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** ................................. 7
2. **Begreper og definisjoner i epidemiologi** .................................................. 9
   - Epidemiologiske grunnbegreper .................................................. 9
   - Alder, periode og kohort .................................................. 12
3. **Data** .................................................. 14
   - Innhenting av data .................................................. 14
   - Grafiske fremstillinger .................................................. 16
   - Lexis-diagram .................................................. 20
   - Data i S-plus .................................................. 21
4. **Modeller** .................................................. 23
   - Generaliserte lineære modeller (GLM) .................................................. 23
   - Poissonmodeller .................................................. 25
   - ”Maximum likelihood”-estimering i GLM .................................................. 27
   - Alder-periode-kohort-modeller .................................................. 28
5. **Parametrerisering av modeller med én og to variabler** .................................................. 31
   - Estimering av alderseffekter .................................................. 32
   - Parametrerisering av alder-periode-modeller .................................................. 37
   - Parametrerisering av alder-kohort-modeller .................................................. 42
   - Parametrerisering av alder-drift-modeller .................................................. 46
6 Parametrisering av alder-periode-kohort-modeller 49
   Metode 1: Førsteordensdifferanser ......................... 51
   Metode 2: Andreordensdifferanser ......................... 58

7 Alternative metoder 63
   Holfords metode .............................................. 63
   Carstensens metode ........................................... 67

8 Diskusjon 72

A 76
B 77
C 79

Litteratur 84
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 parametrises. 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 8 får vi en oppsummering og en diskusjon av de ulike metodene. Til slutt gjøres det et forsøk på å formulere en konklusjon.
Begreper og definisjoner i epidemiologi

Epidemiologiske grunnbegreper

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 \textit{frekvens}. Et av de mest brukte målene i epidemiologi er \textit{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 finns 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 $\Delta t_i$ 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 $\Delta t$ 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 $\Delta t$ 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.

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

Forventet antall døde i hver aldersgruppe fås ved å multiplisere de aldersspesifikke ratene med antall personer i den tilsvarende aldersgruppen i referansepopulasjonen. I dette eksempelen 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.


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 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.
Data

Innhenting av data


Eksempel

Tabell 2 viser insidensraten for blærekreft hos menn i kretfregisteret for Birmingham i perioden 1960–1976 (Clayton & Schifflers 1987a). Dataene i tabellen er lagt ut slik at de reflekerer 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.


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

Eksempel

Tabell 3 viser antall dødsfall av prosatakreft 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.

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

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


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


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 opp 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

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

\[ \text{populasjon} = \text{antall} \cdot 100000 / \text{rate} \]

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

\[ f(y; \theta) = \frac{\theta^y \cdot e^{-\theta}}{y!} \quad y = 0, 1, 2, \ldots \]
Dette kan omskrives til
\[ f(y; \theta) = \exp(y \cdot \log \theta - \theta - \log y!) \]
som viser at poissonfordelingen tilhører den eksponentielle familie med \( a(y) = y \) (alså på kanonisk form), \( b(\theta) = \log \theta \), \( c(\theta) = -\theta \) og \( d(y) = -\log y! \).

**Eksempel**

Tettheten for normalfordelingen er gitt ved
\[ f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right) \]
hvor \( \mu \) er den parameteren vi er interesserte i og \( \sigma^2 \) 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 fordelen fra den eksponentielle familie og er på kanonisk form, men parameterne \( \theta_i \) trenger ikke være like alle sammen.

2. I modellene bruker vi vanligvis ikke \( \theta_i \), men et mindre antall parametere \( \beta_1, \ldots, \beta_p \) (hvor \( p < N \)). Vi kan samle parameterne i en vektor
\[
\beta = \begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_p 
\end{bmatrix}
\]

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 \( X \), hvor

\[
X = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{N1} & x_{N2} & \cdots & x_{Np} 
\end{bmatrix}
\]

4. Det fins en monoton linkfunksjon \( g \) slik at

\[ g(\mu_i) = x_i^T \beta \]

hvor

\[ x_i^T = [x_{i1} \ x_{i2} \ \cdots \ x_{ip}] \] 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 **tededata** 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 \( \Delta t \). 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 $\mu$ forventet antall tilfeller i løpet av tidsrommet $\Delta t$, det vil si at $E(Y) = \mu$. For poissonfordelingen er variansen lik forventningen, det vil si at $\text{Var}(Y) = \mu$.

**Log-lineære modeller**


Effekten av forklaringsvariablene på responsvariabelen $Y$ modelleres ved hjelp av parameteren $\mu$. 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 $\theta_i$ som $\theta_i = \exp(x_i^T \beta)$, slik at den generaliserte lineære modellen blir

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

Logaritmefunksjonen er den naturlige linkfunksjonen. Det gir modellen

$$\log E(Y_i) = \log \mu_i = \log n_i + x_i^T \beta$$

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

Hele modellen kan skrives på matriseform som

$$\log E(Y) = \log \mu = \log N + X\beta$$

hvor $X$ er designmatrisen og $\beta$ er vektoren som inneholder alle parameterne. Disse to er definert tidligere, dessuten er
\[
Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix} \quad \text{og} \quad N = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{bmatrix}
\]

"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 \( \theta \) 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 \( \beta_1, \beta_2, \ldots, \beta_p \) som er relatert til \( Y_i \) gjennom \( E(Y_i) = \mu_i \) og \( g(\mu_i) = x_i^T \beta \). 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

\[
l = \sum_{i=1}^{N} l_i
\]

Vi får likningene

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

Disse likningene må løses numerisk.

Generelt for generaliserte lineære modeller får vi følgende likninger på matriseform

\[
X^T W X b = X^T W z
\]

Her er \( X \) designmatrisen og \( b \) er estimator for \( \beta \). Disse likningene må løses iterativt, fordi både \( W \) og \( z \) normalt vil avhenge av \( b \). \( W \) er en diagonalmatrise \((N \times N)\) hvor elementene er gitt ved
\[ w_{ii} = \frac{1}{\text{Var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \]

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

\( z \) har elementene gitt ved

\[ z_i = \sum_{k=1}^{\mu} x_{ik} b_k + (y_i - \mu_i) \left( \frac{\partial \eta}{\partial \mu_i} \right) . \]

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

\[ \frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial \log(\mu_i)}{\partial \mu_i} = \frac{1}{\mu_i} \] og

\[ \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\partial \eta_i} = \frac{1}{\mu_i} = \mu_i . \]

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

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

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

\[ z_i = \sum_{k=1}^{\mu} x_{ik} b_k + (y_i - \mu_i) \left( \frac{\partial \eta}{\partial \mu_i} \right) = \sum_{k=1}^{\mu} x_{ik} b_k + (y_i - \mu_i) \cdot \frac{1}{\mu_i} = \sum_{k=1}^{\mu} x_{ik} b_k + \frac{y_i}{\mu_i} - 1 . \]

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 \( \lambda \) stå for raten, der telleren antas å være poissonfordelt. La videre \( \alpha \) stå for effekten av alder, \( \beta \) for effekten av periode og \( \gamma \) 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.


\[
\log E(\lambda) = \alpha + g(p)
\]

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 av \( k \) (= 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.
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

\[ D = 2 \cdot [l (b_{max}, y) - l (b, y)] \]
hvor \( l(b_{\text{max}}; y) \) er log-likelihood-funksjonen for en mettet (maksimal) modell, mens \( l(b; y) \) 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 parametreringen. 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

\[
D_1 \sim \chi^2_{N-p} \quad \text{og} \quad D_2 \sim \chi^2_{N-q}.
\]

Da vil \( \Delta D \sim \chi^2_{p-q} \) \((p > q)\). Mer om egenskapene til deviansen kan man finne i for eksempel Dobson (2002).

De ulike parametreringene vil bli belyst med eksempler. Jeg vil starte med den enklaste modellen, med alder som eneste forklaringsvariabel. Sammen med denne enkle modellen vil jeg ta opp noen generelle problemer i forbindelse med parametrering 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

\[
\log E(\lambda_{\text{ap}}) = \alpha_a
\]
hvor \( \lambda_{ap} \) er rate i aldersgruppe \( a \) og kalenderperiode \( p \) og hvor \( \alpha_a \) er effekten av alder i aldersgruppe \( a \). Når modellen skal parametrises er det vanlig å utnytte at

\[ \lambda_{ap} = \frac{Y_{ap}}{n_{ap}} \]

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 \( \alpha_a \) for \( a = 1, 2, \ldots, A \), 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 \( \alpha_a \). Tolkningen av parameterne for denne parametriseringen er rett frem, idet \( \alpha_a \) 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 \( \mu \) som gennomsnittlig \( \log \)rate) for alle aldersgruppene, mens \( \alpha_a \) er avvik i \( \log \)rate) i gruppe \( a \) i forhold til \( \mu \). Denne parametriseringen vil gi \((A + 1)\) parametere \((\mu, \alpha_1, \ldots, \alpha_A)\), men bare \( A \) uavhengige likninger. Rangen til designmatrisen \( X \) blir også bare \( A \), og det fører til at \( X^TX \) 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 $a_1 = 0$, eller generelt med gruppe $a$ som referansegruppe setter vi $a_\alpha = 0$. De andre gruppende 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 $\sum_{a=1}^{A} a \alpha = 0$. 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

$$\log Y_{ap} = \mu + a \alpha + \log n_{ap}$$

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 $\mu$, 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 $C$. 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

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 $\alpha_a$ 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 $\mu$.

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 $\mu$ som gjennomsnittsverdien i aldersgruppe 1. Videre kan $\alpha_a$ 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 $\mu$ kan tolkes som gjennomsnittlig log(rate) for alle gruppene samlet. Videre kan vi tolke $\alpha_a$ som avvik i log(rate) i gruppe $a$ i forhold til gjennomsnittet for alle gruppene. Også i dette tilfellet vil log(rate) i gruppe 2 kunne regnes ut som $\mu + \alpha_2$.

Legg merke til at $\alpha_{11}$ 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 $\mu$ representere gjennomsnittlig log(rate) for alle gruppene samlet. Derimot er det vanskelig å finne noen meningsfull direkte tolkning av $\alpha_a$. 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. Ortagonale 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.


<table>
<thead>
<tr>
<th>Parameter</th>
<th>Uten $\mu$</th>
<th>Kontrast &quot;treatment&quot;</th>
<th>Kontrast &quot;sum&quot;</th>
<th>Kontrast &quot;Helmert&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0,000</td>
<td>- 11,972</td>
<td>- 8,828</td>
<td>- 8,828</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>- 11,972</td>
<td>0,000</td>
<td>- 3,144</td>
<td>0,226</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>- 11,521</td>
<td>0,451</td>
<td>- 2,693</td>
<td>0,387</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>- 10,585</td>
<td>1,387</td>
<td>- 1,757</td>
<td>0,349</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>- 9,962</td>
<td>2,010</td>
<td>- 1,134</td>
<td>0,403</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>- 8,996</td>
<td>2,976</td>
<td>- 0,168</td>
<td>0,356</td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td>- 8,474</td>
<td>3,498</td>
<td>0,354</td>
<td>0,331</td>
</tr>
<tr>
<td>$\alpha_7$</td>
<td>- 7,933</td>
<td>4,039</td>
<td>0,894</td>
<td>0,316</td>
</tr>
<tr>
<td>$\alpha_8$</td>
<td>- 7,392</td>
<td>4,580</td>
<td>1,436</td>
<td>0,285</td>
</tr>
<tr>
<td>$\alpha_9$</td>
<td>- 7,041</td>
<td>4,931</td>
<td>1,787</td>
<td>0,263</td>
</tr>
<tr>
<td>$\alpha_{10}$</td>
<td>- 6,690</td>
<td>5,282</td>
<td>2,138</td>
<td>0,229</td>
</tr>
<tr>
<td>$\alpha_{11}$</td>
<td>- 6,540</td>
<td>5,432</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

I kolonne 3 er sammenhengen mellom $\alpha_i$ og $\alpha_i^*$ gitt ved

$$\alpha_i^* = \alpha_i - \alpha_1$$

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 $\alpha_i$ og $\alpha_i^*$ gitt ved
\[ \alpha_a^* = \alpha_a - \overline{\alpha} \quad \text{hvor} \quad \overline{\alpha} = \frac{1}{11} \sum_{a=1}^{11} \alpha_a \]

For eksempel er
\[ \alpha_1^* = \alpha_1 - \overline{\alpha} = -11,972 - (-8,828) = -3,144 \]

I kolonne 5 er sammenhengen mellom \( \alpha_a \) og \( \alpha_a^* \) 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
\[ \alpha_1^* = \frac{1}{1+1} \left( \alpha_{1+1} - \frac{1}{1} \left( \sum_{i=1}^{1} \alpha_i \right) \right) = \frac{1}{2} \left( \alpha_2 - \alpha_1 \right) = \frac{1}{2} \left( -11,521 - (-11,972) \right) = 0,226 \text{ og } \]
\[ \alpha_2^* = \frac{1}{2+1} \left( \alpha_{2+1} - \frac{1}{2} \left( \sum_{i=1}^{2} \alpha_i \right) \right) = \frac{1}{3} \left( \alpha_3 - \frac{1}{2} \left( \alpha_1 + \alpha_2 \right) \right) = \frac{1}{3} \left( -10,585 - \frac{1}{2} (-11,972 - 11,521) \right) = 0,387 \]

### 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 \( \mu, \alpha_a \) for \( a = 1, 2, \ldots, A \) og \( \beta_p \) 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 ubestemtethetsproblemet 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 \( \mu \) betraktes som gjennomsnittlig log(rate) for referansegruppen, det vil si log(rate) i aldersgruppe 1 og kalenderperiode 1, mens \( \alpha_a \) 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 \( \alpha_a \) for \( a = 1, 2, \ldots, A \) og \( \beta_p \) 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 ubestemtshetsproblemet 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 eksempeliet gjelder blærekreft hos menn i Birmingham i perioden 1960–1976. Dette eksempeliet 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.


<table>
<thead>
<tr>
<th>Alder</th>
<th>( \alpha_a )</th>
<th>Periode</th>
<th>( \beta_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 – 29</td>
<td>-12,314</td>
<td>1960 – 62</td>
<td>0,000</td>
</tr>
<tr>
<td>30 – 34</td>
<td>-11,848</td>
<td>1963 – 66</td>
<td>0,088</td>
</tr>
<tr>
<td>35 – 39</td>
<td>-10,898</td>
<td>1968 – 72</td>
<td>0,485</td>
</tr>
<tr>
<td>40 – 44</td>
<td>-10,279</td>
<td>1973 – 76</td>
<td>0,503</td>
</tr>
<tr>
<td>45 – 49</td>
<td>-9,318</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 – 54</td>
<td>-8,796</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55 – 59</td>
<td>-8,257</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 – 64</td>
<td>-7,728</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65 – 69</td>
<td>-7,392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 – 74</td>
<td>-7,036</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 – 79</td>
<td>-6,874</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Eksempel**


<table>
<thead>
<tr>
<th>Alder</th>
<th>$\alpha$</th>
<th>Periode</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 – 29</td>
<td>-14,916</td>
<td>1955 – 59</td>
<td>0,000</td>
</tr>
<tr>
<td>30 – 34</td>
<td>-13,892</td>
<td>1960 – 64</td>
<td>0,157</td>
</tr>
<tr>
<td>35 – 39</td>
<td>-12,913</td>
<td>1965 – 69</td>
<td>0,317</td>
</tr>
<tr>
<td>40 – 44</td>
<td>-11,792</td>
<td>1970 – 74</td>
<td>0,461</td>
</tr>
<tr>
<td>45 – 49</td>
<td>-10,797</td>
<td>1975 – 79</td>
<td>0,580</td>
</tr>
<tr>
<td>50 – 54</td>
<td>-9,889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55 – 59</td>
<td>-9,176</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 – 64</td>
<td>-8,614</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65 – 69</td>
<td>-8,155</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 – 74</td>
<td>-7,842</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 – 79</td>
<td>-7,596</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Eksempel


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>25 – 29</td>
<td>0,19 (3)</td>
<td>0,13 (2)</td>
<td>0,50 (7)</td>
<td>0,19 (3)</td>
<td>0,70 (10)</td>
</tr>
<tr>
<td>30 – 34</td>
<td>0,66 (11)</td>
<td>0,98 (16)</td>
<td>0,72 (11)</td>
<td>0,71 (10)</td>
<td>0,57 (7)</td>
</tr>
<tr>
<td>35 – 39</td>
<td>0,78 (11)</td>
<td>1,32 (22)</td>
<td>1,47 (24)</td>
<td>1,64 (25)</td>
<td>1,32 (15)</td>
</tr>
<tr>
<td>40 – 44</td>
<td>2,67 (36)</td>
<td>3,16 (44)</td>
<td>2,53 (42)</td>
<td>3,38 (53)</td>
<td>3,93 (48)</td>
</tr>
<tr>
<td>45 – 49</td>
<td>4,84 (77)</td>
<td>5,60 (74)</td>
<td>4,93 (68)</td>
<td>6,05 (99)</td>
<td>6,83 (88)</td>
</tr>
<tr>
<td>50 – 54</td>
<td>6,60 (106)</td>
<td>8,50 (131)</td>
<td>7,65 (89)</td>
<td>10,59 (142)</td>
<td>10,42 (134)</td>
</tr>
<tr>
<td>55 – 59</td>
<td>10,36 (157)</td>
<td>12,00 (184)</td>
<td>12,68 (189)</td>
<td>14,34 (180)</td>
<td>17,95 (177)</td>
</tr>
<tr>
<td>60 – 64</td>
<td>14,76 (193)</td>
<td>16,37 (232)</td>
<td>18,00 (262)</td>
<td>17,60 (249)</td>
<td>23,91 (239)</td>
</tr>
<tr>
<td>65 – 69</td>
<td>20,53 (219)</td>
<td>22,60 (267)</td>
<td>24,90 (323)</td>
<td>24,33 (325)</td>
<td>32,70 (343)</td>
</tr>
<tr>
<td>70 – 74</td>
<td>26,24 (223)</td>
<td>27,70 (250)</td>
<td>30,47 (308)</td>
<td>36,94 (412)</td>
<td>38,47 (358)</td>
</tr>
<tr>
<td>75 – 79</td>
<td>33,47 (198)</td>
<td>33,61 (214)</td>
<td>36,77 (253)</td>
<td>43,69 (338)</td>
<td>45,20 (312)</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Alder</th>
<th>$\alpha_a$</th>
<th>Periode</th>
<th>$\beta_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 – 29</td>
<td>-12,816</td>
<td>1955 – 59</td>
<td>0,000</td>
</tr>
<tr>
<td>30 – 34</td>
<td>-12,007</td>
<td>1960 – 64</td>
<td>0,107</td>
</tr>
<tr>
<td>35 – 39</td>
<td>-11,430</td>
<td>1965 – 69</td>
<td>0,162</td>
</tr>
<tr>
<td>40 – 44</td>
<td>-10,581</td>
<td>1970 – 74</td>
<td>0,278</td>
</tr>
<tr>
<td>45 – 49</td>
<td>-9,985</td>
<td>1975 – 78</td>
<td>0,423</td>
</tr>
<tr>
<td>50 – 54</td>
<td>-9,548</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55 – 59</td>
<td>-9,124</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 – 64</td>
<td>-8,825</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65 – 69</td>
<td>-8,502</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 – 74</td>
<td>-8,250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 – 79</td>
<td>-8,065</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Eksempel

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

<table>
<thead>
<tr>
<th>Alder</th>
<th>( \alpha_a )</th>
<th>Periode</th>
<th>( \beta_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 – 54</td>
<td>- 7,755</td>
<td>1935 – 39</td>
<td>0,000</td>
</tr>
<tr>
<td>55 – 59</td>
<td>- 6,959</td>
<td>1940 – 44</td>
<td>0,201</td>
</tr>
<tr>
<td>60 – 64</td>
<td>- 6,243</td>
<td>1945 – 49</td>
<td>0,409</td>
</tr>
<tr>
<td>65 – 69</td>
<td>- 5,719</td>
<td>1950 – 54</td>
<td>0,605</td>
</tr>
<tr>
<td>70 – 74</td>
<td>- 5,278</td>
<td>1955 – 59</td>
<td>0,711</td>
</tr>
<tr>
<td>75 – 79</td>
<td>- 4,970</td>
<td>1960 – 64</td>
<td>0,757</td>
</tr>
<tr>
<td>80 – 84</td>
<td>- 4,836</td>
<td>1965 – 69</td>
<td>0,812</td>
</tr>
</tbody>
</table>

Parameterverdiene for \( \beta_p \) som er presentert i disse fire eksemplene er ikke absolutte. Ved å velge en annen parametrisering kunne vi få andre parameterverdier for \( \beta_p \), 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 parametrisering.

**Parametrerisering 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 parametrerises på flere ulike måter. Det er ikke noe prinsipielt nytt ved parametrerisingen 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 \( \alpha_a \) for \( a = 1, 2, \ldots, A \) og \( \gamma_k \) 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 alderssspesifikke 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, riktig nok 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.


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

Eksempel

I det neste eksempel 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,

Tabell 12. Lungekreft hos belgiske kvinner 1955–1978. Alder- ($\alpha_a$) og kohort- ($\gamma_k$) parametere estimert fra ratene i tabell 8.

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

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- (\(\alpha\)) og kohort- (\(\gamma_k\))
parametre estimert fra antall dødsfall og antall
personår i tabell 3.

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

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

**Parametrising av alder-drift-modeller**

I forrige underkapittel så vi eksempel på et datasett som viste god tilpasning både til alder-
eksemeplet, og de innfører begrepet *drift* for å beskrive variasjon over tid som ikke skiller

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 \( \alpha_1, \alpha_2, \ldots, \alpha_A \) og \( \delta_p \). Dette gir \((A + 1)\) parametere og \((A + 1)\) uavhengige likninger, og vi slipper dermed ubestemtethetsproblemet og trenger ikke innføre ekstra restriksjoner. For denne parametriseringen kan vi tolke \( \exp(\alpha_a) \) som tilpassete alderssspesifikke rater for referanseperioden. Parameteren \( \delta_p \) 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 \( \alpha_1, \alpha_2, \ldots, \alpha_A \) og \( \delta_k \). 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 alderssspesifikke rater for referansekohorten. Parameteren \( \delta_k \) 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**
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 unenigheten skyldes vel først og fremst at de har ulike utgangspunkt og forskjellige oppfatninger av hva en modell er.

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

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 $\alpha_a$ for $a = 1, 2, \ldots, A$, $\beta_p$ for $p = 1, 2, \ldots, P$ og $\gamma_k$ 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, $X^TX$ 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 restriksjonen 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 flett.

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 $\delta_p$ og kohort-drift $\delta_k$, 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 $\beta_p$, vi skal fjerne trenden fra disse parameterne og lage nye parameterer $\beta_p^*$ 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 \( \delta \) er driftsparameteren. Problemet nå er å velge \( \delta \) slik at \( \beta_p^* \) 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 sukssessive førsteordensdifferanser eller ”førstedifferanser”. Vi ser først på periodeeffekter.

Førstedifferansene blir \( (\beta_2 - \beta_1), (\beta_3 - \beta_2), \ldots, (\beta_p - \beta_{p-1}) \). 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 \( \delta \), 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 \( (\gamma_2 - \gamma_1), (\gamma_3 - \gamma_2), \ldots, (\gamma_K - \gamma_{K-1}) \). Vi ønsker å fjerne trenden fra kohortkurven, og innfører nye parametere \( \gamma_k^* \). Sammenhengen mellom \( \gamma_k^* \) og \( \gamma_k \) er nå gitt ved

\[ \gamma_k^* = \gamma_k + \delta(k - k_0) \]
hvor \( \delta \) er driftsparameteren og \( k_0 \) er referansekohorten. Vi velger igjen \( \delta \), 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 $\beta_i = 0$ (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

$$\beta_1 = 0 \implies \beta_1^* = 0 \text{ og } \beta_p^* = 0$$

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

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

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

$$\log Y = \log N + X\beta$$

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

$$\log Y - \log N = X\beta$$

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

$$b^{(1)} = (X^TX)^{-1}X^Tz$$

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

$$b^{(2)} = [(X^TWX)^{(1)}]^{-1}(X^TWz)^{(1)}$$

Generelt kan vi finne iterasjon nummer $m$ som

$$b^{(m)} = [(X^TWX)^{(m-1)}]^{-1}(X^TWz)^{(m-1)}$$

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 eksempelen i S-plus er vist i appendiks C.

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

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 $\beta_p$ er uavhengige av hvilken parametrisering som er brukt som utgangspunkt. Vi ser at de tre parameterverdiene $\beta_2$ til $\beta_4$ 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 $\beta_p$ er rett frem, i vårt eksempel blir $\beta_p$ 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 $\alpha_a$ som log(rate) i referansekohorten etter justering for periodeeffekter.
Alder- ($\alpha_a$), periode- ($\beta_p$) og kohort- ($\gamma_k$) parametre estimeret fra ratene i tabell 15.

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

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 parametrerisering vi har startet med. Vi ser at alle parameterverdiene for $\gamma_k$ 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 $\gamma_k$ 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

$$\delta = -\frac{\beta_p - \beta_k}{P-1} = -\frac{0,272 - 0}{5-1} = -0,068$$

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.


| Alder | \( \alpha_a \) | \( \exp(\alpha_a) \cdot 100,000 \) | Periode | \( \beta_p \) | Kohort | \( \gamma_k \) *
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>25 – 29</td>
<td>-12,170</td>
<td>0,518</td>
<td>1955 – 59</td>
<td>0,000</td>
<td>1875 – 84</td>
<td>0,000</td>
</tr>
<tr>
<td>30 – 34</td>
<td>-10,745</td>
<td>2,155</td>
<td>1960 – 64</td>
<td>-0,010</td>
<td>1880 – 89</td>
<td>-0,090</td>
</tr>
<tr>
<td>35 – 39</td>
<td>-9,974</td>
<td>4,660</td>
<td>1965 – 69</td>
<td>0,052</td>
<td>1885 – 94</td>
<td>-0,211</td>
</tr>
<tr>
<td>40 – 44</td>
<td>-9,440</td>
<td>7,948</td>
<td>1970 – 74</td>
<td>0,175</td>
<td>1890 – 99</td>
<td>-0,305</td>
</tr>
<tr>
<td>45 – 49</td>
<td>-9,062</td>
<td>11,599</td>
<td>1975 – 79</td>
<td>0,272</td>
<td>1895 – 04</td>
<td>-0,362</td>
</tr>
<tr>
<td>50 – 54</td>
<td>-8,822</td>
<td>14,745</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55 – 59</td>
<td>-8,732</td>
<td>16,134</td>
<td>1900 – 09</td>
<td>0,328</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 – 64</td>
<td>-8,740</td>
<td>16,005</td>
<td>1910 – 19</td>
<td>-0,226</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65 – 69</td>
<td>-8,722</td>
<td>16,296</td>
<td>1915 – 24</td>
<td>-0,182</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 – 74</td>
<td>-8,692</td>
<td>16,792</td>
<td>1920 – 29</td>
<td>-0,173</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 – 79</td>
<td>-8,624</td>
<td>17,974</td>
<td>1925 – 34</td>
<td>-0,195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80 – 84</td>
<td>-8,556</td>
<td>19,364</td>
<td>1930 – 39</td>
<td>-0,234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85 – 89</td>
<td>-8,488</td>
<td>21,064</td>
<td>1935 – 44</td>
<td>-0,198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90 – 94</td>
<td>-8,420</td>
<td>22,964</td>
<td>1940 – 49</td>
<td>-0,116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95 – 99</td>
<td>-8,352</td>
<td>25,064</td>
<td>1945 – 54</td>
<td>0,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Modell</th>
<th>$D$</th>
<th>$df$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alder</td>
<td>1096</td>
<td>44</td>
</tr>
<tr>
<td>Alder + drift</td>
<td>298</td>
<td>43</td>
</tr>
<tr>
<td>Alder + periode</td>
<td>215</td>
<td>40</td>
</tr>
<tr>
<td>Alder + kohort</td>
<td>85,8</td>
<td>30</td>
</tr>
<tr>
<td>Alder + periode + kohort</td>
<td>30,5</td>
<td>27</td>
</tr>
</tbody>
</table>

Eksempel

Det neste eksempelen 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 $\beta_p \ast$ 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 $\beta_p \ast$ 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 $\gamma_k \ast$ er positive, det vil si at også kohortkurven er konkav. Vi ser videre at kohortkurven først stiger jevnt til et toppunkt før det etter å synke jevnt.

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

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

med parameterverdier fra tabell 20. For kohort får vi

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


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

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

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 $X^TX$ 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 \( \beta = 0 \) og \( \gamma = 0 \) vil vi få de samme parameterverdiene som i tabell 20.


<table>
<thead>
<tr>
<th>Alder</th>
<th>( \alpha_a )</th>
<th>Periode</th>
<th>( \beta_p )</th>
<th>Kohort</th>
<th>( \gamma_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 – 54</td>
<td>- 8,028</td>
<td>1935 – 39</td>
<td>0,000</td>
<td>1850 – 59</td>
<td>0,000</td>
</tr>
<tr>
<td>55 – 59</td>
<td>- 7,308</td>
<td>1940 – 44</td>
<td>0,111</td>
<td>1855 – 64</td>
<td>0,259</td>
</tr>
<tr>
<td>60 – 64</td>
<td>- 6,670</td>
<td>1945 – 49</td>
<td>0,255</td>
<td>1860 – 69</td>
<td>0,300</td>
</tr>
<tr>
<td>65 – 69</td>
<td>- 6,189</td>
<td>1950 – 54</td>
<td>0,401</td>
<td>1865 – 74</td>
<td>0,428</td>
</tr>
<tr>
<td>70 – 74</td>
<td>- 5,734</td>
<td>1955 – 59</td>
<td>0,482</td>
<td>1870 – 79</td>
<td>0,549</td>
</tr>
<tr>
<td>75 – 79</td>
<td>- 5,388</td>
<td>1960 – 64</td>
<td>0,539</td>
<td>1875 – 84</td>
<td>0,587</td>
</tr>
<tr>
<td>80 – 84</td>
<td>- 5,170</td>
<td>1965 – 69</td>
<td>0,640</td>
<td>1880 – 89</td>
<td>0,647</td>
</tr>
<tr>
<td>85 – 89</td>
<td></td>
<td></td>
<td></td>
<td>1885 – 94</td>
<td>0,751</td>
</tr>
<tr>
<td>90 – 94</td>
<td></td>
<td></td>
<td></td>
<td>1890 – 99</td>
<td>0,697</td>
</tr>
<tr>
<td>95 – 99</td>
<td></td>
<td></td>
<td></td>
<td>1895 – 04</td>
<td>0,639</td>
</tr>
<tr>
<td>100 – 09</td>
<td></td>
<td></td>
<td></td>
<td>1900 – 09</td>
<td>0,415</td>
</tr>
<tr>
<td>105 – 14</td>
<td></td>
<td></td>
<td></td>
<td>1905 – 14</td>
<td>0,232</td>
</tr>
<tr>
<td>110 – 19</td>
<td></td>
<td></td>
<td></td>
<td>1910 – 19</td>
<td>0,000</td>
</tr>
</tbody>
</table>

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

\[
\frac{\exp(\beta_1)}{\exp(\beta_2)} \cdot \frac{\exp(\beta_2)}{\exp(\beta_3)} \cdot \frac{\exp(\beta_3)}{\exp(\beta_4)} \cdot \ldots
\]

På logaritmisk skala blir disse kontrastene andreordensdifferanser eller ”andredifferanser” som Clayton & Schifflers kaller det. Dette kan vi skrive som
Andreordensdifferansene er identifiserbare og uavhengige av hvilken parametrisering som er valgt i utgangspunktet. Vi bruker andreordensdifferansene til å definere nye parametere. La $\alpha_a$, $\beta_p$ og $\gamma_k$ være parameterne for en vilkårlig parametrisering av alder-periode-kohort-modellen. Vi definerer da de nye parameterne som

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

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 $\exp(2\alpha_*^e) = 0,519$, 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 $\exp(2\beta_*^e) = 1,075$ 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 eksempelen i S-plus er vist i appendiks C.

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

Eksempel

I det andre eksempelset ser vi nok en gang på forekomsten av prostatakretf 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.


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

Holfords metode

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

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 matematiskt som

\[ \alpha_a = \alpha_{La} + \alpha_{Ca} \]

hvor \( \alpha_{La} \) representerer den lineære komponenten og \( \alpha_{Ca} \) representerer krumningen.

Den lineære komponenten kan skrives som

\[ \alpha_{La} = \varphi_{A0} + \varphi_{A1} x_a \]

hvor \( \varphi_{A0} \) og \( \varphi_{A1} \) 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 \( \varphi_{A0} \). 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 \( \varphi_{P0} \) og \( \varphi_{P1} \) regresjonskoeffisienter, \([p - (P + 1)/2]\) er normalisert periodeindeks og \( \beta_{cp} \) 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 \( \varphi_{K0} \) og \( \varphi_{K1} \) regresjonskoeffisienter, \([k - (K + 1)/2]\) er normalisert kohortindeks og \( \gamma_{ck} \) 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 $\phi_{\text{d1}} + \phi_{\text{p1}}$ estimerbar, og det samme er $\phi_{\text{p1}} + \phi_{\text{k1}}$. 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**


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

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 de 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 $\phi_{\text{p1}} + \phi_{\text{k1}} = 0,008 + 0,107 = 0,115$. På figur 5 er kohort plottet mot $\gamma_k$ og mot $\gamma_{ck}$ (tall fra tabell 23). Vi ser at kurvene har sanne mønster, noe som ikke er overraskende da $\gamma_k$ også er fremkommet ved å fjerne lineær trend (se kapittel 6).

<table>
<thead>
<tr>
<th>Alder</th>
<th>$\alpha_a$</th>
<th>Avvik ($=\alpha_a$)</th>
<th>Periode</th>
<th>$\beta_p$</th>
<th>Avvik ($=\beta_p$)</th>
<th>Kohort</th>
<th>$\gamma_k$</th>
<th>Avvik ($=\gamma_k$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 – 54</td>
<td>- 8,028</td>
<td>-0,242</td>
<td>1935 – 39</td>
<td>0,000</td>
<td>-0,026</td>
<td>1850 – 59</td>
<td>0,000</td>
<td>-0,375</td>
</tr>
<tr>
<td>55 – 59</td>
<td>- 7,308</td>
<td>0,001</td>
<td>1940 – 44</td>
<td>0,111</td>
<td>-0,022</td>
<td>1855 – 64</td>
<td>0,259</td>
<td>-0,124</td>
</tr>
<tr>
<td>60 – 64</td>
<td>- 6,670</td>
<td>0,162</td>
<td>1945 – 49</td>
<td>0,255</td>
<td>0,015</td>
<td>1860 – 69</td>
<td>0,300</td>
<td>-0,091</td>
</tr>
<tr>
<td>65 – 69</td>
<td>- 6,189</td>
<td>0,166</td>
<td>1950 – 54</td>
<td>0,401</td>
<td>0,054</td>
<td>1865 – 74</td>
<td>0,428</td>
<td>0,029</td>
</tr>
<tr>
<td>70 – 74</td>
<td>- 5,734</td>
<td>0,144</td>
<td>1955 – 59</td>
<td>0,482</td>
<td>0,028</td>
<td>1870 – 79</td>
<td>0,549</td>
<td>0,142</td>
</tr>
<tr>
<td>75 – 79</td>
<td>- 5,388</td>
<td>0,012</td>
<td>1960 – 64</td>
<td>0,539</td>
<td>-0,022</td>
<td>1875 – 84</td>
<td>0,587</td>
<td>0,172</td>
</tr>
<tr>
<td>80 – 84</td>
<td>- 5,170</td>
<td>-0,246</td>
<td>1965 – 69</td>
<td>0,640</td>
<td>-0,028</td>
<td>1880 – 89</td>
<td>0,647</td>
<td>0,224</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 alderperiode-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 han 1-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

\[
PT = N_{a,p} \times \Delta t
\]

hvor \(\Delta t\) 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 \( \lambda \) 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 stykkevis 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 scala for rater langs andreaksen, og at det for periode- og kohorteffekter skal være scala 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**


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15 – 19</td>
<td>10 (2321)</td>
<td>7 (2233)</td>
<td>13 (2382)</td>
<td>13 (2919)</td>
<td>15 (3155)</td>
<td>33 (2883)</td>
<td>35 (2858)</td>
<td>37 (3033)</td>
<td>49 (3015)</td>
<td>51 (2789)</td>
<td>41 (2011)</td>
</tr>
<tr>
<td>20 – 24</td>
<td>30 (2439)</td>
<td>31 (2234)</td>
<td>46 (2165)</td>
<td>49 (2313)</td>
<td>55 (2881)</td>
<td>85 (3162)</td>
<td>110 (2902)</td>
<td>140 (2859)</td>
<td>151 (3059)</td>
<td>150 (3052)</td>
<td>112 (2283)</td>
</tr>
<tr>
<td>25 – 29</td>
<td>55 (2372)</td>
<td>62 (2345)</td>
<td>63 (2169)</td>
<td>82 (2096)</td>
<td>87 (2294)</td>
<td>103 (2888)</td>
<td>153 (3168)</td>
<td>201 (2883)</td>
<td>214 (2869)</td>
<td>268 (3095)</td>
<td>194 (2507)</td>
</tr>
<tr>
<td>30 – 34</td>
<td>56 (2398)</td>
<td>66 (2324)</td>
<td>82 (2308)</td>
<td>88 (2135)</td>
<td>103 (2100)</td>
<td>124 (2310)</td>
<td>164 (2881)</td>
<td>207 (3136)</td>
<td>209 (2865)</td>
<td>258 (2871)</td>
<td>251 (2464)</td>
</tr>
<tr>
<td>35 – 39</td>
<td>53 (2308)</td>
<td>56 (2349)</td>
<td>56 (2281)</td>
<td>67 (2281)</td>
<td>99 (2135)</td>
<td>124 (2107)</td>
<td>142 (2302)</td>
<td>152 (2856)</td>
<td>188 (3107)</td>
<td>209 (2846)</td>
<td>199 (2292)</td>
</tr>
<tr>
<td>40 – 44</td>
<td>35 (2082)</td>
<td>47 (2263)</td>
<td>65 (2305)</td>
<td>64 (2250)</td>
<td>67 (2270)</td>
<td>85 (2129)</td>
<td>103 (2090)</td>
<td>119 (2273)</td>
<td>121 (2821)</td>
<td>155 (3071)</td>
<td>126 (2264)</td>
</tr>
<tr>
<td>45 – 49</td>
<td>29 (1866)</td>
<td>30 (2030)</td>
<td>37 (2214)</td>
<td>54 (2260)</td>
<td>45 (2214)</td>
<td>64 (2239)</td>
<td>63 (2095)</td>
<td>66 (2047)</td>
<td>92 (2229)</td>
<td>86 (2770)</td>
<td>96 (2453)</td>
</tr>
<tr>
<td>50 – 54</td>
<td>16 (1618)</td>
<td>28 (1801)</td>
<td>22 (1962)</td>
<td>27 (2146)</td>
<td>46 (2198)</td>
<td>36 (2155)</td>
<td>50 (2173)</td>
<td>49 (2027)</td>
<td>61 (1862)</td>
<td>64 (2163)</td>
<td>51 (2105)</td>
</tr>
<tr>
<td>55 – 59</td>
<td>6 (1413)</td>
<td>14 (1538)</td>
<td>16 (1713)</td>
<td>25 (1868)</td>
<td>26 (2042)</td>
<td>29 (2095)</td>
<td>28 (2051)</td>
<td>43 (2059)</td>
<td>42 (1923)</td>
<td>34 (1883)</td>
<td>45 (1634)</td>
</tr>
<tr>
<td>60 – 64</td>
<td>9 (1210)</td>
<td>12 (1305)</td>
<td>11 (1424)</td>
<td>13 (1584)</td>
<td>20 (1720)</td>
<td>18 (1880)</td>
<td>28 (1930)</td>
<td>23 (1884)</td>
<td>26 (1890)</td>
<td>15 (1772)</td>
<td>10 (1392)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Modell</th>
<th>Devians</th>
<th>Antall frihetsgrader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alder</td>
<td>1371</td>
<td>100</td>
</tr>
<tr>
<td>Alder + drift (kohort)</td>
<td>185</td>
<td>99</td>
</tr>
<tr>
<td>Alder + kohort</td>
<td>120</td>
<td>81</td>
</tr>
<tr>
<td>Alder + periode + kohort</td>
<td>65</td>
<td>72</td>
</tr>
<tr>
<td>Alder + periode</td>
<td>164</td>
<td>90</td>
</tr>
<tr>
<td>Alder + drift (periode)</td>
<td>185</td>
<td>99</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15 – 24</td>
<td>78</td>
<td>121</td>
<td>188</td>
<td>322</td>
<td>401</td>
</tr>
<tr>
<td>(9227)</td>
<td>(9779)</td>
<td>(12081)</td>
<td>(11652)</td>
<td>(11915)</td>
<td></td>
</tr>
<tr>
<td>25 – 34</td>
<td>239</td>
<td>315</td>
<td>417</td>
<td>725</td>
<td>949</td>
</tr>
<tr>
<td>(9436)</td>
<td>(8708)</td>
<td>(9592)</td>
<td>(12068)</td>
<td>(11700)</td>
<td></td>
</tr>
<tr>
<td>35 – 44</td>
<td>191</td>
<td>252</td>
<td>375</td>
<td>516</td>
<td>673</td>
</tr>
<tr>
<td>(9002)</td>
<td>(9117)</td>
<td>(8641)</td>
<td>(9521)</td>
<td>(11845)</td>
<td></td>
</tr>
<tr>
<td>45 – 54</td>
<td>103</td>
<td>140</td>
<td>191</td>
<td>228</td>
<td>303</td>
</tr>
<tr>
<td>(7315)</td>
<td>(8582)</td>
<td>(8806)</td>
<td>(8342)</td>
<td>(9144)</td>
<td></td>
</tr>
<tr>
<td>55 – 64</td>
<td>41</td>
<td>65</td>
<td>93</td>
<td>122</td>
<td>117</td>
</tr>
<tr>
<td>(5466)</td>
<td>(6589)</td>
<td>(7737)</td>
<td>(7924)</td>
<td>(7468)</td>
<td></td>
</tr>
</tbody>
</table>
Tabell 27. Devians og antall frihetsgrader for ulike modeller. I alle modellene er både alder og periode delt inn i 10-års-klasser.

<table>
<thead>
<tr>
<th>Modell</th>
<th>Devians</th>
<th>Antall frihetsgrader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alder</td>
<td>1013</td>
<td>20</td>
</tr>
<tr>
<td>Alder + drift (kohort)</td>
<td>59</td>
<td>19</td>
</tr>
<tr>
<td>Alder + kohort</td>
<td>36</td>
<td>12</td>
</tr>
<tr>
<td>Alder + periode + kohort</td>
<td>8,4</td>
<td>9</td>
</tr>
<tr>
<td>Alder + periode</td>
<td>50</td>
<td>16</td>
</tr>
<tr>
<td>Alder + drift (periode)</td>
<td>59</td>
<td>19</td>
</tr>
</tbody>
</table>

Diskusjon

I denne oppgaven har vi sett at bruk av alderspesifikke 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
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-intervalller både for alder og kalenderperiode, mens det er brukt overlappende 10-års-intervalller 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 intervalller. 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 intervalller, men for eksempel Holford (1983) tar også for seg metoder med ulike lange intervalller. Han har også i en nyere artikkel (Holford 2006) tatt spesielt for seg problemet med ulike lange intervalller. 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.

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.


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 nettdriften hos Holford.


**Konklusjon og videre arbeid**


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


<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>antall</td>
<td>ålder</td>
<td>periode</td>
<td>pop</td>
<td>kohort</td>
<td>rate</td>
</tr>
<tr>
<td>1</td>
<td>3.00</td>
<td>1.00</td>
<td>1.00</td>
<td>10000000.00</td>
<td>11.00</td>
</tr>
<tr>
<td>2</td>
<td>16.00</td>
<td>2.00</td>
<td>1.00</td>
<td>9411765.00</td>
<td>10.00</td>
</tr>
<tr>
<td>3</td>
<td>24.00</td>
<td>3.00</td>
<td>1.00</td>
<td>7500000.00</td>
<td>9.00</td>
</tr>
<tr>
<td>4</td>
<td>79.00</td>
<td>4.00</td>
<td>1.00</td>
<td>7596154.00</td>
<td>8.00</td>
</tr>
<tr>
<td>5</td>
<td>234.00</td>
<td>5.00</td>
<td>1.00</td>
<td>8181818.00</td>
<td>7.00</td>
</tr>
<tr>
<td>6</td>
<td>458.00</td>
<td>6.00</td>
<td>1.00</td>
<td>6897590.00</td>
<td>6.00</td>
</tr>
<tr>
<td>7</td>
<td>720.00</td>
<td>7.00</td>
<td>1.00</td>
<td>5664831.00</td>
<td>5.00</td>
</tr>
<tr>
<td>8</td>
<td>890.00</td>
<td>8.00</td>
<td>1.00</td>
<td>4425659.00</td>
<td>4.00</td>
</tr>
<tr>
<td>9</td>
<td>891.00</td>
<td>9.00</td>
<td>1.00</td>
<td>3651639.00</td>
<td>3.00</td>
</tr>
<tr>
<td>10</td>
<td>920.00</td>
<td>10.00</td>
<td>1.00</td>
<td>2804023.00</td>
<td>2.00</td>
</tr>
<tr>
<td>11</td>
<td>831.00</td>
<td>11.00</td>
<td>1.00</td>
<td>1824769.00</td>
<td>1.00</td>
</tr>
<tr>
<td>12</td>
<td>3.00</td>
<td>1.00</td>
<td>2.00</td>
<td>10000000.00</td>
<td>12.00</td>
</tr>
<tr>
<td>13</td>
<td>17.00</td>
<td>2.00</td>
<td>2.00</td>
<td>9444444.00</td>
<td>11.00</td>
</tr>
<tr>
<td>14</td>
<td>29.00</td>
<td>3.00</td>
<td>2.00</td>
<td>9354839.00</td>
<td>10.00</td>
</tr>
<tr>
<td>15</td>
<td>76.00</td>
<td>4.00</td>
<td>2.00</td>
<td>7230855.00</td>
<td>9.00</td>
</tr>
<tr>
<td>16</td>
<td>185.00</td>
<td>5.00</td>
<td>2.00</td>
<td>7341270.00</td>
<td>8.00</td>
</tr>
<tr>
<td>17</td>
<td>552.00</td>
<td>6.00</td>
<td>2.00</td>
<td>7852063.00</td>
<td>7.00</td>
</tr>
<tr>
<td>18</td>
<td>867.00</td>
<td>7.00</td>
<td>2.00</td>
<td>6474981.00</td>
<td>6.00</td>
</tr>
<tr>
<td>19</td>
<td>1230.00</td>
<td>8.00</td>
<td>2.00</td>
<td>5129274.00</td>
<td>5.00</td>
</tr>
<tr>
<td>20</td>
<td>1266.00</td>
<td>9.00</td>
<td>2.00</td>
<td>3817853.00</td>
<td>4.00</td>
</tr>
<tr>
<td>21</td>
<td>1243.00</td>
<td>10.00</td>
<td>2.00</td>
<td>2937840.00</td>
<td>3.00</td>
</tr>
<tr>
<td>22</td>
<td>937.00</td>
<td>11.00</td>
<td>2.00</td>
<td>1954526.00</td>
<td>2.00</td>
</tr>
<tr>
<td>23</td>
<td>1.00</td>
<td>1.00</td>
<td>3.00</td>
<td>10000000.00</td>
<td>13.00</td>
</tr>
<tr>
<td>24</td>
<td>11.00</td>
<td>2.00</td>
<td>3.00</td>
<td>9166667.00</td>
<td>12.00</td>
</tr>
<tr>
<td>25</td>
<td>33.00</td>
<td>3.00</td>
<td>3.00</td>
<td>9428571.00</td>
<td>11.00</td>
</tr>
<tr>
<td>26</td>
<td>82.00</td>
<td>4.00</td>
<td>3.00</td>
<td>9010989.00</td>
<td>10.00</td>
</tr>
<tr>
<td>27</td>
<td>183.00</td>
<td>5.00</td>
<td>3.00</td>
<td>7011490.00</td>
<td>9.00</td>
</tr>
<tr>
<td>28</td>
<td>450.00</td>
<td>6.00</td>
<td>3.00</td>
<td>6998445.00</td>
<td>8.00</td>
</tr>
<tr>
<td>29</td>
<td>1069.00</td>
<td>7.00</td>
<td>3.00</td>
<td>7326936.00</td>
<td>7.00</td>
</tr>
<tr>
<td>30</td>
<td>1550.00</td>
<td>8.00</td>
<td>3.00</td>
<td>5807419.00</td>
<td>6.00</td>
</tr>
<tr>
<td>31</td>
<td>1829.00</td>
<td>9.00</td>
<td>3.00</td>
<td>4342355.00</td>
<td>5.00</td>
</tr>
<tr>
<td>32</td>
<td>1584.00</td>
<td>10.00</td>
<td>3.00</td>
<td>2966280.00</td>
<td>4.00</td>
</tr>
<tr>
<td>33</td>
<td>1285.00</td>
<td>11.00</td>
<td>3.00</td>
<td>2070911.00</td>
<td>3.00</td>
</tr>
<tr>
<td>34</td>
<td>4.00</td>
<td>1.00</td>
<td>4.00</td>
<td>10000000.00</td>
<td>14.00</td>
</tr>
<tr>
<td>35</td>
<td>8.00</td>
<td>2.00</td>
<td>4.00</td>
<td>10000000.00</td>
<td>13.00</td>
</tr>
<tr>
<td>36</td>
<td>39.00</td>
<td>3.00</td>
<td>4.00</td>
<td>9285714.00</td>
<td>12.00</td>
</tr>
<tr>
<td>37</td>
<td>95.00</td>
<td>4.00</td>
<td>4.00</td>
<td>9134615.00</td>
<td>11.00</td>
</tr>
<tr>
<td>38</td>
<td>267.00</td>
<td>5.00</td>
<td>4.00</td>
<td>8782894.00</td>
<td>10.00</td>
</tr>
<tr>
<td>39</td>
<td>431.00</td>
<td>6.00</td>
<td>4.00</td>
<td>6671827.00</td>
<td>9.00</td>
</tr>
<tr>
<td>40</td>
<td>974.00</td>
<td>7.00</td>
<td>4.00</td>
<td>6653005.00</td>
<td>8.00</td>
</tr>
<tr>
<td>41</td>
<td>1840.00</td>
<td>8.00</td>
<td>4.00</td>
<td>6678766.00</td>
<td>7.00</td>
</tr>
<tr>
<td>42</td>
<td>2395.00</td>
<td>9.00</td>
<td>4.00</td>
<td>5013607.00</td>
<td>6.00</td>
</tr>
<tr>
<td>43</td>
<td>2292.00</td>
<td>10.00</td>
<td>4.00</td>
<td>3472201.00</td>
<td>5.00</td>
</tr>
<tr>
<td>44</td>
<td>1787.00</td>
<td>11.00</td>
<td>4.00</td>
<td>2111045.00</td>
<td>4.00</td>
</tr>
<tr>
<td>45</td>
<td>12.00</td>
<td>1.00</td>
<td>5.00</td>
<td>10000000.00</td>
<td>15.00</td>
</tr>
<tr>
<td>46</td>
<td>8.00</td>
<td>2.00</td>
<td>5.00</td>
<td>8888889.00</td>
<td>14.00</td>
</tr>
<tr>
<td>47</td>
<td>30.00</td>
<td>3.00</td>
<td>5.00</td>
<td>9375000.00</td>
<td>13.00</td>
</tr>
<tr>
<td>48</td>
<td>115.00</td>
<td>4.00</td>
<td>5.00</td>
<td>9055118.00</td>
<td>12.00</td>
</tr>
<tr>
<td>49</td>
<td>285.00</td>
<td>5.00</td>
<td>5.00</td>
<td>9018987.00</td>
<td>11.00</td>
</tr>
<tr>
<td>50</td>
<td>723.00</td>
<td>6.00</td>
<td>5.00</td>
<td>8536009.00</td>
<td>10.00</td>
</tr>
<tr>
<td>51</td>
<td>1004.00</td>
<td>7.00</td>
<td>5.00</td>
<td>6129426.00</td>
<td>9.00</td>
</tr>
<tr>
<td>52</td>
<td>1811.00</td>
<td>8.00</td>
<td>5.00</td>
<td>6347704.00</td>
<td>8.00</td>
</tr>
<tr>
<td>53</td>
<td>3028.00</td>
<td>9.00</td>
<td>5.00</td>
<td>6011515.00</td>
<td>7.00</td>
</tr>
<tr>
<td>54</td>
<td>3176.00</td>
<td>10.00</td>
<td>5.00</td>
<td>4255091.00</td>
<td>6.00</td>
</tr>
<tr>
<td>55</td>
<td>2659.00</td>
<td>11.00</td>
<td>5.00</td>
<td>2551579.00</td>
<td>5.00</td>
</tr>
</tbody>
</table>
Appendiks B


```r
> # 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 -0.683874 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  0.02099212  -392.727257
fald9  -7.608725856  0.02054444  -366.9075581
fald10 -7.143638656  0.02035544  -342.242757
fald11 -6.707365514  0.02028861  -332.132926
fortk1  -0.986968653  0.04304350  -22.9295608
fortk2  -0.907749363  0.03276515  -27.7047182
fortk3  -0.667148138  0.03175308  -20.8529341
fortk4  -0.380849601  0.02833357  -13.4607755
fortk5   0.052809913  0.02214129   2.3651326
fortk6   0.055664404  0.02191338   2.5849060
fortk7   0.059184235  0.02043755   2.9126395
fortk8   0.194384161  0.03895746   4.9896522
fortk9   0.196560040  0.03094162   6.3737301
fortk10  0.326452588  0.09529137   3.4258354
fortk11  0.196560040  0.03094162   6.3737301
fortk12  0.196560040  0.03094162   6.3737301
fortk13  0.196560040  0.03094162   6.3737301
fortk14  0.196560040  0.03094162   6.3737301
fortk15  0.196560040  0.03094162   6.3737301

(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(model.kohort2a)
Appendiks C


```r
> # 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
k11  0.2271237
k12  0.2338228
k13  0.3820233
k14  0.5110127
k15  0.6976336
> #
```

79
# 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
k1 -0.24977767
k2 -0.27208462
k3 -0.32484549
k4 -0.35078146
k5 -0.33987951
k6 -0.23854919
k7 -0.12349548
k8  0.11138342
k9  0.18844311
k10 0.2343903
k11 0.2628578
k12 0.3674931
k13 0.5171570
k14 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
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  -0.39984092
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  -0.39984092
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

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>85.29872</td>
<td>320.8554</td>
<td>577.0248</td>
<td>840.952</td>
<td>1051.898</td>
<td>1047.409</td>
<td>959.7628</td>
<td>804.2654</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>713.2518</td>
<td>621.2812</td>
<td>497.83243</td>
<td>342.9093</td>
<td>676.0791</td>
<td>959.8136</td>
<td>1189.237</td>
<td>1282.944</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>1086.381</td>
<td>890.1751</td>
<td>729.8235</td>
<td>609.6238</td>
<td>481.7188</td>
<td>98.35904</td>
<td>372.7072</td>
<td>783.9653</td>
</tr>
<tr>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
</tr>
<tr>
<td>1210.606</td>
<td>1461.792</td>
<td>1564.175</td>
<td>1429.803</td>
<td>1078.975</td>
<td>867.9149</td>
<td>676.5775</td>
<td>519.1244</td>
</tr>
<tr>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
<td>40</td>
</tr>
<tr>
<td>127.0479</td>
<td>464.5761</td>
<td>912.8659</td>
<td>1497.495</td>
<td>1978.741</td>
<td>2086.753</td>
<td>1874.436</td>
<td>1542.385</td>
</tr>
<tr>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>48</td>
</tr>
<tr>
<td>1155.041</td>
<td>896.324</td>
<td>647.3336</td>
<td>179</td>
<td>587.9521</td>
<td>1103.065</td>
<td>1711.133</td>
<td>2401.331</td>
</tr>
<tr>
<td>49</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2744.718</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>51</th>
<th>52</th>
<th>53</th>
<th>54</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>2445.616</td>
<td>2002.2</td>
<td>1650.969</td>
<td>1182.194</td>
<td>859.8232</td>
</tr>
</tbody>
</table>

> #
> # 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) {
>   j <- j + 1
> }
> c <- 2
> while(c < lk) {
>   c <- c + 1
> }
> #
> # Andreordensdifferanser:
> #
> astart <- aaa[2:(la - 1)]
> pstart <- 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(pstart)
> [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
```r
#
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.