Noen metoder for analyse av alder-periode-kohort-modeller

Frank Helén Pedersen

Masteroppgave i matematisk statistikk

Matematisk institutt
Universitetet i Bergen

Mai 2008

Takk

Jeg vil takke veilederen min, professor Ivar Heuch, for hjelp og støtte under den lange prosessen.
Jeg vil også takke mine medstudenter for å ha bidratt til et positivt miljø på lesesalen.
Spesielt vil jeg takke Karl Ove Hufthammer for at han alltid viste stor velvilje og positiv innstilling når det gjaldt datatekniske spørsmål.
Til slutt vil jeg takke ledelse og øvrige kolleger ved Askøy Videregående Skole for positive tilbakemeldinger underveis, og spesielt vil jeg takke Geir Mjaavatn for hjelp til det datatekniske i innspurten.

Innhold

1

Innledning

I analyse av epidemiologiske data har aldersstandardiserte rater vært mye benyttet for å beskrive forekomsten av en sykdom over tid. De siste tiårene har bruk av aldersspesifikke rater blitt mer og mer vanlig i slike analyser. Ulike metoder for å analysere data som beskriver insidens- og mortalitetsrater fra sykdomsregistre basert på aldersspesifikke rater, har derfor fått mye oppmerksomhet. Flere ulike modeller som beskriver effekten av alder, kalenderperiode og fødselskohort er presentert i litteraturen. Det er særlig ett problem som trolig har bidratt til en rik litteratur og en heftig debatt, nemlig det faktum at de tre variablene alder, periode og kohort er direkte lineært avhengige av hverandre. Dette har ført til en sann overflod av forslag til hvordan modellene bør parametriseres. Jeg vil i denne oppgaven presentere noen av de vanligste modellene og et utvalg av de mange mulige parametriseringene, men oversikten vil langt fra være fullstendig. Jeg vil illustrere bruken av de ulike metodene ved hjelp av eksempler fra ulike kreftregistre.

I kapittel 2 presenteres noen sentrale begreper innen epidemiologi, blant annet aldersstandardisering. Jeg vil også presisere hva som menes med alder, periode og kohort i denne oppgaven.

I kapittel 3 ser vi nærmere på ulike metoder å presentere data på. Vi ser på vanlige tabeller, Lexis-diagrammer og grafiske fremstillinger.

I kapittel 4 ser vi først kort på det teoretiske grunnlaget for modellene som brukes. Vi tar for oss generaliserte lineære modeller (GLM) på generell basis, og deretter ser vi spesielt på poissonmodeller. Vi ser så på "maximum likelihood"-estimering i GLM. Til slutt presenteres kort noen av modellene som blir studert nærmere i de neste kapitlene.

I kapittel 5 tar vi utgangspunkt i modeller presentert av Clayton & Schifflers (1987a), og vi undersøker ulike metoder for å parametrisere deres modeller. Vi ser først på en modell med alder som eneste faktor. Her kommer det også et innslag om restriksjoner og kontraster. Så ser vi på alder-periode-modellen, og videre på alder-kohort-modellen og til slutt på alder-drift-modellen.

I kapittel 6 fortsetter vi med å undersøke ulike metoder for å parametrisere modeller presentert av Clayton & Schifflers (1987b). I dette kapittelet tar vi for oss alder-periode-kohort-modellen, og ser nærmere på to metoder: Metoden med førsteordensdifferanser og metoden med andreordensdifferanser.

I kapittel 7 presenteres først en alternativ metode for å parametrisere alder-periode-kohort-modellen fra kapittel 6 basert på Holford (1983). Deretter presenteres en alternativ alder-periode-kohort-modell basert på Carstensen (2007), sammen med hans forslag til hvordan denne modellen kan parametriseres.

I kapittel 8 får vi en oppsummering og en diskusjon av de ulike metodene. Til slutt gjøres det et forsøk på å formulere en konklusjon.

2

Begreper og definisjoner i epidemiologi

Epidemiologiske grunnbegreper

I denne første delen av kapittelet vil jeg ta for meg noen grunnleggende begreper innen epidemiologi. Men først vil jeg definere hva vi mener med moderne epidemiologi. En mulig definisjon er: Epidemiologiske studier er statistisk kartlegging av sykdommers forekomst og årsaksforhold (Aalen et al. 2006). Epidemiologi er ikke lenger begrenset til studiet av epidemier (smittsomme sykdommer), og strengt tatt behøver ikke en epidemiologisk studie heller være en studie av en sykdom.

Mål for frekvens

La p være sannsynligheten for å bli syk eller dø av en eller annen sykdom. Dersom vi skal lage en sannsynlighetsmodell må vi kunne estimere p. Det virker naturlig å estimere p ved å bruke en eller annen form for relativ hyppighet eller *frekvens*. Et av de mest brukte målene i epidemiologi er *insidensrate*. Vi kan definere insidensrate som antall nye tilfeller i løpet av en gitt tidsperiode delt på samlet populasjonstid under risiko i den samme tidsperioden. Hvis vi kaller insidensraten for IR, populasjonstiden for PT og antall nye tilfeller i tidsrommet (t_0, t) for n, kan vi skrive det som

$$IR = \frac{n}{PT}$$

Det fins flere måter å beregne samlet *populasjonstid* på. Dersom alle individenes forløp i den aktuelle populasjonen er kjent, kan det beregnes temmelig nøyaktig som summen av alle enkeltbidragene fra alle individene under risiko. Dette kan vi skrive som

$$PT = \sum_{i=1}^{N'} \Delta t_i$$

hvor $N' =$ antall individer i "frisk" populasjon og Δt_i er tid under risiko for individ nr. *i*.

Hvis vi ikke kjenner alle individenes forløp, kan vi for stabile populasjoner beregne populasjonstiden som

$$PT = N' \cdot (\Delta t)$$

hvor igjen $N' =$ antall "friske" individer i populasjonen og Δt er tid under risiko (Kleinbaum et al. 1982). For sjeldne sykdommer, som for eksempel kreft, vil $N' \approx N$, hvor N er populasjonsstørrelsen. Samlet populasjonstid vil vanligvis oppgis som totalt antall *personår*. Antall personår kan regnes ut på flere måter, men i forbindelse med kreft er det vanlig å beregne antall personår som middelfolkemengden N i det aktuelle tidsintervallet ganget med antall år Δt i tidsintervallet, det vil si $PT = N \cdot \Delta t$.

Et annet viktig mål i epidemiologien er *mortalitetsrate* eller *dødsrate*, som kan defineres som antall individer som dør i løpet av en gitt tidsperiode delt på samlet populasjonstid under risiko i den samme tidsperioden. Hvis vi kaller mortalitetsraten for *MR*, populasjonstiden for *PT* og antall individer som dør i tidsrommet (t_0, t) for n, kan vi skrive det som

$$MR = \frac{n}{PT}$$

Populasjonstiden beregnes på samme måte som for insidensrate. For sjeldne sykdommer som kreft er det vanlig å oppgi insidensraten og mortalitetsraten per 100 000 personår.

Relativ risiko er et annet begrep som er mye brukt i epidemiologien. La p_1 være sannsynligheten for å bli syk (eller dø) i gruppe 1 og la p_2 være sannsynligheten for å bli syk (eller dø) i gruppe 2. Forholdet mellom p_2 og p_1 blir da den relative risikoen for å bli syk (eller dø) i gruppe 2 i forhold til gruppe 1, altså

$$RR = \frac{p_2}{p_1}$$

hvor *RR* er relativ risiko. I mange sammenhenger vil gruppe 1 være en kontrollgruppe, men langt fra alltid. Relativ risiko estimeres som forholdet mellom estimerte rater ("rate ratio").

Aldersstandardisering

Standardisering av rater brukes for å kunne sammenlikne grupper med ulik alderssammensetning i tid eller rom. Aldersstandardisering brukes særlig ved presentasjon av mortalitets- eller insidensrater for kreftsykdommer, blant annet for å belyse endringer i rater over tid.

To former for standardisering har vært brukt, direkte og indirekte. *Direkte standardisering* er vanlig i kreftstatistikk. For å kunne gjennomføre direkte aldersstandardisering trenger man en referansepopulasjon med kjent alderssammensetning. Videre må man kunne regne ut de aldersspesifikke insidens- eller mortalitetsratene i studiepopulasjonen. Jeg vil illustrere utregningen ved hjelp av et eksempel.

Eksempel

Dette eksempelet er et tenkt eksempel. De nødvendige opplysningene er samlet i tabell 1.

Tabell 1. Tabell for utregning av aldersstandardisert mortalitetsrate i et tenkt eksempel.

Alders-gruppe	Studie-populasjon	Antall døde i studiepopulasjon	Mortalitetsrate i studiepopulasjon	Referanse-populasjon	Forventet antall døde
40 – 49	4 000	40	0,01	1 000 000	10 000
50 – 59	2 000	100	0,05	600 000	30 000
60 – 69	2 000	500	0,25	200 000	50 000
Totalt	8 000	640	0,08	1 800 000	90 000

Forventet antall døde i hver aldersgruppe fås ved å multiplisere de aldersspesifikke ratene med antall personer i den tilsvarende aldersgruppen i referansepopulasjonen. I dette eksempelet blir den aldersstandardiserte mortalitetsraten:

$$MR_a = 90000 : 1800000 = 0{,}05$$

Den reelle mortalitetsraten i studiepopulasjonen er til sammenlikning 0,08. Dette kan forklares ved at det i studiepopulasjonen er en større andel i den eldste aldersgruppen med den

høyeste mortalitetsraten (25 %) sammenliknet med referansepopulasjonen (bare 11 % i den eldste aldersgruppen).

Ved *indirekte standardisering* brukes referansepopulasjonen til å fremskaffe aldersspesifikke rater. Disse ratene brukes til å beregne forventet antall døde i hver aldersgruppe i studiepopulasjonen. Forventet antall døde kan så summeres, og summen sammenliknes med det reelle antall døde i studiepopulasjonen.

Eksempler på referansepopulasjoner er verdensstandarden og europeisk standard. Et av problemene med bruk av aldersstandardiserte rater er tap av informasjon siden disse ratene er basert på en sum. Det kan for eksempel være vanskelig å oppdage aldersspesifikke forskjeller i risiko over tid.

Alder, periode og kohort

Alder, periode og kohort er tre tidsvariabler som brukes mye i epidemiologiske studier. Tid er en kontinuerlig størrelse, og derfor er også alder, periode og kohort i utgangspunktet kontinuerlige størrelser. Det er likevel vanlig i kohortanalyser å anta at alle disse tre tidsvariablene er kategoriske.

Med *alder* mener vi løpende alder, dersom en person for eksempel får stilt en diagnose eller dør dagen før han fyller 35 år, så regnes alderen som 34 år. Det er vanlig å dele opp alder i aldersgrupper, der intervallene i hver gruppe er like lange. Vanlig brukt er 5-års-intervaller, for eksempel 20–24 år, 25–29 år osv. Jeg vil bruke bokstaven a som betegnelse på faktoren alder. Anta at antall aldersgrupper er A, da vil den yngste aldersgruppen svare til $a = 1$, den nest yngste til $a = 2$, osv. til den eldste aldersgruppen som blir $a = A$. Hver aldersgruppe tilsvarer da ett nivå i en faktormodell.

Med *periode* mener vi kalenderperiode, det vil si datoen når eventuelt diagnosen blir stilt eller pasienten dør. Det er vanlig å dele periode opp i intervaller, helst like lange. Også her er det vanlig å bruke 5-års-intervaller, for eksempel 1950–1954, 1955–1959 osv. Jeg vil bruke

bokstaven p for faktoren periode. Anta at det er P perioder, da vil $p = 1$ være den tidligste perioden, $p = 2$ den neste osv. til $p = P$ som blir den siste perioden.

I utgangspunktet er kohort bare en gruppe individer som følges over tid. Når vi bruker begrepet *kohort* i forbindelse med alder og periode er det underforstått at vi mener fødselskohort, det vil si fødselsdatoen til pasienten. Dersom alder og periode er delt i intervaller, er det naturlig også å dele kohort i intervaller. For kohort er det ikke uvanlig med 10-års-intervaller, for eksempel 1930–39, 1940–49 osv. Jeg vil bruke bokstaven k for faktoren kohort. Anta at antall kohorter er K, da er $k = 1$ den eldste kohorten, $k = 2$ den nest eldste osv. til $k = K$ som er den yngste kohorten.

Dersom alder og periode er gitt, kan vi finne kohortene. Jeg vil bruke et tenkt eksempel for å vise hvordan vi gjør dette.

Eksempel
Anta at alder og periode deles i 5-års-intervaller. Anta videre at det er tre aldersgrupper og tre perioder. La den første aldersgruppen være 20–24 år, og de andre gruppene 25–29 år og 30–34 år. La så den første perioden være 1960–1964, og de to andre periodene 1965–1969 og 1970–1974. En pasient i aldersgruppen 20–24 år som får diagnosen i perioden 1960–1964, kan tidligst være født i 1935 (får diagnosen i 1960 før han/hun fyller 25 år). På tilsvarende måte finner vi at det seineste en pasient i samme aldersgruppe og samme periode kan være født er i 1944. Kohorten blir altså 1935–1944.
På samme måte kan vi finne de andre kohortene. Kohortene blir 1925–1934, 1930–1939, 1935–1944, 1940–1949 og 1945–1954. Vi ser at vi får overlappende 10-års-intervaller, og vi ser også at vi får 5 kohorter mot 3 aldersgrupper og 3 perioder.

Anta generelt at alder og periode er delt i like lange intervaller. Da kan vi beregne antall kohorter som

$$K = A + P - 1$$

Videre er det en direkte lineær sammenheng mellom alder, periode og kohort gitt ved

$$k = A - a + p$$

Hvis alder og periode ikke er delt i like lange intervaller, så blir det mer komplisert.
Dette problemet vil bare så vidt bli berørt i denne oppgaven, men vil bli kommentert i kapittel 8.

3

Data

Innhenting av data

Epidemiologiske data fins i stor utstrekning i offentlige registre. Statistikk over forekomst av kreft er ofte samlet i egne kreftregistre. I likhet med mange andre land har Norge et nasjonalt kreftregister. Kreftregisteret i Norge utgir en årlig rapport, den siste som er utgitt er for 2006 (Cancer Registry of Norway 2007). Disse rapportene inneholder blant annet en oversikt over antall nye tilfelle etter krefttype og kjønn. De gir også en oversikt over antall nye tilfelle etter alder, der alderen er gruppert i 5-års-intervaller. De har også flere oversikter over aldersstandardiserte rater med verdensstandarden som referansepopulasjon. I tillegg gir de en oversikt over folketallet i Norge fordelt på aldersklasser og kjønn, samt en standardpopulasjon (verden).

Jeg vil bruke data fra kreftstatistikk for å belyse ulike modeller og parametriseringer. Mine data vil hentes fra diverse artikler hvor slike data presenteres, fremfor direkte fra ulike lands kreftregistre. I disse artiklene er dataene vanligvis lagt ut i tabeller fordelt på aldersklasser og kalenderperioder. I tabellene er enten antall tilfelle oppgitt sammen med ratene i hver aldersklasse, eller antall tilfelle sammen med populasjonsstørrelsen i hver aldersklasse. Nedenfor følger to eksempler som viser hvordan slike tabeller kan se ut. Disse eksemplene vil bli brukt seinere i oppgaven.

Eksempel

Tabell 2 viser insidensraten for blærekreft hos menn i kreftregisteret for Birmingham i perioden 1960–1976 (Clayton & Schifflers 1987a). Dataene i tabellen er lagt ut slik at de reflekterer måten de er samlet inn på, det vil si med kolonner som definerer P kalenderperioder og rader som definerer A aldersgrupper. I dette eksempelet er $A = 11$, og alle aldersgruppene består av 5-års-klasser. Videre er $P = 4$, og klassene for periode varierer fra 3 til 5 år, dessuten mangler det tilsynelatende et år (1967). I selve tabellen er ratene per 100 000 personår oppgitt sammen med antall tilfelle.

Tabell 2. Aldersspesifikke insidensrater (per 100 000 personår) for blærekreft hos menn i Birminghamområdet i perioden 1960–1976. Antall tilfelle i parentes. (Kilde: Cancer Incidence in Five Continents, Vol. 1-Vol. 4).

Alder\periode	1960–62	1963–66	1968–72	1973–76
25 – 29	0,42 (2)	0,31 (2)	0,55 (5)	1,10 (9)
30 – 34	0,00 (0)	0,65 (4)	1,73 (14)	1,15 (8)
35 – 39	2,06 (11)	1,21 (8)	4,02 (31)	2,49 (16)
40 – 44	1,62 (8)	4,03 (28)	6,74 (55)	5,29 (33)
45 – 49	9,40 (48)	7,02 (45)	14,95 (126)	16,80 (107)
50 – 54	13,90 (67)	16,65 (108)	25,73 (199)	24,41 (164)
55 – 59	24,25 (102)	29,15 (171)	41,06 (309)	44,81 (245)
60 – 64	44,50 (141)	50,51 (253)	71,39 (469)	70,25 (372)
65 – 69	60,47 (135)	66,97 (226)	100,69 (514)	101,97 (440)
70 – 74	94,84 (150)	95,73 (210)	141,96 (450)	142,70 (420)
75 – 79	116,08 (116)	118,16 (159)	154,19 (276)	174,42 (270)

Eksempel

Tabell 3 viser antall dødsfall av prostatakreft hos ikke-hvite menn i USA i perioden 1935–1969 (Holford 1983). Dataene i tabellen er lagt ut på samme måte som i forrige eksempel, det vil si med kolonner for $P = 7$ perioder og rader for $A = 7$ aldersgrupper. I dette eksempelet er både alder og periode delt inn i 5-års-intervaller. I denne tabellen er ikke ratene oppgitt, i stedet er midtperiodepopulasjonene oppgitt sammen med antall tilfelle.

Tabell 3. Antall dødsfall av prostatakreft for ikke-hvite menn i USA i perioden 1935–1969. I parentes estimerte midtperiodepopulasjoner (x 10^3). (Kilde: National Center for Health Statistics 1937–1973 (for antall dødsfall), Grove & Hetzel 1968 og Bureau of the Census 1974 (for populasjonstall)).

Alder\periode	1935–39	1940–44	1945–49	1950–54	1955–59	1960–64	1965–69
50 – 54	177 (301)	271 (317)	312 (353)	382 (395)	321 (426)	305 (473)	308 (498)
55 – 59	262 (212)	350 (248)	552 (279)	620 (301)	714 (358)	649 (411)	738 (443)
60 – 64	360 (159)	479 (194)	644 (222)	949 (222)	932 (258)	1292 (304)	1327 (341)
65 – 69	409 (132)	544 (144)	812 (169)	1150 (210)	1668 (230)	1958 (264)	2153 (297)
70 – 74	328 (76)	509 (94)	763 (110)	1097 (125)	1593 (149)	2039 (180)	2433 (197)
75 – 79	222 (37)	359 (47)	584 (59)	845 (71)	1192 (91)	1638 (108)	2068 (118)
80 – 84	108 (19)	178 (22)	285 (32)	475 (39)	742 (44)	992 (56)	1374 (66)

Grafiske fremstillinger

Det kan ofte være nyttig å fremstille dataene grafisk før det lages modeller. Det er fire klassiske plott i epidemiologien:

1) Rater mot alder, hvor observasjoner innen hver periode er forbundet.
2) Rater mot alder, hvor observasjoner innen hver fødselskohort er forbundet.
3) Rater mot periode, hvor observasjoner innen hver aldersklasse er forbundet.
4) Rater mot kohort, hvor observasjoner innen hver aldersklasse er forbundet.

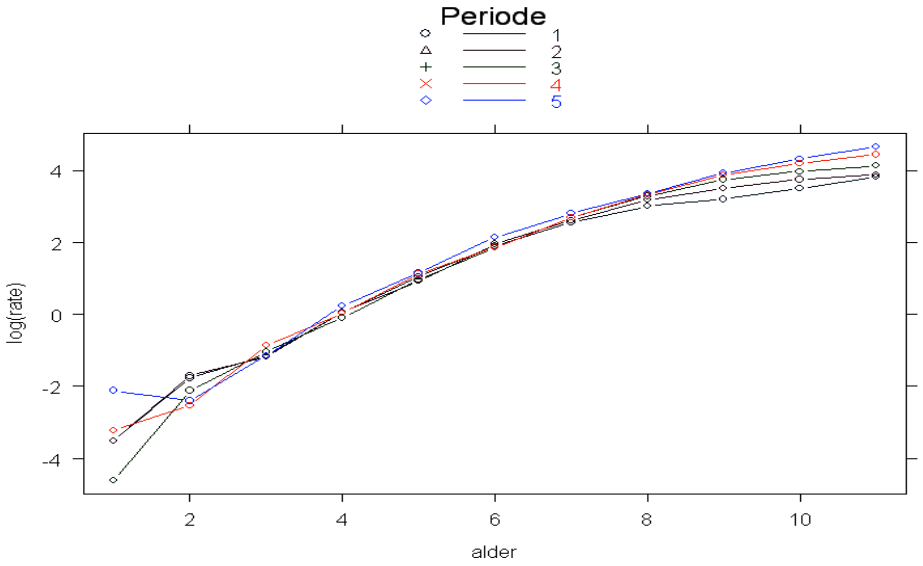
Eksempel

Som eksempel skal vi se på mortalitetsratene for blærekreft hos italienske menn i perioden 1955–1979 (Clayton & Schifflers 1987a). Dataene er samlet i tabell 4. Figur 1 viser de fire klassiske plottene i dette tilfelle. Figuren fortjener noen kommentarer. Jeg har valgt å bruke logaritmisk skala for ratene, men en absolutt skala kan også brukes. Vi ser at kurvene for kohorter (plott 2) er av varierende lengde (figur 1b). Ellers ser vi av plott 1 (figur 1a) og 2 at alder ser ut til å være en viktig faktor. Dersom plott 1 og 3 (figur 1c) har parallelle linjer indikerer det at de aldersspesifikke ratene er proporsjonale mellom periodene (det vil si følger en alder-periode-modell). Dersom plott 2 og 4 (figur 1d) har parallelle linjer indikerer det at de aldersspesifikke ratene er proporsjonale mellom kohortene (det vil si følger en alder-kohort-modell). Det er vanskelig ut fra figuren å si noe sikkert om hvilken av disse modellene som er best. Data fra dette eksempelet vil bli analysert i seinere kapitler.

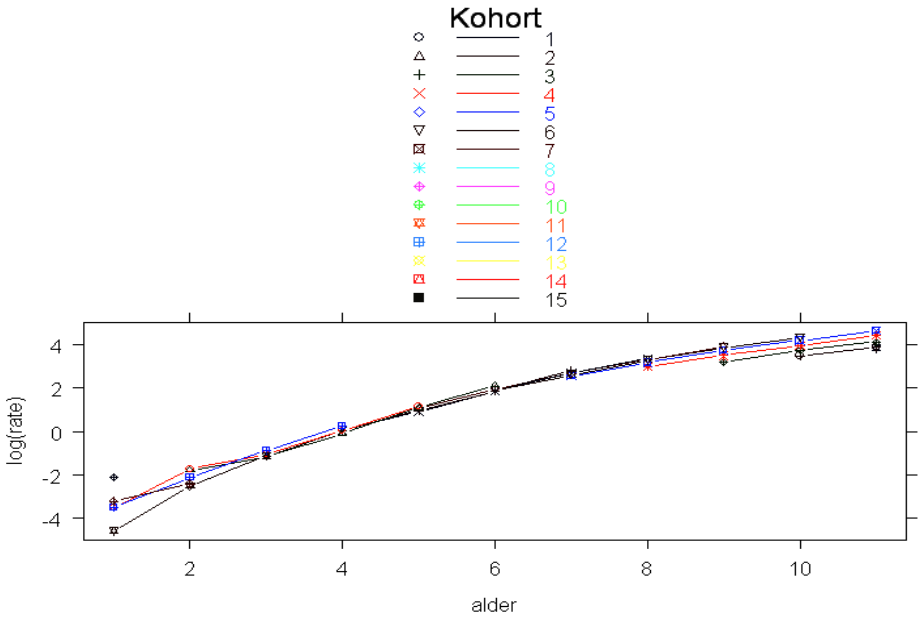
Tabell 4. Aldersspesifikke mortalitetsrater (per 100 000 personår) for blærekreft hos italienske menn i perioden 1955–1979. Antall tilfelle i parentes. (Kilde: WHO mortality database).

Alder\periode	1955–59	1960–64	1965–69	1970–74	1975–79
25 – 29	0,03 (3)	0,03 (3)	0,01 (1)	0,04 (4)	0,12 (12)
30 – 34	0,17 (16)	0,18 (17)	0,12 (11)	0,08 (8)	0,09 (8)
35 – 39	0,32 (24)	0,31 (29)	0,35 (33)	0,42 (39)	0,32 (30)
40 – 44	1,04 (79)	1,05 (76)	0,91 (82)	1,04 (95)	1,27 (115)
45 – 49	2,86 (234)	2,52 (185)	2,61 (183)	3,04 (267)	3,16 (285)
50 – 54	6,64 (458)	7,03 (552)	6,43 (450)	6,46 (431)	8,47 (723)
55 – 59	12,71 (720)	13,39 (867)	14,59 (1069)	14,64 (974)	16,38 (1004)
60 – 64	20,11 (890)	23,98 (1230)	26,69 (1550)	27,55 (1840)	28,53 (1811)
65 – 69	24,40 (891)	33,16 (1266)	42,12 (1829)	47,77 (2395)	50,37 (3028)
70 – 74	32,81 (920)	42,31 (1243)	52,87 (1584)	66,01 (2292)	74,64 (3176)
75 – 79	45,54 (831)	47,94 (937)	62,05 (1285)	84,65 (1787)	104,21 (2659)

(a)

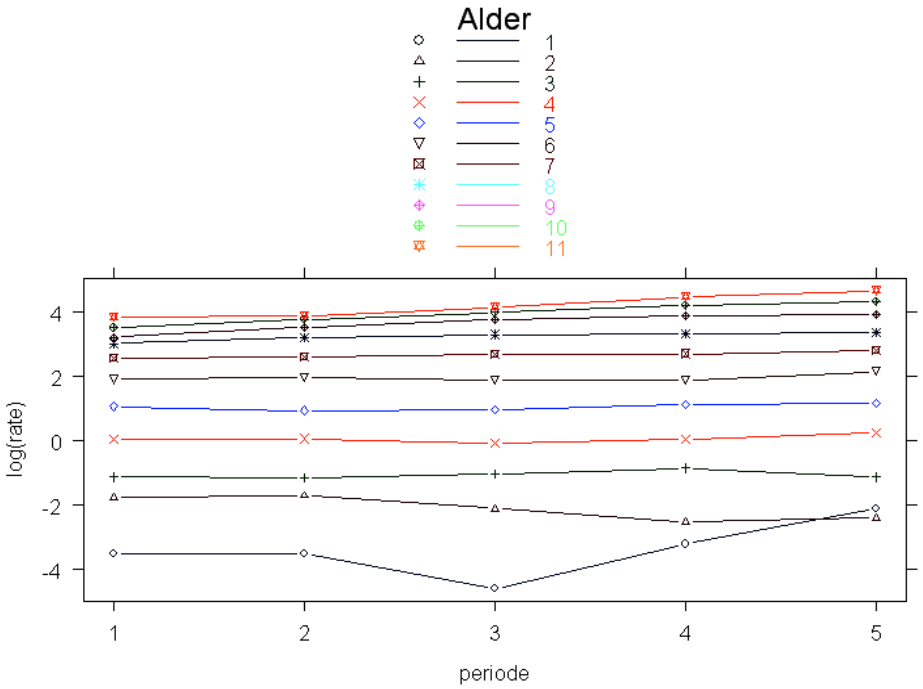


(b)

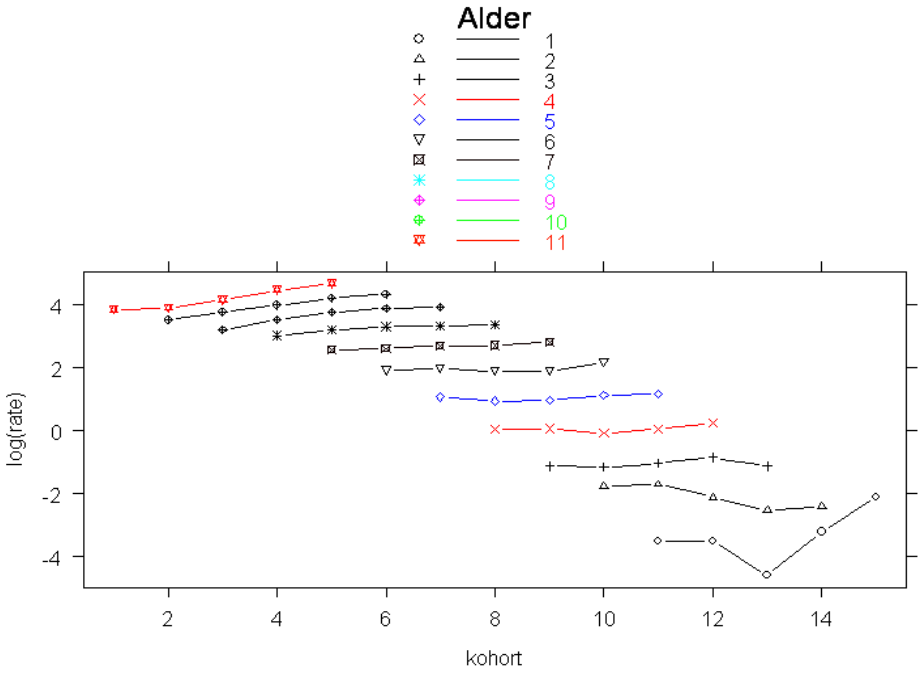


Figur 1. De fire klassiske plottene for rater. Mortalitetsrater (per 100 000 personår) for blærekreft hos italienske menn 1955–1979 i alderen 25–79 år. Øverst (a): aldersspesifikke rater etter periode (plott 1). Nederst (b): aldersspesifikke rater etter kohort (plott 2). Øverst på neste side (c): periodespesifikke rater etter alder (plott 3). Nederst på neste side (d): kohortspesifikke rater etter alder (plott 4).

(c)


Alder
1
2
3
4
5
6
7
8
9
10
11
log(rate)
periode

(d)


Alder
1
2
3
4
5
6
7
8
9
10
11
log(rate)
kohort

Lexis-diagram

Et alternativ til standardtabellene der dataene er ordnet etter alder og periode, er å bruke Lexis-diagram, oppkalt etter Wilhelm Lexis som var en tysk vitenskapsmann som jobbet med samfunnsvitenskap og statistikk. I et Lexis-diagram får vi alle de tre tidsvariablene inn i samme diagram. Et eksempel på et Lexis-diagram er vist på figur 2. I dette diagrammet er kalendertid plottet mot alder, med kalendertid langs førsteaksen og alder langs andreaksen. De horisontale og vertikale linjene representerer gruppeinndelingen for alder og periode. Diagonalene representerer fødselskohorter, kohortene leses på skrå med den eldste kohorten i øvre venstre hjørne og den yngste kohorten i nedre høyre hjørne. Hos Clayton & Schifflers (1987b) har de byttet om på aksene, slik at i dette tilfelle må kohortene leses fra nedre høyre hjørne mot øvre venstre hjørne.

Eksempel

Jeg vil bruke det samme tenkte eksempelet som ble presentert i slutten av kapittel 2, med tre aldersgrupper og tre perioder. Det er Lexis-diagrammet for dette eksempelet som er vist i figur 2. La oss se nærmere på kvadratet i midten. I dette kvadratet finner vi alle som er i alderen 25–29 år i løpet av kalenderperioden 1965–1969 og følgelig er født i kohorten 1935–1944 (i eksempelet er det 7 tilfelle i en populasjon på 9 000 personer). Diagonalen deler dette kvadratet i to, slik at de som er født i 1935–1939 står i trekanten oppe til venstre (i eksempelet 3 tilfelle i en populasjon på 4 000), mens de som er født i 1940–1944 står i trekanten nede til høyre (4 tilfelle i en populasjon på 5 000). Resultatet vil bli at tallene i hver trekant vil representere en aldersgruppe på 5 år, en periode på fem år og en kohort på 5 år. Vi vil altså få ikke-overlappende kohorter av samme lengde som alder og periode (i dette tilfelle 5 år) i motsetning til det vi vil få fra standardtabellene.

I et slikt Lexis-diagram er antall kohorter K gitt ved

$$K = A + P$$

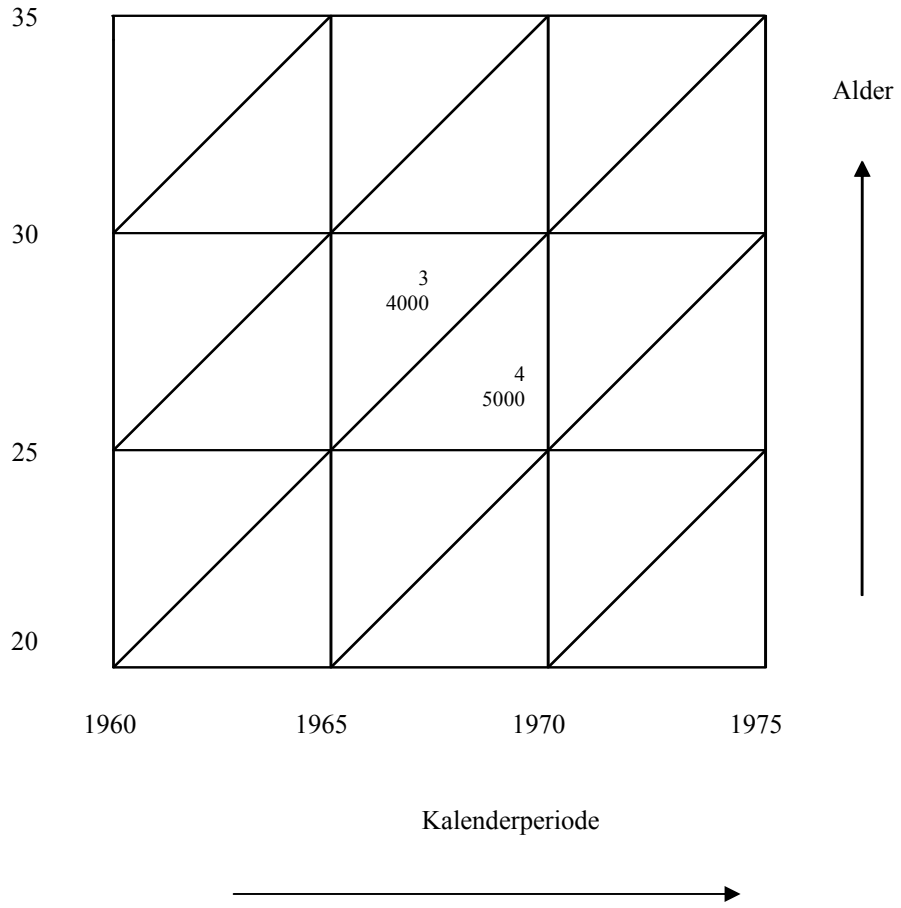
hvor som vanlig A er antall aldersgrupper og P er antall perioder.

Hver kombinasjon av alder og periode gir nå to kohorter. Den eldste av disse er gitt ved

$$k_e = A - a + p$$

og den yngste er gitt ved

$$k_y = A + 1 - a + p$$



Figur 2. Lexis-diagram for et tenkt eksempel med tre aldersgrupper og tre kalenderperioder. I hver av trekantene skal det stå to tall, her er bare tallene for ruten i midten vist. Figuren er forklart videre i hovedteksten.

Data i S-plus

Jeg vil bruke statistikkpakken S-plus (versjonene 6.2 og 7.0 for Windows) når jeg skal implementere modeller og metoder som beskrives seinere i denne oppgaven. Som støttelitteratur i S-plus har jeg brukt Crawley (2002) og Venables & Ripley (2002). Før vi kan foreta statistiske analyser ved hjelp av S-plus må vi sette opp dataene på en annen måte enn slik de står i de opprinnelige tabellene. Den vanligste måten å lagre data på i S-plus er i såkalte datarammer (”data frames”), hvor hovedpoenget er én kolonne for hver variabel.

Eksempel

Vi ser igjen på eksempelet med blærekreft hos italienske menn i perioden 1955–1979 (tabell 4). I denne tabellen er mortalitetsraten oppgitt per 100 000 personår, sammen med antall tilfelle. Derimot er antall personår ikke oppgitt, og antall personår inngår i parametriseringen av modellene som skal testes i seinere kapitler. Men det er ikke noe stort problem siden antall personår enkelt kan beregnes som

$$populasjon = antall \cdot 100000 / rate$$

forutsatt at raten ikke er null. Null-observasjoner takles for øvrig på en grei måte av S-plus, men antall personår må beregnes på en annen måte. Et eksempel på dette er gitt i kapittel 5. Appendiks A viser hvordan den opprinnelige tabellen er blitt omarbeidet til en dataramme.

4

Modeller

Generaliserte lineære modeller (GLM)

Alle modellene brukt i denne oppgaven vil være innenfor rammen av generaliserte lineære modeller. Før jeg forklarer hva vi mener med generaliserte lineære modeller, vil jeg si noe om den eksponentielle familie av fordelinger. Jeg vil i hele dette kapittelet og i resten av oppgaven bruke forkortelsen "log" for den naturlige logaritmen (det vil si $\log(x) = \ln(x)$), og forkortelsen "exp" for antilogaritmen til den naturlige logaritmen (det vil si $\exp(x) = e^x$).

Betrakt en enkelt tilfeldig variabel Y hvis sannsynlighetsfordeling avhenger av en enkelt parameter θ. Fordelingen tilhører den eksponentielle familie hvis tettheten til Y kan skrives på formen

$$f(y;\theta) = \exp[a(y) \cdot b(\theta) + c(\theta) + d(y)] .$$

Hvis $a(y) = y$ sies fordelingen å være på kanonisk form. Uttrykket $b(\theta)$ kalles naturlig parameter.

Eksempel

Tettheten for poissonfordelingen er gitt ved

$$f(y;\theta) = \frac{\theta^y \cdot e^{-\theta}}{y!} \quad \text{y = 0, 1, 2, …}$$

Dette kan omskrives til

$$f(y;\theta) = \exp(y \cdot \log\theta - \theta - \log y!)$$

som viser at poissonfordelingen tilhører den eksponentielle familie med $a(y) = y$ (altså på kanonisk form), $b(\theta) = \log\theta$, $c(\theta) = -\theta$ og $d(y) = -\log y!$.

Eksempel

Tettheten for normalfordelingen er gitt ved

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2 \right)$$

hvor μ er den parameteren vi er interesserte i og σ^2 betraktes som en støyparameter.

Dette kan omskrives til

$$f(y;\mu) = \exp\left(-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right).$$

Dette er på kanonisk form. Videre er $b(\mu) = \frac{\mu}{\sigma^2}$, $c(\mu) = -\frac{\mu^2}{2\sigma^2}$ og

$d(y) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$.

Flere eksempler på fordelinger som tilhører den eksponentielle familie er omtalt i Dobson (2002). Der kan man også lese mer om egenskapene til den eksponentielle familie.

Dobson angir følgende egenskaper som kjennetegn for generaliserte lineære modeller:

1. Responsvariablene $Y_1, Y_2, \ldots, Y_N$ antas å være uavhengige. De deler den samme fordelingen fra den eksponentielle familie og er på kanonisk form, men parameterne θ_i trenger ikke være like alle sammen.

2. I modellene bruker vi vanligvis ikke θ_i, men et mindre antall parametere $\beta_1, \ldots, \beta_p$ (hvor $p < N$). Vi kan samle parameterne i en vektor

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{bmatrix}$$

3. Vi har forklaringsvariabler x_{ij} ($i = 1, 2, \ldots, N$ og $j = 1, 2, \ldots, p$) som kan være målte verdier av kontinuerlige forklaringsvariabler (kovariater) eller nivåer av kategoriske forklaringsvariabler (dummyvariabler). Vi kan samle alle forklaringsvariablene i en designmatrise $\boldsymbol{X}$, hvor

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{N1} & x_{N2} & . & . & x_{Np} \end{bmatrix}$$

4. Det fins en monoton linkfunksjon g slik at

$$g(\mu_i) = \boldsymbol{x}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta}$$

hvor

$$\boldsymbol{x}_{\mathrm{i}}^{\mathrm{T}} = [x_{i1} \quad x_{i2} \quad . \quad . \quad x_{ip}] \text{ og } \mu_i = E(Y_i).$$

Poissonmodeller

Begivenheter som opptrer helt uavhengige av hverandre, kan betraktes som del av en poissonprosess dersom forventet antall begivenheter per tidsenhet er konstant og dersom to begivenheter i tillegg ikke inntrer nøyaktig samtidig. Oversikter over slike begivenheter samles ofte i tabeller. Slike *telledata* modelleres derfor ofte ved hjelp av poissonfordelingen. La for eksempel den stokastiske variabelen Y være antall tilfeller av en sjelden sykdom i løpet av et tidsrom Δt. Det er vanlig å anta at Y i slike sammenhenger er poissonfordelt. Vi har tidligere sett at tettheten til Y er gitt ved

$$f(y) = \frac{\mu^y \cdot e^{-\mu}}{y!} \quad , \quad y = 0, 1, 2, \ldots$$

Her er μ forventet antall tilfeller i løpet av tidsrommet Δt, det vil si at $E(Y) = \mu$. For poissonfordelingen er variansen lik forventningen, det vil si at Var(Y) = μ.

Log-lineære modeller

Log-lineære modeller er generaliserte lineære modeller hvor responsvariabelen Y antas å være poissonfordelt og logaritmen er linkfunksjon. Vi kan skille mellom to hovedvarianter av log-lineære modeller. I den første varianten er minst en av forklaringsvariablene kontinuerlig (betegnes poissonregresjon), i den andre varianten er alle forklaringsvariablene kategoriske. Dobson (2002) begrenser bruken av betegnelsen log-lineære modeller til bare å omfatte den siste varianten, mens jeg vil inkludere også poissonregresjon under betegnelsen log-lineære modeller.

Effekten av forklaringsvariablene på responsvariabelen Y modelleres ved hjelp av parameteren μ. La $Y_1, Y_2, \ldots, Y_N$ være uavhengige stokastiske variable hvor Y_i står for antall begivenheter observert fra eksponering n_i, hvor n_i som regel er en "populasjon" (med en utvidet tolkning av "populasjon"). Vi kan da skrive modellen som

$$E(Y_i) = \mu_i = n_i \cdot \theta_i \ .$$

Det er vanlig å modellere θ_i som $\theta_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, slik at den generaliserte lineære modellen blir

$$E(Y_i) = \mu_i = n_i \cdot \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \text{ hvor } Y_i \sim \text{Poisson } (\mu_i).$$

Logaritmefunksjonen er den naturlige linkfunksjonen. Det gir modellen

$$\log E(Y_i) = \log \mu_i = \log n_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$$

Leddet $\log n_i$ regnes som en kjent konstant og kalles "offset".

Hele modellen kan skrives på matriseform som

$$\log \mathrm{E}(\boldsymbol{Y}) = \log \boldsymbol{\mu} = \log \boldsymbol{N} + \boldsymbol{X}\boldsymbol{\beta}$$

hvor $\boldsymbol{X}$ er designmatrisen og $\boldsymbol{\beta}$ er vektoren som inneholder alle parameterne. Disse to er definert tidligere, dessuten er

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_N \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_N \end{bmatrix} \quad \text{og} \quad \boldsymbol{N} = \begin{bmatrix} n_1 \\ n_2 \\ . \\ . \\ n_N \end{bmatrix}$$

"Maximum likelihood" - estimering i GLM

Det er vanlig å bruke "maximum likelihood" (ML) for å estimere parameterne i generaliserte lineære modeller. Likelihood-funksjonen $L(\theta; y)$ er algebraisk lik tetthetsfunksjonen $f(y; \theta)$. ML-estimatoren for θ maksimerer også log-likelihood-funksjonen $l(\theta; y) = \log L(\theta; y)$.

La $Y_1, Y_2, \ldots, Y_N$ være uavhengige stokastiske variable som oppfyller kravene for en GLM. Anta videre at Y_i er poissonfordelt med $E(Y_i) = \mu_i$. Vi ønsker å estimere parameterne β_1, $\beta_2, \ldots, \beta_p$ som er relatert til Y_i gjennom $E(Y_i) = \mu_i$ og $g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. Log-likelihood-funksjonen til hver Y_i er gitt ved

$$l_i = y_i \cdot \log \mu_i - \mu_i - \log y_i!$$

Log-likelihood-funksjonen for alle Y_i-ene er da

$$l = \sum_{i=1}^{N} l_i \quad .$$

Vi får likningene

$$\frac{\partial l}{\partial \beta_j} = 0 \quad \text{for } j = 1, 2, \ldots, p$$

Disse likningene må løses numerisk.

Generelt for generaliserte lineære modeller får vi følgende likninger på matriseform

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}$$

Her er $\boldsymbol{X}$ designmatrisen og $\boldsymbol{b}$ er estimator for $\boldsymbol{\beta}$. Disse likningene må løses iterativt, fordi både $\boldsymbol{W}$ og $\boldsymbol{z}$ normalt vil avhenge av $\boldsymbol{b}$. $\boldsymbol{W}$ er en diagonalmatrise (N x N) hvor elementene er gitt ved

$$w_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

hvor $\eta_i = g(\mu_i)$.

$\boldsymbol{z}$ har elementene gitt ved

$$z_i = \sum_{k=1}^{p} x_{ik} b_k + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right).$$

Utledning og flere detaljer er vist i Dobson (2002).

Vi antar nå at Y_i er poissonfordelt med $E(Y_i) = \mu_i$. Da er $\text{Var}(Y_i) = \mu_i$, videre er $\eta_i = g(\mu_i) = \log(\mu_i)$. Det gir følgende utregninger

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial \log(\mu_i)}{\partial \mu_i} = \frac{1}{\mu_i} \text{ og}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{\frac{1}{\mu_i}} = \mu_i \ .$$

Vi setter disse resultatene inn i uttrykket for w_{ii}, og får

$$w_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{1}{\mu_i} \cdot (\mu_i)^2 = \mu_i \ .$$

Så setter vi de samme resultatene inn i uttrykket for z_i, det gir

$$z_i = \sum_{k=1}^{p} x_{ik} b_k + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) = \sum_{k=1}^{p} x_{ik} b_k + (y_i - \mu_i) \cdot \frac{1}{\mu_i} = \sum_{k=1}^{p} x_{ik} b_k + \frac{y_i}{\mu_i} - 1 \ .$$

Uttrykkene for w_{ii} og z_i vil bli brukt seinere i oppgaven.

Alder-periode-kohort-modeller

Det er vanlig å bruke poissonmodeller for å modellere virkningen av alder, periode og kohort. Det antas da at raten (insidensraten eller mortalitetsraten) i hver celle i tabellen er konstant. Tradisjonelt har alder, periode og kohort blitt regnet som kategoriske variabler. Variablene

defineres da som faktorer, slik at hver distinkt verdi av de tre variablene representerer ett nivå av faktoren.

La λ stå for raten, der telleren antas å være poissonfordelt. La videre α stå for effekten av alder, β for effekten av periode og γ for effekten av kohort. Den multiplikative alder-periode-kohort-modellen kan da skrives

$$E(\lambda) = \exp(\alpha\beta\gamma)$$

eller som log-lineær faktormodell

$$\log E(\lambda) = \alpha + \beta + \gamma .$$

Clayton & Schifflers (1987a+b) presenterer denne modellen, sammen med flere reduserte modeller. Den enkleste modellen har alder som eneste forklaringsvariabel. Hvis vi bruker de samme symbolene som foran, kan modellen formuleres slik

$$\log E(\lambda) = \alpha$$

Dette er en modell som forutsetter effekt av alder, og der periode og kohort tilsynelatende ikke har noen virkning.

De neste modellene forutsetter virkning av én tidsvariabel i tillegg til alder. Vi ser først på alder-periode-modellen. Den kan skrives

$$\log E(\lambda) = \alpha + \beta$$

Dette er en modell med periodeeffekt, men ingen kohorteffekt. Tilsvarende kan vi skrive alder-kohort-modellen

$$\log E(\lambda) = \alpha + \gamma$$

I denne modellen er det en kohorteffekt, men ingen periodeeffekt.

Clayton & Schifflers (1987a) innfører også begrepet *drift* i forbindelse med et eksempel, der både alder-periode-modellen og alder-kohort-modellen ser ut til å passe like bra. Dette betegner de som en alder-drift-modell eller en log-lineær trend-modell. Den kan formuleres som

$$\log E(\lambda) = \alpha + g(p)$$

der $g(p)$ er en lineær funksjon av p (= periode). Alternativt kan den formuleres som

$$\log E(\lambda) = \alpha + h(k)$$

der $h(k)$ er en lineær funksjon ev k (= kohort).

I de neste kapitlene vil vi se nærmere på disse modellene, og presentere ulike måter modellene kan parametriseres på. For å illustrere de ulike metodene vil modeller og parametriseringer bli testet på konkrete eksempler fra kreftstatistikk.

5

Parametrisering av modeller med én og to variabler

Vi skal se nærmere på hvordan modellene fra forrige kapittel kan parametriseres. I dette kapittelet skal vi se på modeller med én og to variabler. Parametrisering av modeller med tre variabler vil bli gjennomgått i de neste kapitlene. Men før vi starter med det, må begrepet *modell* presiseres, siden modell og parametrisering brukes om hverandre i litteraturen. I denne oppgaven definerer jeg to modeller som ulike dersom 1) antall variabler er forskjellig, 2) hvilke variabler som inngår er forskjellig og/eller 3) en eller flere av variablene ikke er av samme type (kategorisk/kontinuerlig). Ut i fra denne definisjonen kan vi si at en modell kan parametriseres på flere ulike måter, både med hensyn på antall parametere som inngår og hvilke parametere som inngår. For en gitt modell vil likevel de estimerte verdiene være uavhengige av parametriseringen.

Ved vurdering og sammenlikning av ulike modeller er *deviansen* en nyttig størrelse. Vi kan definere deviansen som

$$D = 2 \cdot [l(\boldsymbol{b}_{max}; \boldsymbol{y}) - l(\boldsymbol{b}; \boldsymbol{y})]$$

hvor l ($\boldsymbol{b}_{max}$; $\boldsymbol{y}$) er log-likelihood-funksjonen for en mettet (maksimal) modell, mens l ($\boldsymbol{b}$; $\boldsymbol{y}$) er log-likelihood-funksjonen for en aktuell modell. En mettet modell er en modell som har perfekt tilpasning til data, fordi den har like mange parametere som antall verdier som skal tilpasses. Den aktuelle modellen vil da ha færre parametere enn den mettete modellen.
For de fleste poissonmodeller er deviansen gitt ved

$$D = 2 \cdot \sum_{i=1}^{N} o_i \log\left(\frac{o_i}{e_i}\right)$$

hvor o_i er observert verdi (= y_i) og e_i er estimert verdi (= $\hat{y}_i$). Deviansen for en bestemt modell vil være uavhengig av parametriseringen. Dersom modellen er god vil dessuten deviansen D være tilnærmet kji-kvadratfordelt med ($N - p$) frihetsgrader, hvor N er antall observasjoner og p er antall parametere. Dette kan brukes til å undersøke en modells "godhet" ("goodness of fit"). Deviansen kan også brukes til å sammenlikne modellers "godhet". Dersom vi har to "gode" modeller med henholdsvis p og q parametere, så er $D_1 \sim \chi_{N-p}$ og $D_2 \sim \chi_{N-q}$. Da vil $\Delta D \sim \chi_{p-q}$ (p > q). Mer om egenskapene til deviansen kan man finne i for eksempel Dobson (2002).

De ulike parametriseringene vil bli belyst med eksempler. Jeg vil starte med den enkleste modellen, med alder som eneste forklaringsvariabel. Sammen med denne enkle modellen vil jeg ta opp noen generelle problemer i forbindelse med parametrisering av modeller. Deretter vil jeg studere modeller med to tidsvariabler, mens modeller med alle tre tidsvariablene vil bli behandlet i de neste kapitlene.

Estimering av alderseffekter

Vi skal først se på en modell med alder som eneste faktor. En måte å skrive denne modellen på er

$$\log E(\lambda_{ap}) = \alpha_a$$

hvor λ_{ap} er raten i aldersgruppe *a* og kalenderperiode *p* og hvor α_a er effekten av alder i aldersgruppe *a*. Når modellen skal parametriseres er det vanlig å utnytte at

$$\lambda_{ap} = Y_{ap} / n_{ap}$$

hvor Y_{ap} er antall tilfelle i aldersgruppe *a* og kalenderperiode *p* og n_{ap} er antall personår i aldersgruppe *a* og kalenderperiode *p*. I litteraturen er det vanlig å skrive $\log Y_{ap}$ i stedet for det mer korrekte $\log E(Y_{ap})$ i modeller. Jeg vil tillate meg å gjøre det samme i resten av denne oppgaven. Da kan vi skrive en mulig parametrisering av denne modellen som

$$\log Y_{ap} = \alpha_a + \log n_{ap}$$

Vi vil anta at n_{ap} er en konstant både for denne parametriseringen og alle øvrige parametriseringer. Vi skal estimere α_a for *a* = 1, 2, . . . , *A*, hvor *A* er antall aldersgrupper. Denne parametriseringen vil altså gi *A* parametere, vi vil få *A* uavhengige likninger og rangen til designmatrisen ***X*** blir også *A*. Det vil si at likningene har en entydig løsning, og vi kan derfor finne entydige estimater for parameterverdiene for α_a. Tolkningen av parameterne for denne parametriseringen er rett frem, idet α_a vil representere gjennomsnittlig log(rate) i aldersgruppe *a*.

En annen mulig parametrisering av den samme modellen er

$$\log Y_{ap} = \mu + \alpha_a + \log n_{ap}$$

Vi kan betrakte μ som gjennomsnittlig log(rate) for alle aldersgruppene, mens α_a er avvik i log(rate) i gruppe *a* i forhold til μ. Denne parametriseringen vil gi (*A* + 1) parametere (μ, α_1, . . . , α_A), men bare *A* uavhengige likninger. Rangen til designmatrisen **X** blir også bare *A*, og det fører til at $\mathbf{X}^T\mathbf{X}$ blir singulær. Likningene gir flere mulige løsninger, vi får et ubestemthetsproblem, det vil si at vi får flere mulige verdier av de ulike parameterne. Det er vanlig å løse dette problemet ved å legge inn bestemte *restriksjoner*.

Restriksjoner

For å løse ubestemthetsproblemet, og få identifiserbare parametere, legger vi inn restriksjoner. Restriksjonene består i enten å øke antall likninger eller å redusere antall parametere. Vi skal se på to av de vanligste restriksjonene som brukes.

Den første er "hjørnepunkt-restriksjoner". Den består i at vi velger én gruppe som referansegruppe, ofte gruppe 1, men det er ikke noe i veien for å velge en annen referansegruppe. Dersom gruppe 1 blir referansegruppe setter vi $\alpha_1 = 0$, eller generelt med gruppe *a* som referansegruppe setter vi $\alpha_a = 0$. De andre gruppene blir da uansett sammenliknet med referansegruppen. Med denne restriksjonen blir antall parametere redusert fra $(A + 1)$ til A, vi får entydige løsninger av likningene og vi får identifiserbare parametere.

Den andre er "sum null-restriksjoner". Den består i at summen av parameterne for en faktor settes lik null, det vil si at vi setter $\sum_{a=1}^{A} \alpha_a = 0$. Dette fører til at vi får en ekstra likning. Da har vi (A + 1) likninger, det vil si like mange likninger som parametere, noe som gjør det mulig å finne entydige løsninger og dermed identifiserbare parametere.

Kontraster

Begrepet *kontraster* kommer opprinnelig fra variansanalyse (ANOVA). Her brukes kontraster til å sammenlikne gjennomsnitt eller grupper av gjennomsnitt med andre gjennomsnitt eller grupper av gjennomsnitt. Vi så foran på parametriseringen

$$\log Y_{ap} = \mu + \alpha_a + \log n_{ap}$$

som altså har flere parametere enn det er mulig å regne ut, vi sier at modellen er *overparametrisert*. Dette kan altså løses ved å innføre en eller annen form for restriksjoner. Den tilsynelatende enkleste måten å løse dette på i praksis er å sette $\mu = 0$. Det er i prinsippet det samme som å fjerne μ, og da er vi over i den første parametriseringen. Dette går greit så lenge det bare er én faktor, men løser ikke problemet hvis det er to eller flere faktorer.

En annen måte å få innført restriksjoner på er å sette en eller annen kontrast. De fleste statistikkprogrammer har flere standardkontraster som man kan velge mellom, alternativt kan man lage kontrastene selv ved å lage en kontrastmatrise $\boldsymbol{C}$. I eksempelet nedenfor vil jeg sammenlikne og kommentere tre standardkontraster som brukes i S-plus.

Eksempel

I kapittel 3 (tabell 2) så vi på insidensraten for blærekreft hos menn i kreftregisteret for Birmingham i perioden 1960–1976. Vi skal estimere effekten av alder, først ved å bruke den

første parametriseringen, deretter ved å bruke tre ulike kontraster med utgangspunkt i den andre parametriseringen. Antall personår, som inngår i begge parametriseringene, er ikke oppgitt i tabellen, men kan enkelt beregnes som vist i kapittel 3. Det er ett unntak, i aldersgruppe 2 (30–34 år) og periode 1 (1960–62) er raten lik null, da er det umulig å regne ut antall personår. I stedet har jeg valgt å sette antall personår til gjennomsnittet av de to nabotallene.

Tabell 5 viser de estimerte parameterverdiene for ulike parametriseringer og ulike kontraster. Resultatene for den første parametriseringen står i kolonne 2. I dette tilfellet mangler referansenivået ($\mu = 0$) og α_a kan tolkes direkte som gjennomsnittsverdien av log (rate) i hver aldersgruppe. I de tre siste kolonnene står resultatene for de andre parametriseringene med et konstantledd μ.

I kolonne 3 er det brukt en kontrast som kalles "treatment". Denne kontrasten tilsvarer det å bruke hjørnepunkt-restriksjoner. I S-plus betyr det at gruppe 1 velges som referansegruppe, det vil si $\alpha_1 = 0$. Vi kan tolke μ som gjennomsnittsverdien i aldersgruppe 1. Videre kan α_a tolkes som avvik i log(rate) i gruppe *a* i forhold til gruppe 1. For eksempel vil log(rate) i gruppe 2 kunne regnes ut som $\mu + \alpha_2$.

I kolonne 4 er det brukt en kontrast som kalles "sum", som tilsvarer sum null-restriksjoner. Det vil si at $\sum_{a=1}^{11} \alpha_a = 0$, og μ kan tolkes som gjennomsnittlig log(rate) for alle gruppene samlet. Videre kan vi tolke α_a som avvik i log(rate) i gruppe *a* i forhold til gjennomsnittet for alle gruppene. Også i dette tilfelle vil log(rate) i gruppe 2 kunne regnes ut som $\mu + \alpha_2$. Legg merke til at α_{11} ikke er estimert, men kan finnes som $\alpha_{11} = -\sum_{a=1}^{10} \alpha_a$.

I kolonne 5 er det brukt en kontrast som kalles "Helmert". I likhet med kontrasten foran vil μ representere gjennomsnittlig log(rate) for alle gruppene samlet. Derimot er det vanskelig å finne noen meningsfull direkte tolkning av α_a. Til tross for dette er "Helmert" standardkontrast i S-plus. Grunnen til det er at "Helmert"-kontrastene er *ortogonale*. To kontraster er ortogonale til hverandre hvis sammenlikningen er statistisk uavhengig. Ortogonale kontraster leder til en ortogonal kovariansmatrise. Fordelen med ortogonale

kontraster er at p-verdier indikerer utvetydig om en spesiell kontrast er signifikant. Verken ”treatment”-kontrasten eller ”sum”-kontrasten er ortogonal.

Tabell 5. Blærekreft i Birmingham 1960-1976. Estimerte parameterverdier for virkningen av alder. Fire ulike parametriseringer av samme modell, de tre siste med et konstantledd μ.

Parameter	Uten μ	Kontrast ”treatment”	Kontrast ”sum”	Kontrast ”Helmert”
μ	0,000	- 11,972	- 8,828	- 8,828
α_1	- 11,972	0,000	- 3,144	0,226
α_2	- 11,521	0,451	- 2,693	0,387
α_3	- 10,585	1,387	- 1,757	0,349
α_4	- 9,962	2,010	- 1,134	0,403
α_5	- 8,996	2,976	- 0,168	0,356
α_6	- 8,474	3,498	0,354	0,331
α_7	- 7,933	4,039	0,894	0,316
α_8	- 7,392	4,580	1,436	0,285
α_9	- 7,041	4,931	1,787	0,263
α_{10}	- 6,690	5,282	2,138	0,229
α_{11}	- 6,540	5,432	-	-

Vi skal se på den matematiske sammenhengen mellom de ulike parameterne foran. Vi tar utgangspunkt i parameterne i kolonne 2, og definerer disse som $\alpha_1, \alpha_2, \ldots, \alpha_{11}$. Videre definerer vi parameterne i de øvrige kolonnene (ikke μ) som $\alpha_1 *, \alpha_2 *, \ldots, \alpha_{11} *$, slik at $\alpha_a *$ får ulik tolkning i de tre kolonnene.

I kolonne 3 er sammenhengen mellom α_a og $\alpha_a *$ gitt ved

$$\alpha_a * = \alpha_a - \alpha_1$$

For eksempel er

$$\alpha_1 * = \alpha_1 - \alpha_1 = 0 \text{ og}$$

$$\alpha_2 * = \alpha_2 - \alpha_1 = -11{,}521 - (-11{,}972) = 0{,}451$$

I kolonne 4 er sammenhengen mellom α_a og $\alpha_a *$ gitt ved

$$\alpha_a^* = \alpha_a - \bar{\alpha} \qquad \text{hvor} \qquad \bar{\alpha} = \frac{1}{11}\sum_{a=1}^{11}\alpha_a$$

For eksempel er

$$\alpha_1^* = \alpha_1 - \bar{\alpha} = -11{,}972 - (-8{,}828) = -3{,}144$$

I kolonne 5 er sammenhengen mellom α_a og α_a^* gitt ved

$$\alpha_a^* = \frac{1}{a+1}\left(\alpha_{a+1} - \frac{1}{a}\left(\sum_{i=1}^{a}\alpha_i\right)\right)$$

For eksempel er

$$\alpha_1^* = \frac{1}{1+1}\left(\alpha_{1+1} - \frac{1}{1}\left(\sum_{i=1}^{1}\alpha_i\right)\right) = \frac{1}{2}(\alpha_2 - \alpha_1) = \frac{1}{2}(-11{,}521 - (-11{,}972)) = 0{,}226 \text{ og}$$

$$\alpha_2^* = \frac{1}{2+1}\left(\alpha_{2+1} - \frac{1}{2}\left(\sum_{i=1}^{2}\alpha_i\right)\right) = \frac{1}{3}\left(\alpha_3 - \frac{1}{2}(\alpha_1 + \alpha_2)\right) = \frac{1}{3}(-10{,}585 - \frac{1}{2}(-11{,}972 - 11{,}521)) = 0{,}387$$

Parametrisering av alder-periode-modeller

Jeg vil nå studere modeller med to faktorer, først vil jeg se på en modell med kalenderperiode som faktor i tillegg til alder. Jeg vil først presentere to ulike parametriseringer av alder-periode-modellen, og deretter vil jeg se nærmere på noen konkrete eksempler for å illustrere anvendelsen av modellen.

En mulig parametrisering av denne modellen er

$$\log Y_{ap} = \mu + \alpha_a + \beta_p + \log n_{ap}$$

Igjen står Y_{ap} for antall tilfelle i aldersgruppe *a* og kalenderperiode *p* og n_{ap} står for antall personår i aldersgruppe *a* og kalenderperiode *p*. Vi skal estimere μ, α_a for $a = 1, 2, \ldots, A$ og β_p for $p = 1, 2, \ldots, P$. Dette gir $(A + P + 1)$ parametere, mens antall uavhengige likninger bare er $(A + P - 1)$. Vi får det samme ubestemthetsproblemet som vi tok opp i forrige underkapittel. Vi må innføre en form for restriksjoner. Dersom vi for eksempel bruker hjørnepunkt-restriksjoner, kan vi velge å sette $\alpha_1 = 0$ og $\beta_1 = 0$. Da blir antall parametere redusert til (A + P – 1) som er lik antall likninger, vi får entydige løsninger av

likningene og identifiserbare parametere. For denne parametriseringen og med disse restriksjonene kan μ betraktes som gjennomsnittlig log(rate) for referansegruppen, det vil si log(rate) i aldersgruppe 1 og kalenderperiode 1, mens α_a er endring i log(rate) i aldersgruppe *a* sammenliknet med aldersgruppe 1 og $\exp(\beta_p)$ kan tolkes som relativ risiko i kalenderperiode *p* sammenliknet med kalenderperiode 1.

En alternativ parametrisering av den samme modellen er

$$\log Y_{ap} = \alpha_a + \beta_p + \log n_{ap}$$

Vi skal estimere α_a for $a = 1, 2, \ldots, A$ og β_p for $p = 1, 2, \ldots, P$. Dette gir $(A + P)$ parametere, mens antall uavhengige likninger bare er $(A + P - 1)$. Vi får altså det samme ubestemthetsproblemet her også. I dette tilfelle er det nok å innføre én restriksjon. Dersom vi igjen velger hjørnepunkt-restriksjoner kan vi for eksempel sette $\beta_1 = 0$. Da blir antall parametere redusert til $(A + P - 1)$. For denne parametriseringen kan vi tolke $\exp(\alpha_a)$ som tilpassete aldersspesifikke insidensrater eller mortalitetsrater for referanseperioden, det vil her si periode 1, mens $\exp(\beta_p)$ kan tolkes som tilpasset relativ risiko i periode *p* sammenliknet med referanseperioden.

I eksemplene som følger foretrekker jeg å bruke denne siste parametriseringen med hjørnepunkt-restriksjoner. Jeg vil også velge kalenderperiode 1 som referanseperiode, det vil si $\beta_1 = 0$. Det betyr videre at jeg vil bruke "treatment"-kontrasten i S-plus. Jeg velger denne parametriseringen og denne restriksjonen blant annet fordi det gjør det relativt lett å tolke parameterverdiene.

Eksempel

Det første eksempelet gjelder blærekreft hos menn i Birmingham i perioden 1960–1976. Dette eksempelet ble presentert i kapittel 3, og dataene står i tabell 2. De estimerte parameterverdiene for disse dataene er vist i tabell 6. Vi ser at insidensraten øker med alderen. Vi ser også at den relative risikoen øker med kalenderperiode, og at den gjør et sprang mellom periode 2 og periode 3. For å undersøke om alder-periode-modellen er en modell som viser god tilpasning til data, kan vi se på deviansen. Deviansen $D = 41{,}1$ og antall frihetsgrader $df = 30$. Som nevnt i innledningen av dette kapitlet er antall frihetsgrader *df* lik antall observasjoner *N* minus antall estimerte parametere *p*, slik at her blir $df = N - p =$

44 – 14 =30. Det ble også nevnt at ved god tilpasning er deviansen tilnærmet kji-kvadratfordelt. Her er altså $D = 41{,}1$ og $df = 30$, noe som ikke er signifikant. Dette kan tyde på at alder-periode-modellen er en god modell for disse dataene, og at det er en periodeeffekt. I en modell med bare alder som faktor er $D = 327$ og $df = 33$. Inkludering av periode som faktor i modellen gir altså en signifikant forbedring i tilpasning.

Tabell 6. Blærekreft hos menn i Birmingham 1960–1976. Alder- (α_a) og periode- (β_p) parametere estimert fra ratene i tabell 2.

Alder	α_a	Periode	β_p
25 – 29	- 12,314	1960 – 62	0,000
30 – 34	- 11,848	1963 – 66	0,088
35 – 39	- 10,898	1968 – 72	0,485
40 – 44	- 10,279	1973 – 76	0,503
45 – 49	- 9,318		
50 – 54	- 8,796		
55 – 59	- 8,257		
60 – 64	- 7,728		
65 – 69	- 7,392		
70 – 74	- 7,036		
75 – 79	- 6,874		

Eksempel

Det neste eksempelet vi skal se på er forekomsten av blærekreft hos italienske menn i perioden 1955–1979. Dette eksempelet ble også presentert i kapittel 3, og dataene står i tabell 4. De estimerte parameterverdiene for disse dataene er vist i tabell 7. Vi ser at mortalitetsraten øker med alderen. Vi ser også at den relative risikoen øker med kalenderperiode, men denne gangen uten store sprang. For å undersøke om alder-periode-modellen viser god tilpasning til data, kan vi igjen se på deviansen. Deviansen $D = 512{,}5$ og antall frihetsgrader $df = 40$, hvilket er klart signifikant. Dette tyder på at alder-periode-modellen ikke er noen spesielt god modell for disse dataene, og at vi bør forsøke å finne en bedre modell.

Tabell 7. Blærekreft hos italienske menn 1955–1979. Alder- (α_a) og periode- (β_p) parametere estimert fra ratene i tabell 4.

Alder	α_a	Periode	β_p
25 – 29	-14,916	1955 – 59	0,000
30 – 34	-13,892	1960 – 64	0,157
35 – 39	-12,913	1965 – 69	0,317
40 – 44	-11,792	1970 – 74	0,461
45 – 49	-10,797	1975 – 79	0,580
50 – 54	- 9,889		
55 – 59	- 9,176		
60 – 64	- 8,614		
65 – 69	- 8,155		
70 – 74	- 7,842		
75 – 79	- 7,596		

Eksempel

I det neste eksempelet skal vi se på forekomsten av lungekreft hos belgiske kvinner i perioden 1955–1978 (Clayton & Schifflers 1987a). De aldersspesifikke mortalitetsratene er presentert i tabell 8.

Tabell 8. Aldersspesifikke mortalitetsrater (per 100 000 personår) for lungekreft hos belgiske kvinner i perioden 1955–1978. Antall tilfelle i parentes. (Kilde: WHO mortality database).

Alder\periode	1955 – 59	1960 – 64	1965 – 69	1970 – 74	1975 – 78
25 – 29	0,19 (3)	0,13 (2)	0,50 (7)	0,19 (3)	0,70 (10)
30 – 34	0,66 (11)	0,98 (16)	0,72 (11)	0,71 (10)	0,57 (7)
35 – 39	0,78 (11)	1,32 (22)	1,47 (24)	1,64 (25)	1,32 (15)
40 – 44	2,67 (36)	3,16 (44)	2,53 (42)	3,38 (53)	3,93 (48)
45 – 49	4,84 (77)	5,60 (74)	4,93 (68)	6,05 (99)	6,83 (88)
50 – 54	6,60 (106)	8,50 (131)	7,65 (99)	10,59 (142)	10,42 (134)
55 – 59	10,36 (157)	12,00 (184)	12,68 (189)	14,34 (180)	17,95 (177)
60 – 64	14,76 (193)	16,37 (232	18,00 (262)	17,60 (249)	23,91 (239)
65 – 69	20,53 (219)	22,60 (267)	24,90 (323)	24,33 (325)	32,70 (343)
70 – 74	26,24 (223)	27,70 (250)	30,47 (308)	36,94 (412)	38,47 (358)
75 – 79	33,47 (198)	33,61 (214)	36,77 (253)	43,69 (338)	45,20 (312)

De estimerte parameterverdiene for disse dataene er vist i tabell 9. Vi ser nok en gang at mortalitetsraten øker med alderen, og nok en gang øker den relative risikoen med kalenderperiode. Vi ønsker også denne gang å undersøke hvor god tilpasning alder-periode-modellen viser til de oppgitte dataene ved å se på deviansen. Deviansen $D = 38{,}5$ og antall frihetsgrader $df = 40$, dette er ikke signifikant. Dette kan tyde på at alder-periode-modellen er en god modell for disse dataene, og at det er en periodeeffekt akkurat som i det første eksemplet. I en modell med bare alder som faktor er til sammenlikning $D = 197$ og $df = 44$, med andre ord vil inkludering av periode som faktor i modellen gi en signifikant forbedring i tilpasning.

Tabell 9. Lungekreft hos belgiske kvinner 1955–1978.
Alder- (α_a) og periode- (β_p) parametere estimert
fra ratene i tabell 8.

Alder	α_a	Periode	β_p
25 – 29	-12,816	1955 – 59	0,000
30 – 34	-12,007	1960 – 64	0,107
35 – 39	-11,430	1965 – 69	0,162
40 – 44	-10,581	1970 – 74	0,278
45 – 49	- 9,985	1975 – 78	0,423
50 – 54	- 9, 548		
55 – 59	- 9,124		
60 – 64	- 8,825		
65 – 69	- 8,502		
70 – 74	- 8,250		
75 – 79	- 8,065		

Eksempel

I det siste eksempelet skal vi se på forekomsten av prostatakreft hos ikke-hvite menn i USA i perioden 1935–1969. Dette eksempelet ble først presentert i kapittel 3, og dataene står i tabell 3. De estimerte parameterverdiene for disse dataene er vist i tabell 10. Også i dette siste eksempelet ser vi at mortalitetsraten øker med alderen, og at den relative risikoen øker med kalenderperiode. For å undersøke om alder-periode-modellen viser god tilpasning til data, ser vi enda en gang på deviansen. Deviansen $D = 721$ og antall frihetsgrader $df = 42$, noe som er klart signifikant. Alder-periode-modellen ser med andre ord ikke ut til å beskrive disse dataene spesielt godt, i likhet med det resultatet vi fikk i eksempel nummer 2.

Tabell 10. Prostatakreft hos ikke-hvite menn i USA 1935–1969. Alder- (α_a) og periode- (β_p) parametere estimert fra antall dødsfall og antall personår i tabell 3.

Alder	α_a	Periode	β_p
50 – 54	- 7,775	1935 – 39	0,000
55 – 59	- 6,959	1940 – 44	0,201
60 – 64	- 6,243	1945 – 49	0,409
65 – 69	- 5,719	1950 – 54	0,605
70 – 74	- 5,278	1955 – 59	0,711
75 – 79	- 4,970	1960 – 64	0,757
80 – 84	- 4,836	1965 – 69	0,812

Parameterverdiene for β_p som er presentert i disse fire eksemplene er ikke absolutte. Ved å velge en annen parametrisering kunne vi fått andre parameterverdier for β_p, men førstedifferansene $\beta_{p+1} - \beta_p$ vil være uavhengige av hvilken parametrisering som blir valgt. Det vil si at den relative risikoen mellom to perioder vil være reell, og ikke bare et utslag av den valgte parametriseringen.

Parametrisering av alder-kohort-modeller

I dette underkapittelet vil jeg se nærmere på en modell med fødselskohort som faktor i tillegg til alder. Tabeller over kreftrater er typisk ordnet etter alder og kalenderperiode, men som vist i kapittel 2 kan vi i noen tilfelle finne kohorten på en enkel måte. Dersom for eksempel alder *a* og periode *p* er delt i like lange intervaller er kohorten *k* gitt ved

$$k = A - a + p$$

hvor *A* er antall aldersgrupper.

I likhet med alder-periode-modellen kan alder-kohort-modellen parametriseres på flere ulike måter. Det er ikke noe prinsipielt nytt ved parametriseringen av alder-kohort-modellen i

forhold til alder-periode-modellen. Jeg vil derfor bare se på én parametrisering av alder-kohort-modellen. En mulig parametrisering av denne modellen er

$$\log Y_{ak} = \alpha_a + \gamma_k + \log n_{ak}$$

hvor Y_{ak} er antall tilfelle i aldersgruppe a og fødselskohort k, n_{ak} er antall personår i aldersgruppe a og kohort k, og regnes også her som konstant. Vi skal estimere α_a for $a = 1, 2, \ldots, A$ og γ_k for $k = 1, 2, \ldots, K$. Her er antall parametere lik $(A + K)$, mens antall uavhengige likninger bare er $(A + K - 1)$. Vi får tilsvarende ubestemthetsproblem som i forrige underkapittel. Vi må innføre restriksjoner. Dersom vi velger å bruke hjørnepunkt-restriksjoner, må vi velge en referansegruppe. Det er vanlig å bruke første gruppe som referanse, men den eldste kohorten omfatter bare én observasjon. Jeg foretrekker derfor å bruke en av gruppene som omfatter flere observasjoner som referansekohort. For denne parametriseringen og med denne restriksjonen kan vi tolke $\exp(\alpha_a)$ som tilpassete aldersspesifikke insidensrater eller mortalitetsrater for referansekohorten, mens $\exp(\gamma_k)$ kan tolkes som tilpasset relativ risiko i kohort k sammenliknet med referansekohorten.

Jeg vil bruke de samme eksemplene som i forrige underkapittel, unntatt det første eksempelet. I det første eksempelet er kalenderperiode delt i intervaller av varierende lengde, noe som gjør beregning av kohorter vanskelig, og tolkningen av resultatet uklart. Jeg vil bruke hjørnepunkt-restriksjoner med den midterste kohorten som referansekohort. Standardkontrasten "treatment" i S-plus bruker alltid den første gruppen som referanse. Siden jeg ønsker en av de andre kohortene som referanse må jeg lage kontrastmatrisene selv.

Eksempel

Det første eksempelet vi skal studere er forekomsten av blærekreft hos italienske menn i perioden 1955–1979 (data i tabell 4 i kapittel 3). Antall kohorter er 15, og som referansekohort velger jeg kohort nummer 8 (1910–1919), det vil si at $\gamma_8 = 0$. De estimerte parameterverdiene er vist i tabell 11. Vi så i forrige underkapittel at den relative risikoen øker med kalenderperiode, og vi ser her at den relative risikoen også ser ut til å øke med fødselskohort, riktignok med noen mindre avvik i de yngre kohortene. For å undersøke om alder-kohort-modellen viser god tilpasning til data, ser vi på deviansen. Deviansen $D = 39{,}4$ og antall frihetsgrader $df = 30$, dette er ikke signifikant. Dette kan tyde på at alder-kohort-modellen er en god modell for disse dataene, og at det er en kohorteffekt. Vi så på det samme

eksempelet i forbindelse med alder-periode-modellen, der var $D = 512{,}5$ og $df = 40$, med andre ord ikke spesielt god tilpasning. Alder-kohort-modellen viser en så markant forbedring i tilpasning i forhold til alder-periode-modellen at vi her må kunne si at disse dataene har en kohorteffekt fremfor en periodeeffekt. Implementeringen av dette eksempelet i S-plus er vist i appendiks B.

Tabell 11. Blærekreft hos italienske menn 1955–1979.
Alder- (α_a) og kohort- (γ_k) parametere estimert
fra ratene i tabell 4.

Alder	α_a	Kohort	γ_k
25 – 29	- 15,228	1875 – 84	- 0,987
30 – 34	- 13,705	1880 – 89	- 0,908
35 – 39	- 12,733	1885 – 94	- 0,667
40 – 44	- 11,623	1890 – 99	- 0,381
45 – 49	- 10,563	1895 – 04	- 0,165
50 – 54	- 9,607	1900 – 09	- 0,053
55 – 59	- 8,823	1905 – 14	0,006
60 – 64	- 8,173	1910 – 19	0,000
65 – 69	- 7,609	1915 – 24	0,059
70 – 74	- 7,144	1920 – 29	0,194
75 – 79	- 6,707	1925 – 34	0,197
		1930 – 39	0,327
		1935 – 44	- 0,057
		1940 – 49	- 0,023
		1945 – 54	1,595

Eksempel

I det neste eksempelet skal vi se på forekomsten av lungekreft hos belgiske kvinner i perioden 1955–1978 (data i tabell 8). Som i forrige eksempel er antallet kohorter 15, og akkurat som i forrige eksempel velger jeg kohort nummer 8 (1910–1919) som referansekohort, det vil si at $\gamma_8 = 0$. De estimerte parameterverdiene er vist i tabell 12. Vi har tidligere sett at den relative risikoen øker med kalenderperiode, og det ser ut til at den relative risikoen også øker med fødselskohort, det vil si at de eldste kohortene har minst risiko. Riktignok ser det ut for at det også her er noen mindre avvik fra mønsteret blant de yngre kohortene, i likhet med forrige eksempel. For å undersøke om alder-kohort-modellen viser god tilpasning til disse dataene,

ser vi igjen på deviansen. Deviansen $D = 29{,}7$ og antall frihetsgrader $df = 30$. Dette indikerer at dataene har en god tilpasning til denne modellen. Vi har tidligere sett at disse dataene også har en god tilpasning til alder-periode-modellen med $D = 38{,}5$ og $df = 40$. Vi har altså et eksempel på et datasett som har like god tilpasning til alder-periode-modellen som til alder-kohort-modellen. Vi har altså en tydelig tidseffekt, men problemer med å finne ut om denne effekten er relatert til periode eller kohort. Vi skal studere dette problemet nærmere i neste underkapittel.

Tabell 12. Lungekreft hos belgiske kvinner 1955–1978. Alder- (α_a) og kohort- (γ_k) parametere estimert fra ratene i tabell 8.

Alder	α_a	Kohort	γ_k
25 – 29	- 13,210	1875 – 84	- 0,663
30 – 34	- 12,094	1880 – 89	- 0,649
35 – 39	- 11,522	1885 – 94	- 0,563
40 – 44	- 10,593	1890 – 99	- 0,461
45 – 49	- 9,888	1895 – 04	- 0,331
50 – 54	- 9,331	1900 – 09	- 0,276
55 – 59	- 8,811	1905 – 14	- 0,128
60 – 64	- 8,397	1910 – 19	0,000
65 – 69	- 7,962	1915 – 24	0,139
70 – 74	- 7,606	1920 – 29	0,152
75 - 79	- 7,340	1925 – 34	0,325
		1930 – 39	0,410
		1935 – 44	0,389
		1940 – 49	0,024
		1945 – 54	1,340

Eksempel

I det siste eksempelet skal vi se på forekomsten av prostatakreft hos ikke-hvite menn i USA i perioden 1935–1969 (data i tabell 3 i kapittel 3). Antallet kohorter er 13, og jeg velger kohort nummer 7 (1880–1889) som referansekohort, det vil si at $\gamma_7 = 0$. De estimerte parameterverdiene er vist i tabell 13. Vi ser at den relative risikoen er minst for den eldste kohorten, og at den øker frem til og med den tiende kohorten, deretter avtar den relative risikoen igjen. For å undersøke om alder-kohort-modellen viser god tilpasning til disse dataene, ser vi nok en gang på deviansen. Deviansen $D = 127{,}4$ og antall frihetsgrader $df =$

30, noe som er klart signifikant. Dette indikerer at dataene ikke passer spesielt godt med alder-kohort-modellen, og vi har tidligere sett at heller ikke alder-periode-modellen beskriver disse dataene på en god måte.

Tabell 13. Prostatakreft hos ikke-hvite menn i USA 1935–1969. Alder- (α_a) og kohort- (γ_k) parametere estimert fra antall dødsfall og antall personår i tabell 3.

Alder	α_a	Kohort	γ_k
50 – 54	- 7,319	1850 – 59	- 1,307
55 – 59	- 6,499	1855 – 64	- 0,942
60 – 64	- 5,766	1860 – 69	- 0,784
65 – 69	- 5,185	1865 – 74	- 0,534
70 – 74	- 4,632	1870 – 79	- 0,297
75 – 79	- 4,184	1875 – 84	- 0,156
80 – 84	- 3,863	1880 – 89	0,000
		1885 – 94	0,203
		1890 – 99	0,246
		1895 – 04	0,285
		1900 – 09	0,154
		1905 – 14	0,062
		1910 – 19	- 0,070

I likhet med parameterverdiene for β_p i alder-periode-modellen er heller ikke parameterverdiene for γ_k i alder-kohort-modellen absolutte, men også for alder-kohort-modellen er førstedifferansene $\gamma_{k+1} - \gamma_k$ uavhengige av parametriseringen. Med andre ord er den relative risikoen mellom to kohorter reell, og ikke et resultat av den valgte parametriseringen.

Parametrisering av alder-drift-modeller

I forrige underkapittel så vi eksempel på et datasett som viste god tilpasning både til alder-periode-modellen og til alder-kohort-modellen. Clayton & Schifflers (1987a) diskuterer dette eksempelet, og de innfører begrepet *drift* for å beskrive variasjon over tid som ikke skiller

mellom periodeinnflytelse og kohortinnflytelse, og som kan beskrives like godt av en alder-periode-modell som av en alder-kohort-modell. Clayton & Schifflers bruker betegnelsen log-lineære drift-modeller om modeller som beskriver slik variasjon.

En mulig parametrisering av den log-lineære drift-modellen med alder og periode er

$$\log Y_{ap} = \alpha_a + \delta_P (p - p_0) + \log n_{ap}$$

Her er Y_{ap} antall tilfelle i aldersgruppe *a* og kalenderperiode *p*, n_{ap} er antall personår i aldersgruppe *a* og kalenderperiode *p* og p_0 er referanseperioden. Vi skal estimere α_1, α_2, . . . , α_A og δ_P. Dette gir (*A* + 1) parametere og (*A* + 1) uavhengige likninger, og vi slipper dermed ubestemthetsproblemet og trenger ikke innføre ekstra restriksjoner. For denne parametriseringen kan vi tolke $\exp(\alpha_a)$ som tilpassete aldersspesifikke rater for referanseperioden. Parameteren δ_P kan vi kalle driftsparameteren eller trendparameteren, og $\exp(\delta_P)$ kan tolkes som tilpasset relativ risiko mellom to påfølgende perioder. Etter denne modellen er den relative risikoen fra én periode til neste periode den samme uavhengig av hvilke perioder vi ser på, med andre ord er den relative risikoen konstant etter denne modellen.

En mulig parametrisering av den log-lineære drift-modellen med alder og kohort er

$$\log Y_{ak} = \alpha_a + \delta_K (k - k_0) + \log n_{ak}$$

Her er Y_{ak} antall tilfelle i aldersgruppe *a* og fødselskohort *k*, n_{ak} er antall personår i aldersgruppe *a* og kohort *k* og k_0 er referansekohorten. Vi skal estimere α_1, α_2, . . . , α_A og δ_K. Som for drift-periode-modellen gir dette (*A* + 1) parametere og (*A* + 1) uavhengige likninger. For denne parametriseringen kan vi tolke $\exp(\alpha_a)$ som tilpassete aldersspesifikke rater for referansekohorten. Parameteren δ_K er driftsparameteren eller trendparameteren, og $\exp(\delta_K)$ kan tolkes som tilpasset relativ risiko mellom to påfølgende kohorter. Etter denne modellen er den relative risikoen fra én kohort til neste kohort konstant.

Eksempel

Vi skal se på eksempelet med forekomst av lungekreft hos belgiske kvinner i perioden 1955–1978 (data i tabell 8). Jeg vil velge første kalenderperiode (1955–1959) som referanseperiode, det vil si $p_0 = 1$. Videre vil jeg velge fødselskohort nummer 8 (1910–1919)

som referansekohort, det vil si $k_0 = 8$. De estimerte parameterverdiene for både periode-drift-modellen og kohort-drift-modellen er samlet i tabell 14. Vi ser at $\delta_P = 0{,}1025$. Dette er endringen i log(rate) fra én periode til neste. Videre fører dette til at $\exp(\delta_P) = 1{,}108$ noe som kan tolkes som at den relative risikoen øker med 10,8 % fra én periode til neste. Vi ser også at $\delta_K = 0{,}1025$. Dette er endringen i log(rate) fra én kohort til neste. I eksempelet blir $\exp(\delta_K) = 1{,}108$, som kan tolkes på tilsvarende måte som i periode-drift-modellen, altså at den relative risikoen øker med 10,8 % fra én kohort til neste. Vi legger for øvrig merke til at $\delta_P = \delta_K$. Begge modellene gir den samme deviansen $D = 42{,}3$ og samme antall frihetsgrader $df = 43$. Dataene gir altså en god tilpasning til begge modellene. Det naturlige spørsmålet er da om periode-drift-modellen og kohort-drift-modellen egentlig er samme modell. Clayton & Schifflers (1987a) mener at det ikke er samme modell og argumenterer godt for det. Carstensen (2007) på sin side understreker at det er samme modell, og begrunner også det på en like overbevisende måte. Den tilsynelatende uenigheten skyldes vel først og fremst at de har ulike utgangspunkt og forskjellige oppfatninger av hva en modell er.

Tabell 14. Lungekreft hos belgiske kvinner 1955–1978.
Parameterverdier for alder (α_a) og drift (δ_P, δ_K)
estimert fra ratene i tabell 8.

Alder	Periode-drift	Kohort-drift
25 – 29	- 12,827	- 13,135
30 – 34	- 12,018	- 12,223
35 – 39	- 11,443	- 11,546
40 – 44	- 10,595	- 10,595
45 – 49	- 9,997	- 9,894
50 – 54	- 9,558	- 9,353
55 – 59	- 9,136	- 8,829
60 – 64	- 8,838	- 8,428
65 – 69	- 8,515	- 8,002
70 – 74	- 8,262	- 7,647
75 – 79	- 8,077	- 7,359
Drift	$\delta_P = 0{,}1025$	$\delta_K = 0{,}1025$

6

Parametrisering av alder-periode-kohort-modeller

I forrige kapittel så vi på modeller med to variabler, både alder-periode- og alder-kohort-modeller. Vi presenterte eksempler som viste god tilpasning til en av disse modellene, eller til og med til begge modellene (driftmodellen), men vi viste også et eksempel der ingen av modellene var tilfredsstillende. Det er da naturlig å utvide modellen til en modell med tre variabler, det vil si en alder-periode-kohort-modell.

En mulig parametrisering av alder-periode-kohort-modellen er

$$\log Y_{ap} = \alpha_a + \beta_p + \gamma_k + n_{ap}$$

Her er Y_{ap} antall tilfelle i aldersgruppe *a* og kalenderperiode *p*, n_{ap} er antall personår i aldersgruppe *a* og kalenderperiode *p*, og dessuten er *k* gitt ved

$$k = A - a + p$$

hvor A er antall aldersgrupper. Denne sammenhengen gjelder strengt tatt bare hvis alder og periode er delt inn i like lange intervaller. Dersom dette kravet ikke er oppfylt, blir det mer problematisk å finne kohortene, i verste fall kan vi få like mange kohorter som det er observasjoner, og mange kohorter vil overlappe hverandre. Mulige løsninger på dette problemet vil bli diskutert i kapittel 8. I dette kapittelet antar vi at betingelsen om like lange intervaller er oppfylt.

Vi skal altså estimere α_a for $a = 1, 2, \ldots, A$, β_p for $p = 1, 2, \ldots, P$ og γ_k for $k = 1, 2, \ldots, K$. Det betyr at vi får (A + P + K) parametere. Det virker naturlig at vi først prøver å innføre noen av de vanlige restriksjonene. Dersom vi velger hjørnepunkt-restriksjoner, kan vi for eksempel sette $\beta_1 = 0$ og $\gamma_k = 0$ for en vilkårlig k, avhengig av hvilken kohort vi ønsker å bruke som referansekohort (for eksempel $k = (K+1)/2$, som i forrige kapittel). Dessverre vil ikke dette fungere, $\mathbf{X^T X}$ vil bli singulær og vi får ikke entydige løsninger av likningene. Faktisk er det uendelig mange løsninger som alle gir like god tilpasning til de oppgitte dataene. Dette understrekes også av Clayton & Schifflers (1987b), som i tillegg illustrerer dette poenget på en utmerket måte. De oppgir tre eksempler på parameterverdier som alle gir like god tilpasning til de oppgitte ratene, og som har vidt forskjellige alderskurver. Parameterne er med andre ord ikke identifiserbare.

Parametriseringen av modellen slik den er spesifisert foran og med de nevnte restriksjoner har én parameter for mye i forhold til det som lar seg estimere fra data. Problemet er at de tre tidsvariablene er direkte lineært avhengige av hverandre, der sammenhengen som tidligere vist, er gitt ved $k = A - a + p$. Vi får altså én uavhengig likning mindre enn det vi ellers ville fått.

Rent teknisk kan vi løse dette problemet ved å legge inn en ekstra, vilkårlig restriksjon. Problemet er å velge denne restriksjonen på en fornuftig måte, slik at parameterverdiene gir mening og resultatene lar seg tolke på en fornuftig måte. Vi skal etter hvert se på noen mulige løsninger av dette problemet. Men først skal vi se nærmere på hva som egentlig er problemet. Ifølge Clayton & Schifflers (1987b) så vil alder-periode-kohort-modellen omfatte følgene effekter: 1) drift, 2) ikke-drift periodeeffekter og 3) ikke-drift kohorteffekter. Problemet er at vi ikke klarer å skille mellom periode-drift δ_p og kohort-drift δ_k, men bare kan beregne en netto drift $\delta = \delta_p + \delta_k$. Dette problemet var vi også innom i slutten av forrige kapittel.

Vi skal se nærmere på to av metodene Clayton & Schifflers (1987b) presenterer for å måle periode- og kohorteffekter. Begge metodene tar utgangspunkt i en vilkårlig parametrisering av modellen. Anta at vi skal se på periodeeffekter. Vi kaller parameterne fra den vilkårlige parametriseringen for β_p, vi skal fjerne trenden fra disse parameterne og lage nye parametere β_p * ved å legge til et log-lineært ledd med drift. Da kan vi skrive de nye parameterne som

$$\beta_p^* = \beta_p + \delta(p - p_0)$$

hvor δ er driftsparameteren. Problemet nå er å velge δ slik at β_p^* er uten drift. I resten av kapittelet skal vi se på hvordan dette er løst hos Clayton & Schifflers (1987b), og illustrere dette ved hjelp av eksempler.

Metode 1: Førsteordensdifferanser

Den første metoden bygger på at drift defineres som gjennomsnittet av suksessive førsteordensdifferanser eller "førstedifferanser". Vi ser først på periodeeffekter. Førstedifferansene blir $(\beta_2 - \beta_1)$, $(\beta_3 - \beta_2)$, . . . , $(\beta_P - \beta_{P-1})$. Fra forrige avsnitt har vi at

$$\beta_p^* = \beta_p + \delta(p - p_0)$$

hvor p_0 er referanseperioden. Dette fører til et valg av δ, gitt ved

$$\delta = -\frac{\beta_P - \beta_1}{P - 1},$$

slik at $\beta_1^* = \beta_P^*$. Uttrykt med ord betyr det at periodekurven tvinges til å returnere til det samme nivået som den startet på.

Nå skal vi anvende den samme metoden på kohorteffekter. Førstedifferansene blir nå $(\gamma_2 - \gamma_1)$, $(\gamma_3 - \gamma_2)$, . . . , $(\gamma_K - \gamma_{K-1})$. Vi ønsker å fjerne trenden fra kohortkurven, og innfører nye parametere γ_k^*. Sammenhengen mellom γ_k^* og γ_k er nå gitt ved

$$\gamma_k^* = \gamma_k + \delta(k - k_0)$$

hvor δ er driftsparameteren og k_0 er referansekohorten. Vi velger igjen δ, gitt ved

$$\delta = -\frac{\gamma_K - \gamma_1}{K - 1},$$

slik at $\gamma_1^* = \gamma_K^*$. På samme måte som for periodekurven betyr dette at kohortkurven returnerer til det samme nivået som den startet på.

Jeg vil teste metoden på to eksempler. I dette kapittelet vil jeg ikke bruke noen av standardprogrammene i S-plus, men selv lage programmene fra grunnen av. Jeg vil bruke

”maximum likelihood” til å estimere parameterne. ”Maximum likelihood”-estimering for poissonmodeller ble beskrevet helt generelt i kapittel 4, og flere av formlene i programmet vil hentes derfra. Jeg vil starte med å lage en designmatrise **X**, og da må jeg aller først velge restriksjoner. Jeg vil igjen velge hjørnepunkt-restriksjoner. For periodeeffekter velger jeg å sette $\beta_1 = 0$ (det vil si $p_0 = 1$) og $\gamma_k = 0$ for $k = (K + 1)/2$. I tillegg må vi bruke restriksjonen $\beta_1^* = \beta_P^*$. Videre har vi

$$\beta_1 = 0 \Rightarrow \beta_1^* = 0 \text{ og } \beta_P^* = 0$$

For kohorteffekter velger jeg også hjørnepunkt-restriksjoner, og setter $\beta_1 = 0$ og $\gamma_1 = 0$ (det vil si $k_0 = 1$). I tillegg bruker jeg restriksjonen $\gamma_1^* = \gamma_K^*$. Da får vi

$$\gamma_1 = 0 \Rightarrow \gamma_1^* = 0 \text{ og } \gamma_K^* = 0$$

La $\boldsymbol{\beta}$ være en vektor som inneholder alle parameterne som skal estimeres, og la $\boldsymbol{b}$ være estimatoren for $\boldsymbol{\beta}$. Likningene vi får må løses iterativt, og da trenger vi startverdier for alle parameterne. Vi kan skrive modellen vår på matriseform som

$$\log \boldsymbol{Y} = \log \boldsymbol{N} + \boldsymbol{X\beta}$$

hvor $\boldsymbol{X}$ er designmatrisen, $\boldsymbol{\beta}$ er definert foran, $\boldsymbol{Y}$ og $\boldsymbol{N}$ er definert tidligere. Vi skriver om modellen og får

$$\log \boldsymbol{Y} - \log \boldsymbol{N} = \boldsymbol{X\beta}$$

I dette tilfelle vil designmatrisen bare inneholde dummyvariabler. Vi setter $\boldsymbol{z} = \log \boldsymbol{Y} - \log \boldsymbol{N}$ og kan finne startverdien for $\boldsymbol{b}$ som

$$\boldsymbol{b}^{(1)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{z}$$

Nå kan vi finne en startverdi for vektmatrisen $\boldsymbol{W}$ og definere $\boldsymbol{z}$ på nytt. Formlene for elementene som inngår i $\boldsymbol{W}$ og $\boldsymbol{z}$ er gitt i kapittel 4. Vi kan finne neste verdi av $\boldsymbol{b}$ som

$$\boldsymbol{b}^{(2)} = [(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{(1)}]^{-1}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z})^{(1)}$$

Generelt kan vi finne iterasjon nummer m som

$$\boldsymbol{b}^{(m)} = [(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{(m-1)}]^{-1}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z})^{(m-1)}$$

Det er denne metoden som er benyttet i eksemplene nedenfor.

Eksempel

Vi ser først på et eksempel med data fra forekomsten av brystkreft i Japan i perioden 1955–1979 (Clayton & Schifflers 1987b). En oversikt over mortalitetsrater er vist i tabell 15. Implementeringen av dette eksempelet i S-plus er vist i appendiks C.

Tabell 15. Aldersspesifikke mortalitetsrater (per 100 000 personår) for brystkreft hos japanske kvinner i perioden 1955–1979. Antall tilfelle i parentes. (Kilde: WHO mortality data base).

Alder\periode	1955–59	1960–64	1965–69	1970–74	1975–79
25 – 29	0,44 (88)	0,38 (78)	0,46 (101)	0,55 (127)	0,68 (179)
30 – 34	1,69 (299)	1,69 (330)	1,75 (363)	2,31 (509)	2,52 (588)
35 – 39	4,01 (596)	3,90 (680)	4,11 (798)	4,44 (923)	4,80 (1056)
40 – 44	6,59 (874)	6,57 (962)	6,81 (1171)	7,79 (1497)	8,27 (1716)
45 – 49	8,51 (1022)	9,61 (1247)	9,96 (1429)	11,68 (1987)	12,51 (2398)
50 – 54	10,49 (1035)	10,80 (1258)	12,36 (1560)	14,59 (2079)	16,56 (2794)
55 – 59	11,36 (970)	11,51 (1087)	12,98 (1446)	14,97 (1828)	17,79 (2465)
60 – 64	12,03 (820)	10,67 (861)	12,67 (1126)	14,46 (1549)	16,42 (1962)
65 – 69	12,55 (678)	12,03 (738)	12,10 (878)	13,81 (1140)	16,46 (1683)
70 – 74	15,81 (640)	13,87 (628)	12,65 (656)	14,00 (900)	15,60 (1162)
75 – 79	17,97 (497)	15,62 (463)	15,83 (536)	15,71 (644)	16,52 (865)

De estimerte parameterverdiene for periodeeffekter er vist i tabell 16. Det er også tatt med en ekstra kolonne for estimerte mortalitetsrater per 100 000 personår i hver aldersgruppe. Vi får en alderskurve som øker monotont. For periode har vi fjernet trend, slik at de nye parameterverdiene β_p * er uavhengige av hvilken parametrisering som er brukt som utgangspunkt. Vi ser at de tre parameterverdiene β_2 * til β_4 * alle er negative, noe som betyr at kurven for periode er konveks. Kohortkurven avtar frem til kohort 4 (1890–1899), deretter øker den jevnt. Problemet er at det ikke er liketil å tolke disse parameterverdiene. For eksempel øker vår alderskurve monotont, men som Clayton & Schifflers (1987b) har demonstrert kan man få alderskurver med en helt annen form ved å velge en annen parametrisering. Heller ikke tolkningen av β_p * er rett frem, i vårt eksempel blir β_p * best tolket som

$$\beta_p^* = (\beta_p - \beta_1) - (p-1)(\beta_P - \beta_1)/(P-1)$$

Clayton & Schifflers demonstrerer at også kohortkurven kan ha forskjellig form avhengig av hvilken parametrisering som er valgt. Videre kan vi vel tolke α_a som log(rate) i referansekohorten etter justering for periodeeffekter.

Tabell 16. Brystkreft hos japanske kvinner 1955–1979. Periodeeffekter med lineær trend fjernet. Alder- (α_a), periode- (β_p *) og kohort- (γ_k) parametere estimert fra ratene i tabell 15.

Alder	α_a	$\exp(\alpha_a)\cdot 100000$	Periode	β_p *	Kohort	γ_k
25 – 29	- 12,599	0,338	1955 – 59	0,000	1875 – 84	- 0,250
30 – 34	- 11,106	1,502	1960 – 64	- 0,078	1880 – 89	- 0,272
35 – 39	- 10,268	3,473	1965 – 69	- 0,084	1885 – 94	- 0,325
40 – 44	- 9,666	6,340	1970 – 74	- 0,028	1890 – 99	- 0,351
45 – 49	- 9,219	9,914	1975 – 79	0,000	1895 – 04	- 0,340
50 – 54	- 8,912	13,476			1900 – 09	- 0,239
55 – 59	- 8,754	15,783			1905 – 14	- 0,123
60 – 64	- 8,694	16,759			1910 – 19	0,000
65 – 69	- 8,608	18,264			1915 – 24	0,111
70 – 74	- 8,510	20,144			1920 – 29	0,188
75 – 79	- 8,375	23,056			1925 – 34	0,234
					1930 – 39	0,263
					1935 – 44	0,367
					1940 – 49	0,517
					1945 – 54	0,701

De estimerte parameterverdiene for kohorteffekter er samlet i tabell 17 sammen med en kolonne for estimerte rater per 100 000 personår. Alderskurven stiger til å begynne med for så å flate mer eller mindre ut, før den igjen stiger svakt. Denne alderskurven avviker betydelig fra alderskurven vi fikk fra tabell 16, noe som igjen understreker hvor vanskelig det er å finne den ”riktige” alderskurven. Periodekurven synker svakt før den stiger. Vi har fjernet trenden for kohort, slik at de nye parameterverdiene er uavhengige av hvilken parametrisering vi har startet med. Vi ser at alle parameterverdiene for γ_k * er negative, det betyr at kohortkurven er konveks, men ellers går den litt opp og ned. Som for periode er ikke tolkningen av parameterverdiene liketil, men γ_k * blir trolig best tolket som

$$\gamma_k^* = (\gamma_k - \gamma_1) - (k-1)(\gamma_K - \gamma_1)/(K-1)$$

For å finne driften bruker vi formlene i starten av dette underkapittelet. For periode får vi

$$\delta = -\frac{\beta_P - \beta_1}{P-1} = -\frac{0{,}272 - 0}{5-1} = -0{,}068$$

hvor parameterverdiene er hentet fra tabell 17. Tilsvarende får vi for kohort

$$\delta = -\frac{\gamma_K - \gamma_1}{K-1} = -\frac{0{,}701 - (-0{,}250)}{15-1} = -0{,}068$$

hvor parameterverdiene er hentet fra tabell 16. Som ventet får vi samme verdien for drift i de to tilfellene. Den verdien vi har funnet for drift ($\delta = -0{,}068$) kan vi kalle for netto drift. For å oppsummere kan vi si at i dette eksempelet har vi både drift, ikke-drift periodeeffekter og ikke-drift kohorteffekter.

Tabell 17. Brystkreft i Japan 1955–1979. Kohorteffekter med lineær trend fjernet.
Alder- (α_a), periode- (β_p) og kohort- (γ_k *) parametere estimert fra ratene i tabell 15.

Alder	α_a	$\exp(\alpha_a)\cdot 100000$	Periode	β_p	Kohort	γ_k *
25 – 29	- 12,170	0,518	1955 – 59	0,000	1875 – 84	0,000
30 – 34	- 10,745	2,155	1960 – 64	- 0,010	1880 – 89	- 0,090
35 – 39	- 9,974	4,660	1965 – 69	0,052	1885 – 94	- 0,211
40 – 44	- 9,440	7,948	1970 – 74	0,175	1890 – 99	- 0,305
45 – 49	- 9,062	11,599	1975 – 79	0,272	1895 – 04	- 0,362
50 – 54	- 8,822	14,745			1900 – 09	- 0,328
55 – 59	- 8,732	16,134			1905 – 14	- 0,281
60 – 64	- 8,740	16,005			1910 – 19	- 0,226
65 – 69	- 8,722	16,296			1915 – 24	- 0,182
70 – 74	- 8,692	16,792			1920 – 29	- 0,173
75 – 79	- 8,624	17,974			1925 – 34	- 0,195
					1930 – 39	- 0,234
					1935 – 44	- 0,198
					1940 – 49	- 0,116
					1945 – 54	0,000

For å undersøke om alder-periode-kohort-modellen er en modell som viser god tilpasning til data, kan vi ennå en gang se på deviansen. Tabell 18 viser deviansen *D* og antall frihetsgrader *df* for testing av ulike modeller med utgangspunkt i dette eksempelet. Det kan se ut som alder-periode-kohort-modellen er den beste, og at vi har både en periodeeffekt og en kohorteffekt i tillegg til en alderseffekt. Det er i hvert fall bare den siste modellen som ikke viser signifikans.

Tabell 18. Brystkreft hos japanske kvinner 1955–1979. Devians *D* og antall frihetsgrader *df* for ulike modeller.

Modell	*D*	*df*
Alder	1096	44
Alder + drift	298	43
Alder + periode	215	40
Alder + kohort	85,8	30
Alder + periode + kohort	30,5	27

Eksempel

Det neste eksempelet studerte vi også i kapittel 5, det dreier seg om forekomsten av prostatakreft hos ikke-hvite menn i USA i perioden 1935–1969 (data i tabell 3 i kapittel 3). De estimerte parameterverdiene for periodeeffekter er vist i tabell 19. Vi ser at alderskurven er monotont voksende. For periode er trend fjernet, og vi ser at alle β_p * er positive, noe som betyr at vi får en konkav periodefunksjon. Kohortkurven øker frem til kohort 10 (1895–1904), deretter avtar den. Som i forrige eksempel er tolkningen av parameterverdiene problematisk, men β_p * kan tolkes på samme måte som i forrige eksempel. De estimerte parameterverdiene for kohorteffekter er samlet i tabell 20. Alderskurven stiger jevnt, men stiger ikke så bratt som kurven fra tabell 19. Også periodekurven stiger jevnt. Vi har fjernet trend fra kohortkurven, og ser at alle γ_k * er positive, det vil si at også kohortkurven er konkav. Vi ser videre at kohortkurven først stiger jevnt til et toppunkt for deretter å synke jevnt.

Vi skal beregne driften for dette eksempelet også. For periode får vi

$$\delta = -\frac{0{,}640 - 0}{7 - 1} = -0{,}107$$

med parameterverdier fra tabell 20. For kohort får vi

$$\delta = -\frac{-0{,}006 - (-1{,}287)}{13 - 1} = -0{,}107$$

med parameterverdier fra tabell 19. Også i dette eksempelet har vi alle de tre effektene som Clayton & Schifflers (1987b) omtaler. Det ser også ut som ikke-drift kohorteffektene er mer markert enn ikke-drift periodeeffektene.

For å se på tilpasningen til data ser vi igjen på deviansen, og finner at $D = 98{,}9$ og antall frihetsgrader $df = 25$. Dette er klart signifikant, og kan tyde på at modellen ikke er så bra likevel. I kapittel 5 så vi at alder-kohort-modellen hadde $D = 127{,}4$ og $df = 30$. En utvidelse fra en alder-kohort-modell til en alder-periode-kohort-modell gir altså ikke så mye bedre tilpasning. Holford (1983) forklarer den tilsynelatende mangelen på tilpasning til data disse modellene viser med det høye antallet døde. Han mener at alder-kohort-modellen også gir en brukbar tilpasning til data, og demonstrerer dette ved hjelp av en figur.

Tabell 19. Prostatakreft hos ikke-hvite menn i USA 1935–1969. Periodeeffekter med lineær trend fjernet. Alder- (α_a), periode- (β_p *) og kohort- (γ_k) parametere estimert fra antall dødsfall og antall personår i tabell 3.

Alder	α_a	Periode	β_p *	Kohort	γ_k
50 – 54	- 7,382	1935 – 39	0,000	1850 – 59	- 1,287
55 – 59	- 6,555	1940 – 44	0,004	1855 – 64	- 0,921
60 – 64	- 5,810	1945 – 49	0,041	1860 – 69	- 0,774
65 – 69	- 5,222	1950 – 54	0,080	1865 – 74	- 0,539
70 – 74	- 4,661	1955 – 59	0,055	1870 – 79	- 0,311
75 – 79	- 4,208	1960 – 64	0,005	1875 – 84	- 0,166
80 – 84	- 3,883	1965 – 69	0,000	1880 – 89	0,000
				1885 – 94	0,211
				1890 – 99	0,264
				1895 – 04	0,312
				1900 – 09	0,195
				1905 – 14	0,119
				1910 – 19	- 0,006

Hva skjer dersom vi prøver å bruke et av standardprogrammene i S-plus til å estimere parameterne i alder-periode-kohort-modellen, og bare innfører de vanlige hjørnepunkt-restriksjonene? Som nevnt foran vil $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ bli singulær, men S-plus takler dette på en måte. Vi vil få frem verdier for alle parameterne unntatt den siste parameteren, som oppgis som ”ikke tilgjengelig”. Til tross for denne manglende verdien vil programmet regne ut tilpassete log(rater) for alle observasjonene, det vil si at den siste parameteren må ha fått en verdi. Ved å kontrollere svarene viser det seg at S-plus i dette tilfelle setter verdien av den siste

parameteren lik null. Med andre ord dersom vi setter $\beta_1 = 0$ og $\gamma_1 = 0$ vil vi få de samme parameterverdiene som i tabell 20.

Tabell 20. Prostatakreft hos ikke-hvite menn i USA 1935–1969. Kohorteffekter med lineær trend fjernet. Alder- (α_a), periode- (β_p) og kohort- (γ_k *) parametere estimert fra tallene i tabell 3.

Alder	α_a	Periode	β_p	Kohort	γ_k *
50 – 54	- 8,028	1935 – 39	0,000	1850 – 59	0,000
55 – 59	- 7,308	1940 – 44	0,111	1855 – 64	0,259
60 – 64	- 6,670	1945 – 49	0,255	1860 – 69	0,300
65 – 69	- 6,189	1950 – 54	0,401	1865 – 74	0,428
70 – 74	- 5,734	1955 – 59	0,482	1870 – 79	0,549
75 – 79	- 5,388	1960 – 64	0,539	1875 – 84	0,587
80 – 84	- 5,170	1965 – 69	0,640	1880 – 89	0,647
				1885 – 94	0,751
				1890 – 99	0,697
				1895 – 04	0,639
				1900 – 09	0,415
				1905 – 14	0,232
				1910 – 19	0,000

Metode 2: Andreordensdifferanser

Slik denne metoden er beskrevet av Clayton & Schifflers (1987b) er dette en metode som bare tar for seg ikke-drift-effekter. De gir ingen beskrivelse av hvordan de eventuelt beregner drift med denne metoden, derfor vil heller ikke jeg omtale drift i forbindelse med denne metoden. Ikke-drift-effekter kan uttrykkes som kontraster mellom relative risikoer. Slike kontraster kan for eksempel være forholdet mellom to påfølgende relative risikoer. For periodeeffekter kan dette uttrykkes matematisk som

$$\frac{\exp(\beta_3)/\exp(\beta_2)}{\exp(\beta_2)/\exp(\beta_1)}, \quad \frac{\exp(\beta_4)/\exp(\beta_3)}{\exp(\beta_3)/\exp(\beta_2)}, \ldots$$

På logaritmisk skala blir disse kontrastene andreordensdifferanser eller ”andredifferanser” som Clayton & Schifflers kaller det. Dette kan vi skrive som

$$(\beta_3 - \beta_2) - (\beta_2 - \beta_1) = \beta_3 - 2\beta_2 + \beta_1 , \qquad \beta_4 - 2\beta_3 + \beta_2 , \ldots$$

Andreordensdifferansene er identifiserbare og uavhengige av hvilken parametrisering som er valgt i utgangspunktet. Vi bruker andreordensdifferansene til å definere nye parametere. La α_a, β_p og γ_k være parameterne for en vilkårlig parametrisering av alder-periode-kohort-modellen. Vi definerer da de nye parameterne som

$$\alpha_a{}^* = \alpha_{a+1} - 2\alpha_a + \alpha_{a-1} \qquad \text{for } a = 2, 3, \ldots (A-1)$$

$$\beta_p{}^* = \beta_{p+1} - 2\beta_p + \beta_{p-1} \qquad \text{for } p = 2, 3, \ldots (P-1)$$

$$\gamma_k{}^* = \gamma_{k+1} - 2\gamma_k + \gamma_{k-1} \qquad \text{for } k = 2,3, \ldots (K-1)$$

Vi ser at i forhold til de opprinnelige parameterne så mangler tilsynelatende første og siste parameter for hver av de tre variablene, dette skyldes at det ikke er mulig å beregne andreordensdifferanser for disse.

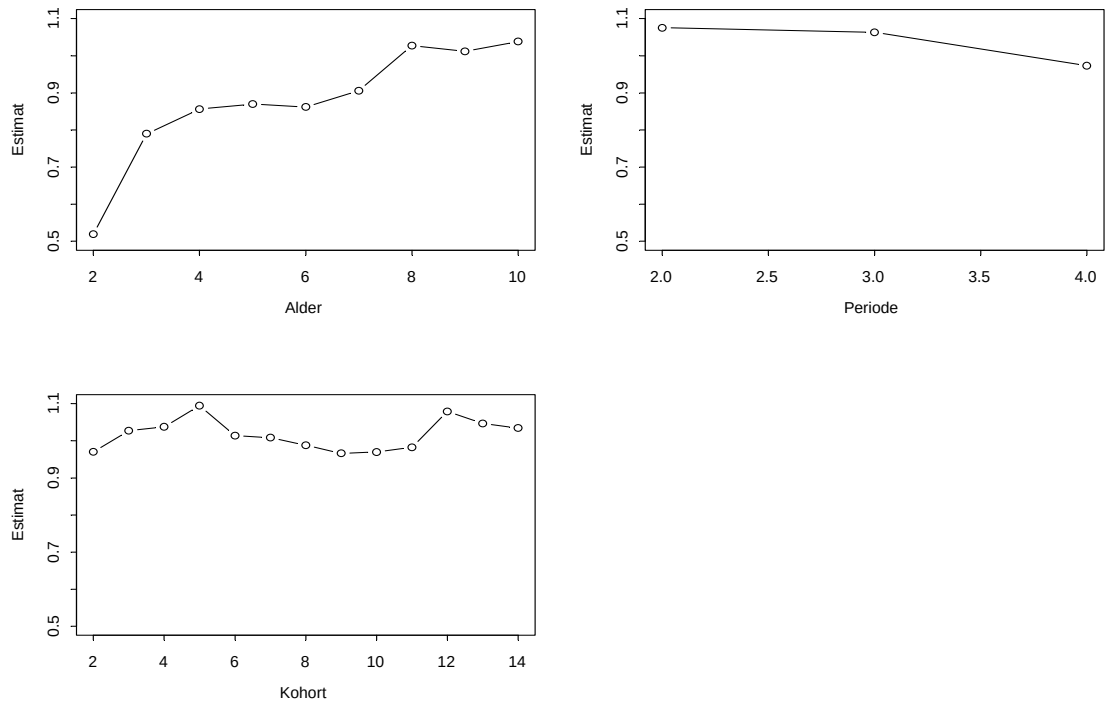
Jeg vil teste denne metoden på de samme to eksemplene som jeg brukte til å teste den første metoden. Jeg vil ta utgangspunkt i en av parametriseringene som er brukt tidligere, og bruke et dataprogram som jeg har laget for å regne ut de nye parameterne.

Eksempel

Det første eksempelet dreier seg om forekomsten av brystkreft hos japanske kvinner i perioden 1955–1979 (data i tabell 15). De estimerte parameterverdiene for andreordensdifferenser er samlet i tabell 21. Figur 3 viser de samme parameterne fremstilt grafisk. Hvordan skal vi så tolke disse parameterne? For eksempel er $\exp(\alpha_2{}^*) = 0{,}519$, dette kan tolkes som at den relative risikoen for aldersgruppe 3 versus aldersgruppe 2 bare er 52 prosent av den relative risikoen for aldersgruppe 2 versus aldersgruppe 1. Vi ser av figuren at alderskurven har en tydelig fordypning rundt menopausen (ca. 50 år), kalt ”Clemmesen’s hook” hos Clayton & Schifflers (1987b). Tilsvarende kan vi tolke $\exp(\beta_2{}^*) = 1{,}075$ som at den relative risikoen for periode 3 versus periode 2 er 7,5 prosent høyere enn den relative risikoen for periode 2 versus periode 1. Kohortkurven har to påfallende uregelmessigheter, som indikerer plutselige forandringer i kohorttrend rundt 1900 og rundt 1935. Implementeringen av dette eksempelet i S-plus er vist i appendiks C.

Tabell 21. Brystkreft hos japanske kvinner 1955–1979. Alder-, periode- og kohorteffekter som andreordensdifferanser.

Alder	$\exp(\alpha_a^*)$	Periode	$\exp(\beta_p^*)$	Kohort	$\exp(\gamma_k^*)$
25 – 29	-	1955 – 59	-	1875 – 84	-
30 – 34	0,519	1960 – 64	1,075	1880 – 89	0,970
35 – 39	0,790	1965 – 69	1,063	1885 – 94	1,027
40 – 44	0,856	1970 – 74	0,973	1890 – 99	1,034
45 – 49	0,870	1975 – 79	-	1895 – 04	1,095
50 – 54	0,862			1900 – 09	1,014
55 – 59	0,906			1905 – 14	1,008
60 – 64	1,027			1910 – 19	0,988
65 – 69	1,011			1915 – 24	0,966
70 – 74	1,038			1920 – 29	0,970
75 – 79	-			1925 – 34	0,983
				1930 – 39	1,079
				1935 – 44	1,047
				1940 – 49	1,034
				1945 – 54	-



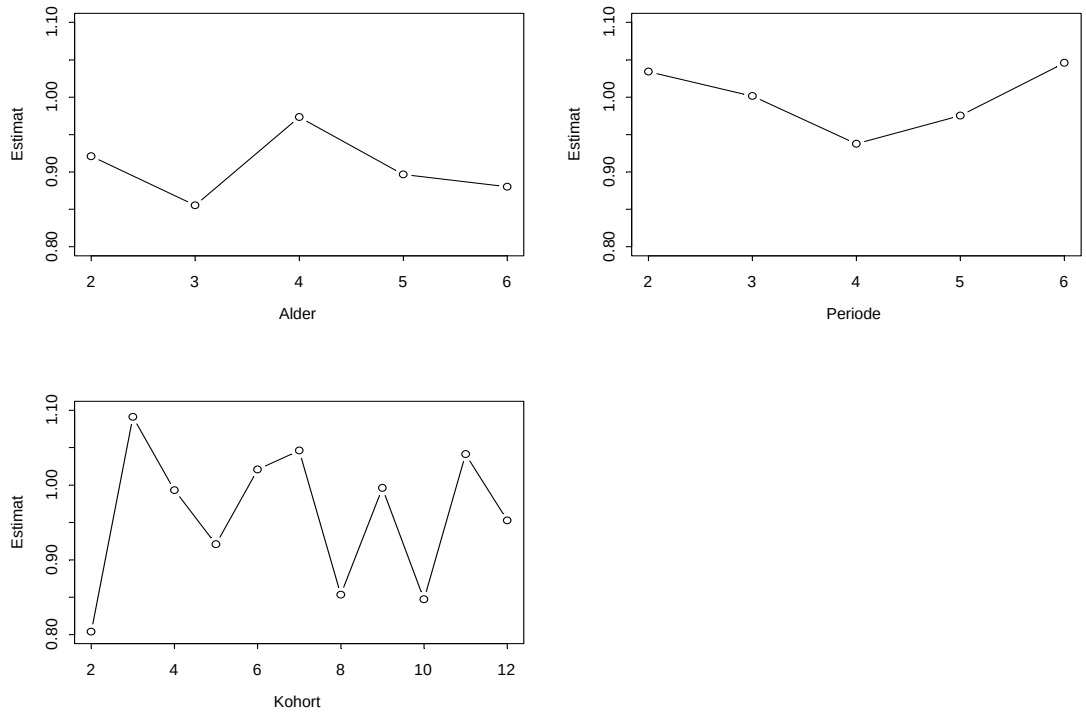
Figur 3. Brystkreft hos japanske kvinner 1955–1979. Estimater av andreordensdifferanser for alder-, periode- og kohorteffekter.

Eksempel

I det andre eksempelet ser vi nok en gang på forekomsten av prostatakreft hos ikke-hvite menn i USA i perioden 1935–1969 (data i tabell 3). De estimerte parameterverdiene for andreordensdifferenser er samlet i tabell 22. Figur 4 viser de samme parameterne fremstilt grafisk. Alderskurven holder seg hele tiden under 1, som kan tolkes som at forholdet mellom den relative risikoen mellom to påfølgende aldersgrupper avtar. Ellers legger vi merke til at alderskurven har en fordypning rundt 60 år. Kohortkurven er svært uregelmessig med store sprang, noe som indikerer flere plutselige forandringer i kohorttrend.

Tabell 22. Prostatakreft hos ikke-hvite menn i USA 1935–1969.
Alder-, periode- og kohorteffekter som andreordensdifferanser.

Alder	$\exp(\alpha_a^*)$	Periode	$\exp(\beta_p^*)$	Kohort	$\exp(\gamma_k^*)$
50 – 54	-	1935 – 39	-	1850 – 59	-
55 – 59	0,921	1940 – 44	1,034	1855 – 64	0,804
60 – 64	0,855	1945 – 49	1,002	1860 – 69	1,091
65 – 69	0,973	1950 – 54	0,938	1865 – 74	0,993
70 – 74	0,897	1955 – 59	0,975	1870 – 79	0,921
75 – 79	0,880	1960 – 64	1,046	1875 – 84	1,021
80 – 84	-	1965 – 69	-	1880 – 89	1,046
				1885 – 94	0,853
				1890 – 99	0,996
				1895 – 04	0,847
				1900 – 09	1,041
				1905 – 14	0,953
				1910 – 19	-



Figur 4. Prostatakreft hos ikke-hvite menn i USA 1935–1969. Estimater av andreordensdifferanser for alder-, periode- og kohorteffekter.

7

Alternative metoder

Holfords metode

Denne metoden ble presentert i Holford (1983). Min fremstilling bygger delvis på denne artikkelen og delvis på en nyere artikkel (Holford 1991). Han tar utgangspunkt i samme modell som ble presentert i forrige kapittel, der de tre variablene alder, periode og kohort betraktes som faktorer. Med min notasjon kan modellen formuleres slik

$$\log Y_{ap} = \mu + \alpha_a + \beta_p + \gamma_k + n_{ap}$$

Her er Y_{ap} antall tilfelle i aldersgruppe *a* og kalenderperiode *p*, og Y_{ap} antas å være poissonfordelt. Videre er n_{ap} antall personår i aldersgruppe *a* og kalenderperiode *p*, og n_{ap} antas å være konstant. Dessuten kan fødselskohort *k* beregnes som

$$k = A - a + p$$

hvor *A* er antall aldersgrupper.

Mens jeg, i likhet med Clayton & Schifflers (1987a og b), har brukt hjørnepunkt-restriksjoner, så bruker Holford sum null-restriksjoner. Det betyr at

$$\sum_{a=1}^{A} \alpha_a = \sum_{p=1}^{P} \beta_p = \sum_{k=1}^{K} \gamma_k = 0$$

hvor *P* er antall perioder og *K* er antall kohorter. Restriksjoner og kontraster er omtalt tidligere, og i kapittel 5 er disse begrepene gitt en relativt grundig gjennomgang.

Hovedpoenget med Holfords metode er at han bruker to komponenter for å representere effekten av hver av de tre faktorene. Den ene komponenten er lineær trend og den andre komponenten er krumning eller avvik fra linearitet. Den siste komponenten kan også kalles residualkomponenten. For effekten av alder kan dette formuleres matematisk som

$$\alpha_a = \alpha_{La} + \alpha_{Ca}$$

hvor α_{La} representerer den lineære komponenten og α_{Ca} representerer krumningen.
Den lineære komponenten kan skrives som

$$\alpha_{La} = \varphi_{A0} + \varphi_{A1} x_a$$

hvor φ_{A0} og φ_{A1} er regresjonskoeffisienter for vanlig lineær regresjon og x_a er regressorvariabel. Holford (1991) bruker en normalisert aldersindeks som regressorvariabel, det vil si

$$x_a = a - (A+1)/2$$

Med dette uttrykket for x_a får vi følgende formel for effekten av alder

$$\alpha_a = \varphi_{A0} + [a - (A+1)/2]\varphi_{A1} + \alpha_{Ca}$$

Vi finner en tilsvarende formel hos Holford (1991). Men det er en forskjell, han har ikke noe konstantledd som tilsvarer leddet φ_{A0}. Men Holford bruker sum null-restriksjoner, og da stemmer det med at $\varphi_{A0} = 0$.

Tilsvarende kan vi finne et uttrykk for effekten av periode som

$$\beta_p = \varphi_{P0} + [p - (P+1)/2]\varphi_{P1} + \beta_{Cp}$$

Her er φ_{P0} og φ_{P1} regresjonskoeffisienter, $[p - (P+1)/2]$ er normalisert periodeindeks og β_{Cp} er krumningen eller residualkomponenten.

På samme måte kan vi også finne et uttrykk for effekten av kohort som

$$\gamma_k = \varphi_{K0} + [k - (K+1)/2]\varphi_{K1} + \gamma_{Ck}$$

Her er da φ_{K0} og φ_{K1} regresjonskoeffisienter, $[k - (K+1)/2]$ er normalisert kohortindeks og γ_{Ck} er krumningen.

Når effekten av alder, periode og kohort skal presenteres, kan den lineære komponenten og krumningen rapporteres hver for seg. Som Holford (1983) har vist er ikke total lineær trend

estimerbar, derimot er for eksempel $\varphi_{A1} + \varphi_{P1}$ estimerbar, og det samme er $\varphi_{P1} + \varphi_{K1}$. Det siste uttrykket, som altså er summen av stigningstallene for periode og kohort, kaller han netto drift (Holford 1991). I tillegg til netto drift vil han rapportere krumningen for både alder, periode og kohort. I motsetning til lineær trend er krumningen estimerbar, og kan finnes ved å fjerne den lineære komponenten fra de estimerte parameterverdiene. Vi vil se på et eksempel for å vise anvendelsen av denne metoden.

Eksempel

Vi skal se nærmere på et eksempel med prostatakreft hos ikke-hvite menn i USA, som er det eksempelet Holford selv bruker (Holford 1983). Jeg vil bruke metoden slik den er beskrevet av Holford (1991). Vi tar utgangspunkt i en vilkårlig parametrisering av modellen, jeg vil bruke de estimerte parameterverdiene i tabell 20 i kapittel 6. Vi starter med å tilpasse en regresjonslinje gjennom de estimerte parameterverdiene for alderseffekter. Det gir linjen

$$l(a) = -6{,}355 + 0{,}477 \cdot (a - 4)$$

På samme måte kan vi tilpasse en regresjonslinje for periodeeffekter, og får

$$l(p) = 0{,}347 + 0{,}107 \cdot (p - 4)$$

Til slutt tilpasser vi også en regresjonslinje for kohorteffekter. Det gir linjen

$$l(k) = 0{,}423 + 0{,}008 \cdot (k - 7)$$

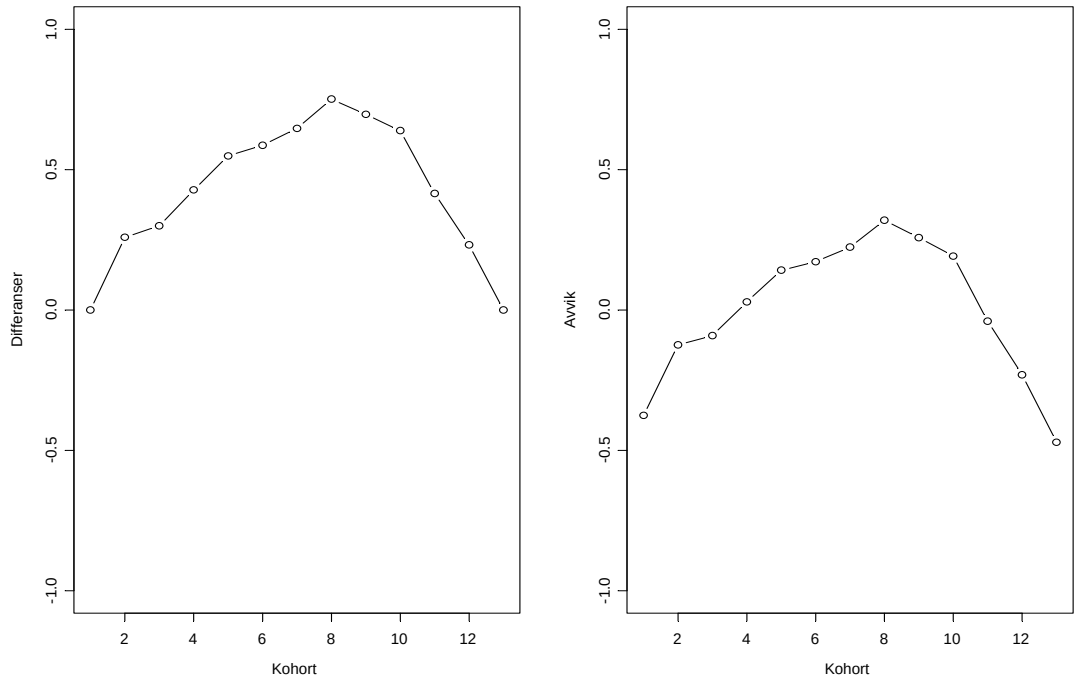
For å finne krumningen fjernes nå den lineære komponenten fra de estimerte parameterverdiene. Dette gjøres ved å trekke den tilpassete linjen fra den estimerte parameterverdien. Som eksempel kan vi vise utregningen for den første aldersgruppen

$$\alpha_{C1} = -8{,}028 - (-6{,}355 + 0{,}477 \cdot (1 - 4)) = -0{,}242$$

Alle avvikene fra linearitet, som er det samme som krumning, er samlet i tabell 23. Hvis vi sammenlikner mine tall med Holfords tall ser vi at avvikene er de samme. Dette til tross for at både estimerte parameterverdier og tilpassete linjer er helt forskjellig fra Holfords', men det bekrefter Holfords påstand om krumningens estimerbarhet. Vi ser også at det han kaller netto drift, $\hat{\gamma}_L + \hat{\pi}_L = 0{,}115$ (Holford 1983), samsvarer med mine tall, der $\varphi_{P1} + \varphi_{K1} = 0{,}008 + 0{,}107 = 0{,}115$. På figur 5 er kohort plottet mot γ_k og mot γ_{Ck} (tall fra tabell 23). Vi ser at kurvene har sanne mønster, noe som ikke er overraskende da γ_k også er fremkommet ved å fjerne lineær trend (se kapittel 6).

Tabell 23. Prostatakreft hos ikke-hvite menn i USA 1935–1969. Estimerte parameterverdier er hentet fra tabell 20 i kapittel 6. Avvik fra linearitet er beregnet etter Holfords metode.

Alder	α_a	Avvik ($=\alpha_{Ca}$)	Periode	β_p	Avvik ($=\beta_{Cp}$)	Kohort	γ_k	Avvik ($=\gamma_{Ck}$)
50 – 54	- 8,028	-0,242	1935 – 39	0,000	-0,026	1850 – 59	0,000	-0,375
55 – 59	- 7,308	0,001	1940 – 44	0,111	-0,022	1855 – 64	0,259	-0,124
60 – 64	- 6,670	0,162	1945 – 49	0,255	0,015	1860 – 69	0,300	-0,091
65 – 69	- 6,189	0,166	1950 – 54	0,401	0,054	1865 – 74	0,428	0,029
70 – 74	- 5,734	0,144	1955 – 59	0,482	0,028	1870 – 79	0,549	0,142
75 – 79	- 5,388	0,012	1960 – 64	0,539	-0,022	1875 – 84	0,587	0,172
80 – 84	- 5,170	-0,246	1965 – 69	0,640	-0,028	1880 – 89	0,647	0,224
						1885 – 94	0,751	0,320
						1890 – 99	0,697	0,258
						1895 – 04	0,639	0,192
						1900 – 09	0,415	-0,040
						1905 – 14	0,232	-0,231
						1910 – 19	0,000	-0,471



Figur 5. Prostatakreft hos ikke-hvite menn i USA 1935–1969. Førsteordensdifferanser etter Clayton & Schifflers (1987b) (til venstre), sammenliknet med avvik fra linearitet etter Holford (1983) (til høyre). Estimerte verdier fra tabell 23.

Carstensens metode

Alle modellene vi har studert så langt har vært faktormodeller, det vil si at både alder, kalenderperiode og fødselskohort betraktes som faktorer med én parameter for hver distinkt verdi av alder *a*, periode *p* og kohort *k*. Nå skal vi se nærmere på en metode der alder-periode-kohort-modellen ikke betraktes som en faktormodell. Metoden presenteres av Carstensen (2007) i en helt fersk artikkel. Jeg vil legge frem noen av hovedpoengene hans. Han understreker at dersom vi skal bruke en statistisk modell til å beskrive rater fra et sykdomsregister, så er det egentlig tre separate emner som skal betraktes. De tre emnene er:

Data: Hvordan skal data tabuleres?

Modell: Skal vi bruke en faktormodell eller en modell med glatte funksjoner?

Parametrisering: Hvilke restriksjoner skal vi bruke? Hvordan skal resultatene presenteres?

Jeg vil ta for meg ett og ett emne, og legge frem hans idéer og anbefalinger.

Data

Han mener at tabellene skal være så detaljerte som mulig, bare begrenset av tilgjengelige populasjonstall. Femårsintervaller som er vanlig i litteraturen blir som oftest for grovt, i eksempelet som han bruker for å illustrere metoden sin bruker han1-års-intervaller. I prinsippet er det derimot ikke noe i veien for å bruke enda kortere intervaller. Data skal fortrinnsvis presenteres i et Lexis-diagram, der både alder, periode og kohort inngår. Intervallene for de tre variablene trenger ikke være av samme lengde. Hver celle i Lexis-diagrammet skal inneholde antall tilfelle og et mål for risikotid. Han presenterer en alternativ formel for å regne ut total risikotid fra et Lexis-diagram. La $L_{a,p}$ være populasjonsstørrelsen i aldersgruppe *a* i begynnelsen av kalenderperiode *p*. Det er vanlig å regne ut middelfolkemengden som

$$N_{a,p} = \frac{1}{2} L_{a,p} + \frac{1}{2} L_{a,p+1} \ .$$

Han mener det er bedre å estimere middelfolkemengden i aldersgruppe *a* og periode *p* som

$$N_{a,p} = \frac{1}{6} L_{a-1,p} + \frac{1}{3} L_{a,p} + \frac{1}{3} L_{a,p+1} + \frac{1}{6} L_{a+1,p+1} \ .$$

Den totale risikotiden (i praksis som oftest antall personår) blir da

$$PT = N_{a,p} * \Delta t$$

hvor Δt er lengden av tidsintervallet, for eksempel antall år.

En annen viktig forutsetning er at det for alle celler i Lexis-diagrammet gjelder at

$$k = p - a$$

hvor k er kohort.

Modell

På generell form kan modellen hans formuleres slik

$$\log[\lambda(a, p)] = f(a) + g(p) + h(k)$$

Her er $\lambda(a, p)$ ratene ved alder a i periode p for personer i fødselskohort $k = p - a$. I denne modellen antas det at a, p og k representerer henholdsvis gjennomsnittsalder, gjennomsnittsperiode og gjennomsnittskohort for hver observasjonsenhet. Modellen tillater effektene av hver av de tre variablene å være ikke-lineær. Han vil altså ikke bruke den tradisjonelle faktormodellen, men en modell der de tre variablene betraktes som kontinuerlige kovariater. Raten må antas å være konstant innen hver celle i Lexis-diagrammet. Modeller for λ kan da tilpasses ved å bruke poissonregresjon for uavhengige observasjoner med et offset-ledd for antall personår. Effektene av alder, periode og kohort modelleres ved hjelp av parametrisk glatte funksjoner.

Parametrisering

Han understreker at parametriseringen må velges med omhu, slik at relevante trekk kan oppdages. Han vil bruke parametriske funksjoner for å beskrive effektene. Han demonstrerer metoden ved hjelp av et eksempel hvor han ser på forekomsten av testikkelkreft hos danske menn i perioden 1943–1996. Som parametrisk glatte funksjoner har han i dette eksempelet valgt å bruke ”natural splines”, som er stykkevise polynomfunksjoner av tredje grad, som begrenses til å være lineære utenfor de ytterste knutene (”knots”). Neste punkt er å trekke ut drift eller lineær trend, for eksempel ved ortogonal projeksjon. Videre er det viktig å oppgi hvordan man velger referansekohort eller referanseperiode. Han anbefaler at estimatene skal rapporteres som linjegrafer. Videre anbefaler han at det for alderseffekter skal være skala for rater langs andreaksen, og at det for periode- og kohorteffekter skal være skala for relativ risiko langs andreaksen.

Det ville nå vært naturlig å anvende Carstensens metode på et av eksemplene fra kapittel 6, men ingen av disse eksemplene har oppgitt tilstrekkelig detaljerte data til at det lar seg gjøre. I stedet for vil jeg anvende metodene som er beskrevet i tidligere kapitler på Carstensens eksempel, og sammenlikne mine resultater med Carstensens resultater.

Eksempel

Carstensen (2007) bruker et eksempel som går på forekomsten av testikkelkreft i Danmark i perioden 1943–1996. Innledningsvis oppgir han en tabell over antall tilfelle og antall personår fordelt på 5-års-klasser for både alder og periode. Denne tabellen er gjengitt i tabell 24. Disse dataene er analysert med metodene som er beskrevet kapittel 5 og 6. Tabell 25 gir en oversikt over deviansen for de forskjellige faktormodellene.

Tabell 24. Antall tilfelle av testikkelkreft i Danmark i årene 1943–1996 i 5-års-klasser. Antall personår i tusen (menn) i parentes. Etter Carstensen (2007).

Alder\ periode	1943–47	1948–52	1953–57	1958–62	1963–67	1968–72	1973–77	1978–82	1983–87	1988–92	1993–96
15 – 19	10 (2321)	7 (2233)	13 (2382)	13 (2919)	15 (3155)	33 (2883)	35 (2858)	37 (3033)	49 (3015)	51 (2789)	41 (2011)
20 – 24	30 (2439)	31 (2234)	46 (2165)	49 (2313)	55 (2881)	85 (3162)	110 (2902)	140 (2859)	151 (3059)	150 (3052)	112 (2283)
25 – 29	55 (2372)	62 (2345)	63 (2169)	82 (2096)	87 (2294)	103 (2888)	153 (3168)	201 (2883)	214 (2869)	268 (3095)	194 (2507)
30 – 34	56 (2398)	66 (2324)	82 (2308)	88 (2135)	103 (2100)	124 (2310)	164 (2881)	207 (3136)	209 (2865)	258 (2871)	251 (2464)
35 – 39	53 (2308)	56 (2349)	56 (2281)	67 (2281)	99 (2135)	124 (2107)	142 (2302)	152 (2856)	188 (3107)	209 (2846)	199 (2292)
40 – 44	35 (2082)	47 (2263)	65 (2305)	64 (2250)	67 (2270)	85 (2129)	103 (2090)	119 (2273)	121 (2821)	155 (3071)	126 (2264)
45 – 49	29 (1866)	30 (2030)	37 (2214)	54 (2260)	45 (2214)	64 (2239)	63 (2095)	66 (2047)	92 (2229)	86 (2770)	96 (2453)
50 – 54	16 (1618)	28 (1801)	22 (1962)	27 (2146)	46 (2198)	36 (2155)	50 (2173)	49 (2027)	61 (1982)	64 (2163)	51 (2105)
55 – 59	6 (1413)	14 (1538)	16 (1713)	25 (1868)	26 (2042)	29 (2095)	28 (2051)	43 (2059)	42 (1923)	34 (1883)	45 (1634)
60 – 64	9 (1210)	12 (1305)	11 (1424)	13 (1584)	20 (1720)	18 (1880)	28 (1930)	23 (1884)	26 (1890)	15 (1772)	10 (1392)

For sammenlikningens skyld er dataene i tabell 24 gruppert på nytt, og nå fordelt på 10-års-klasser. For at det skal gå opp er den siste perioden (1993–1996) utelatt. Dataene gruppert i 10-års-klasser er vist i tabell 26. Dataene er så analysert på nytt med de samme metodene, og tabell 27 gir en oversikt over deviansen for de ulike modellene. Hvis vi sammenlikner tabell 25 og tabell 27 ser vi at både for 5-års-intervaller og 10-års-intervaller ser det ut som alder-periode-kohort-modellen er den klart beste modellen. Resultatene fra metoden med andreordensdifferanser er vist grafisk i figur 6. Hvis vi sammenlikner med kurvene i Carstensen (2007) ser vi at vi får en brå endring på kohortkurven samme sted som han har en markert bunn, nemlig rundt andre verdenskrig.

Tabell 25. Devians og antall frihetsgrader for ulike modeller.
I alle modellene er både alder og periode delt inn i 5-års-klasser.

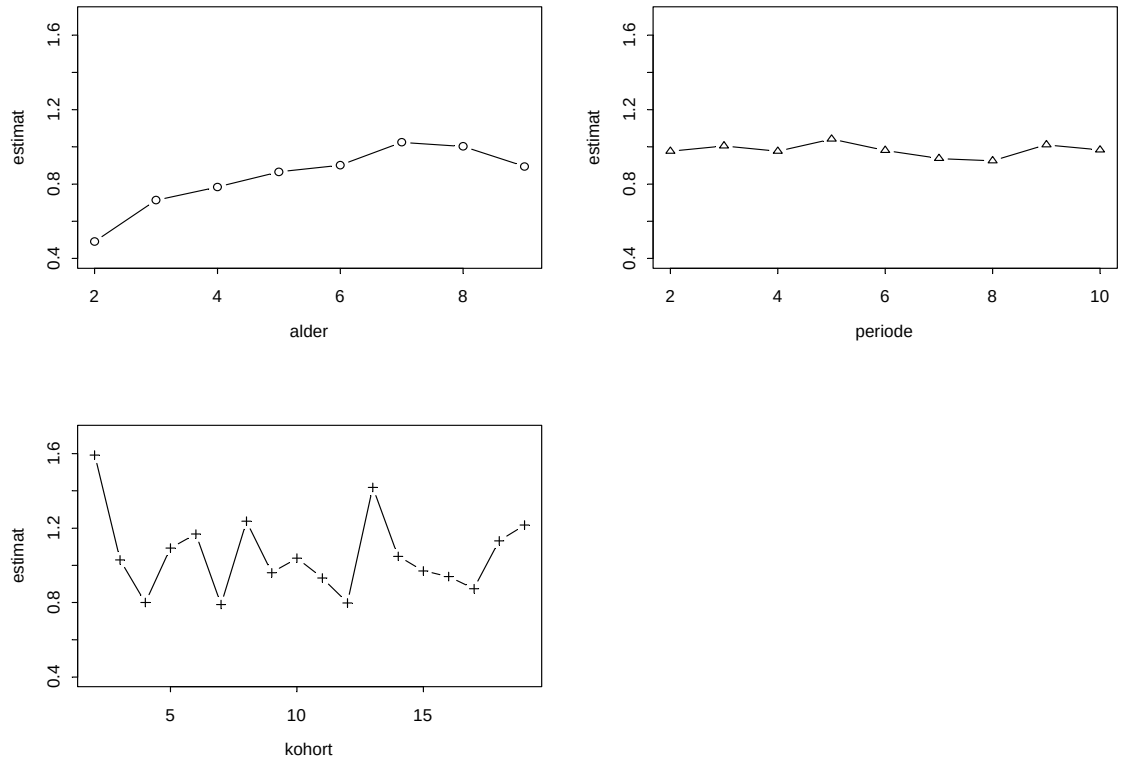
Modell	Devians	Antall frihetsgrader
Alder	1371	100
Alder + drift (kohort)	185	99
Alder + kohort	120	81
Alder + periode + kohort	65	72
Alder + periode	164	90
Alder + drift (periode)	185	99

Tabell 26. Antall tilfelle av testikkelkreft i Danmark i årene 1943–1992 fordelt på 10-års-klasser med utgangspunkt i tabell 24. Antall personår i tusen (menn) i parentes.

Alder\periode	1943–52	1953–62	1963–72	1973–82	1983–92
15 – 24	78 (9227)	121 (9779)	188 (12081)	322 (11652)	401 (11915)
25 – 34	239 (9436)	315 (8708)	417 (9592)	725 (12068)	949 (11700)
35 – 44	191 (9002)	252 (9117)	375 (8641)	516 (9521)	673 (11845)
45 – 54	103 (7315)	140 (8582)	191 (8806)	228 (8342)	303 (9144)
55 – 64	41 (5466)	65 (6589)	93 (7737)	122 (7924)	117 (7468)

Tabell 27. Devians og antall frihetsgrader for ulike modeller.
I alle modellene er både alder og periode delt inn i 10-års-klasser.

Modell	Devians	Antall frihetsgrader
Alder	1013	20
Alder + drift (kohort)	59	19
Alder + kohort	36	12
Alder + periode + kohort	8,4	9
Alder + periode	50	16
Alder + drift (periode)	59	19



Figur 6. Testikkelkreft i Danmark 1943–1996. Estimater av andreordensdifferanser for alder-, periode- og kohorteffekter.

8

Diskusjon

I denne oppgaven har vi sett at bruk av aldersspesifikke rater er et alternativ til aldersstandardiserte rater når vi skal studere forekomsten av kroniske sykdommer, som for eksempel kreft, over tid. For å finne eventuelle tendenser i sykdomsforekomsten har vi brukt de tre tidsvariablene alder, kalenderperiode og fødselskohort. Det faktum at det er en direkte lineær sammenheng mellom de tre variablene skaper ekstra problemer i statistiske analyser. I denne oppgaven har vi presentert noen av metodene som er foreslått for å løse disse problemene. Metodene kan deles inn i flere trinn, fra innhenting av data og valg av modell til parametrisering og valg av restriksjoner og til slutt presentasjon av resultatene.

Oppsummering og diskusjon

Det er ulike oppfatninger om hvordan data skal fremstilles. I nesten alle eksemplene i denne oppgaven er det brukt standardtabeller der alder og periode er oppgitt. Dette synes å være den klart vanligste måten å presentere slike data på, en måte som kan sammenliknes med måten data presenteres på i tverrsnittsstudier. Det er også mulig å fremstille data i tabeller der alder og kohort er oppgitt, et eksempel på dette er gitt i Clayton & Schifflers (1987a). Dette kan sammenliknes med datapresentasjonen i longitudinelle studier. Et alternativ til å fremstille data i en vanlig tabell er å bruke et Lexis-diagram, der både alder, periode *og* kohort er oppgitt. Bruken av Lexis-diagram er demonstrert hos Carstensen (2007). Et Lexis-diagram krever at vi har tilgang til fødselsår i tillegg til alder og periode, eller i det minste mer detaljerte opplysninger om alder og periode slik at fødselsdato kan beregnes. Slike

opplysninger er sjelden oppgitt i vanlig kreftstatistikk, det er for eksempel ikke oppgitt i rapportene fra Kreftregisteret i Norge (se for eksempel Cancer Registry of Norway 2007).

Hvor detaljerte dataene er fremstilt vil til en viss grad styre videre valg av for eksempel modeller. I denne oppgaven er det gjennomgående brukt 5-års-intervaller både for alder og kalenderperiode, mens det er brukt overlappende 10-års-intervaller for fødselskohort. Alle metodene som er lagt frem i denne oppgaven, med unntak av Carstensens metode i kapittel 7, tar utgangspunkt i at alder og periode er delt i like lange intervaller. Dersom alder og periode er delt i ulike lange intervaller skaper dette ekstra problemer, noe som ikke er undersøkt nærmere i denne oppgaven. Mange av metodene som er omtalt i litteraturen forutsetter også like lange intervaller, men for eksempel Holford (1983) tar også for seg metoder med ulike lange intervaller. Han har også i en nyere artikkel (Holford 2006) tatt spesielt for seg problemet med ulike lange intervaller. Her foreslår han blant annet, som én av metodene, å bruke ”splines”, og dermed nærmer han seg Carstensens metode.

Alle modellene som er brukt i denne oppgaven er poissonmodeller, noe som går igjen i det meste av litteraturen som er skrevet om hvordan effekten av alder, periode og kohort skal modelleres. Det har også vært vanlig å bruke faktormodeller, og det har også jeg gjort i alle eksemplene mine. Carstensen (2007) på sin side anbefaler å modellere effekten av alder, periode og kohort ved hjelp av parametrisk glatte funksjoner. Da slipper han også unna problemet med ulike intervallengder for alder og periode som ble kommentert i forrige avsnitt.

Som nevnt i begynnelsen av kapittelet er det et problem at de tre variablene er direkte lineært avhengige. En måte dette problemet kan omgås på, er å se på bare to variabler om gangen, det vil si enten å se på en alder-periode-modell eller en alder-kohort-modell, slik som hos Clayton & Schifflers (1987a). Jeg har også sett på slike modeller, problemet er å vurdere om modellene beskriver dataene på en tilfredsstillende måte. Det har vært vanlig å bruke deviansen for å undersøke ulike modellers ”godhet”, se for eksempel Clayton & Schifflers (1987a+b). Imidlertid hevder Carstensen (2007) at deviansen er en størrelse som avhenger mer av den valgte tabelleringen enn av om modellen beskriver data på en adekvat måte. For å se nærmere på denne påstanden har jeg brukt eksempelet med forekomsten av testikkelkreft i Danmark 1943–1996, og delt inn data på to måter: i 5-års-klasser og 10-års-klasser (tabell 24 og 26). Deretter har jeg sammenliknet deviansen for ulike modeller for både 5-års-klasser og

10-års-klasser (tabell 25 og 27). Hvis vi først ser på alder-periode-kohort-modellen i de to tilfellene, så ser vi at verken 5-års-klasser eller 10-års-klasser gir signifikans, noe som kan tolkes som at det er en "god" modell i begge eksemplene. Hvis vi nå ser på alder-kohort-modellen og sammenlikner deviansen for 5- og 10-års-intervaller, så får vi signifikans i begge tilfellene, men med betydelig forskjell i p-verdi (hhv. $3 \cdot 10^{-3}$ og $3 \cdot 10^{-4}$). Carstensen har nok et poeng, men deviansen kan vel i det minste brukes til å sammenlikne modeller.

I alder-periode-kohort-modellen med alle tre variablene til stede samtidig, må det gjøres et vilkårlig valg av to absolutte nivå og en trend dersom man ønsker å rapportere estimerte effekter, uansett hvilken metode som velges. Faktormodellen er den mest brukte modellen, og det er skrevet mye litteratur om hvordan denne modellen kan parametriseres på best mulig måte. Vi har sett på metoden med førsteordensdifferanser etter Clayton & Schifflers (1987b), hvor for eksempel verdien av én kohort blir satt lik null, og verdien av to perioder blir satt lik null, noe som fjerner lineær trend eller drift fra periode. En annen metode vi har sett på, også beskrevet av Clayton & Schifflers (1987b), er metoden med andreordensdifferanser, hvor bare den estimerbare krumningen blir rapportert. Vi har også sett på Holfords metode (Holford (1983 og 1991), hvor lineær trend blir fjernet fra alle de tre variablene, og netto drift rapporteres som summen av to stigningstall. Vi gir også en kort innføring i Carstensens metode (Carstensen 2007), men denne metoden blir ikke demonstrert ved hjelp av eksempler. Carstensen bruker ikke faktormodellen, men en modell med parametrisk glatte funksjoner.

Et viktig spørsmål er hvordan resultatene skal tolkes og rapporteres. Krumningen, som er estimerbar, har vært oppgitt på flere ulike måter. I oppgaven har vi sett at krumningen kan oppgis som førsteordensdifferanser, andreordensdifferanser eller avvik fra linearitet. Av figur 5 fremgår det at det er en tydelig sammenheng mellom førsteordensdifferanser hos Clayton & Schifflers (1987b) og avvik fra linearitet hos (Holford (1983). Driften eller lineær trend kan også oppgis på forskjellige måter. Vi har sett at driften kan oppgis som et gjennomsnitt av førsteordensdifferanser (Clayton & Schifflers 1987b) eller som netto drift (Holford (1983). Hvis vi ser på eksempelet med forekomst av prostatakreft hos ikke-hvite menn i USA, så er driften δ etter den første metoden $\delta = -0{,}107$ og etter den andre metoden $\delta = 0{,}115$. Vi ser at driften tilsynelatende er negativ etter den første metoden, mens den er positiv etter den andre metoden. Dette skyldes bare ulik skrivemåte, idet jeg bare har fulgt skrivemåten til Clayton & Schifflers (1987b) som adderer driftsleddet, i stedet for å gjøre som Holford (1983)

som trekker fra lineær trend. Det siste er vel mer naturlig, og da ser vi at driften hos Clayton & Schifflers er omtrent det samme som nettodriften hos Holford.

Det har vært vanlig å oppgi de estimerte parameterverdiene i en tabell. Dette er også mest brukt i denne oppgaven, selv om det noen figurer innimellom. Carstensen (2007) vil heller oppgi absolutt nivå med aldersparameterne, det vil si rapportere aldersspesifikke rater. Carstensen ønsker også å presentere estimatene som linjegrafer med konfidensgrenser.

Konklusjon og videre arbeid

Er det så mulig å trekke noen konklusjoner? Kan vi plukke ut én metode, og si at dette er den beste metoden? Carstensen (2007) mener i hvert fall at en modell med parametrisk glatte funksjoner er bedre enn de tradisjonelle faktormodellene. Robertson & Boyle (1998) sammenlikner flere metoder basert på en faktormodell, og anser metoder basert på estimerbare funksjoner som krumning og avvik fra linearitet som de mest egnede. Jeg tror at datatilgjengeligheten til en viss grad kan begrense valg av metode.

Hvis jeg til slutt skal forsøke å komme med en konklusjon, så vil jeg si: Velg en metode der både krumning og drift blir rapportert i en eller annen form, men uansett hvilken metode man velger er det viktig å oppgi forutsetninger og valg som er gjort underveis.

I fremtidige arbeider med å sammenlikne ulike metoder, kan man for eksempel bruke samme eksempel, eller aller helst flere eksempler, på alle metodene. Det kan være både metoder som bygger på faktormodeller og metoder som bruker glatte funksjoner. I tillegg bør man variere intervallengdene. Siste ord er sikkert ikke sagt i debatten om hvilken metode som er best.

Appendiks A

Eksempel på dataramme i S-plus. Dataramme for tabell 4 (mortalitetsrater for blærekreft hos italienske menn 1955–1979).

		1	2	3	4	5	6
		antall	alder	periode	pop	kohort	rate
1		3.00	1.00	1.00	10000000.00	11.00	0.03
2		16.00	2.00	1.00	9411765.00	10.00	0.17
3		24.00	3.00	1.00	7500000.00	9.00	0.32
4		79.00	4.00	1.00	7596154.00	8.00	1.04
5		234.00	5.00	1.00	8181818.00	7.00	2.86
6		458.00	6.00	1.00	6897590.00	6.00	6.64
7		720.00	7.00	1.00	5664831.00	5.00	12.71
8		890.00	8.00	1.00	4425659.00	4.00	20.11
9		891.00	9.00	1.00	3651639.00	3.00	24.40
10		920.00	10.00	1.00	2804023.00	2.00	32.81
11		831.00	11.00	1.00	1824769.00	1.00	45.54
12		3.00	1.00	2.00	10000000.00	12.00	0.03
13		17.00	2.00	2.00	9444444.00	11.00	0.18
14		29.00	3.00	2.00	9354839.00	10.00	0.31
15		76.00	4.00	2.00	7238095.00	9.00	1.05
16		185.00	5.00	2.00	7341270.00	8.00	2.52
17		552.00	6.00	2.00	7852063.00	7.00	7.03
18		867.00	7.00	2.00	6474981.00	6.00	13.39
19		1230.00	8.00	2.00	5129274.00	5.00	23.98
20		1266.00	9.00	2.00	3817853.00	4.00	33.16
21		1243.00	10.00	2.00	2937840.00	3.00	42.31
22		937.00	11.00	2.00	1954526.00	2.00	47.94
23		1.00	1.00	3.00	10000000.00	13.00	0.01
24		11.00	2.00	3.00	9166667.00	12.00	0.12
25		33.00	3.00	3.00	9428571.00	11.00	0.35
26		82.00	4.00	3.00	9010989.00	10.00	0.91
27		183.00	5.00	3.00	7011494.00	9.00	2.61
28		450.00	6.00	3.00	6998445.00	8.00	6.43
29		1069.00	7.00	3.00	7326936.00	7.00	14.59
30		1550.00	8.00	3.00	5807419.00	6.00	26.69
31		1829.00	9.00	3.00	4342355.00	5.00	42.12
32		1584.00	10.00	3.00	2996280.00	4.00	52.87
33		1285.00	11.00	3.00	2070911.00	3.00	62.05
34		4.00	1.00	4.00	10000000.00	14.00	0.04
35		8.00	2.00	4.00	10000000.00	13.00	0.08
36		39.00	3.00	4.00	9285714.00	12.00	0.42
37		95.00	4.00	4.00	9134615.00	11.00	1.04
38		267.00	5.00	4.00	8782894.00	10.00	3.04
39		431.00	6.00	4.00	6671827.00	9.00	6.46
40		974.00	7.00	4.00	6653005.00	8.00	14.64
41		1840.00	8.00	4.00	6678766.00	7.00	27.55
42		2395.00	9.00	4.00	5013607.00	6.00	47.77
43		2292.00	10.00	4.00	3472201.00	5.00	66.01
44		1787.00	11.00	4.00	2111045.00	4.00	84.65
45		12.00	1.00	5.00	10000000.00	15.00	0.12
46		8.00	2.00	5.00	8888889.00	14.00	0.09
47		30.00	3.00	5.00	9375000.00	13.00	0.32
48		115.00	4.00	5.00	9055118.00	12.00	1.27
49		285.00	5.00	5.00	9018987.00	11.00	3.16
50		723.00	6.00	5.00	8536009.00	10.00	8.47
51		1004.00	7.00	5.00	6129426.00	9.00	16.38
52		1811.00	8.00	5.00	6347704.00	8.00	28.53
53		3028.00	9.00	5.00	6011515.00	7.00	50.37
54		3176.00	10.00	5.00	4255091.00	6.00	74.64
55		2659.00	11.00	5.00	2551579.00	5.00	104.21

Appendiks B

Eksempel på program i S-plus. Blærekreft hos italienske menn 1955–1979. Standardprogram for generaliserte lineære modeller med poissonfordeling, men med egenprodusert kontrastmatrise for kohort. Både selve programmet og resultatet av kjøring av programmet er vist.

```
> #  Eks. 2:  Blærekreft Italia
#  alder + kohort
#
#
attach(blarekreft.italia)
> attach(kontr.kohort2)
> fald <- factor(alder)
> fort <- factor(kohort)
> options(contrasts = c("contr.treatment", "contr.poly"))
> contrasts(fort) <- kontr.kohort2
> modell.kohort2a <- glm(antall ~ fald - 1 + fort + offset(log(pop)), family =
   poisson)
> summary(modell.kohort2a, correlation = F)


Call: glm(formula = antall ~ fald - 1 + fort + offset(log(pop)), family = poisson)
Deviance Residuals:
       Min         1Q        Median        3Q      Max
 -2.090273 -0.6838741 4.140271e-007 0.7347972 1.501223

Coefficients:
                Value Std. Error      t value
  fald1 -15.228439290 0.31753086  -47.9589274
  fald2 -13.704750360 0.15159111  -90.4060282
  fald3 -12.733382944 0.09562480 -133.1598413
  fald4 -11.622762149 0.05923564 -196.2123190
  fald5 -10.563463145 0.03777007 -279.6781454
  fald6  -9.607457117 0.02656116 -361.7107517
  fald7  -8.823364159 0.02137665 -412.7571968
  fald8  -8.173032762 0.01909212 -428.0840096
  fald9  -7.608727885 0.02258402 -336.9075581
 fald10  -7.143638656 0.02363544 -302.2427257
 fald11  -6.707365514 0.02548281 -263.2113294
 fortk1  -0.986968653 0.04304350  -22.9295608
 fortk2  -0.907749363 0.03276515  -27.7047182
 fortk3  -0.667148148 0.02790330  -23.9092913
 fortk4  -0.380849601 0.02483357  -15.3360775
 fortk5  -0.164991132 0.02297279   -7.1820227
 fortk6  -0.052809913 0.02214129   -2.3851326
 fortk7   0.005664404 0.02191338    0.2584906
 fortk9   0.059184235 0.03094375    1.9126395
fortk10   0.194384161 0.03895746    4.9896522
fortk11   0.196560040 0.06004162    3.2737301
fortk12   0.326452588 0.09529138    3.4258354
fortk13  -0.056591931 0.18295227   -0.3093262
              Value Std. Error     t value
fortk14 -0.02990751  0.3233421 -0.09249493
fortk15  1.59525029  0.4291377  3.71733883

(Dispersion Parameter for Poisson family taken to be 1)
```

```
    Null Deviance: 531430.5 on 55 degrees of freedom

Residual Deviance: 39.39531 on 30 degrees of freedom

Number of Fisher Scoring Iterations: 4
```

- #

```
fitted(modell.kohort2a)
        1        2        3        4        5       6       7        8
 2.963042 12.77008 23.48155 68.06005 212.6454 439.842 707.284 853.2456

        9      10  11       12       13       14       15       16       17
 929.6523 893.528 831 3.374034 12.84233 33.52883 68.80598 189.7219 530.8578

       18       19       20       21      22       23       24      25
 904.4094 1227.151 1294.165 1190.821 963.472 2.300357 14.19353 33.8667

       26       27       28       29       30       31       32       33
 98.05979 192.2473 470.4743 1085.036 1554.337 1826.594 1617.107 1298.527

       34       35       36       37       38       39       40       41
 2.362567 10.55662 37.97991 99.62165 275.6791 475.8637 979.6693 1895.192

       42       43       44 45       46       47       48       49       50
 2359.319 2325.451 1762.482 12 9.637433 26.14302 112.4525 283.7063 696.9622

       51       52       53       54       55
 957.6015 1791.074 2999.269 3188.093 2643.519
```

Appendiks C

Eksempel på program i S-plus. Brystkreft hos japanske kvinner 1955–1979. Egenprodusert program bygget opp fra grunnen av. Programmet beregner både førsteordensdifferanser og andreordensdifferanser, det siste også med figur. Både selve programmet og resultatet av kjøring av programmet er vist.

```
> #  Eks.4:  Brystkreft Japan
#  alder + periode + kohort
#  med designmatrise
#
#
attach(brystkreft)
> attach(xmatr2.eks4)
> xx <- cbind(a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, p2, p3, p4, k1,
    k2, k3, k4, k5, k6, k7, k9, k10, k11, k12, k13, k14, k15)
> xtx <- t(xx) %*% xx
> zz <- log(antall) - log(pop)
> beta <- solve(xtx) %*% (t(xx) %*% zz)
> beta
numeric matrix: 28 rows, 1 columns.
            [,1]
 a1 -12.59622159
 a2 -11.10570932
 a3 -10.25415634
 a4  -9.65185697
 a5  -9.21203290
 a6  -8.90940031
 a7  -8.75024542
 a8  -8.68999201
 a9  -8.61169807
a10  -8.50748199
a11  -8.37770522
 p2  -0.08430399
 p3  -0.07990092
 p4  -0.01952348
 k1  -0.24651654
 k2  -0.27358247
 k3  -0.31877681
 k4  -0.35358362
 k5  -0.34188144
 k6  -0.24130832
 k7  -0.13020261
 k9   0.10512960
k10   0.16424445
         [,1]
k11 0.2271237
k12 0.2338228
k13 0.3820233
k14 0.5110127
k15 0.6976336
> #
```

```
#  Iterasjoner:
#
enmat <- diag(55)
> n <- 1
> while(n < 5) {
    eksbe <- xx %*% beta
    w <- pop * exp(eksbe)
    ww <- w * enmat
    zzz <- eksbe + antall/w - 1
    bb <- solve(t(xx) %*% ww %*% xx) %*% (t(xx) %*% ww %*% zzz)
    print(bb)
    beta <- bb
    n <- n + 1
}
numeric matrix: 28 rows, 1 columns.
            [,1]
 a1 -12.59932354
 a2 -11.10588400
 a3 -10.26785426
 a4  -9.66587148
 a5  -9.21937314
 a6  -8.91200124
 a7  -8.75354237
 a8  -8.69410261
 a9  -8.60765379
a10  -8.50986240
a11  -8.37444410
 p2  -0.07817781
 p3  -0.08372801
 p4  -0.02821774
 k1  -0.24977767
 k2  -0.27208462
 k3  -0.32484549
 k4  -0.35078146
 k5  -0.33987951
 k6  -0.23854919
 k7  -0.12349548
 k9   0.11138342
k10   0.18844311
         [,1]
k11 0.2343903
k12 0.2628578
k13 0.3674931
k14 0.5171570
k15 0.7007356
numeric matrix: 28 rows, 1 columns.
            [,1]
 a1 -12.59944312
 a2 -11.10600184
 a3 -10.26790202
 a4  -9.66591786
 a5  -9.21936862
 a6  -8.91208328
 a7  -8.75359600
 a8  -8.69414331
 a9  -8.60772797
a10  -8.50992433
a11  -8.37451555
 p2  -0.07818147
 p3  -0.08370402
 p4  -0.02819571
 k1  -0.24970622
 k2  -0.27204284
 k3  -0.32478959
 k4  -0.35075366
 k5  -0.33984092
 k6  -0.23850125
 k7  -0.12347992
```

```
 k9   0.11141236
k10   0.18835190
         [,1]
k11 0.2343598
k12 0.2627864
k13 0.3671969
k14 0.5172537
k15 0.7008552
numeric matrix: 28 rows, 1 columns.
            [,1]
 a1 -12.59944313
 a2 -11.10600185
 a3 -10.26790202
 a4  -9.66591786
 a5  -9.21936861
 a6  -8.91208328
 a7  -8.75359600
 a8  -8.69414331
 a9  -8.60772797
a10  -8.50992433
a11  -8.37451555
 p2  -0.07818147
 p3  -0.08370402
 p4  -0.02819571
 k1  -0.24970622
 k2  -0.27204284
 k3  -0.32478959
 k4  -0.35075366
 k5  -0.33984092
 k6  -0.23850125
 k7  -0.12347992
 k9   0.11141236
k10   0.18835189
         [,1]
k11 0.2343598
k12 0.2627864
k13 0.3671969
k14 0.5172538
k15 0.7008552
numeric matrix: 28 rows, 1 columns.
            [,1]
 a1 -12.59944313
 a2 -11.10600185
 a3 -10.26790202
 a4  -9.66591786
 a5  -9.21936861
 a6  -8.91208328
 a7  -8.75359600
 a8  -8.69414331
 a9  -8.60772797
a10  -8.50992433
a11  -8.37451555
 p2  -0.07818147
 p3  -0.08370402
 p4  -0.02819571
 k1  -0.24970622
 k2  -0.27204284
 k3  -0.32478959
 k4  -0.35075366
 k5  -0.33984092
 k6  -0.23850125
 k7  -0.12347992
 k9   0.11141236
k10   0.18835189
         [,1]
k11 0.2343598
k12 0.2627864
k13 0.3671969
```

```
k14 0.5172538
k15 0.7008552
> #
#  y-hatt:
#
w
        1        2        3       4        5        6        7        8
 85.29872 320.8554 577.0248 840.952 1051.898 1047.409 959.7628 804.2654

        9       10  11      12       13       14       15       16       17
 713.2518 621.2812 497 83.2943 342.9093 676.0791 959.8136 1189.237 1282.944

       18       19       20       21       22       23       24       25
 1086.381 890.1751 729.8235 609.6238 481.7188 98.35904 372.7072 783.9653

       26       27       28       29       30       31       32       33
 1210.606 1461.792 1564.175 1429.803 1078.975 867.9149 676.5775 519.1244

       34       35       36       37       38       39       40       41
 127.0479 464.5761 912.8659 1497.495 1978.741 2086.753 1874.436 1542.385

       42      43       44  45       46       47       48       49       50
 1155.041 896.324 647.3336 179 587.9521 1103.065 1711.133 2401.331 2744.718

       51     52       53       54       55
 2445.616 2002.2 1650.969 1182.194 859.8232
> #
#
aa <- beta[1:11]
> la <- length(aa)
> pp <- c(0, beta[12:14], 0)
> lp <- length(pp)
> kk <- c(beta[15:21], 0, beta[22:28])
> lk <- length(kk)
> aaa <- aa
> ppp <- pp
> kkk <- kk
> i <- 2
> while(i < la) {
   aaa[i] <- aa[i + 1] - 2 * aa[i] + aa[i - 1]
   i <- i + 1
}
> j <- 2
> while(j < lp) {
   ppp[j] <- pp[j + 1] - 2 * pp[j] + pp[j - 1]
   j <- j + 1
}
> c <- 2
> while(c < lk) {
   kkk[c] <- kk[c + 1] - 2 * kk[c] + kk[c - 1]
   c <- c + 1
}
> #
#  Andreordensdifferanser:
#
astar <- aaa[2:(la - 1)]
> pstar <- ppp[2:(lp - 1)]
> kstar <- kkk[2:(lk - 1)]
> #
#
exp(astar)
[1] 0.5192647 0.7896893 0.8560428 0.8699984 0.8617431 0.9057114 1.0273294
[8] 1.0114534 1.0383212
> exp(pstar)
[1] 1.075364 1.062932 0.973057
> exp(kstar)
 [1] 0.9700476 1.0271446 1.0375652 1.0946415 1.0137757 1.0084945 0.9880050
 [8] 0.9661146 0.9695418 0.9825724 1.0789451 1.0467042 1.0341135
```

```
> #
#
xalder <- c(2:10)
> xper <- c(2:4)
> xkohort <- c(2:14)
> yalder <- exp(astar)
> yper <- exp(pstar)
> ykohort <- exp(kstar)
> par(mfrow = c(2, 2))
> plot(xalder, yalder, type = "b", ylim = c(0.5, 1.1), xlab = "Alder", ylab =
   "Estimat")
> plot(xper, yper, type = "b", ylim = c(0.5, 1.1), xlab = "Periode", ylab =
   "Estimat")
> plot(xkohort, ykohort, type = "b", ylim = c(0.5, 1.1), xlab = "Kohort", ylab
    = "Estimat")
```

Litteratur

Aalen, O.O. (red.), Frigessi, A., Moger, T.A., Scheel, I., Skovlund, E. og Veierød, M.B. 2006. Statistiske metoder i medisin og helsefag. Gyldendal. ISBN 82-05-34685-2.

Cancer Registry of Norway. 2007. Cancer in Norway 2006. Oslo, Norway.

Carstensen, B. 2007. Age-period-cohort models for the Lexis diagram. *Statistics in Medicine* **26**: 3018-3045.

Clayton, D. & Schifflers, E. 1987a. Models for temporal variation in cancer rates. I: Age-period and age-cohort models. *Statistics in Medicine* **6**: 449-467.

Clayton, D. & Schifflers, E. 1987b. Models for temporal variation in cancer rates. II: Age-period-cohort models. *Statistics in Medicine* **6**: 469-481.

Crawley, M.J. 2002. Statistical computing. An introduction to data analysis using S-plus. Wiley. ISBN 0-471-56040-5.

Dobson, A. 2002. An introduction to generalized linear models. 2nd ed. Chapman & Hall. ISBN 1-58488-165-8.

Holford, T.R. 1983. The estimation of age, period and cohort effects for vital rates. *Biometrics* **39**: 311-324.

Holford, T.R. 1991. Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Revue of Public Health* **12**: 425-457.

Holford, T.R. 2006. Approaches to fitting age-period-cohort models with unequal intervals. *Statistics in Medicine* **25**: 977-993.

Kleinbaum, D.G., Kupper, L.L. and Morgenstern, H. 1982. Epidemiologic research. Principles and quantitative methods. Van Nostrand Reinhold. ISBN 0-534-97950-5.

Robertson, C. & Boyle, P. 1998. Age-period-cohort analysis of chronoc disease rates. I: Modelling approach. *Statistics in Medicine* **17**: 1305-1323.

Venables, W.N. & Ripley, B.D. 2002. Modern applied statistics with S. Fourth ed. Springer. ISBN 0-387-95457-0.