NTNU
Det skapende universitet

# Betydning av feilspesifisert underliggende hasard for estimering av regresjonskoeffisienter og avhengighet i frailty-modeller

**Bjørnar Tumanjan Mortensen**

Master i fysikk og matematikk
Oppgaven levert: Mai 2007
Hovedveileder: Bo Henry Lindqvist, MATH
Biveileder(e): Tron Anders Moger, Biostatistisk institutt, UiO

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgavetekst

I denne oppgaven skal jeg undersøke betydningen av feilspesifisert underliggende hasard for estimering av regresjonskoeffisienter og avhengighet i frailty-modeller. Dette gjøres ved å simulere data med en annen underliggende hasard enn den som antas under estimeringen.

Oppgaven gitt: 15. januar 2007
Hovedveileder: Bo Henry Lindqvist, MATH

# Forord

Med denne oppgaven har jeg fullført den femårige masterutdanningen industriell matematikk ved NTNU. Oppgaven er gjort gjennom vårsemesteret 2007 ved biostatistisk institutt, UiO, som har lagt forholdene svært godt til rette og som jeg har fått et meget godt inntrykk av. Ekstern veileder Tron Anders Moger ved biostatistisk institutt har vært meget hjelpsom og engasjert og en stor takk rettes derfor til ham.

Bjørnar Mortensen, Oslo, mai 2007

## Sammendrag

Med levetidsdata for et stort antall familier kan man bruke frailty-modeller til å finne risikofaktorer og avhengighet innad i familien. En måte å gjøre dette på er å anta en realistisk fordeling for frailty-variabelen og en fordeling for den underliggende hasarden. Det er ikke gjort noen store undersøkelser om betydningen av feilspesifisert underliggende hasard i frailty-modeller tidligere. Grunnen til dette er at det har vært vanlig å anta en ikke-parametrisk underliggende hasard. Dette er mulig for enkle frailty-modeller, men for frailty-modeller med ulik grad av korrelasjon innen en familie blir dette straks svært vanskelig. Derfor er det interessant å undersøke betydningen av feilspesifisert underliggende hasard. I hele denne oppgaven antar vi at den underliggende hasarden er Weibullfordelt. Frailty-fordelingen antas å være enten gamma- eller stablefordelt. Vi simulerer data der den underliggende hasarden er enten Gompertzfordelt, badekarformet eller log-logistisk fordelt. Basert på sannsynlighetsmaksimeringsestimatoren for avhengigheten og regresjonsparametrene undersøker vi betydningen av feilspesifisert underliggende hasard.

Simuleringene viser at dersom det er et stor variasjon i levetidene og et stort sprik mellom virkelig og tilpasset underliggende hasard, underestimeres både risikofakorene og avhengigheten i relativt stor grad. Dette gjelder både når frailty-variabelen er stablefordelt og når den er gammafordelt. Enda mer alvorlig er det dersom også frailty-fordelingen er feilspesifisert.

# Innhold

# 1 Innledning

I medisinsk forskning støter man ofte på levetidsdata. Levetidsdata inneholder informasjon om tiden fra et startpunkt og frem til en hendelse, for eksempel en bestemt sykdom, inntreffer. Som oftest er det mange individer i kohorten der hendelsen aldri har inntruffet. Det vil si at vi har høyresensurerte data, og dette gjør at man må analysere slike data ved hjelp av spesielle metoder. Gjennomsnittlig levetid er for eksempel sjelden særlig nyttig å regne ut, i og med at vi ikke har levetiden til samtlige individer. Dessuten er levetidsdata som oftest ikke normalfordelte. På grunn av dette er det svært vanlig å bruke Cox-regresjon (Cox, 1972) istedenfor lineær regresjon for å finne ut hvilken effekt potensielle risikofaktorer har på levetiden.

I Cox-regresjon antar man ingen fordelingsform for levetidene. Imidlertid gjør man en sterk antagelse om at effekten av risikofaktorene er konstant over tid, og at effekten er lineær på en logaritmisk skala. Istedenfor å benytte levetiden som avhengig variabel av risikofaktorene, benytter man heller hasarden. Hasarden kan tolkes som risikoen (per tidsenhet) for at hendelsen inntreffer på et bestemt tidspunkt, gitt at den ikke har inntruffet fram til dette tidspunktet. Den enkleste Cox-regresjonsmodellen i levetidsanalyse definerer hasarden for individ $i$ som

$$h(t) = h_0(t) \exp\left(\boldsymbol{\beta}^T \mathbf{Z}_i\right). \tag{1}$$

Her er $h_0(t)$ den underliggende hasarden som er felles for alle individer, $\boldsymbol{\beta}$ er en vektor av ukjente regresjonsparametre og $\mathbf{Z}_i$ er kovariatvektoren for individ $i$.

For en bestemt sykdom kan man ha korrelasjon mellom levetidene innen familier. Korrelasjonen kan skyldes faktorer som er vanskelige å måle direkte. For eksempel kan det være vanskelig å måle genetiske komponenter og bruke dette som en risikofaktor. Med familiedata kan man introdusere en frailty-variabel $Y$ til å si noe om korrelasjonen i levetider innen familier. Hasarden for individ $i$ er da gitt ved

$$h\left(t\right) = Y h_0\left(t\right) \exp\left(\boldsymbol{\beta}^T \mathbf{Z}_i\right). \tag{2}$$

Her vil den uobserverte heterogeniteten fanges opp i frailty-variabelen $Y$, mens Cox-leddet, $\exp(\boldsymbol{\beta}^T \mathbf{Z}_i)$, fanger opp den observerte heterogeniteten. Noen tidlige og viktige artikler om frailty-modeller er gjort av Clayton (1978), Vaupel and Yashin (1985a) og Hougaard (1986b). Selve begrepet frailty ble først introdusert av Vaupel *et al.* (1979). Frailty-modeller er fremdeles i kraftig utvikling.

Den enkleste frailty-modellen innebærer at alle individer innen en familie har lik verdi av frailty-variabelen, denne modellen blir kalt shared frailty-modell. Ved å benytte denne modellen kan man se på graden av korrelasjon etter å ha trukket fra virkninger av kovariater som slektningene har felles. At

det er en opphopning av for eksempel lungekreft innen en familie, kan skyldes at det er mange i familien som røyker. Etter at vi har trukket fra virkningen av denne risikofaktoren, vil vi se at korrelasjonen i tid til lungekreft mellom slektningene reduseres, siden lungekreft hovedsaklig skyldes røyking. Ser vi imidlertid på en arvelig sykdom, vil vi mest sannsynlig finne korrelasjon mellom slektningene selv etter å ha trukket fra virkninger av kovariater. Dersom frailtyleddet viser at vi har korrelasjon, må det altså være slik at det fins uobserverte miljømessige eller genetiske komponenter som er felles for individer i samme familie. Vi antar at $Y$ er konstant over tid. Man kan si at mens den underliggende hasarden, $h_0(t)$, beskriver individvariasjon, så beskriver frailty-variabelen gruppevariasjon. Dersom $Y$ for en bestemt familie er stor, kan dette tolkes som at denne familien har store uobserverte risikofaktorer, noe som gir en høy risiko for sykdom. Tilsvarende kan lav $Y$ for en bestemt familie tolkes som at det er liten sjanse for at individene i denne familien får sykdommen. Dette vil si at dersom vi har stor variasjon i $Y$ for de ulike familiene, så kan det finnes uobserverte risikofaktorer som er felles for de i samme familie. Tid til sykdommen vil være avhengig innad i familiene og uavhengig mellom familiene. Det er $Y$ alene som modellerer korrelasjonen. Man kan også benytte en frailty-modell uten kovariater, kun for å få et mål på avhengigheten. Dette kan være aktuelt for eksempel på registerdata der få eller ingen kovariater er tilgjengelig. Denne modellen blir da et spesialtilfelle av (2).

Ofte er målet vårt å estimere korrelasjonen og regresjonsparametrene, og da må man spesifisere $Y$ nærmere. En nærliggende mulighet er å benytte en ikke-parametrisk $Y$. Dette kan være hensiktsmessig dersom man ser på $Y$ som støy og kun er ute etter å finne effekten av kovariatene, men som regel er man mer interessert i å finne et godt estimat også for korrelasjonen. Dette er vanskelig når man benytter en ikke-parametrisk $Y$ fordi man da generelt må bruke asymptotisk statistisk inferens (Hougaard, 2000) som kan fungere godt når antall individer i hver familie er stort, men dårlig når antall individer i hver familie er forholdsvis lavt. Derfor er det mer fornuftig å anta en apriorifordeling for $Y$. Det er praktisk å anta en fordeling for $Y$ som har en enkel Laplacetransformasjon. Grunnen til dette er at overlevelsesfunksjonen, $S(t)$, det vil si sannsynligheten for at individ $i$ lever lenger enn $t$, er gitt som

$$\begin{aligned} S(t) &= P(T > t) \\ &= E\left[S\left(t|Y\right)\right] \\ &= E\left[\exp\left(-H\left(t\right)\right)\right] \\ &= E\left[\exp\left(-Y H_0\left(t\right)\exp\left(\boldsymbol{\beta}^T \mathbf{Z}_i\right)\right)\right] \\ &= L_Y\left(H_0\left(t\right)\exp\left(\boldsymbol{\beta}^T \mathbf{Z}_i\right)\right), \end{aligned} \tag{3}$$

der $L_Y(\bullet) = \int \exp(-sy)\, f(y) dy$ er den Laplacetransformerte av $Y$ og $H(t) = \int_0^t h(u)du$. Tilsvarende er $H_0(t) = \int_0^t h_0(u)du$. Den vanligste antagelsen er at

$Y$ er gammafordelt, siden dette fører til enkel regning og gir en enkel Laplace-transformasjon. En annen vanlig fordeling er PVF-fordelingen (power variance function) (Hougaard, 2000). PVF-fordelingen har en ekstra parameter og inkluderer både gammafordelingen, invers-normalfordelingen og positive stable-fordelingen. Den kan derfor benyttes for å gi en bedre tilpasning til data. I denne oppgaven brukes gammafordelingen og stablefordelingen som frailty-fordelinger.

Som for frailty-variabelen må man også spesifisere den underliggende hasarden $h_0(t)$. I en shared frailty-modell er det enkelt å regne med en ikke-parametrisk underliggende hasardfunksjon, og dette er implementert i statistisk programvare som S-plus og R. I mer kompliserte modeller tar man imidlertid hensyn til at ikke alle individer i samme familie har lik grad av korrelasjon. Et eksempel er en kjernefamilie med mor, far og barn. Her kan man dele frailtyen inn i komponenter for felles og individuelt miljø i tillegg til genetikk. Felles miljø kan ha samme verdi for alle i familien. Den genetiske komponenten vil la mor og far være uavhengige, mens foreldre og barn har korrelasjon $\frac{1}{2}$ siden de i gjennomsnitt deler halvparten av genene. Komponenten for individuelt miløe er uavhengig for alle. Da er $Y$ sammensatt av flere fordelinger. Ulempen med slike modeller er at en ikke-parametrisk $h_0(t)$ fører til at estimeringen av parametrene blir vanskelig sammenlignet med en parametrisk $h_0(t)$.

Det er vist at en feilspesifisert gammafordeling for frailty-variabelen $Y$ har liten betydning for estimeringen av parametrene i den underliggende hasarden og regresjonsparametrene, i verste fall blir estimatene rundt 10 % feil (Hsu og Gorfine, 2007). I andre tilfeller er ikke det viktigste at $h_0(t)$ er korrekt bare man får gode estimater av korrelasjonen og regresjonsparametrene. En praktisk fordeling som ofte brukes som $h_0(t)$ er Weibullfordelingen, hovedsaklig fordi hasardene for mange sykdommer ser nettopp Weibullfordelte ut. Erfaringer fra tidligere analyser kan indikere at feilspesifisert $h_0(t)$ ikke har så stor betydning (Moger *et al.,* submitted og Moger, 2004), men det er ønskelig med en grundigere og mer systematisk kartlegging av skjevheten og standardfeilen til parametrene. Hvis vi kan vise at feilspesifisert Weibullhasard har liten betydning i en shared frailty-modell, indikerer dette at feilspesifisert Weibullhasard i en mer komplisert modell også kan ha liten betydning. Dermed kan man stole mer på resultater man får ved å bruke modeller med Weibullfordelt hasard.

I kapittel 2 skal vi først se på noen grunnleggende begreper i levetidsanalyse. I tillegg skal vi se hvordan vi kan estimere parametrene og standardavvikene til disse. I neste kapittel presenteres noen vanlige fordelinger for hasarden, nemlig Weibullfordelingen, Gompertzfordelingen, log-logistisk fordeling og badekarfordeling. I kapittel 4 går vi litt nøyere inn på frailty-modeller og utleder likelihooden for en shared frailty-modell. I kapittel 5 undersøker vi hvor godt denne modellen fungerer på genererte data, både når frailty-variabelen er stablefordelt og når den er gammafordelt. Vi skal

anta at $h_0(t)$ er Weibullfordelt, men genererer data der $h_0(t)$ ikke er Weibullfordelt. Til sist i dette kapittelet undersøker vi hvor godt modellen fungerer dersom både frailty-variabelen $Y$ og $h_0(t)$ er feilspesifisert. Til slutt følger en diskusjon i kapittel 6.

# 2 Noen grunnleggende begreper i levetidsanalyse

I dette kapittelet skal vi se på noen grunnleggende begreper i levetidsanalyse og sammenhengen mellom disse. Jeg har brukt kapittel 2-4 i boka til Klein og Moeschberger (2002). La $T$ være tiden inntil en bestemt hendelse inntreffer, for eksempel tiden til sykdom. Da har vi som vanlig en sannsynlighetsfordeling $f(t)$ som viser den relative sannsynligheten for å bli syk ved et bestemt tidspunkt. Videre er det to svært vanlige størrelser, overlevelsesfunksjonen $S(t)$ og hasardfunksjonen (kalles også hasardraten eller hasarden) $h(t)$. Overlevelsesfunksjonen gir sannsynligheten for at hendelsen, for eksempel sykdommen, ikke har inntruffet ved tid $T$, mens hasardfunksjonen gir sannsynligheten for at hendelsen inntreffer i løpet av neste "øyeblikk". Mellom disse tre funksjonene har vi en entydig transformasjon, det vil si at dersom vi kjenner en av disse funksjonene, kan de to andre eksakt bestemmes. Sammenhengen mellom sannsynlighetsfordelingen og overlevelsesfunksjonen er

$$S(t) = P(T > t) = \int_t^\infty f(t)dt,$$

dersom $T$ er kontinuerlig. Hasardfunksjonen er gitt som

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}$$

og ved hjelp av Bayes' formel kan vi skrive

$$h(t) = \frac{f(t)}{S(t)}. \tag{4}$$

Sammenhengen mellom hasarden og overlevelsesfunksjonen er

$$S(t) = \exp\left[-\int_0^t h(u)du\right]. \tag{5}$$

Høyresensurering er vanlig for levetidsdata. Det vil si at for noen individer har hendelsen ikke inntruffet ved undersøkelsens slutt. Det er da opplagt at likelihooden ikke kan konstrueres på vanlig måte når det gjelder høyresensurerte levetidsdata, siden likelihooden vanligvis er gitt som

$$l(\boldsymbol{\eta}) = \prod_i P(T = t_i|\boldsymbol{\eta}) = \prod_i f(t_i|\boldsymbol{\eta}) \underset{(4)}{=} \prod_i S(t_i|\boldsymbol{\eta}) h(t_i|\boldsymbol{\eta}),$$

der $\boldsymbol{\eta}$ er alle parametrene i modellen. Det eneste vi vet om de høyresensurerte dataene er at de har overlevd fram til tid $t_i$, det vil si at de har overlevd *lenger* enn $t_i$. De gir dermed et bidrag til likelihooden lik overlevelsesfunksjonen ved tid $t$, $S(t_i|\boldsymbol{\eta})$. Vi lar $\delta_i = 0$ hvis individ $i$ ikke har opplevd hendelsen. Motsatt

lar vi $\delta_i = 1$ dersom hendelsen har inntruffet for individ $i$. Likelihooden blir dermed

$$\begin{aligned} l(\boldsymbol{\eta}) &= \prod_i \left[P(T = t_i|\boldsymbol{\eta})\right]^{\delta_i} P(T > t_i|\boldsymbol{\eta})^{1-\delta_i} \\ &= \prod_i \left[S(t_i|\boldsymbol{\eta})h\left(t_i|\boldsymbol{\eta}\right)\right]^{\delta_i} S(t_i|\boldsymbol{\eta})^{1-\delta_i} \\ &= \prod_i S(t_i|\boldsymbol{\eta})h\left(t_i|\boldsymbol{\eta}\right)^{\delta_i}. \end{aligned} \tag{6}$$

Likelihooden angir altså sannynligheten for å få dataene som er fått, gitt parametrene i modellen. Det vil si at den er en funksjon av parametrene. Ved å maksimere likelihooden med hensyn på parametrene, finner vi altså de estimatorene for parametrene som maksimerer sannsynligheten for å få akkurat disse dataene. Disse kalles gjerne ML(maximum likelihood)-estimatorer. Av rent praktiske grunner er det vanlig å maksimere log-likelihooden istedenfor likelihooden. Dette gjøres ved å sette $U(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \log l(\boldsymbol{\eta}) = 0$. Siden den logaritmiske funksjonen er en monotont voksende funksjon, forandres ikke ML-estimatorene.

I tillegg til estimatorer er det også interessant å finne estimert kovariansmatrise til parametrene. La $\widehat{\boldsymbol{\eta}}$ være ML-estimatorene for parametrene. Ved Taylorutvikling om de sanne verdiene av parametrene, $\boldsymbol{\eta}_0$, har vi at

$$0 = U(\widehat{\boldsymbol{\eta}}) \approx U(\boldsymbol{\eta}_0) + \frac{\partial}{\partial \boldsymbol{\eta}} U(\boldsymbol{\eta}_0)\left(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\right).$$

Det vil si at

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \approx \mathbf{i}(\boldsymbol{\eta}_0)^{-1} U(\boldsymbol{\eta}_0),$$

der $\mathbf{i}(\boldsymbol{\eta}_0) = -\frac{\partial}{\partial \boldsymbol{\eta}} U(\boldsymbol{\eta}_0)$ er observert informasjonsmatrise. Videre har vi at

$$Var\left(\mathbf{i}(\boldsymbol{\eta}_0)^{-1} U(\boldsymbol{\eta}_0)\right) = \mathbf{i}(\boldsymbol{\eta}_0)^{-1} Var\left(U(\boldsymbol{\eta}_0)\right) \mathbf{i}(\boldsymbol{\eta}_0)^{-1}$$

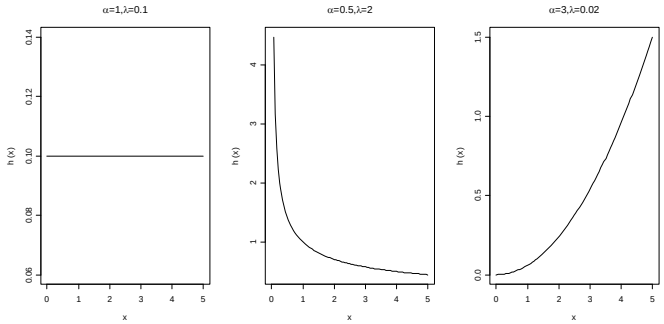
Man kan vise at $U(\boldsymbol{\eta}_0)$ er normalfordelt med forventning 0 og kovariansmatrise $\mathbf{i}(\boldsymbol{\eta}_0)$. Følgelig får $\widehat{\boldsymbol{\eta}}$ kovariansmatrise $\mathbf{i}(\boldsymbol{\eta}_0)^{-1}$. Estimert kovariansmatrise blir da $\mathbf{i}(\widehat{\boldsymbol{\eta}})^{-1}$ og

$$\widehat{\boldsymbol{\eta}} \sim N\left(\boldsymbol{\eta}_0, \mathbf{i}(\boldsymbol{\eta}_0)^{-1}\right).$$

# 3 Vanlige fordelinger for hasarden

| | Hasardfunksjon $h(x)$ | Betingelser |
|---|---|---|
| Weibull | $\alpha\lambda x^{\alpha-1}$ | $\alpha, \lambda > 0, x \geq 0$ |
| Log-logistisk | $\frac{\alpha x^{\alpha-1}\lambda}{1+\lambda x^{\alpha}}$ | $\alpha, \lambda > 0, x \geq 0$ |
| Gompertz | $\lambda e^{\alpha x}$ | $\alpha, \lambda > 0, x \geq 0$ |

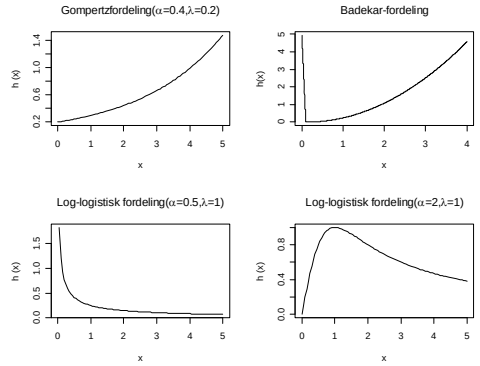
Tabell 1: Aktuelle hasarder




Figur 1: Weibullhasarder

Vi skal undersøke estimatene av korrelasjonen og regresjonsparametrene dersom vi antar at den underliggende hasarden kommer fra Weibullfordelingen, når den i virkeligheten kommer fra en annen fordeling. Det er tre gode grunner til å bruke Weibullfordelingen som underliggende hasard. For det første har den et enkelt matematisk uttrykk, som vist i tabell 1. For det andre har hasarden den egenskapen at avhengig av om $\alpha < 1$, $\alpha = 1$ eller $\alpha > 1$, vil hasardfunksjonen være henholdsvis konstant, avtagende eller økende, så vi slipper å gjøre noen antagelser om dette. Tre eksempler er vist i figur 1. Den tredje og kanskje viktigste grunnen er rett og slett at de observerte hasardene for mange sykdommer ser Weibullfordelte ut.

Det er også andre fordelinger som den underliggende hasarden kan tenkes å komme fra. Vi skal se spesielt på Gompertzfordelingen, den log-logistiske fordelingen og badekarfordelingen. Hasarden for Gompertzfordelingen er vist øverst til venstre i figur 2 og det matematiske uttrykket i tabell **??**. Dersom den underliggende hasarden i virkeligheten er Gompertzfordelt og vi bruker en modell med Weibullfordelt underliggende hasard, kan vi få problemer. Grunnen til dette er at Gompertzhasarden har en startverdi lik $\lambda$, mens Weibullhasarden enten starter i 0 og er økende eller starter i uendelig og er avtagende, eller den kan være konstant. Dermed får Weibullhasarden en dårlig tilpasning ved lave tidspunkt. Ettersom Weibullhasarden vil forsøke å tilpasse seg tidlige tidspunkt så godt som mulig, kan vi dermed få en dårlig




Figur 2: Noen aktuelle hasarder

tilpasning også ved sene tidspunkt.

Nederst i figur 2 ser vi at vi også kan få problemer dersom den underliggende hasarden i virkeligheten er log-logistisk fordelt fordi denne hasarden blant annet først kan øke og deretter avta. Da kan en tilpasning ved hjelp av Weibullfordelingen bli rimelig unøyaktig og dermed også estimatene. Den siste fordelingen vi skal undersøke er badekarfordelingen, vist øverst til høyre i figur 2. Grunnen til navnet er at hasarden kan minne om tverrsnittet av et badekar. Det som kjennetegner denne hasarden er stor risiko for sykdom i starten. Med tiden minsker risikoen, før den flater ut og blir større igjen. I praksis vil hasarden til en rekke levetider forholde seg på denne måten. Et eksempel er hasarden for et menneskes levetid. Fra mennesket er 0 til 1-2 år har det en relativt stor risiko for å dø. Deretter synker risikoen drastisk og er svært lav før den begynner å ta seg opp igjen ved 30-50-årsalderen. Siden en Weibullhasard enten vil øke, avta eller være konstant vil også slike data kunne bli et problem for en modell med Weibullfordelt underliggende hasard.

# 4 Frailty-modeller

Det er vanlig å bruke en multiplikativ frailty-modell, der hasarden for hvert individ er gitt som produktet av frailty-variabelen $Y$, den underliggende hasarden $h_0(t)$ og et Cox regresjonsledd, $\exp(\boldsymbol{\beta}^T\mathbf{Z})$, som i ligning (2). Som nevnt i introduksjonen vil vi benytte en shared frailty-modell, i første omgang med gammafordelt frailty-variabel $Y$, i neste omgang med stablefordelt frailty-variabel. Vi kan ha positiv avhengighet for levetidene innen en familie, mens det er mer komplisert å utvide modellene til negativ avhengighet. I de aller fleste eksempler er dette heller ikke nødvendig, siden det mest naturlige er at levetidene for individer i samme gruppe har ingen eller positiv korrelasjon. Individene er uavhengige gitt $Y$. Som mål for graden av avhengighet er det en nærliggende løsning å bruke variansen til $Y$. Ulempen med denne løsningen er at variansen til $Y$ ikke nødvendigvis er sammenlignbar for to ulike frailty-fordelinger. Derfor må vi finne et annet mål, og et mye brukt mål er Kendall's $\tau$, som er gitt ved

$$\tau = 4\int_0^\infty sL\left(s\right)L^{(2)}(s)ds - 1. \tag{7}$$

Kendall's $\tau$ ligger mellom 0 og 1, og stor $\tau$ vil si høy grad av avhengighet mellom individene i en familie.

For å finne estimatorer for parametrene, regner vi ut log-likelihooden til observasjonene og maksimerer denne, som vist i kapittel 2. La $H_0(t) = \int_0^t h_0(u)du$ være den kumulative underliggende hasarden. Fra ligning (5) får vi at den betingede overlevelsesfunksjonen for $k$ individer i samme familie gitt ved

$$S(t_1,\ldots,t_k|Y) = S(t_1|Y)\cdot\ldots\cdot S(t_n|Y) = \exp\left\{-Y\sum_{i=1}^{k}\exp(\boldsymbol{\beta}^T\mathbf{Z}_i)H_0(t_i)\right\}.$$

Vi antar at vi har høyresensurerte data. Vi lar $\delta_i = 1$ hvis individ $i$ har fått sykdommen og $\delta_i = 0$ hvis individ $i$ ikke har fått sykdommen. Fra ligning (6) får vi dermed at den betingede likelihooden for $k$ individer i en familie er

$$\begin{aligned} l|Y &= \prod_i S(t_i|Y)h\left(t_i|Y\right)^{\delta_i} \\ &= \exp\left\{-Y\sum_{i=1}^{k}\exp(\boldsymbol{\beta}^T\mathbf{Z}_i)H_0(t_i)\right\}\left\{\prod_{i=1}^{k}[Yh_0\left(t_i\right)\exp(\boldsymbol{\beta}^T\mathbf{Z}_i)]^{\delta_i}\right\} \end{aligned} \tag{8}$$

Først antar vi at $Y$ er gammafordelt med sannsynlighetsfordeling $f(y) = \theta^\theta y^{\theta-1}\exp(-\theta y)/\Gamma(\theta)$, $\theta > 0$. Fordelingen har lik form- og skalaparameter og dermed forventning 1. Dette gjøres for at parametrene i modellen skal

være identifiserbare når $h_0(t)$ inneholder både skala- og formparameter. Den $k$'te deriverte av Laplacetransformen til $Y$ med denne fordelingen, er

$$L_Y^{(k)}(s) = (-1)^k \frac{\Gamma(k+\theta)}{\Gamma(\theta)} \frac{\theta^\theta}{(\theta+s)^{k+\theta}}. \tag{9}$$

Samtidig er det kjent at $E\{Y^k \exp(-Ys\} = (-1)^k L_Y^{(k)}(s)$. Da blir den ubetingede likelihooden for en familie

$$\begin{aligned}
l &= S(t_1,\ldots,t_k) \prod_{i=1}^{k} [h(t_i)]^{\delta_i} \\
&= E\left[ \left\{ \exp\left[ -Y \sum_{i=1}^{k} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) H_0(t_i) \right] \right\} \left\{ \prod_{i=1}^{k} [Y h_0(t_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)]^{\delta_i} \right\} \right] \\
&= \left\{ \prod_{i=1}^{k} [h_0(t_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)]^{\delta_i} \right\} (-1)^{\delta_\cdot} L_Y^{(\delta_\cdot)} \left( \sum_{i=1}^{k} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) H_0(t_i) \right) \\
&\underset{(9)}{=} \left\{ \prod_{i=1}^{k} [h_0(t_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)]^{\delta_i} \right\} \frac{\Gamma(\delta_\cdot + \theta)}{\Gamma(\theta)} \times \\
&\quad \times \frac{\theta^\theta}{\left[ \theta + \sum_{i=1}^{k} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) H_0(t_i) \right]^{\delta_\cdot + \theta}},
\end{aligned} \tag{10}$$

der $\delta_\cdot = \sum_{i=1}^{k} \delta_i$. I tillegg antar vi en Weibullfordelt hasard, $h_0(t) = \alpha\lambda t^{\alpha-1}$, og dermed $H(t) = \int_0^t h_0(u)du = \lambda t^\alpha$. Vi setter dette inn i (10) og får at

$$\begin{aligned}
l &= \left\{ \prod_{i=1}^{k} \left[ \alpha\lambda t_i^{\alpha-1} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) \right]^{\delta_i} \right\} \frac{\Gamma(\delta_\cdot + \theta)}{\Gamma(\theta)} \times \\
&\quad \times \frac{\theta^\theta}{\left[ \theta + \sum_{i=1}^{k} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) \lambda t_i^\alpha \right]^{\delta_\cdot + \theta}}.
\end{aligned}$$

Siden individer som ikke er i slekt er uavhengige, blir likelihooden for de $n$ familiene i kohorten

$$\begin{aligned}
l &= \prod_{j=1}^{n} \left\{ \prod_{i=1}^{k_j} [\alpha\lambda t_{i,j}^{\alpha-1} \exp(\boldsymbol{\beta}^T \mathbf{Z}_{i,j})]^{\delta_{i,j}} \right\} \frac{\Gamma(\delta_{\cdot,j} + \theta)}{\Gamma(\theta)} \times \\
&\quad \times \frac{\theta^\theta}{\left[ \theta + \sum_{i=1}^{k_j} \exp(\boldsymbol{\beta}^T \mathbf{Z}_{i,j}) \lambda t_{i,j}^\alpha \right]^{\delta_{\cdot,j} + \theta}},
\end{aligned}$$

der subskript $j$ står for familie $j$. Neste skritt er som vanlig å finne logaritmen av dette, slik at parameterestimeringen blir enklere og mer stabil. Parameterverdiene som maksimerer likelihooden forblir de samme siden $\log x$ er en

monotont voksende funksjon. Log-likelihooden blir

$$\begin{aligned}\log l &= \sum_{j=1}^{n}\left\{\sum_{i=1}^{k_j}\delta_{i,j}\left[\log(\alpha\lambda)+(\alpha-1)\log(t_{i,j})+\boldsymbol{\beta}^T\mathbf{Z}_{i,j}\right]+\right.\\ &\quad +\log\left[\Gamma(\delta_{.,j}+\theta)\right]-\log\left[\Gamma(\theta)\right]+\\ &\quad \left.+\theta\log(\theta)-(\delta_{.,j}+\theta)\log(\theta+\sum_{i=1}^{k_j}\exp\left(\boldsymbol{\beta}^T\mathbf{Z}_{i,j}\right)\lambda t_{i,j}^{\alpha})\right\}.\end{aligned}$$

Når vi antar at frailty-variabelen er stablefordelt, kan log-likelihooden regnes ut på tilsvarende måte. Stablefordelingen uttrykkes som en uendelig sum,

$$f(y)=-\frac{1}{\pi y}\sum_{q=1}^{\infty}\frac{\Gamma(q\theta+1)}{q!}\left(-\frac{\nu}{\theta}\right)^q y^{-q\theta}\sin(\theta q\pi).$$

Av samme grunn som at vi velger å ha kun en parameter i gammafordelingen, velger vi også her å sette $\theta=\nu$. Denne fordelingen har uendelig forventning og varians og er veldig skjev. Fra Hougaard (2000) har vi at den $\delta.$'te deriverte av Laplacetransformen til $Y$ er

$$L_Y^{(\delta.)}=Q\exp\left[-\left(\sum_{i=1}^{k}\exp\left(\boldsymbol{\beta}^T\mathbf{Z}_{i,j}\right)H_0(t_i)\right)^{\theta}\right],$$

der

$$Q=\sum_{m=1}^{\delta.}C_{\delta.,m}(\theta)\,\theta^m\left[\sum_{i=1}^{k}\exp\left(\boldsymbol{\beta}^T\mathbf{Z}_{i,j}\right)H_0(t_i)\right]^{m\theta-\delta.},$$

der

$$\begin{aligned}C_{\delta.,1}(\theta) &= \frac{\Gamma(\delta.-\theta)}{\Gamma(1-\theta)}, C_{\delta.,\delta.}(\theta)=1\\ C_{\delta.,m}(\theta) &= C_{\delta.-1,m-1}(\theta)+[(\delta.-1)-m\theta]\,C_{\delta.-1,m}(\theta).\end{aligned}$$

Ved å bruke ligning (8) og ligning (10) får vi at likelihooden for en familie kan skrives som

$$l=\left\{\prod_{i=1}^{k}[h_0(t_i)\exp(\boldsymbol{\beta}^T\mathbf{Z}_i)]^{\delta_i}\right\}Q(\theta)\exp\left\{-\left[\lambda\sum_{i=1}^{k_j}t_{i,j}^{\alpha}\exp\left(\boldsymbol{\beta}^T\mathbf{Z}_{i,j}\right)\right]^{\theta}\right\}.$$

Dermed blir log-likelihooden, når vi antar stablefordelt frailty-variabel og

Weibullfordelt $h_0(t)$,

$$\begin{aligned}\log l &= \sum_{j=1}^{n}\left\{\sum_{i=1}^{k_j}\delta_{i,j}\left[\log(\alpha\lambda)+(\alpha-1)\log(t_{i,j})+\boldsymbol{\beta}^T\mathbf{Z}_{i,j}\right]+\right.\\ &\left.+\log Q_j(\theta)-\left[\lambda\sum_{i=1}^{k_j}t_{i,j}^{\alpha}\exp\left(\boldsymbol{\beta}^T\mathbf{Z}_{i,j}\right)\right]^{\theta}\right\}. \qquad (11)\end{aligned}$$

Fra ligning (7) får vi at Kendall's $\tau$ når $Y$ er gammafordelt blir $\tau=\frac{1}{1+2\theta}$ og når $Y$ er stablefordelt blir $\tau=1-\theta$. Dermed kan vi finne et estimat for $\tau$ utifra estimatet for $\theta$. Ved å bruke at $Var(f(\theta))\approx\left(\frac{df(\theta)}{d\theta}\right)^2 Var(\theta)$, finner vi dessuten at

$$Var(\tau)\approx\frac{4}{(1+2\theta)^4}Var(\theta)$$

når $Y$ er gammafordelt, og

$$Var(\tau)=Var(\theta)$$

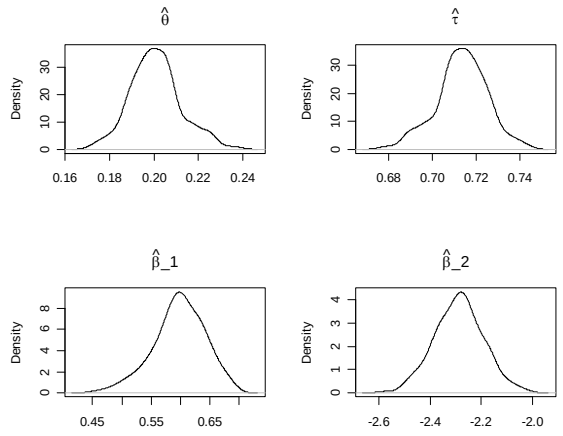
når $Y$ er stablegfordelt. Ved å bruke dette kan vi finne estimert standardavvik til $\tau$.

# 5 Simuleringer

Vi skal undersøke skjevheten og standardavviket til frailty-parameteren $\theta$ og regresjonsparametrene $\beta_i$, $i = 1, \ldots, p$. Vi simulerer $m = 200$ kohorter med $n = 5000$ familier i hver kohort fra modellen (2). Grunnen til at $n$ velges såpass stor er at dette er nødvendig for å generere nok familier med mer enn en hendelse/sykdom. Dermed kan avhengigheten estimeres. Dersom sykdommen er sjelden er det også i virkeligheten nødvendig med stor $n$. Derfor er det vanlig at familiedatasett er relativt store. Med 200 kohorter tar da en simulering omtrent ett døgn.

For å se effekten av både kontinuerlige og dikotome kovariater med både uavhengighet og avhengighet innad i familier, har vi to kovariater i modellen, det vil si $p = 2$. Den første kovariaten, med regresjonsparameter $\beta_1$, er gammafordelt med formparameter lik 0.5 og skalaparameter lik 1, det vil si at den er skjevfordelt. Denne kovariaten er uavhengig for alle individer. Den andre kovariaten, med regresjonsparameter $\beta_2$, er dikotom, det vil si enten 0 eller 1, og er avhengig for individer i samme familie. Denne avhengigheten simuleres på en enkel måte. For hver familie settes det første familiemedlemmet, det vil si et tilfeldig familiemedlem, til å ha verdien 0 med 50% sannsynlighet eller 1 med 50% sannsynlighet. Det andre familiemedlemmet, et annet tilfeldig familiemedlem, får samme verdi på kovariaten som det første med 70% sannsynlighet. Det tredje familiemedlemmet får deretter samme verdi som det andre familiemedlemmet med 70% sannsynlighet og så videre. Antall individer i den $j$'te familien, $k_j$, trekkes slik at vi har 15% sannsynlighet for ett individ i familie $j$, 30% for to individer, 25% for tre, 20% for fire, 6% for fem og 4% for seks. Deretter velges regresjonsparametrene og frailty-parameteren $\theta$, og frailty-variablene $Y$ trekkes fra enten gammafordelingen med form- og skalaparameter $\theta$ eller stablefordelingen med parameter $\theta$. Videre velger vi en fordeling for den underliggende hasarden, inkludert verdier for parametrene.

Gitt modellen i (2) kan vi nå simulere levetidene til samtlige individer. En måte å gjøre dette på er å trekke $u_i \sim Uniform\,(0,1)$ og deretter velge $t_i$ slik at $S\,(t_i|Y) = u_i$ for alle individene. Det vil si at vi må finne $t_i = S^{-1}\,(u_i|Y)$. Den inverse av overlevelsesfunksjonen er enkel å finne når $h_0\,(t)$ er Weibull-, Gompertz-, log-logistisk- eller badekarfordelt. For å få høyresensurerte data trekker vi en normalfordelt grense for hvert individ, $g_i \sim N\left(\mu, \sigma^2\right)$, og dersom levetiden til et individ er høyere enn denne grensen, så sier vi at hendelsen ikke har inntruffet for dette individet. Dersom vi velger å sensurere mange individer, kan vi velge en lav forventning. Da bør vi også passe på å ha en relativt stor varians for i det hele tatt å få hendelser som skjer etter lang tid. For lav varians kan føre til at vi i praksis kun får svært lave tidspunkt for hendelsene og dermed vil formen til den underliggende hasarden bare være relevant ved lave tidspunkt. Etter en simulering bør vi derfor undersøke om hendelsen inntreffer på tidspunkt der den underliggende hasarden har fått




Figur 3: Fordeling til estimater når $h_0(t)$ ∼Weibull($\alpha = 3$, $\lambda = 0.1$), $\tau = 0.714$ og $Y$ er gammafordelt.

sin karakteristiske form. I relle situasjoner er det vanlig med en høy grad av sensurering. I disse simuleringene sensurerer vi derfor omtrent 90 % av individene.

Ved å bruke optim-rutinen i R, finner vi log-likelihoodens maksimum og dermed estimatene for parametrene. De kan så settes inn ligningen for informasjonsmatrisen, slik at vi finner estimerte standardavvik. Disse kan sammenlignes med de empiriske standardavvikene. For for eksempel parameteren $\theta$ er de empiriske standardavvikene gitt som $\frac{1}{m-1}\sum_{k=1}^{m}\left(\widehat{\theta}_k^{MLE} - \overline{\theta}\right)^2$, der $\widehat{\theta}_k^{MLE}$ er estimatet av $\theta$ i den k'te kohorten basert på likelihood-estimatoren (MLE) og $\overline{\theta} = \frac{1}{m}\sum_{k=1}^{m}\widehat{\theta}_k^{MLE}$ er det gjennomsnittlige estimatet for $\theta$. Det er tilsvarende for de andre parametrene. Dersom de empiriske standardavvikene er mye høyere enn de estimerte standardavvikene, kan dette føre til at vi underestimerer usikkerheten til estimatene når vi benytter et reelt datasett.

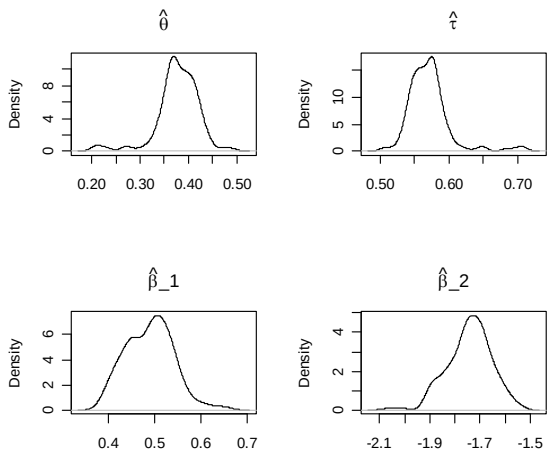
## 5.1 Gammafordelt frailty-variabel

På grunn av de matematiske egenskapene og den enkle Laplacetransformen er det vanlig å anta at frailty-variabelen $Y$ er gammafordelt med form- og skalaparameter $\theta$. I de følgende simuleringene antar vi denne fordelingen for $Y$, og dataene blir simulert med denne frailty-fordelingen.

For å ha noe å sammenligne standardavvikene med, undersøker vi standardavviket for parametrene dersom $h_0(t)$ virkelig er Weibullfordelt (at parametrene blir korrekt estimert er opplagt). Resultatene er vist i tabell 2

| Parameter | Par | Mpar | estSE | empSe |
|---|---|---|---|---|
| $h_0(t)$ Weibull($\alpha = 3, \lambda = 0.1$) | | | | |
| $\theta$ | 0.2 | 0.201 | 0.0116 | 0.0115 |
| $\lambda$ | 0.1 | 0.100 | 0.0065 | 0.0067 |
| $\alpha$ | 3 | 3.000 | 0.0681 | 0.0689 |
| $\beta_1$ | 0.6 | 0.598 | 0.0434 | 0.0443 |
| $\beta_2$ | $-2.3$ | $-2.288$ | 0.0955 | 0.0952 |
| $\tau$ | 0.714 | 0.714 | 0.0117 | 0.0116 |
| $h_0(t)$ Gompertz($\alpha = 1.5, \lambda = 0.005$) | | | | |
| $\theta$ | 0.2 | 0.3775 | 0.0275 | 0.0452 |
| $\beta_1$ | 0.6 | 0.4898 | 0.0431 | 0.0515 |
| $\beta_2$ | $-2.3$ | $-1.742$ | 0.0866 | 0.0876 |
| $\tau$ | 0.714 | 0.5715 | 0.0293 | 0.0321 |
| $h_0(t)$ Log-logistisk($\alpha = 1.3, \lambda = 0.2$) | | | | |
| $\theta$ | 0.2 | 0.192 | 0.0117 | 0.0127 |
| $\beta_1$ | 0.6 | 0.604 | 0.0402 | 0.0388 |
| $\beta_2$ | $-2.3$ | $-2.348$ | 0.0941 | 0.0966 |
| $\tau$ | 0.714 | 0.723 | 0.0133 | 0.0132 |
| $h_0(t)$ badekarformet | | | | |
| $\theta$ | 0.2 | 0.7736 | 0.0770 | 0.0723 |
| $\beta_1$ | 0.6 | 0.3598 | 0.0350 | 0.0323 |
| $\beta_2$ | $-2.3$ | $-1.304$ | 0.0680 | 0.0581 |
| $\tau$ | 0.714 | 0.3938 | 0.0223 | 0.0217 |

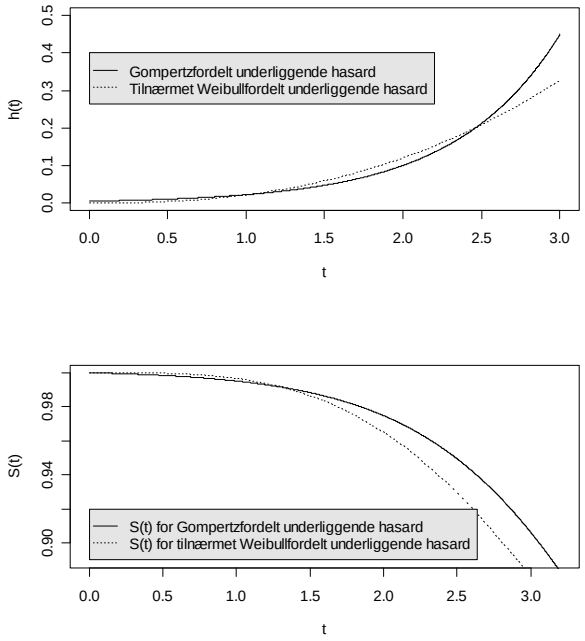
Tabell 2: Resultater fra simuleringer der $\tau = 0.714$ og $Y$ er gammafordelt. Par=parametre, Mpar=gjennomsnittlige parameterestimater basert på de $m = 200$ kohortene, estSE=gjennomsnittlig estimert standardavvik basert på de $m = 200$ kohortene, empSE=empirisk standardavvik basert på de $m = 200$ kohortene.



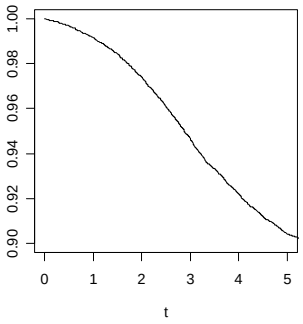

Figur 4: Fordeling til estimater når $h_0(t)$ ∼Gompertz($\alpha = 1.5$, $\lambda = 0.005$), $\tau = 0.714$ og $Y$ er gammafordelt.

og de estimerte fordelingene er vist i figur 3. Standardavviket for målet på avhengigheten er omtrent 0.01, mens det for regresjonsparametrene er litt høyere. Alle de estimerte fordelingene, i figur 3, ser tilnærmet normalfordelte ut. Dette gjelder i stort sett alle simuleringene som er gjort. I de følgende simuleringene vises de estimerte fordelingene i all hovedsak bare dersom det ser ut som at de er dårlig tilpasset en normalfordeling.

Selv om en Gompertzfordelt hasard har ganske lik form som en Weibullfordelt hasard, er det langt fra opplagt at estimatene blir brukbare dersom $h_0(t)$ er Gompertzfordelt. Som tidligere nevnt er det problematisk at for Gompertz er $h_0(0) = \lambda$, mens for Weibull er alltid $h_0(0) = 0$. Dersom $\lambda$ er stor kan dette føre til at den tilnærmede Weibullhasarden raskt vil gå opp mot $\lambda$ og dermed blir parameteren $\alpha$ i Weibullfordelingen gjerne estimert til å være mellom 1 og 2. Det vil si at den dobbeltderiverte av Weibullhasarden ofte vil bli negativ selv om den dobbeltderiverte av Gompertzhasarden er positiv. Det er gjort simuleringer med forholdsvis stor $\lambda$, blant annet $\lambda = 0.05$, men sensureringen fører da til at nesten alle hendelsene inntreffer på tidlige tidspunkt der $h_0(t)$ er omtrent konstant. Dermed blir selve formen til Gompertzhasarden mindre viktig og estimeringen av parametrene blir svært god, siden Weibullhasarden har den egenskapen at den kan øke svært raskt opp til en verdi og derfra opptre tilnærmet konstant i ganske lang tid. For å få hendelser ved tidspunkt der Gompertzhasarden har fått sin karakteristiske form har vi derfor valgt $\lambda = 0.005$ og $\alpha = 1.5$. Resultatet av simuleringen er vist i tabell 2. Vi ser at både avhengigheten og betydningen av kovariatene er underestimert med $20 - 25\%$. Videre ser vi at estimert




Figur 5: $Y$ gammafordelt, $\tau = 0.714$. Øverst: Sammenligning av korrekt Gompertz $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt Gompertz overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$. $\tau = 0.714$




Figur 6: Kaplan-Meier, $h_0(t) \sim$Gompertz$(\alpha = 1.5, \lambda = 0.005)$, $\tau = 0.714$, $Y$ er gammafordelt.

standardavvik er ganske nært empirisk standardavvik for alle parametrene. Standardavviket til $\tau$ (og $\theta$) er endel høyere i denne simuleringen enn i simuleringen med korrekt spesifisert Weibull $h_0(t)$ mens standardavvikene til regresjonsparametrene er omtrent de samme. Figur 4 viser at de estimerte fordelingene ikke er like symmetriske som i første simulering, og det kan se ut som om estimatene er litt mer ustabile i denne simuleringen.

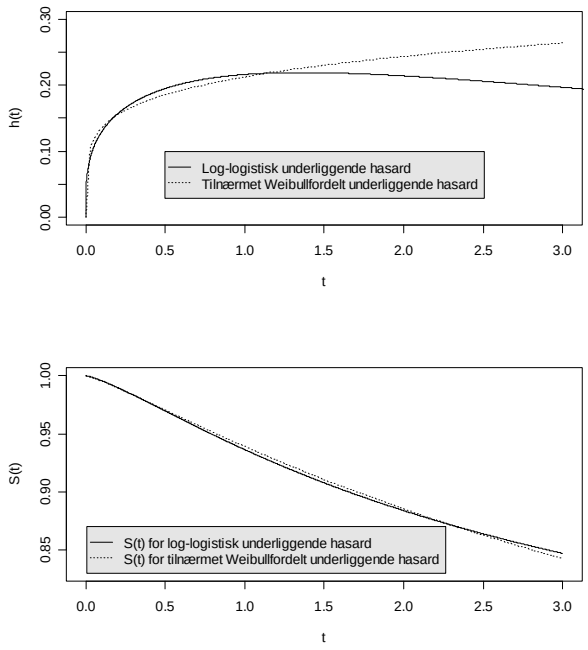
Ligning (3) kan brukes til å utlede overlevelsesfunksjonen, $S(t)$, for et individ med gitte kovariater. Dette blir da overlevelse for hele modellen, ikke underliggende overlevelsesfunksjon. Når $h_0(t)$ er feilspesifisert, er det opp til frailty-fordelingen å kompensere for dette. Overlevelsesfunksjonen viser i hvor stor grad det lykkes. Nederst i figur 5 ser vi $S(t)$ basert på virkelige parametre og Gompertzhasard, plottet mot $S(t)$ som er basert på gjennomsnittlige estimerte parametre og Weibullhasard. Begge kovariatene er valgt til å ha forventningsverdien 0.5. Vi ser at den virkelige overlevelsesfunksjonen avviker endel i forhold til den estimerte overlevelsesfunksjonen som følge av at $h_0(t)$ avviker. Figur 6 viser Kaplan-Meier-plottet for den første kohorten. Dette gjelder i alle følgende Kaplan-Meier-plott. Det er tydelig at vi har utnyttet den karakteristiske formen til Gompertzhasarden siden Kaplan-Meier-plottet i figur 6 viser at vi har hendelser ved tidspunkt der den underliggende hasarden er sterkt økende.

Dersom den underliggende hasarden i virkeligheten er log-logistisk fordelt ($\alpha = 1.3$, $\lambda = 0.2$), ser det ut til at estimatene blir mye bedre enn når den er Gompertzfordelt, tabell 2. Skjevheten til $\tau$ er kun 0.01. Dette kan skyldes at den log-logistiske hasarden har samme startverdi som Weibullhasarden. Dermed slipper den tilpassede hasarden å få en for rask stigning i starten som fører til dårlig tilpasning ved senere tidspunkt. Dessuten blir den tilnærmede hasarden svært lik ved alle relevante tidspunkt. Heller ikke her er de estimerte standardavvikene langt unna de empiriske standardavvikene, og de ligger ganske nært til standardavvikene i simuleringen med Weibullfordelt hasard. De estimerte fordelingene er i større grad enn i forrige simulering symmetriske og godt tilpasset en normalfordeling. I dette eksempelet fungerer altså antagelsen om Weibullfordelt $h_0(t)$ godt. Det er også gjort simuleringer med log-logistisk $h_0(t)$ der $\alpha = 0.8$ og $\lambda = 0.05$, det vil si at $h_0(t)$ er avtagende. Også her blir estimatene meget gode, og de empiriske og estimerte standardavvikene forblir nesten uforandret.

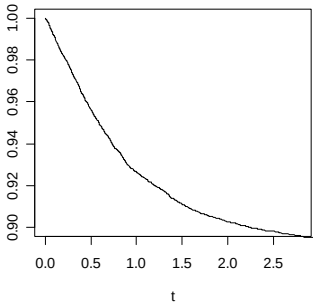
Når vi antar at $h_0(t)$ er Weibullfordelt, er det intuitivt at om hasarden i virkeligheten er badekarformet, så kan det gi dårlig tilpasning og dårlige estimater. Et mulig valg for en badekarhasard er

$$h_0(t) = \begin{cases} 0.5 - 5t \text{ for } 0 < t < 0.1 \\ 0.005\,(t - 0.1)^4 \text{ for } t > 0.1 \end{cases}$$

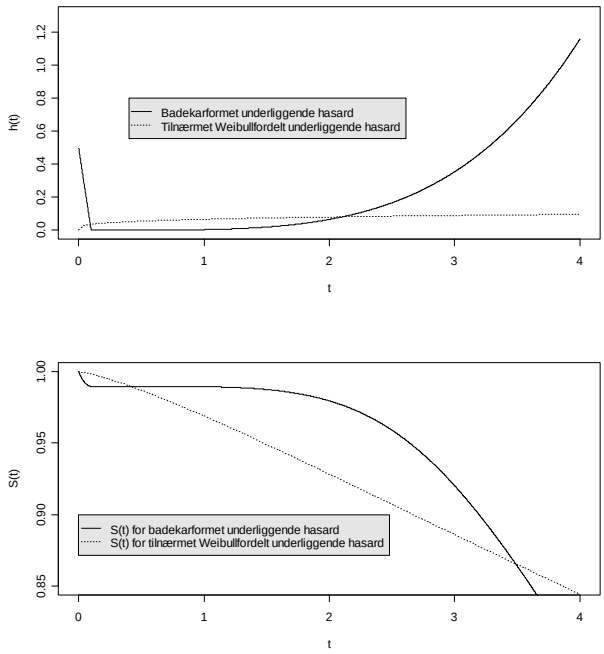
Det vil si at hasarden er lineært avtagende for $0 < t < 0.1$, mens for $t > 0.1$ er hasarden forskjøvet Weibullfordelt med $\lambda = 0.001$ og $\alpha = 5$. Av Kaplan-Meier-plottet i figur 10 ser vi at for omtrent 2% av individene inntreffer hen-



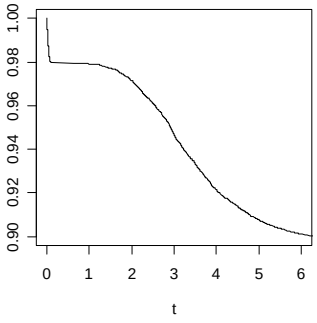

Figur 7: $Y$ gammafordelt, $\tau = 0.714$. Øverst: Sammenligning av korrekt log-logistisk $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt log-logistisk overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$. $\tau = 0.714$




Figur 8: Kaplan-Meier, $h_0(t) \sim$log-logistisk$(\alpha = 1.3, \lambda = 0.2)$, $\tau = 0.714$, $Y$ er gammafordelt.




Figur 9: $Y$ gammafordelt, $\tau = 0.714$. Øverst: Sammenligning av korrekt badekarformet $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt badekar overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$. $\tau = 0.714$



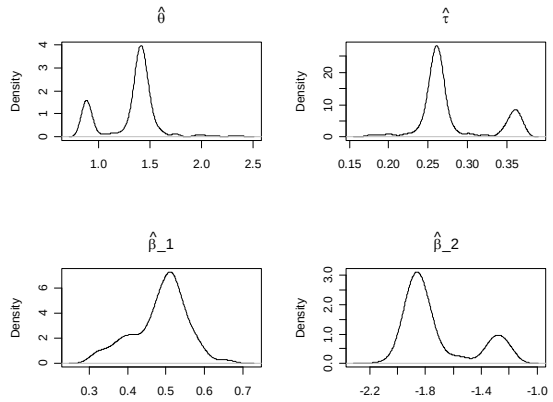

Figur 10: Kaplan-Meier, badekarformet $h_0(t)$, $\tau = 0.714$, $Y$ er gammafordelt.

| Parameter | Par | Mpar | estSE | empSe |
|---|---|---|---|---|
| $h_0(t)$ Gompertz($\alpha = 1.5, \lambda = 0.005$) | | | | |
| $\theta$ | 0.75 | 1.3291 | 0.1358 | 0.2704 |
| $\beta_1$ | 0.6 | 0.4855 | 0.0365 | 0.0710 |
| $\beta_2$ | $-2.3$ | $-1.7233$ | 0.0737 | 0.2483 |
| $\tau$ | 0.4 | 0.2797 | 0.0404 | 0.0444 |
| $h_0(t)$ badekarformet | | | | |
| $\theta$ | 0.75 | 119.9 | 287.0 | 95.24 |
| $\beta_1$ | 0.6 | 0.3751 | 0.0273 | 0.0263 |
| $\beta_2$ | $-2.3$ | $-1.4791$ | 0.0633 | 0.0613 |
| $\tau$ | 0.4 | 0.0087 | 0.0160 | 0.0094 |

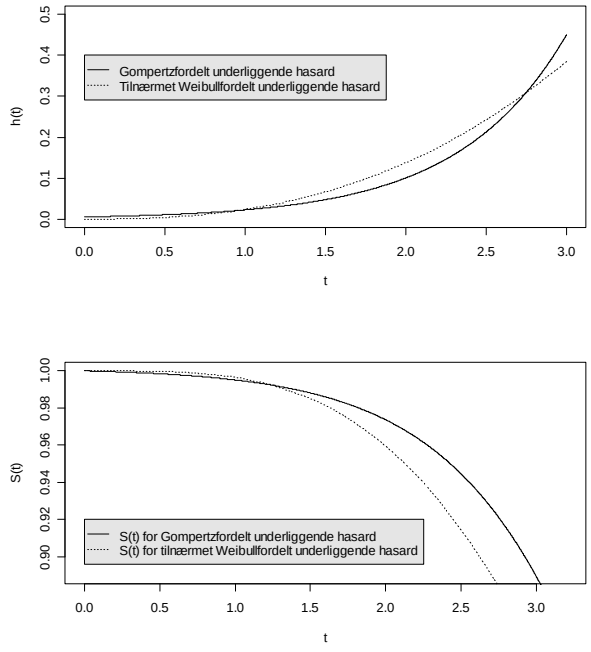
Tabell 3: Resultater fra simuleringer der $\tau = 0.4$ og $Y$ er gammafordelt. Se tabell 2 for forklaring.

delsen før den underliggende hasarden synker ned til 0, for 8% av individene inntreffer hendelsen når hasarden igjen begynner å øke, mens 90% av individene blir sensurert. Tabell 2 viser at både avhengigheten og kovariatenes virkning i større grad enn i de tidligere simuleringene blir underestimert. Figur 9 gir oss en indikasjon på hvorfor estimatene blir så dårlige. Av hensyn til de tidlige sykdomstilfellene øker Weibullhasarden relativt raskt i starten. Siden Weibullhasarden dermed ikke kan avta, holder den et tilnærmet konstant nivå for tidspunkt der den virkelige hasarden er omtrent 0. Idet den virkelige hasarden begynner å øke raskt igjen, kan ikke den Weibullfordelte hasarden gjøre annet enn å holde seg på tilnærmet samme nivå som tidligere. Dette gjenspeiles også i overlevelsesfunksjonen nederst i samme figur. De estimerte og empiriske standardavvikene er svært like og estimatene for fordelingene er symmetriske og tilsynelatende tilnærmet normalfordelte, men likevel gir det i dette tilfellet relativt dårlige resultater å bruke en Weibullfordelt hasard. Kaplan-Meier-plottet viser tydelig at $h_0(t)$ er tilnærmingsvis badekarformet. For virkelige data kan det derfor være viktig å undersøke Kaplan-Meier-plottet før man eventuelt bruker en Weibullfordelt hasard.

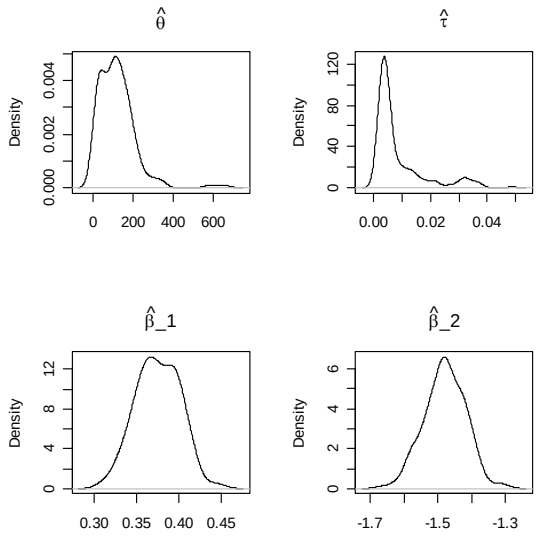
I de forrige simuleringene har vi valgt en relativt stor avhengighet innad i familiene, $\tau = 0.714$. Vi skal nå se hva som skjer dersom vi velger en mindre avhengighet. Vi setter $\lambda = 0.005$ og $\alpha = 1.5$ i den Gompertzfordelte $h_0(t)$, velger en litt lavere avhengighet, $\tau = 0.4$, og simulerer nye kohorter. Vi ser at regresjonsparametrene blir underestimert i like stor grad som når $\tau = 0.714$. Dette kan tyde på at avhengigheten påvirker estimatene for regresjonsparametrene i liten grad når $h_0(t)$ er Gompertzfordelt. Vi ser at avhengigheten i litt større grad blir underestimert enn når vi hadde $\tau = 0.714$, her blir den underestimert med 30%. Vi ser også at den estimerte fordelingen til estimatene, særlig for avhengigheten, har to topper. Begge disse problemene




Figur 11: Fordeling til estimater når $h_0(t)$ ∼Gompertz($\alpha = 1.5$, $\lambda = 0.005$), $\tau = 0.4$ og $Y$ er gammafordelt.



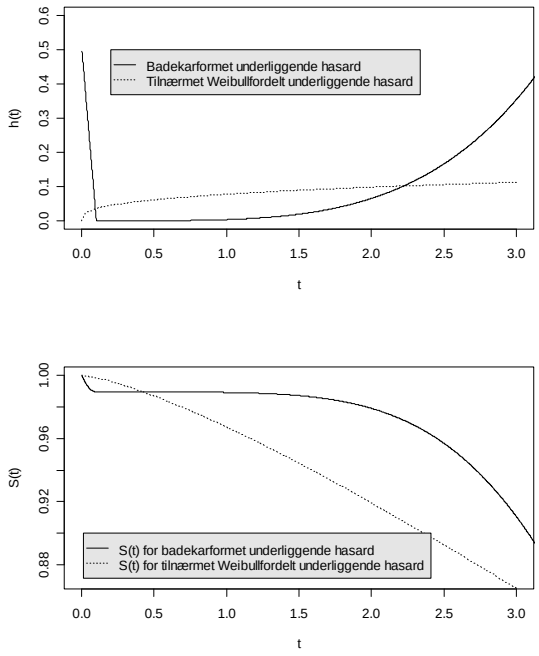

Figur 12: $Y$ gammafordelt, $\tau = 0.4$. Øverst: Sammenligning av korrekt Gompertzfordelt $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt Gompertz overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$. $\tau = 0.4$



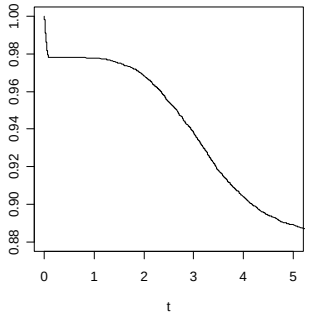

Figur 13: Fordeling til estimater når $h_0(t)$ er badekarformet, $\tau = 0.4$ og $Y$ er gammafordelt.

kan skyldes at med høy grad av sensurering og lav grad av avhengighet vil noen kohorter ha få familier med mer enn ett sykdomstilfelle og dermed blir estimeringen mer ustabil. Dette gjelder intuitivt ikke estimeringen av regresjonsparametrene.

Med samme badekarformede hasard som tidligere og lavere avhengighet, $\tau = 0.4$, blir esimatet for avhengigheten svært dårlig, tabell 3. Selv om $\tau$ er såpass stor som 0.4, estimeres parameteren til omtrent 0. Et mulig problem kan være at for noen kohorter blir avhengigheten lav og dermed estimatet for $\theta$ svært høyt. Dermed blir gjennomsnittet av estimatene for $\theta$ også høyt. Dette problemet vil intuitivt være mindre for estimatet av $\tau$, siden estimatet av $\tau$ må være mellom 0 og 1. I figur 13 ser vi også at det ikke er dette som er problemet i dette tilfellet. Estimatene for $\tau$ er stort sett mindre enn 0.04 og aldri i nærheten av den sanne verdien. Estimatene for regresjonsparametrene blir derimot omtrent like gode som tidligere. Dette kan tyde på at avhengigheten påvirker estimatene for regresjonsparametrene i liten grad, ikke bare for Gompertzfordelt $h_0(t)$, men også når det gjelder badekarformet $h_0(t)$. Dermed kan det være grunn til å tro at dette også gjelder mer generelt. I likhet med simuleringen der $\tau = 0.714$, er den dobbeltderiverte av den tilpassede Weibullhasarden også her negativ. Vi ser i figur 14 at dette fører til en svært dårlig tilpasning for overlevelsesfunksjonen.




Figur 14: $Y$ gammafordelt, $\tau = 0.4$. Øverst: Sammenligning av korrekt badekarformet $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt badekar overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$. $\tau = 0.4$




Figur 15: Kaplan-Meier, badekarformet $h_0(t)$, $\tau = 0.4$, $Y$ er gammafordelt.

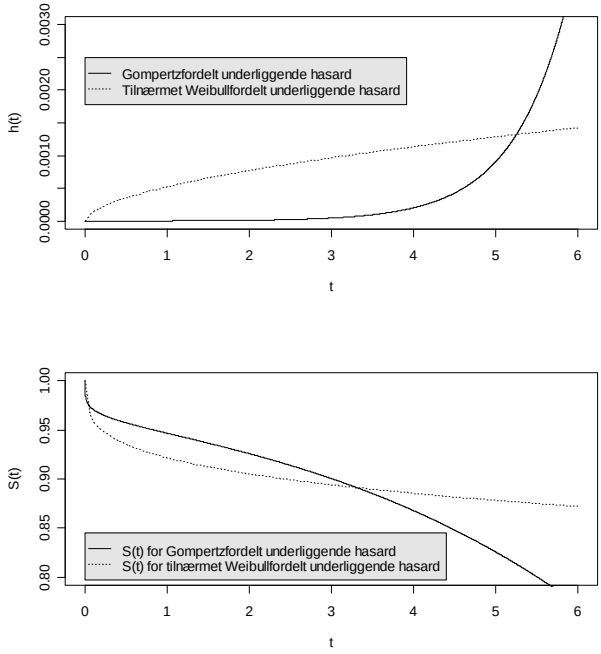
| Parameter | Par | Mpar | estSE | empSe |
|---|---|---|---|---|
| $h_0(t)$ Weibull($\alpha = 3, \lambda = 0.1$) | | | | |
| $\theta$ | 0.2 | 0.2009 | 0.0077 | 0.0080 |
| $\lambda$ | 0.1 | 0.1065 | 0.0338 | 0.0375 |
| $\alpha$ | 3 | 2.9984 | 0.0751 | 0.0769 |
| $\beta_1$ | 0.6 | 0.5994 | 0.0508 | 0.0541 |
| $\beta_2$ | $-2.3$ | $-2.2928$ | 0.1270 | 0.1366 |
| $\tau$ | 0.8 | 0.7991 | 0.0077 | 0.0080 |
| $h_0(t)$ Gompertz($\alpha = 1.5, \lambda = 5 \cdot 10^{-7}$) | | | | |
| $\theta$ | 0.2 | 0.2880 | 0.0109 | 0.0139 |
| $\beta_1$ | 0.6 | 0.4894 | 0.0523 | 0.0507 |
| $\beta_2$ | $-2.3$ | $-1.8931$ | 0.1283 | 0.1351 |
| $\tau$ | 0.8 | 0.7120 | 0.0109 | 0.0139 |
| $h_0(t)$ badekarformet | | | | |
| $\theta$ | 0.2 | 0.2557 | 0.0097 | 0.0117 |
| $\beta_1$ | 0.6 | 0.4795 | 0.0495 | 0.0506 |
| $\beta_2$ | $-2.3$ | $-1.923$ | 0.1235 | 0.1442 |
| $\tau$ | 0.8 | 0.7443 | 0.0097 | 0.0117 |

Tabell 4: Resultater fra simuleringer der $\tau = 0.8$ og $Y$ er stablefordelt. Se tabell 2 for forklaring.

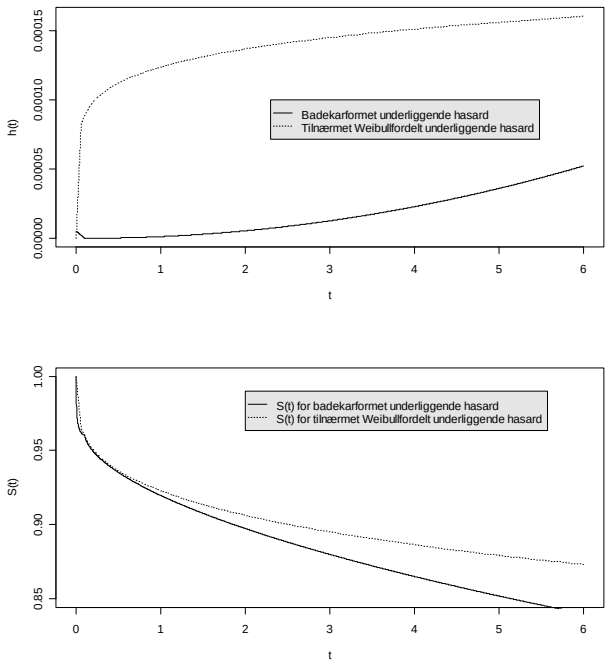
### 5.2 Stablefordelt frailty-variabel

En annen vanlig frailty-fordeling er stablefordelingen. I de følgende simuleringene antar vi at $Y$ er stablefordelt mens $h_0(t)$ fremdeles antas å være Weibullfordelt. For å finne estimater for parametrene i modellen optimeres dermed log-likelihooden i ligning (11). Vi genererer data med stablefordelt $Y$, slik at denne antagelsen er korrekt og undersøker om resultatene blir omtrent like gode med denne frailty-fordelingen. Først gjøres en simulering der alle antagelser er korrekte, det vil si at $h_0(t)$ er Weibullfordelt. Målet for avhengigheten, $\tau$ settes lik 0.8 mens de andre parametrene forblir uforandret, tabell 4. Vi ser at standardavviket til $\theta$ ble større da vi antok gammafordelt frailty-variabel og valgte $\tau = 0.714$. Grunnen til dette er nok at standardavviket til $\theta$ avtar når $\theta$ nærmer seg 1. Dette gjenspeiles også i simuleringen med gammafordelt frailty-variabel der $\tau = 0.4$. Her er standardavviket til $\theta$ enda større. Standardavviket til regresjonsparametrene forblir uforandret.

Videre simulerer vi data der $h_0(t) \sim \text{Gompertz}\left(\alpha = 1.5, \lambda = 5 \cdot 10^{-7}\right)$. Grunnen til at $\lambda$ settes svært lav er at vi fremdeles ønsker å utnytte forskjellen i formen til en Gompertzhasard og en Weibullhasard. Når $Y$ er stablefordelt, så er forventningsverdien til $Y$ uendelig og dermed blir hasarden, det vil si risikoen for å dø i den fulle modellen, større enn når $Y$ er gammafordelt. Dette kan kompenseres ved å la $h_0(t)$ være lav inntil den får sin karakter-




Figur 16: $Y$ stablefordelt, $\tau = 0.8$. Øverst: Sammenligning av korrekt Gompertzfordelt $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt Gompertz overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.

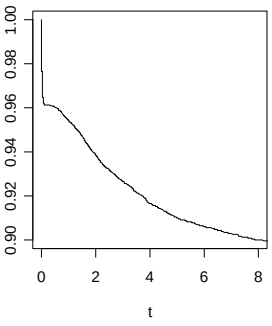



Figur 17: $Y$ stablefordelt, $\tau = 0.8$. Øverst: Sammenligning av korrekt badekarformet $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt badekar overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.

istiske form. Vi ser at i dette tilfellet blir estimatene, i tabell 4, forholdsvis gode. Parameteren som beskriver avhengigheten, $\tau$, estimeres 11% feil, mens regresjonsparametrene estimeres i underkant av 20% feil. Som før blir alle parametrene underestimert. Skjevheten er i omtrent samme skala som simuleringen med gammafordelt frailty-variabel og $\tau = 0.714$. Det viser at om frailty-variabelen er gammafordelt eller stablefordelt spiller liten rolle når $h_0(t)$ er Gompertzfordelt og avhengigheten forholdsvis stor. De empiriske og estimerte standardavvikene til parametrene stemmer overens i tilfredsstillende grad. Dessuten er de i omtrent samme størrelsesorden som når vi simulerer med korrekte antagelser.

En annen interessant observasjon, øverst i figur 16, er at den estimerte underliggende Weibullhasarden i dette tilfellet ikke ligner Gompertzhasarden i det hele tatt. Siden $h_0(t)$ ser ut til å ha en mye bedre tilpasning når $Y$ er gammafordelt, er det grunn til å tro at det stablefordelingen som er skyld i dette. Stablefordelingen er svært skjev og dette er nok den mest sannsynlige grunnen.

Det er også gjort en simulering med stablefordelt frailtyvariabel og badekar-




Figur 18: Kaplan-Meier, badekarformet $h_0(t)$, $\tau = 0.8$, $Y$ er stablefordelt.

formet $h_0(t)$. Avhengigheten er den samme som for simuleringen med en Gompertzfordelt $h_0(t)$, nemlig $\tau = 0.8$. Som før er det viktig å velge den underliggende hasarden slik at vi får et passe antall hendelser før $h_0(t)$ blir 0. Jeg har valgt

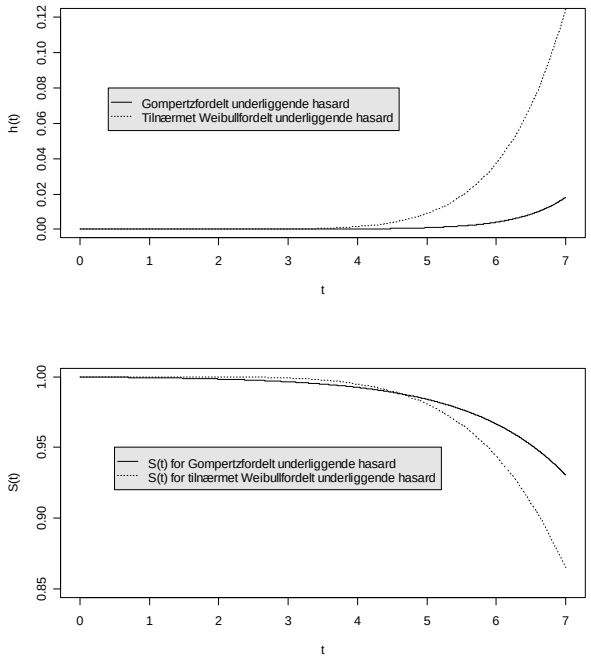
$$h_0(t) = \begin{cases} 5 \cdot 10^{-6} - 5 \cdot 10^{-5} t \text{ for } 0 < t < 0.1 \\ 1.5 \cdot 10^{-6} (t - 0.1)^2 \text{ for } t > 0.1 \end{cases}.$$

Det vil si at hasarden er lineært avtagende for $t < 0.1$, mens den er forskjøvet Weibullfordelt med $\lambda = 5 \cdot 10^{-7}$ og $\alpha = 3$ for $t \geq 0.1$. Av Kaplan-Meier-plottet ser vi at omtrent 4% av levetidene inntreffer når $t < 0.1$, omtrent 6% inntreffer når $t > 0.1$, mens omtrent 90% av levetidene som vanlig blir sensurert. Også her blir både avhengigheten og regresjonsparametrene underestimert. I tabell 4 ser vi at regresjonsparametrene blir underestimert i omtrent samme grad som når vi antok stablefordelt frailtyvariabel og gompertzfordelt $h_0(t)$, mens $\tau$ i dette tilfellet faktisk blir bedre estimert. Øverst i figur 17 ser vi at dette gjelder selv om den estimerte underliggende hasarden er svært forskjellig fra den virkelige. Grunnen til dette er at overlevelsesfunksjonen, nederst i samme figur, ser ut til å bli bedre tilpasset enn i den forrige simuleringen og at det er dette som er viktig når det gjelder å få et godt estimat for avhengigheten.

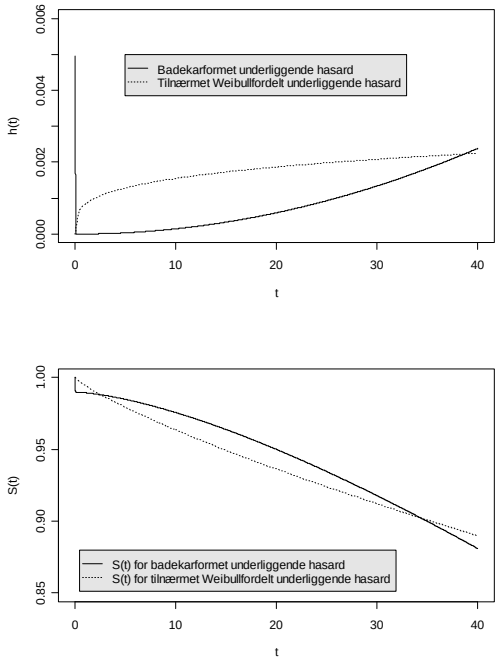
Det er også gjort to simuleringer med en lavere avhengighet, $\tau = 0.5$. I den første er $h_0(t)$ Gompertzfordelt med samme verdier på parametrene som i den forrige simuleringen med Gompertzfordelt $h_0(t)$, nemlig $\alpha = 1.5$ og $\lambda = 5 \cdot 10^{-7}$. I tabell 5 ser vi at estimatet for både avhengigheten og regresjonsparametrene blir endel dårligere i dette tilfellet. Målet på avhengigheten, $\tau$, blir underestimert med 35%, mens regresjonsparametrene blir underestimert med henholdsvis 28% og 34%. At målet på avhengigheten blir underestimert i større grad når $\tau$ er rundt 0.5 er ikke overraskende og stemmer med det

| Parameter | Par | Mpar | estSE | empSe |
|---|---|---|---|---|
| $h_0(t)$ Gompertz($\alpha = 1.5, \lambda = 5 \cdot 10^{-7}$) | | | | |
| $\theta$ | 0.5 | 0.6768 | 0.0221 | 0.0338 |
| $\beta_1$ | 0.6 | 0.4321 | 0.0484 | 0.0490 |
| $\beta_2$ | $-2.3$ | $-1.515$ | 0.0976 | 0.1132 |
| $\tau$ | 0.5 | 0.3232 | 0.0221 | 0.0338 |
| $h_0(t)$ badekarformet | | | | |
| $\theta$ | 0.5 | 0.6495 | 0.0169 | 0.0186 |
| $\beta_1$ | 0.6 | 0.4487 | 0.0400 | 0.0403 |
| $\beta_2$ | $-2.3$ | $-1.769$ | 0.1035 | 0.1057 |
| $\tau$ | 0.5 | 0.3505 | 0.0169 | 0.0186 |

Tabell 5: Resultater fra simuleringer der $\tau = 0.5$ og $Y$ er stablefordelt. Se tabell 2 for forklaring.




Figur 19: $Y$ stablefordelt, $\tau = 0.5$. Øverst: Sammenligning av korrekt Gompertzfordelt $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt Gompertz overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.

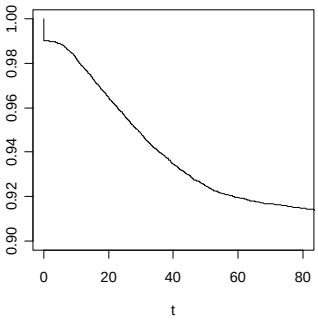



Figur 20: $Y$ stablefordelt, $\tau = 0.5$. Øverst: Sammenligning av korrekt badekarformet $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt badekar overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.

vi har funnet ut for gammafordelt frailty-variabel. Men at regresjonsparametrene blir underestimert i signifikant større grad når $\tau = 0.5$ i forhold til når $\tau = 0.8$ er overraskende. Når $Y$ er gammafordelt ser det ut til at avhengigheten har liten innvirkning på estimatene for regresjonsparametrene. I dette tilfellet, hvor $Y$ er stablefordelt, påvirkes estimatet for regresjonsparametrene av avhengigheten i større grad.

En annen vesentlig forskjell på denne simuleringen og simuleringen der $\tau = 0.8$ er at i dette tilfellet ser den estimerte underliggende hasarden, øverst i figur 19, ut til å være bedre tilpasset en Gompertzhasard. En sannsynlig årsak til at $h_0(t)$ estimeres så ulikt avhengig av om $\tau = 0.8$ eller $\tau = 0.5$ er at vi får mange svært lave levetider når $\tau$ er høy. I tilfellet når $\tau = 0.5$ får vi nesten ingen lave levetider.

Til slutt undersøker vi estimatene dersom vi velger en lavere avhengighet, $\tau = 0.5$, når $h_0(t)$ er badekarformet. For å få et passe antall hendelser som inntreffer før hasarden synker ned til 0, må vi her velge en litt annen




Figur 21: Kaplan-Meier, badekarformet $h_0(t)$, $\tau = 0.5$, $Y$ er stablefordelt.

badekarfordeling enn når $\tau = 0.8$. Vi velger

$$h_0(t) = \begin{cases} 5 \cdot 10^{-3} - 5 \cdot 10^{-2} t \text{ for } 0 < t < 0.1 \\ 1.5 \cdot 10^{-6} (t - 0.1)^2 \text{ for } t > 0.1 \end{cases}.$$

I Kaplan-Meier-plottet ser vi at omtrent 1% av levetidene inntreffer ved $t < 0.1$, omtrent 9% av hendelsene inntreffer ved $t > 0.1$, mens omtrent 90% av levetidene blir sensurert. I denne simuleringen ser det ut som om vi tilpasser $h_0(t)$ bedre enn i simuleringen der $\tau = 0.8$. Grunnen kan være at når $\tau = 0.8$ var $h_0(t)$ svært lav for alle relevante tidspunkt. Dette kan ha vært problematisk for estimeringen av $h_0(t)$. En interessant observasjon er imidlertid at estimatene for både avhengigheten og regresjonsparametrene, i tabell 5, er litt dårligere i denne simuleringen. Avhengigheten er feilestimert med 30%. Selv om en avhengighet på $\tau = 0.5$ i forhold til en avhengighet på $\tau = 0.8$ nok kan føre til dårligere estimater, tyder dette på at tilpasningen til $h_0(t)$ ikke er helt avgjørende for hvor gode estimater man får. Likevel er det betryggende at estimatene blir såpass gode også i denne simuleringen. Dessuten er de observerte standardavvikene og de empiriske standardavvikene nesten identiske.

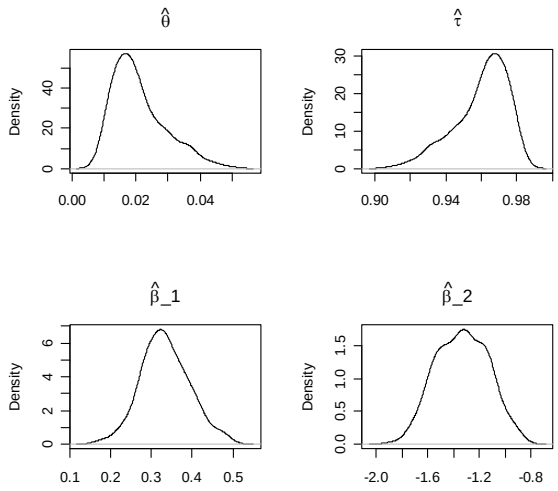
## 5.3 Feilspesifisert frailty-variabel og underliggende hasard

Hittil har vi undersøkt estimatene dersom $h_0(t)$ har vært feilspesifisert. Samtidig har vi antatt at frailty-variabelen er enten gamma- eller stablefordelt og i alle simuleringene har denne antagelsen vært korrekt. Vi skal nå undersøke estimatene dersom vi feilspesifiserer *både* frailty-variabelen $Y$ og den underliggende hasarden $h_0(t)$. Dette vil intuitivt føre til dårligere resultater.

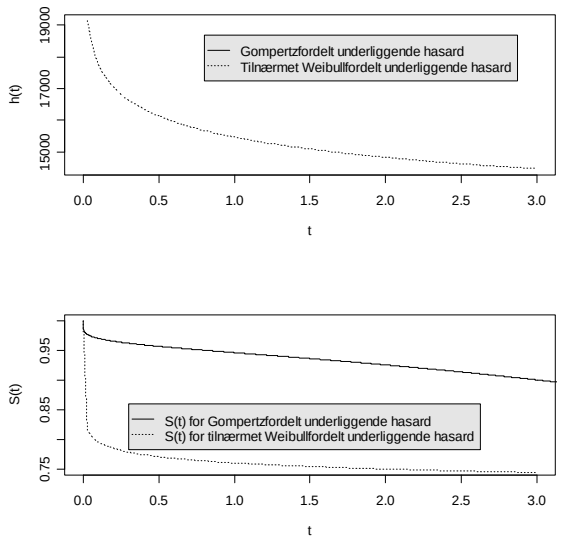
Som i kapittel 5.2 genererer vi først data med Gompertz eller badekarformet $h_0(t)$ og stablefordelt $Y$, men nå tilpasses modellen med gammafordelt

| Parameter | Par | Mpar | estSE | empSe |
|---|---|---|---|---|
| $h_0(t)$ Gompertz($\alpha = 1.5, \lambda = 5 \cdot 10^{-7}$) | | | | |
| $\beta_1$ | 0.6 | 0.3341 | 0.0498 | 0.0595 |
| $\beta_2$ | $-2.3$ | $-1.334$ | 0.1176 | 0.1987 |
| $\tau$ | 0.8 | 0.9600 | 0.0013 | 0.0150 |
| $h_0(t)$ badekarformet | | | | |
| $\beta_1$ | 0.6 | 0.4076 | 0.0472 | 0.0443 |
| $\beta_2$ | $-2.3$ | $-1.587$ | 0.1095 | 0.1506 |
| $\tau$ | 0.8 | 0.9745 | 0.0007 | 0.0035 |

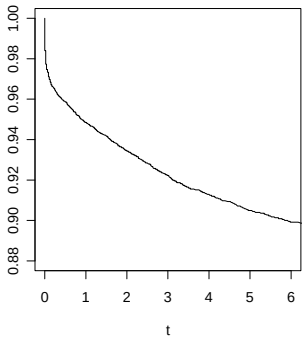
Tabell 6: Resultater fra simuleringer der $\tau = 0.8$ og $Y$ er stablefordelt. Modellen er tilpasset gammafordelt $Y$. Se tabell 2 for forklaring.




Figur 22: Fordeling til estimater når $h_0(t)$ er Gompertzfordelt og $\tau = 0.8$. $Y$ er stablefordelt, men antas å være gammafordelt.




Figur 23: $Y$ er stablefordelt, men antas å være gammafordelt, $\tau = 0.8$. Øverst: Sammenligning av korrekt Gompertz $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt Gompertz overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.




Figur 24: Kaplan-Meier, Gompertzfordelt $h_0(t)$, $\tau = 0.8$. $Y$ er stablefordelt, men antas å være gammafordelt.

frailty-variabel. Den underliggende hasarden antas som før å være Weibull-fordelt. Resultatene vises i tabell 6. I den første simuleringen bruker vi samme $h_0(t)$ som i den andre simuleringen i kapittel 5.2, det vil si $h_0(t) \sim$ Gompertz($\alpha = 1.5, \lambda = 5 \cdot 10^{-7}$). Det vil si at vi genererer data på nøyaktig samme måte, men her antar vi at $Y$ er gammafordelt. I denne simuleringen, hvor fordelingen til $Y$ er feilspesifisert, overestimeres $\tau$ til 0.96. Regresjonsparametrene underestimeres med i overkant av 40%. I simuleringen der vi antok riktig frailty-fordeling ble regresjonsparametrene underestimert med i underkant av 20%. Dette er en relativt stor forskjell som utelukkende skyldes de ulike egenskapene til gamma- og stablefordelingen. Stablefordelingen har uendelig varians. Dette gjør at spredningen til $Y$ blir større, og dermed er det intuitivt at avhengigheten blir overestimert dersom vi antar at $Y$ er gammafordelt. Siden det legges for stor vekt på avhengigheten er det naturlig at regresjonsparametrene mister sin betydning i større grad enn tidligere.

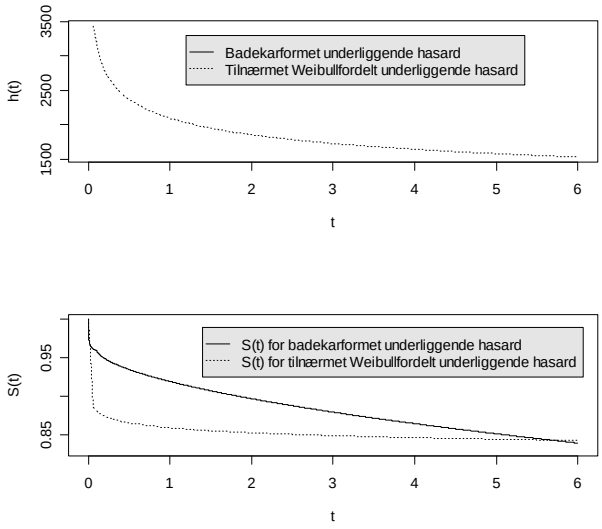
I figur 22 ser vi at estimatene for $\tau$ blir skjevfordelt. Dette skyldes naturligvis at estimatene må være mindre enn 1 og gjennomsnittet av estimatene ligger svært nært 1.

En annen konsekvens av feilspesifisert fraily-variabel er at den estimerte underliggende hasarden blir svært stor ved lave tidspunkt. Dette vises både i figur 23 og i figur 25, der $h_0(t)$ er badekarformet. I disse figurene er det ikke engang mulig å se formen på den virkelige underliggende hasarden på grunn av skaleringen av y-aksen. Dette henger også sammen med at gammafordelingen ikke er like skjevfordelt og kan få like stor varians som stablefordelingen. Siden $Y$ er stablefordelt vil hendelser ofte inntreffe etter svært kort tid. Dette kommer klart fram i Kaplan-meier-plottetene i figur 24 og i figur 26, der $h_0(t)$ er badekarformet. Måten modellen med gammafordelt frailty-variabel tilpasser dette på, er å sette $h_0(t)$ svært høy ved svært lave tidspunkt.

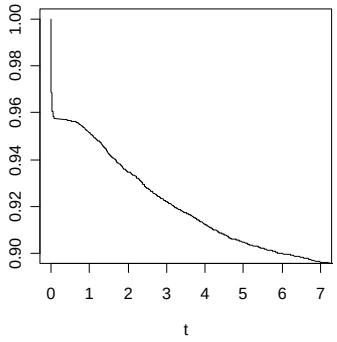
I simuleringen der $h_0(t)$ er badekarformet med de samme parametrene som i den tredje simuleringen i kapittel 5.2, er tendensen den samme, om enn i litt mindre grad. Tabell 6 viser at regresjonsparametrene blir underestimert med i overkant av 30%, i motsetning til simuleringen med korrekt spesifisert frailty-variabel der de ble underestimert med omtrent 20%. Målet for avhengigheten, $\tau$, blir også her overestimert, nå til 0.9745.

Disse simuleringene viser at med både feilspesifisert $h_0(t)$ og feilspesifisert frailty-variabel der vi antar at $Y$ er gammafordelt blir resultatene i enda mindre grad pålitelige. Avhengigheten vil generelt bli overestimert mens regresjonsparametrene i enda større grad blir underestimert.

I de siste simuleringene har vi antatt at $Y$ er stablefordelt, mens dataene genereres med gammafordelt $Y$. Den underliggende hasarden, $h_0(t)$, er henholdsvis Gompertzfordelt og badekarformet, og parametrene er de samme som i henholdsvis den andre og den fjerde simuleringen i kapittel 5.1. Dermed kan vi bruke resultatene til å si noe om hvordan feilspesifisert stable frailty-fordeling virker inn på estimatene, dersom $Y$ i virkeligheten er gammafordelt.




Figur 25: $Y$ er stablefordelt, men antas å være gammafordelt, $\tau = 0.8$. Øverst: Sammenligning av korrekt badekarformet $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt badekar overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.



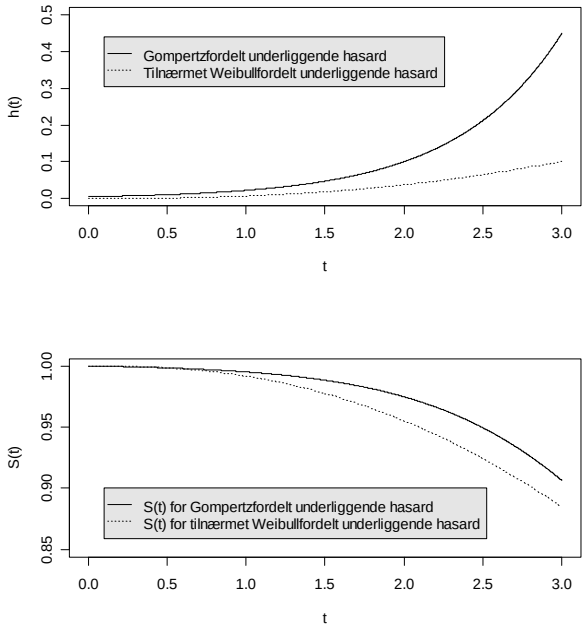

Figur 26: Kaplan-Meier, badekarformet $h_0(t)$, $\tau = 0.8$. $Y$ er stablefordelt, men antas å være gammafordelt.

| Parameter | Par | Mpar | estSE | empSe |
|---|---|---|---|---|
| $h_0(t)$ Gompertz($\alpha = 1.5, \lambda = 0.005$) | | | | |
| $\beta_1$ | 0.6 | 0.4248 | 0.0393 | 0.0497 |
| $\beta_2$ | $-2.3$ | $-1.551$ | 0.0886 | 0.0957 |
| $\tau$ | 0.714 | 0.2976 | 0.0137 | 0.0181 |
| $h_0(t)$ badekarformet | | | | |
| $\beta_1$ | 0.6 | 0.3372 | 0.0318 | 0.0341 |
| $\beta_2$ | $-2.3$ | $-1.294$ | 0.0692 | 0.0773 |
| $\tau$ | 0.714 | 0.1242 | 0.0116 | 0.0129 |

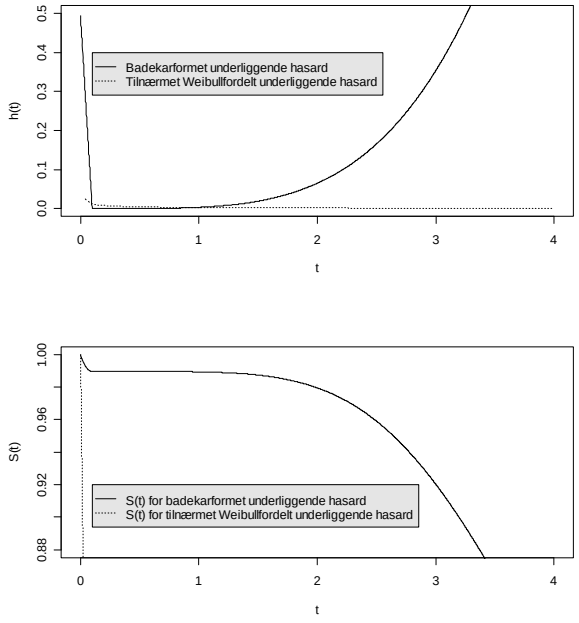
Tabell 7: Resultater fra simuleringer der $\tau = 0.714$ og $Y$ er gammafordelt. Modellen er tilpasset stablefordelt $Y$. Se tabell 2 for forklaring.

Når $h_0(t)$ er Gompertzfordelt, viser tabell 6 at en slik feilspesifisering av frailty-fordelingen gjør at regresjonsparametrene underestimeres i større grad, omtrent 30% nå mot 20% med korrekt gamma frailty-fordeling. Også avhengigheten underestimeres i mye større grad i denne simuleringen, $\tau$ estimeres til ca. 0.30. Med korrekt frailty-fordeling ble denne estimert til omtrent 0.57. Det virker fornuftig at avhengigheten underestimeres i større grad enn tidligere, nettopp av samme grunn til at avhengigheten blir overestimert når $Y$ er stablefordelt og vi antar gammafordeling. Siden stablefordelingen har uendelig stor varians vil gammafordelingen ikke klare å fange opp all variansen.

I simuleringen der $h_0(t)$ er badekarformet ser det ut til at en feilspesifisering av frailty-fordelingen på denne måten spiller liten rolle for estimeringen av regresjonsparametrene. Vi sammenligner med den fjerde simuleringen i kapittel 5.1. Avhengigheten, $\tau$, underestimeres derimot i mye større grad enn før. Mens estimatet av $\tau$ var 0.39 med korrekt frailty-fordeling, er det nå 0.12.




Figur 27: $Y$ er gammafordelt, men antas å være stablefordelt, $\tau = 0.714$. Øverst: Sammenligning av korrekt Gompertz $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt Gompertz overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.




Figur 28: $Y$ er gammafordelt, men antas å være stablefordelt, $\tau = 0.714$. Øverst: Sammenligning av korrekt badekarformet $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt badekar overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.

# 6 Diskusjon

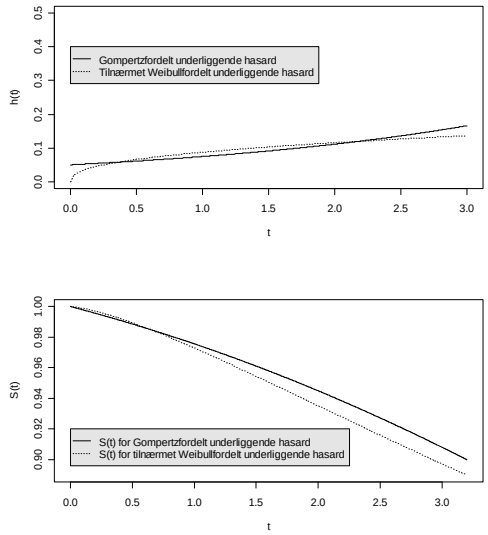
En viktig forutsetning for å få såpass skjeve estimater som i denne oppgaven er å la hendelsene inntreffe over et stort tidsrom, det vil si å bruke en stor del av den underliggende hasarden, $h_0(t)$. Dersom vi bruker en Gompertz- eller badekarfordelt $h_0(t)$ som er slik at svært mange av hendelsene inntreffer på relativt tidlige tidspunkter, det vil si tidspunkter før hasarden har fått sin karakteristiske form, blir estimatene mye bedre enn det som er vist her. Et eksempel for Gompertz $h_0(t)$ er vist i tabell 8 og i figur 29. Her er estimatene for avhengigheten og regresjonsparametrene $4-8\%$ feil og generelt vil skjevheten i stor grad holde seg under $5-10\%$. Vi ser at den estimerte overlevelsesfunksjonen, $S(t)$, avviker fra den virkelige og likevel får vi altså såpass gode estimater. Når vi nesten utelukkende har tidlige hendelser blir det imidlertid vanskelig å identifisere $h_0(t)$ som Gompertz- eller badekarfordelt, og likheten med en Weibullfordeling vil naturlig nok bli større. Derfor har jeg i denne oppgaven valgt å fokusere på underliggende hasarder som gjør at vi får stor variasjon i levetidene. Ved å velge en sensureringstid med lav forventning og høy varians får vi ytterligere variasjon i levetidene.

| Parameter | Par | Mpar | estSE | empSe |
|---|---|---|---|---|
| $h_0(t)$ Gompertz($\alpha = 0.4, \lambda = 0.05$) | | | | |
| $\theta$ | 0.2 | 0.246 | 0.0160 | 0.0165 |
| $\beta_1$ | 0.6 | 0.576 | 0.0402 | 0.0387 |
| $\beta_2$ | $-2.3$ | $-2.093$ | 0.0880 | 0.0791 |
| $\tau$ | 0.714 | 0.670 | 0.0148 | 0.0147 |

Tabell 8: Resultater fra simuleringer der $\tau = 0.714$ og $Y$ er gammafordelt. Se tabell 2 for forklaring.

Det er viktig å ha i bakhodet at disse grepene fører til at simuleringene i denne oppgaven er worst case-scenarier. For virkelige datasett har man, som tidligere nevnt, ofte forkunnskap om $h_0(t)$. Hasarden for mange sykdommer kan rett og slett ikke være badekarformet. Dessuten kan det også hende at utbruddene for sykdom ikke er så mye spredt i tid som det vi har antatt her. Man kan derfor ikke uten videre si at frailty-modeller med Weibull $h_0(t)$ er lite robuste.

I og med at estimatene er såpass skjeve, er det mindre interessant å studere empirisk og estimert standardavvik inngående. Noe som imidlertid er verdt å legge merke til er at empirisk og estimert standardavvik ofte stemmer overraskende godt overens med tanke på hvor skjeve estimatene er. Dersom modellen er bedre tilpasset så vil sannsynligvis estimert standardavvik ligge enda tettere mot empirisk og vi kan dermed slå fast at estimert




Figur 29: $Y$ er gammafordelt, $\tau = 0.714$. Øverst: Sammenligning av korrekt Gompertz $h_0(t)$ mot estimert Weibull $h_0(t)$. Nederst: Sammenligning av korrekt Gompertz overlevelsesfunksjon $S(t)$ mot estimert Weibull $S(t)$.

standardavvik for parametrene i en frailty-modell generelt er svært pålitelig.

I utgangspunktet skulle man kanskje tro at estimatet for regresjonsparameteren for den dikotome og avhengige kovariaten ville bli mer skjevt enn estimatet for regresjonsparameteren for den kontinuerlige og uavhengige kovariaten, siden noe av denne avhengigheten kanskje kunne bli fanget opp av frailty-variabelen. Det er imidlertid vanskelig å se noen forskjell på skjevheten i de to regresjonsparametrene. Siden avhengighet i en kovariat er svært vanlig innen familier er dette oppløftende.

I denne oppgaven har vi generert data med tre ulike fordelinger for $h_0(t)$ i tillegg til Weibullfordelingen, nemlig Gompertzfordelingen, den log-logistiske fordelingen og badekarfordelingen. Det er kun vist resultater fra èn simulering med log-logistisk fordelt $h_0(t)$. Grunnen til dette er for det første at dataene generert med denne underliggende hasarden tilpasser seg veldig godt modellen med Weibull $h_0(t)$. Denne fordelingen har flere likhetstrekk med Weibullfordelingen, og itillegg er det vanskelig å sette parametrene i fordelingen slik at levetidene blir spredt utover et stort tidsrom og samtidig ha 90% sensurering. Dermed får vi ikke utnyttet den karakteristiske forskjellen mellom en Weibull- og log-logistiskfordelt hasard. Den andre grunnen er at Gompertzfordelingen og badekarfordelingen i virkeligheten oftere opptrer som hasardfordelinger.

Når $h_0(t)$ er Gompertzfordelt får vi markant skjevere estimater enn når

$h_0(t)$ er log-logistisk fordelt. Grunnen til dette kan være at for Gompertz-fordelingen er det enklere å benytte en større del av hasarden, samtidig som det er større forskjeller mellom de to hasardene. For eksempel er Weibull-hasarden alltid 0 ved tid 0, mens Gompertzhasarden har verdien $\lambda$ ved tid 0. Når vi genererer data med gammafordelt frailty-variabel, får vi generelt bedre estimater dersom $h_0(t)$ er Gompertzfordelt og ikke badekarformet. Dette gjelder uavhengig av om vi antar korrekt frailty-fordeling eller ikke. Når frailty-variabelen er stablefordelt, er det imidlertid en badekarformet hasard som gir best estimater. Grunnen til dette kan være at stablefordelingen er svært skjevfordelt og har mange levetider svært nær 0. Da vil badekarformen være av mindre betydning, siden vi får mange svært tidlige levetider uansett. Siden badekarfordelingen som er definert her er Weibullfordelt for $t > 0.1$, så vil ikke feilen her bli like stor.

Når $Y$ er gammafordelt spiller graden av avhengighet innad i familier liten rolle for estimeringen av regresjonsparametrene. Når $Y$ er stablefordelt ser det imidlertid ut til at større grad av avhengighet fører til bedre estimater for regresjonsparametrene. Selve graden av avhengigheten blir underestimert i prosentvis større grad når avhengigheten er lav $(0.4 - 0.5)$ enn når den er høy $(0.7 - 0.8)$. Dette gjelder særlig når $Y$ er stablefordelt, men også når den er gammafordelt. Dette skyldes sannsynligvis at lav avhengighet fører til færre familier med mer enn ett sykdomstilfelle og dermed blir estimeringen mer unøyaktig.

Som nevnt er det vist at feilspesifisert frailty-fordeling i verste fall fører til 10% skjevhet i estimatene (Hsu og Gorfine, 2007). Dette viser at feilspesifisert frailty-fordeling har relativt liten betydning dersom $h_0(t)$ er korrekt spesifisert. Til slutt i denne oppgaven har vi undersøkt betydningen av feilspesifisert frailty-fordeling dersom også $h_0(t)$ er feilspesifisert. Vi har gjort to ulike simuleringer der dataene er generert på nøyaktig samme måte, men der vi i den første simuleringen antar korrekt frailty-fordeling og i den andre simuleringen antar feil frailty-fordeling. Det viser seg da at dersom $Y$ er gammafordelt, mens vi antar stablefordeling, fører dette til markant dårligere estimater. Estimatene for regresjonsparametrene blir litt dårligere, men det mest alvorlige er estimatet for avhengigheten som underestimeres i svært stor grad. Dersom $Y$ er stablefordelt, mens vi antar gammafordeling, blir estimatene for regresjonsparametrene underestimert i langt større grad enn med korrekt frailty-fordeling. Grunnen til dette kan være at avhengigheten blir sterkt overestimert og dermed får tillegges regresjonsparametrene desto mindre betydning. Avhengigheten blir overestimert fordi variansen i stablefordelingen er uendelig, og dermed får vi en dårlig tilpasning til gammafordelingen. Disse undersøkelsene er basert på både badekar- og Gompertzfordelt underliggende hasard.

Generelt kan man si at når levetidene er spredd utover et lite tidsrom, slik at den karakteristriske formen til $h_0(t)$ ikke er utnyttet, så har små avvik i $h_0(t)$ relativt liten betydning. Når det motsatte er tilfelle, det vil si

at levetidene er spredd utover et stort tidsrom slik at en stor del av $h_0(t)$ blir benyttet og den karakteristiske formen til hasarden er tydelig i dette tidsrommet, har vi sett en rekke eksempler på at det kan gå svært galt.

# Referanser


[1] Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**: 439–454.

[2] Hougaard,P. (1986b). A class of multivariate failure time distributions. *Biometrika*, **73**: 671-678.

[3] Clayton, D. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**:141-151.

[4] Vaupel, J.W. and Yashin A.I. (1985a). Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *The American Statistician*, **39**: 176-185

[5] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer

[6] Hsu, L. and Gorfine, M. (2007). Effect of Frailty Distribution Misspecification on Marginal Regression Estimates and Hazard Functions in Multivariate Survival Analysis. *Statistics in medicine.*

[7] Moger, T. A., Pawitan, Y. and Borgan, Ø. (submitted). Case-cohort methods for survival data on families from routine registers. *Statistics in medicine.*

[8] Moger, T. A. (2004). *Frailty models based on PVF distributions, with emphasis on models for analysing family data*. Unipub forlag

[9] Klein, J.P. and Moeschberger, M.L. (2002). *Survival analysis*. New York: Springer