Logistisk regression

Susanne Rosthøj

26. oktober 2020

Outline

Outcome er *binært* (0/1, ja/nej eller case/kontrol).

- Tabeller
- Risk, odds og odds-ratio
- Simpel logistisk regression:
  - Binær
  - Kvantitativ
  - En binær og en kvantitativ
    - Interaktion
  - Flere kovariater
- Modelkontrol
  - Diagnostics
- Prædiktion

Typer af outcome

- Kvantitativ - *den generelle lineære model*
- **Binær** - *logistisk regression*
- *Ordinal* - *proportional odds regression*
- Tælletal - *Poisson regression*
- Levetid - *Cox regression*

Eksempler på binære outcomes

- Farveblindhed
- Komplikationer ved operation
- Udskrivning af astmapatienter efter første undersøgelse
- Astma
- Vitamin D-deficiens

Prostata kræft

380 patienter, variable målt ved baseline eksamination.

Outcome:

spredning	Tumor er trængt igennem prostatakapslen (0/1, 1="spredning")

Hvordan afhænger risikoen for spredning af

Forklarende variable:

psa	Prostatic Specific Antigen Value mg/ml
involvering	Kapsel involvering ved rektal eksploration (0/1, 1="involvering")
knude	Knudes placering på lap ("ingen","venstre","hoejre","begge")
alder65	Under / over 65 år ("Under"/"Over")
gleason	Gleason score, 0-10.

Kan vi ud fra disse variable prædiktere spredning?

Involvering og spredning

		Spredning				
		0		1		total
Involvering	0: nej	217	(64.0)	122	(36.0)	339
	1: ja	10	(24.4)	31	(75.6)	41
	total	227		153		380

Er risikoen for spredning den samme uanset involvering eller ej?

p_{uden} Risiko for patienter *uden* involvering

p_{med} Risiko for patienter *med* involvering

Hypotese: $p_{uden} = p_{med}$

Test for uafhængighed

Hypotese: $p_{uden} = p_{med}$ testes med

- Chi-i-anden test (χ^2-test)
  - når tabellen ikke er 'for tynd' (forventede værdier ≥ 5)
- Fishers eksakte test
  - som altid kan bruges

Disse er *test for uafhængighed* mellem kovariat (involvering) og outcome (spredning)

Her er p$<$.0001 ($\chi^2 = 23.9$, df=1) og vi konkluderer at risikoen for spredning er *forskellig* for patienter med og uden involvering.

Kvantificering af effekten I

Risikodifferens:

$$p_{med} - p_{uden} = 75.6 - 36.0 = 39.6$$

*Der er 39.6 procent**point** flere patienter med involvering der får spredning end patienter uden (95% CI 24.2 til 55.1)*

Kvantificering af effekten II

Relativ risiko:

$$\frac{p_{med}}{p_{uden}} = \frac{75.6}{36.0} = 2.10$$

Risikoen for spredning er 2.1 gange større for patienter med involvering ifht patienter uden involvering (95% CI 1.68 til 2.63)

Patienter med involvering har 110% større risiko for spredning end patienter uden (95% CI 68 til 163%)

Kvantificering af effekten III

Odds for patienter med involvering:

$$odds_{med} = \frac{p_{med}}{1 - p_{med}} = \frac{31/41}{1 - 31/41} = \frac{31}{10} = 3.1$$

Odds for patienter uden involvering:

$$odds_{uden} = \frac{p_{uden}}{1 - p_{uden}} = \frac{122}{217} = 0.56$$

Odds ratio:

$$\text{OR} = \frac{odds_{med}}{odds_{uden}} = \frac{3.1}{0.56} = 5.51$$

Odds for spredning er 5.5 gange større for patienter med involvering end for patienter uden (95% CI 2.61 til 11.63)

Formålet med logistisk regression

For et **binært outcome**, e.g.

$$Y_i = \begin{cases} 1 & \text{hvis patient } i \text{ havde spredning} \\ 0 & \text{hvis patient } i \text{ ikke havde spredning} \end{cases}$$

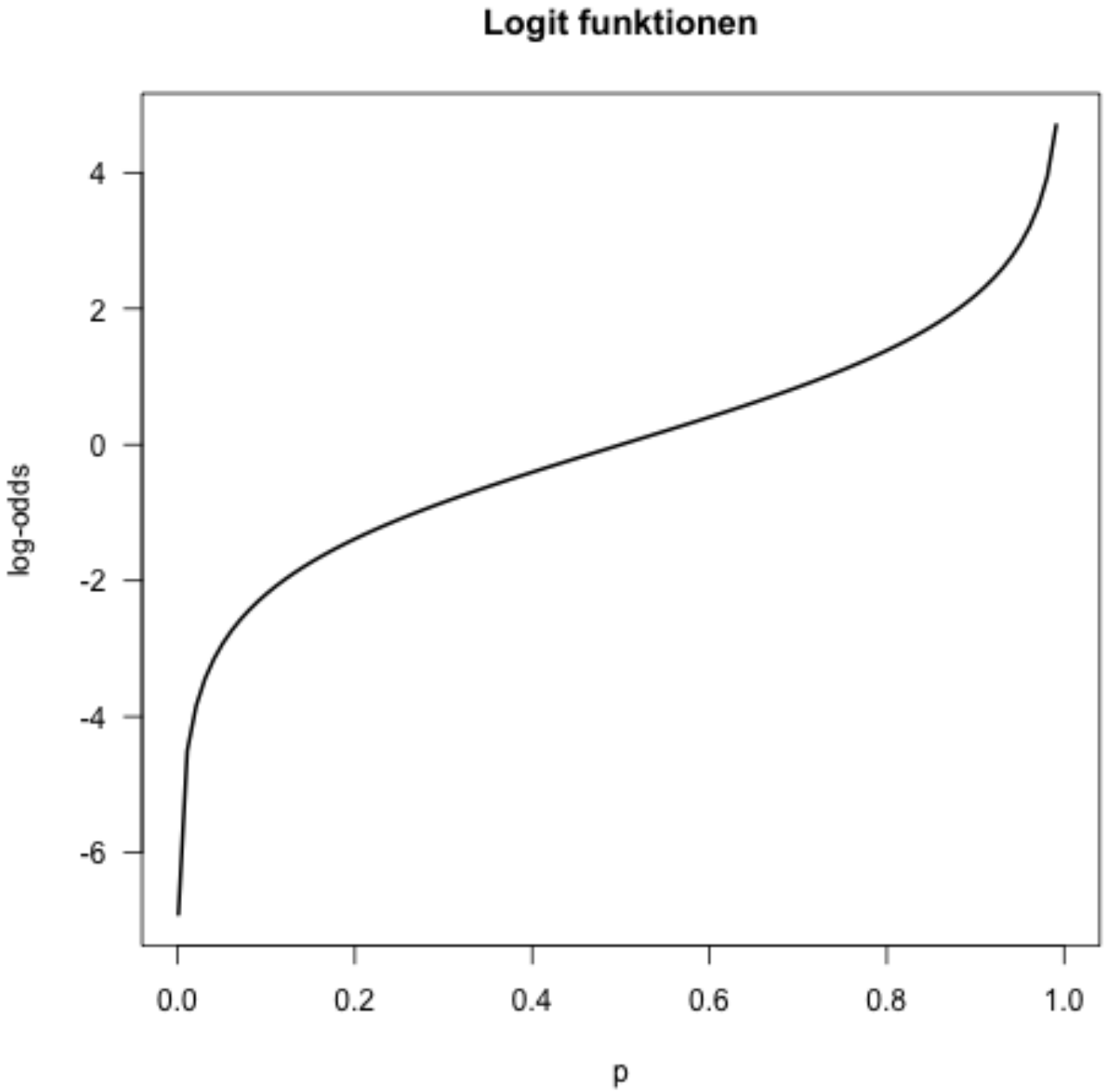
at beskrive sammenhængen med forklarende variable for patient i.

I logistisk regression formulerer vi modeller for **log-odds**:

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

som vi kalder for logit-funktionen.

Sandsynlighed og log-odds



Den logistiske regressionsmodel

Kovariat:

$$\texttt{involvering}_i = \begin{cases} 0 & \text{hvis patient } i \text{ uden involvering} \\ 1 & \text{hvis patient } i \text{ med involvering} \end{cases}$$

Model:

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b \cdot \texttt{involvering}_i = \begin{cases} a & i \text{ uden involvering} \\ a+b & i \text{ involvering} \end{cases}$$

$$= \begin{cases} \log(\frac{122}{217}) \\ \log(\frac{31}{10}) \end{cases} = \begin{cases} -0.58 \\ 1.13 \end{cases}$$

$$= \begin{cases} -0.58 \\ -0.58 + (1.13 + 0.58) \end{cases} = \begin{cases} -0.58 \\ -0.58 + 1.71 \end{cases}$$

Forskellen i **log-odds** mellem patienter med og uden involvering er b=1.71 (?!)

OR fra logistisk regression

$$\log\left(\frac{p_i}{1-p_i}\right) \quad = \quad a + b \cdot \texttt{involvering}_i = \begin{cases} a & i \text{ uden involvering} \\ a+b & i \text{ med involvering} \end{cases}$$

Effekten af involvering er givet ved b:

$$\begin{aligned} b &= (a+b) - a \\ &= \log \text{ (odds med involvering) - log (odds uden involvering)} \\ &= \log\left(\frac{\text{odds med involvering}}{\text{odds uden involvering}}\right) = \log(\text{OR}) \end{aligned}$$

ie.

$$\exp(b) = \text{OR} = \exp(1.71) = 5.51$$

Hvad er OR for ingen involvering vs involvering?

Logistisk regression i R / SAS

Logistisk regression hører under klassen af **G**eneraliserede **L**ineære **M**odeller (GLM - ikke at forveksle med *Generel* LM)

R-kode:

```
glm1 <- glm( spredning ~ involvering, data=d,
             family = 'binomial')
summary( glm1 )
```

SAS:

```
proc logistic data = d;
   class involvering (ref='0') / param=glm;
   model spredning(event='1') = involvering;
run;
```

Logistisk regression i SPSS

Analyze -> Generalized Linear Models

5 faner skal udfyldes:

Type of Model: Vælg Binary Logistic (under Binary Response)

Response: Sæt `spredning` som Dependent Variable. Under Reference category vælges First (lowest value), således at `spredning`=1 vurderes mod `spredning`=0 (dvs risikoen for spredning i stedet for risikoen for ikke at have spredning)

Predictors: Her sættes kvalitative (faktorer, her `involvering`) i Factors (i Options afkrydse descending under Category Order for Factors for at få 0 som reference). Kvantitative sættes i Covariates.

Model: Som GLM

Statistics: Afkryds Include exponential parameter estimates (under Print)

Output i R - I

```r
glm1 <- glm( spredning ~ involvering, data=d,
             family = 'binomial')
summary( glm1 )
```

```
Call:
glm(formula = spredning ~ involvering, family = "binomial", data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6799  -0.9446  -0.9446   1.4297   1.4297

Coefficients:
            Estimate Std. Error z value   Pr(>|z|)
(Intercept)  -0.5759     0.1132  -5.089 0.00000036 ***
involvering   1.7073     0.3809   4.483 0.00000738 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 512.29  on 379  degrees of freedom
Residual deviance: 488.53  on 378  degrees of freedom
AIC: 492.53

Number of Fisher Scoring iterations: 4
```

Output i R - II

```
coef( glm1 )
```

```
(Intercept) involvering
 -0.5758763   1.7072784
```

```
exp( coef(glm1) )
```

```
(Intercept) involvering
   0.562212    5.513934
```

```
exp( confint.default(glm1) )
```

```
                 2.5 %    97.5 %
(Intercept) 0.4503797  0.701813
involvering 2.6137465 11.632143
```

Output i SAS

Analysis of Maximum Likelihood Estimates

Parameter	DF		Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1		-0.5759	0.1132	25.8987	<.0001
involvering	1	1	1.7073	0.3809	20.0933	<.0001
involvering	0	0	0	.	.	.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
involvering 1 vs 0	5.514	2.614	11.632

Output - SPSS

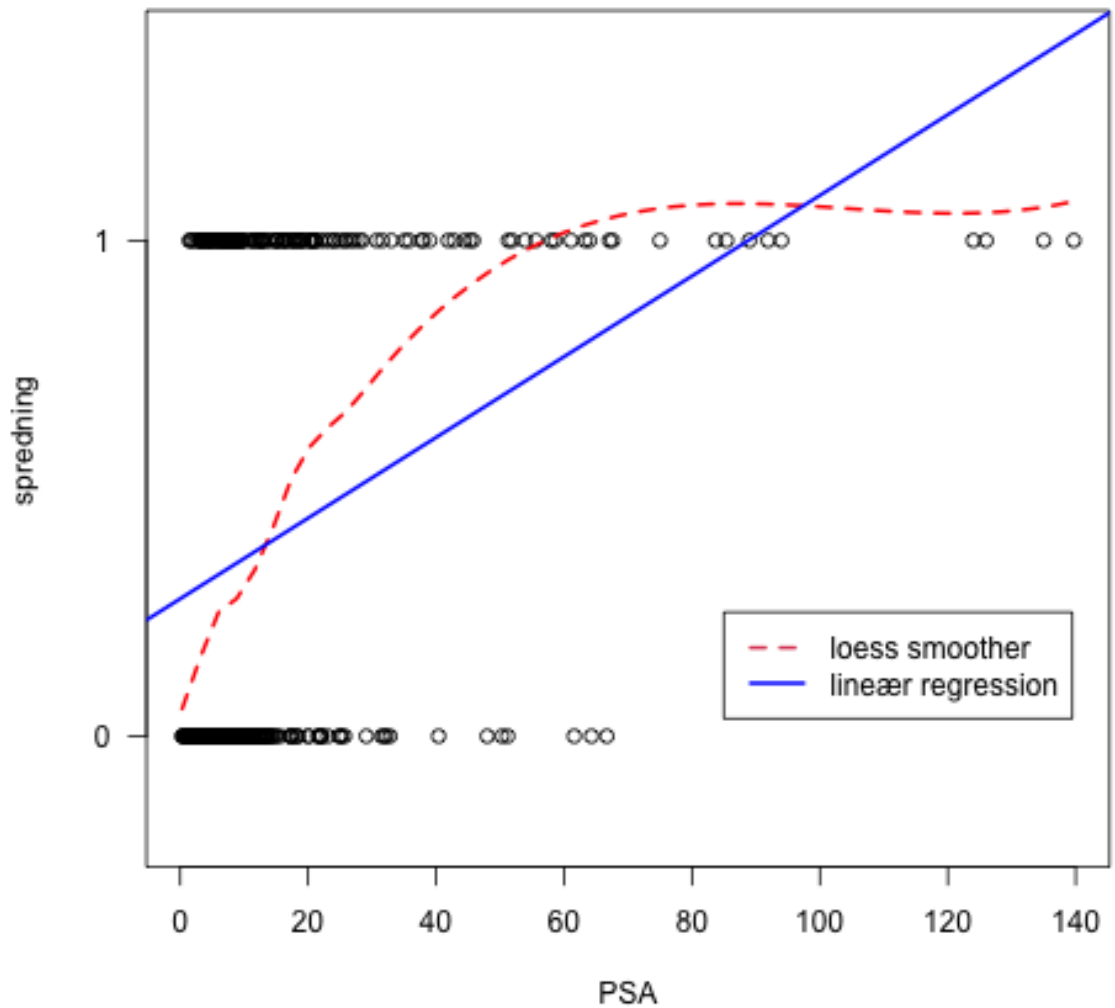
Fylder alt for meget til at det kan være her . . .

Under blokken med 'Parameter Estimates' aflæses

- a og b i første søjle (B)
- OR i 3. sidste søjle (Exp(B))
- CI i de to sidste søjler
- p-værdi i 4. sidste søjle

Kvantitativ kovariat

Lineær regression går ikke:



Logistisk regression med en kvantitativ kovariat

Model for log-odds er lineær: $\log\left(\frac{p_i}{1-p_i}\right) = a + b \cdot \texttt{psa}_i$

Sammenlign to patienter med en forskel på 1 μg/L PSA, feks 51 vs 50:

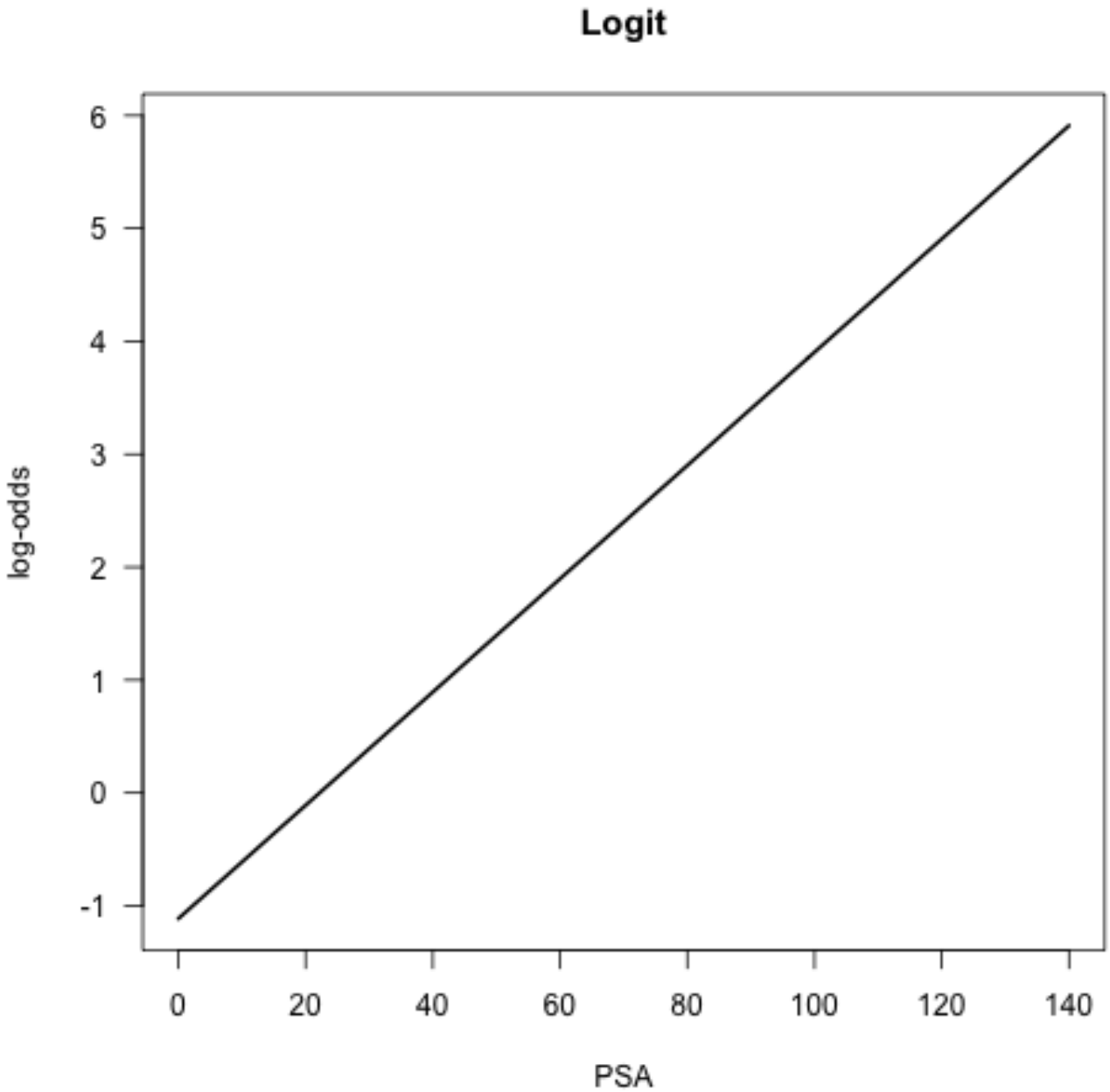
$$\text{OR} = \frac{\text{odds PSA} = 51}{\text{odds PSA} = 50}$$

$$\begin{aligned} \log(\text{OR}) &= \log(\text{odds PSA} = 51) - \log(\text{odds PSA} = 50) && (1) \\ &= (a + 51 \cdot b) - (a + 50 \cdot b) && (2) \\ &= b && (3) \end{aligned}$$

i.e.

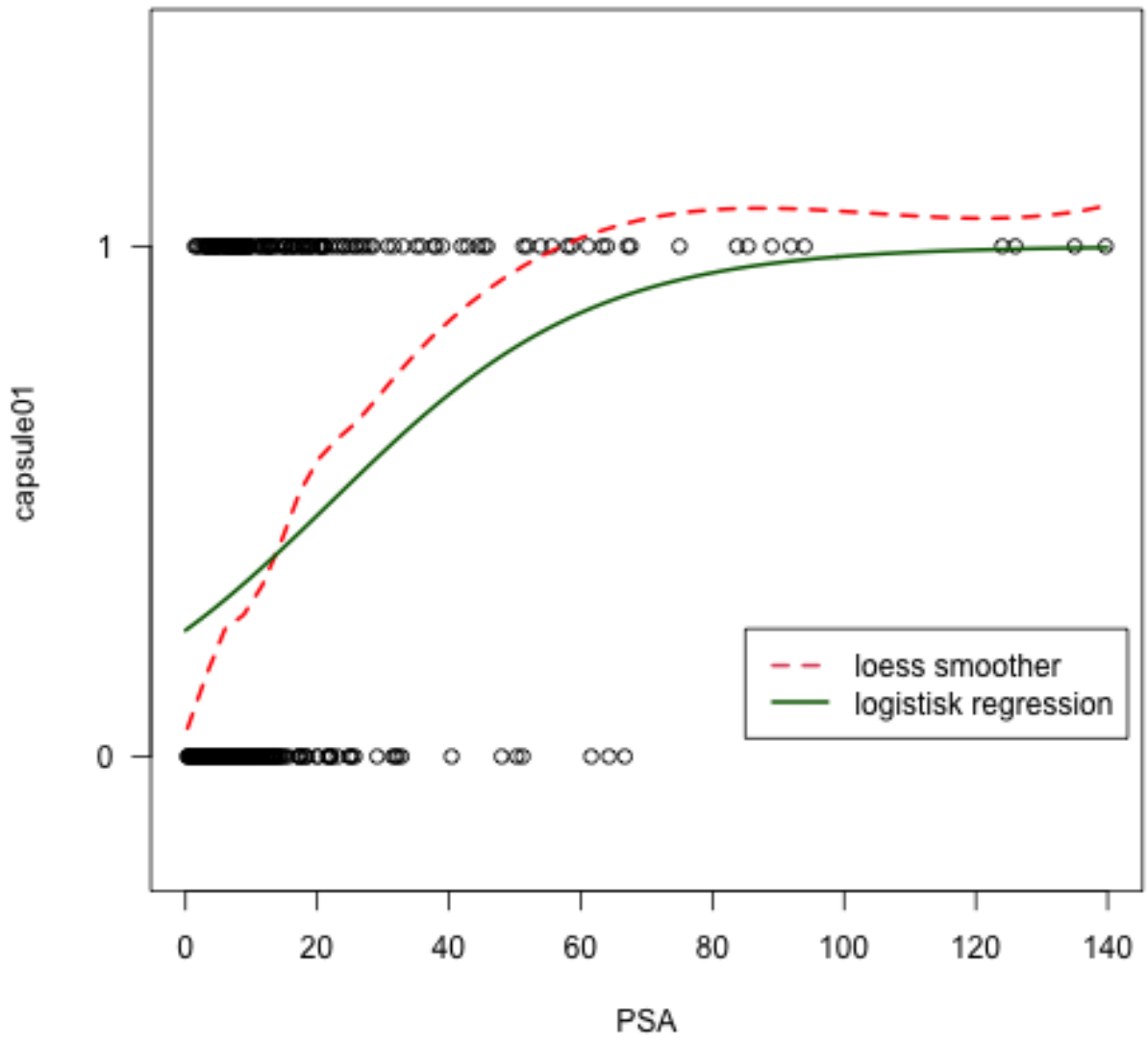
$$\text{OR} = \exp(b) = \exp(0.05) = 1.05.$$

Plot på logit-skala



Prædikterede sandsynligheder

$$p(x) = \frac{\exp(a + b \cdot x)}{1 + \exp(a + b \cdot x)}$$



Øvelse: Rapportering pr 10 enheder

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= a + b \cdot \texttt{psa}_i, \qquad b = 0.0502 \\ OR &= \exp(0.0502) = 1.0513 \end{aligned}$$

Vi vil næppe rapportere effekten af PSA pr 1 enhed μg/L.

Hvad er effekten af PSA pr 10 μg/L?

Man finder en OR på:

1. 1.50
2. 1.65
3. 1.89
4. 10.51

Modelkontrol

Vi har antaget at effekten af PSA lineær (på logit-skala). Er det rimeligt?

Numerisk modelkontrol:

- **Overall goodness of fit:** Hosmer-Lemeshow test
- **Linearitet:** Lineære splines, tilføj logaritmeret eller kvadreret kovariat
- **Grafisk modelkontrol:** Residualplot
- **Diagnostics:** Cook, dfbetas

Overall test for Goodness-of-fit

Hosmer-Lemeshow goodness-of-fit test:

- Observationerne inddeles i 10 ca. lige store grupper, baseret på stigende prædikteret sandsynlighed for gennemtræning
- I hver gruppe sammenlignes observerede og forventede antal af spredninger og størrelserne

$$\frac{(\text{observeret antal} - N\hat{p})^2}{N\hat{p}(1-\hat{p})}$$

sammenlægges til en approksimativ χ^2-teststørrelse med 8 frihedsgrader (antal grupper minus 2)

Test af goodness-of-fit

	y0	y1	yhat0	yhat1
[0.25,0.272]	36	4	29.489598	10.51040
(0.272,0.291]	25	11	25.851318	10.14868
(0.291,0.305]	28	10	26.690339	11.30966
(0.305,0.322]	25	15	27.435075	12.56493
(0.322,0.337]	20	16	24.114498	11.88550
(0.337,0.363]	34	10	28.603123	15.39688
(0.363,0.395]	21	12	20.447832	12.55217
(0.395,0.474]	15	22	20.927518	16.07248
(0.474,0.635]	16	22	17.595647	20.40435
(0.635,0.997]	7	31	5.845053	32.15495

Her finder vi $\chi^2 = 15.95$ og dermed p=0.04. Dermed halter modellen lidt.

Overall test af goodness-of-fit i praksis

Vi har kun én forklarende variabel. Hosmer-Lemeshow testet kan også benyttes på multivariable modeller.

I tilfælde af sparsomme data kan inddelingen have en del indflydelse på testet, dvs. det er meget **ustabilt**. SAS giver her f.eks. pga anden inddeling p=0.001...

Desuden kan det ændre sig, hvis man skifter til at se på det modsatte outcome altså spredning="0" (i R er p=0.09 vs p=0.04).

Residualplot

Pearson residualer

$$res_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i \cdot (1 - \hat{p}_i)}}$$

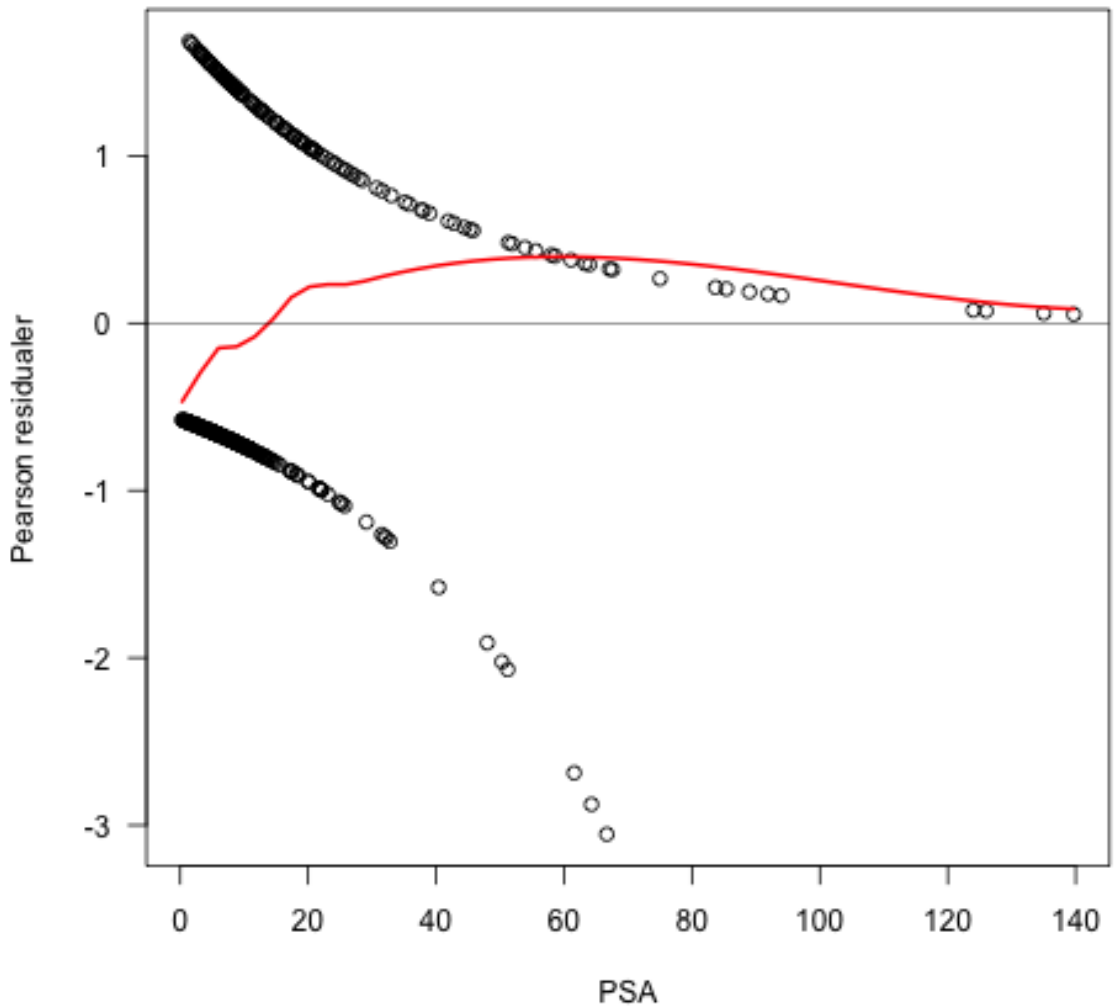
plottes vs

- kovariater
- fittede værdier

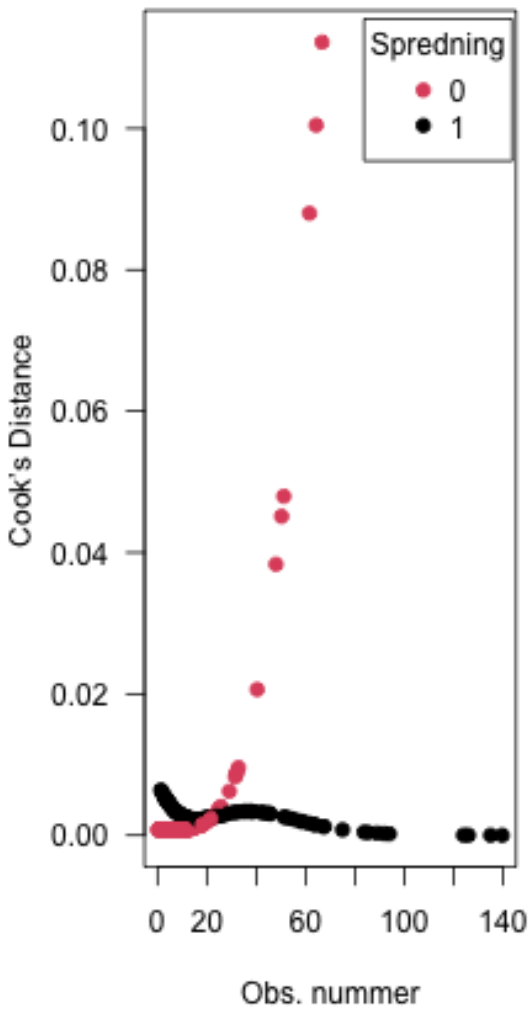
for at se efter krumninger som tegn på manglende linearitet.

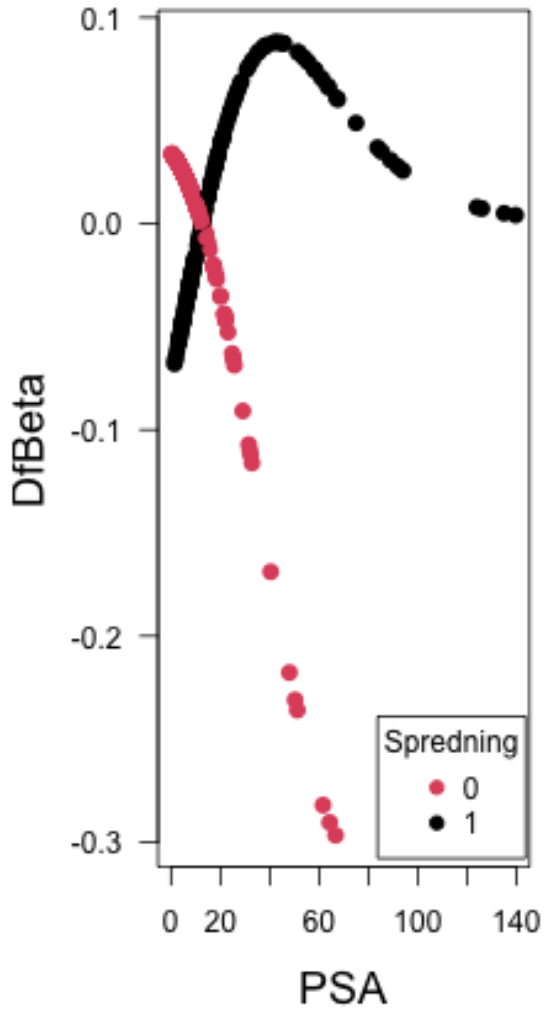
Residualplot

Bliver skøre at se på . . .



Diagnostics plots





PSA som lineær spline

Tærskelværdier bestemmes ud fra kvartiler for PSA blandt 'cases':

25%: 7.4 50%: 13.2 75%: 26.0

Splinevariable:

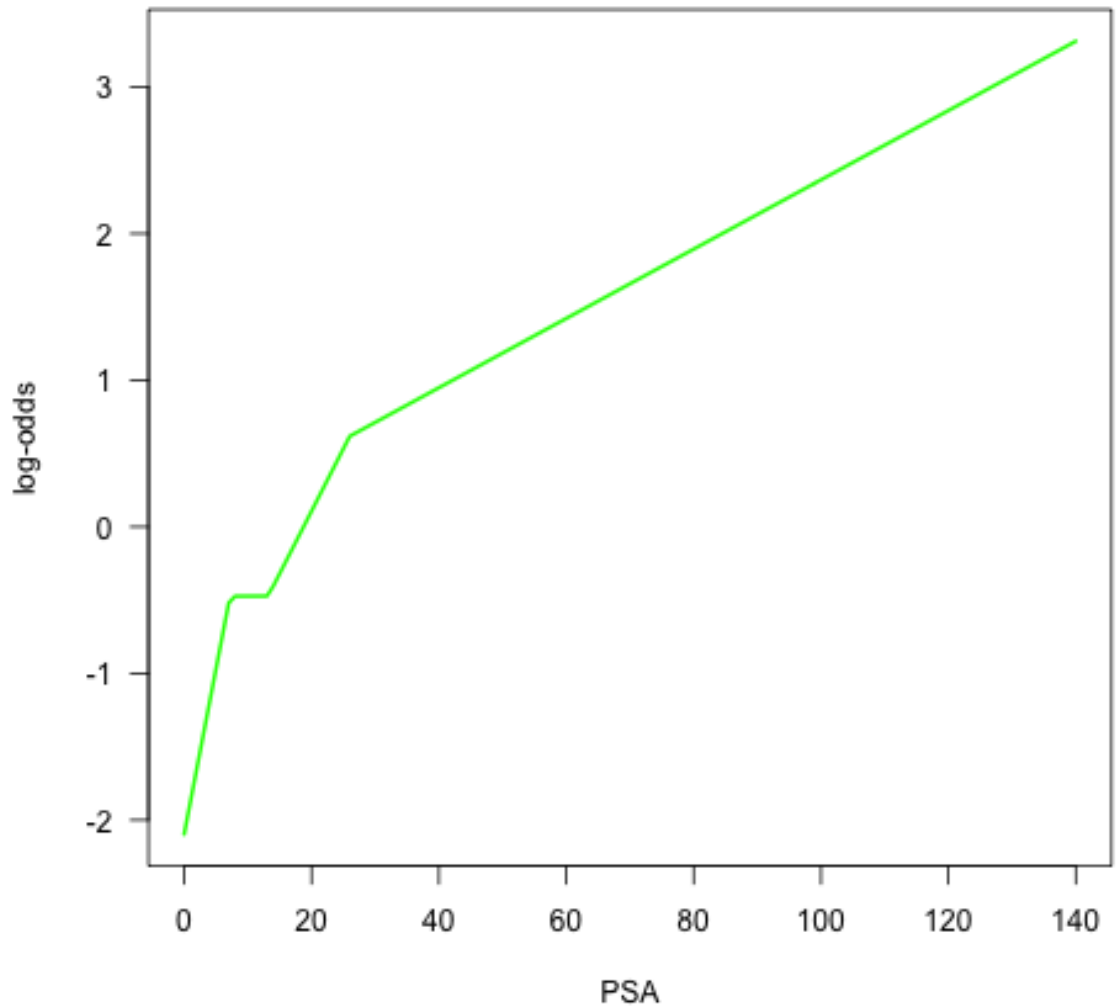
$$\text{psa\_7\_4} = \begin{cases} \text{psa} - 7.4 & \text{psa} > 7.4 \\ 0 & \text{psa} \leq 7.4 \end{cases}$$

Tilsvarende defineres psa_13_2 og psa_26

Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0697	0.4651	-4.4500	0.0000
psa	0.2178	0.0817	2.6661	0.0077
psa_7_4	-0.2209	0.1352	-1.6336	0.1023
psa_13_2	0.0887	0.1035	0.8568	0.3915
psa_26	-0.0619	0.0498	-1.2444	0.2134

Effekt af PSA som lineær spline



Test af linearitet: H_0: psa_7_4=psa_13_2=psa_26=0 giver p=0.057 (df=3)

Rapportering af splinemodellen

Med en lineær spline afhænger OR af PSA-værdierne:

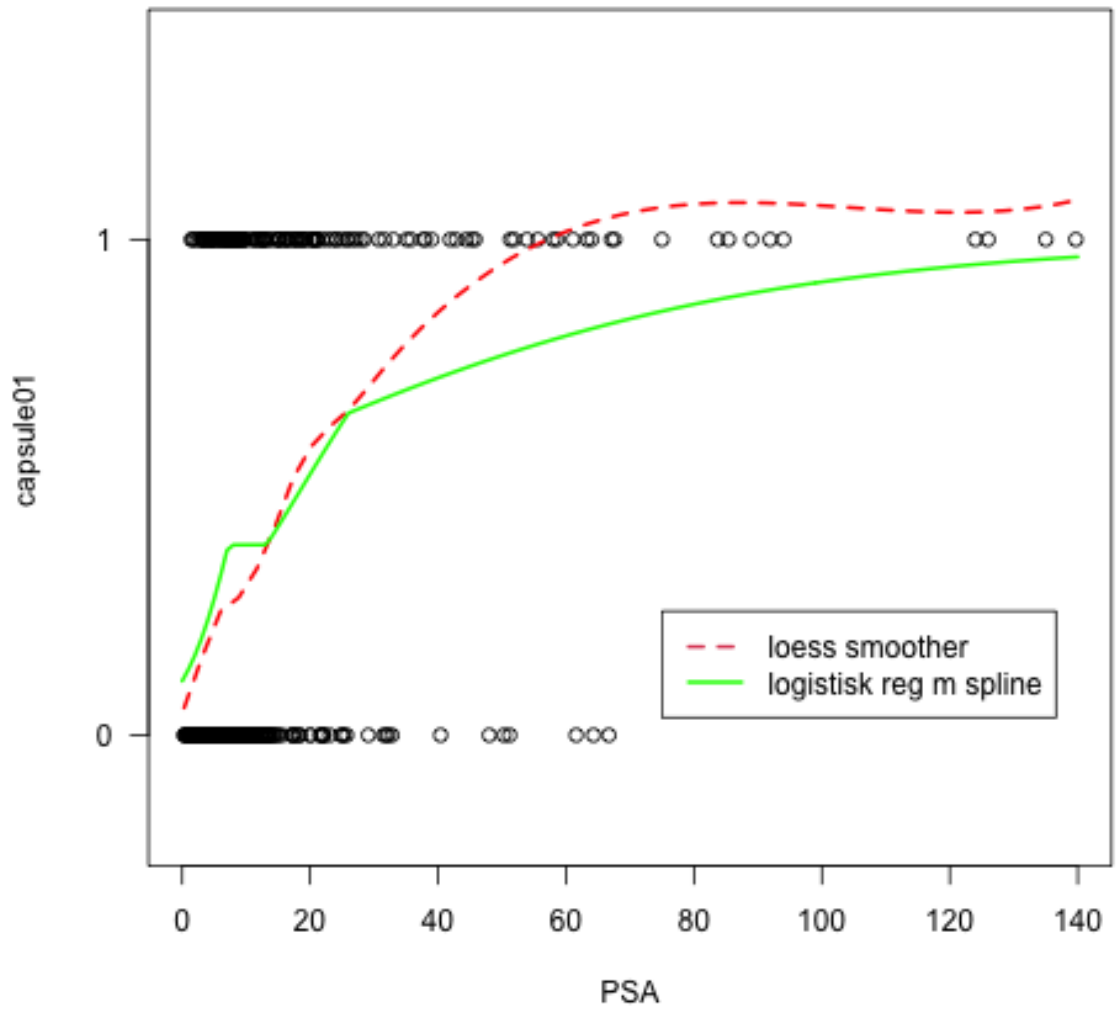
I intervallerne er

PSA	log-odds bidrag	OR	CI
$<$ 7.2:	.2178	1.24	1.06-1.48
7.2-13.2:	.2178-.2209	1.00	0.86-1.15
13.2-26.0:	.2178 -.2209+.0887	1.09	1.01-1.18
$>$ 26.0:	.2178-.2209+.0887-.0619	1.02	1.00-1.05

pr 1 μg/L PSA.

Prædikterede sandsynligheder

Hosmer-Lemeshow p=0.50.



Test for lineær effekt af log(PSA)

Definér ny variabel: log2psa=$\log_2$(PSA)

Model: $$\log\left(\frac{p_i}{1-p_i}\right) = a + b \cdot \texttt{psa}_i + c \cdot \texttt{log2psa}_i$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0707	0.4119	-5.0270	0.0000
psa	0.0147	0.0138	1.0598	0.2892
log2psa	0.4459	0.1651	2.7010	0.0069

Med $\log_2$(PSA) i modellen bliver PSA overflødig, p=0.29

Model med log(PSA)

Model: $$\log\left(\frac{p_i}{1-p_i}\right) = a + c \cdot \log_2(\text{psa}_i)$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.3756	0.3256	-7.2969	0
log2psa	0.6020	0.0909	6.6232	0

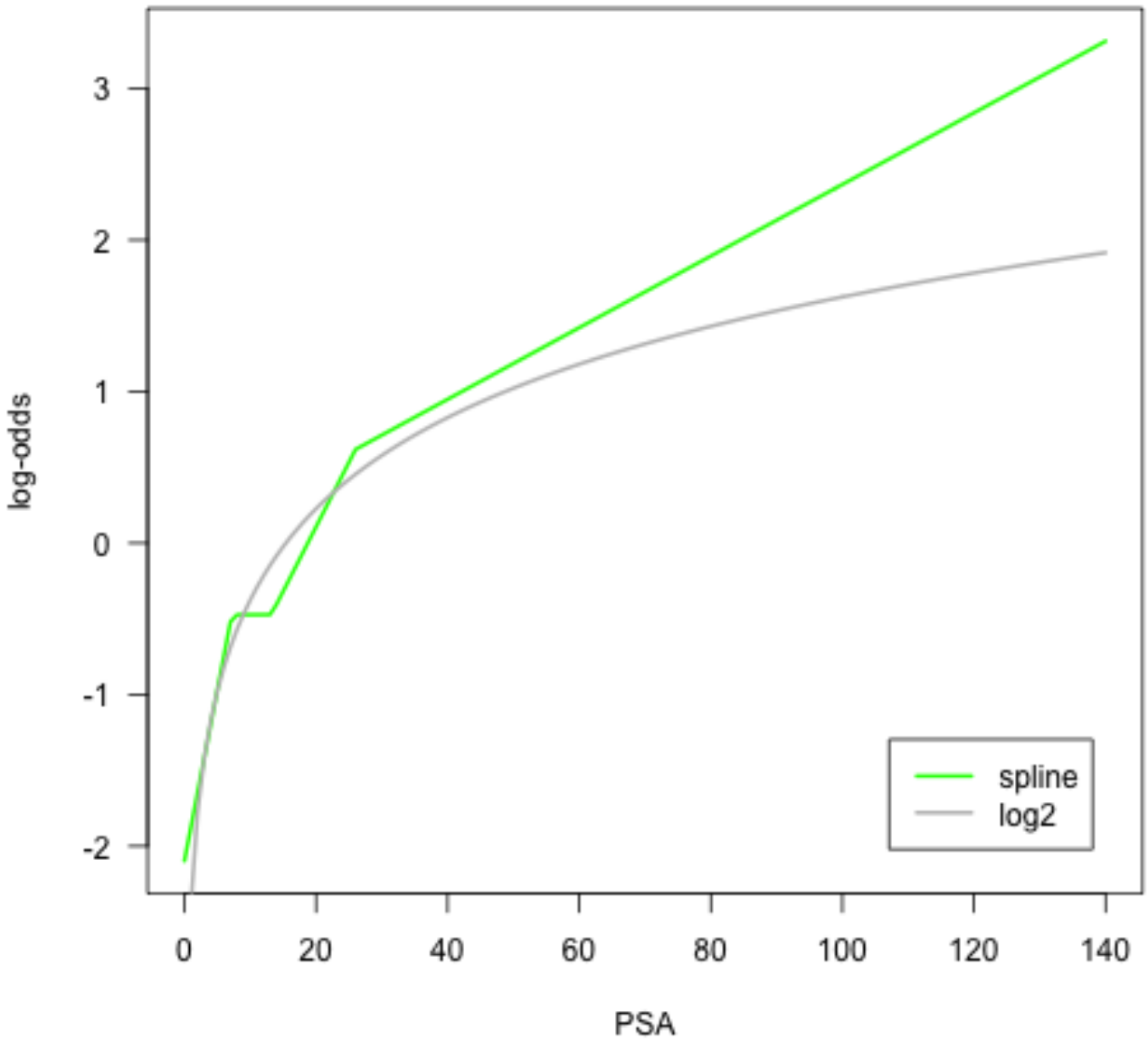
Effekten af $\log_2$(PSA) er beskrevet ved $c = 0.602$ (95 %CI 0.424 til 0.780)

*Dvs ved en **fordobling** af PSA ...*

... er OR = exp(0.602) = 1.83 (95% CI 1.53 til 2.18)

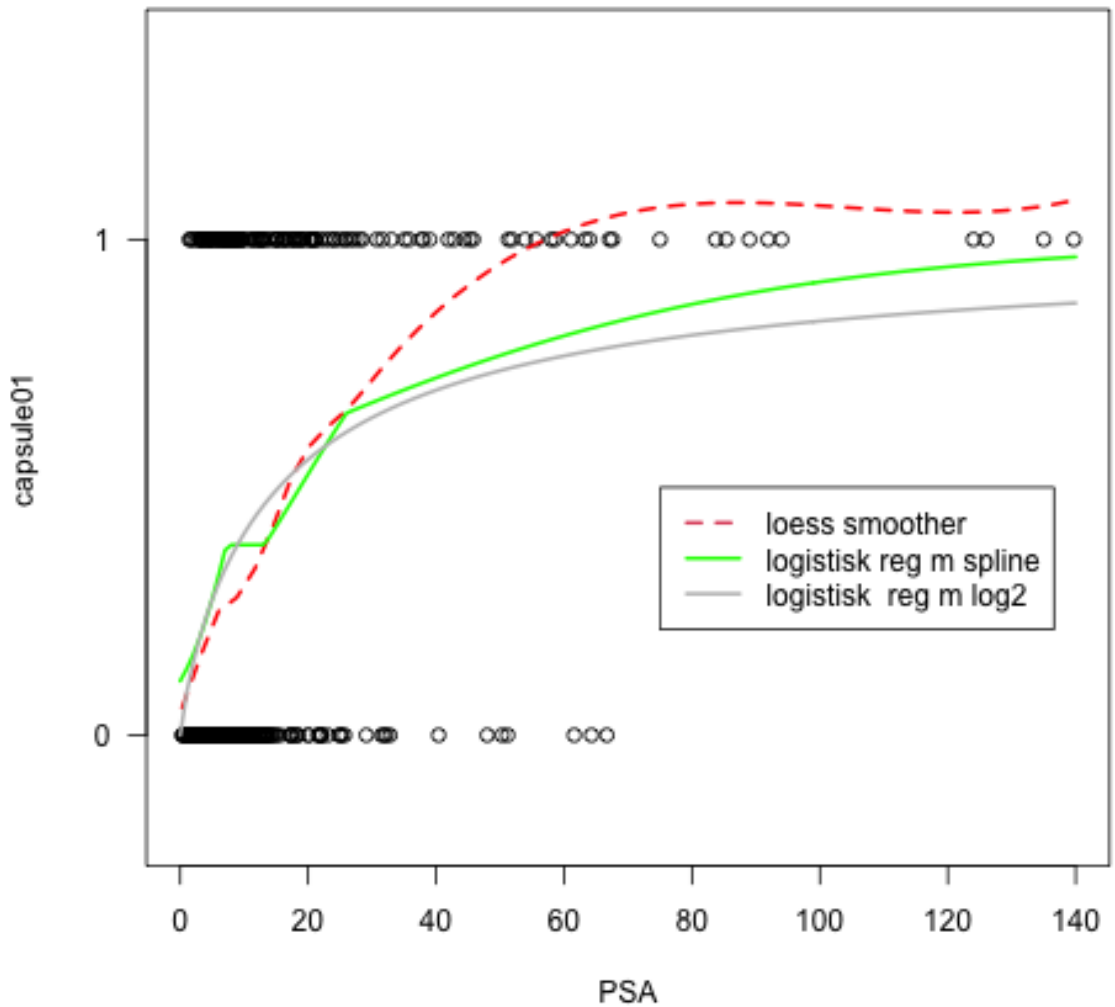
... øges odds for spredning med 83% (95% CI 53 til 118%)

Effekt af log2(PSA) på logit-skala



Prædikterede sandsynligheder

Hosmer-Lemeshow p=0.17.



En kvalitativ og en kvantitativ kovariat

Model:

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= a + b \cdot \texttt{involvering}_i + c \cdot \log_2(\texttt{psa}_i) \\
&= \begin{cases} a + c \cdot \log_2(\texttt{psa}_i) & \text{hvis involvering}_i = 0 \\ a + b + c \cdot \log_2(\texttt{psa}_i) & \text{hvis involvering}_i = 1 \end{cases}
\end{aligned}$$

Hvad forventer vi sker med OR for involvering (ujusteret OR=5.51)?

Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.3242	0.3276	-7.0936	0.000
involvering	1.2530	0.4053	3.0914	0.002
log2psa	0.5487	0.0927	5.9163	0.000

For fastholdt PSA er OR=exp(1.25)=3.50 for involvering vs ingen involvering

For fastholdt involvering er OR=exp(0.55)=1.73 ved en fordobling af PSA

Fiktivt eksempel

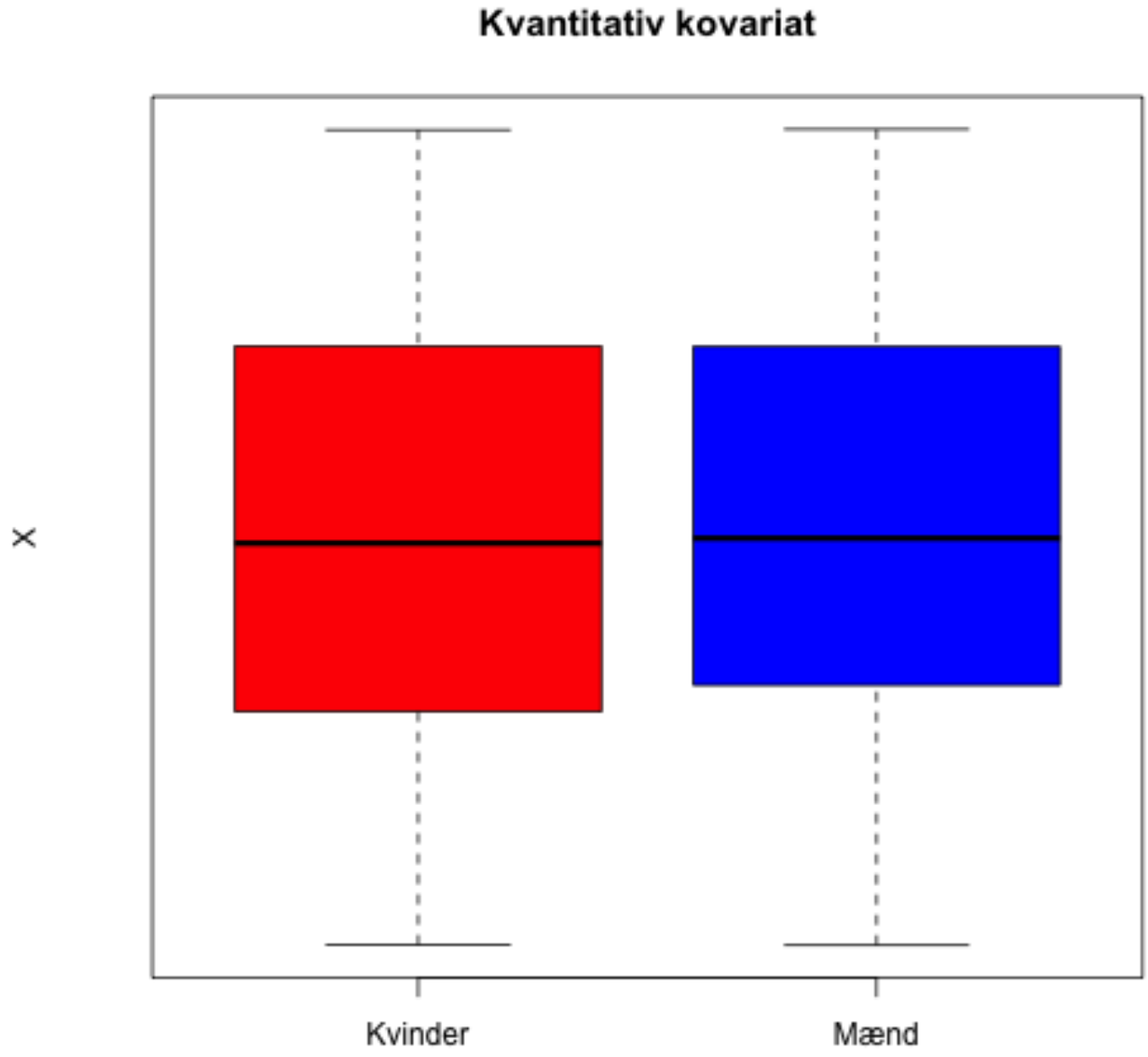
Hvornår skal man justere?

Tænkt eksempel: Køn, kvantitativ kovariat X og binært outcome Y.

Både køn og X er prædiktive for Y, men der er ***ingen*** association mellem køn og X.

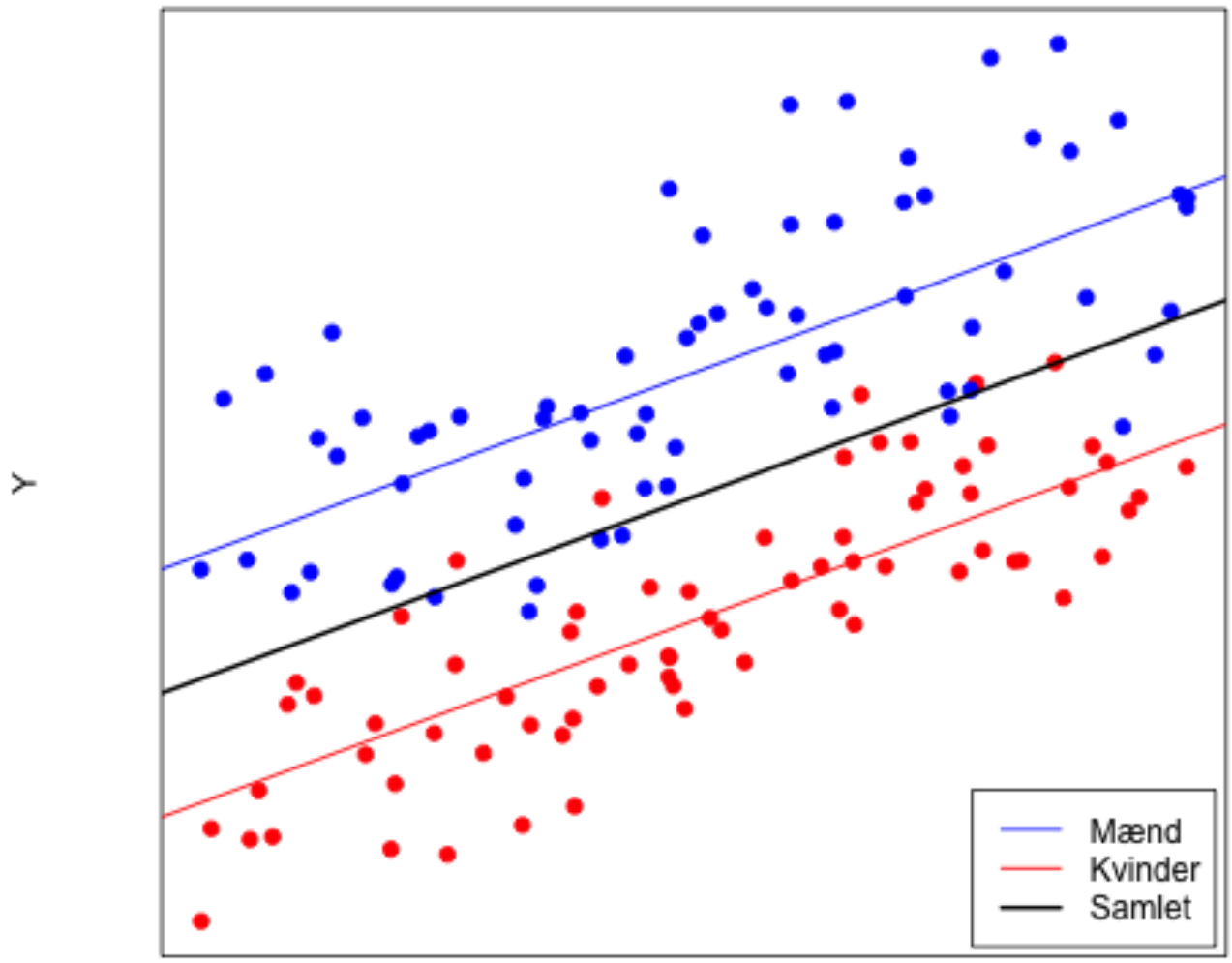
Dvs der er ingen confounding.

X vs køn

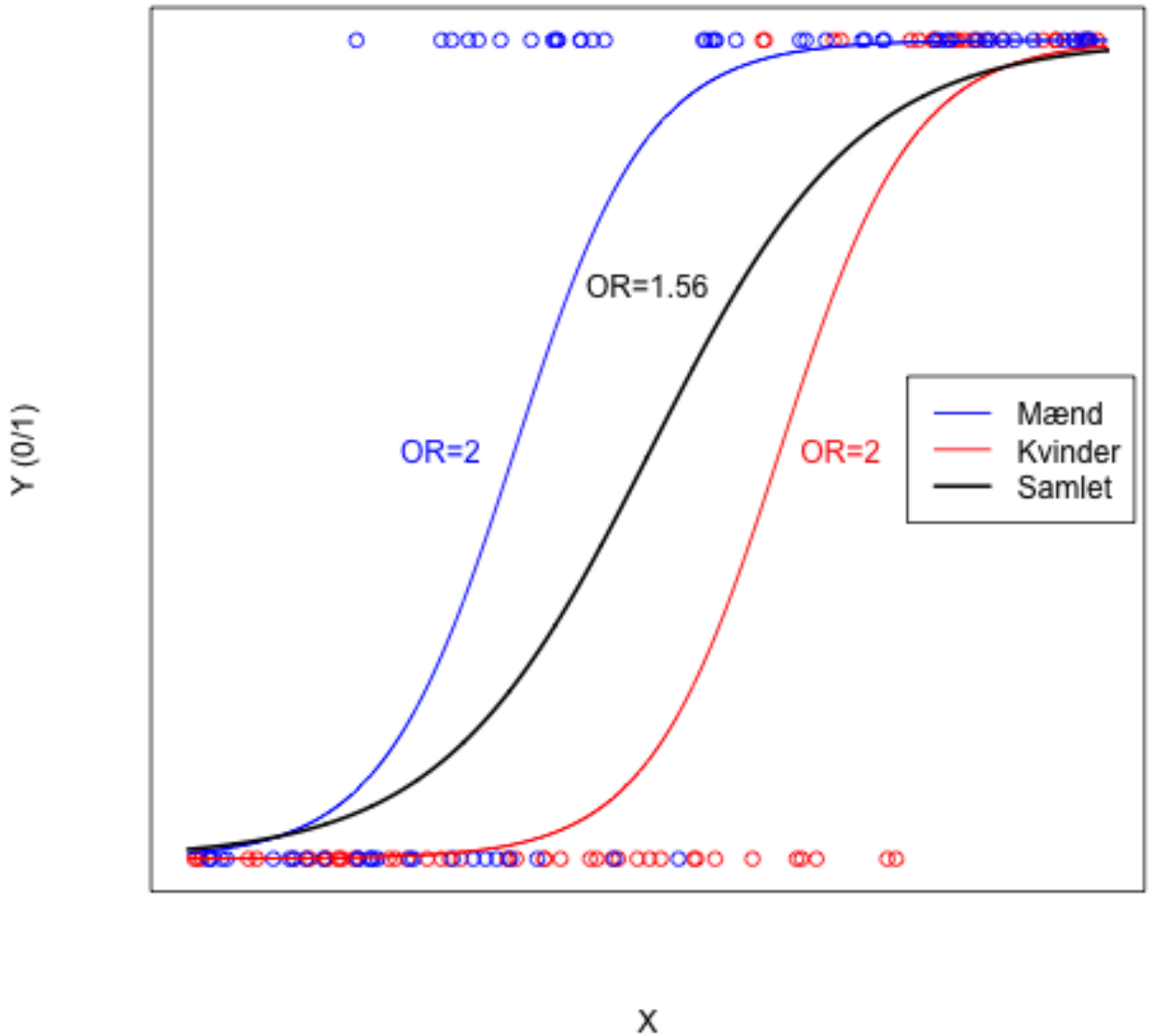


Tilbageblik på lineær regression

Her er Y kvantitativ:



Logistisk regression med / uden betydende kovariat



Konsekvens af manglende kovariat

Mangler vi en vigtig prædiktor for outcome bliver OR i det samlede materiale (/ujusteret) *mindre* end i subgrupperne (/justeret).

Det skyldes (*teknisk!*) at logit-funktionen ikke er lineær og at vi ved udeladelse af kovariaten tager gennemsnit over *inhomogene* populationer.

Randomiserede studier:

- Vi har sjældent alle kovariater
- Er der strenge inklusionskriterier har vi en ret homogen population, og OR bliver formentligt stor
- Er der ingen inklusionskriterier har vi en ret inhomogen population, og OR bliver formentligt lille

Vi kan *ikke* direkte sammenligne OR'er, hvor vi ikke justerer for det samme!

Interaktion

Spredning vs Gleason:

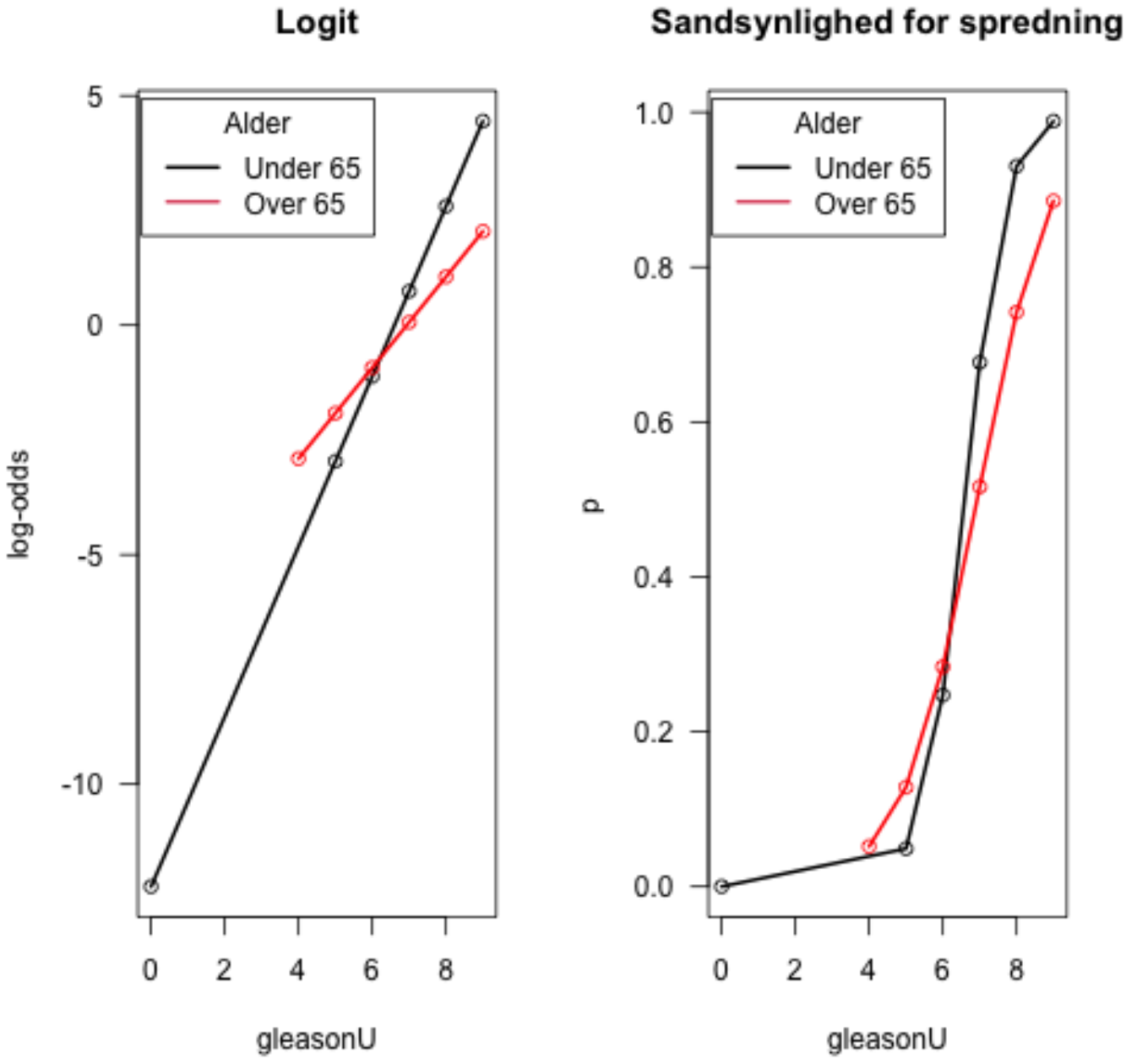
	gleason	Antal	antal0	antal1	Andele	andel0	andel1
1	0		2	0		1.00	0.00
2	4		1	0		1.00	0.00
3	5		61	6		0.91	0.09
4	6		101	38		0.73	0.27
5	7		55	73		0.43	0.57
6	8		6	24		0.20	0.80
7	9		1	12		0.08	0.92

Har Gleason score lige stor betydning for 'unge' og 'ældre'?

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.8693	1.1711	-5.8659	0.0000
alder65Under	-5.3611	2.3484	-2.2829	0.0224
gleason	0.9906	0.1759	5.6303	0.0000
alder65Under:gleason	0.8623	0.3619	2.3828	0.0172

Muligvis - vi har en p-værdi på 0.02.

Prædikterede sandsynligheder



Interaktionsmodeller i R / SAS / SPSS

Syntaks i R

```
glm( spredning ~ alder65*gleason, data=d, family='binomial')
glm( spredning ~ alder65+alder65:gleason, data=d, family='binomial')
```

Syntaks i SAS (proc logistic)

```
proc logistic data=d;
  class alder65 (ref='Under') / param=glm;
  model spredning(event='1') = alder65 gleason alder65*gleason;
  oddsratio gleason / at ( alder65="Under" "Over");
run;
```

SPSS

```
Når modellen bygges tilføjes alder65 og interaktionen mellem alder65 og gleason
```

Rapportering af modellen

Vi påstår, at der er en interaktion mellem aldersgruppe og gleason score, p=0.02

Vi må derfor rapportere effekten af gleason for hver aldersgruppe. Her finder vi

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.8693	1.1711	-5.8659	0.0000
alder65Under	-5.3611	2.3484	-2.2829	0.0224
alder65Over:gleason	0.9906	0.1759	5.6303	0.0000
alder65Under:gleason	1.8529	0.3162	5.8595	0.0000

For de yngre (<= 65 år) er OR=exp(1.85)=6.37 pr 1 i gleason score, 95% CI 3.43 til 11.85.

For de ældre (> 65 år) er OR=exp(0.99)=2.69 pr 1 i gleason score, 95% CI 1.91 til 3.80.

En multivariabel model

Med udgangspunkt i de forrige modeller kan vi formulere en model med alle variable, f.eks.

	Estimate	Pr(>|z|)	OR	lower	upper
(Intercept)	-6.834	0.000	0.00	0.00	0.00
involvering	0.618	0.174	1.85	0.76	0.76
log2psa	0.316	0.003	1.37	1.11	1.11
alder65Under	-5.031	0.034	0.01	0.00	0.00
gleason	0.690	0.000	1.99	1.38	1.38
knudeBegge	1.390	0.002	4.01	1.67	1.67
knudeHoejre	1.463	0.000	4.32	2.08	2.08
knudeVenstre	0.721	0.044	2.06	1.02	1.02
alder65Under:gleason	0.796	0.029	2.22	1.08	1.08

Hosmer-Lemeshow goodness-of-fit p=0.25

Fortolkning:

> *Ved en fordobling af PSA øges odds for spredning med 37% (OR=exp(0.32)=1.37) for fastholdt involvering, alder, gleason og knudeplacering.*
>
> ...

Kan vi prædiktere?

$$\hat{p}_i = \frac{\exp(-6.83 + 0.62 \cdot \texttt{involvering} + 0.32 \cdot \texttt{log2psa} + \cdots)}{1 + \exp(-6.83 + 0.62 \cdot \texttt{involvering} + 0.32 \cdot \texttt{log2psa} + \cdots)}$$

Vi kan vælge en tærskelværdi, f.eks. p=0.5, og definere:

person i som **case** hvis $\hat{p}_i > 0.5$
person i som **kontrol** hvis $\hat{p}_i \leq 0.5$

	Prædiktion	
Spredning	case	kontrol
0	36	191
1	103	50

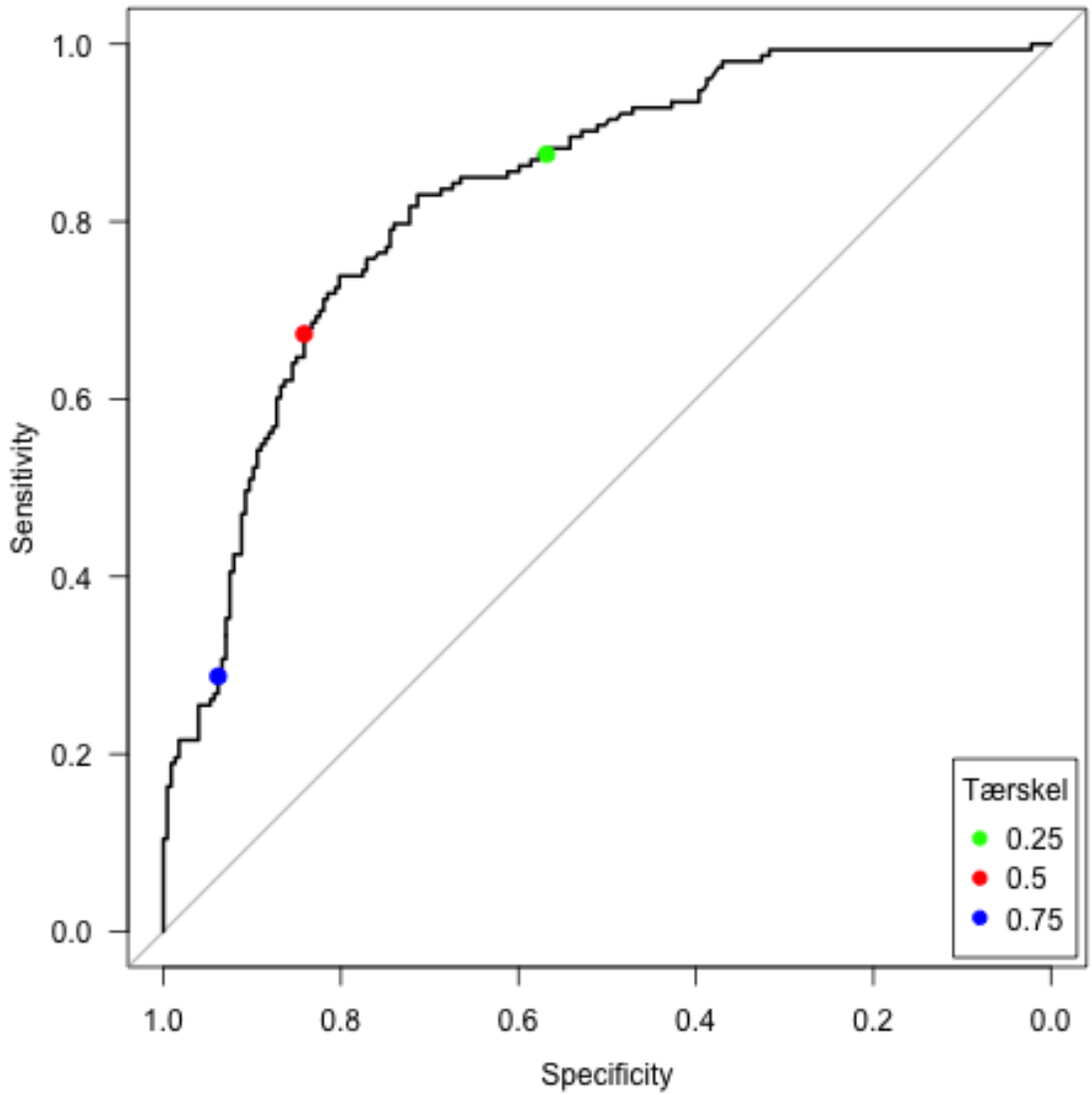
Sensitivitet: $\frac{103}{103+50}$=0.67

Specificitet: $\frac{191}{191+36}$=0.84

Andel korrekt klassificeret: $\frac{191+103}{380}$=0.77

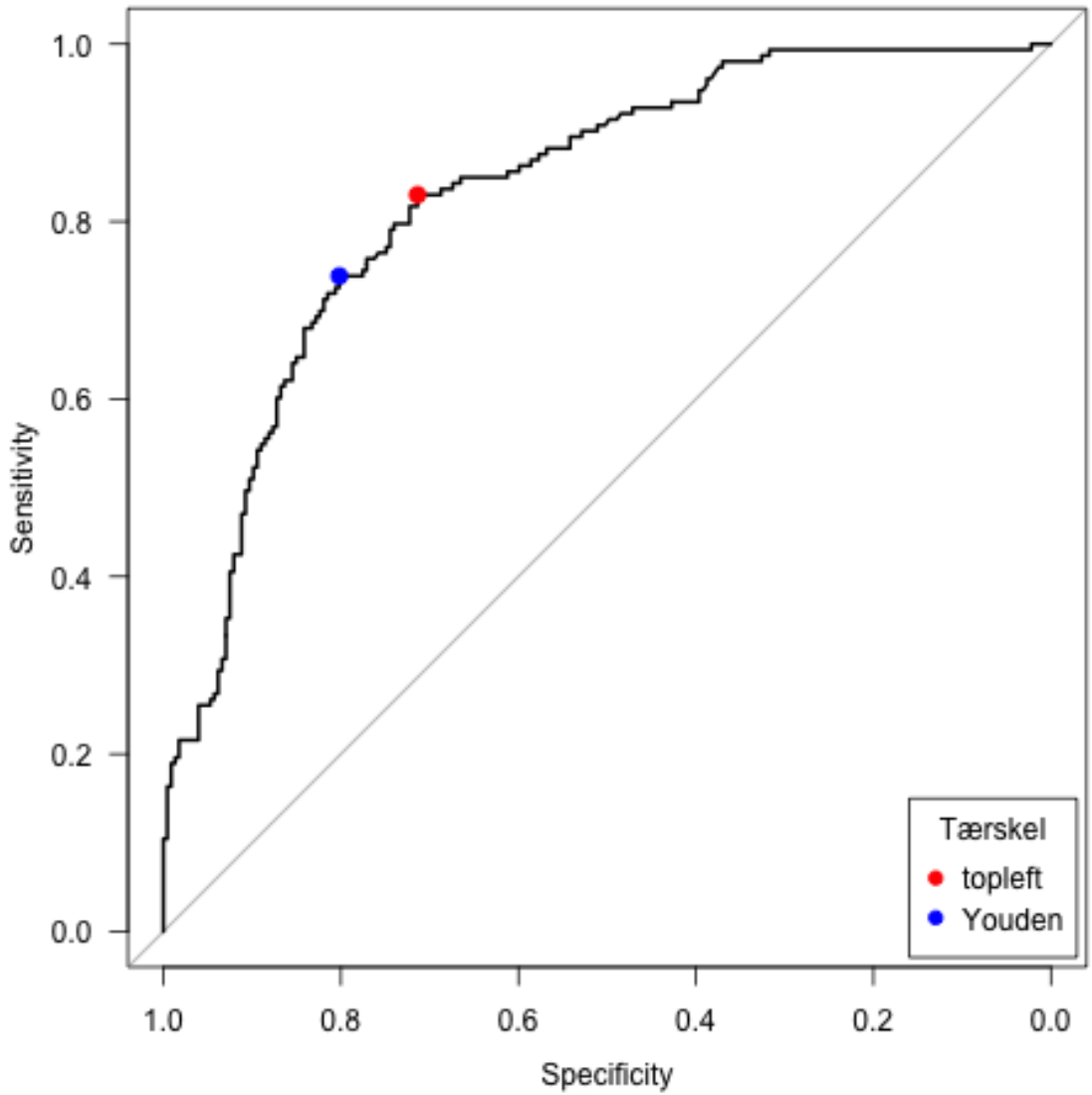
ROC kurve

Receiver Operating Characteristics kurve:



Optimal tærskelværdi

Closest top-left: 0.412 Youden: 0.323



AUC

Arealet under kurven (AUC) bruges som et mål for, hvor godt vi kan prædiktere:

1	Perfekt prædiktion!
0.5	Svarer til at vi kaster en mønt
0	Fuldstændig forkert prædiktion

Her er AUC=0.83 (95% CI 0.79-0.87)

ADVARSEL: Vi kan ***ikke*** evaluere hvor godt en model prædikterer på baggrund af de data, modellen er baseret på. AUC skal derfor korrigeres for *"optimisme"*.

- Training og validation sets
- Bootstrap
- Shrinkage (nedskalering) af koefficenterne fra modellen

Forslag til ekstra litteratur

Om brug af residualer og diagnostics:

> Zhang, Z. Residuals and regression diagnostics: focusing on logistic regression (2016) *Annals of Translational Medicine* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4885900/ (baseret på R-kode)

Korrektion for optimisme:

> Moons et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker (2012) *Heart*