En empirisk studie av kumulative insidensfunksjoner estimert ved Cox regresjon og Fine-Gray metoden

AV

Nawroz Khalaf

MASTEROPPGAVE
for graden

Master i Modellering og dataanalyse

Det matematisk- naturvitenskapelige fakultet
Universitetet i Oslo

Mai 2013

Til mine foreldre Shahnaz & Kader

Forord

Jeg vil begynne med å takke min fantastiske veileder Ørnulf Borgan for den gode oppfølgningen jeg har fått gjennom hele skriveprosessen. Tusen takk for tålmodige forklaringer og konstruktive tilbakemeldinger. Jeg er takknemlig for all støtten jeg har fått.

Jeg vil også takke mine hyggelige kolleger i avdelingen Aktuar Liv i SpareBank1 Forsikring. Jeg har lært mye av dere gjennom deltidsjobben.

Takk til venner og familie. Takk til mine søsken Narin, Sliva, Lelan og Zana for deres oppmuntringer. Størst takk til min mine foreldre som alltid har motivert meg, og hatt troen på meg. Uten deres støtte hadde jeg ikke klart dette.

Oslo, mai 2013
Nawroz Khalaf

Innhold

1 Innledning **1**
1.1 Beskrivelse av datasettet . . . 2

2 Modellering uten kovariater **5**
2.1 Levetidsanalyse . . . 5
2.1.1 Begreper . . . 5
2.1.2 Data . . . 6
2.1.3 Estimater . . . 7
2.2 Konkurrerende dødsårsaker . . . 8
2.2.1 Begreper . . . 9
2.2.2 Data . . . 10
2.2.3 Estimater . . . 10
2.3 Markov kjeder . . . 12
2.3.1 Begreper . . . 12
2.3.2 Data . . . 13
2.3.3 Estimater . . . 14
2.4 Programvare . . . 14

3 Modellering med kovariater **17**
3.1 Cox regresjon . . . 17
3.1.1 Estimater . . . 19
3.2 Cox-regresjon for konkurrerende dødsårsaker . . . 22
3.2.1 Estimater . . . 22
3.3 Programvare . . . 29

4 Modellering med subfordelinger **35**
4.1 Subfordelingshasardrate uten kovariater . . . 35
4.1.1 Begreper . . . 35
4.1.2 Høyre-sensurering . . . 36
4.1.3 Venstre-trunkering . . . 37
4.1.4 Data . . . 38
4.1.5 Estimering med høyre-sensurerte og venstre-trunkerte data . . . 38
4.2 Proporsjonal subfordelingshasard modell . . . 39
4.3 Programvare . . . 44

5 Sammenligning **49**

5.1 Generering av levetider 50
5.2 Simuleringsmodeller 51
5.2.1 Estimering av parametrene i Weibull fordelingen 51
5.2.2 Cox-modell med parametrisk baseline 53
5.2.3 Additiv modell med parametrisk baseline 53
5.3 Resultater og diskusjon 54
5.3.1 Cox-modell 54
5.3.2 Additiv modell 61

6 Oppsummering og konklusjon **69**

A Tillegg til Kapittel 4 **73**
A.1 Tabeller 73
A.2 Plott av kumulative insidensfunksjoner 77
A.3 Standardfeil 79

B Tillegg til Kapittel 5 **83**
B.1 Tabeller 83
B.2 Plott av kumulative insidensrater 86
B.3 Estimater additiv modell 90

Kapittel 1

Innledning

Levetidsanalyse brukes innenfor mange områder; medisin, økonomi, ingeniørfag og forsikring. For sistnevnte er det viktig å analysere hvordan dødssannsynligheter endres over tid, og hvilke faktorer som har innvirkning. Dette påvirker beregning av premier og avsetninger til fremtidige utbetalinger. Vi skal i denne oppgaven bruke begrepet *levetid* om tid til død. I andre sammenhenger brukes begrepet også om tid til andre hendelser enn død. For å beskrive fordelingen av levetider, kan en bruke *overlevelsesfunksjoner* og *hasardrater*. Overlevelsesfunksjonen angir sannsynligheten for at levetiden er større enn en gitt verdi, mens hasardraten er sannsynligheten for dødsfall på et gitt tidspunkt.

Det enkleste tilfellet er å kun modellere én dødsårsak. Utvider vi dette til flere årsaker som kan stoppe levetiden til et individ, får vi *konkurrerende dødsårsaker*. Her må hasardraten for hver dødsårsak modelleres, og de kalles *årsaksspesifikke hasardrater*. Disse brukes videre til å beregne *kumulative insidensfunksjoner*, som angir sannsynlighetene for å dø av de ulike dødsårsakene.

Årsaksspesifikke hasardrater og kumulative insidensfunksjoner har ulike tolkninger og bruksområder. De årsaksspesifikke hasardratene kan brukes til å finne årsaker til en sykdom. Man ser da på hvordan ulike faktorer, eller *kovariater*, påvirker hasardratene. I motsetning til hasardraten, så vil den kumulative insidensfunksjonen for en bestemt årsak påvirkes av hasardratene for de konkurrerende dødsårsakene. Hvis færre personer dør av en årsak, så vil dødssannsynlighetene øke for de resterende årsakene. Dersom en er interessert i å undersøke betydningen av kovariatene for sannsynligheten for å dø av en bestemt sykdom, er det de kumulative insidensfunksjonene som er mest relevante. Vi vil i denne oppgaven fokusere på de kumulative insidensfunksjonene. Det er interessant å studere hvordan kovariater som røykevaner, blodtrykk og BMI påvirker dødssannsynlighetene for bestemte årsaker.

I denne oppgaven skal vi studere og sammenligne to modeller for konkurrerende dødsårsaker; *Cox proporsjonale modell* for de årsaksspesifikke hasardratene (Cox 1972) og *proporsjonal subfordelingsmodell* (Fine & Gray 1999). For begge modellene skal vi beregne kumulative insidensfunksjoner, og se hvordan de påvirkes av kovariatene.

Cox-modellen er den mest anvendte regresjonsmodellen innen levetidsanalyse. I denne modellen har kovariatene en log-lineær effekt på de årsaksspesifikke hasardratene. I en

Cox-modell for konkurrerende dødsårsaker påvirker ikke nødvendigvis en kovariat den årsaksspesifikke hasardraten og den tilhørende kumulative insidensfunksjonen i samme retning. Dette skyldes at den kumulative insidensfunksjonen for en bestemt årsak ikke bare avhenger av hasardraten for den årsaken, men også av de konkurrerende dødsårsakene. I beregning av dødssannsynligheten for en bestemt årsak med subfordelingsmodellen, består den kumulativ insidensfunksjonen av subfordelingshasardraten for kun den årsaken. Dermed vil en kovariat påvirke hasarden og insidensfunksjonen i samme retning.

Målet med oppgaven er å sammenligne de to metodene for å studere kumulative insidensfunksjoner for situasjoner som kan forekomme i praksis. Vi vil derfor ta utgangspunkt i et virkelig datasett om dødelighet og dødsårsaker i den norske befolkningen. Datasettet som vi skal bruke er et utvalg på 4000 personer hentet fra *fylkesundersøkelsene*, se neste avsnitt for beskrivelse. I tillegg skal vi simulere data som er av lignende type som data fra fylkesundersøkelsene.

Inndelingen av oppgaven er slik :

- Kapittel 2 : I dette kapitlet studerer vi modellering uten kovariater. Vi introduserer grunnleggende begreper innen levetidsanalyse og konkurrerende dødsårsaker. Siden konkurrerende dødsårsaker er et spesialtilfelle av Markov-kjeder, har vi i tillegg et avsnitt med en kort innføring i Markov-kjeder.
- Kapittel 3 : Dette kapitlet presenterer Cox proporsjonale regresjonsmodell med tidsuavhengige kovariater, og videre for konkurrerende dødsårsaker. I tillegg har vi med noen eksempler som illustrerer teorien.
- Kapittel 4 : Vi beskriver først modellering av subfordelingshasard uten kovariater, og forklarer hvordan vekter beregnes for høyre-sensurerte og venstre-trunkerte data. Deretter introduserer vi den proporsjonale subfordelingshasard modellen. På slutten har vi med figurer og tabeller som sammenligner denne modellen med Cox regresjonsmodell.
- Kapittel 5 : I dette kapitlet simulerer vi data fra to modeller, og bruker disse til å beregne kumulative insidensfunksjoner med Cox- og subfordelingsmodell, og sammenligner resultatene.
- Kapittel 6 : Her gir vi en oppsummering og diskusjon av resultatene i oppgaven.

På slutten har vi Tillegg med tabeller og figurer. Disse er referert til i kapitlene 4 og 5. Statistikkprogrammet **R** (R Development Core Team 2010) er brukt til å uføre alle beregningene. På slutten av kapitlene 2, 3 og 4 har vi egne avsnitt om programvare. I disse avsnittene beskrives funksjonene som er brukt.

1.1 Beskrivelse av datasettet

I perioden 1974-1978 ble alle menn og kvinner i alderen 35-49 år, som var bosatt i fylkene Oppland, Sogn og Fjordane og Finnmark, invitert til å delta i en helseundersøkelse med

vekt på hjertesykdom. Dette var de første hjerte- og karundersøkelsene i regi av Statens helseundersøkelser, og ble kalt *fylkesundersøkelsene*. Bakgrunnen for undersøkelsene var høy dødelighet av hjerte- og karsykdommer[1]. De som deltok måtte oppgi informasjon om sine røykevaner, og målinger av blodtrykk og kroppsmasseindeks (BMI) ble foretatt. Oppfølgingen av omtrent 50 000 personer varte frem til slutten av år 2000. Tidspunkt for død og dødsårsak ble registrert.

Det er få personer som er yngre enn 40 år, så disse har blitt venstre-trunkert ved 40 år. De fleste er mellom 40 og 50 år når de deltar i undersøkelsen. I tillegg er de høyre-sensurerte ved 70 år, siden vi er interessert i å studere dødelighet for middelaldrede menn og kvinner. En person er altså fulgt opp til alderen 70, død eller sensurering før 70 år. Datasettet er hentet fra `http://folk.uio.no/borgan/abg-2008/data/data.html`. Dette er et tilfeldig utvalg på 4000 personer fra fylkesundersøkelsene, med 2086 menn og 1914 kvinner. Dødsårsakene er delt inn i disse fire gruppene :

1. Kreft
2. Hjerte- og karsykdommer, inkludert plutselig død
3. Andre medisinske årsaker
4. Alkoholmisbruk, kronisk leversykdom og ulykker og vold

Tabell 1.1 viser en oversikt over antall døde og prosentandel for hver dødsårsak delt inn etter kjønn. For menn er det flest dødsfall grunnet hjerte- og karsykdommer inkludert plutselig død, og deretter kreft. Den vanligste dødsårsaken for kvinner er kreft, etterfulgt av hjerte- og karsykdommer. Det var like mange menn og kvinner som døde av andre medisinske årsaker. Alkoholmisbruk, kronisk leversykdom og ulykker og vold er en vanligere dødsårsak blant menn enn kvinner.

Tabell 1.1: Antall døde og prosentandel av hver dødsårsak for menn og kvinner.

Dødsårsak	Menn		Kvinner	
	Antall	%-andel	Antall	%-andel
1	129	6.2 %	88	4.6%
2	186	8.8 %	54	2.8 %
3	34	1.6 %	34	1.8 %
4	49	2.3 %	12	0.6 %
Totalt	398	19.0 %	188	9.8 %

I Tabell 1.2 og Tabell 1.3 har vi laget en oversikt over variablene i datasettet for hhv menn og kvinner. For de numeriske variablene er gjennomsnitt og standardavvik beregnet, mens for de kategoriske har vi funnet antall og prosentandel. Gjennomsnittlig målt blodtrykk for menn er 136. For kvinner er gjennomsnittsverdien 132, som er litt lavere. Gjennomsnittlig BMI-verdi er 25 for begge kjønn. Menn og kvinner har derimot ulike røykevaner. Omtrent halvparten av kvinnene var ikke-røykere, mens blant menn var andelen kun 21%. Blant menn var det 32% som hadde vært tidligere røykere, og utgjør den største røykegruppen for menn.

[1] http://www.fhi.no

Tabell 1.2: Oversikt over variablene for menn.

Variabel		
	Gjennomsnitt	Standardavvik
Blodtrykk	136	17
BMI	25	3
Alder ved røykestart	23	7
	Antall	%-andel
Fylke		
Oppland	1033	49.5 %
Sogn og Fjordane	589	28.2 %
Finnmark	464	22.3 %
Røykegruppe		
Aldri røykt	445	21.3%
Tidligere røyker	676	32.4 %
1-9 sigaretter per dag	199	9.5%
10-19 sigaretter per dag	457	21.9 %
20+ sigaretter per dag	241	11.6 %
Pipe eller sigar	68	3.3 %

Tabell 1.3: Oversikt over variablene for kvinner.

Variabel		
	Gjennomsnitt	Standardavvik
Blodtrykk	132	18
BMI	25	4
Alder ved røykestart	27	8
	Antall	%-andel
Fylke		
Oppland	1028	53.7%
Sogn og Fjordane	501	26.2 %
Finnmark	385	20.1 %
Røykegruppe		
Aldri røykt	947	49.5 %
Tidligere røyker	333	17.4 %
1-9 sigaretter per dag	225	11.8 %
10-19 sigaretter per dag	341	17.8 %
20+ sigaretter per dag	65	3.4 %
Pipe eller sigar	3	0.2 %

Kapittel 2

Modellering uten kovariater

Vi skal først ta for oss det enkleste tilfellet, som er levetidsmodeller uten kovariater. Teorien er basert på Kapittel 3 i Aalen, Borgan og Gjessing (2008), og er hovedsakelig et kort sammendrag av deler av dette kapitlet. Vi innfører grunnleggende teori om *levetidsanalyse* og *Markov kjeder*. Dette brukes til beregning av overgangssannsynlighetene i modeller for *konkurrerende dødsårsaker*.

2.1 Levetidsanalyse

Levetiden til et individ er tiden vi måler for individet fra et gitt startpunkt til et gitt sluttpunkt. Et eksempel er tid fra fødsel til død. Begrepet levetid brukes ikke bare i forbindelse med død, men er et generelt begrep for å måle tiden til en bestemt hendelse inntreffer. Andre eksempler er tiden fra pasienter blir symptomfrie til tilbakefall av sykdommen eller tid fra ekteskap til skilsmisse.

2.1.1 Begreper

Vi betegner levetiden med T, som er en ikke-negativ tilfeldig variabel, og definerer først *overlevelsesfunksjonen*

$$S(t) = P(T > t). \tag{2.1}$$

Dette er sannsynligheten for at levetiden T blir større enn t. $S(t)$ starter alltid i 1; $S(0) = 1$, og den er en synkende funksjon med tiden.

Hasardraten er definert som

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.2}$$

Dette er momentan dødelighet på tid t. Brøken viser dødsrate per tidsenhet. Når tidsinkrementet Δt er lite, blir $\alpha(t)\Delta t$ en tilnærming til sannsynligheten for at individet ikke overlever til tiden $t + \Delta t$, gitt at levetiden er større enn t. Den kumulative hasardraten

$$A(t) = \int_0^t \alpha(s)ds \tag{2.3}$$

er integralet av hasardraten fra 0 til tid t. Det kan vises at sammenhengen mellom overlevelsesfunksjonen og hasardraten er

$$S(t) = \exp\left(-\int_0^t \alpha(s)ds\right). \tag{2.4}$$

2.1.2 Data

Vi har nå sett på levetiden til ett individ. Her skal vi utvide dette til data for n uavhengige individer. Vi lar T_i være levetiden til individ i, der $i = 1, 2, ..., n$.

Når levetiden ikke kan måles, kalles det *sensurering*. Dette er vanlig i levetidsanalyse, og kan f.eks. skyldes at hendelsen ikke har inntruffet før studien avsluttes, eller at en person trekker seg fra studien etter startpunktet. Det er altså ufullstendige observasjoner. *Høyre-sensurering* er en type sensurering som oppstår når startpunktet er målt, men ikke sluttpunktet. Når denne typen sensurering ikke påvirker levetiden, kalles det *uavhengig høyre-sensurering*. Sannsynligheten for overlevelse for et individ endres ikke ved å ha sensurering. Vi antar at det er tilfellet.

Vi lar $\widetilde{T}_i$ være observert tid for individ i, og definerer indikatorvariabelen

$$D_i = \begin{cases} 1 & \text{hvis } \widetilde{T}_i = T_i \ , \\ 0 & \text{hvis } \widetilde{T}_i < T_i \ . \end{cases} \tag{2.5}$$

D_i angir om den observerte tiden til individ i er en levetid ($D_i = 1$) eller sensureringstid ($D_i = 0$). Siden vi observerer høyst én hendelse per individ, kan vi definere *telleprosessen*

$$N_i(t) = I\left(\widetilde{T}_i \leq t, D_i = 1\right) \quad \text{for } i = 1, 2, ..., n. \tag{2.6}$$

$N_i(t) = 1$ hvis den faktiske overlevelsestiden er observert, og den ligger i intervallet $[0, t]$. Videre trenger vi en risikoindikator

$$Y_i(t) = \begin{cases} 1 & \text{hvis } \widetilde{T}_i \geq t \ , \\ 0 & \text{hvis } \widetilde{T}_i < t \ , \end{cases} \tag{2.7}$$

som viser om individ i fortsatt er under observasjon rett før tid t. Vi innfører også

$$Y(t) = \sum_{i=1}^{n} Y_i(t). \tag{2.8}$$

Dette er antall individer som er under observasjon rett før tid t.

Ikke alle levetider observeres fra starten. *Venstre-trunkering* går ut på å kun inkludere individer som har levetid større eller lik en bestemt verdi, altså et tidspunkt som individene må ha overlevd til for å kunne delta i studien. Formlene over kan modifiseres til å gjelde for venstre-trunkering. Dersom s_i er trunkeringstiden for individ i, kan vi i (2.6) begrense $\widetilde{T}_i$ til å gjelde for $s_i < \widetilde{T}_i \leq t$;

$$N_i(t) = I\left(s_i < \widetilde{T}_i \leq t, D_i = 1\right) \quad \text{for } i = 1, 2, ..., n. \tag{2.9}$$

Modifisert risikoindikator blir

$$Y_i(t) = \begin{cases} 1 & \text{hvis } s_i < t \leq \widetilde{T}_i \ , \\ 0 & \text{ellers.} \end{cases} \tag{2.10}$$

2.1.3 Estimater

Vi skal nå introdusere to kjente metoder for å estimere overlevelsesfunksjonen og den kumulative hasardraten. Disse er hhv *Kaplan-Meier* og *Nelson-Aalen*, som begge er ikke-parametriske estimeringsmetoder. Vi antar at hendelsestidspunktene er forskjellige for alle individene, dvs. at flere hendelser ikke kan inntreffe på samme tidspunkt. Det finnes formler som er tilpasset sammenfallende hendelser for begge metodene, men disse tar vi ikke med her.

Det er enklere å estimere den kumulative hasardraten $A(t)$ enn å estimere hasardraten $\alpha(t)$ direkte. *Nelson-Aalen* metoden estimerer $A(t)$, og er gitt ved

$$\widehat{A}(t) = \sum_{i:\widetilde{T}_i \leq t, D_i=1} \frac{1}{Y(\widetilde{T}_i)}. \tag{2.11}$$

Estimatoren øker med $1/Y(\widetilde{T}_i)$ for hver hendelse, og summen av disse gir en estimator for den kumulative hasardfunksjonen. Nelson-Aalen estimatoren er tilpasset høyresensurerte data, og er derfor mye brukt i overlevelsessanalyse. Den gjelder også ved venstretrunkering, og er en høyre-kontinuerlig funksjon. Variansen estimeres med

$$\widehat{\sigma}^2(t) = \sum_{i:\widetilde{T}_i \leq t, D_i=1} \frac{1}{Y(\widetilde{T}_i)^2}. \tag{2.12}$$

Kaplan-Meier estimatoren estimerer overlevelsesfunksjonen $S(t)$, og er gitt ved

$$\widehat{S}(t) = \prod_{i:\widetilde{T}_i \leq t, D_i=1} \left(1 - \Delta\widehat{A}(\widetilde{T}_i)\right) = \prod_{i:\widetilde{T}_i \leq t, D_i=1} \left(1 - \frac{1}{Y(\widetilde{T}_i)}\right), \tag{2.13}$$

der $\Delta\widehat{A}(\widetilde{T}_i)$ er inkrementet i Nelson-Aalen estimatoren på tid $\widetilde{T}_i$. Variansen kan estimeres som produktet av $\widehat{S}(t)^2$ og $\widehat{\sigma}^2(t)$;

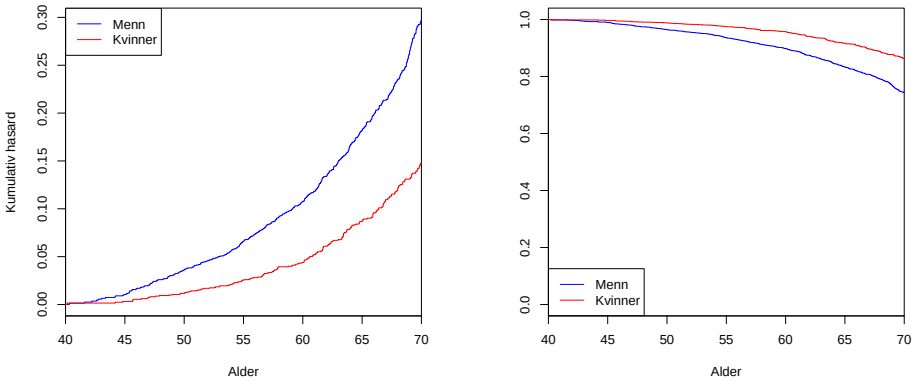
$$\widehat{\tau}^2(t) = \widehat{S}(t)^2 \sum_{i:\widetilde{T}_i \leq t, D_i=1} \frac{1}{Y(\widetilde{T}_i)^2}. \tag{2.14}$$

Et alternativ til denne estimatoren er *Greenwood*-estimatoren, men denne tas ikke med her.

Hvis vi har store utvalg, er Nelson-Aalen og Kaplan-Meier estimatorene tilnærmet normalfordelte for en gitt t. Dette kan brukes til å finne konfidensintervaller.

Eksempel 2.1. I dette eksemplet skal vi beregne estimatene (2.11) og (2.13) for datasettet, som er beskrevet i Kapittel 1, separat for menn og kvinner. Vi skal se på dødeligheten for alle dødsårsakene samlet. I datasettet er det 2086 menn og 1914 kvinner.

I Figur 2.1a har vi plottet de empiriske kumulative hasardratene. De er beregnet ved å bruke Nelson-Aalen estimatoren (2.11). Her må vi se på stigningstallene til plottene. De viser at menn har høyere dødsintensitet enn kvinner. Mellom 40 og 55 år er stigningstallene tilnærmet konstante for begge kjønn. I denne aldersgruppen er hasardraten omtrent 0.005



(a) (b)

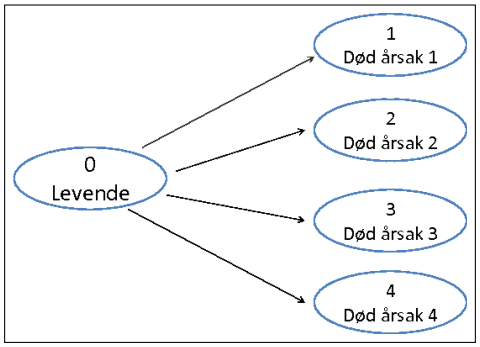
Figur 2.1: (a) Nelson-Aalen estimater av de kumulative hasardratene og (b) Kaplan-Meier estimater for total dødelighet separat for menn og kvinner. I begge figurene har vi brukt datasettet som er beskrevet i Eksempel 2.1.

per år for menn og 0.002 per år for kvinner. Etter 55 år ser vi at dødsintensitetene øker med alderen.

Figur 2.1b viser de estimerte overlevelsesfunksjonene for menn og kvinner. Siden dataene er vestre-trunkerte, så ser vi på betinget overlevelse. Vi ser at menn har lavere overlevelsessannsynlighet enn kvinner for alle aldre, og forskjellen øker med alderen. Gitt overlevelse til 40 år, så er sannsynligheten for å overleve til 50 år estimert til 96% og 99% for hhv menn og kvinner. Ved alderen 70 har de betingede sannsynlighetene sunket til 74% for menn og 86% for kvinner.

2.2 Konkurrerende dødsårsaker

Vi har sett på tilfellet der én hendelse kan inntreffe. Nå skal vi utvide til en modell med flere mulige hendelser, dvs. at vi kan ha overganger til flere tilstander. For et individ som er under risiko, kan levetiden stoppes av flere årsaker. Dette kalles *konkurrerende dødsårsaker*. Konkurrerende dødsårsaker er et spesialtilfelle av *Markov kjeder* (se avsnitt 2.3). Dersom vi har k ulike dødsårsaker, får vi en modell med $k+1$ tilstander. Vi lar den stokastiske prosessen $X(t)$ beskrive tilstanden til prosessen på tid t, og $X(t-)$ tilstanden rett før tid t. Tilstanden *levende* betegnes med 0. I Figur 2.2 ser vi et eksempel med mulige overganger for $k = 4$ konkurrerende dødsårsaker.



Figur 2.2: Tilstander for fire konkurrerende dødsårsaker.

2.2.1 Begreper

Når vi studerer konkurrerende dødsårsaker, må vi spesifisere dødsårsaken når vi skal finne hasardraten. Hasardratene her kalles *årsaksspesifikke hasardrater*, og er definert som

$$\alpha_{0h}(t) = \lim_{\Delta t \to 0} \frac{P\left(X(t+\Delta t) = h | X(t-) = 0\right)}{\Delta t}. \tag{2.15}$$

Denne brøken har samme tolkning som (2.2); telleren viser sannsynligheten for at prosessen er i tilstand h på tid $t+\Delta t$, gitt at prosessen befinner seg i tilstand 0 rett før tiden t. $\alpha_{0h}(t)$ viser altså den momentane overgangssannsynligheten på tid t fra tilstanden levende til død av årsak h. Ved å integrere over $\alpha_{0h}(s)$ fra 0 til t, får vi den kumulative årsaksspesifikke hasardraten

$$A_{0h}(t) = \int_0^t \alpha_{0h}(u)du. \tag{2.16}$$

De $k+1$ ulike overgangssannsynlighetene mellom to tidspunkter s og t, for $s < t$, skrives som

$$P_{0h}(s,t) = P\{X(t) = h | X(s) = 0\} \quad \text{for } h = 0, ..., k. \tag{2.17}$$

Overlevelsessannsynligheten fra tid s til t er gitt som

$$P_{00}(s,t) = \exp\left[-\int_s^t \sum_{h=1}^{k} \alpha_{0h}(u)du\right], \tag{2.18}$$

der $\sum_{h=1}^{k} \alpha_{0h}(u)$ er den totale hasarden på tid u når alle dødsårsakene tas i betraktning. Dette er overlevelsesfunksjonen uttrykt med hasardrater, tilsvarende (2.4). Videre har vi de resterende overgangssannsynlighetene

$$P_{0h}(s,t) = \int_s^t P_{00}(s,u)\alpha_{0h}(u)du \quad \text{for } h = 1, ..., k. \tag{2.19}$$

Disse sannsynlighetene kalles *kumulative insidensfunksjoner*, og viser sannsynligheten for å dø av de k forskjellige dødsårsakene. Integralet i (2.19) kan forklares intuitivt; sannsynligheten for å være levende mellom s og u, $P_{00}(s,u)$, multipliseres med sannsynligheten for å dø av årsak h på tid u, $\alpha_{0h}(u)du$. Siden overgangen kan skje til alle tider u mellom s og t, integrerer vi over dette intervallet.

2.2.2 Data

Som beskrevet i avsnitt 2.1.2, er $\widetilde{T}_i$ observert levetid for individ i, der $i = 1, ..., n$. Tilsvarende (2.5) har vi her

$$D_i = \begin{cases} h & \text{hvis død av årsak } h, \\ 0 & \text{hvis sensurering.} \end{cases} \tag{2.20}$$

Denne variabelen viser om $\widetilde{T}_i$ er sensureringstidspunkt eller levetid. Hvis sistnevnte er tilfellet, viser den dødsårsaken. Siden vi ser på overganger til forskjellige tilstander, har vi nå k telleprosesser for individ i;

$$N_{ih}(t) = I\left(\widetilde{T}_i \leq t, D_i = h\right) \quad \text{for } h = 1, 2, ..., k. \tag{2.21}$$

$N_{ih}(t) = 1$ hvis individ i dør av årsak h i intervallet $[0, t]$. Riskindikatoren blir

$$Y_i(t) = \begin{cases} 1 & \text{hvis } \widetilde{T}_i \geq t\ , \\ 0 & \text{hvis } \widetilde{T}_i < t\ . \end{cases} \tag{2.22}$$

Dette er en indikatorvariabel som viser om individ i er under risiko rett før tid t. Summen

$$Y(t) = \sum_{i=1}^{n} Y_i(t) \tag{2.23}$$

er antall individer som fortsatt er under risiko rett før tid t.

2.2.3 Estimater

Nelson-Aalen estimatoren estimerer de kumulative årsaksspesifikke hasardratene;

$$\widehat{A}_{0h}(t) = \sum_{i:\widetilde{T}_i \leq t, D_i = h} \frac{1}{Y(\widetilde{T}_i)}. \tag{2.24}$$

Kaplan-Meier estimatoren brukes til å estimere $P_{00}(s, t)$, så vi får

$$\widehat{P}_{00}(s, t) = \prod_{i:s < \widetilde{T}_i \leq t, D_i \neq 0} \left(1 - \frac{1}{Y(\widetilde{T}_i)}\right). \tag{2.25}$$

Vi skriver de ordnede observerte levetidene som $\widetilde{T}_{(1)} \leq \widetilde{T}_{(2)} \leq ... \leq \widetilde{T}_{(n)}$, og de tilhørende indikatorvariablene som $D_{(1)}, D_{(2)}, ..., D_{(n)}$. Estimatoren (2.25) benyttes i estimeringen av de kumulative insidensfunksjonene $P_{0h}(s, t)$, og den naturlige estimatoren er

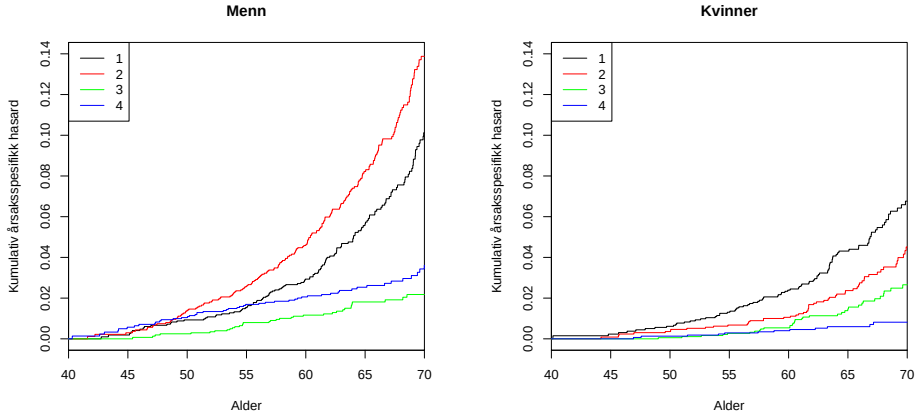
$$\widehat{P}_{0h}(s, t) = \sum_{i:s < \widetilde{T}_{(i)} \leq t, D_{(i)} \neq 0} \widehat{P}_{00}(s, \widetilde{T}_{(i-1)}) \Delta \widehat{A}_{0h}(\widetilde{T}_{(i)}), \tag{2.26}$$

der $\Delta \widehat{A}_{0h}(\widetilde{T}_{(i)}) = \Delta N_{0h}(\widetilde{T}_{(i)}) / Y_0(\widetilde{T}_{(i)})$. Vi får denne estimatoren ved å erstatte $P_{00}(s, u)$ med (2.25) for $t = \widetilde{T}_{(i-1)}$, og $\alpha_{0h}(u)du$ med $\Delta \widehat{A}_{0h}(\widetilde{T}_{(i)})$. Varians-estimatorene til $\widehat{P}_{00}(s, t)$ og $\widehat{P}_{0h}(s, t)$ er gitt i hhv seksjonene 3.2.1 og 3.4.5 i Aalen, Borgan og Gjessing (2008).

Eksempel 2.2. Vi skal her se på samme datasett som ble brukt i Eksempel 2.1. Dødsårsakene er kategorisert slik :

1. Kreft
2. Hjerte- og karsykdommer, inkludert plutselig død
3. Andre medisinske årsaker
4. Alkoholmisbruk, kronisk leversykdom og ulykker og vold

Figur 2.3 viser Nelson-Aalen estimatene av de kumulative årsaksspesifikke hasardratene for menn og kvinner. Vi har brukt formel (2.24) for å beregne disse. For menn ser vi at de kumulative hasardratene for årsak 3 og årsak 4 danner tilnærmede rette linjer, og er nesten parallelle etter 50-års alderen. Dette tyder på konstant hasardrate for disse dødsårsakene. Stigningstallet er ca. 0.001 for dødsårsak 3 og 4 etter 50 år. Det er dødsårsak 2, hjerte- og karsykdommer, som har høyest hasardrate fra 50 år.

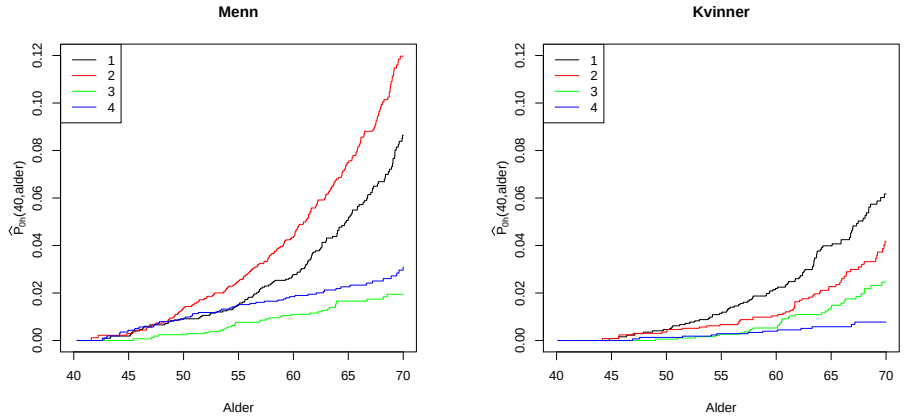


(a) (b)

Figur 2.3: Nelson-Aalen estimater av de kumulative årsaksspesifikke hasardratene for de fire dødsårsakene beskrevet i Eksempel 2.2.

For kvinner er det kreft som har forårsaket flest dødsfall, og som har høyest hasardrate for alle aldre. Hasardraten er tilnærmet konstant frem til 55 år. Da er den ca. 0.01 per år. Etter denne alderen ser det ut til at raten øker hvert år. Kurven for årsak 4 danner en tilnærmet rett horisontal linje, så det er veldig lav risiko.

Siden dataene er venstre-trunkerte og høyre-sensurerte ved 70 år, estimerer vi $P_{0h}(40,t)$ for $h = 1,2,3,4$ og $t \in [40,70]$. I Figur 2.4 ser vi de estimerte overgangssannsynlighetene, dvs. de kumulative insidensfunksjonene, fra tilstand 0 når vi tar hensyn til alle dødsårsakene. Disse har samme form som plottene i Figur 2.3, men vi leser av de estimerte overgangssannsynlighetene mellom to tidspunkter direkte på y-aksen. Blant menn er estimert kumulativ insidensfunksjon mellom 40 og 70 år høyest for dødsårsak 2. Mellom 40 og 70 år er den estimerte sannsynligheten 12% for å dø av denne årsaken. For kvinner



(a) (b)

Figur 2.4: Estimat av de kumulative insidensfunksjonene for de fire dødsårsakene beskrevet i Eksempel 2.2

er denne sannsynligheten lavere; omtrent 4%. Estimert sannsynlighet for å dø av kreft er 8.6% og 6.3% for hhv menn og kvinner.

2.3 Markov kjeder

En *Markov-kjede* i kontinuerlig tid er en stokastisk prosess $\{X(t)\}$ for $t \geq 0$ som beveger seg mellom ulike *tilstander*. Hvis prosessen er i tilstand h på tid t, skrives dette som $X(t) = h$. Vi antar at prosessen består av et endelig antall mulige tilstander; $\mathbb{C} = \{0, 1, ..., k\}$. En viktig egenskap med Markov-kjeder er at prosessen kun avhenger av nå-tilstanden, og ikke av tidligere tilstander som prosessen har vært i. Varighet i en tilstand har heller ikke betydning for overgangssannsynlighetene.

2.3.1 Begreper

Generelt er den momentane overgangssannsynligheten eller intensiteten fra tilstand g til h på tid t definert som

$$\alpha_{gh}(t) = \lim_{\Delta t \to 0} \frac{P\left(X(t + \Delta t) = h | X(t-) = g\right)}{\Delta t} \quad \text{for } g \neq h. \tag{2.27}$$

Overgangssannsynligheten mellom to tilstander g og h skrives som

$$P_{gh}(s, t) = P\{X(t) = h | X(s) = g\} \quad \text{for } s < t \ . \tag{2.28}$$

Dette er sannsynligheten for at prosessen befinner seg i tilstand h på tid t, gitt at den er i tilstand g på et tidligere tidspunkt s. For konkurrerende dødsårsaker uttrykte vi disse

sannsynlighetene med hasardrater. Generelt kan vi ikke finne uttrykk for disse overgangssannsynlighetene. Med $(k+1)$ tilstander, har vi $(k+1)\times(k+1)$ overgangssannsynligheter. Overgangssannsynlighetene mellom tidspunktene s og t ordnes i en matrise

$$\mathbf{P}(s,t) = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & \dots & k \end{array} \\ \begin{pmatrix} P_{00}(s,t) & P_{01}(s,t) & \dots & P_{0k}(s,t) \\ P_{10}(s,t) & P_{11}(s,t) & \dots & P_{1k}(s,t) \\ \vdots & \vdots & \ddots & \vdots \\ P_{k0}(s,t) & P_{k1}(s,t) & \dots & P_{kk}(s,t) \end{pmatrix} \end{array} \tag{2.29}$$

for $s < t$. For konkurrerende dødsårsaker er det kun mulig med overganger fra tilstand 0. Dermed består første rad i matrisen av sannsynligheter. I tillegg er $P_{gg}(s,t) = 1$ for $g = 2, 3, ..., k$. Resten av sannsynlighetene i matrisen er 0.

$\mathbf{P}(s,t)$ kan skrives som et matriseprodukt (Ross(2007))

$$\mathbf{P}(s,t) = \mathbf{P}(t_0,t_1) \times \mathbf{P}(t_1,t_2) \times \cdots \times \mathbf{P}(t_{K-1},t_K), \tag{2.30}$$

der $s = t_0 < t_1 < ... < t_K = t$, som vi får hvis vi deler opp intervallet $(s,t]$ i K deler.

På tilsvarende måte som for overgangssannsynlighetene, kan vi definere en matrise $\boldsymbol{\alpha}(t)$ med alle intensitetene på tid t;

$$\boldsymbol{\alpha}(t) = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & \dots & k \end{array} \\ \begin{pmatrix} \alpha_{00}(t) & \alpha_{01}(t) & \dots & \alpha_{0k}(t) \\ \alpha_{10}(t) & \alpha_{11}(t) & \dots & \alpha_{1k}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1}(t) & \alpha_{k2}(t) & \dots & \alpha_{kk}(t) \end{pmatrix} \end{array}. \tag{2.31}$$

Summen av hver rad i matrisen skal være 0, så diagonalelementene er gitt som $\alpha_{gg}(t) = -\sum_{h\neq g} \alpha_{gh}(t)$ for $g = 0, 1, ..., k$. Vi kan tilnærme overgangssannsynlighetene over et lite intervall $(u, u+du]$ med

$$\mathbf{P}(u, u+du) \approx \mathbf{I} + \boldsymbol{\alpha}(u)du, \tag{2.32}$$

der $\mathbf{I}$ er identitetsmatrisen med samme dimensjon som $\boldsymbol{\alpha}(u)$. Dersom vi øker antall subintervaller K, og samtidig lar tidsinkrementene du bli mindre, så kan vi bruke (2.32) til å skrive (2.30) som produktintegralet

$$\mathbf{P}(s,t) = \prod_{u\in(s,t]} \{\mathbf{I} + \boldsymbol{\alpha}(u)du\}. \tag{2.33}$$

2.3.2 Data

Vi ser på n individer som observeres, og lar $T_1 < T_2 < ...$ betegne målte tidspunkter for observerte overganger mellom to tilstander. Vi antar at flere overganger ikke kan skje på samme tidspunkt. Telleprosessen $N_{gh}(t)$ viser antall overganger fra tilstand g til h i intervallet $[0,t]$. Indikatorvariabelen $Y_{ig}(t)$ viser om individ i er i tilstand g rett før tid t, mens $Y_g(t) = \sum_{i=1}^{n} Y_{ig}(t)$ viser hvor mange individer som befinner seg i tilstand g rett før tid t.

2.3.3 Estimater

Nelson-Aalen estimatoren kan også her brukes til å estimere de kumulative overgangsintensitetene. Disse ordnes i matrisen $\widehat{\mathbf{A}}(t)$, og elementene er

$$\widehat{A}_{gh}(t) = \begin{cases} \int_0^t dN_{gh}(s)/Y_g(s) & \text{for } h \neq g, \\ -\sum_{h \neq g} \widehat{A}_{gh}(t) & \text{for } h = g. \end{cases} \tag{2.34}$$

Estimatoren til matrisen $\mathbf{P}(s,t)$ er

$$\widehat{\mathbf{P}}(s,t) = \prod_{s<T_i \leq t} \left(\mathbf{I} + \Delta \widehat{\mathbf{A}}(T_i) \right). \tag{2.35}$$

Estimatorene (2.25) og (2.26) for konkurrerende dødsårsaker er spesialtilfeller av det generelle tilfellet (2.35). For store utvalg er estimatoren tilnærmet normalfordelt. Vi kan også finne formler for å estimere kovarianser. Disse egenskapene er gitt i seksjonene 3.4.4 og 3.4.5 i Aalen, Borgan og Gjessing (2008).

2.4 Programvare

For beregning av estimatene i Eksempel 2.2, bruker vi pakken `mstate` i **R** (R Development Core Team 2010). Vi skal her gi en kort gjennomgang av hvordan pakken virker. Vi har tatt utgangspunkt i Wreede *et al.* (2011). Artikkelen inneholder dokumentasjon til `mstate`-pakken, som er utviklet til å beregne bl.a. elementene i matrisen (2.35).

Funksjonen `transMat()` oppretter en overgangsmatrise. De mulige overgangene betegnes med tall fra 1 og oppover, mens de resterende elementene i matrisen betegnes med `NA`. For spesialtilfellet konkurrerende dødsårsaker, er det laget en egen funksjon, `trans.comprisk()`, som gjør dette. Overgangsmatrisen i Eksempel 2.2 ser slik ut :

```
        to
from    alive dead1 dead2 dead3 dead4
  alive    NA     1     2     3     4
  dead1    NA    NA    NA    NA    NA
  dead2    NA    NA    NA    NA    NA
  dead3    NA    NA    NA    NA    NA
  dead4    NA    NA    NA    NA    NA
```

For å videre kunne bruke funksjonene i pakken, må dataene være ordnet i et bestemt format; *long format*. Dataene blir ordnet slik at for hvert individ i datasettet, vil alle mulige overganger bli listet nedover, så hvert individ får flere rader. Funksjonen `msprep()` konverterer til dette formatet. For konkurrerende årsaker får hver person k rader; én rad for hver dødsårsak. Siden datasettet vi bruker har fire dødsårsaker, ser long-formatet for første individ slik ut:

```
   id from to trans Tstart Tstop  time status sex county sbp  bmi smkstart
1  1    1  2     1  40.00 60.80 20.80      0   2     14 110 2.18       NA
2  1    1  3     2  40.00 60.80 20.80      0   2     14 110 2.18       NA
3  1    1  4     3  40.00 60.80 20.80      0   2     14 110 2.18       NA
4  1    1  5     4  40.00 60.80 20.80      0   2     14 110 2.18       NA
```

For å bruke funkjonene som beskrives videre, må dataene være i dette formatet.

Funksjonen `msfit()` estimerer kumulative intensiteter svarende til matrisen (2.31). Et objekt av typen `coxph` brukes som input. Det kan være en Cox-modell med eller uten kovariater. Her har vi ikke med noen kovariater. Dette objektet brukes videre som parameter i funksjonen `probtrans()`, som estimerer elementene i matrisen (2.35), dvs. overgangssannsynlighetene. I denne funksjonen må vi spesifisere hvilken metode som skal brukes for å beregne varians-estimater. Valget "aalen" gir varians-estimatene i seksjon 3.4.5 i Aalen, Borgan og Gjessing (2008).

Kapittel 3

Modellering med kovariater

I Kapittel 2 tilpasset vi modeller for levetidsdata uten å ta hensyn til kovariater. Her skal vi ta de med. Vi antar at det er n individer som observeres, og ser igjen kun på tilfellet der maksimalt én hendelse kan inntreffe for hvert individ. Vi betrakter kun tidsuavhengige kovariater, og antar at venstre-trunkering og høyre-sensurering er uavhengige av levetidene til individene.

3.1 Cox regresjon

Som beskrevet i avsnitt 2.1.2, så er $Y_i(t)$ risikoindikator for individ i rett før tid t. Vi betegner T_i som den faktiske levetiden til individ i, og lar $\widetilde{T}_i$ være observert levetid. $D_i = I(\widetilde{T}_i = T_i)$ er den tilhørende indikatorvariabelen som viser om $\widetilde{T}_i$ er den faktiske levetiden, og $N_i(t)$ er antall hendelser for dette individet i tidsrommet $[0\ ,\ t]$. Hvis vi har p kovariater for individ i, ordnes de i vektoren

$$\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T, \tag{3.1}$$

mens vektoren med regresjonskoeffisientene skrives som

$$\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T. \tag{3.2}$$

For å definere hasardraten, må vi spesifisere kovariatene vi studerer. I en Cox-regresjonsmodell er hasardraten for individ i definert som

$$\alpha(t|\mathbf{x}_i) = \alpha_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}, \tag{3.3}$$

der $\alpha_0(t)$ kalles *baseline hasard*. Kovariatene har i denne modellen en log-lineær effekt på hasardraten. Forholdet mellom hasardratene til to individer med kovariater $\mathbf{x}_1$ og $\mathbf{x}_2$ blir

$$\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = \exp\{\boldsymbol{\beta}^T(\mathbf{x}_2 - \mathbf{x}_1)\}, \tag{3.4}$$

og kalles *hasard ratio* (HR). Den er konstant og modellen kalles *proporsjonal*. Effekten av kovariatene endres altså ikke over tid. Hvis vi sammenligner to individer der kovariat j

for den ene er én enhet større enn den andre, dvs. $x_{2j} = x_{1j} + 1$, og alle de resterende kovariatene har samme verdi for begge individene, får vi forholdet

$$\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = e^{\beta_j}, \tag{3.5}$$

som er et spesialtilfelle av (3.4). Ved å øke kovariat j med én enhet, så multipliseres hasardraten med e^{β_j}.

Siden Cox-modellen ikke er parametrisk, kan vi ikke bruke ordinær likelihood funksjon til å estimere regresjonskoeffisientene. Vi beregner isteden den *partielle* likelihood funksjonen

$$L(\boldsymbol{\beta}) = \prod_{i:D_i=1} \frac{Y_i(\widetilde{T}_i)\exp\{\boldsymbol{\beta}^T\mathbf{x}_i\}}{\sum_{l=1}^{n} Y_l(\widetilde{T}_i)\exp\{\boldsymbol{\beta}^T\mathbf{x}_l\}}. \tag{3.6}$$

Dette er et produkt over alle observerte hendelsestidspunkter. Brøken viser sannsynligheten for at hendelsen inntreffer for individ i på tid $\widetilde{T}_i$, gitt at det har inntruffet en hendelse på dette tidspunktet blant alle individene som er under risiko rett før $\widetilde{T}_i$. Se seksjon 4.1 i Aalen, Borgan og Gjessing (2008) for utledning. Ved å maksimere den partielle likelihood-funksjonen (3.6), så finner vi de estimerte regresjonskoeffisientene $\widehat{\boldsymbol{\beta}}$. Det kan vises at denne vektoren er multivariat normalfordelt for store utvalg med kovariansmatrise gitt som den inverse av *observert informasjonsmatrise*

$$\mathbf{I}(\boldsymbol{\beta}) = \{\frac{-\partial^2}{\partial\beta_h\partial\beta_j}\log L(\boldsymbol{\beta})\} \text{ for } h,j = 1,2,...,p\,. \tag{3.7}$$

Fra dette kan vi lage konfidensintervaller, og utføre tester for $\boldsymbol{\beta}$. Se seksjon 4.1.1 i Aalen, Borgan og Gjessing (2008).

Eksempel 3.1. Fra Tabell 1.2 og Tabell 1.3 legger vi merke til at det var få personer som røykte pipe eller sigar. Dette gir usikre estimater, så vi velger å fjerne denne gruppen fra alle analysene videre i oppgaven. Totalt utgjør de 77 personer. For å illustrere teorien i dette avsnittet tilpasser vi to Cox-modeller, separat for menn og kvinner, og tar med kovariatene *blodtrykk*, *BMI* og *røykegruppe*.

I en Cox-modell antar man at de kontinuerlige variablene har log-lineær effekt på dødeligheten. Her er det BMI og blodtrykk som er registrert som numeriske. Vi har sjekket denne antagelsen for disse variablene, og den virker ikke rimelig for BMI. Vi velger derfor å gruppere denne variabelen. Vi har tatt utgangspunkt i den generelle BMI-skalaen [1], og justerer den for de laveste og høyeste BMI-verdiene. Dette gjør vi fordi det er få data for lave og høye BMI-verdier. Vi deler variabelen inn i gruppene mindre enn 20 (undervektig), 20-24.9 (normal vekt), 25-29.9 (overvektig) og 30 eller over (fedme). Tabell 3.1 viser en oversikt over antall menn og kvinner i hver gruppe. Det er flest personer under kategorien normal vekt. For BMI bruker vi normal vekt som referansegruppe, og for røykevaner er det personer som aldri har røykt som er referansegruppen. I tillegg sentrerer vi variabelen blodtrykk ved å trekke fra 135 for begge kjønn, så referansepersonene har denne verdien.

Fra regresjonsanalysen for menn, som er oppsummert i Tabell 3.2, ser vi at blodtrykk er en tydelig signifikant variabel. Ved å øke blodtrykk med 10, og samtidig holde verdiene

[1] http://www.nlm.nih.gov/medlineplus/ency/article/007196.htm

Tabell 3.1: Fordelingen av antall menn og kvinner i de fire BMI-gruppene.

Kategori	BMI	Antall	
		Menn	Kvinner
Undervektig	< 20.0	50	98
Normal vekt	20.0-24.9	994	1036
Overvektig	25.0-29.9	835	557
Fedme	≥ 30.0	98	195
Totalt		1977	1886

av de resterende kovariatene faste, så ser vi at dødsrisikoen øker med ca. 20%. BMI-gruppen fedme den eneste kovariaten som ikke er signifikant på nivå 5%. For en person som er undervektig er dødsrisikoen omtrent dobbelt så høy sammenlignet med en som har normal vekt. For røyking er de estimerte regresjonskoeffisientene tilnærmet like store for røykegruppene 1-9, 10-19 og 20+ sigaretter per dag. I forhold til en som aldri har røykt, er estimert dødelighet over tre ganger høyere for en mann som er i den siste røykegruppen. For tidligere røykere er dødsintensiteten 53% høyere.

Tabell 3.2: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med total dødelighet for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.180	1.198	0.031	0.000
BMI				
Undervektig	0.745	2.107	0.271	0.006
Overvektig	0.249	1.283	0.111	0.025
Fedme	-0.050	0.951	0.251	0.841
Røykegruppe				
Tidligere røyker	0.426	1.531	0.186	0.022
1-9 sigaretter per dag	1.086	2.962	0.211	0.000
10-19 sigaretter per dag	1.088	2.969	0.182	0.000
20+ sigaretter per dag	1.191	3.291	0.201	0.000

I Tabell 3.3 ser vi at blodtrykk omtrent samme effekt på totaldødeligheten for kvinner som for menn. Her er det BMI-gruppen overvektig som ikke er signifikant. Gruppene undervektig og fedme har tilnærmet samme effekt. Vi legger merke til at tidligere røyking blant kvinner ikke har signifikant effekt. Største estimerte HR er for de som røyker 20+ pr dag. Sammenlignet med en kvinne som ikke røyker, er dødsintensiteten ca. 4.5 ganger høyere.

3.1.1 Estimater

I dette avsnittet skal vi oppgi estimatene for kumulativ baseline, kumulativ hasard og overlevelse for gitte kovariater. Disse er hentet fra seksjon 4.1.2 i Aalen, Borgan og Gjessing (2008).

Tabell 3.3: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med total dødelighet for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.171	1.186	0.032	0.000
BMI				
Undervektig	0.678	1.971	0.283	0.017
Overvektig	0.073	1.076	0.180	0.685
Fedme	0.610	1.840	0.216	0.005
Røykegruppe				
Tidligere røyker	0.353	1.423	0.221	0.111
1-9 sigaretter per dag	0.768	2.156	0.217	0.000
10-19 sigaretter per dag	0.705	2.024	0.205	0.001
20+ sigaretter per dag	1.500	4.483	0.301	0.000

Kumulativ baseline hasard

$$A_0(t) = \int_0^t \alpha_0(u)du \tag{3.8}$$

kan estimeres med *Breslow-estimatoren*

$$\widehat{A}_0(t) = \sum_{i:\widetilde{T}_i \le t, D_i=1} \frac{1}{\sum_{l=1}^n Y_l(\widetilde{T}_i)\exp(\widehat{\boldsymbol{\beta}}^T \mathbf{x}_l)}. \tag{3.9}$$

Videre har vi kumulativ hasard for en gitt vektor $\mathbf{x}_0$ med kovariater

$$A(t|\mathbf{x}_0) = \exp(\boldsymbol{\beta}^T \mathbf{x}_0) A_0(u). \tag{3.10}$$

Ved å sette inn de estimerte regresjonskoeffisientene og estimere baseline hasard med (3.9), har vi estimatoren

$$\widehat{A}(t|\mathbf{x}_0) = \exp(\widehat{\boldsymbol{\beta}}^T \mathbf{x}_0)\widehat{A}_0(t). \tag{3.11}$$

Tilsvarende Kaplan-Meier estimatoren (2.13), så er

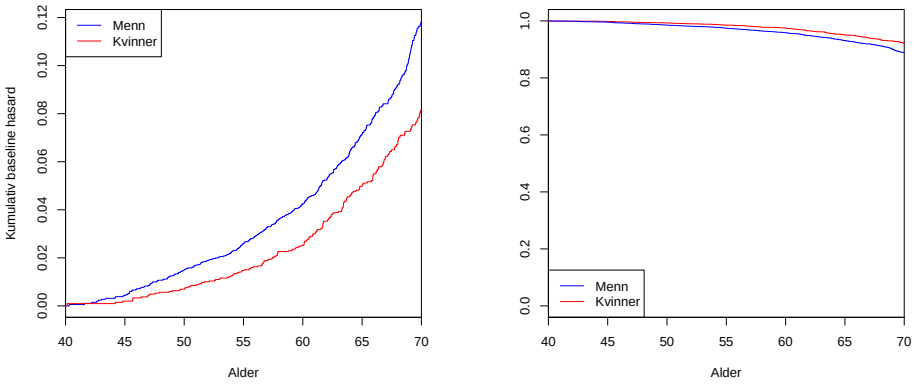
$$\widehat{S}(t|\mathbf{x}_0) = \prod_{i:\widetilde{T}_i \le t, D_i=1} \{1 - \Delta\widehat{A}(\widetilde{T}_i|\mathbf{x}_0)\} \tag{3.12}$$

estimator for overlevelse gitt kovariatene $\mathbf{x}_0$. Denne kan utledes ved å skrive $S(t|\mathbf{x}_0)$ som et produktintegral.

Eksempel 3.2. Vi skal her estimere baseline hasard for de to Cox-modellene som er estimert i Eksempel 3.1. Deretter estimerer vi overlevelse for noen kombinasjoner av kovariatene. Dette blir en utvidelse av Eksempel 2.1, der vi ikke tok med kovariatene.

I Figur 3.1a har vi plottet Breslow-estimatene av de kumulative baseline hasardene for menn og kvinner. For begge kjønn har referansepersonen blodtrykk 135, er ikke-røyker og tilhører BMI gruppen normal. Mellom 40 og 55 år er dødsintensitetene tilnærmet konstante, og verdiene er omtrent 0.004 pr år for menn og 0.001 for kvinner. Figur 3.1b viser plott av estimert betinget overlevelse fra 40 år for referansepersonene. Ved 50 år er estimert

overlevelse omtrent 99% for begge kjønn. For alderen 70 år, er de estimerte sannsynlighetene 89% og 92% for hhv menn og kvinner. Vi observerer det samme som i Figur 2.1; dødeligheten er generelt høyere blant menn, men det er mindre forskjell mellom kjønnene.



(a) (b)

Figur 3.1: (a) Breslow-estimater av de kumulative baseline hasardene den tilpassede Cox-modellen i Eksempel 3.1. (b) Estimert overlevelse for referansepersoner.

Videre skal vi plotte de estimerte overlevelsessannsynlighetene for noen kombinasjoner av kovariatene. Vi skal sammenligne normal vekt og overvekt, ikke-røyker og røykegruppen 10 - 19 sigaretter per dag, og to verdier av blodtrykk; 125 og 145. Dette blir åtte kombinasjoner, og disse er listet i Tabell 3.4. Verdiene for blodtrykk er valgt utfra nedre og øvre kvartiler.

Tabell 3.4: Åtte kombinasjoner av kovariatene blodtrykk, BMI-gruppe og røykegruppe.

Person	Blodtrykk	BMI-gruppe	Røykegruppe
A	125	Normal	Aldri røykt
B	125	Normal	10-19 sigaretter pr dag
C	145	Normal	Aldri røykt
D	145	Normal	10-19 sigaretter pr dag
E	125	Overvektig	Aldri røykt
F	125	Overvektig	10-19 sigaretter pr dag
G	145	Overvektig	Aldri røykt
H	145	Overvektig	10-19 sigaretter per dag

Figur 3.2 viser plott av estimert overlevelse for de åtte personene A-H i Tabell 3.4. Kvinner har lavere dødelighet enn menn i alle plottene. Forskjellen mellom overlevelseskurvene er størst for kombinasjon H, som er en person med relativt høyt blodtrykk, er overvektig og røyker 10-19 sigaretter pr dag. Ved 70 år er overlevelsessannsynligheten kun 58% for menn med denne kombinasjonen av kovariatene, mens for kvinner er den 81%. Denne

kombinasjonen gir lavest overlevelse for begge kjønn. Generelt ser vi at røyking reduserer overlevelsen. Økt blodtrykk og BMI har ikke så stor effekt på dødeligheten for kvinner. Endring av disse kovariatene har større effekt for menn.

3.2 Cox-regresjon for konkurrerende dødsårsaker

Vi skal gi en kort innføring i *stratifiserte* Cox-modeller for konkurrerende dødsårsaker. Stratifisering av en Cox-modell går ut på å gruppere populasjonen og tilpasse en baseline for hver gruppe, som kalles *stratum*. Dersom f.eks. antagelsen om proporsjonale hasarder ikke er oppfylt, kan en stratifisert modell være et alternativ. For konkurrerende dødsårsaker skal vi stratifisere utfra de k dødsårsakene, så vi tilpasser en baseline for hver dødsårsak. Hasardraten for individ i som tilhører stratum h er gitt som

$$\alpha_h(t|\mathbf{x}_i) = \alpha_{h,0}(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}_i\} \text{ for } h = 1, 2, ..., k\,, \tag{3.13}$$

der $\alpha_{h,0}$ er baseline hasard for dødsårsak h og $\boldsymbol{\beta}$ er vektoren med p regresjonskoeffisienter. For å finne likelihood-funksjonen, må vi skille mellom hendelsestidspunktene for hvert stratum. Hvis vi betegner de observerte tidspunktene i stratum h som $\tilde{T}_{h1}, \tilde{T}_{h2}, ...$, blir den partielle likelihood-funksjonen

$$L(\boldsymbol{\beta}) = \prod_{h=1}^{k}\prod_{i:D_i=h}\frac{Y_i(\tilde{T}_{hi})\exp\{\boldsymbol{\beta}^T\mathbf{x}_i\}}{\sum_{l=1}^{n}Y_l(\tilde{T}_{hi})\exp\{\boldsymbol{\beta}^T\mathbf{x}_l\}}. \tag{3.14}$$

Dette er et produkt over de observerte hendelsestidspunktene for alle dødsårsakene.

I modellen (3.13) har kovariatene samme effekt for alle dødsårsakene. Vi kan la kovariatene variere for hver dødsårsak, og bruker da *årsaksspesifikke* kovariater for individ i. For dødsårsak h blir regresjonskoeffisientene

$$\boldsymbol{\beta}_h = (\beta_{h1}, ..., \beta_{hp})^T. \tag{3.15}$$

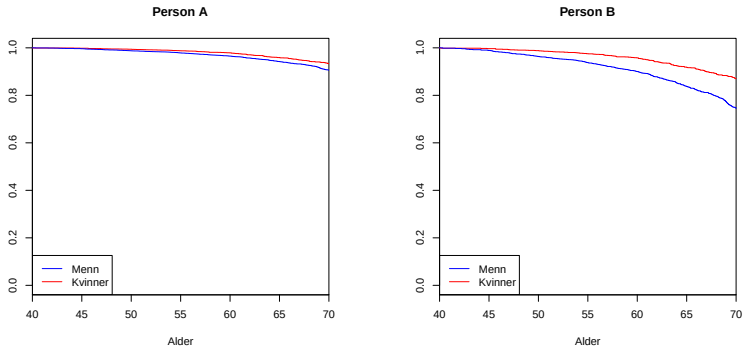
Siden vil tillater de k dødsårsakene å ha ulike baseline og kovariater, tilsvarer dette å tilpasse k forskjellige Cox-modeller. Vi setter $\boldsymbol{\beta}_h$ inn i modellen (3.13), og den årsaksspesifikke Cox-modellen for dødsårsak h blir

$$\alpha_h(t|\mathbf{x}_i) = \alpha_{h,0}(t)\exp\{\boldsymbol{\beta}_h^T\mathbf{x}_i\} \text{ for } h = 1, 2, ..., k\,, \tag{3.16}$$

der $\mathbf{x}_i$ er de opprinnelige kovariatene for individ i. I denne modellen tilpasser vi en baseline for hver overgang, og hver kovariat kan ha forskjellig effekt på overgangene til de k dødsårsakene. Den partielle likelihood-funksjonen for dødsårsak h er på samme form som (3.6), og vi får den ved å erstatte $\boldsymbol{\beta}$ med $\boldsymbol{\beta}_h$.

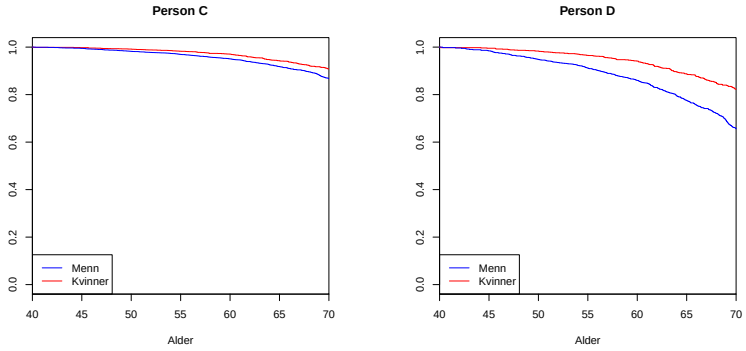
3.2.1 Estimater

Estimatene av de kumulative baseline hasardene og overgangssannsynlighetene beregnes på tilsvarende måte som i avsnitt 2.2.3, men her må vi i tillegg spesifisere kovariatene.



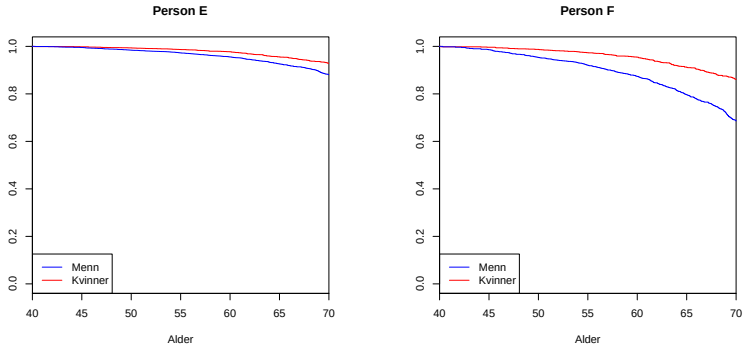
(a)

(b)



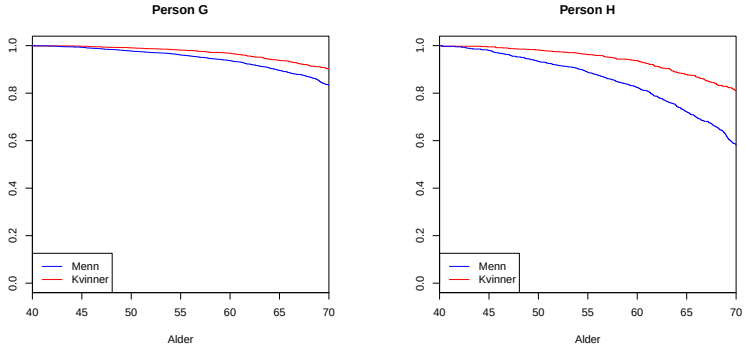
(c)

(d)



(e)

(f)



(g)

(h)

Figur 3.2: Estimert overlevelse fra 40 år for menn og kvinner med kombinasjonene av kovariatene gitt i Tabell 3.4.

Til å estimere kumulativ baseline hasard for en bestemt dødsårsak, kan Breslow-estimatoren (3.9) brukes. Estimatoren til baseline hasard for årsak h på tid t er

$$\widehat{A}_{h,0}(t) = \sum_{i:\widetilde{T}_i \leq t, D_i = h} \frac{1}{\sum_{l=1}^{n} Y_l(\widetilde{T}_i) \exp(\widehat{\boldsymbol{\beta}}_h^T \mathbf{x}_l)}. \tag{3.17}$$

Tilsvarende (3.10), blir estimatoren for kumulativ hasardrate for dødsårsak h i modell (3.16) lik

$$\widehat{A}_h(t|\mathbf{x}_0) = \exp(\widehat{\boldsymbol{\beta}}_h^T \mathbf{x}_0) \widehat{A}_{h,0}(t). \tag{3.18}$$

Estimatoren for overlevelse mellom tid s og t for $s < t$ blir

$$\widehat{P}_{00}(s,t|\mathbf{x}_0) = \prod_{i:s<\widetilde{T}_i \leq t, D_i \neq 0} \{1 - \sum_{h=1}^{k} \Delta \widehat{A}_h(\widetilde{T}_i|\mathbf{x}_0)\}, \tag{3.19}$$

der $\Delta \widehat{A}_h(t|\mathbf{x}_0)$ er inkrementet i estimatoren (3.18) for dødsårsak h på tid t. De ordnede observerte levetidene skrives som $\widetilde{T}_{(1)} \leq \widetilde{T}_{(2)} \leq ... \leq \widetilde{T}_{(n)}$, og de tilhørende indikatorvariablene som $D_{(1)}, D_{(2)}, ..., D_{(n)}$. Den estimerte sannsynligheten for å få en overgang i tidsintervallet $(s,t]$ er på samme form som (2.26);

$$\widehat{P}_{0h}(s,t|\mathbf{x}_0) = \sum_{i:s<\widetilde{T}_{(i)} \leq t, D_{(i)} \neq 0} \widehat{P}_{00}(s, \widetilde{T}_{(i-1)}|\mathbf{x}_0) \Delta \widehat{A}_h(\widetilde{T}_i|\mathbf{x}_0). \tag{3.20}$$

Dette er estimatoren for den kumulative insidensfunksjonen med kovariater $\mathbf{x}_0$.

Eksempel 3.3. Her skal vi tilpasse en stratifisert Cox-modell for konkurrerende dødsårsaker. For hver dødsårsak skal vi estimere overgangs-spesifikke effekter av kovariatene i modellen (3.16). F.eks. er det rimelig å tenke at BMI har større effekt på risikoen for død av hjerte- og karsykdommer enn den har på kreft. Personene grupperes inn etter de fire dødsårsakene

1. Kreft
2. Hjerte- og karsykdommer, inkludert plutselig død
3. Andre medisinske årsaker
4. Alkoholmisbruk, kronisk leversykdom og ulykker og vold

Tabell 3.5 viser antall døde for hver dødsårsak delt inn etter kjønn. Dette er samme oversikt som i Tabell 1.1, men her har vi ekskludert de som røykte pipe eller sigar. I tabellene 3.6 og 3.7 har vi delt inn antall dødsfall etter de fire dødsårsakene og tilhørighet til BMI- og røykegruppe for hhv menn og kvinner. For å få en oversikt over hvordan fordelingen av blodtrykk er, har vi i disse tabellene gruppert den også. I kombinasjonen mellom BMI og død forårsaket av hjerte- og karsykdommer for menn, så var det flest under kategorien overvektig. Blant kvinnene som døde av kreft, hadde de fleste normal vekt. For hjerte- og karsykdommer er det flest dødsfall for personer med høyt blodtrykk.

Tabell 3.5: Antall døde av hver dødsårsak for menn og kvinner.

Dødsårsak	Menn	Kvinner
1	123	87
2	175	53
3	30	34
4	48	12
Totalt	376	186

Tabell 3.6: Oversikt over antall dødsfall blant menn for hver av de fire dødsårsakene, delt inn etter kovariatene.

Variabel	Antall	Dødsårsak 1	Dødsårsak 2	Dødsårsak 3	Dødsårsak 4
Blodtrykk					
< 125	487	32	24	12	8
125-134	549	29	37	5	13
135-144	467	29	40	5	6
≥ 145	512	33	74	8	21
BMI					
Undervektig	50	2	9	2	2
Normal vekt	992	63	65	15	13
Overvektig	835	53	86	11	29
Fedme	98	2	12	1	3
Røykegruppe					
Aldri røykt	444	13	15	5	9
Tidligere røyker	675	32	52	9	8
1-9 sigaretter per dag	199	14	29	2	6
10-19 sigaretter per dag	457	38	50	10	19
20+ sigaretter per dag	240	26	29	4	6

Tabell 3.7: Oversikt over antall dødsfall blant kvinner for hver av de fire dødsårsakene, delt inn etter kovariatene.

Variabel	Antall	Dødsårsak 1	Dødsårsak 2	Dødsårsak 3	Dødsårsak 4
Blodtrykk					
< 125	765	36	10	12	5
125-134	459	16	6	6	2
135-144	321	16	11	5	2
≥ 145	365	19	26	11	3
BMI					
Undervektig	98	4	2	7	2
Normal vekt	1035	42	21	14	4
Overvektig	557	24	19	7	3
Fedme	195	12	11	5	3
Røykegruppe					
Aldri røykt	947	31	18	13	4
Tidligere røyker	333	16	8	6	1
1-9 sigaretter per dag	225	14	10	6	3
10-19 sigaretter per dag	340	20	16	5	1
20+ sigaretter per dag	65	6	1	4	3

Vi skal videre estimere regresjonskoeffisientene. Referansepersonene er de samme som i eksemplene 3.1 og 3.2. Vi lager én tabell for hver dødsårsak, så det blir åtte tabeller totalt. Resultatene fra regresjonsanalysen for menn er vist i tabellene 3.8, 3.9, 3.10 og 3.11. I Tabell 3.8 ser vi at røyking er den eneste kovariaten som har virkning på dødsintensiteten for kreft. Effekten av røyking er ikke signifikant for tidligere røykere, men har en klar effekt på dødeligheten for de resterende røykegruppene. Forholdet mellom hasardratene for en som røyker mer 20+ sigaretter per dag og ikke-røyker er omtrent 4.

For hjerte- og karsykdommer, oppsummert i Tabell 3.9, er det normalt at blodtrykk påvirker død av denne årsaken. For hver 10. enhets økning av blodtrykk, øker dødsintensiteten av hjerte- og karsykdommer med 33%. Av BMI-gruppene er det gruppen undervektig som har en klar signifikant virkning på dødeligheten. Røyking har tydelig effekt her også.

Tabell 3.8: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *kreft* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.069	1.072	0.058	0.231
BMI				
Undervektig	-0.337	0.714	0.719	0.639
Overvektig	-0.011	0.989	0.189	0.955
Fedme	-1.226	0.294	0.721	0.089
Røykegruppe				
Tidligere røyker	0.423	1.527	0.329	0.198
1-9 sigaretter per dag	0.928	2.530	0.386	0.016
10-19 sigaretter per dag	1.092	2.980	0.323	0.001
20+ sigaretter per dag	1.353	3.869	0.345	0.000

Tabell 3.9: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *hjerte- og karsykdommer inkludert plutselig død* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.282	1.326	0.042	0.000
BMI				
Undervektig	1.128	3.089	0.358	0.002
Overvektig	0.332	1.393	0.167	0.047
Fedme	0.311	1.365	0.317	0.327
Røykegruppe				
Tidligere røyker	0.771	2.162	0.294	0.009
1-9 sigaretter per dag	1.536	4.645	0.321	0.000
10-19 sigaretter per dag	1.262	3.531	0.296	0.000
20+ sigaretter per dag	1.449	4.260	0.318	0.000

Fra Tabell 3.10, som er for andre medisinske årsaker, ser vi at ingen av kovariatene er signifikante. Blant menn er det færrest dødsfall her. For siste dødsårsak, som er alko-

holmisbruk, kronisk leversykdom og ulykker og vold, er det kun BMI-gruppen overvektig som har signifikant effekt. I forhold til en mann med normal vekt, er dødeligheten i denne gruppen 2.7 ganger høyere.

Tabell 3.10: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *andre medisinske årsaker* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.051	1.053	0.118	0.663
BMI				
Undervektig	1.070	2.915	0.755	0.157
Overvektig	-0.140	0.869	0.402	0.728
Fedme	-0.463	0.629	1.040	0.656
Røykegruppe				
Tidligere røyker	0.406	1.500	0.602	0.500
1-9 sigaretter per dag	0.163	1.178	0.867	0.850
10-19 sigaretter per dag	0.915	2.497	0.592	0.122
20+ sigaretter per dag	0.721	2.057	0.708	0.308

Tabell 3.11: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *alkoholmisbruk, kronisk leversykdom og ulykker og vold* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.118	1.125	0.089	0.188
BMI				
Undervektig	1.123	3.075	0.760	0.140
Overvektig	0.994	2.701	0.338	0.003
Fedme	0.753	2.123	0.648	0.246
Røykegruppe				
Tidligere røyker	-0.665	0.514	0.505	0.187
1-9 sigaretter per dag	0.556	1.743	0.528	0.293
10-19 sigaretter per dag	0.839	2.313	0.405	0.039
20+ sigaretter per dag	0.351	1.421	0.527	0.506

Som tidligere skrevet, er kreft hovedårsaken til død blant kvinner. Røyking har en signifikant effekt på død av kreft også for kvinner. Dette kan vi se fra Tabell 3.12. Forholdet mellom hasardratene for en som røyker 10-19 sigaretter per dag og en som ikke røyker er ca. 2. For dødsårsaken hjerte- og karsykdommer har blodtrykk omtrent samme effekt for begge kjønn. For andre medisinske årsaker er det stor forskjell mellom hasard ratioene og p-verdiene for kvinner og menn. En kvinne som er undervektig har 5.8 ganger høyere risiko for død av denne årsaken i forhold til en kvinne med normal vekt. For dødsårsakene alkoholmisbruk, kronisk leversykdom og ulykker og vold, så er estimert HR for siste røykegruppe ca. 14.

I Figur 3.3 har vi plottet de estimerte kumulative insidensfunksjonene for de åtte kombina-

Tabell 3.12: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *kreft* for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.007	1.007	0.063	0.907
BMI				
Undervektig	-0.041	0.960	0.525	0.938
Overvektig	0.070	1.072	0.260	0.788
Fedme	0.508	1.663	0.339	0.133
Røykegruppe				
Tidligere røyker	0.488	1.629	0.315	0.121
1-9 sigaretter per dag	0.747	2.110	0.328	0.023
10-19 sigaretter per dag	0.700	2.014	0.305	0.022
20+ sigaretter per dag	1.336	3.802	0.455	0.003

Tabell 3.13: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *hjerte- og karsykdommer inkludert plutselig død* for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.284	1.328	0.045	0.000
BMI				
Undervektig	0.096	1.101	0.744	0.897
Overvektig	0.263	1.301	0.326	0.419
Fedme	0.785	2.193	0.377	0.037
Røykegruppe				
Tidligere røyker	0.221	1.247	0.432	0.609
1-9 sigaretter per dag	0.754	2.126	0.402	0.060
10-19 sigaretter per dag	1.122	3.070	0.348	0.001
20+ sigaretter per dag	0.371	1.449	1.034	0.720

Tabell 3.14: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *andre medisinske årsaker* for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.256	1.292	0.065	0.000
BMI				
Undervektig	1.757	5.797	0.473	0.000
Overvektig	-0.324	0.723	0.472	0.492
Fedme	0.426	1.530	0.529	0.421
Røykegruppe				
Tidligere røyker	0.238	1.269	0.507	0.639
1-9 sigaretter per dag	0.563	1.755	0.506	0.266
10-19 sigaretter per dag	0.122	1.129	0.540	0.822
20+ sigaretter per dag	1.745	5.725	0.598	0.003

Tabell 3.15: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en Cox-modell med død grunnet *alkoholmisbruk, kronisk leversykdom og ulykker og vold* for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.033	1.033	0.171	0.849
BMI				
Undervektig	1.548	4.701	0.872	0.076
Overvektig	0.339	1.403	0.773	0.661
Fedme	1.649	5.200	0.810	0.042
Røykegruppe				
Tidligere røyker	-0.359	0.698	1.120	0.748
1-9 sigaretter per dag	1.141	3.131	0.768	0.137
10-19 sigaretter per dag	-0.271	0.762	1.128	0.810
20+ sigaretter per dag	2.631	13.885	0.803	0.001

sjonene av kovariatene i Tabell 3.4 for menn mellom 40 og 70 år. Røyking øker dødssannsynligheten for alle årsakene. Hvis vi sammenligner personene A (ikke-røyker) og B (10-19 sigaretter pr dag), som begge har blodtrykk 125 og normal vekt, så er sannsynligheten for å dø av kreft ved 70 år 0.05 for A og 0.1 for B. Sannsynligheten til død av hjerte- og karsykdommer ved 70 år er 0.03 og 0.08 for hhv A og B. For begge disse dødsårsakene har dødssannsynligheten blitt omtrent doblet. For personer som har høyt blodtrykk eller er overvektige, har røyking enda større effekt.

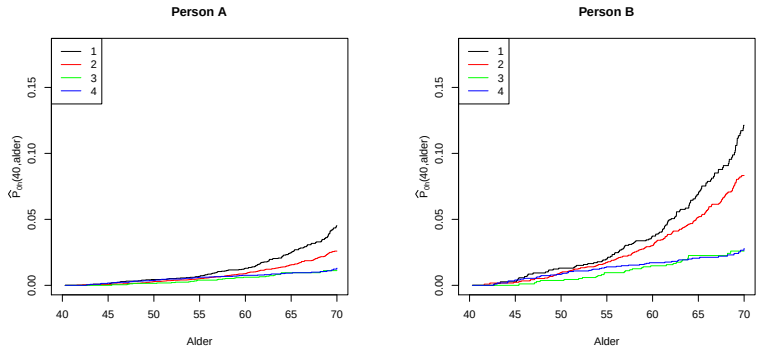
Sannsynligheten til dødsårsaken alkohol, kronisk leversykdom og ulykker og vold er høyere for menn som er overvektige i forhold til de som har normal vekt. F.eks. er estimert sannsynlighet for person B (normal vekt) omtrent 0.03 ved 70 år, mens den er 0.07 for person F (overvektig). For de resterende dødsårsakene, har denne variabelen mindre betydning. Effekten av økt blodtrykk påvirker mest død av hjerte- og karsykdommer.

Hvis vi ser på de kumulative insidensfunksjonene for kvinner, som er vist i Figur 3.4, så er de generelt lavere enn for menn. Alkoholmisbruk, kronisk leversykdom og ulykker og vold har veldig lave sannsynligheter for alle de åtte personene i Tabell 3.4. Forholdet mellom sannsynlighetene for de resterende tre dødsårsakene er omtrent det samme som vi observerte hos menn. Uten kovariater så var det kreft som hadde høyest kumulative insidensfunksjoner, mens for menn så var det hjerte- og karsykdommer, som er vist i Figur 2.4. Når vi tar med effekten av kovariatene, er det dermed ikke så stor forskjell mellom sammensetningen av dødsårsakene for menn og kvinner.

3.3 Programvare

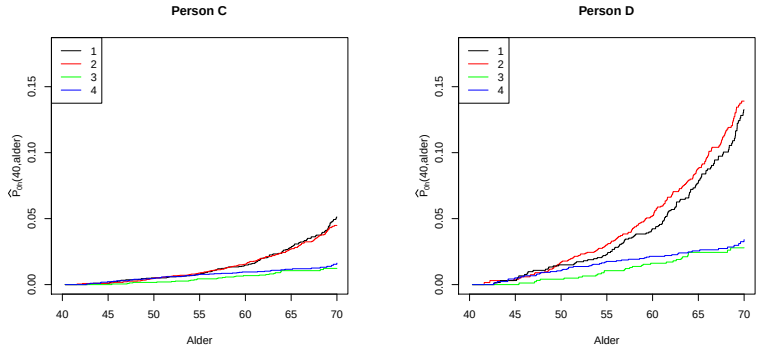
Her skal vi beskrive hvordan estimatene i Eksempel 3.3 er beregnet. Dette er en fortsettelse av avsnitt 2.4 om `mstate`-pakken i **R**, og er basert på avsnitt 4 i Wreede *et al.* (2011) og Putter (2011).

Først estimerer vi regresjonskoeffisientene for hver dødsårsak i modellen (3.16). Som tid-



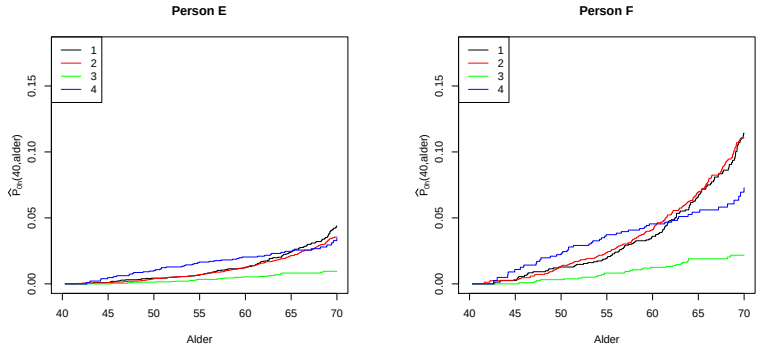
(a)

(b)



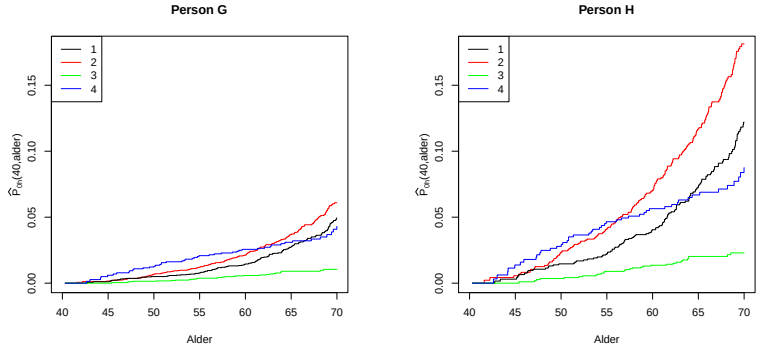
(c)

(d)



(e)

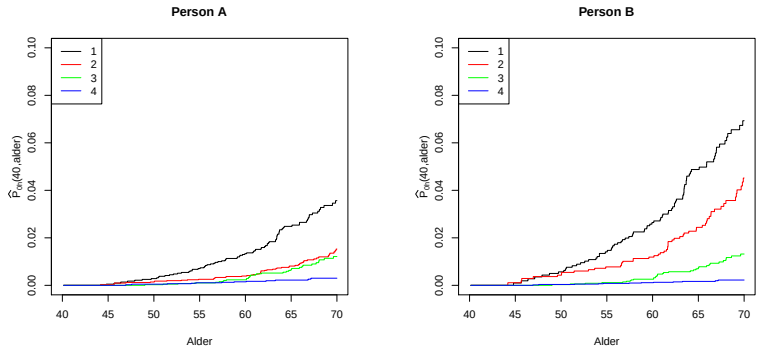
(f)



(g)

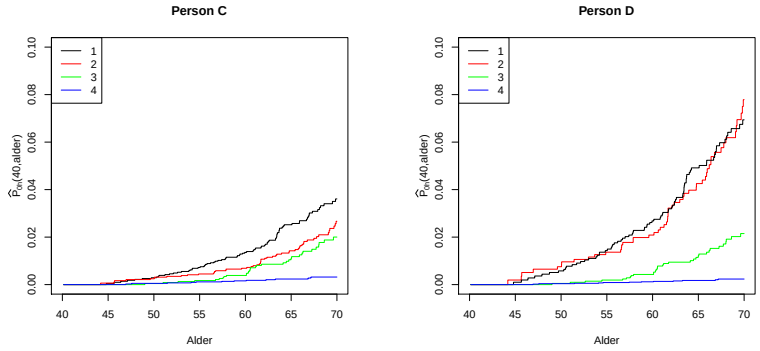
(h)

Figur 3.3: Estimater av de kumulative insidensfunksjonene med de fire dødsårsakene beskrevet i Eksempel 3.3 for menn.



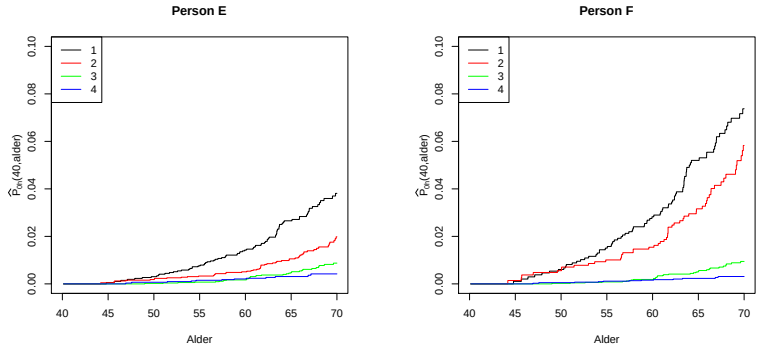
(a)

(b)



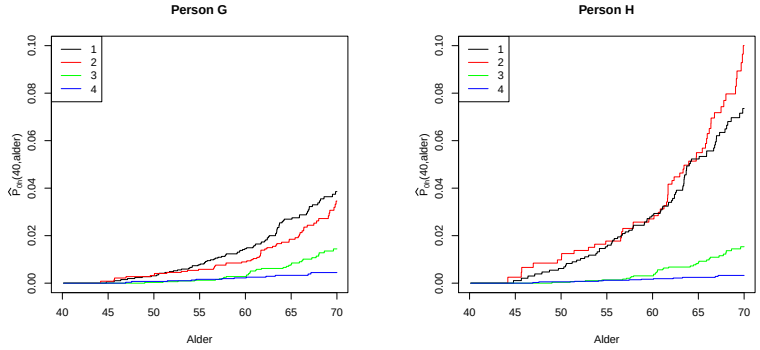
(c)

(d)



(e)

(f)



(g)

(h)

Figur 3.4: Estimater av de kumulative insidensfunksjonene med de fire dødsårsakene beskrevet i Eksempel 3.3 for kvinner.

ligere beskrevet, må vi ha dataene på long-format. Vi legger til de årsaksspesifikke kovariatene i datasettet, og alle regresjonskoeffisientene i de k modellene estimeres samtidig. For dødsårsak h og person i innfører vi vektoren $\mathbf{x}_{ih}$ med lengde $k \times p$, som består av de opprinnelige kovariatene fra indeks h til indeks $h + p$ og 0 ellers. En utvidet vektor med alle regresjonskoeffisientene blir

$$\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, ..., \boldsymbol{\beta}_k^T)^T, \tag{3.21}$$

der $\boldsymbol{\beta}_h$ er vektoren (3.15), som er regresjonskoeffisientene for dødsårsak h. En alternativ måte å skrive modellen (3.16) på er

$$\alpha_h(t|\mathbf{x}_i) = \alpha_{h,0}(t) \exp\{\boldsymbol{\beta}^{*T} \mathbf{x}_{ih}\}. \tag{3.22}$$

For å legge til årsaksspesifikke kovariater i datasettet, brukes funksjonen `covs.expand()`. Hver kovariat deles inn i antall overganger. For konkurrerende dødsårsaker deles hver kovariat i k nye kovariater. Disse er indikatorvariable for kategoriske kovariater. I vårt eksempel får vi 4×8 nye kovariater. For første person i datasettet har vi dette `msdata`-objektet:

```
  id from to trans Tstart Tstop time status csbp10 bmigr smkgr csbp10.1 csbp10.2 csbp10.3
1  1    1  2     1      40  60.8 20.8      0   -2.5     2     1     -2.5      0.0      0.0
2  1    1  3     2      40  60.8 20.8      0   -2.5     2     1      0.0     -2.5      0.0
3  1    1  4     3      40  60.8 20.8      0   -2.5     2     1      0.0      0.0     -2.5
4  1    1  5     4      40  60.8 20.8      0   -2.5     2     1      0.0      0.0      0.0


  csbp10.4 bmigr1.1 bmigr1.2 bmigr1.3 bmigr1.4 bmigr3.1 bmigr3.2 bmigr3.3 bmigr3.4
1      0.0        0        0        0        0        0        0        0        0
2      0.0        0        0        0        0        0        0        0        0
3      0.0        0        0        0        0        0        0        0        0
4     -2.5        0        0        0        0        0        0        0        0


  bmigr4.1 bmigr4.2 bmigr4.3 bmigr4.4 smkgr2.1 smkgr2.2 smkgr2.3 smkgr2.4 smkgr3.1
1        0        0        0        0        0        0        0        0        0
2        0        0        0        0        0        0        0        0        0
3        0        0        0        0        0        0        0        0        0
4        0        0        0        0        0        0        0        0        0


  smkgr3.2 smkgr3.3 smkgr3.4 smkgr4.1 smkgr4.2 smkgr4.3 smkgr4.4 smkgr5.1 smkgr5.2
1        0        0        0        0        0        0        0        0        0
2        0        0        0        0        0        0        0        0        0
3        0        0        0        0        0        0        0        0        0
4        0        0        0        0        0        0        0        0        0


  smkgr5.3 smkgr5.4
1        0        0
2        0        0
3        0        0
4        0        0
```

Denne personen er en kvinne som tilhører BMI gruppen normal, har blodtrykk 110 og er ikke-røyker. Siden BMI-gruppen og røykegruppen hun tilhører er referansegrupper, er de ikke med blant kovariatene.

Videre i Eksempel 3.3 har vi tilpasset modellen (3.16), som er en stratifisert Cox-modell med årsaksspesifikke overganger. Vi må oppgi alle de årsaksspesifikke kovariatene når funksjonen `coxph()` brukes. For å tilpasse en baseline for hver overgang, må vi i tillegg til kovariatene ha med `strata(trans)`, der `trans` er overgangsmatrisen.

Tilsvarende avsnitt 2.4, må vi først estimere overgangs-intensitetene med funksjonen `msfit()`. Men her må de estimeres de for gitte verdier av kovariatene. Vi kan bestemme verdiene av kovariatene enten ved å bruke `newdata`, eller velge en person i datasettet. Til slutt brukes dette objektet i `probtrans()`, som beregner de estimerte overgangssannsynlighetene for denne personen, dvs. elementene i matrisen (3.20).

Kapittel 4

Modellering med subfordelinger

I tilfellet med kun én dødsårsak, så er det en direkte kobling mellom hasardraten $\alpha(t|\mathbf{x})$ og dødssannsynligheten $1 - S(t|\mathbf{x})$. En kovariat påvirker hasardraten og dødssannsynligheten i samme retning. Dersom f.eks. økning av en kovariat gir høyere hasardrate, vil også dødssannsynligheten øke. For konkurrerende dødsårsaker har vi ikke lenger en slik sammenheng. Fra (3.20) ser vi at den kumulative insidensfunksjonen for dødsårsak h er en funksjon av hasardratene for alle dødsårsakene. Dermed kan en kovariat ha ulik effekt på den årsaksspesifikke hasardraten og den kumulative insidensfunksjonen for samme dødsårsak. Et alternativ til de årsaksspesifikke hasardratene er metoden i Fine & Gray (1999), som er modellering med *subfordelingshasarder*. I neste avsnitt beskriver vi denne metoden. Tilsvarende Kapittel 3, skal vi anta at kovariatene ikke er tidsavhengige.

4.1 Subfordelingshasardrate uten kovariater

4.1.1 Begreper

Vi bruker samme notasjon som i de tidligere kapitlene, og betegner levetiden med den stokastiske variabelen T. Vi lar variabelen D betegne dødsårsaken. Med k konkurrerende dødsårsaker, kan vi ha $D = 1, ..., k$ ved død. For dødsårsak h har vi i intervallet $(0, t]$ årsaksspesifikk kumulativ insidensfunksjon

$$F_h(t) = P(T \leq t, D = h) = P_{0h}(0, t). \tag{4.1}$$

Dette er sannsynligheten for å dø av årsak h innen tid t. Fine & Gray (1999) innførte en metode som gjenoppretter en-til-en forholdet mellom effekten av kovariatene for hasardrate og kumulativ insidensfunksjon. Det går ut på modellering av *subfordelingshasard*

$$\alpha_h^{sub}(t) = \lim_{\Delta t \to 0} \frac{P\{t \leq T < t + \Delta t, D = h \mid T \geq t \cup (T < t \cap D \neq h)\}}{\Delta t}. \tag{4.2}$$

Forskjellen mellom denne og den årsaksspesifikke hasardraten (2.15) er i definisjonen av risikomengden. Dersom en person dør av en annen årsak enn h før tiden t, så beholdes

personen i risikomengden for denne dødsårsaken. På tid t består dermed risikomengden av personene som lever, og de som har dødd av andre dødsårsaker enn h før tid t. Dette gjør at subfordelingshasardene blir vanskeligere å tolke. Vi kan alternativt definere subfordelingshasarden for årsak h som den vanlige hasarden for den stokastiske variabelen

$$T^* = T \times I(D = h) + \infty \times I(D \neq h). \tag{4.3}$$

Subfordelingshasarden (4.2) kan omformuleres til

$$\alpha_h^{sub}(t) = \frac{\mathrm{d}F_h(t)/\mathrm{dt}}{1 - F_h(t)} = -\frac{\mathrm{d}}{\mathrm{dt}}\log\{1 - F_h(t)\}. \tag{4.4}$$

Fra (4.4) følger det at den kumulative insidensfunksjonen for dødsårsak h blir

$$F_h(t) = 1 - \exp\left(-\int_0^t \alpha_h^{sub}(u)du\right), \tag{4.5}$$

som er på samme form som $1 - S(t)$ fra avsnitt 2.1.1. Vi ser at den kumulative insidensfunksjonen for dødsårsak h bare er en funksjon av subfordelingshasarden for denne årsaken, og ikke for de konkurrerende dødsårsakene.

Ikke alle levetider observeres fra start til slutt. Vi kan ha både høyre-sensurering og venstre-trunkering. Vi betegner sensureringstiden med C. Videre lar vi L være trunkeringstid. Gitt at individet kommer under observasjon, blir observert levetid $\widetilde{T} = min(T, C)$. Vi antar også her at T er uavhengig av C og L, og at C og L er uavhengige av dødsårsaken. Fordelingen til C og L betegnes med hhv G og H. I Geskus (2011) defineres $\overline{G}(t) = 1-G(t)$, som er sannsynligheten for at det ikke er høyre-sensurering i tidsrommet $(0, t]$.

Hasardraten (4.2) gjelder kun for data med fullstendige levetider. I de neste avsnittene, som er basert på Geskus (2011), skal vi utvide dette til å gjelde for både høyre-sensurering og venstre-trunkering. Vi skal forklare hvordan vekter brukes for å kompensere for levetider som kan være både sensurerte og trunkerte. Vi skal ikke presentere bevisene i artikkelen, men gi intuitive forklaringer til estimatorene (Borgan, privat meddelelse).

4.1.2 Høyre-sensurering

Estimeringen av parametrene for subfordelingshasarder med høyre-sensurerte data avhenger av hva slags type høyre-sensurering det er. Ved *administrativ sensurering*, dvs. når den eneste årsaken til sensurering er at studien avsluttes før individet har dødd, så er den potensielle sensureringstiden C kjent. Personer som dør av andre årsaker enn h regnes som under risiko fram til den administrative sensureringstiden da disse personene fjernes fra risikomengden. Et individ er altså under risiko på tid t hvis $\widetilde{T}$ er større eller lik t, eller hvis det dør av en annen årsak før tid t og $C \geq t$. Risikoindikatoren blir dermed

$$I\{\widetilde{T} \geq t \cup (T < t \cap D \neq h, C \geq t)\} = I(\widetilde{T} \geq t) + I(T < t, D \neq h, C \geq t). \tag{4.6}$$

For tilfeldig høyre-sensurering, er ikke det andre leddet i (4.6) kjent. Fine og Gray foreslo bruk av en metode som går ut på vekting (IPCW; *inverse probability of censoring weighting*). Individene som dør av andre årsaker enn h forblir i risikomengden, men med vekter

i intervallet $(0, 1)$ som avhenger av hendelsestidspunktet og fordelingen til sensureringen. Vi vil se hvordan disse vektene skal være.

Gitt informasjonen vi har om levetiden, kan vi finne den betingede forventingen

$$E\{I(T < t, D \neq h, C \geq t) \,|\, T, D \neq h, C \geq T\}. \tag{4.7}$$

Vi kan nå skrive (4.7) som

$$\begin{aligned} E\{I(T < t, D \neq h)I(C \geq t) \,|\, T, D \neq h, C \geq T\} &= I(T < t, D \neq h)E\{I(C \geq t) \,|\, T, C \geq T\} \\ &= I(T < t, D \neq h)P\{C \geq t \,|\, T, C \geq T\} \\ &= I(T < t, D \neq h)\frac{P(C \geq t)}{P(C \geq T \,|\, T)} \\ &= I(T < t, D \neq h)\frac{\overline{G}(t)}{\overline{G}(T)}. \end{aligned} \tag{4.8}$$

For død av årsak $D \neq h$ på tid $T < t$, har vi på tid t dermed vekten

$$\omega_C(t) = \frac{\overline{G}(t)}{\overline{G}(T)}. \tag{4.9}$$

Den kompenserer for mulig høyre-sensurering av individer som har dødd av konkurrerende årsaker. En intuitiv tolkning av denne vekten er altså at personer som dør av andre årsaker potensielt kunne ha blitt sensurert, så risikomengden blir mindre med disse vektene.

4.1.3 Venstre-trunkering

Her skal vi se på levetider som kan være venstre-trunkerte, men ikke sensurerte. Ved venstre-trunkering kommer individene under observasjon kun hvis levetiden er større enn trunkeringstiden. Risikoindikatoren på tid t for et individ kan formuleres som

$$I\{(L < t \leq T) \cup (L < T < t, D \neq h)\} = I(L < t \leq T) + I(L < T < t, D \neq h). \tag{4.10}$$

På samme måte som i tilfellet for sensurering, finner vi den betingede forventningen for det andre leddet i (4.10). Her betinger vi på at individet har kommet under observasjon til tid t;

$$E\{I(L < T < t, D \neq h) \,|\, T, D \neq h, L < t\}. \tag{4.11}$$

Vi ser videre på dette uttrykket, og får

$$\begin{aligned} E\{I(T < t, D \neq h)I(L < T) \,|\, T, D \neq h, L < t\} &= I(T < t, D \neq h)P(L < T \,|\, T, L < T) \\ &= I(T < t, D \neq h)\frac{P(L < T \,|\, T)}{P(L < t)} \\ &= I(T < t, D \neq h)\frac{H(T)}{H(t)}. \end{aligned} \tag{4.12}$$

Fra (4.12) ser vi at indikatorvariabelen er multiplisert med en faktor. For å kompensere for dette, brukes den inverse som vekt. På tid t for hendelsestidspunktet $T < t$ har vi

$$\omega_L(t) = \frac{H(t)}{H(T)}. \tag{4.13}$$

For $T < t$ er $H(T) < H(t)$, så vekten (4.13) blir større enn 1. Individer som dør av andre årsaker enn h før trunkeringstiden skulle teoretisk sett ha vært med i risikomengden. Den kompenserer vi for ved å la individene som dør av konkurrerende årsaker etter trunkeringstiden få disse vektene.

4.1.4 Data

Som i de tidligere kapitlene lar vi n være antall individer i studien, og de observerte levetidene $\widetilde{T}_i = min(T_i, C_i)$ for $i = 1, 2, ..., n$. De observerte tidene for venstre-trunkering betegnes med $L_1, L_2..., L_n$. Vi antar at de observerte sensureringstidene er distinkte, dvs. at levetidene til to individer ikke sensureres på samme tid, og tilsvarende for trunkering. I Geskus (2011) håndteres dette ved å innføre notasjonen m_i for antall sensureringer på tid C_i, og w_i for antall trunkeringer på tid L_i. For hvert individ har vi i tillegg variabelen $D_i \in \{0, 1, 2, ..., k\}$ som viser om levetiden er sensurert ($D_i = 0$) eller dødsårsaken ($D_i \neq 0$). Som tidligere beskrevet, er $Y_i(t)$ den tradisjonelle risikoindikatoren på tid t for individ i;

$$Y_i(t) = I(L_i < t \leq \widetilde{T}_i). \tag{4.14}$$

4.1.5 Estimering med høyre-sensurerte og venstre-trunkerte data

Før vi innfører de tidsavhengige vektene som justerer for høyre-sensurering og venstre-trunkering, skriver vi hvordan fordelingene til sensurering og trunkering estimeres. Estimatoren for $\overline{G}(t)$ er i Geskus (2011) definert som

$$\widehat{\overline{G}}(t) = \prod_{i:\widetilde{T}_i \leq t, D_i = 0} \left(1 - \frac{1}{Y(\widetilde{T}_i)}\right), \tag{4.15}$$

der $Y(t) = \sum_{l=1}^{n} Y_l(t)$. Ved å multiplisere over $D_i = 0$ istedenfor $D_i = 1$ i Kaplan-Meier estimatoren (2.13) får vi (4.15). Estimatoren for $H(t)$ oppnås på tilsvarende vis som (4.15);

$$\widehat{H}(t) = \prod_{i:L_i > t} \left(1 - \frac{1}{Y(L_i)}\right). \tag{4.16}$$

Dette er estimert sannsynlighet for å delta i studien innen tid t.

Personer som dør av konkurrerende årsaker før t får vekter som er produktet av de estimerte vektene for sensurering (4.9) og trunkering (4.13). For et individ l er vektfunksjonen i Geskus (2011) definert som

$$\omega_l(t) = \begin{cases} 1 & \text{hvis } L_i < t \leq \widetilde{T}_i, \\ \frac{\widehat{\overline{G}}(t)}{\widehat{\overline{G}}(\widetilde{T}_j)} \frac{\widehat{H}(t)}{\widehat{H}(\widetilde{T}_j)} & \text{hvis død av en annen årsak enn } h \\ & \text{på tid } \widetilde{T}_j < t, \\ 0 & \text{ellers.} \end{cases} \tag{4.17}$$

Estimert antall individer som er under risiko rett før tid blir summen av vektene (4.17) for alle individene :

$$Y^*(t) = \sum_{l=1}^{n} \omega_l(t). \tag{4.18}$$

Risikomengden justeres slik at man tar hensyn til høyre-sensurering og venstre-trunkering. Vi antar som tidligere at kun én hendelse kan inntreffe på et bestemt tidspunkt, og ser først på hasardrate uten kovariater. Estimatoren for den kumulative subfordelingshasarden med høyre-sensurering og venstre-trunkering blir

$$\widehat{A}_h^{sub}(t) = \sum_{i:\widetilde{T}_i \le t, D_i = h} \frac{1}{Y^*(\widetilde{T}_i)}. \tag{4.19}$$

Fra Kapittel 2 har vi $\widehat{F}_h(t) = \widehat{P}_{0h}(0,t)$, som er estimatoren (2.26) med $s = 0$. Geskus (2011) innfører estimatoren

$$\widehat{\widehat{F}}_h^{PL}(t) = 1 - \prod_{i:\widetilde{T}_i \le t, D_i = h} \{1 - \Delta \widehat{A}_h^{sub}(\widetilde{T}_i)\}, \tag{4.20}$$

og viser at den er ekvivalent med $\widehat{F}_h(t)$.

4.2 Proporsjonal subfordelingshasard modell

I Fine & Gray (1999) er det foreslått en proporsjonal hasard modell for subfordelingshasardene;

$$\alpha_h^{sub}(t|\mathbf{x}) = \alpha_{h,0}^{sub}(t) \exp\left(\boldsymbol{\beta}_h^T \mathbf{x}\right), \tag{4.21}$$

der $\mathbf{x}$ er vektoren med p kovariater. Den partielle likelihood funksjonen for denne modellen blir på samme form som (3.6);

$$L(\boldsymbol{\beta}) = \prod_{i:D_i = h} \frac{\omega_i(\widetilde{T}_i) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}}{\sum_{l=1}^{n} \omega_l(\widetilde{T}_i) \exp\{\boldsymbol{\beta}^T \mathbf{x}_l\}}. \tag{4.22}$$

Her har vi erstattet $Y_l(t)$ med $\omega_l(t)$ i (3.6). Regresjonskoeffisientene estimeres ved å maksimere (4.22). Det er i Fine & Gray (1999) vist at maksimum likelihood estimatet $\widehat{\boldsymbol{\beta}}$ er multi-normalfordelt. I Geskus (2011) er det videre vist at normalitet for $\widehat{\boldsymbol{\beta}}$ fortsatt gjelder med høyre-sensurerte og venstre-trunkerte data når vektene fra (4.17) brukes.

Fra modellen (4.21) følger det at den kumulative insidensfunksjonen for dødsårsak h med kovariater $\mathbf{x}$ blir

$$F_h(t|\mathbf{x}) = 1 - \exp\{-\int_0^t \alpha_h^{sub}(u|\mathbf{x}) du\}, \tag{4.23}$$

som er på samme form som (2.4). Estimert kumulativ baseline blir

$$\widehat{A}_{h,0}^{sub}(t) = \sum_{i:\widetilde{T}_i \le t, D_i = h} \frac{1}{\sum_{l=1}^{n} \omega_l(\widetilde{T}_i) \exp(\widehat{\boldsymbol{\beta}}_h^T \mathbf{x}_l)}. \tag{4.24}$$

Estimatoren for kumulativ hasardrate for årsak h oppnås ved å sette inn inkrementet til estimatoren (4.24) og de estimerte regresjonskoeffisientene i modellen (4.21). Gitt kovariatene $\mathbf{x}_0$, blir estimert sannsynlighet for overlevelse til tid t lik

$$\widehat{S}(t|\mathbf{x}_0) = \prod_{i:\widetilde{T}_i \leq t, D_i \neq 0} \{1 - \sum_{h=1}^{k} \Delta \widehat{A}_h(\widetilde{T}_i|\mathbf{x}_0)\}. \tag{4.25}$$

Kumulativ insidensfunksjon for årsak h i intervallet $[0, t)$ estimeres med

$$\widehat{F}_h(t|\mathbf{x}_0) = 1 - \prod_{i:\widetilde{T}_i \leq t, D_i = h} \{1 - \Delta \widehat{A}_h(\widetilde{T}_i|\mathbf{x}_0)\}. \tag{4.26}$$

Som tidligere beskrevet, avhenger denne bare av dødsårsak h.

Eksempel 4.1. Her skal vi tilpasse modellen (4.21) for hver av de fire dødsårsakene i datasettet som er beskrevet i Kapittel 1. Dødsårsakene som modelleres er i tabellene og figurene nummerert slik :

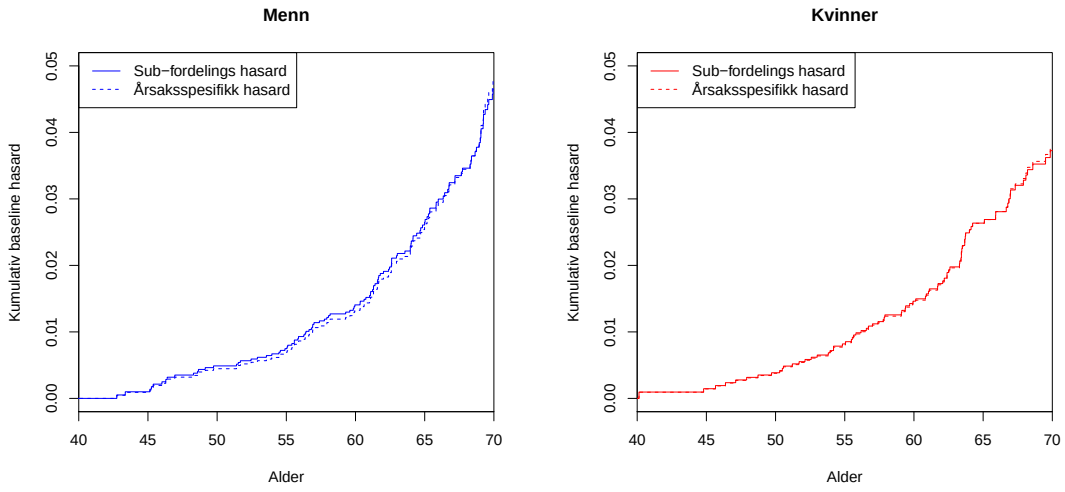
1. Kreft
2. Hjerte- og karsykdommer, inkludert plutselig død
3. Andre medisinske årsaker
4. Alkoholmisbruk, kronisk leversykdom og ulykker og vold

Vi skal først sammenligne baseline subfordelingshasardene med de med de årsaksspesifikke baseline hasardratene fra modellen (3.16). De kumulative baseline hasardene er plottet i figurene 4.1-4.4. For menn ser vi at det er liten forskjell i de to baseline hasardene. For kvinnene er plottene nesten sammenfallende. I Figur 4.4 har baseline hasardene omtrent konstante stigningstall, og de er ca. 0.004 og 0.001 pr 10. år for hhv menn og kvinner.

De estimerte regresjonskoeffisientene i modell (4.21) er lagt ved i Tillegg A.1. Se oppsummering i tabellene A.1-A.4 og A.5-A.8 for hhv menn og kvinner. Som tidligere nevnt, er definisjonen av risikomengde forskjellig for årsaksspesifikk hasard og subfordelingshasard. For sistnevnte er ikke risikomengden naturlig, fordi den ikke reduseres ved død av andre dødsårsaker enn den som modelleres. P-verdiene for Wald-testene er omtrent like for begge metodene. Siden tolkningen av regresjonskoeffisientene er som i Eksempel 3.3, skal vi her fokusere på å finne eventuelle forskjeller mellom koeffisientene for subfordelingshasardene og de tilhørende årsaksspesifikke hasardene, som ligger i tabellene 3.8-3.15.

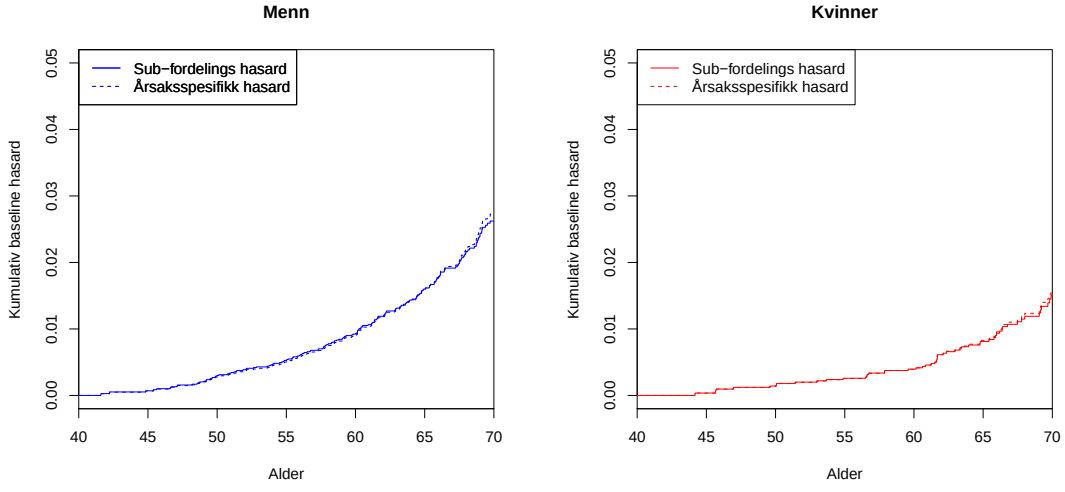
I Tabell 4.1 (menn) og Tabell 4.2 (kvinner) har vi beregnet relativ endring av årsaksspesifikk hasard ratio i modellen (3.16) i forhold til subfordelingshasard ratioene i modellen (4.21). Vi ser først på resultatene for menn. Generelt er de fleste estimerte HR noe høyere med Cox-modellen (3.16) sammenlignet med HR fra modellen (4.21), men det er små endringer. Endringene er størst for røykerne. Her er HR for de årsaksspesifikk hasardene mellom 5% og 11% høyere. Sammenlignet med tilfellet for menn, er de relative endringene lavere for kvinner.

Vi plotter også her de kumulative insidensfunksjonene for personer med kovariater fra Tabell 3.4. De er vist i Tillegg A.2; se figurene A.1 og A.2 for hhv menn og kvinner. Samme



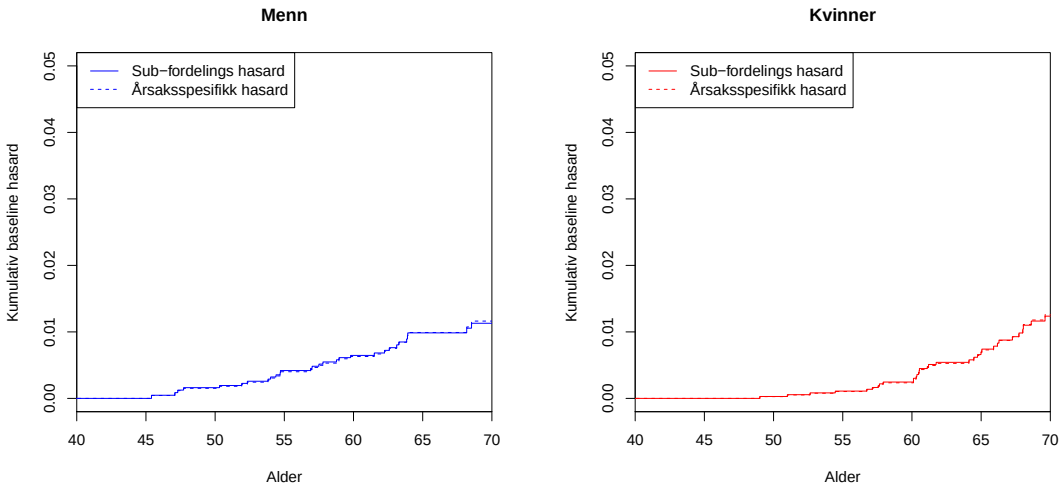
(a) (b)

Figur 4.1: Kumulativ baseline hasard for modellene (4.21) og (3.16) med *kreft* for (a) menn og (b) kvinner.



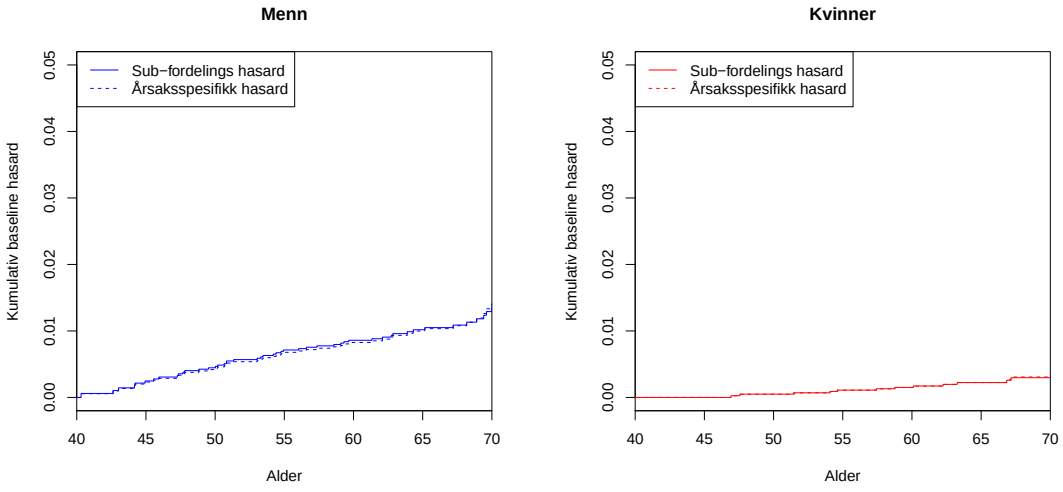
(a) (b)

Figur 4.2: Kumulativ baseline hasard for modellene (4.21) og (3.16) med *hjerte- og karsykdommer inkludert plutselig død* for (a) menn og (b) kvinner.



(a) (b)

Figur 4.3: Kumulativ baseline hasard for modellene (4.21) og (3.16) med *andre medisinske årsaker* for (a) menn og (b) kvinner.



(a) (b)

Figur 4.4: Kumulativ baseline hasard for modellene (4.21) og (3.16) med *alkoholmisbruk, kronisk leversykdom og ulykker og vold* for (a) menn og (b) kvinner.

Tabell 4.1: Relativ endring av årsaksspesifikk hasard ratio i forhold til subfordelingshasard ratio for menn.

Kovariat	Årsak 1	Årsak 2	Årsak 3	Årsak 4
Blodtrykk (per 10)	0.024	0.018	0.014	0.019
BMI				
Undervektig	0.195	0.030	0.106	0.097
Overvektig	0.025	0.002	0.021	-0.014
Fedme	0.007	-0.052	-0.022	-0.055
Røykgruppe				
Tidligere røyker	0.017	-0.012	0.019	0.021
1-9 sigaretter per dag	0.111	0.051	0.113	0.094
10-19 sigaretter per dag	0.091	0.058	0.075	0.063
20+ sigaretter per dag	0.076	0.082	0.097	0.080

Tabell 4.2: Relativ endring av årsaksspesifikk hasard ratio i forhold til subfordelingshasard ratio for kvinner.

Kovariat	Årsak 1	Årsak 2	Årsak 3	Årsak 4
Blodtrykk (per 10)	0.009	0.008	0.014	0.007
BMI				
Undervektig	0.049	0.045	0.043	0.019
Overvektig	0.003	-0.003	0.033	0.032
Fedme	0.015	0.028	0.051	0.012
Røykgruppe				
Tidligere røyker	0.003	-0.004	0.012	0.007
1-9 sigaretter per dag	0.017	-0.026	0.038	-0.008
10-19 sigaretter per dag	0.024	-0.011	0.055	0.025
20+ sigaretter per dag	0.069	0.108	0.096	0.024

skala på y-aksen som figurene i Eksempel 3.3 er brukt. Ved å sammenligne med de kumulative insidensfunksjonene som er plottet i figurene 3.3 og 3.4, så ser vi ikke noen tydelige forskjeller. For å oppnå en mer nøyaktig sammenligning, har vi beregnet relativ endring av de estimerte kumulative insidensfunksjonene (3.20) for årsaksspesifikke hasardrater i forhold til de tilsvarende for subfordelingshasardene med de åtte kombinasjonene av kovariatene som er vist i Tabell 3.4 ved alder 50, 60 og 70 år. Disse er oppsummert i Tabell 4.3 (menn) og Tabell 4.4 (kvinner).

For de fleste personene er det negative endringer, dvs. at de estimerte dødssannsynlighetene er noe høyere med subfordelingshasardene. Relativt sett er det større forskjeller ved alderen 50 enn 60 og 70 år. Dette skyldes at det er lave dødssannsynligheter, slik at små endringer gir stort utslag. Blant kvinner er det størst endring for dødsårsakene alkoholmisbruk, kronisk leversykdom og ulykker og vold, og estimatene er høyere med Cox-modellen. Det er få kvinner som dør av disse årsakene. De fleste kvinnene som dør opplever konkurrerende dødsårsaker. Disse endrer ikke risikomengden for subfordelingshasarden, mens risikomengden for den årsaksspesifikke hasarden reduseres. Forskjellen mellom risikomengdene for de to metodene blir dermed større her enn for de andre dødsårsakene.

De estimerte standardavvikene til de kumulative insidensfunksjonene for begge metodene lagt er i Tillegg A.3. Ved å sammenligne estimatene for subfordelingshasardene og de årsaksspesifikke hasardene, observerte vi ikke så store forskjeller. I datasettet vi har brukt til nå er det generelt lav dødelighet. Risikomengdene i de to metodene blir dermed ikke så ulike. I neste kapittel skal vi simulere data med større dødelighet for å studere dette nærmere.

4.3 Programvare

I dette avsnittet skal vi forklare hvordan subfordelingshasarder kan implementeres i **R**. Funksjonen `crr()` i pakken `cmprsk` estimerer regresjonskoeffisientene i den proporsjonale subfordelingshasard modellen (4.21). Den bruker IPWC-vekting til høyre-sensurerte data, som er foreslått i Fine & Gray (1999). Men denne funksjonen håndterer ikke trunkering. Man tilpasser en modell for hver dødsårsak, og denne brukes som parameter i funksjonen `predict()` til å estimere de kumulative insidensfunksjonene.

For å estimere regresjonskoeffisientene i den proporsjonale subfordelingshasard modellen (4.21) med høyre-sensurerte og venstre-trunkerte data, tar vi utgangspunkt i **R**-koden som er brukt i Geskus (2011). Først må dataene transformeres til vektet form. Siden vektene avhenger av tid, kan informasjon om en person bli fordelt over flere linjer. Individene som dør av andre årsaker enn den som modelleres får vekter. Som beskrevet, så er disse individene fortsatt i risikomengden, men med en vekt som avhenger av tidspunktet for død. Funksjonen `crprep()`, som ligger i `mstate`-pakken, konverterer datasettet til dette formatet. Med vekter for høyre-sensurering og venstre-trunkering ser de første linjene i datasettet for kvinner slik ut :

```
  id Tstart Tstop status weight.cens weight.trunc death.w.csbp10 death.w.bmigr
1  1  40.00 60.80      0   1.0000000            1          -2.5             2
2  2  44.43 57.65      3   1.0000000            1          -1.5             4
3  2  57.65 57.90      3   1.0000000            1          -1.5             4
```

Tabell 4.3: Relativ endring av de estimerte kumulative insidensfunksjonene (3.20) for årsaksspesifikke hasardrater i forhold til de tilsvarende for subfordelingshasardene for menn.

Person	Kovariater	Dødsårsak	Relativ endring v/alder		
			50 år	60 år	70 år
A		1	-0.099	-0.074	-0.004
	Blodtrykk 125	2	-0.062	-0.043	0.002
	Normal vekt	3	-0.073	-0.043	-0.003
	Aldri røykt	4	-0.232	-0.126	-0.031
B		1	-0.005	0.002	0.026
	Blodtrykk 125	2	-0.011	-0.007	-0.004
	Normal vekt	3	-0.004	0.004	0.011
	10-19 sigaretter pr dag	4	-0.163	-0.077	-0.033
C		1	-0.049	-0.029	0.028
	Blodtrykk 145	2	-0.025	-0.008	0.031
	Normal vekt	3	-0.045	-0.020	0.014
	Aldri røykt	4	-0.188	-0.088	-0.004
D		1	0.038	0.033	0.025
	Blodtrykk 145	2	0.023	0.022	0.012
	Normal vekt	3	0.019	0.014	0.004
	10-19 sigaretter pr dag	4	-0.124	-0.049	-0.030
E		1	-0.076	-0.057	0.002
	Blodtrykk 125	2	-0.064	-0.049	-0.009
	Overvektig	3	-0.055	-0.032	0.002
	Aldri røykt	4	-0.249	-0.142	-0.048
F		1	0.011	0.003	0.000
	Blodtrykk 125	2	-0.017	-0.023	-0.032
	Overvektig	3	0.006	-0.002	-0.007
	10-19 sigaretter pr dag	4	-0.178	-0.094	-0.058
G		1	-0.028	-0.017	0.026
	Blodtrykk 145	2	-0.028	-0.017	0.017
	Overvektig	3	-0.029	-0.012	0.013
	Aldri røykt	4	-0.204	-0.104	-0.024
H		1	0.049	0.025	-0.023
	Blodtrykk 145	2	0.014	0.000	-0.025
	Overvektig	3	0.025	-0.001	-0.031
	10-19 sigaretter pr dag	4	-0.140	-0.069	-0.065

Tabell 4.4: Relativ endring av de estimerte kumulative insidensfunksjonene (3.20) for årsaksspesifikke hasardrater i forhold til de tilsvarende for subfordelingshasardene for kvinner.

Person	Kovariater	Dødsårsak	Relativ endring v/alder		
			50 år	60 år	70 år
A		1	-0.359	-0.087	-0.024
	Blodtrykk 125	2	-0.016	-0.010	0.021
	Normal vekt	3	-0.077	-0.060	-0.014
	Aldri røykt	4	-0.023	-0.020	0.013
B		1	-0.328	-0.066	-0.013
	Blodtrykk 125	2	-0.028	-0.026	-0.011
	Normal vekt	3	-0.023	-0.016	0.005
	10-19 sigaretter pr dag	4	-0.001	-0.005	0.013
C		1	-0.335	-0.070	-0.014
	Blodtrykk 145	2	0.000	0.006	0.032
	Normal vekt	3	-0.048	-0.034	0.008
	Aldri røykt	4	-0.010	-0.009	0.020
D		1	-0.306	-0.053	-0.014
	Blodtrykk 145	2	-0.011	-0.010	-0.001
	Normal vekt	3	0.002	0.006	0.014
	10-19 sigaretter pr dag	4	0.010	0.003	0.010
E		1	-0.355	-0.084	-0.021
	Blodtrykk 125	2	-0.019	-0.013	0.017
	Overvektig	3	-0.043	-0.028	0.014
	Aldri røykt	4	0.009	0.012	0.043
F		1	-0.324	-0.064	-0.015
	Blodtrykk 125	2	-0.031	-0.030	-0.016
	Overvektig	3	0.009	0.013	0.027
	10-19 sigaretter pr dag	4	0.030	0.025	0.038
G		1	-0.331	-0.068	-0.012
	Blodtrykk 145	2	-0.003	0.002	0.029
	Overvektig	3	-0.015	-0.003	0.033
	Aldri røykt	4	0.021	0.023	0.049
H		1	-0.304	-0.053	-0.019
	Blodtrykk 145	2	-0.014	-0.014	-0.005
	Overvektig	3	0.032	0.032	0.031
	10-19 sigaretter pr dag	4	0.040	0.031	0.033

```
4  2  57.90 59.12      3   0.9978225           1          -1.5           4
5  2  59.12 59.37      3   0.9874683           1          -1.5           4
6  2  59.37 59.40      3   0.9814704           1          -1.5           4

  death.w.smkgr count failcode
1             1     1        1
2             1     1        1
3             1     2        1
4             1     3        1
5             1     4        1
6             1     5        1
```

Vi tilpasser deretter regresjonsmodellen med `coxph()`. Dødsårsaken som modelleres må spesifiseres. I tillegg må vi her angi vektene som parameteren `weight=weight.cens` for sensurering, og `weight=weight.cens*weight.trunc` dersom dataene også er trunkerte. Funksjonen `survfit()` brukes deretter til å estimere de kumulative insidensratene. Dette gjøres separat for hver dødsårsak.

De to metodene som er beskrevet over gir de samme estimatene for regresjonskoeffisientene og de kumulative insidensfunksjonene. Varians-estimatene beregnes på ulike måter, så de blir noe forskjellige. For å forsikre oss om at fremgangsmåten for beregningene i Eksempel 4.1 er riktige, beregnet vi estimatene uten trunkering med begge metodene.

Kapittel 5

Sammenligning

I dette kapitlet skal vi simulere data fra to modeller, og beregne kumulative insidensfunksjoner med hasardrater fra Cox- og subfordelingsmodell. I datasettet vi har brukt i de foregående kapitlene er det lav dødelighet, og alderen til personene er høyre-sensurert ved 70 år. Ved å simulere levetider, kan vi følge personene lenger og se på effekten av høyere dødelighet. Det er registrert fire dødsårsaker, men i simuleringene skal vi kun ta med de to vanligste; *kreft* og *hjerte- og karsykdommer, inkludert plutselig død*. Vi skal på slutten undersøke hvor godt modellene estimerer de sanne kumulative insidensfunksjonene.

I avsnitt 5.1 skal vi beskrive hvordan vi kan generere venstre-trunkerte levetider for konkurrerende dødsårsaker. På grunnlag av genererte levetider, kan vi bestemme $\widehat{P}_{0h}(s,t|\mathbf{x}_0)$ og $\widehat{P}_{0h}^{sub}(s,t|\mathbf{x}_0)$, som er de estimerte kumulative insidensfunksjonene for dødsårsak h mellom tidspunktene s og t for kovariatvektor $\mathbf{x}_0$ beregnet med hhv Cox- (3.16) og den proporsjonale subfordelingsmodellen (4.21). Dette kan vi gjenta B ganger, og så beregne gjennomsnittet for Cox-modellen

$$\overline{\widehat{P}}_{0h}(s,t|\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B}\widehat{P}_{0h,b}(s,t|\mathbf{x}_0) \tag{5.1}$$

og for subfordelingsmodellen

$$\overline{\widehat{P}}_{0h}^{sub}(s,t|\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B}\widehat{P}_{0h,b}^{sub}(s,t|\mathbf{x}_0). \tag{5.2}$$

Videre kan vi finne de sanne kumulative insidensfunksjonene $P_{0h}(s,t|\mathbf{x}_0)$ for simuleringsmodellen, og sammenligne med gjennomsnittene (5.1) og (5.2). Gitt kovariatene $\mathbf{x}_0$, har vi

$$P_{0h}(s,t|\mathbf{x}_0) = \int_s^t \exp\left(-\int_s^u \sum_{h=1}^{k}\alpha_h(v|\mathbf{x}_0)dv\right)\alpha_h(u|\mathbf{x}_0)du. \tag{5.3}$$

Dette tilsvarer overgangssannsynligheten (2.19) fra Kapittel 2, men her har vi med kovariater. For en gitt tid t, kan vi dermed beregne skjevheten

$$\overline{\widehat{P}}_{0h}(40,t|\mathbf{x}_0) - P_{0h}(40,t|\mathbf{x}_0)\ , \tag{5.4}$$

som er avviket mellom estimert verdi av kumulativ insidensfunksjon med Cox-modellen og sann verdi for en person med kovariater $\mathbf{x}_0$. På samme måte kan vi beregne skjevhet for subfordelingsmodellen.

Det kvadratiske avviket fra den empiriske fordelingen er et mål på variasjonen i estimatene. For $\widehat{P}_{0h}(s,t|\mathbf{x}_0)$ er den gitt ved

$$\tau(s,t|\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B}\{\widehat{P}_{0h,b}(s,t|\mathbf{x}_0) - P_{0h}(s,t|\mathbf{x}_0)\}^2 \,, \tag{5.5}$$

der $\widehat{P}_{0h,b}(s,t|\mathbf{x}_0)$ er estimert kumulativ insidensfunksjon beregnet for simulering b. Det kvadratiske avviket for $\widehat{P}^{sub}_{0h}(s,t|\mathbf{x}_0)$ beregnes på tilsvarende måte.

5.1 Generering av levetider

Vi lar T være levetid med kumulativ hasard rate $A(t)$. Fra Kapittel 2 har vi sammenhengen $P(T > t) = e^{-A(t)}$. Vi setter $V = A(T)$, og finner

$$P\{V > v\} = P\{A(T) > v\} = P\{T > A^{-1}(v)\} = e^{-v}, \tag{5.6}$$

som er overlevelsesfunksjonen til eksponentialfordelingen med parameter lik 1. Tilsvarende kan vi vise at hvis $V \sim \exp(1)$, så har $T = A^{-1}(V)$ kumulativ hasard $A(t)$. Denne sammenhengen kan vi bruke til å generere levetider med gitte hasarder. For å generere n levetider trekker vi først $V_i \sim \exp(1)$ for $i = 1, ..., n$. Så setter vi

$$T_i = A^{-1}(V_i). \tag{5.7}$$

Med venstre-trunkerte data må vi betinge hasardraten på trunkeringstiden L_i. Fra Kapittel 1 i Aalen, Borgan og Gjessing (2008) har vi at hasardraten for venstre-trunkerte data for person i er gitt ved

$$\alpha(t|L_i) = \begin{cases} 0 & \text{hvis } t \le L_i \,, \\ \alpha(t) & \text{hvis } t > L_i \,. \end{cases} \tag{5.8}$$

Gitt at individet har kommet under observasjon, er hasardraten altså lik $\alpha(t)$. Ved å integrerer (5.8) mellom 0 og t, får vi den kumulative hasardraten

$$A(t|L_i) = \begin{cases} 0 & \text{hvis } t \le L_i \,, \\ \int_{L_i}^{t} \alpha(u)du & \text{hvis } t > L_i \,. \end{cases} \tag{5.9}$$

Dersom personen har kommet under observasjon, er den kumulative hasarden integralet av hasardraten mellom trunkeringstiden L_i og t.

Videre kan vi utvide dette til trekning av levetider for konkurrerende dødsårsaker med venstre-trunkering. Vi trekker n trunkeringstider $L_1, ..., L_n$ med tilhørende kovariater $\mathbf{x}_1, ..., \mathbf{x}_n$. For hver person genererer vi potensiell tid for død av årsak h;

$$T_{ih} = A_h^{-1}(V_i|\mathbf{x}_i, L_i) \quad \text{for } h = 1, 2, ..., k, \tag{5.10}$$

der $A_h^{-1}(\cdot|\mathbf{x}_i, L_i)$ er den inverse kumulative hasardraten for dødsårsak $h = 1, ..., k$. Sensureringstiden for person i settes lik en konstant C_i. Ved å velge den minste verdien av $T_{i1}, ..., T_{ik}$ og C_i for hver person, får vi en fullstendig observert eller høyre-sensurert levetid. Dette er oppsummert i Algoritme 1.

Algoritme 1: *Trekning av levetider*

for $i = 1 \rightarrow n$
    Trekk trunkeringsalder L_i med tilhørende kovariater $\mathbf{x}_i$
    for $h = 1 \rightarrow k$
        Trekk potensiell tid for død T_{ih} med (5.10)
    end for
    Trekk eventuell tid for høyre-sensurering C_i
    Hvis $\min(T_{i1}, ..., T_{ik}, C_i) = T_{ih} \rightarrow$ Død av årsak h
    Ellers $\min(T_{i1}, ..., T_{ik}, C_i) = C_i \rightarrow$ Sensurering
end for

5.2 Simuleringsmodeller

Modellene vi skal generere data fra er Cox- og additiv modell. For begge modellene skal vi anta at regresjonskoeffisientene er tidsuavhengige, og at baseline hasardene er Weibull fordelte. Andre fordelinger som også er mye brukt til modellering av hasardrater er eksponential- og Gompertz fordelingen. For Weibull fordelingen har vi

$$\alpha(t; a, b) = (a/b)(t/b)^{a-1}. \tag{5.11}$$

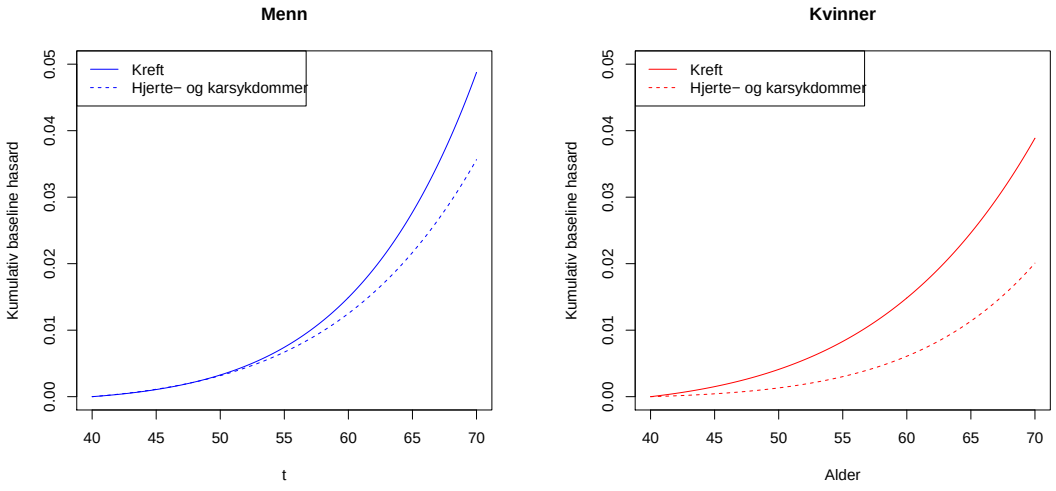
For $a > 1$ øker hasardraten med tiden. I neste avsnitt beskriver vi hvordan vi kan estimere parametrene a og b.

5.2.1 Estimering av parametrene i Weibull fordelingen

For å få rimelige verdier av a og b i (5.11), vil vi tilpasse Weibull baseline hasard for datasettet fra Kapittel 1. Estimatene til parametrene a og b er vist i Tabell 5.1. Funksjonen `weibreg()` i pakken `eha` er brukt til å beregne disse estimatene. Denne funksjonen anvendes på samme måte som `coxph()`. I Figur 5.1 har vi plottet kumulative baseline hasarder med de estimerte parametrene for de to dødsårsakene. For begge kjønn er hasardraten høyere for kreft, men det er større forskjell mellom de to dødsårsakene blant kvinner. Vi kan sammenligne disse plottene med figurene 4.1 og 4.2, som er plott av kumulative baseline hasarder beregnet med Cox- og subfordelingsmodell for hhv dødsårsakene kreft og hjerte- og karsykdommer. Weibull fordelingen er på samme form, og kurvene ligger på omtrent samme nivå, så dette er et rimelig valg av parametrene for baseline.

Tabell 5.1: Estimerte parametre for Weibull baseline separat for menn og kvinner med dødsårsakene *kreft* og *hjerte- og karsykdommer inkludert plutselig død*.

Estimat	Kreft		Hjerte- og karsykdommer	
	Menn	Kvinner	Menn	Kvinner
$\hat{a}$	7.464	5.861	6.481	7.529
$\hat{b}$	104.705	121.030	116.607	117.395



(a) (b)

Figur 5.1: Kumulative Weibull baseline hasarder for (a) menn og (b) kvinner med dødsårsakene *kreft* og *hjerte- og karsykdommer inkludert plutselig død*.

5.2.2 Cox-modell med parametrisk baseline

Vi skal bruke den årsaksspesifikke modellen (3.16) til å generere levetider. Gitt kovariatene $\mathbf{x}$, er den for dødsårsak h definert som

$$\alpha_h(t|\mathbf{x}) = \alpha_{h,0}(t)\exp\{\boldsymbol{\beta}_h^T\mathbf{x}\}. \tag{5.12}$$

Siden Cox-modellen er semi-parametrisk, er den ikke så godt egnet for å generere data. Vi vil derfor anta en parametrisk modell for baseline hasarden, og antar at den er Weibull fordelt. For venstre-trunkerte data blir den årsaksspesifikke hasardraten (5.12) med Weibull baseline lik

$$\alpha_h(t; a_h, b_h|\mathbf{x}, L) = \begin{cases} 0 & \text{hvis } t \leq L\ , \\ (a_h/b_h)(t/b_h)^{a_h-1}\exp\{\boldsymbol{\beta}_h^T\mathbf{x}\} & \text{hvis } t > L\ . \end{cases} \tag{5.13}$$

Betinget kumulativ hasardrate på tid t blir

$$A_h(t|\mathbf{x}, L) = \begin{cases} 0 & \text{hvis } t \leq L\ , \\ b_h^{-a_h}(t^{a_h} - L^{a_h})\exp\{\boldsymbol{\beta}_h^T\mathbf{x}\} & \text{hvis } t > L\ . \end{cases} \tag{5.14}$$

Vi finner (5.10) for (5.14), og får for dødsårsak h den potensielle levetiden

$$T_h = \left(b_h^{a_h} V \exp\left\{-\boldsymbol{\beta}_h^T\mathbf{x}\right\} + L^{a_h}\right)^{1/a_h}. \tag{5.15}$$

Ved å erstatte (5.10) med (5.15) i Algoritme 1, genererer vi levetider fra Cox-modellen (5.13).

5.2.3 Additiv modell med parametrisk baseline

Vi bruker tidligere innført notasjon, og lar $\boldsymbol{\beta}_h = (\beta_{h1}, ..., \beta_{hp})$ være vektoren med de p regresjonskoeffisientene for dødsårsak h, og $\mathbf{x}$ vektoren med kovariater. Den additive regresjonsmodellen med tidsuavhengige regresjonskoeffisienter er gitt ved

$$\alpha_h(t) = \beta_{h,0}(t) + \boldsymbol{\beta}_h^T\mathbf{x} \text{ for } h = 1, 2, ..., k\ , \tag{5.16}$$

der $\beta_{h,0}(t)$ er baseline hasard på tid t. Se seksjon 4.2 i Aalen, Borgan og Gjessing (2008). Ved å øke kovariat j med én enhet, så endres hasardraten med β_{hj} for dødsårsak h. I motsetning til hasardraten for Cox-modellen, kan (5.16) bli negativ. Vi setter inn Weibull baseline (5.11) for baseline i (5.16), og får for venstre-trunkerte data hasardraten

$$\alpha_h(t; a_h, b_h|\mathbf{x}, L) = \begin{cases} 0 & \text{hvis } t \leq L\ , \\ (a_h/b_h)(t/b_h)^{a_h-1} + \boldsymbol{\beta}_h^T\mathbf{x} & \text{hvis } t > L\ , \end{cases} \tag{5.17}$$

Den tilhørende kumulative hasardraten er

$$A_h(t|\mathbf{x}, L) = \begin{cases} 0 & \text{hvis } t \leq L\ , \\ b_h^{-a_h}(t^{a_h} - L^{a_h}) + (t - L)\boldsymbol{\beta}_h^T\mathbf{x} & \text{hvis } t > L\ . \end{cases} \tag{5.18}$$

For å trekke levetider fra denne modellen, må (5.10) løses numerisk.

5.3 Resultater og diskusjon

I de to neste avsnittene skal vi simulere levetider fra modellene (5.13) og (5.17), og bruke de simulerte levetidene til å estimere de kumulative insidensfunksjonene (3.20) og (4.26), som tilhører hhv den årsaksspesifikke Cox-modellen (3.16) og modellen for subfordelinger (4.21) fra alderen $s = 40$ år. Dette beregnes for de åtte kombinasjonene av kovariatene som er gitt i Tabell 3.4. I hver simulering vil vi trekke 1000 levetider, og gjenta dette $B = 100$ ganger. Vi estimerer kumulative insidensfunksjoner slik som i kapitlene 3 og 4. Deretter beregner vi gjennomsnittene (5.1) og (5.2). Videre skal vi for aldrene $t = 50, 60, 70$ og 80 år beregne skjevhet (5.4) og kvadratisk avvik (5.5). Skjevhetene og de kvadratiske avvikene oppsummeres i tabeller. De kvadratiske avvikene blir multiplisert med 10^6. Sensureringstiden settes lik 90 år for alle personene.

5.3.1 Cox-modell

Her skal vi generere levetider fra modellen (5.13). Siden vi trekker levetider fra Cox-modellen, forventer vi at avvikene blir lavere for Cox- (3.16) enn subfordelingsmodellen (4.21). Parameterestimatene fra avsnitt 5.2.1 benyttes som sanne verdier for Weibull baseline hasard. De estimerte regresjonskoeffisientene fra tabellene 3.8-3.9 (menn) og 3.12-3.13 (kvinner) brukes som sanne verdier for $\boldsymbol{\beta}_h$ med $h = 1, 2$. Vi trekker med tilbakelegging tilfeldig trunkeringstid L_i med tilhørende kovariater $\mathbf{x}_i$ fra de faktiske dataene, og genererer potensiell tid for død T_{ih} med (5.15).

For å illustrere variasjonen i resultatene fra trekning til trekning, har vi i figurene 5.2-5.5 plottet $B = 20$ simuleringer for begge metodene, som er de grå linjene. Den mørke linjen viser (5.3), som er fasiten. Vi ser at variasjonen øker med alderen. For død av kreft, ser vi at Fine-Gray overestimerer dødeligheten for personene A og E. Dette er de to personene med lavest dødssannsynligheter. Ellers kan vi ikke fra plottene observere tydelige forskjeller på de to metodene. For å få en mer detaljert sammenligning, ser vi på skjevhetene (5.4) og de kvadratiske avvikene (5.5).

Tabell 5.2 (kreft) og Tabell 5.3 (hjerte- og karsykdommer) viser resultatene for menn. For kreftdød blant menn, så ser vi en antydning til at Cox-modellen overestimerer dødeligheten noe mellom 80 og 90 år for personer med normal vekt (A-D). Skjevheten for subfordelingen er litt høyere. For hjerte- og karsykdommer de fleste estimerte kumulative insidesfunksjonene med Cox-modellen lavere enn sann verdi, men er av ubetydelig størrelse. Forskjellen mellom metodene er mindre for denne dødsårsaken. Som forventet, så er skjevheten lavest for Cox-modellen, men Fine-Gray modellen gir også god tilpasning.

Generelt kan vi se at de kvadratiske avvikene øker med alderen for alle personene, men er størst for personer som røyker. Med noen unntak, er de kvadratiske avvikene noe høyere for subfordelingsmodellen.

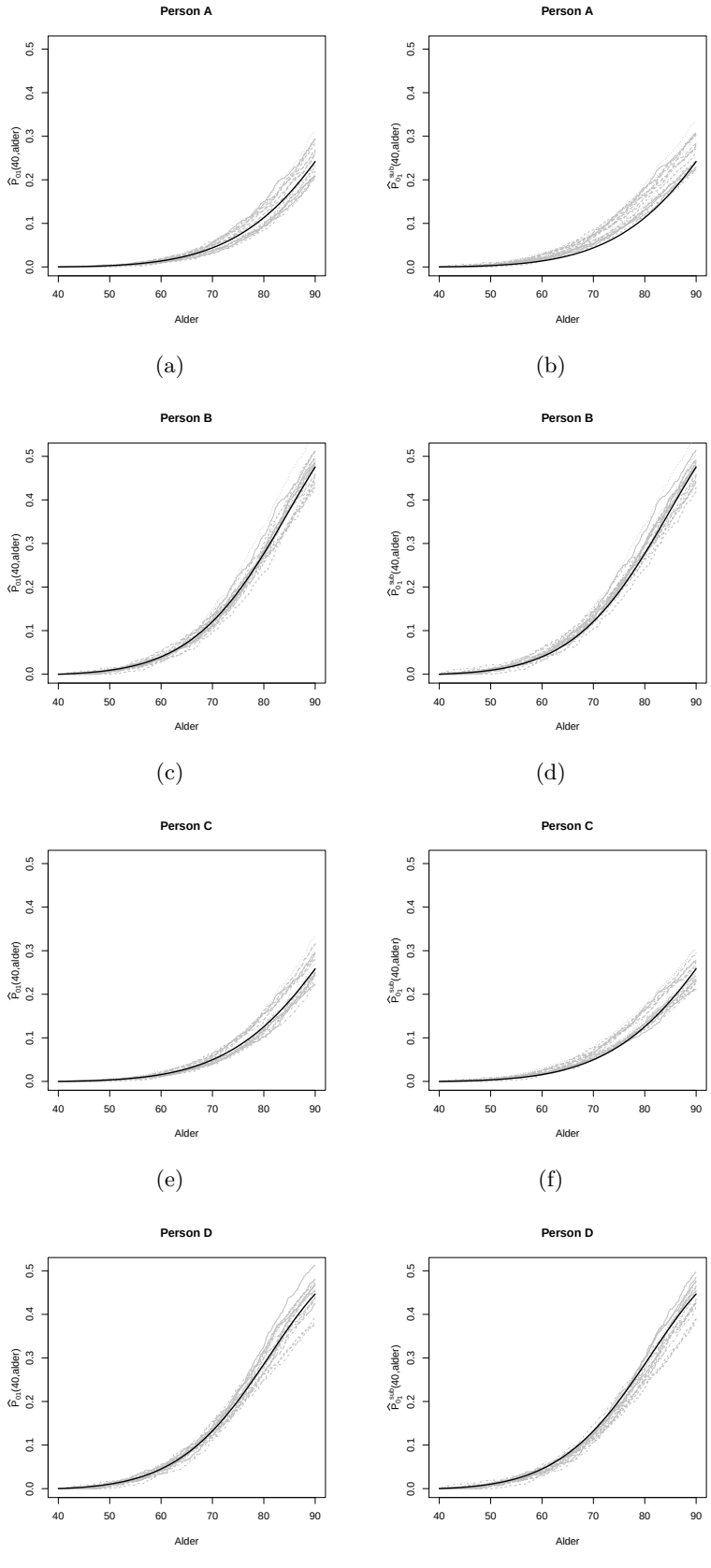
Resultatene for kvinner er lagt i Tillegg B.1 (tabeller) og Tillegg B.2 (figurer). Blant kvinner gir begge modellene god tilpasning. De kvadratiske avvikene og skjevhetene er omtrent på samme størrelse. Dette skyldes at lavere dødelighet blant kvinner enn menn. Vi observerer det samme i plottene.

Tabell 5.2: Skjevhet og kvadratisk avvik for estimerte kumulative insidensfunksjoner beregnet med Cox- og subfordelingsmodell med dødsårsaken *kreft* for menn. Levetidene er trukket fra Cox-modellen (5.13) med $B = 100$ simuleringer, og $n = 1000$ levetider i hver simulering.

Person	Kovariater	Alder	$P_{0h}(t\|\mathbf{x})$	Cox		Fine-Gray	
				Skjevhet	10^6 Kv.avvik	Skjevhet	10^6 Kv. avvik
		50	0.003	-0.000	2.084	0.002	7.609
	Blodtrykk 125	60	0.014	-0.000	11.237	0.008	87.944
A	Normal vekt	70	0.044	0.000	72.962	0.022	605.599
	Aldri røykt	80	0.113	0.002	382.309	0.037	1883.191
		90	0.242	0.006	1190.723	0.024	1803.752
		50	0.009	-0.002	14.891	0.000	17.230
	Blodtrykk 125	60	0.040	-0.001	58.895	0.004	79.095
B	Normal vekt	70	0.121	-0.000	260.419	0.008	322.655
	10-19 sigaretter pr dag	80	0.277	0.001	798.632	0.005	798.159
		90	0.476	0.004	1123.243	-0.008	1195.251
		50	0.003	-0.001	2.722	0.001	4.811
	Blodtrykk 145	60	0.016	-0.000	13.374	0.004	39.008
C	Normal vekt	70	0.050	0.001	80.509	0.011	229.533
	Aldri røykt	80	0.126	0.002	408.858	0.014	609.507
		90	0.259	0.007	1202.839	-0.009	1039.179
		50	0.010	-0.002	19.591	-0.002	17.481
	Blodtrykk 145	60	0.045	-0.001	69.996	-0.005	80.496
D	Normal vekt	70	0.132	0.000	302.607	-0.012	405.737
	10-19 sigaretter pr dag	80	0.286	0.002	896.969	-0.021	1272.020
		90	0.447	0.004	1346.631	-0.003	1431.147
		50	0.003	-0.001	1.924	0.001	4.388
	Blodtrykk 125	60	0.014	-0.001	10.187	0.005	45.654
E	Overvektig	70	0.043	-0.001	59.831	0.014	291.107
	Aldri røykt	80	0.111	-0.002	290.397	0.021	811.199
		90	0.233	-0.001	925.782	0.002	935.439
		50	0.009	-0.002	14.447	-0.001	13.105
	Blodtrykk 125	60	0.039	-0.002	64.044	-0.001	60.420
F	Overvektig	70	0.117	-0.003	287.141	-0.004	287.185
	10-19 sigaretter pr dag	80	0.263	-0.006	920.573	-0.013	1043.561
		90	0.435	-0.006	1624.691	-0.014	1763.761
		50	0.003	-0.001	2.504	0.000	2.888
	Blodtrykk 145	60	0.016	-0.001	12.185	0.002	18.403
G	Overvektig	70	0.049	-0.001	63.684	0.004	92.614
	Aldri røykt	80	0.122	-0.001	296.536	0.001	284.190
		90	0.245	0.001	905.742	-0.024	1290.570
		50	0.010	-0.002	18.959	-0.003	17.998
	Blodtrykk 145	60	0.044	-0.002	78.670	-0.009	131.145
H	Overvektig	70	0.127	-0.003	341.227	-0.021	721.484
	10-19 sigaretter pr dag	80	0.263	-0.005	1053.961	-0.029	1725.474
		90	0.390	-0.005	1814.298	0.008	1796.709

Tabell 5.3: Skjevhet og kvadratisk avvik for estimerte kumulative insidensfunksjoner beregnet med Cox- og subfordelingsmodell med dødsårsaken *hjerte- og karsykdommer* for menn. Levetidene er trukket fra Cox-modellen (5.13) med $B = 100$ simuleringer, og $n = 1000$ levetider i hver simulering.

Person	Kovariater	Alder	$P_{0h}(t\|\mathbf{x})$	Cox		Fine-Gray	
				Skjevhet	10^6 Kv.avvik	Skjevhet	10^6 Kv. avvik
		50	0.002	-0.000	0.722	0.001	1.942
	Blodtrykk 125	60	0.009	-0.001	4.330	0.002	10.635
A	Normal vekt	70	0.026	-0.002	26.781	0.006	65.621
	Aldri røykt	80	0.059	-0.004	117.252	0.008	179.548
		90	0.114	-0.006	377.233	-0.002	327.672
		50	0.008	-0.001	9.039	0.000	8.854
	Blodtrykk 125	60	0.032	-0.002	33.481	-0.001	30.131
B	Normal vekt	70	0.085	-0.001	145.011	-0.002	134.146
	10-19 sigaretter pr dag	80	0.174	-0.001	481.099	-0.003	442.456
		90	0.276	-0.000	852.204	0.000	829.077
		50	0.004	-0.001	2.080	0.001	4.035
	Blodtrykk 145	60	0.016	-0.002	12.263	0.002	19.274
C	Normal vekt	70	0.045	-0.003	70.379	0.006	109.216
	Aldri røykt	80	0.101	-0.006	277.686	0.005	282.945
		90	0.188	-0.010	792.248	-0.013	796.609
		50	0.015	-0.001	26.395	-0.000	22.227
	Blodtrykk 145	60	0.056	-0.003	91.820	-0.006	103.392
D	Normal vekt	70	0.143	-0.002	347.088	-0.011	426.780
	10-19 sigaretter pr dag	80	0.278	-0.002	910.539	-0.014	1047.539
		90	0.404	-0.001	1332.652	0.004	1350.698
		50	0.003	-0.000	1.375	0.001	3.901
	Blodtrykk 125	60	0.013	-0.001	7.794	0.003	20.199
E	Overvektig	70	0.036	-0.002	44.537	0.008	117.532
	Aldri røykt	80	0.082	-0.005	192.459	0.010	315.934
		90	0.156	-0.007	603.172	-0.004	573.486
		50	0.012	-0.001	17.553	0.001	17.951
	Blodtrykk 125	60	0.045	-0.002	67.228	-0.001	65.386
F	Overvektig	70	0.117	-0.001	269.294	-0.002	267.363
	10-19 sigaretter pr dag	80	0.234	0.001	842.781	-0.003	830.869
		90	0.358	0.004	1427.199	0.004	1457.169
		50	0.006	-0.001	4.012	0.001	8.265
	Blodtrykk 145	60	0.023	-0.002	22.236	0.003	36.752
G	Overvektig	70	0.062	-0.004	116.408	0.008	190.924
	Aldri røykt	80	0.138	-0.007	445.921	0.007	483.054
		90	0.251	-0.010	1201.627	-0.017	1328.562
		50	0.020	-0.001	51.336	-0.001	45.275
	Blodtrykk 145	60	0.077	-0.003	184.585	-0.008	215.571
H	Overvektig	70	0.194	-0.001	634.177	-0.015	802.609
	10-19 sigaretter pr dag	80	0.362	0.001	1502.167	-0.016	1731.586
		90	0.502	0.004	2011.614	0.014	2250.681



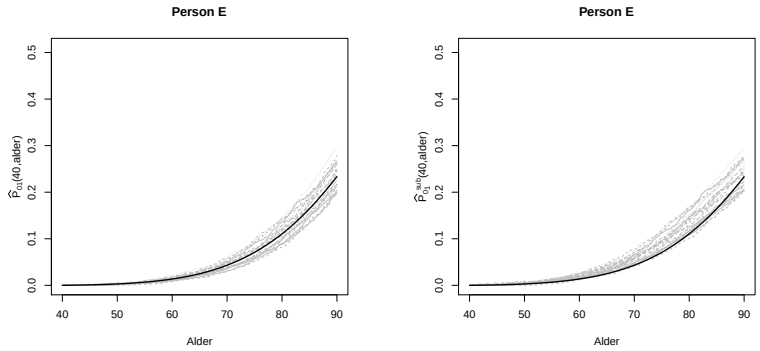
(a) (b)

(c) (d)

(e) (f)

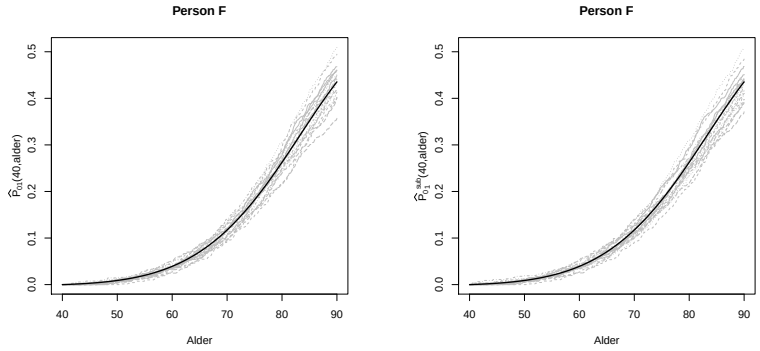
(g) (h)

Figur 5.2: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *kreft* for menn og personene A-D. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



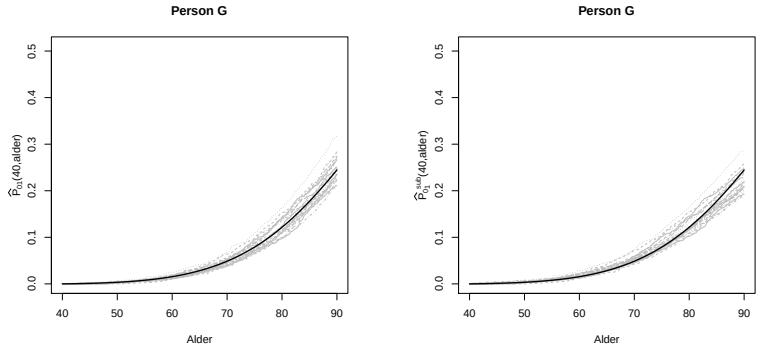
(a)

(b)



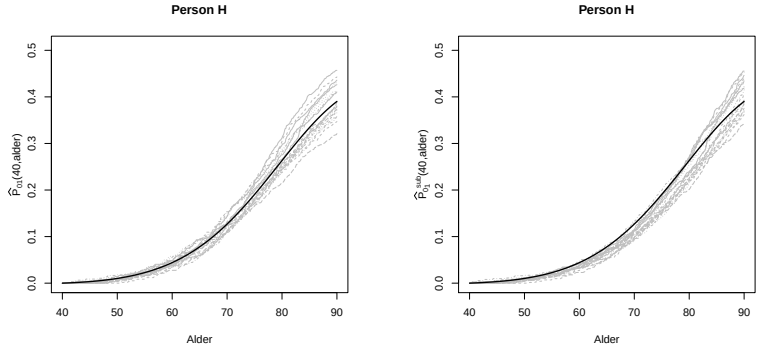
(c)

(d)



(e)

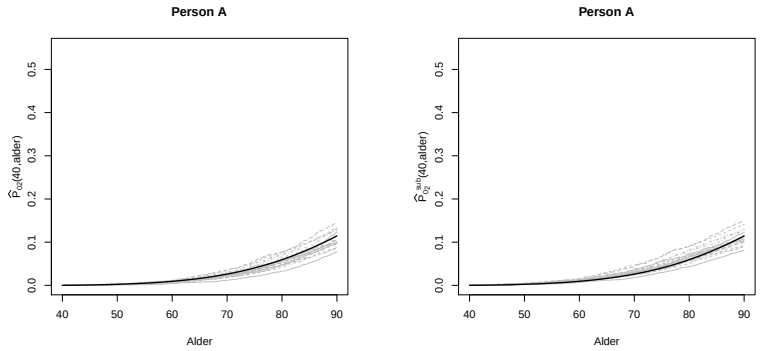
(f)



(g)

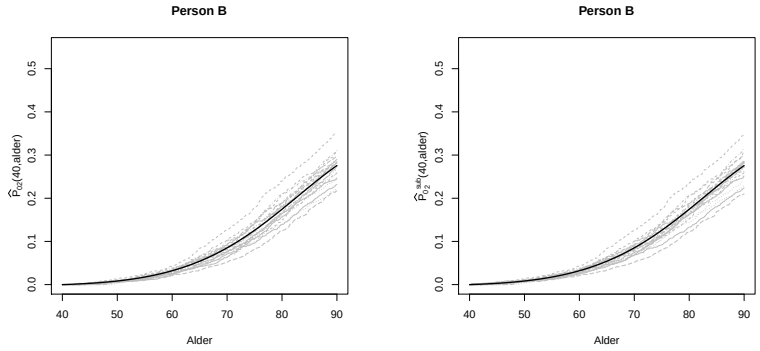
(h)

Figur 5.3: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *kreft* for menn og personene E-H. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



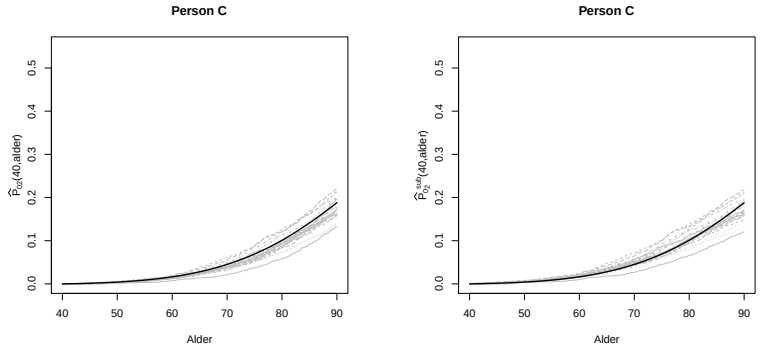
(a)

(b)



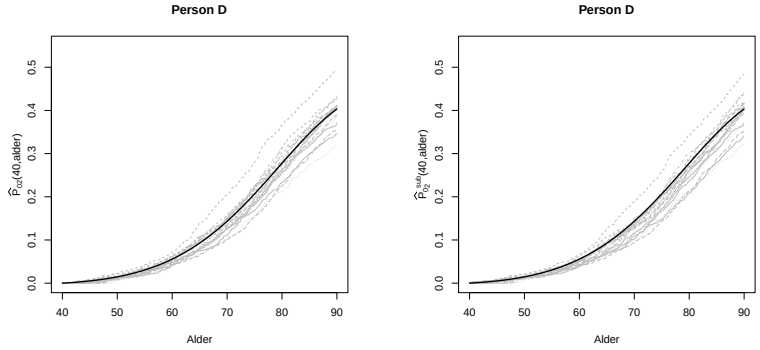
(c)

(d)



(e)

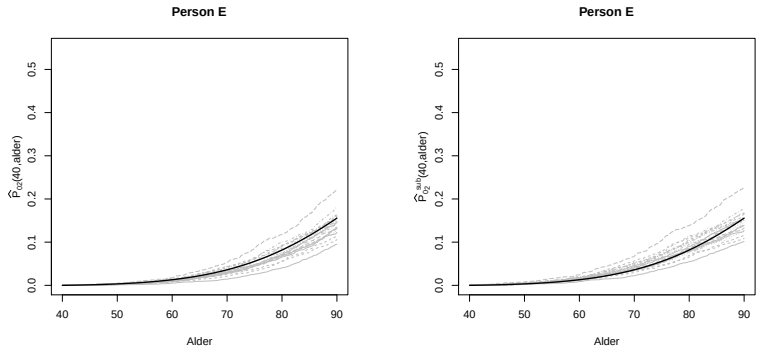
(f)



(g)

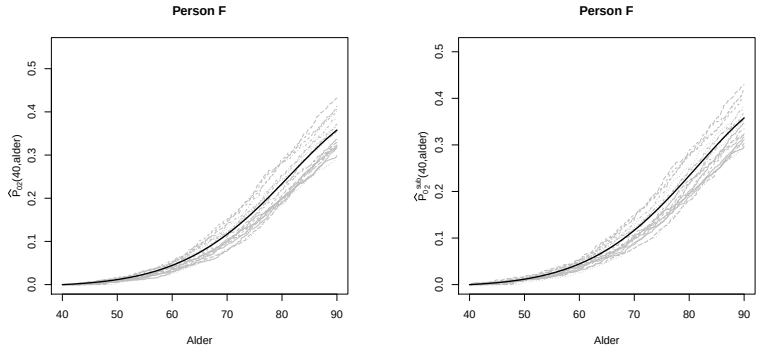
(h)

Figur 5.4: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *hjerte- og karsykdommer* for menn og personene A-D. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



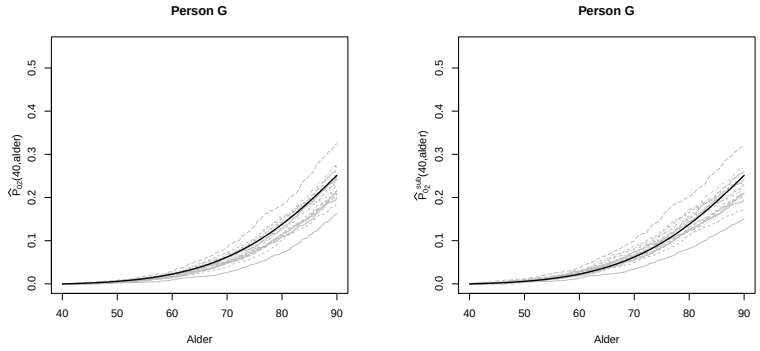
(a)

(b)



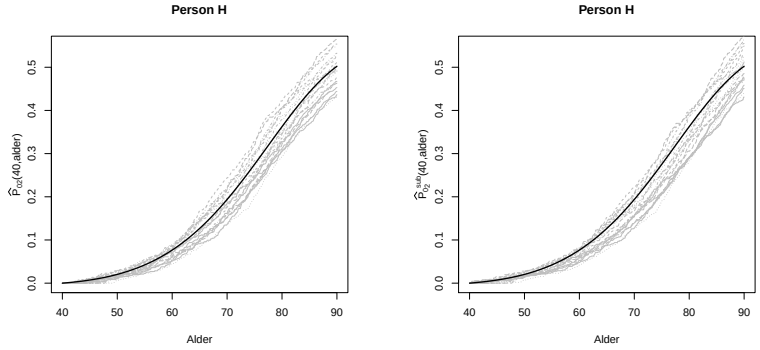
(c)

(d)



(e)

(f)



(g)

(h)

Figur 5.5: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *hjerte- og karsykdommer* for menn og personene E-H. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.

5.3.2 Additiv modell

Vi skal generere data fra den additive modellen (5.17), og estimere kumulative insidensfunksjoner med Cox- og subfordelingsmodell. Her er dermed begge modellene feil. Med trekninger fra den additive modellen, er målet å undersøke hvilken av metodene som best estimerer insidensfunksjonene ved en feilspesifisert modell. Siden det er høyere dødelighet blant menn enn kvinner, velger vi her å beregne estimatene kun for menn.

Først estimerer vi regresjonskoeffisientene for den additive modellen. Vi bruker funksjonen `aalen()` fra pakken `timereg`. Den første parameteren er den samme som i `coxph()`, dvs. et objekt av typen `Surv`. For å spesifisere at regresjonskoeffisientene skal være konstante, skriver vi `const(kovariat)`. Resultatene er gitt i tabellene B.3 (kreft) og B.4 (hjerte- og karsykdommer). Disse estimatene brukes som sanne verdier av regresjonskoeffisientene i den kumulative hasardraten (5.18).

Som nevnt i avsnitt 5.2.3, må vi finne T_{ih} numerisk. For hver person $i = 1, ..., n$ benytter vi funksjonen `uniroot()` til å finne nullpunktet til uttrykket

$$A_h(t|\mathbf{x}_i, L_i) - V_i \ ,$$

der $V_i \sim \exp(1)$ og $\widehat{A}_h(\cdot|\mathbf{x}_i, L_i)$ er den kumulative hasardraten (5.18) med verdiene for a og b for menn fra Tabell 5.1.

Figurene 5.6-5.9 viser resultatene fra $B = 20$ simuleringer. Avvikene er størst for personene A og E, der vi igjen ser at dødeligheten for begge dødsårsakene overestimeres. I Tabell 5.4 (kreft) og Tabell 5.5 (hjerte- og karsykdommer) har vi oppsummert resultatene fra simuleringene. Fra disse resultatene kan vi ikke finne noen systematiske forskjeller mellom de to metodene. De kvadratiske avvikene er på samme nivå.

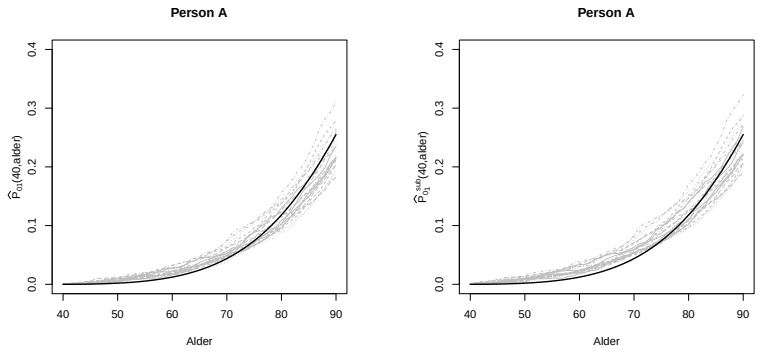
Vi har i dette kapitlet først trukket levetider fra Cox-modellen, og estimert kumulative insidensfunksjoner med både Cox og Fine-Gray. Selv om Fine-Gray var gal modell her, så ble likevel estimatene ganske nøyaktige. Ved å trekke fra den additive modellen, ble resultatene omtrent like. Disse resultatene antyder at metodene er like gode til å estimere kumulative insidensfunksjoner for praktiske anvendelser.

Tabell 5.4: Skjevhet og kvadratisk avvik for estimerte kumulative insidensfunksjoner beregnet med Cox- og subfordelingsmodell med dødsårsaken *kreft* for menn. Levetidene er trukket fra den additive modellen (5.17) med $B = 100$ simuleringer, og $n = 1000$ levetider i hver simulering.

Person	Kovariater	Alder	$P_{0h}(t\|\mathbf{x})$	Cox		Fine-Gray	
				Skjevhet	10^6 Kv.avvik	Skjevhet	10^6 Kv. avvik
		50	0.002	0.006	43.538	0.007	66.898
	Blodtrykk 125	60	0.012	0.011	156.771	0.014	243.868
A	Normal vekt	70	0.044	0.012	235.206	0.017	412.905
	Aldri røykt	80	0.118	0.004	396.853	0.012	579.272
		90	0.255	-0.012	1252.023	-0.006	1193.715
		50	0.016	-0.006	54.051	-0.005	42.866
	Blodtrykk 125	60	0.040	-0.010	144.678	-0.008	122.770
B	Normal vekt	70	0.082	-0.011	290.713	-0.009	257.884
	10-19 sigaretter pr dag	80	0.159	-0.007	532.505	-0.005	515.992
		90	0.285	0.006	1212.682	0.006	1220.915
		50	0.005	0.004	23.487	0.005	32.796
	Blodtrykk 145	60	0.017	0.007	84.543	0.009	112.600
C	Normal vekt	70	0.050	0.007	167.544	0.010	212.177
	Aldri røykt	80	0.121	0.003	443.934	0.006	476.433
		90	0.247	-0.004	1173.172	-0.005	1158.214
		50	0.019	-0.008	87.043	-0.008	79.402
	Blodtrykk 145	60	0.044	-0.013	227.062	-0.013	234.821
D	Normal vekt	70	0.086	-0.014	404.761	-0.015	445.278
	10-19 sigaretter pr dag	80	0.160	-0.007	680.501	-0.010	711.800
		90	0.276	0.010	1543.942	0.007	1456.129
		50	0.002	0.006	44.495	0.007	63.731
	Blodtrykk 125	60	0.012	0.012	168.070	0.014	236.776
E	Overvektig	70	0.043	0.013	278.784	0.017	418.571
	Aldri røykt	80	0.115	0.007	438.767	0.013	592.972
		90	0.248	-0.005	1178.001	-0.002	1176.964
		50	0.016	-0.006	52.392	-0.005	43.227
	Blodtrykk 125	60	0.040	-0.009	137.053	-0.008	126.534
F	Overvektig	70	0.081	-0.009	287.834	-0.009	281.733
	10-19 sigaretter pr dag	80	0.156	-0.004	551.943	-0.004	553.764
		90	0.277	0.011	1523.885	0.010	1514.877
		50	0.005	0.004	23.000	0.004	29.605
	Blodtrykk 145	60	0.017	0.007	87.455	0.008	104.128
G	Overvektig	70	0.049	0.009	179.132	0.009	200.353
	Aldri røykt	80	0.118	0.006	406.775	0.006	415.521
		90	0.241	0.001	1019.543	-0.002	984.268
		50	0.019	-0.008	83.531	-0.008	79.268
	Blodtrykk 145	60	0.044	-0.012	210.329	-0.013	234.220
H	Overvektig	70	0.085	-0.012	372.840	-0.015	441.829
	10-19 sigaretter pr dag	80	0.157	-0.004	616.106	-0.009	653.529
		90	0.269	0.012	1572.866	0.011	1511.532

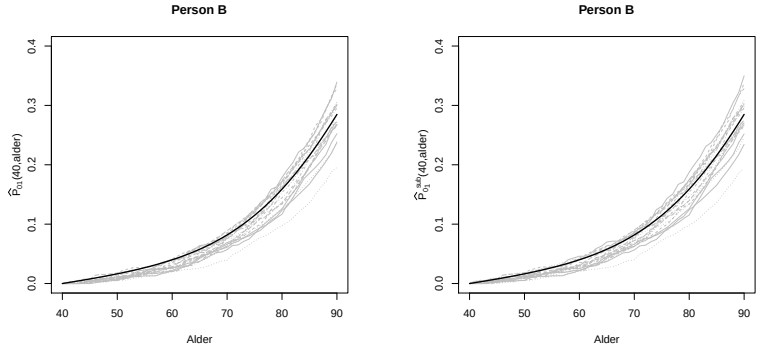
Tabell 5.5: Skjevhet og kvadratisk avvik for estimerte kumulative insidensfunksjoner beregnet med Cox- og subfordelingsmodell modell med dødsårsaken *hjerte- og karsykdommer* for menn. Levetidene er trukket fra den additive modellen (5.17) med $B = 100$ simuleringer, og $n = 1000$ levetider i hver simulering.

Person	Kovariater	Alder	$P_{0h}(t\|\mathbf{x})$	Cox		Fine-Gray	
				Skjevhet	10^6 Kv.avvik	Skjevhet	10^6 Kv. avvik
		50	-0.006	0.017	289.329	0.018	344.534
	Blodtrykk 125	60	-0.005	0.031	972.060	0.033	1109.733
A	Normal vekt	70	0.009	0.037	1473.205	0.040	1694.033
	Aldri røykt	80	0.045	0.031	1189.583	0.034	1417.677
		90	0.112	0.012	631.150	0.014	718.861
		50	0.015	0.002	33.127	0.003	38.109
	Blodtrykk 125	60	0.035	0.004	79.613	0.004	82.466
B	Normal vekt	70	0.066	0.003	161.828	0.003	158.047
	10-19 sigaretter pr dag	80	0.115	-0.001	291.709	-0.002	282.343
		90	0.185	-0.007	597.357	-0.008	584.100
		50	0.012	0.003	42.454	0.005	61.473
	Blodtrykk 145	60	0.030	0.005	103.951	0.007	144.192
C	Normal vekt	70	0.059	0.003	200.127	0.006	248.277
	Aldri røykt	80	0.108	-0.005	449.148	-0.002	465.637
		90	0.181	-0.017	1184.603	-0.014	1155.255
		50	0.032	-0.009	149.331	-0.008	125.200
	Blodtrykk 145	60	0.068	-0.015	369.678	-0.015	354.458
D	Normal vekt	70	0.113	-0.019	662.077	-0.020	687.029
	10-19 sigaretter pr dag	80	0.171	-0.019	917.902	-0.021	983.139
		90	0.245	-0.012	1147.695	-0.013	1155.595
		50	0.002	0.011	141.623	0.013	184.934
	Blodtrykk 125	60	0.009	0.020	458.045	0.022	564.339
E	Overvektig	70	0.030	0.023	674.209	0.026	836.487
	Aldri røykt	80	0.073	0.016	599.320	0.020	757.926
		90	0.142	0.001	784.274	0.003	847.143
		50	0.022	-0.003	50.758	-0.001	43.947
	Blodtrykk 125	60	0.049	-0.004	117.798	-0.003	107.502
F	Overvektig	70	0.086	-0.005	282.054	-0.006	269.304
	10-19 sigaretter pr dag	80	0.140	-0.007	567.701	-0.008	553.482
		90	0.212	-0.007	1114.426	-0.008	1079.555
		50	0.019	-0.002	51.062	0.000	51.698
	Blodtrykk 145	60	0.044	-0.004	123.101	-0.001	123.177
G	Overvektig	70	0.080	-0.008	312.975	-0.005	299.649
	Aldri røykt	80	0.134	-0.014	746.685	-0.011	711.293
		90	0.209	-0.020	1612.563	-0.018	1588.764
		50	0.039	-0.013	245.326	-0.011	207.129
	Blodtrykk 145	60	0.081	-0.020	589.221	-0.020	569.618
H	Overvektig	70	0.131	-0.023	939.011	-0.024	970.864
	10-19 sigaretter pr dag	80	0.194	-0.018	1104.531	-0.021	1154.833
		90	0.270	-0.003	1480.980	-0.005	1425.173



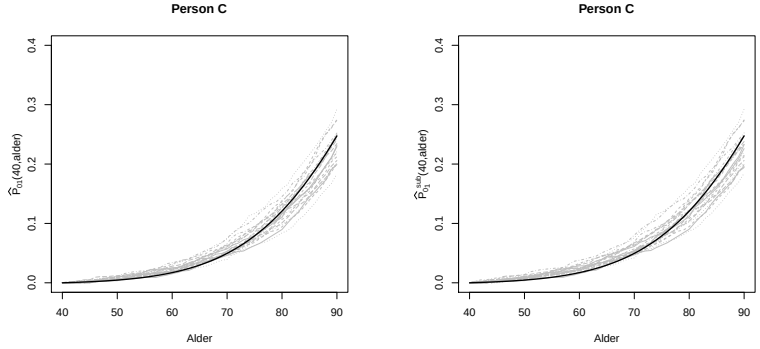
(a)

(b)



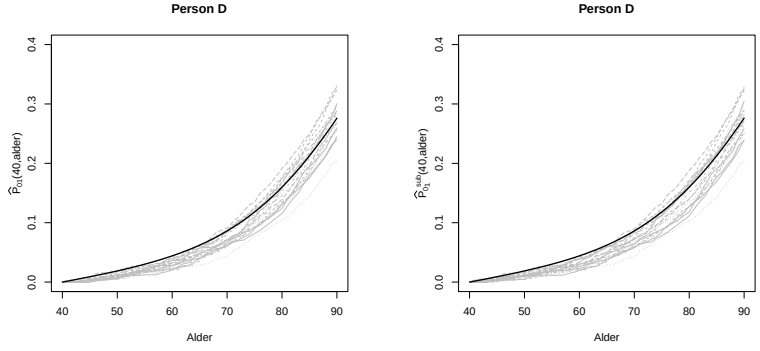
(c)

(d)



(e)

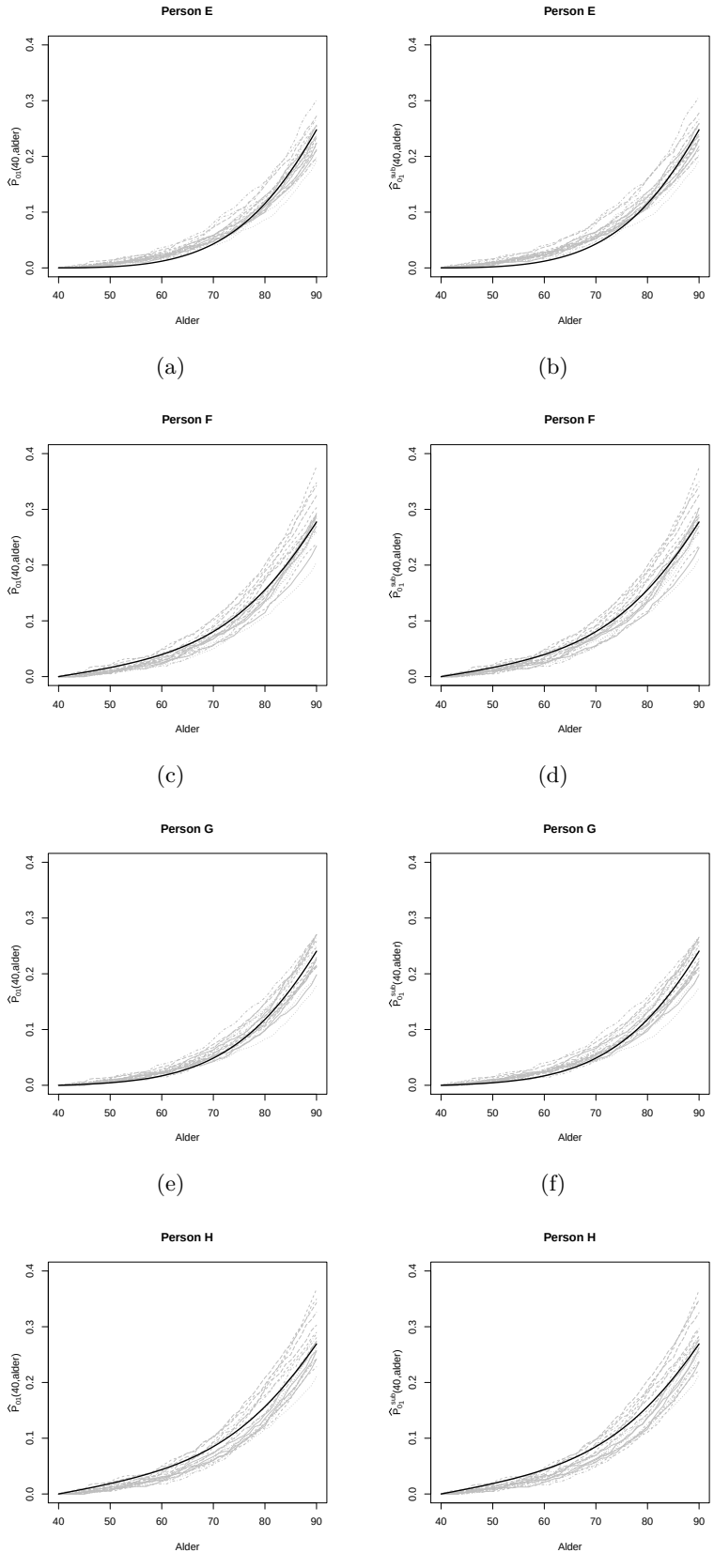
(f)



(g)

(h)

Figur 5.6: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *kreft* for menn og personene A-D. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra den additive-modellen (5.17) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



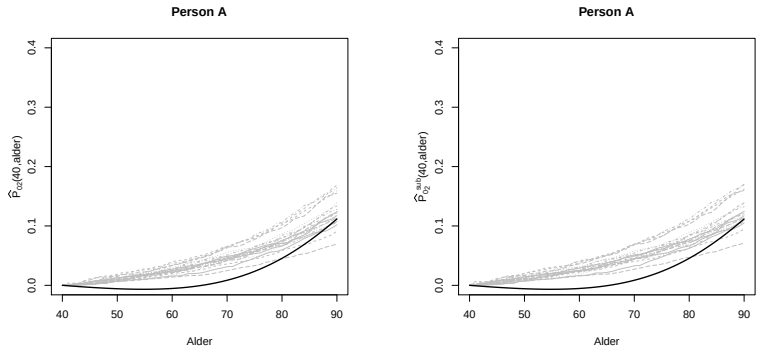
(a) (b)

(c) (d)

(e) (f)

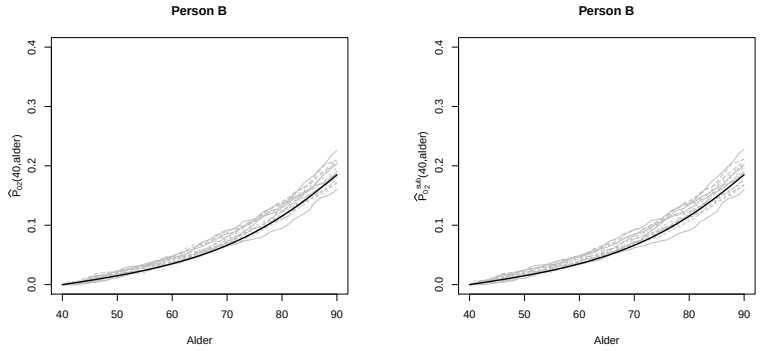
(g) (h)

Figur 5.7: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *kreft* for menn og personene E-H. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra den additive-modellen (5.17) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



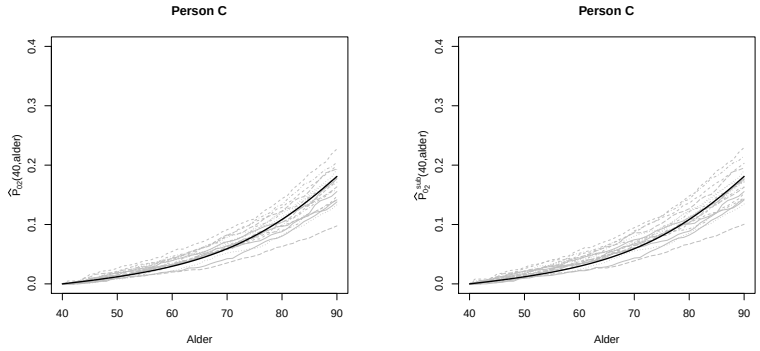
(a)

(b)



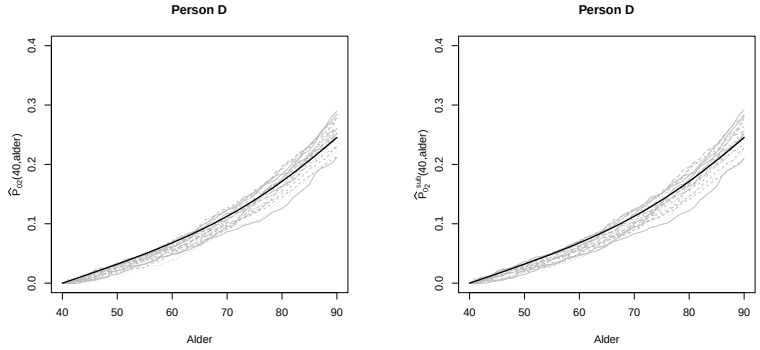
(c)

(d)



(e)

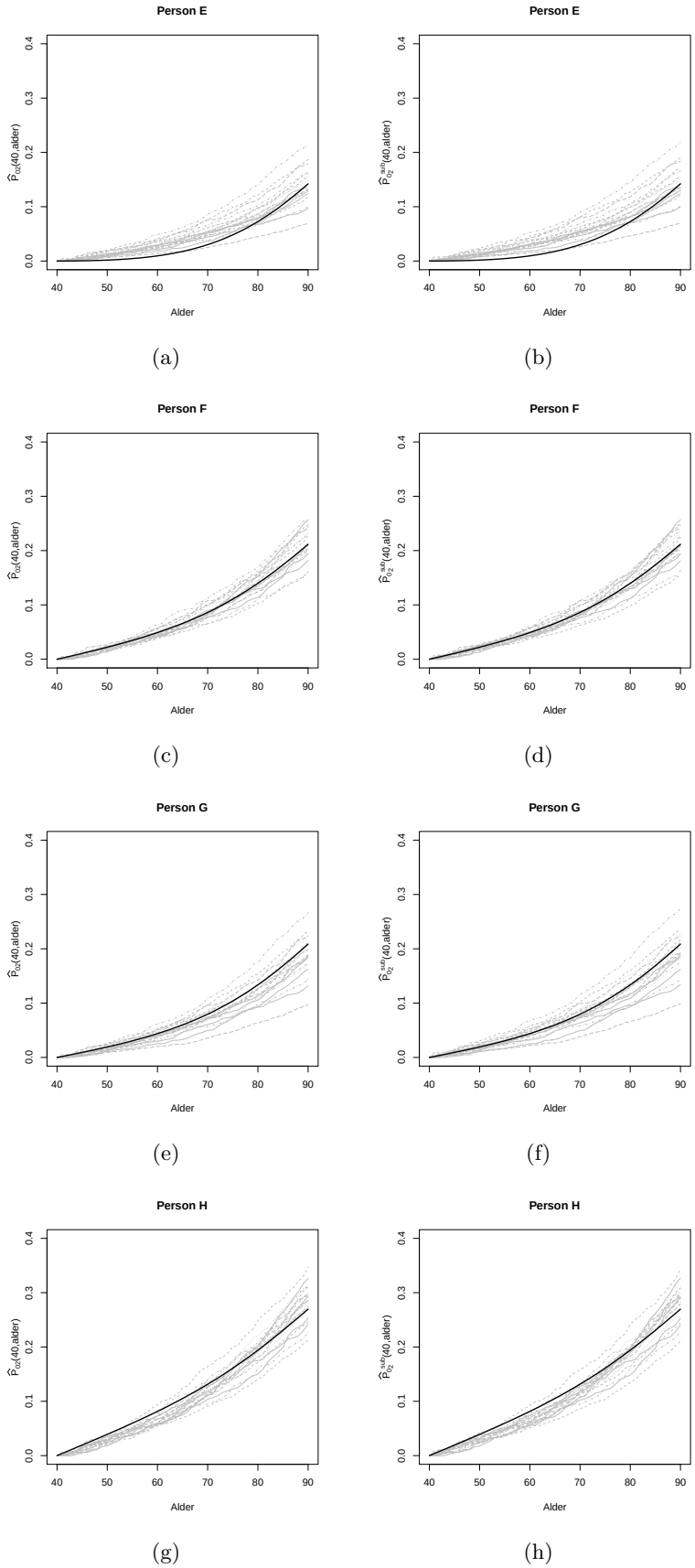
(f)



(g)

(h)

Figur 5.8: kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *hjerte- og karsykdommer* for menn og personene A-D. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra den additive-modellen (5.17) med $B = 20$ simuleringer, g $n = 1000$ levetider i hver simulering.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figur 5.9: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *hjerte- og karsykdommer* for menn og personene E-H. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra den additive-modellen (5.17) med $B = 20$ simuleringer, g $n = 1000$ levetider i hver simulering.

Kapittel 6

Oppsummering og konklusjon

Målet med denne oppgaven har vært å sammenligne modellene Cox- og Fine-Gray for å studere kumulative insidensfunksjoner for realistiske situasjoner. Vi har presentert begge metodene, og anvendt teorien på noen eksempler. Til beregningene har vi først brukt et reelt datasett. Dette var et utvalg på 4000 personer fra fylkesundersøkelsene. Vi valgte å modellere med disse kovariatene fra datasettet: *blodtrykk* (numerisk), *BMI* (fire grupper) og *røykegruppe* (fem grupper). Vi har også simulert levetider, som etterligner virkelige data, med Cox- og additiv modell. Trunkeringstider og kovariater ble tilfeldig trukket fra datasettet. Åtte kombinasjoner av de tre kovariatene, som vi har betegnet med A-H, ble brukt i eksemplene og til sammenligningene. Alle analysene er utført separat for menn og kvinner.

I Kapittel 3 presenterte vi stratifisert Cox-modell for konkurrerende dødsårsaker. Vi estimerte regresjonskoeffisienter for denne modellen, og brukte de videre til å estimere overlevelsessannsynligheter, årsaksspesifikke hasardrater og kumulative insidensfunksjoner for personene A-H. I Kapittel 4 introduserte vi Fine-Gray metoden. I siste del av kapitlet sammenlignet vi resultatene fra datasettet for begge modellene. Vi plottet estimatene for de kumulative baseline hasardene i begge modellene, som tilhører en person med blodtrykk 135, normal vekt og ikke-røyker. Kurvene for de to metodene var omtrent sammenfallende. Vi beregnet relativ endring av de kumulative insidensfunksjonene ved 50, 60 og 70 år for de åtte personene nevnt over, og observerte ikke store forskjeller mellom estimatene. Dette skyldes generelt lav dødelighet i datasettet.

Videre genererte vi data fra Cox- og additiv modell med Weibull fordelt baseline, og satte sensureringsalderen til 90 år, som er høyere enn i datasettet. I den første simuleringsmodellen var målet å undersøke hvor godt Fine-Gray modellen estimerte de kumulative insidensfunksjonene når Cox var den riktige modellen. Avvikene ble noe større, men tilpasningen var god. Ved å også se på gal modell, var hensikten å undersøke om høyere dødelighet resulterte i større avvik mellom modellene og eventuelt hvilken av metodene som best estimerte dødssannsynlighetene. Vi så ikke noen markante forskjeller her heller. Dersom en er interessert i å modellere de kumulative insidensfunksjonene, så kan Fine-Gray metoden brukes. Den kumulative insidensfunksjonen avhenger bare av subfordelingshasardraten for den årsaken som modelleres.

Fine-Gray modellen er ikke like mye brukt som Cox. Resultatene våre viser at denne metoden gir estimater for de kumulative insidensfunksjonene med omtrent samme nøyaktighet som Cox-modellen. Fordelen med modellen til Fine-Gray er at det er en-til-en forhold mellom subfordelingshasard og kumulativ insidensfunksjon for samme dødsårsak, men det er ikke tilfellet for de årsaksspesifikke hasardene for Cox-modellen.

Som beskrevet i Kapittel 4, så er ikke risikomengden for subfordelinger naturlig. Individer som dør av konkurrerende dødsårsaker forblir i risikomengden til sensureringstiden. Dette resulterer i at subfordelingshasarden ikke har en praktisk tolkning. For å studere dødsrater innen medisin, er det derfor mer hensiktsmessig å modellere med de årsaksspesifikke hasardratene fra Cox-modellen.

Som nevnt innledningsvis, har vi kun sett på situasjoner som er interessante i virkeligheten. For slike tilfeller, ble ikke resultatene fra de to metodene så forskjellige. Dersom vi hadde konstruert mer ekstreme situasjoner, er det rimelig å tenke at avvikene mellom modellene ville blitt større.

Bibliografi

Aalen, O.O., Ø.Borgan, og H.K.Gjessing (2008). *Survival and Event History Analysis.* Springer, New York.

Andersen, P.K., R.B.Geskus, T.de Witte, og H.Putter (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology* **41**, 861-870.

Bakoyannis, G., og G.Toulomi (2012). Practical methods for competing risks data; A review. *Statistical Methods in Medical Research* **21**, 257-272.

Cox, D.R (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **34**, 187-220.

de Wreede, L.C., M.Fiocco, og H.Putter (2011). mstate: An R Package for the Analysis of Competing Risks and Multi-State Models.*Journal of Statistical Software* **38**, 1-30.

Fine, J.P, og R.J.Gray (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risks. *Journal of the American Statistical Association* **94**, 496-509.

Geskus, R.B. (2011). Cause-Specific Cumulative Incidence Estimation and the Fine and Gray Model Under Both Left Truncation and Right Censoring. *The International Biometric Society* **67**, 39-49.

Putter, H. (2011). Tutorial in biostatistics: Competing risks and multi-state models Analyses using the mstate package. http://cran.r-project.org/web/packages/mstate/vignettes/Tutorial.pdf,1-43.

Putter, H, M.Fiocco, og R.B.Geskus (2007). Tutorial in biostatistics : Competing risks and multi-state models. *Statistics in Medicine* **26**, 2389-2430.

Ross, S.M. (2007). *Introduction to Probability Models.* Academic Press, London.

Tillegg A

Tillegg til Kapittel 4

Vi har estimert regresjonskoeffisientene i den proporsjonale modellen for subfordelings hasarder (4.21) med disse fire dødsårsakene :

1. Kreft
2. Hjerte- og karsykdommer, inkludert plutselig død
3. Andre medisinske årsaker
4. Alkoholmisbruk, kronisk leversykdom og ulykker og vold

Tabellene A.1-A.4 og A.5-A.8 viser resultatene for hhv menn og kvinner. I Figur A.1 (menn) og Figur A.2 (kvinner) ser vi de estimerte kumulative indisensratene for alder mellom 40 og 70 år med de åtte kombinasjonene av kovariatene fra Tabell 3.4.

A.1 Tabeller

Tabell A.1: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *kreft* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.045	1.046	0.057	0.427
BMI				
Undervektig	-0.554	0.575	0.719	0.441
Overvektig	-0.036	0.965	0.189	0.849
Fedme	-1.233	0.291	0.721	0.088
Røykegruppe				
Tidligere røyker	0.407	1.502	0.329	0.217
1-9 sigaretter per dag	0.811	2.250	0.386	0.036
10-19 sigaretter per dag	0.996	2.707	0.323	0.002
20+ sigaretter per dag	1.274	3.576	0.345	0.000

Tabell A.2: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *hjerte- og karsykdommer inkludert plutselig død* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.264	1.302	0.041	0.000
BMI				
Undervektig	1.097	2.995	0.358	0.002
Overvektig	0.330	1.391	0.167	0.048
Fedme	0.362	1.436	0.317	0.254
Røykegruppe				
Tidligere røyker	0.783	2.187	0.294	0.008
1-9 sigaretter per dag	1.484	4.410	0.320	0.000
10-19 sigaretter per dag	1.202	3.327	0.296	0.000
20+ sigaretter per dag	1.364	3.913	0.318	0.000

Tabell A.3: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *andre medisinske årsaker* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.037	1.038	0.116	0.751
BMI				
Undervektig	0.957	2.604	0.755	0.205
Overvektig	-0.161	0.851	0.403	0.689
Fedme	-0.441	0.643	1.041	0.672
Røykegruppe				
Tidligere røyker	0.386	1.472	0.602	0.521
1-9 sigaretter per dag	0.044	1.045	0.867	0.960
10-19 sigaretter per dag	0.837	2.309	0.592	0.158
20+ sigaretter per dag	0.619	1.857	0.708	0.382

Tabell A.4: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *alkoholmisbruk, kronisk leversykdom og ulykker og vold* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.099	1.104	0.088	0.259
BMI				
Undervektig	1.021	2.776	0.760	0.179
Overvektig	1.008	2.739	0.337	0.003
Fedme	0.806	2.240	0.648	0.214
Røykegruppe				
Tidligere røyker	-0.687	0.503	0.505	0.174
1-9 sigaretter per dag	0.457	1.579	0.528	0.387
10-19 sigaretter per dag	0.774	2.168	0.405	0.056
20+ sigaretter per dag	0.267	1.306	0.527	0.612

Tabell A.5: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *kreft* for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	-0.002	0.998	0.062	0.975
BMI				
Undervektig	-0.091	0.913	0.525	0.862
Overvektig	0.067	1.069	0.260	0.798
Fedme	0.493	1.637	0.339	0.146
Røykegruppe				
Tidligere røyker	0.485	1.624	0.315	0.123
1-9 sigaretter per dag	0.730	2.075	0.328	0.026
10-19 sigaretter per dag	0.675	1.965	0.305	0.027
20+ sigaretter per dag	1.264	3.539	0.456	0.006

Tabell A.6: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *hjerte- og karsykdommer inkludert plutselig død* for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.276	1.317	0.045	0.000
BMI				
Undervektig	0.050	1.051	0.743	0.947
Overvektig	0.266	1.305	0.327	0.415
Fedme	0.757	2.131	0.377	0.045
Røykegruppe				
Tidligere røyker	0.225	1.252	0.431	0.602
1-9 sigaretter per dag	0.780	2.181	0.399	0.051
10-19 sigaretter per dag	1.132	3.103	0.348	0.001
20+ sigaretter per dag	0.257	1.293	1.034	0.803

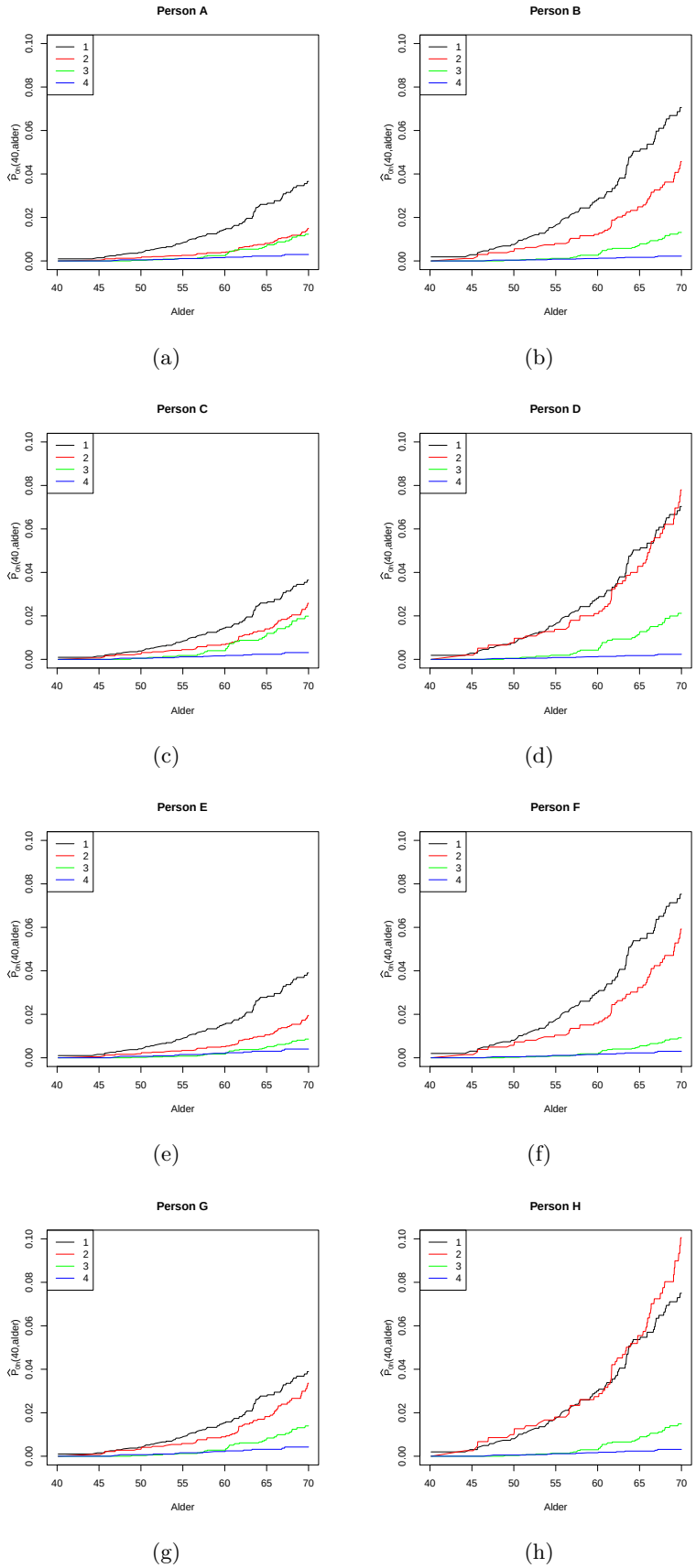
Tabell A.7: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *andre medisinske årsaker* for kvinner.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.242	1.274	0.064	0.000
BMI				
Undervektig	1.713	5.546	0.474	0.000
Overvektig	-0.357	0.700	0.474	0.451
Fedme	0.373	1.452	0.530	0.482
Røykegruppe				
Tidligere røyker	0.225	1.253	0.506	0.656
1-9 sigaretter per dag	0.524	1.688	0.507	0.302
10-19 sigaretter per dag	0.065	1.067	0.541	0.904
20+ sigaretter per dag	1.644	5.174	0.600	0.006

Tabell A.8: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for en subfordelings proporsjonal hasardmodell med død grunnet *alkoholmisbruk, kronisk leversykdom og ulykker og vold for kvinner.*

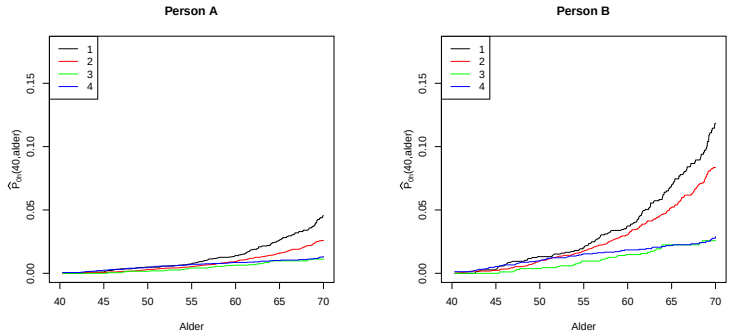
Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.026	1.026	0.167	0.877
BMI				
Undervektig	1.529	4.612	0.872	0.080
Overvektig	0.306	1.358	0.772	0.691
Fedme	1.637	5.138	0.810	0.043
Røykegruppe				
Tidligere røyker	-0.366	0.693	1.120	0.744
1-9 sigaretter per dag	1.149	3.156	0.769	0.135
10-19 sigaretter per dag	-0.297	0.743	1.127	0.792
20+ sigaretter per dag	2.606	13.546	0.802	0.001

A.2 Plott av kumulative insidensfunksjoner

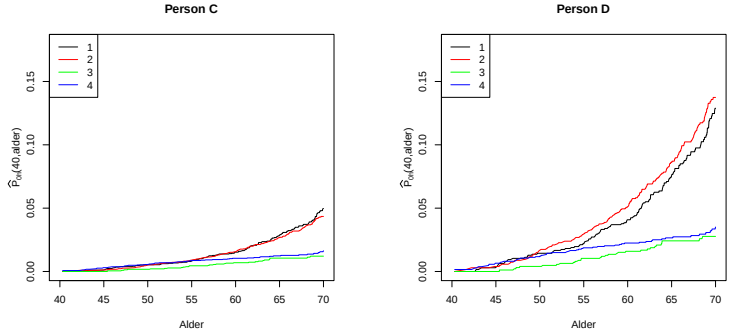


(a) (b) (c) (d) (e) (f) (g) (h)

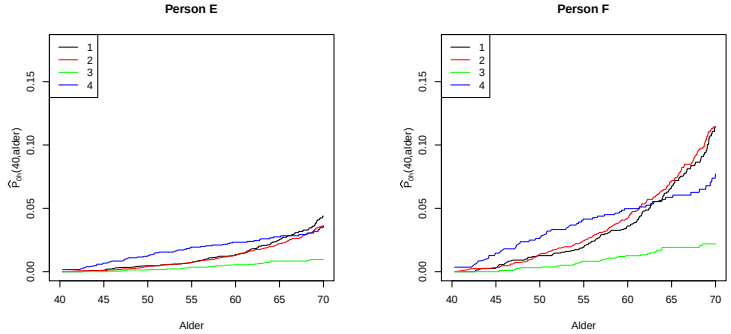
Figur A.2: Estimater av de kumulative insidensfunksjonene med de fire dødsårsakene beskrevet i Eksempel 4.1 for kvinner.



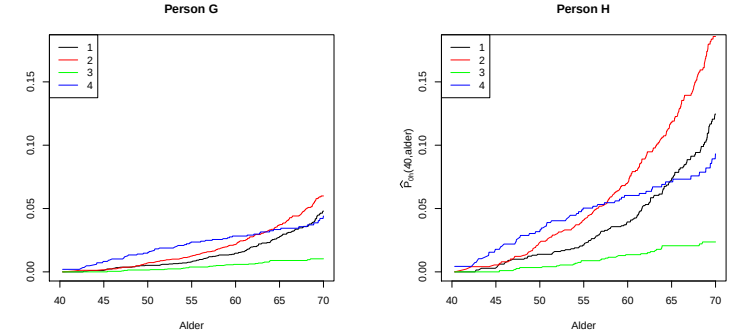
(a) (b)



(c) (d)



(e) (f)



(g) (h)

Figur A.1: av de kumulative insidensfunksjonene med de fire dødsårsakene beskrevet i Eksempel 4.1 for menn.

A.3 Standardfeil

Tabellene A.9-A.10 og A.11-A.12 viser standardfeil for de kumulative insidensfunksjonene beregnet i avsnitt 4.2 for Cox- og subfordelingsmodell.

Tabell A.9: Standardfeil for kumulative insidensfunksjoner i Cox-modellen, menn.

Person	Kovariater	Årsak	Alder		
			50	60	70
A		1	0.002	0.004	0.013
	Blodtrykk 125	2	0.001	0.003	0.007
	Normal vekt	3	0.001	0.003	0.006
	Aldri røykt	4	0.002	0.003	0.006
B		1	0.004	0.008	0.022
	Blodtrykk 125	2	0.003	0.006	0.014
	Normal vekt	3	0.002	0.006	0.010
	10-19 sigaretter pr dag	4	0.004	0.006	0.009
C		1	0.002	0.005	0.015
	Blodtrykk 145	2	0.002	0.004	0.012
	Normal vekt	3	0.001	0.004	0.007
	Aldri røykt	4	0.002	0.004	0.007
D		1	0.005	0.009	0.024
	Blodtrykk 145	2	0.005	0.010	0.022
	Normal vekt	3	0.002	0.006	0.010
	10-19 sigaretter pr dag	4	0.005	0.008	0.011
E		1	0.002	0.004	0.013
	Blodtrykk 125	2	0.001	0.004	0.010
	Overvektig	3	0.001	0.003	0.005
	Aldri røykt	4	0.005	0.008	0.013
F		1	0.004	0.008	0.022
	Blodtrykk 125	2	0.004	0.008	0.019
	Overvektig	3	0.002	0.006	0.009
	10-19 sigaretter pr dag	4	0.009	0.014	0.021
G		1	0.002	0.005	0.014
	Blodtrykk 145	2	0.002	0.006	0.016
	Overvektig	3	0.001	0.003	0.006
	Aldri røykt	4	0.006	0.010	0.015
H		1	0.005	0.009	0.022
	Blodtrykk 145	2	0.006	0.012	0.026
	Overvektig	3	0.002	0.006	0.009
	10-19 sigaretter pr dag	4	0.011	0.016	0.022

Tabell A.10: Standardavvik for kumulative insidensfunksjoner i Cox-modellen, kvinner.

Person	Kovariater	Årsak	Alder		
			50	60	70
A		1	0.001	0.004	0.008
	Blodtrykk 125	2	0.001	0.001	0.005
	Normal vekt	3	0.000	0.001	0.005
	Aldri røykt	4	0.000	0.001	0.002
B		1	0.003	0.004	0.018
	Blodtrykk 125	2	0.002	0.001	0.013
	Normal vekt	3	0.000	0.001	0.007
	10-19 sigaretter pr dag	4	0.000	0.001	0.002
C		1	0.001	0.004	0.009
	Blodtrykk 145	2	0.001	0.001	0.008
	Normal vekt	3	0.000	0.001	0.007
	Aldri røykt	4	0.001	0.001	0.002
D		1	0.003	0.004	0.019
	Blodtrykk 145	2	0.004	0.001	0.022
	Normal vekt	3	0.001	0.001	0.010
	10-19 sigaretter pr dag	4	0.001	0.001	0.003
E		1	0.001	0.004	0.010
	Blodtrykk 125	2	0.001	0.001	0.006
	Overvektig	3	0.000	0.001	0.004
	Aldri røykt	4	0.001	0.001	0.003
F		1	0.003	0.004	0.022
	Blodtrykk 125	2	0.003	0.001	0.019
	Overvektig	3	0.000	0.001	0.006
	10-19 sigaretter pr dag	4	0.001	0.001	0.004
G		1	0.002	0.004	0.010
	Blodtrykk 145	2	0.002	0.001	0.010
	Overvektig	3	0.000	0.001	0.006
	Aldri røykt	4	0.001	0.001	0.003
H		1	0.003	0.004	0.022
	Blodtrykk 145	2	0.006	0.001	0.030
	Overvektig	3	0.000	0.001	0.009
	10-19 sigaretter pr dag	4	0.001	0.001	0.004

Tabell A.11: Standardavvik for kumulative insidensfunksjoner for subfordelingsmodellen, menn.

Person	Kovariater	Årsak	Alder		
			50	60	70
A		1	0.002	0.004	0.014
	Blodtrykk 125	2	0.001	0.003	0.008
	Normal vekt	3	0.001	0.004	0.006
	Aldri røykt	4	0.002	0.004	0.006
B		1	0.004	0.008	0.025
	Blodtrykk 125	2	0.003	0.006	0.016
	Normal vekt	3	0.002	0.006	0.010
	10-19 sigaretter pr dag	4	0.004	0.007	0.010
C		1	0.002	0.005	0.015
	Blodtrykk 145	2	0.002	0.005	0.012
	Normal vekt	3	0.001	0.004	0.007
	Aldri røykt	4	0.003	0.005	0.007
D		1	0.005	0.009	0.027
	Blodtrykk 145	2	0.005	0.010	0.026
	Normal vekt	3	0.002	0.007	0.011
	10-19 sigaretter pr dag	4	0.005	0.008	0.012
E		1	0.002	0.004	0.014
	Blodtrykk 125	2	0.001	0.004	0.010
	Overvektig	3	0.001	0.003	0.006
	Aldri røykt	4	0.005	0.009	0.014
F		1	0.004	0.009	0.026
	Blodtrykk 125	2	0.004	0.009	0.023
	Overvektig	3	0.002	0.006	0.010
	10-19 sigaretter pr dag	4	0.010	0.016	0.024
G		1	0.001	0.003	0.007
	Blodtrykk 145	2	0.001	0.003	0.009
	Overvektig	3	0.000	0.002	0.005
	Aldri røykt	4	0.005	0.008	0.012
H		1	0.005	0.009	0.027
	Blodtrykk 145	2	0.006	0.014	0.034
	Overvektig	3	0.002	0.006	0.010
	10-19 sigaretter pr dag	4	0.012	0.018	0.026

Tabell A.12: Standardavvik for kumulative insidensfunksjoner for subfordelingsmodellen, kvinner.

Person	Kovariater	Årsak	Alder		
			50	60	70
A		1	0.002	0.004	0.009
	Blodtrykk 125	2	0.001	0.001	0.005
	Normal vekt	3	0.000	0.001	0.005
	Aldri røykt	4	0.000	0.001	0.002
B		1	0.003	0.008	0.019
	Blodtrykk 125	2	0.002	0.004	0.014
	Normal vekt	3	0.000	0.001	0.007
	10-19 sigaretter pr dag	4	0.000	0.001	0.002
C		1	0.002	0.004	0.009
	Blodtrykk 145	2	0.001	0.003	0.008
	Normal vekt	3	0.000	0.002	0.007
	Aldri røykt	4	0.001	0.001	0.002
D		1	0.003	0.008	0.020
	Blodtrykk 145	2	0.004	0.007	0.025
	Normal vekt	3	0.001	0.002	0.011
	10-19 sigaretter pr dag	4	0.000	0.001	0.003
E		1	0.002	0.004	0.010
	Blodtrykk 125	2	0.001	0.002	0.006
	Overvektig	3	0.000	0.001	0.004
	Aldri røykt	4	0.001	0.002	0.003
F		1	0.004	0.010	0.024
	Blodtrykk 125	2	0.003	0.006	0.021
	Overvektig	3	0.000	0.001	0.006
	10-19 sigaretter pr dag	4	0.001	0.002	0.003
G		1	0.002	0.004	0.010
	Blodtrykk 145	2	0.002	0.003	0.010
	Overvektig	3	0.000	0.001	0.006
	Aldri røykt	4	0.001	0.002	0.003
H		1	0.004	0.010	0.024
	Blodtrykk 145	2	0.005	0.010	0.034
	Overvektig	3	0.000	0.002	0.009
	10-19 sigaretter pr dag	4	0.001	0.002	0.004

Tillegg B

Tillegg til Kapittel 5

B.1 Tabeller

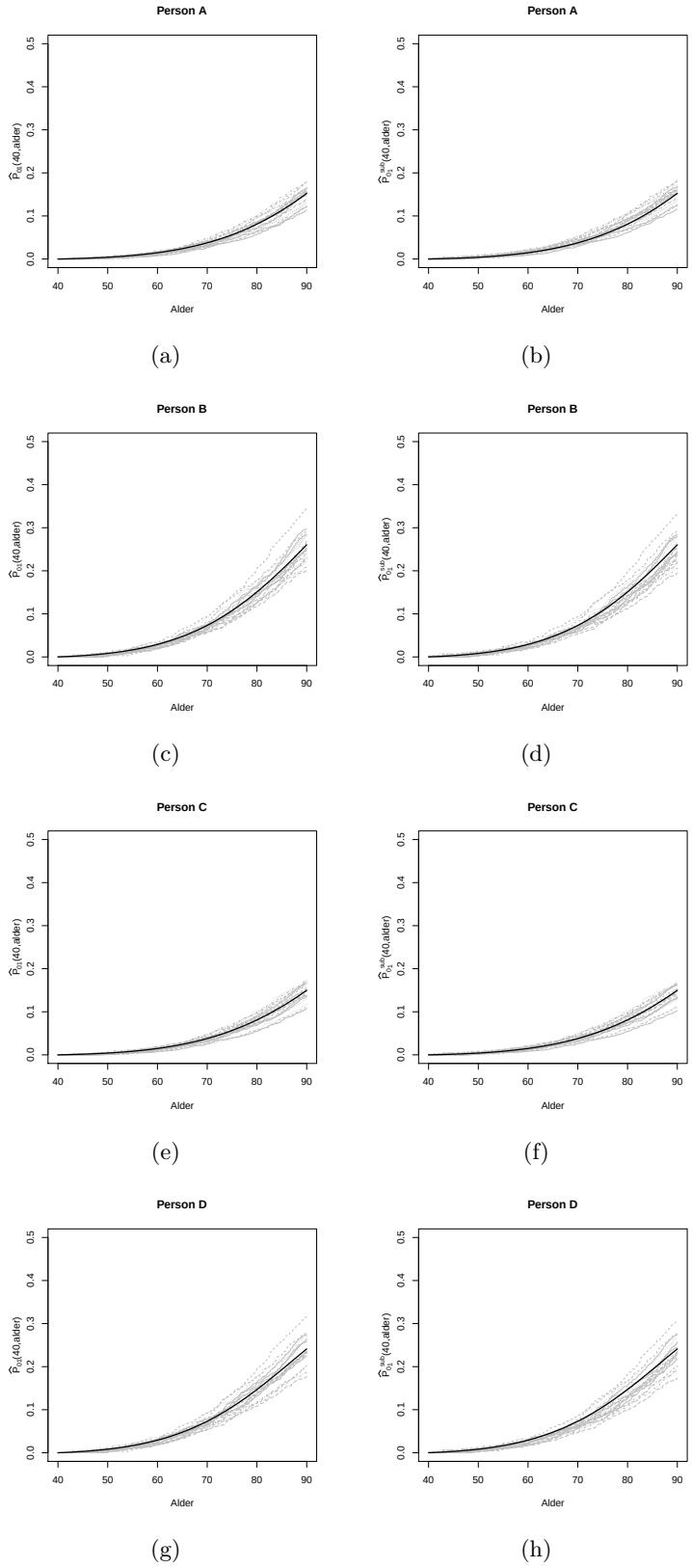
Tabell B.1: Skjevhet og kvadratisk avvik for estimerte kumulative insidensfunksjoner beregnet med Cox- og subfordelingsmodell med dødsårsaken *kreft* for kvinner. Levetidene er trukket fra Cox-modellen (5.13) med $B = 100$ simuleringer og $n = 1000$ levetider i hver simulering.

Person	Kovariater	Alder	$P_{0h}(t\|\mathbf{x})$	Cox		Fine-Gray	
				Skjevhet	10^6 Kv.avvik	Skjevhet	10^6 Kv. avvik
A	Blodtrykk 125 Normal vekt Aldri røykt	50	0.004	-0.001	4.322	0.000	4.669
		60	0.015	-0.001	12.510	0.001	14.662
		70	0.038	-0.002	44.005	0.003	56.947
		80	0.081	-0.002	140.155	0.005	170.708
		90	0.153	-0.002	344.912	0.001	349.141
B	Blodtrykk 125 Normal vekt 10-19 sigaretter pr dag	50	0.008	-0.002	17.468	-0.001	13.946
		60	0.029	-0.003	56.060	-0.002	52.535
		70	0.073	-0.003	197.941	-0.004	198.376
		80	0.151	-0.002	526.703	-0.006	542.890
		90	0.260	-0.001	1141.646	-0.008	1165.604
C	Blodtrykk 145 Normal vekt Aldri røykt	50	0.004	-0.001	4.713	-0.000	4.206
		60	0.015	-0.002	14.804	-0.000	13.910
		70	0.038	-0.002	57.736	-0.001	57.393
		80	0.081	-0.003	200.777	-0.002	198.747
		90	0.150	-0.003	529.814	-0.008	547.933
D	Blodtrykk 145 Normal vekt 10-19 sigaretter pr dag	50	0.008	-0.002	18.251	-0.001	13.409
		60	0.029	-0.003	56.518	-0.005	60.545
		70	0.073	-0.004	206.983	-0.009	253.922
		80	0.147	-0.004	557.201	-0.014	673.219
		90	0.241	-0.004	1114.808	-0.008	1120.560
E	Blodtrykk 125 Overvektig Aldri røykt	50	0.004	-0.001	5.217	0.000	5.609
		60	0.016	-0.001	16.878	0.001	19.263
		70	0.040	-0.001	57.763	0.002	70.493
		80	0.086	-0.001	206.695	0.004	235.891
		90	0.161	-0.001	541.330	0.001	551.113
F	Blodtrykk 125 Overvektig 10-19 sigaretter pr dag	50	0.009	-0.002	22.072	-0.001	18.312
		60	0.031	-0.002	81.347	-0.002	77.540
		70	0.078	-0.002	302.664	-0.004	301.261
		80	0.158	-0.001	854.400	-0.006	862.011
		90	0.266	0.001	1977.461	-0.002	1984.617
G	Blodtrykk 145 Overvektig Aldri røykt	50	0.004	-0.001	5.502	-0.000	4.925
		60	0.016	-0.002	18.143	-0.001	17.596
		70	0.040	-0.002	67.925	-0.002	69.893
		80	0.086	-0.002	246.330	-0.003	251.971
		90	0.156	-0.003	645.261	-0.008	667.746
H	Blodtrykk 145 Overvektig 10-19 sigaretter pr dag	50	0.009	-0.002	22.471	-0.002	17.197
		60	0.031	-0.003	77.364	-0.005	83.455
		70	0.077	-0.003	292.641	-0.010	344.433
		80	0.152	-0.003	795.261	-0.013	878.636
		90	0.241	-0.002	1722.760	0.003	1719.070

Tabell B.2: Skjevhet og kvadratisk avvik for estimerte kumulative insidensfunksjoner beregnet med Cox- og subfordelingsmodell med dødsårsaken *hjerte- og karsykdommer* for kvinner. Levetidene er trukket fra Cox-modellen (5.13) med $B = 100$ simuleringer og $n = 1000$ levetider i hver simulering.

Person	Kovariater	Alder	$P_{0h}(t\|\mathbf{x})$	Cox		Fine-Gray	
				Skjevhet	10^6 Kv.avvik	Skjevhet	10^6 Kv. avvik
		50	0.001	-0.000	0.456	0.000	0.735
	Blodtrykk 125	60	0.005	-0.000	1.647	0.000	2.205
A	Normal vekt	70	0.015	-0.001	9.667	0.001	12.561
	Aldri røykt	80	0.039	-0.001	46.530	0.002	52.132
		90	0.088	-0.002	174.913	-0.001	176.507
		50	0.003	-0.001	4.163	-0.000	5.402
	Blodtrykk 125	60	0.014	-0.001	12.804	-0.001	13.654
B	Normal vekt	70	0.043	-0.002	60.255	-0.002	63.219
	10-19 sigaretter pr dag	80	0.109	-0.005	257.815	-0.005	268.105
		90	0.223	-0.007	827.182	-0.009	845.693
		50	0.002	-0.000	1.418	0.000	2.265
	Blodtrykk 145	60	0.008	-0.001	4.829	0.001	6.709
C	Normal vekt	70	0.026	-0.001	28.816	0.002	38.091
	Aldri røykt	80	0.068	-0.002	139.830	0.003	163.563
		90	0.149	-0.002	496.787	-0.001	525.069
		50	0.005	-0.001	12.965	-0.000	16.611
	Blodtrykk 145	60	0.024	-0.002	37.672	-0.002	41.761
D	Normal vekt	70	0.075	-0.003	181.579	-0.004	191.611
	10-19 sigaretter pr dag	80	0.184	-0.006	716.107	-0.009	763.236
		90	0.354	-0.006	1782.309	-0.010	1884.620
		50	0.001	-0.000	0.786	0.000	1.205
	Blodtrykk 125	60	0.006	-0.000	3.517	0.000	4.365
E	Overvektig	70	0.019	-0.001	20.415	0.001	24.339
	Aldri røykt	80	0.050	-0.002	86.417	0.002	91.203
		90	0.112	-0.002	365.306	-0.001	356.298
		50	0.004	-0.001	7.072	-0.000	9.146
	Blodtrykk 125	60	0.018	-0.001	28.381	-0.001	30.796
F	Overvektig	70	0.056	-0.002	158.406	-0.003	168.123
	10-19 sigaretter pr dag	80	0.139	-0.005	662.648	-0.007	702.520
		90	0.275	-0.007	2262.891	-0.009	2385.997
		50	0.002	-0.000	2.353	0.000	3.608
	Blodtrykk 145	60	0.010	-0.001	9.664	0.001	12.230
G	Overvektig	70	0.033	-0.001	54.303	0.002	65.121
	Aldri røykt	80	0.087	-0.002	222.713	0.003	238.332
		90	0.188	-0.002	827.102	-0.003	824.769
		50	0.007	-0.001	21.257	-0.001	27.294
	Blodtrykk 145	60	0.031	-0.002	78.480	-0.002	87.348
H	Overvektig	70	0.097	-0.003	426.211	-0.005	454.366
	10-19 sigaretter pr dag	80	0.231	-0.006	1554.046	-0.011	1660.265
		90	0.425	-0.006	3752.573	-0.007	4063.670

B.2 Plott av kumulative insidensrater



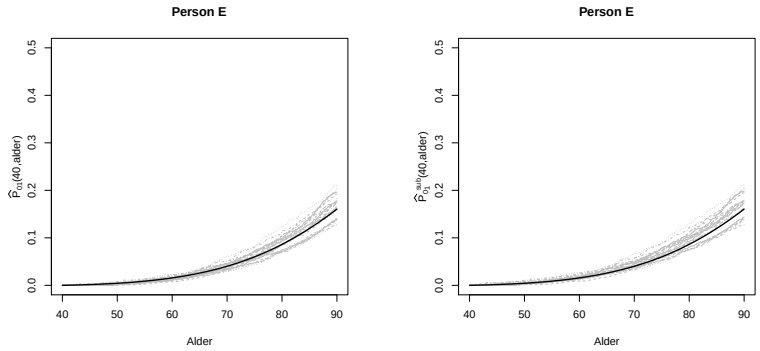
(a) (b)

(c) (d)

(e) (f)

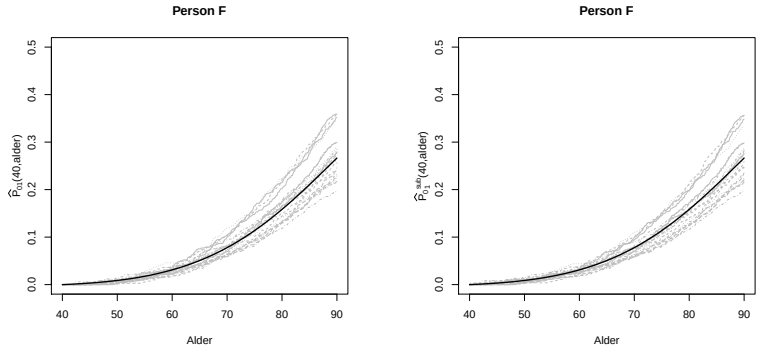
(g) (h)

Figur B.1: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *kreft* for kvinner og personene A-D. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



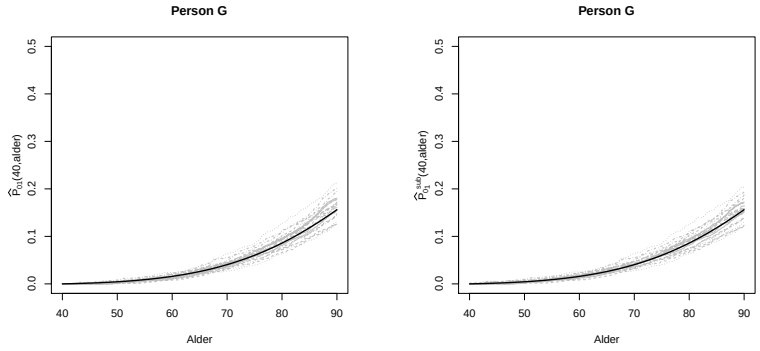
(a)

(b)



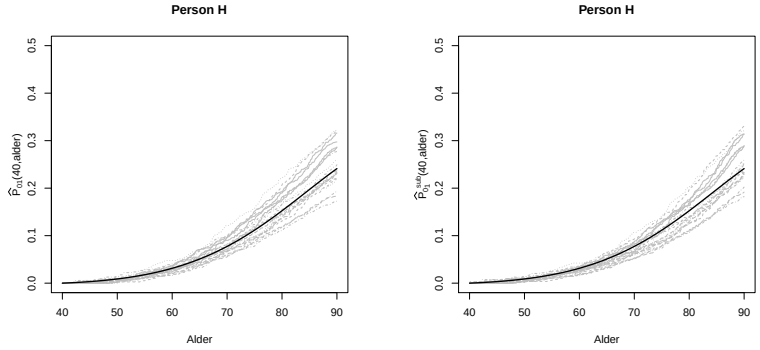
(c)

(d)



(e)

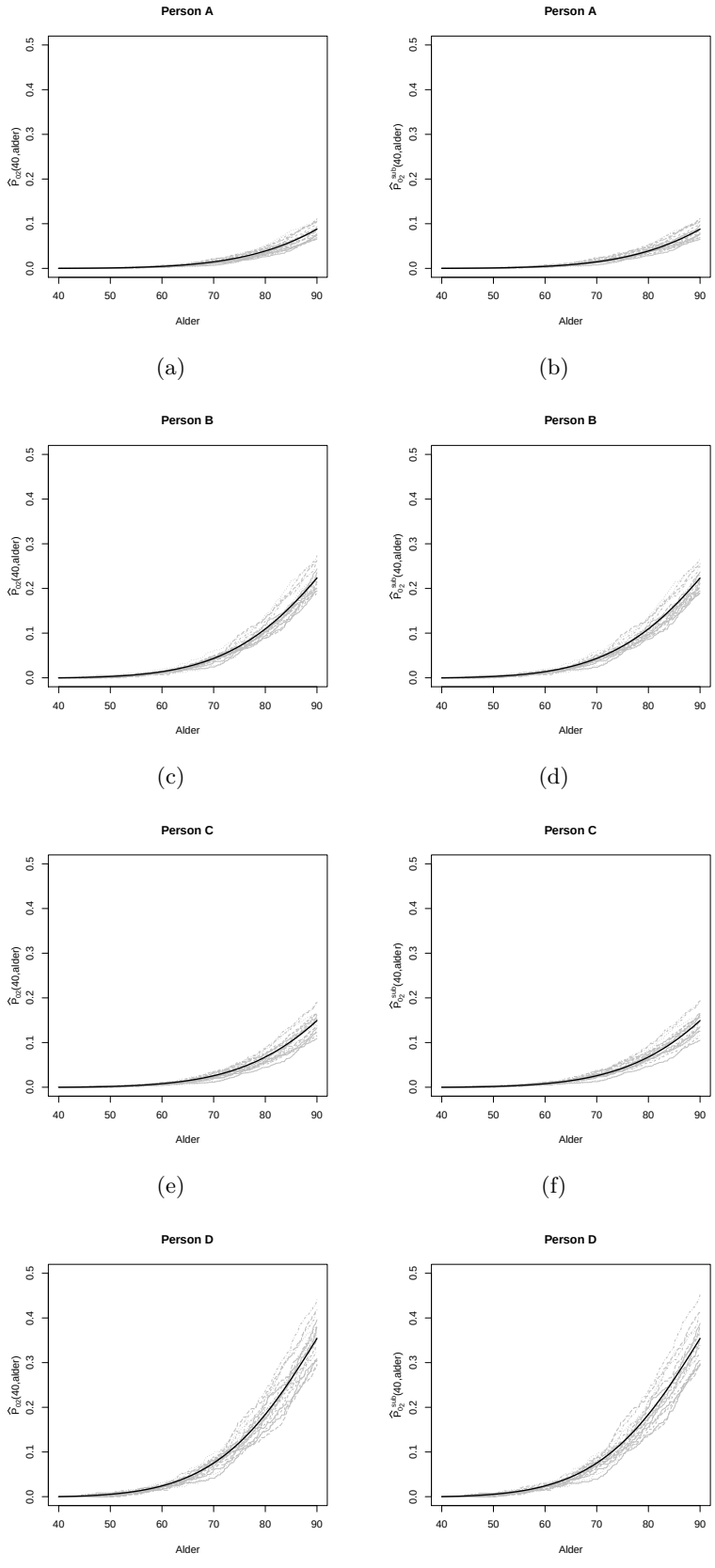
(f)



(g)

(h)

Figur B.2: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *kreft* for kvinner og personene E-H. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



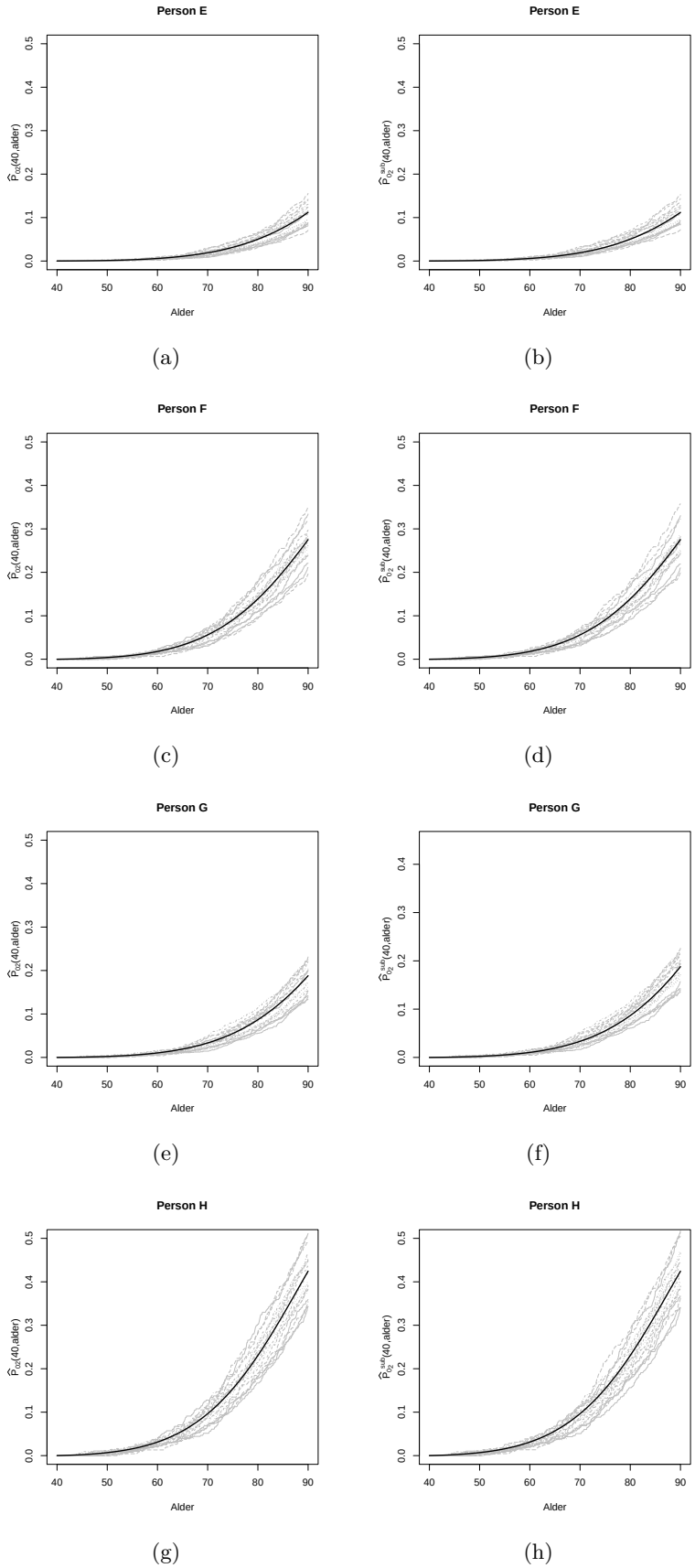
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figur B.3: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *hjerte- og karsykdommer* for kvinner og personene A-D. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figur B.4: Estimerte kumulative insidensfunksjoner beregnet med Cox- (venstre) og subfordelingsmodell (høyre) med dødsårsaken *hjerte- og karsykdommer* for kvinner og personene E-H. Den mørke linjen viser fasiten (5.3) og de grå linjene er simuleringer. Levetidene er trukket fra Cox-modellen (5.13) med $B = 20$ simuleringer, og $n = 1000$ levetider i hver simulering.

B.3 Estimater additiv modell

Tabell B.3: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for den additive modellen (5.17) med død grunnet *kreft* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.000	0.000	0.000	0.290
BMI				
Undervektig	-0.000	0.001	0.001	0.666
Overvektig	-0.000	0.000	0.000	0.996
Fedme	-0.001	0.000	0.001	0.004
Røykegruppe				
Tidligere røyker	0.000	0.000	0.000	0.398
1-9 sigaretter per dag	0.001	0.001	0.001	0.048
10-19 sigaretter per dag	0.001	0.000	0.000	0.001
20+ sigaretter per dag	0.002	0.001	0.001	0.000

Tabell B.4: Estimerte regresjonskoeffisienter, standardfeil og p-verdier (Wald-test) for den additive modellen (5.17) med død grunnet *hjerte- og karsykdommer* for menn.

Kovariat	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	P-verdi
Blodtrykk (per 10)	0.001	0.000	0.000	0.000
BMI				
Undervektig	0.004	0.002	0.002	0.043
Overvektig	0.001	0.000	0.000	0.051
Fedme	0.001	0.001	0.001	0.544
Røykegruppe				
Tidligere røyker	0.001	0.000	0.000	0.045
1-9 sigaretter per dag	0.003	0.001	0.001	0.000
10-19 sigaretter per dag	0.002	0.000	0.000	0.000
20+ sigaretter per dag	0.003	0.001	0.001	0.000