Universiteit Utrecht

Faculteit Bètawetenschappen

Survival analyse met imperfecte data

Bachelor Thesis

Kerim Delic

Studentnummer: 6004261

Studie: Wiskunde en toepassingen

Begeleider:

Dr.M.C.J. Bootsma
Mathematisch Instituut

Samenvatting

In deze thesis wordt het probleem van imperfecte data binnen de *survival analyse* behandeld door een model te construeren voor imperfecte testen. De definities van de *survival functie* en de *hazard functie* worden eerst gegeven. Op basis van de theorie van telprocessen worden er twee schatters afgeleid voor deze functies, de *Nelson-Aalen* schatter en de *Kaplan-Meier* schatter. Vervolgens wordt het effect van covariaten door middel van het *Cox proportional hazards model* toegelicht, waarbij de gedeeltelijke aannemelijkheid zal worden geïntroduceerd. Tot slot wordt het *Cox proportional hazards model* uitgebreid door de aannemelijkheidsfunctie te definiëren voor imperfecte testen.

8 juni 2020

Inhoudsopgave

1 Introductie

Survival analyse is een statistische term voor methoden voor het analyseren van zogeheten *survival data*. Deze data bevat gegevens over de tijdsduur totdat er een bepaalde gebeurtenis plaatsvindt. Een voorbeeld van een gebeurtenis is het ontslag van een patiënt uit het ziekenhuis. De *survival data* bevat dan informatie over de lengte van de periode die een bepaalde patiënt verblijft in het ziekenhuis. Vaak worden in de literatuur dit soort gebeurtenissen aangeduid als *dood*. In deze thesis zal gebruik worden gemaakt van de term gebeurtenis. Bij de evaluatie van de tijdsduur tot deze gebeurtenis moet er ook per patiënt rekening worden gehouden met een aantal bijbehorende variabelen. Deze variabelen die effect kunnen hebben op de uitkomst noemen we covariaten. De covariaten zorgen ervoor dat we bepaalde resultaten in perspectief zien. Covariaten als leeftijd, het aantal jaren dat de patiënt heeft gerookt of hartfunctie kunnen bijvoorbeeld van grote invloed zijn op het tijdstip van het ontslag uit het ziekenhuis en moeten dus worden verwerkt bij het analyseren. Verder neemt *survival data* ook een reeks aan andere complicaties met zich mee, waardoor het verschilt van andere soorten data. Deze complicaties worden geïllustreerd met het volgende voorbeeld van een onderzoek waarin tijdstippen van ontslagen uit het ziekenhuis worden geobserveerd.

Stel dat voor een periode van 30 dagen een aantal patiënten in het ziekenhuis worden gevolgd. Het kan voorkomen dat een patiënt na 30 dagen nog niet is ontslagen uit het ziekenhuis. Het enige gegeven is het feit dat de patiënt na 30 dagen is ontslagen, met de reële aanname dat een patiënt niet oneindig lang in een ziekenhuis kan verblijven. Dit hoeft niet de enige reden te zijn voor complicaties. Het kan ook voorkomen dat na 10 dagen de patiënt verdwenen is uit het onderzoek, maar niet de gebeurtenis 'ontslag' heeft ervaren. In de praktijk weten we dus soms niet wanneer de gebeurtenis exact heeft plaatsgevonden, in dit geval weten we alleen dat de gebeurtenis voor of na een bepaald tijdstip heeft plaatsgevonden. Data met deze eigenschap wordt *censored data* genoemd. Het tijdstip waarvoor er wordt geëvalueerd of een gebeurtenis er voor of na heeft plaatsgevonden wordt ook wel het *censoring tijdstip* genoemd. In het vervolg worden, indien er sprake is van *censored data*, gegevens van steekproeven beschouwd waarbij de gebeurtenis na het *censoring tijdstip* plaatsvindt. In de literatuur wordt er dan gesproken van *rechter censoring*. Bij *linker censoring* is er sprake van een *censoring tijdstip* waarbij de gebeurtenis hiervoor heeft plaatsgevonden. Daarnaast is er nog een variant van *censoring* genaamd *interval censoring*, waarbij de gebeurtenis plaatsvindt binnen een bepaald tijdsinterval. De focus in deze thesis zal liggen op *rechter censoring*, maar sommige uitgewerkte resultaten gelden ook voor steekproeven met een ander type *censoring*.

Eerst worden er een aantal definities, relaties en eigenschappen geïntroduceerd over de *survival functie*, *hazard functie* en telprocessen. Indien het nodig is zullen twee gevallen worden onderscheiden, het continue en het discrete geval. Met deze opgebouwde kennis zullen schatters worden afgeleid en toegelicht.
Vervolgens wordt er een regressiemodel beschouwd dat ook wel bekend staat als het *Cox proportional hazards model*. In dit model zullen covariaten worden belicht. Nadat er een beeld is ontstaan van de wijze waarop dit model werkt, wordt er tot slot een model beschouwd voor imperfecte data. Imperfecte data heeft de extra eigenschap dat de observaties van de gebeurtenissen niet volledig betrouwbaar zijn. Dit houdt in dat de data informatie bevat, waarbij de kans aanwezig is dat deze informatie onjuist is. Het is mogelijk dat er bijvoorbeeld wordt geobserveerd dat een patiënt acht dagen in het ziekenhuis heeft verbleven, terwijl dit in realiteit zeven dagen zijn. De nadruk van deze thesis zal vooral liggen bij resultaten die voortkomen uit een onderzoek met imperfecte testen. Dit soort testen zijn dan niet volledig betrouwbaar en de gevoeligheid van deze test gaat een rol spelen bij het analyseren. In een later stadium zal dit geval verder worden uiteengezet.

Merk op dat deze thesis volledig in het Nederlands is geschreven. Bepaalde wiskundige begrippen zijn in het Engels gelaten, maar grotendeels zijn ook deze begrippen naar het Nederlands vertaald. Voor de lezer staat in appendix A een lijst met de vertaalde begrippen, zodat de koppeling gemaakt kan worden met verdere Engelstalige literatuur over het desbetreffende onderwerp.

2 Survival en hazard

Dit hoofdstuk is geschreven op basis van hoofdstuk 2 van *Mortensen* [2], waarbij een aantal argumenten zijn toegevoegd voor bepaalde beweringen en afleidingen.
Het hoofdstuk bestaat uit twee delen. In het eerste deel wordt de *survival-* en *hazard functie* gedefinieerd voor een continue stochastische variabele en zullen een aantal eigenschappen en relaties worden beschouwd. In het tweede deel van dit hoofdstuk zal er op een analoge wijze het geval van een discrete stochastische variabele worden belicht.

2.1 Continue stochastische variabele

De verdeling van een positieve continue stochastische variabele X kan worden beschreven door de cumulatieve verdelingsfunctie $F(\cdot)$, waarbij in dit geval deze variabele de tijdsduur tot een bepaalde gebeurtenis representeert. Deze cumulatieve verdelingsfunctie wordt gegeven door:

$$F(t) = P(X \leq t), \qquad \text{met } t \geq 0.$$

De cumulatieve verdelingsfunctie kan ook met de kansdichtheid $f(\cdot)$ worden beschreven:

$$F(t) = \int_0^t f(\tau)d\tau.$$

De functie $F(t)$ bevat dus informatie over de grootte van de kans dat de gebeurtenis voor een tijdstip t plaatsvindt. De *survival functie* is het complement van deze kans en wordt in de onderstaande paragraaf uiteengezet.

2.1.1 Survival functie

De *survival functie* is een niet stijgende functie die in tegenstelling tot de cumulatieve verdelingsfunctie de kans weergeeft dat een individu de gebeurtenis nog niet ervaren heeft tot en met het tijdstip t.

Definitie 2.1.1. *Survival functie*

$$S(t) = P(X > t)$$

Merk op dat de eerder benoemde relatie tussen de *survival functie* en de cumulatieve verdelingsfunctie als volgt uitgedrukt kan worden:

$$S(t) = P(X > t) = 1 - P(X \leq t) = 1 - F(t). \tag{2.1}$$

Een eigenschap van de kansdichtheid $f(\cdot)$ is dat de integraal over het hele domein $[0, \infty)$ genomen gelijk is aan één:

$$\int_0^\infty f(\tau)d\tau = 1.$$

Met deze eigenschap kunnen we de *survival functie* ook weergeven door de vergelijking (2.1) verder uit te schrijven op de volgende manier:

$$S(t) = 1 - F(t) = \int_t^\infty f(\tau)d\tau.$$

Hieruit kunnen we concluderen dat de *survival functie* op tijdstip nul gelijk is aan één. Dus in het begin van de studie weten we met zekerheid dat de gebeurtenis nog niet heeft plaatsgevonden. Daarnaast zien we wel dat het zeker is dat op een gegeven moment de gebeurtenis wel zal plaatsvinden, want de integraal convergeert naar nul:

$$\lim_{t\to\infty} S(t) = \lim_{t\to\infty} \int_t^\infty f(\tau)d\tau = 0.$$

Zoals we ook eerder al opmerkten zal elke patiënt op een gegeven moment het ziekenhuis verlaten.

Merk op dat we de kansdichtheid verder ook kunnen schrijven als de afgeleide van de functie $-S(\cdot)$:

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}(1 - S(t)) = -\frac{d}{dt}S(t). \tag{2.2}$$

Veelal is men in een onderzoek geïnteresseerd in de verwachting van de tijdsduur totdat de gebeurtenis plaatsvindt (denk aan het plannen van bedden voor patiënten in een ziekenhuis, hierbij is het goed om te weten hoe lang een patiënt een bed gemiddeld bezet houdt), ofwel een *gemiddelde tijdsduur*.

Definitie 2.1.2. Voor de continue positieve stochastische variabele X met kansdichtheid $f(x)$ is de verwachting:

$$E[X] = \lim_{n\to\infty} \int_0^n xf(x)dx.$$

We nemen aan dat deze verwachting bestaat, oftewel dat de integraal van de bovenstaande definitie convergeert. Deze verwachting kan dan worden uitgedrukt in termen van de *survival functie* zoals hieronder wordt afgeleid:

$$\begin{aligned}
E[X] &= \lim_{r\to\infty} \int_0^r tf(t)dt \\
&= \lim_{r\to\infty} [tF(t)]_{t=0}^{t=r} - \int_0^r F(t)dt \\
&= \lim_{r\to\infty} rF(r) - \int_0^r 1 - S(t)dt \\
&= \lim_{r\to\infty} r(1 - S(r)) - r + \int_0^r S(t)dt \\
&= \lim_{r\to\infty} \int_0^r S(t)dt - rS(r) \\
&= \lim_{r\to\infty} \int_0^r S(t)dt - r\int_r^\infty f(x)dx.
\end{aligned} \tag{2.3}$$

Vervolgens bekijken we de tweede term in de laatste vergelijking. Deze integraal wordt namelijk geëvalueerd tussen r en oneindig, oftewel $0 \leq r \leq x < \infty$. Uit het majorantie-kenmerk voor integreerbaarheid volgt dan dat de ongelijkheid $|rf(x)| \leq xf(x)$ impliceert dat $|r\int_r^\infty f(x)dx| \leq \int_r^\infty xf(x)dx$. Met het feit dat de verwachting eindig is, kunnen we dan met dit majorantie-kenmerk het volgende afleiden:

$$0 \leq \lim_{r\to\infty} r\int_r^\infty f(x)dx \leq \lim_{r\to\infty} \int_r^\infty xf(x)dx = 0$$

Zo valt bij (2.3) in de laatste vergelijking de tweede term weg en wordt het resultaat:

$$E[X] = \lim_{r\to\infty} \int_0^r S(t)dt.$$

In het geval dat er *censored data* aanwezig is in de dataverzameling kunnen we niet direct op deze manier de *gemiddelde tijdsduur* berekenen. *Rechter censoring* veroorzaakt namelijk dat we alleen weten dat de gebeurtenis na een *censoring tijdstip* plaatsvindt, waardoor het exacte tijdstip van de gebeurtenis onbekend is. Deze *censoring tijdstippen* vormen dan een soort ruis, wat bij verdere analyse problemen kan veroorzaken. Om dit op te lossen wordt er een restrictie opgelegd op de *gemiddelde tijdsduur*. Hiermee wordt dan de periode tot de gebeurtenis geëvalueerd met als grens een gegeven tijdstip τ, dit noemen we dan de *gemiddelde begrensde tijdsduur*. Verder nemen we aan dat de *censoring* onafhankelijk is van de uitkomst.

Definitie 2.1.3. De *gemiddelde begrensde tijdsduur* is als volgt gedefinieerd,

$$\mu_\tau = E[min(X, \tau)]$$

waarbij $\tau > 0$ de grens is.

De *gemiddelde begrensde tijdsduur* kan ook worden beschreven als de integraal van de *survival functie*, hiervoor introduceren we de stochastische variabele $Y = min(X, \tau)$ met de bijbehorende *survival functie* $S_Y(y)$. De *gemiddelde begrensde tijdsduur* kan dan ook worden geformuleerd als:

$$\begin{aligned}
\mu_\tau = E[Y] &= \lim_{r\to\infty} \int_0^r S_Y(u)du \\
&= \lim_{r\to\infty} \int_0^r P(Y > u)du \\
&= \lim_{r\to\infty} \int_0^r P(X > u, \tau > u)du \\
&= \lim_{r\to\infty} \int_0^r P(X > u)P(\tau > u)du \qquad (X \text{ en } \tau \text{ zijn onafhankelijk van elkaar}) \\
&= \lim_{r\to\infty} \int_0^r P(X > u)1\{\text{als } \tau > u\}du \\
&\text{Als } u \geq \tau \text{ dan } P(\tau > u) = 0, \text{ dus wordt deze integraal alleen geëvalueerd van 0 tot } \tau \\
&= \int_0^\tau P(X > u)du \\
&= \int_0^\tau S(u)du.
\end{aligned} \tag{2.4}$$

Op deze manier wordt *censored data* verwerkt door gebruik te maken van de *gemiddelde begrensde tijdsduur* in plaats van de algemene *gemiddelde tijdsduur*.

Het is benoemd dat de *survival functie* de kans weergeeft op het plaatsvinden van de gebeurtenis na een tijdstip t. Deze functie slaagt er alleen niet in om het risico op het lopen van een gebeurtenis in een bepaald tijdsinterval (of op een bepaald tijdstip) in kaart te brengen, dit wordt wel bewerkstelligd met de *hazard functie* die in de volgende paragraaf aan bod komt.

2.1.2 Hazard functie

De *hazard functie* is niet een kans zoals de *survival functie*, maar kan worden gezien als een soort kracht waarmee aan een individu getrokken wordt tot het ervaren van de gebeurtenis. Des te groter de waarde van de *hazard functie* op een tijdstip, des te groter het risico op het ervaren van de gebeurtenis op dit tijdstip. De *hazard functie* kan worden opgevat als een voorwaardelijke kans dat de gebeurtenis in een interval plaatsvindt, waarbij de lengte van dit interval in de limiet naar nul toe gaat. De voorwaarde hierbij is dat de gebeurtenis nog niet heeft plaatsgevonden. Dit staat uitgewerkt in de definitie hieronder.

Definitie 2.1.4. *Hazard functie*

$$h(t) = \lim_{\Delta\tau\to 0} \frac{P(t \leq X < t + \Delta\tau \mid X \geq t)}{\Delta\tau}$$

Op deze wijze wordt dezelfde *survival data* op twee verschillende manieren (met *survival-* en *hazard functie*) uitgedrukt. De *hazard functie* kan weer worden herschreven in termen van de *survival functie*, hiervoor is het volgende inzicht over de kansdichtheid $f(\cdot)$ nodig:

$$\begin{aligned}
f(t) &= \frac{d}{dt}F(t) \\
&= \lim_{\Delta\tau\to 0} \frac{F(t + \Delta\tau) - F(t)}{\Delta\tau} \\
&= \lim_{\Delta\tau\to 0} \frac{P(X \leq t + \Delta\tau) - P(X \leq t)}{\Delta\tau} \\
&= \lim_{\Delta\tau\to 0} \frac{P(t \leq X < t + \Delta\tau)}{\Delta\tau}.
\end{aligned} \tag{2.5}$$

Merk op dat in het continue geval de gelijkheid $P(X < t) = P(X \leq t)$ geldt, omdat dan de kans dat de gebeurtenis exact plaatsvindt op een tijdstip t gelijk is aan nul. Met dit inzicht kan de *hazard functie* $h(t)$ worden uitgedrukt in termen van de kansdichtheid $f(t)$ en de *survival functie* $S(t)$:

$$\begin{aligned} h(t) &= \lim_{\Delta\tau \to 0} \frac{P(t \leq X < t + \Delta\tau)}{\Delta\tau P(X \geq t)} \\ &= \frac{f(t)}{P(X \geq t)} \\ &= \frac{f(t)}{S(t)} \\ &= -\frac{d}{dt} ln[S(t)]. \end{aligned} \tag{2.6}$$

De laatste vergelijking wordt afgeleid met het resultaat van (2.2).

Verder is er naast de *hazard functie* ook een *cumulatieve hazard functie*. In het volgende hoofdstuk zal duidelijk worden waarom het gebruik van een *cumulatieve hazard functie* nuttig is vanuit het oogpunt van het vinden van een geschikte schatter voor zowel de *hazard-* als *survival functie*.

Definitie 2.1.5. De *cumulatieve hazard functie* van een continue stochastische variabele X wordt uitgedrukt als:

$$H(t) = \int_0^t h(\tau) d\tau.$$

De *cumulatieve hazard functie* op tijdstip t is de totale kracht waarmee aan een individu getrokken is tot het ervaren van de gebeurtenis tot t. Des te groter $H(t)$, des te groter de waarschijnlijkheid dat de gebeurtenis voor tijdstip t heeft plaatsgevonden. De *cumulatieve hazard functie* kan ook worden uitgedrukt in termen van de *survival functie* vanuit de laatste vergelijking van (2.6), waardoor het volgende resultaat is af te leiden:

$$H(t) = -ln[S(t)]. \tag{2.7}$$

Hieruit volgt dus ook,

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(\tau) d\tau} \tag{2.8}$$

waaruit met de derde vergelijking van (2.6) en (2.7) de kansdichtheid $f(t)$ weer uitgedrukt kan worden in termen van de *hazard functie*:

$$f(t) = h(t) e^{-\int_0^t h(\tau) d\tau}. \tag{2.9}$$

2.2 Discrete stochastische variabele

In dit deel worden de *survival-* en *hazard functie* beschouwd in het geval dat de stochastische variabele X discreet is. De variabele X neemt dan de waarden t_i aan zodanig dat deze tijdstippen geordend kunnen worden op basis van grootte ($t_0 < t_1 < ... < t_i < ...$). Verder kunnen we de kans $p(\cdot)$ als volgt definiëren,

$$p(t_i) = P(X = t_i)$$

met $i = 0, 1, 2, ...$
De *survival functie* wordt dan beschreven op de volgende manier:

$$S(t) = \sum_{t_i > t} p(t_i). \tag{2.10}$$

De *hazard functie* wordt gezien als de kans dat de gebeurtenis plaatsvindt op tijdstip t_i, gegeven dat de gebeurtenis nog niet heeft plaatsgevonden tot tijdstip t_i:

$$h(t_i) = P(X = t_i \mid X \geq t_i) = \frac{p(t_i)}{S(t_{i-1})},$$

waarbij $S(t_0) = 1$.

Merk verder op dat $p(t_i) = S(t_{i-1}) - S(t_i)$, zodat de *hazard functie* weer herschreven kan worden in termen van de *survival functie*:

$$h(t_i) = 1 - \frac{S(t_i)}{S(t_{i-1})}. \tag{2.11}$$

De discrete *survival functie* van (2.10) op tijdstip t kan ook worden uitgedrukt in termen van de *survival functies* van tijdstippen ($t_i \leq t$) met als voorwaarde dat de gebeurtenis nog niet heeft plaatsgevonden tot tijdstip t_{i-1}, oftewel $P(X > t_i \mid X > t_{i-1})$. Door het product van deze voorwaardelijke *survival functies* te nemen, leidt dit tot de *survival functie* van t:

$$S(t) = \prod_{t_i \leq t} \frac{S(t_i)}{S(t_{i-1})}.$$

De *survival functie* kan met (2.11) en het laatste resultaat als volgt worden geformuleerd:

$$S(t) = \prod_{t_i \leq t} (1 - h(t_i)).$$

Op grond van deze laatste relatie wordt de schatter van de *survival functie* (**Kaplan-Meier schatter**) afgeleid, waarbij er eerst een schatter wordt bepaald voor de *cumulatieve hazard functie* (**Nelson-Aalen schatter**). De totstandkoming van deze schatters wordt in het volgende hoofdstuk verder behandeld met een benadering vanuit telprocessen.

3 Telprocessen

Dit hoofdstuk is geschreven op basis van hoofdstuk 3 van *Mortensen* [2], waarbij een aantal argumenten zijn toegevoegd voor bepaalde beweringen en afleidingen.
Zoals eerder is benoemd, bevat *survival data* informatie over de tijdstippen van gebeurtenissen en of er eventueel sprake is van *censoring*. Deze data kan hierdoor worden gerepresenteerd met een telproces, mits de *censoring tijdstippen* op een juiste manier worden verwerkt. In dit hoofdstuk zal de benodigde theorie van telprocessen worden geïntroduceerd. Op basis van deze resultaten zal een schatter worden beschreven voor de *cumulatieve hazard functie* en zo ook de *survival functie*. Deze twee schatters worden ook wel de **Nelson-Aalen schatter** en de **Kaplan-Meier schatter** genoemd.

In de benadering van *survival data* met telprocessen staan drie processen centraal: telprocessen, *intensiteitsprocessen* en *risicoprocessen*. Eerst beschouwen we het telproces.

Definitie 3.0.1. Een stochastisch proces $\{N(t),\, t \geq 0\}$ is een telproces wanneer het voldoet aan de volgende drie eigenschappen:

- $N(t) \geq 0$
- $N(t)$ is een geheel getal
- Als $s \leq t$ dan $N(s) \leq N(t)$.

Gegeven is een steekproef van n individuen waarbij er in de dataverzameling sprake is van *rechter censoring*. Stel dat X_j het tijdstip is waarop de gebeurtenis plaatsvindt en C_j het *censoring tijdstip* voor individu $j = 1, 2, ..., n$. Verder nemen we aan dat het tijdstip van de gebeurtenis X_j en *censoring* C_j onafhankelijk zijn van elkaar en dat beide variabelen continu zijn. We definiëren dan de volgende positieve stochastische variabele $T_j = min(X_j, C_j)$ met als extra informatie:

$$\delta_j = \begin{cases} 1, & \text{als } X_j \leq C_j \\ 0, & \text{anders} \end{cases}$$

De variabele δ_j kan opgevat worden als een binaire variabele die kennis geeft over het feit of er wel of niet is geobserveerd dat individu j de gebeurtenis heeft ervaren. Het telproces op individueel niveau wordt dan beschreven als:

$$N_j(t) = \begin{cases} 1, & \text{als } T_j \leq t \text{ en } \delta_j = 1 \\ 0, & \text{anders} \end{cases}$$

De som van deze individuele processen geeft dan het aantal gebeurtenissen in totaal die hebben plaatsgevonden en geobserveerd zijn tot en met tijdstip t:

$$N(t) = \sum_{j=1}^{n} N_j(t). \tag{3.1}$$

Het verschil $N(t) - N(s)$ geeft vervolgens het aantal gebeurtenissen die hebben plaatsgevonden in het interval $(s, t]$. De geschiedenis van dit telproces op tijdstip t geeft informatie over de voorgaande tijdstippen (of er wel of geen sprake is van een *censoring tijdstip* voor t). De geschiedenis voor tijdstip t wordt genoteerd als G_t. Verder geldt dat als tijdstip s strikt kleiner is dan tijdstip t dat de geschiedenis van tijdstip s een deelverzameling is van de geschiedenis van tijdstip t:

$$G_s \subseteq G_t, \qquad \text{als } s < t.$$

Nu kan de verandering van het telproces over een interval $[t, t + dt)$ met $dt > 0$ worden gedefinieerd door gebruik te maken van de notatie t^-. Deze notatie geeft het tijdstip vlak voor t weer, oftewel een limiet waarbij t van links wordt benaderd. De verandering schrijven we dan als volgt:

$$dN(t) = N([t + dt]^-) - N(t^-).$$

Eerder is opgemerkt dat de variabelen X_j en C_j beide continu zijn. Vanwege deze assumptie is de stochastische variabele T_j ook continu en is de kans dat twee gebeurtenissen op hetzelfde tijdstip plaatsvinden gelijk aan nul. In het geval dat we dt dan willekeurig klein nemen, kan ook $dN(t)$ worden gezien als een binaire variabele die de waarde nul of één aanneemt. Deze uitkomsten zijn dan equivalent aan of er wel of geen gebeurtenis plaatsvindt in het interval. Op basis hiervan introduceren we het volgende proces dat centraal staat binnen deze benadering. Dit proces zal in deze thesis het *intensiteitsproces* worden genoemd.

Definitie 3.0.2. Het *intensiteitsproces* $\lambda(\cdot)$ van een gegeven telproces $N(\cdot)$ is als volgt gedefinieerd:

$$\lambda(t) = \lim_{dt \to 0} \frac{P(dN(t) = 1 \mid G_{t^-})}{dt}.$$

Het *intensiteitsproces* is op een soortgelijke wijze geconstrueerd als de *hazard functie*. Dit proces geeft dus ook een kracht weer waarmee aan een individu getrokken wordt tot het ervaren van de gebeurtenis. De voorwaardelijke kans dat individu j (gegeven de geschiedenis op tijdstip t) de gebeurtenis ervaart in het interval $[t, t + dt)$ wordt dan beschreven als:

$$P(t \leq T_j < t + dt, \delta_j = 1 \mid G_{t^-}).$$

Dit kan weer worden herschreven in termen van de *hazard functie* op de volgende manier, mits dt willekeurig klein is:

$$\begin{aligned} P(t \leq T_j < t + dt, \delta_j = 1 \mid G_{t^-}) &= P(t \leq X_j < t + dt, C_j \geq t + dt \mid X_j \geq t, C_j \geq t) \\ &= P(t \leq X_j < t + dt \mid X_j \geq t) P(C_j \geq t + dt \mid C_j \geq t) \\ &= \frac{F(t + dt) - F(t)}{S(t)} P(C_j \geq t + dt \mid C_j \geq t) \\ &= \frac{f(t)dt}{S(t)} P(C_j \geq t + dt \mid C_j \geq t) \\ &\approx h(t)dt. \end{aligned} \tag{3.2}$$

Het laatste resultaat volgt uit hoofdstuk 2 en de aanname dat dt naar nul gaat, zodat de kans $P(C_j \geq t+dt \mid C_j \geq t)$ naar één gaat.

Het *intensiteitsproces* kan ook worden uitgedrukt in termen van het *risicoproces*. Voordat dit resultaat wordt afgeleid, zal eerst nog een indicatorfunctie $\gamma_j(\cdot)$ en de definitie van het *risicoproces* worden geïntroduceerd:

$$\gamma_j(t) = \begin{cases} 1, & \text{als } T_j \geq t \\ 0, & \text{anders} \end{cases}$$

Deze functie geeft een antwoord op de vraag of individu j risico loopt op het ervaren van de gebeurtenis op tijdstip t door simpelweg te evalueren of het individu de gebeurtenis/*censoring* voor of na t heeft ervaren. Het *risicoproces* is vervolgens de som over de bovenstaande functies $\gamma_j(\cdot)$ met $j = 1, 2, ..., n$ en stelt dus het aantal individuen voor die risico lopen op het ervaren van de gebeurtenis op een bepaald tijdstip.

Definitie 3.0.3. Het *risiscoproces* $Y(\cdot)$ wordt gedefinieerd als,

$$Y(t) = \sum_{j=1}^{n} \gamma_j(t).$$

Vervolgens wordt de voorwaardelijke kans die deel uitmaakt van het *intensiteitsproces* weer herzien, wat tot het volgende inzicht resulteert:

$$\begin{aligned} P(dN(t) = 1 \mid G_{t^-}) &= E[dN(t) \mid G_{t^-}] \\ &= E[\#\{j : T_j \in [t, t + dt), \delta_j = 1\} \mid G_{t^-}] \\ &= Y(t)h(t)dt. \end{aligned} \tag{3.3}$$

Oftewel, het blijkt dat het *intensiteitsproces* als product van het *risicoproces* en *hazard functie* geschreven kan worden:

$$\lambda(t) = Y(t)h(t).$$

Verder geldt ook dat het *intensiteitsproces* uitgedrukt kan worden met de verwachting van het aantal individuen dat in het interval $[t, t + dt)$ de gebeurtenis ervaart,

$$\lambda(t) = \frac{E[dN(t) \mid G_{t^-}]}{dt}$$

hierdoor kan het totale aantal verwachte gebeurtenissen worden beschreven met het *cumulatief intensiteitsproces*.

Definitie 3.0.4. Het *cumulatief intensiteitsproces* is gedefinieerd als,

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

Het verschil tussen het totale aantal gebeurtenissen en het totale aantal verwachte gebeurtenissen op een tijdstip t noteren we dan als $M(t)$.

Definitie 3.0.5. Een stochastisch proces $\{M(t), t \geq 0\}$ wordt een martingaal genoemd wanneer het proces de volgende twee eigenschappen heeft:

- $E[M(t)] < \infty$
- $E[M(t) \mid G_s] = M(s)$, voor alle $s \leq t$.

Om te laten zien dat een proces een martingaal is, moeten de twee eigenschappen in de bovenstaande definitie gelden voor dit proces.

Definitie 3.0.6. De telproces martingaal wordt voor een telproces $N(\cdot)$ en *intensiteitsproces* $\Lambda(\cdot)$ gedefinieerd als:

$$M(t) = N(t) - \Lambda(t).$$

De bewering is dat het verschil $M(\cdot)$ een martingaal is. We zullen dit de telproces martingaal noemen, zoals is aangegeven in de bovenstaande definitie. De telproces martingaal voldoet duidelijk aan de eerste eigenschap. Om te laten zien dat de tweede eigenschap ook geldt, wordt eerst het volgende afgeleid:

$$\begin{aligned} E[dM(t) \mid G_{t^-}] &= E[dN(t) - d\Lambda(t) \mid G_{t^-}] \\ &= E[dN(t) \mid G_{t^-}] - E[\lambda(t)dt \mid G_{t^-}] \\ &= 0. \end{aligned} \tag{3.4}$$

Dit resultaat is genoeg om in te zien dat het proces $M(\cdot)$ de tweede eigenschap van een martingaal bezit, dit leiden we als volgt af voor tijdstip $s \leq t$:

$$\begin{aligned} E[M(t) \mid G_s] - M(s) &= E[M(t) - M(s) \mid G_s] \\ &= E[\int_s^t dM(u)du \mid G_s] \\ &= \int_s^t E[dM(u) \mid G_s]du \\ &= \int_s^t E[E[dM(u) \mid G_{u^-}] \mid G_s]du \\ &= 0. \end{aligned} \tag{3.5}$$

De conclusie is dat het verschil $M(\cdot)$ inderdaad een martingaal is.

In deze context wordt het *cumulatief intensiteitsproces* $\Lambda(\cdot)$ ook wel een *compensator* genoemd van een proces.

Definitie 3.0.7. Een *compensator* is een proces $\hat{X}(\cdot)$, waarvoor voor een gegeven proces $X(\cdot)$ het verschil $X(t) - \hat{X}(t)$ op een gegeven tijdstip $t \geq 0$ een martingaal is.

Het *cumulatief intensiteitsproces* $\Lambda(\cdot)$ is de *compensator* van het telproces $N(\cdot)$. Verder is de variantie van de telproces martingaal ook een *compensator* voor het proces $M^2(\cdot)$ en wordt deze *compensator* genoteerd als $\langle M(\cdot)\rangle$.
Om dit te laten zien, beschouwen we nu eerst de verandering $dM^2(t)$ van $M^2(t)$ over het interval $[t, t+dt)$:

$$\begin{aligned} dM^2(t) &= M^2([t+dt]^-) - M^2(t^-) \\ &= (dM(t) + M(t^-))^2 - M^2(t^-) \\ &= (dM(t))^2 + 2M(t^-)dM(t). \end{aligned} \tag{3.6}$$

Verder weten we uit (3.4) dat de verwachting van de verandering van de telproces martingaal op tijdstip t (gegeven de geschiedenis op tijdstip t^-) gelijk is aan 0:

$$E[2M(t^-)dM(t) \mid G_{t^-}] = 2M(t^-)E[dM(t) \mid G_{t^-}] = 0.$$

Uit de vergelijking van (3.6) en het bovenstaande resultaat volgt dan de gelijkheid:

$$E[dM^2(t) \mid G_{t^-}] = E[(dM(t))^2 \mid G_{t^-}].$$

Dus de verandering van de *compensator* is gelijk aan de variantie van de verandering van de telproces martingaal,

$$d\langle M\rangle(t) = E[(dM(t))^2 \mid G_{t^-}] = Var[dM(t) \mid G_{t^-}].$$

Vanwege deze gelijkheid geldt hetzelfde resultaat als in (3.4), de verwachting van de verandering van $M^2(t) - \langle M\rangle(t)$ is gelijk aan nul. Zoals we in (3.5) hebben gezien, is dit equivalent aan de martingaal eigenschap. Dus is $M^2(t) - \langle M\rangle(t)$ ook daadwerkelijk een martingaal. Voor een willekeurig klein interval $[t, t+dt)$ hebben we toegelicht dat de verandering van het telproces als een binaire variabele kan worden gezien (neemt waarde nul of één aan), zodoende is de variantie van de verandering van de telproces martingaal ongeveer gelijk aan het *intensiteitsproces*:

$$\begin{aligned} Var[dM(t) \mid G_{t^-}] &= Var[dN(t) - d\Lambda(t) \mid G_{t^-}] \\ &= Var[dN(t) - E[dN(t) \mid G_{t^-}] \mid G_{t^-}] \\ &= Var[dN(t) \mid G_{t^-}] \\ &= E[(dN(t))^2] - E[dN(t)]^2 \\ &= d\Lambda(t) - d\Lambda(t)^2 \qquad\qquad (dN(t) = 0 \text{ of } 1, \text{ dus } E[(dN(t))^2] = E[(dN(t)]) \\ &\approx d\Lambda(t). \end{aligned} \tag{3.7}$$

Met de kennis die we nu hebben opgebouwd over telprocessen kunnen we de constructie van de eerder benoemde schatters behandelen. Dit wordt in de volgende twee paragrafen toegelicht.

3.1 Nelson-Aalen schatter

Op basis van de bovenstaande benadering vanuit telprocessen kunnen we een schatter vinden van de *cumulatieve hazard functie* $H(\cdot)$, deze schatter wordt ook wel de Nelson-Aalen schatter genoemd. We hebben in het vorige deel een telproces *martingaal* gedefinieerd als $M(t) = N(t) - \Lambda(t)$ en hieruit blijkt dat de verandering van het telproces als volgt kan worden geschreven:

$$dN(t) = Y(t)h(t)dt + dM(t).$$

Laat nu $Y(t) > 0$, dan kan deze gelijkheid ook worden geschreven als,

$$\frac{dN(t)}{Y(t)} = h(t)dt + \frac{dM(t)}{Y(t)}.$$

De Nelson-Aalen schatter $\hat{H}(\cdot)$ van de *cumulatieve hazard functie* wordt gedefinieerd als de integraal van deze gelijkheid, waarbij de term $\frac{dM(t)}{Y(t)}$ zal worden geïnterpreteerd als de foutterm van de schatter:

$$\hat{H}(t) := \int_0^t \frac{dN(u)}{Y(u)} du.$$

Merk op dat als $Y(t) = 0$, dan ook $\hat{H}(t) = 0$.
De term $\frac{dN(t)}{Y(t)}$ is gelijk aan $\frac{\#\text{gebeurtenissen in } [t,t+dt)}{\#\text{individuen die risico lopen}}$. Dus is de schatter van de *cumulatieve hazard functie* op tijdstip t in principe een som van deze relatieve waarden over deze kleine intervallen tot en met tijdstip t.

We zien daarnaast dat de Nelson-Aalen schatter uit twee termen bestaat. Een term is de integraal $\int_0^t h(u)du$ en dit is de definitie van de *cumulatieve hazard functie*. De andere term kan dan gezien worden als een soort foutterm die het verschil geeft tussen de schatter en de daadwerkelijke *cumulatieve hazard functie*. Wat kunnen we nu zeggen over deze foutterm?
Ten eerste kijken we naar de verwachting van de foutterm om een uitspraak te kunnen doen over de zuiverheid van de schatter,

$$E[\frac{dM(t)}{Y(t)} \mid G_{t^-}] = \frac{E[dM(t) \mid G_{t^-}]}{Y(t)} = 0.$$

Het *risicoproces* $Y(t)$ voor een gegeven tijdstip t met de kennis van de geschiedenis tot tijdstip t^- is dan een constante. Hieruit volgt dat de uitkomst van deze voorwaardelijk verwachting gelijk is aan het proces zelf. Verder weten we met (3.4) dat de voorwaardelijke verwachting van de verandering van de telproces martingaal gelijk is aan 0. Oftewel, de schatter is in dit geval ongeveer zuiver. Naast de verwachting zijn we ook geïnteresseerd in de grootte van de variantie, want des te kleiner de variantie des te betrouwbaarder de schatter:

$$Var[\frac{dM(t)}{Y(t)} \mid G_{t^-}] = \frac{Var[dM(t) \mid G_{t^-}]}{Y(t)^2} = \frac{d\langle M\rangle(t)}{Y(t)^2}.$$

Dit resultaat is af te leiden met (3.7).

Vanuit de theorie van telprocessen is er dus een benadering gevonden van *survival data* en met name een schatter van de *cumulatieve hazard functie*. Op basis van deze schatter wordt vervolgens de Kaplan-Meier schatter in het volgende deel geïntroduceerd.

3.2 Kaplan-Meier schatter

De Kaplan-Meier schatter geeft een schatting van de *survival functie* $S(\cdot)$. Deze schatter is gebaseerd op de discrete definitie van de *survival functie* ($S(t) = \prod_{t_j \leq t}(1 - h(t_j))$) en de Nelson-Aalen schatter $\hat{H}(t)$:

$$\hat{S}(t) := \prod_{T_j \leq t} [1 - d\hat{H}(T_j)].$$

Het tijdsinterval wordt in partities verdeeld, laten we zeggen dat elke partitie één tijdseenheid (één dag/uur) is. Vervolgens wordt er voor elke tijdseenheid de kans dat de gebeurtenis niet plaatsvindt bepaald. Hierdoor heeft de Kaplan-Meier schatter een soort trapvorm. Op het moment dat er een gebeurtenis plaatsvindt, wordt er een sprong naar beneden gemaakt en anders blijft de schatter constant.

Tot nu toe hebben we dus een manier gevonden om de functies te schatten, maar hebben we covariaten buiten beschouwing gelaten. In het volgende hoofdstuk wordt een model beschreven waarin de covariaten wel worden verwerkt om zo betrouwbaardere uitspraken te kunnen doen over de kansen op het ervaren van de gebeurtenis van een individu met bekende covariaten.

4 Cox proportional hazards model

Dit hoofdstuk is geschreven op basis van *Cox* [1] en hoofdstuk 4 (4.1, 4.1.1, 4.1.2 en 4.1.4) van *Mortensen* [2], waarbij naast een voorbeeld ook een aantal argumenten zijn toegevoegd voor bepaalde beweringen en afleidingen.
In dit hoofdstuk wordt een regressiemodel beschouwd van de *survival data* genaamd het *Cox proportional hazards model*. In dit model worden zoals eerder is benoemd de covariaten behandeld. Een voorbeeld van een covariaat binnen het onderzoek van het ziekenhuis (beschreven in de introductie) kan het aantal jaren zijn dat een individu heeft gerookt. Men is dan vooral geïnteresseerd in de relatie tussen de rook-jaren en de kans op ontslag. Oftewel, wordt de kans op onstlag groter naarmate een individu minder jaren heeft gerookt. Dit soort relaties kunnen in kaart worden gebracht met het *Cox proportional hazards model*, waarbij er gekeken wordt naar de invloed van de covariaten op de *hazard functie*.

Nogmaals spreken we ook in dit hoofdstuk van een steekproef met *rechter censoring*. Gegeven is een n aantal individuen waarvan de covariaten bekend zijn. De data bestaat dan uit tupels (T_j, δ_j, Z_j) met $j = 1, 2, ..., n$. Hierbij zijn de variabelen T_j en δ_j gedefinieerd zoals in hoofdstuk 3. Daarnaast nemen we aan dat niet alleen de gebeurtenissen X_j en C_j (tijdstippen van gebeurtenis en *censoring*) onafhankelijk zijn van elkaar, maar dat ook de gebeurtenissen T_j (voor $j = 1, 2, ..., n$) onderling onafhankelijk zijn. Verder is $Z_j = [Z_{j1}, Z_{j2}, ..., Z_{jp}]^\top$ een vector van p covariaten gekoppeld aan individu j. De voorwaardelijke *hazard functie* $h(t|Z_j)$ voor een individu (met de bijbehorende covariaten als voorwaarde) wordt in het algemeen als een semi-parametrisch *hazard model* beschreven op de volgende manier:

$$h(t|Z_j) = h_0(t)c(\beta^\top Z_j). \tag{4.1}$$

Hierbij is de functie $h_0(t)$ een onbekende (de verdeling is onbekend en daarom wordt dit model ook gezien als gedeeltelijk parametrisch, ofwel semi-parametrisch) positieve functie die we de *baseline hazard functie* noemen. De tweede term beschreven als $c(\beta^\top Z_j)$ is een positieve functie van de covariaten die de *link functie* genoemd wordt, waarbij $\beta = [\beta_1, \beta_2, ..., \beta_p]^\top$ een vector van onbekende parameters is. We nemen verder ook aan dat deze covariaten tijdsinvariant zijn.

Het *Cox proportional hazards model* behoort tot de klasse van modellen die gegeven is in (4.1). De *link functie* is voor dit specifieke model gegeven als:

$$c(\beta^\top Z_j) = e^{\sum_{l=1}^{p} \beta_l Z_{jl}} = e^{\beta^\top Z_j}.$$

Dit resulteert vanuit (4.1) tot in het model beschreven als volgt,

$$h(t|Z_j) = h_0(t)e^{\beta^\top Z_j} \tag{4.2}$$

en dit kan worden herschreven als een lineair model van de logaritme van de breuk $\frac{h(t|Z_j)}{h_0(t)}$, mits $h_0(t) > 0$:

$$log[\frac{h(t \mid Z_j)}{h_0(t)}] = \beta^\top Z_j.$$

We zien dan dat een verandering van één eenheid in Z_{ji} voor een verandering van β_i eenheden zorgt ten opzichte van de logaritme van $\frac{h(t|Z_j)}{h_0(t)}$.
Een eigenschap van het model is dat twee individuen (met verschillende waarden van de covariaten, respectievelijk Z' en Z'') met elkaar vergeleken kunnen worden door naar de verhouding te kijken tussen de twee uitkomsten:

$$\frac{h(t|Z')}{h(t|Z'')} = \frac{h_0(t)e^{\beta^\top Z'}}{h_0(t)e^{\beta^\top Z''}} = e^{\beta^\top (Z'-Z'')}. \tag{4.3}$$

Merk op dat de verhouding niet varieert over de tijd. We noemen de verhouding ook wel het *relatieve risico* van een individu met covariaten Z' op het ervaren van de gebeurtenis ten opzichte van een ander individu met covariaten Z''.

Met de relaties die we hebben afgeleid in hoofdstuk 2 ((2.8) en (2.9)) kunnen we vanuit het *Cox proportional hazards model* ook de kansdichtheid en *survival* functie (beide geconditioneerd op Z) afleiden:

$$f(t|Z) = h(t|Z)e^{-\int_0^t h(u|Z)du} = h_0(t)e^{\beta^\top Z}e^{-e^{\beta^\top Z}\int_0^t h_0(u)du},$$

$$S(t|Z) = e^{-e^{\beta^\top Z}\int_0^t h_0(u)du}.$$

Om een uitspraak te kunnen doen over de onbekende parameters binnen het model is het een logische stap om de meest aannemelijke schatter (MLE) te behandelen. Alleen is het voor het *Cox proportional hazards model* onmogelijk om de gezamenlijke aannemelijkheidsfunctie te bepalen vanwege het feit dat de verdeling van de *baseline hazard functie* onbekend is. Om toch tot een schatting te komen van de onbekende parameter β wordt de *gedeeltelijke aannemelijkheid* gemaximaliseerd.

4.1 Gedeeltelijke aannemelijkheid

Gegeven is een dataverzameling bestaande uit steekproeven van de stochastische vector Y met kansdichtheid $f(y, \beta, \theta)$. Stel nu dat er een transformatie (onafhankelijk van β) van Y naar een verzameling stochastische variabelen (U, V) bestaat. Om precies te zijn wordt Y getransformeerd naar een reeks $(U_1, V_1, U_2, V_2, ..., U_m, V_m)$, waarbij deze componenten van de reeks weer vectoren kunnen zijn. We nemen aan dat de gezamenlijke aannemelijkheid als volgt uitgedrukt kan worden,

$$\prod_{i=1}^{m} f(u_i|u^{(i-1)}, v^{(i-1)}, \beta, \theta) \prod_{i=1}^{m} f(v_i|u^{(i)}, v^{(i-1)}, \beta) \tag{4.4}$$

met θ als een parameter waar we in eerste opzicht niet in geïnteresseerd zijn, maar wel rekening mee moeten houden. De parameter waar we wel wat over willen weten is β. Verder is $u^{(i)} = (u_1, u_2, ..., u_i)$ en $v^{(i)} = (v_1, v_2, ..., v_i)$. De aannemelijkheid is voor simpele gevallen zoals *Cox* in [1] beschrijft: "de gezamenlijke kansdichtheid van de geobserveerde data weergegeven als een functie van de onbekende parameters".

Voorbeeld 4.1.1. Om een beeld te krijgen van de keuze van het beschrijven van de gezamenlijke aannemelijkheid zoals hierboven is gedefinieerd, herzien we ons onderzoek in het ziekenhuis. Voor het gemak gaan we ervan uit dat een individu kan worden ontslagen uit het ziekenhuis indien de persoon in kwestie geen beademingsapparatuur nodig heeft. Stel nu dat Y informatie representeert over het feit of een individu uit het ziekenhuis is ontslagen. De transformatie die we hebben gedefinieerd kan in dit geval tot twee stochastische variabelen U en V leiden die enerzijds informatie geven over de longfunctie en anderzijds over de noodzaak van het gebruiken van een beademingsapparaat. Hierbij hangt de longfunctie op dag i af van:

- De longfunctie tijdens de voorgaande dagen (u^{i-1})
- Of de persoon de laatste dagen een beademingsapparaat gebruikte (v^{i-1})
- Een aantal parameters β die horen bij de covariaten
- Parameter θ behorend bij het aantal jaren dat het individu heeft gerookt

We zijn in eerste instantie niet geïnteresseerd in de laatste parameter, maar het is mogelijk dat θ gerelateerd is aan β en onze schattingen van β verstoort. De noodzaak van een beademingsapparaat op dag i hangt af van:

- Het feit of het individu de voorgaande dagen ook gebruik maakte van het beademingsapparaat (v^{i-1})
- De longfunctie tijdens de voorgaande dagen en dag i (u^i)
- Een aantal β parameters die horen bij de covariaten

We zien dat het het aantal rook-jaren niet direct van invloed is op de vraag of de persoon een beademingsapparaat nodig heeft. De longfunctie hangt wel af van het aantal rook-jaren en is daarentegen wel van invloed. Op deze wijze wordt de aannemelijkheid als het ware gesplitst in een deel waar een variabele als rook-jaren wel van belang speelt en een deel waar dit niet direct een rol heeft op het gebruik van een beademingsapparaat. Zo is er een deel van de aannemelijkheid dat niet afhangt van parameter θ en dus kunnen we dit gedeelte wel gebruiken voor het maken van een adequate schatting van de parameter β. △

Het rechter product (wat niet afhangt van θ) van de gezamenlijke aannemelijkheid wordt gezien als de *gedeeltelijke aannemelijkheid*:

$$L(\beta) = \prod_{i=1}^{m} f(v_i|u^{(i)}, v^{(i-1)}, \beta). \tag{4.5}$$

4.1.1 Gedeeltelijke aannemelijkheid met unieke tijdstippen van gebeurtenissen

Nu we de *gedeeltelijke aannemelijkheid* hebben geïntroduceerd, beschouwen we deze aannemelijkheid voor het *Cox proportional hazards model*. Een belangrijk concept bij de constructie van de *gedeeltelijke aannemelijkheid* is dat het interval (en lengte van het interval) tussen twee opeenvolgende gebeurtenissen geen informatie geeft over de invloed van de covariaten op de *hazard functie*. Hierbij nemen we verder aan dat de tijdstippen van de gebeurtenissen op een continue wijze zijn gemeten, waardoor het niet mogelijk is dat twee gebeurtenissen op exact hetzelfde tijdstip plaatsvinden. Het kan overigens wel voorkomen dat twee gebeurtenissen op eenzelfde tijdstip plaatsvinden en hiervoor wordt de *gedeeltelijke aannemelijkheid* aangepast. In deze thesis zullen we alleen het continue geval bespreken (unieke tijdstippen). Voor een uitleg van de *gedeeltelijke aannemelijkheid* met een discrete tijdmeting wordt de lezer verwezen naar hoofdstuk 4.1.3. van *Mortensen* [2].

We gaan uit van een steekproef van n individuen, waarbij *rechter censoring* van toepassing is. Van deze n individuen hebben in totaal k (met $k \leq n$) individuen de gebeurtenis ervaren en $n-k$ individuen ervaren de gebeurtenis na een *censoring tijdstip*. We ordenen de tijdstippen van de gebeurtenissen als volgt, $t_1 < t_2 < ... < t_k$ en definiëren de verzameling R_i (met $i = 1, 2, ..., k$):

$$R_i = \{j : T_j \geq t_i\}.$$

Deze verzameling noemen we de *risicoverzameling* en geeft het aantal individuen weer waarvan het tijdstip van *censoring* of van de gebeurtenis na tijdstip t_i plaatsvindt.

De *gedeeltelijke aannemelijkheid* is gebaseerd op de voorwaardelijke kans dat individu i de gebeurtenis ervaart in het interval $[t_i, t_i + dt_i)$ gegeven de *risicoverzameling*. Om precies te zijn, is de *gedeeltelijke aannemelijkheid* in dit geval het product van de voorwaardelijke kansen dat de gebeurtenis is ervaren door het individu i op tijdstip t_i gegeven dat er is geobserveerd dat een individu uit de *risicoverzameling* de gebeurtenis ervaart op tijdstip t_i. In het algemeen is dus de gedeeltelijke aannemelijkheid zoals in (4.5) een product van de termen,

$$f(v_i|u^i, v^{i-1}, \beta) \tag{4.6}$$

met $i = 1, 2, ..., k$. Voor het *Cox proportional hazards model* representeert V_i het feit of in het interval $[t_i, t_i + dt_i)$ de gebeurtenis plaatsvindt en is U_i de informatie of er *censoring tijdstippen* zijn in $[t_{i-1}, t_i)$. Dus geeft (4.6) daadwerkelijk de voorwaardelijke kans aan dat individu i de gebeurtenis ervaart in $[t_i, t_i + dt_i)$ gegeven de risico verzameling (want door u^i en v^{i-1} weten we of er al gebeurtenissen/*censoring tijdstippen* zijn plaatsgevonden). Deze kansen worden beschreven door de *hazard functie*, waardoor de *gedeeltelijke aannemelijkheid* op basis hiervan uitgeschreven kan worden:

$$\begin{aligned} L_i(\beta) &= \frac{h(t_i|Z_i)dt_i}{\sum_{j \in R_i} h(t_i|Z_j)dt_i} \\ &= \frac{h_0(t_i)e^{\beta^\top Z_i}dt_i}{\sum_{j \in R_i} h_0(t_i)e^{\beta^\top Z_j}dt_i} \\ &= \frac{e^{\beta^\top Z_i}}{\sum_{j \in R_i} e^{\beta^\top Z_j}}. \end{aligned} \tag{4.7}$$

Dus heeft elk individu waarvan is geobserveerd dat de gebeurtenis heeft plaatsgevonden een bijdrage aan de *gedeeltelijke aannemelijkheid*. Het resultaat van de algemene *gedeeltelijke aannemelijkheid* is dan het product van deze individuele bijdragen:

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{\beta^\top Z_i}}{\sum_{j \in R_i} e^{\beta^\top Z_j}}.$$

4.2 Een discrete versie van het Cox proportional hazards model

In hoofdstuk 2.2 zijn er een aantal relaties uiteengezet tussen de *survival functie* en de *hazard functie* voor het geval van een discrete variabele. Op basis hiervan zal de discrete versie van het *Cox proportional hazards model* beschouwd worden.

De *survival* functie wordt in dit geval weergegeven als,

$$S(t|Z) = S_0(t)^{e^{\beta^\top Z}} \tag{4.8}$$

met de term $S_0(t) = e^{-\int_0^t h_0(u)du}$ als de *baseline survival functie*. Laat verder X een discrete stochastische variabele die de waarden t_i, met $(t_0 < t_1 < ... < t_i < ...)$ aanneemt. In hoofdstuk (2.2) is afgeleid dat in het discrete geval de *survival functie* (in ons geval de *baseline survival functie*) op de volgende manier kan worden uitgedrukt in termen van de *hazard functie*:

$$S_0(t) = \prod_{t_i \leq t} (1 - h(t_i)). \tag{4.9}$$

Nu vervangen we $S_0(t)$ in (4.8) met de uitdrukking van (4.9) en levert dit het volgende resultaat op:

$$S(t|Z) = \prod_{t_i \leq t} (1 - h(t_i))^{e^{\beta^\top Z}}. \tag{4.10}$$

Laat nu de discrete versie van de voorwaardelijke (met covariaten Z als voorwaarde) *hazard functie* $h(t_i|Z) = P(X = t_i \mid X \geq t_i, Z)$, dan is het *Cox proportional hazards model* ook uit te drukken in termen van de *hazard functie* $h(t_i)$:

$$\begin{aligned} 1 - h(t_i|Z) &= P(X > t_i \mid X \geq t_i, Z) && (= P(X > t_i \mid Z)) \\ &= \frac{S(t_i|Z)}{S(t_{i-1}|Z)} && \text{(zie (2.10))} \\ &= \frac{\prod_{t_j \leq t_i}(1 - h(t_j))^{e^{\beta^\top Z}}}{\prod_{t_j \leq t_{i-1}}(1 - h(t_j))^{e^{\beta^\top Z}}} \\ &= (1 - h(t_i))^{e^{\beta^\top Z}}. \end{aligned} \tag{4.11}$$

Met (4.11) komen we dus tot het resultaat:

$$h(t_i|Z) = 1 - (1 - h(t_i))^{e^{\beta^\top Z}}. \tag{4.12}$$

Laat nu

$$dH(t|Z) = H([t + dt]^-|Z) - H(t^-|Z) = P(X \in [t, t + dt) \mid Z)$$

voor dt willekeurig klein, dan kan het model in (4.12) ook geschreven worden als:

$$dH(t|Z) = 1 - (1 - dH_0(t))^{e^{\beta^\top Z}}. \tag{4.13}$$

In dit hoofdstuk is de aannemelijkheidsfunctie uitgewerkt en is de manier waarop het model is opgebouwd uiteengezet. Op grond van dit fundament kunnen verdere vraagstukken worden behandeld. Een soort vraagstuk dat in het volgende hoofdstuk zal worden toegelicht, is het verwerken van onbetrouwbare observaties in dit soort modellen.

5 Imperfecte data

Dit hoofdstuk is gebaseerd op eigen afleidingen van functies en voorbeelden.
Tot nu toe hebben we alleen data beschouwd met de aanname dat we deze data volledig kunnen vertrouwen. Op het moment dat een patiënt uit het ziekenhuis wordt ontslagen, nemen we dit ook aan als waarheid. Data waarvan we niet weten of iets zeker heeft plaatsgevonden op een bepaald tijdstip noemen we imperfecte data. De aanwezigheid van dit soort data moet worden verwerkt bij het modelleren. Het voornaamste is dat de aannemelijkheidsfunctie wordt geconstrueerd in deze context, zodat we ook met imperfecte data de coëfficiënten kunnen schatten. In dit stuk zal een specifiek geval van dit soort data worden besproken, waarbij we resultaten van imperfecte testen zullen belichten. Het hoofdstuk is opgebouwd uit twee delen. Het eerste deel behandelt het model zonder covariaten en het tweede deel neemt het model met één covariaat in beschouwing.

5.1 Model zonder covariaten met perfecte testen

We bekijken opnieuw de studie die plaatsvindt in een ziekenhuis (benoemd in de introductie en 4.1). Deze studie wordt nu aangepast door m patiënten te volgen waarvan zeker is dat ze drager zijn van een bepaalde bacterie. Elk individu dat aanwezig is in het begin van het onderzoek is dus drager van de bacterie, bovendien is een individu dat wellicht in een later stadium van het onderzoek het ziekenhuis binnenkomt ook met zekerheid drager van de bacterie. Naarmate het onderzoek vordert, wordt er regelmatig getest op de bacterie.

Eerst zullen we de aannemelijkheidsfunctie beschrijven in termen van *survival functies* op basis van volledig betrouwbare testen en zonder rekening te houden met enige covariaten. Deze aannemelijkheidsfunctie zullen we per individu opstellen. Dit noemen we dan de individuele aannemelijkheidsfunctie of de individuele bijdrage aan de aannemelijkheidsfunctie in zijn geheel. Voordat we verder gaan met het opstellen van deze aannemelijkheid moeten er een aantal assumpties worden gemaakt. Elk individu heeft dezelfde tijdstippen waarop er wordt getest. Dus elke patiënt heeft n test-momenten $t_0, t_1, ..., t_n$. Daarnaast nemen we aan dat zodra de patiënt de bacterie heeft verloren het niet mogelijk is dat de patiënt weer drager wordt. In het geval dat de test volledig betrouwbaar is, geeft een negatieve testuitslag aan dat de patiënt de bacterie heeft verloren. De kans op het verliezen van de bacterie tussen tijdstip t_{i-1} en t_i wordt dan opgevat als de kans dat de patiënt tot en met tijdstip t_{i-1} drager is van de bacterie, maar tussen t_{i-1} en t_i de bacterie verliest:

$$S(t_{i-1})(1 - \frac{S(t_i)}{S(t_{i-1})}).$$

Merk op dat dit inhoudt dat er op tijdstip t_{i-1} een positief testresultaat werd geconstateerd en op tijdstip t_i een negatieve uitslag. Met het volgende voorbeeld wordt duidelijk op welke manier deze individuele aannemelijkheid wordt geconstrueerd.

Voorbeeld 5.1.1. Stel dat individu i op tijdstip t_1 positief test op de bacterie en op t_2 een negatieve uitslag krijgt. De aannemelijkheidsfunctie is dan de som van de kansen dat het individu tussen t_0 en t_1 de bacterie verliest, dat het individu tussen t_1 en t_2 de bacterie verliest en dat het individu de bacterie na t_2 verliest. De bijdrage van individu i aan de aannemelijkheidsfunctie schrijven we dan als L_i en ziet er als volgt uit:

$$\begin{aligned} L_i &= S(t_0)(1 - \frac{S(t_1)}{S(t_0)})0 + S(t_1)(1 - \frac{S(t_2)}{S(t_1)}) + S(t_2)0 \\ &= S(t_1) - S(t_2). \end{aligned} \tag{5.1}$$

Merk op dat $S(t_0) = 1$. Vanwege het feit dat er op tijdstip t_1 een positieve testuitslag was, vermenigvuldigen we de eerste kans met nul. Het is niet mogelijk dat de patiënt de bacterie heeft verloren in het eerste tijdsinterval. De kans dat het individu de bacterie na t_2 heeft verloren wordt ook vermenigvuldigd met nul, omdat op tijdstip t_2 er een negatieve uitslag is geobserveerd. △

Zoals we in het voorbeeld zien is de individuele aannemelijkheid in principe een som van kansen, maar vanwege de betrouwbaarheid van de test blijft er maar een term over op basis van de observaties. Zodra een negatieve test is geobserveerd dan weten we met zekerheid dat de patiënt tussen het tijdstip t_i van de negatieve test

en het tijdstip t_{i-1} van de laatste positieve test de bacterie heeft verloren. In het algemeen kunnen we de individuele aannemelijkheid op de volgende manier opstellen:

$$L_j = \begin{cases} S(t_{i-1})(1 - \frac{S(t_i)}{S(t_{i-1})}), & \text{als op tijdstip } t_i \text{ met } 1 \leq i \leq n \text{ een negatieve testuitslag plaatsvindt} \\ S(t_n), & \text{anders} \end{cases} \tag{5.2}$$

De gezamenlijke aannemelijkheid is dan het product van deze individuele aannemelijkheidsfuncties.

5.2 Imperfecte testen

Stel nu dat de uitslagen van dit soort testen niet volledig betrouwbaar zijn. Dit soort testen noemen we dan imperfecte testen. In het geval dat de patiënt geen drager is, kan de patiënt geen positieve testuitslag krijgen. Indien de patiënt wel een drager is, bestaat er een kans dat de uitslag negatief is. De waarschijnlijkheid dat een patiënt positief is getest gegeven dat de patiënt een drager is van de bacterie noteren we dan als,

$$\phi = P(\text{positieve test} \mid \text{patiënt drager}) \tag{5.3}$$

met $i = 1, 2, ..., m$. Het complement van deze kans geeft dus de waarschijnlijkheid dat de uitslag negatief is, terwijl de patiënt een drager van de bacterie is. Verder nemen we aan dat de test onafhankelijk is van de tijd (en covariaten). De voorwaardelijke kans weergegeven in (5.3) is dus een constante en noemen we de gevoeligheid van de test. Deze gevoeligheid zal weer opgenomen worden in de individuele aannemelijkheidsfunctie die in de volgende delen geformuleerd zal worden.

5.2.1 Model zonder covariaten met imperfecte testen

Eerst zullen we het model voor imperfecte testen beschouwen zonder rekening te houden met covariaten. Met het feit dat de testen imperfect zijn, moet dus de gevoeligheid van deze testen worden verwerkt bij de kansen op verlies of geen verlies.
Stel dat op tijdstip t_i de patiënt negatief test op de bacterie. Voor deze negatieve testuitslag zijn er dan bij de patiënt in totaal $k \leq i - 1$ positieve uitslagen geobserveerd en zo ook $(i - 1) - k$ negatieve uitslagen. Hoe vaak de kansen ϕ en $1 - \phi$ voorkomen in de waarschijnlijkheid op het verliezen van de bacterie tussen t_{i-1} en t_i hangt dan af van het aantal positieve (en zo ook negatieve) testen:

$$S(t_{i-1})(1 - \frac{S(t_i)}{S(t_{i-1})})\phi^k(1 - \phi)^{((i-1)-k)}.$$

Er volgt dus een wijziging van de aannemelijkheid. Deze wijziging wordt in het onderstaande voorbeeld in context geplaatst en vervolgens in een algemene vorm van deze aannemelijkheidsfunctie verwerkt.

Voorbeeld 5.2.1. Stel dat individu i op tijdstippen t_1, t_2 en t_3 wordt getest. De uitkomsten van de testen zijn: positief, negatief en negatief. In dit geval is de bijdrage van dit individu aan de aannemelijkheidsfunctie te beschrijven door de volgende termen te sommeren:

- De kans dat het individu de bacterie heeft verloren tussen t_0 en t_1 is gelijk aan nul. De testuitslag op tijdstip t_1 is positief en dit houdt in dat het onmogelijk is dat het individu de bacterie heeft verloren.

- De kans dat het individu de bacterie heeft verloren tussen t_1 en t_2 is gelijk aan: $S(t_1)(1 - \frac{S(t_2)}{S(t_1)})\phi$. Hierbij wordt de kans dat het individu nog steeds drager is vermenigvuldigd met ϕ vanwege de positieve test.

- De kans dat het individu de bacterie heeft verloren tussen t_2 en t_3 is gelijk aan: $S(t_2)(1 - \frac{S(t_3)}{S(t_2)})\phi(1 - \phi)$. Hierbij wordt er vermenigvuldigd met ϕ voor de positieve test. Voor de negatieve test wordt er vermenigvuldigd met $(1 - \phi)$, wat inhoudt dat de tweede test een vals negatief resultaat moet zijn en dus deze kans opgenomen moet worden.

- De kans dat het individu de bacterie heeft verloren na t_3 is gelijk aan: $S(t_3)\phi(1-\phi)^2$. In dit geval moeten er dus twee incorrecte negatieve testen hebben plaatsgevonden en die kans is gelijk aan $(1-\phi)^2$.

De bijdrage noteren we dan als:

$$\begin{aligned} L_i &= S(t_1)(1-\frac{S(t_2)}{S(t_1)})\phi + S(t_2)(1-\frac{S(t_3)}{S(t_2)})\phi(1-\phi) + S(t_3)\phi(1-\phi)^2 \\ &= S(t_1)\phi - S(t_2)\phi^2 - S(t_3)\phi^2 + S(t_3)\phi^3. \end{aligned} \tag{5.4}$$

Merk op dat in het geval dat de test volledig betrouwbaar is ($\phi = 1$) de individuele aannemelijkheid hetzelfde wordt als in 5.1. △

Voor elk tijdstip waarop er getest wordt, houden we bij of de uitslag van de test negatief of positief is. Deze indicatorfunctie noteren we voor $k = 1, 2, ..., n$ als:

$$\alpha_k = \begin{cases} 1, & \text{als op tijdstip } t_k \text{ het resultaat positief is} \\ 0, & \text{anders} \end{cases}$$

In het bovenstaande voorbeeld zien we dat voor de aannemelijkheidsfunctie het van belang is om bij te houden hoeveel positieve uitslagen zijn voorgekomen. Dit doen we door de indicatorfuncties te sommeren tot en met het tijdstip dat geëvalueerd wordt,

$$A_i = \sum_{k=1}^{i} \alpha_k$$

hierbij laten we verder $A_0 = 0$. Deze som geeft het aantal positieve testuitslagen aan tot en met tijdstip t_i en noteren we als A_i. Om het aantal negatieve testuitslagen tot en met tijdstip t_i te bepalen kunnen we simpelweg het aantal positieve uitslagen van het aantal tijdstippen in totaal afhalen $i - A_i$. Naast het bijhouden van het aantal negatieve of positieve testuitslagen is het ook van belang om te weten op welk tijdstip de laatste positieve test plaatsvond, want de kans op het verliezen van de bacterie tussen twee tijdstippen is nul als er na het tweede tijdstip nog een positief testresultaat is geobserveerd.
De uitslagen en tijdstippen van de testen zijn geobserveerd en hiervan verzamelen we de tijdstippen van de positieve uitslagen. Het tijdstip van de laatste positieve test is dan het grootste element van deze 'positieve' tijdstippen en noteren we dan als t_N met N als volgt gedefinieerd:

$$N = max\{i : \text{test op } t_i \text{ is positief}\}.$$

Met deze aantallen wordt de individuele aannemelijkheid voor patiënt j op de volgende manier geformuleerd:

$$L_j = \sum_{i>N} S(t_{i-1})(1-\frac{S(t_i)}{S(t_{i-1})})\phi^{A_{i-1}}(1-\phi)^{((i-1)-A_{i-1})} + S(t_n)\phi^{A_n}(1-\phi)^{(n-A_n)}. \tag{5.5}$$

5.2.2 Model met één covariaat en imperfecte testen

Stel dat we nu één covariaat Z_1 in beschouwing nemen. Deze covariaat kan bijvoorbeeld het geslacht of de leeftijd van de patiënt voorstellen. Hoe wordt in dit geval dan de aannemelijkheidsfunctie gedefinieerd? Deze functie hangt nu naast de *survival functie* en de gevoeligheid ϕ ook af van de covariaat Z_1. Voordat we deze functie gaan construeren, behandelen we eerst de gevolgen van het betrekken van een covariaat.

Eerst moet de *baseline survival functie* bepaald worden. Stel dat we alleen de covariaat leeftijd beschouwen, dan bepalen we eerst voor een specifieke groep binnen de studie de *survival functie*. We kunnen dit bijvoorbeeld doen voor de groep die bestaat uit 18-jarige patiënten. De *survival functie* voor deze specifieke groep noteren we dan als $S_0(\cdot)$ en wordt onze *baseline survival functie*. Zoals we hebben gezien in hoofdstuk 4 wordt, in het geval dat we covariaten opnemen in het model, de *baseline hazard functie* vermenigvuldigd met een *link functie*. Voor het *Cox proportional hazards model* wordt deze *link functie* geschreven als $e^{\beta_1 Z_1}$, waarbij Z_1 de bekende covariaat is en β_1 een onbekende parameter. Nu willen we bijvoorbeeld de *survival functie*

voor 80-jarige patiënten bepalen. Met de kennis van hoofdstuk 4 kan deze *survival functie* voor patiënten van 80 als volgt worden beschreven:

$$S(t|Z_1 = 80) = S_0(t)^{e^{\beta_1 80}}.$$

Het doel is vervolgens om de parameter β_1 te schatten. Dit kan gedaan worden door de opgestelde functie te maximaliseren. Zoals in het vorige deel zal eerst een soortgelijk voorbeeld worden gegeven waarbij de data van de groep van 18-jarige patiënten gebruikt wordt om de *baseline survival functie* te bepalen om zo de *survival functie* voor de groep van 80-jarige patiënten te benaderen.

Voorbeeld 5.2.2. Stel dat een 80-jarig individu i op tijdstippen t_1, t_2 en t_3 wordt getest. De uitkomsten van de testen zijn: positief, negatief en negatief. In dit geval is de bijdrage van dit individu aan de aannemelijkheidsfunctie te beschrijven door de volgende termen te sommeren:

- De kans dat het individu de bacterie heeft verloren tussen t_0 en t_1 is gelijk aan nul.
- De kans dat het individu de bacterie heeft verloren tussen t_1 en t_2 is gelijk aan:

$$S_0(t_1)^{e^{\beta_1 80}}\phi(1 - \frac{S_0(t_2)^{e^{\beta_1 80}}}{S_0(t_1)^{e^{\beta_1 80}}}).$$

- De kans dat het individu de bacterie heeft verloren tussen t_2 en t_3 is gelijk aan:

$$S_0(t_2)^{e^{\beta_1 80}}\phi(1 - \frac{S_0(t_3)^{e^{\beta_1 80}}}{S_0(t_2)^{e^{\beta_1 80}}})(1-\phi).$$

- De kans dat het individu de bacterie heeft verloren na t_3 is gelijk aan:

$$S_0(t_3)^{e^{\beta_1 80}}\phi(1-\phi)^2.$$

De aannemelijkheid is dan de som van deze termen. De parameters kunnen geschat worden door de volgende functie te maximaliseren:

$$\begin{aligned} L_i &= S_0(t_1)^{e^{\beta_1 80}}\phi(1 - \frac{S_0(t_2)^{e^{\beta_1 80}}}{S_0(t_1)^{e^{\beta_1 80}}}) + S_0(t_2)^{e^{\beta_1 80}}\phi(1 - \frac{S_0(t_3)^{e^{\beta_1 80}}}{S_0(t_2)^{e^{\beta_1 80}}})(1-\phi) \\ &\quad + S_0(t_3)^{e^{\beta_1 80}}\phi(1-\phi)^2 \\ &= S_0(t_1)^{e^{\beta_1 80}}\phi - S_0(t_2)^{e^{\beta_1 80}}\phi^2 - S_0(t_3)^{e^{\beta_1 80}}\phi^2 + S_0(t_3)^{e^{\beta_1 80}}\phi^3. \end{aligned} \tag{5.6}$$

△

De aannemelijkheidsfunctie voor individu j kunnen we ook weer in een algemene vorm uitschrijven. Hiervoor gebruiken we de *link functie* en dezelfde functies die zijn gedefinieerd voor het opstellen van deze aannemelijkheid zonder covariaten. Deze algemene functie L_j ziet er dan zo uit:

$$L_j = \sum_{i>N} S_0(t_{i-1})^{e^{\beta_1 Z_1}}(1 - \frac{S_0(t_i)^{e^{\beta_1 Z_1}}}{S_0(t_{i-1})^{e^{\beta_1 Z_1}}})\phi^{A_{i-1}}(1-\phi)^{((i-1)-A_{i-1})} + S_0(t_n)^{e^{\beta_1 Z_1}}\phi^{A_n}(1-\phi)^{(n-A_n)}. \tag{5.7}$$

De aannemelijkheidsfunctie is dan in principe een functie van $n+3-N$ parameters. De parameters bestaan uit de $(n+1)-N$ *survival waarden*, gevoeligheid van de test ϕ en de coëfficiënt β_1 die hoort bij de covariaat. In het geval dat elke patiënt op dezelfde tijdstippen getest wordt, dan kunnen we voor deze tijdstippen de *baseline survival functie* bepalen. Met deze waarden van de *baseline functie*, bekende gevoeligheid van de test en coëfficiënt wordt de aannemelijkheidsfunctie opgesteld en kan deze gemaximaliseerd worden om de MLE van de coëfficiënt β_1 te vinden. Hiermee is er in dit hoofdstuk een procedure gegeven voor het modelleren van imperfecte testen.

6 Discussie en conclusie

Survival data heeft een andere structuur dan de data waar we normaal gesproken gewend zijn om mee te werken. Dit heeft mede te maken met de aanwezigheid van *censored data*, hierdoor zijn de standaard statistische modellen niet geschikt voor het analyseren van dit soort data. Belangrijke aspecten van deze data worden beschreven met de *hazard-* en *survival functie*, waarbij elke functie een andere hoeveelheid weergeeft. De *survival functie* stelt de kans voor op het ervaren van de gebeurtenis en de *hazard functie* geeft een hoeveelheid weer die het risico op het ervaren van de gebeurtenis voorstelt. Er zijn dus twee verschillende manieren om *survival data* te specificeren. Een belangrijke assumptie hierbij is dat er alleen sprake is van *rechter censoring* binnen de dataverzameling. De keuze hiervoor is gemaakt op basis van de praktische toepassing van de thesis. In de praktijk komt dit type *censoring* het meest voor. Andere typen censoring die kort zijn genoemd, worden minder gebruikt. In het geval dat de andere soorten *censoring* van toepassing zijn, zou er op een andere manier de data moeten worden verwerkt.

Het schatten van deze functies wordt gerealiseerd vanuit de benadering van telprocessen. De Nelson-Aalen schatter benadert de *cumulatieve hazard functie* en de Kaplan-Meier schatter benadert de *survival functie*. Deze schatters zijn gebaseerd op het aantal gebeurtenissen die op een tijdstip plaatsvinden en het aantal individuen die op dit tijdstip risico lopen. Het zijn vrijwel zuivere schatters, maar in deze thesis is dat niet uitgebreid behandeld. Daarentegen is de manier waarop vanuit telprocessen de data verwerkt kan worden wel toegelicht.

De resultaten van onderzoeken waar de data uit voortkomt zijn vaak afhankelijk van covariaten, in dit geval is er uitgegaan van tijdsinvariante covariaten. We zijn in de praktijk vaak geïnteresseerd in het effect van de covariaten op deze resultaten. In algemene zin willen we in dit geval met regressie analyse de invloed van covariaten achterhalen. Het meest voorkomende model voor deze analyse is het *Cox proportional hazards model*. Het model is uitgewerkt voor zowel het continue als het discrete geval. Vervolgens is de totstandkoming van de gedeeltelijke aannemelijkheid toegelicht met een voorbeeld. Een opmerking hierop is dat voor het formuleren van de aannemelijkheid van het *Cox proportional hazards model* er vanuit is gegaan dat gebeurtenissen op unieke tijdstippen plaatsvinden. In de praktijk komt het voor dat de data deze eigenschap niet heeft en daarvoor is er een aanpassing nodig aan de aannemelijkheidsfunctie. Voor deze aanpassing wordt de lezer verwezen naar de twee bronnen die zijn opgenomen in de referenties.

Een groot deel van de literatuur over *survival analyse* houdt geen rekening met problemen als imperfecte data. In deze thesis wordt er een handvat geboden voor het modelleren van dit probleem voor het specifieke geval van imperfecte testen. In deze thesis zijn fout positieve testuitslagen uitgesloten, omdat dit soort resultaten zeldzaam voorkomen. In het geval dat we dit wel willen verwerken bij de analyses zou de waarschijnlijkheid op fout positieve testuitslagen ook kunnen worden behandeld op een soortgelijke manier door dit op te nemen in de aannemelijkheidsfunctie. Het zou dan mogelijk kunnen zijn dat de patiënt de bacterie verliest voor een positief resultaat. Dus zou de functie voor n test-momenten uit $n + 3$ parameters bestaan.

Merk op dat er geen experimenten zijn gedaan, maar dat de benodigde constructie van het model wel is gegeven. Hiervoor is de aannemelijkheidsfunctie uitgedrukt in termen van de *survival functie* met extra parameters als de gevoeligheid van de test en covariaten. De aannemelijkheidsfunctie is afgeleid op dezelfde manier als dit gedaan zou worden voor betrouwbare testen door tussen twee opeenvolgende tijdstippen de kans op het plaatsvinden van de gebeurtenis te evalueren. Deze functie is op individueel niveau gedefinieerd, voor de algemene aannemelijkheidsfunctie zou het product van deze individuele bijdragen moeten worden genomen.

Voor experimenten zou dan voor een vast aantal test-momenten op willekeurige wijze de testuitslagen kunnen worden gegenereerd van individuen gegroepeerd op basis van de covariaat. Door één groep als controlegroep aan te duiden en hiervan de *survival functie* te bepalen, hebben we een functie die als *baseline* genomen kan worden. Met software naar eigen keuzen kan dan de aannemelijkheidsfunctie voor een gegeven covariaat en test-gevoeligheid worden gemaximaliseerd. Verder is er van één covariaat uitgegaan binnen het model, maar kan er op soortgelijke wijze ook hetzelfde model worden uitgebreid naar een groter aantal covariaten. In dit geval zijn er meer vrijheidsgraden waardoor het numeriek complexer wordt.

A Appendix

Aannemelijkheidsfunctie = Likelihood function
Cumulatieve verdelingsfunctie = Cumulative distribution function
Gebeurtenis = Death/Event
Gedeeltelijke aannemelijkheid = Partial likelihood
Gemiddelde tijdsduur = Mean survival time
Gemiddelde begrensde tijdsduur = Restricted mean survival time
Intensiteitsproces = Intensity process
Kansdischtheid = Probability density function
Martingaal = Martingale
Meest aannemelijke schatter = Maximum likelihood estimator (MLE)
Risico- proces/verzameling = Risk set
Steekproef = Sample
Stochastische variabele = Random variable
Telproces = Counting process
Tijdstip = Time
Variantie = Variance
Verwachting = Expected value
Voorwaardelijke kans = Conditional probability
Zuiver = Unbiased

Referenties

[1] David R Cox. „Partial likelihood". In: *Biometrika* 62.2 (1975), p. 269–276.

[2] Rikke Nørmark Mortensen. „Pseudo-observations in survival analysis". Master thesis, Aalborg University, 2013.