefter vi kan benytte funktioner såsom

```
summary(model1)
confint(model1)
drop1(model1, test = "Chisq")
```

23 / 114



Output fra logistisk regression

```
> summary(model1)
```

```
Call:
glm(formula = infektion ~ factor(over2timer) + alder, family = binomial(link = "logit"),
    data = inf)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.73658	1.07867	-4.391	1.13e-05 ***
factor(over2timer)1	1.32921	0.47295	2.810	0.00495 **
alder	0.03548	0.01491	2.379	0.01734 *

Selve parameterestimerne er på log-odds skala, og derfor ikke umiddelbart fortolkelige, pånær fortegnet.

Se i stedet på Odds-ratio herunder (jvf s. 18-19)

```
> cbind(exp(coef(model1)),exp(confint(model1)))
```

Waiting for profiling to be done...

		2.5 %	97.5 %
(Intercept)	0.008768596	0.0007057854	0.05281793
factor(over2timer)1	3.778073411	1.4916906798	9.68216684
alder	1.036119878	1.0093289803	1.07145168

24 / 114



Kommentarer til output

Fokus er på Odds-ratio (dvs. **forholdet mellem odds**) for komplikation **mellem to niveauer** af en kovariat, **for fastholdt værdi af de øvrige kovariater**

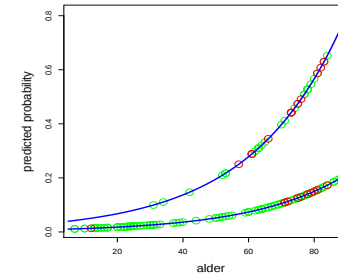
- ▶ Personer med en operationsvarighed over 2 timer har en odds for sårinfektion, der er 3.78 gange højere end dem med en operationsvarighed under 2 timer **på samme alder** (CI: 1.49, 9.68).
- ▶ For en patientpopulation, der har alderen $X+10$ år, vil odds (forholdet mellem patienter der får hhv. ikke får sårinfektion) være en faktor $1.036^{10} = 1.42$ større end hos en patientpopulation, der har alderen X år (forudsat at de to populationer har **lige lang operationstid**). Det svarer til, at odds er øget med 42%, med CI: 9.7%-99.4% (tilbagetransformeret fra 1.009^{10} hhv. 1.071^{10})

25 / 114

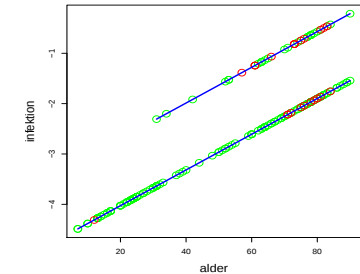


Predikterede sandsynligheder for postoperativ sårinfektion

På sandsynligheds-skala



På logit-skala



Kurverne er (parallelle) rette linier på logit-skala

Kode til dette s. 98-99

26 / 114



*Inhomogene populationer

Ofte er der uerkendte kovariater (kunne f.eks. være køn), og så er der større forskel på personerne end modellen angiver (overspredning)

Effekt af manglende kovariater

- ▶ De “sædvanlige” pga confounding
- ▶ De “nye”, som skyldes ikke-linearitet af logit-funktionen
 - ▶ **Undervurdering af effekter**, her alder
 - ▶ Undervurdering af standard errors

Det, modellen udtaler sig om, er nemlig en sammenligning af 2 personpopulationer (**af blandet køn**) med forskellig alder – og det giver ikke nødvendigvis det samme som en tilsvarende sammenligning, opdelt på køn.

27 / 114



Et simpelt eksempel

med tænkte sandsynligheder (f.eks. for infektion) ved alder 30 og 40 år, for hhv. mænd og kvinder:

Køn	30 år	40 år	Differens	log(OR)	OR
Mænd	0.2	0.4	0.2	0.981	2.67
Kvinder	0.6	0.8	0.2	0.981	2.67
Alle	0.4	0.6	0.2	0.811	2.25

Begge køn har $OR=2.67$ for infektion ved 40 år i forhold til ved 30 år, **men** på “populationsniveau” har vi $OR=2.25$

Gennemsnittet af subpopulationernes OR'er (som her for nemheds skyld er ens) er altid større end OR udregnet for hele populationen.

28 / 114



Konsekvens for randomiserede undersøgelser

Her er der pr. definition *ingen* confounding, men størrelsen af Odds Ratio for behandlingseffekt *kan alligevel afhænge af populationssammensætningen!*

- ▶ Hvis der er restriktive inklusionskriterier, bliver OR formentlig stort
- ▶ Hvis der "blot" er samlet revl og krat, bliver OR formentlig mindre

Det er altså helt og holdent spørgsmålets præcisering, der afgør svaret....

Det er ikke *evidensen* (P-værdien), der refereres til, men selve estimatet.

29 / 114



Confounding og inhomogeniteter

Hvad hvis vi sammenligner to populationer med 10 års forskel *uden at tage hensyn til operationstid*, dvs. kun med alder som kovariat?

Så får vi formentlig noget

- ▶ *større*, fordi der er *confounding* mellem alder og operationstid, idet de ældre generelt har længere operationstid

	over2timer	alder.mean	alder.min	alder.max
1	0	52.40132	7	90
2	1	68.88095	31	90

- ▶ *mindre*, fordi vi tager gennemsnit over nogle mere inhomogene populationer

(Forskellen er dog ikke stor, se venstre kolonne af tabel s. 31)

30 / 114



OR estimater for infektion i forskellige modeller

Outcome: infektion	Effekt af	
Kovariater i model	10 år ældre OR (CI)	operationstid > 2 timer OR (CI)
<i>kun</i> alder	1.48 (1.13, 1.93) P=0.0043	–
<i>kun</i> over2timer	–	5.13 (2.07, 12.71) P=0.0004
alder <i>og</i> over2timer	1.42 (1.097, 1.994) P=0.017	3.78 (1.49, 9.68) P=0.0050

31 / 114



Tilbage til modellen

Var den fornuftig ud fra et anvendelses-synspunkt?

1. *Fik vi stillet et relevant spørgsmål?*
– det stod forhåbentlig i protokollen
2. *Fik vi opstillet en model, der tillod at besvare dette spørgsmål?*
– var det f.eks. en *interaktion*, der var i fokus?
3. *Har vi leveret et fyldestgørende svar?*
og har vi husket at kvantificere, med konfidensinterval?

Og så er der jo lige det med *modelkontrol*.....:

Har vi modelleret det så tilpas godt, så vi også tror på svaret?

32 / 114



Interaktion

Har operationsvarigheden en *større* betydning for ældre mennesker end for yngre?

Inkluder interaktionen mellem operationsvarighed og alder, i form af effekten `over2timer:alder` (kode s. 100)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.94854	1.28631	-3.847	0.00012	***
factor(over2timer)1	2.19979	2.47794	0.888	0.37468	
alder	0.03853	0.01779	2.165	0.03037	*
factor(over2timer)1:alder	-0.01230	0.03445	-0.357	0.72103	

Her ser vi kun på P-værdierne

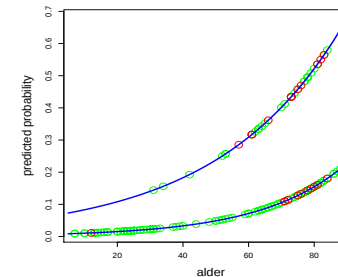
Hvis (når) vi vil vide noget om Odds Ratioer, må vi selv specificere yderligere...eller opdele

33 / 114

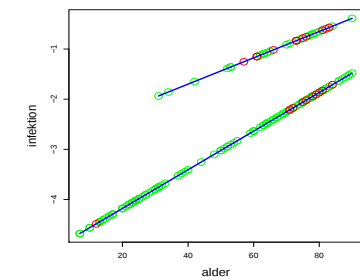


Interaktion mellem alder og operationsvarighed

På sandsynligheds-skala



På logit-skala



Der ses ikke nogen tegn på interaktion ($P=0.72$), se s. 33

34 / 114



Modelkontrol

Vi havde oprindeligt antaget **additivitet** mellem alder og operationsvarighed, samt **linearitet** for alderseffekten (på logit-skala).

Numerisk modelkontrol:

- ▶ Overall goodness of fit (s. 36-38, kode s. 101)
- ▶ **Linearitet:**
Tilføj logaritmeret kovariat, Lineære splines, eller evt. kvadratled ("velkendt", s. 39-42, kode s. 102-103)
- ▶ **Additivitet:**
Check for interaktion (har vi lige gjort s. 33-34, "velkendt")

Grafisk modelkontrol:

- ▶ Residualplots (s. 43-45)
- ▶ Diagnostic plots (s. 46-47)

35 / 114



* Overall Goodness-of-fit

for modellen s. 23 (se kode s. 101)

- ▶ Observationerne inddeles i 10 ca. lige store grupper, baseret på stigende predikteret sandsynlighed for infektion
- ▶ I hver gruppe sammenlignes observerede og forventede antal af infektioner, og
- ▶ Størrelserne $\frac{(OBS - N\hat{p})^2}{N\hat{p}(1-\hat{p})}$ sammenlægges til en **approximativ** χ^2 -teststørrelse med 8 frihedsgrader (antal grupper minus 2)

Her finder vi $\chi^2 = 7.50 \sim \chi^2(8) \Rightarrow P = 0.48$ og dermed intet tegn på problemer med modellen

36 / 114



Goodness-of-fit i praksis

sædvanligvis kaldt **Hosmer-Lemeshow**-testet. Dette kræver en "add-on" pakke:

```
install.packages("ResourceSelection")
library(ResourceSelection)

hl = hoslem.test(inf$infektion, fitted(model2), g=10)
```

Bemærk:

I tilfælde af sparsomme data kan inddelingen have en del indflydelse på testet, dvs. *det er meget ustabilt*. F.eks. kan det ændre sig, hvis man skifter til at se på det modsatte outcome, altså event="0" (her $P=0.13$ vs. $P=0.48$).

37 / 114



Output fra Hosmer-Lemeshow testet

Se kode foregående side

```
> hl
Hosmer and Lemeshow goodness of fit (GOF) test

data:  inf$infektion, fitted(model2)
X-squared = 7.5046, df = 8, p-value = 0.4833

> cbind(hl$observed, hl$expected)
      y0 y1  yhat0  yhat1
[0.0092,0.0151] 21  1 21.73221 0.2677906
(0.0151,0.0182] 21  0 20.63792 0.3620831
(0.0182,0.0296] 15  0 14.66526 0.3347357
(0.0296,0.0749] 20  0 18.93914 1.0608598
(0.0749,0.102] 18  3 19.11427 1.8857349
(0.102,0.124] 15  2 15.04375 1.9562545
(0.124,0.139] 19  4 19.99388 3.0061220
(0.139,0.191] 15  1 13.45090 2.5490981
(0.191,0.303] 13  7 14.87914 5.1208598
(0.303,0.404] 14  5 12.54354 6.4564615
```

Med en P-værdi på 0.48 ses ingen generel afvigelse fra modellen
men testet er som nævnt meget ustabilt ved små datasæt

38 / 114



Alderseffekt modelleret med både alder og log(alder)

$\text{logit}(p) = \beta_0 + \beta_1 \cdot 1_{\text{optid} > 2} + \beta_2 \cdot (\text{alder} - 50) + \beta_3 \cdot (\log_{1.1}(\text{alder}) - \log_{1.1}(50))$ Se kode s. 102

```
> summary(model3)

Call:
glm(formula = infektion ~ factor(over2timer) + alder.minus50 +
    logalder50, family = binomial(link = "logit"),
    data = inf, na.action = na.exclude)

Coefficients:
    Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.20987    0.57313  -5.601 2.14e-08 ***
factor(over2timer)1  1.39310    0.48747   2.858 0.00427 **
alder.minus50     0.06927    0.05104   1.357 0.17472
logalder50       -0.15144    0.21136  -0.716 0.47368

> cbind(exp(coef(model3)), exp(confint(model3)))

              2.5 %      97.5 %
(Intercept)  0.0403619 0.01174104 0.1141373
factor(over2timer)1  4.0273194 1.55451983 10.6900590 <-----
alder.minus50     1.0717232 0.95754902 1.1792414
logalder50       0.8594711 0.59978978 1.4497293
```

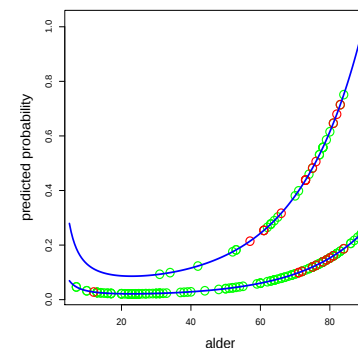
Meget ringe indikation af afvigelse fra linearitet ($P = 0.47$)

39 / 114

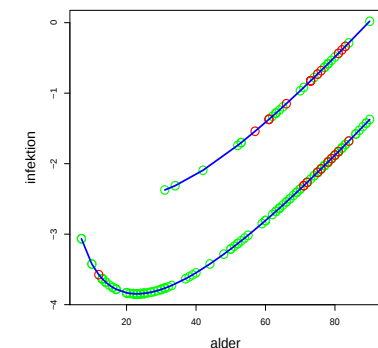


Effekten af alder og log(alder)

På sandsynligheds-skala



På logit-skala



40 / 114



Effekten af alder som lineær spline

med knæk ved 60 og 75 år får vi **outputtet** (kode s. 103):

```
> summary(model4)

Call:
glm(formula = infektion ~ factor(over2timer) + alder + alder.over60 +
    alder.over75, family = binomial(link = "logit"),
    data = inf, na.action = na.exclude)

Coefficients:      Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.69328    1.60934  -2.916  0.00354 **
factor(over2timer)1  1.31426    0.49053   2.679  0.00738 **
alder           0.02841    0.03323   0.855  0.39260
alder.over60     0.07302    0.08186   0.892  0.37240
alder.over75    -0.20019    0.12292  -1.629  0.10340

> cbind(exp(coef(model4)),exp(confint(model4)))
              2.5 %      97.5 %
(Intercept)  0.009156558 0.0001282662 0.1133436
factor(over2timer)1 3.721984866 1.4281607678 9.9363844 <-----
alder         1.028819014 0.9696274894 1.1141968
alder.over60  1.075753440 0.9185984486 1.2746073
alder.over75  0.818578804 0.6324346966 1.0293964
```

En vis indikation for afvigelse fra linearitet ved 75 år ($P = 0.10$),

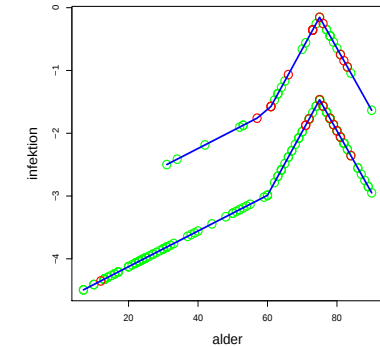
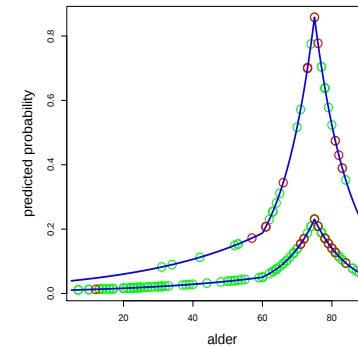
men...
41 / 114



Effekten af alder som lineær spline

På sandsynligheds-skala

På logit-skala



Intet signifikant knæk, men dog en tendens....

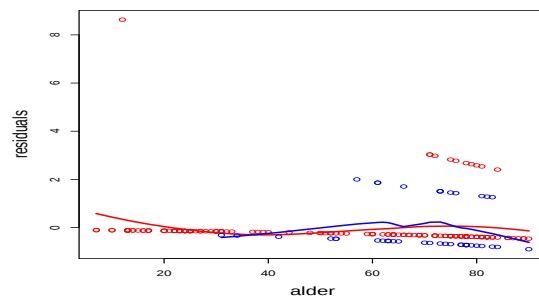
Tror vi på den?

42 / 114



Modelkontrol

Den sædvanlige konstruktion af residualer ($r_i = y_i - \hat{p}_i$) giver nogle forfærdeligt grimme figurer, f.eks. (se kode s. 104)



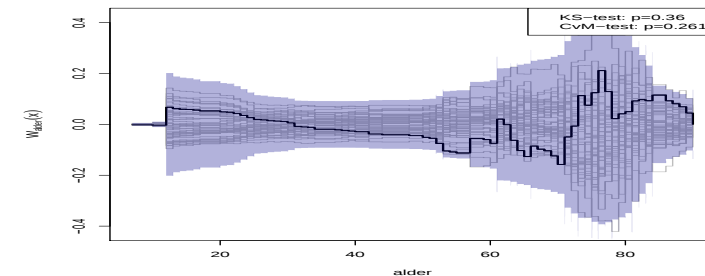
og her er endda standardiseret til $e_i = \frac{r_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$

43 / 114



Modelkontrol, fortsat

Der kan være en fordel i at se på **kumulerede residualer**, kumuleret fra lave til høje aldre (se kode s. 105):



Her er indlagt 40 forløb, simuleret under forudsætning af en korrekt model. **Falder det aktuelle (fed streg) forløb udenfor?**

44 / 114



Modelkontrol, fortsat

Figuren på forrige side kan vurderes visuelt, men man kan også lave et **numerisk check** på den maksimale afvigelse fra 0, her baseret på 1000 simulationer.

Vi finder $P=0.36$ (fremgår af selve figuren), og dermed ingen evidens for en gal modellering af alderseffekten.

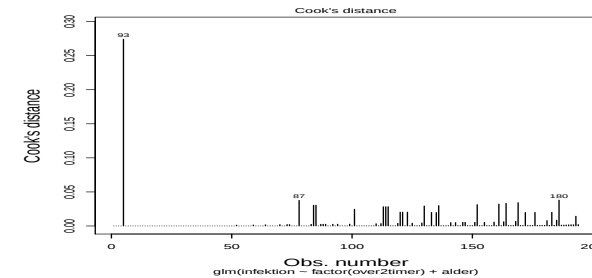
Koden ses på s. 105

45 / 114



Diagnostics – Cooks afstand

Cooks afstand: (mål for den enkelte observations indflydelse på estimerterne), her plottet op mod observationsnummeret (kode s. 106)



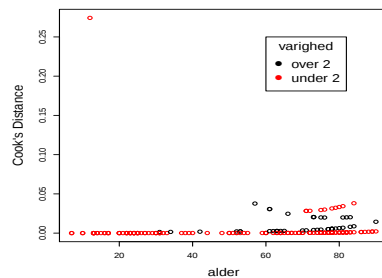
Bemærk den enkelte observation med stor Cook, som i R placeres for sig selv.

46 / 114



Diagnostic plots (kode s. 106)

Cooks afstand vs. alder
blå: lang operationstid

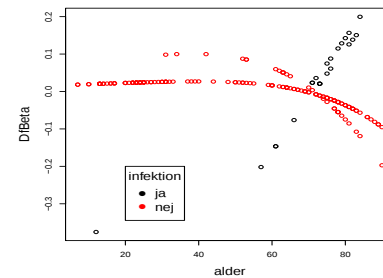


Her kan vi se, at observationen med stor indflydelse er ung, med en kort operationstid, **men med infektion**

47 / 114



Effekt på alders-estimat
vs. alder, blå: infektion



Den suspekte person

findes at være person nr. 93:

```
> inf[cook1==max(cook1),]
  id infektion optid alder over2timer alder.minus50 logalder50 alder_over60
93 93         1   55   12         0        -38  -14.97339             0
  alder_over75 alder.over60 alder.over75 resid1
93           0           0           0  8.631275
```

Det drejer sig altså om en 12-årig patient, der til trods for kort operationstid, *alligevel* får en infektion.

Det er usædvanligt, og derfor ændres estimerterne en del, hvis denne patient pilles ud.

Og hvilken begrundelse kunne der også være for det?

48 / 114



Konklusion

baseret på resultater fra s. 24, reproduceret her:

```
> cbind(exp(coef(model1)),exp(confint(model1)))
              2.5 %      97.5 %
(Intercept)    0.008768596 0.0007057854 0.05281793
factor(over2timer)1 3.778073411 1.4916906798 9.68216684
alder           1.036119878 1.0093289803 1.07145168
```

- ▶ Sample size er lidt for lille til, at man kan udtale sig meget sikkert omkring alderseffekten, men risikoen *ser ud til* at stige med alderen.
- ▶ Det er bedst at have en lav operationsvarighed (*surprise* 🤔) hvor meget bedre er dog ret usikkert (bredt konfidensinterval)

49 / 114



Odds ratio, Risk ratio og Risk differenser

Den **link-funktion**, der anvendes, afgør hvilket mål, der kommer ud af det

- ▶ logit-link er *default*, giver **Odds Ratio**
det brugte vi for at modellere infektioner
Dette link skal *altid* benyttes ved case-control studier.
- ▶ log-link giver **Risk Ratio** (relativ risiko)
det kunne vi f.eks. bruge på data vedr. **farveblindhed** (s. 51-52)
- ▶ identity-link giver **Risk difference**
det kan vi bruge til at lave trend test for **kejsersnit vs. skostørrelse** (s. 53ff)

logit er default, fordi det sikrer, at vi ikke kommer udenfor intervallet (0,1) med sandsynlighederne

50 / 114



Farveblindhed - igen

Risk Ratio (relativ risiko) estimeres ved at benytte **log-link**:

```
gender = as.factor(c(rep("Piger",120),rep("Drenge",150)))
farve = c(rep("nej",119),rep("ja",1),rep("nej",144),rep("ja",6))

farveblind=(farve=="ja")

model1 = glm(farveblind ~ relevel(gender,ref="Piger"),
             family=binomial(link="log"))
```

Nederste linie af output på næste side viser, at den **relative risiko for farveblindhed for drenge vs. piger** er estimeret til 4.80, med konfidensgrænser (0.84, 90.01), altså *ingen signifikant forskel* (P=0.14), og en **meget usikker** bestemmelse af den relative risiko.

51 / 114



Output fra analyse med log-link

```
> summary(model1)

Call:
glm(formula = farveblind ~ relevel(gender, ref = "Piger"),
    family = binomial(link = "log"))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.7875     0.9955  -4.809 1.52e-06 ***
relevel(gender, ref = "Piger")Drenge  1.5686     1.0729   1.462   0.144
(Dispersion parameter for binomial family taken to be 1)

> cbind(exp(coef(model1)),exp(confint(model1)))
              2.5 %      97.5 %
(Intercept)    0.008333335 0.000477164 0.0361345
relevel(gender, ref = "Piger")Drenge 4.799999225 0.835307130 90.0068868

# eller med Wald-baseret konfidensinterval:
> cbind(exp(coef(model1)),exp(confint.default(model1)))
              2.5 %      97.5 %
(Intercept)    0.008333335 0.001184171 0.05864398
relevel(gender, ref = "Piger")Drenge 4.799999225 0.586129194 39.30872716
```

Se kommentarer til output på forrige side

52 / 114



Eksempel om kejsersnit vs. skostørrelse

351 fødende kvinder har fået registreret deres skostørrelse, samt om et kejsersnit blev aktuelt ved fødslen.

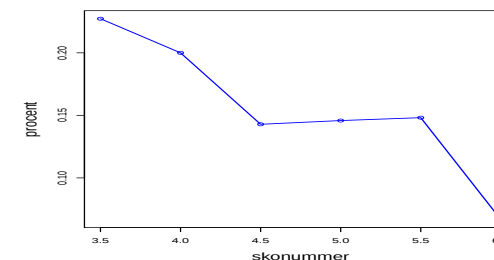
Kejsersnit	< 4		4		Skonummer 4.5 5		5.5	≥ 6	Total
	Ja	Nej	Ja	Nej	Ja	Nej	Ja	Nej	
Ja	5	7	6	7	8	10	43		
Nej	17	28	36	41	46	140	308		
I alt	22	35	42	48	54	150	351		
% kejsersnit	22.7	20.0	14.3	14.6	14.8	6.7	12.3		

Vi kunne have en hypotese om **faldende sandsynlighed for kejsersnit med stigende (sko)størrelse**

53 / 114



Frekvens af kejsersnit som funktion af skostørrelse



Måske er der linearitet i skostørrelse?

Dette *kunne* gå an her, fordi vi ikke er tæt på 0....

54 / 114



Test for trend

Modellen *antager* linearitet i skostørrelse: $p_i = \alpha - \beta \times \text{sko}_i$

og estimationen udføres som en regression i Binomial-fordelingen, med identity-link.

```
skonummer = c(rep(3.5,22),rep(4,35),rep(4.5,42),rep(5,48),
               rep(5.5,54),rep(6,150))
snit = c(rep("ja",5),rep("nej",17),rep("ja",7),rep("nej",28),rep("ja",6),rep("nej",36),
          rep("ja",7),rep("nej",41),rep("ja",8),rep("nej",46),rep("ja",10),rep("nej",140))
```

```
> table(skonummer,snit)
```

```
      snit
skonummer  ja  nej
3.5      5   17
4        7   28
4.5      6   36
5        7   41
5.5      8   46
6       10  140
```

```
kejsersnit=(snit=="ja")
```

```
model2 = glm(kejsersnit ~skonummer,
              family=binomial(link="identity"))
```

55 / 114



Output: Test for trend

altså med linearitet på selve sandsynligheds-skalaen

```
> summary(model2)
```

Call:

```
glm(formula = kejsersnit ~ skonummer, family = binomial(link = "identity"))
```

Coefficients:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.45607    0.12662   3.602 0.000316 ***
skonummer   -0.06356    0.02281  -2.786 0.005337 **
```

```
> confint(model2)
```

```
              2.5 %      97.5 %
(Intercept)  0.2162037  0.7221782
skonummer   -0.1107865 -0.0190242
```

Ekstra risiko for kejsersnit, når skoene er et nummer **mindre**: 0.0636, CI=(0.0190, 0.1108), altså mellem 1.9% og 11.1%

56 / 114



Trick: Modelkontrol af lineariteten

Lineariteten i skonummer var jo *en antagelse*.

Trick (i alle former for regression):

Når kovariaten kun antager så få værdier, kan linearitetsantagelsen checkes ved at sætte en kopi af skonummer ind som faktor oveni, og så teste, om denne kan undværes:

```
model3 = glm(kejsersnit ~ skonummer + factor(skonummer),
             family=binomial(link="identity"))
```

```
anova(model3, test="LR")
```

Modellen ville være den samme, hvis skonummer blev udeladt, men så ville vi ikke få testet, om modellen med **kun skonummer** var fornuftig

57 / 114



Output fra check af linearitet

```
> anova(model3, test="LR")
Analysis of Deviance Table
```

```
Model: binomial, link: identity
Response: kejsersnit
```

```
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			350	261.07	
skonummer	1	8.0502	349	253.02	0.00455 **
factor(skonummer)	4	1.2940	345	251.72	0.86240

Bemærk:

- ▶ Man kan *ikke* teste skonummer væk fra modellen, da dette ikke ændrer modellen.
Det ovenstående test er i en model med *kun* skonummer
- ▶ Man kan *godt* teste factor(skonummer) væk fra modellen, da man derved reducerer til en lineær effekt af skonummer

Lineariteten i skonummer ser rimelig ud (P=0.86)

58 / 114



Typer af outcome

Hidtil

- ▶ Kvantitative, Den generelle lineære model
- ▶ Dikotom (0/1), Logistisk regression

Nu følger en lille smule om:

- ▶ **Ordinale outcomes**
f.eks. fysisk formåen på en skala 1-2-3-4
- ▶ **Tælleletal** (uden øvre grænse), f.eks. antal metastaser

Poisson regression eller log-lineære modeller

59 / 114



Eksempel om leverfibrose

Ordinalt outcome: Leverfibrose, grad 0,1,2 eller 3

Kovariater:

3 blodmarkører relateret til fibrose: **HA, YKL40, PIIINP**

Problemstilling:

Hvad kan vi sige om fibrosegrad ud fra måling af disse 3 blodmarkører?

```
> summary(fib[3:6],)
      ykl40      piiinp      ha      fibrosegrad
Min.   : 50.0   Min.   : 1.70   Min.   : 21.0   0:27
1st Qu.: 175.0   1st Qu.: 5.15   1st Qu.: 30.0   1:40
Median : 330.0   Median : 9.00   Median : 105.5   2:42
Mean    : 533.5   Mean    :13.41   Mean    : 318.5   3:20
3rd Qu.: 718.0   3rd Qu.:15.85   3rd Qu.: 250.5
Max.    :4850.0   Max.    :70.00   Max.    :4730.0
NA's    :2       NA's    :1
```

(Julia Johansen, KKHH)

60 / 114



Ordinale data

- ▶ data på en rangskala
- ▶ afstand mellem responskategorier kendes ikke, eller er udefineret
- ▶ evt. en underliggende kvantitativ skala

Vi sidder mellem to stole:

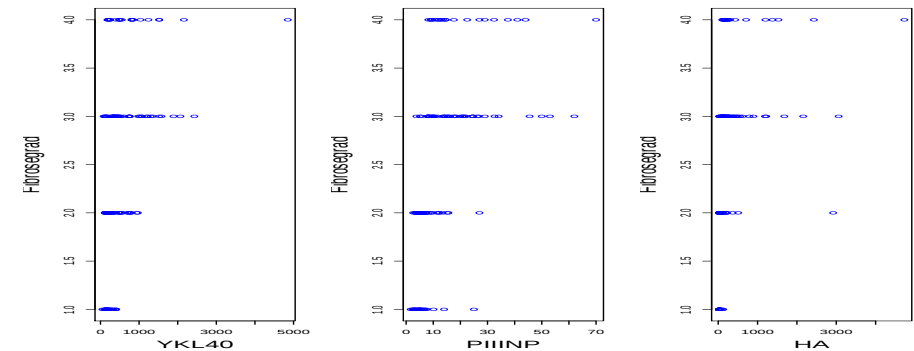
- ▶ Vi kan reducere til et binært respons og lave logistisk regression
 - men vi kan opdele flere steder
- ▶ Vi kan 'lade som om' det er normalfordelt
 - virker naturligvis bedst hvis der er mange responskategorier

61 / 114



Fordeling af blodmarkører

vist for hver fibrose kategori



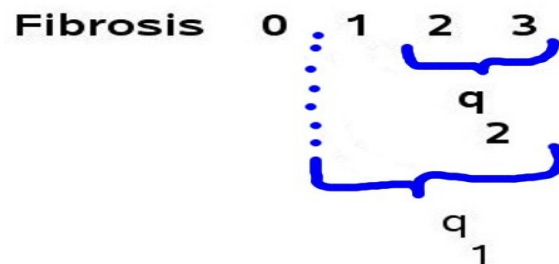
Er der noget problem ved de meget skæve fordelinger?

62 / 114



Kumulerede sandsynligheder

Sandsynlighed for *dette eller værre*



Tilbageregning til sandsynligheder for de enkelte fibrosegrader:

$$p_0 = 1 - q_1, \quad p_1 = q_1 - q_2$$

$$p_s = q_s - q_3, \quad p_3 = q_3$$

63 / 114



Proportional odds model

Logistisk regression for hver tærskel,
med **samme afhængighed** af kovariaterne, dvs.

$$\text{logit}(q_3) = \beta_{03} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\text{logit}(q_2) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\text{logit}(q_1) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Her er kovariaterne de 3 biomarkører, alle $\log_2$ -transformerede:

Odds ratio vil **ikke afhænge af cutpoint**,
og denne antagelse testes automatisk som

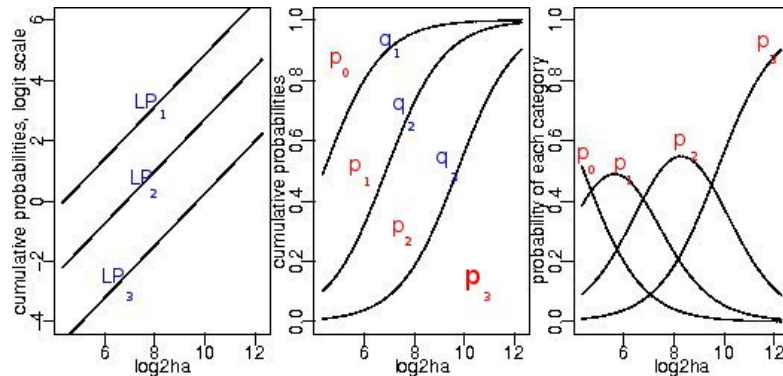
Proportional odds assumption

64 / 114



Illustration af predikterede sandsynligheder

i tilfælde af en enkelt kovariat



65 / 114

Model med alle tre kovariater

alle logaritmetransformeret ($\log_2$):

```
library(MASS)
```

```
m <- polr(fibrosegrad ~ lha + lpiinp + lykl40, data = fib, Hess=TRUE)
```

som giver outputtet på næste side

Vi ville egentlig også gerne have et test for antagelsen om proportional odds, men det vil R ikke umiddelbart give, da det siges at give for stor sandsynlighed for forkastelse. I SAS får vi:

```
Score Test for the Proportional Odds Assumption
Chi-Square    DF    Pr > ChiSq
      9.6967      6      0.1380
```

som siger, at antagelsen om *proportional odds* ikke er direkte forkert

66 / 114

Output fra ordinal regression i R

```
> summary(m)
Call:
polr(formula = fibrosegrad ~ lha + lpiinp + lykl40, data = fib,
      Hess = TRUE)

Coefficients:
              Value Std. Error t value
lha      0.3891    0.1660    2.344
lpiinp   0.8224    0.2602    3.161
lykl40   0.5429    0.1667    3.257

Intercepts:
              Value Std. Error t value
0|1  7.5924    1.3651    5.5617
1|2 10.0120    1.5135    6.6151
2|3 12.7770    1.6907    7.5573

Residual Deviance: 246.4663
AIC: 258.4663
(3 observations deleted due to missingness)
```

Koefficienterne for de tre blodmarkører skal tilbagetransformeres for at kunne fortolkes, se næste side.

67 / 114

Output, fortsat

Tabel med P-værdier:

```
> ctable=coef(summary(m))
> p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
> ptable <- cbind(ctable, "p value" = p)

> ptable[1:3,]
              Value Std. Error  t value      p value
lha      0.3890613    0.1660076  2.343635  0.019096839
lpiinp   0.8223711    0.2601810  3.160766  0.001573551
lykl40   0.5429406    0.1666785  3.257411  0.001124334
```

Tabel med odds ratio'er:

```
> cbind(exp(coef(m)),exp(confint(m)))
              2.5 %    97.5 %
lha      1.475595  1.073350  2.062009
lpiinp   2.275890  1.375084  3.829320
lykl40   1.721060  1.246188  2.402576
```

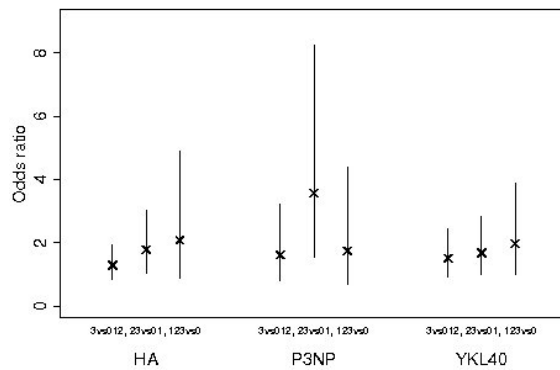
Eksempel på fortolkning:

En fordobling af markøren ykl40 giver 72.1% større odds for at være i en høj kategori

68 / 114

Afvigelse fra proportional odds antagelsen?

Helt *frie* logistiske regressioner for hver tærskel giver disse odds ratio'er:



Måske er der lidt forskel på estimerne for piiiin...

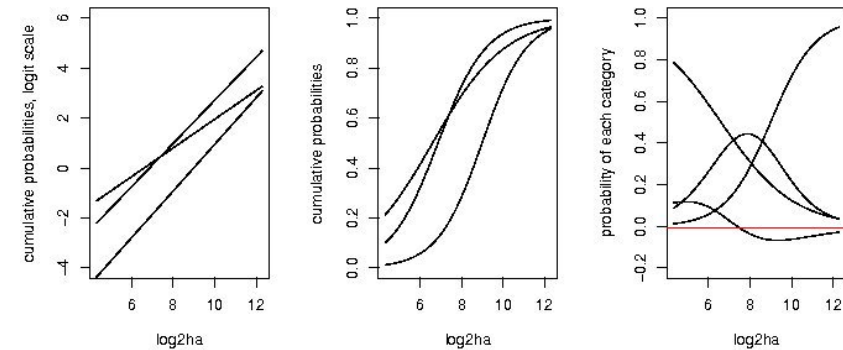
men det var altså ikke signifikant ($P=0.13$, se s. 66)

69 / 114



Hvis der ikke er proportional odds

kan vi få groteske resultater, i form af negative predikterede sandsynligheder....



70 / 114

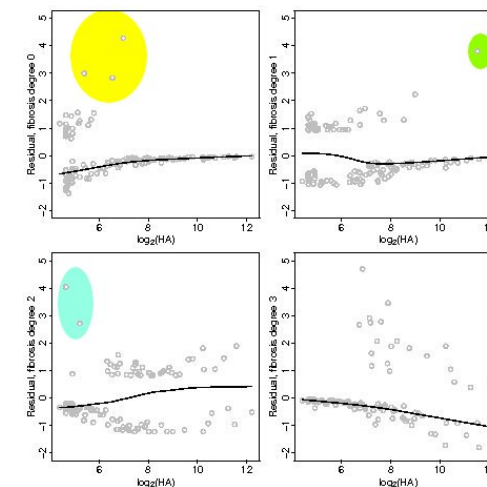


Modelkontrol

Der er flere ting, der skal checkes:

- ▶ **lineariteten** af kovariat-effekterne, på logit-skalaen, *ligesom for logistisk regression*, men der er mange residualplots: Et for hver kombination af fibrosegrad og kovariat, dvs. 12 i alt.
- ▶ **proportional odds** antagelsen
- ▶ modellens evne til faktisk at prediktere en fornuftig fibrosegrad

Residualplot for ha



Svag afvigelse fra linearitet

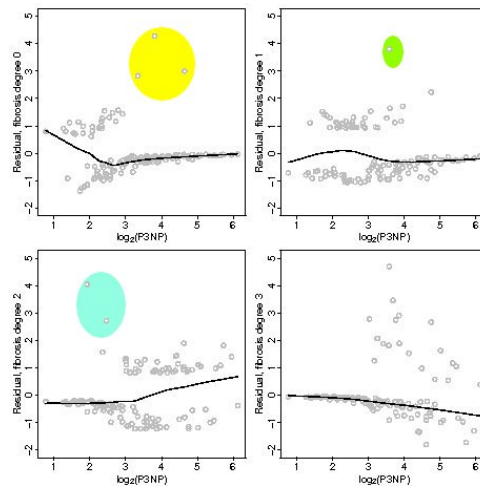
71 / 114



72 / 114



Residualplot for p3np

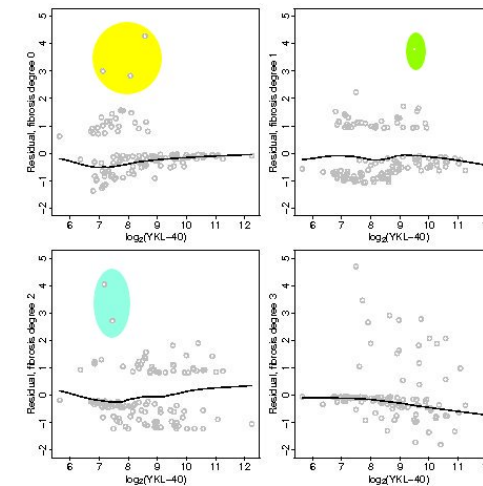


Nogen afvigelse fra linearitet

73 / 114



Residualplot for ykl40



Rimelig linearitet

74 / 114



Diagnostics

De farvede observationer skal man måske se lidt nærmere på:

- ▶ Tre gule for grad 0
som *burde* have haft en højere grad,
baseret på deres ret høje kovariat-værdier
– specielt for p3np
- ▶ En enkelt grøn for grad 1
som *burde* have haft en højere grad,
baseret på de meget høje kovariat-værdier
– specielt for ha
- ▶ To turkise for grad 2
som *burde* have haft en lavere grad,
baseret på de forholdsvis lave kovariat-værdier
– specielt for ha

75 / 114



Predikterede sandsynligheder

For hver person kan værdierne af de 3 blodmarkører benyttes til at prediktere sandsynligheder for hver af de 4 fibrosegrader.

Vi ser gerne, at disse *passer med* de faktiske fibrosegrader, altså at den predikterede sandsynlighed for den observerede fibrosegrad er så høj som muligt.

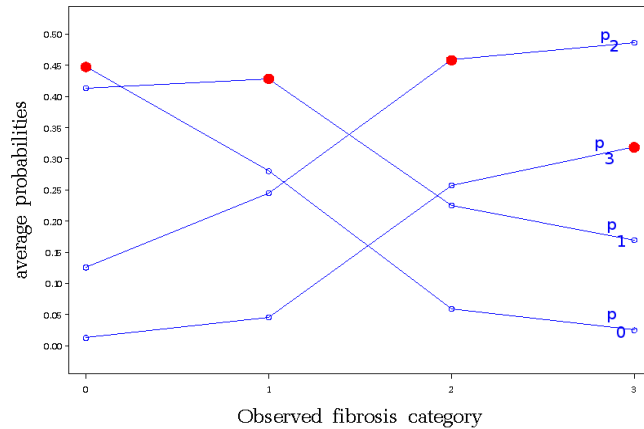
På s. 77 ses en figur, hvor

- ▶ X-aksen viser faktisk observeret fibrosegrad
- ▶ For alle personer med en bestemt observeret fibrosegrad (f.eks. 2) udregnes nu gennemsnitlig predikteret sandsynlighed for hver af de 4 fibrosegrader, og disse afsættes som punkter op ad Y-aksen. Det røde punkt svarer til den predikterede sandsynlighed for grad 2, som jo svarer til den observerede og derfor gerne skulle være høj.

76 / 114



Predikterede sandsynligheder, II



Værdierne svarende til den **korrekte fibrosegrad** er røde, og de skulle gerne være høje

77 / 114



Observeret vs. predikteret fibrosegrad

Table of degree_fibr by pred_grad

degree_fibr pred_grad

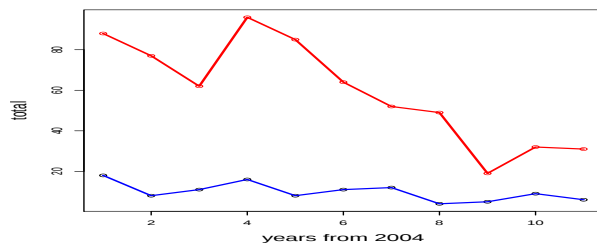
Frequency						
Col Pct	0	1	2	3	Total	
0	13	12	1	0	26	
59.09	29.27	1.96	0.00			
1	8	21	9	0	38	
36.36	51.22	17.65	0.00			
2	1	7	27	7	42	
4.55	17.07	52.94	58.33			
3	0	1	14	5	20	
0.00	2.44	27.45	41.67			
Total	22	41	51	12	126	

78 / 114



Tællelatal

Antal dræbte i trafikken fra 2004-2014, kønsopdelt, fra Danmarks Statistik



Vi lader Y_{gt} (total) betegne antallet af trafikdræbte med køn g (gender) og tid t (years_from_2004), i alt 22 observationer og 3 variable.

79 / 114



Binomialfordeling, med approksimationer

Hvis vi *vidste*, hvor mange personer, der var udsat for at blive dræbt i trafikken (N_{gt} , afhængig af køn g og årstal t), kunne vi modellere antal dræbte som en Binomialfordeling

$$Y_{gt} \sim \text{Bin}(N_{gt}, p_{gt})$$

men vi kender ikke N_{gt} , vi ved bare, at den er stor, og at p_{gt} er lille. I sådan et tilfælde **approksimeres Binomialfordelingen af en Poisson-fordeling**:

$$P(Y_{gt} = m) = \frac{\lambda_{gt}^m}{m!} \exp(-\lambda_{gt})$$

hvor $\lambda_{gt} = N_{gt}p_{gt}$ er **middelværdien, og variansen!!**.

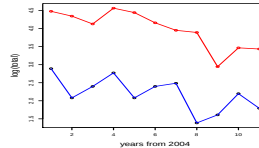
80 / 114



Regression i Poisson-fordelingen

Da middelværdien af tælleantal skal være ikke-negativ, men er uden øvre grænse, modellerer vi den på log-skala (**log-link**):

$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ hvor kovariaterne her er hhv. køn (gender) og årstal, her angivet som år siden 2004 (years_since_2004).



Figuren kunne gøre det rimeligt at se på en lineær effekt af tid, og evt. inkludere en interaktion mellem køn og tid:

Er der sket et større fald for mænd end for kvinder?

81 / 114



Regression i Poisson-fordelingen, II

Testvenlig kode:

```
model1 = glm(antal ~ gender + years.from2004
             + gender:years.from2004,
             family=poisson(link="log"),data=total)
summary(model1)
anova(model1,test="LR")
```

Estimationsvenlig kode:

```
model2 = glm(antal ~ gender + gender:years.from2004,
             family=poisson(link="log"),data=total)
summary(model2)
cbind(exp(coef(model2)),exp(confint(model2)))
```

82 / 114



Output fra Poisson-regression

med to separate lineære effekter af tid (mænd og kvinder), dvs. en **interaktion**

```
> anova(model1,test="LR")
Analysis of Deviance Table
Model: poisson, link: log
Response: antal
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			21	572.77	
gender	1	435.50	20	137.27	<2e-16 ***
years.from2004	1	88.09	19	49.18	<2e-16 ***
gender:years.from2004	1	0.52	18	48.66	0.4688

```
> cbind(exp(coef(model2)),exp(confint(model2)))[3:4,]
```

		2.5 %	97.5 %
genderK:years.from2004	0.9153843	0.8605032	0.9724935
genderM:years.from2004	0.8932410	0.8709227	0.9158761

Vi ser et estimat på ca. 10% reduktion pr. år (**faktorerne 0.8932 hhv. 0.9154**), nogenlunde ens for mænd og kvinder (P=0.47).

83 / 114



Model uden interaktion

Vi udelader den insignifikante interaktion:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.88145	0.11262	25.585	<2e-16 ***
genderM	1.80250	0.10386	17.356	<2e-16 ***
years.from2004	-0.10939	0.01186	-9.219	<2e-16 ***

```
> cbind(exp(coef(model3)),exp(confint(model3)))[2:3,]
Waiting for profiling to be done...
```

		2.5 %	97.5 %
genderM	6.0648148	4.9714262	7.4724675
years.from2004	0.8963825	0.8756712	0.9173715

Begge effekter (køn og årstal) er signifikante:

- ▶ Ca. 10% fald i risiko pr. år
- ▶ Mænd har en ca. 6 gange så stor risiko som kvinder

84 / 114



Problemer med Poisson-fordelingen

Det er mere reglen end undtagelsen, at Poisson-fordelingen passer dårligt, fordi variansen er større end middelværdien (i Poisson-fordelingen er disse *ens*, som nævnt s. 80)

Vi har altså meget ofte **overspredning**:

- ▶ formentlig pga oversete kovariater
- ▶ med den konsekvens, at standard errors undervurderes
- ▶ og der findes **alt for stærke signifikanser**

Korrektion for overspredning (baseret på deviance)

```
model3.quasi = glm(antal ~ gender + years.from2004,
  family=quasipoisson(link="log"),data=total)
```

85 / 114



Output ved korrektion for overspredning

Først selve estimatet for overspredningen:

$$1.5830 = \sqrt{2.5059}$$

(Dispersion parameter for quasipoisson family taken to be 2.505865)

Residual deviance: 49.181 on 19 degrees of freedom

Når der korrigeres for overspredningen, får man lidt bredere konfidensintervaller (sammenlign til s. 84):

```
> cbind(exp(coef(model3.quasi)),exp(confint(model3.quasi)))[2:3,]
```

		2.5 %	97.5 %
genderM	6.0648148	4.4455397	8.4824737
years.from2004	0.8963825	0.8637131	0.9297486

86 / 114



Sammenlign med normalfordelingsanalyse

Da vi ikke har nogen antal på 0, kan vi tillade os at analysere **logaritmer**, og da antallene er rimeligt store, kan vi forsøge os med en *almindelig* model, dvs. en model baseret på en normalfordelingsantagelse:

Efter tilbagetransformation finder vi så

```
> model3.nf = lm(logantal ~ gender + years.from2004, data=total)
> cbind(exp(coef(model3.nf)),exp(confint(model3.nf)))[2:3,]
              2.5 %    97.5 %
genderM      6.002605 4.4266945 8.1395410
years.from2004 0.897187 0.8550093 0.9414455
```

hvilket ses kun at være en anelse anderledes end det, vi fandt ved Poisson-analysen s. 86, (inkluderende overspredning).

87 / 114



Konklusion vedr. trafikuheld

- ▶ Der er flere mænd end kvinder, der bliver dræbt i trafikken, ca. 6 gange så mange
 - ▶ De færdes måske mere?
 - ▶ De er måske mere uforsigtige?
 - ▶ Alder er måske en skjult confounder?

Det kunne jo være, at der var mange flere mandlige bilister i de yngre aldersklasser... men dette er ikke registreret

Se mere detaljerede figurer på de næste sider

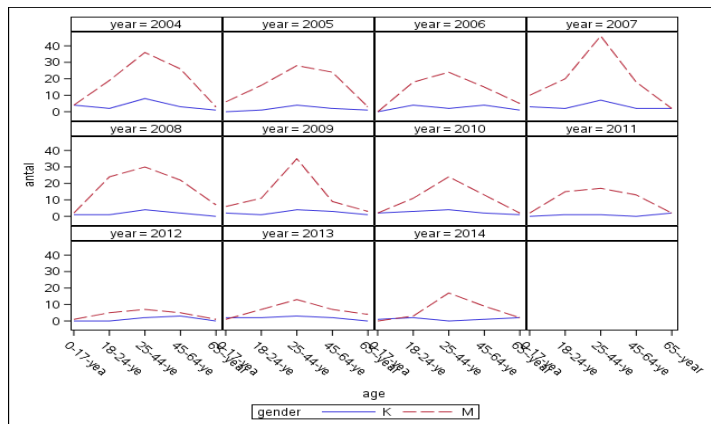
- ▶ Der er sket en reduktion i antallet af trafikdræbte over årene 2004-2014, ca. 10% pr. år

88 / 114



Trafikuheld, opdelt på årstal

som funktion af alder:

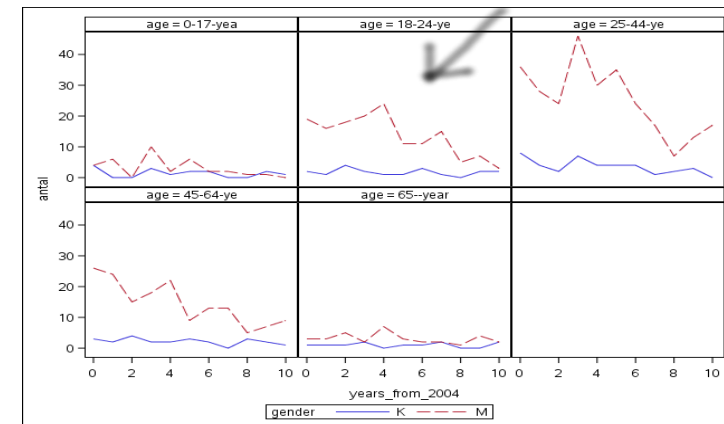


89 / 114



Trafikuheld, opdelt efter alder

som funktion af år siden 2004:



90 / 114



Trafikuheld, kun for 18-24-årige

med korrektion for overspredning på $1.126 = \sqrt{1.268}$
(se kode s. 114)

```
> summary(m1.quasi)
```

```
Call:
glm(formula = antal ~ gender + years.from2004 + gender:years.from2004,
     family = quasipoisson(link = "log"), data = trafik18.24)
```

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.74892    0.45635   1.641   0.118
genderM         2.39799    0.47985   4.997 9.34e-05 ***
years.from2004 -0.04226    0.08215  -0.514   0.613
genderM:years.from2004 -0.08084    0.08764  -0.922   0.368
(Dispersion parameter for quasipoisson family taken to be 1.268374)
Residual deviance: 23.199 on 18 degrees of freedom
```

Interaktionen er ikke signifikant ($P=0.37$), men vi vil alligevel gerne se de to separate estimater (næste side).

91 / 114



Trafikuheld for 18-24-årige, fortsat

Estimationsvenlig model:

```
> m2.quasi = glm(antal ~ gender + gender:years.from2004,
+   family=quasipoisson(link="log"), data=trafik18.24)
> cbind(exp(coef(m2.quasi)), exp(confint(m2.quasi)))[3:4,]
              2.5 %      97.5 %
genderK:years.from2004 0.9586225 0.8125242 1.1259631
genderM:years.from2004 0.8841774 0.8320417 0.9379807
```

Det ser ud som om der er sket en noget større reduktion for mænd end for kvinder, men forskellen er som tidligere nævnt ikke signifikant ($P=0.36$)

92 / 114



Sammenlign med normalfordelingsanalyse

Nu har vi et enkelt år, hvor der ikke er nogen kvinder, der er dræbt i trafikken, og så kan vi ikke bruge en normalfordelingsmodel på logaritmerne....[fordi man ikke kan tage logaritmen til 0](#)

Desuden er antallene generelt så små, når vi kun ser på en enkelt aldersklasse, at en normalfordelingsapproximation ikke ville være særlig god alligevel....

93 / 114



APPENDIX

Programbidder svarende til diverse slides:

- ▶ To-gange-to tabeller, s. 95
- ▶ Plot af binære data med loess-udglatning, s. 96
- ▶ Logistisk regression, s. 97-98
- ▶ Modelkontrol, s. 101-105
- ▶ Diagnostics, s. 106
- ▶ Alternative links, s. 107-109
- ▶ Ordinale data, s. 110
- ▶ Tælleletal, s. 111-114

94 / 114



To-gange-to tabeller

Slide 7-10

```
fb = matrix(c(1,6,119,144), nrow=2)
rownames(fb) = c("Piger","Drenge")
colnames(fb) = c("ja","nej")
```

```
addmargins(fb)
```

```
prop.table(fb,1)*100
```

```
chisq.test(fb)$expected
chisq.test(fb)
fisher.test(fb)
prop.test(fb)
```

```
install.packages("epitools")
library("epitools")
```

```
epitab(fb,method="riskratio",rev="columns")$tab
```

```
epitab(fb,method="oddsratio",rev="rows")$tab
```

95 / 114



Figurer

Slide 13

Plot af infektion vs. alder, med loess-smoother:

```
inff <-
read.table("http://publicifsv.sund.ku.dk/~lts/basal/Rdata/bremmel.txt",
header=T,na.strings=c("."))
```

```
inf = inff[order(inff$alder),]
inf$over2timer = as.numeric(inf$optid > 120)
```

```
inf0 <- subset(inf, over2timer==0)
inf1 <- subset(inf, over2timer==1)
```

```
loess0 <- loess(infektion ~ alder, data=inf0, span=0.80)
pred0 = predict(loess0)
loess1 <- loess(infektion ~ alder, data=inf1, span=0.80)
pred1 = predict(loess1)
```

```
plot(inf0$alder, inf0$infektion,
     ylab="infektion", xlab="alder", col="red", cex.lab=1.5)
lines(inf0$alder, pred0, col = "red", lwd = 3, lty = 1)
points(inf1$alder, inf1$infektion, col="blue")
lines(inf1$alder, pred1, col = "blue", lwd = 3, lty = 1)
legend(20, 0.8, legend=c("over 2","under 2"), pch=16,
      col=c("blue", "red"), cex=1.5)
```

96 / 114



Logistisk regression

Slide 23-24

```
model1 = glm(formula=infektion ~
  factor(over2timer)+alder,
  family=binomial(link="logit"), data=inf)
```

hvorefter vi kan benytte funktioner såsom

```
summary(model1)
confint(model1)
drop1(model1, test = "Chisq")

cbind(exp(coef(model1)),exp(confint(model1)))
```

97 / 114



Figur af predikterede sandsynligheder

Slide 26

Prediktion og figur til højre

```
model1 = glm(formula=infektion ~
  factor(over2timer)+alder,
  family=binomial(link="logit"), data=inf,na.action=na.exclude)

pred.0 <- predict(model1)[inf$over2timer==0]
pred.1 <- predict(model1)[inf$over2timer==1]

pred.00 <- predict(model1)[inf$over2timer==0 & inf$infektion==0]
pred.01 <- predict(model1)[inf$over2timer==0 & inf$infektion==1]
pred.10 <- predict(model1)[inf$over2timer==1 & inf$infektion==0]
pred.11 <- predict(model1)[inf$over2timer==1 & inf$infektion==1]

plot(inf$alder, predict(model1),type="n",
  ylab="infektion", xlab="alder", cex.lab=1.5)
points(inf00$alder, pred.00, col="green", cex=2,lwd=2)
points(inf01$alder, pred.01, col="red", cex=2, lwd=2)
lines(inf0$alder, pred.0, col = "blue", lwd = 3, lty = 1)
points(inf10$alder, pred.10, col="green", cex=2, lwd=2)
points(inf11$alder, pred.11, col="red", cex=2, lwd=2)
lines(inf1$alder, pred.1, col = "blue", lwd = 3, lty = 1)
```

98 / 114



Figur af predikterede sandsynligheder, fortsat

Slide 26

Figur til venstre, tilbagetransformeret til sandsynligheder

```
ny.0 = data.frame(alder=seq(5,90,0.5),over2timer=0)
ny.1 = data.frame(alder=seq(5,90,0.5),over2timer=1)

ny.0$pred = exp(predict(model1, ny.0))
ny.1$pred = exp(predict(model1, ny.1))

plot(inf$alder, exp(predict(model1)),type="n",
  ylab="predicted probability", xlab="alder", col="red", cex.lab=1.5)
points(inf00$alder, exp(pred.00), col="green", cex=2,lwd=2)
points(inf01$alder, exp(pred.01), col="red", cex=2, lwd=2)
lines(ny.0$alder, ny.0$pred, col = "blue", lwd = 3, lty = 1)
points(inf10$alder, exp(pred.10), col="green", cex=2, lwd=2)
points(inf11$alder, exp(pred.11), col="red", cex=2, lwd=2)
lines(ny.1$alder, ny.1$pred, col = "blue", lwd = 3, lty = 1)
```

99 / 114



Interaktionsmodel

Slide 33

Testvenlig kode:

```
model2 = glm(formula=infektion ~
  factor(over2timer)+alder+factor(over2timer):alder,
  family=binomial(link="logit"), data=inf,na.action=na.exclude)

summary(model2)
cbind(exp(coef(model2)),exp(confint(model2)))
```

Estimationsvenlig kode:

```
model2 = glm(formula=infektion ~
  factor(over2timer)+factor(over2timer):alder,
  family=binomial(link="logit"), data=inf,na.action=na.exclude)

summary(model2)
cbind(exp(coef(model2)),exp(confint(model2)))
```

100 / 114



Goodness of fit

Slide 36-38

```
install.packages("ResourceSelection")
library(ResourceSelection)

hl = hoslem.test(inf$infektion, fitted(model2), g=10)

> hl

> cbind(hl$observed,hl$expected)
```

101 / 114



Check af linearitetsantagelsen for alder

Slide 39

- ▶ Alderseffekt modelleret med både alder og log(alder) (output s. 39-40). Definer:

```
inf$alder.minus50 = inf$alder - 50
inf$logalder50=log(inf$alder)/log(1.1)-log(50)/log(1.1)
```

og brug modellen

```
model3 = glm(formula=infektion ~
factor(over2timer)+alder.minus50+logalder50,
family=binomial(link="logit"), data=inf,na.action=na.exclude)
```

```
summary(model3)
cbind(exp(coef(model3)),exp(confint(model3)))
```

102 / 114



Check af linearitetsantagelsen for alder, II

Slide 42

- ▶ Alderseffekt modelleret som lineær spline (output s. 41-42):
Definer

```
inf$alder.over60=(inf$alder>60)*(inf$alder-60)
inf$alder.over75=(inf$alder>75)*(inf$alder-75)
```

og brug modellen

```
model4 = glm(formula=infektion ~
factor(over2timer)+alder+alder.over60+alder.over75,
family=binomial(link="logit"), data=inf,na.action=na.exclude)
```

103 / 114



Plot af standardiserede residualer

Slide 43

Modellen er den på s. 23,

```
inf$resid1 <- residuals(model1, "pearson")
```

```
inf0 <- subset(inf, over2timer==0)
inf1 <- subset(inf, over2timer==1)
```

```
loess0 <- loess(resid1 ~ alder, data=inf0, span=0.80)
pred0 = predict(loess0)
loess1 <- loess(resid1 ~ alder, data=inf1, span=0.80)
pred1 = predict(loess1)
```

```
plot(inf$alder, inf$resid1,
ylab="residuals", xlab="alder", type="n", cex.lab=1.5)
points(inf0$alder, inf0$resid1, col="red")
lines(inf0$alder, pred0, col = "red", lwd = 3, lty = 1)
points(inf1$alder, inf1$resid1, col="blue")
lines(inf1$alder, pred1, col = "blue", lwd = 3, lty = 1)
```

104 / 114



Plot af kumulerede residualer

Slide 44

Figuren med simulationer af de kumulerede residualer er dannet ved at benytte pakken gof:

```
install.packages("gof")
library(gof)

cumcheck1 = cumres(model1,R=1000,variable="alder")

plot(cumcheck1)
```

105 / 114



Diagnostics

Slide 46-47

```
plot(model1,which=4, cex.lab=1.5)

cook1=cooks.distance(model1)

plot(inf$alder,cook1,col=2-inf$over2timer,
      ylab="Cook's Distance", xlab="alder", cex.lab=1.5, lwd=2)
legend(60, 0.25, title="varighed", legend=c("over 2","under 2"), pch=16,
      col=c("black", "red"), cex=1.5)

plot(inf$alder,dfbetas(model1)[,3],col=2-inf$infektion,
      ylab="DfBeta", xlab="alder", cex.lab=1.5, lwd=2)
legend(20, -0.2, title="infektion",legend=c("ja","nej"), pch=16,
      col=c("black", "red"), cex=1.5)

inf[cook1==max(cook1),]
```

106 / 114



Relativ risiko for farveblindhed

Slide 51-52

```
gender = as.factor(c(rep("Piger",120),rep("Drenge",150)))
farve = c(rep("nej",119),rep("ja",1),rep("nej",144),rep("ja",6))

farveblind=(farve=="ja")

model1 = glm(farveblind ~ relevel(gender,ref="Piger"),
             family=binomial(link="log"))

> summary(model1)

> cbind(exp(coef(model1)),exp(confint(model1)))

> cbind(exp(coef(model1)),exp(confint.default(model1)))
```

107 / 114



Test for trend

Slide 55-56

Modellen antager linearitet i skostørrelse: $p_i = \alpha - \beta \times \text{sko}_i$

```
skonummer = c(rep(3.5,22),rep(4,35),rep(4.5,42),rep(5,48),
               rep(5.5,54),rep(6,150))
snit = c(rep("ja",5),rep("nej",17),rep("ja",7),rep("nej",28),
          rep("ja",6),rep("nej",36),rep("ja",7),rep("nej",41),
          rep("ja",8),rep("nej",46),rep("ja",10),rep("nej",140))

> table(skonummer,snit)
      snit
skonummer  ja  nej
      3.5    5  17
       4     7  28
      4.5    6  36
       5     7  41
      5.5    8  46
       6    10 140

kejersersnit=(snit=="ja")
model2 = glm(kejersersnit ~skonummer, family=binomial(link="identity"))
```

108 / 114



Modelkontrol af lineariteten, kejsersnit

Slide 57-58

Trick (i alle former for regression):

Når kovariaten kun antager så få værdier, kan linearitetsantagelsen checkes ved at sætte en kopi af `skonummer` ind som faktor oveni, og så teste, om denne kan undværes:

```
model3 = glm(kejsersnit ~ skonummer + factor(skonummer),
             family=binomial(link="identity"))
anova(model3, test="LR")
```

Modellen ville være den samme, hvis `skonummer` blev udeladt, men så ville vi ikke få testet, om modellen med **kun skonummer** var fornuftig

109 / 114



Proportional odds model

Slide 66-67

Logistisk regression for hver tærskel, med **samme afhængighed** af kovariaterne, som her alle er $\log_2$ -transformerede:

```
library(MASS)

m <- polr(fibroseggrad ~ lha + lpiiinp + lykl40, data = fib, Hess=TRUE)
```

Odds ratio vil **ikke afhænge af cutpoint**,

```
> summary(m)

> ctable=coef(summary(m))
> p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
> ptable <- cbind(ctable, "p value" = p)

> ptable[1:3,]

> cbind(exp(coef(m)), exp(confint(m)))
```

110 / 114



Poisson analyse, med interaktion

Slide 82-83

Testvenlig kode:

```
model1 = glm(antal ~ gender + years.from2004
             + gender:years.from2004,
             family=poisson(link="log"), data=total)
summary(model1)
anova(model1, test="LR")
cbind(exp(coef(model2)), exp(confint(model2))) [3:4,]
```

Estimationsvenlig kode:

```
model2 = glm(antal ~ gender + gender:years.from2004,
             family=poisson(link="log"), data=total)
summary(model2)
cbind(exp(coef(model2)), exp(confint(model2)))
```

111 / 114



Poisson analyse

Slide 85-86

Her bruges `family=quasipoisson` for at korrigere for overspredning, ellers helt som s. 111

```
model3.quasi = glm(antal ~ gender + years.from2004,
                  family=quasipoisson(link="log"), data=total)

summary(model3.quasi)

cbind(exp(coef(model3.quasi)), exp(confint(model3.quasi))) [2:3,]
```

112 / 114



Normalfordelingsmodel for tællel

Slide 87

foretages på logaritmetransformerede data:
total\$logantal=log(total\$antal):

```
model3.nf = lm(logantal ~ gender + years.from2004, data=total)
summary(model3.nf)
cbind(exp(coef(model3.nf)),exp(confint(model3.nf)))[2:3,]
```

113 / 114



Poisson analyse, kun for 18-24 årige

Slide 91-92

```
trafik18.24 <- subset(trafik, age=="18-24-ye")

m1.quasi = glm(antal ~ gender + years.from2004 + gender:years.from2004,
               family=quasipoisson(link="log"),data=trafik18.24)
summary(m1.quasi)

m2.quasi = glm(antal ~ gender + gender:years.from2004,
               family=quasipoisson(link="log"),data=trafik18.24)

cbind(exp(coef(m2.quasi)),exp(confint(m2.quasi)))[3:4,]
```

114 / 114

