Faculty of Health Sciences

Basal Statistik

Logistisk regression mm.

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`

Typer af outcome

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

Eksempel fra tidligere: Farveblindhed og køn

	Farveblindhed?		
	nej	ja	Total
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: $p_d = p_p$

Mål for kønseffekt på farveblindhed

f.eks. drenge vs. piger:

$$\text{Differens:} \quad p_d - p_p$$

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

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

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)

Test af hypotesen om uafhængighed

Se kode s. 93

gender	farveblind		
Frequency			
Expected			
Row Pct	ja	nej	Total
dreng	6	144	150
	3.8889	146.11	
	4.00	96.00	
pige	1	119	120
	3.1111	116.89	
	0.83	99.17	
Total	7	263	270

Output, fortsat

```
Statistics for Table of gender by farveblind

Statistic                       DF       Value      Prob
------------------------------------------------------------
Chi-Square                       1      2.6472    0.1037
Likelihood Ratio Chi-Square      1      3.0022    0.0832
Continuity Adj. Chi-Square       1      1.5418    0.2144

 WARNING: 50% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.

       Fisher's Exact Test
---------------------------------
Two-sided Pr <= P            0.1364
```

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

Output, fortsat

Differens mellem sandsynlighederne
$\hat{p}_d - \hat{p}_p = 0.0317, CI = (-0.0890, 0.1517)$:

Statistics for Table of gender by farveblind

Column 1 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.0400	0.0160	0.0053	0.0747	0.0148	0.0850
Row 2	0.0083	0.0083	0.0000	0.0288	0.0002	0.0456
Total	0.0259	0.0097	0.0051	0.0467	0.0105	0.0527
Difference	0.0317	0.0180	-0.0112	0.0745	-0.0890	0.1517

Difference is (Row 1 - Row 2)

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

Output, fortsat

Odds ratio (OR=4.96) og risk ratio = relativ risiko (RR=4.80), option `relrisk` (`Col1` er *farveblind="ja"*):

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	4.9583	0.5887	41.7605
Cohort (Col1 Risk)	4.8000	0.5858	39.3291
Cohort (Col2 Risk)	0.9681	0.9333	1.0041

Der er altså ca. 5 gange så mange drenge som piger, der er farveblinde.

Bemærk, at OR≈RR, da farveblindhed er sjældent forekommende

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

Variable	N	Mean	Std Dev	Minimum	Maximum
infektion	194	0.1185567	0.3241026	0	1.0000000
optid	194	93.9432990	63.7111510	5.0000000	390.0000000
alder	194	55.9690722	25.0143808	7.0000000	90.0000000

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:

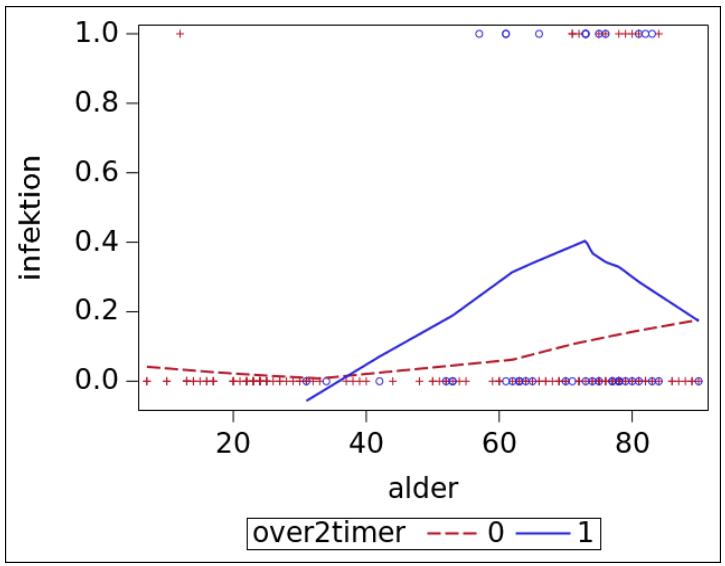
- Patientens alder, som lineær effekt
  (evt. som alder pr. 10-år ved at skalere til `alder10=alder/10`)
- Indikator for operationsvarighed over 2 timer:

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

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

Figurer

Plot af infektion vs. alder, med `loess`-smoother
Blå: over 2 timer, Rød: under 2 timer




```
proc sgplot data=infektion;
    loess Y=infektion X=alder /
    group=over2timer smooth=0.8;
run;
```

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

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

Model for p_i'erne, fortsat

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

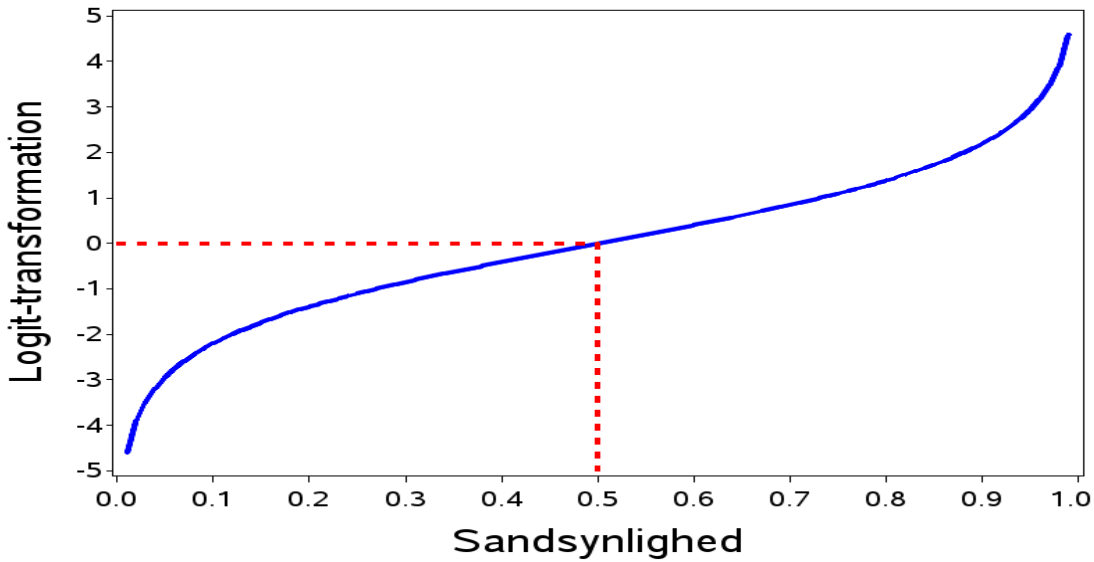
- Derfor bruger man en funktion $g(p_i)$, kaldet link-funktionen, og antager linearitet på denne skala.

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

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

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

Logit funktionen

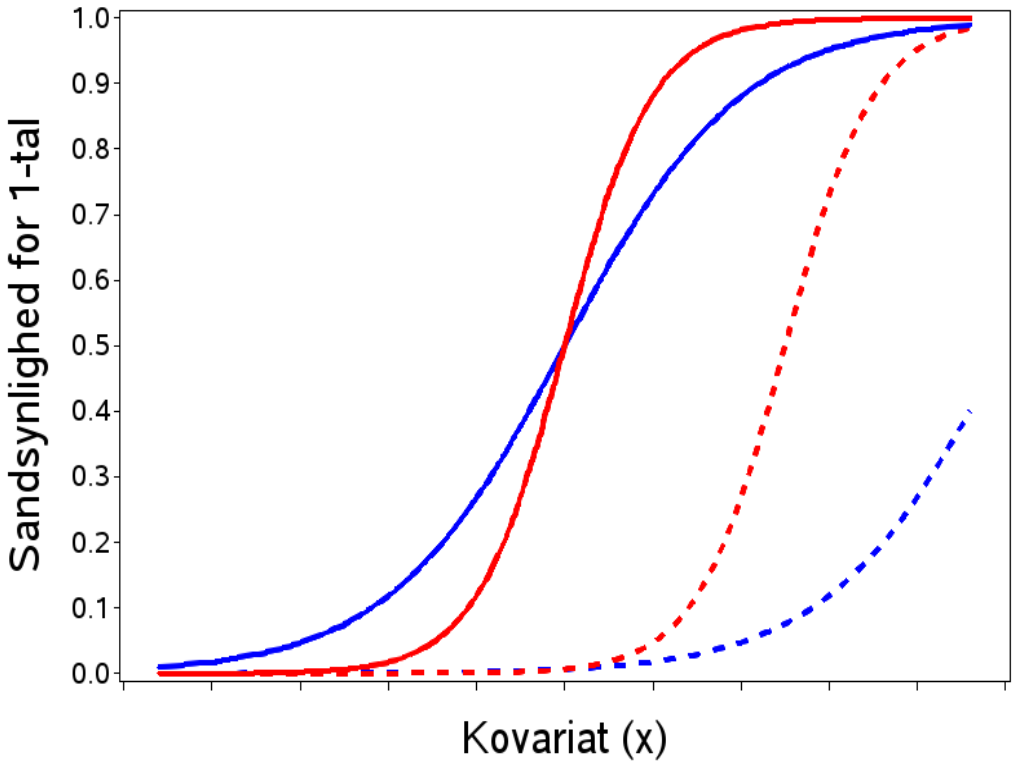


Sandsynligheder tæt på 0 giver store negative `logit`-værdier,
Sandsynligheder tæt på 1 giver store positive `logit`-værdier

Så på denne skala giver det mening at bygge lineære modeller

Logistiske kurver – for en enkelt kovariat, x

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



Parameterværdier:

- β_0: -5, 0
- β_1: 1, 2

Fortolkning af parametre - for en enkelt kovariat

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

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

$$\text{Person B: } \text{logit}(p_B) = \beta_o + \beta_1 \times (a + 10)$$
$$\text{Person A: } \text{logit}(p_A) = \beta_o + \beta_1 \times a$$
$$\text{Forskel: } \text{logit}(p_B) - \text{logit}(p_A) = \beta_1 \times 10$$

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.

Fortolkning af parametre - for to kovariater

Nu er

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

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) $x_{2A} = x_{2B} = b$:

$$\begin{aligned}
\text{Person B:}\quad & \text{logit}(p_B) = \beta_o + \beta_1 \times (a + 10) + \beta_2 \times b \\
\text{Person A:}\quad & \text{logit}(p_A) = \beta_o + \beta_1 \times a \qquad\quad + \beta_2 \times b \\
\text{Forskel:}\quad & \text{logit}(p_B) - \text{logit}(p_A) = \beta_1 \times 10
\end{aligned}$$

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

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

`alder_minus50=alder-50`

eller (i SAS) benytte estimate-sætninger

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

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) (`dist=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 (`event=''1''`)

Logistisk regression i praksis

De to procedurer i SAS

- Den *nemme*: `PROC LOGISTIC`:

```
proc logistic data=infektion;
  class over2timer / param=glm;
  model infektion(event="1") = over2timer alder;
run ;
```

- Den *der kan noget mere:* `PROC GENMOD`:

```
proc genmod descending data=infektion;
  class over2timer;
  model infektion = over2timer alder
        / dist = binomial link = logit ;
run ;
```

Yderligere kode bagest, s. 95-102

Output fra logistisk regression

her fra PROC LOGISTIC

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.7362	1.0786	19.2823	<.0001
over2timer 1	1	1.3292	0.4729	7.8986	0.0049
over2timer 0	0	0	.	.	.
alder	1	0.0355	0.0149	5.6610	0.0173

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
over2timer 1 vs 0	3.778	1.495	9.546
alder	1.036	1.006	1.067

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

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

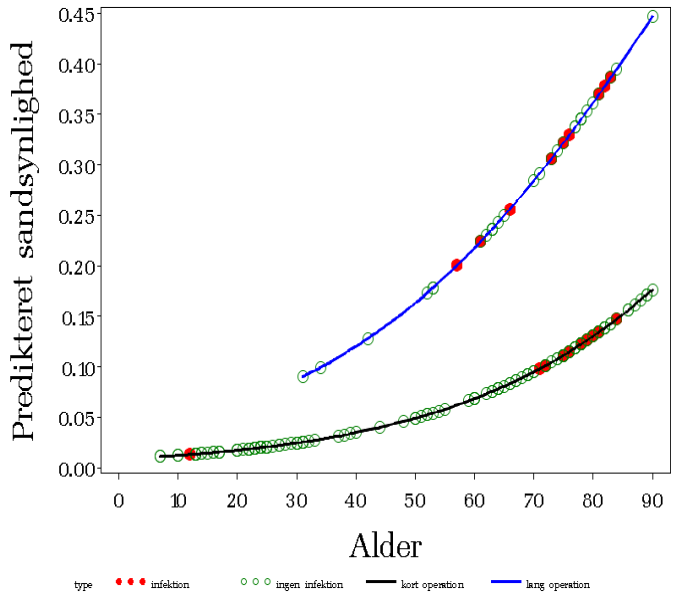
Kommentarer til output

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

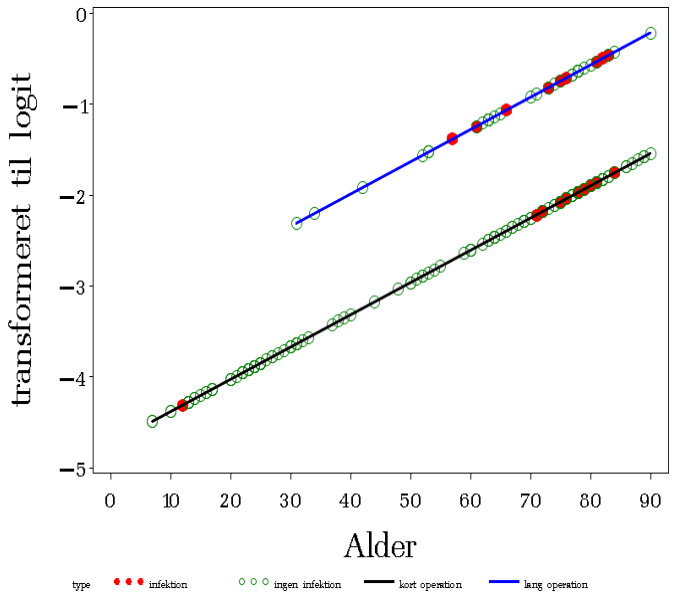
- Personer med en operationsvarighed over 2 timer har en odds for sårinfektion, der er 3.78 gange højere end dem med en operationsvarighed under 2 timer på samme alder (CI: 1.50, 9.55).
- For en patientpopulation, der har alderen X+10 år, vil odds (forholdet mellem patienter der får hhv. ikke får sårinfektion) 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: 6%-91% (tilbagetransformeret fra 1.006^{10} hhv. 1.067^{10})

Predikterede sandsynligheder for postoperativ sårinfektion

På sandsynligheds-skala



På logit-skala



Kurverne er (parallelle) rette linier på logit-skala

En figur svarende til kurverne til venstre genereres automatisk i SAS, `PROC LOGISTIC`, (den til højre er kun for forståelsen).

*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.

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.

Konsekvens for randomiserede undersøgelser

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

- Hvis der er restriktive inklusionskriterier,
  bliver OR formentlig stort
- Hvis der "blot" er samplet 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.

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

Analysis Variable : alder

over2timer	N Obs	Mean	Minimum	Maximum
0	152	52.4013158	7.0000000	90.0000000
1	42	68.8809524	31.0000000	90.0000000

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

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

OR estimater for infektion i forskellige modeller

Outcome: `infektion`		Effekt af	
Kovariater i model		10 år ældre OR (CI)	operationstid $>$ 2 timer OR (CI)
kun `alder`		1.48 (1.13, 1.93) P=0.0043	–
kun `over2timer`		–	5.13 (2.07, 12.71) P=0.0004
`alder` og `over2timer`		1.43 (1.06, 1.91) P=0.017	3.78 (1.50, 9.55) P=0.0049

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?

Interaktion

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

Inkluder interaktionen mellem operationsvarighed og alder, i form af effekten `over2timer*alder`

Analysis of Maximum Likelihood Estimates

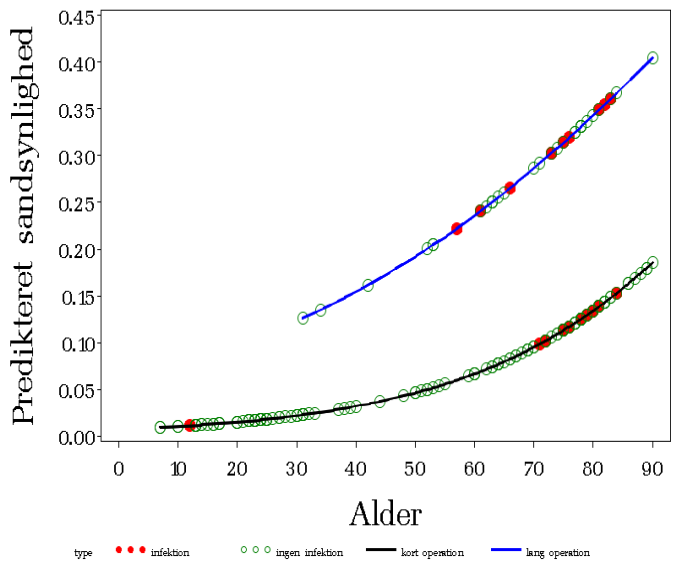
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.9485	1.2864	14.7981	0.0001
over2timer	1	1	2.1998	2.4780	0.7881	0.3747
over2timer	0	0	0	.	.	.
alder		1	0.0385	0.0178	4.6879	0.0304
alder*over2timer	1	1	-0.0123	0.0344	0.1275	0.7210
alder*over2timer	0	0	0	.	.	.

Her ser vi kun på P-værdierne

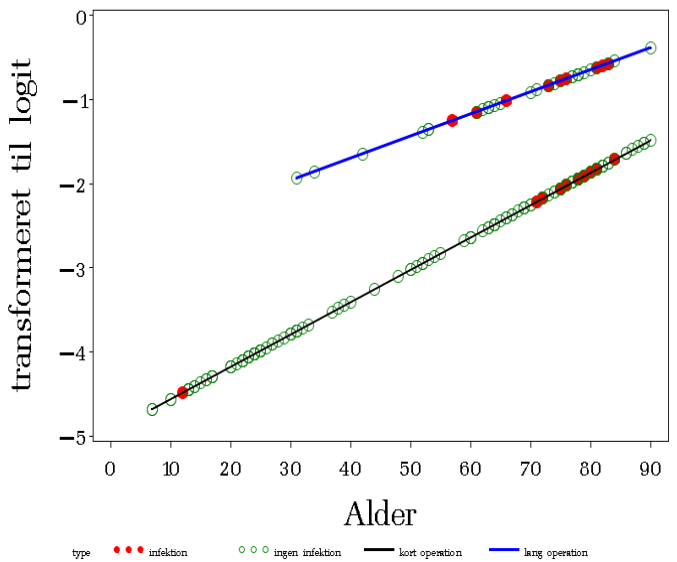
Hvis (når) vi vil vide noget om Odds Ratioer, må vi selv specificere yderligere...eller opdele

Interaktion mellem alder og operationsvarighed

På sandsynligheds-skala



På logit-skala



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

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. 103)
- Linearitet:
  Tilføj logaritmeret kovariat, Lineære splines,
  eller evt. kvadratled ("velkendt", s. 39-42, kode s. 104)
- Additivitet:
  Check for interaktion (har vi lige gjort s. 33-34, "velkendt")

Grafisk modelkontrol:

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

* Overall Goodness-of-fit

for modellen s. 23 (se kode s. 103)

- 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 $\frac{(\text{OBS}-N\hat{p})^2}{N\hat{p}(1-\hat{p})}$ sammenlægges til en approksimativ χ^2-teststørrelse med 8 frihedsgrader (antal grupper minus 2)

Her finder vi $\chi^2 = 6.67 \sim \chi^2(8) \Rightarrow P = 0.57$ og dermed intet tegn på problemer med modellen

Goodness-of-fit i praksis

sædvanligvis kaldt Hosmer-Lemeshow-testet:

```
proc logistic data=infektion;
  class over2timer / param=GLM;
  model infektion(event="1") = over2timer alder
        / lackfit link=logit ;
run ;
```

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=0.57).

Output fra *lackfit* i LOGISTIC

Se kode foregående side

Partition for the Hosmer and Lemeshow Test

		infektion = 1		infektion = 0	
Group	Total	Observed	Expected	Observed	Expected
1	19	1	0.26	18	18.74
2	19	0	0.36	19	18.64
3	19	0	0.45	19	18.55
4	19	0	0.99	19	18.01
5	20	2	1.73	18	18.27
6	18	3	1.96	15	16.04
7	21	3	2.61	18	18.39
8	19	2	2.79	17	16.21
9	21	7	5.07	14	15.93
10	19	5	6.79	14	12.21

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.6674	8	0.5729

Med en P-værdi på 0.57 ses ingen generel afvigelse fra modellen

men testet er som nævnt meget ustabilt ved små datasæt

Alderseffekt modelleret med både alder og log(alder)

$$\text{logit}(p) = \beta_o + \beta_1 \cdot 1_{\text{optid}>2} + \beta_2 \cdot (\text{alder-50}) + \beta_3 \cdot (\log_{1.1}(\text{alder}) - \log_{1.1}(50))$$

Output (se kode s. 104):

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.2099	0.5731	31.3681	<.0001
over2timer	1	1	1.3931	0.4875	8.1671	0.0043
over2timer	0	0	0	.	.	.
alder_minus50		1	0.0693	0.0510	1.8426	0.1746
logalder50		1	-0.1515	0.2113	0.5138	0.4735

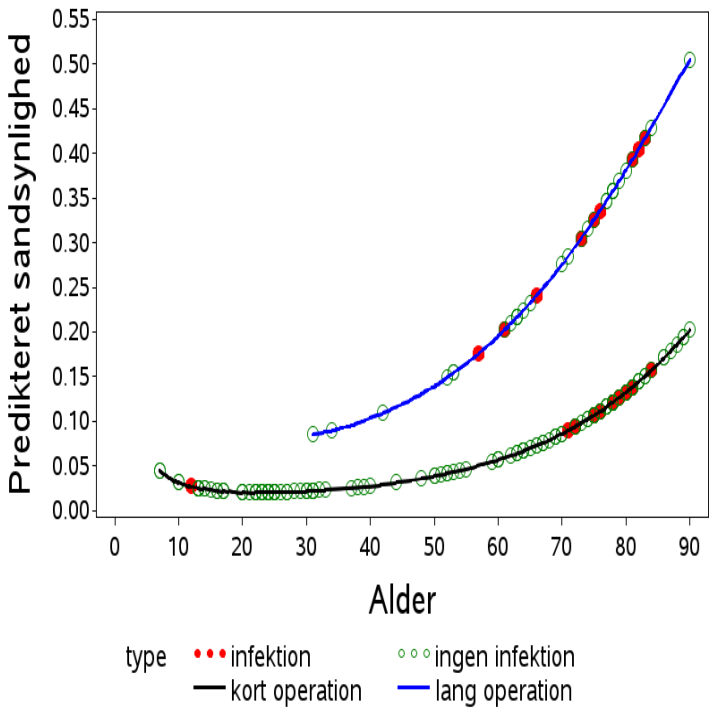
Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
over2timer	1 vs 0	4.027	1.549	10.470

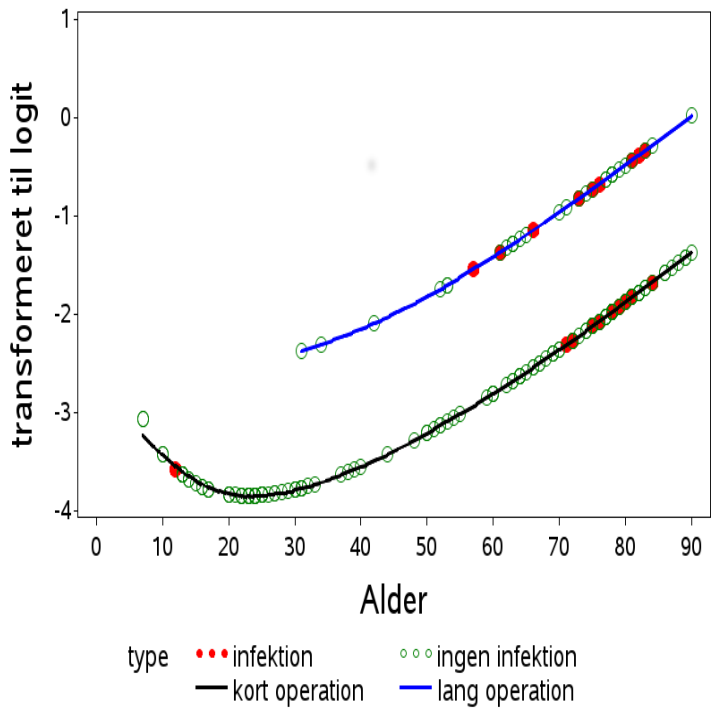
Meget ringe indikation af afvigelse fra linearitet ($P = 0.47$)

Effekten af alder og log(alder)

På sandsynligheds-skala



På logit-skala



Effekten af alder som lineær spline

med knæk ved 60 og 75 år får vi outputtet (kode s. 104):

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.6933	1.6096	8.5023	0.0035
over2timer	1	1	1.3143	0.4905	7.1783	0.0074
over2timer	0	0	0	.	.	.
alder		1	0.0284	0.0332	0.7307	0.3927
alder_over60		1	0.0730	0.0819	0.7956	0.3724
alder_over75		1	-0.2002	0.1229	2.6523	0.1034

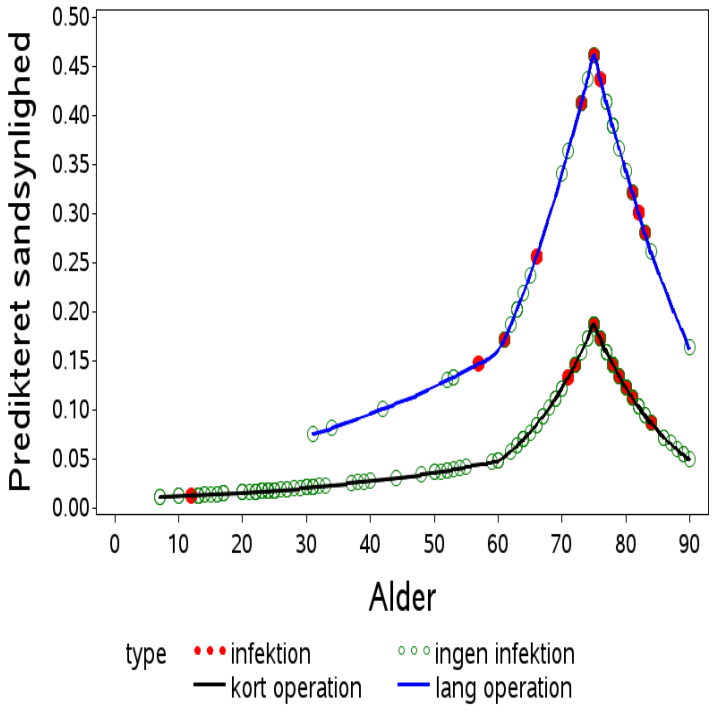
Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
over2timer	1 vs 0	3.722	1.423	9.735

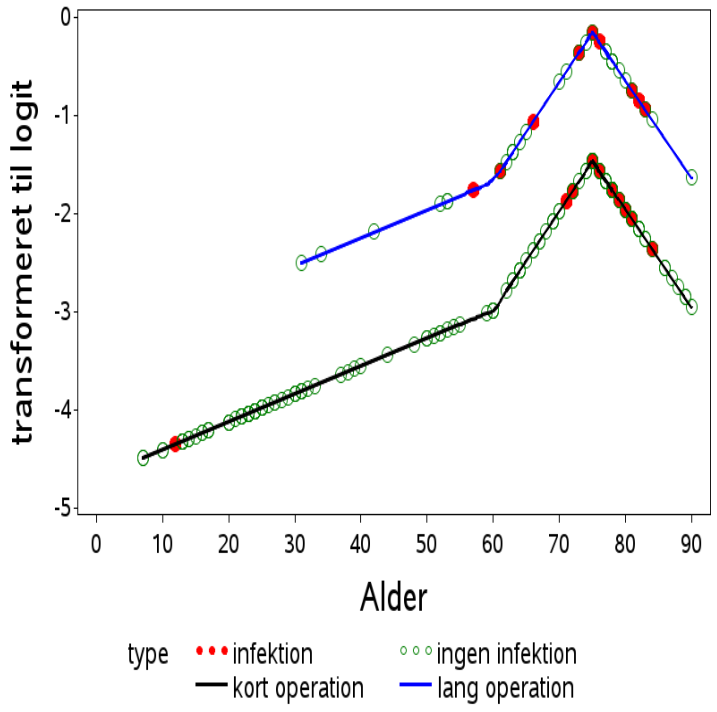
En vis indikation for afvigelse fra linearitet ved 75 år ($P = 0.10$), men...

Effekten af alder som lineær spline

På sandsynligheds-skala



På logit-skala

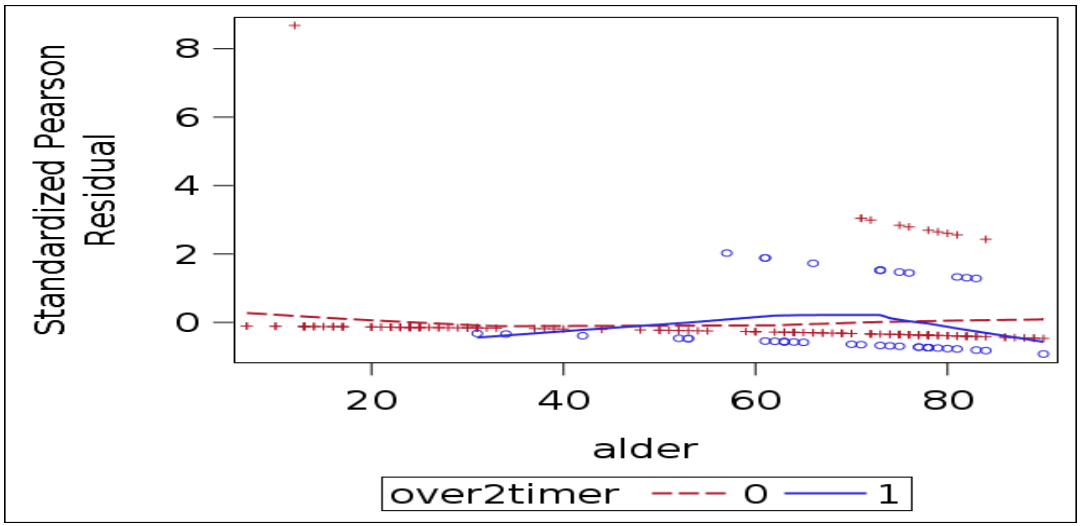


Intet signifikant knæk, men dog en tendens.....

Tror vi på den?

Modelkontrol

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



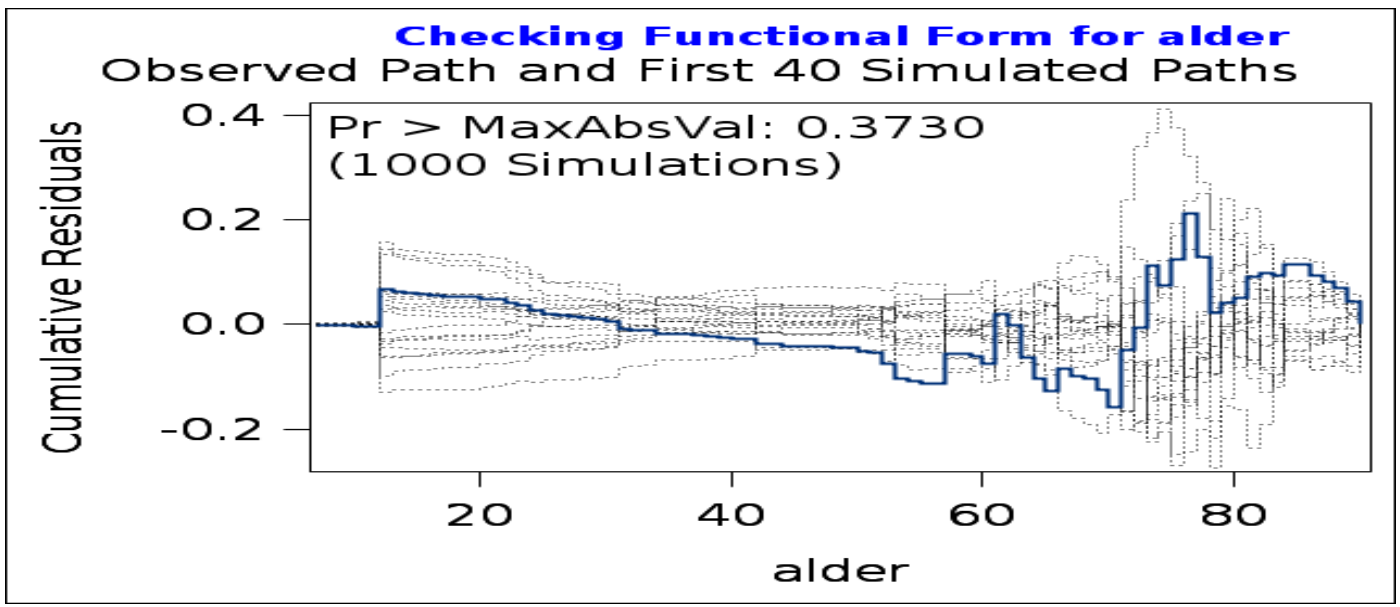
og her er endda standardiseret til

$$e_i = \frac{r_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

Se kode s. 105

Modelkontrol,fortsat

Der kan være en fordel i at se på kumulerede residualer, kumuleret fra lave til høje aldre (se kode s. 106):



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

Modelkontrol, fortsat

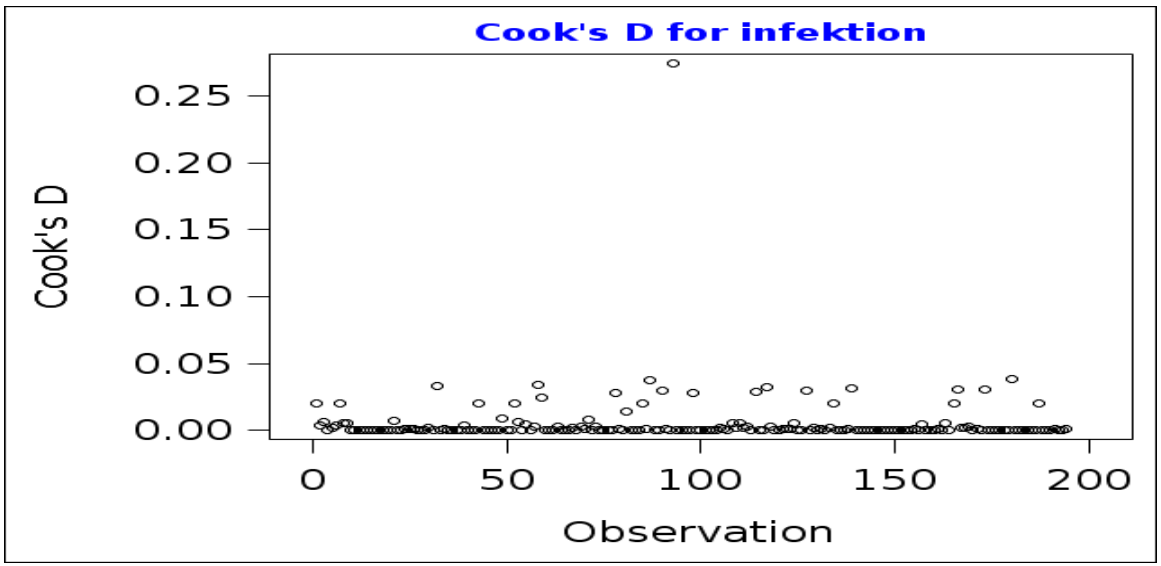
Figuren på forrige side kan vurderes visuelt, men man kan også lave et numerisk check på den maksimale afvigelse fra 0, her baseret på 1000 simulationer.

Vi finder P=0.37 (fremgår af selve figuren), og dermed ingen evidens for en gal modellering af alderseffekten.

Koden ses på s. 106

Diagnostics – Cooks afstand

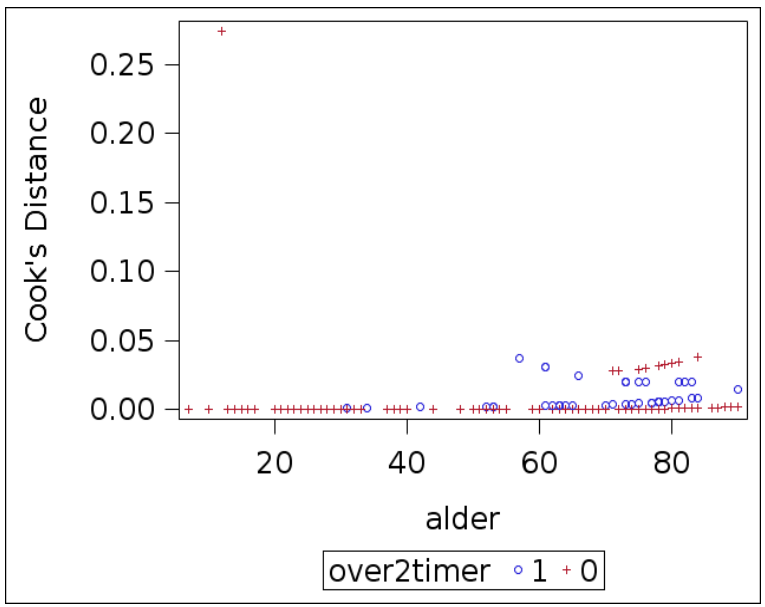
Cooks afstand: (mål for den enkelte observations indflydelse på estimaterne), her plottet op mod observationsnummeret (kode s. 107)



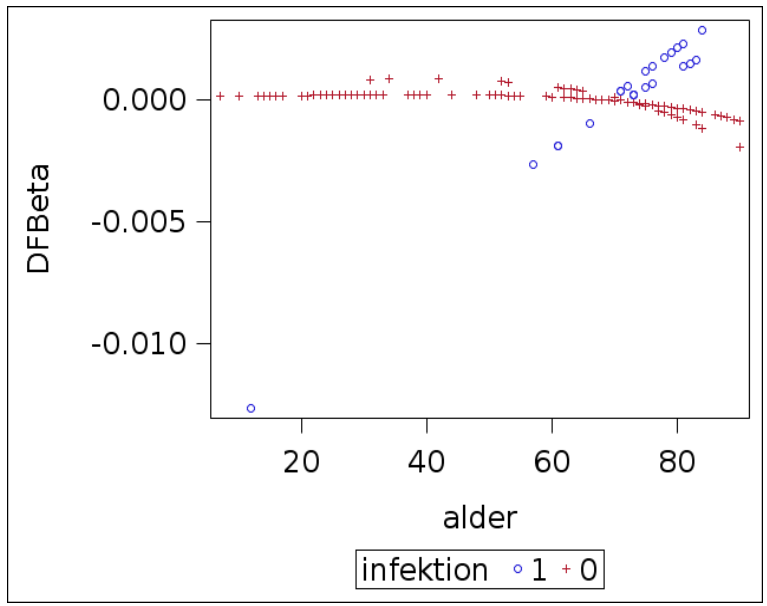
Bemærk den enkelte observation med stor Cook

Diagnostic plots (kode s. 107)

Cooks afstand vs. alder
blå: lang operationstid



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



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

Den suspekte person

Person nr. 93:

Obs	id	infektion	optid	alder	phat	reschi	stdreschi
93	93	1	55	12	0.013245	8.63127	8.67829

Obs	cookd	DFBeta1	DFBeta2	DFBeta3	DFBeta4
93	0.27423	0.27218	0.071224	.	-0.012666

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 estimaterne en del, hvis denne patient pilles ud.

Og hvilken begrundelse kunne der også være for det?

Konklusion

baseret på resultaterne fra s. 24

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.7362	1.0786	19.2823	<.0001
over2timer 1	1	1.3292	0.4729	7.8986	0.0049
over2timer 0	0	0	.	.	.
alder	1	0.0355	0.0149	5.6610	0.0173

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
over2timer 1 vs 0	3.778	1.495	9.546
alder	1.036	1.006	1.067

- 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 konfidensinterva

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

Farveblindhed - igen

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

```
proc genmod data = farveblind;
class gender;
weight antal;
model farveblind = gender / dist=binomial link=log;
estimate "boys vs girls" gender 1 -1 / exp;
contrast "boys vs girls" gender 1 -1;
run;
```

Nederste linie af output på næste side viser, at den relative risiko for farveblindhed for drenge vs. piger er estimeret til 4.8, med konfidensgrænser (0.59, 39.33), altså *ingen signifikant forskel* (P=0.14), og en *meget usikker* bestemmelse af den relative risiko.

Output fra analyse med log-link

```
PROC GENMOD is modeling the probability that farveblind='ja'. One way to
change this to model the probability that farveblind='nej' is to specify
the DESCENDING option in the PROC statement.

          Analysis Of Maximum Likelihood Parameter Estimates

                              Standard     Wald 95%          Wald
Parameter         DF  Estimate   Error  Confidence Limits Chi-Square Pr > ChiSq
Intercept          1   -4.7875  0.9958  -6.7393  -2.8357     23.11     <.0001
gender     dreng   1    1.5686  1.0732  -0.5347   3.6720      2.14     0.1438
gender     pige    0    0.0000  0.0000   0.0000   0.0000         .          .
Scale              0    1.0000  0.0000   1.0000   1.0000
NOTE: The scale parameter was held fixed.

                     Contrast Estimate Results

                       L'Beta         L'Beta
Label                 Estimate    Confidence Limits
boys vs girls          1.5686     -0.5347     3.6720
Exp(boys vs girls)     4.8000      0.5858    39.3291

Contrast             DF   Chi-Square    Pr > ChiSq    Type
boys vs girls         1       3.00        0.0832       LR
```

Se kommentarer til output på forrige side

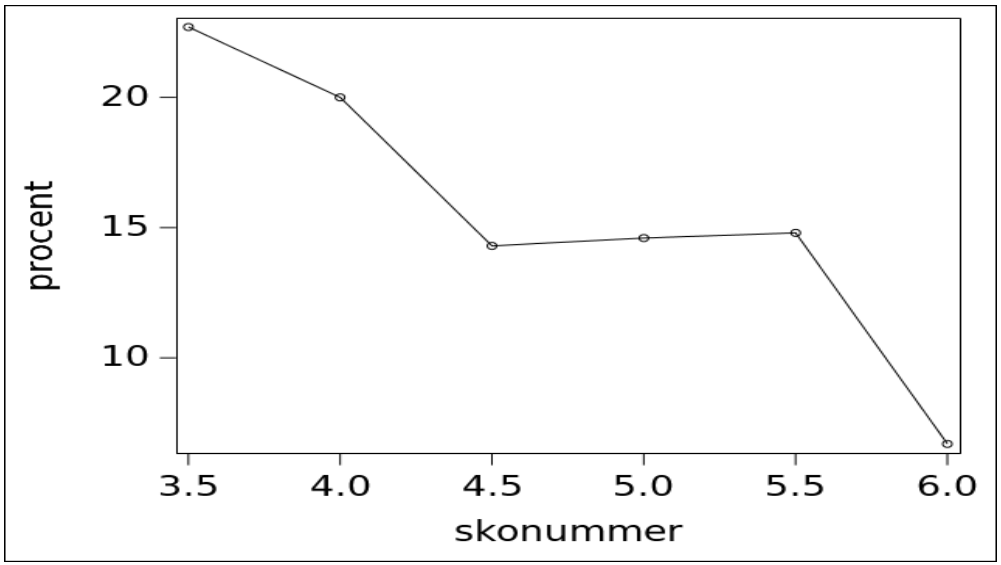
Eksempel om kejsersnit vs. skostørrelse

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

	Skonummer						
Kejsersnit	< 4	4	4.5	5	5.5	≥ 6	Total
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

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....

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.

Data og kode:

```
skonummer total kejsersnit
3.5  22  5
4.0  35  7
4.5  42  6
5.0  48  7
5.5  54  8
6.0 150 10
```

```
proc genmod data = sko;
model kejsersnit/total = skonummer /
    dist=binomial link=identity;
run;
```


Output: Test for trend

altså med linearitet på selve sandsynligheds-skalaen

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Pearson Chi-Square	4	1.3741	0.3435

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4561	0.1296	0.2021	0.7100	12.39	0.0004
skonummer	1	-0.0636	0.0234	-0.1095	-0.0176	7.36	0.0067
Scale	0	1.0000	0.0000	1.0000	1.0000		

Ekstra risiko for kejsersnit, når skoene er et nummer mindre: 0.0636, CI=(0.0176, 0.1095), altså mellem 1.8 og 11.0%point

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 (`skostr=skonummer;`) ind som faktor oveni, og så teste, om denne kan undværes:

```
proc genmod descending data = sko;
class skostr;
model kejsersnit/total = skonummer skostr /
    dist=binomial link=identity type3;
run;
```

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

Output fra check af linearitet

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
skonummer	0	0.00	.
skostr	4	1.29	0.8624

Bemærk:

- Man kan *ikke* teste `skonummer` væk fra modellen, da dette ikke ændrer modellen
- Man kan *godt* teste `skostr` væk fra modellen, da man derved reducerer til en lineær effekt af `skonummer`

Lineariteten i skonummer ser rimelig ud (P=0.86)

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

  og måske kan vi også nå (s. 78ff)
- Tælletal (uden øvre grænse), f.eks. antal metastaser

  Poisson regression eller *log-lineære modeller*

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?

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
fibrosegrad	129	1.4263566	0.9903850	0	3.0000000
ykl40	129	533.5116279	602.2934049	50.0000000	4850.00
piiinp	127	13.4149606	12.4887192	1.7000000	70.0000000
ha	128	318.4531250	658.9499624	21.0000000	4730.00

(Julia Johansen, KKHH)

Ordinale data

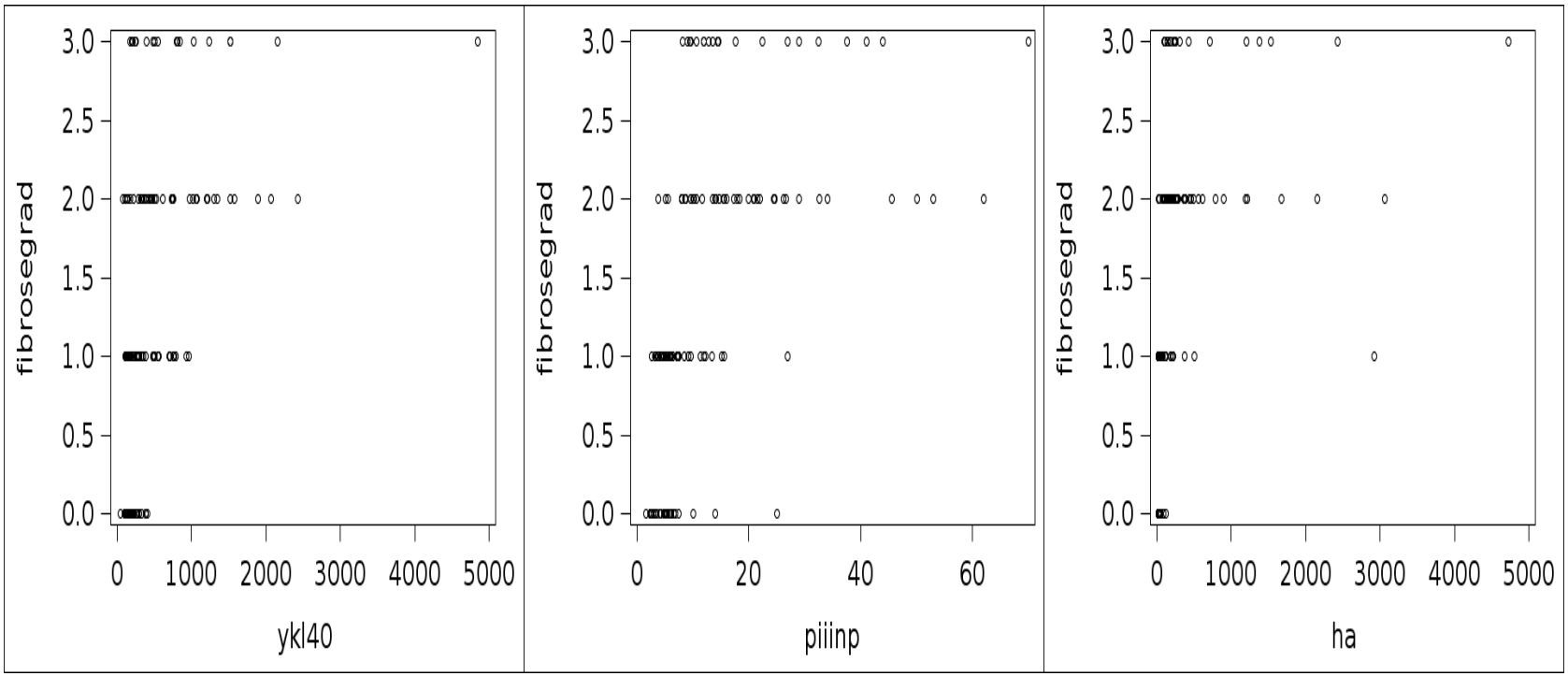
- data på en rangskala
- afstand mellem responskategorier kendes ikke, eller er udefineret
- evt. en underliggende kvantitativ skala

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

Fordeling af blodmarkører

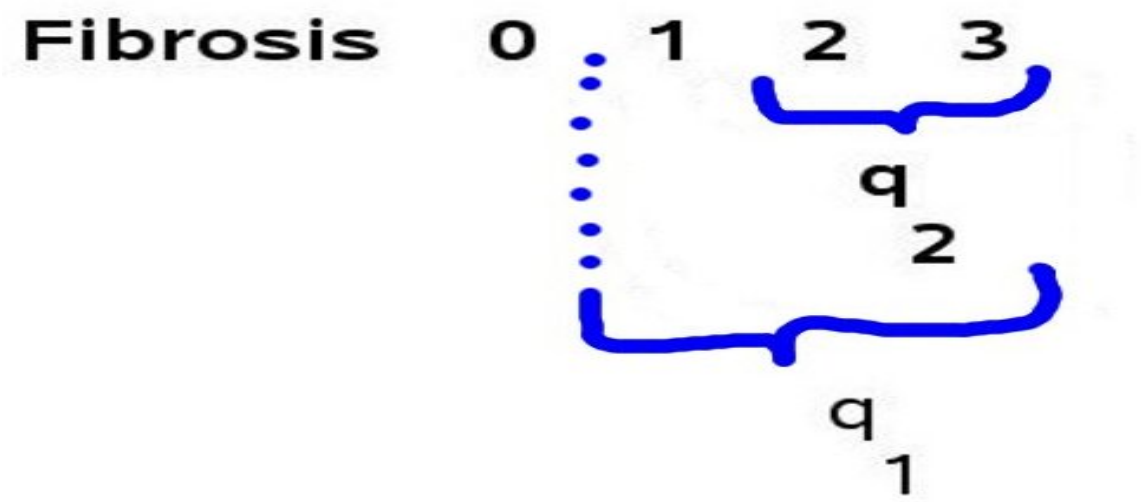
vist for hver fibrose kategori



Er der noget problem ved de meget skæve fordelinger?

Kumulerede sandsynligheder

Sandsynlighed for *dette eller værre*



Tilbageregning til sandsynligheder for de enkelte fibrosegrader:

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

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

Proportional odds model

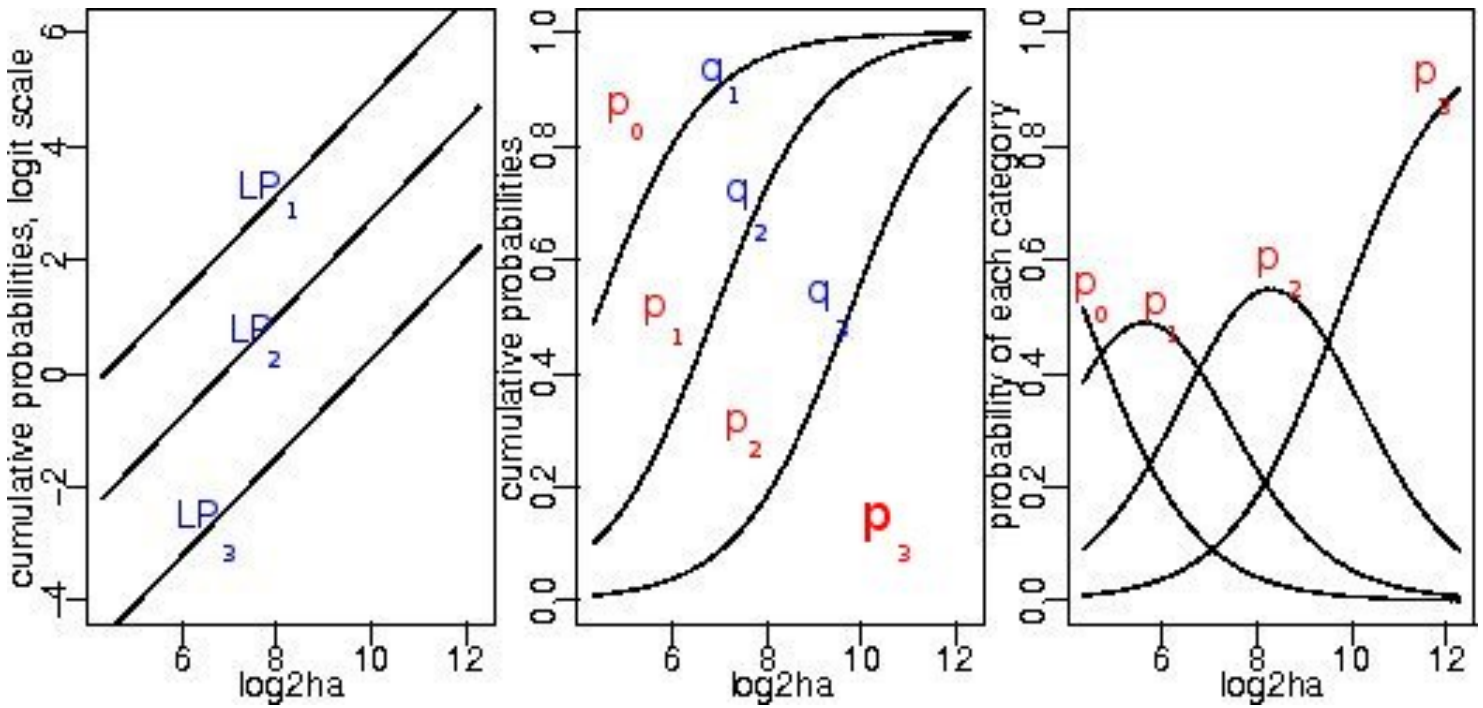
Logistisk regression for hver tærskel,
med samme afhængighed af kovariaterne, dvs.

$$\begin{aligned}
\text{logit}(q_3) &= \beta_{03} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
\text{logit}(q_2) &= \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
\text{logit}(q_1) &= \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3
\end{aligned}$$

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

Illustration af predikterede sandsynligheder

i tilfælde af en enkelt kovariat



Model med alle tre kovariater

alle logaritmetransformeret ($\log_2$):

```
proc logistic data=fibrosis descending;
  model degree_fibr=lha lpiiinp lykl40;
run;
```

som giver outputtet

```
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

Output, fortsat

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept 3	1	-12.7767	1.6959	56.7592	<.0001	
Intercept 2	1	-10.0117	1.5171	43.5506	<.0001	
Intercept 1	1	-7.5922	1.3748	30.4975	<.0001	
lha	1	0.3889	0.1600	5.9055	0.0151	0.4174
lpiiinp	1	0.8225	0.2524	10.6158	0.0011	0.5231
lykl40	1	0.5430	0.1700	10.2031	0.0014	0.3750

Odds Ratio Estimates

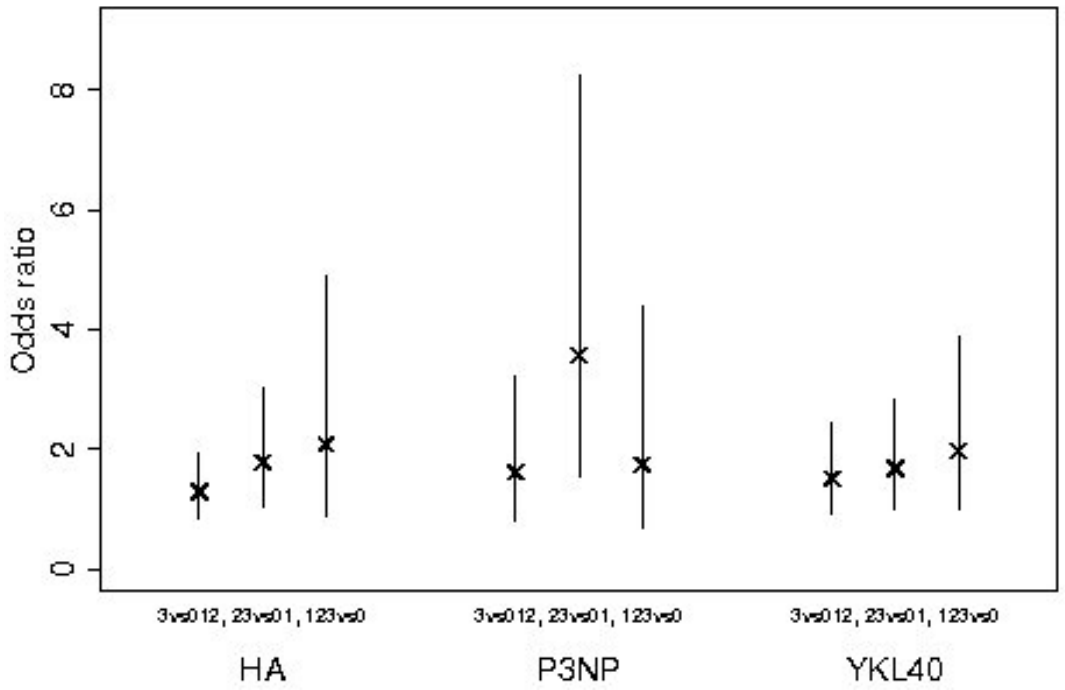
Effect	Point Estimate	95% Wald Confidence Limits	
lha	1.475	1.078	2.019
lpiiinp	2.276	1.388	3.733
lykl40	1.721	1.233	2.402

Eksempel på fortolkning:

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

Afvigelse fra proportional odds antagelsen?

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

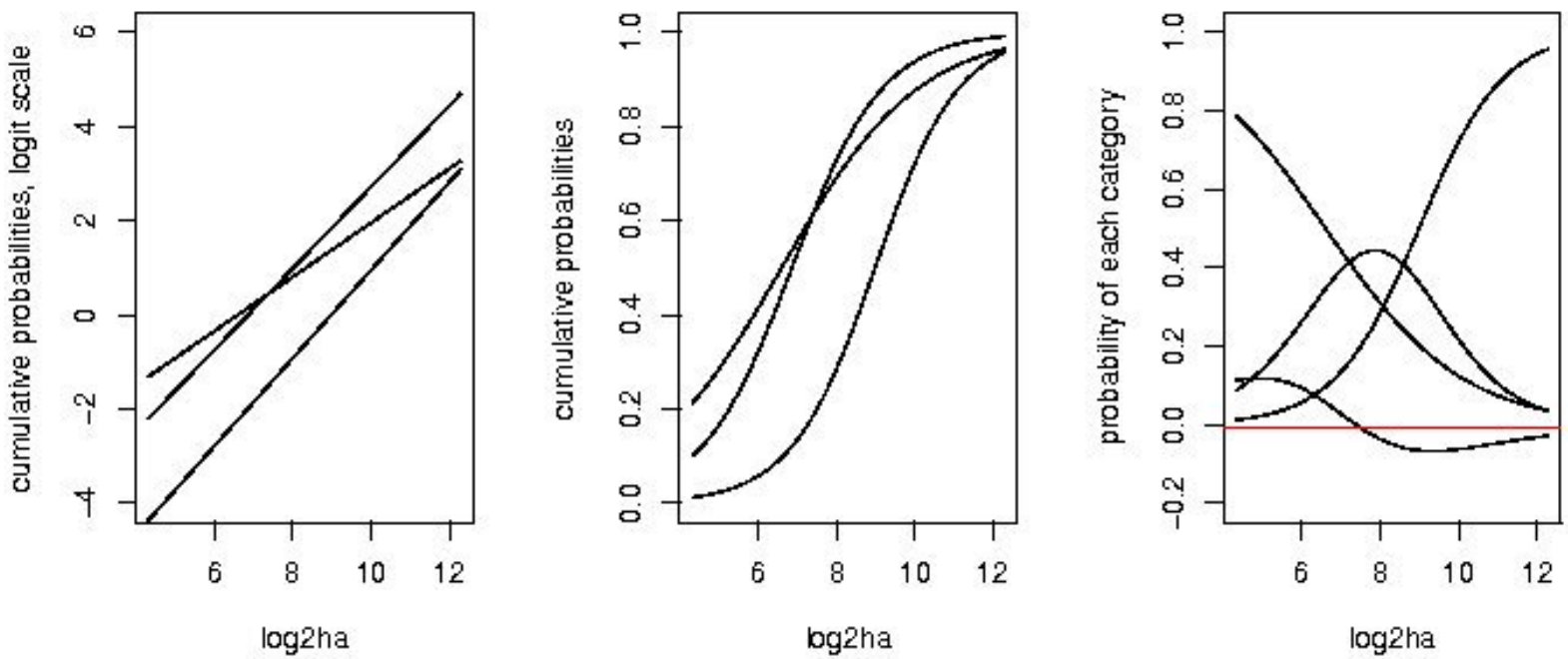


Måske er der lidt forskel på estimaterne for piiinp...

men det var altså ikke signifikant (P=0.14, se s. 66)

Hvis der ikke er proportional odds

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

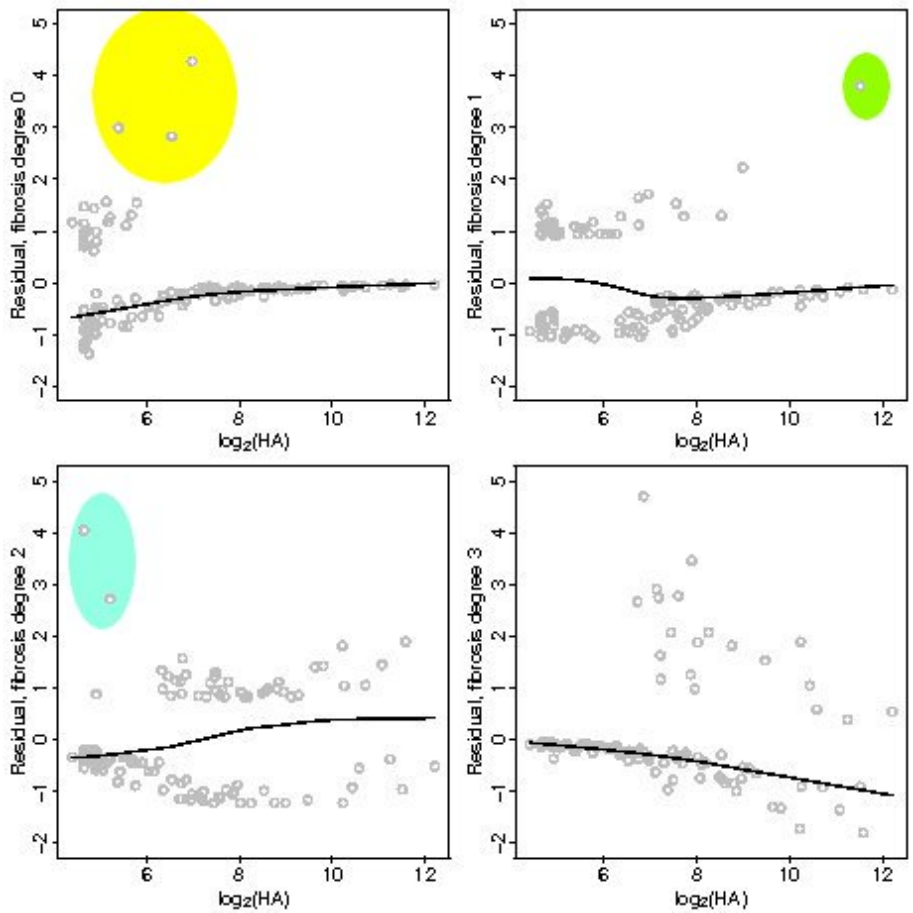


Modelkontrol

Der er flere ting, der skal checkes:

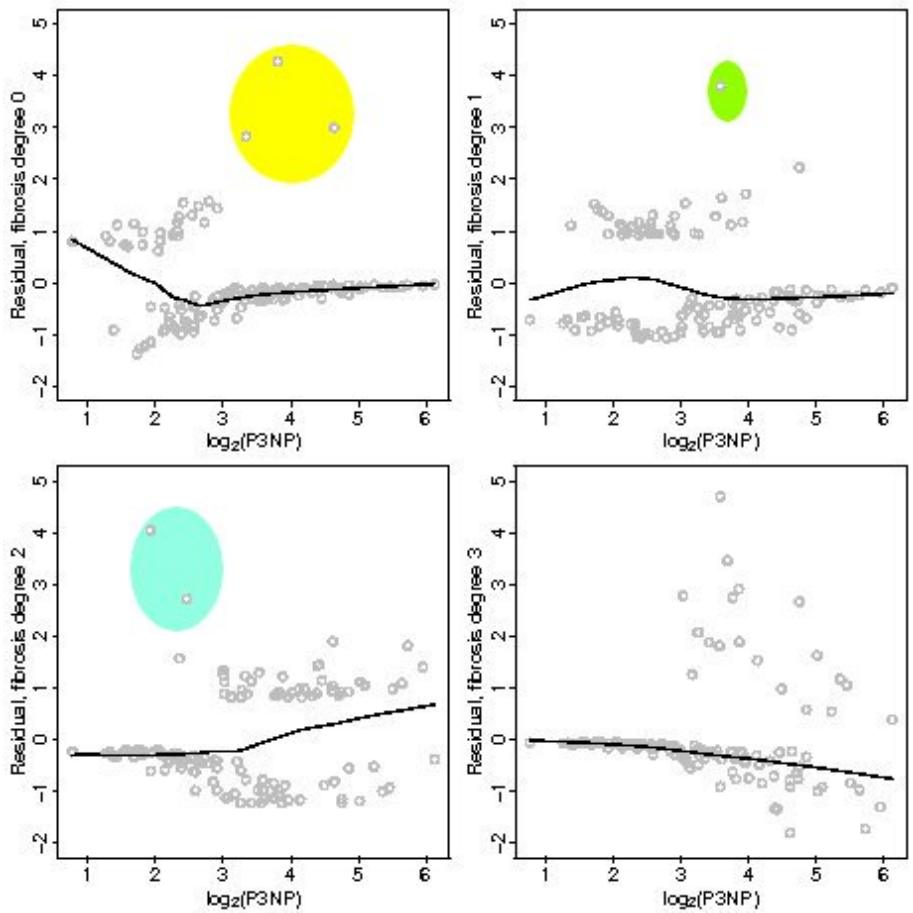
- lineariteten af kovariat-effekterne,
  på logit-skalaen, *ligesom for logistisk regression*,
  men der er mange residualplots:
  Et for hver kombination af fibrosegrad og kovariat,
  dvs. 12 i alt.
- proportional odds antagelsen
- modellens evne til faktisk at prediktere en fornuftig fibrosegrad

Residualplot for ha



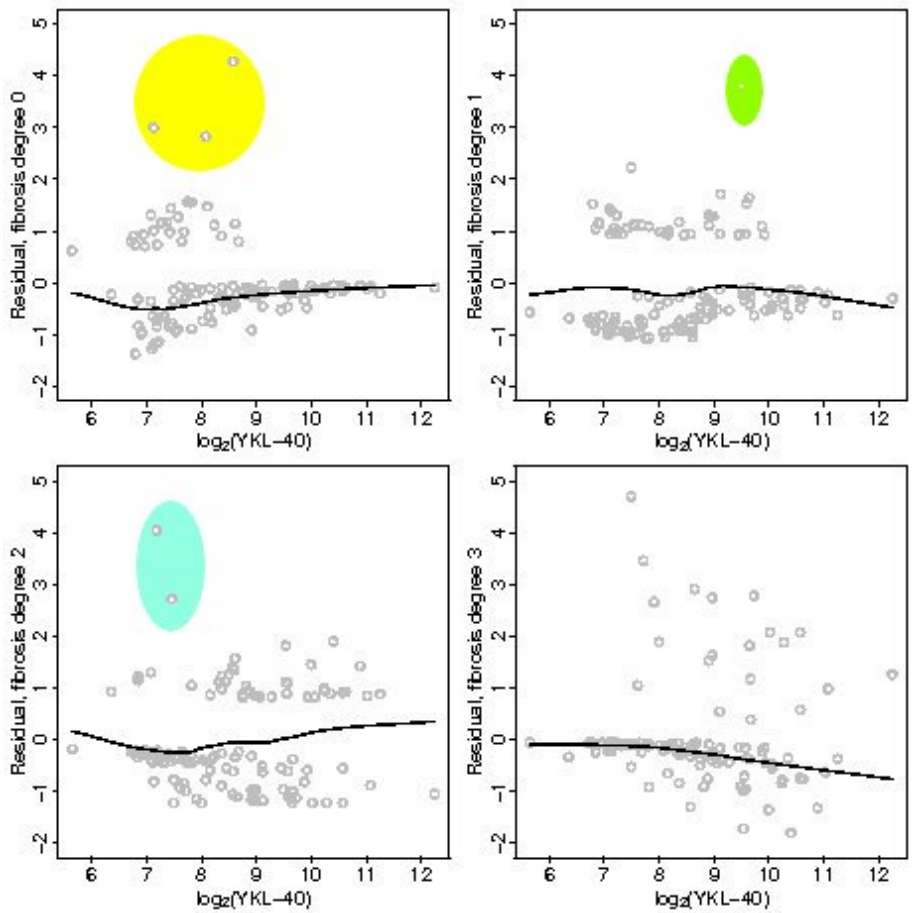
Svag afvigelse fra linearitet

Residualplot for p3np



Nogen afvigelse fra linearitet

Residualplot for ykl40



Rimelig linearitet

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

Predikterede sandsynligheder

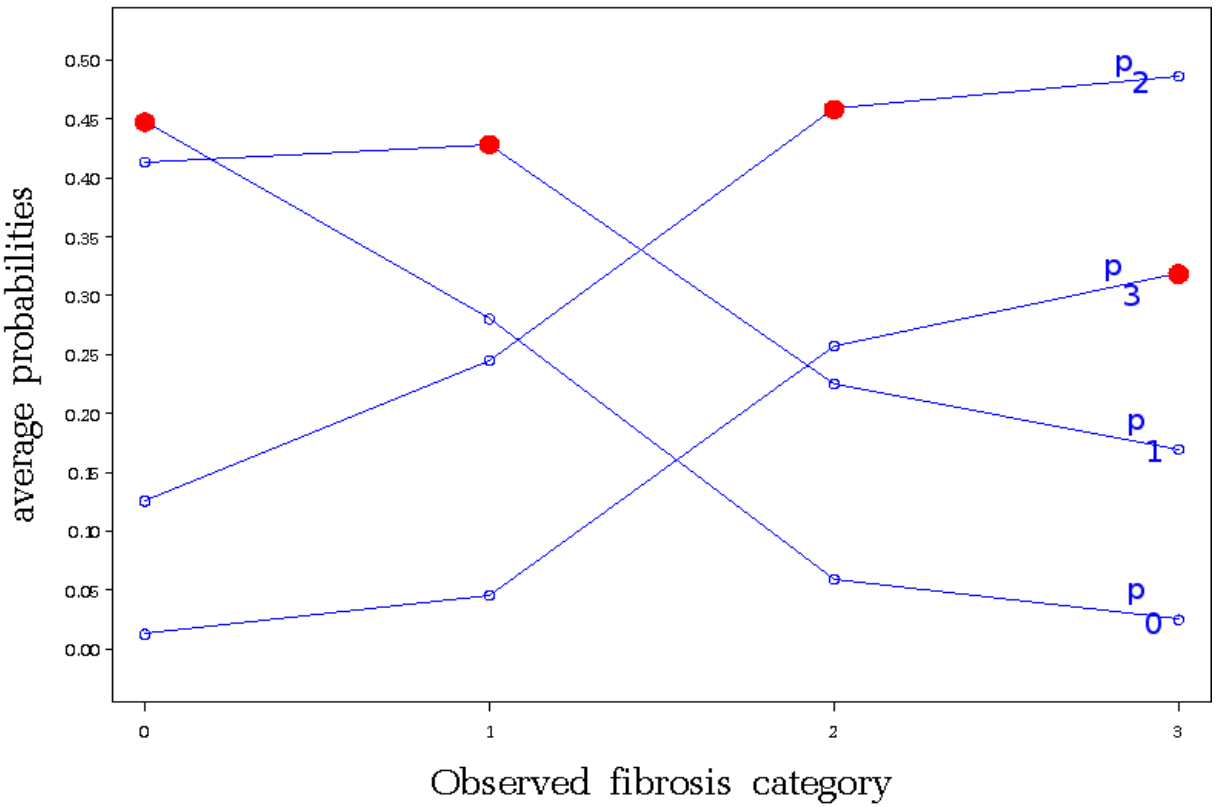
For hver person kan værdierne af de 3 blodmarkører benyttes til at prediktere sandsynligheder for hver af de 4 fibrosegrader.

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. 76 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.

Predikterede sandsynligheder, II



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

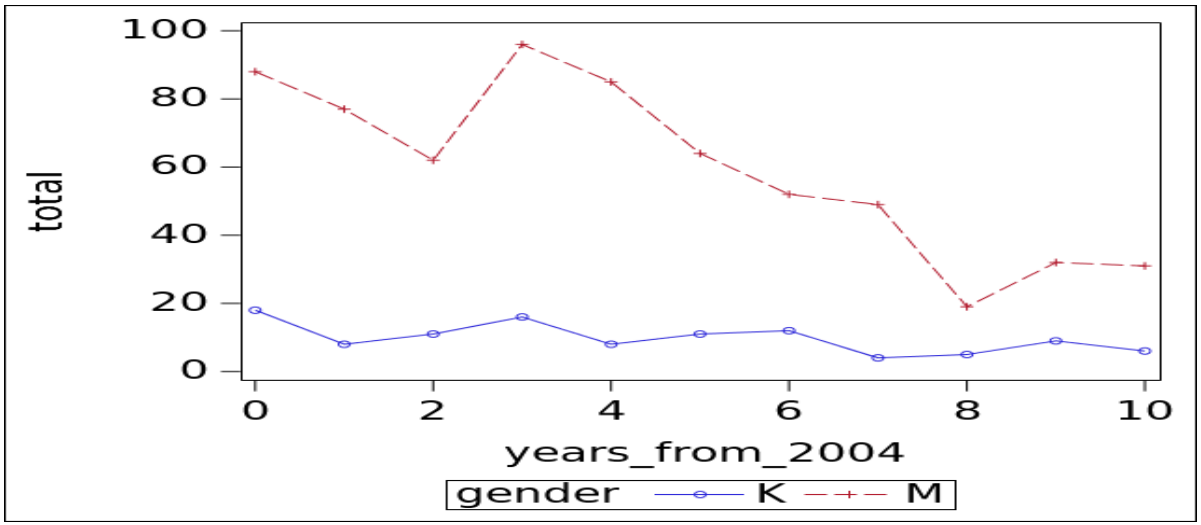
Observeret vs. predikteret fibrosegrad

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

Tælletal

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 linier med 3 variable.

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

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

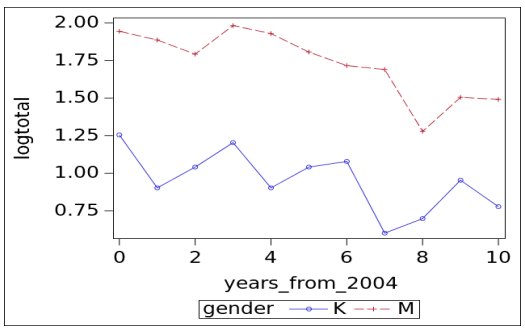
men vi kender ikke N_{gt}, vi ved bare, at den er stor, og at p_{gt} er lille. I sådan et tilfælde approksimeres Binomialfordelingen af en *Poisson*-fordeling:

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

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

Regression i Poisson-fordelingen

Da middelværdien af tælletal 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?

Regression i Poisson-fordelingen, II

Vi anvender nu `PROC GENMOD`:

```
proc genmod data=trafik;
class gender;
model total=gender years_from_2004 gender*years_from_2004 /
dist=poisson link=log type3;
estimate 'reduktion pr. aar, M' years_from_2004 1
          gender*years_from_2004 0 1 / exp;
estimate 'reduktion pr. aar, F' years_from_2004 1
          gender*years_from_2004 1 0 / exp;
run;
```

På s. 113 er koden udbygget, så man kun får udskrevet relevante dele af output fra `estimate`-sætningerne.

Output fra Poisson-regression

med to separate lineære effekter af tid (mænd og kvinder), dvs. en interaktion (se også kode s. 113)

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
gender	1	164.73	<.0001
years_from_2004	1	36.73	<.0001
years_from_20*gender	1	0.52	0.4688

Contrast Estimate Results

Obs	Label	LBeta Estimate	LBeta LowerCL	LBeta UpperCL
2	Exp(reduktion pr. aar, M)	0.8932	0.8711	0.9160
4	Exp(reduktion pr. aar, F)	0.9154	0.8612	0.9730

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).

Output fra Poisson-regression, II

Vi udelader den insignifikante interaktion:

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
gender	1	435.50	<.0001
years_from_2004	1	88.09	<.0001

Obs	Label	LBeta Estimate	LBeta LowerCL	LBeta UpperCL
2	Exp(reduktion pr. aar, begge)	0.8964	0.8758	0.9175
4	Exp(gender, M vs. F)	6.0648	4.9478	7.4339

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

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. 79)

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 (`scale=p`)

```
proc genmod data=trafik;
class gender;
model total=gender years_from_2004 /
dist=poisson link=log scale=p type3;
run;
```

Output ved korrektion for overspredning

Først selve estimatet for overspredningen (1.5830)

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Pearson Chi-Square	19	47.6114	2.5059

Scale	0	1.5830	0.0000	1.5830	1.5830

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

Når der korrigeres for overspredningen, får man lidt bredere konfidensintervaller (jvf s. 83):

Contrast Estimate Results

Obs	Label	LBeta Estimate	LBeta LowerCL	LBeta UpperCL
2	Exp(reduktion pr. aar, begge)	0.8964	0.8640	0.9300
4	Exp(gender, M vs. F)	6.0648	4.3942	8.3706

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 (kode s. 115)

Obs	Parameter	back_ estimate	back_ lower	back_ upper	Probt
1	reduktion pr. aar, begge	0.89719	0.85501	0.94145	0.0002
2	gender, M vs. F	6.00260	4.42669	8.13954	<.0001

hvilket ses kun at være en anelse anderledes end det, vi fandt ved Poisson-analysen s. 85, (inkluderende overspredning).

Konklusion vedr. trafikuheld

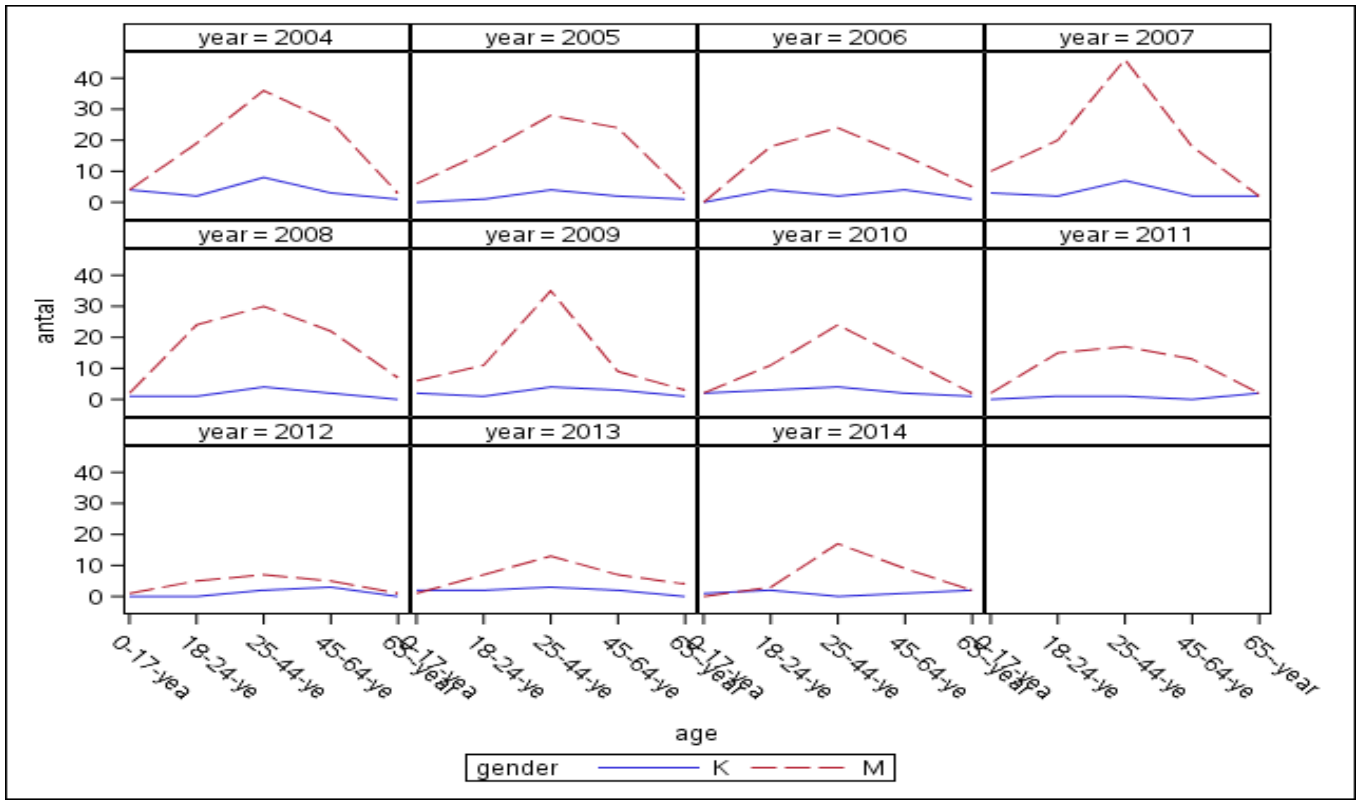
- Der er flere mænd end kvinder, der bliver dræbt i trafikken, ca. 6 gange så mange
  - De færdes måske mere?
  - De er måske mere uforsigtige?
  - Alder er måske en skjult confounder?
    Det kunne jo være, at der var mange flere mandlige bilister i de yngre aldersklasser... men dette er ikke registreret

  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

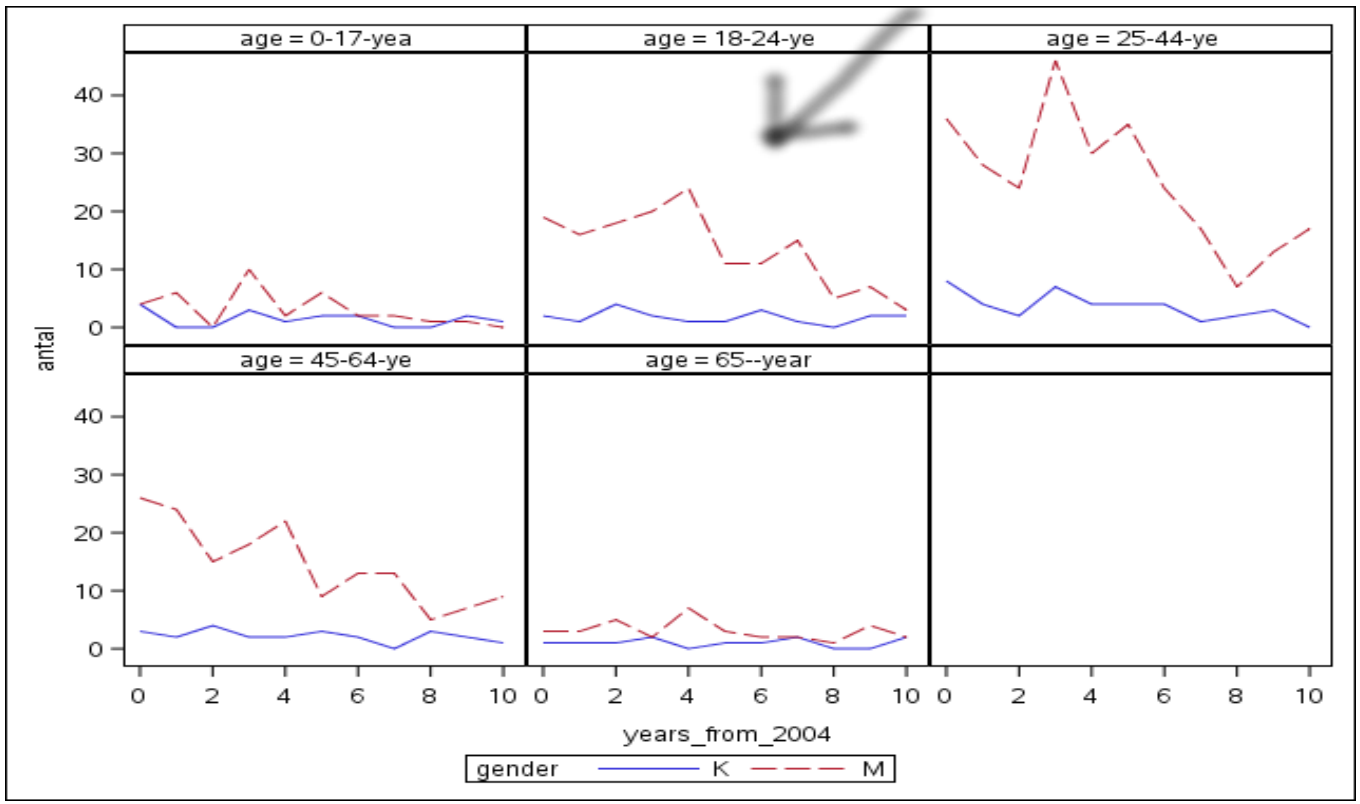
Trafikuheld, opdelt på årstal

som funktion af alder (kode s. 116)



Trafikuheld, opdelt efter alder

som funktion af år siden 2004 (kode s. 116)



Trafikuheld, kun for 18-24-årige

med korrektion for overspredning på 1.1262
(`scale=p`, se kode s. 117)

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
gender	1	18	38.93	<.0001	38.93	<.0001
years_from_2004	1	18	3.58	0.0749	3.58	0.0586
years_from_20*gender	1	18	0.84	0.3719	0.84	0.3598

Contrast Estimate Results

Obs	Label	LBeta Estimate	LBeta LowerCL	LBeta UpperCL
2	Exp(reduktion pr. aar, M)	0.8842	0.8328	0.9387
4	Exp(reduktion pr. aar, F)	0.9586	0.8161	1.1261

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

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 normalfordelingsapproksimation ikke ville være særlig god alligevel....

APPENDIX

Programbidder svarende til diverse slides:

- To-gange-to tabeller, s. 93
- Plot af binære data med `loess`-udglatning, s. 94
- Logistisk regression, s. 95-96
- GENMOD vs. LOGISTIC, s. 97-102
- Modelkontrol, s. 103-106
- Diagnostics, s. 107
- Alternative links, s. 108-111
- Ordinale data, s. 112
- Tælletal, s. 113-117

To-gange-to tabeller

Slide 7-10

```
proc freq data=farveblind;
   tables gender*farveblind /
   chisq nopercent nocol expected riskdiffc relrisk;
weight antal;
run;
```

hvor datasættet `farveblind` indeholder 4 linier:

```
gender farveblind antal
pige nej 119
pige ja 1
dreng nej 144
dreng ja 6
```

Figurer

Slide 13

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

```
proc sgplot data=infektion;
    loess Y=infektion X=alder /
    group=over2timer smooth=0.8;
run;
```

PROC LOGISTIC i SAS

Slide 23-24

Den skrabede kode i LOGISTIC til den beskrevne logistiske regression:

```
proc logistic data=infektion;
  class over2timer / param=glm;
  model infektion(event="1") = over2timer alder;
run ;
```

PROC LOGISTIC i SAS

Slide 23-24

Udbygget kode
med estimate-sætninger
og konstruktion af output-datasæt:

```
ods graphics on;
proc logistic plots=ALL data=infektion;
  class over2timer(ref="0") / param=glm;
  model infektion(event="1") = over2timer alder
        / clparm=wald lackfit aggregate /*scale=p*/ link = logit ;
  estimate "10 years" alder 10 / exp;
  estimate "more than two hours" over2timer 1 -1 / exp;
  estimate "10 years and more than two hours" over2timer 1 -1 alder 10 / exp;
output out=pred lower=lower pred=phat resdev=resdev reschi=reschi upper=upper;
run ;
ods graphics off;
```

PROC GENMOD i SAS

Den skrabede kode i GENMOD til den beskrevne logistiske regression:

```
proc genmod descending data=infektion;
  class over2timer;
  model infektion = over2timer alder
        / dist = binomial link = logit ;
run ;
```

- descending-option sikrer, at man modellerer sandsynligheden for et 1-tal (ikke et 0)
- class fungerer som i GLM

PROC GENMOD i SAS

Udbygget kode med estimate-sætninger og konstruktion af output-datasæt:

```
proc genmod descending data=infektion;
  class over2timer;
  model infektion = over2timer alder
        / dist = binomial link = logit ;
  estimate "10 years" alder 10 / exp;
  estimate "more than two hours" over2timer -1 1 / exp;
  estimate "10 years and more than two hours"
                 over2timer -1 1 alder 10 / exp;
  output out=pred lower=lower pred=phat upper=upper
                 reschi=reschi stdreschi=stdreschi
                 cooksd=cooksd dfbeta=_all_;
run ;
```

Kommentarer til GENMOD-koden

Model-options

- dist angiver fordelingen, her en Binomialfordeling
  (med antalsparameter 1, dvs. en Bernoulli fordeling)
  Denne er vigtig at have med, ellers benyttes en Normalfordelingsantagelse.
- link angiver link-funktionen g
  (logit er default, så det behøver man faktisk ikke at skrive)
- estimate-sætninger, med exp-option
  fordi vi så, at exponentialfunktionen til parametrene fortolkes som odds ratio'er

Noter til GENMOD-koden

- Outcome infektion er en 0/1-variabel, altså en Binomialfordelt størrelse med $n = 1$.
  Dette kan skrives eksplicit ved at definere en variabel `"antal=1;"` og så benytte konstruktionen

```
model infektion/antal = over2timer alder
      / dist = binomial link = logit ;
```

- Hvis der er tale om grupperede data, vil variablen `antal` angive, hvor mange personer, der har et bestemt mønster af kovariater, medens `infektion` så angiver, hvor mange af disse, der fik infektion

Nu er option `dist=binomial` strengt taget ikke nødvendig mere...

LOGISTIC eller GENMOD?

Fordele ved LOGISTIC:

- giver automatisk odds ratioer (uden `estimate`-sætninger)
- har flere automatiske plots (ikke en ubetinget fordel), men har dog plottet s. 26 (venstre del) som default
- har en ekstra modelkontrol, Hosmer-Lemeshow testet (option `lackfit`, se s. 37-38)
- har lettere variabelnavne for `DFBETA`, nemlig

```
DFBETA_Intercept DFBETA_over2timer0
DFBETA_over2timer1  DFBETA_alder
```

- har et test for antagelsen om proportional odds ved modellering af ordinale data

LOGISTIC eller GENMOD?

Ulemper ved LOGISTIC:

- pas på med selve parameterestimaterne (brug `PARAM=GLM`)
- kan ikke håndtere så mange forskellige outcome typer som `GENMOD`
- har kun få mulige link-funktioner
  (Vi har her mest brugt den traditionelle *logit*,
  men benytter dog et *log*-link og *identity*-link sidst i forelæsningen)
- mangler `ASSESS`-sætningen til modelkontrol (s. 44, 45, 106)

Goodness of fit

Slide 36-38

udføres i LOGISTIC med option `lackfit`:

```
proc logistic data=infektion;
  class over2timer / param=GLM;
  model infektion(event="1") = over2timer alder
        / lackfit link=logit ;
run ;
```

Check af linearitetsantagelsen for alder

Slide 39-42

- Alderseffekt modelleret med både alder og log(alder) (output s. 39-40). Definer:
  `logalder50=log(alder)/log(1.1)-log(50)/log(1.1);`
  og brug modellen

```
model infektion = over2timer alder_minus50 logalder50
      / dist = binomial link = logit;
```

- Alderseffekt modelleret som lineær spline (output s. 41-42):
  Definer `alder_over60=(alder>60)*(alder-60)` og
  `alder_over75=(alder>75)*(alder-75)`
  og brug modellen

```
model infektion/antal = over2timer alder
        alder_over60 alder_over75 / dist = binomial link = logit;
```

Plot af standardiserede residualer

Slide 43

Modellen er den på s. 23, hvor vi har output-sætningen:

```
output out=pred lower=lower pred=phat upper=upper
  reschi=reschi stdreschi=stdreschi cooksd=cooksd dfbeta=_all_;
```

Ud fra datasættet pred kan man nu selv styre plottene, f.eks. plottet s. 43:

```
proc sgplot data=pred_linear;
    loess Y=stdreschi X=alder / group=over2timer smooth=0.7;
run;
```

Plot af kumulerede residualer

Slide 44

Figuren med simulationer af de kumulerede residualer er dannet ved at benytte linien

```
assess var=(alder) / resample npaths=40 seed=106165;
```

i PROC GENMOD konstruktionen

Værdien for seed er tilfældigt valgt, jeg bruger sædvanligvis denne, fordi jeg kan huske den

```
proc genmod descending data=infektion;
  class over2timer;
  model infektion = over2timer alder
        / dist = binomial link = logit;
  assess var=(alder) / resample npaths=40 seed=106165;
run ;
```

Diagnostics

Slide 46-47

Modellen er den på s. 23 (øverste del), hvor vi har tilføjet `output`-sætningen:

```
output out=pred lower=lower pred=phat upper=upper
  reschi=reschi stdreschi=stdreschi cooksd=cooksd dfbeta=_all_;
```

Ud fra datasættet `pred` kan man nu selv styre plottene, f.eks.

```
proc sgplot data=pred;
    scatter Y=cooksd X=alder / group=over2timer;
run;
proc sgplot data=pred;
    scatter Y=DFBeta4 X=alder / group=infektion;
run;
```

Bemærk: Variablen `DFBeta4` betegner det 4. parameterestimat i rækken, dvs. koefficienten til `alder` (først er der intercept, samt 2 parametre svarende til grupperingen `over2timer`).

Relativ risiko for farveblindhed

Slide 51-52

Risk Ratio (relativ risiko) estimeres i GENMOD med log-link.
Bemærk brugen af `weight` option:

Data:

```
gender farveblind antal
pige nej 119
pige ja 1
dreng nej 144
dreng ja 6
```

```
proc genmod data = farveblind;
class gender;
weight antal;
model farveblind = gender /
    dist=binomial link=log;
estimate "boys vs girls"
         gender 1 -1 / exp;
run;
```

Bemærk den manglende `descending`-option, fordi `farveblind=''ja''` er først i alfabetet

Alternativ datastruktur for farveblindhed

```
gender total farveblinde
pige 120 1
dreng 150 6
```

Så bruges i stedet programbidden

```
proc genmod data = farveblind;
class gender;
model farveblinde/total = gender /
    dist=binomial link=log;
estimate "boys vs girls" gender 1 -1 / exp;
run;
```

Test for trend

Slide 55-56

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

kan udføres som en regression i Binomial-fordelingen,
med `identity`-link i GENMOD,
her med med mulig overspredning (option `pscale`)

```
proc genmod data = sko;
model kejsersnit/total = skonummer /
    dist=binomial link=identity aggregate pscale;
run;
```

Bemærk skrivemåden i Model-sætningen: `kejsersnit/total`

Modelkontrol af lineariteten, kejsersnit

Slide 57-58

Check af linearitet kan foretages ved at sætte en kopi (`skostr=skonummer`) ind som ekstra faktor i modellen:

```
proc genmod descending data = sko;
class skostr;
model kejsersnit/total = skonummer skostr /
    dist=binomial link=identity type3;
run;
```

Proportional odds model

Slide 66-67

Logistisk regression for hver tærskel,
med samme afhængighed af kovariaterne (link=cumlogit),
som her alle er $\log_2$-transformerede:

```
proc logistic data=fibrosis descending;
  model degree_fibr=lha lpiiinp lykl40;
run;
```

Odds ratio vil ikke afhænge af cutpoint,
og denne hypotese testes automatisk
(*proportional odds assumption*)

Poisson analyse

Slide 81-82

```
proc genmod data=trafik;
class gender;
model total=gender years_from_2004 gender*years_from_2004 /
dist=poisson link=log type3;
estimate 'reduktion pr. aar, M' years_from_2004 1
          gender*years_from_2004 0 1 / exp;
estimate 'reduktion pr. aar, F' years_from_2004 1
          gender*years_from_2004 1 0 / exp;
ods output Estimates=estimater;
run;

proc print data=estimater; where substr(Label,1,3)="Exp";
var Label LBetaEstimate LBetaLowerCL LBetaUpperCL;
run;
```

Poisson analyse

Slide 84-85

Her bruges `scale=p` for at korrigere for overspredning, ellers helt som s. 113

```
proc genmod data=trafik;
class gender;
model total=gender years_from_2004 gender*years_from_2004 /
dist=poisson link=log scale=p type3;
run;
```

Normalfordelingsmodel for tælletal

bør foretages på logaritmetransformerede data: `logtotal=log10(total)`:

Slide 86

```
proc glm data=ny;
class gender;
model logtotal=gender years_from_2004 /
               solution clparm;
estimate 'reduktion pr. aar, begge' years_from_2004 1;
estimate 'gender, M vs. F' gender -1 1;
ods output Estimates=estimater;
run;

data ud;
set estimater;

back_estimate=10**Estimate;
back_lower=10**LowerCL;
back_upper=10**UpperCL;
run;

proc print data=ud;
var Parameter back_estimate back_lower back_upper Probt;
run;
```

Opdelte plot (panels)

Slide 88-89

```
proc sort data=a1; by gender;
run;

proc sgpanel data=a1;
panelby gender;
series Y=antal X=years_from_2004 / group=age;
run;

proc sort data=a1; by year;
run;

proc sgpanel data=a1;
panelby year / rows=3 columns=4;
series Y=antal X=age / group=gender;
run;
```

Poisson analyse, kun for 18-24 årige

Slide 90

```
proc genmod data=a1; where age="18-24-ye";
class gender age;
model antal=gender years_from_2004 gender*years_from_2004 /
dist=poisson link=log scale=p type3;
estimate 'reduktion pr. aar, M' years_from_2004 1
          gender*years_from_2004 0 1 / exp;
estimate 'reduktion pr. aar, F' years_from_2004 1
          gender*years_from_2004 1 0 / exp;
ods output Estimates=estimater;
run;

proc print data=estimater; where substr(Label,1,3)="Exp";
var Label LBetaEstimate LBetaLowerCL LBetaUpperCL;
run;
```