Kapittel 14 Logistisk regresjon

Logistisk regresjon

- Logistisk regresjon med én forklaringsvariabel
  - Odds og odds-ratio
  - Tolkning av koeffisienter
  - Inferens – konfidensintervaller og Wald-test
- Multippel logistisk regresjon

14.1 Logistisk regresjon

- I kapittel 10 og 11 (lineær regresjon) hadde vi

y $x_1\ x_2\ \ldots\ x_p$

kvantitativ kvantitative og/eller kategoriske

- I kapittel 14 (logistisk regresjon) har vi

y $x_1\ x_2\ \ldots\ x_p$

kategorisk 0/1 kvantitative og/eller kategoriske

Eks. røykestatus,
x_{i3} = 1 hvis røyker,
= 0 hvis ikke

$$y = \begin{cases} 1 & \text{"suksess"} \quad \text{(f.eks. syk, kort hale, er på instagram, olje!)} \\ 0 & \text{ellers} \quad \text{(f.eks. frisk, lang hale, er ikke på insta, tørr)} \end{cases}$$

Vi har data (x_1,y_1) , ... , (x_n,y_n).

Her er y_i en binær respons (0 eller 1) for individ i,
og x_i er en forklaringsvariabel (binær eller kvantitativ).

Vi er interesserte i hvordan 'suksess'-sannsynligheten' $P(y=1)$ påvirkes av x.

Vi lar suksess-sannsynligheten være en funksjon av x,

$$p(x) = E(y \mid x) = P(y = 1 \mid x)$$

og ønsker en modell som beskriver forholdet mellom $p(x)$ og x

En mulig modell for p(x) kunne være en enkel lineær modell $p(x) = \beta_0 + \beta_1 x$

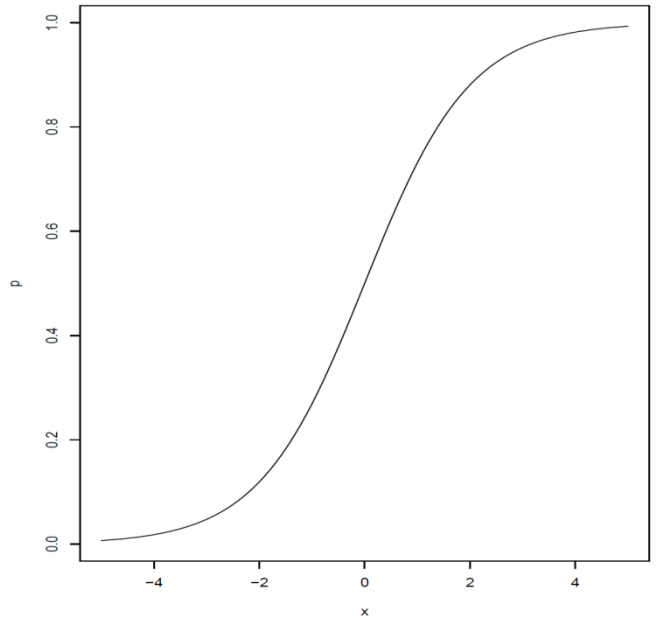
Dette er en additiv risiko-modell, som kan brukes i enkelte situasjoner.

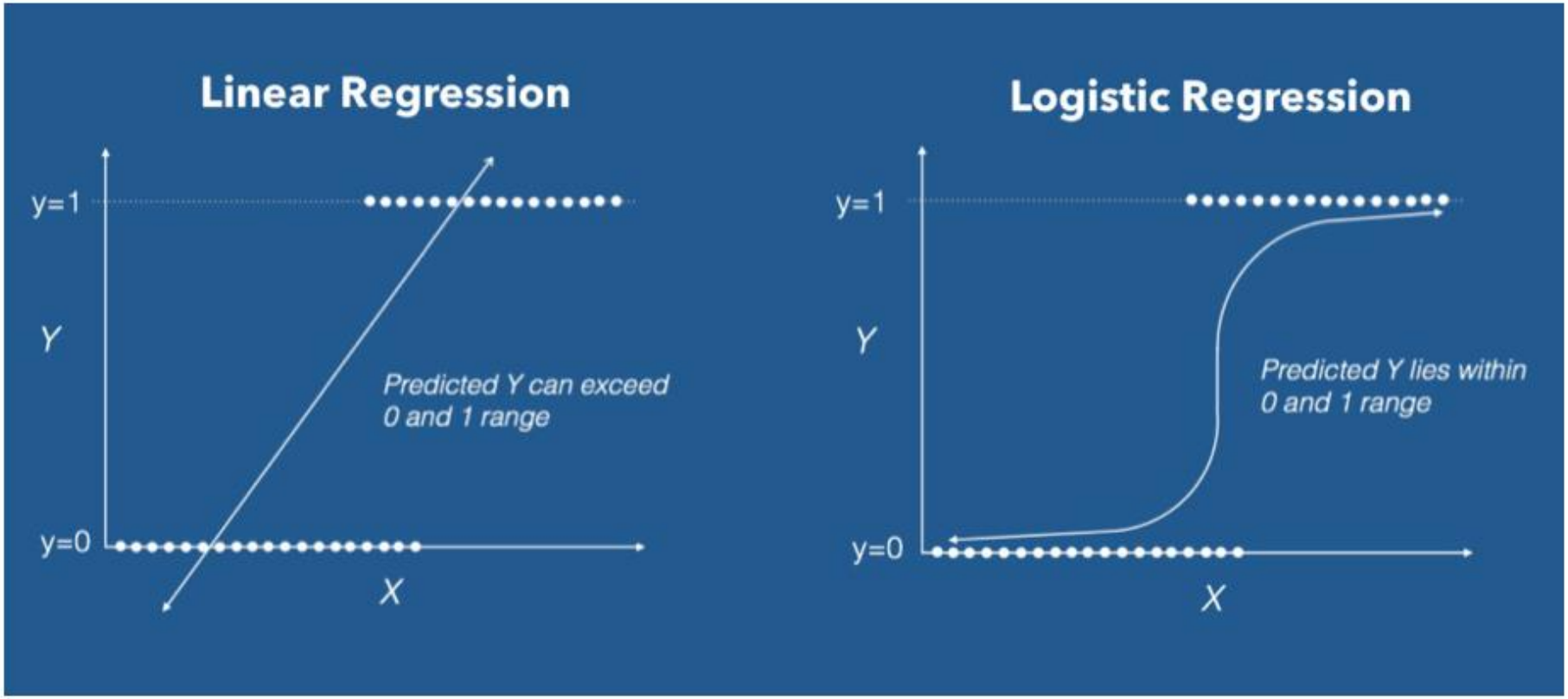
Husk at vi må ha 0 ≤ p(x) ≤1 for alle x, for at p(x) skal være er en sannsynlighet for alle x.
Vi kan ikke sikre oss mot at en lineær modell kan gi ulovlige verdier for p(x)

Vi kan sikre lovlige verdier for p(x) ved å bruke en logistisk regresjonsmodell gitt ved

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Dette gir en "S-formet" sammenheng mellom p(x) og x



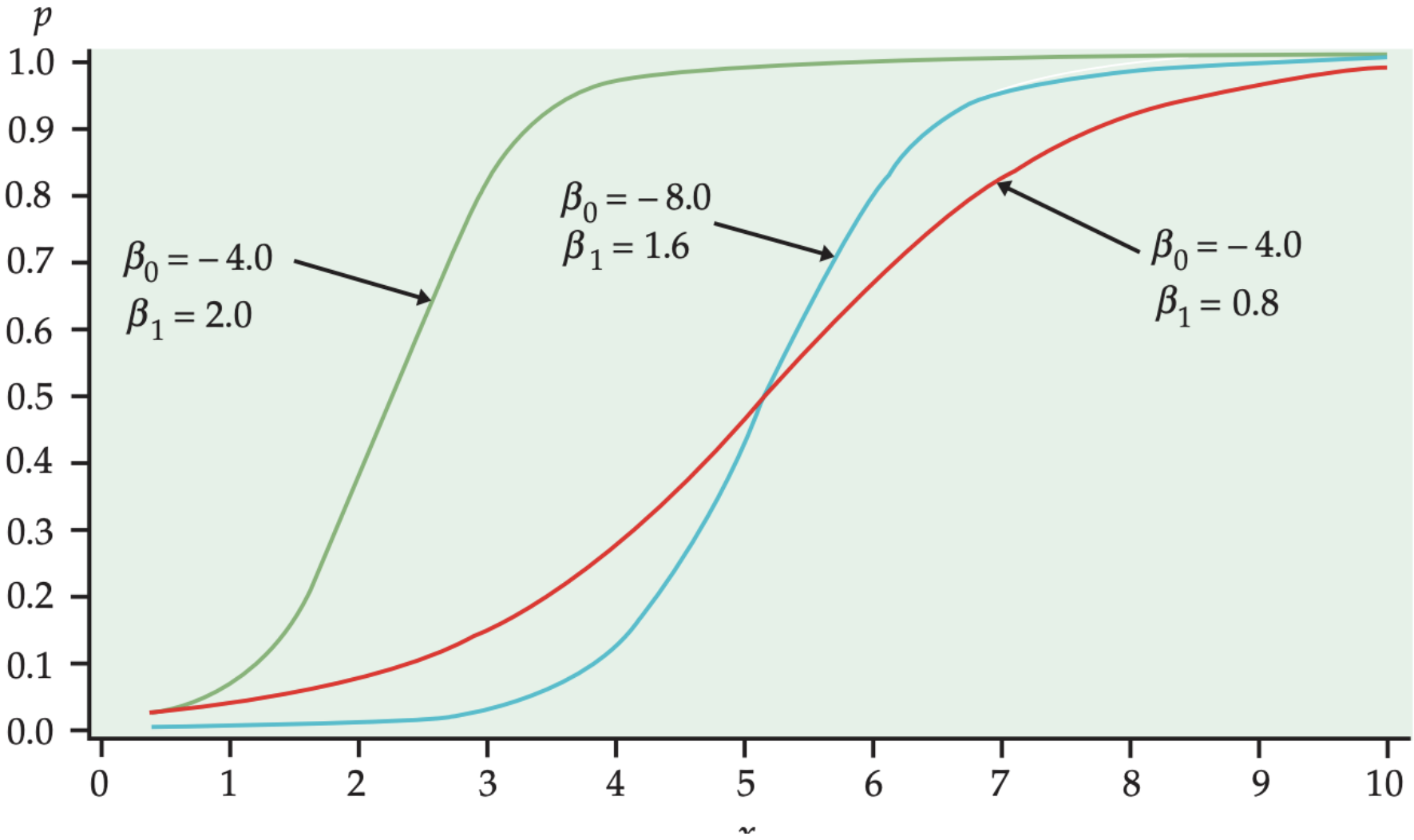


$$p(x) = \beta_0 + \beta_1 x$$

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Vi ser her kurven for $p(x) = \dfrac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$

plottet for noen ulike kombinasjoner av verdier for parameterne β_0 og β_1



Tre ekvivalente måter å formulere den logistiske regresjonsmodellen på

Vi har sett på formelen for p(x):

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Logistisk regresjon kan også formuleres vha odds:

Odds for suksess = p / (1-p)

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

Læreboka vår formulerer modellen ved å si at log-odds følger en lineær modell

$$\log\left[\frac{p(x)}{1 - p(x)}\right] = \beta_0 + \beta_1 x$$

Logistisk regresjon med en forklaringsvariabel

LOGISTIC REGRESSION MODEL

The **statistical model for logistic regression** is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where p is a binomial proportion and x is the explanatory variable. The parameters of the logistic regression model are β_0 and β_1.

Merk: her skriver vi log for den naturlige logaritmen (også kalt ln)

Odds og odds-ratio eksempel

Tilfeldig utvalg på 1069 unge bestående av 537 kvinner og 532 menn

La p_k = P(tilfeldig kvinne i denne aldersgruppen er Instagrambruker)
La p_m = P(tilfeldig mann i denne aldersgruppen er Instagrambruker)

Andel kvinner i utvalget som er Instagrambrukere: $\hat{p}_k = \frac{328}{537} = 0.6108$

Estimert odds Instagrambruker for kvinner: $\frac{\hat{p}_k}{1-\hat{p}_k} = \frac{0.6108}{1-0.6108} = 1.5694$

Andel menn i utvalget som er Instagrambrukere: $\hat{p}_m = \frac{234}{532} = 0.4398$

Estimert odds Instagrambruker for menn: $\frac{\hat{p}_m}{1-\hat{p}_m} = \frac{0.4398}{1-0.4398} = 0.7851$

Estimert *odds-ratio* for kvinner i forhold til menn:

OR = $\text{Odds}_{\text{kvinner}}$ / $\text{Odds}_{\text{menn}}$ = 1.5694 / 0.7851 = 1.999

Omtrent dobbelt så stor odds for kvinner som for menn for å være Instagrambruker for disse utvalgene

Odds-ratio: Ratioen mellom oddsene for to ulike verdier av forklaringsvariabelen

Hvis vi ser på to individer som har kovariater henholdsvis $x + \Delta$ og x , blir odds-ratioen mellom dem (OR):

$$\frac{p(x+\Delta)/[1-p(x+\Delta)]}{p(x)/[1-p(x)]} = \frac{\exp(\beta_0 + \beta_1 (x+\Delta))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1 \Delta)$$

Tolkning av regresjonskoeffisienter

Når Δ =1, ser vi at e^{β_1} er odds-ratioen som tilsvarer en enhets økning i verdien av forklaringsvariabelen.

For eksempel, hvis vi lar x=1 for kvinner og x=0 for menn blant instagrambrukerne, er odds-ratioen for kvinner sammenligna med menn e^{β_1}

Eksempel: WCGS er en stor epidemiologisk studie designet for å studere risikofaktorer for hjertesykdom blant middelaldrende menn.

Mennene ble fulgt opp i 10 år, og for hver mann ble det registrert om han utviklet hjertesykdom (y = 1) eller ikke (y = 0) i løpet av perioden.

Hvordan påvirker mennenes alder (ved oppstart av studien) risikoen (sannsynligheten) for å utvikle hjertesykdom?

Sammendrag av data

Aldersgr. v/ oppstart	35-40	41-45	46-50	51-55	56-60
Totalantall	543	1091	750	528	242
Antall syke	31	55	70	65	36
Andel syke	5.7 %	5.0 %	9.3 %	12.3 %	14.9 %

La oss analysere dataene fra WCGS i R

Husk at x_i er mannens alder ved start og at $y_i = 1$ hvis mann i har vært hjertesyk i løpet av perioden, $y_i = 0$ ellers.

```
url = "https://www.uio.no/studier/emner/matnat/math/STK1000/data/wcgs.txt"
wcgs = read.table(url, sep="\t",header=T,na.strings=".")
# R-kommando for logistisk regresjon:
fit=glm(chd69~age, data=wcgs, family=binomial)
summary(fit)
```

R output (editert):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9395	0.5493	-10.813	< 2e-16
age	0.0744	0.0113	6.585	4.56e-11

Odds-ratioen for ett års økning i alder er $e^{0.0744} = 1.077$

mens ratio for en ti-års-økning er $e^{0.0744 \times 10} = 2.10$

14.2 Konfidensintervall for β_1 og odds-ratio

95% konfidensintervall for β_1 (basert på tilnærming til normalfordelinga):

$$b_1 \pm 1.96 \times SE(b_1)$$

$OR = \exp(\beta_1)$ er odds-ratio for en enhets endring i x

Vi kan lage et 95% konfidensintervall for *OR* ved å transformere nedre og øvre grense for konfidensintervallet for β_1

I hjerte-eksempelet har vi

R output (editert):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9395	0.5493	-10.813	< 2e-16
age	0.0744	0.0113	6.585	4.56e-11

$b_1 = 0.0744$ og $SE(b_1) = 0.0113$ og dermed 95% konfidensinterval for β_1:

$0.0744 \pm 1.96 \times 0.0113$ *dvs.* fra 0.052 til 0.096

Estimat for odds-ratio: $OR = \exp(0.0744) = 1.077$

95% konfidensintervall for *OR* : $(e^{b_1 - 1.96 \times SE(b_1)}, e^{b_1 + 1.96 \times SE(b_1)})$, dvs:

fra $\exp(0.052) = 1.053$ til $\exp(0.096) = 1.101$

R-funksjon for å beregne odds-ratio med 95% konfidensgrenser

```
expcoef=function(glmobj)
{
regtab=summary(glmobj)$coef
expcoef=exp(regtab[,1])
lower=expcoef*exp(-1.96*regtab[,2])
upper=expcoef*exp(1.96*regtab[,2])
cbind(expcoef,lower,upper)
}

expcoef(fit)
```

R output (editert):

	expcoef	lower	upper
(Intercept)	0.0026	0.0009	0.0077
age	1.077	1.054	1.101

Wald test for $H_0 : \beta_1 = 0$

For å teste nullhypotesen $H_0 : \beta_1 = 0$ mot det tosidige $H_A : \beta_1 \neq 0$ bruker vi Wald-test-observator:

$$z = \frac{b_1}{SE(b_1)}$$

Vi forkaster H_0 for store verdier av $|z|$

Under H_0 er testobservatoren *tilnærma standard normalfordelt.*

P-verdi (to-sidig): $P = 2\,P(Z > |z|)$ der Z er standard normal.

I hjerte-eksemplet har vi $b_1 = 0.0744$ og $SE(b_1) = 0.0113$

Wald-test-observator

$$z = 0.0744 / 0.0113 = 6.58$$

som er meget signifikant (den bittelille p-verdien står angitt i R-utskriften på slide 13)

Fra enkel til **multippel logistisk regresjon**

Antar nå at vi for hvert individ har

- en binær respons y
- forklaringsvariable $x_1, x_2, ..., x_p$

Vi lar $p(x_1, x_2, ..., x_p) = E(y \mid x_1, x_2, ..., x_p) = P(y = 1 \mid x_1, x_2, ..., x_p)$

Logistisk regresjonsmodell:

$$p(x_1, x_2, ..., x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}$$

Alternativ definisjon, som i boka:

$$\log\left(\frac{p(x_1, x_2, ..., x_p)}{1 - p(x_1, x_2, ..., x_p)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

En tredje måte å beskrive modellen på er gjennom odds:

$$\frac{p(x_1, x_2, ..., x_p)}{1 - p(x_1, x_2, ..., x_p)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

Hvis vi sammenligner to individer med verdiene hhv. $x_1 + \Delta$ og x_1 for den første forklaringsvariabelen, og parvis identiske verdier for alle de andre forklaringsvariablene, blir odds-ratioen deres

$$\frac{p(x_1+\Delta, x_2, ..., x_p)\Big/\left[1-p(x_1+\Delta, x_2, ..., x_p)\right]}{p(x_1, x_2, ..., x_p)\Big/\left[1-p(x_1, x_2, ..., x_p)\right]}$$

$$= \frac{\exp(\beta_0 + \beta_1(x_1+\Delta) + \beta_2 x_2 + ... + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)} = \exp(\beta_1 \Delta)$$

For Δ=1 ser vi at e^{β_1} er odds-ratioen for en enhets økning i verdien for første kovariat, *dersom alle de andre kovariatene holdes fast*.

Tilsvarende tolkning holder også for de andre regresjonskoeffisientene.

La oss se videre på WCGS-studien

med hjertesykdom som respons (som før) og alder, kolesterol (mg/dL), systolisk blodtrykk (mmHg), body mass indeks (kg/m^2), og røyking (ja, nei) som forklaringsvariable

Merk at vi har ekskludert et individ med usedvanlig høy verdi av kolesterol-målinga.

R-kommando:

```
wcgs.mult=glm(chd69~age+chol+sbp+bmi+smoke, data=wcgs, family=binomial,
                        subset=(chol<600))
summary(wcgs.mult)
```

R output (editert):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.3110	0.9773	-12.598	< 2e-16
age	0.0644	0.0119	5.412	6.22e-08
chol	0.0107	0.0015	7.079	1.45e-12
sbp	0.0193	0.0041	4.716	2.40e-06
bmi	0.0574	0.0264	2.179	0.0293
smoke	0.6345	0.1401	4.526	6.01e-06

Odds-ratios med 95% konfidensintervaller

R-kommando (her kaller vi funksjonen fra slide 15):

expcoef(wcgs.mult)

R output (editert):

	expcoef	lower	upper
(Intercept)	4.50e-06	6.63e-07	3.06e-05
age	1.067	1.042	1.092
chol	1.011	1.008	1.014
sbp	1.019	1.011	1.028
bmi	1.059	1.006	1.115
smoke	1.886	1.433	2.482

Hvert av konfidensintervallene over er funnet ved først å beregne 95% konfidensintervall for β_j, dvs

$$b_j \pm 1.96 \times SE(b_j)$$

og fra dette beregnes 95% konfidensintervall for odds-ratioen for en enhets økning i x_j, *når alle andre kovariater holdes fast*:

$$(e^{b_j - 1.96 \times SE(b_j)}, e^{b_j + 1.96 \times SE(b_j)})$$

Klassifikasjon med logistisk regresjon

- Vi har sett på logistisk regresjon som en regresjonsmodell for kategoriske responsvariable med to kategorier (respons=0 eller 1)

- I situasjoner der man har mer enn to kategorier eksisterer det naturlige utvidelser, men for enkelhets skyld har vi kun sett på binære situasjoner.

- Når vi bruker en estimert/tilpasset logistisk modell til å predikere en kategorisk respons predikerer vi først sannsynligheten for hver av kategoriene, deretter kan vi bruke disse predikerte sannsynlighetene til å velge en predikert kategori. På denne måten kan logistisk regresjon brukes til *klassifikasjon.*

- Logistisk regresjon er en av de mest brukte klassifikatorene, og er en grunnleggende byggestein for mange statistiske/maskinlærings-metoder.