Innføring i medisinsk statistikk
del 2

Multippel logistisk regresjon

(Rosner, 2000; kap. 13.7 og
2006 presentasjon av Tom Ivar Lund Nilsen)

Gunnar Taraldsen
Institutt for nevromedisin

Logistisk regresjon

- **Hvorfor brukes logistisk regresjon?**

1. **Av samme grunn som en bruker lineær regresjon, men for tilfeller hvor prediksjonsvariabelen kun tar 2 verdier** (f.eks. død eller levende, gruppe 1 eller 2).

2. **Noen vanlige generaliserte lineære modeller** (Nelder og Wedderburn 1972) gitt av type prediksjonsvariabel

   1. Kontinuerlig: Lineær regresjon.
   2. Indikator (=Dikotom, tar 2 verdier): Logistisk.
   3. Diskret (tar 2 eller flere verdier): Poisson, Logistisk
   4. ++ (positiv + mange varianter av overnevnte + nye)

Regresjonsmodeller

- Hvorfor vil man bruke regresjonsmodeller?

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

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

Prediksjon versus 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

En epidemiologisk problemstilling

- en typisk epidemiologisk problemstilling kan være:
  - hva er sammenhengen mellom en eksponering (E) og en sykdom (S)
  - S er et dikotomt utfall hvor 0 betyr ikke-syk og 1 betyr frisk
    - det dikotome utfallet kan f.eks. være koronar hjertesykdom (CHD)
  - E er en dikotom eksponeringsvariabel, f.eks. fysisk aktivitet, klassifisert som "inaktiv" eller "aktiv"
- problemstillingen i denne studien blir derfor å vurdere om fysisk aktivitet (aktiv vs. inaktiv) er assosiert med CHD-status (0,1)
- for å vurdere i hvilken grad en eksponering er assosiert med et utfall må vi ofte kontrollere for effekten av andre faktorer, slik som alder (C_1), kjønn (C_2) og røykestatus (C_3)

Et multivariabelt problem

- i dette eksempelet utgjør variabelen E (fysisk aktivitet), sammen med et sett med kontrollvariable (C_1-C_3) en samling med **uavhengige** variabler som vi vil bruke for å beskrive den **avhengige** variabelen D (CHD-status)
- generelt kan faktorene skrives som X_1, X_2, opp til X_k, hvor k er antall faktorer (variable) som studeres
  - For eksempel kan:
    - X_1 = eksponeringsvariabelen E
    - X_2, X_3 og X_4 kan være kontrollvariablene C_1, C_2 og C_3
- når vi ønsker å relatere et sett av X'er til en avhengig variabel D har vi en **multivariabel** problemstilling
  - i en analyse av et slikt problem benyttes ofte en matematisk modell

Logistisk regresjon

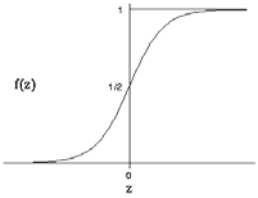
- logistisk regresjon er en matematisk modellering som kan benyttes for å beskrive sammenhengen mellom et sett av uavhengige X'er og et dikotomt avhengig utfall
- andre matematiske modeller er også mulig, men logistisk regresjon er den overlegent mest populære ved analyse av epidemiologiske data med et dikotomt utfall
- hvorfor er logistisk regresjon blitt så populær...?
  - Enkel modell
  - Tolkning av lineære koeffisienter som OR (odds ratio)
  - Den matematiske formen kan begrunnes fundamentalt ved en underliggende modell i heldige tilfeller

7

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:

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

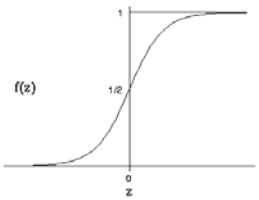


hvor z kan uttrykkes lineært ved de uavhengige variablene (X'ene). (Tenk på z som en eksponering/dose, og p=f(z) som sannsynligheten for et av to mulige utfall, for eksempel risikoen for å dø.)

8

Den logistisk funksjonen (2)

- vi ser at verdien til f(z) varier kun mellom 0 og 1 uansett hvilken verdi z har
- i tillegg ser vi at funksjonen har en utpreget S-form



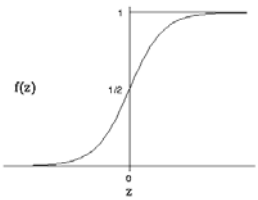
- disse to egenskapene er 2 av grunnene til at den logistiske modellen er så populær i epidemiologiske studier...

(Enhver kumulativ fordelingsfunksjon til en symmetrisk kontinuerlig sannsynlighetsfordeling har disse 2 egenskapene)

9

Den logistisk funksjonen (3)

- 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

10

Lineær regresjon

- for kanskje lettere å forstå den logistiske modellen starter vi med å repetere den lineære regresjonsmodellen
  - en kontinuerlig utfallsvariabel **Y**
  - denne kan relateres til et sett av forklaringsvariabler X_1, …, 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"
    - E(y|x) kan anta alle verdier fra $-\infty$ til $+\infty$
  - x-ene representerer de ulike forklaringsvariablene
  - β-ene representerer størrelsen på sammenhengen mellom X og Y
  - α-en representerer et konstantledd

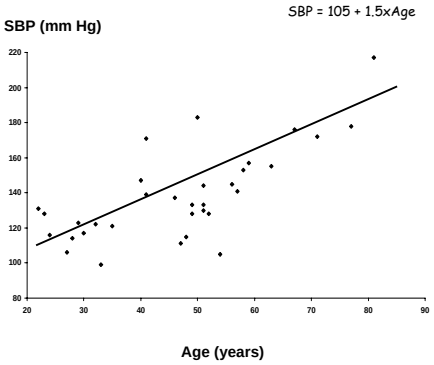
11

Enkel lineær regression

Table 1 Age and systolic blood pressure (SBP) among 33 adult women

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

12



13

Den logistiske regresjonsmodellen

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

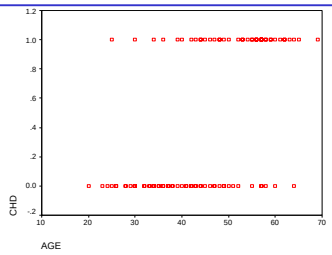
14

Eksempel

- et datasett med informasjon om alder (20-69 år) og forekomst av hjertesykdom (42 av 100) blant 100 deltakere
- vi vil undersøke sammenhengen mellom alder (AGE) som uavhengig variabel og hjertesykdom (CHD) som avhengig variabel
  - AGE er en kontinuerlig variabel med verdier fra 20 til 69, mens CHD bare kan ha to verdier - syk eller ikke-syk (kodet som 1 eller 0)
  - dersom utfallsvariabelen hadde vært kontinuerlig og ikke binær ville vi sannsynligvis startet med et scatter-plot av AGE vs CHD
  - et slikt plott ville vært til hjelp når vi skal bestemme sammenhengen mellom AGE og CHD

15

Eksempel
(forts.)



- alle punktene ligger langs en av to parallelle linjer som tilsvarer CHD = 0 og CHD = 1
  - tendens til at individer med CHD = 0 er yngre enn de med CHD = 1
  - plottet illustrerer den dikotome egenskapen til utfallsvariabelen, men gir ikke noe tydelig bilde på sammenhengen mellom CHD og AGE

16

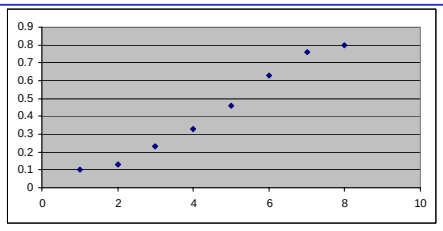
Eksempel
(forts.)

- for å undersøke denne sammenhengen bedre kan vi lage intervaller av AGE og beregne gjennomsnittet av CHD
  - CHD er kodet 0,1 og da får vi ssh (p) for CHD i hver aldersgruppe

		CHD		
aldersgruppe	n	nei	ja	% CHD
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.23
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
	100	57	43	

17

Eksempel
(forts.)



- vi ser av plottet at ssh. for CHD (*p*) er mellom 0 og 1
  - endringen i *p* for én enhets endring i x blir gradvis mindre når andelen med sykdom nærmer seg 0 og 1
  - men kurven kan aldri krysse 0 eller 1
  - Ligner på logistisk kurve…

18

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 *p* som er mindre enn 0 eller større enn 1, og det er ikke mulig
- i stedet kan man gjøre en logistisk (logit) transformasjon av *p*

$$\text{logit}(p) = z = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

logit(*p*)

- i motsetning 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 for eksempel kan være sannsynligheten for sykdom ved en gitt x
- Den logistiske funksjonen er den inverse av logit funksjonen:

$$\text{logit}(p) = z \Leftrightarrow p = 1/(1+\exp(-z))$$

Odds

- 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)

Naturlig logaritmen av oddsen

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

Estimering av β-verdiene

- for å tilpasse den logistiske regresjonsmodellen til et datasett kreves det at man estimerer β-verdiene i ligningen

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

- i lineær regresjon bruker man minste kvadratsums metode ("least squares method") for å estimere β-verdiene
  - man velger de verdiene for β-ene som minimerer avstanden mellom de observerte dataene og regresjonslinja
- i logistisk regresjon benytter man en tilsvarende metode som kalles **maksimum likelihood (=rimelighetsestimering)**
  - man velger β-verdiene slik at sannsynligheten for å oppnå de observerte dataene maksimeres

Regresjonsparametrene - ß

- dette er en studie av CVD-død (0,1) og fysisk aktivitet (inaktiv = 1 vs aktiv = 2), justert for kjønn (menn = 1, kvinner = 2) og alder (år)

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	AGE	.141	.005	776.391	1	.000
	GENDER(1)	-.373	.105	12.714	1	.000
	PYSACT2(1)	-.658	.105	39.595	1	.000
	Constant	-9.655	.360	720.475	1	.000

a. Variable(s) entered on step 1: AGE, GENDER, PYSACT2.

- får regresjonskoeffisienter (ß-estimater) for hver uavhengig variabel som er inkludert i modellen
  - ß-verdiene sier noe om størrelsen på sammenhengen
  - ß-verdiene er justert for hverandre

Resulterende modell

- den naturlige logaritmen til oddsen for CVD-død er dermed uttrykt ved følgende ligning:

$$\ln\left(\frac{p}{1-p}\right) = -9.66 + (-0.658)PYSACT + (-0.373)GENDER + (0.141)AGE$$

25

exp(ß)

- i tillegg gir SPSS exp(β)
  - dette er det samme som OR, og vi kan be om å få konfidensintervall rundt OR-estimatet
  - Signifikant dersom intervallet ikke inneholder 1

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
AGE	.141	.005	776.391	1	.000	1.151	1.140	1.163
GENDER(1)	-.373	.105	12.714	1	.000	.688	.561	.845
PYSACT2(1)	-.658	.105	39.595	1	.000	.518	.422	.636
Constant	-9.655	.360	720.475	1	.000	.000		

ariable(s) entered on step 1: AGE, GENDER, PYSACT2.

26

Odds ratio

- i en case-kontrollstudie vet vi ikke prevalens av sykdom, men vi har syke (case) og ikke-syke (kontroller)
- da 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)

27

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 pr 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

28

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 (OR ulik RR generelt!)
  - 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

29

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 (=ca RR her?)
- røykere har altså ca 0.7 ganger så stor risiko for å dø av hjertesykdom som ikke-røykere...

30

Vurdering av resultatet?

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

31

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

32

Rapporter p-verdi og konfidensintervall

- man kan få 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
- Konfidensintervallet kan vise om effekten av faktoren er relevant i sammenhengen

33

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 forteller det 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 for dette er p-verdien mer informativ

34

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
  - konfidensintervallet 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, men det riktige er
  - konfidensintervallet vil i 95% av tilfellene (dersom hele undersøkelsen gjentas mange ganger) omslutte den "sanne" OR i populasjonen

35

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?

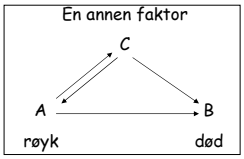
36

Konfundering?

- vi har ennå ikke vurdert om resultatet kan være påvirket av (konfundering med) andre variable
  - dvs. variable som er forbundet både med røyking og med hjertekardød...
    - alder
    - kjønn
    - fysisk aktivitet
    - Etc
- Konfundering=Effektforveksling=Faktorforvirring='Aliasing'= Folding=....

Kriterier for konfundering

- kriteriet for at en faktor skal være en konfunder 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
- da er begge kriteriene for at alder kan være en konfunder oppfylt, men hva gjør vi så...?

Justering for konfundering

- dersom man har informasjon om potensielle konfundere så kan man justere bort den konfunderende effekten i statistiske analyser
- dette vil i hovedsak si i regresjonsmodeller, og det er her logistisk regresjon kommer inn i vårt eksempel...

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[a]	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 utvider vi analysen ved å legge til en aldersvariabel i regresjonsmodellen sammen med røykevariabelen, og vi får 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[a]	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 (gitt at OR ca = RR: bør diskuteres!)
- illustrerer viktigheten av å justere for potensielle konfundere
  - det hjelper ikke å ha lave p-verdier og smale konfidensintervall dersom man har målt en feilaktig sammenheng
  - Gal modell gir gale svar

Detaljert informasjon

- men selv om man har informasjon om potensielle konfundere 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[a]	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)

43

Hvordan vurdere konfundering?

- konfundering vurderes ut fra om det skjer en endring i risikoestimatet (OR) for den faktoren man studerer når man legger til mulige konfunderende variable 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! konfundering vurderes ikke ut fra om det er en signifikant sammenheng mellom konfunderen og sykdommen – men om det er en **endring** i den assosiasjonen man studerer...

44

Flere mulige konfundere ?

- 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.

45

Flere mulige konfundere ?

- 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.

46

Simpson's paradoks (neste sider)

- Konfundering (to confound= å forvirre, å villede, å ...)
- Effektforveksling
- Faktorforvirring
- 'Aliasing'
- Folding
- Konvolusjon

47

Modelltesting (Rosner, neste sider)

48

Fra Gunnar Løvås: Statistikk for universiteter og høgskoler. Universitetsforlaget.

Simpson's paradox

Våre statistiske konklusjoner må alltid ledsages av sunn fornuft og god kunnskap om det fenomenet vi studerer. I noen tilfeller kan vi faktisk få helt feil bilde dersom vi utelater en viktig variabel, slik at den riktige sammenhengen er motsatt av den vi har oppdaget. Dette kalles Simpsons paradoks og illustreres med det følgende eksemplet. Slike feilkonklusjoner skyldes ofte at data fra ulike kilder (tabeller) er slått sammen i en tabell. Les eksemplet, og husk budskapet neste gang du uttaler deg på bakgrunn av statistiske undersøkelser.

Eksempel 258 *Det hadde vært interessant å sende ut et spørreskjema til bilførere for å kartlegge hvor mange som har fått en bilskade de siste årene. Resultatet av undersøkelsen kunne blitt slik:*

	Bilskade	*Ikke bilskade*	*Total*
Mann	233	323	556
Kvinne	87	194	281
Total	320	517	837

Her er det en signifikant forskjell på kjønnenes evne til å kjøre skadefritt. Andelen $233/556 = 0.42$ *av mennene var involvert i en skade, mens kvinnenes andel bare var* $87/281 = 0.31$. *Betyr dette at kvinner er dyktigere sjåfører enn menn?*

Løsning: Ved å undersøke spørreskjemaene nærmere kunne vi foreta en ekstra kategorisering, avhengig av hvor stor bil personene kjører. Da kunne resultatene presenteres slik:

	Store biler			*Små biler*		
	Skade	*Ikke skade*	*Total*	*Skade*	*Ikke skade*	*Total*
Mann	150	35	185	82	288	370
Kvinne	16	2	18	72	192	264
Total	166	37	203	154	480	634

Nå er bildet snudd på hodet. For store biler var 88 % av kvinnene mot 81 % av mennene involvert i en skade. For små biler var 27 % av kvinnene mot 22 % av mennene involvert. Kvinnene har høyest skadeandel både for store og små biler!

Assessing Goodness of Fit of Logistic-Regression Models

We can use the predicted probabilities for individual subjects to define residuals and assess goodness of fit of logistic-regression models.

Residuals in Logistic Regression If our data are in ungrouped form—that is, each subject has a unique set of covariate values (as in Table 13.19)—then we can define the **Pearson residual** for the ith observation by

$$r_i = \frac{y_i - \hat{p}_i}{se(\hat{p}_i)}$$

where

$y_i = 1$ if the ith observation is a success and $= 0$ if it is a failure

If our data are in grouped form—i.e., if the subjects with the same covariate values have been grouped together (as in Table 13.15)—then the Pearson residual for the ith group of observations is defined by

$$r_i = \frac{y_i - \hat{p}_i}{se(\hat{p}_i)}$$

where

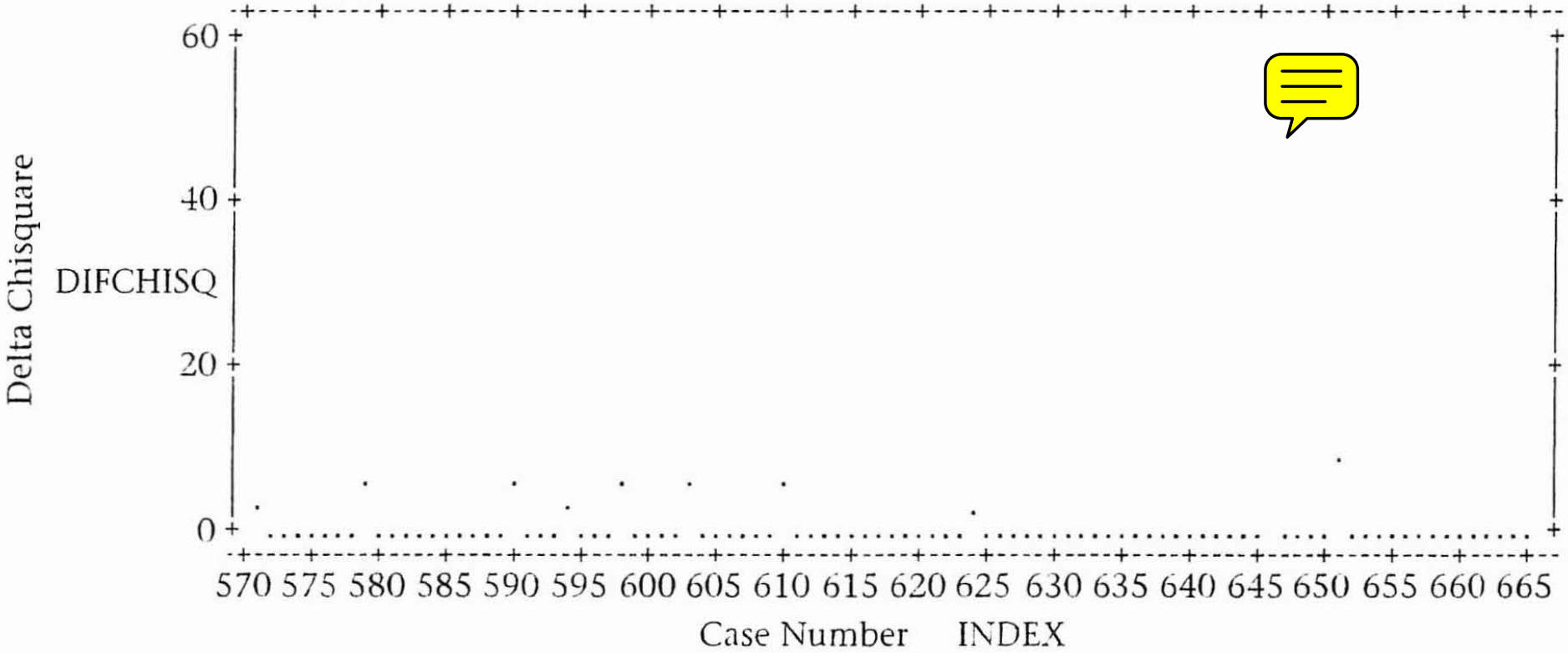
y_i = proportion of successes among the ith group of observations

$\hat{p}_i = \dfrac{e^{\hat{L}_i}}{1 + e^{\hat{L}_i}}$ as defined for ungrouped data

$$se(\hat{p}_i) = \sqrt{\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}}$$

n_i = number of observations in the ith group

particularly if the data are in ungrouped form. Nevertheless, Pearson residuals with large absolute values are worth further checking to be certain that the values of the dependent and independent variables are correctly entered and possibly to identify patterns in covariate values that consistently lead to outlying values. For ease of observation, the square of the Pearson residuals (referred to as DIFCHISQ) are displayed in Figure 13.5 for a subset of the data. The largest Pearson residual in this data set is for observation 646, which corresponds to a young smoker with no other risk factors who had a predicted probability of CHD of approximately 2% and developed CHD during the 10 years. He had a Pearson residual of 7.1, corresponding to a DIFCHISQ of $7.1^2 \approx 50$. If there are several other young smokers with no other risk factors who had events, we may want to modify our model to indicate interaction effects between smoking and age; i.e., to allow the effect of smoking to be different for younger versus older men.


Delta Chisquare
DIFCHISQ
60
40
20
0
570 575 580 585 590 595 600 605 610 615 620 625 630 635 640 645 650 655 660 665
Case Number INDEX