Innføring i medisinsk statistikk
del 2

multippel logistisk regresjon

(Rosner, 2000; kap. 13.7)

Tom Ivar Lund Nilsen
Institutt for samfunnsmedisin

regresjonsmodeller

- Hvorfor vil man bruke regresjonsmodeller?

1. **predikere et utfall** (f.eks. sykdom, død, blodtrykk) basert på et sett av forklaringsvariabler (kovariater)
   - lager statistiske modeller og vurderer modellens GOF
   - inklusjon av variabler baseres på statistisk signifikans (p-verdi)

2. **måle en sammenheng** (assosiasjonen) mellom et utfall og en forklaringsvariabel (eksponering) og samtidig **kontrollere** for effekten av andre variabler (konfounding og/eller interaksjon)
   - vurderer om sammenhengen endres når man justerer for potensielle konfoundere
   - vurdere om sammenhengen er forskjellig innenfor ulike nivå av en annen faktor
   - inklusjon av variabler baseres ikke på statistisk signifikans

prediksjon vs assosiasjon

- analysestrategien når det gjelder prediksjon er ikke den samme som når man holder på med en assosiasjonsstudie
  - denne forskjellen er forskeren ofte ikke bevisst på når han/hun analyserer kliniske eller populasjonsbaserte epidemiolgiske data
  - ofte følger man en prediksjonsstrategi selv om hensikten med studien er å måle en/flere assosiasjoner
  - husk at "stepwise regression"er fy-fy i assosiasjonsstudier…

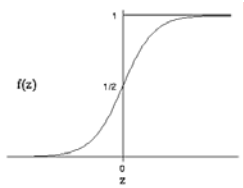
logistisk regresjon

- logistisk regresjon er en matematisk modellering som kan benyttes for å beskrive sammenhengen mellom en eller flere uavhengige variabler (X1, X2, …, Xk) og et dikotomt (todelt) avhengig utfall
- andre matematiske modeller er også mulig, men logistisk regresjon er den overlegent mest populære ved analyse av epidemiologiske data med et dikotom utfall
- hvorfor er logistisk regresjon blitt så populær…?

den logistisk funksjonen (1)

- den logistiske funksjonen viser den matematiske formen som den logistiske modellen er basert på
- funksjonen kalles f(z), og er gitt av uttrykket:

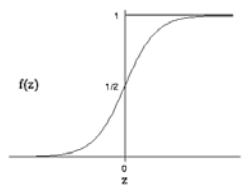
$$f(z) = \frac{1}{1+e^{-z}}$$



hvor z er et uttrykk for de uavhengige variablene (X'ene)

den logistisk funksjonen (2)

- modellen er laget for å beskrive sannsynligheter, og sannsynligheter kan bare ha verdier mellom 0 og 1
- dermed gir modellen sannsynligheten (eller risikoen) for at et individ skal ha en sykdom



- S-formen er også tiltalende, fordi den viser at for lave verdier av z er risikoen lav, inntil en viss grenseverdi hvor risikoen stiger raskt, og for høye verdier av z er risikoen høy
- dette illustrerer hvordan man ofte tenker at ulike eksponeringer virker på forekomsten av sykdom

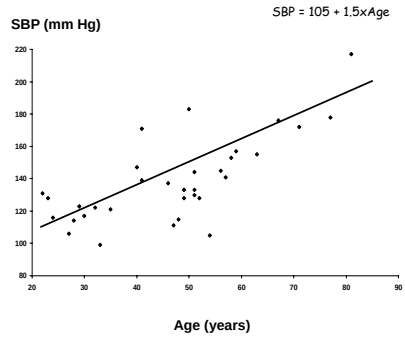
lineær regresjon

- ta utgangspunkt i den lineære regresjonen
  - en kontinuerlig utfallsvariabel **Y**
  - denne kan relateres til et sett av forklaringsvariabler $\mathbf{X_1}$, …, $\mathbf{X_k}$
- den matematiske modellen er gitt ved uttrykket:

$$E(y \mid x) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - E(y|x) leses som "forventningsverdien av Y, gitt verdien av x"
    - denne kan anta alle verdier fra -∞ til +∞
  - x-ene representerer de ulike forklaringsvariablene
  - β-ene representerer størrelsen på sammenhengen mellom X og Y
  - α-en representerer et konstantledd

7



8

den logistiske regresjonsmodellen

- i **logistisk regresjon** har vi
  - en dikotom utfallsvariabel Y (syk = 1, frisk = 0)
  - derfor kan forventningsverdien E(y|x) bare ha verdier mellom 0 og 1
  - forventningsverdien til et dikotomt utfall vil være det samme som *sannsynlighet* for utfallet
    - derfor kan E(y|x) uttrykkes som *p*

9

logistisk transformasjon

- dersom man vil undersøke sammenhengen mellom et dikotomt utfall Y og et sett av forklaringsvariabler (X_1, …, X_k) kan man ikke bruke ligningen for lineær regresjon

$$p = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- dette fordi ved gitte x-verdier kan høyresiden av formelen gi en forventningsverdi (*p*) som er mindre enn 0 og større enn 1, og det er ikke mulig
- i stedet kan man gjøre en logistisk (logit) transformasjon av *p*

$$\text{logit(p)} = \ln\left(\frac{p}{1-p}\right)$$

10

logit(*p*)

- i motestning til *p* må ikke logit(*p*) ha verdier mellom 0 og 1, men kan anta alle verdier fra -∞ til +∞
- da kan vi skrive følgende uttrykk for den **logistiske modellen**

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n$$

  - hvor p er sannsynligheten for sykdom ved en gitt x

11

odds vs sannsynlighet

- sannsynligheten for å få en "ener" er 1 av 6 = 1/6 = 0.17
- odds for å få en "ener" ved terningkast er 1 til 5 eller

$$= \frac{1/6}{5/6} = 1/5 = 0.2$$

- odds er sannsynligheten (risiko) for at en egenskap er til stede dividert med sannsynligheten for at egenskapen ikke er til stede

odds = p / (1 - p)

hvor p = sannsynligheten for egenskapen
(f.eks. ssh for å være røyker)

12

naturlig logaritmen av oddsen

- logit(p) = $\ln\left(\frac{p}{1-p}\right)$
- **logit(p) er altså den naturlig logaritmen av odds for sykdom**

odds ratio

- i en case-kontrollstudie kan vi beregne odds for å være eksponert blant de syke, og odds for å være eksponert blant de friske, og forholdet mellom oddsene kalles odds ratio (OR)

		Case	Kontroll
Eksponering	Ja	a	b
	Nei	c	d

- odds ratio vil være odds for eksponering blant casene (a/c) dividert med odds for eksponering blant kontorollene (b/d)

eksempel
røyking og hjertekardød

- vi vil studere sammenhengen mellom røyking og risiko for å dø av hjertekarsykdom i HUNT 1 (1984-86)
  - 1502 døde (case), og selektert 2 kontroller per case
  - vi får følgende data som settes inn i en 2x2 tabell

		Død: Ja	Død: Nei
Røyker	Ja	321	832
	Nei	1181	2172

- ut fra disse dataene kan man beregne odds for å være røyker blant de casene

  a/c = 321/1181 = 0.27
- og odds for å være røyker blant kontrollene

  b/d = 832/2172 = 0.38

odds ratio for eksponering

- dette gir en odds ratio på 0.27/0.38 = 0.7
  - odds for å være røyker blant casene er 0.7 ganger så stor som oddsen for å være røyker blant kontrollene
  - dvs. at casene har ~30% lavere odds (risiko) for å være røyker sammenlignet med kontrollene
  - men ønsker ikke si noe om risikoen for å være røyker dersom man dør av hjertesykdom, vi vil heller si noe om risikoen for å dø dersom man røyker...
- odds ratio for å være eksponert gitt at man er syk er det samme som odds ratio for å være syk gitt at man er eksponert

OR for sykdom

		Død: Ja	Død: Nei
Røyker	Ja	321	832
	Nei	1181	2172

- ut fra disse dataene kan man beregne odds for å være case blant røykere

  a/b = 321/832 = 0.39
- og odds for å være case blant ikke-røykere

  c/d = 1181/2172 = 0.54
- OR = 0.39 / 0.54 = 0.7
- røykere har altså 0.7 ganger så stor risiko for å dø av hjertesykdom som ikke-røykere

vurdering av resultatet?

- hvor pålitelig er dette resultat?
- vi kan sjekke konfidensintervall og/eller p-verdi
  - dvs. er det en signifikant sammenheng

p-verdi

- p-verdien er sannsynligheten for å ha fått det resultatet du har (OR = 0.7) gitt at det egentlig ikke er noen sammenheng (dvs. at OR egentlig er 1.0)
  - i dette tilfellet er p-verdien <0.001
  - det er mindre enn 1 ‰ sannsynligheten for å ha fått en OR på 0.7 basert på dette utvalget dersom den sanne OR (i populasjonen) var 1.0
  - basert på p-verdien ser det jo ut til å være en pålitelig sammenheng vi har målt
- men vær oppmerksom på at p-verdien er konfundert:
  - presisjonen blandes sammen med effektstørrelsen

den konfunderte p-verdi

- man kan får en lav p-verdi både når effekten er liten og utvalget stort, og når effekten er stor men utvalget lite:
  - 1500 case, 3000 kontroller, OR = 0.7, p <0.001
  - 150 case, 3000 kontroller, OR = 0.2, p < 0.001
  - 3000 case, 6000 kontroller, OR = 0.9, p <0.001
- en feil mange gjør er å bruke p-verdien som en dikotom variabel (± 0.05), og kun rapportere sig. vs non-sig.
- dersom man skal bruke p-verdier må man oppgi hele tallet
  - mange forskningsresultat er gått i søppelbøtta fordi man har vært for opphengt i p-verdien som ikke var signifikant, og mange forskningsresultater har blitt overdrevet fordi p-verdien har vært signifikant

konfidensintervall

- konfidensintervall blir også brukt for å si noe om presisjonen på den målte sammenhengen
  - et smalt intervall betyr høy presisjon, og et bredt betyr lav presisjon
- ettersom intervallet angis av to tall og ikke bare ett (slik som p), er det ikke konfundert
  - det forteller oss noe om både presisjon og effektstørrelse
- blir ofte anvendt som en hypotesetest hvor man bare bedømmer om verdien gitt av H0 er inkludert i intervallet eller ikke
  - men bruk en "romslig" fortolkning

konfidensintervallet i eksempelet

- 95% konfidensintervallet i denne analysen hvor odds ratio er 0.7 har en nedre verdi på 0.6 og en øvre verdi på 0.8
  - konfindesintervallet er smalt, og vi har god presisjon
  - konfidensintervallet omslutter ikke nullverdien 1.0
    - en statistisk signifikant sammenheng med a = 0.05
- en vanlig fortolkning er at vi med 95% sikkerhet kan si at OR for populasjonen som utvalget kommer fra ligger innenfor dette intervallet
  - konfidensintervallet vil i 95% av tilfellene omslutte den "sanne" OR i populasjonen

fortsatt usikkert?

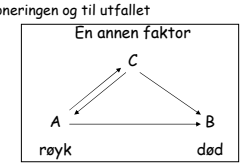
- da har vi altså vurdert resultatene våre ut fra statistiske kriterier, men er vi likevel sikre på at dette er et pålitelig resultat?

konfounding?

- vi har ennå ikke vurdert om resultatet kan være påvirket av konfounding med andre variabler
  - dvs. variabler som er forbundet både med røyking og med hjertekardød...
    - alder
    - kjønn
    - fysisk aktivitet
    - etc.

kriterier for konfounding

- kriteriet for at en faktor skal være en konfounder er at den skal være relatert både til eksponeringen og til utfallet



- alder
  - alder øker risikoen for hjertekardød
  - røyking vil være ulikt fordelt i ulike aldersgrupper
- begge kriteriene for en konfounder oppfylt, men hva gjør vi så...?

justering for konfounding

- dersom man har informasjon om potensielle confoundere så kan man justere bort den konfunderende effekten i statistiske analyser
- dette vil i hovedsak si i regresjonsmodeller, som f.eks. logistisk regresjon...

logistisk regresjon

- først kjører vi en analyse hvor bare røykevariabelen er med i regresjonsmodellen...

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1	røyk_2	-.343	.075	20.931	1	.000	.710	.613	.822
	Constant	-.266	.098	7.422	1	.006	.766		

a. Variable(s) entered on step 1: røyk_2.

- vi ser at vi får samme resultat som da vi regnet for hånd:
  - OR = 0.7 med tilhørende p-verdi og konfidensintervall

justering for alder

- deretter gjentar vi den samme analysen en gang til, men nå legger vi til en aldersvariabel (her: alder som en kontinuerlig variabel) i regresjonsmodellen sammen med røykevariabelen, og vifår følgende resultat:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1	røyk_2	.723	.104	48.474	1	.000	2.061	1.681	2.526
	alder	.147	.005	1031.024	1	.000	1.158	1.148	1.169
	Constant	-11.474	.386	882.044	1	.000	.000		

a. Variable(s) entered on step 1: røyk_2, alder.

hva betyr dette?

- etter at vi legger inn alder i modellen forandrer odds ratio estimatet seg dramatisk
  - ujustert hadde vi en OR = 0.7, som vi fortolket som at røykere hadde lavere risiko for å dø av hjertekarsykdom enn ikke-røykere
  - justert for alder er OR nå forandret til 2.0, dvs. røykere har dobbelt så stor risiko for å dø av hjertekarsykdom som ikke-røykere
- illustrerer viktigheten av å justere for potensielle confoundere
  - det hjelper ikke å ha lave p-verdier og smale konfidensintervall dersom man har målt en feilaktig sammenheng

detaljert informasjon

- men selv om man har informasjon om potensielle konfoundere er det også nødvendig at denne informasjonen er mest mulig detaljert
  - la oss tenke oss at vi bare hadde informasjon om alder i to kategorier - over og under 50 år
  - dersom vi gjentar analysen med denne aldersvariabelen får vi følgende resultat:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1	røyk_2	-.060	.084	.500	1	.479	.942	.798	1.112
	alder_50	3.845	.214	323.947	1	.000	46.737	30.750	71.036
	Constant	-7.776	.440	312.394	1	.000	.000		

a. Variable(s) entered on step 1: røyk_2, alder_50.

- vi får en forandring i OR fra 0.7 til 1.0 (dvs. ingen sammenheng)

hvordan vurdere konfounding

- konfounding vurderes ut fra om det skjer en endring i risikoestimatet (OR) for den faktoren man studerer når man legger til mulige konfunderende variabler i regresjonsmodellen
  - i vårt eksempel var det ganske enkelt å vurdere, det skjedde en stor endring - fra 0.7 til 2.0 – når vi inkluderte alder i modellen
- men hvor stor skal endringen være for at man sier at justeringen har hatt en effekt?
  - ikke noe fasitsvar, en tommelfingerregel er at en endring i OR (eller RR) på 20% eller mer bør man ta hensyn til
- Viktig! konfounding vurderes ikke ut fra om det er en signifikant sammenheng mellom konfounderen og sykdommen – men om det er en **endring** i den assosiasjonen man studerer...

31

flere mulige konfoundere ?

- kjønn?

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1[a]	røyk_2	.708	.104	45.987	1	.000	2.030	1.654	2.491
	alder	.148	.005	1028.383	1	.000	1.159	1.149	1.170
	gender	-.239	.086	7.682	1	.006	.787	.665	.932
	Constant	-11.184	.399	786.452	1	.000	.000		

a. Variable(s) entered on step 1: røyk_2, alder, gender.

32

flere mulige konfoundere ?

- fysisk aktivitet?

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1[a]	røyk_2	.731	.123	35.430	1	.000	2.077	1.633	2.642
	alder	.149	.005	761.774	1	.000	1.160	1.148	1.173
	gender	-.337	.105	10.269	1	.001	.714	.581	.877
	fysakt_2	-.606	.106	32.904	1	.000	.546	.444	.671
	Constant	-10.208	.507	405.034	1	.000	.000		

a. Variable(s) entered on step 1: røyk_2, alder, gender, fysakt_2.

- dette er blant annet grunnen til at man ikke skal gjøre automatisk stepwise regression
  - hverken kjønn eller fysisk aktivitet trenger å være med når vi ser på effekt av røyking

33