

Uit

NORGES
ARKTISKE
UNIVERSITET

Regresjonsmodeller

HEL – 8020 Analyse av registerdata i forskning

Tom Wilsgaard



Intro

- Mye forskning innen medisin og helsefag dreier seg om å studere
 - assosiasjonen mellom en eller flere eksponeringsvariabler mot en responsvariabel
 - forskjell i respons mellom ulike grupper, justert for kovariater
- Regresjonsmodeller
- For eksempel
 - Er det sammenheng mellom fysisk aktivitet og brystkreft.
 - Finnes det nye biomarkører som kan predikere hjerteinfarkt uavhengig av de kjente kardiovaskulære risikofaktorene
 - Hvilke livsstilsfaktorer er assosiert med ideelle blodtrykksverdier.

Mål

- Introdusere regresjonsanalysemetoder i medisinsk forskning
 - Lineær regresjon
 - Logistisk regresjon
 - Cox regresjon
- Metodevalg avhenger i hovedsak av
 - Type avhengig variabel

Type avhengig variabel

- Måling på en kontinuerlig skala



lineær regresjon

- Binær respons (syk / ikke syk)



Logistisk regresjon

- Tid til hendelse / sykdom



Cox regresjon

Hva er av interesse

- Oftest er det flere uavhengige variable (prediktorer/kovariater) involvert
- Vi kan da være interessert i
 - Den kombinerte effekten av alle de uavhengige variablene
 - Test av modell, p-verdi
 - «goodnes-of-fit»
 - Forklart varians
 - Den individuelle effekten av hver enkelt uavhengig variabel
 - Mål på effekt
 - Signifikans
 - Rangering
 - Interaksjon og «confounding».

Regresjonsmodeller

Lineær regresjon

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \dots + \beta_m \cdot x_m + \varepsilon$$

Logistisk regresjon

$$P(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \dots + \beta_m \cdot x_m)}}$$

Cox regression

$$h(t) = h_0(t) \cdot e^{\beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \dots + \beta_m \cdot x_m}$$

Høyre side

- Variablene er
 - Et antall utvalgte variable
 - Eksponeringsvariable, uavhengige variable, kovariater, prediktorer, x-variable
 - Kan være av alle typer
 - Kontinuerlige
 - Nominale – kategoriske
 - Ordinale – kategoriske
 - De kategoriske kan modelleres som indikatorvariable (dummy).

Fra enkel til multivariabel modell

- Når flere x-variabler legges til
 - Modell fit vil bli bedre (R^2 , verdi av likelihoodfunksjonen)
 - Global test trenger ikke få lavere p-verdi
 - Alle variable er gjensidig justert for hverandre
 - Regresjonskoeffisientene blir ofte mindre (nærmere 0)

Fra enkel til multivariabel modell(2)

Eksempel, y = systolisk BT, n = 31 personer

Modell	x	R^2	Test av modell F (p-verdi)	β (p-verdi)
1	Vekt	0.249	9.64 (.004)	0.633 (.004)
2	Vekt Alder	0.280	5.43 (.010)	0.546 (.018) 0.211 (.288)
3	Vekt Alder Puls	0.385	5.64 (.004)	0.492 (.025) 0.163 (.386) 0.536 (.040)

Utskrift fra statistikkpakker

- Vurder
 - Modell-antakelser
 - Uvanlige observasjoner
 - Observasjoner med sterk innflytelse på modellen
 - Kolinearitet
 - Modelltilpasning
 - Global modell
 - Blokk av variable
 - Hver x-variabel

Enkel lineær regresjon

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

Y = response, avhengig variabel

X = uavhengig variabel

β_0 = intercept

β_1 = stigningstall

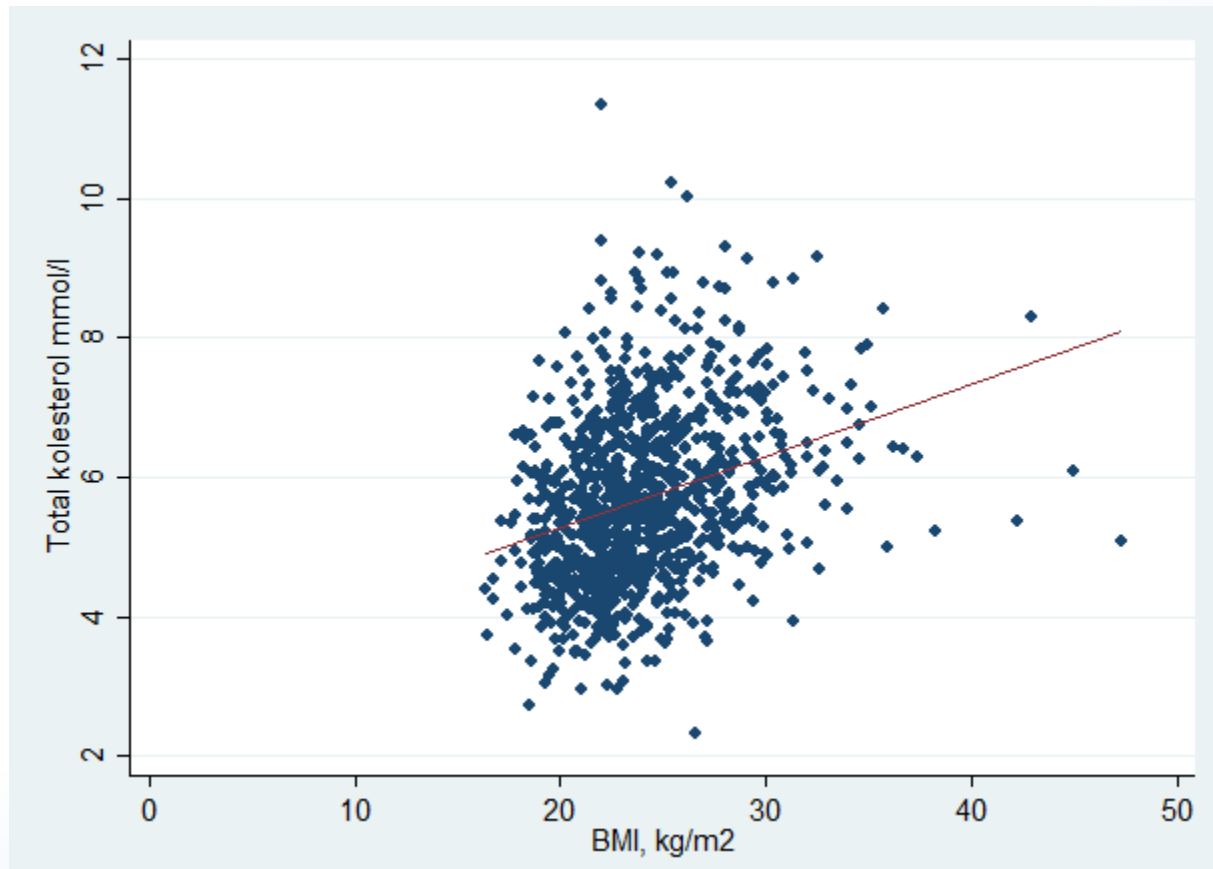
ε = residual

Modellantakelser

- Uavhengighet
- Linearitet
- Homoskedastisitet
- Normalitet

Enkel lineær regresjon (2)

- Data, kolesterol vs BMI: n=1000 personer, 20-61 år
- Vurder et spredningsplott
- Er det lineær sammenheng



Enkel lineær regresjon, utskrift

```
regress chol bmi
```

Source	SS	df	MS	Number of obs	=	999
-----				F(1, 997)	=	110.80
Model	143.73978	1	143.73978	Prob > F	=	0.0000
Residual	1293.38325	997	1.29727507	R-squared	=	0.1000
-----				Adj R-squared	=	0.0991
Total	1437.12303	998	1.44000303	Root MSE	=	1.139

chol	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

bmi	.1031713	.0098014	10.53	0.000	.0839377	.122405
_cons	3.219536	.2370584	13.58	0.000	2.754345	3.684726

Multivariabel lineær regresjon, utskrift

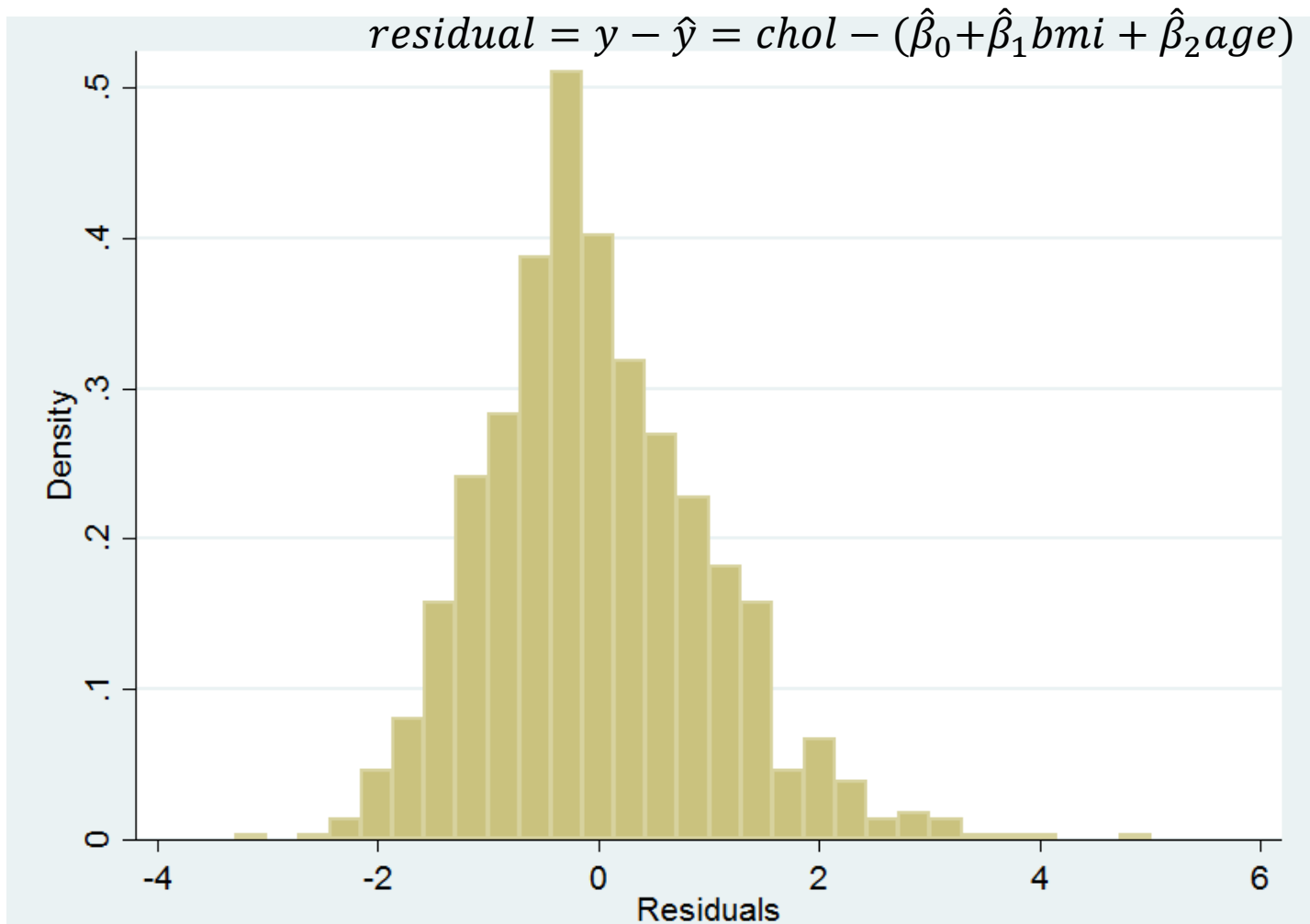
```
regress chol bmi age
```

Source	SS	df	MS	Number of obs	=	999
-----				F(2, 996)	=	177.59
Model	377.779272	2	188.889636	Prob > F	=	0.0000
Residual	1059.34375	996	1.06359815	R-squared	=	0.2629
-----				Adj R-squared	=	0.2614
Total	1437.12303	998	1.44000303	Root MSE	=	1.0313

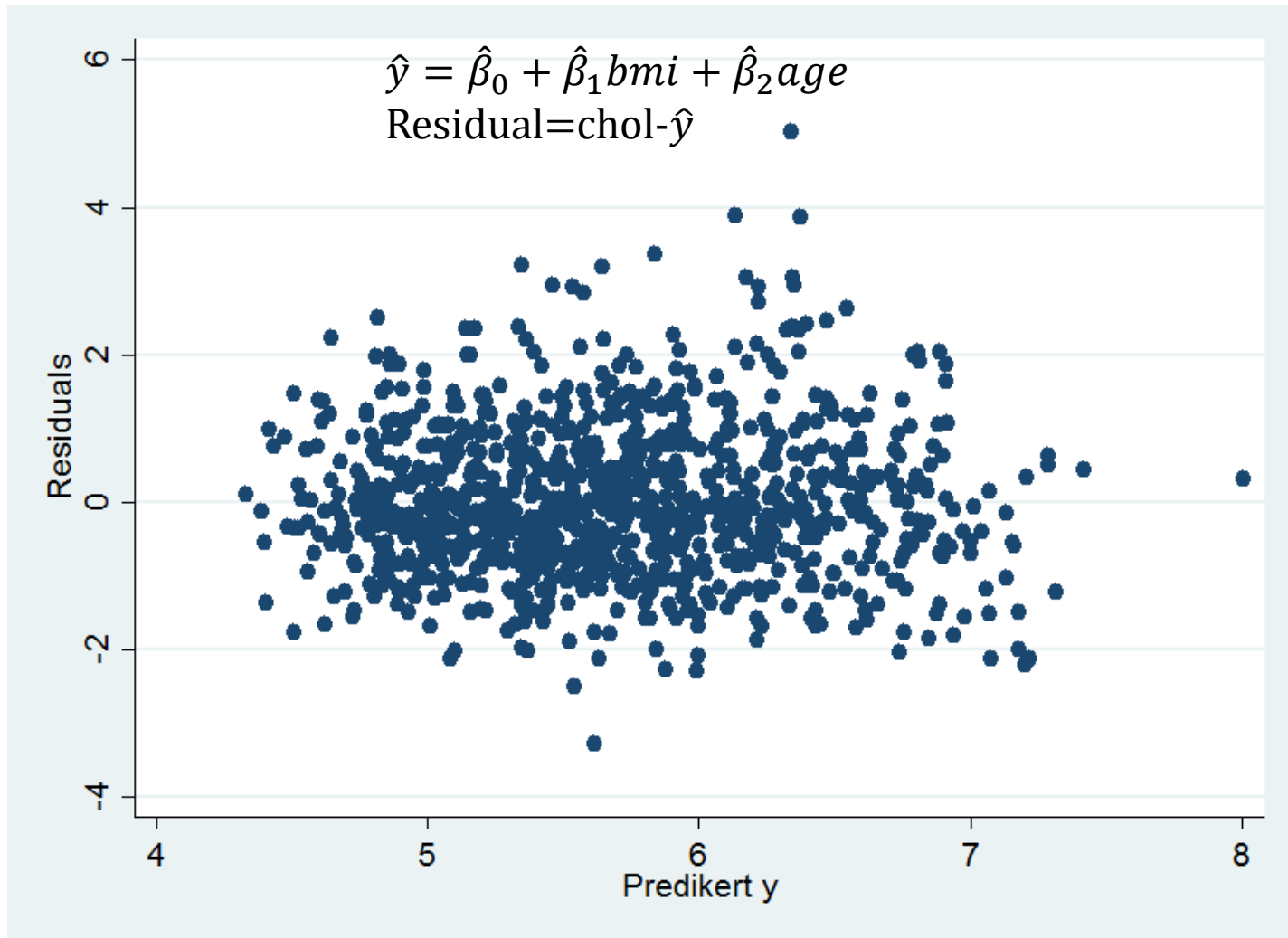
chol	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

bmi	.0611743	.0093155	6.57	0.000	.0428941	.0794545
age	.0487547	.0032867	14.83	0.000	.0423051	.0552044
_cons	2.417997	.2213453	10.92	0.000	1.98364	2.852353

Normalitetsantakelsen, histogram av residualene



Homoskedastisitet, plott residualene vs predikert



Logistisk regresjon

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

p	= P[y=1], sannsynlighet for y = 1
y	= respons, 1=syk; 0=ikke syk
x _i 'ene	= uavhengige variabler
β _i 'ene	= regresjons-koeffisienter
exp(β _i)	= odds ratio

Logistisk regresjon, utskrift

logit status smoke

status	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke	.6037231	.2056024	2.94	0.003	.2007497	1.006697
_cons	-2.397895	.1592796	-15.05	0.000	-2.710077	-2.085713

logit status smoke, or

status	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke	1.828915	.3760295	2.94	0.003	1.222319	2.736546
_cons	.0909091	.01448	-15.05	0.000	.0665317	.1242185

Logistisk regresjon, fysisk aktivitet

tab phys status

phys	status		Total
	0	1	
1	217	41	258
2	504	60	564
3	138	11	149
4	29	0	29
Total	888	112	1,000

tab phys status,nofreq row

phys	status		Total
	0	1	
1	84.11	15.89	100.00
2	89.36	10.64	100.00
3	92.62	7.38	100.00
4	100.00	0.00	100.00
Total	88.80	11.20	100.00

logistic status age phys

status	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.132871	.0137844	10.25	0.000	1.106174	1.160213
phys	.5851344	.1034604	-3.03	0.002	.413762	.8274858
_cons	.0018091	.0011352	-10.06	0.000	.0005289	.0061885

Logistisk regresjon, fysisk aktivitet med dummy

```
tab phys status
```

phys	Freq.
1	258
2	564
3	149
4	29
Total	1,000

```
gen phys2 = phys==2
```

```
gen phys3 = phys > 2
```

```
tab1 phys2
```

phys2	Freq.	Percent
0	436	43.60
1	564	56.40

```
tab1 phys3
```

phys3	Freq.	Percent
0	822	82.20
1	178	17.80

Logistisk regresjon, fysisk aktivitet med dummy(2)

```
logistic status age phys2 phys3
```

status	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.133368	.0138028	10.28	0.000	1.106635	1.160746
phys2	.5365943	.1310052	-2.55	0.011	.3325309	.8658848
phys3	.3785521	.1460737	-2.52	0.012	.1776923	.80646
_cons	.0010733	.0006167	-11.90	0.000	.0003481	.0033096

```
logistic status age phys
```

status	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.132871	.0137844	10.25	0.000	1.106174	1.160213
phys	.5851344	.1034604	-3.03	0.002	.413762	.8274858
_cons	.0018091	.0011352	-10.06	0.000	.0005289	.0061885

Haugnes HS et al. Components of the metabolic syndrome in long-term survivors of testicular cancer. *Ann Oncol.* 2007;18:241-248

Table 5. Logistic regression, with metabolic syndrome (≥ 2 components present) as the dependent variable

	Multiple-adjusted results ^{a,b}			Age-adjusted results ^c			Multiple-adjusted results ^{b,c}		
	OR	95% CI	P value	OR	95% CI	P value	OR	95% CI	P value
Treatment group ^d			0.002			0.002			0.003
Control group	–	–		1.00	Reference		1.00	Reference	
Surgery	1.00	Reference		0.75	0.54–1.03		0.80	0.57–1.12	
Radiotherapy	1.22	0.84–1.78		0.89	0.71–1.12		0.96	0.75–1.24	
Cis \leq 850 mg	1.44	0.98–2.12		1.11	0.85–1.43		1.15	0.87–1.52	
Cis $>$ 850 mg	3.05	1.72–5.40		2.10	1.31–3.36		2.43	1.46–4.04	
Total testosterone	0.96	0.93–0.98	0.001	0.95	0.93–0.96	$<$ 0.001	0.95	0.93–0.96	$<$ 0.001
Pack-years ^e			0.273			0.40			0.27
0 (never smoker)	1.00	Reference		1.00	Reference		1.00	Reference	
0.1–9.9	1.05	0.73–1.50		1.09	0.86–1.38		1.16	0.90–1.49	
10–19.9	1.10	0.75–1.60		1.02	0.80–1.31		1.04	0.80–1.35	
≥ 20	1.48	1.00–2.18		1.23	0.96–1.58		1.29	0.98–1.69	
Physical activity			0.200			0.041			0.12
No activity	1.00	Reference		1.00	Reference		1.00	Reference	
Moderate activity	0.96	0.64–1.44		1.08	0.83–1.41		1.19	0.90–1.58	
Hard activity	0.75	0.50–1.15		0.85	0.65–1.11		0.97	0.73–1.30	
Educational level			0.61			0.018			0.065
Low	1.00	Reference		1.00	Reference		1.00	Reference	
High (college/university)	0.93	0.70–1.23		0.81	0.67–0.96		0.83	0.68–1.01	
Family status			0.57			0.99			0.91
Living alone	1.00	Reference		1.00	Reference		1.00	Reference	
Married/cohabitant	0.91	0.66–1.26		1.00	0.80–1.25		1.01	0.80–1.29	

Odds ratio (OR) and 95% confidence interval (CI) for different predictors of the metabolic syndrome.

^aFor testicular cancer survivors only.

^bAdjusted for age and all listed variables

^cFor control population and testicular cancer survivors.

^dThere are missing data for 18 patients and nine controls in age-adjusted analyses. There are missing data for 125 patients and 117 controls in the multiple-adjusted analyses.

^eNumber of cigarette packs per day multiplied by number of smoking years.

OR for ≥ 2 metabolic syndrom components

	OR*	95% CI*	P-value
Control group	1.00	Reference	0.003
Surgery	0.80	0.57-1.12	
Radiotherapy	0.96	0.75-1.24	
Cis \leq 850 mg	1.15	0.87-1.52	
Cis $>$ 850 mg	2.43	1.46-4.04	

*Adjusted for age and all listed variables

OR for daglig røyking etter utdanning, kvinner

Utdannelse, nivå	1994-95	2007-08
1	1.00	1.00
2	0.69 (0.63, 0.76)	0.70 (0.62, 0.80)
3	0.36 (0.34, 0.41)	0.39 (0.33, 0.47)
4	0.22 (0.19, 0.25)	0.22 (0.19, 0.27)
P for trend	< 0.001	< 0.001

*Justert for alder

Eggen AE, et al. Trends in CVD risk factors across levels of education. Is the educational gap increasing? The Tromsø Study 1994-2008. JECH 2014

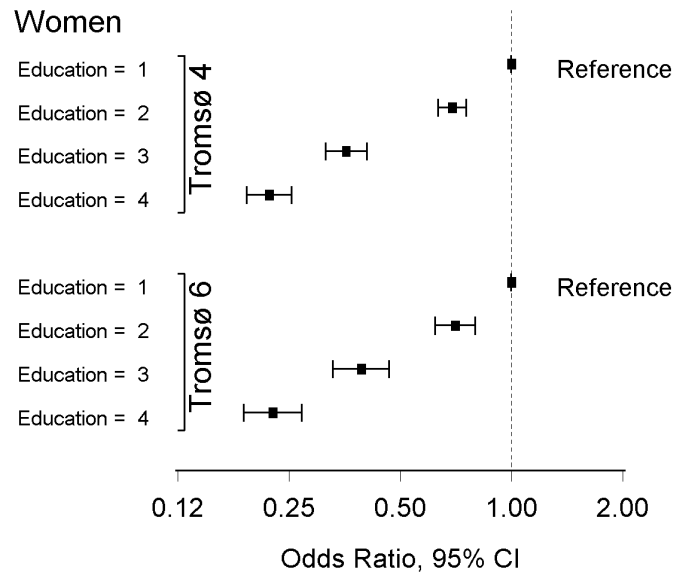
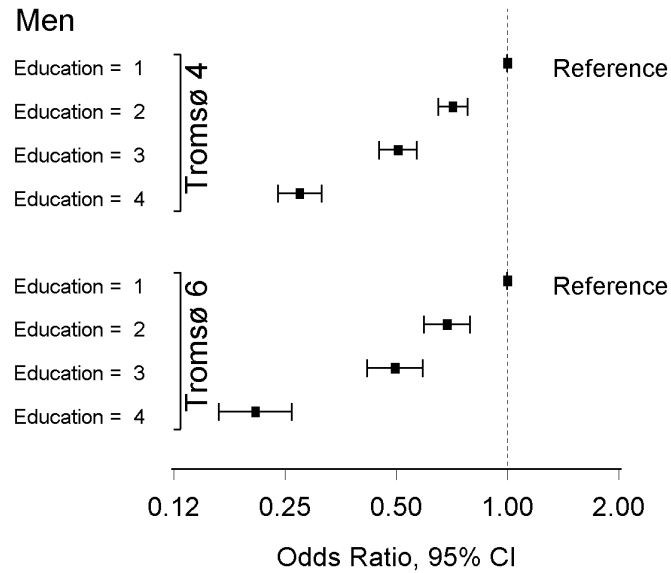
OR for daglig røyking etter utdanning, kvinner

Utdannelse, nivå	1994-95	2007-08
1	1.00	1.00
2	0.69 (0.63, 0.76)	0.70 (0.62, 0.80)
3	0.36 (0.34, 0.41)	0.39 (0.33, 0.47)
4	0.22 (0.19, 0.25)	0.22 (0.19, 0.27)
P for trend	< 0.001	< 0.001
Pr nivå	0.57 (0.54, 0.59)	0.59 (0.55, 0.64)
P for forskjell i trend	0.67	

*Justert for alder

Eggen AE, et al. Trends in CVD risk factors across levels of education. Is the educational gap increasing? The Tromsø Study 1994-2008. JECH 2014

OR for daglig røyking



Cox proporsjonale hasard regresjon

$$h(t) = h_0(t) \cdot e^{\beta_1 x_1 + \beta_2 x_2}$$

t	= tid til hendelse
h(t)	= hasard som funksjon av tid
$h_0(t)$	= baseline hasard, dvs. hasard når alle $x=0$
x_i 'ene	= uavhengige variabler
β_i 'ene	= regresjons-koeffisienter
$\exp(\beta_i)$	= hasard-ratio (HR)

Modellantakelser

- Uavhengighet
- Linearitet, logskala
- Konstant HR over tid (proporsjonale hasardantakelsen)

Cox regresjon, utskrift

```
stset obstime, failure(status)
```

```
stcox age smoke
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.123149	.0116965	11.15	0.000	1.100456	1.146309
smoke	1.965788	.3824019	3.47	0.001	1.342622	2.878191

```
stcox age smoke, nohr
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.1161362	.010414	11.15	0.000	.0957251	.1365473
smoke	.6758931	.1945286	3.47	0.001	.2946241	1.057162

Cox regresjon, kjønnsspesifikk

```
stcox age smoke if sex==0
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.123518	.0230406	5.68	0.000	1.079255	1.169597
smoke	2.169104	.7722115	2.18	0.030	1.079564	4.358251

```
stcox age smoke if sex==1
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.113783	.0136012	8.82	0.000	1.087442	1.140762
smoke	1.840968	.4286248	2.62	0.009	1.166446	2.905547

Cox regresjon, test av kjønnsforskjell

- Interaksjon mellom kjønn og røyk
- Inkluder kryssprodukt mellom kjønn og røyk i modellen

```
generate int_sex_smoke= sex*smoke
stcox age smoke sex int_sex_smoke
```

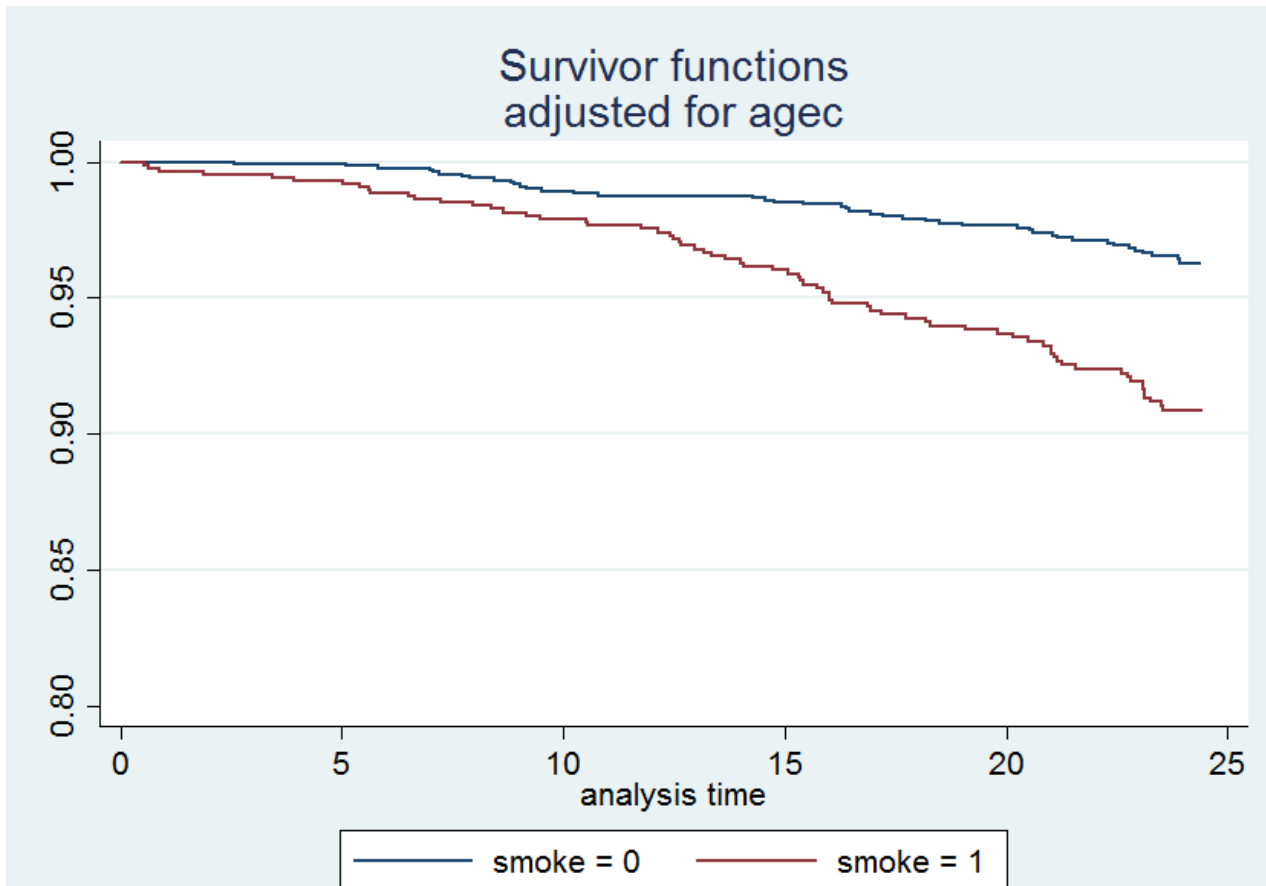
_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.11637	.0117325	10.47	0.000	1.09361	1.139604
sex	1.785369	.5965232	1.73	0.083	.9275272	3.436603
smoke	2.146293	.7592013	2.16	0.031	1.072991	4.293206
int_sex_smoke	.8547785	.3618855	-0.37	0.711	.3728057	1.959859

Ikke signifikant

Overlevelseskurver

```
gen agec=age-37.03
```

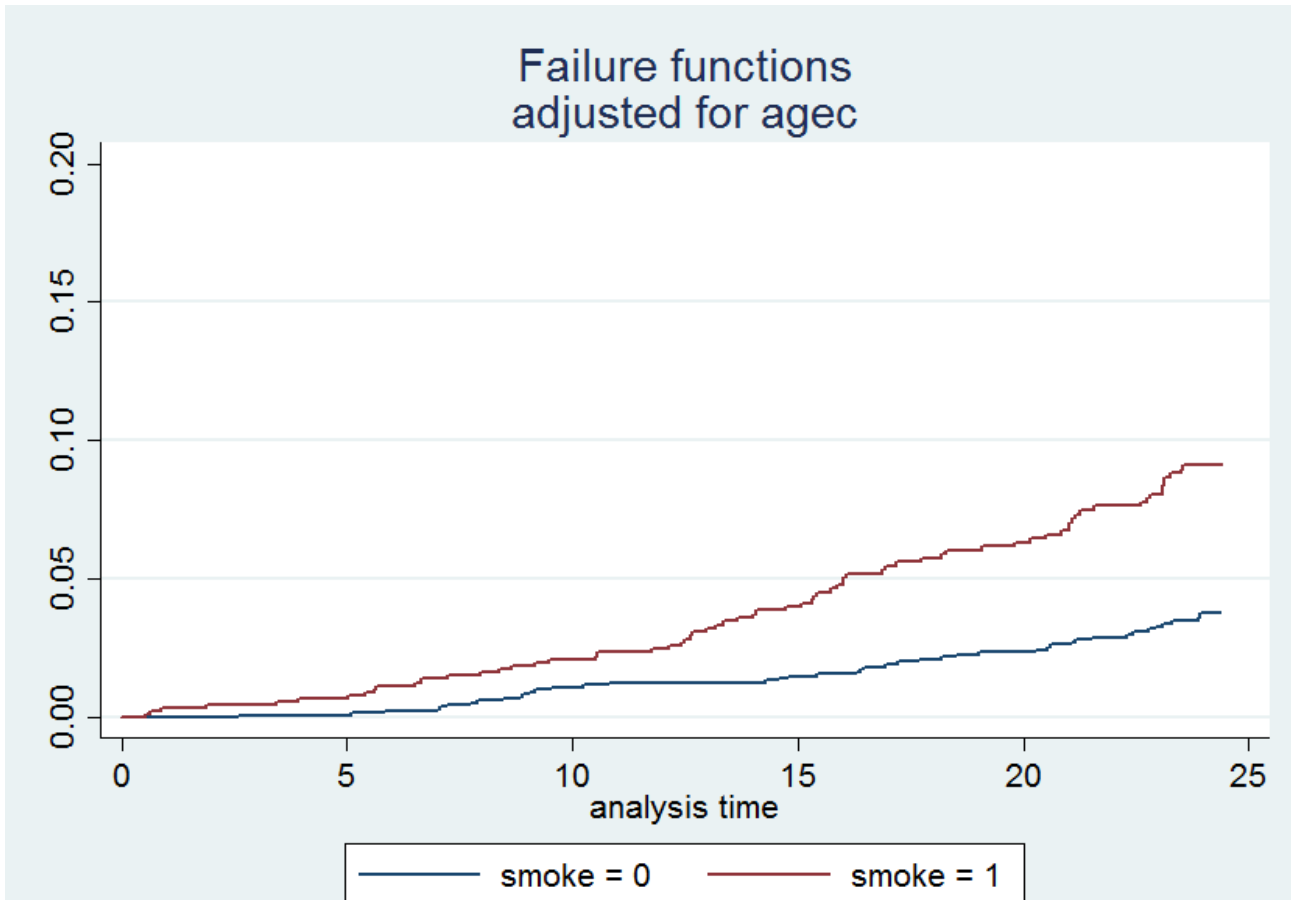
```
sts graph, by(smoke) adjustfor(agec) ylabel(0.80(0.05)1.0)
```



Failure-kurver

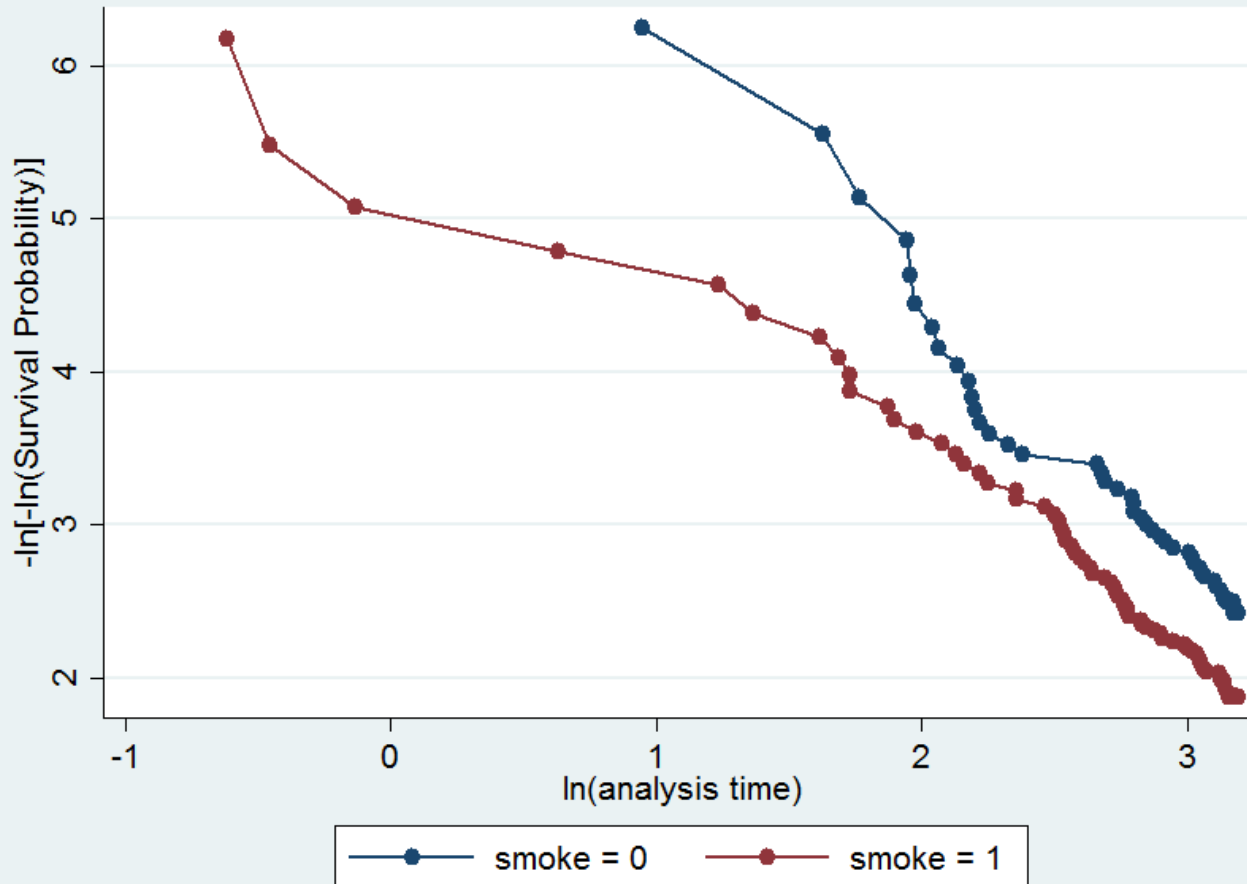
```
gen agec=age-37.03
```

```
sts graph failure, by(smoke) adjustfor(agec) ylabel(0.80(0.05)1.0)
```



Proporsjonal hasard-antakelse

stphplot,by(smoke)



Hazard ratios for mortality by BMI change in n=4881 men

10 yr BMI change	Person-years	Deaths	MR*	HR†	95% CI†
< -1	1,928	45	7.8	1.94	1.35, 2.80
-1 to 0	7,046	77	4.3	1.25	0.93, 1.68
0 to 1	16,187	101	3.2	1.00	ref
1 to 2	14,916	67	3.0	0.92	0.67, 1.25
2 to 3	7,841	31	3.1	0.91	0.61, 1.36
3 +	3,881	20	4.9	1.22	0.75, 1.99
P value difference					0.0027

*Age adjusted mortality rate per 1,000 person-years using poisson regression models.

†Adjusted for initial BMI, age, smoking status, and leisure time physical activity using Cox proportional hazard regression models.

Fractional polynomials

- Anta en uavhengig variabel x
- Cox model, lineær sammenheng (på log skala)

$$h(t) = h_0(t) \cdot e^{\beta_1 x}$$

- Cox model ikke-lineær sammenheng

$$h(t) = h_0(t) \cdot e^{\beta_1 x^p + \beta_2 x^q},$$

hvor p og $q = (-2, -1, -0.5, 0, 0.5, 1, 2, 3)$

- Dersom $p=0$ blir transformasjonen $\ln(x)$ (ikke x^0)
- Dersom $p = q$, f.eks. = 2

$$h(t) = h_0(t) \cdot e^{\beta_1 x^2 + \beta_2 \ln(x) \cdot x^2}$$

Eksempel «BMI change»

- Eksponeringsvariabel, ΔBMI
- I frac pol kan ikke variabelen ≤ 0
- Sentrerte variabelen på -11
 - $\Delta\text{BMIC} = \Delta\text{BMI} + 11$
- Beste powers ble estimert til $p = 2$ og $q = 3$
- Cox modell:

$$h(t) = h_0(t) \exp(\beta_1 \Delta\text{BMIC}^2 + \beta_2 \Delta\text{BMIC}^3 + \beta_3 x)$$

Beregning av HR

$$h(t) = h_0(t) \exp(\beta_1 \Delta BMI c^2 + \beta_2 \Delta BMI c^3 + \beta_3 x)$$

- Anta vi ønsker å beregne HR mellom $\Delta BMI = -2$ mot $\Delta BMI = 1$
- $\Delta BMI = -2$
 - $\Delta BMI c = \Delta BMI + 11 = -2 + 11 = 9$
- $\Delta BMI = 1$
 - $\Delta BMI c = \Delta BMI + 11 = 1 + 11 = 12$
- $HR = h(t \text{ gitt } \Delta BMI c = 9) / h(t \text{ gitt } \Delta BMI c = 12)$
- $HR = \frac{e^{(\beta_1 \cdot 9^2 + \beta_2 \cdot 9^3)}}{e^{(\beta_1 \cdot 12^2 + \beta_2 \cdot 12^3)}} = 2.09$

Hazard ratios* for mortality by levels of BMI change in n=4881 men

Modell: $h(t) = h_0(t) \exp(\beta_1 \Delta BMI c^2 + \beta_2 \Delta BMI c^3 + \beta_3 x)$

10 yr BMI change	Men		Women	
-2	2.09	1.56, 2.81	1.19	0.81, 1.75
-1	1.53	1.28, 1.83	1.11	0.88, 1.41
0	1.19	1.10, 1.29	1.05	0.94, 1.17
1	1.00	ref	1.00	ref
2	0.92	0.85, 0.99	0.97	0.88, 1.07
3	0.95	0.79, 1.13	0.96	0.80, 1.15
4	1.10	0.80, 1.52	0.97	0.74, 1.29
P value BMI change†		<0.0001		0.63

*Adjusted for initial BMI, age, smoking status, and leisure time physical activity.

†The joint effect of two fractional polynomial terms of BMI change.

Takk

