# Statistiske modellar for alder-periode-kohort-effektar i epidemiologi

Gunhild Forland

Masteroppgåve i statistikk - dataanalyse
Matematisk institutt
Universitetet i Bergen


20. mai 2011

# Takk

Eg vil takke rettleiaren min, Ivar Heuch, for god hjelp i prosessen med denne masteroppgåva. Han har vore oppmuntrande og motiverande.

Takk til familien min som har vist interesse for det eg har jobba med og spesielt til mamma som har korrekturlese og gitt meg språklege råd undervegs.

Eg vil takke medstudentane mine for hjelp med LaTeX og andre faglege spørsmål og for sosiale lunsjpausar. Ein spesiell takk til Kristine Kvangarsnes for å ha lese gjennom oppgåva og kome med mange gode råd. Takk til vennene mine for ei fin studietid i Bergen.

# Innhald

# Innleiing

Denne oppgåva tek utgangspunkt i artikkelen Holford [13] har skrive om modellar for alder-periode-kohort-effektar og korleis desse kan estimerast. Artikkelen er kortfatta og kan difor vere vanskeleg å forstå. Eitt av måla med denne oppgåva er å presentere Holford sin teori på ein slik måte at han er enklare å forstå. I kapittel 1 og 3 ser vi på korleis Holford har sett opp alder-periode-kohort-modellen og i kapittel 2 ser vi på dei to matematiske metodane ortogonale polynom og generalisert invers, som Holford nyttar i artikkelen sin.

Kapittel 4 viser korleis Carstensen har implementert alder-periode-kohort-modellar i programpakken `Epi` i R [4]. Programmet `apc.fit` frå denne pakken blir samanlikna med programmet `glm` for å finne ut korleis det fungerer.

I kapittel 5 ser vi på korleis splinefunksjonar kan brukast i samband med alder-periode-kohort-modellar med utgangspunkt i artikkelen av Heuer [12]. I programmet `apc.fit` er det òg ulike modellar for splinefunksjonar.

I kapittel 6 ser vi på tre artiklar som bruker alder-periode-kohort-modellar og programmet `apc.fit` i praksis for å analysere data for kreft.

Til slutt vil vi i kapittel 7 oppsummere og sjå på nokre moglege retningar for vidare arbeid.

# Kapittel 1

# Introduksjon til alder-periode-kohort-modellen

## 1.1 Generell forklaring av modellen

Epidemiologi er definert som studiet av helsetilstand og sjukdomsutbreiing i ei folkemengd, og av årsaker til sjukdom og død [1]. Denne oppgåva ser på modellar som viser kva effektar faktorane alder, periode og kohort har på kor mange diagnosar som vert stilte per personår. Alderen det her er snakk om er pasienten sin alder når diagnosen blir stilt. Periode er definert som tidspunktet diagnosen blir stilt, og kohort er tidspunktet pasienten er fødd. Desse tre variablane; alder, periode og kohort, er lineært avhengige med samanhengen

$$\text{kohort} = \text{periode} - \text{alder}. \tag{1.1}$$

Personår er definert som summen av år kvar person er med i undersøkinga fram til eit eventuelt sjukdomsutbrot eller død [21]. I praksis blir ofte storleiken til populasjonen midt i perioden brukt fordi denne kan bli sett på som å vere proporsjonal med personår [13].

Sjukdomstilfelle og dødsfall blir rapporterte inn til for eksempel Kreftregisteret med dei tre tidsfaktorane alder, periode og kohort. Ofte er ein interessert i å vite kva for ein tidsfaktor som har innverknad på førekomstane av sjukdommen ein ser på. Ein kan sjå på ulike kombinasjonar av dei tre faktorane og finne ut kva kombinasjon som forklarer førekomstane best.

Alderen bør alltid takast med i modellen fordi det er den faktoren som vanlegvis har størst innverknad på utvikling av sjukdom. Periodeeffektar kjem stort sett av forbetringar i diagnostisering og behandling eller endring i klassifisering av sjukdomen på eit visst tidspunkt. Kohorteffekten kan sjåast på som ein generasjonseffekt som speglar ulik livsstil eller ulik miljøpåverknad

frå generasjon til generasjon.

Ein alder-periode-kohort-modell forklarer diagnoseraten i ei folkemengd som produktet av ein alderseffekt, ein periodeeffekt og ein kohorteffekt. Målet med modellen er å sjå på storleiken til ratane, utviklinga over tid og variasjonane for ulike aldrar. Det er vanleg å gruppere tidsfaktorane i 5-års intervall for å få relativt glatte kurver og ikkje altfor mange parametrar.

Holford [13] representerer aldersgruppene ved $i$ som går frå 1 til $I$, og periodegruppene er representerte ved $j$ som går frå 1 til $J$. Alder og periode er uttrykt som intervall og dermed blir kohort òg uttrykt som intervall. Til kohortgruppene bruker vi indeksen $k$ som går frå 1 til $K$. Samanhengen mellom alder, periode og kohort (1.1) kan skrivast som

$$k = j + I - i. \tag{1.2}$$

Dersom alder og periode er inndelt i 5-års intervall, vil kohort få overlappande 10-års intervall.

Ser vi på personane i gruppa med alder 30-34 år som fekk diagnosen i perioden 1995-1999, vil dei høyre til 1960-1969-kohorten. Ser vi heller på aldersgruppa 35-39 år med same periode, vil vi ha kohorten 1955-1964 som overlappar med det førre kohort-intervallet. Ser vi på dei same gruppene med gruppeindeksar, får vi at $i$ går frå 1 til 2. Det er berre éi periodegruppe, så ho får nummer $j = 1$. Kohorten 1955-1964 blir $k = 1$ og 1960-1969 får nummer $k = 2$. Set vi inn $j = 1$ og $I = 2$ i (1.2), får vi $k = 2$ når $i = 1$ og $k = 1$ når $i = 2$. Dette stemmer med dei tilhøyrande gruppene.

Ein generell alder-periode-kohort-modell på multiplikativ form er

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \pi_j + \gamma_k \tag{1.3}$$

der $\lambda_{ijk}$ er diagnoseratane, $\alpha_i$ er alderseffektane, $\pi_j$ er periodeeffektane og $\gamma_k$ er kohorteffektane [13]. I denne modellen nyttar Holford grensekrava $\sum_i \alpha_i = \sum_j \pi_j = \sum_k \gamma_k = 0$. På grunn av den lineære avhengigheita mellom faktorane, (1.2), vil fridomsgradene for testing av nullhypotesen om at $\pi_j = 0$ vere $J - 2$ i staden for $J - 1$ når alle tre faktorane er med i modellen. Tilsvarande endring gjeld ved testing av alders- og kohorteffektane.

Carstensen [3] har i staden for å gi gruppene indeksar $i$, $j$ og $k$, brukt middelverdiane i kvart intervall $a$, $p$ og $c$, for å identifisere gruppene. Modellen (1.3) kan då uttrykkjast som

$$\log(\lambda(a, p)) = f(a) + g(p) + h(c) \tag{1.4}$$

der pasienten har alderen $a$, perioden $p$ og fødselskohorten er $c = p - a$. Dette er det same uttrykket som (1.3) med ein annan notasjon. $f(a)$ tilsvarer $\alpha_i$,

$g(p)$ tilsvarer $\pi_j$ og $h(c)$ tilsvarer $\gamma_k$. Parameteren $\mu$ i (1.3) er uavhengig av alder, periode og kohort, og kan vere inkludert i $f$, $g$ eller $h$ eller delt på fleire av dei slik at $\log(\lambda_{ijk}) = \log(\lambda(a,p))$.

## 1.2 Parametrisering

Holford [13] har i sitt arbeid brukt ein regresjonsmodell som tilsvarer (1.3). Kvar av tidseffektane blir delt i to komponentar: lineær trend og krumming eller avvik frå den rette linja. Den lineære trenden til alderseffekten blir uttrykt ved

$$\alpha_L = C \sum_i c_i \alpha_i \tag{1.5}$$

der $c_i = i - \frac{1}{2}(I+1)$ er gruppenummeret $i$ minus middelverdien av gruppenummera og

$$C = \frac{1}{\sum_i c_i^2} \tag{1.6}$$

er inversen av summen av dei kvadrerte gruppenummeravvika. Det Holford [13] kallar krummingskomponentar, men som eigentleg representerer ikkje-lineær trend, er for alderseffektane gitt ved

$$\tilde{\alpha}_i = \alpha_i - c_i \alpha_L. \tag{1.7}$$

Her er den lineære trenden trekt frå den opprinnelege alderseffekten. Alderseffektane i (1.3) kan då uttrykkjast som summen av den lineære komponenten og krummingskomponenten:

$$\alpha_i = c_i \alpha_L + \tilde{\alpha}_i \tag{1.8}$$

På tilsvarande måte kan vi uttrykkje lineær- og krummingskomponentane for periode- og kohorteffektane. Denne måten å uttrykkje tidseffektane på, kjem av at det er brukt ortogonale polynom. Dette kjem vi tilbake til i avsnitt 2.1.

Designmatrisa $\mathbf{X}$ blir sett opp ved hjelp av ei parametrisering ut frå desse komponentane. Holford [13] innfører ny notasjon for dei uavhengige variablane, som til saman utgjer designmatrisa. Den lineære komponenten for gruppe $i$ er $A_L(i) = c_i$ og krummingskomponenten for gruppe $i$ blir kalla $A_{Cl}(i)$ der $l = 1, ..., I-2$. Det er altså éi kolonne i designmatrisa for den lineære komponenten og $I-2$ kolonner for krummingskomponenten. $A_{Cl}$-ane er ortogonale til $A_L$-ane, det vil seie at $\sum_i A_L(i) A_{Cl}(i) = 0$. Krummingsparametrane kan då uttrykkjast ved $\tilde{\alpha}_i = \sum_l A_{Cl} \alpha_{Cl}$ der $\alpha_{Cl}$ er parameteren som korresponderer til kolonne $A_{Cl}(\cdot)$ i designmatrisa. På tilsvarande måte får vi kolonnene for periode, $\mathbf{P}_L$ og $\mathbf{P}_C$, og kohort, $\mathbf{C}_L$ og $\mathbf{C}_C$, med tilhøyrande parametrar

$\pi_L$, $\pi_C$, $\gamma_L$ og $\gamma_C$. På grunn av (1.2) vil $\mathbf{C}_L = \mathbf{P}_L - \mathbf{A}_L$, og designmatrisa har difor ikkje full kolonnerang. Designmatrisa blir

$$\mathbf{X} = \left[ \begin{array}{cccccc} \mathbf{1} & \mathbf{A}_C & \mathbf{P}_C & \mathbf{C}_C & \mathbf{A}_L & \mathbf{P}_L & \mathbf{C}_L \end{array} \right] \tag{1.9}$$

med korresponderande parametrar $\boldsymbol{\beta}' = \left[ \begin{array}{cccccc} \mu & \boldsymbol{\alpha}_C' & \boldsymbol{\pi}_C' & \boldsymbol{\gamma}_C' & \alpha_L & \pi_L & \gamma_L \end{array} \right]$. Eit eksempel på ei slik designmatrise finn du i tabell 2.1.

Modellen i (1.3) kan no skrivast som

$$\begin{aligned} \log(\lambda_{ijk}) = \mu &+ \sum_l A_{Cl}(i)\alpha_{Cl} + \sum_m P_{Cm}(j)\pi_{Cm} + \sum_n C_{Cn}(k)\gamma_{Cn} \\ &+ A_L(i)\alpha_L + P_L(j)\pi_L + C_L(k)\gamma_L + \epsilon_{ijk}. \end{aligned} \tag{1.10}$$

Leddet $\alpha_i$ er her bytt ut med $A_L(i)\alpha_L + \sum_l A_{Cl}(i)\alpha_{Cl}$. Dette stemmer overeins med (1.8) fordi $A_L(i) = c_i$ og $\sum_l A_{Cl}(i)\alpha_{Cl} = \tilde{\alpha}_i$. Ledda $\pi_j$ og $\gamma_k$ er bytt ut på tilsvarande måte. I tillegg er det lagt til eit feilledd $\epsilon_{ijk}$ slik at (1.10) er på forma til ein generell regresjonsmodell $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Responsvariabelen $\mathbf{y}$ er her $\log(\boldsymbol{\lambda})$.

Clayton og Schifflers [5] skriv i den første artikkelen sin om alder-periode- og alder-kohort-modellar. Denne artikkelen tek opp problemet med drift, som er variasjon som blir forklart like godt av alder-periode-modellen som av alder-kohort-modellen. Dette kjem av den lineære avhengigheita (1.2). Dersom vi har den log-lineære modellen

$$Y_{ap} = \alpha_a + \delta_p(p - p_0) \tag{1.11}$$

der $p_0$ er referansekohorten og $\delta_p$ er endringa i logratane frå éin periode til den neste, kan vi bruke notasjonen $c = A - a + p$ for (1.2) og med det byte ut $p$ i (1.11) og få

$$Y_{ac} = \alpha_a + \delta_p(c + a - A - p_0). \tag{1.12}$$

Dette er ein alder-kohort-modell som med $c_0 = A - a_0 + p_0$ kan omformulerast til

$$Y_{ac} = [\alpha_a + \delta_p(a - a_0)] + \delta_p(c - c_0). \tag{1.13}$$

Den opprinnelege alder-periode-modellen kan altså skrivast som ein alder-kohort-modell med ein annan alderssamanheng for ratane. Tilsvarande kunne vi ha reversert argumentet og gått frå ein alder-kohort-modell til ein alder-periode-modell. Clayton og Schifflers [6] ser i den andre artikkelen sin på alder-periode-kohort-modellar. Når ein skal ha med alle tre tidsfaktorane i éin modell, får ein problem med å identifisere effektane. Ein kan få identiske forklaringar av data frå ulike parameterverdiar på grunn av drift.

Vi kjem tilbake til Holford sin modell i kapittel 3 etter ei forklaring av nokre omgrep han nyttar.

# Kapittel 2

# Matematiske metodar Holford bruker

## 2.1 Ortogonale polynom

Draper og Smith [7] forklarer ortogonale polynom for ein regresjonsmodell med observerte data $(X_i, Y_i)$ $(i = 1, \ldots, n)$ der $Y_i$-ane er avhengige av dei uavhengige $X_i$-ane. Modellen har forma

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \epsilon.$$

Viss ein vil leggje til eller ta bort eit ledd frå modellen, vil estimata til dei andre koeffisientane endre seg. Dette kjem av at ledda er avhengige av kvarandre sidan den same $X$-en inngår alle stader. For å unngå dette problemet, kan ein bruke ortogonale polynom. Først konstruerer ein polynom som har den eigenskapen at når ein multipliserer eitt av polynoma med eitt av dei andre og summerer, blir produktet null. Desse nye ortogonale polynoma blir kalla $\psi_i(X)$ [7]. Dei må altså oppfylle likninga

$$\sum_{i=1}^{n} \psi_j(X_i)\psi_l(X_i) = 0 \qquad (j \neq l). \tag{2.1}$$

Slike polynom kan finnast ved hjelp av Gram-Schmidt-prosessen. Denne prosessen har Lay [16] forklart for vektorar, men det fungerer på same måte for polynom. Gram-Schmidt-prosessen tek bort den lineære avhengigheita mellom polynoma slik at ein får ei mengde av ortogonale polynom. Lay forklarer korleis ein kan lage ein ortogonal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ for eit underrom $W$ av $\mathbb{R}^n$ ved hjelp av ein basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ som ikkje er ortogonal, for det same underrommet $W$ [16, side 420]. Denne ortogonale basisen konstruerer ein på

følgjande måte:

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{x}_1 \\
\mathbf{v}_2 &= \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 \\
\mathbf{v}_3 &= \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{v}_2}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 \\
&\vdots \\
\mathbf{v}_p &= \mathbf{x}_p - \frac{\mathbf{x}_p \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 - \cdots - \frac{\mathbf{x}_p \cdot \mathbf{v}_{p-1}}{\mathbf{v}_{p-1} \cdot \mathbf{v}_{p-1}}\mathbf{v}_{p-1}
\end{aligned} \tag{2.2}$$

Denne framgangsmåten kan bli brukt på rom av polynom ved å byte ut vektorane med polynom. Her tilsvarer vektoren $\mathbf{v}_i$ polynomet $\psi_i(X)$, og $\mathbf{x}_i$ blir bytt ut med $Y$. Skalarprodukta er summen av produkta av korresponderande element i vektorane. Det vil seie at $\mathbf{x} \cdot \mathbf{v}_j$ med polynomnotasjon blir $\sum_i Y_i \psi_j(X_i)$. Tilsvarande får skalarproduktet $\mathbf{v}_j \cdot \mathbf{v}_j$ forma $\sum_i (\psi_j(X_i))^2$.

Når $\psi$-ane er funne på denne måten, blir modellen

$$Y = \alpha_0\psi_0(X) + \alpha_1\psi_1(X) + \cdots + \alpha_p\psi_p(X) + \epsilon$$

der $\psi_0(X) = 1$. Den minste kvadrats-estimatoren for $\alpha_j$ blir

$$\hat{\alpha}_j = \frac{\sum_{i=1}^n Y_i\psi_j(X_i)}{\sum_{i=1}^n (\psi_j(X_i))^2},$$

og designmatrisa $\mathbf{X}$ blir på forma

$$\mathbf{X} = \begin{bmatrix} 1 & \psi_1(X_1) & \psi_2(X_1) & \cdots & \psi_p(X_1) \\ 1 & \psi_1(X_2) & \psi_2(X_2) & \cdots & \psi_p(X_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \psi_1(X_n) & \psi_2(X_n) & \cdots & \psi_p(X_n) \end{bmatrix}.$$

Denne vil ha ortogonale kolonner slik at $\mathbf{X}'\mathbf{X}$ berre får verdiar på diagonalen på grunn av (2.1). Diagonalelementa til $\mathbf{X}'\mathbf{X}$ blir $\sum_{i=1}^n (\psi_j(X_i))^2$ [7]. Med denne modellen eksisterer matrisa $(\mathbf{X}'\mathbf{X})^{-1}$. Ho har inversen til elementa i $\mathbf{X}'\mathbf{X}$ på diagonalen og 0 elles. Eit døme på ei slik designmatrise er

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 3 \\ 1 & -1 & -3 \\ 1 & 1 & 1 \end{bmatrix} \text{ med } \mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 20 \end{bmatrix}.$$

| $i$ | $j$ | $k$ | 1 | $A_C$ | | $P_C$ | | $C_C$ | | | | $A_L$ | $P_L$ | $C_L$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 4 | 1 | 1 | -1 | 1 | -1 | -4 | 0 | 6 | 0 | -3 | -3 | 0 |
| | 2 | 5 | 1 | 1 | -1 | -1 | 3 | -3 | -1 | 1 | 5 | -3 | -1 | 1 |
| | 3 | 6 | 1 | 1 | -1 | -1 | -3 | 0 | -1 | -7 | -4 | -3 | 1 | 2 |
| | 4 | 7 | 1 | 1 | -1 | 1 | 1 | 5 | 1 | 3 | 1 | -3 | -3 | 3 |
| 2 | 1 | 3 | 1 | -1 | 3 | 1 | -1 | -3 | 1 | 1 | -5 | -1 | -3 | -1 |
| | 2 | 4 | 1 | -1 | 3 | -1 | 3 | -4 | 0 | 6 | 0 | -1 | -1 | 0 |
| | 3 | 5 | 1 | -1 | 3 | -1 | -3 | -3 | -1 | 1 | 5 | -1 | 1 | 1 |
| | 4 | 6 | 1 | -1 | 3 | 1 | 1 | 0 | -1 | -7 | -4 | -1 | 3 | 2 |
| 3 | 1 | 2 | 1 | -1 | -3 | 1 | -1 | 0 | 1 | -7 | 4 | 1 | -3 | -2 |
| | 2 | 3 | 1 | -1 | -3 | -1 | 3 | -3 | 1 | 1 | -5 | 1 | -1 | -1 |
| | 3 | 4 | 1 | -1 | -3 | -1 | -3 | -4 | 0 | 6 | 0 | 1 | 1 | 0 |
| | 4 | 5 | 1 | -1 | -3 | 1 | 1 | -3 | -1 | 1 | 5 | 1 | 3 | 1 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | 5 | -1 | 3 | -1 | 3 | -3 | -3 |
| | 2 | 2 | 1 | 1 | 1 | -1 | 3 | 0 | 1 | -7 | 4 | 3 | -1 | -2 |
| | 3 | 3 | 1 | 1 | 1 | -1 | -3 | -3 | 1 | 1 | -5 | 3 | 1 | -1 |
| | 4 | 4 | 1 | 1 | 1 | 1 | 1 | -4 | 0 | 6 | 0 | 3 | 3 | 0 |

Tabell 2.1: Designmatrise for likt inndelte alders- og periodeintervall med $I = J = 4$ ($K = 7$)

Dei to siste kolonnene i denne matrisa er $P_C$ for $i = 1$ i tabell 2.1. Viss den observerte Y-vektoren er $\mathbf{y}' = \begin{bmatrix} 2 & 1 & 3 & -1 \end{bmatrix}$, blir det minste kvadratsestimatet for $\alpha_j$

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{n} y_i X_{ij}}{\sum_{i=1}^{n} X_{ij}^2}.$$

Vi får

$$\hat{\alpha}_1 = \frac{2 \cdot 1 + 1 \cdot 1 + 3 \cdot 1 + (-1) \cdot 1}{1^2 + 1^2 + 1^2 + 1^2} = \frac{5}{4}$$

$$\hat{\alpha}_2 = \frac{2 \cdot 1 + 1 \cdot (-1) + 3 \cdot (-1) + (-1) \cdot 1}{1^2 + (-1)^2 + (-1)^2 + 1^2} = -\frac{3}{4}$$

$$\hat{\alpha}_3 = \frac{2 \cdot (-1) + 1 \cdot 3 + 3 \cdot (-3) + (-1) \cdot 1}{(-1)^2 + 3^2 + (-3)^2 + 1^2} = -\frac{9}{20}.$$

For å konstruere $\psi_j(X_i)$ når ein har $X_i$ med lik avstand, kan ein nytte tabellar. I tabell 2.1 er tabell XXIII frå [9] brukt for å finne designmatrisa. Kolonnene for $A_L$ og $P_L$ er funne ved å nytte $\xi_1$, og $A_C$ og $P_C$ er funne ved å nytte $\xi_2$ og $\xi_3$ for $n = 4$. Kolonnene for $C_L$ og $C_C$ er funne ved å nytte kolonnene gitt for $n = 7$ i tabellen.

## 2.2 Generalisert invers

Holford [13] bruker ein generalisert invers for å finne ut om dei lineære komponentane og krummingskomponentane i avsnitt 1.2 er estimerbare. Her kjem ei forklaring av kva ein generalisert invers er og korleis vi kan finne han. I avsnitt 3.2 kjem vi tilbake til Holford sin bruk av denne inversen.

### 2.2.1 Generell definisjon

Ein generalisert invers til ei matrise $\mathbf{A}$ definerer Searle [20] som ei matrise $\mathbf{G}$ som oppfyller likninga

$$\mathbf{AGA} = \mathbf{A}. \tag{2.3}$$

Ein kan finne generaliserte inversar til alle matriser sjølv om dei er singulære eller rektangulære. $\mathbf{G}$ er ikkje eintydig bestemt med denne definisjonen. Ein måte å finne $\mathbf{G}$ på er å skrive $\mathbf{A}$ på diagonal form:

$$\mathbf{PAQ} = \mathbf{\Delta} = \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Då blir $\mathbf{A} = \mathbf{P}^{-1}\mathbf{\Delta}\mathbf{Q}^{-1}$. Så uttrykkjer vi $\mathbf{G}$ ved hjelp av $\mathbf{\Delta}^-$ som er definert ved

$$\mathbf{\Delta}^- = \begin{bmatrix} \mathbf{D}_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Vi får då

$$\mathbf{G} = \mathbf{Q}\mathbf{\Delta}^-\mathbf{P}$$

som ikkje er eintydig fordi $\mathbf{P}$ og $\mathbf{Q}$ ikkje er eintydige. Ved hjelp av desse uttrykka kan vi vise at (2.3) er oppfylt og at G difor er ein generalisert invers til A:

$$\mathbf{AGA} = \mathbf{P}^{-1}\mathbf{\Delta}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{\Delta}^-\mathbf{P}\mathbf{P}^{-1}\mathbf{\Delta}\mathbf{Q}^{-1} = \mathbf{P}^{-1}\mathbf{\Delta}\mathbf{\Delta}^-\mathbf{\Delta}\mathbf{Q}^{-1} = \mathbf{P}^{-1}\mathbf{\Delta}\mathbf{Q}^{-1} = \mathbf{A}$$

Viss A er ei $p \times q$-matrise med rang $r$ og ei ikkje-singulær $r \times r$-matrise i øvre venstre hjørne, kan vi finne $\mathbf{G}$ ved hjelp av ein annan metode [20].

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

der $\mathbf{A}_{11}$ er ei $r \times r$-matrise med rang $r$. Då er

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

ein generalisert invers til $\mathbf{A}$. $\mathbf{0}$-matrisene har orden slik at $\mathbf{G}$ blir $q \times p$.

$$\begin{aligned}\mathbf{AGA} &= \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix}\end{aligned} \tag{2.4}$$

Sidan $\mathbf{A}$ og $\mathbf{A}_{11}$ har rang $r$, kan vi skrive

$$\begin{bmatrix} \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \end{bmatrix}$$

for ei matrise $\mathbf{K}$. For at (2.4) skal vere lik $\mathbf{A}$, må $\mathbf{K} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$. Då får vi

$$\mathbf{A}_{22} = \mathbf{K}\mathbf{A}_{12} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}.$$

Denne metoden for å finne ein generalisert invers har Holford [13] brukt.

Som døme skal vi bruke matrisa

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix}$$

med rang 2.

$$\mathbf{A}_{11} = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}, \text{ og då blir } \mathbf{A}_{11}^{-1} = \frac{1}{3}\begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}.$$

Ein generalisert invers til $\mathbf{A}$ blir

$$\mathbf{G} = \frac{1}{3}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

med metoden skissert over.

$$\begin{aligned}\mathbf{AGA} &= \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 4/3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2/3 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} = \mathbf{A}\end{aligned}$$

Vi ser her at (2.3) er oppfylt.

### 2.2.2 Moore-Penrose-invers

Viss vi i staden for denne generaliserte inversen bruker Moore-Penrose-inversen, får vi ei eintydig matrise $\mathbf{K}$ [20]. For å vere ein Moore-Penrose-invers, må matrisa oppfylle krava

$$\mathbf{AKA} = \mathbf{A} \tag{i}$$
$$\mathbf{KAK} = \mathbf{K} \tag{ii}$$
$$(\mathbf{KA})' = \mathbf{KA} \tag{iii}$$
$$(\mathbf{AK})' = \mathbf{AK}. \tag{iv}$$

Ved å kombinere (i) og (iii), får vi

$$\mathbf{A} = \mathbf{AKA} = \mathbf{A}(\mathbf{KA})' = \mathbf{AA'K'}.$$

Viss vi her multipliserer med $\mathbf{K}$ på venstre side, bruker (iii) og til slutt dividerer med $\mathbf{K'}$ på høgre side, får vi

$$\mathbf{KAA'} = \mathbf{A'} \tag{2.5}$$

som er ekvivalent med $\mathbf{AA'K'} = \mathbf{A}$. På same måte får vi frå (ii) og (iv)

$$\mathbf{KK'A'} = \mathbf{K}. \tag{2.6}$$

Inversen $\mathbf{K}$ kan finnast ved å anta at $\mathbf{K} = \mathbf{TA'}$ for ei matrise $\mathbf{T}$. Det karakteristiske polynomet til ei $n \times n$-matrise $\mathbf{A}$ er definert som $p(\lambda) = \det(\lambda\mathbf{I}_n - \mathbf{A})$. Cayley-Hamiltons teorem seier at viss ein byter ut $\lambda$ med $\mathbf{A}$ i det karakteristiske polynomet, får ein nullmatrisa, $p(\mathbf{A}) = \mathbf{0}$ [18, side 8]. Dette kan ein bruke for å finne $\mathbf{T}$ som multiplisert med $\mathbf{A'}$ gir Moore-Penrose-inversen til $\mathbf{A}$. Viss $\mathbf{A}$ har dimensjon $p \times q$, er $\mathbf{A'A}$ kvadratisk med dimensjon $q \times q$. Vi får det karakteristiske polynomet

$$p(\lambda) = \det(\lambda\mathbf{I}_q - \mathbf{A'A}),$$

og ved Cayley-Hamiltons teorem er

$$\lambda_1(\mathbf{A'A}) + \lambda_2(\mathbf{A'A})^2 + \cdots + \lambda_t(\mathbf{A'A})^t = \mathbf{0} \tag{2.7}$$

for eit heiltal $t$ og skalarar $\lambda_1, \ldots, \lambda_t$ som ikkje alle kan vere 0. Viss $\lambda_r$ er den første skalaren som er ulik 0, er $\mathbf{T}$ definert ved

$$\mathbf{T} = -\frac{1}{\lambda_r}[\lambda_{r+1}\mathbf{I} + \lambda_{r+2}(\mathbf{A'A}) + \cdots + \lambda_t(\mathbf{A'A})^{t-r-1}]. \tag{2.8}$$

Viss vi multipliserer med $(\mathbf{A}'\mathbf{A})^{r+1}$ på begge sider av (2.8) og bruker (2.7), får vi

$$\begin{aligned}\mathbf{T}(\mathbf{A}'\mathbf{A})^{r+1} &= -\frac{1}{\lambda_r}[\lambda_{r+1}(\mathbf{A}'\mathbf{A})^{r+1} + \lambda_{r+2}(\mathbf{A}'\mathbf{A})^{r+2} + \cdots + \lambda_t(\mathbf{A}'\mathbf{A})^t] \\ &= -\frac{1}{\lambda_r}[-\lambda_1\mathbf{A}'\mathbf{A} - \lambda_2(\mathbf{A}'\mathbf{A})^2 - \cdots - \lambda_r(\mathbf{A}'\mathbf{A})^r].\end{aligned} \tag{2.9}$$

Sidan $\lambda_r$ er den første $\lambda$-en som er ulik 0, blir (2.9)

$$\mathbf{T}(\mathbf{A}'\mathbf{A})^{r+1} = (\mathbf{A}'\mathbf{A})^r$$

som kan kortast ned til $\mathbf{T}\mathbf{A}'\mathbf{A}\mathbf{A}' = \mathbf{A}'$ som er (2.5) med $\mathbf{K} = \mathbf{T}\mathbf{A}'$ [20].

Som døme skal vi bruke den same matrisa $\mathbf{A}$ som i førre avsnitt.

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix}$$

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 26 & 8 & 22 \\ 8 & 3 & 10 \\ 22 & 10 & 38 \end{bmatrix}$$

Det karakteristiske polynomet til $\mathbf{A}'\mathbf{A}$ blir

$$\begin{aligned}\det(\lambda\mathbf{I} - \mathbf{A}'\mathbf{A}) &= \begin{vmatrix} \lambda - 26 & -8 & -22 \\ -8 & \lambda - 3 & -10 \\ -22 & -10 & \lambda - 38 \end{vmatrix} \\ &= (\lambda - 26)(\lambda - 3)(\lambda - 38) - 1760 - 1760 \\ &\quad - 484(\lambda - 3) - 100(\lambda - 26) - 64(\lambda - 38) \\ &= \lambda^3 - 38\lambda^2 - 3\lambda^2 + 114\lambda - 26\lambda^2 + 988\lambda + 78\lambda - 2964 \\ &\quad - 3520 - 484\lambda + 1452 - 100\lambda + 2600 - 64\lambda + 2432 \\ &= 532\lambda - 67\lambda^2 + \lambda^3.\end{aligned}$$

Med Cayley-Hamiltons teorem får vi

$$532(\mathbf{A}'\mathbf{A}) - 67(\mathbf{A}'\mathbf{A})^2 + (\mathbf{A}'\mathbf{A})^3 = \mathbf{0}.$$

Vi kan no finne $\mathbf{T}$ og $\mathbf{K}$ med metoden forklart over.

$$\mathbf{T} = -\frac{1}{532}(-67\mathbf{I} + (\mathbf{A}'\mathbf{A})) = \frac{1}{532}\begin{bmatrix} 41 & -8 & -22 \\ -8 & 64 & -10 \\ -22 & -10 & 29 \end{bmatrix}$$

$$
\begin{aligned}
\mathbf{K} &= \mathbf{T}\mathbf{A}' \\
&= \frac{1}{532}\begin{bmatrix} 41 & -8 & -22 \\ -8 & 64 & -10 \\ -22 & -10 & 29 \end{bmatrix}\begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix} \\
&= \frac{1}{532}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix}
\end{aligned}
$$

For å vere sikker på at dette er ein Moore-Penrose-invers, kan vi sjekke at (2.5) og (2.6) er oppfylte.

$$
\begin{aligned}
\mathbf{K}\mathbf{A}\mathbf{A}' &= \frac{1}{532}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix}\begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix}\begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix} \\
&= \frac{1}{532}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix}\begin{bmatrix} 21 & 15 & 19 \\ 15 & 27 & 19 \\ 19 & 19 & 19 \end{bmatrix} \\
&= \frac{1}{532}\begin{bmatrix} 2128 & 532 & 1596 \\ 532 & 532 & 532 \\ 1064 & 2660 & 1596 \end{bmatrix} \\
&= \begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix} = \mathbf{A}'
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{K}\mathbf{K}'\mathbf{A}' &= \frac{1}{532}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix}\frac{1}{532}\begin{bmatrix} 112 & 12 & -40 \\ -77 & 6 & 113 \\ 49 & 10 & 11 \end{bmatrix}\begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix} \\
&= \frac{1}{532^2}\begin{bmatrix} 20874 & 1372 & -12642 \\ 1372 & 280 & 308 \\ -12642 & 308 & 14490 \end{bmatrix}\begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix} \\
&= \frac{1}{532}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix} = \mathbf{K}
\end{aligned}
$$

$\mathbf{K}\mathbf{A}\mathbf{A}' = \mathbf{A}'$ og $\mathbf{K}\mathbf{K}'\mathbf{A}' = \mathbf{K}$, altså er dette ein Moore-Penrose-invers.

### 2.2.3 Kan parametrane estimerast?

Vi har modellen

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$$

der $\mathbf{y}$ er ein $N\times 1$-vektor av observasjonar, $\mathbf{b}$ er ein $p\times 1$-vektor av parametrar, $\mathbf{X}$ er ei $N \times p$-matrise med kjente verdiar og $\boldsymbol{\epsilon}$ er ein vektor av feilledd som er normalfordelt med forventning 0 og varians $\sigma^2\mathbf{I}$. Med denne modellen er forventninga til $\mathbf{y}$ lik $\mathbf{Xb}$.

Ein funksjon av parametrane er estimerbar viss han er lik ein lineær funksjon av forventninga til $\mathbf{y}$ [20]. Det vil seie at $\mathbf{q}'\mathbf{b}$ er estimerbar viss det eksisterer ein vektor $\mathbf{t}'$ slik at $\mathbf{q}'\mathbf{b} = \mathbf{t}'E(\mathbf{y})$. Sidan $E(\mathbf{y}) = \mathbf{Xb}$, er $\mathbf{q}'\mathbf{b}$ estimerbar når

$$\mathbf{q}' = \mathbf{t}'\mathbf{X}. \tag{2.10}$$

Vi har normallikninga

$$\mathbf{X}'\mathbf{X}\mathbf{b}^0 = \mathbf{X}'\mathbf{y} \tag{2.11}$$

der $\mathbf{X}'\mathbf{X}$ kan vere singulær. Normallikninga kan difor ha uendeleg mange løysingar $\mathbf{b}^0$. Viss $\mathbf{G}$ er ein generalisert invers til $\mathbf{X}'\mathbf{X}$, er $\mathbf{b}^0 = \mathbf{G}\mathbf{X}'\mathbf{y}$ ei løysing til (2.11). Vi definerer $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$. Vi kan då vise at $\mathbf{q}'\mathbf{b}$ er estimerbar viss og berre viss $\mathbf{q}'\mathbf{H} = \mathbf{q}'$. Først antek vi at $\mathbf{q}'\mathbf{b}$ er estimerbar. Då er (2.10) oppfylt, og vi får

$$\mathbf{q}'\mathbf{H} = \mathbf{t}'\mathbf{X}\mathbf{H} = \mathbf{t}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{t}'\mathbf{X} = \mathbf{q}'.$$

Likskapen $\mathbf{q}'\mathbf{H} = \mathbf{q}'$ medfører at $\mathbf{q}'\mathbf{b}$ er estimerbar fordi $\mathbf{q}' = \mathbf{q}'\mathbf{H} = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{t}'\mathbf{X}$ for $\mathbf{t}' = \mathbf{q}'\mathbf{G}\mathbf{X}'$. Vi kan velje $\mathbf{t}'$ slik fordi det er nok at det eksisterer ein $\mathbf{t}'$ som oppfyller (2.10) for at $\mathbf{q}'\mathbf{b}$ skal vere estimerbar. Dette har Holford brukt i avsnitt 2.2 i [13]. Vektoren $\mathbf{y}$ er der log $\boldsymbol{\lambda}$ som er vektoren av logaritmen til ratane.

Vi skal bruke ein modell med tovegs anova utan samspel som døme. Modellen er på forma

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}.$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 & 2 & 2 \\ 2 & 2 & 0 & 1 & 1 \\ 2 & 0 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 \\ 2 & 1 & 1 & 0 & 2 \end{bmatrix}$$

med generalisert invers

$$\mathbf{G} = \begin{bmatrix} 3/4 & -1/2 & 0 & -1/2 & 0 \\ -1/2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

og dermed blir

$$\begin{aligned} \mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X} &= \begin{bmatrix} 3/4 & -1/2 & 0 & -1/2 & 0 \\ -1/2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 & 2 & 2 & 2 & 2 \\ 2 & 2 & 0 & 1 & 1 \\ 2 & 0 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 \\ 2 & 1 & 1 & 0 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Vi har $\mathbf{q}_1' = \begin{bmatrix} 0 & 1 & -1 & 1 & -1 \end{bmatrix}$ og $\mathbf{q}_2' = \begin{bmatrix} 0 & 1 & 1 & -1 & -1 \end{bmatrix}$. Vi vil finne ut om $\mathbf{q}_1'\mathbf{b}$ og $\mathbf{q}_2'\mathbf{b}$ er estimerbare ved hjelp av metoden over.

$$\begin{aligned} \mathbf{q}_1'\mathbf{H} &= \begin{bmatrix} 0 & 1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & -1 & 1 & -1 \end{bmatrix} = \mathbf{q}_1' \end{aligned}$$

$$\begin{aligned} \mathbf{q}_2'\mathbf{H} &= \begin{bmatrix} 0 & 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & -1 & -1 & 1 \end{bmatrix} \neq \mathbf{q}_2' \end{aligned}$$

Sidan $\mathbf{q}_1'\mathbf{H}$ er lik $\mathbf{q}_1'$, er $\mathbf{q}_1'\mathbf{b}$ estimerbar. Vektoren $\mathbf{q}_2'\mathbf{H}$ er ikkje lik $\mathbf{q}_2'$, så $\mathbf{q}_2'\mathbf{b}$ er ikkje estimerbar.

# Kapittel 3

# Meir om modellen til Holford

## 3.1 Ortogonale polynom i modellen

Vi kan no sjå på Holford [13] sin modell i lys av forklaringa av ortogonale polynom i 2.1. Vi skal vise at krummingskomponentane (1.7) i avsnitt 1.2 er på same form som $\mathbf{v}_2$ i (2.2). $\mathbf{x}_2$ er i (2.2) den opprinnelege vektoren. Dette tilsvarer det opprinnelege leddet for alderseffekten, $\alpha_i$, i (1.7). For at krummingskomponenten $\tilde{\alpha}_i$ skal stå ortogonalt på den lineære komponenten $\alpha_L$, må vi trekkje projeksjonen av $\alpha_L$ frå $\alpha_i$ for å få $\tilde{\alpha}_i$. Bruker vi (1.5) og (1.6), kan vi skrive (1.7) som

$$\tilde{\alpha}_i = \alpha_i - c_i C \sum_i c_i \alpha_i = \alpha_i - \frac{\sum_i c_i \alpha_i}{\sum_i c_i^2} c_i. \tag{3.1}$$

Viss vi no ser på $\tilde{\alpha}_i$, $\alpha_i$ og $c_i$ som element i vektorane $\tilde{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha}$ og $\mathbf{c}$ som går frå 1 til $I$, kan vi bruke skalarprodukt og skrive (3.1) på vektorform.

$$\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} - \frac{\mathbf{c} \cdot \boldsymbol{\alpha}}{\mathbf{c} \cdot \mathbf{c}} \mathbf{c} = \boldsymbol{\alpha} - \frac{\sum_i c_i \alpha_i}{\sum_i c_i^2} \mathbf{c} \tag{3.2}$$

Ser vi berre på element $i$ i (3.2), får vi (3.1). I (3.2) ser vi at $\boldsymbol{\alpha}$ tilsvarer $\mathbf{x}_2$ og $\mathbf{c}$ tilsvarer $\mathbf{v}_1$ i (2.2). Krummingskomponenten $\tilde{\alpha}_i$ står ortogonalt på lineærkomponenten $\alpha_L$ ved Gram-Schmidt-prosessen (2.2).

## 3.2 Estimerbare funksjonar av parametrane

Designmatrisa (1.9) har ikkje full kolonnerang fordi alder, periode og kohort er lineært avhengige. Ein må difor bruke ein generalisert invers for å teste

for estimerbarheit. Fordi $\mathbf{C}_L = \mathbf{P}_L - \mathbf{A}_L$, lagar Holford [13] partisjonen $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{C}_L \end{bmatrix}$. Vi får

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{C}_L' \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{C}_L \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{C}_L \\ \mathbf{C}_L'\mathbf{X}_1 & \mathbf{C}_L'\mathbf{C}_L \end{bmatrix}.$$

No er det ei ikkje-singulær matrise i øvre venstre hjørne, $[\mathbf{X}_1'\mathbf{X}_1]$, som har inversen $(\mathbf{X}_1'\mathbf{X}_1)^{-1}$. Den andre metoden forklart i avsnitt 2.2.1 kan då bli brukt for å finne ein generalisert invers til $\mathbf{X}'\mathbf{X}$. Denne generaliserte inversen blir

$$\mathbf{G} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix}.$$

Vi kan rekne ut $\mathbf{H}$ som vi treng for å finne ut om parametrane kan estimerast.

$$\begin{aligned} \mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X} &= \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{C}_L \\ \mathbf{C}_L'\mathbf{X}_1 & \mathbf{C}_L'\mathbf{C}_L \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1 & (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_L \\ \mathbf{0} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_L \\ \mathbf{0} & 0 \end{bmatrix} \end{aligned}$$

Dersom alle kolonnene i $\mathbf{X}_1$ er ortogonale, får $\mathbf{X}_1'\mathbf{X}_1$ berre verdiar på diagonalen og $\mathbf{X}_1'\mathbf{C}_L$ får verdiar på dei to siste plassane i vektoren. Vektoren oppe til høgre i $\mathbf{H}$ blir då

$$\begin{aligned} &(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_L \\ &= \begin{bmatrix} (\mathbf{1}'\mathbf{1})^{-1} & 0 & 0 & \cdots & 0 \\ 0 & (\mathbf{A}_C'\mathbf{A}_C)^{-1} & 0 & \ddots & \vdots \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & \ddots & 0 & (\mathbf{A}_L'\mathbf{A}_L)^{-1} & 0 \\ 0 & \cdots & 0 & 0 & (\mathbf{P}_L'\mathbf{P}_L)^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -\mathbf{A}_L'\mathbf{A}_L \\ \mathbf{P}_L'\mathbf{P}_L \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ 1 \end{bmatrix} \end{aligned}$$

Problemet er at kolonnene for $C_C$ ikkje står ortogonalt på resten av kolonnene i $\mathbf{X}_1$ fordi ikkje alle kohortgruppene opptrer like mange gonger. Ser vi tilbake på tabell 2.1, ser vi at alders- og periodegruppene opptrer fire gonger kvar, medan det varierer med plasseringa av kohortgruppene kor mange gonger dei opptrer. Dei kohortgruppene som er nærmast midten, er brukte

oftare ein dei ytste. Dette kjem av den lineære samanhengen mellom gruppene. Matrisa $\mathbf{H}$ blir difor ikkje på den fine forma som Holford seier [13, side 314].

Vi skal prøve å rekne ut $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_L$ for eksempelet i tabell 1 hos Holford [13].

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & -2 & 0 & 6 & -1 & -1 & 0 \\ 1 & 1 & -2 & -1 & -2 & -4 & -1 & 0 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 & -1 & 1 & 2 \\ 1 & -2 & 1 & -1 & 2 & -4 & 0 & -1 & -1 \\ 1 & -2 & -2 & -2 & 0 & 6 & 0 & 0 & 0 \\ 1 & -2 & 1 & -1 & -2 & -4 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2 & -1 & 1 & 1 & -1 & -2 \\ 1 & 1 & -2 & -1 & 2 & -4 & 1 & 0 & -1 \\ 1 & 1 & 1 & -2 & 0 & 6 & 1 & 1 & 0 \end{bmatrix}$$

Den siste kolonna er $\mathbf{C}_L$ og resten av matrisa er $\mathbf{X}_1$. Vi får

$$\mathbf{X}_1'\mathbf{X}_1 = \begin{bmatrix} 9 & 0 & 0 & -6 & 0 & 4 & 0 & 0 \\ 0 & 18 & 0 & 6 & 0 & 10 & 0 & 0 \\ 0 & 0 & 18 & 6 & 0 & 10 & 0 & 0 \\ -6 & 6 & 6 & 24 & 0 & -16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 18 & 0 & 2 & -2 \\ 4 & 10 & 10 & -16 & 0 & 174 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & 0 & 6 \end{bmatrix} \text{ og } \mathbf{X}_1'\mathbf{C}_L = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -4 \\ 0 \\ -6 \\ 6 \end{bmatrix}.$$

Matrisa $\mathbf{X}_1'\mathbf{X}_1$ har verdiar utanom diagonalen, og vektoren $\mathbf{X}_1'\mathbf{C}_L$ har ikkje berre verdiar i dei to siste posisjonane. Vektoren $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_L$ får ikkje den fine forma vi forventa, og matrisa $\mathbf{H}$ får difor ikkje forma

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & \ddots & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}. \tag{3.3}$$

For å teste om den lineære alderskomponenten er estimerbar, bruker Holford [13] matrisa (3.3). Sidan (3.3) ikkje stemmer for alder-periode-kohort-modellen, kan denne $\mathbf{H}$-matrisa ikkje brukast for å finne ut om parametrane kan estimerast. Vi vil likevel bruke (3.3) for å vise korleis Holford har kome

fram til kva funksjonar av parametrane som er estimerbare. Han nyttar framgangsmåten forklart i avsnitt 2.2.3 for å vise at $\alpha_L$ ikkje kan estimerast. Med $\mathbf{q}' = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & 0 \end{bmatrix}$ blir

$$\mathbf{q}'\mathbf{H} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ -1 \end{bmatrix} \neq \mathbf{q},$$

og $\mathbf{q}'\boldsymbol{\beta} = \alpha_L$ er dermed ikkje estimerbar. På same måte kan det visast at lineærkomponentane til periode og kohort ikkje er estimerbare. Generelt er funksjonar på forma $d_1\alpha_L + d_2\pi_L + (d_2 - d_1)\gamma_L$ med vilkårlege $d_1$ og $d_2$ estimerbare. Då er $\mathbf{q}' = \begin{bmatrix} 0 & \cdots & 0 & d_1 & d_2 & d_2 - d_1 \end{bmatrix}$ og $\mathbf{q}'\mathbf{H} = \mathbf{q}$. For krummingskomponentane er alle funksjonar på forma $\begin{bmatrix} \mathbf{q}_C^* & 0 & 0 & 0 \end{bmatrix}\boldsymbol{\beta}$ med vilkårleg $\mathbf{q}_C^*$ estimerbare fordi $\mathbf{q}'$ då berre blir multiplisert med den delen av $\mathbf{H}$ som er lik $\mathbf{I}$.

## 3.3 Modell med ulik storleik på tidsintervalla

Dersom ein i modellen vil fokusere meir på alderen enn perioden, kan ein ha finare inndeling av aldersintervalla enn av periodeintervalla. Holford [13, avsnitt 3] har sett på tilfellet der periodeintervalla er $R$ gonger breiare enn aldersintervalla. Når periodeindeksen aukar med éi eining, tilsvarer det ein auke av aldersindeksen med $R$ einingar. Kvart aldersintervall kan då bli referert til med dobbel indeks $(i, r)$, der $i = 1, \ldots, I$ og $r = 1, \ldots, R$. Med dette systemet aukar vi alderen ved å auke først $r$ og så $i$ slik at vi kjem til høgare aldersintervall. Det skal vere ei balansert fordeling av gruppene slik at alle $i$-grupper er delte inn i like mange $r$-grupper. $R$ er altså lik for alle $i$. Då blir det totalt $IR$ aldersintervall. Kohortintervalla blir òg grupperte slik at dei kan representerast ved $(k, r)$ der $k = 1, \ldots, K$ og $r = 1, \ldots, R$. For aukande kohortintervall minkar vi $r$ og aukar $k$.

Parametriseringa for denne modellen er ulik den vi hadde i avsnitt 1.2. Her må vi ta omsyn til grupperinga av kvart alders- og kohortintervall. Alders- og periodefaktorane blir delte inn i fire komponentar for å få med effektane av grupperinga, lineariteten, parallelliteten og krumminga. Holford ser på variablane for alderskomponentane. Kohorteffekten er delt opp på tilsvarande måte, medan periodeeffekten har dei same komponentane som i avsnitt 1.2 fordi han ikkje er delt opp i nye grupper.

Komponenten for grupperinga av alderseffekten definerer Holford ved

$$A_{Gh}(i,r) = \begin{cases} 1, & h = r \\ -1, & h = R \\ 0, & \text{elles} \end{cases} \tag{3.4}$$

for $h = 1, \ldots, R-1$. Med denne definisjonen vil $A_{Gh}$ aldri bli $-1$ fordi $h$ har $R-1$ som maksimalverdi. Holford har likevel fått både 1 og $-1$ som verdiar i gruppekolonna i tabell 6 i [13]. Det ville vere naturleg å uttrykkje gruppekomponenten som

$$A_{Gh}(i,r) = \begin{cases} 1, & r = h \\ -1, & r = R \\ 0, & \text{elles} \end{cases} \tag{3.5}$$

for $h = 1, \ldots, R-1$ i staden. Då vil $A_G$ få verdien $-1$ når $r$ er maksimal, altså i den siste gruppa i kvart aldersintervall.

Den lineære trenden kan som før uttrykkjast ved $A_L^*(i,r) = i - \frac{1}{2}I - \frac{1}{2}$, men fordi kvar $i$ er delt inn i $R$ undergrupper er det betre å bruke

$$A_L(i,r) = R\left(i - \tfrac{1}{2}I - \tfrac{1}{2}\right) + \left(r - \tfrac{1}{2}R - \tfrac{1}{2}\right). \tag{3.6}$$

Parallellkomponenten fortel om det er skilnad på undergruppene innanfor kvart intervall. Altså om utviklinga frå til dømes $r = 1$ til $r = 2$ er lik for alle intervall $i$. Holford uttrykkjer denne komponenten som

$$A_{Ph}(i,r) = A_{Gh}(i,r)A_L^*(i,r) = \begin{cases} A_L^*(i,r), & h = r \\ -A_L^*(i,r), & h = R \\ 0, & \text{elles} \end{cases} \tag{3.7}$$

for $h = 1, \ldots, R-1$, men vi byter her òg ut $h = r$ med $r = h$ og $h = R$ med $r = R$. Desse ledda er ortogonale til gruppe- og linearitetsledda fordi $\sum_{1,r} A_L(i,r)A_{Ph}(i,r) = \sum_{i,r} A_{Gh}(i,r)A_{Ph}(i,r) = 0$.

Krummingskomponenten i dette tilfellet med gruppering av aldersintervalla, har Holford definert ved

$$A_{Chl}(i,r) = \begin{cases} A_{Cl}(i), & h = r \\ 0, & \text{elles} \end{cases} \tag{3.8}$$

for $h = 1, \ldots, R$, $i = 1, \ldots, I-2$, og $A_{Cl}(i)$ er definert som ortogonal til den lineære komponenten som i avsnitt 1.2. I tabell 3.1 er aldersvariablane frå designmatrisa sett opp for $I = 4$ og $R = 2$.

| $(i,r)$ | $A_G$ | $A_L$ | $A_P$ | $A_C$ | | | |
|---|---|---|---|---|---|---|---|
| (1,1) | 1 | -3,5 | -3 | 1 | 0 | -1 | 0 |
| (1,2) | -1 | -2,5 | 3 | 0 | 1 | 0 | -1 |
| (2,1) | 1 | -1,5 | -1 | -1 | 0 | 3 | 0 |
| (2,2) | -1 | -0,5 | 1 | 0 | -1 | 0 | 3 |
| (3,1) | 1 | 0,5 | 1 | -1 | 0 | -3 | 0 |
| (3,2) | -1 | 1,5 | -1 | 0 | -1 | 0 | -3 |
| (4,1) | 1 | 2,5 | 3 | 1 | 0 | 1 | 0 |
| (4,2) | -1 | 3,5 | -3 | 0 | 1 | 0 | 1 |

Tabell 3.1: Komponentar for gruppering ($A_G$), linearitet ($A_L$), parallellitet ($A_P$) og krumming ($A_C$) for variabelen alder med $i = 1, \ldots, 4$ og $r = 1, 2$.

# Kapittel 4

# R-programmet apc.fit

## 4.1 Generelt

Carstensen [3] har laga R-programmet `apc.fit` [4] som finst i pakken `Epi`. Programmet kan tilpasse alder-periode-kohort-modellar for tabellerte data. Det må minimum ha inndata for dei numeriske vektorane for aldersintervall (`A`), periodeintervall (`P`), tal diagnosetilfelle eller dødsfall (`D`) og personår (`Y`). Kohort `C` blir rekna ut ved $\texttt{C} = \texttt{P} - \texttt{A}$. Intervalla for alder, periode og kohort treng ikkje vere like lange og er representerte ved gjennomsnittet i kvart intervall. I tillegg til desse fire vektorane kan ein gi inn ein del andre argument til programmet som avgjer korleis modellen skal vere.

Argumenta `ref.p` og `ref.c` blir brukte for å oppgi referanseperiode og -kohort. Med argumentet `dist` kan ein velje fordelinga som skal bli brukt i modellen. Alternativa her er `"poisson"` og `"binomial"`. Argumentet `model` avgjer kva type modell som blir brukt. Ein kan velje mellom faktormodell (`"factor"`) med éin parameter for kvar verdi av `A`, `P` og `C` og tre ulike splinemodellar: naturleg spline (`"ns"`), B-spline (`"bs"`) og lineær spline (`"ls"`). For splinemodellane kan ein med argumentet `npar` bestemme kor mange parametrar som skal bli brukte for kvart ledd i modellen eller kva knutepunkt som skal bli brukte. Splinefunksjonar og korleis dei kan bli brukte i `apc.fit` blir forklart i kapittel 5.

For å avgjere parametriseringa til effektane blir argumentet `parm` brukt. Her kan ein velje mellom mange ulike variantar med to eller tre av tidseffektane og med og utan driftledd (sjå avsnitt 1.2). For å bestemme korleis driftparamteren skal bli funnen i modellen, bruker vi argumentet `dr.extr`. Ein kan velje mellom `"weighted"` som lèt det vekta gjennomsnittet for tal tilfelle, `D`, av dei estimerte periode- og kohorteffektane ha stigningstal null, eller `"Holford"` som bruker gjennomsnittet over alle verdiane for dei estimerte

effektane uavhengig av kor mange tilfelle det er. For dei parametriseringane som ikkje har driftledd, blir dette argumentet ignorert. I tillegg til desse argumenta kan ein med `alpha` velje signifikansnivået, med `scale` skalere ratane og med `print.AOV` få utskrift av deviansanalysetabellar for modellane.

Som utdata gir programmet `apc.fit` eit objekt av klassen apc som programma `apc.frame`, `apc.lines` og `apc.plot` kan bruke til å lage eit plott. Dette objektet inneheld matriser for alder-, periode-, kohort- og driftestimata (`Age`, `Per`, `Coh` og `Drift`), ein vektor med referanseperiode og -kohort (`Ref`), ein deviansanalysetabell som samanliknar dei fem klassiske modellane (`AOV`), forklaring av modellen og parametriseringa (`Type`) og ein vektor for knutepunkta dersom `model` er `"ns"` eller `"bs"` (`Knots`).

## 4.2 Tofaktormodell

For å finne ut korleis dette programmet fungerer, bruker vi eit enkelt datasett med fire alders- og periodeintervall, sjå vedlegg A.1. Dette datasettet inneheld vektorane for dei fire variablane programmet `apc.fit` treng. Verdiane i D-vektoren aukar for aukande alder, og eg har prøvd å ikkje la det vere nokon periodeeffekt. Datasettet er brukt som inndata i programmet `apc.fit` med `parm="AP-C"`. Med denne parametriseringa er alderseffektane ratar for referanseperioden i alder-periode-modellen. Periodeeffektane er forholdsratar relativt til referanseperioden. Kohorteffektane blir med denne parametriseringa rekna ut etterpå slik at estimata for alder- og periodeeffektane er uavhengige av kohorten. Ratane det her er snakk om er forholdet mellom tilfelle (`D`) og personår (`Y`) [4].

| Alder | Alderseffekt | Periode | Periodeeffekt |
|---|---|---|---|
| 32 | $8{,}839774 \cdot 10^{-5}$ | 1977 | 1,000000 |
| 37 | $1{,}148586 \cdot 10^{-4}$ | 1982 | 1,039809 |
| 42 | $1{,}314593 \cdot 10^{-4}$ | 1987 | 1,113994 |
| 47 | $1{,}575187 \cdot 10^{-4}$ | 1992 | 1,131792 |

Tabell 4.1: Estimatorar for alders- og periodeeffektar

Her ser vi berre på alders- og periodeeffektane. Med denne alder-periode-modellen får ein 14 fridomsgrader og residualdevians 0,80950. Med alder som einaste faktor, får ein 17 fridomsgrader og residualdevians 1,20036. 95 %-konfidensintervalla for periodeeffektane inneheld 1 for alle periodeverdiane, så eg har lukkast i å lage eit datasett utan signifikant periodeeffekt. Vi ser òg av tabell 4.1 at estimatorane for alderseffektane aukar med verdien for alderen.

Det same datasettet er brukt med programmet `glm` for å samanlikne verdiane som kjem ut. Dette programmet blir brukt til å tilpasse generaliserte lineære modellar. Med ulike argument kan ein spesifisere korleis modellen skal vere. To ulike faktormodellar med `D` som responsvariabel og `log(Y)` som offset-ledd er køyrt, først med berre alder som faktor og etterpå med både alder og periode. Som fordeling for feilen er Poisson brukt slik som i `apc.fit` med logaritmisk linkfunksjon. Med programmet `glm` får vi òg ut residualdevians og tal fridomsgrader. Desse verdiane er like dei vi fekk med `apc.fit`.

## 4.3 Trefaktormodell

Vi skal no sjå på ein full alder-periode-kohort-modell med det same datasettet som i førre avsnitt (A.1). Argumentet `parm="ACP"` er brukt i programmet `apc.fit`. Denne parametriseringa gir ut maksimum likelihood-estimat. Alderseffektane er ratar for referansekohorten, og kohorteffektane er forholdsratar relativt til referansekohorten. Periodeeffektane er sett til å ha gjennomsnitt og stigningstal lik null. Vi går altså ut frå at perioden ikkje har innverknad på ratane, og drifta blir inkludert i kohorteffekten. Både `dr.extr="weighted"` og `dr.extr="Holford"` er brukt. Sjå A.2 for utskrift av programmet. Med fem desimalar gir desse to variantane av modellen like mange fridomsgrader og lik residualdevians. Referanseperioden er 1977 og 1940 er referansekohort. Det er små skilnader mellom estimatorane for alder, periode og kohort for dei to variantane av modellen.

Etterpå vart programmet `glm` brukt på det same datasettet for å samanlikne resultata. Utskrifta frå dette programmet finn du i A.3. Med `D` som responsvariabel, `log(Y)` som `offset` og `family=poisson(link="log")` får ein ut dei same fridomsgradene og same residualdevians som med `apc.fit`. `A`, `P` og `C` er brukte som faktorar.

## 4.4 Eksempel med lungekreft-data

Når ein lastar ned pakken `Epi` som programmet `apc.fit` er ein del av, får ein med datasettet lungDK som inneheld data for lungekrefttilfelle blant menn i Danmark. Dette datasettet har vi òg brukt for å finne ut korleis programmet `apc.fit` fungerer. Vi skal igjen samanlikne med det vi får ut ved å bruke programmet `glm`.

Parametriseringa `parm="AP-C"` er den same som i avsnitt 4.2, men no skal vi sjå på den fulle alder-periode-kohort-modellen òg. Ifølgje programfor-

| | F.grader | Res. dev. |
|---|---|---|
| Alder | 200 | 15420,7 |
| Alder-drift | 199 | 6806,1 |
| Alder-kohort | 162 | 910,6 |
| Alder-periode-kohort | 144 | 284,7 |
| Alder-periode | 180 | 2999,0 |

Tabell 4.2: Utdata frå apc.fit med datasettet lungDK

klaringa er kohorteffektane her frå modellen med kohort åleine. Logaritmen av dei tilpassa verdiane frå alder-periode-modellen er brukt som offset. For å få ein tilsvarande modell med `glm`, er det først køyrt ein modell med alder og periode som faktorar. Logaritmen av dei tilpassa verdiane i denne modellen er brukt som offset i ein modell med kohort som faktor. Med denne glm-modellen blir residualdeviansen lik 10548 med 180 fridomsgrader. Dette er like mange fridomsgrader som ein får med alder-periode-modellen (sjå tabell 4.2) med `apc.fit`, men residualdeviansen liknar ikkje på noko vi har fått med `apc.fit`. Dersom vi ser på modellen med berre alder og periode som faktorar med programmet `glm`, får vi ut same tal fridomsgrader og devians (180 og 2999,0) som for alder-periode-modellen med programmet `apc.fit` med `parm="AP-C"`. Når vi prøver å køyre ein glm-modell med alder, periode og kohort som faktorar, blir residualdeviansen lik 284,73 med 144 fridomsgrader. Dette er likt det ein får med `parm="AP-C"` i `apc.fit`. Det kan sjå ut som programmet ikkje fungerer på den måten det er forklart i R-dokumentasjonen [4].

Dersom ein køyrer programmet med `parm="ACP"`, får ein ut driftestimata gitt i tabell 4.3. APC er maksimum likelihood-estimatet for drifta, og A-d er estimatet frå alder-drift-modellen. Estimatet frå alder-drift-modellen blir det same uansett kva parametrisering ein bruker og korleis driftparameteren blir funnen. Maksimum likelihood-estimatet for drifta er større med Holford sin måte å finne driftparameteren enn når det vekta gjennomsnittet har stigningstal null. Fridomsgradene og residualdeviansen med parametriseringa `ACP` er lik uavhengig av `dr.extr` og er gitt i tabell 4.2.

| | weighted | Holford |
|---|---|---|
| APC | 1,019787 | 1,033864 |
| A-d | 1,023496 | 1,023496 |

Tabell 4.3: Driftestimat for lungDK med parm=“ACP”

# Kapittel 5

# Spline

## 5.1 Generelle definisjonar

Heuer [12] definerer ein splinefunksjon i regresjon som ein funksjon på eit intervall $(a, b)$ som er samansett av polynom. Desse polynoma deler $(a, b)$ i $m + 1$ bitar definert av $m$ knutepunkt $\xi_1 < \xi_2 < \cdots < \xi_m$. Vi seier at denne splinefunksjonen har orden $q$ dersom den høgaste graden til polynoma er $q - 1$. Polynoma og dei deriverte må vere kontinuerlege i knutepunkta slik at overgangane blir glatte. Knutepunkta treng ikkje vere jamt fordelte. Det finst fleire typar splinefunksjonar. Vi skal her sjå på nokre av dei.

Gerald og Wheatley [11] definerer ein lineær spline som ein splinefunksjon der alle polynoma har grad 1. Kurva er difor samansett av rette linjer med diskontinuitetar i knutepunkta. Splinefunksjonar av høgare orden enn 2 har ikkje dette diskontinuitetsproblemet.

Det vanlegaste er å bruke kubiske splinefunksjonar, det vil seie splinefunksjonar av orden 4. Desse må ha polynom som er kontinuerlege i knutepunkta. Stigningstalet og krumminga må òg vere like der polynoma møtest. Det vil seie at den første- og andrederiverte må vere kontinuerlege. For polynoma i endane er det ikkje så strenge krav fordi dei berre har eitt anna polynom dei er knytte saman med.

Friedman, Hastie og Tibshirani [10, side 143] har sett opp ein basis for ein kubisk spline med to knutepunkt, $\xi_1$ og $\xi_2$:

$$\begin{gathered} h_1(X) = 1,\ h_2(X) = X,\ h_3(X) = X^2,\ h_4(x) = X^3, \\ h_5(X) = (X - \xi_1)_+^3,\ h_6(X) = (X - \xi_2)_+^3. \end{gathered} \tag{5.1}$$

Den trunkerte potens-basis-funksjonen $(u)_+^q$ er definert ved

$$(u)_+^q = \begin{cases} u^q & \text{viss } u > 0 \\ 0 & \text{elles.} \end{cases} \tag{5.2}$$

Det er i (5.1) seks basisfunksjonar. Det vil seie at vi har eit seksdimensjonalt lineært funksjonsrom. For å sjekke at dette stemmer, kan vi telje opp parametrane. Med to knutepunkt blir splinefunksjonen delt i tre område. For kvart område har polynomet fire parametrar. I kvart av dei to knutepunkta er det tre avgrensingar: Polynoma og dei første- og andrederiverte må vere kontinuerlege. Vi får då (3 område) $\times$ (4 parametrar) $-$ (2 knutepunkt) $\times$ (3 avgrensingar) $= 6$.

Naturlege splinar er splinefunksjonar med lineære halar. Dei andrederiverte til polynoma i kvar ende er då null. Dette gjer til at ein unngår store svingingar i halane som ein fort kan få med polynom av høg grad.

## 5.2 B-spline

B-splinen har den eigenskapen at dersom eit datapunkt blir endra, blir berre ein del av splinekurva endra. Vi får ein lokal effekt i motsetnad til vanleg kubisk spline der ei endring av eitt datapunkt fører til at heile kurva blir endra. Ein B-splinefunksjon av orden $q$ for ein variabel $x$ er gitt ved

$$f(x) = \sum_{i=-(q-1)}^{m} B_{i,q}(x)\vartheta_i \tag{5.3}$$

der $B_{i,q}(x)$ er basisvektorar som er definerte rekursivt og $\vartheta_i$ er korresponderande koeffisientar. Indeksen $q$ er ordenen til B-splinefunksjonen [12].

Heuer forklarer i artikkelen sin to typar splinefunksjonar: trunkert potensbasis-spline (TP-spline) og B-spline. Desse bruker han seinare i Holford sin modell.

TP-splinefunksjonen for ein variabel $x$ er definert ved

$$f(x) = \sum_{j=0}^{q-1} \beta_{0j} \cdot x^{j-1} + \sum_{i=1}^{m} \beta_i \cdot (x - \xi_i)_+^{q-1} \tag{5.4}$$

der $\beta_{0j}$ $(j = 0, \ldots, q-1)$ og $\beta_{1i}$ $(i = 1, \ldots, m)$ er ukjente koeffisientar til TP-splinefunksjonen. Ein slik splinefamilie er definert ved ordenen $q$, talet på knutepunkt $m$ og posisjonen til knutepunkta $(\xi_1, \xi_2, \ldots, \xi_m)$. Når parametriseringa med variabelen $x$ blir innført, får ein $m + q$ nye variablar, $x^j$ for $j = 0, \ldots, q-1$ og $(x - \xi_i)_+^{q-1}$ for $i = 1, \ldots, m$. Dette er basisfunksjonane til TP-splinen. Ein ulempe med denne typen splinefunksjonar er at basisfunksjonane kan vere sterkt korrelerte viss det er mange knutepunkt eller knutepunkta ligg tett. For å unngå dette kan ein heller bruke B-splinefunksjonar. Basisvektorane til ein TP-spline og ein B-spline spenner

det same rommet og gir same kurveestimering om ein ser bort frå numeriske avvik. Basisfunksjonane til B-splinar er ikkje null for eit intervall med $q-1$ knutepunkt. Vi seier at B-splinefunksjonar er velkondisjonerte. Vi får då ei samanbunden regresjonsmatrise og stabile estimatorar. Heuer [12] held fram med å sjå på B-splinefunksjonar fordi dei har betre eigenskapar enn TP-splinefunksjonar, men han poengterer at utrekningane kunne ha vore gjort for TP-splinefunksjonar utan å forvente ustabilitet på grunn av korrelerte basisfunksjonar i denne samanhengen.

Heuer bruker Eubank [8] sin rekursive definisjon av B-splinefunksjonen. Lat $(a,b) \subset \mathbb{R}$ og lat $\xi_1, \ldots, \xi_m$ med $a < \xi_1 < \xi_2 < \cdots < \xi_m < b$ vere $m$ kjente, fastlagde indre knutepunkt. Definer $2q$ ekstra endepunkt $\xi_{-(q-1)}, \ldots, \xi_{-1}, \xi_0$ og $\xi_{m+1}, \ldots, \xi_{m+q}$ der den første sekvensen er lik $a$ og den andre er lik $b$. Lat $B_{-(q-1),q}, \ldots, B_{m,q}$ vere basisfunksjonar som er rekursivt definerte ved

$$B_{i,q}(x) = \frac{x-\xi_i}{\xi_{i+q-1}-\xi_i} B_{i,q-1}(x) + \frac{\xi_{i+q}-x}{\xi_{i+q}-\xi_{i+1}} B_{i+1,q-1}(x) \tag{5.5}$$

med $i = -(q-1), \ldots, m$ og startverdiar

$$B_{j,1}(x) = \begin{cases} 1, & x \in [\xi_j, \xi_{j+1}) \\ 0, & \text{elles} \end{cases}$$

for $j = -(q-1), \ldots, m+q-2$ og

$$B_{m+q-1,1}(x) = \begin{cases} 1, & x \in [\xi_{m+q-1}, \xi_{m+q}] \\ 0, & \text{elles.} \end{cases}$$

Då kan vi kalle ein funksjon $f : (a,b) \to \mathbb{R}$ som oppfyller (5.3) for ein basis-spline (B-spline) av orden $q$ med $m$ indre knutepunkt som er parvis ulike. Variablane $\vartheta_i \in \mathbb{R}, i = -(q-1), \ldots, m$ er dei ukjente koeffisientane til B-splinen. B-spline-basisen har som TP-spline-basisen dimensjonen $m+q$. Dei $2q$ knutepunkta som blir lagde til, er endepunkta i intervallet $(a,b)$. Dei trengst for å få B-splinen til å gå gjennom endepunkta, men har ingen innverknad på fasongen til sjølve kurva. Det er vanskeleg å fortolke denne rekursive definisjonen.

For å sjå korleis dette blir i praksis, kjem her eit eksempel på ein B-splinebasis for ein lineær spline ($q = 2$). Det blir då $2q = 4$ ekstra knutepunkt $\xi_{-1} = \xi_0 = a$ og $\xi_{m+1} = \xi_{m+2} = b$. For å gjere det enkelt, set vi her $m$ til å vere 1. Då er det berre eitt indre knutepunkt. Med $x$ i intervallet $[a,b]$ blir

dei tre basisfunksjonane

$$\begin{aligned} B_{-1,2}(x) &= \frac{x-\xi_{-1}}{\xi_0-\xi_{-1}}B_{-1,1}(x) + \frac{\xi_1-x}{\xi_1-\xi_0}B_{0,1}(x) \\ &= \frac{\xi_1-x}{\xi_1-a}B_{0,1}(x) \\ B_{0,2}(x) &= \frac{x-\xi_0}{\xi_1-\xi_0}B_{0,1}(x) + \frac{\xi_2-x}{\xi_2-\xi_1}B_{1,1}(x) \\ &= \frac{x-a}{\xi_1-a}B_{0,1}(x) + \frac{b-x}{b-\xi_1}B_{1,1}(x) \\ B_{1,2}(x) &= \frac{x-\xi_1}{\xi_2-\xi_1}B_{1,1}(x) + \frac{\xi_3-x}{\xi_3-\xi_2}B_{2,1}(x) \\ &= \frac{x-\xi_1}{b-\xi_1}B_{1,1}(x). \end{aligned} \tag{5.6}$$

Set vi knutepunkta til å vere $a = -1$, $\xi_1 = 0$ og $b = 1$, kan basisfunksjonane uttrykkjast som

$$\begin{aligned} B_{-1,2}(x) &= -xB_{0,1}(x) = \begin{cases} -x & \text{for } x \in [-1,0) \\ 0 & \text{elles} \end{cases} \\ B_{0,2}(x) &= (x+1)B_{0,1}(x) + (1-x)B_{1,1}(x) = \begin{cases} x+1 & \text{for } x \in [-1,0) \\ 1-x & \text{for } x \in [0,1) \\ 0 & \text{elles} \end{cases} \\ B_{1,2}(x) &= xB_{1,1}(x) = \begin{cases} x & \text{for } x \in [0,1) \\ 0 & \text{elles.} \end{cases} \end{aligned} \tag{5.7}$$

Det er tre viktige matematiske eigenskapar ved B-splinefunksjonar [2, kapittel IX]:

$$\begin{aligned} &\text{(i) } B_{i,q}(x) = 0 \text{ for } x \notin [\xi_i, \xi_{i+q}] \\ &\text{(ii) } B_{i,q}(x) = \sum_{-(q-1)}^{m} B_{i,q}(x) = 1 \text{ for } x \in (a,b) \\ &\text{(iii) } B_{i,q}(x) > 0 \text{ for } x \in [\xi_i, \xi_{i+q}] \end{aligned} \tag{5.8}$$

Eigenskapane (i) og (iii) i (5.8) fortel at ein B-splinefunksjon er basert på $q$ ikkje-negative basisfunksjonar, $B_{i-q+1,q}, \ldots, B_{i,q}$, for kvart intervall $[\xi_i, \xi_{i+q}]$. Resten av basisfunksjonane er null på dette intervallet. Eigenskapen (ii) i (5.8) viser at summen av basisfunksjonane på heile intervallet $(a, b)$ er 1, og vektoren $\mathbf{1}$ er difor i spennet til B-spline-basisen.

## 5.3 Spline i alder-periode-kohort-modellen

For å kunne bruke splinefunksjonar i alder-periode-kohort-modellen, treng vi splinefunksjonar som ikkje er like ustabile i halane som vanlege TP-splinefunksjonar og B-splinefunksjonar. Vi må òg ta omsyn til problemet med å identifisere effektestimatorane. Heuer [12] går over til å sjå på kubiske regresjonssplinar fordi dei er glatte og fleksible nok for dei fleste bruksområde. Indeksen $q$ er 4. Vi kan droppe denne slik at $B_i(x)$ er basisvektoren til ein kubisk B-spline. For å unngå ustabilitet i halane, nyttar Heuer naturlege splinefunksjonar.

Vi har knutepunktsekvensen $\xi_{-3}, \ldots, \xi_{m+4}$ med $\xi_{-3} = \cdots = \xi_0 = a$ og $\xi_{m+1} = \cdots = \xi_{m+4} = b$ der $(a, b)$ er observasjonsintervallet. Resten av knutepunkta er fordelte over $(a, b)$ med $a < \xi_1 < \cdots < \xi_m < b$. Splinefunksjonen er lineær på intervalla $(a, \xi_1]$ og $[\xi_m, b)$. Det kjem an på plasseringa av $\xi_1$ og $\xi_m$ kor lange dei lineære delane av splinefunksjonen er. For at funksjonen skal bli lineær på desse intervalla, må den andrederiverte til splinefunksjonen vere null her.

Basisen vi nyttar for splineparametriseringa av tidsvariablane i alder-periode-kohort-modellen er

$$\mathbf{1}, x, \tilde{B}_0(x), \ldots, \tilde{B}_{m-3}(x) \tag{5.9}$$

med dimensjon $m$. Dette er ei endring av den originale B-spline-basisen (5.3) som gjer at halane blir meir stabile. Basisen kan delast opp i konstante, lineære og ikkje-lineære komponentar for å få kurveestimat som kan fortolkast.

Med splinemodellar treng ein ikkje gruppere data i 5- eller 10-års intervall for å få glatte estimat for effektane. Ein får utnytta den tilgjengelege informasjonen betre når ein ikkje grupperer data. Heuer set opp korleis Holford sin modell kan parametriserast med B-splinefunksjonar dersom ratane er gitt i 1-års alders- og periodeintervall [12, avnsitt 4]. Han endrar indeksane til Holford for å indikere at vi ser på årlege data. Aldersintervalla får indeksane $s = 1, \ldots, S$, periodeintervalla $t = 1, \ldots, T$ og kohortintervalla $u = 1, \ldots, U$. Tilsvarande (1.2) blir $u = S - s + t$. Kohortintervalla får lengde 2 år med 1 års overlapping med årlege data. Lineærkomponentane som tilsvarer (1.5) er her definert ved

$$A_L(s) = s - \frac{S+1}{2},\ P_L(t) = t - \frac{T+1}{2},\ C_L(u) = u - \frac{U+1}{2}. \tag{5.10}$$

Desse komponentane står ortogonalt på konstantleddet. For krummingseffektane, dei ikkje-lineære effektane, bruker vi dei ikkje-lineære komponentane frå (5.9), $\tilde{B}_0(x), \ldots, \tilde{B}_{m-3}(x)$. Vi får følgjande krummingskomponentmatriser

for dei tre tidsvariablane i modellen:

$$\begin{aligned} \text{alder: } Z^a(s) &= [\tilde{B}_0^a(s), \ldots, \tilde{B}_{m_a-3}^a(s)],\ s = 1, \ldots, S \\ \text{periode: } Z^p(t) &= [\tilde{B}_0^p(t), \ldots, \tilde{B}_{m_p-3}^p(t)],\ t = 1, \ldots, T \\ \text{kohort: } Z^c(u) &= [\tilde{B}_0^c(u), \ldots, \tilde{B}_{m_c-3}^c(u)],\ u = 1, \ldots, U \end{aligned} \tag{5.11}$$

Her er $m_a$, $m_p$ og $m_c$ talet på indre knutepunkt for alder, periode og kohort. Når vi kombinerer (5.10) og (5.11), får vi den log-lineære alder-periode-kohort-modellen

$$\log(\lambda_{st}) = \mu + A_L(s)\bar{\alpha}_L + P_L(t)\bar{\pi}_L + f^a(s) + f^p(t) + f^c(u) \tag{5.12}$$

der $f^a(s) = Z^a(s)\vartheta^a$, $f^p(t) = Z^p(t)\vartheta^p$, $f^c(u) = Z^c(u)\vartheta^c$, $\mu$ er skjæringspunktet, $\bar{\alpha}_L$ og $\bar{\pi}_L$ er parametrane for lineær trend for alder og periode, $\vartheta^a$, $\vartheta^p$ og $\vartheta^c$ er parametervektorane for dei ikkje-lineære basisvektorane i (5.11) og $\lambda_{st}$ er den årlege Poisson-raten [12].

## 5.4 Spline i apc.fit

Som i avnsitt 4.4 skal vi bruke datasettet lungDK for å sjå på programmet `apc.fit`. I staden for å bruke opsjonen `model="factor"`, bruker vi her `"ls"`, `"ns"` og `"bs"`. Som parametrisering av effektane er `parm="AC-P"` brukt, og vi ser på resultata for modellen med berre alder og alder-kohort-modellen.

| | npar=2 | npar=3 | npar=5 | npar=8 |
|---|---|---|---|---|
| Alder | 217 | 216 | 214 | 211 |
| Alder-kohort | 215 | 213 | 209 | 203 |

Tabell 5.1: Fridomsgrader med ulikt tal parametrar for splinemodell i apc.fit

Vi ser i tabell 5.1 at for modellen med berre alder mistar vi ei fridomsgrad for kvar parameter vi legg til i modellen. I alder-kohort-modellen mistar vi to fridomsgrader for kvar parameter vi legg til. Dette er fordi vi legg til parametrane for kvart ledd og har difor lagt til to parametrar, éin for aldersleddet og éin for kohortleddet, i den siste modellen. Det er like mange fridomsgrader uavhengig av kva type splinemodell som blir brukt.

Først ser vi på lineære splinefunksjonar med `npar` lik 3 og 5. Denne opsjonen fortel kor mange parametrar som skal bli brukt for kvart ledd i modellen. Det vil seie at det er eit uttrykk for kor mange stykke kurva blir delt opp i. Resultata frå deviansanalysen står i tabell 5.2. Vi ser her at residualdevianSen minkar kraftig når `npar` aukar. Det blir mindre usikkerheit i modellen når parametertalet aukar.

| | npar=2 | npar=3 | npar=5 |
|---|---|---|---|
| Alder | 16335,6326 | 15751,6192 | 15465,6696 |
| Alder-kohort | 2121,6127 | 1492,9335 | 1109,0113 |

Tabell 5.2: Residualdevians med model="ls"

| | npar=2 | npar=3 | npar=5 | npar=8 |
|---|---|---|---|---|
| Alder | 15523,5991 | 15463,7394 | 15433,8604 | 15433,0281 |
| Alder-kohort | 1297,9119 | 1209,0713 | 1032,0022 | 989,3221 |

Tabell 5.3: Residualdevians med model="ns"

Deretter ser vi på naturlege splinefunksjonar med `npar` lik 2, 3, 5 og 8. Resultata står i tabell 5.3. Residualdeviansen minkar når talet parametrar aukar. Modellen blir altså betre når vi legg til parametrar, men vi mistar fridomsgrader. For kvar av desse modellane er det eitt indre knutepunkt mindre enn parametertalet. For modellen med `npar=5` vil det seie at vi har fire indre knutepunkt og i tillegg endepunkta. Splinefunksjonen er difor sett saman av fem polynom.

| | npar=3 | npar=5 | npar=8 |
|---|---|---|---|
| Alder | 15433,9621 | 15433,7360 | 15430,2690 |
| Alder-kohort | 1304,9881 | 1065,1509 | 997,0611 |

Tabell 5.4: Residualdevians med model="bs"

Til slutt ser vi på B-splinefunksjonar med `npar` lik 3, 5 og 8. Med `npar=2` kjem det ei åtvaringsmelding frå R om at minimumsverdien for B-splinefunksjonar er 3. Dette kjem av at ein med B-splinefunksjonar legg til fleire endepunkt. For å ha to knutepunkt i kvar ende av splinekurva må det minst vere tre parametrar. Med fire eller fleire parametrar får ein òg indre knutepunkt. I tabell 5.4 er residualdeviansen for B-splinefunksjonane vist. Desse er òg viste i figur 5.1. Vi ser at for alder og periode overlappar dei tre kurvene nesten totalt. For kohort er det ganske stor skilnad i endane. Det er større svingingar med fleire parametrar.

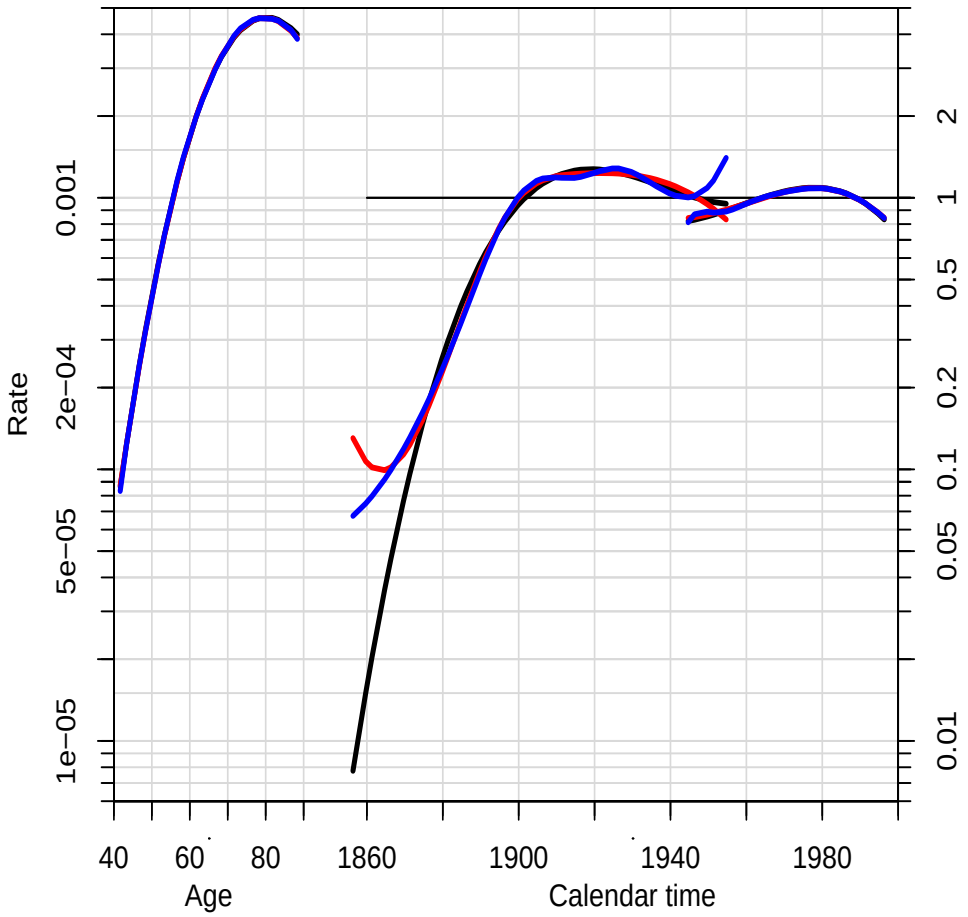
I figur 5.2 er ratane for dei tre ulike splinefunksjonane med fem parametrar viste. Bortsett frå i endane for kohort-kurva er det ikkje vesentlege skilnader mellom dei tre variantane.

For å samanlikne `apc.fit` med `glm` når splinefunksjonar er brukte, ser vi på ein modell for kubisk spline med to knutepunkt ut frå Carstensen si framstilling [3, side 3033]. Knutepunkta er dei same som for naturleg spline med tre parametrar og B-spline med fem parametrar. Resultata frå analysen

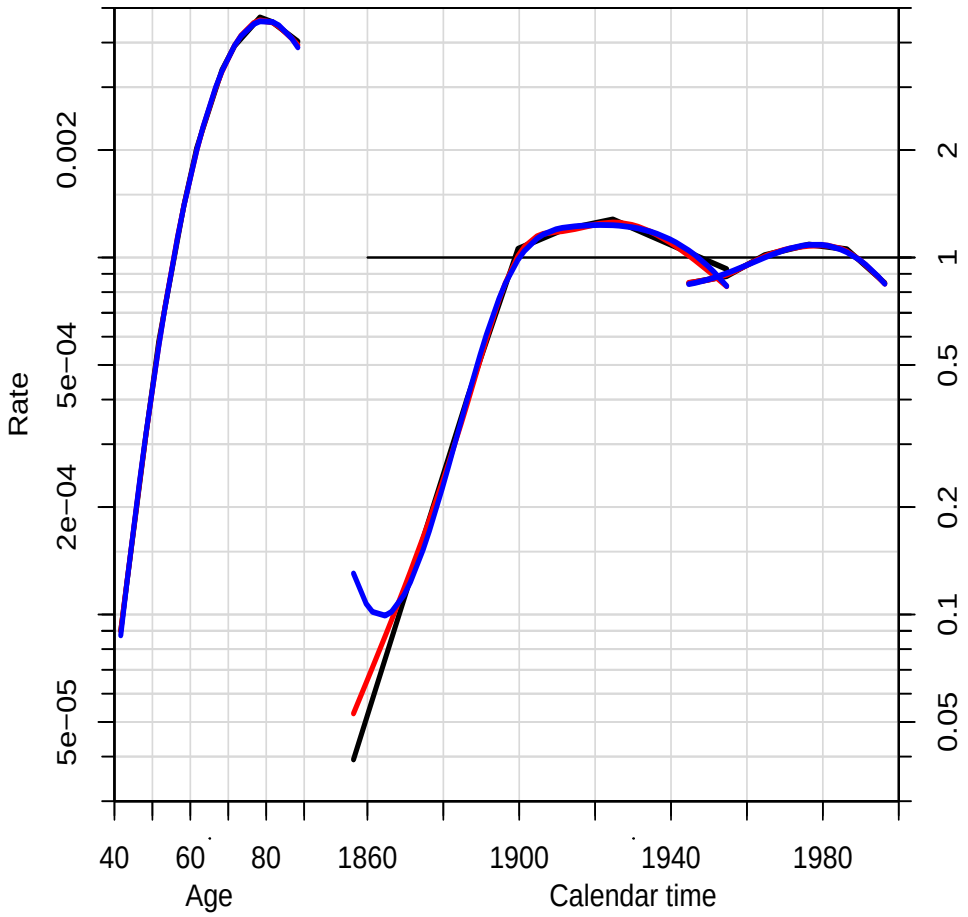
| | F.grader | Res.dev. |
|---|---|---|
| Alder | 216 | 15751,619 |
| Alder-kohort | 213 | 1492,9335 |

Tabell 5.5: Utdata frå splinemodell med to knutepunkt med programmet glm

står i tabell 5.5. Dette er nøyaktig same resultat som for lineær spline med tre parametrar. Dette kan tyde på at definisjonen av lineær spline som er brukt i programmet, ikkje er den same som i avsnitt 5.1.




Figur 5.1: B-splinefunksjonar for alder mot rate (til venstre) og kohort og periode mot forholdsrate (til høgre). 3 parametrar (svart), 5 parametrar (raud) og 8 parametrar (blå) er vist.




Figur 5.2: Splinefunksjonar for alder mot rate (til venstre) og kohort og periode mot forholdsrate (til høgre). Lineær spline (svart), naturleg spline (raud) og B-spline (blå) med 5 parametrar.

# Kapittel 6

# Praktisk bruk av modellen

Vi vil her sjå på nokre artiklar som nyttar alder-periode-kohort-modellar og programmet `apc.fit` for å analysere utviklinga til ulike krefttypar.

## 6.1 Kolorektal kreft

Larsen og Bray [15] har sett på tilfelle av kolorektal kreft i Noreg i perioden 1962-2006. Kolorektal kreft er eit felles namn på kreft i tjukktarm og endetarm [1]. Krefttilfella er klassifiserte i tre etter lokalisering: i endetarmen og på høgre og venstre side i tjukktarmen. I tillegg til diagnosetidspunkt og anatomisk plassering er det henta informasjon om kjønn, alder og topografikodar frå Kreftregisteret og populasjonsdata frå Statistisk sentralbyrå. Alderen er avgrensa til 35-74 år og er inndelt i 5-års intervall. Periodane er òg delte inn i 5-års intervall, og fødselskohortane får dermed 10-års intervall der det første er 1888-1897 og det siste er 1968-1977. Det er brukt ein faktormodell med alder, periode og kohort som faktorar. Driftleddet, summen av stigningstala til perioden og kohorten, har blitt brukt til å uttrykkje utviklinga for plassering, kjønn og alder. Gjennomsnittet og stigningstalet til periodeeffekten er sett lik null fordi ein antek at endringar i diagnoseraten skuldast fødselskohorten. Dette er rimeleg fordi miljøfaktorar er viktige årsaker til kolorektal kreft.

Programmet `apc.fit` er brukt for å gjennomføre analyse av alder-periode-kohort-modellen. Det er brukt ein naturleg splinefunksjon med fem parametrar for alder og periode og tolv parametrar for fødselskohorten fordi ein på førehand var mest interessert i generasjonseffektar. Knutepunkta er sette slik at det er like mange hendingar mellom dei. Sidan det står at berre periodeeffekten er avgrensa til å ha gjennomsnitt og stigningstal lik null, er `parm="ACP"` brukt som argument for parametriseringa av effektane i `apc.fit`. Medianane for periode og kohort er brukte som referansekategoriar. Periode-

og kohorteffektane er presenterte som forholdsratar relativt til kvar sin referansekategori. Alderseffektane er presenterte som ratar for referansekohorten.

Resultata frå [15] viser at det sidan 1960-åra har vore ein auke i tilfelle av kolorektal kreft for begge kjønn i Noreg. Auken ser ut til å minke frå 1980-85, spesielt for endetarmskreft blant menn. Andelen av kreftsvulstar på høgre side i forhold til alle tilfella, aukar med tida for både kvinner og menn. På 1960- og 1970-talet var det ein rask auke i tilfelle av kolorektal kreft for både kvinner og menn. Denne auken minka frå 1980 for kvinner og omtrent fem år seinare for menn. Det er berre nedgang for venstresidig tjukktarmskreft og endetarmskreft. Frå 1987 til 2006 har ratane vore ganske stabile. For høgresidig tjukktarmskreft har det vore ein aukande trend for begge kjønn i heile perioden.

Dersom ein samanliknar dei observerte trendane i ratane mot fødselskohort og kalenderperiode, ser ein at det er ein kraftig auke med tida for alle lokaliseringane. Linjene for påfølgjande generasjonar er parallelle, men det er ikkje så tydeleg for periodane. Det er ein korttidsnedgang i ratane for kohorten 1938-1947. Tilfella av venstre- og høgresidig tjukktarmskreft stabiliserer seg for generasjonane fødde etter 1955, medan tilfella av endetarmskreft aukar for etterkrigsgenerasjonane spesielt blant kvinner [15].

Uavhengig av lokaliseringa er det ein signifikant ikkje-lineær effekt for både periode og kohort. Den fulle alder-periode-kohort-modellen beskriv datasettet best. Set ein stigningstalet for perioden til null, slik at drift er inkludert i kohorttrenden, får vi ein tydeleg auke i risikoen for generasjonane frå 1900 til 1940. Utanom effekten av andre verdskrigen, impliserer denne parametriseringa ein uniform auke i risikoen for endetarmskreft for begge kjønn for påfølgjande kohortar frå 1910 til slutten av 1960-åra. For endetarms- og venstresidig tjukktarmskreft er det ein perioderelatert nedgang [15].

Finland opplever den same aukande trenden som Noreg, medan ratane har flata ut dei siste 25 åra i Danmark og Sverige. Dette reflekterer at utbreiinga er ulik, at risikofaktorar er ulikt fordelte og at det er ulike årsaker til at pasientar får kolorektal kreft.

Det er ikkje lett å forklare kva som er dei underliggjande årsakene til trendane. Tidleg på 1970-talet vart koloskopi introdusert som ein diagnosereiskap i Noreg, men dette samanfell ikkje i tid med store endringar i ratane.

## 6.2 Endometriekreft

Lindemann, Eskild, Vatten og Bray [17] har sett på førekomstar av endometriekreft i Noreg i perioden 1953-2007. Dei prøver òg å seie noko om korleis utviklinga kan bli i perioden 2008-2027. Endometriekreft eller livmorkreft er

definert som ein vondarta svulst i slimhinna i livmora [1]. Lindemann et al. [17] har sett på kvinner i alderen 35 til 79 år delte inn i 5-års aldersgrupper. Fødselskohortar med 10-års overlappande intervall frå 1874-1883 til 1964-1973 er danna ved å trekkje midtpunkta i aldersgruppene frå midtpunkta i 5-års periodegrupper. For å sjå på effektane alder, periode og kohort har på førekomstane av livmorkreft, er kommandoen `apc.fit` i R brukt. Periode- og kohorteffektane er presenterte som forholdsratar relativt til den korresponderande mediankohorten eller -perioden. Det kjem ikkje klart fram kva parametrisering som er brukt i programmet eller om to ulike parametriseringar er brukte.

For å sjå framover på kor mange sjukdomstilfelle det kan bli i 2015 og 2025, er det brukt alder-periode-kohort-modellar for to ulike scenario. Dei fire siste 5-års periodane som er observerte, er brukte til å ekstrapolere sjukdomsutviklinga for dei fire 5-års periodane frå 2008 til 2027. Scenario A er rekna ut ved hjelp av ein potensfunksjon og ein projeksjon av den lineære trenden dei siste ti åra (1998-2007). Dette representerer eit konservativt framtidig mønster. I scenario B går ein ut frå at den lineære trenden ikkje blir svekt i framtidige periodar. Talet på nye tilfelle er rekna ut ved å multiplisere dei utrekna insidensratane med mediumvarianten av folketalsprognosane frå Statistisk sentralbyrå. Endringar i talet på forventa nye tilfelle kjem både av demografiske endringar (aldring og vekst) og risikoendringar.

Det er naturleg å dele aldersgruppene i to, før og etter menopausen som er sett til 55 år. For kvinner under 55 år er ratane for livmorkreft høgare i Noreg enn i dei andre nordiske landa, men for eldre kvinner har ratane i Noreg, Sverige og Finland konvergert i løpet av dei siste 20 åra. Risikoen for å få livmorkreft har auka med tida og er estimert til omtrent 2,5 % for kvinner diagnostisert mellom 2003 og 2007. Mellom 1988 og 2007 var den årlege auken i tilfelle av livmorkreft 1,7 % blant norske kvinner under 80 år [17].

Blant norske kvinner under 55 år var det ein nedgang i tilfelle av livmorkreft, både med omsyn til periode og alder fram til 1998, deretter auka førekomsten. Blant kvinner over 55 år har ratane auka for etterfølgjande fødselskohortar frå slutten av 1800-talet og for periodar frå tidleg på 1960-talet.

Den fulle alder-periode-kohort-modellen gir best tilpassing for datasettet med devians 26,6 og 27 fridomsgrader. Det er ikkje signifikant forskjell mellom alder-drift-modellen og alder-periode-modellen. Uansett om den lineære trenden for periode eller kohort er sett til å vere null, er det aukande ratar for generasjonane fødde mellom 1880 og 1925. For generasjonane fødde frå 1925 til 1945 er ratane stabile, medan det for dei seinare kohortane er ein nedgang i ratane for livmorkreft. Dersom vi går ut frå at den underliggjande lineære trenden kjem av kalenderperioden, viser analysen av periodeeffekta-

ne at ratane aukar uniformt utover i perioden studien varte med ein mogleg akselerasjon frå 1998 [17].

Til slutt har Lindemann et al. [17] sett på forventa utvikling av trendane for livmorkreft fram til 2027. Ratane er i Noreg forventa å ha ein topp rundt 2020 og deretter ha ein nedgang med scenario A. Frå 2005 til 2015 vil det vere ein 35 % auke som kjem av underliggjande risiko (18 %) og aldrande populasjon (17 %). I 2025 vil det vere 57 % fleire tilfelle enn i 2005. Omtrent ein tredel av denne auken kjem av aldrande populasjon. Dersom ein ser på scenario B der det er konstant drift, vil ratane auke omtrent lineært fram til midten av 2020-åra. Det vil då bli ei dobling av nye tilfelle i perioden frå 2005 til 2025. Andelen tilfelle diagnostisert ved 80 år eller eldre, er estimert til 21 % i 2025 medan han i 2005 var 18 %. Dette kjem ikkje berre av aldrande populasjon, men òg fordi det er observert ein relativt høg gjennomsnittleg årleg auke i tilfelle blant kvinner som er 80 år eller eldre frå 1998 til 2007.

## 6.3 Non-Hodgkins lymfom

Viel, Fournier og Danzon [23] har sett på tilfelle av non-Hodgkins lymfom i den franske regionen Doubs i perioden 1980-2005. Non-Hodgkins lymfom (NHL) er ei gruppe av kreftsjukdommar som høyrer inn under hovudgruppa malignt lymfom (kreft i lymfesystemet) [1]. Viel et al. [23] har sett på pasientar i alderen 20-89 år. Den log-lineære modellen (1.4) er brukt for å finne effektane av alder, periode og kohort. Det er brukt B-splinefunksjonar med restriksjonar (naturlege splinar) med sju parametrar for ledda for alder, periode og kohort. For å teste for signifikans av effektar mellom modellane, er skilnaden i deviansen mellom dei ulike modellane samanlikna ved bruk av F-testen. Testen er signifikant dersom den tosidige p-verdien er mindre enn 0,05.

For å analysere data er programmet `apc.fit` i R brukt med `model="ns"`, `parm="AP-C"` og `dr.extr="weighted"`. På førehand går ein ut frå at endringa i ratane hovudsakleg skuldast periodeeffektane. Difor er parametriseringa `AP-C` brukt. Der er først ein alder-periode-modell tilpassa med det første året (1980) som referanseperiode. Deretter er logaritmane til dei tilpassa verdiane frå denne modellen brukte som offset-variablar i ein modell med kohorteffektar. Driftparameteren er funnen ved å bruke vekta gjennomsnitt med `dr.extr="weighted"`.

I perioden mellom 1980 og 2005 vart det i Doubs-regionen registrert 1457 tilfelle av NHL i aldersgruppa 20-89 år [23]. Den korresponderande populasjonen var i 1999 på 367 842. Det er aukande tal tilfelle av NHL med aukande alder gjennom heile analyseperioden. Det ser ikkje ut til å vere noko samspel

mellom alder og periode. Alder-drift-modellen har signifikant betre tilpassing enn modellen med berre alder. Store krummingseffektar for perioden viser at det ikkje berre er ein lineær trend for perioden. Signifikante krummingseffektar for kohorten er det verken i alder-periode-kohort-modellen eller i alder-kohort-modellen. Periodeeffekten gjorde eit hopp i 1983 og stabiliserte seg i 1992 [23].

Auken i periodeeffektane kan ha samanheng med betring i NHL-oppdaging og aukande bruk av nye metodar og teknikkar. Dersom dette stemmer, vil ratane halde seg på noverande nivå fordi moderne diagnostiseringsprosedyrar no er tekne i bruk overalt. Endringar i klassifikasjon og betre diagnostisering kan ikkje forklare den jamne auken insidensratane til NHL har fram til slutten av 1990-talet. Det er meir sannsynleg at ei aukande eksponering for risikofaktorar er ei god forklaring. Viel et al. [23] har ein hypotese om at auken i periodeeffekten i tidsrommet 1983-1992 kjem av miljømessige faktorar frå 1960-talet som har hatt ein 20-års latensperiode. Dei konkluderer med at auken i NHL-tilfelle i Doubs-regionen er meir avhengig av alder og periode enn av kohort.

# Kapittel 7

# Konklusjon og vidare arbeid

Trass i at alder-periode-kohort-modellen kan vere vanskeleg å forstå ut frå Holford [13] si forklaring, er modellen brukt av mange forskarar innan kreft-epidemiologi [15, 17, 23]. Det er ein fordel at modellen tek omsyn til alder, periode og kohort samstundes, men dette er òg veikskapen til modellen. Sidan tidsvariablane er avhengige av kvarandre, kan det vere vanskeleg å skilje mellom effektane av faktorane. Dette problemet med å identifisere kva den lineære trenden kjem av blir teke opp av Clayton og Schifflers [5, 6] og Carstensen [3].

Holford [13] parametriserer modellen slik at kvar tidseffekt blir delt i to komponentar: lineær og ikkje-lineær trend. Desse komponentane står ortogonalt på kvarandre. For å teste om parametrane er estimerbare, bruker Holford ein generalisert invers. Holford sin konklusjon på dette spørsmålet stemmer ikkje fordi han undervegs har brukt ein føresetnad som ikkje stemmer. Vi veit difor ikkje kva som skal til for at ein skal ha estimerbare parametrar.

Holford [13] sin modell der det er ulik storleik på intervalla for tidsfaktorane kan vere nyttig dersom ein vil vektleggje effekten av éin faktor meir enn dei to andre. I artikkelen [14] ser Holford nærmare på ulike måtar å parametrisere modellen med ulik storleik på tidsintervalla. Det kunne ha vore interessant å sjå meir på slike modellar.

R-programmet `apc.fit` er laga av Carstensen for å kunne implementere alder-periode-kohort-modellar. Programmet er enkelt å køyre med éin kommando med fleire argument. Dokumentasjonen [4] er på nokre punkt mangelfull slik at det er vanskeleg å vite korleis modellen ein køyrer eigentleg er parametrisert. Programmet er ikkje brukt i så mange artiklar. Det kan kome av at programmet er ganske nytt, at ein del bruker andre statistikkprogram som SAS i staden for R og at dokumentasjonen ikkje er utfyllande nok.

Det kan vere nyttig å bruke splinefunksjonar i alder-periode-kohort-modellar slik Heuer [12] innfører i artikkelen sin. Ved bruk av splinefunskjonar treng

ein ikkje gruppere data, og ein får dermed utnytta den tilgjengelege informasjonen betre. I dokumentasjonen til `apc.fit` [4] står det ikkje korleis dei ulike splinemodellane er definerte. Dette gjer det vanskeleg for brukaren å vite kva modell han skal velje og korleis han skal tolke resultata.

Robertson, Gandini og Boyle [19] samanliknar metodar for å ta omsyn til problemet med å identifisere om den lineære trenden kjem av endring i periode eller kohort. Dei konkluderer med ei anbefaling av metodar basert på krumming eller andre estimerbare funksjonar. Dei nemner framgangsmåten til Clayton og Schifflers [6] og Holford [13] som døme på slike metodar.

Tarone og Chu [22] har brukt ikkje-parametriske regresjonsmetodar for å analysere innverknaden fødselskohorten har på brystkreft. Eit mogleg vidare arbeid frå denne oppgåva kan vere å sjå på alder-periode-kohort-modellar med slike ikkje-parametriske metodar og samanlikne dette med modellar med splinefunksjonar.

Det kan òg vere aktuelt å analysere eit datasett for ein type kreft ved hjelp av ulike alder-periode-kohort-modellar. Ein kan då samanlikne resultata ein får frå analysar med ulike parametriseringar eller med og utan splinefunksjonar.

# Litteratur


[1] *Store medisinske leksikon.* Kunnskapsforlaget, Oslo, 2006.

[2] C. de Boor. *A practical guide to splines.* Springer Verlag, New York, 1978.

[3] B. Carstensen. Age-period-cohort models for the Lexis diagram. *Statistics in Medicine*, 26(15):3018–3045, 2007.

[4] B. Carstensen. Fit an age-period-cohort model to tabular data. `http://finzi.psych.upenn.edu/R/library/Epi/html/apc.fit.html`, februar 2010.

[5] D. Clayton og E. Schifflers. Models for temporal variation in cancer rates .1. age-period and age-cohort models. *Statistics in Medicine*, 6(4):449–467, 1987.

[6] D. Clayton og E. Schifflers. Models for temporal variation in cancer rates .2. age-period-cohort models. *Statistics in Medicine*, 6(4):469–481, 1987.

[7] N. R. Draper og H. Smith. *Applied regression analysis.* Wiley, New York, 1966.

[8] R. L. Eubank. *Spline smoothing and nonparametric regression.* Dekker, New York, 1988.

[9] R. A. Fisher og F. Yates. *Statistical tables for biological, agricultural and medical research.* Longman, London, 1963.

[10] J. Friedman, T. Hastie og R. Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag New York, New York, NY, 2009.

[11] C. F. Gerald og P. O. Wheatley. *Applied numerical analysis.* Addison-Wesley, Boston, 2004. 7th ed.

[12] C. Heuer. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, 53(1):161–177, 1997.

[13] T. R. Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39(2):311–324, 1983.

[14] T. R. Holford. Approaches to fitting age-period-cohort models with unequal intervals. *Statistics in Medicine*, 25(6):977–993, 2006.

[15] I. K. Larsen og F. Bray. Trends in colorectal cancer incidence in Norway 1962-2006: an interpretation of the temporal patterns by anatomic subsite. *International Journal of Cancer*, 126(3):721–732, 2010.

[16] D. C. Lay. *Linear algebra and its applications*. Pearson education, Boston, 2006.

[17] K. Lindemann, A. Eskild, L. J. Vatten og F. Bray. Endometrial cancer incidence trends in Norway during 1953-2007 and predictions for 2008-2027. *International Journal of Cancer*, 127(11):2661–2668, 2010.

[18] C. R. Rao og S. K. Mitra. *Generalized inverse of matrices and its applications*. Wiley, New York, 1971.

[19] C. Robertson, S. Gandini og P. Boyle. Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology*, 52(6):569–583, 1999.

[20] S. R. Searle. *Linear models*. Wiley, New York, 1971.

[21] I. dos Santos Silva. *Cancer epidemiology: principles and methods*. International Agency for Research on Cancer, Lyon, 1999.

[22] R. E. Tarone og K. C. Chu. Implications of birth cohort patterns in interpreting trends in breast-cancer rates. *Journal of the National Cancer Institute*, 84(18):1402–1410, 1992.

[23] J. F. Viel, E. Fournier og A. Danzon. Age-period-cohort modelling of non-Hodgkin's lymphoma incidence in a French region: a period effect compatible with an environmental exposure. *Environmental Health*, 9, 2010.

# Tillegg A

# Programmeringskode frå R

## A.1 Enkelt datasett

```
A P D Y
32 1977 10 105854
37 1977 13 108632
37 1982 12 103548
42 1982 14 106482
42 1987 15 108573
47 1987 19 107435
47 1992 18 103547
32 1977 8 103684
32 1982 11 109482
37 1982 12 103145
37 1987 14 106482
42 1987 16 106884
42 1992 15 102648
47 1992 19 104823
32 1982 10 105486
32 1987 9 103549
37 1987 14 103575
37 1992 13 103843
42 1992 17 104982
32 1987 11 106842
37 1992 13 103548
```

## A.2 Program med apc.fit

```
> datasett <- read.table("datasett4.txt", header=T)
> attach(datasett)
> library(Epi)

Attaching package: 'Epi'


        The following object(s) are masked from package:base :

         as.Date.numeric,
         merge.data.frame

>
> vekta <- apc.fit(datasett, ref.c=1940, ref.p=1977, dist="poisson",
model="factor", dr.extr="weighted", parm="ACP", scale=1)
[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
Age                      17    1.20036
Age-drift                16    0.85254  1  0.34782    0.5553
Age-Cohort               14    0.59348  2  0.25906    0.8785
Age-Period-Cohort        12    0.53948  2  0.05400    0.9734
Age-Period               14    0.80950 -2 -0.27002    0.8737
Age-drift                16    0.85254 -2 -0.04303    0.9787
>
> vekta$Age
  Age         Rate         2.5%        97.5%
1  32 8.572023e-05 5.221875e-05 0.0001407149
2  37 1.161709e-04 7.465310e-05 0.0001807784
4  42 1.368795e-04 8.778009e-05 0.0002134426
6  47 1.751097e-04 1.151213e-04 0.0002663575
> vekta$Per
   Per      P-RR      2.5%    97.5%
1 1977 1.0000000 1.0000000 1.000000
3 1982 0.9801912 0.6395520 1.502262
5 1987 1.0149217 0.7274398 1.416015
7 1992 0.9884469 0.8082504 1.208817
```

```
> vekta$Coh
    Coh     C-RR      2.5%    97.5%
2  1940 1.000000 1.0000000 1.000000
1  1945 1.023299 0.7206089 1.453132
9  1950 1.142121 0.7708484 1.692214
16 1955 1.092161 0.6897694 1.729297
> vekta$Drift
    exp(Est.)      2.5%    97.5%
APC  1.008768 0.9802785 1.038086
A-d  1.008575 0.9803178 1.037646
>
> holford <- apc.fit(datasett, ref.c=1940, ref.p=1977, dist="poisson",
model="factor", dr.extr="Holford", parm="ACP", scale=1)
[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
Age                      17    1.20036
Age-drift                16    0.85254  1  0.34782    0.5553
Age-Cohort               14    0.59348  2  0.25906    0.8785
Age-Period-Cohort        12    0.53948  2  0.05400    0.9734
Age-Period               14    0.80950 -2 -0.27002    0.8737
Age-drift                16    0.85254 -2 -0.04303    0.9787
>
> holford$Age
  Age         Rate         2.5%        97.5%
1  32 8.566293e-05 5.178252e-05 0.0001417107
2  37 1.161709e-04 7.465310e-05 0.0001807784
4  42 1.369711e-04 8.862728e-05 0.0002116852
6  47 1.753441e-04 1.171111e-04 0.0002625331
> holford$Per
   Per      P-RR      2.5%    97.5%
1 1977 1.0000000 1.0000000 1.000000
3 1982 0.9795360 0.6464342 1.484282
5 1987 1.0135653 0.7500668 1.369631
7 1992 0.9864661 0.8287829 1.174150
> holford$Coh
    Coh     C-RR      2.5%    97.5%
2  1940 1.000000 1.0000000 1.000000
1  1945 1.023983 0.7198289 1.456654
```

```
9  1950 1.143649 0.7711831 1.696010
16 1955 1.094354 0.6883736 1.739770
> holford$Drift
    exp(Est.)      2.5%    97.5%
APC  1.008525 0.9797409 1.038156
A-d  1.008575 0.9803178 1.037646
>
> detach()
```

## A.3 Program med glm

```
> datasett <- read.table("datasett4.txt", header=T)
> attach(datasett)
>
> C = P-A
>
> summary(glm(D ~ factor(A) + offset(log(Y)),
family = poisson(link="log")))

Call:
glm(formula = D ~ factor(A) + offset(log(Y)),
family = poisson(link = "log"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.54286  -0.19967   0.02897   0.11585   0.43616

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.2837     0.1302 -71.310  < 2e-16 ***
factor(A)37   0.2899     0.1671   1.735 0.082794 .
factor(A)42   0.4477     0.1730   2.587 0.009671 **
factor(A)47   0.6462     0.1866   3.463 0.000533 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 14.5865  on 20  degrees of freedom
Residual deviance:  1.2004  on 17  degrees of freedom
```

```
AIC: 102.13

Number of Fisher Scoring iterations: 3

> summary(glm(D ~ factor(A) + factor(P) + offset(log(Y)),
family = poisson(link="log")))

Call:
glm(formula = D ~ factor(A) + factor(P) + offset(log(Y)),
family = poisson(link = "log"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.39358  -0.12633   0.03401   0.14651   0.34429

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.33366    0.19300 -48.360   <2e-16 ***
factor(A)37    0.26185    0.17556   1.492   0.1358
factor(A)42    0.39685    0.19357   2.050   0.0404 *
factor(A)47    0.57770    0.22116   2.612   0.0090 **
factor(P)1982  0.03904    0.22730   0.172   0.8636
factor(P)1987  0.10795    0.22154   0.487   0.6261
factor(P)1992  0.12380    0.24020   0.515   0.6063
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 14.5865  on 20  degrees of freedom
Residual deviance:  0.8095  on 14  degrees of freedom
AIC: 107.74

Number of Fisher Scoring iterations: 3

> summary(glm(D ~ factor(A) + factor(C) + offset(log(Y)),
family = poisson(link="log")))

Call:
glm(formula = D ~ factor(A) + factor(C) + offset(log(Y)),
family = poisson(link = "log"))
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.36026  -0.09013  -0.03357   0.09822   0.30660

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.36374    0.21901 -42.755  < 2e-16 ***
factor(A)37    0.30156    0.16892   1.785 0.074220 .
factor(A)42    0.46812    0.18479   2.533 0.011303 *
factor(A)47    0.71376    0.21231   3.362 0.000774 ***
factor(C)1945  0.01880    0.17764   0.106 0.915711
factor(C)1950  0.12836    0.19912   0.645 0.519180
factor(C)1955  0.08864    0.23397   0.379 0.704814
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 14.58648  on 20  degrees of freedom
Residual deviance:  0.59348  on 14  degrees of freedom
AIC: 107.52

Number of Fisher Scoring iterations: 3

> summary(glm(D ~ factor(A) + factor(P) + factor(C)
+ offset(log(Y)), family = poisson(link="log")))

Call:
glm(formula = D ~ factor(A) + factor(P) + factor(C)
+ offset(log(Y)), family = poisson(link = "log"))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.370731  -0.079682  -0.006405   0.106474   0.262616

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.335036   0.212802 -43.867   <2e-16 ***
factor(A)37    0.274587   0.177131   1.550   0.1211
factor(A)42    0.409240   0.198544   2.061   0.0393 *
```

```
factor(A)47    0.626165   0.244402   2.562   0.0104 *
factor(P)1982  0.009379   0.234180   0.040   0.9681
factor(P)1987  0.073584   0.234504   0.314   0.7537
factor(P)1992  0.076538   0.259373   0.295   0.7679
factor(C)1945 -0.006355   0.148399  -0.043   0.9658
factor(C)1950  0.074115   0.163886   0.452   0.6511
factor(C)1955        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 14.58648  on 20  degrees of freedom
Residual deviance:  0.53948  on 12  degrees of freedom
AIC: 111.47

Number of Fisher Scoring iterations: 3

>
> detach()
```