Logistisk regression

Susanne Rosthøj
Section of Biostatistics
sro@sund.ku.dk

Outline

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

- Tabeller
  - Risiko / odds og relativ risiko / odds-ratio
- Simpel logistisk regression
  - Binær
  - Kvantitativ
- Modelkontrol
  - Diagnostics
- Linearitet
- Heterogenitet
- Interaktion
  - En binær og en kvantitativ
  - To kvalitative
- Prædiktion
  - ROC kurver, AUC

Regressionsanalyse

Kvantitativt outcome:

- Den generelle lineære model (uge 1-2, 4-7)
- Model for gennemsnitligt niveau - effekter beskrives ved ***forskelle i gennemsnit***

Binært outcome (0/1):

- Logistisk regression
- Model for sandsynligheder - effekter beskrives ved ***Odds Ratio***'er (OR)

Levetid:

- Cox regression
- Model for hazardfunktionen - effekter beskrives ved ***Hazard Ratio***'er (HR)

Ordinalt outcome (Proportional odds), **tælletal** (Poisson)

Eksempler på binære outcomes

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

Prostata kræft

380 patienter med prostatakræft: Data: R / SAS / SPSS

Outcome:

- `gennemtraengning01`: Tumor er trængt igennem prostatakapslen (0/1)

Forklarende variable målt ved baseline eksamination:

- `PSA`: Prostatic Specific Antigen Value, ng/mL
- `knude`: Knudes placering på lap ("ingen","venstre","hoejre","begge")
- `involvering`: Kapsel involvering ved Digital Rektal Eksploration (0/1, 1="involvering")
- `alder65`: Under / over 65 år ("Under"/"Over")
- `gleason`: Gleason score, 0-10.

Hvordan afhænger risikoen for at tumor er trængt igennem kapslen af disse variable?

Kan vi ud fra disse variable prædiktere om tumor er trængt igennem kapslen?

Binær kovariat og binært outcome

Tabelanalyse

Involvering og gennemtrængning

		Gennemtrængning				
		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 gennemtrængning (0/1) den samme uanset involvering eller ej?

- p_0 : Risiko for patienter *uden* involvering
- p_1 : Risiko for patienter *med* involvering

Hypotese: $p_0 = p_1$

Test for uafhængighed

Hypotese: $p_0 = p_1$ testes med

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

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

Her er p<.0001 ($\chi^2 = 23.9$, df=1) og vi konkluderer:

- risikoen for gennemtrængning er *forskellig* for patienter med og uden involvering.

Kvantificering af effekten I

Risikodifferens:

$$p_1 - p_0 = 75.6 - 36.0 = 39.6$$

Konklusion:

- Der er 39.6 procent**point** flere patienter med involvering der har gennemtrængning af prostatakapslen end patienter uden (95% CI 24.2 til 55.1)

Kvantificering af effekten II

Relativ risiko:

$$\frac{p_1}{p_0} = \frac{75.6}{36.0} = 2.10$$

Konklusion:

- Risikoen for gennemtrængning af prostatakapslen er 2.1 gange større for patienter med involvering ifht patienter uden involvering (95% CI 1.68 til 2.63)

eller

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

Kvantificering af effekten III

Odds for patienter med involvering:

$$odds_1 = \frac{p_1}{1-p_1} = \frac{31/41}{1-31/41} = \frac{31/41}{(41-31)/41} = \frac{31}{10} = 3.1$$

Odds for patienter uden involvering:

$$odds_0 = \frac{p_0}{1-p_0} = \frac{122}{217} = 0.56$$

Odds ratio:

$$\text{OR} = \frac{odds_1}{odds_0} = \frac{3.1}{0.56} = 5.51$$

- Odds for gennemtrængning af prostatakapslen er 5.5 gange større for patienter med involvering end for patienter uden (95% CI 2.61 til 11.63)

Binær kovariat og binært outcome

Logistisk regression

Formålet med logistisk regression

For et **binært outcome**, e.g.

$$Y_i = \begin{cases} 0 & \text{ej gennemtrængning} \\ 1 & \text{gennemtrængning} \end{cases}$$

at beskrive sammenhængen med forklarende variable for patient i.

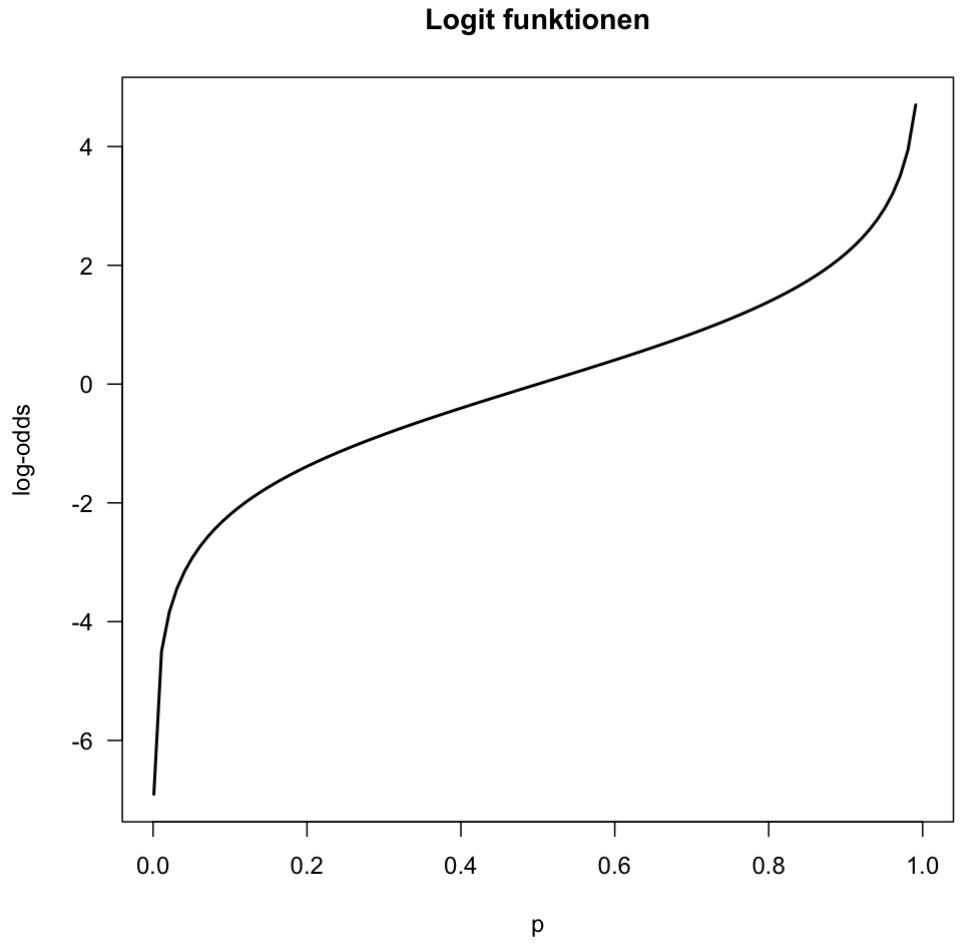
Lad p_i angive sandsynligheden for gennemtrængning for patient i.

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: `Involvering` (0/1)

$$\log(\frac{p_i}{1-p_i}) = a + b \cdot \texttt{involvering}_i$$

$$= \begin{cases} a & \text{ej involvering} \\ a+b & \text{involvering} \end{cases}$$

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

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

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

OR fra logistisk regression

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

Effekten af involvering er givet ved b:

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

Dvs.

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

Output

Parameter estimater:

	Estimate	Std.Error	OR	Lower	Upper	Z.Value	p
(Intercept)	-0.576	0.113	0.562	0.450	0.702	-5.089	0
involvering	1.707	0.381	5.514	2.614	11.632	4.483	0

R / SAS / SPSS

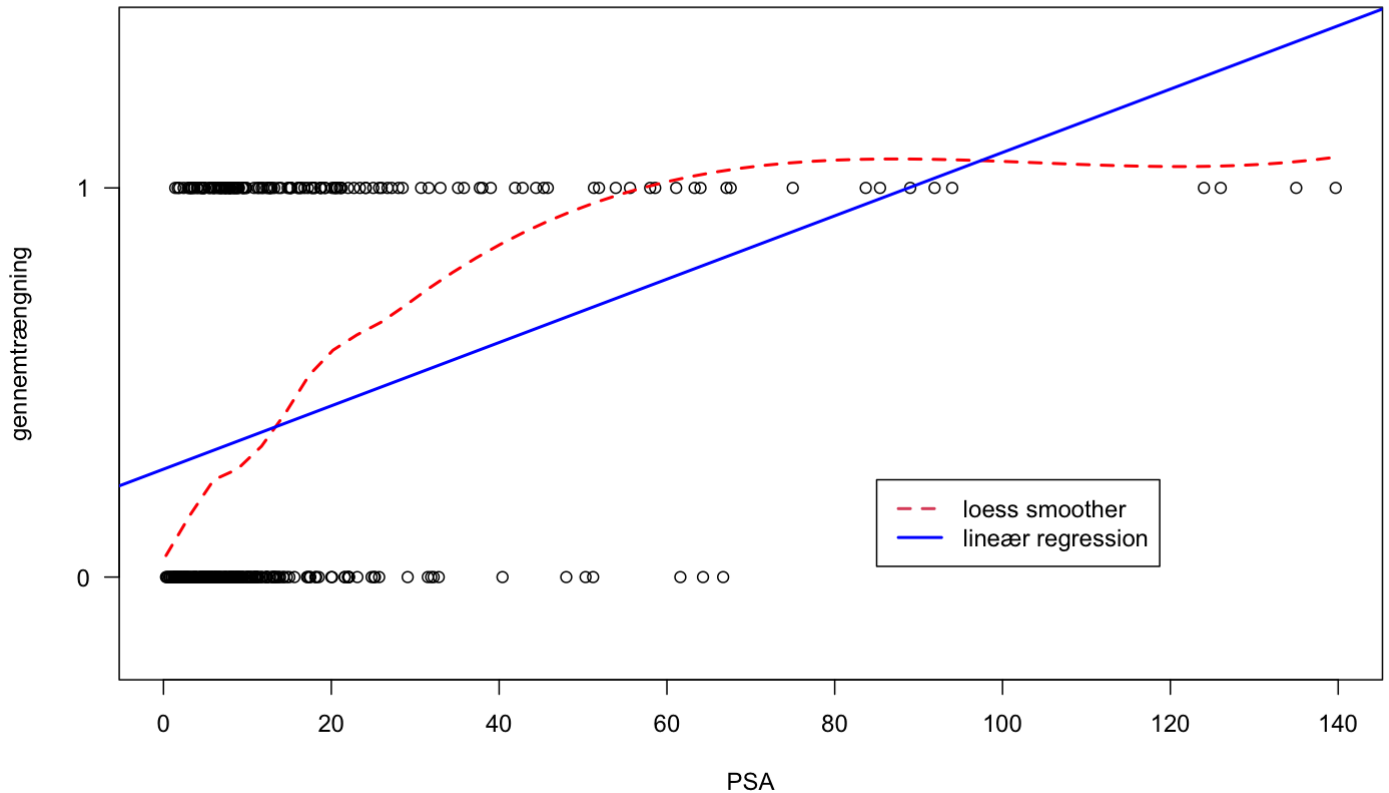
Quiz: Hvad er OR for ingen involvering vs involvering?

Kvantitativ kovariat og binært outcome

Logistisk regression

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 ng/mL PSA, feks 51 vs 50:

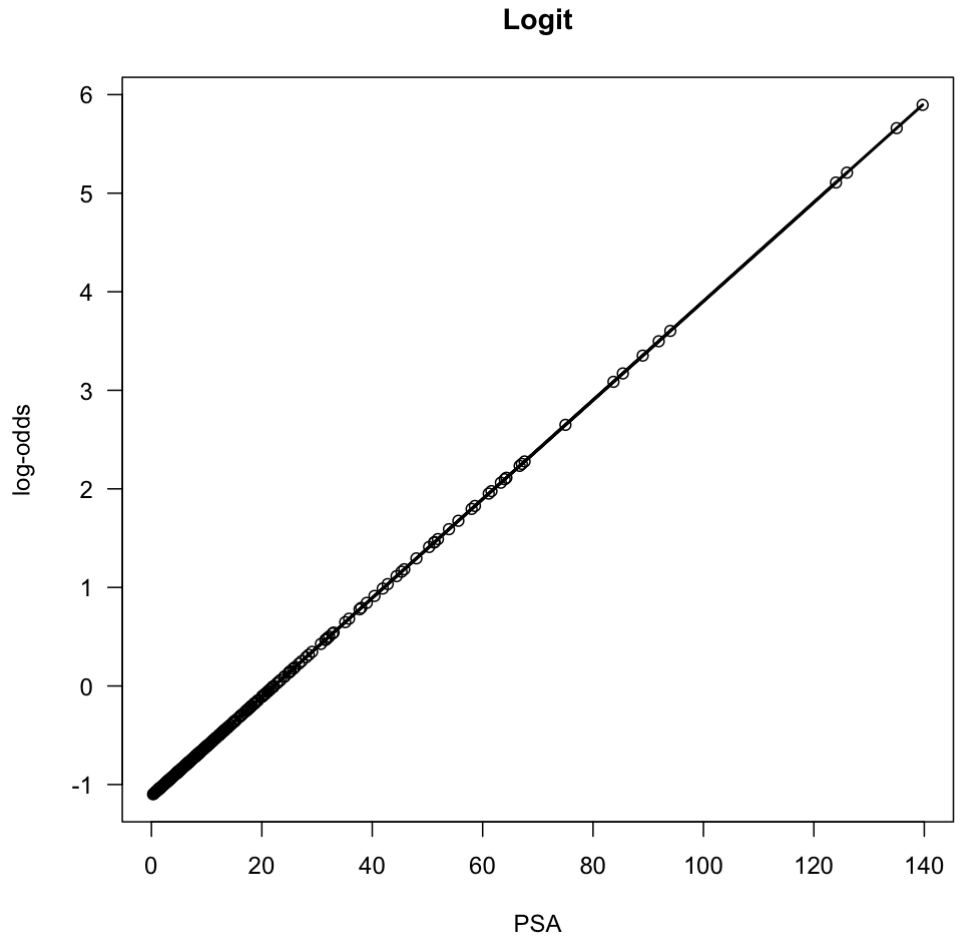
$$\text{OR} = \frac{\text{odds PSA} = 51}{\text{odds PSA} = 50}$$

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

Dvs.

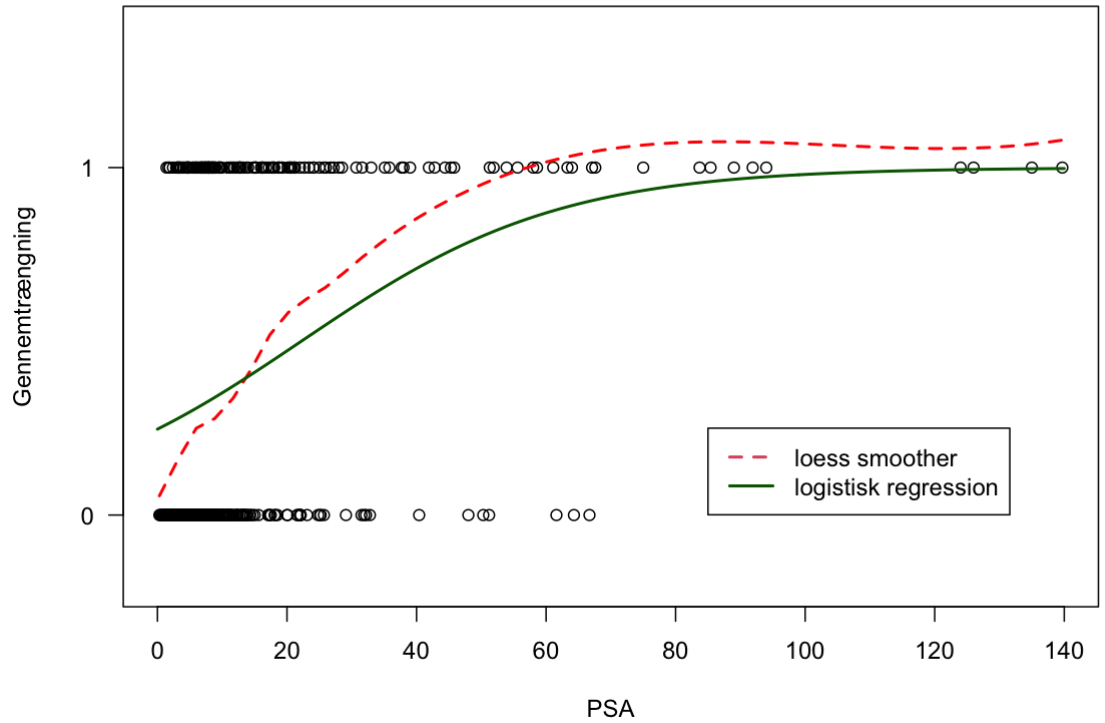
$$\text{OR} \;=\; \exp(b) \;=\; \exp(0.0502) \;=\; 1.0515.$$

Plot på log-odds-skala



Prædikterede sandsynligheder

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



Rapportering pr 10 enheder

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b \cdot \texttt{psa}_i, \qquad b = 0.0502$$

$$\text{OR} = \exp(0.0502) = 1.0513$$

Hvad er effekten af PSA pr 10 ng/mL?

Quiz: Man finder en OR på:

- 1.51
- 1.65
- 1.89
- 10.51

Modelkontrol

Modelkontrol

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

Numerisk modelkontrol:

- Overall goodness of fit: Hosmer-Lemeshow test

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 gennemtrængning
- I hver gruppe sammenlignes observeret og forventede antal:

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

- Størrelserne lægges 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.

R / SAS / SPSS

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 og SPSS giver her pga anden inddeling p=0.001...

Desuden kan det ændre sig, hvis man skifter til at se på det modsatte outcome altså `gennemtraengning01`=0 (i R er *p*=0.09 vs *p*=0.04).

Residualplot

Pearson residualer

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

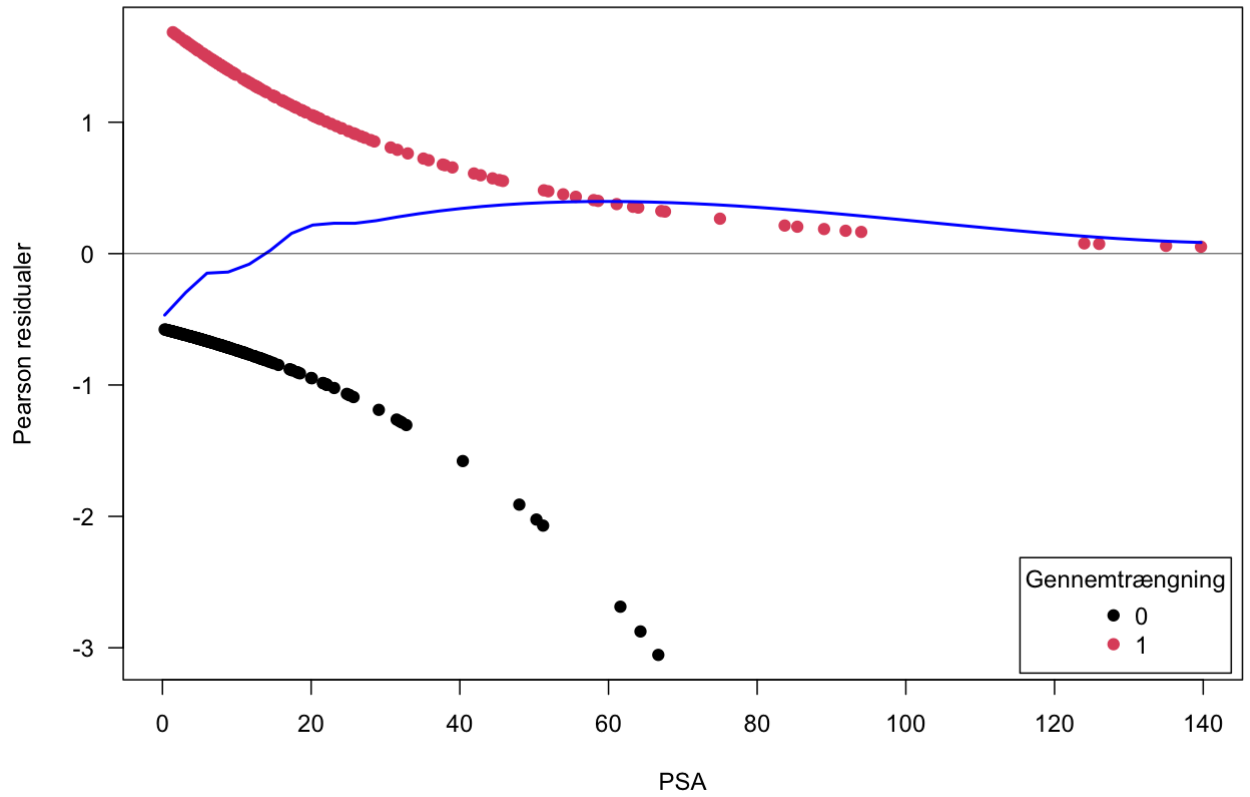
plottes vs

- kovariater
- fittede værdier

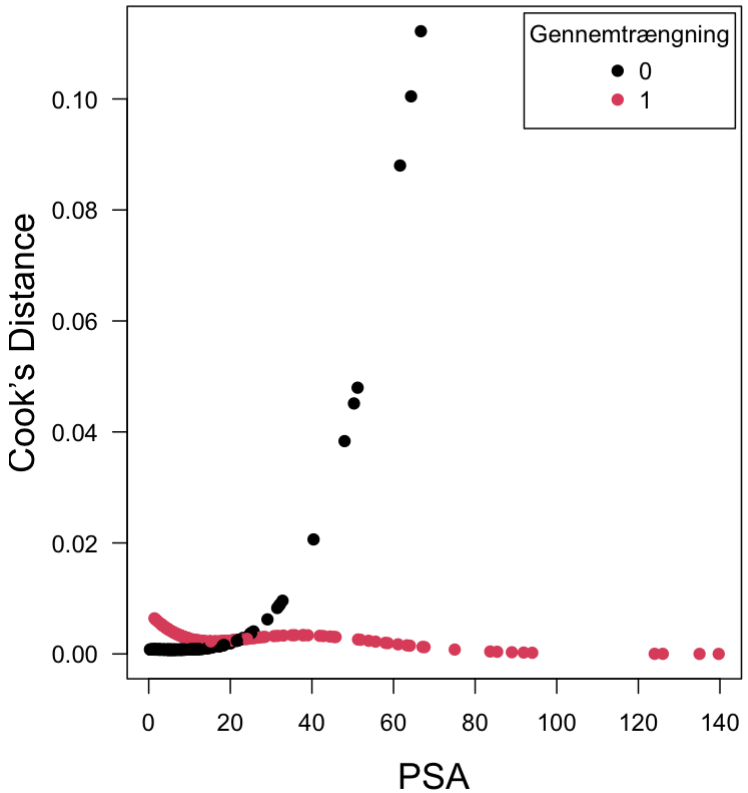
for at se efter krumninger som tegn på manglende linearitet.

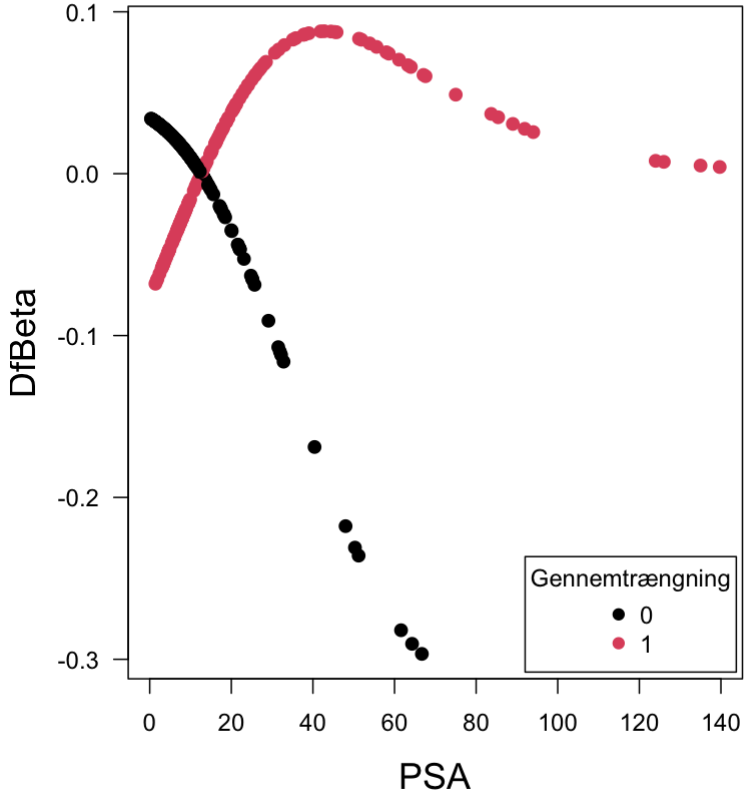
Residualplot

Bliver lidt skøre at se på ... R / SAS / SPSS



Diagnostics plots





Når linearitet ikke holder

Splines, log-transformation

Udgangspunktet

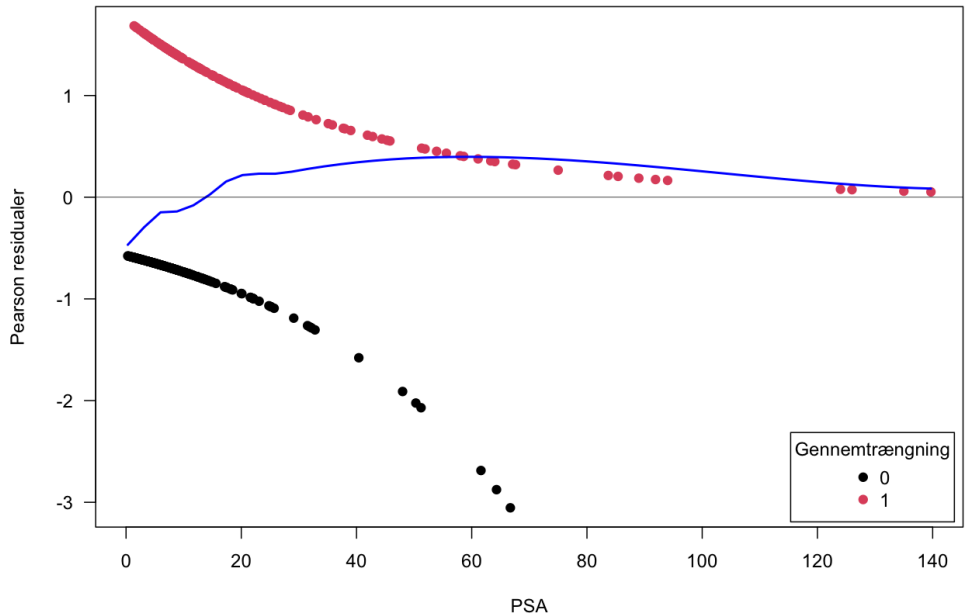
Lineær sammenhæng mellem PSA og log-odds for gennemtrængning:

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

$$b = 0.0502$$

OR per 10 PSA: $\exp(10 \times 0.0502) = 1.65$.

Hosmer-Lemeshow p=0.04 (0.001, 0.09 ...)



PSA som lineær spline

Tærskelværdier bestemmes ud fra kvartiler for PSA blandt cases (`gennemtraengning01=1`):

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

Splinevariable:

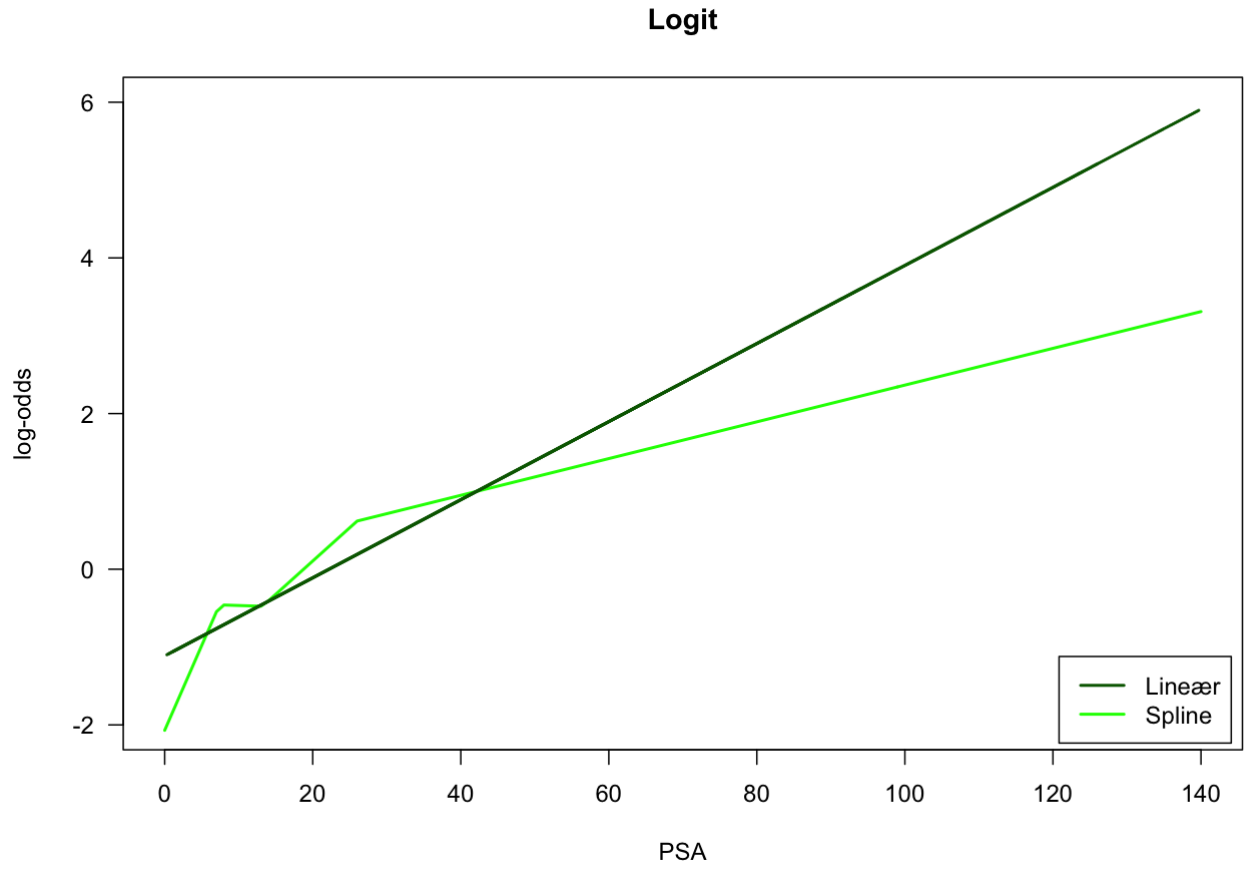
$$\texttt{psa\_spline1} = \begin{cases} \texttt{psa} - 7.4 & \texttt{psa} > 7.4 \\ 0 & \texttt{psa} \leq 7.4 \end{cases}$$

Tilsvarende defineres `psa_spline2` og `psa_spline3`

Output

	Estimate	Std.Error	OR	Lower	Upper	Z.Value	p
(Intercept)	-2.070	0.465	0.126	0.051	0.314	-4.450	0.0000
psa	0.218	0.082	1.243	1.059	1.459	2.666	0.0077
psa_spline1	-0.221	0.135	0.802	0.615	1.045	-1.634	0.1023
psa_spline2	0.089	0.103	1.093	0.892	1.338	0.857	0.3915
psa_spline3	-0.062	0.050	0.940	0.853	1.036	-1.244	0.2134

Effekt af PSA som lineær spline



Test af linearitet: H_0 : psa_spline1 = psa_spline2 = psa_spline3=0 giver p=0.057 (df=3)

Rapportering af splinemodellen

Med en lineær spline afhænger OR af PSA-værdierne:

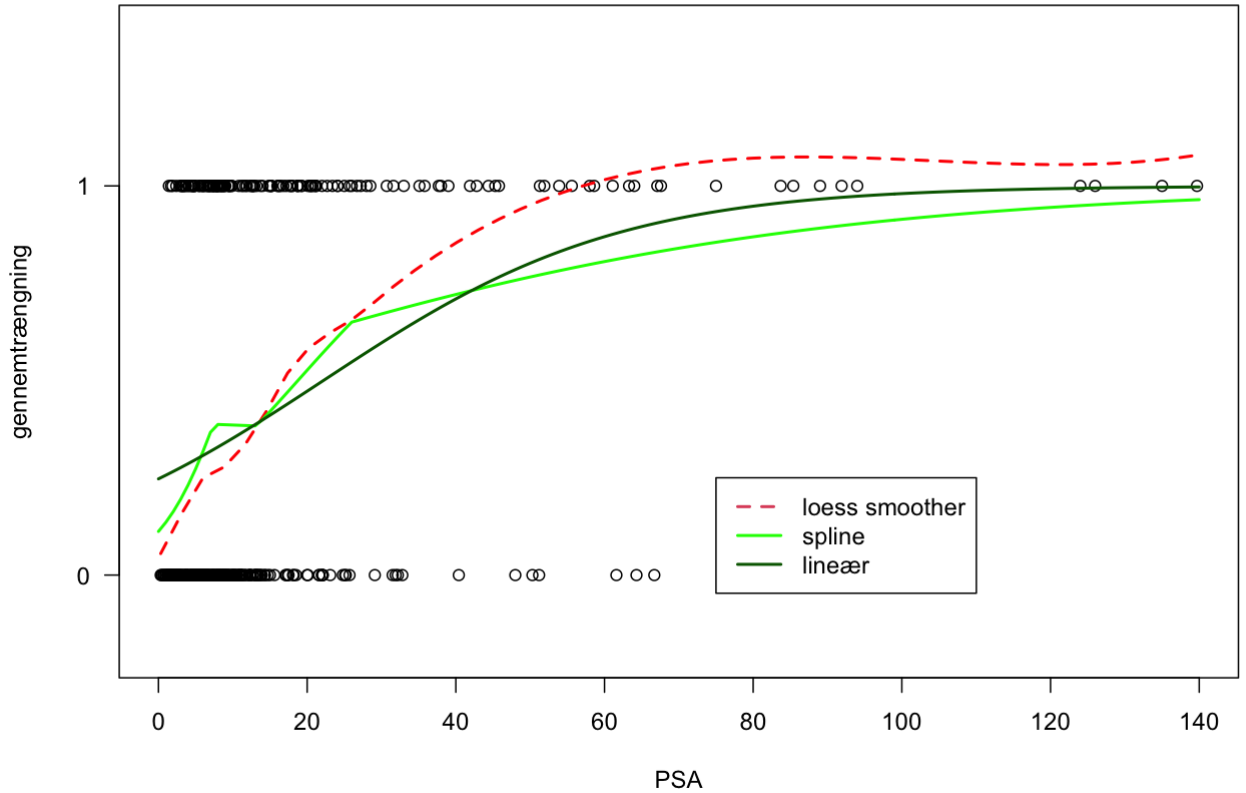
I intervallerne er

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 ng/mL PSA.

Prædikterede sandsynligheder

Hosmer-Lemeshow p=0.50.



Test for lineær effekt af log(PSA)

Definér ny variabel: `log2psa`$= \log_2$ (`psa`)

Model:

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

	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 $\log_2$(PSA)

Model:

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

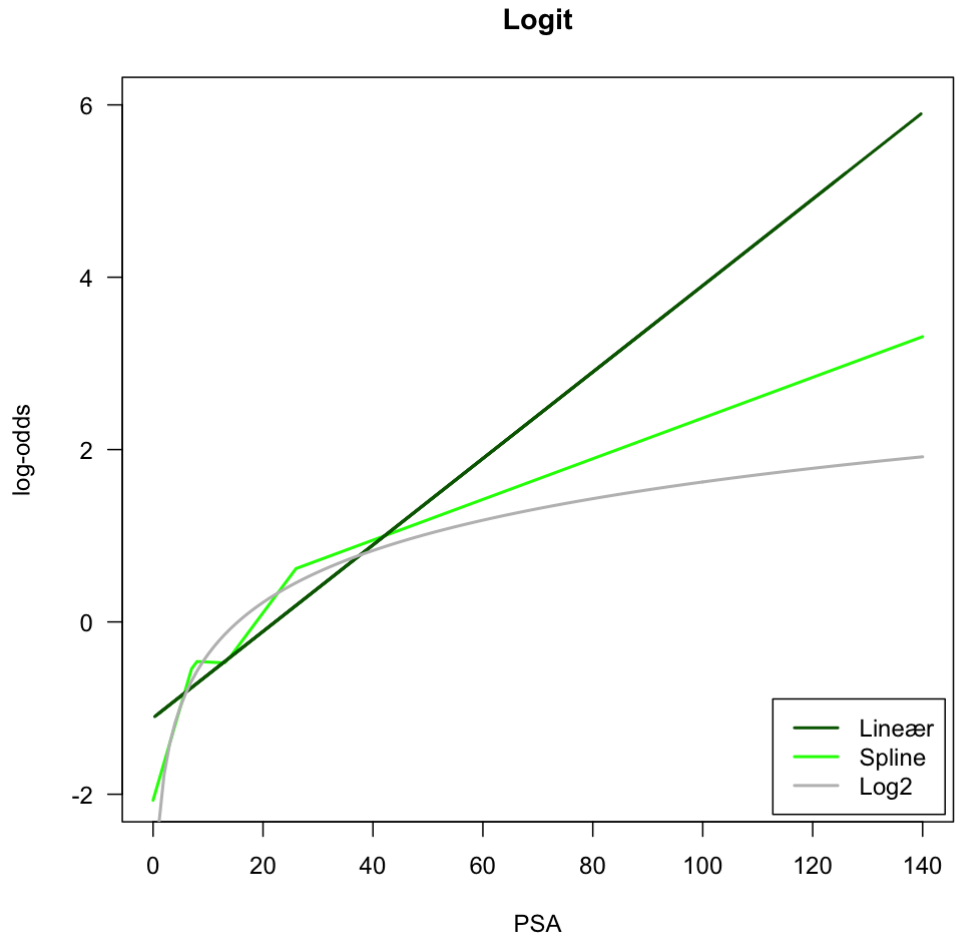
	Estimate	Std.Error	OR	Lower	Upper	Z.Value	p
(Intercept)	-2.376	0.326	0.093	0.049	0.176	-7.297	0
log2psa	0.602	0.091	1.826	1.528	2.182	6.623	0

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

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

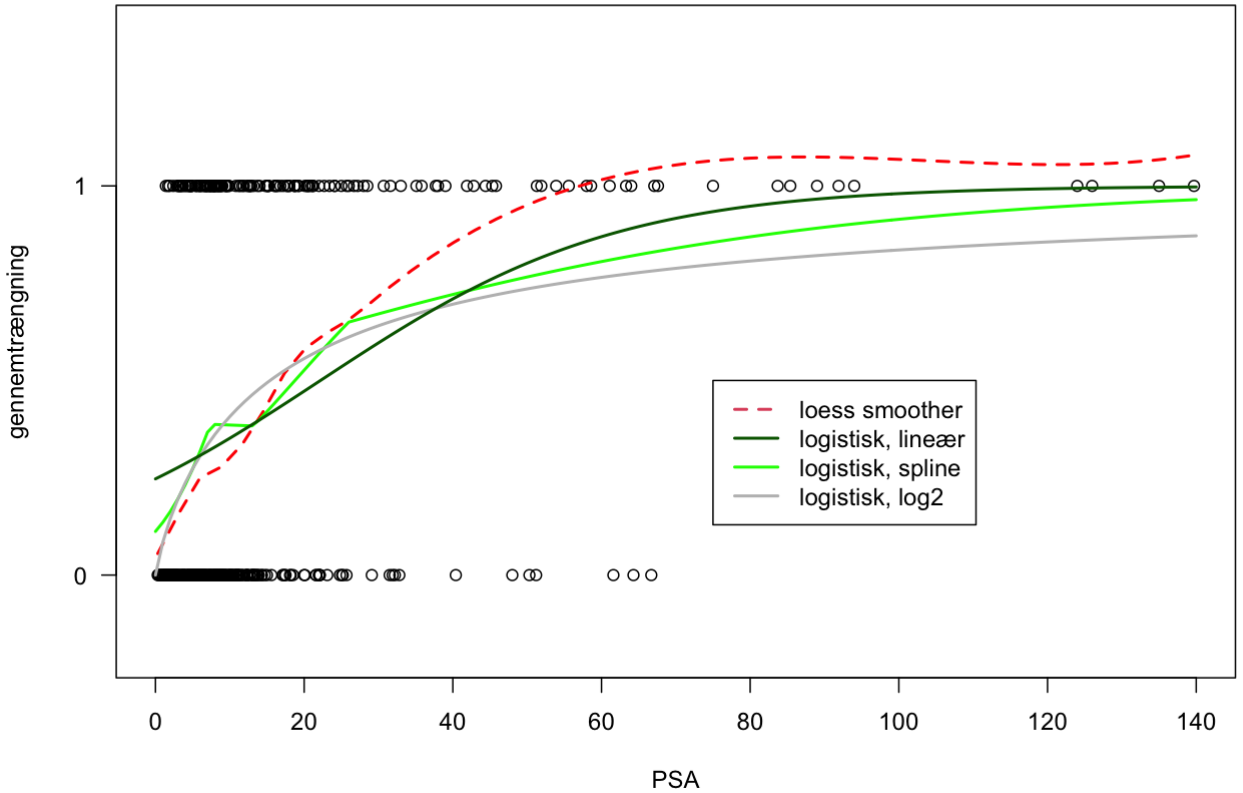
... er OR = exp(0.602) = 1.83 (95% CI 1.53 til 2.18)
... øges odds for gennemtrængning med 83% (95% CI 53 til 118%)

Effekt af $\log_2$(PSA) på log-odds-skala



Prædikterede sandsynligheder

Hosmer-Lemeshow p=0.17.



En multivariabel model

En kvalitativ og en kvantitativ kovariat

En kvalitativ og en kvantitativ kovariat

Model:

$$\log(\frac{p_i}{1-p_i}) = a + b \cdot \texttt{involvering}_i + c \cdot \log_2(\texttt{psa}_i)$$

$$= \begin{cases} a + c \cdot \log_2(\texttt{psa}_i) & \text{hvis involvering}_i = 0 \\ a + b + c \cdot \log_2(\texttt{psa}_i) & \text{hvis involvering}_i = 1 \end{cases}$$

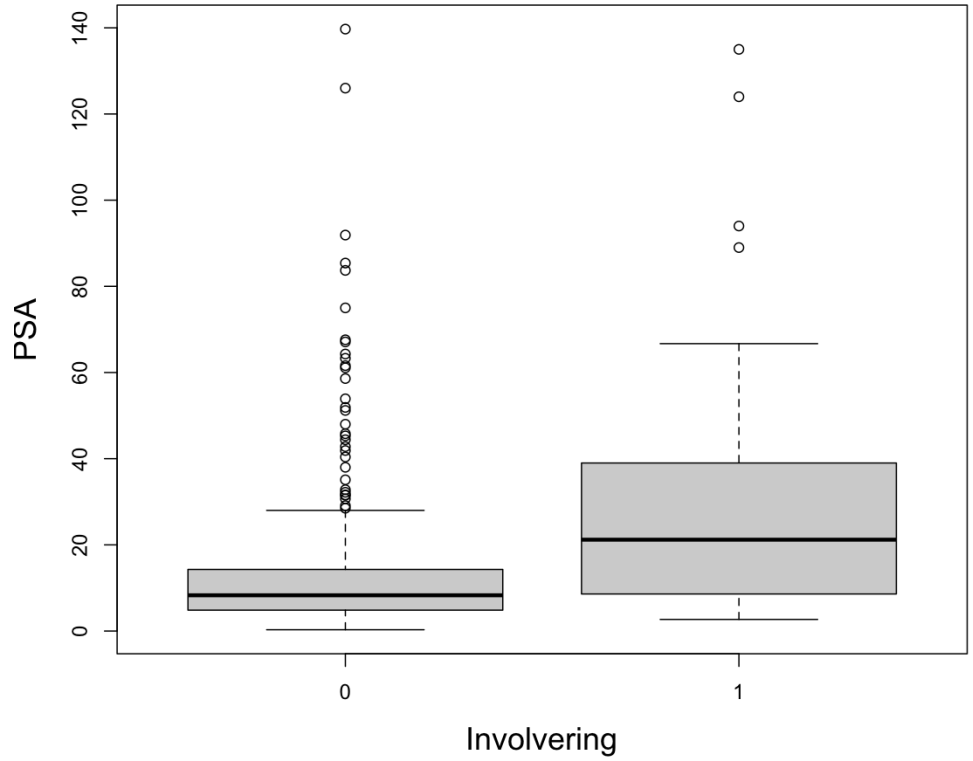
Hvad forventer vi sker med OR for involvering (ujusteret OR=5.51)?

Confounding?

Er der en association mellem PSA og involvering?

Median 8.3 vs 21.2

Wilcoxon: $p<.0001$



Output

	Estimate	Std.Error	OR	Lower	Upper	Z.Value	p
(Intercept)	-2.324	0.328	0.098	0.051	0.186	-7.094	0.000
involvering	1.253	0.405	3.501	1.582	7.748	3.091	0.002
log2psa	0.549	0.093	1.731	1.443	2.076	5.916	0.000

For fastholdt PSA:

OR = 3.50 for involvering vs ingen involvering (95% CI 1.58-7.75)

For fastholdt involvering:

OR = 1.73 ved en fordobling af PSA (95%CI 1.44-2.08).

Heterogenitet

Inhomogene populationer

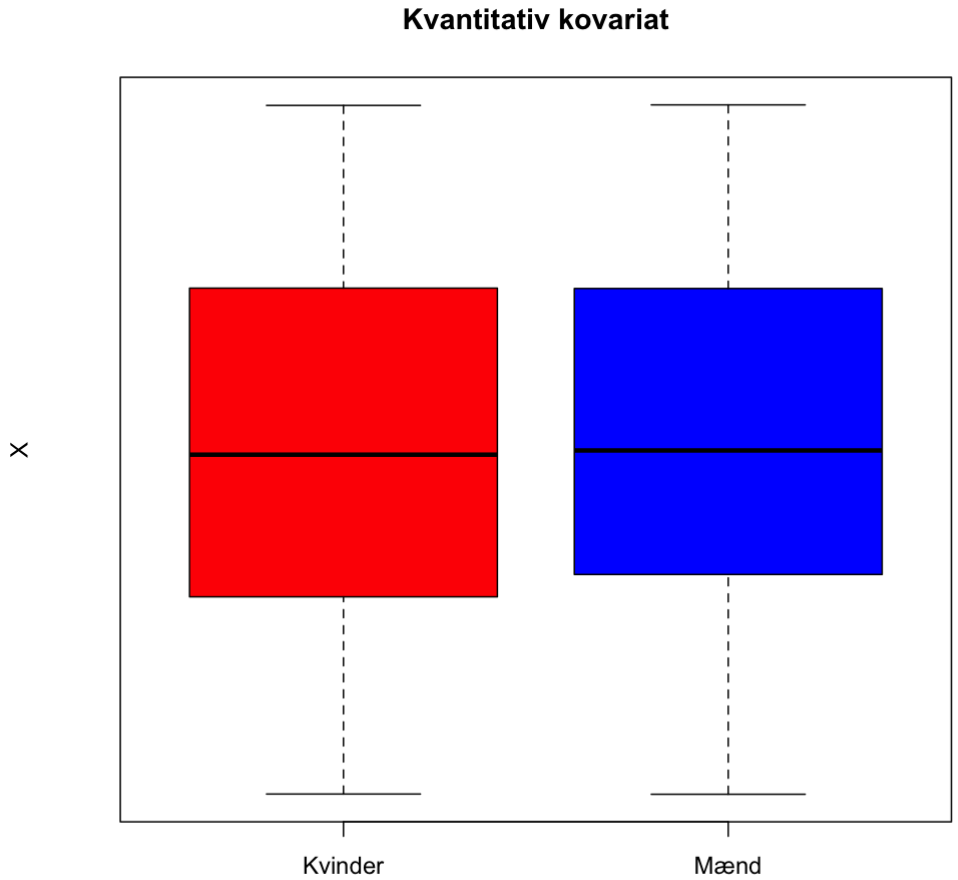
Fiktivt eksempel

Hvornår skal man justere?

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

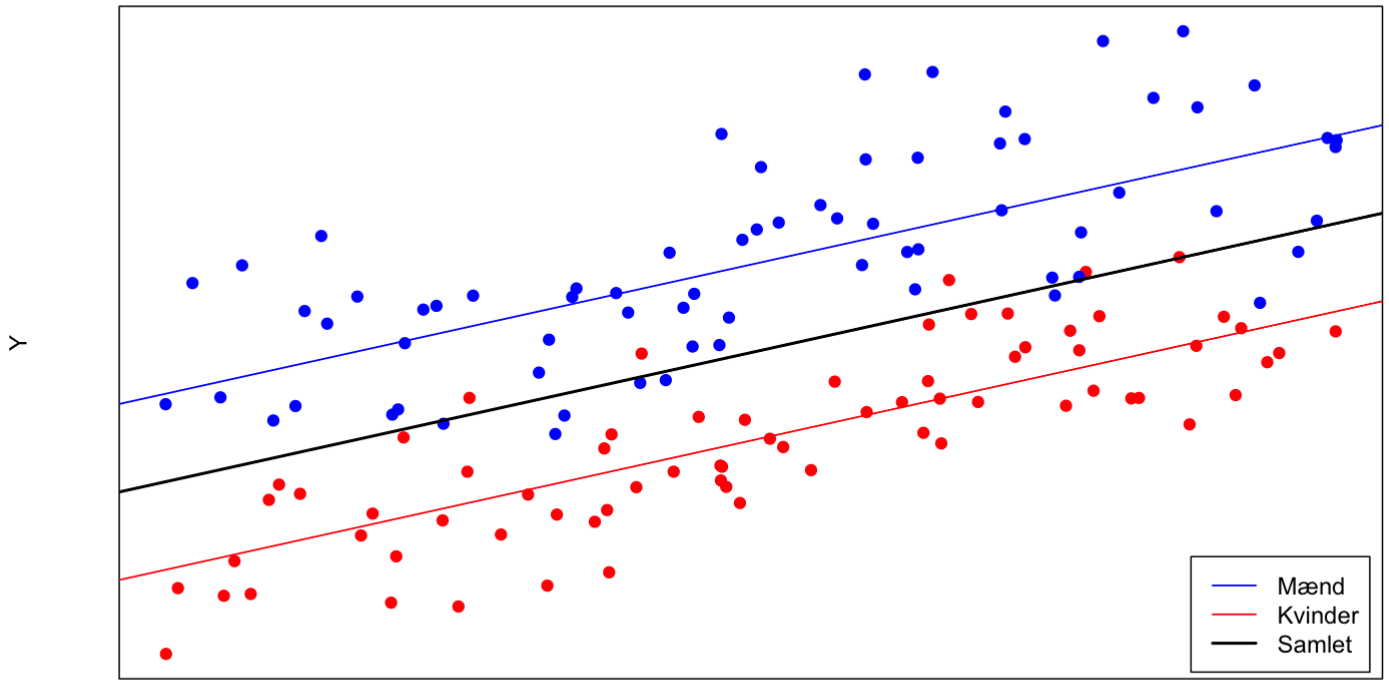
Både køn og X er prædiktive for Y, men der er ***ingen*** association mellem køn og X (dvs. ingen confounding).

X vs køn

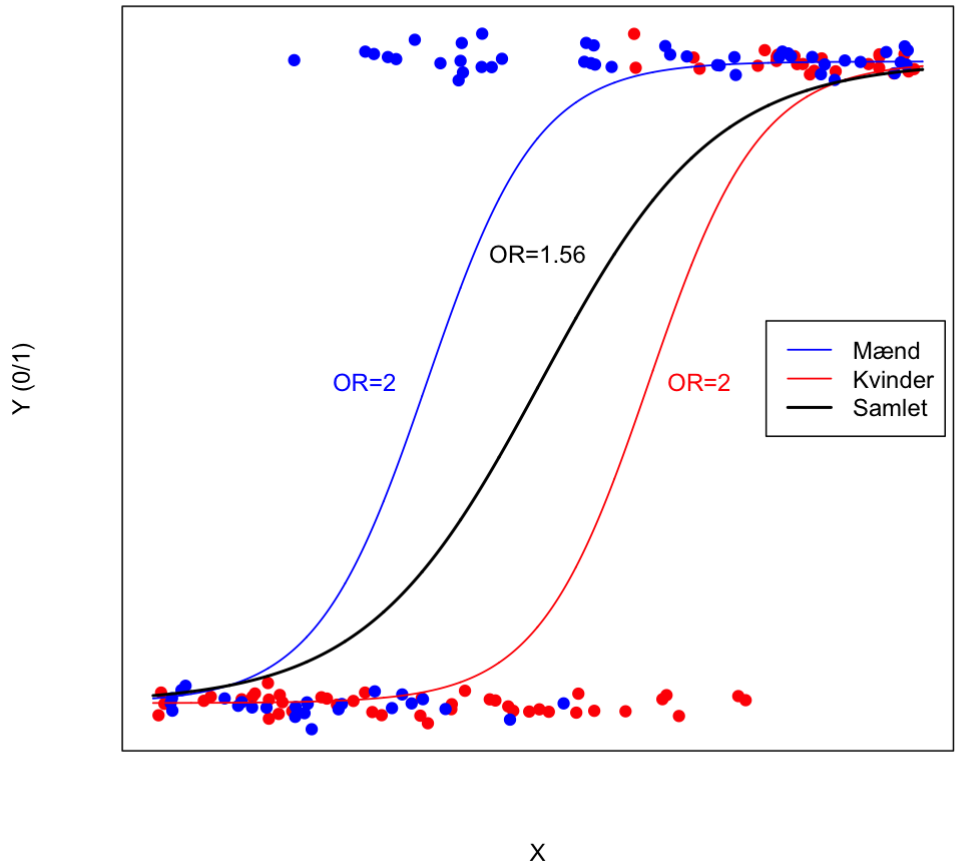


Tilbageblik på lineær regression

Her er Y kvantitativ:



Logistisk regression m/u betydende kovariat



Randomiserede studier

Her er eksponeringen randomiseringsgruppen (e.g. "Kontrol"/"Intervention").

Mænd OR = 2.67:

	p	*odds*
Kontrol	0.20	0.250
Intervention	0.40	0.667

Kvinder OR = 2.67:

	p	*odds*
Kontrol	0.60	1.5
Intervention	0.80	4

Alle (50% M/K) OR=2.25

	p	*odds*
Kontrol	0.40	0.667
Intervention	0.6	1.5

Subpopulationernes OR er altid større end OR for hele populationen.

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

Skal vi have alle kovariater med?

- 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

Interaktion: Kvantitativ og kvalitativ

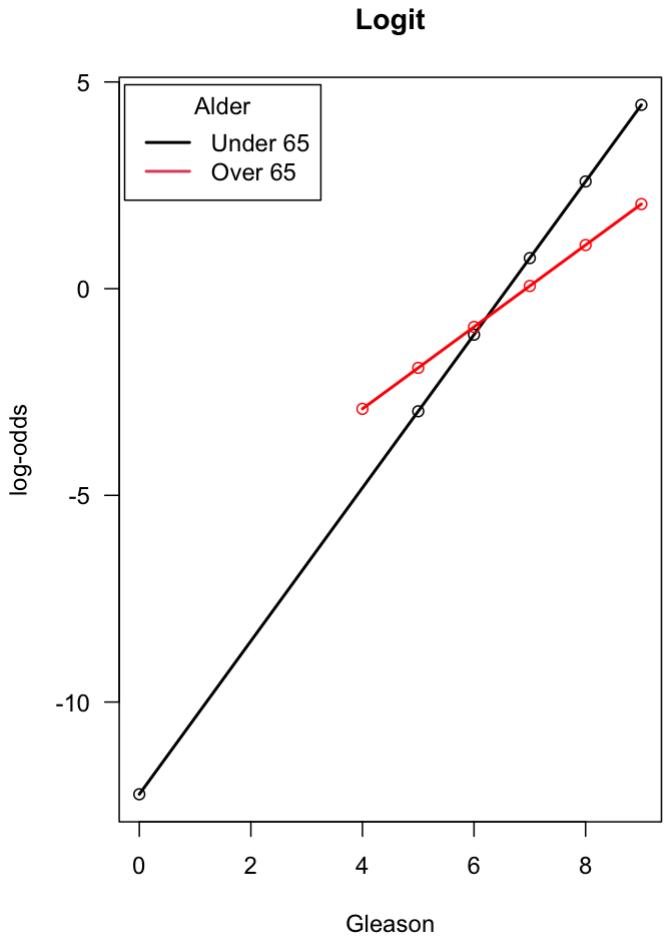
Gennemtrængning vs Gleason:

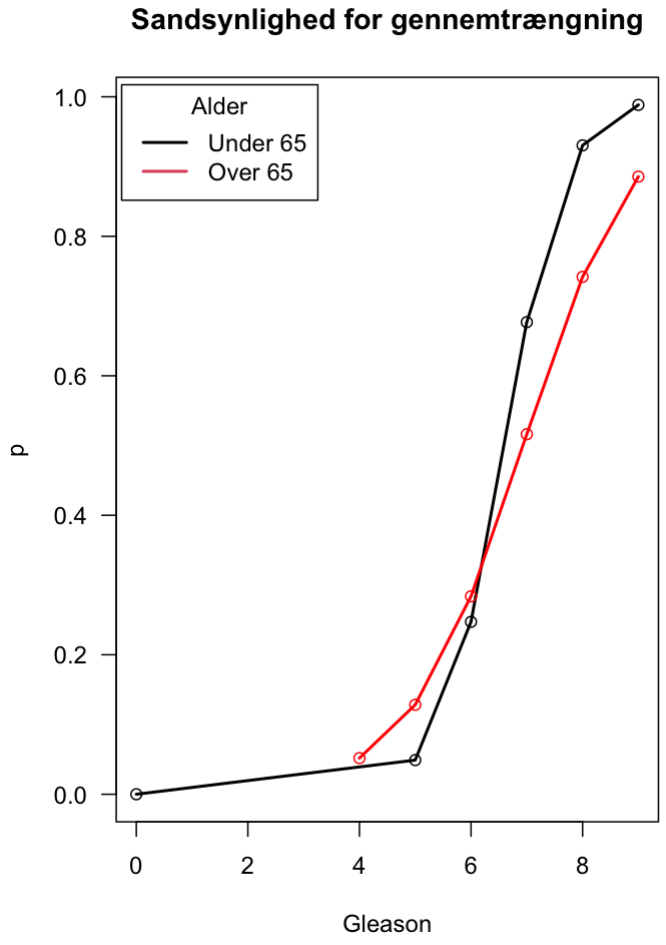
	gleason	Antal	antal0	antal1	Andele	andel0	andel1
1	0		2	0		1.00	0.00
2	4		1	0		1.00	0.00
3	5		61	6		0.91	0.09
4	6		101	38		0.73	0.27
5	7		55	73		0.43	0.57
6	8		6	24		0.20	0.80
7	9		1	12		0.08	0.92

Har Gleason score forskellig betydning for unge og ældre? Muligvis - vi har en *p*-værdi på 0.02:

	Estimate	Std. Error	z value	p
(Intercept)	-6.869	1.171	-5.866	0.0000
alder65Under	-5.361	2.348	-2.283	0.0224
gleason	0.991	0.176	5.630	0.0000
alder65Under:gleason	0.862	0.362	2.383	0.0172

Prædikterede sandsynligheder





Rapportering af modellen

Vi finder 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	OR	Lower	Upper	Z.Value	p
(Intercept)	-6.869	1.171	0.001	0.000	0.010	-5.866	0.0000
alder65Under	-5.361	2.348	0.005	0.000	0.468	-2.283	0.0224
alder65Over:gleason	0.991	0.176	2.693	1.907	3.802	5.630	0.0000
alder65Under:gleason	1.853	0.316	6.378	3.432	11.854	5.859	0.0000

- For de **yngre**: OR = 6.37 pr 1 i Gleason (95% CI 3.43-11.85)
- For de **ældre**: OR = 2.69 pr 1 i Gleason (95% CI 1.91-3.80)

To kvalitative med interaktion

Andele med gennemtrængning:

PSA	Ej involvering	Involvering
< 20	86/282 (46%)	13/19 (68%)
>= 20	36/57 (63%)	18/22 (82%)

Log-odds for gennemtrængning:

PSA	Ej Involvering	Involvering
< 20	-0.82	0.77
>= 20	0.54	1.50

Odds for gennemtrængning:

PSA	Ej involvering	Involvering
< 20	86/196 (0.44)	13/6 (2.17)
>= 20	36/21 (1.71)	18/4 (4.50)

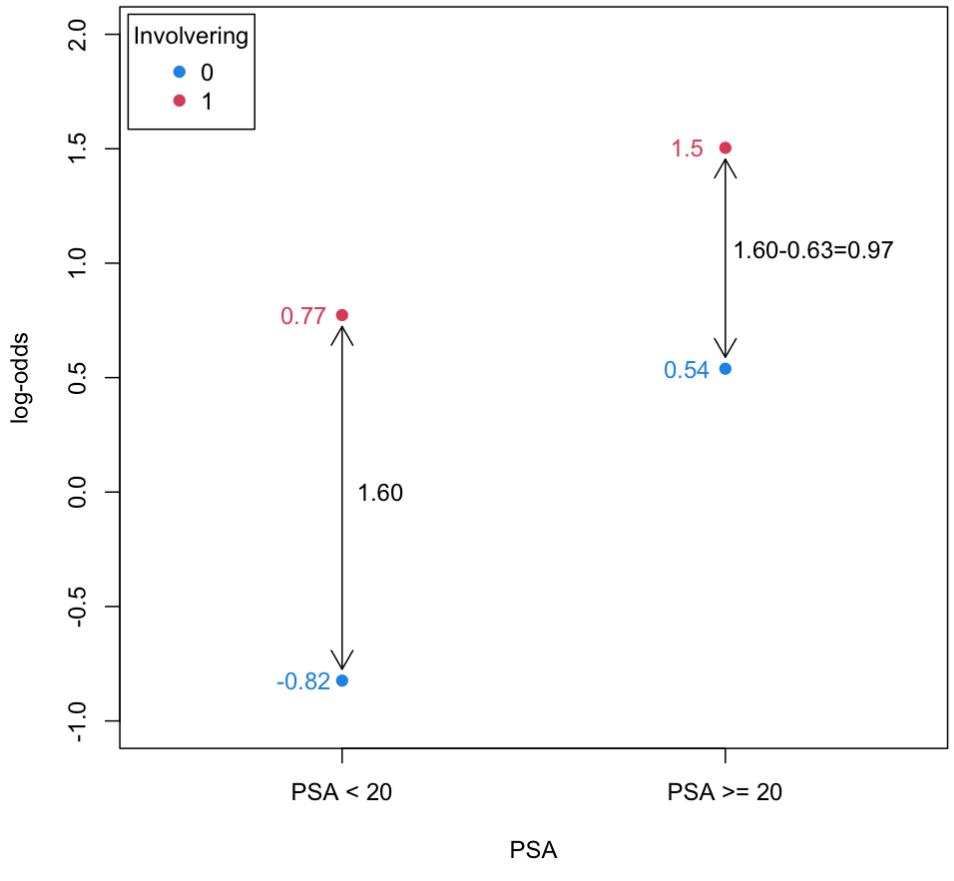
```
                            Estimate
(Intercept)                  -0.8238
involvering                   1.5970
psa01 >= 20                   1.3628
involvering:psa01 >= 20      -0.6319
```

Illustration

```
                            Estimate Pr(>|z|)
(Intercept)                  -0.8238   0.0000
involvering                   1.5970   0.0017
psa01 >= 20                   1.3628   0.0000
involvering:psa01 >= 20      -0.6319   0.4301
```

Log-odds for gennemtrængning:

PSA	Ej Involvering	Involvering
< 20	-0.82	0.77
>= 20	0.54	1.50



Rapportering interaktionsmodel

Effekten af involvering:

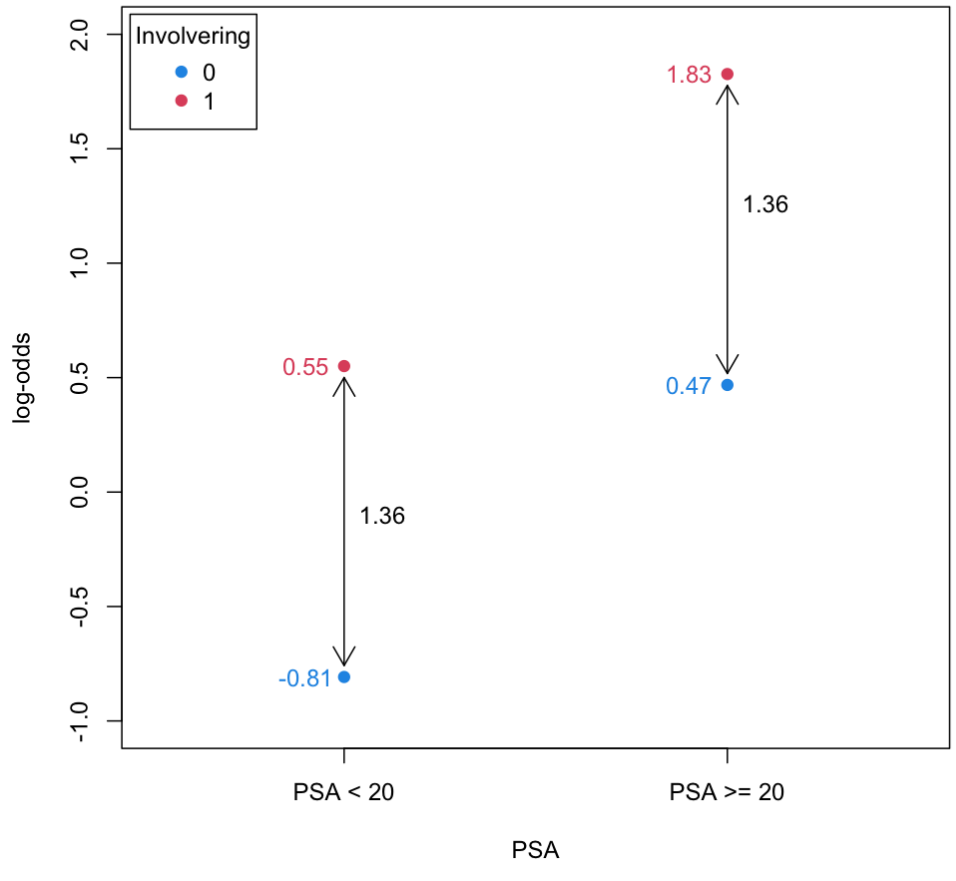
- For patienter med PSA < 20: OR = exp(1.60) = 4.94 (95%CI 1.82-13.42)
- For patienter med PSA ≥ 20: OR = exp(1.60-0.63) = 2.63 (95%CI 0.78-8.80)

Effekten af PSA over 20:

- For patienter uden involvering: OR = exp(1.36) = 3.91 (95% CI 2.16-7.08)
- For patienter med involvering: OR = exp(1.36-0.63) = 2.08 (95% CI 0.49-8.88)

To kvalitative uden interaktion

	Estimate	Std. Error
(Intercept)	-0.8079	0.1274
involvering	1.3584	0.3982
psa01 PSA >= 20	1.2760	0.2809



Prædiktion

En multivariabel model

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

	Estimate	Std.Error	OR	Lower	Upper	Z.Value	p
(Intercept)	-6.834	1.231	0.001	0.000	0.012	-5.552	0.0000
involvering	0.618	0.454	1.855	0.762	4.517	1.360	0.1738
log2psa	0.316	0.108	1.372	1.111	1.694	2.940	0.0033
alder65Under	-5.031	2.369	0.007	0.000	0.678	-2.124	0.0337
gleason	0.690	0.189	1.994	1.378	2.886	3.661	0.0003
knudeBegge	1.390	0.448	4.013	1.667	9.658	3.101	0.0019
knudeHoejre	1.463	0.374	4.321	2.076	8.993	3.913	0.0001
knudeVenstre	0.721	0.358	2.057	1.019	4.148	2.014	0.0440
alder65Under:gleason	0.796	0.365	2.218	1.084	4.538	2.180	0.0293

Hosmer-Lemeshow goodness-of-fit p=0.25

- Ved involvering øges odds for gennemtrængning med 86% (95% CI -24 til 352%) 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** (`pred`=1) hvis $\hat{p}_i > 0.5$
- person i som **kontrol** (`pred`=0) hvis $\hat{p}_i \leq 0.5$

	Gennemtrængning	
Prædiktion	0	1
0	191	50
1	36	103

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

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

ROC kurve

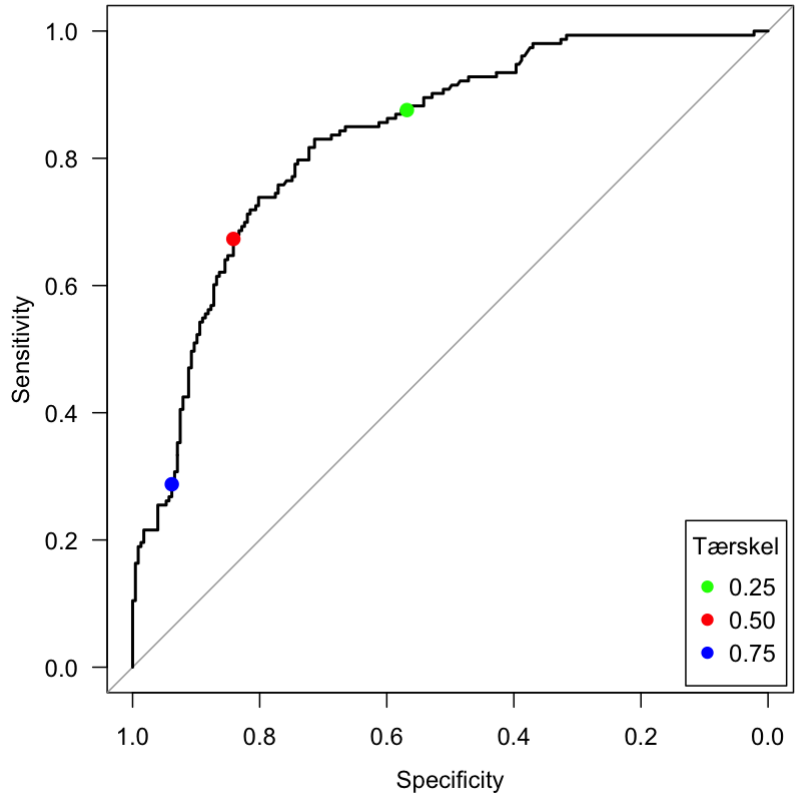
Receiver Operating Characteristics kurve: R / SAS / SPSS

Tærskelværdi p=0.25:

	Gennemtrængning	
Prædiktion	0	1
0	129	19
1	98	134

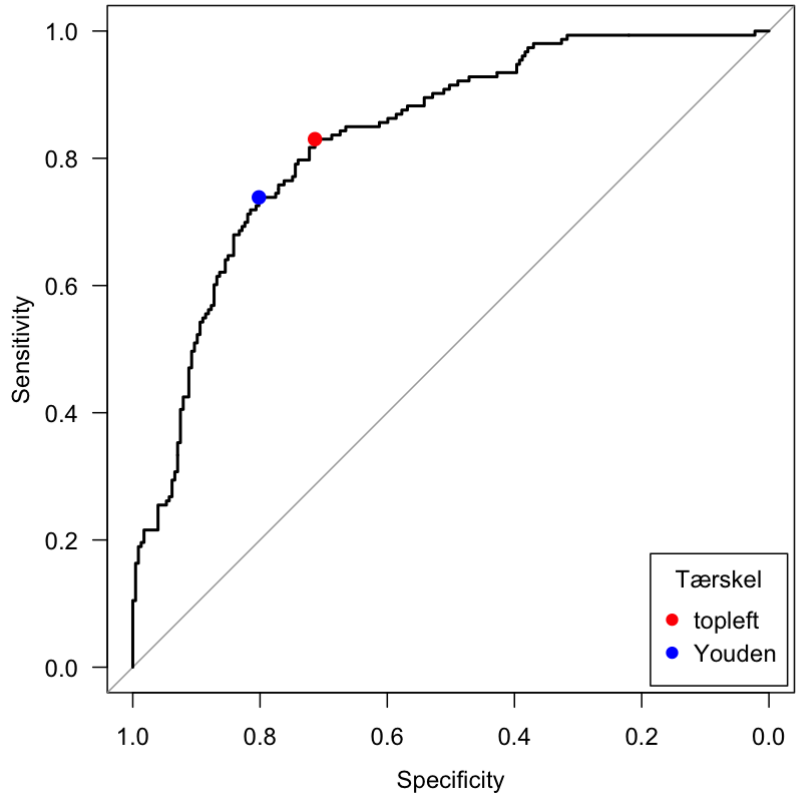
Tærskelværdi p=0.75:

	Gennemtrængning	
Prædiktion	0	1
0	213	109
1	14	44



Optimal tærskelværdi

Closest top-left: 0.412 Youden: 0.323



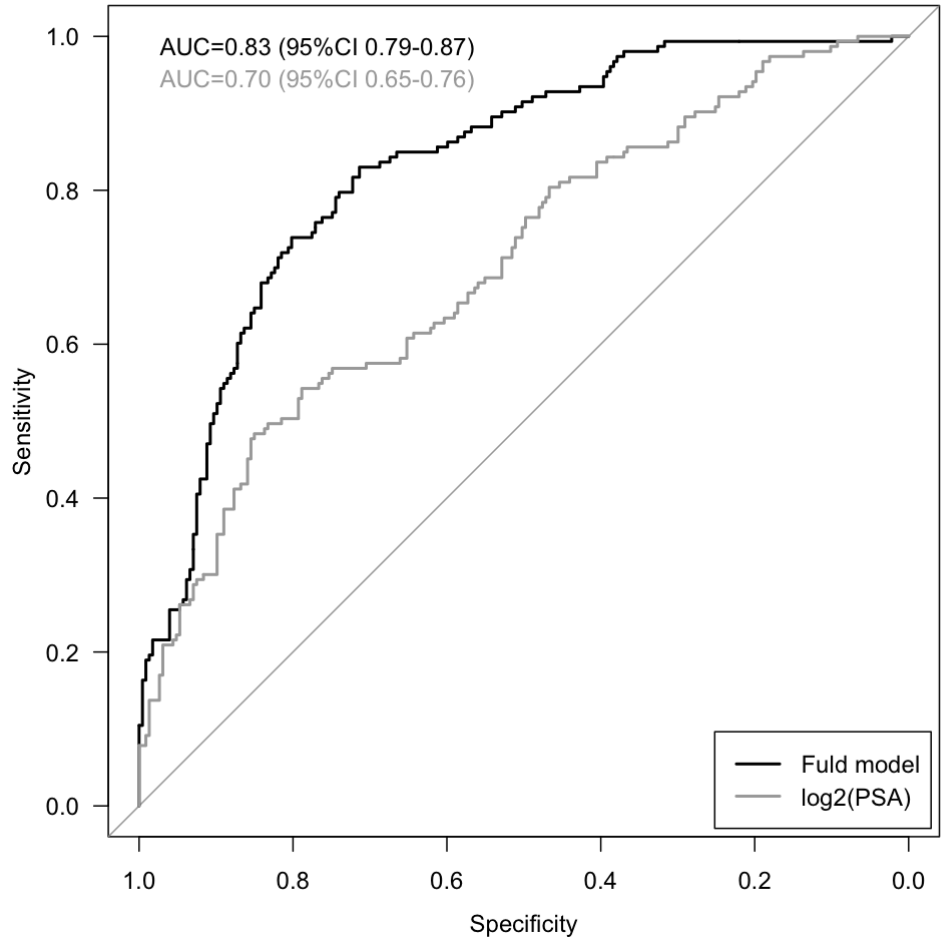
AUC

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)

Sammenligning af prædiktionsmodeller



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 koefficienterne 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*