Statistiske metoder for alder-periode-kohort-analyser

En sammenligning av nyere metoder med konvensjonelle generaliserte lineære modeller

Masteroppgave i statistikk – Dataanalyse

Olaug Margrete Askeland

Universitetet i Bergen

Matematisk institutt

20. november 2013

”All statistical models are wrong, but some are useful.”
George E. P. Box

Takk

Aller først vil jeg rette en stor takk til Professor Ivar Heuch som har vært veilederen min på denne oppgaven. Vi har hatt gode møter gjennom den tiden jeg har arbeidet med denne oppgaven, og han har gitt meg god inspirasjon og gode tilbakemeldinger på oppgaven. Det har vært en lærerik prosess der jeg har kunnet rådføre meg med en dyktig og hjelpsom veileder.

Videre fortjener også mannen min, Rune, en stor takk for å ha støttet meg og motivert meg gjennom masterstudiet. Både han og våre barn, Maria og Torstein, har vært tålmodige gjennom hele prosessen. Det har vært både travle dager og litt mindre travle dager, og det har alltid vært godt å komme hjem til dere for avkobling.

Bergen, 20.11.2013
Olaug Margrete Askeland

Innhold

Innledning

I studier av forekomsten av tidsrelaterte hendelser finner epidemiologer, demografer og samfunnsvitere det ofte nyttig å skille mellom tre forskjellige tidsrelaterte dimensjoner, nemlig alder, tidsperiode og fødselskohort. Alder-periode-kohort (APC) analyser forsøker å separere den påvirkningen som skyldes alder, fra den påvirkningen som er assosiert med tidsperiode, og den påvirkning som er assosiert med fødselstidspunkt (kohort). Den velkjente sammenhengen mellom de tre faktorene, periode – alder = kohort, gjør parameterestimeringen vanskelig, og et generelt dilemma ved APC-analyser er problemstillingen med å separere de simultane effektene til alder, periode og kohort. Identifikasjonsproblemet med parameterestimering har blitt studert siden 1970-tallet, og er fremdeles debattert.

Artikkelen til Robert M. O'Brien [1] danner utgangspunktet for denne oppgaven. I artikkelen til O'Brien tar han for seg APC-problemstillingen, og forsøker å dele den inn i to fundamentale problemer. Det ene problemet er konfunderingen av de lineære effektene til alder med effektene til periode og kohort, de lineære effektene til periode med alder og kohort, og de lineære effektene til kohort med periode og alder. Det andre problemet omhandler modellidentifikasjon. O'Brien benytter seg av to metoder for å løse modellidentifikasjonsproblemet og introduserer leseren for *Constrained Generalized Linear Models estimator* (CGLIM) og den nyere metoden *Intrinsic Estimator* (IE). I det videre arbeidet har jeg fokusert på modellidentifikasjon og ulike metoder som er introdusert for å kunne løse dette problemet.
Det første målet med masteroppgaven var å forstå og utdype artikkelen til O'Brien, og deretter arbeide videre med noen av problemstillingene som ble introdusert i artikkelen. Spesielt ønsket jeg å undersøke nærmere hvor robust IE-metoden er. Denne metoden har i nyere tid blitt introdusert som et alternativ til mer tradisjonelle APC-metoder, og IE-metoden forsøker å oppnå modellidentifikasjon med minimale antagelser [2]. IE-metoden produserer estimater som har ønskede statistiske egenskaper, og den gir en unik løsning. I CGLIM-metoden settes det krav til betingelser, ved at to effektkoeffisienter settes lik hverandre. Valget av betingelse må basere seg på tidligere teoretisk eller empirisk informasjon, noe som sjelden eksisterer. Estimatene til koeffisientene er sensitiv for valget av betingelse, og ulike valg kan produsere vidt forskjellige estimater for effektene til aldersgrupper, perioder og fødselskohorter.

Kapittel 1 og 2 gir en kort innføring i epidemiologi og alder-periode-kohort-modellen, samt en introduksjon til teori som er nødvendig for å sette seg inn i oppgavens tematikk.

I Kapittel 3 beskrives den konvensjonelle CGLIM-metoden og den nyere introduserte IE-metoden.

Kapittel 4 viser bruken av CGLIM- og IE-metoden i praksis for et enkelt eksempel med 3 aldersgrupper, 3 perioder og følgelig 5 fødselskohorter.

Kapittel 5 omhandler artikkelen til Robert M. O'Brien [1], der alder-periode-kohort-problemstillingen deles inn i 2 fundamentale problemer.

I Kapittel 6 beskrives metode og simuleringsoppsettet som benyttes for modellvalidering videre i oppgaven. Kapittelet omhandler også de ulike målene som benyttes for senere sammenligning av ulike metoder. Dette er observatorer som sier noe om hvor god en modell er, *goodness-of-fit-statistics*.

Jeg har benyttet simuleringsanalyser for å undersøke hvor godt de ulike metodene gjenskaper den sanne formen til en underliggende modell som genererer dataene. I Kapittel 7 har jeg introdusert 7 ulike modeller for generering av data, og benyttet meg av CGLIM og IE for å analysere disse dataene. Slik kan jeg undersøke om disse estimatorene gir numeriske estimater for alder-, periode- og kohortkoeffisientene som er valide, og avdekker de sanne effektene.

I det videre arbeidet ønsket jeg også å gjøre en tilsvarende sammenligning mellom IE-metoden og andre metoder som er velkjente og som er beskrevet som mulige løsninger for modellidentifikasjonsproblemet til APC-analyser. Kapittel 8 gir leseren kjennskap til artiklene til Clayton og Schifflers [3, 4]. I den første artikkelen tar forfatterne for seg modeller som beskriver variasjon over tid uten å skille mellom periodeinnflytelse og kohortinnflytelse. Forfatterne innfører begrepet *drift* for variasjon som kan beskrives like godt av en alder-periode-modell, som av en alder-kohort-modell. I sin andre artikkel omtaler forfatterne i hvilken logisk rekkefølge de mener man skal vurdere ulike modeller ved analyse av et datasett. Ved å vurdere forbedring i tilpasning av dataene, kan en vurdere hvilken modell som er best egnet for et gitt datasett. Ved å benytte seg av mål som devians, kan en vurdere om en bør benytte seg av en redusert modell som kun inkluderer noen av faktorene (alder, periode og kohort) eller om en bør benytte den fulle APC-modellen. Clayton og Schifflers nevner bl.a. førsteordensdifferanser og andreordensdifferanser for å presentere periode- og kohorteffekter. I siste del av kapittelet har jeg benyttet meg av simuleringsanalyser for å sammenligne IE-metoden med Clayton og Schifflers metoder. For generering av data har jeg benyttet meg av 4 av modellene fra Kapittel 7, og i tillegg er det introdusert 2 nye modeller for generering av data.

Under arbeidet med den aktuelle problemstillingen for oppgaven, har det underveis dukket opp nyere publikasjoner [5] som foreslår Partial Least Squares (PLS) som løsning til modellidentifikasjonsproblemet. Det ble derfor naturlig å vie litt plass til denne tematikken i Kapittel 9. Der har jeg analysert ulike datasett med både IE- og PLS-

metoden for å kunne sammenligne estimatene som beregnes. Til slutt i kapittelet har jeg også gjennomført simuleringsanalyser på en av modellene fra Kapittel 8 for å kunne sammenligne IE-metoden med PLS-metoden.

Underveis i de ulike simuleringsdelene fra Kapittel 7, 8 og 9 er resultatene diskutert. I Kapittel 10 oppsummerer jeg mer generelt resultatene som er presentert tidligere i oppgaven, og introduserer kort eventuelle retninger for et videre arbeid.
Siden IE-metoden blir introdusert som en metode som skal løse problemene knyttet til modellidentifikasjon, har jeg også sett kort på om denne metoden er tatt i bruk i publikasjoner de senere årene.

Vedlegg A inneholder eksempel på R-koder som er benyttet ved simuleringsanalysene. Videre er R-koder for den praktiske bruken av CGLIM, IE og PLS på et gitt datasett gjengitt. I Vedlegg B er noen av resultatene og tabellene som er utelatt fra oppgaven tatt med.

Alle analysene er utført i statistikkprogrammet R, versjon 2.15.2.

1. Innledende teori

1.1 Epidemiologi

Epidemiologi er definert som studiet av helsetilstand og sykdomsutbredelse i en befolkning, og av årsaker til sykdom og død [6]. Epidemiologi omfatter alle former for helserelaterte emner og sykdom, ikke bare smittsomme tilstander, slik navnet kan antyde. I epidemiologisk forskning kartlegger man statistisk sykdommers forekomst og årsaksforhold og man konsentrerer seg om befolkninger. Den aktuelle befolkningen kan avgrenses på ulike måter. Noen ganger er man opptatt av alle som bor eller oppholder seg i et bestemt geografisk område, mens man andre ganger kan være opptatt av dem som arbeider i en bestemt bedrift eller har et spesielt yrke. Epidemiologiske data finnes i stor utstrekning i helseregistre, helseundersøkelser og andre befolkningsbaserte forskningsprosjekter.

I epidemiologiske studier er det vanlig å dele befolkningen inn i visse undergrupper, f.eks. etter kjønn og alder. Dette er vesentlig for forståelsen av variasjoner i sykdomsforekomsten. Om man ikke kjenner aldersfordelingen, kan man således lett trekke feilaktige slutninger om risikoforhold. Metoder der man splitter dataene opp i aldersgrupper kalles aldersstandardisering, og tilsvarende kalles oppsplitting og fordeling av data etter kjønn for kjønnsstandardisering.

I epidemiologi [7] blir nesten alle kvantitative data angitt i forhold til størrelsen på den aktuelle befolkningen, som *rater*. Raten er et uttrykk for frekvensen eller relativ hyppighet av et observert fenomen. Et av de mest brukte målene i epidemiologi er *insidensrate*, som kan defineres som antall nye tilfeller i løpet av en gitt tidsperiode delt på samlet populasjonstid under risiko i den samme tidsperioden. Samlet populasjonstid vil vanligvis oppgis som totalt antall *personår*. Personår er definert som summen av år hver person er med i undersøkelsen frem til eventuell sykdom eller død. Et annet viktig mål i epidemiologien er *dødsrate* eller *mortalitetsrate*, som kan defineres som antall individer som dør i løpet av en gitt tidsperiode delt på samlet populasjonstid under risiko i den samme tidsperioden. Antallet av hendelser slik som død eller sykdomsinsidens følger generelt en poissonfordeling, dvs. responsvariabelen y antas å være poissonfordelt. Responsvariabelen y er den avhengige variabelen, og den kan også være mer lik vanlige normalfordelte variable, spesielt i anvendelser av ikke-medisinsk type.
Variasjonen i responsvariabelen er relatert til en eller flere forklaringsvariabler, som omtales som de uavhengige x-variablene. Eksempler på slike kan være alder, periode og kohort, som er tidsvariabler mye brukt i epidemiologiske studier. I utgangspunktet er disse variablene kontinuerlige størrelser, men det er vanlig i kohortanalyser å anta at variablene er kategoriske. En kategorisk variabel er en faktor med to eller flere

nivåer. Det er vanlig å gruppere tidsvariablene i 5-års intervall for ikke å få altfor mange parametre.

Alder er den løpende alderen til personen når diagnosen blir stilt eller når han dør. Periode er det tidspunktet, den kalenderperiode, der diagnosen til personen blir stilt eller ved død. Og kohort er tidspunktet personen er født, også omtalt som fødselskohort. Alder, periode og kohort er lineært avhengige, og sammenhengen kan uttrykkes som:

$$kohort = periode - alder$$

Disse tre effektene har et felles identifikasjonsproblem siden de har et eksakt lineært forhold mellom seg. Kjenner man verdien på to av variablene, kan en regne ut den tredje. Problemet oppstår fordi hver av de tre faktorene kan skrives som en lineærkombinasjon av de to andre. Det er dermed ingen variasjon i verdiene for hver bestemt variabel for en gitt verdikombinasjon av de to andre, og det blir umulig å skille effektene av disse eksponeringsfaktorene i en analyse med gjensidig justering.

Alderen bør alltid være med i modellene fordi det er en faktor som vanligvis har størst innvirkning på utvikling av sykdom eller død. Periodeeffektene kommer ofte fra forbedringer i diagnostisering og behandling eller ved at man ved et visst tidspunkt har endret på klassifisering av en sykdom. Kohorteffektene gjenspeiler ofte generasjonseffekter, der ulike generasjoner har hatt ulik påvirkning fra livsstil eller miljø. Hvis alder og periode er inndelt i 5-årsintervall, vil kohortinndelingen få overlappende 10-årsintervall. Hvis alder og periode ikke er delt i like lange intervaller, vil det komplisere analysene.

1.2 Alder-periode-kohort-modell

En alder-periode-kohort (APC) modell er et nyttig modelleringsverktøy som kan benyttes til å oppsummere informasjon som er samlet i et register. APC-analyser har spilt en avgjørende rolle i studiet av tidsspesifikke fenomen i sosiologi, demografi og epidemiologi gjennom de siste 80 årene, og denne typen analyser skiller mellom 3 typer av tidsrelatert variasjon i dataene en er interessert i.

Alderseffekter – variasjon assosiert med ulike aldersgrupper
Periodeeffekter – variasjon over tidsperioder som påvirker alle aldersgrupper samtidig
Kohorteffekter – forandringer på tvers av grupper av individer som er født samtidig

Alderseffektene gir informasjon om sykdomsratene, dødsratene eller andre rater i form av ulike aldersgrupper, og representerer ulike risikoer assosiert med ulike aldersgrupper. Periode- og kohorteffekter reflekterer påvirkningen fra sosiale krefter. Periodeeffektene representerer variasjonen i rater over tid som er assosiert med alle

aldersgrupper samtidig. Periodeeffektene kan belyse forandringer i behandlingen som kan affisere alle aldersgrupper samtidig. Periodevariasjon er ofte et resultat fra et skifte i sosialt, historisk og kulturelt miljø. Kohorteffektene er assosiert med langtidseksponering, der ulike generasjoner er eksponert for ulike risikoer. Kohortvariasjon kan reflektere effektene av tidlig eksponering til sosioøkonomiske, atferdsmessige og miljømessige faktorer som handler vedvarende over tid, og gir forskjeller i livsløputfallet for spesifikke kohorter.

APC-modellen benyttes til analyse av tabeller med rater av hendelser som bl. a. fødsler, dødsfall, sykdomsinsidens og kriminalitet. Disse modellene tilpasser effektene av alder, periode og kohort som faktorer. En APC-modell forsøker å dekomponere tidsforandringene til en avhengig variabel i alderseffekter, periodeeffekter og kohorteffekter. Det er vanlig praksis å rapportere alder- og periodeeffekter i 5-årsintervall, noe som resulterer i 10-års overlappende intervaller for de relevante kohortene.

APC-modellene har et identifikasjonsproblem. Fødselsdatoen kan beregnes direkte fra alderen ved en gitt hendelse og datoen for den samme hendelsen. I en generalisert lineær modell (GLM) vil dette føre til overparametrisering og konsekvensen blir ekskludering av et av uttrykkene. Det er derfor nødvendig å sette begrensninger for modellen for å kunne trekke ut identifiserbare resultater for hver av parametrene. Dette er nødvendig siden hver av komponentene (alder, periode og kohort) i modellen gir ulik innsikt i trendene over tid, og ofte er det ønskelig å ha alle disse faktorene med i modellen.

Identifikasjonsproblemet fører til store metodiske utfordringer og det er vanskelig å estimere de sanne separate effektene av alder, periode og kohort samtidig. Dette problemet har vært mye omtalt i litteraturen fra demografi, epidemiologi og biostatistikk, og i løpet av de siste tiårene har en rekke løsninger vært foreslått. Eksempler fra epidemiologi er artiklene til Clayton og Schifflers [4] og Osmond og Gardner [8]. CGLIM, *constrained generalized linear models estimator*, har lenge vært konvensjonell blant demografer og andre samfunnsvitere. Mason et al. [9] introduserte denne teknikken for sosiologer i 1973. Ved å innføre en enkelt betingelse i modellen, kan de individuelle koeffisientene i alder-periode-kohort-modellen identifiseres. Men estimatene til koeffisientene er sensitiv for valget av betingelse, og valget bør derfor basere seg på tidligere teoretisk eller empirisk informasjon, noe som dessverre sjelden eksisterer. Senere utvikling i APC-metodikken i biostatistikk har understreket nytten av estimerbare funksjoner som er upåvirket av valget av begrensninger på parametrene. Clayton og Schifflers [4] og Holford [10] er noen av de som omtaler denne tilnærmingen. Andre som også tar opp dette er Kupper et al. [11] og Robertson et al. [12]. I artiklene [13, 14] introduserer Fu, Yang og Land en ny APC-estimator som er basert på estimerbare funksjoner, og den singulære verdi dekomposisjonen av matriser. Denne nye APC-estimatoren kalles *intrinsic estimator*, IE.

1.3 APC-klassifisering

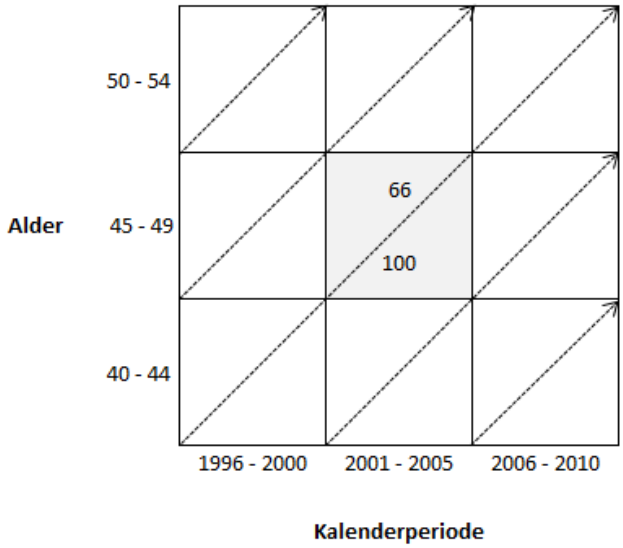
Dataene som benyttes i APC-analyser er ofte ordnet i standardtabeller fordelt på aldersklasser og kalenderperioder. I tabellene er ofte enten antall tilfeller oppgitt sammen med ratene i hver aldersklasse, eller antall tilfeller er oppgitt sammen med populasjonsstørrelsen i hver aldersklasse. Aldersintervallene definerer radene og tidsperiodene definerer kolonnene. I Tabell 1-1 er det vist et eksempel på hvordan en slik tabell kan se ut (utdrag fra en større tabell [15]).

Tabell 1-1: Dødelighet kvinner, etter alder. Døde pr. 100 000 middelfolkemengde. Årlig gjennomsnitt. Kilde: Statistisk sentralbyrå (Tabell 79).

		Periode		
		1996-2000	2001-2005	2006-2010
Aldersgruppe	40-44	114	104	85
	45-49	182	166	150
	50-54	310	272	252

For eksempel var dødsraten for aldersgruppen 45-49 år målt i perioden 1996-2000 på 0,00182, dvs. 182 dødsfall per 100 000 kvinner i befolkningen. Til sammenligning var dødsraten til den samme aldersgruppen i 2006-2010 på 150 dødsfall per 100 000 kvinner. Radene definerer a aldersgrupper og kolonnene definerer p kalenderperioder. I dette eksempelet er $a = 3$ og $p = 3$. Den første dødsraten nevnt over korresponderer også med dødsraten til kvinner født i årene 1946-1955, dvs. fødselskohorten 1946-1955. Kohortene er representert langs diagonalene fra nedre venstre celle til øvre høyre celle. Det er bare én observasjon for den første kohorten, siden medlemmer av denne kohorten vil være i en aldersgruppe som ikke er representert i tabellen den neste perioden. Den neste kohorten har to observasjoner: en i periode 1 – alder 2 og en i periode 2 – alder 3. I dette eksempelet er antall kohorter $k = 5$.

Et alternativ til standardtabellene er et Lexis-diagram, der alle tre tidsvariablene er med i samme diagram. De horisontale og vertikale linjene representerer aldersgruppe- og periodeinndelingen. Fødselskohortene korresponderer til diagonalene. Et eksempel på et slikt diagram for eksempelet med 3 aldersgrupper og 3 periodegrupper er vist i Figur 1-1.



Figur 1-1: Lexis-diagram for eksempelet med 3 aldersgrupper og 3 kalenderperioder.

De diagonale linjene avgrenser fødselskohortene i intervaller som er lik de som er benyttet for alder og kalenderperiode. Hver (alder × periode)-celle inneholder 2 trekantete regioner som refererer til tilstøtende fødselskohorter. Kohortene leses på skrå med den eldste kohorten i øvre venstre hjørne og den yngste kohorten i nedre høyre hjørne. Det skraverte feltet i Figur 1-1 refererer til aldersgruppen 45 – 49 år i kalenderperioden 2001 – 2005. Disse kvinnene vil være født i kohorten 1951 – 1960. Diagonalen deler dette kvadratet i to slik at antallet av de som er født i 1951 – 1955 er gjengitt i trekanten oppe til venstre, mens antallet av de som er født i 1956 – 1960 er gjengitt i trekanten nede til høyre. Lexis-diagrammet gir kohorter som nå er av samme lengde som alder og periode, og de er nå ikke-overlappende.

Den multiple APC-klassifikasjonsmodellen ble beskrevet av Mason et al. [9] for demografi- og samfunnsforskning for 40 år siden.
For mortalitetsratene kan denne modellen bli beskrevet på formen av en lineær regresjon:

$$M_{ij} = \frac{D_{ij}}{P_{ij}} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij} \qquad (1\text{–}1)$$

Her står M_{ij} for den observerte raten i aldersgruppe i (for $i = 1, \dots, a$) ved periode j (for $j = 1, \dots, p$). D_{ij} angir antall tilfeller, mens P_{ij} angir populasjonsstørrelsen. Leddet μ står for intercepten eller den justerte gjennomsnittsraten, α_i angir alderseffekten eller koeffisienten for aldersgruppe i, β_j angir periodeeffekten eller koeffisienten for periode j, γ_k angir den diagonale kohorteffekten eller koeffisienten til kohort $k = a - i + j$ og ε_{ij} angir den tilfeldige feilen med forventning $E(\varepsilon_{ij}) = 0$. Antall kohorter er gitt ved $k = 1, \dots, (a + p - 1)$.

1.4 En lineær regresjonsmodell

En regresjonsmodell er en modell der en ser på sammenhengen mellom en respons og faktorer som kan påvirke eller forklare verdien til denne responsen. En lineær regresjonsmodell er en modell der responsen og parametrene til forklaringsvariablene er bundet sammen i et lineært forhold [16]. Modellen kan skrives på formen:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ \dots \ + \beta_p X_{ip} + \varepsilon_i \qquad (1\text{–}2)$$

der $Y_1, \dots, Y_n$ er uavhengige stokastiske variabler, og $X_{i1}, \dots, X_{ip}$er forklaringsvariabler der $i = 1, \dots, n$. $\beta_0, \dots, \beta_p$ er de ukjente konstantene som en kaller parametre. ε_i angir tilfeldig feil.

Modellen kan også skrives på vektorform som:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (1\text{–}3)$$

Der $\boldsymbol{Y}$ er en $n \times 1$ vektor av observasjonene til den avhengige variabelen, $\boldsymbol{\beta}$ er en $p \times 1$ vektor av parametre og $\boldsymbol{\varepsilon}$ er en vektor av feilledd som er normalfordelt med forventning 0 og varians $\sigma^2\boldsymbol{I}$. Med denne modellen er forventningen til $\boldsymbol{Y}$ lik $\boldsymbol{X\beta}$.
I designmatrisen $\boldsymbol{X}$ er forklaringsvariablene samlet.

Vi har forklaringsvariabler x_{ij} $(i = 1, 2, \dots, n \text{ og } j = 1, 2, \dots, p)$

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix}$$

For å kunne estimere de ukjente parametrene må tallet på observasjoner være større enn tallet på forklaringsvariabler, og dersom dette ikke er oppfylt vil vi få en overparametrisert modell. I en slik overparametrisert modell vil en ha flere parametre enn ligninger og det vil dermed være uendelig mange løsninger for de ukjente parametrene.

Forventningen til responsen er gitt ved:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \ \dots \ + \beta_p X_{ip} = \mu_i$$

Ved å tilpasse en regresjonsmodell på denne formen til et datasett, kan en finne sammenhengen mellom en respons og ulike forklaringsvariabler. Modellen kan også benyttes til å predikere fremtidige verdier til responsen basert på de estimerte parameterverdiene og observasjoner for forklaringsvariablene.

1.5 Generaliserte lineære regresjonsmodeller (GLM)

Den lineære regresjonsmodellen kan utvides til å kunne modellere responsvariabler med fordelinger fra hele den eksponentielle familie. Dette omtales som en *generalisert lineær regresjonsmodell*, GLM.
Dobson [17] definerer at en fordeling tilhører den eksponentielle familie om tettheten til Y kan skrives på formen:

$$f(y;\theta) = \exp\{a(y)\cdot b(\theta) + c(\theta) + d(y)\}$$

der θ er parameteren i fordelingen. Om $a(y) = y$ er fordelingen på kanonisk form (standard form). Eksempler på fordelinger som tilhører den eksponentielle familie er normalfordelingen, poissonfordelingen, gammafordelingen og binomialfordelingen.

Følgende egenskaper kjennetegner en generalisert lineær modell:

- Responsvariablene $Y_1, \dots, Y_n$ kommer fra den samme eksponentielle klasse, har samme fordeling, og de antas å være uavhengige.
- Vektorer av forklaringsvariabler x_{ij} $(i = 1, 2, \dots, n \text{ og } j = 1, 2, \dots, p)$ som kan samles i en designmatrise $\boldsymbol{X}$ som omtalt i forrige avsnitt. Forklaringsvariablene kan være målte verdier av kontinuerlige forklaringsvariabler, eller nivåer av kategoriske forklaringsvariabler, også omtalt som *dummyvariabler*. Dette begrepet kommer jeg tilbake til senere.
- Lineære regresjonskomponenter (prediktorer) kan uttrykkes på formen: $\eta_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$
- Monoton linkfunksjon $g(\,)$ som kobler forventningen til en lineær komponent ved at: $g(\mu_i) = \eta_i$, der $\mu_i = E(Y_i)$

En generalisert lineær modell kan binde en funksjon g av forventningen til forklaringsvariablene. En kaller gjerne linkfunksjonen for bindeleddet mellom den lineære prediktoren η_i og forventningen μ_i. Linkfunksjonen er ofte ikke-lineær, og det kreves at den er deriverbar. GLM består altså av responsen med fordeling fra eksponentiellfamilien, en linkfunksjon g og forklaringsvariabler.

1.6 Poissonregresjon er en GLM

Poissonfordelingen er en diskret sannsynlighetsfordeling som anvendes for å beskrive hendelser som inntreffer helt uavhengig av hverandre. Poissonprosess er en heltallsverdi og stokastisk prosess i kontinuerlig tid som anvendes for å beskrive tilfeldige hendelser som skjer med en viss intensitet. Oversikter over slike hendelser samles ofte i tabeller og man kaller det *telledata*.

Poissonfordelingen har punktsannsynligheten:

$$f(y;\lambda) = \begin{cases} \dfrac{e^{-\lambda}\lambda^{y}}{y!}, & y = 0, 1, 2 \ldots \\ 0, & \text{ellers} \end{cases}$$

For poissonfordelingen er $E(Y) = \mathrm{Var}(Y) = \lambda > 0$

I dette tilfellet har den generaliserte lineære modellen egenskapene:

- Responsene er poissonfordelt $Y_i \sim Po(\mu_i)$
- Lineær komponent $\eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$
- $E(Y_i) = \mu_i = \exp(\eta_i)$, dvs. linkfunksjonen $g(\mu_i) = \log(\mu_i)$ er (den naturlige) logaritmefunksjonen

Forventningen til den poissonfordelte tilfeldige variabelen kan skrives som:

$$E(Y_i) = \mu_i = e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}}$$

Modeller av denne typen er allment brukt i demografisk og epidemiologisk forskning, hvor antallet av hendelser slik som død eller sykdomsinsidens generelt følger en poissonfordeling. Ratene blir da estimert gjennom *log-lineære modeller*, som er generaliserte lineære modeller hvor responsvariabelen antas å være poissonfordelt og logaritmen er linkfunksjonen.
Ligning (1–1) kan ta en log-lineær regresjonsform via en log-link, og skrives på formen:

$$\log(E_{ij}) = \log\left(P_{ij}\right) + \mu + \alpha_i + \beta_j + \gamma_k$$

Her angir E_{ij} forventningen til antall hendelser i celle (i, j) som er antatt å være poissonfordelt. $\mathrm{Log}\left(P_{ij}\right)$ er logaritmen til eksponeringen P_{ij} i modellen og blir omtalt som offset ledd, og regnes som en kjent konstant.
For eksempel vil en demograf modellere dødsrater i et geografisk område som antallet av døde individer delt på personår. I dette tilfellet vil eksponeringen være personår, og i poissonregresjon vil dette bli behandlet som et offset-ledd, hvor eksponeringsvariabelen vil komme inn i ligningen på høyre side.

2. Alder-periode-kohort-analyse

I dette kapittelet gis en kort innføring i en del av teorien for APC-analyser som vil være sentral i etterfølgende kapitler, og eksempel på hvordan en APC-analyse kan settes opp. Uavhengig av hvilken metode som velges for å analysere data, trenger man kode om variablene og innføre ulike parametriseringer. Teori om generaliserte inverser og Moore-Penrose-inversen er presentert på slutten av kapittelet.

2.1 Dummykoding

For kategoriske forklaringsvariabler er det parametere for ulike nivå av en faktor. De korresponderende elementene i en designmatrise er valgt for å ekskludere eller inkludere de hensiktsmessige parametrene for hver observasjon; de kalles *dummyvariabler*. Om de består kun av 0-er og 1-ere, kan begrepet indikatorvariabler benyttes. Dummykoding gir oss en måte å benytte kategoriske prediktorvariabler i ulike typer av estimeringsmodeller, slik som lineær regresjon. Dummykoding bruker bare 0-er og 1-ere til å formidle all den nødvendige informasjonen til en gruppe. Et enkelt eksempel illustrerer dette med 4 observasjoner i hver av 4 grupper.

	Observasjoner for en gitt respons			
	Gruppe1	**Gruppe2**	**Gruppe3**	**Gruppe4**
	1	2	5	10
	3	3	6	10
	2	4	4	9
	2	3	5	11
Mean	2	3	5	10

I dette eksempelet vil en trenge å lage 3 dummykodede variabler. Generelt, med k grupper, vil det bli $k-1$ kodede variabler. Dummyvariablene som lages kalles $d1$, $d2$ og $d3$. For $d1$ vil hver observasjon i gruppe 1 bli kodet som 1, mens for de andre gruppene kodes det 0. Deretter kodes $d2$ med 1 hvis observasjonen er i gruppe 2, og 0 ellers. Tilsvarende gjøres for gruppe 3. $d4$ trengs ikke fordi de andre dummyvariablene har all informasjonen som trengs for å bestemme hvilken observasjon som er i hvilken gruppe.

Slik kan dataene arrangeres for bruk i en regresjonsprosedyre.

Respons y	Gruppe	d1	d2	d3
1	1	1	0	0
3	1	1	0	0
2	1	1	0	0
2	1	1	0	0
2	2	0	1	0
3	2	0	1	0
4	2	0	1	0
3	2	0	1	0
5	3	0	0	1
6	3	0	0	1
4	3	0	0	1
5	3	0	0	1
10	4	0	0	0
10	4	0	0	0
9	4	0	0	0
11	4	0	0	0

Hver av gruppene er definert ved å ha 1-ere for dummyvariabelen som er lik gruppen, med unntak av en gruppe som har kun 0-er. Gruppen med kun 0-er er referansegruppen, som i dette eksempelet er gruppe 4. Når det så gjøres en regresjonsanalyse, vil en med dummykoding få et konstantledd som er lik gjennomsnittet til referansegruppen (mean=10). Koeffisientene til hver av dummyvariablene er lik forskjellen mellom mean til gruppen som er kodet 1 og mean til referansegruppen. I det enkle eksempelet vist over, har mean til gruppe 1 verdien 2, og differansen blir -8, som er regresjonskoeffisienten en får for $d1$.

Et alternativ til dummykoding er effektkoding, og denne kodingen er ganske lik til dummykodingen. Forskjellen ligger i at der referansegruppen kodes med 0 for dummykodingen, vil en benytte seg av en gruppe som kodes -1 ved effektkoding. Ved effektkoding vil konstantleddet være lik samlet mean, og koeffisientene er forskjellen mellom en gitt gruppe og samlet mean.

2.2 APC-analyse

For illustrative hensikter er spesialtilfellet med $a = 3$ og $p = 4$ vist i Tabell 2-1 under. Eksempelet er hentet fra Kupper et al. [11]. I tabellen er en typisk celleoppføring (i, j) gitt ved:

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{a-i+j} \qquad (2\text{–}1)$$

Tabell 2-1: Skjematisk presentasjon av modellen for spesialtilfellet $a = 3, p = 4$.

		Periode *(j)*			
		$j = 1$	$j = 2$	$j = 3$	$j = 4$
Aldersgruppe *(i)*	$i = 1$	$\mu + \alpha_1 + \beta_1 + \gamma_3$	$\mu + \alpha_1 + \beta_2 + \gamma_4$	$\mu + \alpha_1 + \beta_3 + \gamma_5$	$\mu + \alpha_1 + \beta_4 + \gamma_6$
	$i = 2$	$\mu + \alpha_2 + \beta_1 + \gamma_2$	$\mu + \alpha_2 + \beta_2 + \gamma_3$	$\mu + \alpha_2 + \beta_3 + \gamma_4$	$\mu + \alpha_2 + \beta_4 + \gamma_5$
	$i = 3$	$\mu + \alpha_3 + \beta_1 + \gamma_1$	$\mu + \alpha_3 + \beta_2 + \gamma_2$	$\mu + \alpha_3 + \beta_3 + \gamma_3$	$\mu + \alpha_3 + \beta_4 + \gamma_4$

Tabellen viser mønsteret til effektene til de uavhengige variablene i en standard APC-analyse, der μ angir konstantleddet, α_i angir alderseffekten, β_j angir periodeeffekten og γ_{a-i+j} angir kohorteffekten. Aldersgruppene kodes med dummyvariabler langs radene, periodene kodes med dummyvariabler for kolonnene. Kohortene kodes med dummyvariabler langs diagonalen fra nedre venstre celle til øvre høyre celle. Cellene representerer forventede verdier for den alder-periode-spesifikke avhengige variabelen i ligning (2–1). Effekten for rad i og effekten for kolonne j bidrar additivt til responsvariabelen i en celle (i, j). I tillegg bidrar den gitte effekten for en spesifikk kohort additivt til denne verdien.

Den lineære modellen $Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij}$ kan skrives på vektorform som:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

og tilsvarende får vi at ligning (2–1) kan skrives på formen:

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} \qquad (2\text{–}2)$$

hvor $\boldsymbol{Y}$ er vektoren med dødelighetsratene eller de log-transformerte ratene. Designmatrisen $\boldsymbol{X}$ består av dummyvariabler, som markerer med 1-ere og 0-er hvilken gruppe observasjonen tilhører, og $\boldsymbol{\beta}$ er vektoren med modellparametre (og må ikke forveksles med periodeparameteren β_j).

Fra modellen gitt ved ligning (2–2) med tilfellet fra Tabell 2-1 har vi at:

$$\boldsymbol{Y^T} = (Y_{11}, Y_{12}, Y_{13}, Y_{14}; Y_{21}, Y_{22}, Y_{23}, Y_{24}; Y_{31}, Y_{32}, Y_{33}, Y_{34})$$

og

$$\boldsymbol{\beta^T} = (\mu; \alpha_1, \alpha_2, \alpha_3; \beta_1, \beta_2, \beta_3, \beta_4; \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6)$$

Det følger da at designmatrisen $\boldsymbol{X}$ kan skrives som en (12×14) matrise med følgende struktur:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Matrisen presentert i tabellform:

Tabell 2-2: Eksempel på oppsett av en APC-modell.

Respons	alder	periode	kohort	intercept	ald1	ald2	ald3	per1	per2	per3	per4	koh1	koh2	koh3	koh4	koh5	koh6
Y_{11}	1	1	3	1	1	0	0	1	0	0	0	0	0	1	0	0	0
Y_{12}	1	2	4	1	1	0	0	0	1	0	0	0	0	0	1	0	0
Y_{13}	1	3	5	1	1	0	0	0	0	1	0	0	0	0	0	1	0
Y_{14}	1	4	6	1	1	0	0	0	0	0	1	0	0	0	0	0	1
Y_{21}	2	1	2	1	0	1	0	1	0	0	0	0	1	0	0	0	0
Y_{22}	2	2	3	1	0	1	0	0	1	0	0	0	0	1	0	0	0
Y_{23}	2	3	4	1	0	1	0	0	0	1	0	0	0	0	1	0	0
Y_{24}	2	4	5	1	0	1	0	0	0	0	1	0	0	0	0	1	0
Y_{31}	3	1	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0
Y_{32}	3	2	2	1	0	0	1	0	1	0	0	0	1	0	0	0	0
Y_{33}	3	3	3	1	0	0	1	0	0	1	0	0	0	1	0	0	0
Y_{34}	3	4	4	1	0	0	1	0	0	0	1	0	0	0	1	0	0

Den første linjen i Tabell 2-2 korresponderer med $E(Y_{11}) = 1 \cdot \mu + 1 \cdot \alpha_1 + 1 \cdot \beta_1 + 1 \cdot \gamma_3$. Tilsvarende korresponderer den siste linjen med $E(Y_{34}) = 1 \cdot \mu + 1 \cdot \alpha_3 + 1 \cdot \beta_4 + 1 \cdot \gamma_4$.

Den ordinære minste kvadraters estimatoren til regresjonsmodellen i ligning (2–2) er løsningen $\widehat{\boldsymbol{\beta}}$ på normalligningen:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \qquad (2\text{–}3)$$

For å finne parameterestimatene minimeres kvadratsummen av differansen mellom observasjonene og de forventede verdier. Minste kvadraters metoden (OLS) er beskrevet nærmere i Kapittel 9. Minste kvadraters metoden gir unike løsninger for regresjonskoeffisientene når designmatrisen for de uavhengige variablene er av full

rang. $\widehat{\boldsymbol{\beta}}$ er da en kolonne vektor med unike regresjonskoeffisienter for de uavhengige variablene, vanligvis med det første elementet som angir intercept.
Rangen til en matrise er det største antallet lineært uavhengige rader eller kolonner i matrisen. På grunn av det lineære forholdet mellom alder-, periode- og kohorteffekter, er designmatrisen $\boldsymbol{X}$ singulær med en mindre enn full kolonnerang [11]. Dette betyr at noen av kolonnene i $\boldsymbol{X}$ kan danne lineære kombinasjoner eller summeres for å danne en kolonne identisk til andre kolonner i $\boldsymbol{X}$**.** Derfor eksisterer det ikke en regulær invers, $(\boldsymbol{X^T X})^{\boldsymbol{-1}}$, og estimatoren $\widehat{\boldsymbol{\beta}}$ eksisterer ikke. Det finnes ikke en unikt definert vektor med koeffisientestimat. Det medfører at det er et uendelig antall av mulige løsninger til ligningen (2–3). Det er derfor ikke mulig å separat estimere effektene til alder, periode og kohort uten å innføre visse begrensninger på koeffisientene. Dette er modellidentifikasjonsproblemet til APC-analyser. For å finne en unik løsning til de individuelle forklaringsvariablene må en lineær begrensning velges.

2.3 Parametrisering

Eksempel på en enkel modell er gitt i Kapittel 2.4 i Dobson [17]:

$$E(Y_{1j}) = \beta_1$$
$$E(Y_{2j}) = \beta_2$$

der $j = 1, \dots, J$. Her er det 2 parametre som skal estimeres fra 2 sett av observasjoner.

Dette kan bli skrevet på formen til ligning (2–2) som:

$$\begin{bmatrix} E(Y_{11}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{1J}) \\ E(Y_{21}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{2J}) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 0 \\ 0 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

En annen modell har formen:

$$E(Y_{1j}) = \mu + \alpha_1$$
$$E(Y_{2j}) = \mu + \alpha_2$$

der $j = 1, \dots, J$. I denne modellen er μ det samlede gjennomsnitt, og α_1 og α_2 representerer gruppedifferansene fra μ.

Dette kan bli skrevet på formen til ligning (2–2) som:

$$\begin{bmatrix} E(Y_{11}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{1j}) \\ E(Y_{21}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{2j}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

I dette oppsettet er det 3 parametre som skal estimeres fra 2 sett av observasjoner. Det er for mange parametre og vi får en overparametrisert modell. Derfor trengs det en modifikasjon av modellen.
For å løse ubestemthetsproblemet, og få identifiserbare parametre, legger vi inn restriksjoner. Restriksjonene består i å enten øke antall ligninger eller å redusere antall parametre. De vanligste restriksjonene som benyttes er *hjørnepunkts-restriksjoner* og *sum-lik-null-restriksjoner*.

2.4 Hjørnepunktsparametrisering

Med denne parametriseringen velger man en gruppe som referansegruppe, ofte gruppe 1, men man kan også velge en annen gruppe som referansegruppe [17]. Det er det samme hvilken variabel som brukes som referanse, men det er likevel mest hensiktsmessig å velge den verdien en ønsker å måle de andre opp mot. Dersom gruppe 1 blir referansegruppe setter vi $\alpha_1 = 0$, eller generelt med gruppe a som referansegruppe setter vi $\alpha_a = 0$. De andre gruppene blir da sammenlignet med referansegruppen. Med denne restriksjonen blir antallet parametre redusert og vi får entydige løsninger av ligningene og dermed får vi identifiserbare parametre.

Forventningen kan skrives som $\mu + \alpha_i$ med restriksjonen $\alpha_1 = 0$
Eksempel på en slik modell:

$$E(Y_{1j}) = \mu + \alpha_1 = \mu$$
$$E(Y_{2j}) = \mu + \alpha_2$$

der $j = 1, \dots, J$. Her er gruppe 1 satt som referansekategori, og α_2 representerer tilleggseffekten til gruppe 2.

Denne modellen kan bli skrevet på formen til ligning (2–2) som:

$$\begin{bmatrix} E(Y_{11}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{1J}) \\ E(Y_{21}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{2J}) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 0 \\ 1 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} \mu \\ \alpha_2 \end{bmatrix}$$

I hjørnepunktsparametrisering er gruppeeffektene definert som differanser fra en referansekategori som kalles *hjørnepunktet*. Restriksjonen kalles ofte treatment-kontrast. Hjørnepunktsparametrisering/treatment kontrast er default i R.

2.5 Sum-lik-null parametrisering

I denne parametriseringen settes summen av parametrene til en faktor lik null, slik at vi får innført en ekstra ligning: $\sum_{i=1}^{I} \alpha_i = 0$. En får da like mange ligninger som parametre, og det er dermed mulig å finne entydige løsninger for å identifisere parametrene.

Eksempel på en slik modell:

$$E(Y_{1j}) = \mu + \alpha$$
$$E(Y_{2j}) = \mu - \alpha$$

der $j = 1, \dots, J$. Denne versjonen behandler de to gruppene symmetrisk, og μ er den samlede gjennomsnittseffekten og α representerer gruppedifferansene. I dette tilfellet er restriksjonen $\alpha_1 + \alpha_2 = 0$, som er ekvivalent med $\alpha_2 = -\alpha_1$.

Denne modellen kan bli skrevet på formen til ligning (2–2) som:

$$\begin{bmatrix} E(Y_{11}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{1J}) \\ E(Y_{21}) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_{2J}) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 1 \\ 1 & -1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & -1 \end{bmatrix} \times \begin{bmatrix} \mu \\ \alpha_1 \end{bmatrix}$$

2.6 Generalisert invers

En invers til en kvadratisk matrise $\boldsymbol{A}$ er definert som den entydig bestemte matrisen $\boldsymbol{A}^{-1}$ som oppfyller ligningene $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, der $\boldsymbol{I}$ er identitetsmatrisen [18]. Matrisen $\boldsymbol{A}$ er invertibel hvis determinanten til $\boldsymbol{A}$ er ulik null: det($\boldsymbol{A}$) ≠ 0. I motsatt fall er matrisen singulær.

En generalisert invers $\boldsymbol{G}$ til en generell matrise $\boldsymbol{A}$ er en matrise som oppfyller ligningen:

$$\boldsymbol{AGA} = \boldsymbol{A} \qquad (2\text{–}4)$$

En kan finne generaliserte inverser [19] til alle matriser selv om de er singulære eller rektangulære. Den generaliserte inversen er generelt ikke entydig. Dersom $\boldsymbol{A}$ er en kvadratisk $(n \times n)$-matrise med rang n, så er $\boldsymbol{G} = \boldsymbol{A}^{-1}$.

Den generaliserte inversen $\boldsymbol{G}$ for en gitt matrise $\boldsymbol{A}$ er ikke unik. Det er uendelig mange matriser som oppfyller $\boldsymbol{AGA} = \boldsymbol{A}$, og derfor blir hele gruppen av disse matrisene referert til som generaliserte inverser av $\boldsymbol{A}$.

Matrisen $\boldsymbol{A}$ kan skrives på en ekvivalent diagonalform:

$$\boldsymbol{PAQ} = \boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{D_r} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

$\boldsymbol{P}$ og $\boldsymbol{Q}$ er produkt fra elementære operatører, r er rangen til matrisen $\boldsymbol{A}$, og $\boldsymbol{D_r}$ er diagonalmatrisen av orden r. En benytter seg av egenverdier og egenvektorer for å diagonalisere matrisen.
Matrisen $\boldsymbol{A}$ kan da skrives på formen $\boldsymbol{A} = \boldsymbol{P^{-1}\Delta Q^{-1}}$

$\boldsymbol{\Delta^-}$ kan defineres ved $\boldsymbol{\Delta^-} = \begin{bmatrix} \boldsymbol{D_r^{-1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$, og da kan $\boldsymbol{G}$ uttrykkes ved:

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

$\boldsymbol{G}$ er en generalisert invers av $\boldsymbol{A}$. På grunn av at $\boldsymbol{P}$, $\boldsymbol{Q}$ og $\boldsymbol{\Delta^-}$ ikke er entydige, er heller ikke $\boldsymbol{G} = \boldsymbol{Q\Delta^- P}$ entydig.

Ved hjelp av disse uttrykkene kan vi vise at ligning (2–4) er oppfylt og at $\boldsymbol{G}$ derfor er en generalisert invers av $\boldsymbol{A}$:

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

Et eksempel:

$$\boldsymbol{A} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} \qquad \boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 0 \\ -\frac{2}{3} & -\frac{1}{3} & 1 \end{bmatrix} \qquad \boldsymbol{Q} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{PAQ} = \boldsymbol{\Delta} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{\Delta}^{-} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{G}$ er en generalisert invers til $\boldsymbol{A}$.

$$\boldsymbol{G} = \boldsymbol{Q\Delta}^{-}\boldsymbol{P} = \frac{\mathbf{1}}{\mathbf{3}}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{AGA} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} \times \frac{\mathbf{1}}{\mathbf{3}}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \end{bmatrix} \times \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} = \boldsymbol{A}$$

Vi ser at ligning (2–4) er oppfylt.

2.7 Moore-Penrose-inversen

Det finnes mange generaliserte inverser til en matrise $\boldsymbol{A}$, men for å finne en entydig matrise $\boldsymbol{K}$ benyttes Moore-Penrose-inversen [19]. For å være en Moore-Penrose-invers må en matrise oppfylle 4 ulike betingelser.

1. $\boldsymbol{AKA} = \boldsymbol{A}$
2. $\boldsymbol{KAK} = \boldsymbol{K}$
3. $(\boldsymbol{KA})' = \boldsymbol{KA}$
4. $(\boldsymbol{AK})' = \boldsymbol{AK}$

Moore og Penrose reduserte et uendelig antall av generaliserte inverser til en unik løsning ved å innføre 4 rimelige algebraiske betingelser. En vanlig bruk av Moore-Penrose-inversen er å beregne minste kvadraters løsningen til et sett med lineære ligninger som mangler en unik løsning. Moore-Penrose-inversen er definert og unik for alle matriser som inneholder reelle eller komplekse tall. Moore-Penrose-inversen kan beregnes ved å benytte singulær verdi dekomposisjon (SVD).

For talleksempelet i forrige avsnitt kan beregninger for SVD utføres i R, samt kontrollregninger for betingelsene. Moore-Penrose-inversen for matrise $\boldsymbol{A}$ er gitt ved:

$$\boldsymbol{K} = \frac{\mathbf{1}}{\mathbf{532}}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix}$$

Ved å kombinere 1) og 3) får vi:

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

som er ekvivalent med:

$$\boldsymbol{KAA'} = \boldsymbol{A'}$$

På samme måte får vi fra 2) og 4):

$$\boldsymbol{KK'A'} = \boldsymbol{K}$$

Fra talleksempelet har vi:

$$\boldsymbol{KAA'} = \frac{\mathbf{1}}{\mathbf{532}}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix} \times \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix} \times \begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix}$$

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

og tilsvarende:

$$\boldsymbol{KK'A'} = \frac{\mathbf{1}}{\mathbf{532}}\begin{bmatrix} 112 & -77 & 49 \\ 12 & 6 & 10 \\ -40 & 113 & 11 \end{bmatrix} \times \frac{\mathbf{1}}{\mathbf{532}}\begin{bmatrix} 112 & 12 & -40 \\ -77 & 6 & 113 \\ 49 & 10 & 11 \end{bmatrix} \times \begin{bmatrix} 4 & 1 & 3 \\ 1 & 1 & 1 \\ 2 & 5 & 3 \end{bmatrix}$$

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

Betingelsene 1) – 4) er oppfylt og dette er Moore-Penrose-inversen i det gitte eksempelet.

3. To APC-metoder: CGLIM og IE

Den konvensjonelle metoden (CGLIM) innen APC-analyse er beskrevet kort innledningsvis, og deretter presenteres teori og algoritmen for anvendelse av den nyere metoden, intrinsic estimator (IE).

3.1 Constrained generalized linear models estimator (CGLIM)

I alder-periode-kohort-modeller har vi $a + p + k$ parametre i tillegg til parameteren for intercept-leddet som skal estimeres. Det er vanlig å innføre restriksjoner enten ved å sette en av hver av kategoriene som referansegruppe (hjørnepunktsparametrisering), eller benytte seg av sum-lik-null restriksjonene $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0$. Med denne parametriseringen av APC-modellen vil en ha en parameter for mye i forhold til det som lar seg estimere fra data. Matrisen $\boldsymbol{X^T X}$ vil være singulær, og dermed eksisterer det ikke en invers av matrisen. Det vil ikke være en entydig løsning av ligningene og uendelig mange løsninger vil gi like god tilpasning til dataene. Problemet er som tidligere nevnt, at de tre tidsvariablene er lineært avhengige av hverandre. En av tidsvariablene kan uttrykkes ved hjelp av de to andre, og en vil få en uavhengig ligning mindre enn om ikke tidsvariablene hadde vært lineært avhengige.

Den konvensjonelle tilnærmingen i demografi for å løse modellidentifikasjonsproblemet i APC-modeller har vært å benytte seg av CGLIM-metoden [9]. CGLIM står for *constrained generalized linear models estimator*, og metoden innfører en eller flere ekstra betingelser i modellen. I CGLIM settes det krav til betingelser ved at to effektkoeffisienter settes lik hverandre. For eksempel kan de to første periodeeffektene settes lik hverandre og vi får da: $\beta_1 = \beta_2$. Med denne tilleggsbetingelsen blir matrisen $(\boldsymbol{X^T X})$ ikke-singulær og minste kvadratersestimatoren eksisterer.

Hovedproblemet med denne metoden er at den avhenger av forhåndsinformasjon om dataene for å sette disse betingelsene. Estimatene til effektkoeffisientene er sensitiv for valget av identifiserbare betingelser [14], og en bør derfor være forsiktig med modellspesifikasjonene. Valget av betingelse for modellidentifisering må basere seg på tidligere teoretisk eller empirisk informasjon, noe som dessverre sjelden eksisterer. Analytikeren må stole på informasjon som neppe eksisterer eller kan verifiseres. Ulike valg av identifiserende betingelser kan produsere vidt forskjellige estimater for trendene til alder-, periode- og kohortkategoriene. Alle CGLIM-modellene vil produsere samme nivå av goodness-of-fit-data, og dermed gjøre det umulig å benytte seg av modelltilpasning som kriterie for å velge den beste betingede metoden. De ulike CGLIM-modellene vil gi identiske mål for f.eks. devians og AIC.
Ellers i statistikken kan en ofte benytte seg av ulike goodness-of-fit-mål for å skille mellom hvor godt ulike modeller tilpasser et datasett. Noen vanlige goodness-of-fit-mål er beskrevet i Kapittel 6.4.

3.2 Intrinsic estimator (IE)

Nyere utvikling i APC-metodikken har lagt vekt på utnyttingen av estimerbare funksjoner som er upåvirket av valget av betingelse for parametrene. Et lovende alternativ til de konvensjonelle metodene ble beskrevet og evaluert av Yang et al. [14]. Forfatterne tok utgangspunkt i intrinsic estimator som Fu skrev om i sin artikkel [13], og studerte denne estimatoren som den unike estimerbare funksjonen for den multiple APC-klassifikasjonsmodellen i ligning (1–1). I dette kapittelet har jeg benyttet tilsvarende symboler som Yang et al.
Ved å innlemme nyere metodisk utvikling innen estimerbare funksjoner i biostatistikk, ble en ny metode utviklet for estimering av de simultane effektene til alder, periode og kohort. Metoden blir omtalt som *intrinsic estimator* (IE) og den gir en unik løsning til modellen:

$$\boldsymbol{Y} = \boldsymbol{Xb} \qquad (3\text{–}1)$$

som blir bestemt av Moore-Penrose-inversen. Metoden fjerner tilfeldigheten til de lineære betingelsene på parametrene. Denne løsningen er den eneste estimerbare funksjonen til parametervektoren, og den estimerer både lineære og ikke-lineære komponenter [20]. IE-metoden oppnår modellidentifikasjon med minimale antagelser. Hovedideen til IE er å fjerne påvirkningen fra designmatrisen på koeffisientestimatene. Designmatrisen er fastsatt av antall alders- og periodegrupper, og er ikke relatert til responsen, Y_{ij}. IE-metoden produserer estimater som har ønskede statistiske egenskaper.

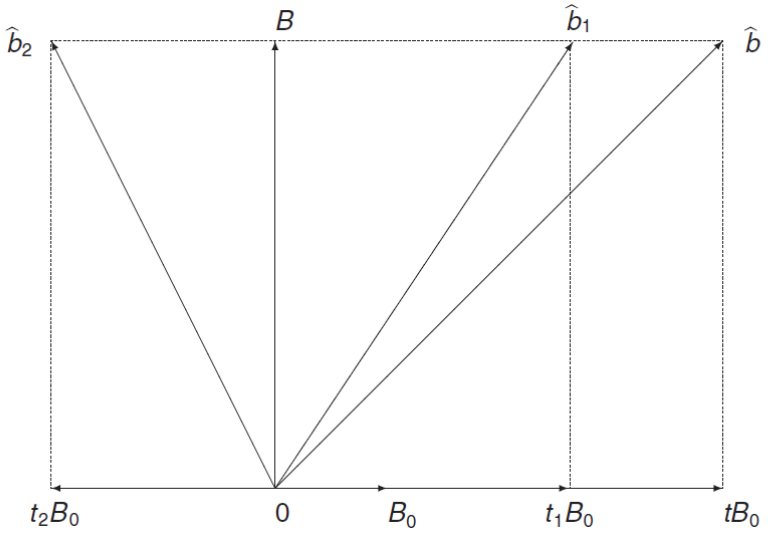
På grunn av at det er en lineær avhengighet mellom alder, periode og kohort i ligning (3–1), eksisterer det en ikke-null vektor $\boldsymbol{B_0}$, en lineær funksjon av designmatrisen $\boldsymbol{X}$, slik at produktet av designmatrisen og vektoren er lik 0:

$$\boldsymbol{XB_0} = 0 \qquad (3\text{–}2)$$

Det eksisterer en lineær kombinasjon av kolonnene i designmatrisen $\boldsymbol{X}$ som er lik 0-vektoren, og $\boldsymbol{X}$ er singulær. $\boldsymbol{B_0}$ representerer nullrommet til designmatrisen $\boldsymbol{X}$. På grunn av at designmatrisen $\boldsymbol{X}$ er en mindre enn full kolonnerang, kan parameterrommet til den ubetingede APC-modellen dekomponeres i to lineære underrom som står vinkelrett på hverandre. Det ene underrommet korresponderer til den unike egenverdien 0 til matrisen $\boldsymbol{X^T X}$ og har dimensjon lik 1. Det kalles for nullrommet til designmatrisen. Det andre ikke-null underrommet er det komplementere underrommet ortogonalt til nullrommet. Den ortogonale dekomposisjonen til parameterrommet gjør at hver av de uendelige mange løsningene til den ubetingede APC-modellen kan skrives som:

$$\widehat{\boldsymbol{b}} = \boldsymbol{B} + t\boldsymbol{B_0} \qquad (3\text{–}3)$$

Parameteren $\boldsymbol{b}$ fra ligning (3–1) er blitt dekomponert i to ortogonale underrom, hvor t er et reelt tall som korresponderer til en spesifikk løsning. $\boldsymbol{B_0}$ er en unik egenvektor av lengde 1. Egenvektoren $\boldsymbol{B_0}$ avhenger bare av designmatrisen og bestemmes fullstendig av antallet aldersgrupper og periodegrupper. Den avhenger ikke av de observerte ratene $\boldsymbol{Y}$. $\boldsymbol{B}$ korresponderer til projeksjonen av $\widehat{\boldsymbol{b}}$ på ikke-nullrommet til designmatrisen $\boldsymbol{X}$, ortogonalt til nullrommet. Figur 3-1 viser den geometriske representasjonen av sammenhengene.



Figur 3-1: Projeksjon av estimatoren $\widehat{\boldsymbol{b}}$ på den vertikale aksen for å gi estimatoren $\boldsymbol{B}$.

Ulike betingelser benyttet ved CGLIM-estimatorer, slik som $\widehat{\boldsymbol{b}}_1$ og $\widehat{\boldsymbol{b}}_2$ gir ulike verdier av t. $\widehat{\boldsymbol{b}}_1$og $\widehat{\boldsymbol{b}}_2$ er ulike sett av parametervektorer, og er vilkårlige estimatorer. $t\boldsymbol{B_0}$ representerer det vilkårlige uttrykket i estimatoren $\widehat{\boldsymbol{b}}$. Projeksjonen av $\widehat{\boldsymbol{b}}_1$og $\widehat{\boldsymbol{b}}_2$ på den vertikale aksen gir den samme estimerbare funksjonen $\boldsymbol{B}$. Den vertikale aksen er bestemt av ratene og er uavhengig av den vilkårlige t. Den vertikale vektoren $\boldsymbol{B}$ er ortogonal til og derfor uavhengig av komponenten $t\boldsymbol{B_0}$. E($\boldsymbol{B}$) er estimerbar selv om ingen av de andre estimatorene er det.

Med dekomposisjonen i ligning (3–3) kan en vise at ulike estimatorer gir samme tilpassede verdier.

$$\boldsymbol{X}\widehat{\boldsymbol{b}} = \boldsymbol{X}(\boldsymbol{B} + t\boldsymbol{B_0}) = \boldsymbol{XB} + t\boldsymbol{XB_0} = \boldsymbol{XB} + 0 = \boldsymbol{XB}$$

$\boldsymbol{XB_0} = 0$ og det følger at $t\boldsymbol{XB_0} = 0$, slik at ligningen er sann for alle verdier av t. Intrinsic estimator defineres til å være $\boldsymbol{B}$ og den bestemmer unikt parametrene til alder-, periode- og kohorteffektene i parameterrommet som er vinkelrett på nullrommet til den singulære designmatrisen. Tilfeldigheten i de ulike estimatorene $\widehat{\boldsymbol{b}}$ langs nullrommet, $t\boldsymbol{B_0}$, er dermed fjernet.

Intrinsic estimator $\boldsymbol{B}$ er estimerbar på den måten at den er projeksjonen til enhver vilkårlig estimator $\widehat{\boldsymbol{b}}$ som kan fås ved å sette en begrensning på det unike parameterrommet vinkelrett på nullrommet. På denne måten avhenger ikke $\boldsymbol{B}$ av begrensninger som er satt til $\widehat{\boldsymbol{b}}$. Metoden med intrinsic estimator gir en unik måte å få en spesiell begrensning gjennom en estimerbar funksjon.

IE-metoden er en anvendelse av Moore-Penrose generaliserte invers på APC-problemet. Det kan også bli sett på som en spesiell form for prinsipal komponent analyse (PCA), en teknikk som kan brukes til å håndtere identifikasjonsproblem når forklaringsvariabler er sterkt korrelert. Denne teknikken er nærmere omtalt i Kapittel 9. Ved å transformere korrelerte forklaringsvariabler til et sett av ortogonale lineære kombinasjoner av disse variablene, prinsipal komponentene, er PCA et nyttig verktøy for å redusere overflødige data og utvikle prediktive modeller. Hensikten med IE-metoden er hverken å fjerne overflødige data eller prediksjon, men derimot estimere effektene til og fange den generelle trenden til alder, periode og kohort.
Algoritmen som benyttes for å beregne IE-estimatene baserer seg på ortonormal transformasjon av en prinsipal komponent regresjon og inkluderer flere steg.

<u>Algoritmen som benyttes for IE-metoden:</u>

1. Egenvektorene $\boldsymbol{u_1}, \dots, \boldsymbol{u_m}$ til matrisen $(\boldsymbol{X^T X})$ beregnes, hvor m angir antallet av rader/kolonner i matrisen. Antall egenvektorer er gitt ved $m = 1 + (a-1) + (p-1) + (a+p-2)$. Egenvektorene normaliseres til å ha unit lengde med $\|u_m\|$. Den ortonormale matrisen $\boldsymbol{U}$ er gitt ved $\boldsymbol{U} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_m})^{\boldsymbol{T}}$.

2. Den spesielle egenvektoren som korresponderer til den unike egenverdien 0 identifiseres, $\boldsymbol{B_0}$. Denne betegnes som $\boldsymbol{u_1}$, $(\boldsymbol{u_1} = \boldsymbol{B_0})$.

3. Prinsipal komponentene velges ut til å være alle egenvektorene som ikke har egenverdi 0, dvs. $\boldsymbol{u_2}, \dots, \boldsymbol{u_m}$. Prinsipal komponentene utgjør kolonnevektorene i designmatrisen $\boldsymbol{V}$, $\boldsymbol{V} = (\boldsymbol{u_2}, \dots, \boldsymbol{u_m})$.

4. En prinsipal komponent regresjon utføres med den avhengige variabelen som respons, der designmatrisen $\boldsymbol{V}$ benyttes for å få koeffisientene $(\boldsymbol{w_2}, \dots, \boldsymbol{w_m})$.

5. Koeffisienten $\boldsymbol{w_1}$ settes til 0, $(\boldsymbol{w_1} = 0)$. Koeffisientvektoren $\boldsymbol{w} = (\boldsymbol{w_1}, \dots, \boldsymbol{w_m})^{\boldsymbol{T}}$ transformeres med den ortonormale matrisen av alle egenvektorene $\boldsymbol{U} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_m})^{\boldsymbol{T}}$, for å omforme koeffisientene fra prinsipal komponent regresjonen til intrinsic estimatoren $\boldsymbol{B} = \boldsymbol{Uw}$.

I stedet for å bruke referansekategorier benytter IE-metoden seg av at summene til de respektive alder-, periode- og kohortkoeffisientene er lik null, også omtalt som effektkoding. Algoritmen som benyttes for IE-metoden, beregner estimatene for effektkoeffisienter for $a-1, p-1$ og $a+p-2$ kategorier for henholdsvis alder, periode og kohort. Deretter benytter IE-metoden sum-lik-null parametrisering for å få de numeriske verdiene for de utelatte alder-, periode- og kohortkategoriene.

IE-metoden kan bli sett på som en spesial variant av prinsipal komponent analysen, som fjerner påvirkningen av nullrommet til designmatrisen $\boldsymbol{X}$ på estimatorene.
I prosedyren til IE-metoden benyttes prinsipal komponent analysen til å finne prinsipal komponentene til variablene. Det benyttes et ekstra steg med invers ortonormal transformasjon av koeffisientestimatene til prinsipal komponent regresjonen for å komme tilbake til det originale rommet med alder-, periode- og kohortkoordinatene. Det er den inverse transformasjonen som gjør IE-metoden til et spesialtilfelle av prinsipal komponent estimatoren. Den vil gi koeffisienter som er direkte fortolkbare som alder-, periode- og kohorteffekter og som kan bli sammenlignet med korresponderende effekter som er estimert ved de konvensjonelle metodene.

IE-metoden benytter seg av at det settes en spesiell betingelse for modellparametrene, men denne betingelsen er ikke avhengig av en analytikers personlige oppfatning. Basert på estimerbare funksjoner og singulær verdi dekomposisjonen av matriser, gir IE robuste estimater for trender til alder, periode og kohort, og bestemmer unikt de ulike koeffisientestimatene.

4. Praktisk bruk av APC-metoder

4.1 Et enkelt eksempel

For å illustrere bruken til noen av metodene som benyttes til analyse av alder-periode-kohort-modeller tas det utgangspunkt i et enkelt eksempel. I dette eksempelet er det med 3 aldersgrupper og 3 perioder, og vi får da 5 kohorter. Tabellen under viser hvordan en slik modell kan settes opp skjematisk.

Tabell 4-1: Skjematisk presentasjon av en alder-periode-kohort-modell for spesialtilfellet $a = 3, p = 3$.

		Periode *(j)*		
		j = 1	*j = 2*	*j = 3*
Aldersgruppe *(i)*	i = 1	$\mu + \alpha_1 + \pi_1 + \chi_3$	$\mu + \alpha_1 + \pi_2 + \chi_4$	$\mu + \alpha_1 + \pi_3 + \chi_5$
	i = 2	$\mu + \alpha_2 + \pi_1 + \chi_2$	$\mu + \alpha_2 + \pi_2 + \chi_3$	$\mu + \alpha_2 + \pi_3 + \chi_4$
	i = 3	$\mu + \alpha_3 + \pi_1 + \chi_1$	$\mu + \alpha_3 + \pi_2 + \chi_2$	$\mu + \alpha_3 + \pi_3 + \chi_3$

Dette er samme type oppsett som i Tabell 2-1, men med kun 3 perioder og andre symbol for parametrene. Her har vi med en parameter for intercept-leddet (μ), 3 parametre for alder ($\alpha_1, \alpha_2, \alpha_3$), 3 parametre for periode (π_1, π_2, π_3) og 5 parametre for kohort ($\chi_1, \chi_2, \chi_3, \chi_4, \chi_5$). Totalt har vi 12 ukjente parametre som kan estimeres i denne modellen. Hver celle i tabellen tilsvarer en respons. Et typisk datasett, en respons, kan være gitt som i tabellen under. Her er det dødelighet blant norske kvinner som er gruppert etter alder og periode. Antallet i hver celle er det årlige gjennomsnittet pr. 100 000 middelfolkemengde. Dette er utdrag fra en større tabell [15] som er benyttet senere i oppgaven i Kapittel 9.6, som omhandler Partial Least Squares-metoden.

Tabell 4-2: Dødelighet kvinner, etter alder. Døde pr. 100 000 middelfolkemengde. Årlig gjennomsnitt. Kilde: Statistisk sentralbyrå (Tabell 79).

		Periode		
		1996-2000	2001-2005	2006-2010
Aldersgruppe	40-44	114	104	85
	45-49	182	166	150
	50-54	310	272	252

4.2 Betingelser

For å løse ubestemthetsproblemet og få identifiserbare parametre, benyttes betingelser for CGLIM-metoden. Betingelsene har som formål å redusere antall parametre som skal estimeres. Først analyseres dataene ved å benytte CGLIM-metoden. Det benyttes 3 ulike varianter av denne metoden, en som setter 2 alderskoeffisienter lik hverandre, en som setter 2 periodekoeffisienter lik hverandre og en som setter 2 kohortkoeffisienter lik hverandre. Deretter benyttes IE-metoden på de samme dataene. Første kategori for hver gruppe er referanse, dvs. $\alpha_1 = \pi_1 = \chi_1 = 0$

4.3 Tradisjonell alder-periode-kohort-modell med betingelser

CGLIM_A:
I den første modellen, aldersmodellen, innføres betingelsen at aldersgruppe 2 settes lik aldersgruppe 1, og dermed får vi: $\alpha_2 = \alpha_1 = 0$. De ukjente parametrene som skal estimeres er da gitt ved vektoren $\boldsymbol{\beta}$, og responsvektoren $\boldsymbol{y}$ i det gitte talleksempelet kan skrives på formen:

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_3 \\ \pi_2 \\ \pi_3 \\ \chi_2 \\ \chi_3 \\ \chi_4 \\ \chi_5 \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} 114 \\ 104 \\ 85 \\ 182 \\ 166 \\ 150 \\ 310 \\ 272 \\ 252 \end{bmatrix}$$

I dette tilfellet får vi designmatrisen $\boldsymbol{X}$ og videre $\boldsymbol{X^T X}$:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{X^T X} = \begin{bmatrix} 9 & 3 & 3 & 3 & 2 & 3 & 2 & 1 \\ 3 & 3 & 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 1 & 3 & 0 & 1 & 1 & 1 & 0 \\ 3 & 1 & 0 & 3 & 0 & 1 & 1 & 1 \\ 2 & 1 & 1 & 0 & 2 & 0 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

GLM-regresjon utføres så for å finne beta-estimatene $\widehat{\boldsymbol{\beta}}$.

CGLIM_P:

I den neste modellen, periodemodellen, innføres betingelsen at periode 2 settes lik periode 1, og dermed får vi: $\pi_2 = \pi_1 = 0$. De ukjente parametrene som skal estimeres er da gitt ved vektoren $\boldsymbol{\beta}$, og i dette tilfellet får vi designmatrisen $\boldsymbol{X}$ og videre $\boldsymbol{X^T X}$:

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \pi_3 \\ \chi_2 \\ \chi_3 \\ \chi_4 \\ \chi_5 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} \quad \boldsymbol{X^T X} = \begin{bmatrix} 9 & 3 & 3 & 3 & 2 & 3 & 2 & 1 \\ 3 & 3 & 0 & 1 & 1 & 1 & 1 & 0 \\ 3 & 0 & 3 & 1 & 1 & 1 & 0 & 0 \\ 3 & 1 & 1 & 3 & 0 & 1 & 1 & 1 \\ 2 & 1 & 1 & 0 & 2 & 0 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

GLM-regresjon utføres så for å finne beta-estimatene $\widehat{\boldsymbol{\beta}}$.

CGLIM_C:

I den neste modellen, kohortmodellen, innføres betingelsen at kohort 2 settes lik kohort 1, og dermed får vi: $\chi_2 = \chi_1 = 0$. De ukjente parametrene som skal estimeres er da gitt ved vektoren $\boldsymbol{\beta}$, og i dette tilfellet får vi designmatrisen $\boldsymbol{X}$ og videre $\boldsymbol{X^T X}$:

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \pi_2 \\ \pi_3 \\ \chi_3 \\ \chi_4 \\ \chi_5 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad \boldsymbol{X^T X} = \begin{bmatrix} 9 & 3 & 3 & 3 & 3 & 3 & 2 & 1 \\ 3 & 3 & 0 & 1 & 1 & 1 & 1 & 0 \\ 3 & 0 & 3 & 1 & 1 & 1 & 0 & 0 \\ 3 & 1 & 1 & 3 & 0 & 1 & 1 & 0 \\ 3 & 1 & 1 & 0 & 3 & 1 & 1 & 1 \\ 3 & 1 & 1 & 1 & 1 & 3 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

GLM-regresjon utføres så for å finne beta-estimatene $\widehat{\boldsymbol{\beta}}$.

4.4 IE-metoden

Med denne metoden trenger vi ikke innføre noen betingelser og de ukjente parametrene som skal estimeres er da gitt ved vektoren $\boldsymbol{\beta}$, og responsvektoren $\boldsymbol{y}$ i det gitte talleksempelet kan som tidligere skrives på formen:

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \pi_2 \\ \pi_3 \\ \chi_2 \\ \chi_3 \\ \chi_4 \\ \chi_5 \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} 114 \\ 104 \\ 85 \\ 182 \\ 166 \\ 150 \\ 310 \\ 272 \\ 252 \end{bmatrix}$$

I dette tilfellet får vi designmatrisen $\boldsymbol{X}$ og videre $\boldsymbol{X^T X}$:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{X^T X} = \begin{bmatrix} 9 & 3 & 3 & 3 & 3 & 2 & 3 & 2 & 1 \\ 3 & 3 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 3 & 0 & 3 & 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 1 & 1 & 3 & 0 & 1 & 1 & 1 & 0 \\ 3 & 1 & 1 & 0 & 3 & 0 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 & 0 & 2 & 0 & 0 & 0 \\ 3 & 1 & 1 & 1 & 1 & 0 & 3 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Egenverdiene og de tilhørende egenvektorene (prinsipal komponentene) til $\boldsymbol{X^T X}$-matrisen identifiseres. De normaliseres til å ha unit lengde og egenvektoren $\boldsymbol{B_0}$ som korresponderer til den unike egenverdien 0 identifiseres. En regresjonsmodell estimeres med responsvektor $\boldsymbol{y}$ og designmatrise $\boldsymbol{V}$, hvor kolonnevektorene er prinsipal komponentene bestemt av egenvektorene til ikke-null egenverdiene (dvs. det estimeres en prinsipal komponent regresjonsmodell). Den ortonormale matrisen $\boldsymbol{U}$ består av alle egenvektorene og benyttes til å transformere koeffisientene fra prinsipal komponent regresjonsmodellen til regresjonskoeffisientene til intrinsic estimator $\boldsymbol{B}$.

4.5 Estimater for de ulike metodene

IE-metoden benytter to ulike parametriseringer. Enten benyttes effektkoding på de første kategoriene, eller de siste kategoriene. CGLIM-metoden har 3 ulike betingelser. Estimatene for CGLIM-metoden er sentrert, slik at de kan sammenlignes direkte med estimatene fra IE-metoden. Sentrering er beskrevet i metodedelen, Kapittel 6.2. Tabell 4-3 gjengir verdiene til alder-, periode- og kohorteffektkoeffisientene for IE-metoden og CGLIM-metoden. Verdiene er log-koeffisienter.

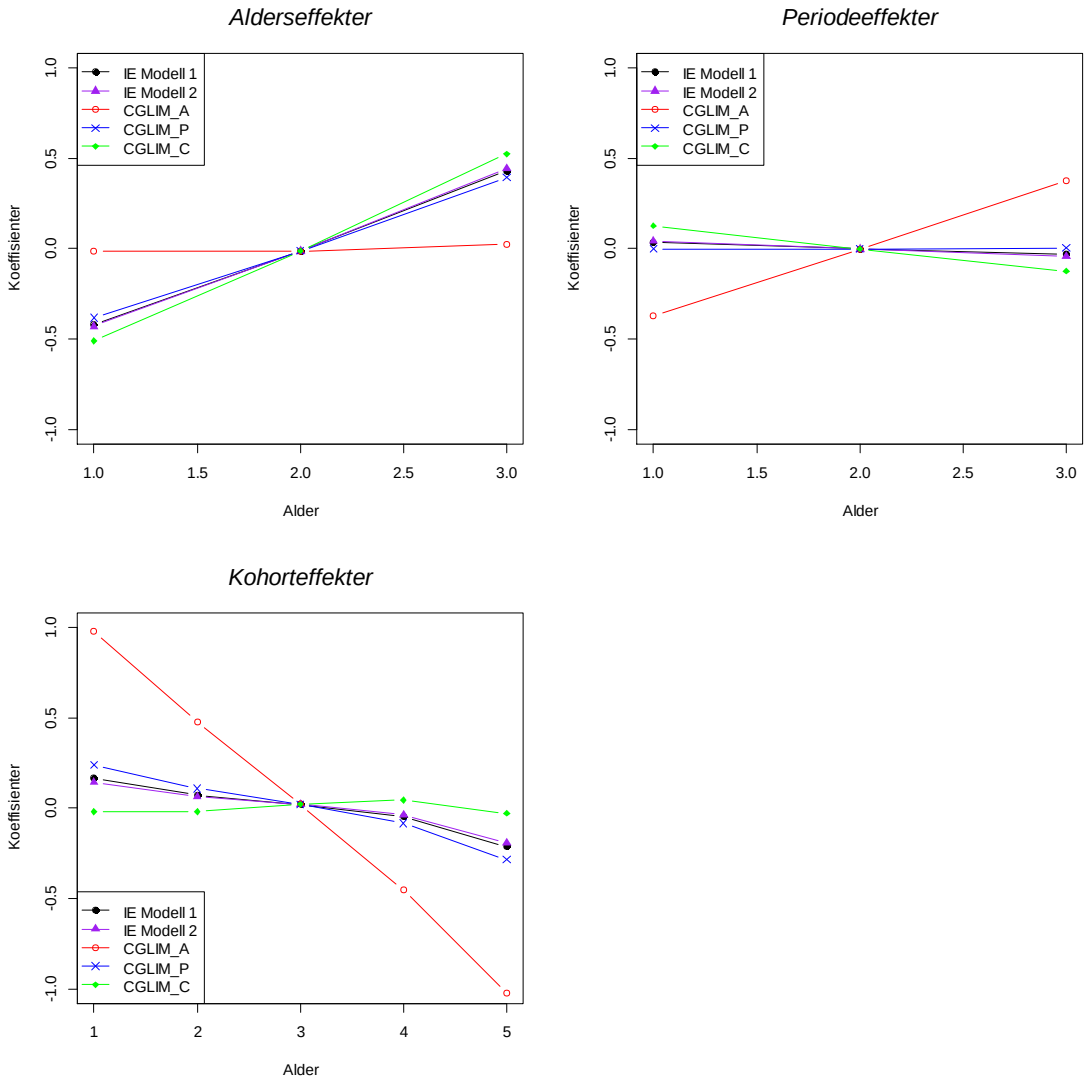
Tabell 4-3: Estimater for henholdsvis alder-, periode- og kohorteffekter med IE-metoden og CGLIM-metoden.

	IE og CGLIM estimat, Dødelighet kvinner, 1996 - 2010				
	IE Modell 1	IE Modell 2	CGLIM_A (A1=A2)	CGLIM_P (P1=P2)	CGLIM_C (C1=C2)
Intercept	-6,407	-6,407	-5,814	-6,554	-6,811
Alder:					
40 – 44	-0,419	-0,429	-0,013	-0,383	-0,511
45 – 49	-0,013	-0,013	-0,013	-0,013	-0,013
50 – 54	0,431	0,442	0,025	0,395	0,524
Periode:					
1996 – 2000	0,034	0,044	-0,372	-0,002	0,126
2001 – 2005	-0,002	-0,002	-0,002	-0,002	-0,002
2006 – 2010	-0,031	-0,042	0,375	0,005	-0,124
Kohort:					
1946	0,166	0,145	0,978	0,238	-0,019
1951	0,073	0,063	0,479	0,109	-0,019
1956	0,021	0,021	0,021	0,021	0,021
1961	-0,047	-0,036	-0,453	-0,083	0,045
1966	-0,213	-0,192	-1,025	-0,285	-0,029

Den første CGLIM-modellen (CGLIM_A) setter betingelsen at effektkoeffisienten til aldersgruppen 45-49 (A2) skal være den samme som til aldersgruppen 40-44 (A1). Den neste CGLIM-modellen (CGLIM_P) setter betingelsen at effektkoeffisientene til perioden 2001-2005 (P2) skal være lik den til 1996-2000 (P1). Tilsvarende setter den siste CGLIM-modellen (CGLIM_C) betingelsen om at effektkoeffisientene til 1951-kohorten (C2) skal være lik den til 1946-kohorten (C1).

De ulike modellene gir ulike estimater for hvordan dødeligheten endrer seg med alderen, med perioden og med kohorten. Mens aldersmodellen gir oss estimater som sier at det omtrent ikke er noen endring i dødeligheten med økende alder, gir de andre modellene oss estimater som indikerer at dødeligheten stiger med alderen. De fleste modellene gir estimater som indikerer at det er lite eller moderat endring i tidsperioden, mens aldersmodellen gir estimater som indikerer at dødeligheten stiger for de ulike tidsperiodene. Resultatene fra aldersmodellen indikerer at dødeligheten synker kraftig med fødselskohort, mens de andre modellene gir estimater som indikerer en mye svakere nedgang. IE-metoden har tilnærmet identiske estimater for de to ulike parametriseringene.

De ulike modellene gir ulike trender, og vi kan ikke vite hvilken modell som gir oss riktige estimater. Dersom en vet noe mer om dataene en analyserer, slik at en har grunnlag for å si om en betingelse er antatt å være riktig, kan man foretrekke en modell foran de andre. Om en derimot ikke vet noe om dataene sine, vil valg av feil modell kunne medføre feil ved estimeringen av effekter. De samme estimatene er illustrert i Figur 4-1. Den horisontale aksen (x-aksen) angir henholdsvis aldersgruppe, periode og kohort, mens den vertikale aksen (y-aksen) angir log-koeffisientene.



Figur 4-1: Illustrasjon av estimater for henholdsvis alder-, periode- og kohorteffekter med IE-metoden og CGLIM-metoden.

Hovedkritikken mot CGLIM-metoden er mangelen på kunnskap om betingelsene som innføres, for valget av betingelse kan produsere vidt forskjellige estimater for effektkoeffisientene, mens de vil gi lik modelltilpasning. Og modellestimatene er sensitiv for valget av betingelse. Mason anbefaler i sin artikkel [9] sammenligning av modeller som reflekterer ulike betingelser for å teste på stabiliteten til resultatene. Gitt begrensningene og usikkerheten som er assosiert med den konvensjonelle metoden for å gi valide estimater til alder-, periode- og kohorteffektene, har IE-metoden som har en ulik tilnærming til problemene blitt tatt i bruk. Den nyere metoden gir en unik løsning til alder-periode-kohort-analysen ved å justere for den lineære avhengigheten mellom alder, periode og kohort [14]. IE-metoden unngår identifikasjonsproblemet og gjør det mulig å estimere effektene uten å innføre betingelser på dataene, og gir derfor estimater som er forventningsrette for alder-, periode- og kohorteffektene [14].

5. Artikkelen til Robert M. O'Brien

Et utgangspunkt for masteroppgaven var å forstå og utdype artikkelen til Robert M. O'Brien [1], og deretter arbeide videre med noen av problemstillingene som introduseres i artikkelen. I dette kapittelet presenteres tematikken til O'Brien, og jeg har valgt å beholde forfatterens overskrifter og benytte tilsvarende symbol for parametrene.

5.1 The age-period-cohort conundrum as two fundamental problems

En generell problemstilling ved alder-periode-kohort-analyser (APC-analyser) er utfordringen med å separere effektene til aldersgrupper, perioder og kohorter. Denne formuleringen unnlater imidlertid å skille to grunnleggende problem i APC-analysene. I artikkelen til O'Brien [1] tar han for seg denne problemstillingen, og forsøker å dele den inn i to fundamentale problem:

1) Problemet med effektforveksling (confounding) av de lineære effektene til alder med effektene til periode og kohort, de lineære effektene til kohort med periode og alder, og de lineære effektene til periode med alder og kohort
2) Problemet med modellidentifikasjon

Det første O'Brien gjør i sin artikkel er å differensiere mellom de to problemene, for deretter å vise hvordan disse innvirker på de ulike konklusjonene som blir tatt utfra de forskjellige løsningene/tilnærmingene til APC-problemet.

I et avsnitt om estimerbare funksjoner undersøker han innvirkningen av fullstendig lineær konfundering på metoder der modellidentifikasjon ikke er et problem. Her ser han på metoder som forsøker å tilskrive den unike variansen i den avhengige variabelen til aldersgrupper, perioder eller kohorter, og han ser på avvik fra linearitet for effektene til hver aldersgruppe, periode eller kohort. Tilslutt i dette avsnittet ser han på metoder som benytter en målbar egenskap som variabel i stedet for f.eks. kohort, for å kunne karakterisere effektene til alder, periode eller kohortegenskaper.

Videre tar han for seg to tilnærminger til modellidentifikasjonsproblemet. Den tradisjonelle c*onstrained generalized linear model* tilnærmingen, omtalt som CGLM i denne artikkelen, og den nyere metoden *intrinsic estimator* (IE).

Multiple classification (dummy variable) coding for APC analysis

I alder-periode-kohort-analyser blir ofte aldersgrupper, perioder og kohorter kodet med dummyvariabler (nivåer av kategoriske forklaringsvariabler). Ved å benytte seg av dummyvariabler for alder, periode og kohort tillates formen på forholdet mellom disse variablene og den avhengige variabelen å være ikke-lineært. O'Brien benytter seg av den multiple klassifikasjonsmodellen gitt ved ligningen:

$$Y_{ij} = \mu + \alpha_i + \pi_j + \chi_{a-i+j} \qquad (5\text{–}1)$$

Y_{ij} er den alder-periode-spesifikke verdien til den avhengige variabelen, μ representerer intercepten, α_i representerer effekten til den i-te aldersgruppen, π_j angir effekten til den j-te perioden og χ_{a-i+j} representerer effektene til kohort $(a - i + j)$, der a er lik antallet aldersgrupper.

Det statistiske problemet med en slik modell er velkjent. Det er en lineær avhengighet mellom dummyvariabler for alder, periode og kohort. Hvis en kjenner aldersgruppen og en periode assosiert med en spesifikk rate, kan en bestemme fødselskohorten assosiert med den raten. Tilsvarende hvis en kjenner perioden og fødselskohorten, kan en bestemme aldersgruppen, og hvis en kjenner aldersgruppen og fødselskohorten er perioden kjent. Den lineære avhengigheten tillater ikke den simultane estimeringen av α_i, π_j og χ_{a-i+j} koeffisientene i ligning (5–1).

Tabellen under illustrerer mønsteret til effektene til de uavhengige variablene i en standard APC-analyse. Cellene representerer forventningen til den alder-periode-spesifikke avhengige variabelen i ligning (5–1).

$$E(Y_{ij}) = \mu + \alpha_i + \pi_j + \chi_{a-i+j}$$

Tabell 5-1: Skjematisk presentasjon av modellen for spesialtilfellet $a = 3, p = 4$.

		Periode *(j)*			
		j = 1	*j = 2*	*j = 3*	*j = 4*
Alders-gruppe *(i)*	i = 1	$\mu + \alpha_1 + \pi_1 + \chi_3$	$\mu + \alpha_1 + \pi_2 + \chi_4$	$\mu + \alpha_1 + \pi_3 + \chi_5$	$\mu + \alpha_1 + \pi_4 + \chi_6$
	i = 2	$\mu + \alpha_2 + \pi_1 + \chi_2$	$\mu + \alpha_2 + \pi_2 + \chi_3$	$\mu + \alpha_2 + \pi_3 + \chi_4$	$\mu + \alpha_2 + \pi_4 + \chi_5$
	i = 3	$\mu + \alpha_3 + \pi_1 + \chi_1$	$\mu + \alpha_3 + \pi_2 + \chi_2$	$\mu + \alpha_3 + \pi_3 + \chi_3$	$\mu + \alpha_3 + \pi_4 + \chi_4$

Denne tabellen er tilsvarende som Tabell 2-1, men med O'Brien sine symboler. Effekten for rad i og effekten for kolonne j bidrar additivt til responsvariabelen i en celle (i, j). I tillegg bidrar den gitte effekten for en spesifikk kohort additivt til denne verdien. Dette er av signifikant betydning i konfunderingen av de lineære effektene til en faktor (alder eller periode eller kohort) med effektene til de andre faktorene.

First fundamental problem: the confounding of linear effects

Det første fundamentale problemet i APC-analyser involverer lineære forhold mellom tid og effektene av aldersgrupper, perioder og kohorter. Det forstås da effekter til alder (eller periode eller kohort) som øker eller minker over tid. Det gjør det dermed umulig å separere de lineære effektene til alder, periode og kohort. Dette problemet er ulikt fra identifikasjonsproblemet.
O'Brien illustrerer dette problemet i Tabell 2 i artikkelen. Deler av tabellen er gjengitt i Tabell 5-2.

Tabell 5-2: The confounding problem of linear effects in APC-models for ages, periods and cohorts.

The confounding problem of linear effects in APC-models for ages, periods and cohorts					
	Periode1	Periode2	Periode3	Periode4	Periode5
Alder1	5,0	5,5	6,0	6,5	7,0
Alder2	4,5	5,0	5,5	6,0	6,5
Alder3	4,0	4,5	5,0	5,5	6,0
Alder4	3,5	4,0	4,5	5,0	5,5
Alder5	3,0	3,5	4,0	4,5	5,0

I dette eksempelet er det en lineær effekt til kohortene som øker med tiden. Responsen i dette tilfellet er en rate som indikerer hvor utsatt en gruppe er for selvmord. For hver ny kohort øker raten med 0,5 per 100 000. Den første kohorten tilsvarer cellen i nedre venstre hjørne, og er representert ved den første perioden og den eldste aldersgruppen. Det er bare en observasjon for denne kohorten, siden gruppen ved neste periode vil tilhøre en aldersgruppe som ikke er tatt med i tabellen. Kohort 2 er representert ved to observasjoner, som begge har en selvmordsrate på 3,5. Slik fortsetter tabellen med en økning av raten med 0,5 for hver ny kohort. Man kan anta at man har en ren kohorteffekt som er lineær med tiden, og at det ikke er en alder- eller periodeeffekt til stede.
Om man heller ser på disse alder-periode-spesifikke ratene med fokus på hvordan periodene endrer seg, ser man at periodeeffekten øker med 0,5 for hver periode fra den første til den siste i tabellen. Fra et aldersperspektiv ser vi at ratene minker med 0,5 for hver aldersgruppe fra den yngste til den eldste. Man kan da spørre seg om dette er en ren kohorteffekt, eller en effekt av alder og periode uten kohorteffekt.

Tilsvarende får vi om vi har en ren lineær alderseffekt hvor raten øker med 0,5 når alderen øker. O'Brien viser også disse verdiene i sin tabell, men de er utelatt her. Denne lineære alderseffekten kan også fullt tilskrives til periode- og kohorteffekter (for økende kohort minker raten med 0,5 og for økende periode øker raten med 0,5 for hver periode). På samme måte kan en ren lineær periodeeffekt også fullt tilskrives som alder- og kohorteffekter. Det er strukturen til APC-datamatrisen som skaper den

fullstendige konfunderingen til en lineær effekt av alder (eller periode eller kohort) med de andre to settene av dummyvariabler. Det blir umulig å separere de lineære effektene til alder, periode og kohort. Vi omtaler dette som et Y-side problem siden det skapes av oppsettet til responsvariabelen.

Second fundamental problem: model identification

Det andre fundamentale problemet O'Brien omtaler i sin artikkel [1] er modellidentifikasjon. Dette problemet oppstår på grunn av den lineære avhengigheten mellom alder, periode og kohortvariablene.

Med O'Brien sine symboler kan ligning (5–1) skrives på formen:

$$\boldsymbol{Y} = \boldsymbol{Xb} + \boldsymbol{\varepsilon} \qquad (5\text{–}2)$$

Der $\boldsymbol{Y}$ er en $ap \times 1$ vektor av observasjonene til den avhengige variabelen, $\boldsymbol{b}$ er en $p \times 1$ vektor av parametre og $\boldsymbol{\varepsilon}$ er en vektor av feilledd som er normalfordelt med forventning 0 og varians $\sigma^2\boldsymbol{I}$. Med denne modellen er forventningen til $\boldsymbol{Y}$ lik $\boldsymbol{Xb}$.
$\boldsymbol{X}$ er designmatrisen hvor de kategoriske forklaringsvariablene (dummyvariablene) er samlet, det er variabler med 1-ere og 0-ere som markerer hvilken gruppe observasjonen tilhører. $\boldsymbol{X}$-matrisen er en $(ap \times 2(a+p)-3)$-matrise.
Rekkefølgen på kolonnene kan bli satt opp som $(1, (a-1), (p-1), (a+p-2))$.
Første kolonne er en kolonne med enere, $(a-1)$ angir antallet av dummyvariabler for alder, $(p-1)$ angir antallet av dummyvariabler for periode og $(a+p-2)$ angir antallet av dummyvariabler for kohort.

Den ordinære minste kvadraters estimatoren til modellen i ligning (5–2) er gitt ved:

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y} \qquad (5\text{–}3)$$

Som tidligere beskrevet i Kapittel 2.2 finnes det et uendelig antall av mulige løsninger til ligning (5–3), på grunn av det lineære forholdet mellom alder-, periode- og kohorteffekter. Det eksisterer ikke en regulær invers og det er ikke mulig å estimere de simultane effektene til alder, periode og kohort uten å innføre begrensninger på koeffisientene.
Figuren under illustrerer dette modellidentifikasjonsproblemet.

$$b_1\begin{bmatrix}1\\1\\.\\.\\1\end{bmatrix} + b_2\begin{bmatrix}x\\x\\.\\.\\x\end{bmatrix} + b_3\begin{bmatrix}x\\x\\.\\.\\x\end{bmatrix} + \ldots + b_{m-1}\begin{bmatrix}x\\x\\.\\.\\x\end{bmatrix} + b_m\begin{bmatrix}x\\x\\.\\.\\x\end{bmatrix} = \begin{bmatrix}0\\0\\.\\.\\0\end{bmatrix}$$

Figur 5-1: Modellidentifikasjonsproblemet i APC-modeller.

Den første vektoren er en vektor av enere og de andre vektorene representerer dummyvariablene for aldersgruppene, periodene og kohortene (med unntak av referansekategoriene). b-ene representerer koeffisienter som multipliseres med kolonnevektorene for å danne 0-vektoren. Når en slik vektor av b-er eksisterer, der ikke alle b-ene er null, indikerer dette en lineær avhengighet i kolonnene i $\boldsymbol{X}$-matrisen. Det eksisterer bare en slik unik vektor for APC-modellen når det ikke er satt noen spesielle betingelser for modellen. Det er en ikke-triviell løsning på de $a \times p$ homogene ligningene, og vi kan skrive $\boldsymbol{Xv} = 0$, der $\boldsymbol{X}$ er designmatrisen og $\boldsymbol{v}$ er vektoren som inneholder b-ene. Denne vektoren, $\boldsymbol{v}$, er i nullrommet til $\boldsymbol{X}$ og benevnes som nullvektoren. Det er den lineære kombinasjonen av kolonnene i $\boldsymbol{X}$ som resulterer i 0-vektoren. At det bare er én slik vektor indikerer at rangen til $\boldsymbol{X}$-matrisen er kun en mindre enn full kolonnerang, og at innføring av en enkelt lineær betingelse bør kunne løse identifikasjonsproblemet. For å kunne finne en unik løsning for de individuelle dummyvariablene, må det velges en lineær betingelse. Enhver slik betingelse assosieres med en generalisert invers som lar oss finne en løsning gitt den spesifiserte lineære betingelsen. Den generaliserte inversen, omtalt som $\boldsymbol{G_c}$ hos O'Brien, multipliseres med $\boldsymbol{X^T Y}$ i ligning (5–3) for å finne en løsning:

$$\widehat{\boldsymbol{b}}_{\boldsymbol{c}} = \boldsymbol{G_c X^T Y}$$

Løsningen som finnes, dvs. parameterestimatene i $\widehat{\boldsymbol{b}}_{\boldsymbol{c}}$ vil være unike for den spesifikke betingelsen.

En typisk betingelse som benyttes ved den tradisjonelle CGLM-metoden er å sette to av koeffisientene assosiert med alder, periode eller kohort lik hverandre. Med denne ene tilleggsbetingelsen blir matrisen $\boldsymbol{X^T X}$ ikke-singulær og minste kvadraters estimatoren eksisterer. Antagelsen er at $\boldsymbol{c^T \beta} = 0$. Eksempelvis er $\boldsymbol{c^T} = (0,1,-1,0,\ldots,0)$ vektoren for betingelsene, der aldersgruppene 1 og 2 blir satt til å ha den samme effekten. Enhver slik betingelse er assosiert med en generalisert invers som benyttes til å løse for aldersgruppe-, periode- og kohortkoeffisientene.

For IE-metoden involverer betingelsen vektoren $\boldsymbol{v}$, og antagelsen er at $\boldsymbol{v^T \beta} = 0$. $\boldsymbol{v}$ er nullvektoren, som når den blir multiplisert med kolonnene i $\boldsymbol{X}$-matrisen resulterer i 0-vektoren. Den generaliserte inversen som er assosiert med dette er Moore-Penrose generaliserte invers.

Identifikasjonsproblemet blir kalt et X-side problem, siden det involverer designmatrisen og begrenser kolonnene i den matrisen for å få en unik løsning på problemet.

Confounding: variance decomposition, deviations from linearity, and cohort characteristics

I dette avsnittet ser O'Brien [1] på estimerbare funksjoner og undersøker virkningen av fullstendig lineær konfundering på metoder hvor modellidentifikasjon ikke er et problem. Det finnes andre tilnærminger til APC-analysene som avhenger av estimerbare funksjoner, dvs. funksjoner som kan identifiseres. Disse metodene unngår identifikasjonsproblemet ved å ikke fokusere på estimeringen av de spesifikke alder-, periode- og kohortkoeffisientene. Disse spesifikke koeffisientene er ikke estimerbare uten å innføre en lineær betingelse på $\boldsymbol{X}$-matrisen. En vanlig betingelse som benyttes for å identifisere de individuelle alder-, periode- og kohortkoeffisientene er å sette to av alderskoeffisientene like hverandre, eller to av periodekoeffisientene like, eller tilsvarende to av kohortkoeffisientene like. Men selv om det ikke innføres noen betingelse, kan vi likevel få viktig informasjon om alder-, periode- og kohorteffekter fra disse estimerbare funksjonene. Vi kan dele opp den unike variansen til den alder-periode-spesifikke avhengige variabelen (responsen) i alderseffekter, periodeeffekter og kohorteffekter. Det er også mulig å estimere avvik fra linearitet for alder-, periode- og kohorteffektkoeffisientene, og andreordensdifferansene kan estimeres for disse effektkoeffisientene. En annen mulighet er å benytte seg av en eller flere kohortkarakteristikker som erstatning for kohorteffektene og dermed kunne få identifiserte effekter for alder, periode og kohortkarakteristikker.

1) Varians dekomposisjon

Siden $\boldsymbol{X}$-matrisen er singulær, vil de individuelle alder-, periode- og kohorteffektene i APC-modellen ikke være estimerbare. Hvert sett av estimater for regresjonskoeffisientene vil være forskjellig ved å benytte ulike generaliserte inverser. Men hvert sett vil gi de samme predikerte verdiene for den avhengige variabelen. De predikerte verdiene til den avhengige variabelen er i denne situasjonen estimerbare, mens de individuelle regresjonskoeffisientene ikke er det. Vi kan utnytte disse estimerbare funksjonene til å få nyttig informasjon om alder-, periode- og kohorteffekter. Effektene som estimeres er identifiserbare.
Variansen til den alder-periode-spesifikke avhengige variabelen kan deles opp i alderseffekter, periodeeffekter eller kohorteffekter uten å innføre noen spesielle betingelser. De predikerte verdiene til den avhengige variabelen kan benyttes til å estimere den totale variansen til aldersgruppene, periodene og kohortene. Så kan variansen som tilskrives unikt til en av faktorene beregnes fra den totale variansen minus variansen som tilskrives de andre to faktorene. Dette kan gjøres ved hjelp av separate regresjonsanalyser der ulike faktorer er med. Signifikanstester hjelper oss å avgjøre om aldersgrupper, perioder og kohorter hver bidrar signifikant til den avhengige variabelen, og om alle tre faktorene trengs for å modellere verdiene til den avhengige variabelen.
Varians dekomposisjonen baserer seg på å tilskrive den unike variansen til aldersgrupper, eller til perioder, eller til kohorter som ikke er assosiert med de andre to faktorene. Metoden vil bare oppdage effekter som ikke er lineært assosiert med tid. Lineære effekter til kohort assosiert med tid vil bli absorbert av alder og periode, lineære effekter til periode assosiert med tid vil bli absorbert av alder og kohort, og lineære effekter til aldersgrupper assosiert med tid, vil bli absorbert av periode og kohort. Det første fundamentale problemet gjør det dermed vanskelig å tilskrive variansen til alder, periode eller kohort.
Det lineære forholdet mellom tid og kohort, tid og periode, og tid og aldersgruppe gjør oppdelingen av variansen i alder, periode og kohorteffekter problematisk, selv om disse komponentene er estimerbare.

2) Avvik fra linearitet og andreordensdifferanser

Avvik fra lineariteten kan estimeres for alder-, periode- og kohorteffektkoeffisientene og andreordensdifferansene kan også estimeres for disse effektkoeffisientene. Holford [10, 21] estimerer avvik fra linearitet ved å kontrollere for lineære trender i alder, periode og kohort, og benytter seg av ortogonale polynom for å estimere avvikene fra linearitet for de individuelle alder-, periode- og kohorteffektene. Avvikene er estimerbare og lar oss måle formen på avvikene. Holford parametriserer modellen slik at hver tidseffekt blir delt i to komponenter: lineær og ikke-lineær trend, dvs. krumning eller avvik fra den rette linjen. Disse komponentene står ortogonalt på hverandre. Alderseffektene kan da uttrykkes som summen av den lineære komponenten og krumningskomponenten, og på tilsvarende måte kan vi uttrykke periode- og kohorteffektene. Designmatrisen blir satt opp ved hjelp av en parametrisering ut fra disse komponentene. For å teste om parametrene er estimerbare, bruker Holford en generalisert invers. Når effekten av alder, periode og kohort skal presenteres, kan den lineære komponenten og krumningen rapporteres hver for seg. Holford viser at den totale lineære trenden ikke er estimerbar, mens i motsetning er krumningen estimerbar, og kan finnes ved å fjerne den lineære komponenten fra de estimerte parameterverdiene. Holfords metode baserer seg på at lineær trend blir fjernet fra alle de tre variablene, og netto drift rapporteres som summen av to stigningstall.
Andreordensdifferansen til alder-, periode- og kohortkoeffisientene er estimerbare, og Clayton og Schifflers [4] viser at man dermed kan bestemme om raten til forandring i forandringene er økende eller minkende. Man kan således se på om hellingen på en kurve blir brattere eller mindre bratt med tiden. I denne metoden er det bare den estimerbare krumningen som blir rapportert. Det kan for eksempel være forholdet mellom to påfølgende relative risikoer.
Avvikene fra linearitet og forandringen i ratene avhenger av ikke-lineære effekter, og de lineære effektene til aldersgrupper, perioder og kohorter vil ikke bli oppdaget med disse metodene. Det første fundamentale problemet gjør også tilnærmingen med å se på de ulike effektenes avvik fra linearitet problematisk.

3) Kohortkarakteristikker

En eller flere kohortkarakteristikker kan benyttes som erstatning for kohorteffektene og gi identifiserbare effekter for alder, periode og kohortkarakteristikker. Dette er mye benyttet i sosiologi og samfunnsvitenskapelige artikler. Et eksempel er O'Brien [22] som fokuserer på to kohortkarakteristikker som han har en teori om at kan ha en innvirkning på mordarrestratene, i tillegg til de faste effektene til aldersgrupper og perioder. Denne metoden benytter seg av målbare variabler som er sterkt knyttet til kohortene, og de blir erstatninger for kohorteffektene. Siden kohortkarakteristikken ikke er lineært avhengig av forklaringsvariablene for alder og periode, unngår man problemet med lineær avhengighet. I analysen har man den avhengige variabelen gitt ved responsen, og man har de uavhengige variablene aldersgrupper, tidsperiode, og mål på variabler assosiert med kohortene. Denne metoden lar oss se på forandringer i den avhengige variabelen fra spesielle kohortkarakteristikker mens vi kontrollerer for både alder og periode. Ved å inkludere variabler som måler kohortkarakteristikker i stedet for dummyvariabler for hver kohort, unngår man problemet med lineær avhengighet. Eksempler på kohortkarakteristikker kan være relativ størrelse på fødselskohortene, familiestruktur i oppveksten, etc. I undersøkelsene får hver kohort tilegnet en verdi for hver av kohortkarakteristikkene. Problemet med lineær konfundering er også et problem ved denne metoden. Kohortkarakteristikkene kan ikke oppdage lineære effekter til kohortene, men bare avvik fra linearitet.

Metodene som har blitt beskrevet i dette avsnittet avhenger alle av avvik fra linearitet for å påvise signifikante effekter. Når man har problemet med lineær konfundering, er den eneste konklusjonen man kan trekke ved å benytte disse metodene, at det er ingen ikke-lineære effekter. Lineære tidseffekter til kohortene vil bli absorbert av alder og periode, og tilsvarende for de andre komponentene. Lineære effekter kan derfor ikke oppdages. Man kan bare få estimert avvik fra linearitet, eller den unike variansen som kan tilskrives hvert sett av dummyvariabler.

Estimating age, period, and cohortcoefficients: CGLM and IE approaches

O'Brien benytter seg av to metoder for å løse modellidentifikasjonsproblemet.

1) Den tradisjonelle metoden: C*onstrained Generalized Linear Models estimator (CGLM)*
2) Den nyere metoden: *Intrinsic Estimator (IE)*

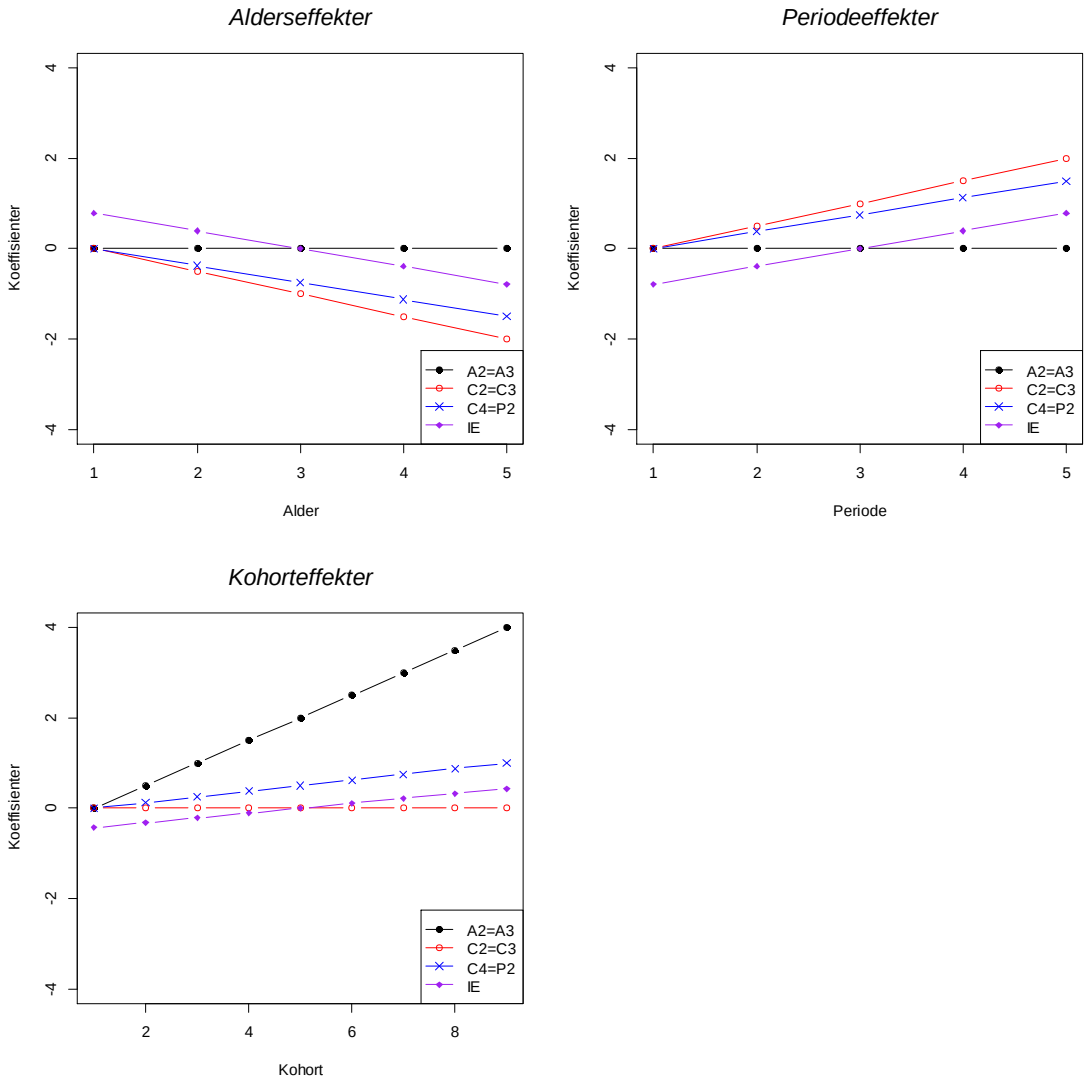
CGLM- og IE-metodene er designet for å estimere hver av alder-, periode- og kohorteffektkoeffisientene, slik at de spesifikke effektene til hver aldersgruppe, hver periode og hver kohort kan bestemmes. Både CGLM- og IE-tilnærmingen baserer seg på å innføre en lineær betingelse på kolonnene i $\boldsymbol{X}$-matrisen, som ikke er av full rang. Denne betingelsen assosieres med en generalisert invers, som gir en unik løsning til alder-, periode- og kohortkoeffisientene, gitt den betingelsen. Den generaliserte inversen som er assosiert med IE-estimatoren er Moore-Penrose generaliserte invers, og den har noen spesielle tekniske egenskaper som gjør den til det absolutt beste valget for en generalisert invers i fravær av annen informasjon. Om en derimot har sikker informasjon om at to av koeffisientene blant alder, periode og kohort er like i populasjonen som undersøkes, så kan valget av CGLM-metoden være å foretrekke.

O'Brien benytter seg av både CGLM- og IE-metoden for å undersøke APC-data som er generert av en ren lineær kohorteffekt, tilsvarende dataene fra Tabell 5-2. For CGLM-metoden benytter han seg av 3 ulike betingelser. Først setter han alder 2 lik alder 3, deretter kohort 2 lik kohort 3 og så kohort 4 lik periode 2. Den siste varianten har han tatt med for å vise hva som kan skje om man setter betingelser utfra observerte data. De observerte verdiene for kohort 4 er alle lik 4,5, og gjennomsnittseffekten for periode 2 er også 4,5. Ved tolking av dataene kan det således virke som om disse 2 effektene kan være den samme i populasjonen. Resultatene fra analysene er gjengitt i Tabell 5-3.

Tabell 5-3: Estimater fra CGLM og IE ved analyse av ren lineær kohorteffekt.

	CGLM (A2=A3)	CGLM (C2=C3)	CGLM (C4=P2)	IE
Intercept	3,000	5,000	4,500	5,000
Alder:				
1	0,000	0,000	0,000	0,786
2	0,000	-0,500	-0,375	0,393
3	0,000	-1,000	-0,750	0,000
4	0,000	-1,500	-1,125	-0,393
5	0,000	-2,000	-1,500	-0,786
Periode:				
1	0,000	0,000	0,000	-0,786
2	0,000	0,500	0,375	-0,393
3	0,000	1,000	0,750	0,000
4	0,000	1,500	1,125	0,393
5	0,000	2,000	1,500	0,786
Kohort:				
1	0,000	0,000	0,000	-0,429
2	0,500	0,000	0,125	-0,321
3	1,000	0,000	0,250	-0,214
4	1,500	0,000	0,375	-0,107
5	2,000	0,000	0,500	0,000
6	2,500	0,000	0,625	0,107
7	3,000	0,000	0,750	0,214
8	3,500	0,000	0,875	0,321
9	4,000	0,000	1,000	0,429

De samme estimatene er også plottet i Figur 5-2.



Figur 5-2: Estimater fra CGLM og IE ved analyse av ren lineær kohorteffekt.

I situasjonen hvor alderseffektene betinges å være like (A2=A3) får vi estimater for effektkoeffisientene som reflekterer kohortgenereringsprosessen. Betingelsen er riktig i forhold til hvordan dataene var generert, og de ulike effektene estimeres riktig. O'Brien utfører også simuleringer (n=1000) og rapporterer resultatene som gjennomsnittseffekter med tilhørende feil. Når han tilfører feil i prosessen finner han i hovedsak identiske resultater.

I den neste situasjonen betinges effektene til 2 kohorter å være like (C2=C3), og estimatene fra denne analysen er ikke i tråd med hvordan dataene antas å være generert. Dette gir mening siden ingen av kohorteffektene var like i datagenereringsprosessen. Her er alle kohorteffektene estimert til 0, mens alderseffektene minker med 0,5 med alderen, og periodeeffektene øker med 0,5 med tiden. På grunn av den fullstendige konfunderingen av de lineære effektene til kohortene, har enhver lineær effekt til kohortene blitt absorbert av alder- og

periodeeffektkoeffisientene. Betingelsen i denne situasjonen er feil i forhold til hvordan dataene ble generert, men på den annen side kan de samme effektene bli generert fullstendig av alder og periode, og da vil den gitte betingelsen være korrekt. Lineære effekter til kohortene kan forklares fullt ut av alder- og periodedummyvariabler, og datagenereringsprosessen kunne ha vært basert på alder- og periodeeffekter. Det ville da ha vært hensiktsmessig å sette to av kohorteffektene lik hverandre.

Den neste situasjonen i tabellen viser hva som skjer når en av kohortene betinges å være lik en av periodene (C4=P2). Når denne betingelsen benyttes, ser vi et mønster hvor den rene lineære effekten til kohortene i genereringsprosessen fordeles til aldersgrupper, perioder og kohorter. O'Brien viser også at ved innføring av tilfeldig feil i simuleringsmodellen, vil mange av koeffisientene for aldersgrupper, perioder og kohorter avvike signifikant fra hverandre.

Når en benytter seg av CGLM-metoden, kan en få en rekke ulike estimater som vil avhenge av hvilken betingelse som innføres. For IE-metoden er det ikke slik, her er det tilstrekkelig med bare én betingelse som resulterer i en enkelt generalisert invers. Betingelsen gir en løsning som er ortogonal til nullvektoren i nullrommet til $\boldsymbol{X}$-matrisen. Under denne betingelsen vil IE-prosedyren fordele effektene til den rene lineære kohorteffekten på aldersgrupper, perioder og kohorter. IE-metoden klarer ikke å bestemme hvordan disse dataene ble generert, om de ble generert som en ren lineær kohorteffekt eller om de ble generert som de lineære effektene av periode og aldersgrupper. IE-metoden krever at løsningsvektoren er ortogonal til nullvektoren. Når O'Brien tilfører feil i prosessen, finner han i hovedsak identiske resultater.

Om betingelsen som innføres ved CGLM-metoden er konsistent eller inkonsistent med datagenereringsprosessen vil avgjøre om estimatene som produseres er korrekte eller ikke. Når betingelsen er inkonsistent, vil ikke estimatene stemme overens med genereringsprosessen. Noen betingelser vil gi resultater som tilskriver all effekt til bare en faktor (alder eller periode eller kohort), og andre vil tilskrive all effekt til de andre to faktorene. Den veldige ustabiliteten skyldes den fullstendige konfunderingen til de lineære effektene.

Simuleringene med rene lineære effekter er de verste situasjonene med tanke på konfundering, men situasjoner der en har delvis lineære effekter skaper også konfundering. I artikkelen [1] tar O'Brien også for seg analyser der en bare har delvis lineære effekter. Det kan for eksempel være at de første kohortene har de samme effektene, mens de resterende kohortene har en lineær økning med økende kohort. I denne situasjonen gir 3 ulike varianter av CGLM-metoden ulike estimerte parametre for alder-, periode- og kohorteffektene utfra valgte betingelse. Igjen viser resultatene at når betingelsen som velges er konsistent med datagenereringsprosessen for alder og periode (f.eks. alder 2 settes lik alder 3), så får vi koeffisienter som er konsistent

med ligningen dataene genereres fra. En betingelse som er inkonsistent med datagenereringen vil gi estimater som feilaktig gjengir datagenereringsprosessen. Til slutt vises også estimatene fra IE-metoden, som også her fordeler de delvis lineære kohorteffektene på koeffisienter for aldersgrupper, perioder og kohorter. Koeffisientene reflekterer ikke genereringsprosedyren. De er kun løsningene som betinges å være ortogonal til nullvektoren.

Conclusions

Helt til slutt oppsummerer O'Brien de to grunnleggende problemene ved APC-analyser. Y-side problemet som involverer den fullstendige konfunderingen av de lineære effektene til kohort med alder og periode, alder med periode og kohort, og periode med alder og kohort. Dette problemet oppstår fra responsvariabelen og det spesifikke mønsteret av lineære effekter i hvilken som helst av faktorene (alder, periode eller kohort). Problemet påvirker også metoder som ikke avhenger av identifiseringen av de individuelle effektkoeffisientene. De lineære effektene skaper problemer for metoder som er designet for å undersøke den unike variansen som kan tilskrives til hver av de tre settene med dummyvariabler. Det skaper også problemer med å estimere avvik fra linearitet for alder-, periode- og kohorteffektene, som igjen medfører at disse metodene er langt mindre nyttig enn om de lineære effektene ikke var til stede i samme grad.

X-side problemet har med modellidentifikasjon å gjøre, og involverer et uendelig antall av løsninger for alder, periode og kohorteffektkoeffisienter når ingen spesiell betingelse er gitt i modellen. O'Brien benytter seg både av analyser der dataene har en ren lineær kohorteffekt og analyser der dataene har en delvis lineær kohorteffekt. Basert på disse resultatene konkluderer han med at om en betingelse velges slik at den er konsistent med datagenereringsprosessen, så vil en få koeffisienter som er konsistent med denne prosessen. På den annen side vil en betingelse som er inkonsistent med datagenereringsprosessen gi koeffisienter som ikke er konsistent med prosessen. Og dette vil gjelde både CGLM- og IE-metoden.

6. Metode

6.1 Modellvalidering

Jeg har benyttet meg av simuleringsanalyser for å undersøke hvor godt de ulike metodene (IE- og CGLIM-estimatorer) gjenskaper den sanne formen til den underliggende modellen som genererer dataene. Videre er det også benyttet simuleringsanalyser for å sammenligne IE-metoden med metodene som Clayton og Schifflers omtaler i sine artikler [3, 4] og for sammenligning med metoden som baserer seg på Partial Least Squares [5]. Det er velkjent at det har vært problemer knyttet til de tradisjonelle CGLIM-metodene for APC-analyser, og det er rimelig å undersøke om disse estimatorene gir numeriske estimater for alder-, periode- og kohorteffektkoeffisientene som er valide, og avdekker de sanne effektene. Samtidig får man undersøkt hvor godt IE-metoden fungerer og hvor robust metoden er mot endringer i effektene.
Analysene utført med IE-metoden er programmert i statistikkprogrammet R etter algoritmen som beskrives i Kapittel 3.2. Programmeringskoder for en av simuleringsmodellene er med som vedlegg. I artikkel [2] blir leseren henvist til en e-postadresse og en internettadresse for å få tilgjengelig programvare for å estimere med IE-metoden. Jeg kontaktet Wenjiang Fu på den aktuelle e-postadressen for å spørre om programvare tilgjengelig for R. Det var ikke tilgjengelig på det aktuelle tidspunktet, men han sendte link til en internettside. På den kunne en legge inn responsdataene og antall aldersgrupper og perioder, og så få de ulike estimatene beregnet. Dette var ikke aktuelt for en simuleringssituasjon, og jeg har derfor programmert IE-algoritmen i R selv. Denne algoritmen er benyttet for alle analyser med IE-metoden i denne oppgaven.

6.2 Simuleringsmodell

Til simuleringsoppsettet mitt har jeg hentet inspirasjon fra Yang et al. [2]. Først benytter jeg meg av en simuleringsmodell som er lignende den Yang benytter, deretter introduserer jeg flere nye modeller hvor jeg endrer effekten for alder, periode og/eller kohort. Slik får jeg modeller med ønskede effekter for de ulike kategoriene. Antallet aldersgrupper, perioder og fødselskohorter som er med i modellene er fast. I oppsettet mitt benytter jeg meg av 9 aldersgrupper, 5 perioder og 13 fødselskohorter. For hver simuleringsmodell genererer jeg 10 000 datasett, hvor responsen er fordelt etter ligningen:

$$y_{ij} \sim \text{Poisson}\{\exp[\text{intercept} + \text{alderseffekt}_{ij} + \text{periodeeffekt}_{ij} + \text{kohorteffekt}_{ij}]\}$$

Denne ligningen for datagenereringsprosessen forteller oss hva de sanne alder-, periode- og kohorteffektene er.

I stedet for å bruke referansekategorier benytter IE-metoden seg av sum-lik-null parametrisering. For å kunne sammenligne estimatene direkte, benytter jeg sentrerte verdier for CGLIM-metoden. Dvs. at jeg trekker fra konstanter fra de ordinære effektene, slik at de nye effektene normaliseres til å ha gjennomsnitt på 0 for hver kategori. Da vil hver kategori summeres til null. For å forhindre at de predikerte verdiene skal endres, må den samme konstanten legges til intercept. Konstantene beregnes som gjennomsnittseffekten for hver kategori.
De simulerte dataene ble analysert i en regresjonsmodell for poissonfordeling, og jeg får estimater for alder-, periode- og kohorteffekter i hvert simulert datasett ved å benytte IE-metoden og 3 ulike CGLIM-estimatorer. En som betinger de to første aldersgruppene til å være like (CGLIM_A), en som betinger de to første periodene til å være like (CGLIM_P) og en som betinger de to første kohortene til å være like (CGLIM_C). Estimatene sammenlignes med de sanne regresjonskoeffisientene.
I Kapittel 8 utføres det simuleringsanalyser for å sammenligne IE-metoden med ulike reduserte modeller, modeller som inkluderer en driftparameter, og fulle APC-modeller basert på førsteordensdifferansene. I Kapittel 9 utføres det også simuleringsanalyser for å sammenligne IE-metoden med PLS-metoden.

6.3 Mean Square Error (MSE)

I statistikken er MSE (Mean Square Error) til en estimator en av mange måter å kvantifisere forskjellen mellom en estimator og den sanne verdien den har. Det er et kvalitetsmål for hvor mye de predikerte verdiene avviker fra de virkelige verdiene. MSE måler gjennomsnittet til den kvadrerte feilen. Feilen er den forskjellen som estimatoren har fra den størrelsen som skal estimeres [23].

MSE til en estimator $\hat{\theta}$ med hensyn på den estimerte parameteren θ kan defineres som:

$$\mathrm{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^2) \qquad (6\text{–}1)$$

MSE har minst to fordeler foran andre avstandsmål. Den er ganske medgjørlig analytisk, og den innlemmer både variansen til estimatoren og biasen. MSE er lik summen til variansen og den kvadrerte bias.

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}, \theta))^2 \qquad (6\text{–}2)$$

MSE har to komponenter, en som måler variabiliteten til estimatoren (presisjon) og den andre som måler dens bias (nøyaktighet). For en forventningsrett (unbiased) estimator er MSE lik variansen. En estimator som har gode MSE-egenskaper, har liten kombinert varians og bias. På samme måte som variansen har MSE også samme måleenhet som kvadratet til størrelsen som blir estimert.

6.4 Modellere *Goodness-of-fit*-tester

Goodness-of-fit til en statistisk modell beskriver hvor godt modellen tilpasser seg et sett av observasjoner. Målene oppsummerer typisk avviket mellom observerte verdier og de forventede verdiene for den aktuelle modellen. Den essensielle ideen i de fleste goodness-of-fit-testene er å sammenligne den aktuelle modellen med en modell som fullstendig forklarer dataene vi er interessert i, en såkalt mettet modell. I tilfellet med log-lineær analyse vil den mettede modellen bestå av en parameter for hver celle, slik at den predikerer perfekt de observerte cellefrekvensene.

6.4.1 Devians

Devians er en nyttig størrelse ved vurdering og sammenligning av ulike modeller. Devianstesten er en mye brukt test, og den kan bli ansett som et mål på manglende tilpasning mellom modell og data. Generelt har man at dess større deviansen er for en modell, dess dårligere tilpasser modellen dataene. Deviansen tolkes vanligvis ikke direkte, men sammenlignes heller med deviansene fra andre modeller som er tilpasset til de samme dataene. Differansen mellom deviansene har en kji-kvadratfordeling med antall frihetsgrader lik differansen i antallet av parametre som estimeres. Kji-kvadratstatistikken er forskjellen mellom deviansen i den aktuelle modellen og deviansen i den mettede modellen. Store verdier for kji-kvadratstatistikken tas som bevis på at nullhypotesen er usannsynlig. Nullhypotesen er at den aktuelle modellen tilpasser dataene godt nok.

Dobson [17] definerer deviansen til en poissonmodell som 2 ganger differansen mellom log-likelihood til modellene og den er gitt ved:

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

I modellen er $l(\boldsymbol{b}_{\max}; \boldsymbol{y})$ log-likelihoodfunksjonen for en maksimal modell, mettet modell, mens $l(\boldsymbol{b}; \boldsymbol{y})$ er log-likelihoodfunksjonen til den aktuelle modellen. En mettet modell er en modell som har perfekt tilpasning til data, fordi den har like mange parametre som antall verdier som skal tilpasses. Den aktuelle modellen vil ha færre parametre enn den mettede modellen. En modell som gir en god tilpasning vil ha en log-likelihood verdi som er nær log-likelihood verdien til den mettede modellen. Dermed følger det at dess mindre verdien for deviansen er, dess mindre er differansen i log-likelihood mellom modellene, og tilpasningen for en gitt modell er bedre. Deviansen er likelihood-ratio statistikken for å teste nullhypotesen om at modellen holder sammenlignet med den mettede modellen.

Maksimum likelihood estimatene er $\hat{\lambda}_i = y_i$, slik at maksimumverdien til log-likelihoodfunksjonen er:

$$l\,(\boldsymbol{b}_{\max}; \boldsymbol{y}) = \sum y_i \cdot \log y_i - \sum y_i - \sum \log\, y_i!$$

Maksimumverdien til log-likelihoodfunksjonen til den aktuelle modellen er:

$$l\,(\boldsymbol{b}; \boldsymbol{y}) = \sum y_i \cdot \log \hat{y}_i - \sum \hat{y}_i - \sum \log\, y_i!$$

Deviansen kan dermed skrives som:

$$D = 2 \cdot (\sum y_i \cdot \log\,(y_i/\hat{y}_i) - \sum (y_i - \hat{y}_i))$$

For de fleste modellene kan det vises at $\sum y_i = \sum \hat{y}_i$, og deviansen kan skrives på formen:

$$D = 2 \cdot \sum o_i \cdot \log\,(\frac{o_i}{e_i})$$

hvor o_i er den observerte verdien, y_i, og e_i er den estimerte forventede verdien, $\hat{y}_i$. Om en modell er god, vil deviansen, D, være tilnærmet kji-kvadratfordelt med $(N - p)$ frihetsgrader, der N står for antallet av observasjoner, og p er antallet parametre. Ved sammenligning av to modeller med henholdsvis p og q parametre, vil differansen være kji-kvadratfordelt med $(p - q)$ frihetsgrader. Vi har da at:

$$\Delta D \sim \chi^2_{p-q}, (p > q)$$

6.4.2 AIC (Akaike Information Criterion)

Akaike Information Criterion (AIC) er en måte å velge ut en modell fra et sett av modeller. AIC er et mål på den relative kvaliteten til en statistisk modell, for et gitt sett med data. Den valgte modellen er den som minimerer Kullback-Leibler avstanden mellom modellen og sannheten [24]. Det kan sees på som et kriterie som søker en modell som har så god tilpasning til sannheten som mulig, men med få parametre. AIC søker å finne en avveining mellom goodness-of-fit til modellen, og kompleksiteten til modellen. Den gir et relativt estimat av informasjonen som går tapt når en gitt modell blir brukt til å representere den prosessen som genererer data.

AIC defineres som:

$$\mathrm{AIC} = -2(l\,(\boldsymbol{b}; \boldsymbol{y}) - l\,(\boldsymbol{b}_{\max}; \boldsymbol{y})) + 2K$$

hvor log-likelihoodfunksjonen er sannsynligheten til dataene i en maksimal modell

eller i en aktuell modell, og Ker antallet frie parametre i modellen. AIC-målet er ofte vist som et ΔAIC-mål, som forskjellen mellom den beste modellen (lavest AIC) og hver modell. Gitt et sett med mulige modeller for dataene, er den foretrukne modellen den med minst AIC-verdi. AIC tar ikke bare hensyn til goodness-of-fit, men inkluderer også en straff som er en økende funksjon av antallet estimerte parametre. Denne straffen forhindrer overtilpasning (en økning av antallet frie parametre i modellen vil forbedre tilpasningen, uavhengig av antall frie parametre i den datagenererende prosessen).

AIC og likelihood-ratio-test (LRT) baserer seg på ganske ulike prosedyrer. Mens AIC baserer seg på modellseleksjon ved å minimere den forventede Kullback-Leibler avstanden, baserer LRT seg på nullhypotesetesting. LRT baserer seg på det faktum at minus to ganger differansen i log-likelihood er kji-kvadratfordelt med antall frihetsgrader lik forskjellen i antallet frie parametre.

6.4.3 Frihetsgrader

Definisjonen av frihetsgrader i statistikk er antallet av verdier i en studie som er frie til å variere. For eksempel, om man trenger å ta 10 ulike kurs i en grad, og det bare er 10 ulike kurs som tilbys, da har man 9 frihetsgrader. I 9 semestre kan man velge hvilket kurs man vil ta, men det siste semesteret er det bare 1 kurs igjen, og det er ikke noe valg.
Frihetsgrader er vanligvis diskutert i relasjon med kji-kvadrat og andre former for hypotesetesting i statistikk. Det er viktig å kalkulere antall frihetsgrader når man bestemmer signifikansen til en kji-kvadratstatistikk og gyldigheten til en nullhypotese.

6.4.4 Ulike mål som er oppgitt for modellene

Df
Angir antall av frihetsgrader i de ulike modellene.

Devians
For hver simulering får jeg ut deviansen for hvor godt den aktuelle modellen tilpasser det genererte datasettet. Verdiene som er oppgitt i tabellene er mean for de 10 000 deviansene for den aktuelle modellen.

AIC
For hver simulering får jeg også ut AIC for hvor godt den aktuelle modellen tilpasser det genererte datasettet. Verdiene som er oppgitt i tabellene er mean for de 10 000 AIC-verdiene for den aktuelle modellen.

Total MSE

For hver simulering får jeg ut et sett med parameterverdier for de ulike effektene. MSE beregnes for hvert sett ved å trekke fra de forventede parameterverdiene (sanne verdier) og deretter kvadrere denne differansen. Verdiene som er plottet for MSE i figurene er mean for MSE til hver parameter (basert på 10 000 simuleringer). I tabellene er totMSE et mål på summen av de ulike mean-verdiene.

Sum avvik

For hver simulering får jeg ut et sett med predikerte verdier for responsen $\hat{\boldsymbol{y}}$**.** Avviket beregnes for hvert sett ved å trekke fra den forventede responsen $\boldsymbol{y}$, (de ulike lambdaverdiene), og deretter kvadreres disse residualene. Summen av avvikene beregnes ved:

$$\text{Sum avvik} = \sum (y_i - \hat{y}_i)^2$$

Sum avvik som oppgis i tabellene baserer seg på mean for 10 000 simuleringer.

Sig

For hver simulering beregnes deviansene. For de modellene som ikke er en full APC-modell er differansen i devians beregnet for hver simulering, dvs. deviansen i full APC-modell minus deviansen i redusert modell. Jeg får dermed ut 10 000 differanser av devians. Differansene i devians kan ses på i forhold til forventningen i kji-kvadratfordelingen. For en variabel X som er kji-kvadratfordelt med k frihetsgrader, $X \sim \chi_k^2$, er forventningen til denne gitt ved:

$$E(X) = k$$

Dette vil si at dersom endringen i deviansen er større enn forventningen i kji-kvadratfordelingen, som da vil si antall frihetsgrader, så er faktoren signifikant. Utfra disse verdiene kan jeg vurdere om inkludering av alle faktorer i modellen (den fulle APC-modellen) vil gi signifikant forbedring i tilpasning av data i forhold til den reduserte modellen. Små endringer i devians mellom modellene vil ikke gi signifikante utslag, og dette indikerer at en redusert modell kan gi likså god tilpasning av dataene som en full APC-modell.

For hver simulering får jeg ut en *p*-verdi for kji-kvadratfordelingen, som indikerer om det er en signifikant forbedring mellom modellene for det aktuelle datasettet. Til sammen får jeg ut 10 000 *p*-verdier og verdien som er oppgitt under sig er mean av disse *p*-verdiene. I tabellene er det også oppgitt hvor stor prosentandel av *p*-verdiene som er mindre enn 0,05. Det er vanlig å benytte seg av et signifikansnivå på 0,05 (5 %), men andre nivåer er også hyppig benyttet, f.eks. 0,01 (1 %) og 0,10 (10 %). Et signifikansnivå gir uttrykk for hvor stor sjanse man ønsker å ta for feilaktig å forkaste en gyldig nullhypotese. Resultatet fra en hypotesetest (*p*-verdien), er statistisk signifikant dersom det er under signifikansnivået.

7. Simuleringsresultater CGLIM vs. IE

7.1 Simuleringsoppsett

Jeg benytter simuleringer hvor den sanne modellen er en full APC-modell hvor det er både alder-, periode- og kohorteffekter til stede. Det ble generert 10 000 datasett for hver modell hvor responsen var fordelt etter en gitt ligning. Ved å generere data på denne måten kan jeg vite hva de sanne alder-, periode- og kohorteffektene er. Antallet av aldersgrupper er 9, antallet av perioder er 5, og vi får dermed 13 kohorter.
Jeg estimerer alder-, periode- og kohorteffektene i hvert simulert datasett ved å benytte IE-metoden og 3 ulike CGLIM-estimatorer: en som betinger at de to første alderseffektene er like, en som betinger de to første periodeeffektene til å være like og en som betinger de to første kohorteffektene til å være like.
For IE-metoden er resultatene fra to ulike parametriseringer tatt med. IE-metoden benytter seg av effektkoding. I IE-Modell 1 kodes de siste gruppene med -1 (alder 9, periode 5 og kohort 13), mens i IE-Modell 2 kodes de første gruppene med -1 (alder 1, periode 1 og kohort 1).

For hver simuleringsmodell er ligningen som datasettene genereres fra tatt med. Videre vises det hvordan hver av alder-, periode- og kohorteffektene beregnes, samt en tabell med de sanne effektene. De sanne effektene er også vist i de påfølgende figurene. For hver ny simuleringsmodell vises det kun figurer for effekter som er endret i forhold til den originale simuleringsmodellen.
Simuleringsresultatene er tatt med i tabellform for den originale simuleringsmodellen (Modell 1) og for simuleringsmodell 7 som vedlegg. Tabellene inneholder estimater for hver koeffisient, standardavviket for hver koeffisient, og MSE for hver koeffisient. Estimatene som oppgis er gjennomsnittet for de 10 000 simulerte estimatene og MSE er gjennomsnittet for de 10 000 beregnede MSE-verdiene. Ved å sammenligne mean for de simulerte estimatene med de sanne verdiene, kan jeg vurdere graden av skjevhet (bias) for hver estimator. Standardavviket for de simulerte estimatene viser hvor mye de estimerte parametrene varierer fra gang til gang. Mindre varians relaterer til statistisk effisiens. MSE er den gjennomsnittlige kvadrerte differansen mellom de estimerte parametrene og sannheten, og tar hensyn til både bias og varians.

De samme estimatene er plottet i påfølgende figurer. Den horisontale aksen (x-aksen) angir henholdsvis aldersgruppe, periode og kohort, mens den vertikale aksen (y-aksen) angir log-koeffisientene. De sanne effektene er også plottet i figurene. Modeller som har en kurve som ligger seg nærmest opp til kurven som angir de sanne koeffisientene klarer å gjenskape de sanne effektene best. På grunn av plassbegrensning er ikke simuleringsresultatene tatt med i tabellform for alle simuleringsmodellene. Personlig synes jeg at det er lettere å lese fra figurene hvilke metoder som klarer å gjenskape de sanne effektene best, og jeg har derfor valgt å presentere kun figurer for en del av

simuleringsmodellene.

De ulike figurene er skalert likt gjennom hele kapittelet for å kunne sammenligne modellene. For alle modellene er det oppgitt ulike mål som samlet MSE for hver kategori, samt den totale MSE for modellen. I tillegg oppgis mål som devians, AIC og summen av avvik mellom de predikerte verdiene og den forventede responsen.

7.2 Modell 1: Den originale simuleringsmodellen

Det ble generert 10 000 datasett hvor responsen var fordelt etter ligningen:

$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + 0.1\sin\left(\text{periode}_{ij}\right) + 0.1\cos\left(\text{kohort}_{ij}\right) + 0.1\sin\left(10 \cdot \text{kohort}_{ij}\right)\right]\}$$

De sanne effektene for hver aldersgruppe, periode og kohort kan beregnes fra ligningene gitt i Tabell 7-1.

Tabell 7-1: Alder-, periode- og kohorteffekter i simuleringsmodell 1.

Alder-, periode- og kohorteffekter i simuleringsmodellen		
Alderseffekter ved alder a	$a = 1, \dots, 9$	$0.1(a - 5)^2$
Periodeeffekter i periode p	$p = 1, \dots, 5$	$0.1\sin(p)$
Kohorteffekter i kohort c	$c = 1, \dots, 13$	$0.1\cos(c) + 0.1\sin(10 \cdot c)$

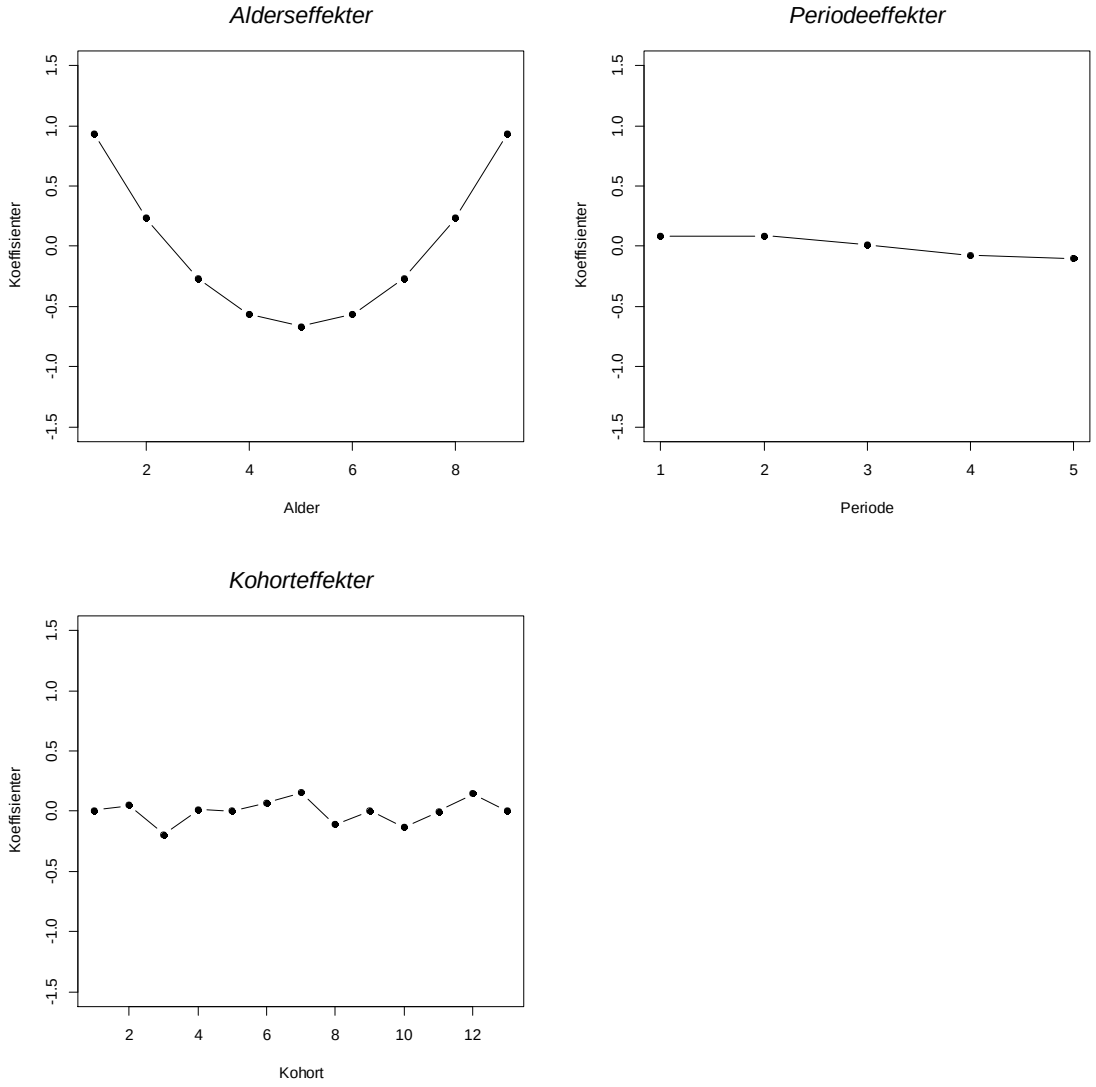
For den gitte modellen er de sanne effektene gjengitt i Tabell 7-2.

Tabell 7-2: Sanne alder-, periode- og kohorteffekter i simuleringsmodell 1.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = 0.933$	$p1 = 0.081$	$c1 = 0.002$
$a2 = 0.233$	$p2 = 0.087$	$c2 = 0.052$
$a3 = -0.267$	$p3 = 0.011$	$c3 = -0.195$
$a4 = -0.567$	$p4 = -0.079$	$c4 = 0.012$
$a5 = -0.667$	$p5 = -0.099$	$c5 = 0.005$
$a6 = -0.567$		$c6 = 0.068$
$a7 = -0.267$		$c7 = 0.155$
$a8 = 0.233$		$c8 = -0.111$
$a9 = 0.933$		$c9 = 0.001$
		$c10 = -0.132$
		$c11 = -0.001$
		$c12 = 0.145$
		$c13 = 0.000$

$intercept = 0.968$

Figur 7-1 illustrerer de sanne effektene.



Figur 7-1: Sanne effekter for henholdsvis alder, periode og kohort i simuleringsmodell 1.

Dette er den originale simuleringsmodellen, og den tar utgangspunkt i den modellen Yang et al. [2] introduserte for generering av data i sin simulering. I denne modellen er det stor forskjell på alderseffektene, mens periode- og kohorteffektene i mindre grad er ulik hverandre. I artikkelen er kun verdier oppgitt for alderseffekter, mens periode- og kohorteffektene presenteres i figurer. Det kan virke som om forfatterne benytter $p = 10, \dots, 14$ for å generere periodeeffekter i sin modell, siden kurven ikke stemmer overens med bruk av $p = 1, \dots, 5$. Grunnen til at jeg mistenker dette er at de gjør noe tilsvarende i senere arbeider, Kap 5.5 i [25]. I praksis vil det ikke bety noen forskjell hva som benyttes, da de ulike metodene vil prøve å gjenskape effektene til den simuleringsmodellen som genererte dataene.

Resultatene fra simuleringene er gjengitt i tabellene under, for to varianter av IE-metoden og 3 varianter av CGLIM-metoden. For hver kategori er det oppgitt et samlet mål for MSE nederst i tabellene.

Tabell 7-3: Simuleringsresultater fra IE- og CGLIM-estimatorer, alderseffekter.

Simuleringsresultat IE- og CGLIM-estimatorer (n=10 000), alderseffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	Intrinsic estimator IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
Alder 1	Mean	0,933	0,931	1,083	-1,944	0,947	1,381
	SD		0,260	0,283	1,278	1,269	3,362
	MSE		0,068	0,102	9,912	1,611	11,503
Alder 2	Mean	0,233	0,212	0,326	-1,944	0,224	0,550
	SD		0,295	0,312	1,278	0,996	2,535
	MSE		0,088	0,106	6,374	0,992	6,527
Alder 3	Mean	-0,267	-0,296	-0,220	-1,734	-0,288	-0,071
	SD		0,396	0,399	0,813	0,739	1,719
	MSE		0,157	0,161	2,814	0,547	2,994
Alder 4	Mean	-0,567	-0,625	-0,587	-1,344	-0,621	-0,513
	SD		0,637	0,636	0,727	0,715	1,060
	MSE		0,409	0,405	1,132	0,514	1,127
Alder 5	Mean	-0,667	-0,724	-0,724	-0,724	-0,724	-0,724
	SD		0,719	0,719	0,719	0,719	0,719
	MSE		0,520	0,520	0,520	0,520	0,520
Alder 6	Mean	-0,567	-0,601	-0,638	0,118	-0,604	-0,713
	SD		0,625	0,626	0,721	0,698	1,043
	MSE		0,392	0,397	0,989	0,489	1,109
Alder 7	Mean	-0,267	-0,246	-0,322	1,192	-0,254	-0,471
	SD		0,361	0,362	0,809	0,725	1,718
	MSE		0,131	0,134	2,783	0,526	2,994
Alder 8	Mean	0,233	0,300	0,186	2,457	0,288	-0,037
	SD		0,305	0,305	1,114	1,021	2,568
	MSE		0,097	0,095	6,184	1,045	6,669
Alder 9	Mean	0,933	1,048	0,896	3,923	1,032	0,598
	SD		0,273	0,281	1,448	1,261	3,341
	MSE		0,087	0,080	11,034	1,600	11,274
Total MSE			1,949	2,001	41,741	7,844	44,718

Tabell 7-4: Simuleringsresultater fra IE- og CGLIM-estimatorer, periodeeffekter.

Simuleringsresultat IE- og CGLIM-estimatorer (n=10 000), periodeeffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	Intrinsic estimator IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
Periode 1	Mean	0,081	0,114	0,039	1,552	0,107	-0,110
	SD		0,206	0,222	0,710	0,482	1,643
	MSE		0,044	0,051	2,670	0,233	2,735
Periode 2	Mean	0,087	0,110	0,073	0,829	0,107	-0,002
	SD		0,203	0,200	0,401	0,482	0,900
	MSE		0,042	0,040	0,711	0,233	0,818
Periode 3	Mean	0,011	0,011	0,011	0,011	0,011	0,011
	SD		0,206	0,206	0,206	0,206	0,206
	MSE		0,042	0,042	0,042	0,042	0,042
Periode 4	Mean	-0,079	-0,099	-0,061	-0,817	-0,095	0,014
	SD		0,204	0,210	0,412	0,372	0,861
	MSE		0,042	0,044	0,715	0,138	0,750
Periode 5	Mean	-0,099	-0,137	-0,061	-1,575	-0,129	0,088
	SD		0,232	0,223	0,697	0,665	1,692
	MSE		0,055	0,051	2,663	0,444	2,899
Total MSE			0,225	0,229	6,801	1,090	7,244

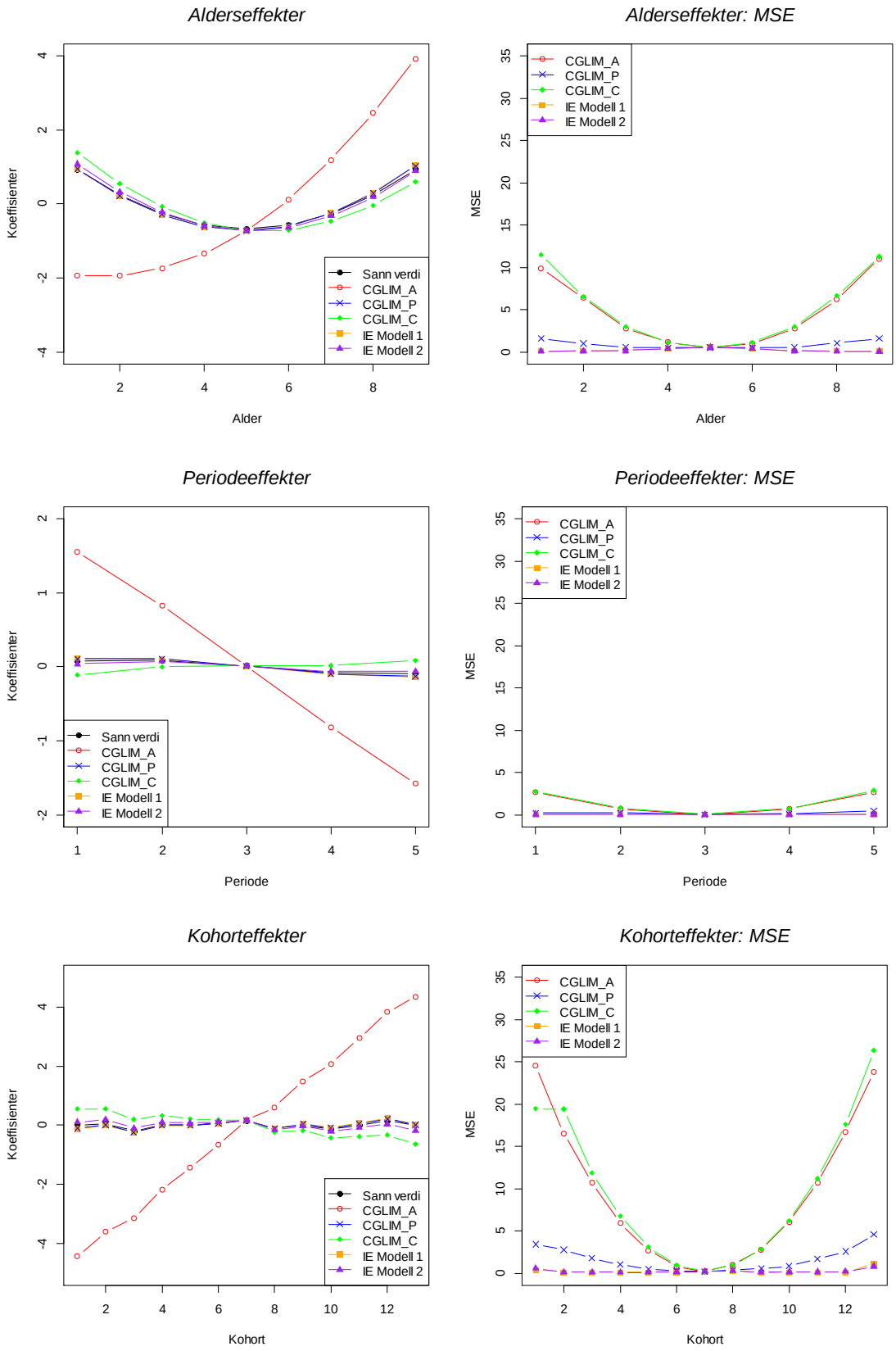
Tabell 7-5: Simuleringsresultater fra IE- og CGLIM-estimatorer, kohorteffekter.

Simuleringsresultat IE- og CGLIM-estimatorer (n=10 000), kohorteffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	Intrinsic estimator IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
Kohort 1	Mean	0,002	-0,122	0,106	-4,435	-0,098	0,553
	SD		0,616	0,737	2,211	1,850	4,377
	MSE		0,395	0,554	24,575	3,431	19,464
Kohort 2	Mean	0,052	-0,009	0,181	-3,603	0,011	0,553
	SD		0,360	0,324	1,781	1,663	4,377
	MSE		0,133	0,122	16,534	2,768	19,411
Kohort 3	Mean	-0,195	-0,252	-0,101	-3,128	-0,237	0,197
	SD		0,366	0,329	1,441	1,330	3,420
	MSE		0,137	0,117	10,677	1,770	11,846
Kohort 4	Mean	0,012	-0,011	0,103	-2,167	0,001	0,326
	SD		0,335	0,304	1,089	1,013	2,580
	MSE		0,113	0,101	5,935	1,026	6,756
Kohort 5	Mean	0,005	-0,004	0,072	-1,442	0,004	0,221
	SD		0,321	0,304	0,780	0,708	1,751
	MSE		0,103	0,097	2,701	0,501	3,113
Kohort 6	Mean	0,068	0,061	0,099	-0,658	0,065	0,174
	SD		0,385	0,377	0,515	0,504	0,969
	MSE		0,148	0,143	0,791	0,254	0,949
Kohort 7	Mean	0,155	0,161	0,161	0,161	0,161	0,161
	SD		0,465	0,465	0,465	0,465	0,465
	MSE		0,216	0,216	0,216	0,216	0,216
Kohort 8	Mean	-0,111	-0,116	-0,154	0,603	-0,120	-0,228
	SD		0,541	0,549	0,706	0,644	0,973
	MSE		0,293	0,304	1,008	0,415	0,960
Kohort 9	Mean	0,001	0,047	-0,029	1,485	0,039	-0,178
	SD		0,304	0,329	0,778	0,731	1,673
	MSE		0,095	0,109	2,807	0,536	2,832
Kohort 10	Mean	-0,132	-0,084	-0,198	2,072	-0,096	-0,421
	SD		0,328	0,370	1,101	0,913	2,475
	MSE		0,110	0,142	6,072	0,834	6,208
Kohort 11	Mean	-0,001	0,065	-0,087	2,940	0,049	-0,385
	SD		0,325	0,385	1,439	1,288	3,322
	MSE		0,110	0,156	10,723	1,661	11,183
Kohort 12	Mean	0,145	0,230	0,041	3,825	0,211	-0,332
	SD		0,341	0,409	1,785	1,606	4,171
	MSE		0,124	0,178	16,724	2,582	17,621
Kohort 13	Mean	0,000	0,034	-0,194	4,347	0,010	-0,641
	SD		1,054	0,879	2,209	2,140	5,094
	MSE		1,112	0,810	23,774	4,578	26,358
Total MSE			3,088	3,049	122,537	20,571	126,916

Tabell 7-6: Mål på de ulike modellene.

Mål på de ulike modellene					
	IE Modell 1	IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
Samlet MSE	5,262	5,279	171,079	29,506	178,879
Intercept	0,838	0,838	0,838	0,838	0,838
Frihetsgrader: 21	*Devians: 25,37*		*AIC: 189,9*		*Sum avvik: 92,2*

Resultatene er også illustrert i Figur 7-2. Den svarte kurven viser de sanne effektene.



Figur 7-2: Simuleringsresultater for IE- og CGLIM-estimatorer.

Kommentarer

I Tabell 7-3, Tabell 7-4, Tabell 7-5 og Tabell 7-6 er simuleringsresultatene gjengitt. For hver estimator vises mean, SD og MSE for de estimerte effektene, basert på alle 10 000 simuleringene. Det oppgis også mål for summert MSE for hver av alder-, periode- og kohortkategoriene, sammen med det totale MSE-målet. Figur 7-2 sammenligner mean for IE- og CGLIM-estimatene fra tabellene, og viser hvilke metoder som klarer å gjenskape profilen til de sanne alder-, periode- og kohorteffektene best. Figuren viser også MSE for hver kategori.
I CGLIM_A er betingelsen at alder 1 (0,933) er lik alder 2 (0,233), en antagelse som er uriktig i dette tilfellet, og fra figurene til alder-, periode- og kohorteffektene ser vi at CGLIM_A gir estimater som er langt unna sannheten. De andre CGLIM-modellene har betingelser som ikke er så avvikende, og gir mer riktige estimater. Differansen for de effektene som betinges å være like er 0,700 for CGLIM_A, mens den for CGLIM_P og CGLIM_C er henholdsvis 0,006 og 0,050. Tabellene og figurene viser at IE-metoden gir minst MSE av alle metodene. Selv om CGLIM_C gir estimater som er mye mer korrekt enn CGLIM_A, ser vi at den likevel gir MSE på samme størrelse som CGLIM_A. Dette skyldes at MSE også tar hensyn til spredningen i resultatene og CGLIM_C har størst spredning i sine resultater. IE-metoden har mindre variasjon i sine estimater enn alle CGLIM-modellene. Alle metodene har større MSE for de yngste og eldste kohortene. Disse kohortene er lokalisert i øvre-høyre og nedre-venstre hjørne i alder-periode tabellen og består av færrest observasjoner.
Fra Tabell 7-6 ser vi at intercept er identisk for alle variantene av IE- og CGLIM-metoden, og dette skyldes at vi sentrerer koeffisientene. Som vist i Tabell 4 i [2] vil sentrering av de ulike variantene av CGLIM-metoden føre til at intercept blir identisk for alle variantene og lik til IE-metoden. Videre i simuleringsanalysene er derfor ikke resultatene for intercept tatt med, da det ikke har noen betydning for sammenligning av metodene.

I simuleringsmodellen til Yang et al. [2] antar jeg at periodeeffektene er generert utfra $p = 10, \dots, 14$, og differansen mellom de to første periodene som betinges å være like er da 0,046. Derfor avviker de estimerte effektene mer fra sannheten for CGLIM_P-modellen i artikkelen enn i mitt simuleringsoppsett.

7.3 Modell 2: Alderseffekt endret

I denne modellen er genereringen av alderseffekter endret i forhold til Modell 1.

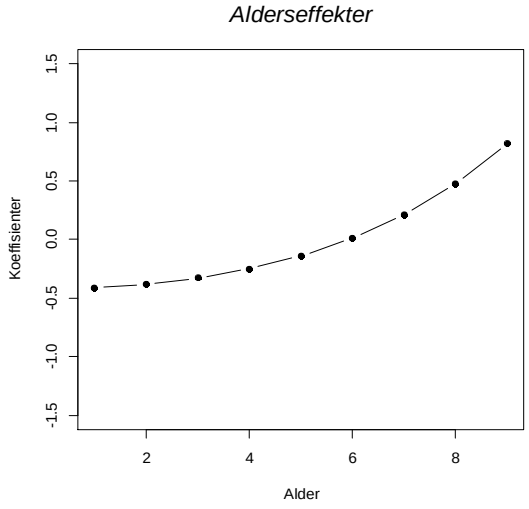
$$y_{ij} \sim \text{Poisson}\{\exp[0.3 + (\exp(0.01 \cdot \text{alder}_{ij}{}^2)) + 0.1\sin(\text{periode}_{ij}) + 0.1\cos(\text{kohort}_{ij}) + 0.1\sin(10 \cdot \text{kohort}_{ij})]\}$$

Tabell 7-7: Alder-, periode- og kohorteffekter i simuleringsmodell 2.

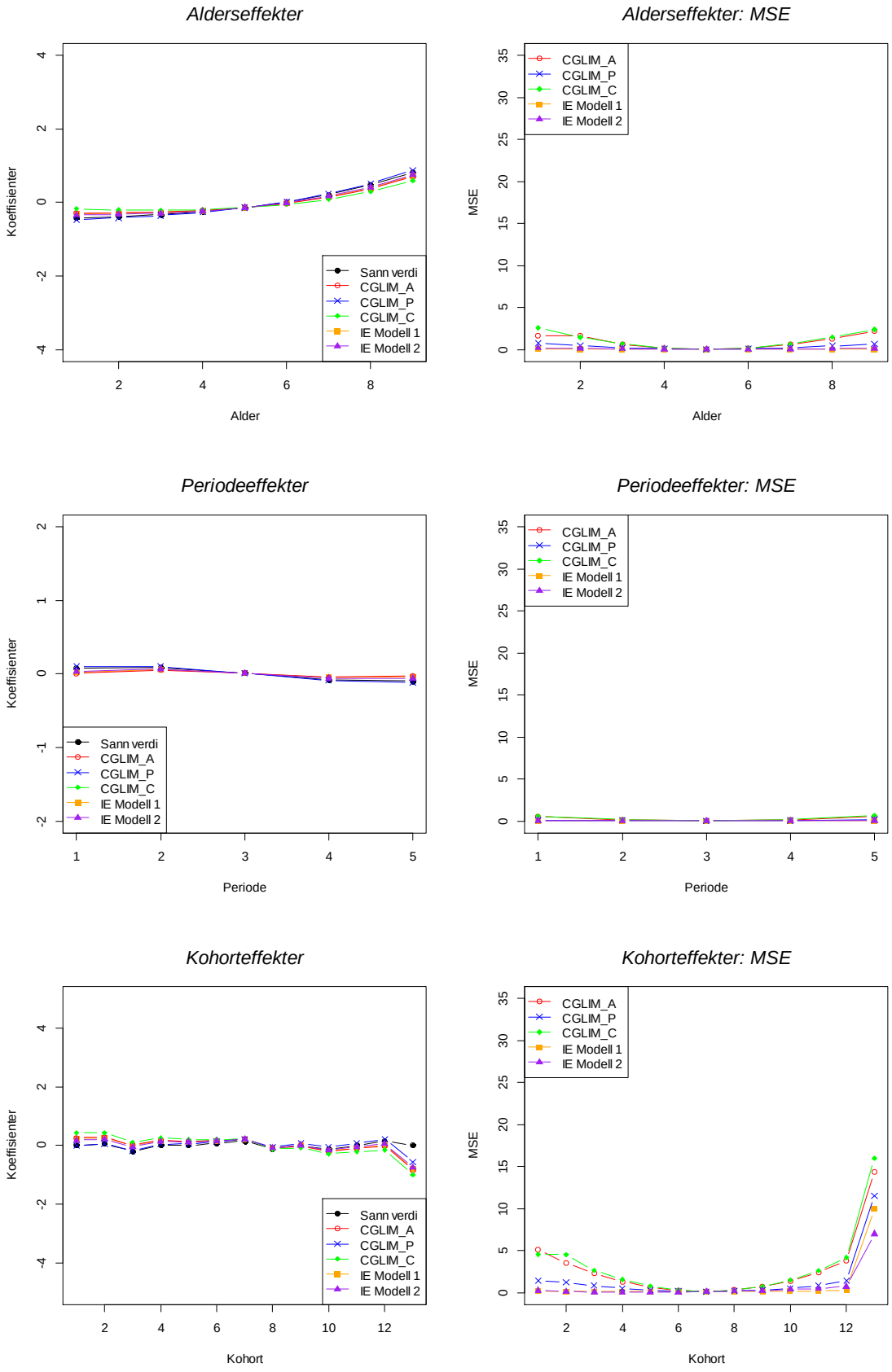
Alder-, periode- og kohorteffekter i simuleringsmodellen		
Alderseffekter ved alder a	$a = 1, \dots, 9$	$\exp(0.01 \cdot a^2)$
Periodeeffekter i periode p	$p = 1, \dots, 5$	$0.1\sin(p)$
Kohorteffekter i kohort c	$c = 1, \dots, 13$	$0.1\cos(c) + 0.1\sin(10 \cdot c)$

Tabell 7-8: Sanne alder-, periode- og kohorteffekter i simuleringsmodell 2.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = -0.414$	$p1 = 0.081$	$c1 = 0.002$
$a2 = -0.383$	$p2 = 0.087$	$c2 = 0.052$
$a3 = -0.329$	$p3 = 0.011$	$c3 = -0.195$
$a4 = -0.250$	$p4 = -0.079$	$c4 = 0.012$
$a5 = -0.140$	$p5 = -0.099$	$c5 = 0.005$
$a6 = 0.010$		$c6 = 0.068$
$a7 = 0.209$		$c7 = 0.155$
$a8 = 0.473$		$c8 = -0.111$
$a9 = 0.824$		$c9 = 0.001$
		$c10 = -0.132$
		$c11 = -0.001$
		$c12 = 0.145$
		$c13 = 0.000$



Figur 7-3: Sanne effekter for alder i simuleringsmodell 2.



Figur 7-4: Simuleringsresultater for IE- og CGLIM-estimatorer.

Tabell 7-9: Mål på de ulike modellene.

	Mål på de ulike modellene				
	IE Modell 1	IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
MSE alder	0,43	0,92	8,42	2,94	9,71
MSE periode	0,10	0,18	1,46	0,46	1,59
MSE kohort	11,77	9,76	36,15	19,19	39,68
Samlet MSE	12,30	10,87	46,02	22,59	50,98
Frihetsgrader: 21	*Devians: 22,78*		*AIC: 227,2*		*Sum avvik: 158,8*

Kommentarer

I denne simuleringsmodellen er alderseffekten endret i forhold til den originale modellen, slik at forskjellen mellom effektene til aldersgruppe 1 og aldersgruppe 2 er mindre. I CGLIM_A er betingelsen at alder 1 (-0,414) er lik alder 2 (-0,383), en antagelse som er mye mer korrekt enn i den forrige simuleringsmodellen. Fra Figur 7-4 ser vi at CGLIM_A tilpasser de sanne effektene mye bedre nå. Differansen for de effektene som betinges å være like er 0,031 for CGLIM_A, mens den for CGLIM_P og CGLIM_C er henholdsvis 0,006 og 0,050. Figurene til alder-, periode- og kohorteffektene og tabellen med MSE-verdier viser at det er IE-metoden som gir de laveste MSE-verdiene også i denne simuleringsmodellen.

7.4 Modell 3: Periodeeffekt endret versjon 1

I denne modellen er datasettene generert fra ligningen:

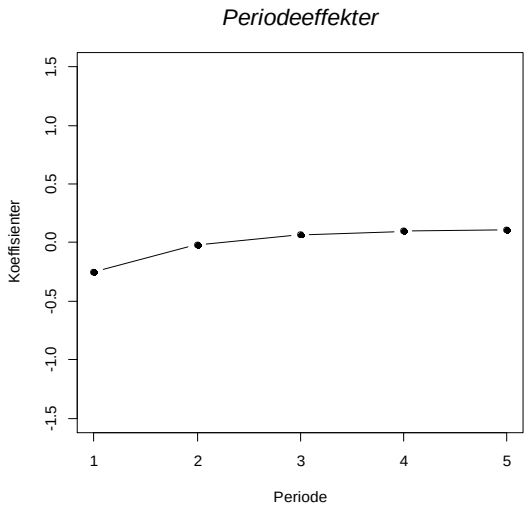
$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + (1 - \exp(-\text{periode}_{ij})) + 0.1\cos\left(\text{kohort}_{ij}\right) + 0.1\sin\left(10 \cdot \text{kohort}_{ij}\right)\right]\}$$

Tabell 7-10: Alder-, periode- og kohorteffekter i simuleringsmodell 3.

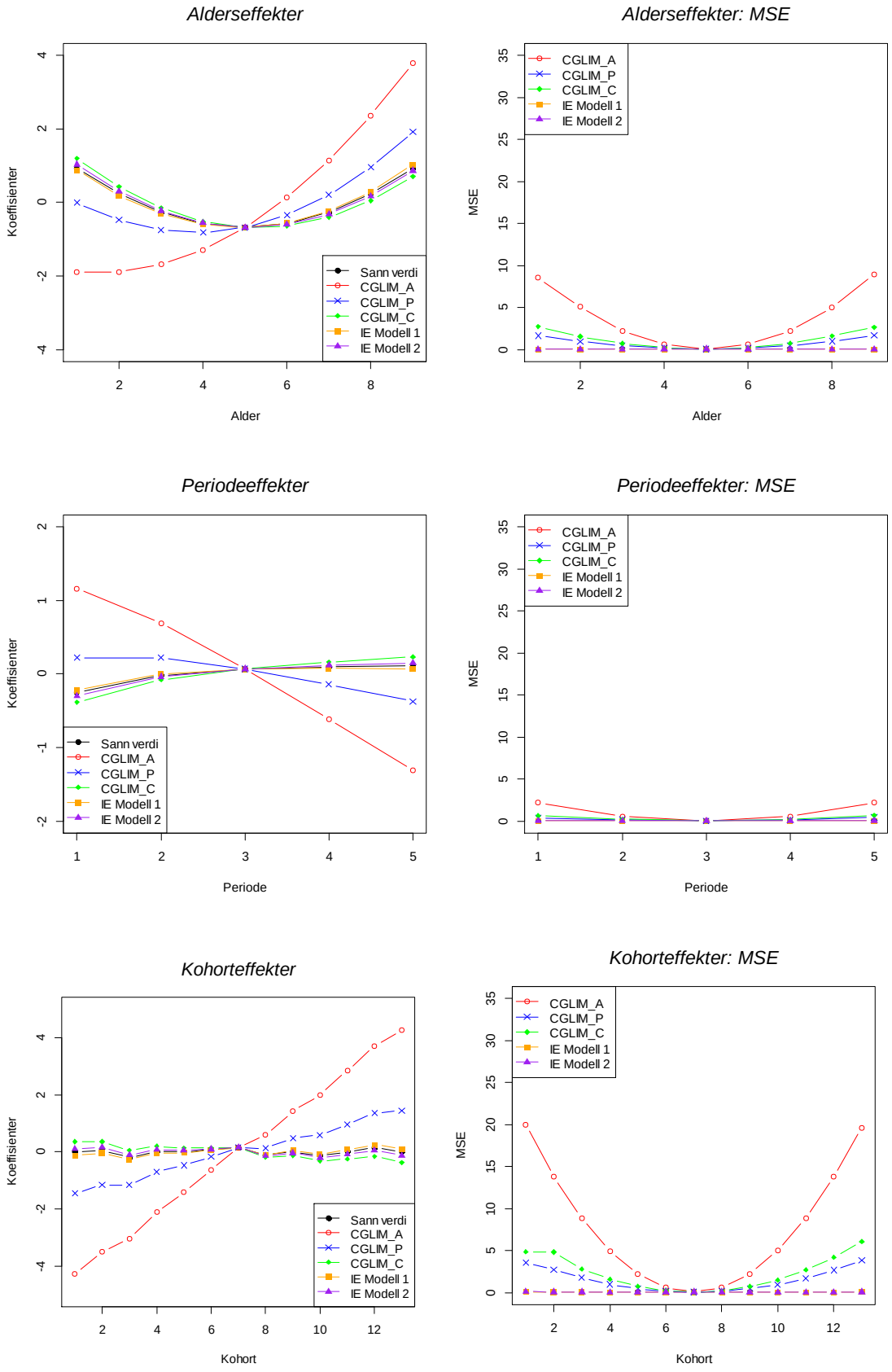
Alder-, periode- og kohorteffekter i simuleringsmodellen		
Alderseffekter ved alder a	$a = 1, \dots, 9$	$0.1(a - 5)^2$
Periodeeffekter i periode p	$p = 1, \dots, 5$	$1 - \exp(-p)$
Kohorteffekter i kohort c	$c = 1, \dots, 13$	$0.1\cos(c) + 0.1\sin(10 \cdot c)$

Tabell 7-11: Sanne alder-, periode- og kohorteffekter i simuleringsmodell 3.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = 0.933$	$p1 = -0.252$	$c1 = 0.002$
$a2 = 0.233$	$p2 = -0.020$	$c2 = 0.052$
$a3 = -0.267$	$p3 = 0.066$	$c3 = -0.195$
$a4 = -0.567$	$p4 = 0.097$	$c4 = 0.012$
$a5 = -0.667$	$p5 = 0.109$	$c5 = 0.005$
$a6 = -0.567$		$c6 = 0.068$
$a7 = -0.267$		$c7 = 0.155$
$a8 = 0.233$		$c8 = -0.111$
$a9 = 0.933$		$c9 = 0.001$
		$c10 = -0.132$
		$c11 = -0.001$
		$c12 = 0.145$
		$c13 = 0.000$



Figur 7-5: Sanne effekter for periode i simuleringsmodell 3.



Figur 7-6: Simuleringsresultater for IE- og CGLIM-estimatorer.

Tabell 7-12: Mål på de ulike modellene.

	Mål på de ulike modellene				
	IE Modell 1	IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
MSE alder	0,37	0,38	33,35	6,61	10,50
MSE periode	0,09	0,09	5,54	1,05	1,71
MSE kohort	0,69	0,70	100,55	19,48	30,49
Samlet MSE	1,15	1,17	139,43	27,14	42,70
Frihetsgrader: 21	*Devians: 22,80*		*AIC: 232,9*	*Sum avvik: 221,1*	

Kommentarer

I denne simuleringsmodellen er periodeeffekten endret i forhold til den originale modellen, slik at forskjellen mellom effektene til periode 1 og periode 2 er noe større. I CGLIM_P er betingelsen at periode 1 (-0,252) er lik periode 2 (-0,020). Differansen for de effektene som betinges å være like er 0,700 for CGLIM_A, mens den for CGLIM_P og CGLIM_C er henholdsvis 0,232 og 0,050. Fra Figur 7-6 ser vi nå at estimatene fra CGLIM_P avviker mer fra sannheten enn for simuleringsmodell 1. Figurene til alder-, periode- og kohorteffektene og tabellen med MSE-verdier viser at det er IE-metoden som har lavest MSE-verdier, og vi ser også at MSE for CGLIM_C er redusert mest i forhold til den originale modellen.

7.5 Modell 4: Periodeeffekt endret versjon 2

I denne modellen er datasettene generert fra ligningen:

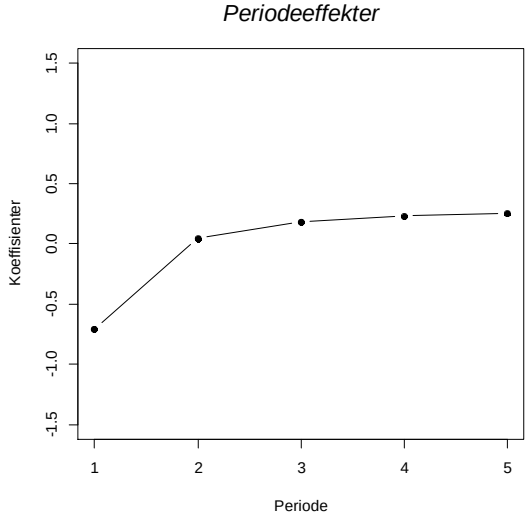
$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + \left(1.5 - \left(1/\text{periode}_{ij}{}^2\right)\right) + 0.1\cos\left(\text{kohort}_{ij}\right) + 0.1\sin\left(10 \cdot \text{kohort}_{ij}\right)\right]\}$$

Tabell 7-13: Alder-, periode- og kohorteffekter i simuleringsmodell 4.

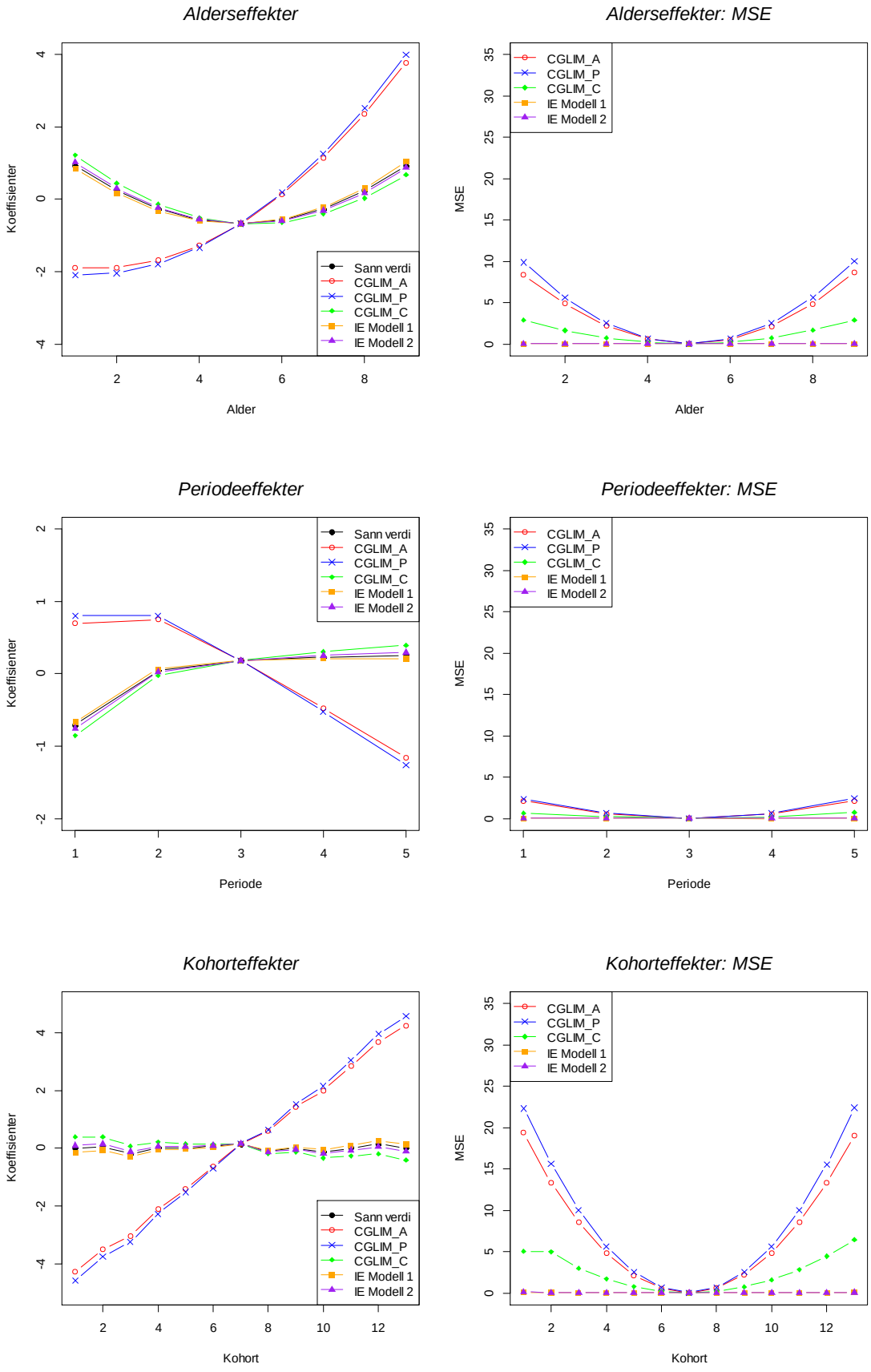
Alder-, periode- og kohorteffekter i simuleringsmodellen		
Alderseffekter ved alder a	$a = 1, \dots, 9$	$0.1(a - 5)^2$
Periodeeffekter i periode p	$p = 1, \dots, 5$	$1.5 - (1/p^2)$
Kohorteffekter i kohort c	$c = 1, \dots, 13$	$0.1\cos(c) + 0.1\sin(10 \cdot c)$

Tabell 7-14: Sanne alder-, periode- og kohorteffekter i simuleringsmodell 4.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = 0.933$	$p1 = -0.707$	$c1 = 0.002$
$a2 = 0.233$	$p2 = 0.043$	$c2 = 0.052$
$a3 = -0.267$	$p3 = 0.182$	$c3 = -0.195$
$a4 = -0.567$	$p4 = 0.230$	$c4 = 0.012$
$a5 = -0.667$	$p5 = 0.253$	$c5 = 0.005$
$a6 = -0.567$		$c6 = 0.068$
$a7 = -0.267$		$c7 = 0.155$
$a8 = 0.233$		$c8 = -0.111$
$a9 = 0.933$		$c9 = 0.001$
		$c10 = -0.132$
		$c11 = -0.001$
		$c12 = 0.145$
		$c13 = 0.000$



Figur 7-7: Sanne effekter for periode i simuleringsmodell 4.



Figur 7-8: Simuleringsresultater for IE- og CGLIM-estimatorer.

Tabell 7-15: Mål på de ulike modellene.

	Mål på de ulike modellene				
	IE Modell 1	IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
MSE alder	0,28	0,26	32,19	37,55	11,08
MSE periode	0,07	0,07	5,32	6,14	1,79
MSE kohort	0,58	0,55	97,26	113,51	32,01
Samlet MSE	0,93	0,88	134,77	157,21	44,87
Frihetsgrader: 21	*Devians: 22,36*		*AIC: 248,0*		*Sum avvik: 324,5*

Kommentarer

I denne simuleringsmodellen er periodeeffekten endret i større grad enn i simuleringsmodell 3, slik at det er enda større forskjell mellom effektene til periode 1 og periode 2. I CGLIM_P er betingelsen at periode 1 (-0,707) er lik periode 2 (0,043). Differansen for de effektene som betinges å være like er 0,700 for CGLIM_A, mens den for CGLIM_P og CGLIM_C er henholdsvis 0,750 og 0,050. Fra Figur 7-8 ser vi nå at estimatene fra CGLIM_P avviker like mye fra sannheten som estimatene fra CGLIM_A. Differansen mellom de sanne effektene til periode 1 og periode 2 er nå i samme størrelsesorden som differansen mellom aldersgruppe 1 og aldersgruppe 2. MSE for CGLIM_P er nå i samme størrelsesorden som for CGLIM_A. Som i simuleringsmodell 3, er MSE for CGLIM_C redusert endel i forhold til den originale simuleringsmodellen. IE-metoden gir lavest MSE-verdier.

Det ble også utført simuleringsanalyser for denne modellen uten kohorteffekter. Jeg ønsket å undersøke hvor godt estimatorene fungerer når det ikke er sanne periode- eller kohorteffekter tilstede. Resultatene er ikke gjengitt her siden de er tilnærmet lik resultatene fra den opprinnelige modellen. MSE-verdiene er omtrent uendret, og spredningen er den samme. IE-metoden har fremdeles lavest MSE-verdier.

7.6 Modell 5: Kohorteffekt endret versjon 1

I denne modellen er datasettene generert fra ligningen:

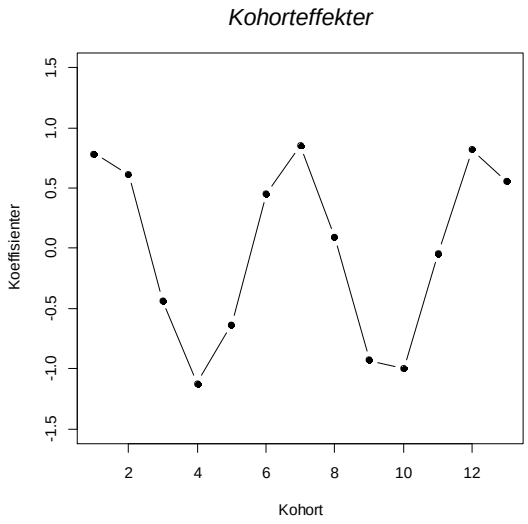
$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + 0.1\sin\left(\text{periode}_{ij}\right) + \sin\left(20 \cdot \text{kohort}_{ij}\right)\right]\}$$

Tabell 7-16: Alder-, periode- og kohorteffekter i simuleringsmodell 5.

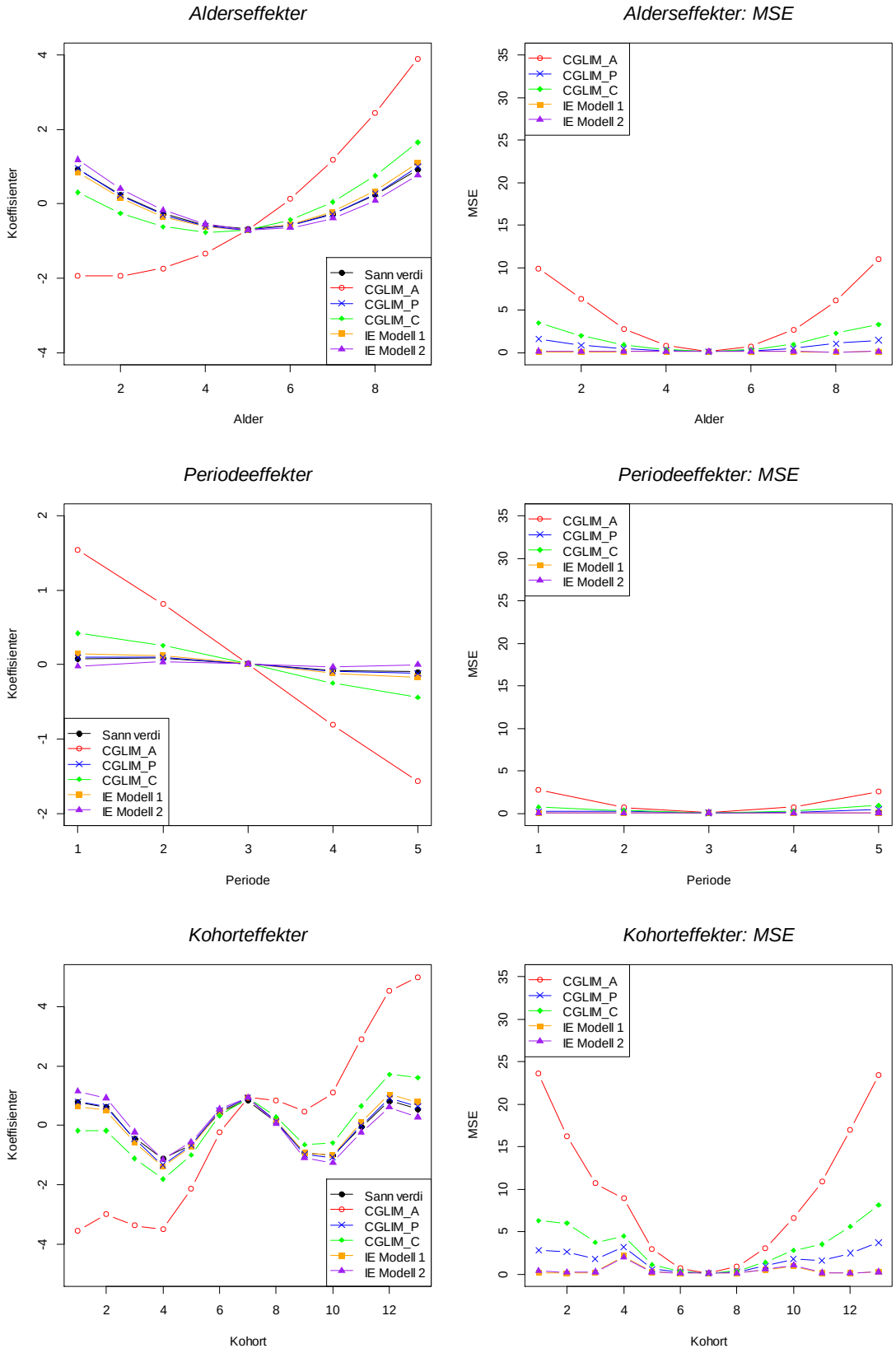
Alder-, periode- og kohorteffekter i simuleringsmodellen		
Alderseffekter ved alder a	$a = 1, \dots, 9$	$0.1(a - 5)^2$
Periodeeffekter i periode p	$p = 1, \dots, 5$	$0.1\sin(p)$
Kohorteffekter i kohort c	$c = 1, \dots, 13$	$\sin(20 \cdot c)$

Tabell 7-17: Sanne alder-, periode- og kohorteffekter i simuleringsmodell 5.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = 0.933$	$p1 = 0.081$	$c1 = 0.784$
$a2 = 0.233$	$p2 = 0.087$	$c2 = 0.616$
$a3 = -0.267$	$p3 = 0.011$	$c3 = -0.434$
$a4 = -0.567$	$p4 = -0.079$	$c4 = -1.123$
$a5 = -0.667$	$p5 = -0.099$	$c5 = -0.635$
$a6 = -0.567$		$c6 = 0.452$
$a7 = -0.267$		$c7 = 0.851$
$a8 = 0.233$		$c8 = 0.091$
$a9 = 0.933$		$c9 = -0.930$
		$c10 = -1.002$
		$c11 = -0.041$
		$c12 = 0.817$
		$c13 = 0.554$



Figur 7-9: Sanne effekter for kohort i simuleringsmodell 5.



Figur 7-10: Simuleringsresultater for IE- og CGLIM-estimatorer.

Tabell 7-18: Mål på de ulike modellene.

	Mål på de ulike modellene				
	IE Modell 1	IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
MSE alder	0,97	1,08	40,56	6,54	13,78
MSE periode	0,24	0,26	6,83	1,12	2,34
MSE kohort	5,35	5,60	125,12	22,16	43,86
Samlet MSE	6,56	6,94	172,51	29,82	59,98
Frihetsgrader: 21	*Devians: 24,57*		*AIC: 188,8*		*Sum avvik: 130,6*

Kommentarer

I denne simuleringsmodellen er kohorteffekten endret i forhold til den originale simuleringsmodellen, slik at forskjellen mellom effektene til kohort 1 og kohort 2 er noe større. I CGLIM_C er betingelsen at kohort 1 (0,784) er lik kohort 2 (0,616). Differansen for de effektene som betinges å være like er 0,700 for CGLIM_A, mens den for CGLIM_P og CGLIM_C er henholdsvis 0,006 og 0,168. Fra Figur 7-10 ser vi nå at estimatene fra CGLIM_C avviker mer fra sannheten enn for simuleringsmodell 1. Figurene til alder-, periode- og kohorteffektene og tabellen med MSE-verdier viser at det er IE-metoden som har lavest MSE-verdier, og vi ser også at MSE for CGLIM_C er redusert i forhold til den originale simuleringsmodellen. Spredningen i resultatene for CGLIM_C er lavere enn for den originale modellen, og påvirker at MSE for CGLIM_C er redusert selv om effektestimatene avviker mer fra sannheten.

7.7 Modell 6: Kohorteffekt endret versjon 2

I denne modellen er datasettene generert fra ligningen:

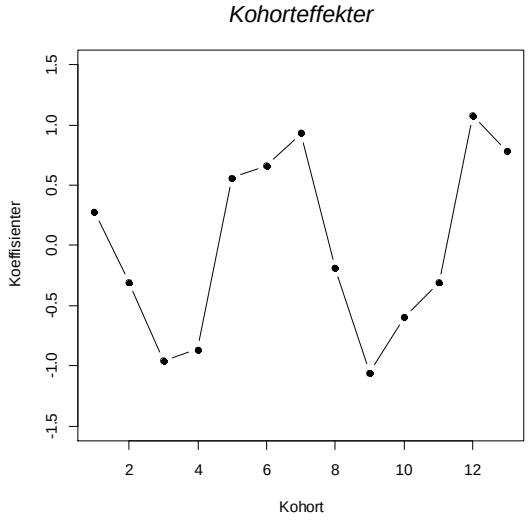
$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + 0.1\sin\left(\text{periode}_{ij}\right) + \cos\left(\text{kohort}_{ij}\right) + 0.3\cos\left(10 \cdot \text{kohort}_{ij}\right)\right]\}$$

Tabell 7-19: Alder-, periode- og kohorteffekter i simuleringsmodell 6.

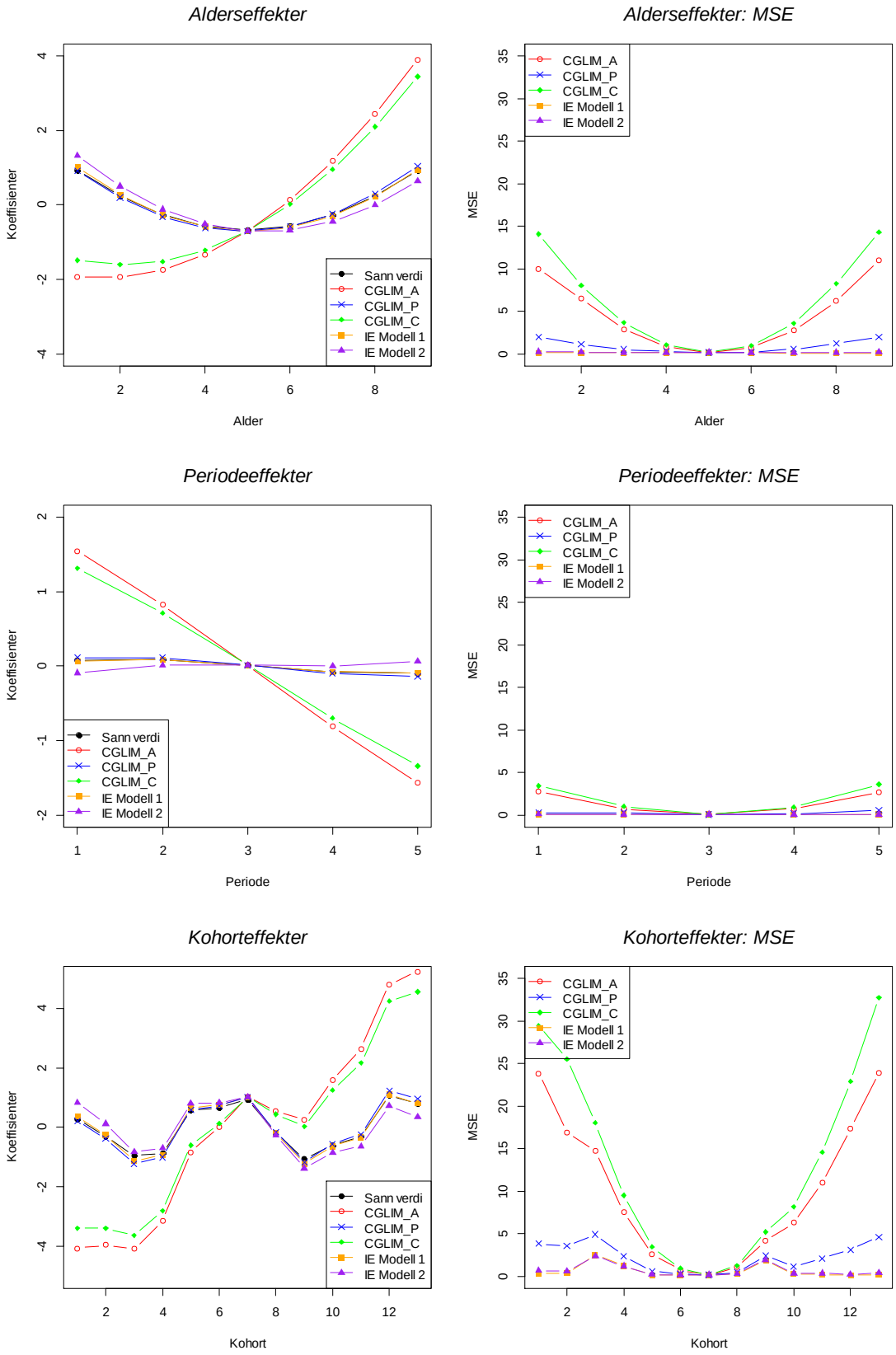
Alder-, periode- og kohorteffekter i simuleringsmodellen		
Alderseffekter ved alder a	$a = 1, \dots, 9$	$0.1(a - 5)^2$
Periodeeffekter i periode p	$p = 1, \dots, 5$	$0.1\sin(p)$
Kohorteffekter i kohort c	$c = 1, \dots, 13$	$\cos(c) + 0.3\cos(10 \cdot c)$

Tabell 7-20: Sanne alder-, periode- og kohorteffekter i simuleringsmodell 6.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = 0.933$	$p1 = 0.081$	$c1 = 0.275$
$a2 = 0.233$	$p2 = 0.087$	$c2 = -0.307$
$a3 = -0.267$	$p3 = 0.011$	$c3 = -0.957$
$a4 = -0.567$	$p4 = -0.079$	$c4 = -0.867$
$a5 = -0.667$	$p5 = -0.099$	$c5 = 0.560$
$a6 = -0.567$		$c6 = 0.661$
$a7 = -0.267$		$c7 = 0.930$
$a8 = 0.233$		$c8 = -0.192$
$a9 = 0.933$		$c9 = -1.059$
		$c10 = -0.594$
		$c11 = -0.309$
		$c12 = 1.075$
		$c13 = 0.784$



Figur 7-11: Sanne effekter for kohort i simuleringsmodell 6.



Figur 7-12: Simuleringsresultater for IE- og CGLIM-estimatorer.

Tabell 7-21: Mål på de ulike modellene.

	Mål på de ulike modellene				
	IE Modell 1	IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
MSE alder	1,00	1,46	41,01	7,94	54,08
MSE periode	0,24	0,31	6,88	1,31	9,03
MSE kohort	7,66	8,78	130,13	29,43	171,85
Samlet MSE	8,90	10,55	178,02	38,68	234,96
Frihetsgrader: 21	*Devians: 24,40*		*AIC: 187,5*		*Sum avvik: 118,5*

Kommentarer

I denne simuleringsmodellen er kohorteffekten endret i større grad enn i simuleringsmodell 5, slik at det er enda større forskjell mellom effektene til kohort 1 og kohort 2. I CGLIM_C er betingelsen at kohort 1 (0,275) er lik kohort 2 (-0,307). Differansen for de effektene som betinges å være like er 0,700 for CGLIM_A, mens den for CGLIM_P og CGLIM_C er henholdsvis 0,006 og 0,582. Fra Figur 7-12 ser vi nå at estimatene fra CGLIM_C avviker tilnærmet like mye fra sannheten som estimatene fra CGLIM_A. Differansen mellom de sanne effektene til kohort 1 og kohort 2 er nå i samme størrelsesorden som differansen mellom aldersgruppe 1 og aldersgruppe 2. MSE for CGLIM_C er nå enda større enn for CGLIM_A, og dette skyldes at CGLIM_C har større spredning i sine estimater. Igjen ser vi at det er IE-metoden som gir lavest MSE-verdier.

Det ble også utført simuleringsanalyser for denne modellen uten periodeeffekter. Resultatene er ikke gjengitt her. MSE-verdiene og spredningen i resultatene er omtrent uendret for alle modellene foruten CGLIM_C. Når periodeeffekten er borte, får CGLIM_C større spredning i sine resultater, og følgelig større MSE-verdi (Samlet MSE = 300,51). IE-metoden gir fremdeles lavest MSE-verdier.
Vi vil også se i neste simuleringsmodell at når periodeeffekten øker ytterligere, vil MSE og spredningen i resultatene til CGLIM_C reduseres enda mer. Det kan virke som at dess større periodeeffekt det er, dess mindre variasjon er det i CGLIM_C.

7.8 Modell 7: Periode- og kohorteffekt endret versjon 2

I denne modellen er datasettene generert fra ligningen:

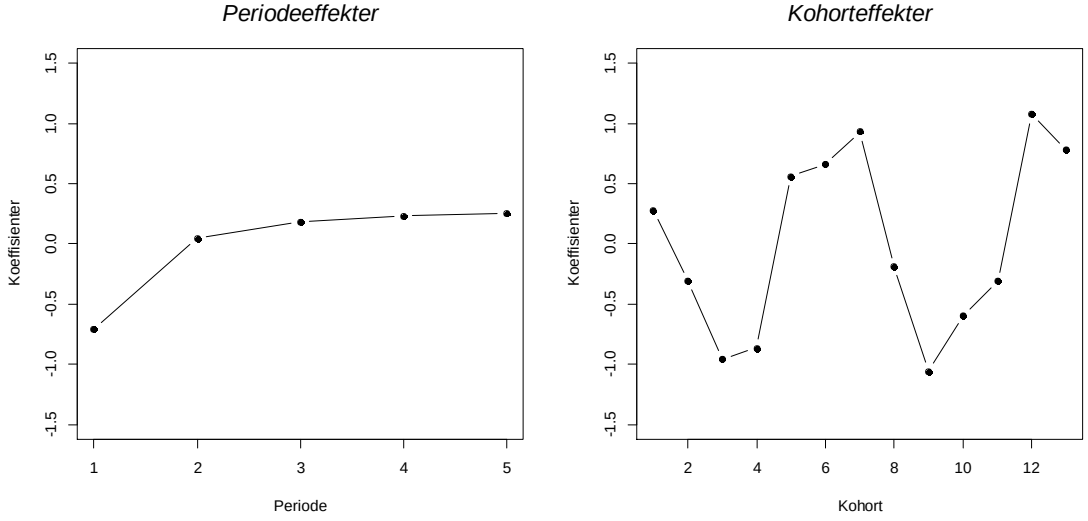
$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + (1.5 - (1/\text{periode}_{ij}{}^2)) + \cos\left(\text{kohort}_{ij}\right) + 0.3\cos\left(10 \cdot \text{kohort}_{ij}\right)\right]\}$$

Tabell 7-22: Alder-, periode- og kohorteffekter i simuleringsmodell 7.

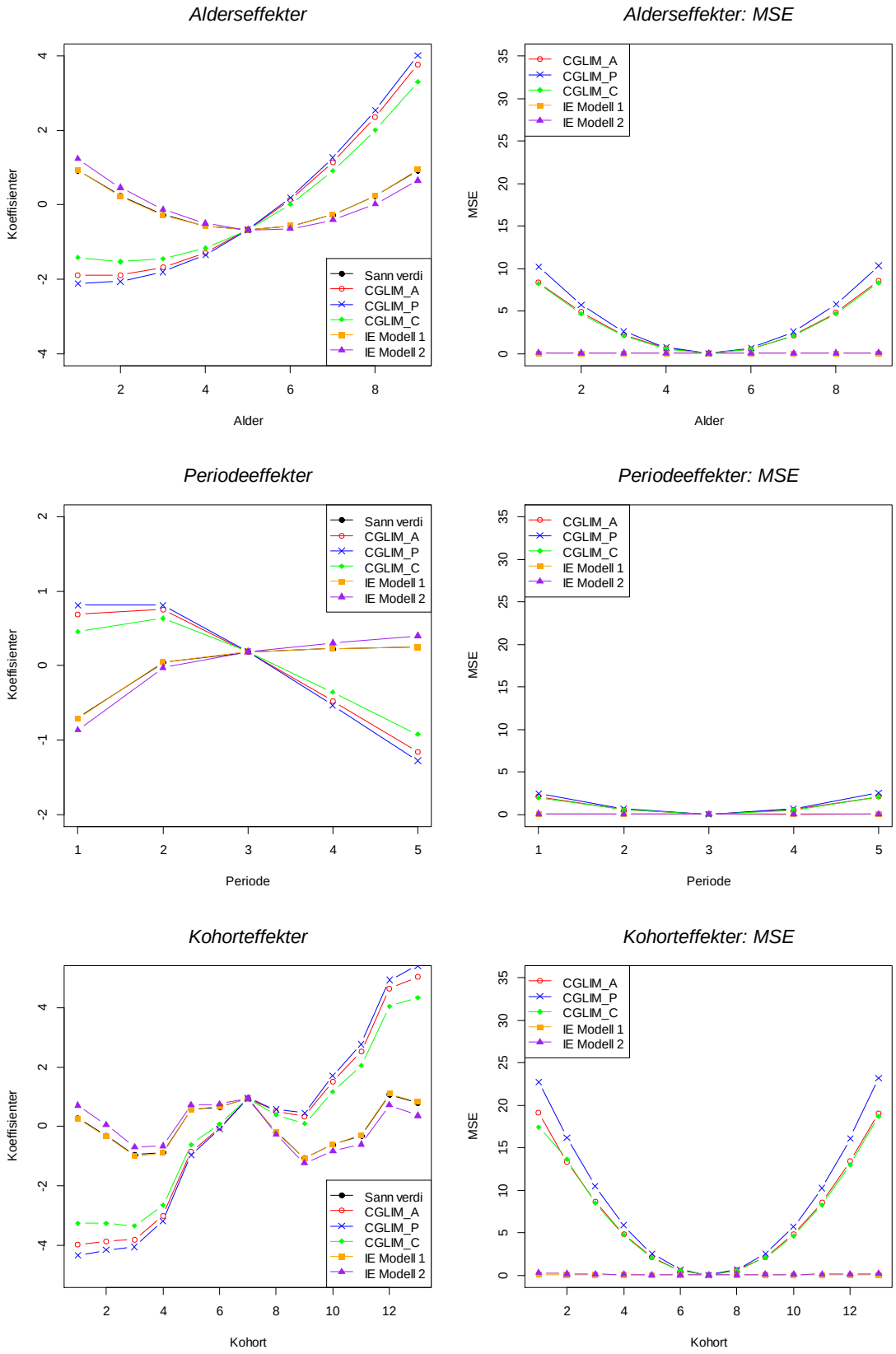
Alder-, periode- og kohorteffekter i simuleringsmodellen		
Alderseffekter ved alder a	$a = 1, \dots, 9$	$0.1(a - 5)^2$
Periodeeffekter i periode p	$p = 1, \dots, 5$	$1.5 - (1/p^2)$
Kohorteffekter i kohort c	$c = 1, \dots, 13$	$\cos(c) + 0.3\cos(10 \cdot c)$

Tabell 7-23: Sanne alder-, periode- og kohorteffekter i simuleringsmodell 7.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = 0.933$	$1 = -0.707$	$c1 = 0.275$
$a2 = 0.233$	$p2 = 0.043$	$c2 = -0.307$
$a3 = -0.267$	$p3 = 0.182$	$c3 = -0.957$
$a4 = -0.567$	$p4 = 0.230$	$c4 = -0.867$
$a5 = -0.667$	$p5 = 0.253$	$c5 = 0.560$
$a6 = -0.567$		$c6 = 0.661$
$a7 = -0.267$		$c7 = 0.930$
$a8 = 0.233$		$c8 = -0.192$
$a9 = 0.933$		$c9 = -1.059$
		$c10 = -0.594$
		$c11 = -0.309$
		$c12 = 1.075$
		$c13 = 0.784$



Figur 7-13: Sanne effekter for henholdsvis periode og kohort i simuleringsmodell 7.



Figur 7-14: Simuleringsresultater for IE- og CGLIM-estimatorer.

Tabell 7-24: Mål på de ulike modellene.

	Mål på de ulike modellene				
	IE Modell 1	IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
MSE alder	0,23	0,54	32,10	38,61	31,23
MSE periode	0,07	0,13	5,31	6,32	5,12
MSE kohort	0,53	1,47	97,12	116,99	94,11
Samlet MSE	0,82	2,13	134,53	161,92	130,46
Frihetsgrader: 21	*Devians: 22,67*		*AIC: 245,6*		*Sum avvik: 450,1*

Kommentarer

Tabeller for simuleringsresultatene til denne modellen er med i vedleggsdelen.
I denne simuleringsmodellen er både periode- og kohorteffekten endret tilsvarende som i simuleringsmodell 4 og 6. I CGLIM_P er betingelsen at periode 1 (-0,707) er lik periode 2 (0,043), og tilsvarende for CGLIM_C er betingelsen at kohort 1 (0,275) er lik kohort 2 (-0,307). Differansen for de effektene som betinges å være like er 0,700 for CGLIM_A, mens den for CGLIM_P og CGLIM_C er henholdsvis 0,750 og 0,582. Fra Figur 7-14 ser vi nå at avviket fra de sanne verdiene er tilnærmet like stort for både CGLIM_A, CGLIM_P og CGLIM_C, og MSE-verdiene for alle disse modellene er store, og i samme størrelsesorden. Spredningen i resultatene for CGLIM_C er mindre enn i simuleringsmodell 6, og det fører til at MSE-verdien blir lavere. Igjen ser vi at det er IE-metoden som gir lavest MSE-verdier og som gir estimater som er mest lik sannheten.

7.9 Resultater

IE-metoden ser ut til å være robust i alle de ulike simuleringsmodellene jeg har sett på i dette kapittelet. Den påvirkes i mye mindre grad av hvilken simuleringsmodell som velges for å generere datasettene enn de ulike variantene av CGLIM-metoden. Dette gjelder også dersom periode- eller kohorteffektene er fraværende. IE-metoden har minst varians og klarer i stor grad å produsere estimater som er tilnærmet de sanne effektene. De to parametriseringene av IE-metoden gir tilnærmet identiske estimater og mål. I alle de ulike simuleringseksemplene som er vist her, er det IE-metoden som gir lavest MSE. Men IE-metoden gir ikke nødvendigvis alltid de mest korrekte estimatene for alle parametrene. Når CGLIM-metoden har en betingelse som stemmer overens med virkeligheten, dvs. at de effektene som betinges å være lik hverandre virkelig er tilnærmet lik hverandre, kan denne metoden også gi estimater som er like gode eller bedre enn IE-metoden. Vi ser at dess større forskjell det er på de effektene som betinges å være lik hverandre i CGLIM-metoden, dess større avvik fra de sanne verdiene ser vi at estimatene får. Så når en ikke vet noe om de dataene en analyserer, vil IE-metoden i mange tilfeller kunne gi sikrere estimater enn ved valg av en av de andre CGLIM-modellene. Valg av en vilkårlig betingelse i CGLIM-metoden kan i verste fall gi estimater som er langt unna sannheten.

8. Artiklene til Clayton og Schifflers

Clayton og Schifflers sine artikler [3, 4] er sentrale innen litteraturen som omhandler modellidentifikasjonsproblemet til APC-analyser. I den første artikkelen tar de for seg modeller som beskriver variasjon over tid uten å skille mellom periodeinnflytelse og kohortinnflytelse, og omtaler begrepet *drift*. I den andre artikkelen sammenligner forfatterne ulike reduserte modeller med de fulle APC-modellene for å vurdere hvilken modell som tilpasser et gitt datasett best. De nevner også førsteordensdifferansene og andreordensdifferansene som mulige løsninger for å presentere periode- og kohorteffekter. Clayton og Schifflers forsøker ikke å løse identifikasjonsproblemet. De innfører ikke noen ekstra antagelser og benytter seg heller ikke av betingelser. Under slike omstendigheter er det kun mulig å separere ut ikke-lineære effekter til periode og kohort. I siste delen av dette kapittelet er det utført simuleringsanalyser for å sammenligne IE-metoden med Clayton og Schifflers sine metoder.

8.1 Models for temporal variation in cancer rates I: Age-Period and Age-Cohort models

8.1.1 Regulære trender: Den log-lineære drift-modellen

I artikkelen til Clayton og Schifflers [3] brukes betegnelsen log-lineære drift-modeller om modeller som beskriver variasjon over tid uten å skille mellom periodeinnflytelse og kohortinnflytelse. Forfatterne innfører begrepet *drift* for variasjon som kan beskrives like godt av en alder-periode-modell, som av en alder-kohort-modell. Artikkelen viser et eksempel med et datasett som blir beskrevet like godt av begge modellene. Først ser de på alderseffektene alene, deretter innfører de periodeeffekter i modellen, og de får resultater som indikerer at det er signifikante periodeeffekter. Tilsvarende resultater finner de når de sammenligner modellen med bare alderseffekter med modellen som både tar hensyn til alders- og kohorteffekter. Paradoksalt beskrives datasettet like godt med begge modellene, og det oppstår vanskeligheter med å tilskrive regulære trender som enten periode- eller kohortpåvirkning. Den eneste mulige løsningen på dette paradokset er at det må være en temporal variasjon i ratene som ikke skiller mellom periode- eller kohortpåvirkning. Variasjonen over tid kan ikke predikeres av hverken alder-periode-modellen eller alder-kohort-modellen. Her kommer begrepet *drift* inn for å beskrive variasjonen.

Estimering av alder- og drifteffekter. 2 mulige parametriseringer av den log-lineære driftmodellen.

Med alder og periode får vi:

$$\log Y_{ap} = \mu + \alpha_a + \delta_p(p - p_0) + \log n_{ap} \qquad (8\text{–}1)$$

Her står Y_{ap} for forventningen til antall tilfeller i aldersgruppe a og periode p, og n_{ap} er tilsvarende antall personår. I artikkelen [3] omtaler forfatterne den logaritmiske aldersspesifikke raten som Y_{ap}, og de benytter ikke intercept-ledd i sin parametrisering. I modellen står p_0 for referanseperioden, og parameteren δ_p er den konstante forandringen i lograte fra en periode til den neste. Parameteren δ_p blir gjerne omtalt som driftparameteren eller trendparameteren.

På samme måte får man en log-lineær versjon av alder-kohort-modellen:

$$\log Y_{ak} = \mu + \alpha_a + \delta_k(k - k_0) + \log n_{ak} \qquad (8\text{–}2)$$

Her står Y_{ak} for forventningen til antall tilfeller i aldersgruppe a og kohort k, og n_{ak} er tilsvarende antall personår. I modellen står k_0 for referansekohorten, og parameteren δ_k er den konstante forandringen i lograte fra en kohort til den neste. I modellene kan α_a tolkes som de tilpassede aldersspesifikke ratene for henholdsvis referanseperioden og referansekohorten.

8.1.2 Eksempel fra Clayton og Schifflers første artikkel

I artikkelen [3] tar forfatterne for seg et eksempel der de ser på forekomsten av lungekreft hos belgiske kvinner i perioden 1955 – 1978. I Tabell 8-1 er de aldersspesifikke mortalitetsratene presentert.

Tabell 8-1: Aldersspesifikke mortalitetsrater (pr. 100 000 personår) for lungekreft hos belgiske kvinner i perioden 1955 – 1978. Antall tilfeller er angitt i parentes. (Datakilde: WHO mortality database).

Alder/Periode	1955 - 1959		1960 - 1964		1965 - 1969		1970 - 1974		1975 - 1978	
25 – 29	0,19	(3)	0,13	(2)	0,50	(7)	0,19	(3)	0,70	(10)
30 – 34	0,66	(11)	0,98	(16)	0,72	(11)	0,71	(10)	0,57	(7)
35 – 39	0,78	(11)	1,32	(22)	1,47	(24)	1,64	(25)	1,32	(15)
40 – 44	2,67	(36)	3,16	(44)	2,53	(42)	3,38	(53)	3,93	(48)
45 – 49	4,84	(77)	5,60	(74)	4,93	(68)	6,05	(99)	6,83	(88)
50 – 54	6,60	(106)	8,50	(131)	7,65	(99)	10,59	(142)	10,42	(134)
55 – 59	10,36	(157)	12,00	(184)	12,68	(189)	14,34	(180)	17,95	(177)
60 – 64	14,76	(193)	16,37	(232)	18,00	(262)	17,60	(249)	23,91	(239)
65 – 69	20,53	(219)	22,60	(267)	24,90	(323)	24,33	(325)	32,70	(343)
70 – 74	26,24	(223)	27,70	(250)	30,47	(308)	36,94	(412)	38,47	(358)
75 – 79	33,47	(198)	33,61	(214)	36,77	(253)	43,69	(338)	45,20	(312)

For disse dataene er ulike modeller analysert og resultatene er samlet i Tabell 8-2.

Tabell 8-2: Lungekreft mortalitet hos belgiske kvinner. Alder-, periode- og kohorteffekter estimert med ulike modeller fra de aldersspesifikke mortalitetsratene. Goodness-of-fit til forskjellige log-lineære modeller.

	Modell				
Klasser	**Alder**	**Alder + periode**	**Alder + kohort**	**Alder + drift (periode)**	**Alder + drift (kohort)**
25 – 29	-12,615	-12,816	-13,541	-12,827	-13,442
30 – 34	-11,818	-12,007	-12,426	-12,018	-12,531
35 – 39	-11,238	-11,430	-11,854	-11,443	-11,853
40 – 44	-10,381	-10,581	-10,925	-10,595	-10,902
45 – 49	-9,786	-9,985	-10,219	-9,997	-10,202
50 – 54	-9,354	-9,548	-9,663	-9,558	-9,661
55 – 59	-8,942	-9,124	-9,143	-9,136	-9,136
60 – 64	-8,633	-8,825	-8,729	-8,838	-8,735
65 – 69	-8,298	-8,502	-8,294	-8,515	-8,310
70 – 74	-8,039	-8,250	-7,938	-8,262	-7,955
75 – 79	-7,852	-8,065	-7,671	-8,077	-7,667
1955 – 1959		0,000			
1960 – 1964		0,107			
1965 – 1969		0,162			
1970 – 1974		0,278			
1975 – 1978		0,423			
1880			-0,331		
1885			-0,317		
1890			-0,231		
1895			-0,105		
1900			0,000		
1905			0,056		
1910			0,204		
1915			0,331		
1920			0,470		
1925			0,484		
1930			0,656		
1935			0,741		
1940			0,720		
1945			0,356		
1950			1,672		
Drift				**0,103**	**0,103**
Devians	**196,6**	**38,47**	**29,67**	**42,32**	**42,32**
Frihetsgrader	**44**	**40**	**30**	**43**	**43**

Ved analysering er modellene parametrisert på samme måte som i artikkelen til Clayton og Schifflers, og inkluderer derfor ikke intercept-ledd. Første kalenderperiode (1955 - 1959) er satt som referanseperiode i modellen med periode, dvs. $p_0 = 1$. Fødselskohort nr. 5 (1900) er satt som referansekohort i modellen med kohorter, dvs. $k_0 = 5$.

Den enkleste modellen har alder som eneste forklaringsvariabel, og forutsetter at periode og kohort ikke har noen tilsynelatende virkning.

Denne modellen kan parametriseres som:

$$\log Y_{ap} = \mu + \alpha_a + \log n_{ap} \quad (8\text{–}3)$$

I alder-periode-modellen forutsetter man at en tidsvariabel, periode, har en virkning i tillegg til alder. Her står β_p for effekten av periode.

Parametrisering av modellen:

$$\log Y_{ap} = \mu + \alpha_a + \beta_p + \log n_{ap} \quad (8\text{–}4)$$

I alder-kohort-modellen forutsetter man at en tidsvariabel, kohort, har en virkning i tillegg til alder. Her står γ_k for effekten av kohort, og i denne ligningen er alder og kohort inkludert i modellen. Indeksen p blir unikt fastlagt av alderen a og kohorten k.

Parametrisering av modellen:

$$\log Y_{ap} = \mu + \alpha_a + \gamma_k + \log n_{ap} \quad (8\text{–}5)$$

Ved å se på deviansen til de ulike modellene kan en undersøke hvor god tilpasning en modell har til de oppgitte dataene. Fra Tabell 8-2 ser vi at i modellen med bare alder som faktor, er deviansen 196,6 og antall frihetsgrader er 44. Til sammenligning har modellen med både alder- og periodeeffekter en devians på 38,5 og 40 frihetsgrader. Dette indikerer at inkludering av periode som faktor i modellen gir en signifikant forbedring i tilpasning til dataene, noe som indikerer en signifikant periodeeffekt. Modellen med både alder- og kohorteffekter har en devians på 29,7 og 30 frihetsgrader, som viser at denne modellen også har god tilpasning til dataene. Det er en signifikant kohorteffekt.
Dette datasettet har like god tilpasning til alder-periode-modellen som til alder-kohort-modellen. Det er en tidseffekt til stede, men det er vanskelig å relatere denne til enten periode eller kohort. Det er her forfatterne innfører begrepet *drift*, for å beskrive denne variasjonen. I tabellen er de estimerte effektene for de ulike parametrene oppgitt. I både alder-periode-modellen og alder-kohort-modellen ser man en nesten monoton økning i mortaliteten med en gjennomsnittsrate på rundt 10 prosent per fem-års periode eller kohort. Dette antyder at en log-lineær drift-modell kan være nyttig. Deviansen til begge modellene som inkluderer drift er 42,3 med 43 frihetsgrader. De to modellene gir identiske prediksjoner og estimatene til den lineære trendkoeffisienten, δ, er den samme (0,103). I tabellen er de estimerte alderseffektene for de to modellene også gjengitt. Estimatene for $\alpha_a$vil være ulike for de to modellene, og vil avhenge av valgene for referanseperiode og referansekohort. Men det er ikke noe informasjon i dataene som kan hjelpe oss å skille mellom hvilken av de to modellene som er best. Forfatterne foreslår derfor begrepet *driftparameter* for koeffisienten δ til den log-lineære drift-modellen, siden det ikke er koblet spesifikt til

en av modellene.

I diskusjonsdelen anbefaler forfatterne at den log-lineære drift-modellen er den neste mulige modellen som vurderes etter at den enkle modellen med kun alderseffekter er benyttet. De mener at drift er de regulære trendene som ikke kan tilskrives til enten periode- eller kohortpåvirkning, og at kun når en observerer uvanlige eller plutselige endringer trenger vurdere alder-periode-modeller eller alder-kohort-modeller. Hvis det er påvirkning som fører til en plutselig endring i alle aldersgruppene samtidig, vil alder-periode-modellen sannsynligvis beskrive dataene godt. Hvis en påvirkning differensierer mellom ulike aldersgrupper, kan en alder-kohort-modell gi en bedre tilpasning. I sin andre artikkel [4] omtaler forfatterne i hvilken logisk rekkefølge de mener man skal vurdere ulike modeller ved analyse av et datasett. Ved å vurdere forbedring i tilpasning av dataene kan en vurdere hvilken modell som er best egnet for et gitt datasett. Den enkleste modellen med kun alderseffekter er nullhypotesen om at det ikke er noen tidseffekt, mens den neste modellen inneholder drift som ikke kan tilskrives til periode- eller kohortpåvirkning. Bare hvis denne modellen ikke beskriver dataene tilstrekkelig mener forfatterne at man trenger gå videre med andre og mer omfattende modeller.

8.2 Models for temporal variation in cancer rates II: Age-Period-Cohort models

I sin andre artikkel [4] tar Clayton og Schifflers for seg problemstillingen der hverken alder-periode-modellen eller alder-kohort-modellen gir en tilstrekkelig tilpassing til dataene.

I artikkelen tar forfatterne for seg et eksempel der de ser på forekomsten av brystkreft hos japanske kvinner i perioden 1955 – 1979. I Tabell 8-3 er de aldersspesifikke mortalitetsratene presentert.

Tabell 8-3: Aldersspesifikke mortalitetsrater (pr. 100 000 personår) for brystkreft hos japanske kvinner i perioden 1955 – 1979. Antall tilfeller er angitt i parentes. (Datakilde: WHO mortality database).

Alder/Periode	1955 - 1959		1960 - 1964		1965 - 1969		1970 - 1974		1975 - 1979	
25 – 29	0,44	(88)	0,38	(78)	0,46	(101)	0,55	(127)	0,68	(179)
30 – 34	1,69	(299)	1,69	(330)	1,75	(363)	2,31	(509)	2,52	(588)
35 – 39	4,01	(596)	3,90	(680)	4,11	(798)	4,44	(923)	4,80	(1056)
40 – 44	6,59	(874)	6,57	(962)	6,81	(1171)	7,79	(1497)	8,27	(1716)
45 – 49	8,51	(1022)	9,61	(1247)	9,96	(1429)	11,68	(1987)	12,51	(2398)
50 – 54	10,49	(1035)	10,80	(1258)	12,36	(1560)	14,59	(2079)	16,56	(2794)
55 – 59	11,36	(970)	11,51	(1087)	12,98	(1446)	14,97	(1828)	17,79	(2465)
60 – 64	12,03	(820)	10,67	(861)	12,67	(1126)	14,46	(1549)	16,42	(1962)
65 – 69	12,55	(678)	12,03	(738)	12,10	(878)	13,81	(1140)	16,46	(1683)
70 – 74	15,81	(640)	13,87	(628)	12,65	(656)	14,00	(900)	15,60	(1162)
75 – 79	17,97	(497)	15,62	(463)	15,83	(536)	15,71	(644)	16,52	(865)

For disse dataene er ulike modeller analysert og resultatene er samlet i Tabell 8-4.

Tabell 8-4: Brystkreft mortalitet hos japanske kvinner. Alder-, periode- og kohorteffekter estimert med ulike modeller fra de aldersspesifikke mortalitetsratene. Goodness-of-fit til forskjellige log-lineære modeller.

	Modell				
Klasser	**Alder**	**Alder + periode**	**Alder + kohort**	**Alder + drift (periode)**	**Alder + drift (kohort)**
25 – 29	-12,182	-12,305	-13,081	-12,374	-12,891
30 – 34	-10,809	-10,931	-11,569	-11,000	-11,431
35 – 39	-10,057	-10,182	-10,714	-10,253	-10,597
40 – 44	-9,523	-9,653	-10,096	-9,723	-9,981
45 – 49	-9,142	-9,274	-9,635	-9,343	-9,515
50 – 54	-8,920	-9,052	-9,316	-9,121	-9,208
55 – 59	-8,865	-8,996	-9,146	-9,065	-9,065
60 – 64	-8,902	-9,036	-9,076	-9,106	-9,020
65 – 69	-8,893	-9,031	-8,981	-9,100	-8,928
70 – 74	-8,844	-8,982	-8,879	-9,051	-8,792
75 – 79	-8,723	-8,863	-8,747	-8,931	-8,586
1955 – 1959		0,000			
1960 – 1964		-0,017			
1965 – 1969		0,050			
1970 – 1974		0,186			
1975 – 1979		0,298			
1880			0,122		
1885			0,063		
1890			-0,004		
1895			-0,023		
1900			0,000		
1905			0,107		
1910			0,229		
1915			0,362		
1920			0,486		
1925			0,576		
1930			0,637		
1935			0,682		
1940			0,807		
1945			0,978		
1950			1,183		
Drift				**0,086**	**0,086**
Devians	**1096**	**215,2**	**85,82**	**297,9**	**297,9**
Frihetsgrader	**44**	**40**	**30**	**43**	**43**

Ved analysering er modellene parametrisert på samme måte som i artikkelen til Clayton og Schifflers, og inkluderer derfor ikke intercept-ledd. Første kalenderperiode (1955 - 1959) er satt som referanseperiode i modellen med periode, dvs. $p_0 = 1$. Fødselskohort nr. 5 (1900) er satt som referansekohort i modellen med kohorter, dvs. $k_0 = 5$.

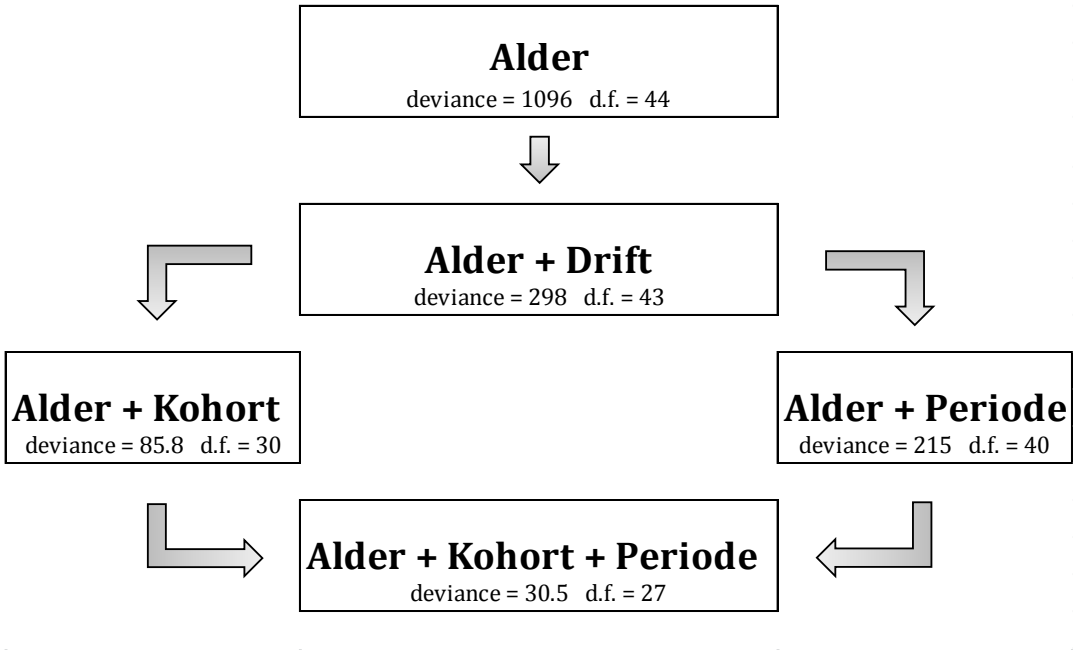
Deviansene til de ulike modellene blir benyttet som mål på hvor god tilpasning en modell har til de oppgitte dataene. Fra Tabell 8-4 ser vi at i modellen med bare alder som faktor er deviansen 1096 og antall frihetsgrader er 44. Av modellene som er vurdert er det alder-kohort-modellen som gir best tilpasning til dataene, men den gir likevel ikke en tilstrekkelig tilpasning. En devians på 86 for 30 frihetsgrader indikerer signifikans, og kan tyde på at modellen ikke er god nok.

Det blir da naturlig å vurdere en modell hvor både periode- og kohorteffekter er inkludert, og forfatterne innfører alder-periode-kohort-modellen. Med samme parametrisering som benyttet tidligere kan denne modellen skrives på formen:

$$\log Y_{ap} = \mu + \alpha_a + \beta_p + \gamma_k + \log n_{ap} \qquad (8\text{–}6)$$
$$k = A - a + p$$

der k står for diagonalene i en $(alder \times periode)$-tabell.

Ved å vurdere forbedring i tilpasning av dataene til alle tilgjengelige modeller kan en vurdere hvilken modell som er best egnet for et gitt datasett. I Figur 8-1 vises rekkefølgen av modeller som er benyttet i eksempelet fra Tabell 8-3 og Tabell 8-4.



Figur 8-1: Devians og antall frihetsgrader for ulike modeller benyttet på brystkreftdata fra japanske kvinner.

Alder-periode-kohort-modellen gir en devians på 30,5 og antallet frihetsgrader er 27. Denne modellen viser ikke signifikans og den indikerer at vi har både en periodeeffekt og en kohorteffekt i tillegg til en alderseffekt. Ved å sammenligne alder-periode-modellen med alder-periode-kohort-modellen konkluderer forfatterne med at etter å justere for periodeeffekter er kohorteffekten statistisk signifikant. På samme måte når de sammenligner alder-kohort-modellen med alder-periode-kohort-modellen etter å ha justert for kohorteffekter, er effekten for periode signifikant.

Som tidligere nevnt i teorien om APC-modeller er det et fundamentalt problem med å tolke parameterestimater i alder-periode-kohort-modeller, fordi det ikke er en unik løsning for modellen, men derimot uendelig mange løsninger. Problemet ligger i at modellen har flere parametre enn det som kan estimeres fra modellen. I Tabell II i artikkel [4] har forfatterne satt opp 3 ulike sett med parameterestimater, som alle gir den samme prediksjonen for den observerte tabellen med data. De ulike settene viser ulike trender i effektene for alder, periode og kohort. Alle de 3 settene med parameterverdier gir like god tilpasning til dataene, men de gir f.eks helt ulike alderskurver, og parametrene kan derfor ikke identifiseres.

Årsaken til disse tilsynelatende paradoksale resultatene ligger i problemet med drift, som jeg tidligere har omtalt, og som ikke kan tilskrives spesifikt til periode- eller kohorteffekter. En klarer ikke å skille mellom drift relatert til periode (δ_p) og drift relatert til kohort (δ_k), og kan derfor bare beregne en netto drift $\delta = \delta_p + \delta_k$. Driften blir beskrevet med en enkelt parameter δ i tillegg til aldersparameteren α_a. Når en innfører alder-periode-modellen, legger en til $(P-2)$ ekstra parametre som uttrykker uregelmessige periodeeffekter. På samme måte legger en til $(C-2)$ parametre når en innfører alder-kohort-modellen i forhold til alder-drift-modellen.
Da vil tilslutt alder-periode-kohort-modellen omfatte effektene: (i) *drift,* (ii) *ikke-drift periodeeffekter* og (iii) *ikke-drift kohorteffekter*.

De 3 settene med ulike parameterestimater i Tabell II i artikkel [4], representerer ulike parametriseringer av den fulle alder-periode-kohort-modellen. Disse inneholder både drifteffekter, ikke-drift periodeeffekter og ikke-drift kohorteffekter. Alle de 3 løsningene som er vist har den samme netto driften, men varierer i henhold til hvordan den er delt opp mellom periode- og kohortkomponenter. Dermed predikerer alle de 3 settene med parametre identiske rater, men foreslår ulike aldersforhold.

Clayton og Schifflers nevner flere metoder for å presentere periode- og kohorteffekter. Metodene tar utgangspunkt i en vilkårlig parametrisering av modellen, f.eks. et av settene med parameterverdier fra tabellen. Først tar de for seg metodene i forhold til periodeeffekter, og kaller parametrene fra den valgte parametriseringen for β_p. De fjerner trenden fra parametrene ved å legge til et log-lineært drift ledd og de nye parametrene β_p^* kan skrives på formen:

$$\beta_p^* = \beta_p + \delta(p - p_0) \qquad (8\text{–}7)$$

hvor en velger driftparameteren δ slik at β_p^* er uten drift og $p_0$er referanseperioden.

De nevner at Holford [10] foreslår å tolke dette slik at den lineære regresjonslinjen til parametrene β_p^* plottet mot periodene, p, skal ha 0 som stigningstall. Da vil β_p^* være identifiserbare og ikke avhenge av oppdelingen av drift.

Den første metoden som Clayton og Schifflers selv beskriver kaller de for et enklere alternativ. Denne metoden baserer seg på at drift er definert som gjennomsnittet til de påfølgende førsteordensdifferansene, som er gitt ved:

$$(\beta_2 - \beta_1), (\beta_3 - \beta_2), \ldots, (\beta_p - \beta_{p-1}) \qquad (8\text{–}8)$$

En velger δ i ligning (8–7) slik at periodekurven tvinges å komme tilbake til samme nivå som den startet på, dvs. at $\beta_1^* = \beta_p^*$. Den første perioden er ofte satt som referanse slik at $\beta_1^* = 0$, noe som da vil resultere i at også $\beta_p^* = 0$. Denne egenskapen er en fordel i forhold til beregninger. Men riktig tolking av parameterverdier som estimeres kan være problematisk med denne metoden. β_p^* bør tolkes som:

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

De samme betraktninger gjelder også for kohorteffekter. Førsteordensdifferansene blir da:

$$(\gamma_2 - \gamma_1), (\gamma_3 - \gamma_2), \ldots, (\gamma_k - \gamma_{k-1}) \qquad (8\text{–}9)$$

For å fjerne trenden fra kohortkurven innføres parametrene γ_k^* som er gitt ved:

$$\gamma_k^* = \gamma_k + \delta(k - k_0) \qquad (8\text{–}10)$$

hvor en velger driftparameteren δ slik at γ_k^* er uten drift og $k_0$er referansekohorten. En velger δ slik at kohortkurven tvinges å komme tilbake til samme nivå som den startet på, dvs. at $\gamma_1^* = \gamma_k^*$. På samme måte som for periode-modellen er tolkingen av parametrene γ_k^* ikke helt enkel.

En annen metode Clayton og Schifflers omtaler er en metode som tar for seg ikke-drift effekter. Ikke-drift effekter virker på en slik måte at forholdet mellom de estimerte ratene til tilstøtende perioder ikke er identisk. Forholdet mellom estimerte rater omtales som relativ risiko (RR), og er mye benyttet i epidemiologi. Relativ risiko er et forholdstall som angir hvor mye større sannsynlighet det er for en hendelse i én gruppe i forhold til en annen. Ikke-drift effekter kan derfor uttrykkes som forskjeller mellom relative risikoer. For periodeeffekter kan forholdet mellom to påfølgende relative risikoer skrives på formen:

$$\frac{\exp(\beta_3)/\exp(\beta_2)}{\exp(\beta_2)/\exp(\beta_1)}, \ldots, \frac{\exp(\beta_p)/\exp(\beta_{p-1})}{\exp(\beta_{p-1})/\exp(\beta_{p-2})} \qquad (8\text{–}11)$$

Og tilsvarende for kohorteffekter får vi:

$$\frac{\exp(\gamma_3)/\exp(\gamma_2)}{\exp(\gamma_2)/\exp(\gamma_1)}, \ldots, \frac{\exp(\gamma_p)/\exp(\gamma_{p-1})}{\exp(\gamma_{p-1})/\exp(\gamma_{p-2})} \qquad (8\text{–}12)$$

For alle de 3 ulike parametriseringene i Tabell II i artikkel [4] er forholdene identiske, og er derfor uavhengig av hvilken parametrisering som velges i utgangspunktet. Et lite utdrag fra tabellen i artikkelen er vist under i Tabell 8-5.

Tabell 8-5: Utdrag fra Tabell II i Clayton and Schifflers II: Three sets of age, period and cohort effects that give identical best fitting expected rates.

Sett nr.	**1**	**2**	**3**
Kohort $(\exp(\gamma_k) \times 100)$			
1880	190,3	149,7	117,8
1885	162,0	135,3	113,0
1890	133,9	118,7	105,3

For de 3 første kohortene er forholdene fra eksempelet gitt ved:

$$\frac{\exp(\gamma_3)/\exp(\gamma_2)}{\exp(\gamma_2)/\exp(\gamma_1)} = \frac{1{,}339/1{,}620}{1{,}620/1{,}903} = \frac{1{,}187/1{,}353}{1{,}353/1{,}497} = \frac{1{,}053/1{,}130}{1{,}130/1{,}178} = 0{,}971$$

Den relative risikoen for kohort 3 versus kohort 2 er i dette tilfellet mindre enn den for kohort 2 versus kohort 1. En får et slags mål på akselerasjonen av kohorttrenden rundt kohort 2.
På logaritmisk skala er disse forholdene andreordensdifferanser.

$$\log\frac{\exp(\gamma_3)/\exp(\gamma_2)}{\exp(\gamma_2)/\exp(\gamma_1)} = \ (\gamma_3 - \gamma_2) - (\gamma_2 - \gamma_1) = \ \gamma_3 - 2\gamma_2 + \gamma_1$$

Andreordensdifferansene er identifiserbare og er uavhengig av den valgte parametriseringen. Vi får følgende andreordensdifferanser rundt kohort 2 til kohort $k-1$:

$$(\gamma_3 - 2\gamma_2 + \gamma_1), \dots, (\gamma_k - 2\gamma_{k-1} + \gamma_{k-2}) \qquad (8\text{–}13)$$

Det er ikke mulig å beregne andreordensdifferanser for den første parameteren eller den siste parameteren, hhv. $k = 1$ og $k = K$. Andreordensdifferansene defineres med nye parametre som:

$$\gamma_k^* = \gamma_{k+1} - 2\gamma_k + \gamma_{k-1}, \qquad k = 2, 3, \dots, (K-1) \qquad (8\text{–}14)$$

Tilsvarende for periode og alder får vi også

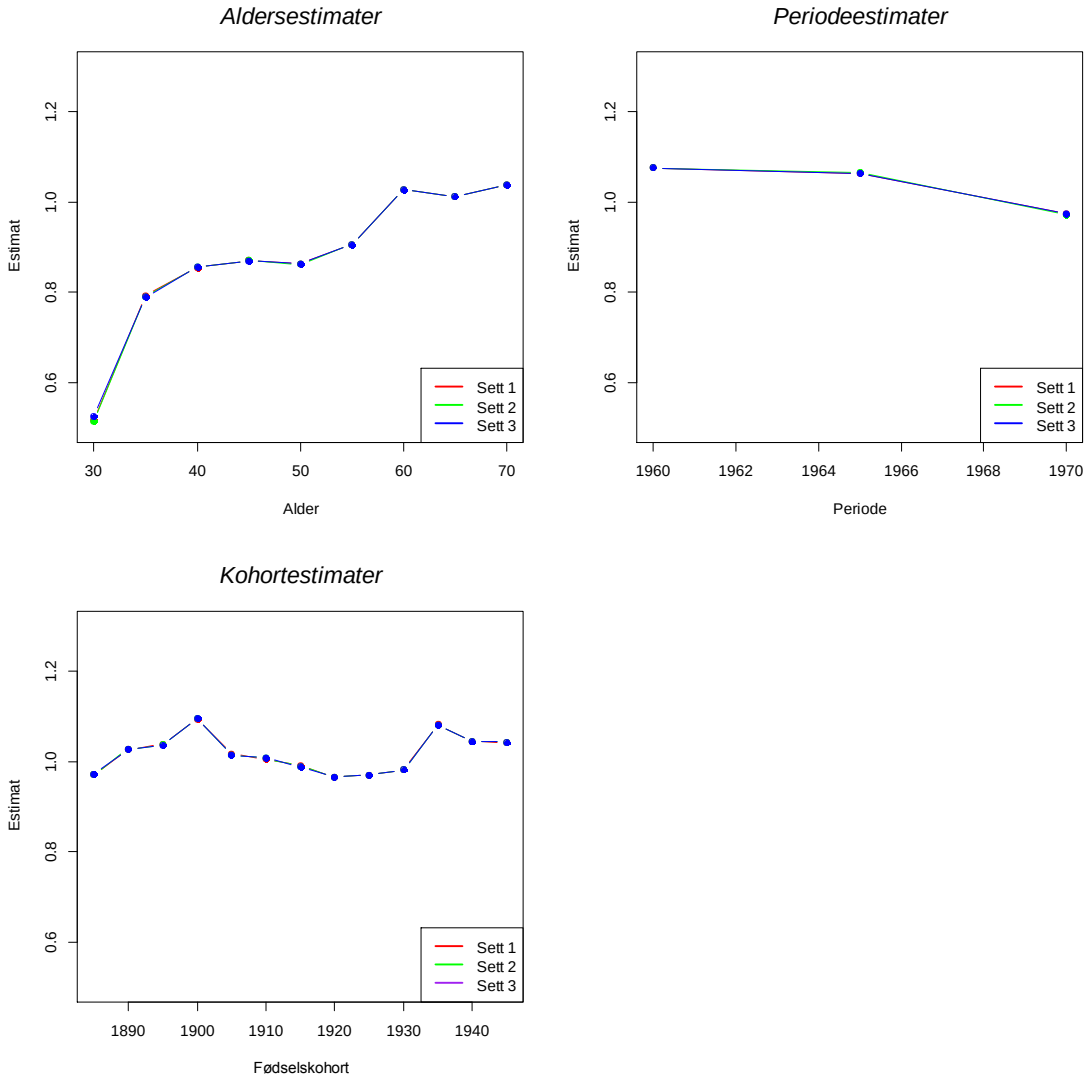
$$\beta_p^* = \beta_{p+1} - 2\beta_p + \beta_{p-1}, \qquad p = 2, 3, \dots, (P-1)$$
$$\alpha_a^* = \alpha_{a+1} - 2\alpha_a + \alpha_{a-1}, \qquad a = 2, 3, \dots, (A-1)$$

Verdiene til andreordensdifferansene blir kun påvirket av dataene som er ved siden av, noe som er en viktig praktisk fordel.

Tabell 8-6: Estimerte parametre for andreordensdifferansene til dataene fra sett 1 gitt i Tabell II i Clayton and Schifflers II.

Alder	$\exp(\alpha_a^*)$	Periode	$\exp(\beta_p^*)$	Kohort	$\exp(\gamma_a^*)$
25 – 29	-	1955 - 1959	-	1875 – 1884	-
30 – 34	0,516	1960 – 1964	1,076	1880 – 1889	0,971
35 – 39	0,792	1965 – 1969	1,062	1885 – 1894	1,026
40 – 44	0,855	1970 – 1974	0,974	1890 – 1899	1,038
45 – 49	0,870	1975 – 1979	-	1895 – 1904	1,094
50 – 54	0,863			1900 – 1909	1,016
55 – 59	0,905			1905 – 1914	1,006
60 – 64	1,027			1910 – 1919	0,990
65 – 69	1,012			1915 – 1924	0,965
70 – 74	1,038			1920 – 1929	0,970
75 – 79	-			1925 – 1934	0,982
				1930 – 1939	1,081
				1935 – 1944	1,044
				1940 – 1949	1,042
				1945 – 1954	-

De estimerte parametrene for andreordensdifferansene til dataene gitt i Tabell II i artikkel [4] er vist i Figur 8-2. Resultater for hver av de 3 settene med effekter fra tabellen er vist hver for seg. Estimatene for andreordensdifferansene er like og de 3 kurvene ligger derfor oppå hverandre.



Figur 8-2: Brystkreft mortalitet hos japanske kvinner: estimater av identifiserbare ikke-drift effekter.

Den estimerte parameterverdien for andreordensdifferansen $\exp(\beta_2^*) = 1{,}076$ kan tolkes som at den relative risikoen for periode 3 versus periode 2 er 8 prosent høyere enn den relative risikoen for periode 2 versus periode 1. Den estimerte parameterverdien for andreordensdifferansen $\exp(\gamma_{10}^*) = 0{,}970$ kan tolkes som at den relative risikoen for kohort 11 versus kohort 10 er 97 prosent av den relative risikoen for kohort 10 versus kohort 9.

8.3 Simulering med metoder fra Clayton og Schifflers artikler, samt IE

Tilsvarende som i Kapittel 7 har jeg generert data ved hjelp av en gitt simuleringsmodell, slik at jeg vet hva de sanne alder-, periode- og kohorteffektene er. Jeg har generert 10 000 datasett for hver simuleringsmodell. Fra hvert simulert datasett har jeg estimert alder-, periode- og/eller kohorteffekter ved hjelp av ulike modeller. Jeg har fått estimater fra modellen som kun tar hensyn til alderseffekter (A), modellen som tar hensyn til både alder- og periodeeffekter (AP), modellen som tar hensyn til både alder- og kohorteffekter (AC), samt de to modellene som inkluderer drift. I tillegg har jeg beregnet estimater med IE-metoden og metoden med førsteordensdifferansene. Jeg har presentert resultater for IE-metoden som benytter effektkoding for aldersgruppe 9, periode 5 og kohort 13. I den ene varianten av metoden med førsteordensdifferanser er første og siste periode satt lik hverandre (P1=P5). Da tvinges periodekurven til å returnere til det samme nivået den startet på. Tilsvarende i den andre varianten av metoden med førsteordensdifferanser er første og siste kohort satt lik hverandre (C1=C13). Da tvinges kohortkurven til å returnere til det samme nivået den startet på. For å kunne sammenligne estimatene med tidligere simuleringer benyttes også her sentrerte verdier.
Modell 1, 4, 6 og 7 fra Kapittel 7 er tatt med her, og i tillegg blir to nye simuleringsmodeller introdusert. På grunn av plassbegrensninger er det også i dette kapittelet kun gjengitt estimater i tabellform for den første simuleringsmodellen. Ellers blir estimatene presentert som kurver i figurer tilsvarende i Kapittel 7. For å unngå å ha altfor mange kurver i hver figur er det estimatene for de fulle APC-modellene som presenteres. Unntaket er at estimatene for alderseffektene er presentert for modellene som inkluderer alder og en driftparameter. Ellers er ikke estimatene for effektkoeffisientene tatt med for de reduserte modellene.
Ulike mål for de reduserte modellene er presentert i tabellform for å kunne sammenlignes mot de fulle APC-modellene. Ulike goodness-of-fit-mål vil kunne hjelpe i vurderingen av om inkludering av alle faktorer i modellen vil gi signifikant forbedring i tilpasning av data.
Andreordensdifferansene er også presentert for de fulle APC-modellene.

8.3.1 Den originale simuleringsmodellen (Modell 1)

Datasettene er generert fra ligningen:

$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + 0.1\sin\left(\text{periode}_{ij}\right) + 0.1\cos\left(\text{kohort}_{ij}\right) + 0.1\sin\left(10 \cdot \text{kohort}_{ij}\right)\right]\}$$

Tabellene under viser koeffisientestimatene for de ulike metodene. I kolonnen etter IE-metoden gjengis estimatene for alder-drift-modellen som baserer seg på periode. Deretter er det alder-drift-modellen som baserer seg på kohort. De to siste kolonnene er førsteordensdifferansene som baserer seg på henholdsvis periode og kohort.

Tabell 8-7: Simuleringsresultater fra IE-metoden, og modellene som inkluderer drift og førsteordensdifferanser, alderseffekter.

Simuleringsresultat IE-metoden og modellene som inkluderer drift og førsteordensdifferanser (n=10 000), alderseffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	Alder + drift (periode)	Alder + drift (kohort)	1.diff (P1=P5)	1.diff (C1=C13)
Alder 1	Mean	0,933	0,931	0,973	1,136	1,183	0,983
	SD		0,260	0,218	0,324	0,426	0,501
	MSE		0,068	0,049	0,146	0,244	0,254
Alder 2	Mean	0,233	0,212	0,229	0,351	0,401	0,251
	SD		0,295	0,277	0,331	0,431	0,468
	MSE		0,088	0,077	0,123	0,214	0,219
Alder 3	Mean	-0,267	-0,296	-0,276	-0,195	-0,170	-0,270
	SD		0,396	0,384	0,403	0,449	0,462
	MSE		0,157	0,148	0,168	0,211	0,213
Alder 4	Mean	-0,567	-0,625	-0,595	-0,554	-0,562	-0,612
	SD		0,637	0,612	0,615	0,646	0,655
	MSE		0,409	0,376	0,379	0,417	0,431
Alder 5	Mean	-0,667	-0,724	-0,685	-0,685	-0,724	-0,724
	SD		0,719	0,695	0,695	0,719	0,719
	MSE		0,520	0,484	0,484	0,520	0,520
Alder 6	Mean	-0,567	-0,601	-0,577	-0,617	-0,663	-0,613
	SD		0,625	0,607	0,610	0,639	0,637
	MSE		0,392	0,369	0,375	0,417	0,408
Alder 7	Mean	-0,267	-0,246	-0,262	-0,343	-0,372	-0,272
	SD		0,361	0,346	0,368	0,413	0,440
	MSE		0,131	0,120	0,142	0,181	0,193
Alder 8	Mean	0,233	0,300	0,242	0,121	0,111	0,261
	SD		0,305	0,281	0,333	0,417	0,475
	MSE		0,097	0,079	0,123	0,189	0,227
Alder 9	Mean	0,933	1,048	0,950	0,788	0,796	0,996
	SD		0,273	0,220	0,326	0,422	0,512
	MSE		0,087	0,049	0,128	0,197	0,266
Total MSE			1,949	1,749	2,067	2,590	2,731
Drift				-0,041	-0,041		

Modellene som inkluderer drift har kun estimater for alderseffektene.

Tabell 8-8: Simuleringsresultater fra IE-metoden, og modellene med førsteordensdifferanser, periodeeffekter.

		Simuleringsresultat IE-metoden og modellene med førsteordensdifferanser (n=10 000), periodeeffekter					
		Sann verdi	Intrinsic estimator IE Modell 1	Alder + drift (periode)	Alder + drift (kohort)	1.diff (P1=P5)	1.diff (C1=C13)
Periode 1	Mean	0,081	0,114			-0,011	0,089
	SD		0,206			0,130	0,279
	MSE		0,044			0,025	0,078
Periode 2	Mean	0,087	0,110			0,048	0,098
	SD		0,203			0,211	0,232
	MSE		0,042			0,046	0,054
Periode 3	Mean	0,011	0,011			0,011	0,011
	SD		0,206			0,206	0,206
	MSE		0,042			0,042	0,042
Periode 4	Mean	-0,079	-0,099			-0,036	-0,086
	SD		0,204			0,220	0,237
	MSE		0,042			0,050	0,056
Periode 5	Mean	-0,099	-0,137			-0,011	-0,111
	SD		0,232			0,130	0,288
	MSE		0,055			0,025	0,083
Total MSE			0,225			0,189	0,314

Tabell 8-9: Simuleringsresultater fra IE-metoden, og modellene med førsteordensdifferanser, kohorteffekter.

		Simuleringsresultat IE-metoden og modellene med førsteordensdifferanser (n=10 000), kohorteffekter					
		Sann verdi	Intrinsic estimator IE Modell 1	Alder + drift (periode)	Alder + drift (kohort)	1.diff (P1=P5)	1.diff (C1=C13)
Kohort 1	Mean	0,002	-0,122			0,256	-0,044
	SD		0,616			0,807	0,593
	MSE		0,395			0,716	0,354
Kohort 2	Mean	0,052	-0,009			0,306	0,056
	SD		0,360			0,530	0,635
	MSE		0,133			0,345	0,403
Kohort 3	Mean	-0,195	-0,252			-0,001	-0,201
	SD		0,366			0,496	0,553
	MSE		0,137			0,283	0,306
Kohort 4	Mean	0,012	-0,011			0,178	0,028
	SD		0,335			0,419	0,447
	MSE		0,113			0,203	0,200
Kohort 5	Mean	0,005	-0,004			0,122	0,022
	SD		0,321			0,397	0,381
	MSE		0,103			0,171	0,145
Kohort 6	Mean	0,068	0,061			0,124	0,074
	SD		0,385			0,406	0,396
	MSE		0,148			0,168	0,157
Kohort 7	Mean	0,155	0,161			0,161	0,161
	SD		0,465			0,465	0,465
	MSE		0,216			0,216	0,216
Kohort 8	Mean	-0,111	-0,116			-0,179	-0,129
	SD		0,541			0,563	0,570
	MSE		0,293			0,322	0,326
Kohort 9	Mean	0,001	0,047			-0,079	0,021
	SD		0,304			0,394	0,419
	MSE		0,095			0,161	0,176
Kohort 10	Mean	-0,132	-0,084			-0,273	-0,123
	SD		0,328			0,440	0,510
	MSE		0,110			0,213	0,260
Kohort 11	Mean	-0,001	0,065			-0,187	0,013
	SD		0,325			0,486	0,604
	MSE		0,110			0,271	0,365
Kohort 12	Mean	0,145	0,230			-0,084	0,165
	SD		0,341			0,539	0,711
	MSE		0,124			0,343	0,505
Kohort 13	Mean	0,000	0,034			-0,344	-0,044
	SD		1,054			1,095	0,593
	MSE		1,112			1,317	0,354
Total MSE			3,088			4,729	3,766

Ulike mål for de fulle APC-modellene og de reduserte modellene presenteres i tabellene under.

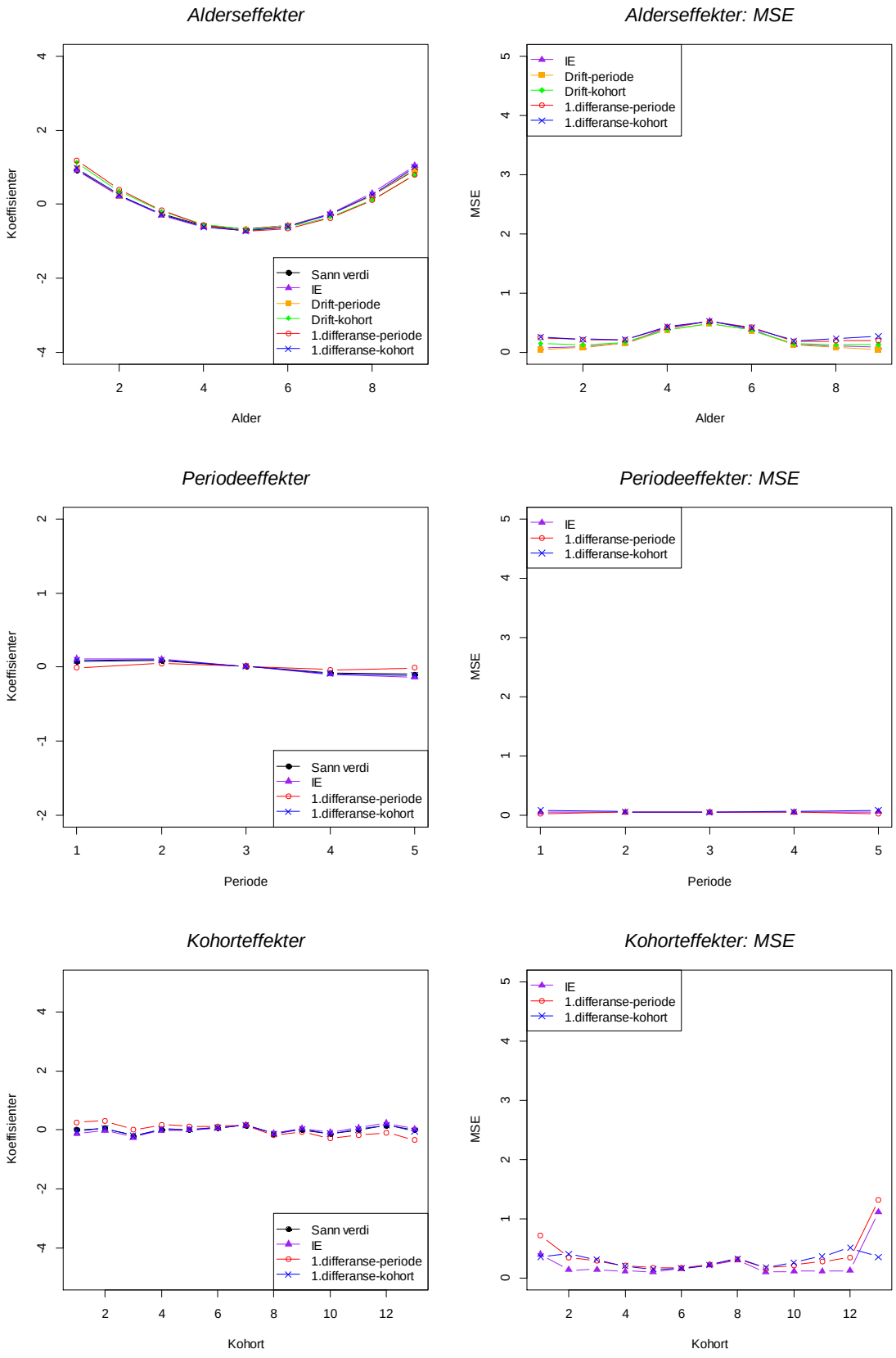
Tabell 8-10: Mål på de ulike modellene.

	Mål på de ulike APC-modellene		
	IE Modell 1	1.diff (P1=P5)	1.diff (C1=C13)
Samlet MSE	5,262	7,508	6,811
Frihetsgrader: 21	*Devians: 25,37*	*AIC: 189,9*	*Sum avvik: 92,2*

Tabell 8-11: Mål på de ulike reduserte modellene.

	Mål på de ulike reduserte modellene				
	Alder	Alder + periode	Alder + kohort	Alder + drift (periode)	Alder + drift (kohort)
Devians	42,35	37,83	28,54	40,87	40,87
Frihetsgrader	36	32	24	35	35
AIC	176,9	180,4	187,1	177,4	177,4
Sum avvik	34,6	50,6	82,1	37,5	37,5
Sig.	0,407	0,420	0,481	0,427	0,427
Andel *p*-verdier < 0,05	10,6 %	9,6 %	5,9 %	9,3 %	9,3 %
Drift				-0,041	-0,041

Resultatene er også illustrert i Figur 8-3. Den svarte kurven viser de sanne effektene.



Figur 8-3: Simuleringsresultater for de ulike modellene.

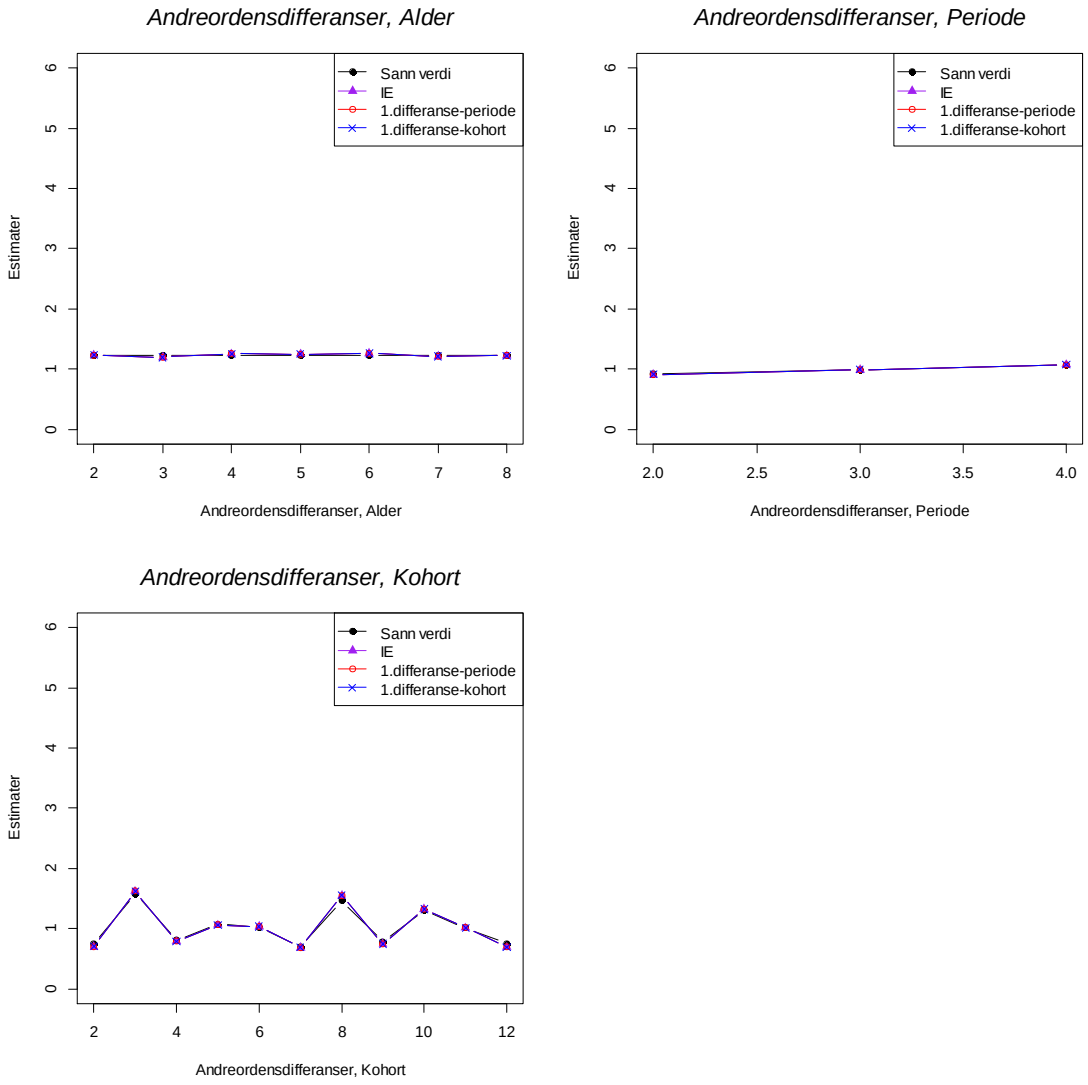
Kommentarer

Dette er den originale simuleringsmodellen som tilsvarer Modell 1 i Kapittel 7. I metoden med førsteordensdifferanser er første og siste periode satt lik hverandre (P1=P5), og tilsvarende er første og siste kohort satt lik hverandre (C1=C13). De sanne effektene i denne simuleringsmodellen er 0,081 for periode 1 og -0,099 for periode 5, samt 0,002 for kohort 1 og 0,000 for kohort 13. Differansene for de sanne effektene er 0,180 og 0,002 for henholdsvis (P1=P5) og (C1=C13). I Tabell 8-7, Tabell 8-8, Tabell 8-9 og Tabell 8-10 er mål for summert MSE for hver av alder-, periode- og kohortkategoriene oppgitt, sammen med det totale MSE-målet. Her ser vi at IE-metoden gir lavest MSE, men at MSE til modellene med førsteordensdifferanser nesten er like lav. IE-metoden og førsteordensdifferansen (C1=C13) gir estimater som er ganske nær sannheten, mens førsteordensdifferansen (P1=P5) avviker litt mer i estimatene sine. IE-metoden har stort sett litt mindre variasjon i sine estimater enn de andre metodene.

Modellene som inkluderer drift i tillegg til alderseffekter, har en driftparameter på $\delta = -0{,}041$. Estimatene for alderseffektene ligger nærmest opp til sannheten for drift-modellen som baserer seg på periode.

I Tabell 8-11 er ulike mål for de reduserte modellene tatt med. Deviansen som er oppgitt er mean for de 10 000 deviansene for den aktuelle modellen. For hver simulering er differansen i devians beregnet for full APC-modell i forhold til redusert modell. Jeg får dermed ut 10 000 *p*-verdier for om inkludering av alle faktorer i modellen (den fulle APC-modellen) vil gi signifikant forbedring i tilpasning av data i forhold til den reduserte modellen. Verdien som er oppgitt under sig. er mean av disse *p*-verdiene.

Når vi sammenligner deviansene for denne simuleringsmodellen, kan vi ikke påstå at de fulle APC-modellene gir signifikant forbedring i tilpasning til dataene på 95 % signifikansnivå, i forhold til de reduserte modellene. Av de 10 000 simulerte datasettene er den fulle APC-modellen signifikant bedre å benytte enn de reduserte modellene i 11 % eller færre av datasettene.



Figur 8-4: Andreordensdifferansene, estimater av identifiserbare ikke-drift effekter.

Andreordensdifferansene kan betraktes som et mål på akselerasjon i periodetrend rundt en gitt periode, eller tilsvarende akselerasjon i kohorttrend rundt en gitt kohort. Andreordensdifferansene viser uregelmessigheter i periode eller kohort som indikerer plutselige endringer i trendene. En viktig praktisk fordel med andreordensdifferansene er at de påvirkes kun av data fra nabogruppene og ikke påvirkes av trender som skjer tidligere eller etter de tilstøtende gruppene. Dersom ratioen mellom 2 tilstøtende relative risikoer er 1 indikerer dette en lokal rett linje, mens verdier større enn 1 indikerer en konveks sammenheng, og tilsvarende verdier mindre enn 1 indikerer en konkav sammenheng. Alderseffektene er i dette tilfellet modellert med et andregradspolynom, og andreordensdifferansene vil alle være like for alder.
Som beskrevet i Clayton [4] vil andreordensdifferansene til de ulike parametersettene være like. Forholdet mellom de tilstøtende koeffisientene vil være identisk selv om ikke koeffisientene er like.

Fra Figur 8-4 ser vi at kurvene for andreordensdifferansene er identisk for de 3 fulle APC-modellene som er med, og de ligger seg dermed oppå hverandre i figurene. Kurvene for andreordensdifferansene som gjelder periode viser ikke noen uregelmessigheter, mens kurvene for andreordensdifferansene som gjelder kohort viser mer uregelmessigheter, som tyder på mer forandringer i kohorteffektene.

8.3.2 Modell med endret periodeeffekt (Modell 4)

Datasettene er generert fra samme ligning som i Modell 4 fra Kapittel 7. Tabellene under viser ulike mål for både de fulle APC-modellene og de reduserte modellene.

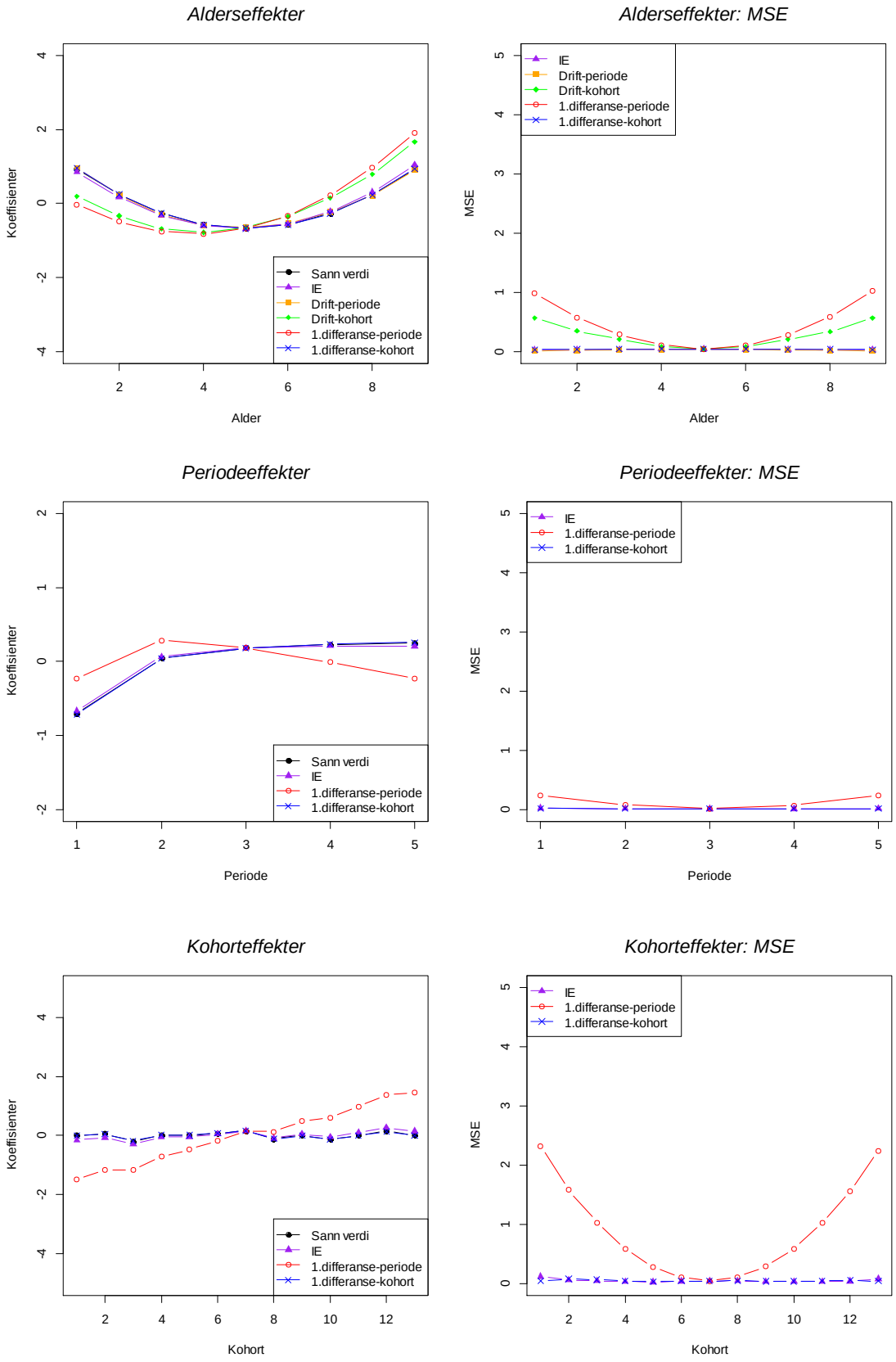
Tabell 8-12: Mål på de ulike modellene.

	Mål på de ulike APC-modellene		
	IE Modell 1	1.diff (P1=P5)	1.diff (C1=C13)
MSE alder	0,28	3,98	0,33
MSE periode	0,07	0,62	0,06
MSE kohort	0,58	11,71	0,57
Samlet MSE	0,93	16,32	0,96
Frihetsgrader: 21	*Devians: 22,36*	*AIC: 248,0*	*Sum avvik: 324,5*

Tabell 8-13: Mål på de ulike reduserte modellene.

	Mål på de ulike reduserte modellene				
	Alder	Alder + periode	Alder + kohort	Alder + drift (periode)	Alder + drift (kohort)
Devians	88,61	37,11	36,79	53,06	53,06
Frihetsgrader	36	32	24	35	35
AIC	284,2	240,7	256,4	250,7	250,7
Sum avvik	807,0	218,4	387,1	332,4	332,4
Sig.	< 0,001	0,312	0,039	0,054	0,054
Andel *p*-verdier < 0,05	100 %	19,6 %	82,4 %	75,1 %	75,1 %
Drift				0,189	0,189

I de følgende figurene vises koeffisientestimatene. Modellene som inkluderer drift har kun koeffisienter for alderseffekter.



Figur 8-5: Simuleringsresultater for de ulike modellene.

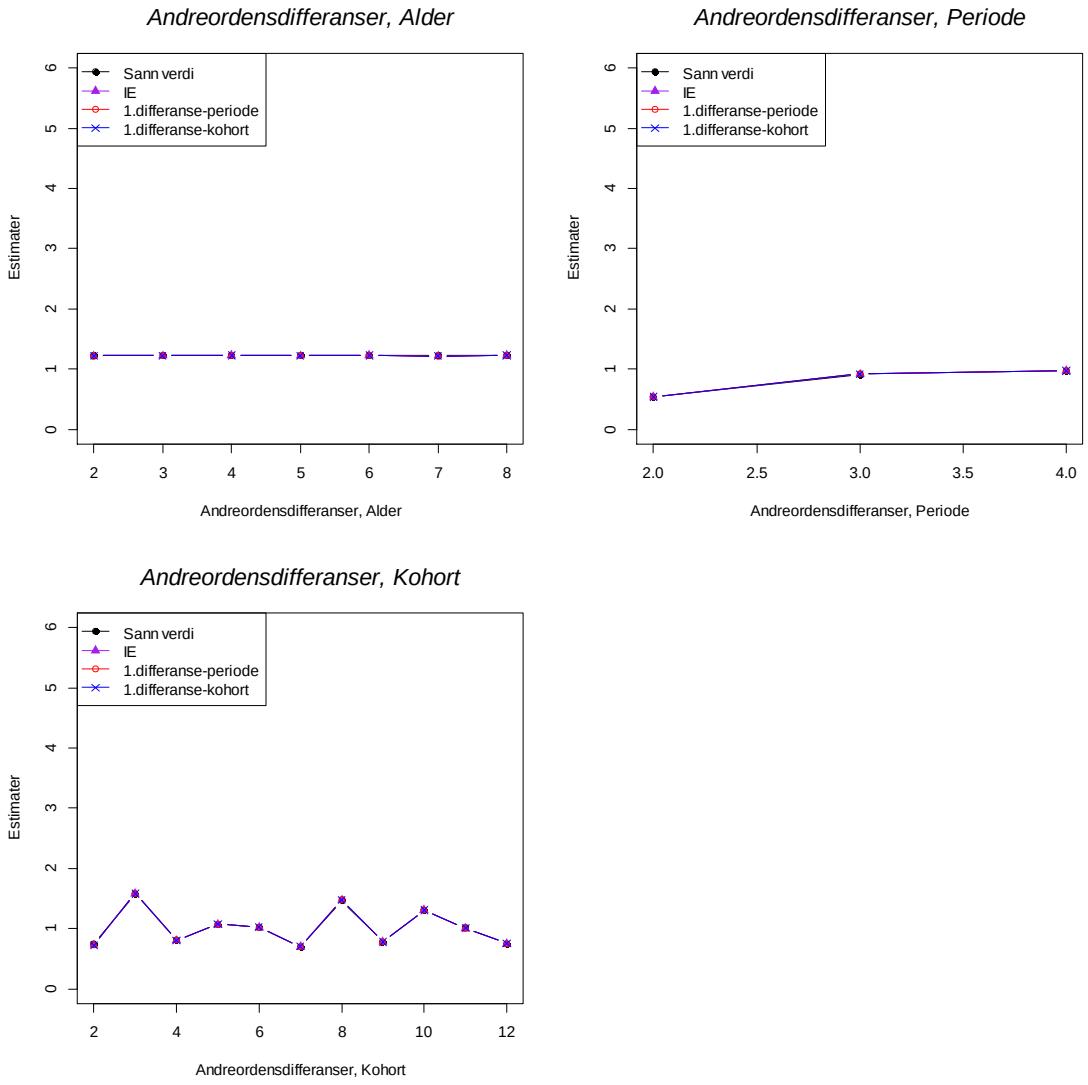
Kommentarer

Dette er simuleringsmodellen som tilsvarer Modell 4 i Kapittel 7. De sanne effektene i denne simuleringsmodellen er -0,707 for periode 1 og 0,253 for periode 5, samt 0,002 for kohort 1 og 0,000 for kohort 13. Differansene for de sanne effektene er 0,960 og 0,002 for henholdsvis (P1=P5) og (C1=C13). Fra Tabell 8-12 ser vi at IE-metoden gir lavest MSE, og at MSE til førsteordensdifferansen (C1=C13) er omtrent like lav. MSE til førsteordensdifferansen (P1=P5) er en god del større. Førsteordensdifferansen (P1=P5) avviker mer i estimatene sine, mens IE-metoden og førsteordensdifferansen (C1=C13) samsvarer godt med sannheten. IE-metoden gir ikke nødvendigvis estimater som er nærmest sannheten for alle koeffisientene, men den har mindre variasjon i sine estimater og har derfor likevel lavest MSE.

Modellene som inkluderer drift i tillegg til alderseffekter, har en driftparameter på $\delta = 0{,}189$. Som i forrige simuleringsmodell ligger estimatene for alderseffektene nærmest opp til sannheten for drift-modellen som baserer seg på periode.

I Tabell 8-13 er ulike mål for de reduserte modellene tatt med. Når vi sammenligner deviansene for denne simuleringsmodellen, kan det se ut som at de fulle APC-modellene gir signifikant forbedring i tilpasning til dataene på 95 % signifikansnivå i forhold til modellen med kun alderseffekter og modellen med alder- og kohorteffekter. Modellene med alder- og drifteffekter har en gjennomsnittlig *p*-verdi på 0,054, og forbedringen med å benytte de fulle APC-modellene er dermed så vidt ikke signifikant på 95 % nivå. De fulle APC-modellene gir heller ikke signifikant bedre tilpasning til dataene i forhold til modellen med alder- og periodeeffekter.

Av de 10 000 simulerte datasettene er den fulle APC-modellen signifikant bedre å benytte i 100 % av datasettene enn modellen med kun alderseffekter, mens den er signifikant bedre for 82 % av datasettene enn modellen med alder- og kohorteffekter. Den fulle modellen er også signifikant bedre å benytte for 75 % av datasettene enn modellen med alder- og drifteffekter. I 20 % av datasettene gir den fulle APC-modellen signifikant bedre tilpasning til dataene i forhold til modellen med alder- og periodeeffekter.



Figur 8-6: Andreordensdifferansene, estimater av identifiserbare ikke-drift effekter.

Fra Figur 8-6 ser vi at kurvene for andreordensdifferansene som gjelder periode viser en svak stigning, mens kurvene for andreordensdifferansene som gjelder kohort viser mer uregelmessigheter tilsvarende som i forrige simuleringsmodell, noe som tyder på mer forandringer i kohorttrenden.

8.3.3 Modell med endret kohorteffekt (Modell 6)

Datasettene er generert fra samme ligning som i Modell 6 i Kapittel 7. Tabellene under viser ulike mål for både de fulle APC-modellene og de reduserte modellene.

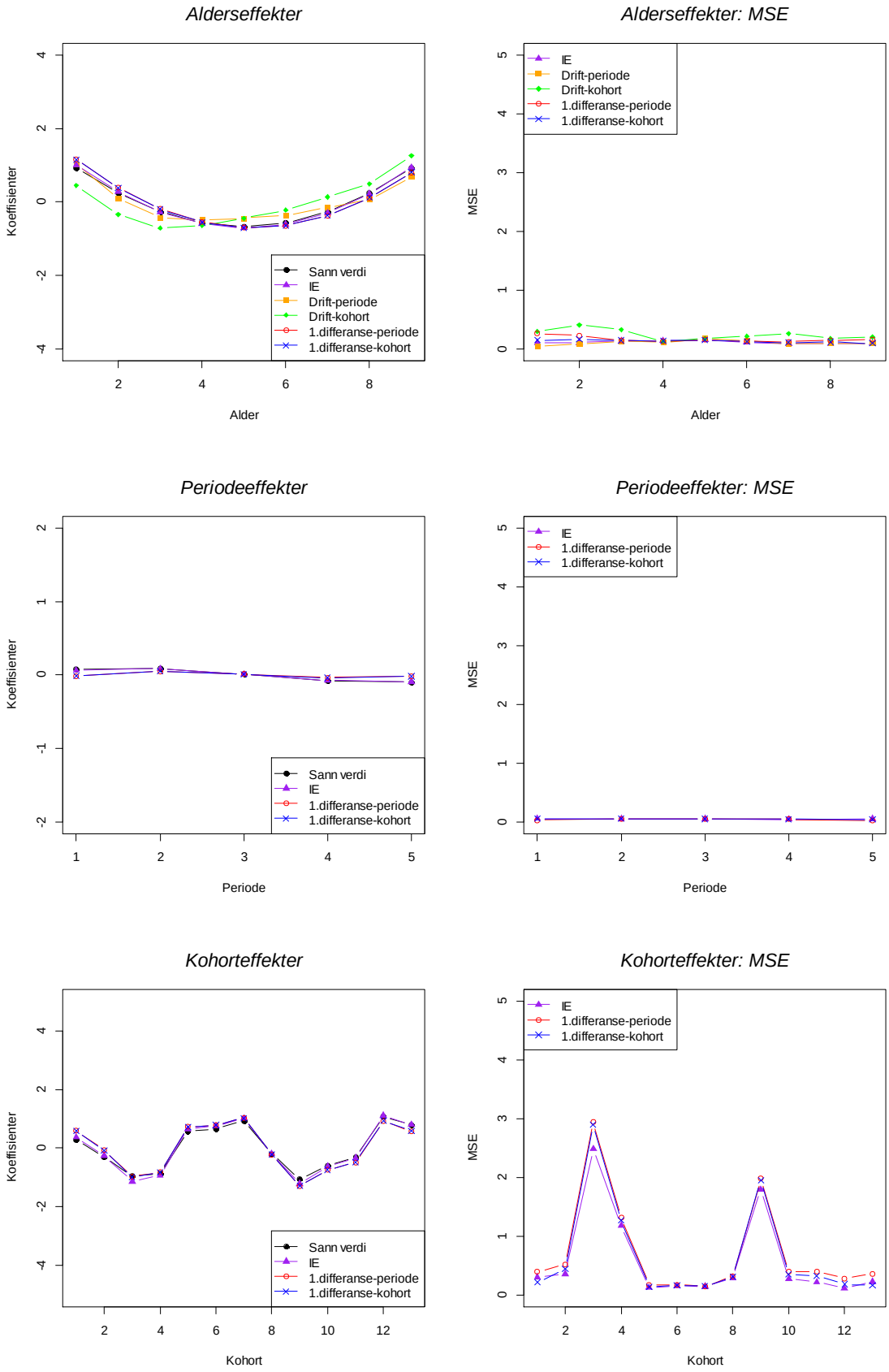
Tabell 8-14: Mål på de ulike modellene.

	Mål på de ulike APC-modellene		
	IE Modell 1	1.diff (P1=P5)	1.diff (C1=C13)
MSE alder	1,00	1,45	1,16
MSE periode	0,24	0,20	0,24
MSE kohort	7,66	9,39	8,53
Samlet MSE	8,90	11,04	9,92
Frihetsgrader: 21	*Devians: 24,40*	*AIC: 187,5*	*Sum avvik: 118,5*

Tabell 8-15: Mål på de ulike reduserte modellene.

	Mål på de ulike reduserte modellene				
	Alder	Alder + periode	Alder + kohort	Alder + drift (periode)	Alder + drift (kohort)
Devians	109,58	93,45	27,56	101,80	101,80
Frihetsgrader	36	32	24	35	35
AIC	242,7	234,6	184,7	236,9	236,9
Sum avvik	391,4	299,5	107,8	319,4	319,4
Sig.	< 0,001	< 0,001	0,484	< 0,001	< 0,001
Andel *p*-verdier < 0,05	100 %	100 %	6,1 %	100 %	100 %
Drift				0,143	0,143

I de følgende figurene vises koeffisientestimatene. Modellene som inkluderer drift har kun koeffisienter for alderseffekter.



Figur 8-7: Simuleringsresultater for de ulike modellene.

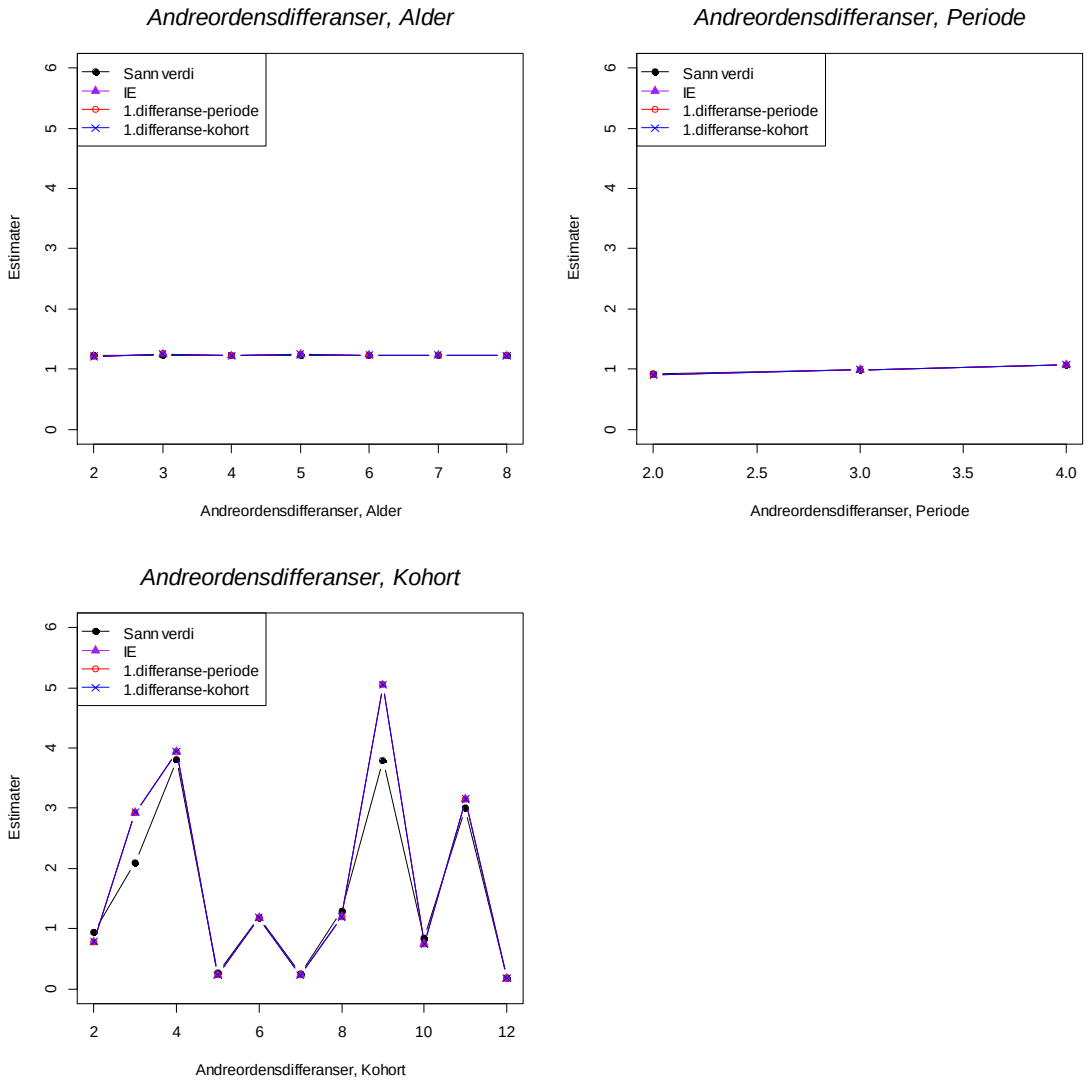
Kommentarer

Dette er simuleringsmodellen som tilsvarer Modell 6 i Kapittel 7. De sanne effektene i denne simuleringsmodellen er 0,081 for periode 1 og -0,099 for periode 5, samt 0,275 for kohort 1 og 0,784 for kohort 13. Differansene for de sanne effektene er 0,180 og 0,509 for henholdsvis (P1=P5) og (C1=C13). Fra Tabell 8-14 ser vi at IE-metoden gir lavest MSE, og at MSE for førsteordensdifferansen (C1=C13) er litt høyere. MSE for førsteordensdifferansen (P1=P5) er enda litt høyere. Modellene med førsteordensdifferanser gir omtrent identiske estimater, mens IE-metoden gir estimater som er litt nærmere sannheten.

Modellene som inkluderer drift i tillegg til alderseffekter, har en driftparameter på $\delta = 0{,}143$. Også her ligger estimatene for alderseffektene nærmest opp til sannheten for drift-modellen som baserer seg på periode.

I Tabell 8-15 er ulike mål for de reduserte modellene tatt med. Når vi sammenligner deviansene for denne simuleringsmodellen, kan det se ut som at de fulle APC-modellene gir signifikant forbedring i tilpasning til dataene på 95 % signifikansnivå i forhold til alle de reduserte modellene med unntak av modellen med alder- og kohorteffekter.

Av de 10 000 simulerte datasettene er den fulle APC-modellen signifikant bedre å benytte i 100 % av datasettene enn alle de reduserte modellene med unntak av modellen med alder- og kohorteffekter. I 6 % av datasettene gir den fulle APC-modellen signifikant bedre tilpasning til dataene i forhold til modellen med alder- og kohorteffekter.



Figur 8-8: Andreordensdifferansene, estimater av identifiserbare ikke-drift effekter.

Fra Figur 8-8 ser vi at kurvene for andreordensdifferansene som gjelder periode ikke viser noen uregelmessigheter, mens kurvene for andreordensdifferansene som gjelder kohort viser store uregelmessigheter, som tyder på at kohorteffektene varierer mye.

8.3.4 Modell med endret periode- og kohorteffekt (Modell 7)

Datasettene er generert fra samme ligning som i Modell 7 fra Kapittel 7. Tabellene under viser ulike mål for både de fulle APC-modellene og de reduserte modellene.

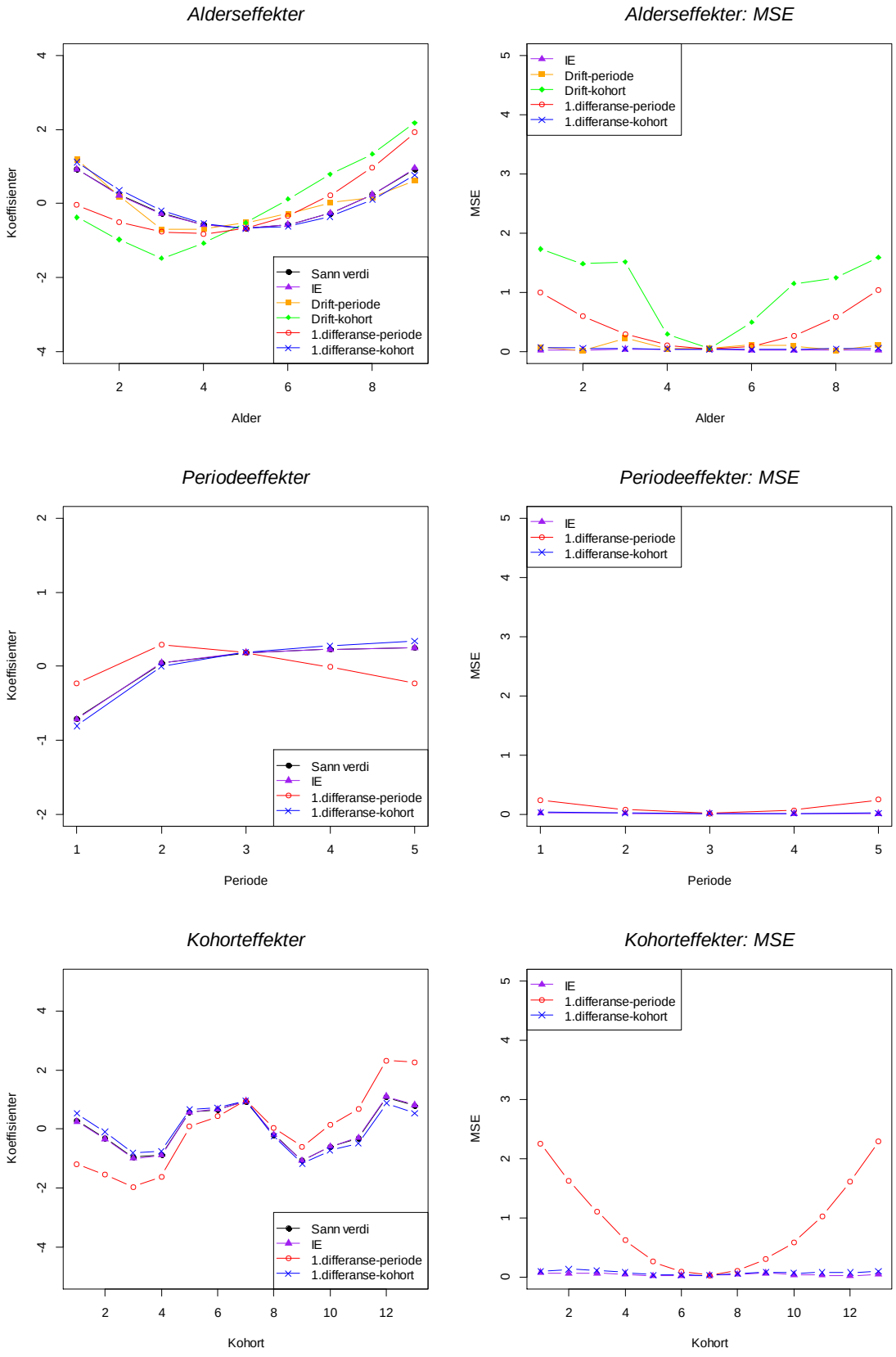
Tabell 8-16: Mål på de ulike modellene.

	Mål på de ulike APC-modellene		
	IE Modell 1	1.diff (P1=P5)	1.diff (C1=C13)
MSE alder	0,23	3,98	0,39
MSE periode	0,07	0,63	0,09
MSE kohort	0,53	11,90	0,93
Samlet MSE	0,82	16,51	1,42
Frihetsgrader: 21	*Devians: 22,67*	*AIC: 245,6*	*Sum avvik: 450,1*

Tabell 8-17: Mål på de ulike reduserte modellene.

	Mål på de ulike reduserte modellene				
	Alder	Alder + periode	Alder + kohort	Alder + drift (periode)	Alder + drift (kohort)
Devians	404,13	224,18	35,56	232,52	232,52
Frihetsgrader	36	32	24	35	35
AIC	597,1	425,1	252,5	427,5	427,5
Sum avvik	8394,8	3279,9	486,5	3548,7	3548,7
Sig.	< 0,001	< 0,001	0,055	< 0,001	< 0,001
Andel *p*-verdier < 0,05	100 %	100 %	75,9 %	100 %	100 %
Drift				0,390	0,390

I de følgende figurene vises koeffisientestimatene. Modellene som inkluderer drift har kun koeffisienter for alderseffekter.



Figur 8-9: Simuleringsresultater for de ulike modellene.

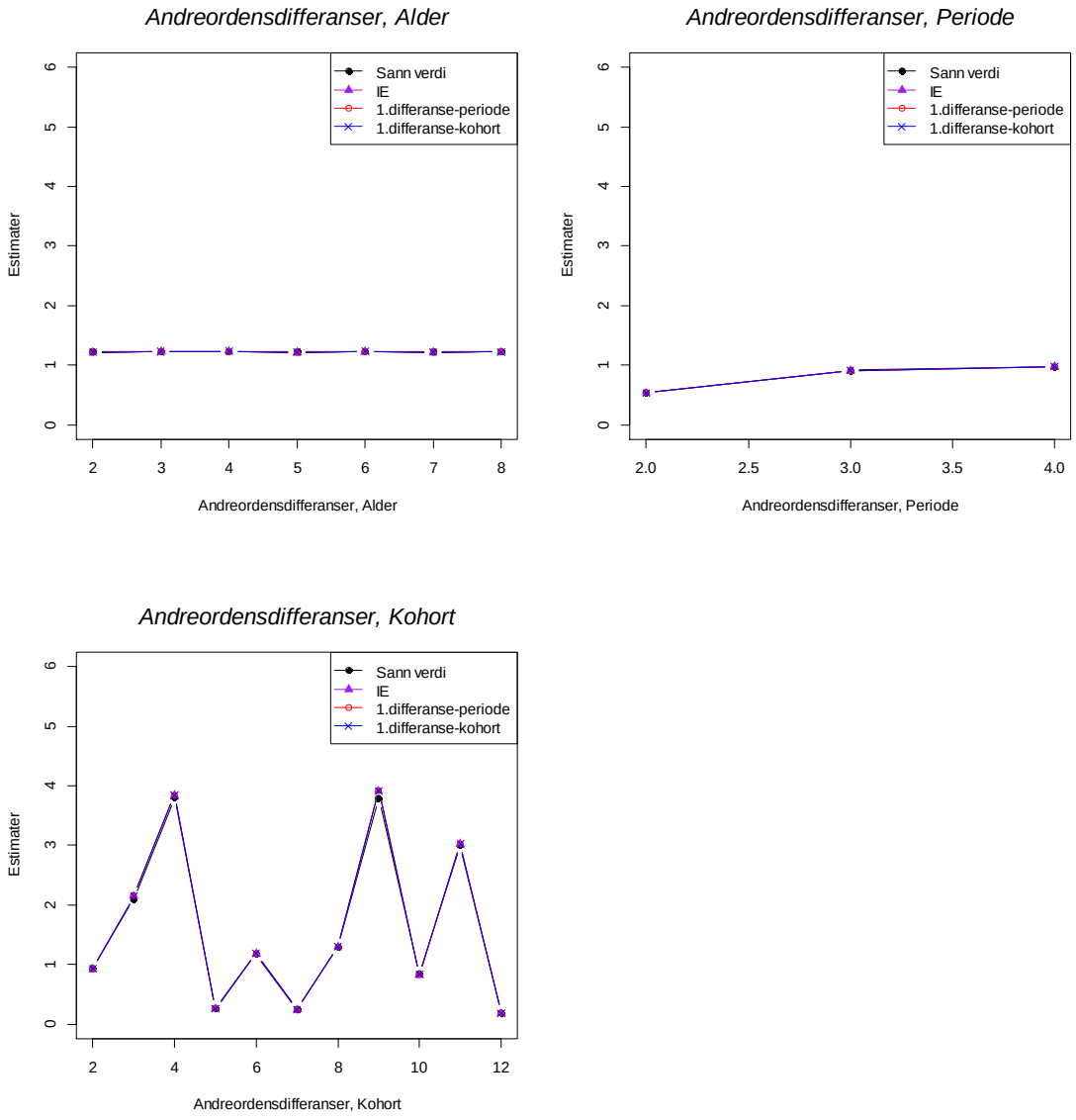
Kommentarer

Dette er simuleringsmodellen som tilsvarer Modell 7 i Kapittel 7. De sanne effektene i denne simuleringsmodellen er -0,707 for periode 1 og 0,253 for periode 5, samt 0,275 for kohort 1 og 0,784 for kohort 13. Differansene for de sanne effektene er 0,960 og 0,509 for henholdsvis (P1=P5) og (C1=C13). Fra Tabell 8-16 ser vi at IE-metoden gir lavest MSE, og at MSE for førsteordensdifferansen (C1=C13) er litt høyere, mens MSE for førsteordensdifferansen (P1=P5) er en god del høyere. Førsteordensdifferansen (P1=P5) avviker mer i estimatene sine. IE-metoden gir estimater som er nærmere sannheten enn førsteordensdifferansen (C1=C13).

Modellene som inkluderer drift i tillegg til alderseffekter, har en driftparameter på $\delta = 0{,}390$. Også her ligger estimatene for alderseffektene nærmest opp til sannheten for drift-modellen som baserer seg på periode, selv om også den avviker en del i sine estimater.

I Tabell 8-17 er ulike mål for de reduserte modellene tatt med. Når vi sammenligner deviansene for denne simuleringsmodellen, kan det se ut som at de fulle APC-modellene gir signifikant forbedring i tilpasning til dataene på 95 % signifikansnivå i forhold til alle de reduserte modellene med unntak av modellen med alder- og kohorteffekter. Modellen med alder- og kohorteffekter har en gjennomsnittlig *p*-verdi på 0,055, og forbedringen med å benytte de fulle APC-modellene er dermed så vidt ikke signifikant på 95 % nivå.

Av de 10 000 simulerte datasettene er den fulle modellen signifikant bedre å benytte i 100 % av datasettene enn alle de reduserte modellene med unntak av modellen med alder- og kohorteffekter. I 76 % av datasettene gir den fulle APC-modellen signifikant bedre tilpasning til dataene i forhold til modellen med alder- og kohorteffekter.



Figur 8-10: Andreordensdifferansene, estimater av identifiserbare ikke-drift effekter.

Fra Figur 8-10 ser vi at kurvene for andreordensdifferansene som gjelder periode viser en svak stigning, mens kurvene for andreordensdifferansene som gjelder kohort viser store uregelmessigheter, som tyder på at kohorteffektene varierer mye.

8.3.5 Modell med endret periode- og kohorteffekt versjon 3

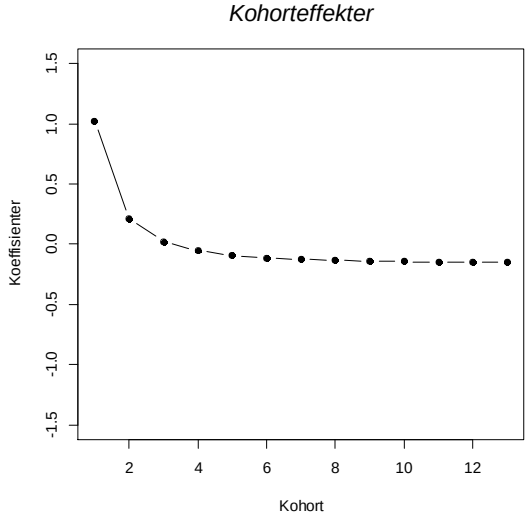
Datasettene er generert fra ligningen:

$$y_{ij} \sim \text{Poisson}\{\exp\left[0.3 + 0.1\left(\text{alder}_{ij} - 5\right)^2 + (1.5 - (1/\text{periode}_{ij}{}^2)) + 2 \cdot \frac{1}{(0.3 + \text{kohort}_{ij})^2}\right]\}$$

Tabell 8-18: Sanne alder-, periode- og kohorteffekter i simuleringsmodellen.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	***periode***	***kohort***
$a1 = 0.933$	$p1 = -0.707$	$c1 = 1.020$
$a2 = 0.233$	$p2 = 0.043$	$c2 = 0.215$
$a3 = -0.267$	$p3 = 0.182$	$c3 = 0.020$
$a4 = -0.567$	$p4 = 0.230$	$c4 = -0.055$
$a5 = -0.667$	$p5 = 0.253$	$c5 = -0.092$
$a6 = -0.567$		$c6 = -0.113$
$a7 = -0.267$		$c7 = -0.126$
$a8 = 0.233$		$c8 = -0.134$
$a9 = 0.933$		$c9 = -0.140$
		$c10 = -0.145$
		$c11 = -0.148$
		$c12 = -0.150$
		$c13 = -0.152$

Alderseffektene er generert tilsvarende som i den originale simuleringsmodellen (Modell 1), periodeeffektene er generert som i simuleringsmodell 4, mens en ny ligning benyttes for å generere kohorteffektene.



Figur 8-11: Sanne effekter for kohort i simuleringsmodellen.

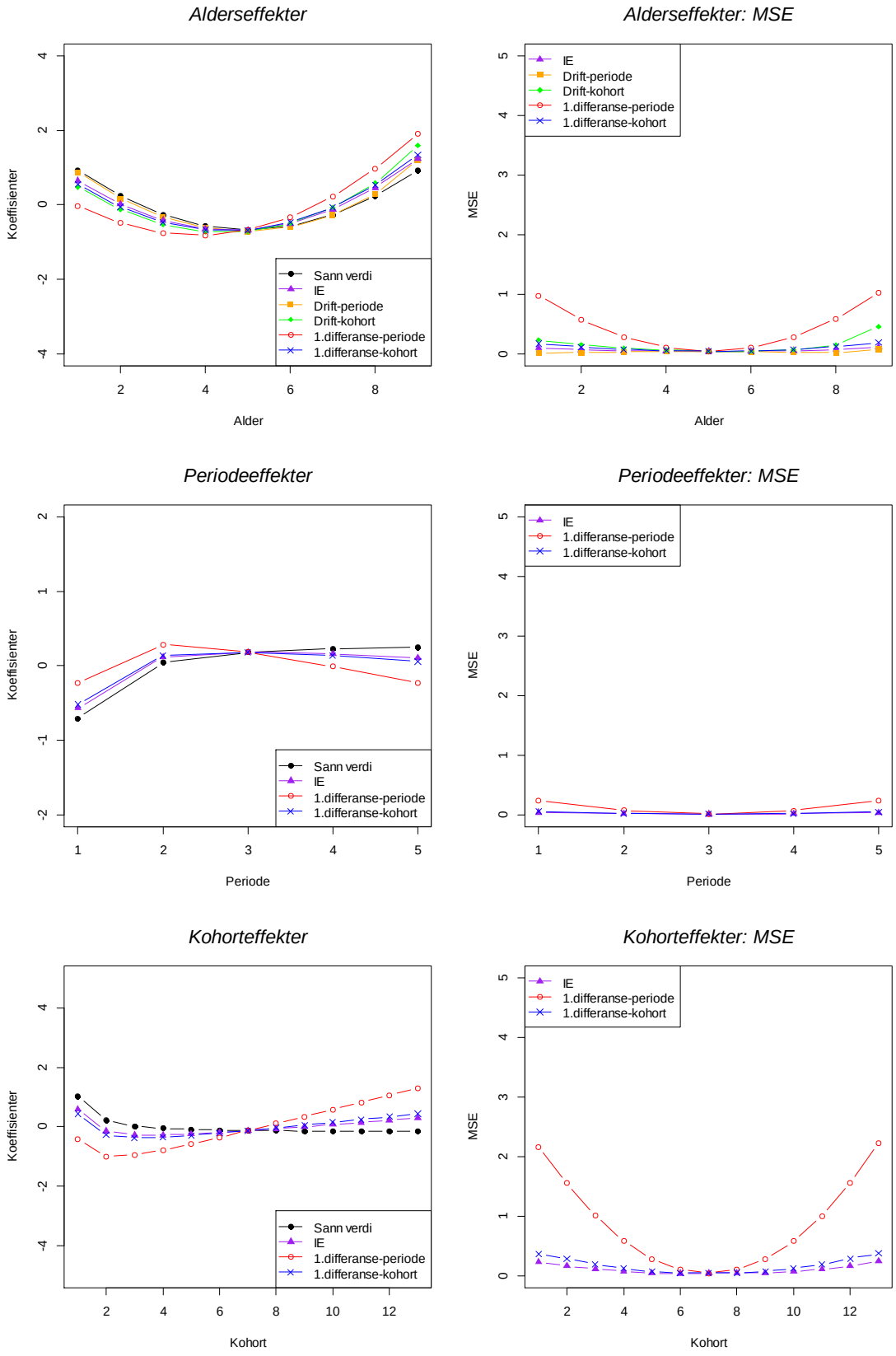
Tabellene under viser ulike mål for både de fulle APC-modellene og de reduserte modellene.

Tabell 8-19: Mål på de ulike modellene.

	Mål på de ulike APC-modellene		
	IE Modell 1	1.diff (P1=P5)	1.diff (C1=C13)
MSE alder	0,56	3,95	0,86
MSE periode	0,11	0,62	0,15
MSE kohort	1,36	11,45	2,19
Samlet MSE	2,03	16,02	3,20
Frihetsgrader: 21	*Devians: 22,25*	*AIC: 252,4*	*Sum avvik: 377,3*

Tabell 8-20: Mål på de ulike reduserte modellene.

	Mål på de ulike reduserte modellene				
	Alder	Alder + periode	Alder + kohort	Alder + drift (periode)	Alder + drift (kohort)
Devians	69,35	47,94	38,14	57,31	57,31
Frihetsgrader	36	32	24	35	35
AIC	269,5	256,1	262,3	259,5	259,5
Sum avvik	508,1	510,7	465,7	494,9	494,9
Sig.	0,004	0,064	0,027	0,026	0,026
Andel *p*-verdier < 0,05	98,4 %	72,5 %	87,2 %	87,5 %	87,5 %
Drift				0,100	0,100



Figur 8-12: Simuleringsresultater for de ulike modellene.

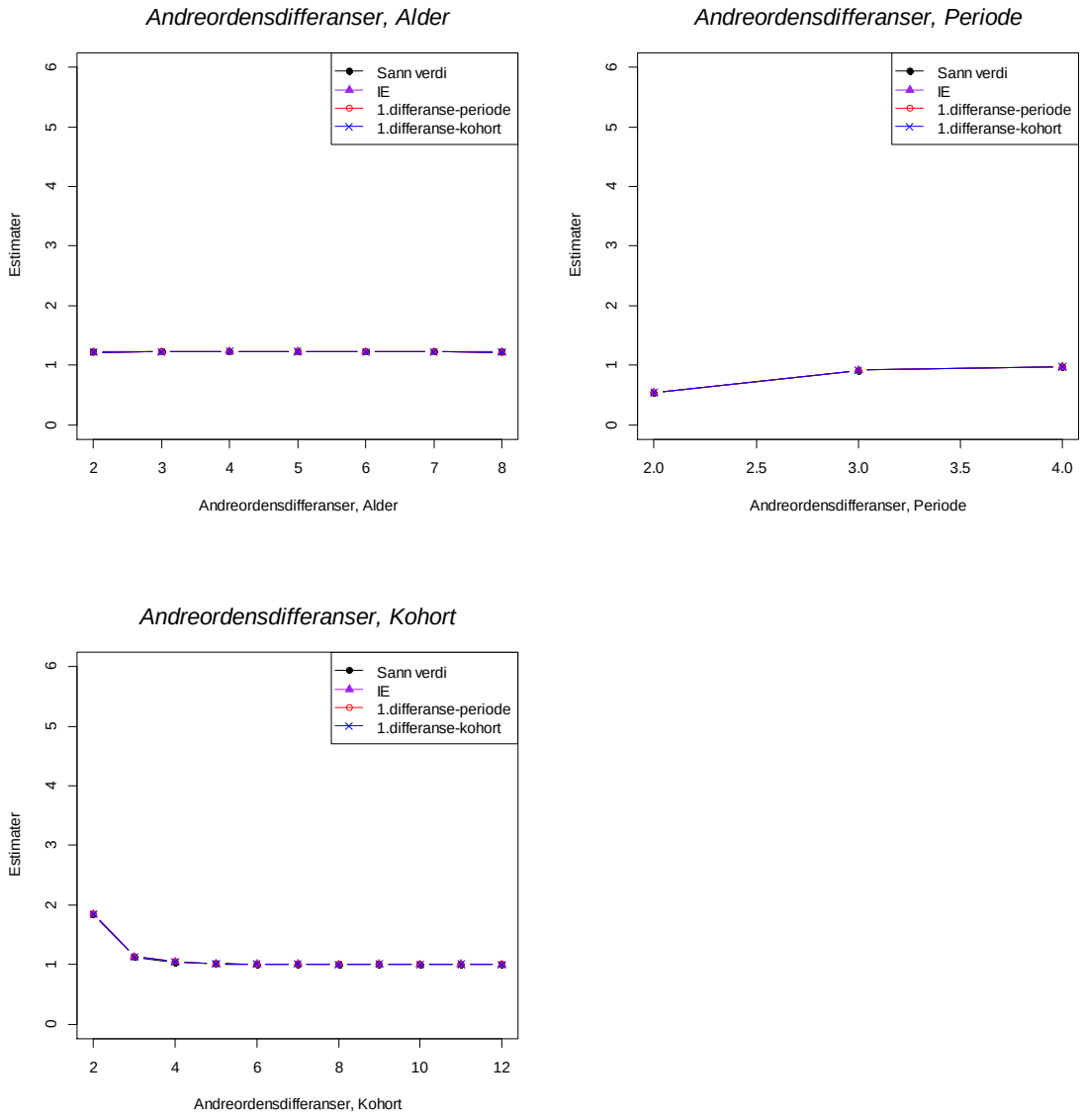
Kommentarer

De sanne effektene i denne simuleringsmodellen er -0,707 for periode 1 og 0,253 for periode 5, samt 1,020 for kohort 1 og -0,152 for kohort 13. Differansene for de sanne effektene er 0,960 og 1,172 for henholdsvis (P1=P5) og (C1=C13). Fra Tabell 8-19 ser vi at IE-metoden gir lavest MSE, og at MSE for førsteordensdifferansen (C1=C13) er litt høyere, mens MSE for førsteordensdifferansen (P1=P5) er en god del høyere. Førsteordensdifferansen (P1=P5) avviker fra sannheten med estimatene sine i større grad enn førsteordensdifferansen (C1=C13). IE-metoden gir estimater som er litt nærmere sannheten.

Modellene som inkluderer drift i tillegg til alderseffekter, har en driftparameter på $\delta = 0{,}100$. Også her ligger estimatene for alderseffektene nærmest opp til sannheten for drift-modellen som baserer seg på periode.

I Tabell 8-20 er ulike mål for de reduserte modellene tatt med. Når vi sammenligner deviansene for denne simuleringsmodellen, kan det se ut som at de fulle APC-modellene gir signifikant forbedring i tilpasning til dataene på 95 % signifikansnivå i forhold til alle de reduserte modellene med unntak av modellen med alder- og periodeeffekter. Modellen med alder- og periodeeffekter har en gjennomsnittlig *p*-verdi på 0,064, og forbedringen med å benytte de fulle APC-modellene er dermed så vidt ikke signifikant på 95 % nivå.

Av de 10 000 simulerte datasettene er den fulle APC-modellen signifikant bedre å benytte i 98 % av datasettene enn modellen med kun alderseffekter, mens den er signifikant bedre for 87 % av datasettene enn modellen med alder- og kohorteffekter, og tilsvarende bedre for 88 % av datasettene enn modellen med alder- og drifteffekter. I 73 % av datasettene gir den fulle APC-modellen signifikant bedre tilpasning til dataene i forhold til modellen med alder- og periodeeffekter.



Figur 8-13: Andreordensdifferansene, estimater av identifiserbare ikke-drift effekter.

Fra Figur 8-13 ser vi at kurvene for andreordensdifferansene som gjelder periode viser en svak stigning, mens kurvene for andreordensdifferansene som gjelder kohort viser en plutselig endring rundt kohort 2 og deretter viser kurven ingen uregelmessigheter. Etter den plutselige endringen rundt kohort 2 varierer ikke kohorteffektene så mye.

8.3.6 Modell med endret alder-, periode- og kohorteffekt

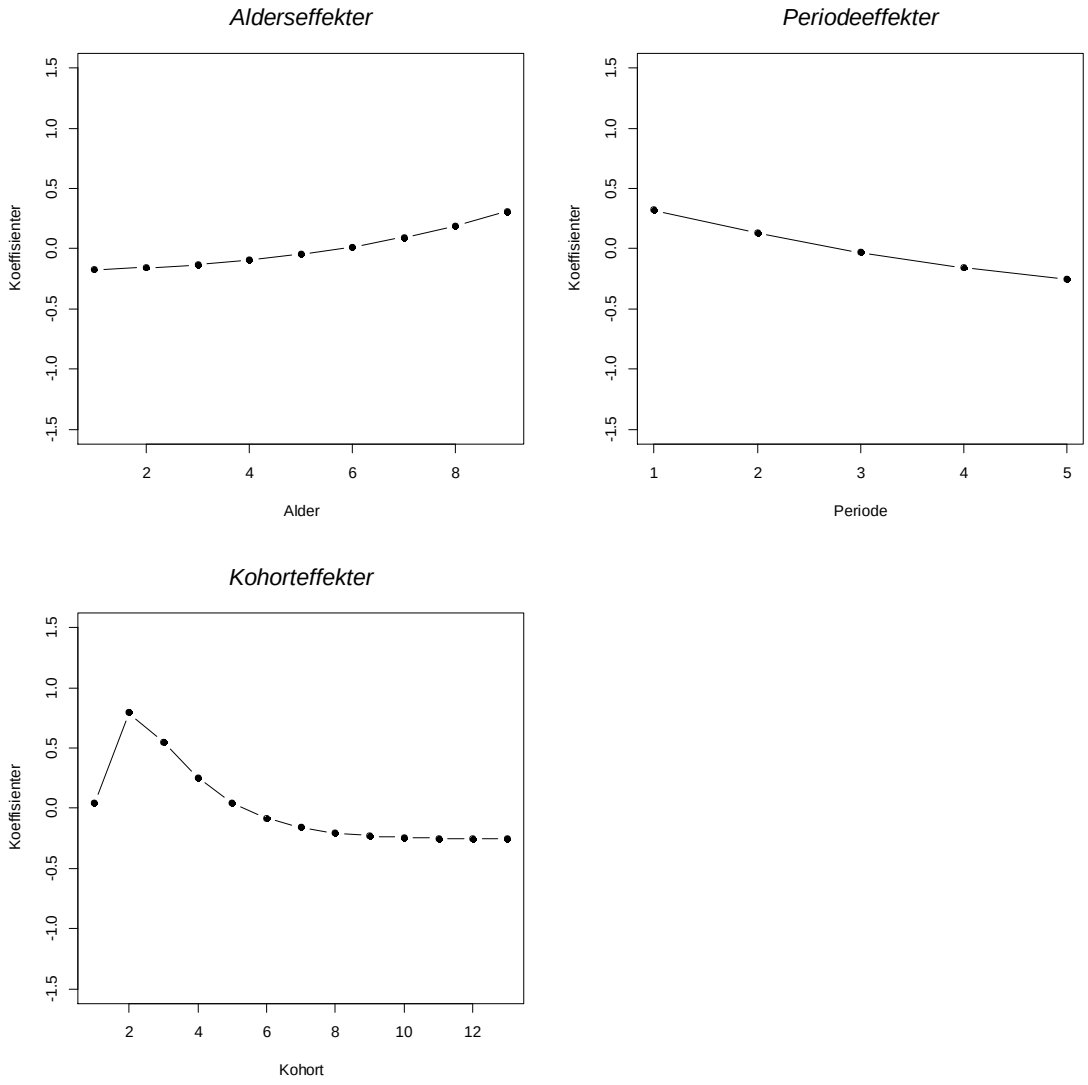
Datasettene er generert fra ligningen:

$$y_{ij} \sim \text{Poisson}\{\exp\left[1.5 + \exp(0.07 \cdot \text{alder}_{ij})^2 + 0.1 \cdot (0.4 \cdot \text{periode}_{ij} - 3)^2 + 2 \cdot (\exp(0.5^{\text{kohort}_{ij}} \cdot \text{kohort}_{ij}) - 1) - \left(\frac{1}{\text{kohort}_{ij}{}^2}\right)\right]\}$$

Denne simuleringsmodellen er ikke benyttet tidligere i oppgaven, men den blir også benyttet i Kapittel 9.8, der IE-metoden sammenlignes med metoden som baserer seg på Partial Least Squares.

Tabell 8-21: Sanne alder-, periode- og kohorteffekter i simuleringsmodellen.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	**periode**	**kohort**
$a1 = -0.173$	$p1 = 0.320$	$c1 = 0.045$
$a2 = -0.158$	$p2 = 0.128$	$c2 = 0.795$
$a3 = -0.133$	$p3 = -0.032$	$c3 = 0.546$
$a4 = -0.096$	$p4 = -0.160$	$c4 = 0.253$
$a5 = -0.048$	$p5 = -0.256$	$c5 = 0.045$
$a6 = 0.015$		$c6 = -0.084$
$a7 = 0.093$		$c7 = -0.161$
$a8 = 0.190$		$c8 = -0.205$
$a9 = 0.309$		$c9 = -0.230$
		$c10 = -0.243$
		$c11 = -0.250$
		$c12 = -0.254$
		$c13 = -0.256$



Figur 8-14: Sanne effekter for henholdsvis alder, periode og kohort i simuleringsmodellen.

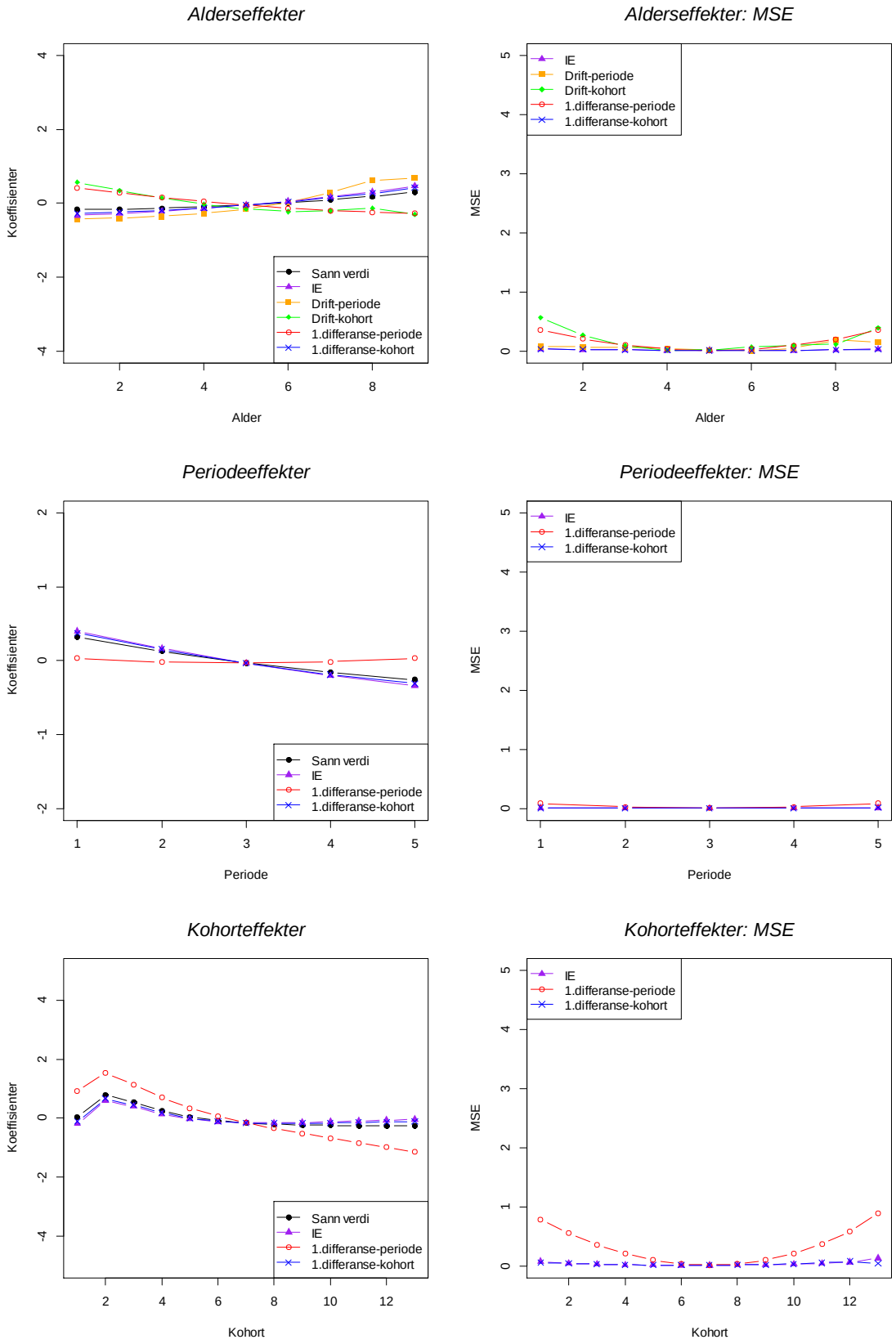
Tabellene under viser ulike mål for både de fulle APC-modellene og de reduserte modellene.

Tabell 8-22: Mål på de ulike modellene.

	Mål på de ulike APC-modellene		
	IE Modell 1	1.diff (P1=P5)	1.diff (C1=C13)
MSE alder	0,17	1,37	0,16
MSE periode	0,04	0,22	0,03
MSE kohort	0,50	4,23	0,40
Samlet MSE	0,70	5,83	0,60
Frihetsgrader: 21	*Devians: 21,24*	*AIC: 297,8*	*Sum avvik: 795,9*

Tabell 8-23: Mål på de ulike reduserte modellene.

	Mål på de ulike reduserte modellene				
	Alder	Alder + periode	Alder + kohort	Alder + drift (periode)	Alder + dri rt)
Devians	243,40	75,93	25,05	80,41	80,41
Frihetsgrader	36	32	24	35	35
AIC	490,0	330,5	295,6	329,0	329,0
Sum avvik	9383,7	3009,2	717,0	3035,3	3035,3
Sig.	< 0,001	< 0,001	0,421	< 0,001	< 0,001
Andel *p*-verdier < 0,05	100 %	99,8 %	10,3%	99,8 %	99,8 %
Drift				-0,249	-0,249



Figur 8-15: Simuleringsresultater for de ulike modellene.

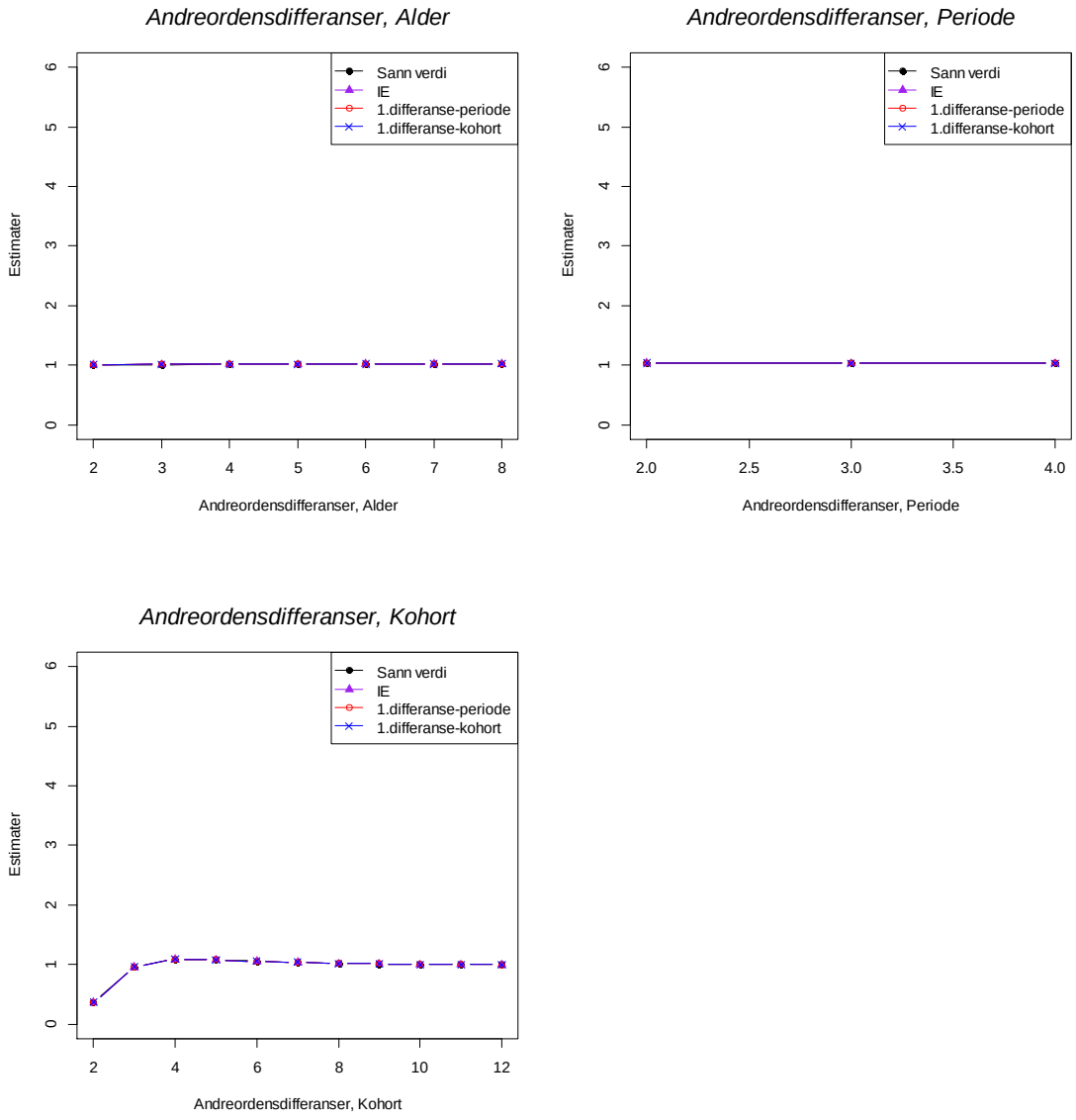
Kommentarer

Dette er den samme simuleringsmodellen som i Kapittel 9.8. De sanne effektene i denne simuleringsmodellen er 0,320 for periode 1 og -0,256 for periode 5, samt 0,045 for kohort 1 og -0,256 for kohort 13. Differansene for de sanne effektene er 0,576 og 0,301 for henholdsvis (P1=P5) og (C1=C13). Fra Tabell 8-22 ser vi at nå er det førsteordensdifferansen (C1=C13) som har lavest MSE, og at IE-metoden har en MSE som er litt høyere, mens MSE for førsteordensdifferansen (P1=P5) er en god del høyere. Førsteordensdifferansen (P1=P5) avviker mest i estimatene sine. Førsteordensdifferansen (C1=C13) gir estimater som er litt nærmere sannheten enn estimatene fra IE-metoden.

Modellene som inkluderer drift i tillegg til alderseffekter, har en driftparameter på $\delta = -0{,}249$. Også her ligger estimater for alderseffektene nærmest opp til sannheten for drift-modellen som baserer seg på periode, selv om også den avviker en del i sine estimater.

I Tabell 8-23 er ulike mål for de reduserte modellene tatt med. Når vi sammenligner deviansene for denne simuleringsmodellen, kan det se ut som at de fulle APC-modellene gir signifikant forbedring i tilpasning til dataene på 95 % signifikansnivå i forhold til alle de reduserte modellene med unntak av modellen med alder- og kohorteffekter.

Av de 10 000 simulerte datasettene er den fulle APC-modellen signifikant bedre å benytte i 100 % av datasettene enn alle de reduserte modellene med unntak av modellen med alder- og kohorteffekter. I 10 % av datasettene gir den fulle APC-modellen signifikant bedre tilpasning til dataene i forhold til modellen med alder- og kohorteffekter.



Figur 8-16: Andreordensdifferansene, estimater av identifiserbare ikke-drift effekter.

Fra Figur 8-16 ser vi at kurvene for andreordensdifferansene som gjelder periode ikke viser noen uregelmessigheter, mens kurvene for andreordensdifferansene som gjelder kohort viser en plutselig endring rundt kohort 2, og deretter viser kurven lite uregelmessigheter. Etter den plutselige endringen rundt kohort 2 varierer ikke kohorteffektene så mye.

8.3.7 Resultater

Tilsvarende som i Kapittel 7 ser vi også her at IE-metoden ser ut til å være robust i alle de ulike simuleringsmodellene vi har sett på i dette kapittelet. Den påvirkes i mye mindre grad av hvilken simuleringsmodell som velges for å generere datasettene enn de andre metodene. IE-metoden har minst varians og klarer i stor grad å produsere estimater som er tilnærmet de sanne effektene. I nesten alle de ulike simuleringseksemplene som er vist her, er det IE-metoden som gir lavest MSE. Men IE-metoden gir ikke nødvendigvis alltid de mest korrekte estimatene for alle parametrene. Når de sanne effektene til første og siste periode og tilsvarende første og siste kohort er ganske lik hverandre, kan metoden med førsteordensdifferanser også gi estimater som er like gode eller bedre enn IE-metoden. Vi ser at dess større forskjell det er på de effektene som betinges å være lik hverandre i metoden med førsteordensdifferanser, dess større avvik fra de sanne verdiene ser vi at estimatene får. Det ser også ut til å være en tendens at metoden med førsteordensdifferanser basert på periodene påvirkes i større grad av dette, og at avvikene må være større før metoden med førsteordensdifferanser basert på kohort påvirkes tilsvarende. I de første simuleringsmodellene der det ikke var så store periode- og kohorteffekter, eller bare den ene effekten var til stede, kan de reduserte modellene også være et godt alternativ til de fulle APC-modellene. Her kan en gjøre seg nytte av å se på mål som devians og AIC for å avgjøre om modellen er god nok for det aktuelle datasettet. Modellene som inkluderer drift gir oss estimatene til den lineære trendkoeffisienten, δ, og denne er identisk for begge parametriseringene av modellen. Driftparameteren angir om det er en tidseffekt til stede, men kan ikke relatere denne effekten til periode eller kohort. Den angir størrelsen på en slags felles trend, og om denne er stigende eller synkende. Men for å kunne tilskrive denne effekten til enten periode, kohort eller både periode og kohort, må vi benytte modeller som inkluderer flere faktorer.
For å oppsummere resultatene i dette kapittelet, ser vi samme tendens som i Kapittel 7 om simulering, at når en ikke vet noe om de dataene en analyserer, vil IE-metoden i mange tilfeller kunne gi sikrere estimater enn ved valg av en av de andre metodene.

9. Partial Least Squares

I de tradisjonelle regresjonsanalysene (generalisert lineær modellering) kreves det at matrisen med de uavhengige variablene er av full rang. Dette er ikke et krav for statistiske metoder som forsøker å redusere antall dimensjoner av data uten å miste viktig informasjon. Eksempler på slike statistiske metoder er prinsipal komponent regresjon (PCR) og Partial Least Squares (PLS) regresjon. Underveis i arbeidet med masteroppgaven har Tu et al. [5, 26] kommet med publikasjoner der de foreslår PLS-regresjon for å separere de simultane effektene til alder, periode og kohort og dermed forsøke å løse problemet med modellidentifikasjon. De viser også relasjonen mellom PLS-regresjon og IE-metoden i et spesifikt eksempel. I dette kapittelet gis det en kort innføring i aktuell teori, og videre benyttes PLS-metoden for analyse av noen datasett. De samme datasettene er også analysert med IE-metoden og i noen tilfeller også CGLIM-metoden. I siste delen av dette kapittelet er det utført simuleringsanalyser for å sammenligne PLS-metoden med IE-metoden.

9.1 OLS

OLS står for *Ordinary Least Squares*, minste kvadraters metode, og er den mest kjente blant de lineære regresjonsmetodene. Metoden tar utgangspunkt i å estimere de ukjente parametrene, $\boldsymbol{\beta}$, i den lineære regresjonsmodellen slik at summen av kvadratavviket mellom responsen og den predikerte responsen minimeres [27]. Minste kvadraters metode går ut på å minimere følgende uttrykk med hensyn på $\boldsymbol{\beta}$:

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (r_i)^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\,\beta_j)^2 \qquad (9\text{–}1)$$

Med RSS menes summen av de kvadrerte residualene, og er en forkortelse for *Residual Sums of Squares*. Metoden minimerer residualene, hvilket er ekvivalent med en maksimering av tilpasningen til responsvektoren $\boldsymbol{y}$. For å kunne estimere $\boldsymbol{\beta}$ krever OLS at $\boldsymbol{X}$-verdiene er lineært uavhengige, dvs. at $(\boldsymbol{X^T X})$ har full rang og at antall variable er mindre eller lik antall observasjoner.
Under forutsetning av at $(\boldsymbol{X^T X})$ er invertibel har regresjonskoeffisientene som minimerer uttrykket en unik løsning som kan beregnes ved:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} \qquad (9\text{–}2)$$

Og minste kvadraters tilpasning til responsvektoren $\boldsymbol{y}$ blir da:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \qquad (9\text{–}3)$$

Dette er ortogonalprojeksjonen av responsen, $\boldsymbol{y}$, ned på søylene i $\boldsymbol{X}$-matrisen.

Når $\boldsymbol{X^T X}$ er invertibel har OLS en unik løsning for estimatene av regresjonskoeffisientene $\boldsymbol{\beta}$. I de tilfellene $\boldsymbol{X^T X}$ ikke er invertibel, vil OLS kunne gi uendelig mange løsninger, og i praksis betyr dette at det finnes uendelig mange kombinasjoner av regresjonskoeffisientene som gir OLS-løsningen. At det finnes uendelig mange løsninger betyr fra et tolkningsperspektiv at det finnes uendelig mange kombinasjoner av måter å forklare relasjonen mellom responsen og forklaringsvariablene. Det er derfor hensiktsmessig å finne alternative metoder som kan produsere modeller som kan gi en sikrere forståelse av relasjonen mellom responsen og forklaringsvariablene.

9.2 PCR

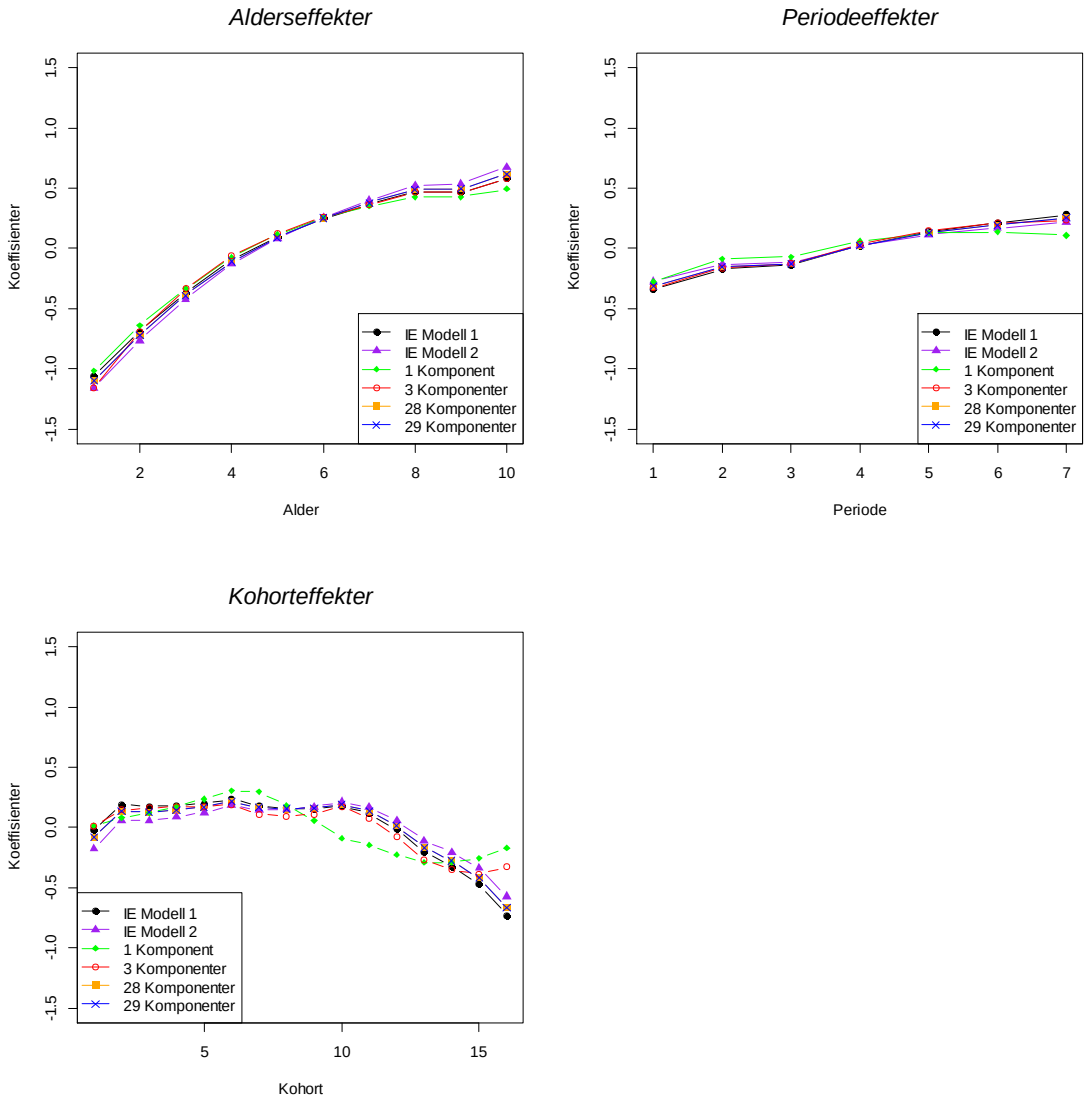
PCR er en forkortelse for *Principal Component Regression*, og er en prinsipal komponent analyse (PCA) anvendt i regresjon. Metoden tar utgangspunkt i komponentene som en finner ved PCA, og utfører regresjon på disse. Om en bruker alle komponentene fra PCA, vil disse spenne ut det samme variabelrommet som den opprinnelige $\boldsymbol{X}$-matrisen og PCR vil i dette tilfellet gi vanlig OLS-regresjon. Antallet av prinsipal komponenter er mindre enn eller likt antallet av originale variabler. Prinsipal komponentene er ordnet slik at den første prinsipal komponenten har den størst mulige variansen blant prinsipal komponentene, dvs. at den står for så mye av variabiliteten i dataene som mulig. Hver påfølgende komponent har den størst mulige variansen under betingelsen at den må være ortogonal (dvs. ukorrelert) til den foregående komponenten. Ved å utnytte at singulær verdi dekomposisjonen gir oss komponenter rangert etter viktighet, kan vi med et mindre antall komponenter beskrive størst mulig variasjon i datasettet og gjøre PCR på dette. Vi får dermed redusert antall variabler som tas med i videre analyse.

9.3 PLS

PLS er en forkortelse for *Partial Least Squares* [28] og oversettes med delvis minste kvadraters metode. Den omtales også som en projeksjon ned på underliggende strukturer. Metoden har klare likhetstrekk til prinsipal komponent regresjon (PCR). For der PCR benytter variansen til forklaringsvariablene som utgangspunkt for å lage komponenter, benytter PLS seg av kovariansen mellom variablene og responsen som utgangspunkt. PLS benytter seg av kovariansen mellom den kontinuerlige responsvektoren $\boldsymbol{y}$ og hver variabelvektor i designmatrisen $\boldsymbol{X}$ som utgangspunkt for en vektingsvektor [28]. Fordelen med å benytte seg av kovarians som grunnlag for vektingsvektoren i PLS sammenlignet med variansen til hver variabel i PCR, er at responsvektoren er med på å påvirke hvordan komponentene lages. Dette fører til at de første PLS-komponentene ofte er mer relevante for prediksjon sammenlignet med de første PCR-komponentene, og at PLS ofte trenger færre komponenter for å produsere gode modeller. I PCR-analysen er utvelgelsen av komponentene uavhengig av responsen $\boldsymbol{y}$, mens i PLS-analysen maksimeres kovariansen med responsen. Komponentene som velges gir et gunstig utgangspunkt for regresjonen, og slik kan kun de komponentene som ansees for å være viktigst i modellen inkluderes. Den første PLS-komponenten har den største kovariansen med responsen, den andre komponenten har den nest største kovariansen, og slik fortsetter det. På samme måte som ved PCR tar metoden utgangspunkt i komponentene som er valgt ut, og utfører regresjon på disse. PLS-regresjon er spesielt egnet når matrisen av prediktorer har flere variabler enn observasjoner, og når det er multikollinearitet blant $\boldsymbol{X}$-verdiene. I slike tilfeller vil ordinær regresjon mislykkes. PLS-regresjon er en teknikk for dimensjonreduksjon og er mye brukt i bioinformatikk.

9.4 Eksempel fra artikkel

I artikkel [5] har forfatterne sett på dødelighetsratene av leverkreft (Hepatocellular Carcinoma) blant Taiwanske menn ≥ 40 år, i perioden 1976 til 2008 (Tabell 3 i artikkel). Tabellen er med i vedleggsdelen. I dette datasettet er det 10 aldersgrupper, 7 perioder og 16 fødselskohorter. Det maksimale antallet av komponenter som kan benyttes i PLS-metoden er 29. Den naturlige log-transformasjonen av dødelighetsratene er responsen i dette eksempelet. Resultatene fra PLS-metoden med ulikt antall komponenter og resultatene fra IE-metoden er vist i figurene under. Forskjellen på de to IE-metodene er parametriseringen. I modell 1 kodes de siste gruppene med -1 (alder 10, periode 7 og kohort 16), mens i modell 2 kodes de første gruppene med -1 (alder 1, periode 1 og kohort 1). De ulike PLS-metodene har med 1, 3, 28 eller 29 komponenter i sin regresjonsanalyse.



Figur 9-1: Estimater for henholdsvis alder-, periode- og kohorteffekter ved bruk av IE og PLS.

Fra Figur 9-1 ser vi at resultatene kan tyde på at 3 komponenter er godt nok for PLS-metoden, og at inkludering av flere komponenter gir liten forbedring i modellen. Resultatene viser en økning i dødelighetsrate med alderen og periode, mens de siste fødselskohortene viser lavere risiko for å dø av leverkreft enn de eldre kohortene. Forskjellene mellom PLS-metodene med ulikt antall komponenter inkludert er liten, men for kohorteffektene synes metoden med 3 komponenter å ha en litt annerledes trend for de siste kohortene sammenlignet med metoden med flere komponenter inkludert.

Forfatterne nevner til slutt identifikasjonsproblemet, som oppstår ved analyse av alder- periode- og kohorteffekter, som følge av den lineære avhengigheten mellom disse 3 faktorene. Ideelt sett bør valget av en betingelse i en modell velges utfra teori man har om materialet, men som oftest er slik informasjon fraværende. Om en matematisk løsning på problemet er forsvarlig vil avhenge av den innførte betingelsen. Forfatterne konkluderer derfor med at når man ikke har informasjon om hvilken betingelse som er mest hensiktsmessig å innføre i modellen, synes resultatene fra PLS- og IE-metodene å være mer tolkbare enn metoder med vilkårlige betingelser.

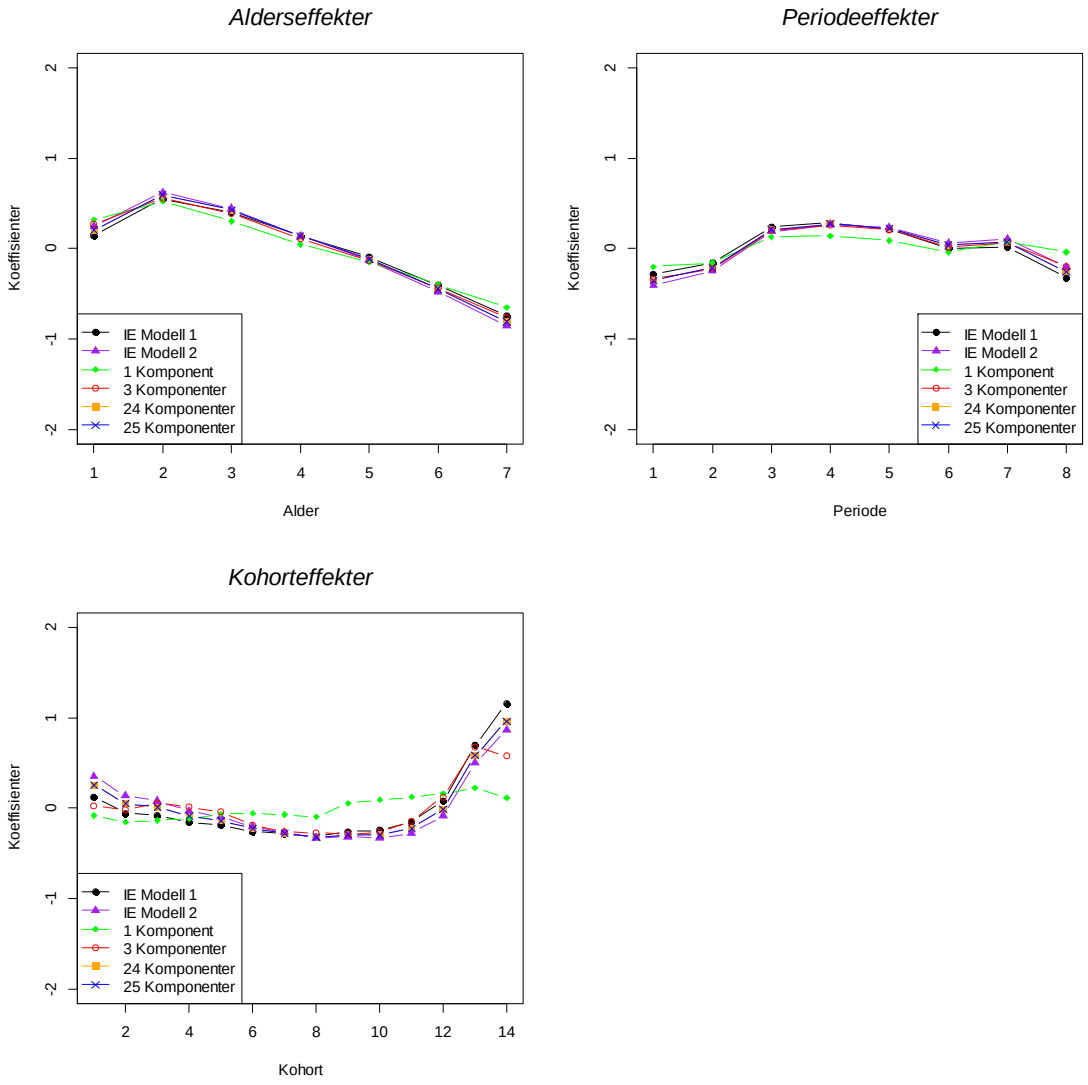
9.5 PLS, IE og CGLIM på datasett fra O'Brien (2000)

Dataene som er benyttet er Homicide arrestrates data fra Figur 1 i [29].

Tabell 9-1: Homicide arrest rates between age 15 and 49 during the years 1960 to 1995.

		Periode							
		1960	*1965*	*1970*	*1975*	*1980*	*1985*	*1990*	*1995*
Aldersgruppe	15-19	8.98	9.07	17.22	17.54	18.02	16.32	36.52	35.24
	20-24	14.00	15.18	23.76	25.62	23.95	21.11	29.10	32.34
	25-29	13.45	14.69	20.09	21.05	18.91	16.79	17.99	16.75
	30-34	10.73	11.70	16.00	15.81	15.22	12.59	12.44	10.05
	35-39	9.37	9.76	13.13	12.83	12.31	9.60	9.38	7.27
	40-44	6.48	7.41	10.10	10.52	8.79	7.50	6.81	5.48
	45-49	5.71	5.56	7.51	7.32	6.76	5.31	5.17	3.67

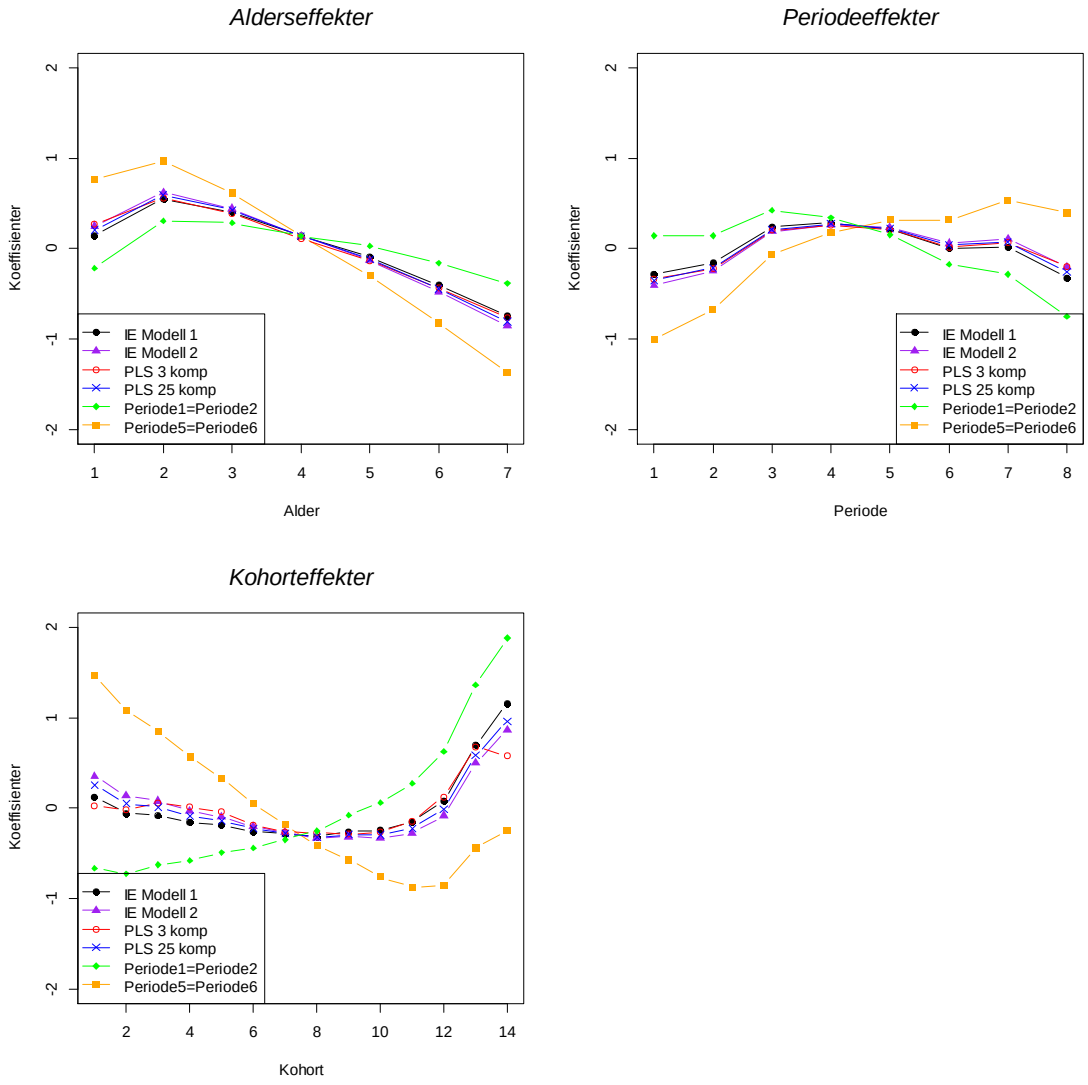
I dette datasettet er det 7 aldersgrupper, 8 perioder og 14 fødselskohorter. Det maksimale antallet av komponenter som kan benyttes i PLS-metoden er 25. Den naturlige log-transformasjonen av mordarrestratene er responsen i dette eksempelet. Resultatene fra PLS-metoder med ulikt antall komponenter og resultatene fra IE-metoden er vist i Figur 9-2. Forskjellen på de to IE-metodene er parametriseringen. I modell 1 kodes de siste gruppene med -1 (alder 7, periode 8 og kohort 14), mens i modell 2 kodes de første gruppene med -1 (alder 1, periode 1 og kohort 1). De ulike PLS-metodene har med 1, 3, 24 eller 25 komponenter i sin regresjonsanalyse.



Figur 9-2: Estimater for henholdsvis alder-, periode- og kohorteffekter for mordarrestdataene.

Fra Figur 9-2 ser vi at resultatene kan tyde på at 3 komponenter er godt nok for PLS-metoden, og at inkludering av flere komponenter gir liten forbedring i modellen. Forskjellene mellom PLS-metodene med ulikt antall komponenter inkludert er liten, men for kohorteffektene synes metoden med 3 komponenter å ha en litt annerledes trend for første og siste kohorten sammenlignet med metodene med flere komponenter inkludert.

Videre er dataene analysert med IE-metoden, to varianter av PLS-metoden der 3 og 25 komponenter er inkludert i analysen, samt med to varianter av CGLIM-metoden. Den ene CGLIM-metoden har betingelsen at periode 1 er lik periode 2 (P1=P2) og den andre har betingelsen at periode 5 er lik periode 6 (P5=P6). For CGLIM-metoden benyttes sentrerte verdier for å lettere kunne sammenligne.



Figur 9-3: Estimater for henholdsvis alder-, periode- og kohorteffekter for mordarrestdataene.

I CGLIM_P settes først periode 1 lik periode 2, dvs. at periodeeffektene settes lik hverandre i 1960 og 1965. I denne modellen øker først periodeeffektene til en topp i 1970 og minker deretter helt frem mot 1995. For alderseffekten øker effekten først fra aldergruppen 15-19 år frem mot 20-24 år, for deretter å avta jevnt frem mot 45-49 år. Kohorteffekten falt litt fra 1915 til det laveste punktet i 1920, deretter var det en gradvis økning frem mot 1965, og deretter en kraftig økning frem mot 1980.
I den andre varianten av CGLIM_P, der periode 5 er satt lik periode 6, forutsetter en å ha lik periodeeffekt i 1980 og 1985. Trenden i periodeeffektene er da økende med unntak av en svak nedgang i de to siste periodene, fra 1990 til 1995. Alderseffekten økte litt fra gruppen 15-19 år til gruppen 20-24 år, og deretter avtok jevnt frem mot 45-49 år. Kohorteffekten hadde en kraftig nedgang fra 1915 til 1965 for deretter å øke kraftig igjen mellom 1970 og 1980. Når en sammenligner de to variantene av CGLIM_P-

metoden, gir de ikke noe ensartet fornuftig trendestimat.
IE-metoden viser en økning i alderseffekten fra alder 15 år til tidlig i 20-årene, og en videre nedgang frem mot alder 49 år. En ser også at periodetrenden svinger fra 1960 til 1995, med en topp mellom 1970 og 1980. Kohorttrenden er svakt synkende fra 1915 til 1950, og øker gradvis videre fra 1950 til 1970, etterfulgt av en kraftig økt trend til 1980.
Effektene som estimeres med PLS-metoden er tilnærmet like de estimatene en får fra IE-metoden.

I artikkel [30] er resultatene fra IE-metoden og CGLIM-metoden sammenlignet med resultatene O'Brien fikk ved analyse av dette datasettet i sin artikkel fra 2000 [29]. De estimerte trendene ved IE og CGLIM er ganske lik de trendene O'Brien estimerte med en APC-karakteristisk modellanalyse (APCC). APCC er en metode for å teste teorier som involverer alder-, periode- og kohorteffekter, og studier som benytter denne metoden fokuserer ofte på kun en enkelt kohortkarakteristikk og kontrollerer for aldersgrupper og perioder. I analysen til O'Brien fokuserer de på to kohortkarakteristikker i tillegg til de faste effektene til aldersgrupper og perioder. Kohortkarakteristikker som forfatterne har en teori om at kan ha innvirkning på mordarrestratene er den relative kohortstørrelsen, og prosentandelen av ikke-ekteskapelige fødsler.
Mønsteret til kurvene for alder- og periodeeffekter vist over, er helt sammenlignbare med de trendene som O'Brien rapporterer. Kohorteffektene plukker også opp virkningene fra de to kohortkarakteristikkene som forfatterne mener kan ha innvirkning. Den relativt store babyboomen på 1950- og 1960-tallet, og den raske økningen i prosentandelen av ikke-ekteskapelige fødsler på 1970- og 1980-tallet.
Fu [30] konkluderer med at IE-metoden leder til fornuftige estimater for mordarrestrate dataene, mens de to variantene av CGLIM-metoden ikke gjør det. I tillegg viser resultatene fra analysene at PLS-metoden også gir fornuftige estimater for trendene i dette datasettet.

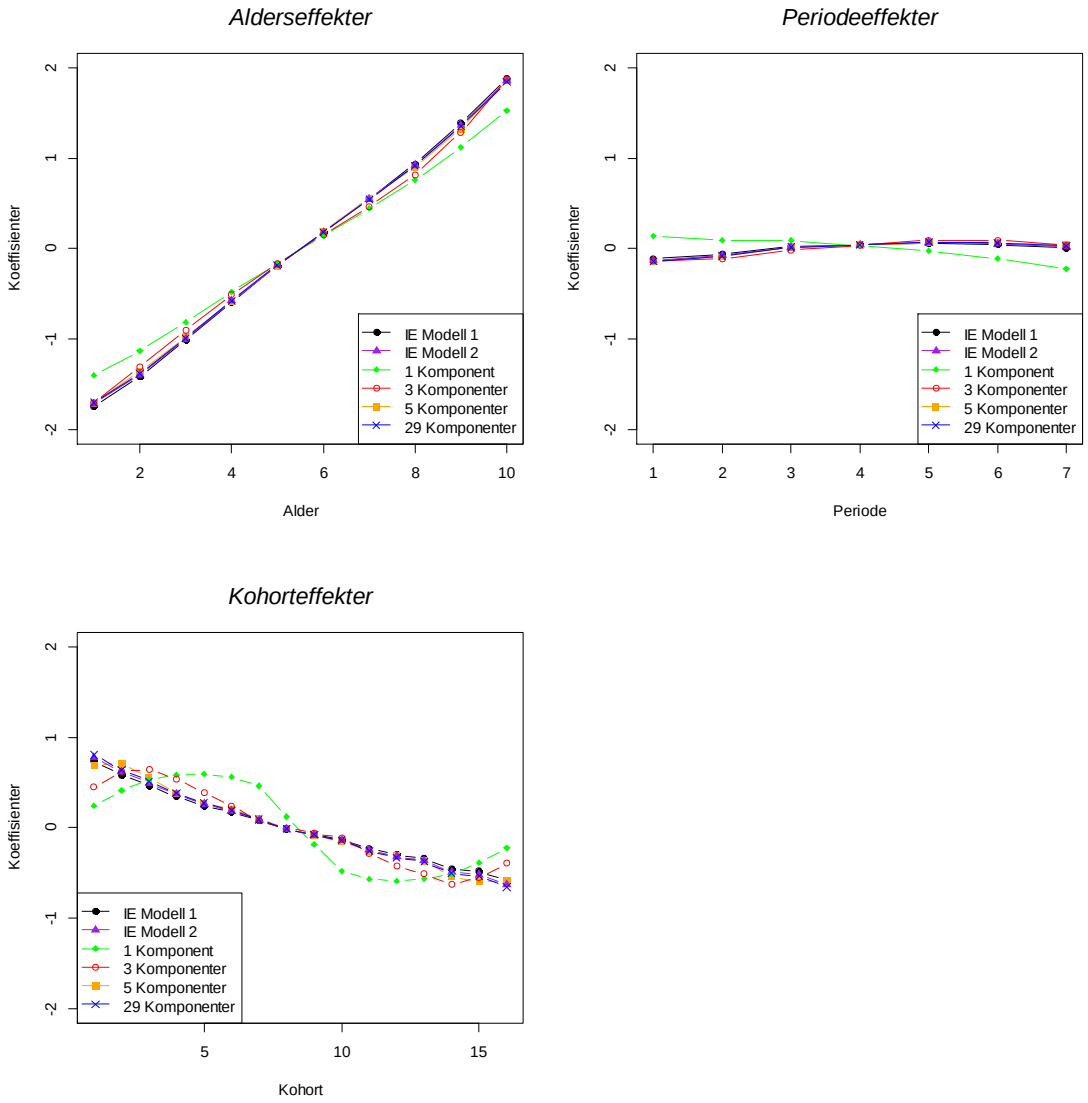
9.6 Analyse av datasett fra Statistisk sentralbyrå

Datasettet er hentet fra Tabell 79 i Statistisk sentralbyrå sin årbok fra 2011 [15]. Tabellen omhandler dødelighet, fordelt etter kjønn og alder. Utvalget fra tabellen er dødelighet for kvinner i aldersgruppen 30 til 79 år, for perioden 1976 til 2010. Både aldersgruppene og periodene er inndelt i 5 års intervaller.

Tabell 9-2: Dødelighet kvinner, etter alder. Døde pr. 100 000 middelfolkemengde. Årlig gjennomsnitt. Kilde: Statistisk sentralbyrå.

		Periode						
		1976-1980	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010
Aldersgruppe	30-34	50	49	52	53	48	45	38
	35-39	79	80	77	71	73	64	58
	40-44	125	127	129	120	114	104	85
	45-49	211	206	214	196	182	166	150
	50-54	353	323	318	299	310	272	252
	55-59	536	524	526	479	454	433	379
	60-64	853	813	818	778	709	658	617
	65-69	1450	1319	1291	1259	1129	1021	940
	70-74	2534	2346	2270	2112	1935	1747	1566
	75-79	4806	4296	4123	3847	3461	3202	2846

I dette datasettet er det 10 aldersgrupper, 7 perioder og 16 fødselskohorter. Det maksimale antallet av komponenter som kan benyttes i PLS-metoden er 29. Den naturlige log-transformasjonen av dødsratene er responsen i dette eksempelet. Resultatene fra PLS-metoder med ulikt antall komponenter og resultatene fra IE-metoden er vist i Figur 9-4. Forskjellen på de to IE-metodene er parametriseringen. I modell 1 kodes de siste gruppene med -1 (alder 10, periode 7 og kohort 16), mens i modell 2 kodes de første gruppene med -1 (alder 1, periode 1 og kohort 1). De ulike PLS-metodene har med 1, 3, 5 eller 29 komponenter i sin regresjonsanalyse.



Figur 9-4: Estimater for henholdsvis alder-, periode- og kohorteffekter for dødelighetsdataene.

Resultatene kan tyde på at 5 komponenter er godt nok for PLS-metoden, og at inkludering av flere komponenter gir liten forbedring i modellen. PLS-metoden med 1 komponent gir ulike estimater i forhold til de andre metodene både for alder-, periode- og kohorteffekter. Forskjellene mellom PLS-metodene med flere komponenter inkludert er liten, men for kohorteffektene synes metoden med 3 komponenter å ha en litt annerledes trend for de første og siste kohortene sammenlignet med metodene med flere komponenter inkludert. Både IE- og PLS-metoden viser de samme effektene for både alder, periode og kohort. Alderseffekten er jevnt stigende med økende alder. Det er en liten endring i periodeeffekten, med en liten økning frem mot periode 5 (1996-2000), for deretter å avta litt for de siste periodene. Trenden for kohorteffektene viser en jevn nedgang for de kohortene som er inkludert i datasettet, fødselskohort 1896 til 1980.

9.7 PLS, IE og CGLIM på datasett fra Yang et al. (2004)

I artikkel [14] har forfatterne sett på dødeligheten blant kvinner i USA i perioden 1960 til 1999. Først tar de for seg ulike reduserte modeller og presenterer estimatene i Tabell 3 i artikkelen, sammen med estimater fra 3 ulike varianter av CGLIM-metoden. I CGLIM_A settes effekten til aldersgruppe 2 lik aldersgruppe 3. I CGLIM_P settes effekten til periode 1 lik periode 2, og i CGLIM_C settes de to siste kohortene lik hverandre (kohort 25 = kohort 26). Deretter har forfatterne anvendt IE-metoden for å analysere de samme dødelighetsdataene, og estimatene er presentert i Tabell 5 i artikkelen. Jeg har analysert dataene med de samme metodene og jeg har også benyttet PLS-metoden på de samme dataene. De ulike PLS-metodene har med 3, 24 og 45 komponenter i sin regresjonsanalyse. Estimatene for log-koeffisientene fra de ulike metodene er vist i tabellene under. IE-metoden benytter ikke referansekategorier for alder-, periode- og kohortkoeffisientene, mens CGLIM-estimatorene gjør det. Dette medfører at jeg har 2 ulike måter å sentrere parametrene på, og forskjellen mellom måtene er omgjøring ved hjelp av en konstant. For eksempel kan man trekke fra IE-estimatet for den første aldersgruppen fra alle alderseffektestimatene for IE-metoden, og dermed få en referansekategori. Tilsvarende kan også gjøres for de andre kategoriene, som forenkler sammenligningen av de ulike koeffisientestimatene. Koeffisientene fra PLS-metoden kan også omgjøres på samme måte for lettere sammenligning. Datasettet og R-kodene er tatt med i vedleggsdelen.

Tabell 9-3: Estimater fra ulike modeller, alderseffekter.

Aldersgrupper	Koeffisienter fra ulike modeller, U.S. Female Mortality						
	CGLIM (A2=A3)	CGLIM (P1=P2)	CGLIM (C25=C26)	Intrinsic Estimator	PLS 3 komp	PLS 24 komp	PLS 45 komp
0 – 4	0,000	0,000	0,000	0,453	0,232	0,458	0,457
5 – 9	-2,387	-2,567	-2,332	-2,144	-2,267	-2,134	-2,134
10 – 14	-2,387	-2,747	-2,277	-2,354	-2,264	-2,334	-2,335
15 – 19	-1,527	-2,068	-1,363	-1,704	-1,421	-1,680	-1,679
20 – 24	-1,243	-1,964	-1,024	-1,630	-1,270	-1,611	-1,611
25 – 29	-0,974	-1,875	-0,700	-1,571	-1,176	-1,548	-1,547
30 – 34	-0,570	-1,651	-0,242	-1,377	-0,985	-1,354	-1,353
35 – 39	-0,074	-1,335	0,309	-1,091	-0,741	-1,071	-1,071
40 – 44	0,477	-0,965	0,914	-0,751	-0,481	-0,741	-0,741
45 – 49	1,040	-0,582	1,532	-0,398	-0,243	-0,393	-0,393
50 – 54	1,591	-0,211	2,137	-0,057	-0,025	-0,053	-0,053
55 – 59	2,124	0,141	2,725	0,266	0,164	0,266	0,265
60 – 64	2,678	0,515	3,334	0,610	0,390	0,603	0,604
65 – 69	3,233	0,891	3,944	0,956	0,643	0,942	0,942
70 – 74	3,819	1,296	4,584	1,331	0,965	1,314	1,313
75 – 79	4,422	1,719	5,242	1,724	1,351	1,705	1,705
80 – 84	5,064	2,181	5,939	2,157	1,830	2,134	2,133
85 – 89	5,708	2,644	6,637	2,590	2,368	2,558	2,558
90 – 94	6,316	3,072	7,300	2,988	2,932	2,938	2,938

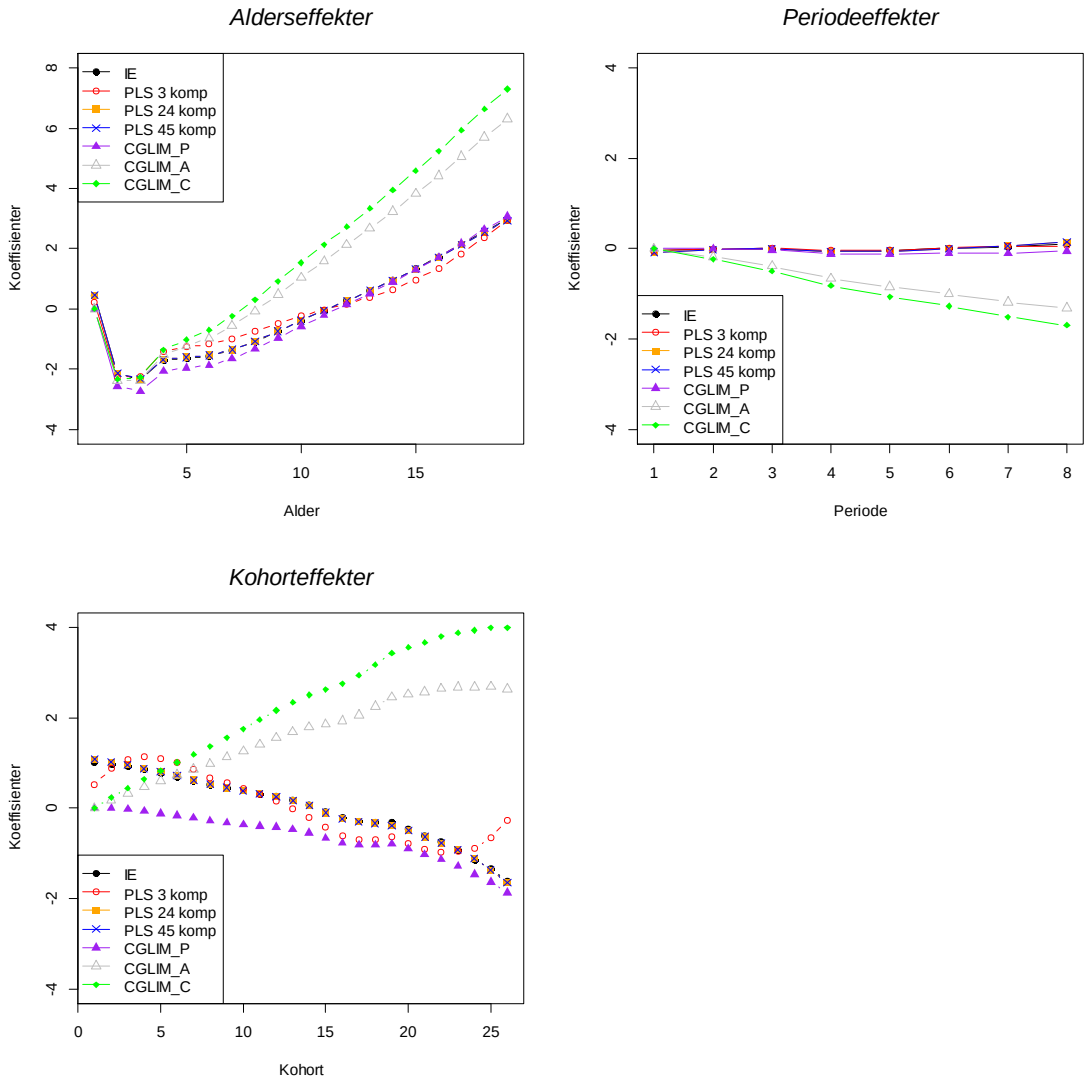
Tabell 9-4: Estimater fra ulike modeller, periodeeffekter.

Periode	Koeffisienter fra ulike modeller, U.S. Female Mortality						
	CGLIM (A2=A3)	CGLIM (P1=P2)	CGLIM (C25=C26)	Intrinsic Estimator	PLS 3 komp	PLS 24 komp	PLS 45 komp
1960 – 1964	0,000	0,000	0,000	-0,039	-0,051	-0,077	-0,078
1965 – 1969	-0,180	0,000	-0,235	-0,009	-0,011	-0,019	-0,020
1970 – 1974	-0,388	-0,028	-0,498	-0,007	0,012	0,005	0,005
1975 – 1979	-0,658	-0,118	-0,822	-0,067	-0,044	-0,059	-0,060
1980 – 1984	-0,844	-0,123	-1,062	-0,043	-0,035	-0,057	-0,057
1985 – 1989	-1,000	-0,099	-1,273	0,011	0,019	0,002	0,002
1990 – 1994	-1,183	-0,102	-1,511	0,038	0,047	0,061	0,062
1995 – 1999	-1,316	-0,054	-1,698	0,115	0,062	0,143	0,142

Tabell 9-5: Estimater fra ulike modeller, kohorteffekter.

Kohort	Koeffisienter fra ulike modeller, U.S. Female Mortality						
	CGLIM (A2=A3)	CGLIM (P1=P2)	CGLIM (C25=C26)	Intrinsic Estimator	PLS 3 komp	PLS 24 komp	PLS 45 komp
1870	0,000	0,000	0,000	1,008	0,529	1,090	1,091
1875	0,178	-0,002	0,233	0,977	0,879	1,028	1,028
1880	0,333	-0,027	0,443	0,922	1,069	0,953	0,953
1885	0,475	-0,066	0,639	0,853	1,135	0,880	0,880
1890	0,607	-0,113	0,826	0,776	1,108	0,805	0,805
1895	0,740	-0,161	1,013	0,698	1,012	0,725	0,724
1900	0,862	-0,219	1,190	0,610	0,854	0,627	0,626
1905	0,983	-0,278	1,366	0,522	0,669	0,528	0,527
1910	1,127	-0,315	1,564	0,455	0,562	0,457	0,456
1915	1,264	-0,357	1,756	0,383	0,437	0,388	0,387
1920	1,409	-0,393	1,955	0,317	0,300	0,321	0,322
1925	1,563	-0,419	2,165	0,262	0,161	0,261	0,260
1930	1,690	-0,473	2,346	0,178	-0,013	0,169	0,168
1935	1,799	-0,544	2,510	0,077	-0,203	0,054	0,054
1940	1,865	-0,658	2,630	-0,067	-0,419	-0,098	-0,098
1945	1,938	-0,766	2,758	-0,204	-0,602	-0,238	-0,238
1950	2,065	-0,818	2,940	-0,287	-0,690	-0,306	-0,306
1955	2,250	-0,813	3,180	-0,312	-0,696	-0,332	-0,333
1960	2,453	-0,790	3,437	-0,319	-0,642	-0,380	-0,381
1965	2,522	-0,902	3,560	-0,460	-0,788	-0,494	-0,494
1970	2,572	-1,032	3,666	-0,620	-0,912	-0,632	-0,631
1975	2,655	-1,130	3,802	-0,748	-0,965	-0,763	-0,764
1980	2,678	-1,287	3,880	-0,934	-0,963	-0,917	-0,916
1985	2,685	-1,460	3,942	-1,137	-0,893	-1,115	-1,114
1990	2,690	-1,635	4,002	-1,342	-0,660	-1,365	-1,366
1995	2,635	-1,870	4,002	-1,607	-0,268	-1,647	-1,646

Som vist i Tabell 4 i [2] vil sentrering av de ulike variantene av CGLIM-metoden føre til at intercept blir identisk for alle variantene og IE-metoden. Siden dette leddet er av mindre interesse ved sammenligning av effektkoeffisientene her, er det utelatt fra tabellene.



Figur 9-5: Estimater for henholdsvis alder-, periode- og kohorteffekter for dataene i artikkelen til Yang et al. (2004).

Forskjellene mellom PLS-metodene med 24 og 45 komponenter inkludert er liten, og inkludering av de ekstra komponentene i modellen gir liten forbedring. Begge de to PLS-metodene og IE-metoden gir tilnærmet identiske estimater.
CGLIM_A setter alder 2 (-2,144(estimat fra IE)) lik alder 3 (-2,354), CGLIM_P setter periode 1 (-0,039) lik periode 2 (-0,009) og CGLIM_C setter kohort 25 (-1,342) lik kohort 26 (-1,607).
I CGLIM_P stemmer betingelsen som er satt bedre overens med virkeligheten enn i de to andre modellene, der effektene som settes å være lik hverandre avviker mer. Som vi har sett tidligere er det også her størst avvik fra de antatte sanne verdiene for de modellene som har mest uriktig betingelser, nemlig CGLIM_A og CGLIM_C.
Estimatene fra IE- og PLS-metoden viser at variasjoner i dødeligheten stammer hovedsakelig fra alder- og kohorteffekter. Alderseffektene skildrer et velkjent mønster for hvordan dødeligheten endres gjennom livsløpet.

9.8 Simuleringsoppsett PLS vs. IE

På samme måte som tidligere er datasettene simulert utfra en oppgitt simuleringsmodell, og resultatene er basert på 10 000 simuleringer.

Datasettene er denne gangen generert utfra:

$$y_{ij} \sim \text{Poisson}\{\exp\left[1.5 + \exp(0.07 \cdot \text{alder}_{ij})^2 + 0.1 \cdot (0.4 \cdot \text{periode}_{ij} - 3)^2 + 2 \cdot (\exp(0.5^{\text{kohort}_{ij}} \cdot \text{kohort}_{ij}) - 1) - (\frac{1}{\text{kohort}_{ij}^{\,2}})\right]\}$$

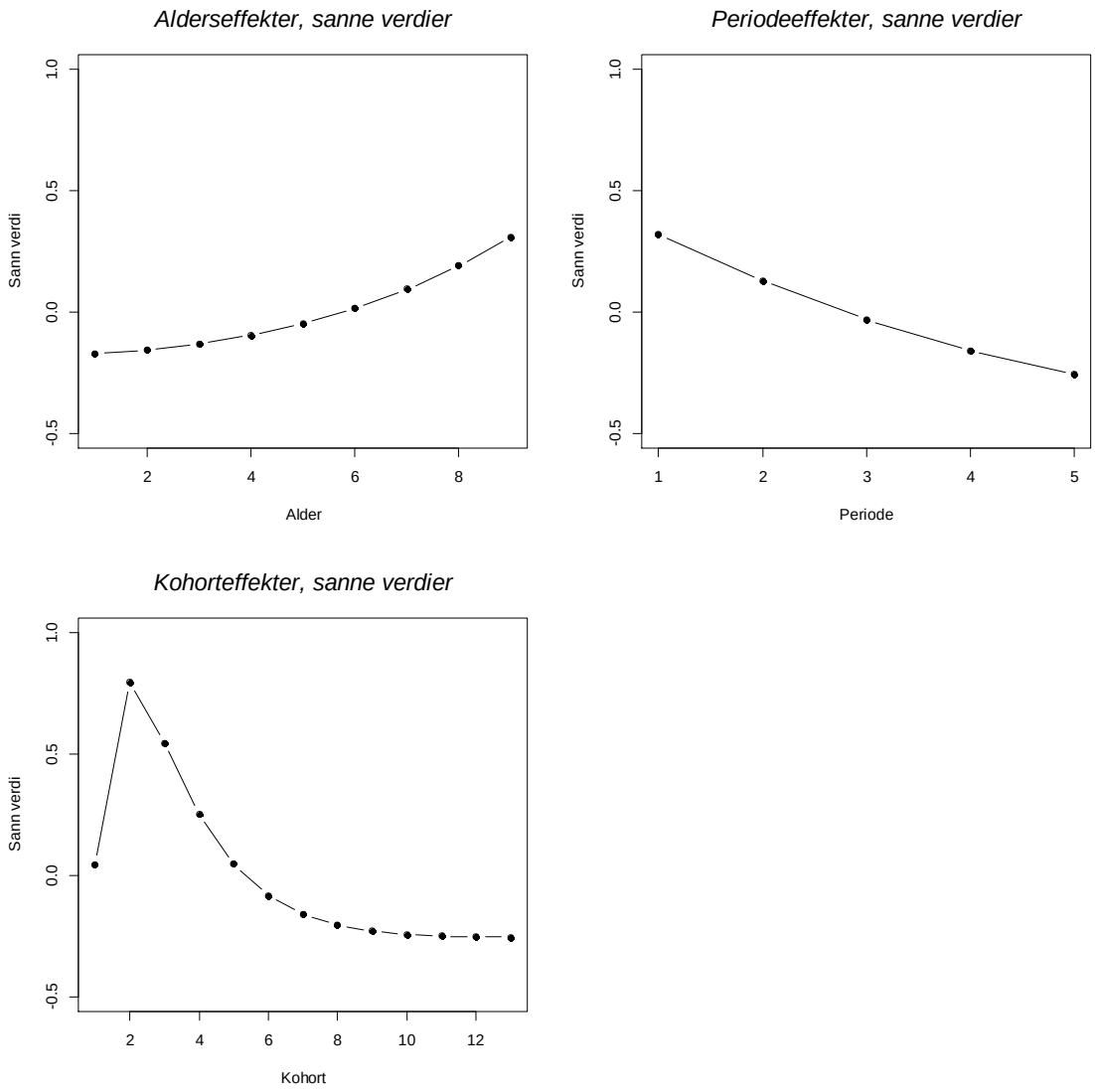
Dette er samme simuleringsmodell som i Kapittel 8.3.6. I simuleringsmodellen er det med 9 aldersgrupper, 5 perioder og 13 fødselskohorter. Den naturlige log-transformasjonen av de simulerte datasettene er responsen i denne modellen. Det maksimale antallet av komponenter som kan benyttes i PLS-metoden er 23. Resultatene fra PLS-metoder med ulikt antall komponenter og resultater fra IE-metoden er vist i tabeller og figurer under. IE-metoden benytter effektkoding, og koder de siste gruppene med -1 (alder 9, periode 5 og kohort 13). De ulike PLS-metodene har med 1, 3, 5 eller 22 komponenter i sin regresjonsanalyse.

I Tabell 9-6 er de sanne effektene for henholdsvis alder, periode og kohort oppgitt.

Tabell 9-6: Sanne alder-, periode- og kohorteffekter i modellen.

Sanne alder-, periode- og kohorteffekter i simuleringsmodellen		
alder	**periode**	**kohort**
$a1 = -0.173$	$p1 = 0.320$	$c1 = 0.045$
$a2 = -0.158$	$p2 = 0.128$	$c2 = 0.795$
$a3 = -0.133$	$p3 = -0.032$	$c3 = 0.546$
$a4 = -0.096$	$p4 = -0.160$	$c4 = 0.253$
$a5 = -0.048$	$p5 = -0.256$	$c5 = 0.045$
$a6 = 0.015$		$c6 = -0.084$
$a7 = 0.093$		$c7 = -0.161$
$a8 = 0.190$		$c8 = -0.205$
$a9 = 0.309$		$c9 = -0.230$
		$c10 = -0.243$
		$c11 = -0.250$
		$c12 = -0.254$
		$c13 = -0.256$

Figur 9-6 illustrerer de sanne effektene.



Figur 9-6: Sanne effekter for henholdsvis alder, periode og kohort i simuleringsmodellen.

Fire ulike PLS-metoder og IE-metoden er benyttet for å estimere de ulike effektene for alder, periode og kohort. I tabellene som følger er verdiene for hver parameter oppgitt. Parameterverdiene er oppgitt som gjennomsnittet av de 10 000 simuleringene. Verdiene for MSE er mean basert på 10 000 simuleringer.

9.9 Simuleringsresultater

Tabell 9-7: Simuleringsresultater fra IE- og PLS-estimatorer, alderseffekter.

Simuleringsresultat IE- og PLS-estimatorer (n=10 000), alderseffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	PLS 1 komp	PLS 3 komp	PLS 5 komp	PLS 22 komp
Alder 1	Mean	-0,173	-0,331	-0,215	-0,302	-0,296	-0,293
	SD		0,105	0,057	0,107	0,102	0,105
	MSE		0,036	0,005	0,028	0,025	0,026
Alder 2	Mean	-0,158	-0,277	-0,201	-0,248	-0,254	-0,249
	SD		0,099	0,057	0,112	0,106	0,110
	MSE		0,024	0,005	0,021	0,021	0,020
Alder 3	Mean	-0,133	-0,212	-0,177	-0,192	-0,201	-0,195
	SD		0,100	0,055	0,113	0,107	0,109
	MSE		0,016	0,005	0,016	0,016	0,016
Alder 4	Mean	-0,096	-0,135	-0,138	-0,126	-0,132	-0,128
	SD		0,099	0,054	0,112	0,105	0,106
	MSE		0,011	0,005	0,014	0,012	0,012
Alder 5	Mean	-0,048	-0,047	-0,079	-0,056	-0,048	-0,048
	SD		0,094	0,052	0,107	0,099	0,100
	MSE		0,009	0,004	0,011	0,010	0,010
Alder 6	Mean	0,015	0,053	0,008	0,022	0,052	0,043
	SD		0,086	0,048	0,098	0,092	0,093
	MSE		0,009	0,002	0,010	0,010	0,009
Alder 7	Mean	0,093	0,172	0,135	0,138	0,163	0,154
	SD		0,076	0,044	0,088	0,084	0,085
	MSE		0,012	0,004	0,010	0,012	0,011
Alder 8	Mean	0,190	0,309	0,293	0,315	0,286	0,283
	SD		0,068	0,040	0,079	0,076	0,078
	MSE		0,019	0,012	0,022	0,015	0,015
Alder 9	Mean	0,309	0,468	0,373	0,450	0,430	0,435
	SD		0,073	0,038	0,069	0,073	0,078
	MSE		0,030	0,005	0,025	0,020	0,022
Total MSE			0,166	0,047	0,156	0,141	0,141

Tabell 9-8: Simuleringsresultater fra IE- og PLS-estimatorer, periodeeffekter.

Simuleringsresultat IE- og PLS-estimatorer (n=10 000), periodeeffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	PLS 1 komp	PLS 3 komp	PLS 5 komp	PLS 22 komp
Periode 1	Mean	0,320	0,401	0,460	0,387	0,386	0,387
	SD		0,053	0,050	0,053	0,055	0,057
	MSE		0,009	0,022	0,007	0,007	0,008
Periode 2	Mean	0,128	0,167	0,237	0,173	0,163	0,161
	SD		0,054	0,056	0,057	0,058	0,061
	MSE		0,004	0,015	0,005	0,005	0,005
Periode 3	Mean	-0,032	-0,032	-0,038	-0,032	-0,030	-0,033
	SD		0,060	0,062	0,061	0,063	0,066
	MSE		0,004	0,004	0,004	0,004	0,004
Periode 4	Mean	-0,160	-0,200	-0,253	-0,196	-0,194	-0,193
	SD		0,065	0,067	0,066	0,068	0,072
	MSE		0,006	0,013	0,006	0,006	0,006
Periode 5	Mean	-0,256	-0,337	-0,406	-0,331	-0,324	-0,322
	SD		0,072	0,069	0,071	0,071	0,073
	MSE		0,012	0,027	0,011	0,010	0,010
Total MSE			0,035	0,081	0,033	0,031	0,033

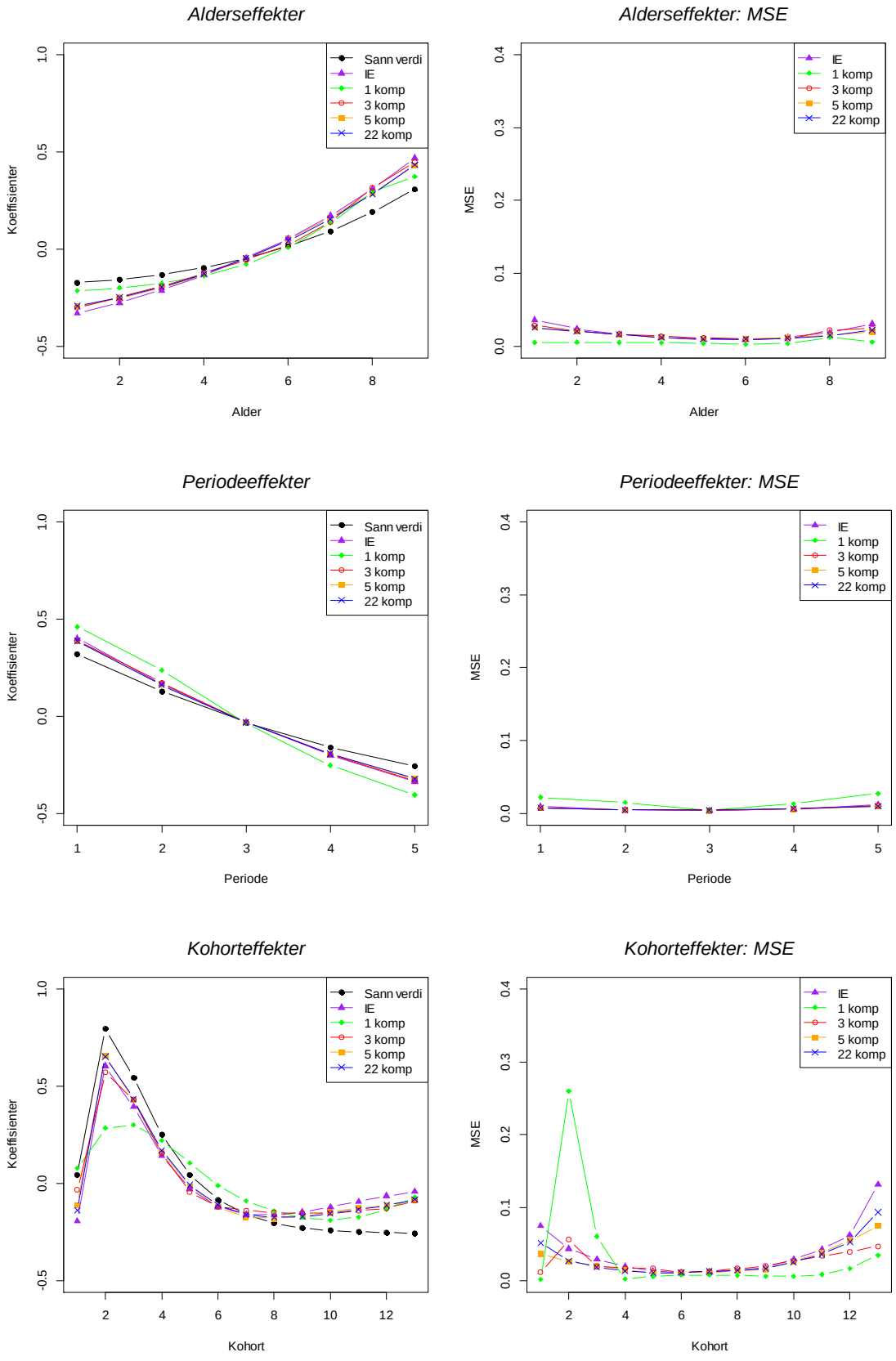
Tabell 9-9: Simuleringsresultater fra IE- og PLS-estimatorer, kohorteffekter.

Simuleringsresultat IE- og PLS-estimatorer (n=10 000), kohorteffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	PLS 1 komp	PLS 3 komp	PLS 5 komp	PLS 22 komp
Kohort 1	Mean	0,045	-0,194	0,077	-0,033	-0,111	-0,138
	SD		0,133	0,015	0,072	0,114	0,136
	MSE		0,075	0,001	0,011	0,037	0,052
Kohort 2	Mean	0,795	0,604	0,285	0,571	0,658	0,652
	SD		0,083	0,018	0,077	0,084	0,084
	MSE		0,043	0,260	0,056	0,026	0,027
Kohort 3	Mean	0,546	0,394	0,301	0,430	0,429	0,434
	SD		0,078	0,025	0,075	0,079	0,077
	MSE		0,029	0,060	0,019	0,020	0,018
Kohort 4	Mean	0,253	0,141	0,221	0,155	0,160	0,171
	SD		0,083	0,034	0,086	0,080	0,082
	MSE		0,019	0,002	0,017	0,015	0,013
Kohort 5	Mean	0,045	-0,028	0,105	-0,046	-0,014	-0,009
	SD		0,085	0,043	0,089	0,083	0,085
	MSE		0,013	0,005	0,016	0,010	0,010
Kohort 6	Mean	-0,084	-0,120	-0,012	-0,120	-0,124	-0,112
	SD		0,097	0,047	0,099	0,095	0,099
	MSE		0,011	0,007	0,011	0,011	0,011
Kohort 7	Mean	-0,161	-0,160	-0,091	-0,140	-0,173	-0,162
	SD		0,108	0,051	0,109	0,106	0,111
	MSE		0,012	0,008	0,012	0,011	0,012
Kohort 8	Mean	-0,205	-0,163	-0,142	-0,150	-0,181	-0,176
	SD		0,110	0,054	0,115	0,111	0,116
	MSE		0,014	0,007	0,016	0,013	0,014
Kohort 9	Mean	-0,230	-0,149	-0,177	-0,154	-0,171	-0,173
	SD		0,105	0,055	0,119	0,110	0,114
	MSE		0,018	0,006	0,020	0,016	0,016
Kohort 10	Mean	-0,243	-0,122	-0,190	-0,154	-0,148	-0,155
	SD		0,121	0,052	0,137	0,130	0,132
	MSE		0,029	0,006	0,027	0,026	0,025
Kohort 11	Mean	-0,250	-0,094	-0,174	-0,143	-0,127	-0,136
	SD		0,136	0,048	0,148	0,153	0,151
	MSE		0,043	0,008	0,034	0,038	0,036
Kohort 12	Mean	-0,254	-0,065	-0,133	-0,128	-0,111	-0,113
	SD		0,164	0,041	0,153	0,185	0,182
	MSE		0,062	0,016	0,039	0,054	0,053
Kohort 13	Mean	-0,256	-0,042	-0,073	-0,090	-0,086	-0,084
	SD		0,293	0,031	0,140	0,216	0,253
	MSE		0,132	0,034	0,047	0,075	0,094
Total MSE			0,499	0,421	0,325	0,353	0,382

Tabell 9-10: Mål på de ulike modellene.

Mål på de ulike modellene					
	IE Modell 1	PLS 1 komp	PLS 3 komp	PLS 5 komp	PLS 22 komp
Samlet MSE	0,701	0,549	0,513	0,525	0,556

Resultatene er også illustrert i Figur 9-7. Den svarte kurven viser de sanne effektene.



Figur 9-7: Simuleringsresultater for IE- og PLS-estimatorer.

9.10 Kommentarer til resultatene

For PLS-metoden er resultater fra 1,3,5 og 22 komponenter tatt med. Simuleringene ble også utført med en PLS-metode med 10 komponenter inkludert, men da resultatene var helt identiske til PLS-metoden med 22 komponenter både når det gjaldt estimater, varians og MSE, blir ikke resultatene gjengitt i tabellen.

Resultatene kan tyde på at 3 eller 5 komponenter er godt nok for PLS-metoden, og at inkludering av flere komponenter gir liten forbedring i modellen. Fra Tabell 9-10 ser vi at den totale MSE for PLS-metoden med 3 komponenter er lavest. Forskjellene mellom estimatene for PLS-metodene med 3, 5 og 22 antall komponenter inkludert er liten, men for kohorteffektene klarer metoden med 3 komponenter i mindre grad å gjenskape den plutselige endringen i kohorteffekt for fødselskohort 2. PLS-metoden som kun har med 1 komponent klarer i mye mindre grad å gjenskape den plutselige endringen, mens PLS-metoden med 5 komponenter klarer like godt som metoden med 22 komponenter å detektere denne endringen. IE-metoden gir tilnærmet like gode estimater for de ulike alder- periode- og kohorteffektene som PLS-metoden, og begge metodene har ganske lik variasjon i sine estimater.

For alle metodene ser vi en tendens til at estimatene for kohortene i endene avviker mest fra sannheten. De første og siste fødselskohortene er dårligst representert i datasettene.

10. Oppsummering og videre arbeid

Utfordringene knyttet til alder-periode-kohort-analyser er velkjente. Problemet med å få pålitelige estimater for de simultane effektene til alder, periode og kohort har lenge vært et tema for diskusjon og forskning. Den lineære avhengigheten mellom alder, periode og kohort gjør at det oppstår problemer med å estimere de separate effektene til parametrene. Modellidentifikasjonsproblemet har ført til utviklingen av en rekke APC-metoder som et forsøk på å løse problemet.

Hovedfokuset i denne oppgaven har vært å se nærmere på en av de nyere metodene som er introdusert for estimering av de simultane effektene til alder, periode og kohort, omtalt som *intrinsic estimator* (IE). Ved hjelp av simuleringsanalyser er IE-metoden sammenlignet med mer tradisjonelle og velkjente metoder i Kapittel 7 og 8, og i Kapittel 9 er det også gjort sammenligninger med en nyere metode som baserer seg på Partial Least Squares.

IE-metoden og CGLIM-metoden ble sammenlignet med simuleringsanalyser i Kapittel 7. Flere ulike simuleringsmodeller ble benyttet for å generere datasett med ulike effekter for de forskjellige parametrene. Målet var at metodene skulle klare å gjenskape disse effektene ved analyse. Simuleringsresultatene viser at IE-metoden viser seg å være robust i alle de ulike simuleringsmodellene som er introdusert, og metoden klarer i stor grad å produsere estimater som er tilnærmet de sanne effektene. IE-metoden har minst varians og er en forventningsrett estimator, og er den metoden som gir minst MSE. For CGLIM-metoden benyttet jeg 3 ulike varianter der ulike betingelser ble innført for å kunne finne en unik løsning. I de tilfellene der betingelsen om at to effekter er lik hverandre stemmer overens med de sanne effektene i simuleringsmodellen, klarer CGLIM-metoden å estimere effektkoeffisienter som er ganske riktige. Vi ser videre at dess større forskjell det er på de effektene som betinges å være lik hverandre i CGLIM-metoden, dess større avvik fra de sanne effektene ser vi at estimatene får. Denne sammenhengen synes å være proporsjonal. Simuleringsanalysene viser også at CGLIM-metoden har større variasjon i sine estimater, og følgelig blir også MSE for denne metoden større enn for IE-metoden. IE-metoden gir ikke nødvendigvis alltid de mest korrekte estimatene for alle parametrene. Når CGLIM-metoden har betingelser som stemmer overens med virkeligheten, kan denne metoden også gi estimater som er like gode eller bedre enn IE-metoden. Hovedproblemet med CGLIM-metoden er at den avhenger av forhåndsinformasjon om dataene for å velge betingelser, og estimatene vil være sensitiv for valget av betingelse. Valget av betingelse må basere seg på tidligere teoretisk eller empirisk informasjon, og dette eksisterer sjelden eller kan neppe verifiseres. Ulike valg av betingelser kan produsere vidt forskjellige estimater for trendene til alder, periode og kohort. Valg av en vilkårlig betingelse i CGLIM-metoden kan i verste fall gi estimater som er langt unna sannheten. Når en ikke vet noe om de

dataene en analyserer, vil IE-metoden i mange tilfeller kunne gi sikrere estimater enn ved valg av en av de andre variantene av CGLIM-metoden. IE-metoden produserer estimerte koeffisienter for alder, periode og kohort og deres standardavvik på en direkte måte, uten å måtte velge blant et utvalg av mulige betingelser på koeffisientene som både kan, og kan ikke være hensiktsmessig for en bestemt analyse.

Videre i Kapittel 8 sammenlignes IE-metoden med noen av metodene som Clayton og Schifflers omtaler i sine artikler [3, 4]. Tilsvarende som i Kapittel 7, ser vi også her at IE-metoden er den metoden som er mest robust av de fulle APC-metodene, og den påvirkes i mye mindre grad av hvilken simuleringsmodell som velges for å generere datasettene enn de andre metodene. IE-metoden har minst varians og klarer i stor grad å produsere estimater som er tilnærmet de sanne effektene, og har dermed også lavest MSE i de fleste simuleringsmodellene. I metoden med førsteordensdifferanser blir trenden fra periodekurven fjernet ved at periodekurven tvinges å komme tilbake til samme nivå den startet på. En innfører betingelsen om at første og siste periode skal være lik hverandre. Tilsvarende benyttes også metoden med førsteordensdifferanser der trenden fra kohortkurven fjernes. Når de sanne effektene til første og siste periode og tilsvarende første og siste kohort er ganske lik hverandre, kan metoden med førsteordensdifferanser også gi estimater som er like gode eller bedre enn IE-metoden. Dess større forskjell det er på de effektene som betinges å være lik hverandre, dess større avvik fra de sanne verdiene ser vi at estimatene får. Metoden med førsteordensdifferanser som baserer seg på periodene påvirkes i større grad av dette, og avvikene må være større før metoden med førsteordensdifferanser basert på kohortene påvirkes tilsvarende. Simuleringsresultatene viser også at i tilfeller der det ikke er så store periode- og kohorteffekter, eller bare den ene effekten er til stede, så kan de reduserte modellene også være et godt alternativ til de fulle APC-modellene. Det er viktig med modellseleksjon der en vurderer ulike modellers tilpasning til et aktuelt datasett før en eventuelt innfører den fulle APC-modellen. Når en skal vurdere de ulike modellene mot hverandre, kan en ha god nytte av ulike goodness-of-fit-mål. Ved å benytte seg av en full APC-modell på et gitt datasett, når en redusert modell tilpasser dataene like godt eller bedre, utgjør det en modellmisspesifikasjon, og bør unngås. I Kapittel 5.5.2 i [25] ser forfatterne på et eksempel med feilaktig bruk av APC-modellen, for et gitt numerisk eksempel. Om effektene til en eller to av de tre faktorene er null, vil de fulle APC-modellene kunne overtilpasse dataene statistisk, og produsere unøyaktige estimater som er biased. På den annen side, når modellseleksjonstester indikerer at både alder-, periode- og kohortkategoriene er delaktig i å produsere et gitt datasett, vil bruken av IE-metoden kunne være nyttig for å gi riktige og stabile estimater.

I Kapittel 9 blir IE-metoden sammenlignet med metoden som baserer seg på Partial Least Squares. Denne metoden benytter seg av kovariansen mellom variablene og responsen som utgangspunkt for å lage komponenter. Slik er responsvektoren med å påvirke hvordan komponentene lages, og de første PLS-komponentene er de mest relevante for prediksjon. Komponentene som velges ut vil gi et gunstig utgangspunkt for regresjon, og slik kan kun de komponentene som ansees for å være viktigst i modellen inkluderes. PLS-regresjon er spesielt egnet når matrisen av prediktorer har flere variabler enn observasjoner. Simuleringsanalysene viser at det kan være nok å inkludere 3-5 komponenter i PLS-metoden, og at inkludering av flere komponenter vil gi liten forbedring i modellen. Sammenlignet med IE-metoden er variasjonen i estimatene ganske lik, og metodene gir tilnærmet like gode estimater for de ulike alder-, periode- og kohorteffektene. At PLS-metoden har vært foreslått for å løse APC-identifikasjonsproblemet, kom opp på slutten av arbeidet med simuleringene. Metoden er derfor kun benyttet for én simuleringsmodell og ved sammenligning av metoder for noen datasett. Det hadde vært interessant å undersøke nærmere om denne metoden viser seg å være robust i andre typer simuleringssituasjoner også. På grunn av omfanget av oppgaven er dette ikke inkludert her, og det ville kunne være en aktuell problemstilling å ta opp i et videre arbeid. Det hadde også vært interessant å jobbe videre med simuleringsanalyser der man inkluderte en eller flere kohortkarakteristikker for å undersøke om APCC-metoden klarer å gjenskape effektene til alder-, periode- og kohortkarakteristikkene. Dette ble vurdert underveis i prosessen med denne oppgaven, men på grunn av at dette er en problemstilling som var mer aktuell før nyere metodikk ble introdusert, ble andre problemstillinger prioritert.

Den samlede konklusjonen er at IE-metoden har passert simuleringstester for validering under ulike omstendigheter, og metoden kan være et nyttig verktøy for å finne kunnskap om de forskjellige effektene til alder, periode og kohort. Den virker robust mot plutselige endringer, og klarer i stor grad å gjenskape effektene for slike endringer. IE-metoden er statistisk effisient, det vil si at dens varians er mindre enn for metodene som benytter seg av en vilkårlig betingelse. Andre metoder som er avhengig av at det innføres en betingelse for å få unike estimater for effektene til alder, periode og kohort kan også gi gode estimater om betingelsen viser seg å stemme overens med virkeligheten. Metodene kan i enkelte tilfeller også gi estimater som er nærmere sannheten enn IE-metoden. Likevel har IE-metoden mindre variasjon i sine estimater og vil oftere gi lavest MSE. Om derimot betingelsen som innføres ikke samsvarer med virkeligheten, kan metodene i verste fall gi helt uriktige estimater. I en analysesituasjon der en ikke har noen forhåndskunnskap om dataene, kan IE-metoden virke som et sikrere metodevalg.
IE-metoden trenger ikke være en universell løsning for APC-identifikasjonsproblemet. Enhver statistisk modell vil bryte ned under visse omstendigheter, og IE-metoden er nok intet unntak her.

I litteratur om kohortanalyser og identifikasjonsproblemet til APC-analyser refereres det ofte til noen retningslinjer og forsiktighetsregler innført av Glenn [31]. Han mener at enhver ny metode som skal være nyttig må tilfredsstille ulike kriterier:

- *metoden må kunne gi tilnærmet korrekte estimater mer ofte enn den ikke gjør det*
- *troverdigheten i estimatene bør vurderes ved å bruke teori og sideinformasjon*
- *konklusjonene om effektene bør holdes tentative*

IE-metoden har vist seg å være en nyttig tilnærming for identifikasjon og estimering i APC-modellen og produserer forventningsrette og effisiente estimater. Metoden har ønskede matematiske og statistiske egenskaper og har bestått både case-studier og simuleringstester for modellvalidering.

Siden IE-metoden har blitt introdusert som en metode som skal løse problemene knyttet til modellidentifikasjon, er det interessant å se på om denne metoden er tatt i bruk i publikasjoner de senere årene. I et søk blant nyere publikasjoner finner jeg både artikler innen epidemiologi, demografi og samfunnsvitenskap som benytter seg av IE-metoden for å analysere sine datasett.
En artikkel [32] som benytter seg av IE-metoden ser på dødelighetsrater for brystkreft. En annen artikkel [33] benytter metoden til å undersøke påvirkningen av alder-, periode- og kohorteffekter på dødeligheten til kronisk obstruktiv lungesykdom (COPD) i Japan i årene 1950-2004. Andre eksempler er studier der en undersøker påvirkningen av alder-, periode- og kohorteffekter på sosial kapital [34] eller effekten fra ulik sysselsetting på arbeidstakernes helse [35]. Miech [36] tar i bruk IE-metoden for å se på trenden i marihuana bruk i USA fra 1985 til 2009. Winkler [37] ser på nedgangen i antall jegere og diskuterer fremtiden til jakting. Metoden er også benyttet til å undersøke endringer i mellommenneskelig tillit, og om hvordan terrorangrep påvirker denne dynamikken [38]. Dette er noen eksempler som viser at metoden er tatt i bruk innen mange ulike forskningsfelt, og med mange ulike typer data.
I tillegg ble det lansert en ny bok om APC-analyser denne våren, der Yang og Land [25] har samlet mye av arbeidet sitt basert på tidligere publikasjoner. Her introduserer de bl.a. leseren til IE-metoden. Dette kan tyde på at metoden er godt etablert og at forfatterne vurderer dette som så aktuelt at de tar det med i en bokutgivelse.

Innledningsvis siterte jeg et berømt utsagn fra George E. P. Box:

"All statistical models are wrong, but some are useful."

Kenneth C. Land sin versjon av dette uttrykket synes jeg kan være beskrivende innen alder-periode-kohort-problematikken:

"All statistical models are wrong, but some have better statistical properties than others – which may make them useful."

Vedlegg

A. R-koder/program

I denne delen følger eksempler på R-koder/program som er benyttet i oppgaven. Kun tekniske detaljer kommenteres i kodene og de bør derfor leses i sammenheng med de aktuelle kapitlene, som forklarer nærmere de spesifikke problemstillingene.

A.1 Simuleringsanalyser CGLIM- og IE-metoden

I Kapittel 7 utføres det simuleringsanalyser. Programmeringen av R-koder for Modell 1 er utført som beskrevet under. Kodene for én variant av CGLIM-metoden og én variant av IE-metoden er vist her. De andre metodene har identiske koder, men andre betingelser er innført og effektkodingen er annerledes.
De andre simuleringsmodellene (Modell 2 – 7) benytter samme program med R-koder, men ligningen for generering av datasettene må endres.

```
### ------- Simuleringsanalyser for Modell 1, CGLIM og IE -------
## --- Oppsett av simuleringsmodellen som datasettene genereres fra ---

# Angir antall grupper for alder og periode
n <- 9 #antall aldersgrupper
m <- 5 #antall periodegrupper

# Kohortmatrise lages
kohortmatrise <- matrix(c(9,10,11,12,13,8,9,10,11,12,7,8,9,10,11,6,7,8,9,10,5,6,7,8,9,4,5,6,7,8,3,4,5,
6,7,2,3,4,5,6,1,2,3,4,5),nrow=9,byrow=T)

# Angir dimensjon for lambdamatrisen
lambdamatrise <- matrix(nrow=n,ncol=m)

# Generering av lambdaverdier til lambdamatrisen
for (i in 1:n)
   for (j in 1:m){
         lambda <- exp(0.3+0.1*((i-5)^2)
         +0.1*(sin(j))+0.1*cos(kohortmatrise[i,j])+0.1*sin(10*(kohortmatrise[i,j])))
         lambdamatrise[i,j]=lambda
}
lambdavektor <- as.vector(t(lambdamatrise))
```

```
# Sanne alderseffekter
vektor_alder <- matrix(ncol=9)
for (i in 1:9){
		alder <- 0.1*((i-5)^2)
		vektor_alder[i]=alder
}
vektor_alder

# Sanne periodeeffekter
vektor_periode <- matrix(ncol=5)
for (j in 1:5){
		periode <- 0.1*sin(j)
		vektor_periode[j]=periode
}
vektor_periode

# Sanne kohorteffekter
vektor_kohort <- matrix(ncol=13)
for (c in 1:13){
		kohort <- 0.1*cos(c)+0.1*sin(10*(c))
		vektor_kohort[c]=kohort
}
vektor_kohort

# Sentrerte sanne effekter
cen_alder <- vektor_alder-mean(vektor_alder)
cen_periode <- vektor_periode-mean(vektor_periode)
cen_kohort <- vektor_kohort-mean(vektor_kohort)

sanne_effekter <- c(cen_alder,cen_periode,cen_kohort)

## Dummykoding til designmatrisene
# Intercept kodes med en vektor med 1-ere
konstant <- rep(1,45)

# Eksempel på koding av de ulike gruppene
# Aldersgruppe 2 kodes
age2 <- c(0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)

# Periode 2 kodes
per2 <- c(0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0)

# Kohort 2 kodes
coh2 <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0)
```

```
## --- CGLIM-metoden ---
# CGLIM_A har betingelse alder1 = alder2 = 0
# Referansegruppe er alder1, periode1 og kohort1

# Lager designmatrise for denne modellen
A1 <- as.matrix(cbind(age3,age4,age5,age6,age7,age8,age9))
P1 <- as.matrix(cbind(per2,per3,per4,per5))
C1 <- as.matrix(cbind(coh2,coh3,coh4,coh5,coh6,coh7,coh8,coh9,coh10,coh11,coh12,coh13))
CGLIM_A <- as.matrix(cbind(konstant,A1,P1,C1))

## --- Simuleringsanalyser ---
# CGLIM_A
# Velger et seed
set.seed(4)

# Angir dimensjonen på matriser som resultatene skal komme i
p<-27
q<-10000
est_matrise_CGLIM_A <- matrix(nrow=p,ncol=q)
devianser_CGLIM_A <- matrix(ncol=q)
AIC_CGLIM_A <- matrix(ncol=q)
pred_CGLIM_A <- matrix(nrow=45,ncol=q)
y_matrise <- matrix(nrow=n,ncol=m)

# For-lokke som gjentar prosedyren 10000 ganger
for (k in 1:10000){
   n <- 9
   m <- 5
   y_matrise <- matrix(nrow=n,ncol=m)

   for (i in 1:n)
      for (j in 1:m){
            #y-ene er poissonfordelt med lambda hentet fra lambdamatrisen
            y=rpois(1,lambdamatrise[i,j])
            y_matrise[i,j]=y
   }
   Y <- y_matrise

   # lager en vektor av y-responsene
   Yrespons <- as.vector(t(Y))

   # Linear regresjon med log som linkfunksjon
   lm_CGLIM_A <- glm(formula=Yrespons~(CGLIM_A-1),family=poisson(link="log"))
   predikert_CGLIM_A <- fitted.values(lm_CGLIM_A)

   # Estimater
   alder <- as.vector(summary(lm_CGLIM_A)$coef[2:8,1])
   mean_a <- mean(sum(alder)/9)
   cen_alder <- alder-mean_a
```

```
    periode <- as.vector(summary(lm_CGLIM_A)$coef[9:12,1])
    mean_p <- mean(sum(periode)/5)
    cen_periode <- periode-mean_p

    cohort <- as.vector(summary(lm_CGLIM_A)$coef[13:24,1])
    mean_c <- mean(sum(cohort)/13)
    cen_cohort <- cohort-mean_c

    age <- c(-mean_a,-mean_a,cen_alder)
    period <- c(-mean_p,cen_periode)
    cohort <- c(-mean_c,cen_cohort)
    estimater_CGLIM_A <- c(age,period,cohort)

    # Hver simulering gir en kolonne med estimater i disse matrisene
    est_matrise_CGLIM_A[,k] <- estimater_CGLIM_A
    devianser_CGLIM_A[,k] <- deviance(lm_CGLIM_A)
    AIC_CGLIM_A[,k] <- AIC(lm_CGLIM_A)
    pred_CGLIM_A[,k] <- predikert_CGLIM_A
}

# Resultater for de 10000 simuleringene
mean_estimater_CGLIM_A <- round(rowMeans(est_matrise_CGLIM_A),3)
mean_devianser_CGLIM_A <- mean(devianser_CGLIM_A)
mean_AIC_CGLIM_A <- mean(AIC_CGLIM_A)

# Standardavvik
t_est_matrise_CGLIM_A <- t(est_matrise_CGLIM_A)
sd_CGLIM_A <- round(apply(t_est_matrise_CGLIM_A,2,sd),3)

# Beregner MSE
MSE_kvadr_CGLIM_A <- (est_matrise_CGLIM_A - sanne_effekter)^2
MSE_CGLIM_A <- rowMeans(MSE_kvadr_CGLIM_A)
MSE_alder_CGLIM_A <- sum(MSE_CGLIM_A[1:9])
MSE_periode_CGLIM_A <- sum(MSE_CGLIM_A[10:14])
MSE_kohort_CGLIM_A <- sum(MSE_CGLIM_A[15:27])
total_MSE_CGLIM_A <- sum(MSE_CGLIM_A)

# Beregner sum avvik
kvadr_avvik_CGLIM_A <- (pred_CGLIM_A - lambdavektor)^2
avvik_CGLIM_A <- rowMeans(kvadr_avvik_CGLIM_A)
sum_avvik_CGLIM_A <- sum(avvik_CGLIM_A)
```

```
## --- IE-metoden (Modell1) ---
A <- as.matrix(cbind(age1,age2,age3,age4,age5,age6,age7,age8))
P <- as.matrix(cbind(per1,per2,per3,per4))
C <- as.matrix(cbind(coh1,coh2,coh3,coh4,coh5,coh6,coh7,coh8,coh9,coh10,coh11,coh12))

# Effektkoding
rA <- age9 * (-1)
rP <- per5 * (-1)
rC <- coh13 * (-1)
cA <- A + rA
cP <- P + rP
cC <- C + rC

# Lager designmatrisen
X <- as.matrix(cbind(konstant,cA,cP,cC))
XtX <- t(X)%*% X

# Singulaer verdi dekomposisjon for XtX
svd(XtX)
eigen(XtX)$values

# Benytter alle egenvektorer som ikke har 0 i egenverdi
V <- as.matrix(eigen(XtX)$vectors[,1:24])

# Designmatrisen transformeres, prinsipal komponentene dannes
IE1 <- X %*% V #den nye designmatrisen

## --- Simuleringsanalyser ---
# Velger det samme seed som for CGLIM-metoden
set.seed(4)

# Angir dimensjonen på matriser som resultatene skal komme i
p <- 27
q <- 10000
est_matrise_IE_1 <- matrix(nrow=p,ncol=q)
devianser_IE_1 <- matrix(ncol=q)
AIC_IE_1 <- matrix(ncol=q)
pred_IE_1 <- matrix(nrow=45,ncol=q)
y_matrise <- matrix(nrow=n,ncol=m)

# For-lokke som gjentar prosedyren 10000 ganger
for (k in 1:10000){
   n <- 9
   m <- 5
   y_matrise <- matrix(nrow=n,ncol=m)

   for (i in 1:n)
      for (j in 1:m){
            y=rpois(1,lambdamatrise[i,j])
            y_matrise[i,j]=y
   }
   Y <- y_matrise
```

```
    # Lager en vektor av y-responsene
    Yrespons <- as.vector(t(Y))

    # Linear regresjon med log som linkfunksjon
    lm_IE_1 <- glm(formula=Yrespons~(IE1-1),family=poisson(link="log"))
    predikert_IE_1 <- fitted.values(lm_IE_1)

    # Estimater
    IE_1 <- as.vector(summary(lm_IE_1)$coef[1:24,1])
    IE_1_estimater <- V %*% IE_1

    a <- IE_1_estimater[2:9]
    alder <- sum(IE_1_estimater[2:9])
    p <- IE_1_estimater[10:13]
    periode <- sum(IE_1_estimater[10:13])
    c <- IE_1_estimater[14:25]
    kohort <- sum(IE_1_estimater[14:25])

    # Estimatene skal summeres til 0, og siste gruppe faar minus summen av de andre
    age <- c(a,-alder)
    period <- c(p,-periode)
    cohort <- c(c,-kohort)
    estimater_IE_1 <- c(age,period,cohort)

    # Hver simulering gir en kolonne med estimater i disse matrisene
    est_matrise_IE_1[,k] <- estimater_IE_1
    devianser_IE_1[,k] <- deviance(lm_IE_1)
    AIC_IE_1[,k] <- AIC(lm_IE_1)
    pred_IE_1[,k] <- predikert_IE_1
}

# Resultater for de 10000 simuleringene
mean_estimater_IE_1 <- round(rowMeans(est_matrise_IE_1),3)
mean_devianser_IE_1 <- mean(devianser_IE_1)
mean_AIC_IE_1 <- mean(AIC_IE_1)

# Standardavvik
t_est_matrise_IE_1 <- t(est_matrise_IE_1)
sd_IE_1 <- round(apply(t_est_matrise_IE_1,2,sd),3)

# Beregner MSE
MSE_kvadr_IE_1 <- (est_matrise_IE_1 - sanne_effekter)^2
MSE_IE_1 <- rowMeans(MSE_kvadr_IE_1)
MSE_alder_IE_1 <- sum(MSE_IE_1[1:9])
MSE_periode_IE_1 <- sum(MSE_IE_1[10:14])
MSE_kohort_IE_1 <- sum(MSE_IE_1[15:27])
total_MSE_IE_1 <- sum(MSE_IE_1)

# Beregner sum avvik
kvadr_avvik_IE_1 <- (pred_IE_1 - lambdavektor)^2
avvik_IE_1 <- rowMeans(kvadr_avvik_IE_1)
sum_avvik_IE_1 <- sum(avvik_IE_1)
```

A.2 Analyse av datasett med ulike metoder (CGLIM, IE og PLS)

I Kapittel 9.7 analyseres et datasett fra Yang [14] med CGLIM-metoden, IE-metoden og metoden som baserer seg på PLS. Programmeringen av R-koder er utført som beskrevet under.

```
### ------- Sammenligning av ulike metoder – CGLIM, IE og PLS -------
## Datasettene leses inn manuelt fra Tabell 1 og 2 i artikkelen siden de ikke er tilgjengelig på fil.
# Tabell 1 angir antall, og Tabell 2 angir populasjon.

rate <- antall/pop
lnrates <- log(rate)

## Dummykoding til designmatrisene
# Intercept kodes med en vektor med 1-ere
konstant <- rep(1,152)

# Eksempel på koding av de ulike gruppene
# Aldersgruppe 2 kodes
a.2 <- c(0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)

# Periode 2 kodes
p.2 <- c(0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,
0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,
0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0)

# Kohort 2 kodes
c.2 <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0)

## --- GCLIM-metoden ---
# setter første kategori som referanse
A1 <- as.matrix(cbind(a.2,a.3,a.4,a.5,a.6,a.7,a.8,a.9,a.10,a.11,a.12,a.13,a.14,a.15,a.16,a.17,a.18,a.19))
P1 <- as.matrix(cbind(p.2,p.3,p.4,p.5,p.6,p.7,p.8))
C1 <- as.matrix(cbind(c.2,c.3,c.4,c.5,c.6,c.7,c.8,c.9,c.10,c.11,c.12,c.13,c.14,c.15,c.16,c.17,c.18,c.19,c.20,
c.21,c.22,c.23,c.24,c.25,c.26))

# CGLIM_A
# Betingelse: alder 2 = alder 3
# Aldersgruppe 2 og 3 får ny koding
a.23 <- c(0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
A23 <- as.matrix(cbind(a.23,a.4,a.5,a.6,a.7,a.8,a.9,a.10,a.11,a.12,a.13,a.14,a.15,a.16,a.17,a.18,a.19))
```

```
# Designmatrisen for denne modellen
A23PC <- as.matrix(cbind(konstant,A23,P1,C1))

# Lineær regresjon (GLM)
CGLIM_A <- glm(formula=antall~(A23PC-1)+offset(log(pop)),family=poisson(link="log"))
A.estimat <- as.vector(summary(CGLIM_A)$coef[1:50,1])

A.alder <- as.vector(summary(CGLIM_A)$coef[2:18,1])
A.periode <- as.vector(summary(CGLIM_A)$coef[19:25,1])
A.cohort <- as.vector(summary(CGLIM_A)$coef[26:50,1])

# Skriver ut estimatene for alder, periode og kohort (inkludert referansegruppene)
A.estimater <- c(0,A.alder[1],A.alder,0,A.periode,0,A.cohort)
round(A.estimater,3)

# CGLIM_P
# Betingelse: periode 1 = periode 2
# Periode 1 er satt til referanse, og tilsvarende for periode 2 i denne modellen
P12 <- as.matrix(cbind(p.3,p.4,p.5,p.6,p.7,p.8))

# Designmatrisen for denne modellen
AP12C <- as.matrix(cbind(konstant,A1,P12,C1))

# Lineær regresjon (GLM)
CGLIM_P <- glm(formula=antall~(AP12C-1)+offset(log(pop)),family=poisson(link="log"))
P.estimat <- as.vector(summary(CGLIM_P)$coef[1:50,1])

P.alder <- as.vector(summary(CGLIM_P)$coef[2:19,1])
P.periode <- as.vector(summary(CGLIM_P)$coef[20:25,1])
P.cohort <- as.vector(summary(CGLIM_P)$coef[26:50,1])

# Skriver ut estimatene for alder, periode og kohort (inkludert referansegruppene)
P.estimater <- c(0,P.alder,0,0,P.periode,0,P.cohort)
round(P.estimater,3)

# CGLIM_C
# Betingelse: kohort 25 = kohort 26
# Kohortgruppe 25 og 26 får ny koding
c.2526 <- c(0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
C2526 <- as.matrix(cbind(c.2,c.3,c.4,c.5,c.6,c.7,c.8,c.9,c.10,c.11,c.12,c.13,c.14,c.15,c.16,c.17,c.18,c.19,
c.20,c.21,c.22,c.23,c.24,c.2526))

# Designmatrisen for denne modellen
APC2526 <- as.matrix(cbind(konstant,A1,P1,C2526))

# Lineær regresjon (GLM)
CGLIM_C <- glm(formula=antall~(APC2526-1)+offset(log(pop)),family=poisson(link="log"))
C.estimat <- as.vector(summary(CGLIM_C)$coef[1:50,1])
```

```
C.alder <- as.vector(summary(CGLIM_C)$coef[2:19,1])
C.periode <- as.vector(summary(CGLIM_C)$coef[20:26,1])
C.cohort <- as.vector(summary(CGLIM_C)$coef[27:50,1])

# Skriver ut estimatene for alder, periode og kohort (inkludert referansegruppene)
C.estimater <- c(0,C.alder,0,C.periode,0,C.cohort,C.cohort[24])
round(C.estimater,3)

## --- IE-metoden ---
# siste gruppe kodes med -1 for hver kategori (effektkoding)
A.IE <- as.matrix(cbind(a.1,a.2,a.3,a.4,a.5,a.6,a.7,a.8,a.9,a.10,a.11,a.12,a.13,a.14,a.15,a.16,a.17,a.18))
P.IE <- as.matrix(cbind(p.1,p.2,p.3,p.4,p.5,p.6,p.7))
C.IE <- as.matrix(cbind(c.1,c.2,c.3,c.4,c.5,c.6,c.7,c.8,c.9,c.10,c.11,c.12,c.13,c.14,c.15,c.16,c.17,c.18,c.19,
c.20,c.21,c.22,c.23,c.24,c.25))

rA <- a.19 * (-1)
rP <- p.8 * (-1)
rC <- c.26 * (-1)

cA <- A.IE + rA
cP <- P.IE + rP
cC <- C.IE + rC

# Designmatrisen for denne modellen
X <- as.matrix(cbind(konstant,cA,cP,cC))
XtX <- t(X)%*% X

# Singulær verdi dekomposisjon for XtX
svd(XtX)
eigen(XtX)$values

# Benytter alle egenvektorene som ikke har 0 i egenverdi
V <- as.matrix(eigen(XtX)$vectors[,1:50])

# Designmatrisen transformeres, prinsipal komponentene dannes
X1 <- X %*% V

# Lineær regresjon (GLM)
IE <- glm(formula=antall~(X1-1)+offset(log(pop)),family=poisson(link="log"))
koeffisienter <- as.vector(summary(IE)$coef[1:50,1])

# Danner regresjonskoeffisientene for originale variabler
IE.estimat <- V %*% koeffisienter

IEa.19 <- sum(IE.estimat[2:19])
IEp.8 <- sum(IE.estimat[20:26])
IEc.26 <- sum(IE.estimat[27:51])

# Skriver ut estimatene for alder, periode og kohort (inkludert referansegruppene)
IE.estimater <- c(IE.estimat[2:19],-IEa.19,IE.estimat[20:26],-IEp.8,IE.estimat[27:51],-IEc.26)
round(IE.estimater,3)
```

```
## --- Partial least squares (PLS) ---
# Matriser med alle gruppene for hhv. alder, periode og kohort
A <- as.matrix(cbind(a.1,a.2,a.3,a.4,a.5,a.6,a.7,a.8,a.9,a.10,a.11,a.12,a.13,a.14,a.15,a.16,a.17,a.18,
a.19))
P <- as.matrix(cbind(p.1,p.2,p.3,p.4,p.5,p.6,p.7,p.8))
C <- as.matrix(cbind(c.1,c.2,c.3,c.4,c.5,c.6,c.7,c.8,c.9,c.10,c.11,c.12,c.13,c.14,c.15,c.16,c.17,c.18,c.19,
c.20,c.21,c.22,c.23,c.24,c.25,c.26))


# Designmatrisen for denne modellen
PLS <- as.matrix(cbind(A,P,C))

library(pls)

# ekstraherer 3 komponenter
plsr3 <- plsr(lnrates ~ PLS, ncomp=3, scale=F, method="oscorespls",validation="LOO", jackknife=T)
round(coef(plsr3),3)

# ekstraherer 45 komponenter
plsr45 <- plsr(lnrates ~ PLS, ncomp=45, scale=F, method="oscorespls",validation="LOO", jackknife=T)
round(coef(plsr45),3)
```

B. Resultater og tabeller som er utelatt fra oppgaven

B.1 Resultater simuleringsmodell 7

I denne delen er resultatene fra simuleringsmodell 7 vist i tabellform. De ulike log-koeffisientene er vist som figur i Kapittel 7, og de samlede målene for MSE er også oppgitt der.

Simuleringsresultater fra IE- og CGLIM-estimatorer, alderseffekter.

Simuleringsresultat IE- og CGLIM-estimatorer (n=10 000), alderseffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	Intrinsic estimator IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
Alder 1	Mean	0,933	0,931	1,235	-1,883	-2,124	-1,413
	SD		0,136	0,133	0,638	0,912	1,651
	MSE		0,019	0,109	8,341	10,178	8,230
Alder 2	Mean	0,233	0,227	0,456	-1,883	-2,064	-1,531
	SD		0,153	0,158	0,638	0,683	1,242
	MSE		0,024	0,074	4,888	5,744	4,655
Alder 3	Mean	-0,267	-0,280	-0,128	-1,687	-1,807	-1,452
	SD		0,186	0,188	0,405	0,464	0,832
	MSE		0,035	0,055	2,181	2,588	2,096
Alder 4	Mean	-0,567	-0,583	-0,506	-1,286	-1,346	-1,168
	SD		0,191	0,190	0,257	0,308	0,459
	MSE		0,037	0,040	0,584	0,703	0,573
Alder 5	Mean	-0,667	-0,674	-0,674	-0,674	-0,674	-0,674
	SD		0,184	0,184	0,184	0,184	0,184
	MSE		0,034	0,034	0,034	0,034	0,034
Alder 6	Mean	-0,567	-0,570	-0,646	0,133	0,193	0,016
	SD		0,159	0,161	0,250	0,260	0,433
	MSE		0,025	0,032	0,552	0,645	0,527
Alder 7	Mean	-0,267	-0,259	-0,411	1,148	1,268	0,913
	SD		0,135	0,136	0,385	0,467	0,831
	MSE		0,018	0,039	2,149	2,574	2,082
Alder 8	Mean	0,233	0,249	0,020	2,359	2,540	2,007
	SD		0,130	0,125	0,547	0,702	1,254
	MSE		0,017	0,061	4,818	5,811	4,718
Alder 9	Mean	0,933	0,960	0,655	3,773	4,014	3,303
	SD		0,131	0,121	0,700	0,916	1,644
	MSE		0,018	0,092	8,556	10,331	8,318
Total MSE			0,226	0,536	32,102	38,607	31,233

Simuleringsresultater fra IE- og CGLIM-estimatorer, periodeeffekter.

Simuleringsresultat IE- og CGLIM-estimatorer (n=10 000), periodeeffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	Intrinsic estimator IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
Periode 1	Mean	-0,707	-0,714	-0,867	0,692	0,813	0,457
	SD		0,147	0,158	0,388	0,325	0,775
	MSE		0,022	0,050	2,110	2,416	1,957
Periode 2	Mean	0,043	0,049	-0,027	0,753	0,813	0,635
	SD		0,120	0,116	0,205	0,325	0,465
	MSE		0,014	0,018	0,546	0,699	0,568
Periode 3	Mean	0,182	0,187	0,187	0,187	0,187	0,187
	SD		0,112	0,112	0,112	0,112	0,112
	MSE		0,013	0,013	0,013	0,013	0,013
Periode 4	Mean	0,230	0,230	0,306	-0,474	-0,534	-0,356
	SD		0,093	0,095	0,220	0,230	0,417
	MSE		0,009	0,015	0,544	0,637	0,518
Periode 5	Mean	0,253	0,249	0,401	-1,158	-1,278	-0,923
	SD		0,099	0,094	0,326	0,456	0,824
	MSE		0,010	0,031	2,096	2,552	2,062
Total MSE			0,067	0,127	5,309	6,317	5,117

Simuleringsresultater fra IE- og CGLIM-estimatorer, kohorteffekter.

Simuleringsresultat IE- og CGLIM-estimatorer (n=10 000), kohorteffekter							
		Sann verdi	Intrinsic estimator IE Modell 1	Intrinsic estimator IE Modell 2	CGLIM (A1=A2)	CGLIM (P1=P2)	CGLIM (C1=C2)
Kohort 1	Mean	0,275	0,254	0,711	-3,966	-4,327	-3,261
	SD		0,271	0,324	1,085	1,251	2,217
	MSE		0,074	0,295	19,168	22,748	17,421
Kohort 2	Mean	-0,307	-0,332	0,049	-3,849	-4,150	-3,261
	SD		0,229	0,213	0,892	1,200	2,217
	MSE		0,053	0,172	13,340	16,206	13,641
Kohort 3	Mean	-0,957	-0,993	-0,688	-3,807	-4,048	-3,337
	SD		0,249	0,236	0,727	0,962	1,681
	MSE		0,063	0,128	8,650	10,476	8,487
Kohort 4	Mean	-0,867	-0,887	-0,658	-2,997	-3,178	-2,645
	SD		0,220	0,211	0,546	0,727	1,262
	MSE		0,049	0,088	4,835	5,867	4,753
Kohort 5	Mean	0,560	0,565	0,717	-0,842	-0,963	-0,607
	SD		0,124	0,122	0,374	0,472	0,837
	MSE		0,015	0,040	2,106	2,540	2,062
Kohort 6	Mean	0,661	0,672	0,748	-0,031	-0,092	0,086
	SD		0,158	0,158	0,243	0,275	0,450
	MSE		0,025	0,033	0,539	0,642	0,533
Kohort 7	Mean	0,930	0,948	0,948	0,948	0,948	0,948
	SD		0,163	0,163	0,163	0,163	0,163
	MSE		0,027	0,027	0,027	0,027	0,027
Kohort 8	Mean	-0,192	-0,194	-0,270	0,510	0,570	0,392
	SD		0,214	0,215	0,291	0,315	0,464
	MSE		0,046	0,052	0,577	0,680	0,556
Kohort 9	Mean	-1,059	-1,075	-1,228	0,332	0,452	0,097
	SD		0,255	0,257	0,448	0,528	0,856
	MSE		0,065	0,095	2,135	2,561	2,068
Kohort 10	Mean	-0,594	-0,591	-0,820	1,519	1,700	1,167
	SD		0,184	0,192	0,570	0,662	1,224
	MSE		0,034	0,088	4,789	5,698	4,596
Kohort 11	Mean	-0,309	-0,294	-0,599	2,520	2,760	2,050
	SD		0,161	0,170	0,727	0,906	1,644
	MSE		0,026	0,113	8,529	10,239	8,263
Kohort 12	Mean	1,075	1,109	0,728	4,626	4,927	4,039
	SD		0,120	0,121	0,882	1,131	2,053
	MSE		0,016	0,135	13,391	16,119	13,000
Kohort 13	Mean	0,784	0,818	0,361	5,039	5,400	4,334
	SD		0,194	0,165	0,964	1,372	2,469
	MSE		0,039	0,206	19,036	23,190	18,698
Total MSE			0,532	1,470	97,121	116,994	94,105

B.2 Tabeller

I dette avsnittet er tabeller som er utelatt fra Kapittel 9 gjengitt.

Tabell 1: Mortality Rates of Hepatocellular Carcinoma in Taiwanese Men ≥ 40 Years of Age Between Years 1976 and 2008. (Number of Deaths Per 10 000).

		Periode						
		1976-1980	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2008
Aldersgruppe	40-44	3,14	3,31	3,32	3,12	3,04	2,91	2,36
	45-49	4,65	5,51	5,29	5,36	4,98	4,77	4,35
	50-54	6,81	7,25	7,46	8,17	8,17	7,42	7,26
	55-59	8,41	9,73	9,90	11,56	12,75	11,77	10,57
	60-64	10,36	12,04	11,57	14,37	16,41	16,82	15,30
	65-69	12,62	14,01	13,88	16,03	18,66	20,43	21,27
	70-74	13,54	14,78	17,04	19,56	20,14	22,30	25,81
	75-79	14,54	17,82	18,67	22,69	24,39	23,46	25,57
	80-84	12,36	16,03	17,51	22,92	26,75	26,46	28,33
	≥ 85	13,40	23,26	21,23	22,76	24,81	29,25	27,06

Tabell 2: Deaths per 100 000: U.S. Females, 1960 - 99.

		Periode							
		1960-1964	1965-1969	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999
Aldersgruppe	0-4	518	444	371	309	259	227	192	160
	5-9	39	36	34	27	23	20	18	16
	10-14	31	30	29	25	22	20	19	18
	15-19	53	58	60	54	48	48	46	43
	20-24	70	73	72	64	57	54	51	48
	25-29	88	86	82	70	64	64	64	59
	30-34	123	123	113	90	80	83	86	82
	35-39	183	185	172	135	115	113	119	119
	40-44	275	282	267	221	186	170	166	170
	45-49	419	422	405	344	304	277	261	252
	50-54	631	625	591	524	483	454	419	395
	55-59	906	905	868	770	741	711	671	631
	60-64	1414	1337	1282	1156	1126	1111	1053	1014
	65-69	2228	2120	1967	1740	1686	1656	1590	1543
	70-74	3608	3400	3112	2698	2606	2563	2445	2428
	75-79	5886	5593	5143	4380	4087	3996	3796	3778
	80-84	9678	9162	8395	7337	6946	6641	6302	6351
	85-89	15665	14993	13757	12016	11487	11223	10444	10934
	90-94	23635	22839	21261	18991	18253	18361	17395	17666

Litteratur

1. O'Brien, R.M., *The age-period-cohort conundrum as two fundamental problems.* Quality & Quantity, 2011. **45**(6): p. 1429-1444.

2. Yang, Y., et al., *The intrinsic estimator for age-period-cohort analysis: What it is and how to use it.* American Journal of Sociology, 2008. **113**(6): p. 1697-1736.

3. Clayton, D. and E. Schifflers, *Models for temporal variation in cancer rates. I: Age-period and age-cohort models.* Statistics in Medicine, 1987. **6**(4): p. 449-467.

4. Clayton, D. and E. Schifflers, *Models for temporal variation in cancer rates. II: Age-period-cohort models.* Statistics in Medicine, 1987. **6**(4): p. 469-481.

5. Tu, Y.K., N. Kramer, and W.C. Lee, *Addressing the identification problem in age-period-cohort analysis: A tutorial on the use of partial least squares and principal components analysis.* Epidemiology, 2012. **23**(4): p. 583-93.

6. Kunnskapsforlaget, *Store Medisinske Leksikon*. 2007: Kunnskapsforlaget, Oslo.

7. Rothman, K.J., *Epidemiology An Introduction*. 2002: Oxford University Press Inc., New York.

8. Osmond, C. and M.J. Gardner, *Age, period and cohort models applied to cancer mortality rates.* Statistics in Medicine, 1982. **1**(3): p. 245-59.

9. Mason, K.O., et al., *Some methodological issues in cohort analysis of archival data.* American Sociological Review, 1973. **38**(2): p. 242-258.

10. Holford, T.R., *The estimation of age, period and cohort effects for vital rates.* Biometrics, 1983. **39**(2): p. 311-324.

11. Kupper, L.L., et al., *Statistical age-period-cohort analysis: A review and critique.* Journal of Chronic Diseases, 1985. **38**(10): p. 811-30.

12. Robertson, C., S. Gandini, and P. Boyle, *Age-period-cohort models: A comparative study of available methodologies.* Journal of Clinical Epidemiology, 1999. **52**(6): p. 569-583.

13. Fu, W.J.J., *Ridge estimator in singular design with application to age-period-cohort analysis of disease rates.* Communications in Statistics-Theory and Methods, 2000. **29**(2): p. 263-278.

14. Yang, Y., W.J.J. Fu, and K.C. Land, *A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models.* Sociological Methodology, 2004, Vol 34, 2004. **34**: p. 75-110.

15. Statistisk Sentralbyrå. *Statistisk årbok 2011*. Available from: http://www.ssb.no/befolkning/artikler-og-publikasjoner/statistisk-aarbok-2011.

16. Johnson, R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*. 6th ed. 2007: Pearson Education, Inc., Upper Saddle River, New Jersey.

17. Dobson, A.J. and A.G. Barnett, *An Introduction to Generalized Linear Models*. 3rd ed. 2008: Chapman & Hall/CRC, Boca Raton, Florida.

18. Lay, D.C., *Linear Algebra and Its Applications*. 3rd ed. 2006: Pearson Education, Inc., Upper Saddle River, New Jersey.

19. Searle, S.R., *Linear Models*. 1971: Wiley, New York.

20. Fu, W.J., *A smoothing cohort model in age-period-cohort analysis with applications to homicide arrest rates and lung cancer mortality rates.* Sociological Methods & Research, 2008. **36**(3): p. 327-361.

21. Holford, T.R., *An alternative approach to statistical age-period-cohort analysis.* Journal of Chronic Diseases, 1985. **38**(10): p. 831-836.

22. O'Brien, R.M., J. Stockard, and L. Isaacson, *The enduring effects of cohort characteristics on age-specific homicide rates, 1960-1995.* American Journal of Sociology, 1999. **104**(4): p. 1061-1095.

23. Casella, G. and R.L. Berger, *Statistical Inference*. 2nd ed. 2002: Duxbury, Pacific Grove, California.

24. Wagenmakers, E.J., *Model selection and multimodel inference: A practical information-theoretic approach.* Journal of Mathematical Psychology, 2003. **47**(5-6): p. 580-586.

25. Yang, Y. and K.C. Land, *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. 2013: Chapman & Hall/CRC, Boca Raton, Florida.

26. Tu, Y.K., G.D. Smith, and M.S. Gilthorpe, *A new approach to age-period-cohort analysis using partial least squares regression: The trend in blood pressure in the Glasgow Alumni Cohort.* Plos One, 2011. **6**(4).

27. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. 2009: Springer, New York.

28. Martens, H. and T. Næs, *Multivariate Calibration*. 1989: Wiley, New York.

29. O'Brien, R.M., *Age period cohort characteristic models.* Social Science Research, 2000. **29**(1): p. 123-139.

30. Fu, W.J.J., K.C. Land, and Y. Yang, *On the intrinsic estimator and constrained estimators in age-period-cohort models.* Sociological Methods & Research, 2011. **40**(3): p. 453-466.

31. Glenn, N.D., *Cohort Analysis*. 2nd ed. 2005: Sage, Thousand Oaks, California.

32. Arnesi, N. and L. Hachuel, *Application of the intrinsic estimator to breast cancer mortality rates.* Revista Panamericana De Salud Publica-Pan American Journal of Public Health, 2011. **30**(3): p. 225-230.

33. Pham, T.M., et al., *Age-period-cohort analysis of chronic obstructive pulmonary disease mortality in Japan, 1950-2004.* Journal of Epidemiology, 2012. **22**(4): p. 302-307.

34. Schwadel, P. and M. Stout, *Age, period and cohort effects on social capital.* Social Forces, 2012. **91**(1): p. 233-252.

35. Nishikitani, M., et al., *Effect of unequal employment status on workers' health: Results from a Japanese national survey.* Social Science & Medicine, 2012. **75**(3): p. 439-451.

36. Miech, R. and S. Koester, *Trends in U.S., past-year marijuana use from 1985 to 2009: An age-period-cohort analysis.* Drug and Alcohol Dependence, 2012. **124**(3): p. 259-267.

37. Winkler, R. and K. Warnke, *The future of hunting: An age-period-cohort analysis of deer hunter decline.* Population and Environment, 2013. **34**(4): p. 460-480.

38. Clark, A.K. and M.A. Eisenstein, *Interpersonal trust: An age–period–cohort analysis revisited.* Social Science Research, 2013. **42**(2): p. 361-375.