Overlevingstafels en longitudinale analyse

Survival-analyse / Duurmodellen

10

Thaya Carolina, Léander Kuijvenhoven en Jan van der Laan

Statistische Methoden (10010)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2010

Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader voorlopig cijfer
x	=	geheim
–	=	nihil
–	=	(indien voorkomend tussen twee getallen) tot en met
0 (0,0)	=	het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	=	een cijfer kan op logische gronden niet voorkomen
2008–2009	=	2008 tot en met 2009
2008/2009	=	het gemiddelde over de jaren 2008 tot en met 2009
2008/'09	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2008 en eindigend in 2009
2006/'07–2008/'09	=	oogstjaar, boekjaar enz., 2006/'07 tot en met 2008/'09

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312
2492 JP Den Haag

Prepress
Centraal Bureau voor de Statistiek - Grafimedia

Omslag
TelDesign, Rotterdam

Inlichtingen
Tel. (088) 570 70 70
Fax (070) 337 59 94
Via contactformulier: www.cbs.nl/infoservice

Bestellingen
E-mail: verkoop@cbs.nl
Fax (045) 570 62 68

Internet
www.cbs.nl

ISSN: 1876-0333

6016510010 X-37

Inhoudsopgave

1. Inleiding op het deelthema

1.1 Algemene beschrijving en leeswijzer

1.1.1 Beschrijving van het deelthema

Dit deelthema beschrijft methoden waarmee duren kunnen worden geanalyseerd. Een duur kan gezien worden als de tijd tussen twee gebeurtenissen: geboorte/sterfte, ontslag/vinden nieuwe baan, huwelijk/scheiding, etc. Enerzijds worden methoden besproken die een beschrijving geven van de duurverdeling. Anderzijds worden methoden besproken die proberen de duurverdeling te modelleren en dan vooral de invloed van achtergrondkenmerken op de duur. Hiermee kunnen vragen als: 'Lopen bedrijven uit klasse A een grotere kans op faillissement dan bedrijven uit klasse B?', 'Zitten vrouwen langer in de WW dan mannen?', beantwoord worden.

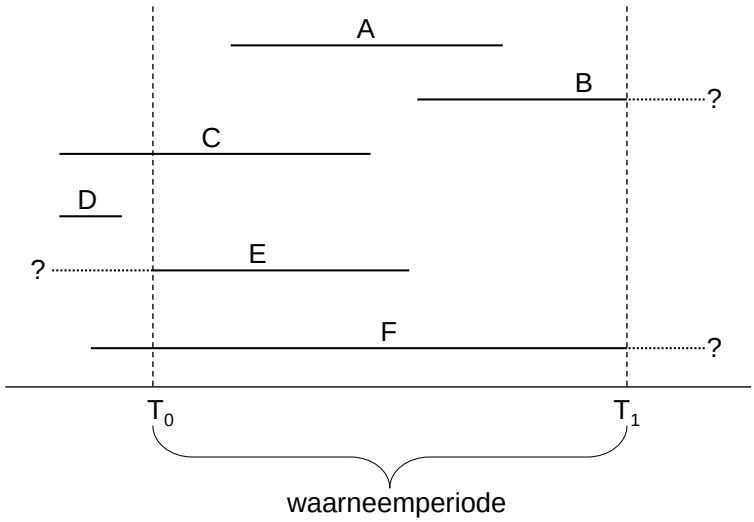
Het grote probleem bij het analyseren van duren zit hem voornamelijk in het feit dat bijna altijd de duur voor een gedeelte van de populatie nog niet is afgerond. Van deze objecten is de duur niet bekend. Er is alleen bekend dat de duur langer is dan een bepaalde waarde. Aangezien dit meestal een selectieve groep is, moeten methodes die gebruikt worden voor het analyseren van duren deze gedeeltelijke informatie op één of andere manier verwerken in hun schattingen. De hier besproken methoden doen dit.

1.1.2 Problemen en oplossingen

Wat maakt omgaan met duren anders dan het omgaan met bijvoorbeeld het inkomen van personen? Men zou toch gewoon de gemiddelde duur kunnen uitrekenen. Twee zaken maken de situatie iets ingewikkelder. Ten eerste, heeft men bijna altijd te maken met censurering en truncatie, en ten tweede, compliceert het tijdsaspect, dat per definitie een rol speelt bij duren, het bepalen van de populatie. Hieronder worden beide punten behandeld.

1.1.2.1 Censurering en truncatie

In het algemeen zal de periode waarin men duren waarneemt beperkt zijn. Zo is aan de ene kant de waarnemingsperiode in ieder geval beperkt tot het 'nu'. Van duren die op dit moment nog niet zijn afgesloten is de eindtijd en daarmee de lengte van de duur niet bekend. Aan de andere kant wordt de waarneming in het algemeen ook beperkt, bijvoorbeeld doordat het register dat gebruikt wordt niet verder terug gaat. Men heeft dus te maken met duren die nog niet afgesloten zijn. Het probleem dat daarbij optreedt is dat het in het algemeen de langere duren zijn die nog niet zijn afgesloten. Deze niet-afgesloten duren kunnen dus niet buiten de analyse gehouden worden.



Figuur 1.1.1. Mogelijke waarnemingen aan duren. Op de horizontale as loopt de tijd; de lijnen geven duren aan.

Figuur 1.1.1 toont de verschillende mogelijkheden die kunnen optreden. De lijnen geven duren aan. De waarneemperiode loopt van T_0 tot T_1. Duren zoals A worden volledig waargenomen. Bij B wordt de duur rechts gecensureerd: het einde van de duur wordt niet waargenomen; er is alleen bekend dat de duur langer is dan een bepaalde waarde. Rechtse censurering komt in de praktijk bijna altijd voor. Stel men is in 2010 geïnteresseerd in de levensduur van personen geboren in 1970. Van de meeste personen is in 2010 de duur nog niet afgelopen en is dus alleen bekend dat de duur langer is dan 39 jaar. Als men bij het bepalen van het gemiddelde negeert dat er duren gecensureerd zijn, zal het gemiddelde nooit meer dan 39 kunnen zijn, wat duidelijk een onderschatting is van de 'werkelijke' levensduur.

Naast rechte censurering kan er natuurlijk ook linkse censurering optreden. Dit de situatie E. In dat geval is alleen bekend dat de duur bij het begin van de waarneemperiode al begonnen is, maar men weet niet hoelang de duur al bezig was aan het begin van de waarneemperiode. Dit is een van de vervelendste gevallen, omdat zonder heel erg sterke aannames te doen er geen goede methoden zijn die hiermee kunnen omgaan.

Gelukkig weet men in het algemeen wel hoelang de duur al gaande was. Zo weet men bijvoorbeeld wel de geboortedatum van personen die geboren zijn voordat het GBA beschikbaar kwam. Deze situatie wordt geschetst door C. Dit wordt (rechtse) truncatie genoemd. Ook immigratie is hier een voorbeeld van: deze personen hebben ook al een levensduur voordat ze in de populatie komen. Op zich is de duur dus bekend van C en men zou dit dus gelijk kunnen stellen aan A. Echter, er zijn ook gevallen als D, die ook getrunceerd zijn, maar die geheel niet worden waargenomen. Getrunceerde duren die wel worden waargenomen zullen in het algemeen langer zijn dan getrunceerde duren die niet worden waargenomen. Men moet dus een of andere correctie toepassen om te voorkomen dat de lengte van de duren wordt overschat.

Er zijn natuurlijk ook allerlei combinaties van linkse en rechte censurering en truncatie mogelijk, zoals F.

Doordat censurering en truncatie in de praktijk bijna altijd voorkomen, moeten bij het maken van statistieken van duren altijd methodes gebruikt worden die op één of andere manier hiermee kunnen omgaan.

1.1.2.2 Populaties in de tijd

Doordat het proces dat men probeert te beschrijven zich in de tijd afspeelt, heeft ook de selectie in de tijd van de populatie die men wil beschrijven een sterke invloed op de uitkomsten.

Een klein voorbeeldje. Stel in een ver land met een veel eenvoudiger rechtssysteem krijgen veroordeelden of 1 jaar gevangenisstraf of 10 jaar en lopen straffen altijd van 1 januari tot 31 december. Ieder jaar worden 1000 mensen veroordeeld voor 1 jaar en 1000 mensen tot 10 jaar. Wat is de gemiddelde gevangenisstraf? Intuïtief zal men $(1000\times1+1000\times10)/(1000+1000)=5,5$ jaar zeggen. Echter, als men de gevangenispopulatie op een bepaalde datum zou onderzoeken, komt men tot een geheel ander resultaat. Op 2 januari zitten er 1000 personen met een straf van 1 jaar. Verder zitten 1000 personen die afgelopen 1 januari zijn veroordeeld tot een straf van 10 jaar, 1000 personen die tot 10 jaar zijn veroordeeld 1 januari van het voorgaande jaar, etc. Op 2 januari zitten er in totaal dus 10×1000 personen die zijn veroordeeld tot 10 jaar in de gevangenis. De personen die op 2 januari in de gevangenis zitten zijn dus gemiddeld tot $(10\ 000\times10+1000\times1)/(10\ 000+1000) = 9,2$ jaar veroordeeld.

In het eerste geval bestaat de populatie uit personen waarvan de duur op een bepaald moment begint: een startcohort. In het tweede geval bestaat de populatie uit personen die op een bepaald moment in een duursituatie zitten: populatie op peilmoment. In het algemeen zal een startcohort meer overeenkomen met wat men intuïtief verwacht dan een populatie op peilmoment. De duurverdeling horende bij een startcohort geeft aan welke duur men kan verwachten als men ook in een duursituatie terecht komt.

Een derde mogelijkheid naast startcohort en populatie op peilmoment, is door te kijken naar duren die binnen een bepaalde periode vallen. Men kijkt dan niet naar de totale duur, maar alleen naar het stuk duur dat binnen de periode valt, of eigenlijk naar de kans op bijvoorbeeld overlijden binnen de periode gegeven de leeftijd. Een voorbeeld hiervan zijn periodelevenstafels welke in de demografie veel gebruikt worden.

1.1.2.3 Leeswijzer

Voor het beschrijven van duurverdelingen zijn een aantal andere maten/functies beschikbaar dan de standaard verdelingsfunctie, gemiddelde, mediaan, etc. Deze worden besproken in hoofdstuk 1. Deze maten zijn ook nodig om de methoden besproken in de daarop volgende hoofdstukken te kunnen begrijpen.

De besproken methoden kunnen in de eerste plaats onderscheiden worden door de aannames die gedaan worden over de verdelingsfunctie. De niet-parametrische methoden maken geen aannames over de verdelingsfunctie. Dit heeft als voordeel dat de verdelingsfunctie beschreven wordt zoals hij wordt waargenomen, maar deze methoden zijn minder toepasbaar als niet heel de verdelingsfunctie wordt waargenomen en zijn alleen beschrijvend. Het is echter wel mogelijk om te toetsen of verdelingen van elkaar verschillen. Dit wordt besproken in hoofdstuk 6.

De parametrische methoden (parametrisch model voor de hazard, hoofdstuk 7, en het logistische model, hoofdstuk 9) nemen aan dat de verdelingsfunctie een bepaalde verdeling volgt die afhangt van bepaalde achtergrondkenmerken van de objecten. Met de parametrische methoden kan dus de invloed van achtergrondkenmerken op de duur bestudeerd worden. Daarnaast kunnen deze methoden, doordat de duurverdeling helemaal bepaald wordt, gebruikt worden om allerlei kenmerken (zoals gemiddelde) van de verdeling af te leiden.

De semi-parametrische methoden (het Coxmodel, hoofdstuk 8, en ook het logistische model, hoofdstuk 9, kan soms hieronder geschaard worden) modelleren wel het effect dat achtergrondkenmerken hebben op de duurverdeling, maar ze proberen zo weinig mogelijk aannames te doen over de vorm van de duurverdeling. Deze methoden kunnen dus voornamelijk gebruikt worden om de invloed van achtergrondkenmerken op de duurverdeling te bestuderen. Om tot schattingen van bijvoorbeeld gemiddelde of mediane duren te komen moeten deze methoden gecombineerd worden met niet-parametrische schatters zoals de Kaplan-Meierschatter.

In het algemeen kan duur als een continue variabele worden gezien. Zo is levensduur in dagen praktisch gezien een continue variabele. De meeste methoden nemen aan dat duur een continue variabele is. Echter, men kan duur vaak ook zien als een discrete variabele. Vaak komt dit doordat de nauwkeurigheid waarmee de duur gemeten wordt een discretisatie veroorzaakt, zoals levensduur in dagen. In enkele gevallen is de duur werkelijk discreet, zoals het aantal sollicitatiebrieven dat iemand moet versturen voordat hij wordt aangenomen. Overlevingstafels kunnen gebruikt worden om discrete duren te beschrijven. Het logistische model kan gebruikt worden om discrete duren te modelleren.

De eigenschappen van de verschillende methoden zijn samengevat in tabel 1.1.1.

Tabel 1.1.1. Toepassingsmogelijkheden voor de verschillende methoden

Methode	Parametrisch	Continu/discreet
Overlevingstabels (hst.3)	nee	discreet
Kaplan-Meierschatter (hst. 4)	nee	continu
Nelson-Aalenschatter (hst. 5)	nee	continu
Vergelijken duurverdelingen (hst. 6)	nee	continu
Parametrisch model (hst. 7)	ja	continu
Coxmodel (hst. 8)	semi	continu
Logistisch model (hst. 9)	semi/ja	discreet

1.2 Afbakening en relatie met andere thema's

Duurdata is een vorm van longitudinale data. In dit deelthema wordt alleen de analyse en beschrijving van duurdata besproken. In de Methodenreeks is er een apart deelthema Overlevingstafels binnen het thema 'Overlevingstafels en longitudinale analyse'.

Een aantal meer complexe methoden worden hier niet besproken, zoals

- Herhaalde gebeurtenissen. Sommige gebeurtenissen kunnen meerdere malen optreden. Zo kunnen personen meerdere keren werkloos worden. Als men specifiek geïnteresseerd in het herhaald optreden zijn de methoden die hier beschreven worden niet geschikt. Complexere methoden zijn dan nodig. Vaak is het mogelijk om dit 'probleem' 'weg te definiëren' door bijvoorbeeld te kijken hoelang het duurt voordat personen die in 2000 werkloos geworden zijn voor het eerst een baan vinden.
- Competing risks. De methoden besproken in dit deelthema gaan ervan uit dat er slechts één soort gebeurtenis kan optreden. In het voorbeeld van werkloosheid kunnen mensen werk vinden of inactief worden. Als het nodig is om onderscheid tussen beide gebeurtenissen te maken, zijn competing risks modellen nodig.

1.3 Plaats in het statistisch proces

Duuranalyses hebben vooral een plaats in de analysestap van het proces. Daarnaast kunnen de methodes ook dienen als input voor bijvoorbeeld regressie-imputatie (zie thema Imputatie).

1.4 Definities

Zaken als hazard, survivalfunctie etc. worden in het volgende hoofdstuk gedefinieerd.

2. Concepten en parameters van duurverdelingen

2.1 Korte beschrijving

- In dit hoofdstuk worden niet zozeer methoden besproken, maar verschillende parameters van duurverdelingen waarmee duurverdelingen kunnen worden aangeduid, zoals de mediaan en het gemiddelde.
- Tevens worden belangrijke concepten die veelvuldig gebruikt worden in de duuranalyse behandeld. Belangrijke concepten zijn ondermeer de survivalfunctie en de hazard.
- Methodes die in latere hoofdstukken worden besproken maken uitvoerig gebruik van de concepten en parameters die in dit hoofdstuk zullen worden besproken.

2.2 Toepasbaarheid

In dit hoofdstuk worden verschillende populatieparameters waarmee duurverdelingen kunnen worden aangeduid besproken. Belangrijke populatieparameters in de survivalanalyse zijn het gemiddelde en de mediaan.

Ook worden er concepten besproken die gangbaar zijn in de duuranalyse, maar minder gangbaar zijn buiten de duuranalyse. Men denkt aan de survivalfunctie, de hazardfunctie en de cumulatieve hazardfunctie.

2.3 Uitgebreide beschrijving

2.3.1 Continue verdelingen

Een erg belangrijk begrip in de duuranalyse is de survivalfunctie. Stel $f(t)$ is de kansdichtheidsfunctie van de stochast T. Merkt op dat $t \geq 0$ omdat duren niet negatief kunnen zijn. De cumulatieve verdelingsfunctie $F(t)$ is dan gedefinieerd als

$$F(t) = P(T \leq t) = \int_0^t f(x)\mathrm{d}x \,. \tag{2.3.1}$$

De survivalfunctie wordt nu als volgt gedefinieerd

$$S(t) = P(T \geq t) = \int_t^\infty f(x)\mathrm{d}x \,. \tag{2.3.2}$$

Voor de survivalfunctie geldt dat $S(0) = 1$ en $S(\infty) = \lim_{t\to\infty} S(t) = 0$. Een praktische interpretatie van de survivalfunctie is dat zij de fractie personen aangeeft die op tijdstip *t* nog geen gebeurtenis hebben ondergaan. Stel dat men begint met 100 personen en dat na een zeker tijdstip *t* 25 mensen uitgestroomd zijn, dan moeten

er dus nog 75 mensen uitstromen. De survivalfunctie is nu gelijk aan 0.75, dus *S(t)*=0.75.

Een ander belangrijk concept in de duuranalyse is de hazardfunctie $h(t)$. De hazardfunctie is gedefinieerd als

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.3.3}$$

De hazardfunctie kan gezien worden als het momentane tempo waarmee een gebeurtenis optreedt op tijdstip t, gegeven dat de gebeurtenis niet opgetreden is voor tijdstip t.

De functies $f(t), S(t),$ en $h(t)$ kunnen onderling in elkaar over worden geschreven. De volgende vergelijking geldt bijvoorbeeld

$$S(t) = \exp(-\int_0^t h(x)\mathrm{d}x). \tag{2.3.4}$$

Het is nu handig om de cumulative hazardfunctie te definiëren

$$H(t) = \int_0^t h(x)\mathrm{d}x. \tag{2.3.5}$$

Het is nu eenvoudig in te zien dat

$$S(t) = \exp(-H(t)). \tag{2.3.6}$$

Tenslotte kan $f(t)$ uitgedrukt worden in $h(t)$ op de volgende wijze

$$f(t) = h(t)\exp(-\int_0^t h(x)\mathrm{d}x). \tag{2.3.7}$$

Deze vergelijkingen lijken misschien op het eerste gezicht abstract, maar veel methoden die nog in de volgende hoofdstukken aan de orde komen maken uitvoerig gebruik van deze vergelijkingen.

2.3.2 Discrete modellen

Als duren bijvoorbeeld afgerond, gegroepeerd of uitgedrukt zijn in het aantal keer dat een bepaalde cyclus is opgetreden kan men gebruik maken van discrete modellen. Stel dat de duur T de waardes $t_1, t_2, \ldots$ kan aannemen met $0 \le t_1 < t_2 < \cdots$. Stel verder dat de kansfunctie gelijk is aan

$$f(t_j) = P(T = t_j) \quad j = 1, 2, \ldots. \tag{2.3.8}$$

De survivalfunctie is dan gelijk aan

$$S(t) = P(T \geq t) = \sum_{j:t_j \geq t} f(t_j). \tag{2.3.9}$$

De hazardfunctie is het discrete geval is

$$h(t_j) = P(T = t_j \mid T \geq t_j) = \frac{f(t_j)}{S(t_j)} \quad j = 1, 2, \ldots. \tag{2.3.10}$$

Net als in het continue geval kunnen $f(t), S(t),$ en $h(t)$ in elkaar worden overgeschreven. Dit betekent dus ook dat het theoretische gezien niet uitmaakt welke van de functies beschikbaar is om de duurverdeling eenduidig vast te leggen. Verder geldt in het discrete geval dat $f(t_j) = S(t_j) - S(t_{j+1})$ en dit impliceert dat

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)} \quad j = 1, 2, \ldots. \tag{2.3.11}$$

Bovendien impliceert dit

$$S(t) = \prod_{j:t_j < t} (1 - h(t_j)). \tag{2.3.12}$$

De cumulatieve hazardfunctie is simpelweg gelijk aan de som van de hazards

$$H(t) = \sum_{j:t_j < t} h(t_j). \tag{2.3.13}$$

Een andere definitie van de cumulatieve hazardfunctie die gangbaar is

$$H(t) = -\log S(t). \tag{2.3.14}$$

Deze vergelijking maakt gebruikt van vergelijking (2.3.12).

2.4 Parameters van duurverdelingen

Verdelingen worden gekenmerkt door verschillende parameters. Enkele bekende parameters zijn het gemiddelde en de mediaan. Bij duuranalyses wordt vaak vanuit de survivalfuncties gedacht. Daarmee wordt bedoeld dat de definitie van de parameters vaak afgeleid is met behulp van de survivalfunctie. Hieronder worden verschillende voorbeelden hiervan gegeven.

Het *p*de kwantiel is gedefinieerd als de kleinste waarde van *t* waarvoor *S(t)* kleiner of gelijk is aan *p*.

$$x_p = \inf\{t : S(t) \leq p\}. \tag{2.4.1}$$

Een belangrijk kwantiel dat in de praktijk veel wordt gebruikt is de mediaan. De mediaan is gedefinieerd als de kleinste waarde van *t* waarvoor *S(t)* kleiner of gelijk is aan 0.5.

$$\xi_{0.5} = \inf\{t : S(t) \le 0.5\}. \tag{2.4.2}$$

Dit is dus de duur waarop *S(t)* van een waarde groter dan 0.5 naar een waarde kleiner dan 0.5 ‘springt’. Andere veel voorkomende kwantielen zijn 0.25 en 0.75.

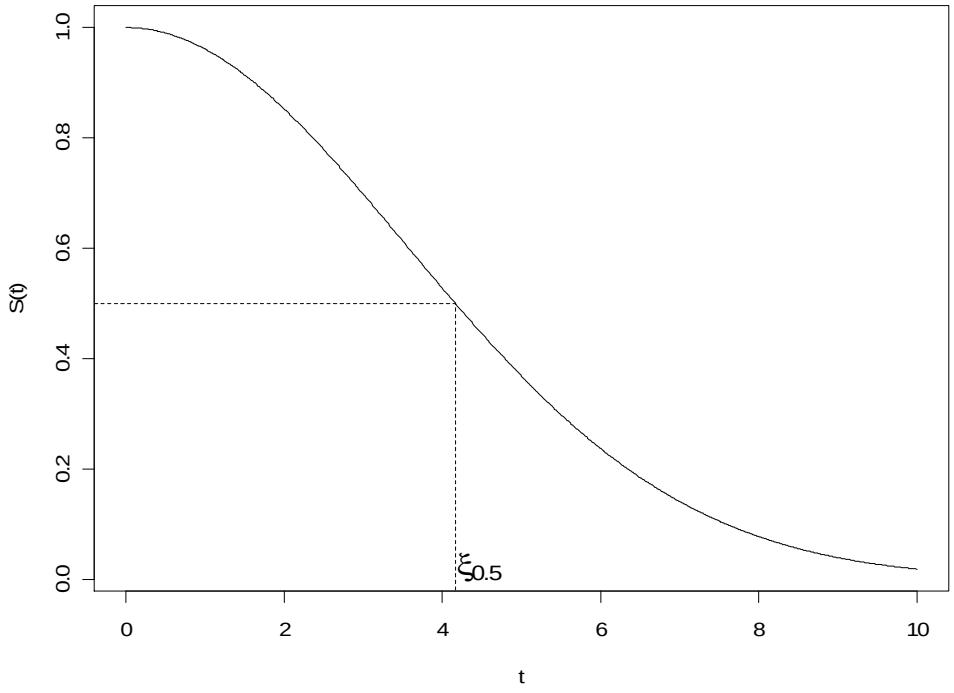
Het gemiddelde is gelijk aan de oppervlakte onder de survivalfunctie

$$\mu = \int_0^\infty S(t)\mathrm{d}t\,. \tag{2.4.3}$$

Bij discrete verdelingen gaat de integraal over in een som.

2.5 Voorbeeld

Hieronder ziet men een typische survivalfunctie afgebeeld. Het is duidelijk te zien dat de functie begint bij 1 en eindigt bij 0. In de grafiek is met een stippellijn de mediaan $\xi_{0.5}$ aangeven. De oppervlakte onder de grafiek is gelijk aan het gemiddelde.



3. Overlevingstafels

3.1 Korte beschrijving

De overlevingstafel – ook sterftetafel genoemd – is een statistische beschrijving van het sterftepatroon in een bevolking (over de jaren heen) en geeft het aantal overlevenden, de cumulatieve hazardfunctie en de hazard rates van de onderliggende bevolking weer. In de demografie wordt de hazard gewoonlijk gedefinieerd als de kans dat iemand uitstroomt tussen x en $x+n$, gegeven dat hij de duur x gehaald heeft.

$$h_x = \Pr(x \le T < x+n \mid T \ge x). \tag{3.1.1}$$

Deze kans, weergeven in formule (3.1.1), kan geschat worden door gebruik te maken van het uitstroomtempo. Het uitstroomtempo m_x is gedefinieerd als het aantal personen dat uitstroomt met een duur tussen x en $x+n$ gedurende de periode (D_x) te delen door het aantal persoonperioden dat risico heeft gelopen binnen deze duurcategorie en binnen de periode (L_x),

$$m_x = \frac{D_x}{L_x}. \tag{3.1.2}$$

Het aantal personen dat aan het begin van het duurinterval dat loopt van x tot $x+n$ risico loopt om uit te stromen in het interval wordt gegeven door l_x. Het verwachte aantal personen dat in het interval uitstroomt, wordt dus gegeven door

$$D_x = l_x h_x, \tag{3.1.3}$$

Stel dat de personen die uitstromen tijdens het interval dit gemiddeld na $a_x n$ doen. Het aantal persoonperioden in het betreffende interval wordt dan gegeven door

$$L_x = (l_x - D_x)n + a_x n D_x = l_x\{1-(1-a_x)h_x\}n. \tag{3.1.4}$$

Door vergelijking (3.1.2), (3.1.3) en (3.1.4) te combineren en op te lossen voor h_x, wordt de relatie tussen de hazard en het uitstroomtempo gevonden:

$$h_x = \frac{nm_x}{1+(1-a_x)nm_x}. \tag{3.1.5}$$

Als a_x gelijk is aan een half, i.e. als wordt aangenomen dat de personen die uitstromen tijdens het interval dit gemiddeld halverwege het interval doen, wordt de actuariële schatter voor de hazard verkregen. Deze hazard kan omgerekend worden naar de survivalfunctie: $S(t) = \prod_{x:x<t}(1-h_x)$.

3.2 Toepasbaarheid

De overlevingstafel is één van de oudste en meest gebruikte methoden om levensduren te beschrijven (zie bijvoorbeeld Berkson en Gage, 1950; Cutler en Ederer, 1958; Gehan, 1969). Het is in wezen een aangepaste en uitgebreide versie van de frequentietabel en kan goed toegepast worden in geval van gecensureerde data. De duurverdeling wordt verdeeld in een bepaald aantal intervallen. Voor elk interval kan vervolgens het aantal of de fractie individuen bepaald worden die aan het begin van het interval in leven zijn, de fractie waarvan de duur in het interval beëindigd wordt (terminal events, sterfte) en het aantal gecensureerde gevallen.

3.3 Uitgebreide beschrijving

3.3.1 Theoretische achtergrond

In het algemeen wordt er een onderscheid gemaakt tussen verschillende soorten overlevingstafels. Deze overlevingstafels worden onderscheiden op basis van de samenstelling van de onderliggende populatie (bevolkingstafels vs. ervaringstafels) en de manier waarop deze populatie wordt waargenomen (generatietafels vs. periodetafels). Hieronder worden een aantal overlevingstafels besproken (zie ook Wolthuis en Bruning, 1996).

Generatietafels (Cohort-tafels)

Generatietafels of cohorttafels beschrijven de sterfte in een bepaalde geboortecohort. Hiervoor wordt het sterfteverloop van een groep personen vanaf geboorte, van jaar tot jaar gevolgd totdat de laatste persoon van deze groep is overleden. Op grond van de waargenomen sterftequotiënten kan een sterftetafel worden geconstrueerd. Deze aanpak leidt ertoe dat er bij ieder geboortejaar een aparte sterftetafel hoort. Omdat de sterftetafel pas geconstrueerd kan worden op het moment dat de laatste persoon van de desbetreffende generatie of geboortecohort is overleden, kunnen de gegevens van de verkregen sterftetafel sterk verouderd zijn.

Periodetafels

Periodetafels worden geconstrueerd op basis van een synthetische geboortecohort, door waarnemingen te verzamelen voor een bepaald kalenderjaar of meerdere kalenderjaren samen. Voor ieder kalenderjaar uit de waarnemingsperiode worden eenjarige sterftequotiënten $\hat{q}_x$ waargenomen voor alle leeftijden. Het sterftequotiënt $\hat{q}_x$ voor de gehele waarnemingsperiode wordt dan vaak gelijk gesteld aan het rekenkundige gemiddelde van de sterftequotiënten voor de afzonderlijke jaren. Op

basis van deze (gemiddelde) sterftequotiënten kan een sterftetafel worden geconstrueerd. Vaak worden de sterftequotiënten eerst nog afgerond om statistische afwijkingen te elimineren (“mortality graduation”). Op basis van de samenstelling van de waargenomen groep of populatie kan voor de periodetafels een onderscheid worden gemaakt in ervaringstafels en bevolkingstafels.

Ervaringstafels

Een ervaringstafel wordt samengesteld aan de hand van het sterfteverloop onder de verzekerden van een verzekeringsmaatschappij en wijkt vaak af van het sterfteverloop van de bevolking. Het doel hierbij is om autoselectie van de kant van de verzekerden in kaart te brengen. In plaats van één sterftequotiënt dat voor eenieder gelijk is, wordt hier verondersteld dat sprake is van een zekere variatie in sterftequotiënten per leeftijd. In de praktijk blijkt namelijk dat mensen, ook al hebben ze een gelijke leeftijd, toch van elkaar verschillen wat betreft sterfterisico. Dit verschil wordt o.a. veroorzaakt door erfelijke aanleg, leefwijze en woon- en werkomgeving van de persoon.

Bevolkingstafels

In het algemeen geldt dat deze sterftetafels worden geconstrueerd met sterftecijfers van een land of van een gedeelte van een land, waarvan de sterfte gedurende een aantal jaren is waargenomen. Hierbij wordt traditioneel een onderscheid gemaakt tussen mannen en vrouwen. Voor elk jaar wordt voor iedere leeftijd afzonderlijk een sterftequotiënt berekend. Bij het berekenen van deze sterftequotiënten wordt er rekening mee gehouden dat in de loop van het jaar de bevolkingssamenstelling verandert door immigratie en emigratie. Dit is belangrijk omdat alleen de mensen die in het land wonen, worden opgenomen in deze sterftetafels. Voor iedere leeftijd wordt het gemiddelde genomen van de bijbehorende waargenomen sterftequotiënten in de periode, waarmee vervolgens de bevolkingstafel wordt geconstrueerd.

3.3.2 Standaard overlevingstafel analyse

In de praktijk zijn er verschillende methoden om de bevolkingstafels op te stellen. In deze paragraaf wordt de standaardmethode, gebaseerd op cohorttafel, beschreven. Een ‘cohort’ wordt gedefinieerd als een groep individuen die een willekeurige steekproef vormt van een populatie. Zoals eerder vermeld kan de overlevingstafel worden gezien als een uitbreiding van de relatieve frequentietabel voor gecensureerde data, niettemin met de overlevingstafel ligt de nadruk op het schatten van de voorwaardelijke of conditionele sterftekans in een interval gegeven overleving aan het begin van het interval, en de overlevingskans tot het eind van het

interval. Voor een uitgebreider uitleg van onderstaande beschrijving, zie Lawless (2003).

Stel, de tijd is verdeeld in $k+1$ intervallen $I_j = [t_{j-1}, t_j)$ voor $j = 1,...,k+1$ en $t_0 = 0$, $t_k = T$ en $t_{k+1} = \infty$, waar T een bovenlimiet is voor de observaties. Voor elke persoon uit een willekeurige steekproef van grootte n, wordt of de levensduur d of de censureringsmoment/tijd L geobserveerd. Echter, voor de gegroepeerde data is alleen bekend in welk interval bepaalde individuen sterven of worden gecensureerd en niet de exacte levensduren en censureringstijden. De data bestaat dan uit het aantal levensduren en censureringstijden die in elk interval I_j voorkomen. Voor het laatste interval I_{k+1} wordt verondersteld dat het alleen levensduren bevat, aangezien alle overlevenden op tijdstip T in het interval I_{k+1} komen te overlijden. Dan kunnen de volgende variabelen gedefinieerd worden:

N_j : de omvang van de risicopopulatie op tijdstip t_{j-1} (i.e. aantal personen in leven en niet gecensureerd),

D_j : aantal doden in interval $I_j = [t_{j-1}, t_j)$ (i.e. aantal levensduren binnen desbetreffende interval),

W_j : aantal censureringsgevallen in $I_j = [t_{j-1}, t_j)$ (i.e. “withdrawals”).

Hieruit volgt, $N_1 = n$ en $N_j = N_{j-1} - D_{j-1} - W_{j-1}$ voor $j = 2,...,k+1$.

Gegeven de survivalfunctie $S(t)$ voor de duurverdeling, worden de (conditionele) overlevings- en bijbehorende sterftekansen gedefinieerd als:

$P_j = S(t_j)$: kans op overleving voorbij tijdsinterval I_j

$p_j = \dfrac{P_j}{P_{j-1}}$: kans op overleving voorbij I_j, gegeven overleving op I_{j-1}

$q_j = 1 - p_j$: kans op sterfte in I_j, gegeven overleving op I_{j-1}

voor $j = 1,...,k+1$. Verder geldt $P_0 = 1$, $P_{k+1} = 0$ en $q_{k+1} = 1$. Uit bovenstaande volgt dat de onvoorwaardelijke of onconditionele overlevingskans P_j kan worden herschreven als

$$P_j = p_1 p_2 \cdots p_j \quad \text{voor} \quad j = 1,...,k+1. \tag{3.3.1}$$

Dat is, de kans op overleving voorbij interval I_j is gelijk aan het product van de conditionele overlevingskansen tot en met I_j. Dit resultaat vormt de basis voor de overlevingstafelbenadering voor survivalanalyse. Bij de overlevingstafelanalyse worden de parameters q_j en p_j geschat en vervolgens wordt met behulp van

vergelijking (3.3.1) P_j geschat. De resultaten worden dan getoond in de vorm van een tabel, die de oorspronkelijke data en de schattingen $\hat{q}_j$ en $\hat{P}_j$ weergeeft. In het algemeen bevatten de kolommen dus, voor elk interval I_j, de waarden voor N_j, D_j, W_j, $\hat{q}_j$ en $\hat{P}_j$. Hieronder worden de twee gevallen besproken voor het berekenen van $\hat{q}_j$, namelijk bij afwezigheid van censurering (i.e. $W_j = 0$) en in het geval van censurering ($W_j > 0$).

3.3.2.1 Schatting q_j als $W_j = 0$

Als een bepaald interval I_j geen censureringsgevallen bevat, wordt q_j geschat als

$$\hat{q}_j = \frac{D_j}{N_j}. \tag{3.3.2}$$

Dit volgt direct uit de definitie van q_j. In dit geval is het aantal doden D_j multinomiaal verdeeld met kansverdeling gegeven door

$$\Pr(D_1,...,D_k) = \frac{n!}{D_1!\cdots D_k!D_{k+1}!}\prod_{j=1}^{k+1}\pi_j^{D_j} \tag{3.3.3}$$

met $D_1+...+D_{k+1} = n$ en $\pi_1+...+\pi_{k+1} = 1$.

De parameter π_j geeft de onconditionele kans op sterfte in interval I_j weer en is gelijk aan

$$\pi_j = P_{j-1} - P_j = p_1\cdots p_{j-1}q_j. \tag{3.3.4}$$

Hieruit volgt een likelihoodfunctie die proportioneel is aan

$$L \propto \prod_{j=1}^{k+1}(p_1\cdots p_{j-1}q_j)^{D_j} = \prod_{j=1}^{k+1}(p_j^{N_j-D_j}q_j^{D_j}) \tag{3.3.5}$$

waarbij $N_j = n - D_1 - \cdots - D_{j-1}$.

Deze functie is maximaal voor $\hat{q}_j = 1-\hat{p}_j = \frac{D_j}{N_j}$, zolang $N_j > 0$, $\forall j \in \{1,...,k+1\}$. Als $N_j = 0$ voor een bepaalde j, dan bestaat er geen maximumlikelihood-schatter voor q_j (de bijbehorende parameters q_j en p_j komen niet voor in de likelihoodfunctie) en wordt verondersteld dat $\hat{q}_j = 1$. De maximumlikelihood-schatter voor P_j is gelijk aan:

$$\hat{P}_j = \hat{p}_1\cdots\hat{p}_j = \frac{N_{j+1}}{n} \tag{3.3.6}$$

waarbij $N_{j+1} = N_j - D_j$.

Dit laatste volgt uit de aanname dat $\hat{p}_j = 0$ als $N_j = 0$. N_{j+1} is binomiaal verdeeld, met kansverdeling

$$\Pr(N_{j+1}) = \binom{n}{x} P_j^x (1-P_j)^{n-x}. \tag{3.3.7}$$

De verdeling van $\hat{P}_j$ volgt dan simpelweg uit bovenstaande, met verwachtingswaarde en variantie gelijk aan $E(\hat{P}_j) = P_j$ en $Var(\hat{P}_j) = \frac{P_j(1-P_j)}{n}$, respectievelijk.

3.3.2.2 Schatting q_j als $W_j > 0$

In geval van censurering, zal de schatter $\hat{q}_j = \frac{D_j}{N_j}$ de echte waarde q_j onderschatten, aangezien de gecensureerde personen in I_j zouden kunnen zijn gestorven vóór het einde van het interval als er geen censurering had plaatsgevonden. De meest gebruikte en voor de hand liggende schatting voor q_j is dan

$$\hat{q}_j = \frac{D_j}{N'_j} = \frac{D_j}{N_j - \frac{1}{2}W_j}. \tag{3.3.8}$$

Vergelijking (3.3.8) vindt men door aan te nemen dat gecensureerde personen gemiddeld gedurende de helft van een interval behoren tot de risicopopulatie. Met andere woorden wordt er aangenomen dat de censureringen uniform verdeeld zijn over het interval.

3.3.3 Alternatieve overlevingstafelanalyse

Een traditionele methode om overlevingstafels te construeren naast de eerder beschreven standaardmethode, staat bekend als de methode van Chiang (1960a, b, 1968, 1984). Bij de methode van Chiang is kennis van de censureringstijden van de duur van alle waargenomen individuen, inclusief sterfte vereist.

De sterftetafel volgens de methode van Chiang wordt aan de hand van een aantal variabelen opgebouwd, die elk een kolom van de tafel weergeven. In het vervolg wordt elk van deze variabelen van de overlevingstafel gedefinieerd en zal de relatie t.o.v. de overige grootheden worden verklaard. Een uitgebreide beschrijving van de toepassing van deze methode op het CBS is te vinden in Van der Meulen (2009).

De eerste kolom van een overlevingstafel bevat de leeftijd x. De hoogst voorkomende leeftijd wordt aangeduid met ω. Alle leeftijden x zijn als een interval

gedefinieerd bestaande uit twee opeenvolgende getallen, namelijk $[x, x+n)$ behalve ω. De hoogst voorkomende leeftijd ω is een open interval. In de meeste gevallen wordt ω gelijk gesteld aan 100 jaar. De leeftijdspecifieke sterftequotiënten ${}_n\hat{q}_x$ worden in de tweede kolom weergegeven. Het geschatte sterftequotiënt kan worden bepaald volgens vergelijking (3.3.9)

$$ {}_n\hat{q}_x = \frac{{}_n d_x}{l_x} = \frac{n_n m_x}{1+(1-{}_n a_x) n \cdot {}_n m_x} \quad \text{voor} \quad x = 0,1,\ldots,\omega. \tag{3.3.9} $$

Hierbij is ${}_n d_x$ gelijk aan het aantal personen dat komt te overlijden op het leeftijdsinterval $[x, x+n)$ en l_x is het aantal overlevenden op leeftijd x. De variabele ${}_n m_x$ is de sterfte-intensiteit op leeftijd x en ${}_n a_x$ is de fractie die geleefd werd door de personen die in het leeftijdsinterval $[x, x+n)$ zijn komen te overlijden. Deze twee grootheden zijn in kolom vijf en zes opgenomen en worden in het vervolg uitgelegd. Het sterftequotiënt ${}_n\hat{q}_x$ geeft de *conditionele* kans weer op sterfte in interval $[x, x+n)$, gegeven overleving aan het begin van het interval. Als het sterftequotiënt van leeftijd x bekend is, kan hieruit het overlevingsquotiënt voor leeftijd x worden afgeleid

$$ {}_n\hat{p}_x = 1 - {}_n\hat{q}_x \quad \text{voor } x = 0,1,\ldots,\omega. \tag{3.3.10} $$

De derde kolom geeft per leeftijd het aantal levenden l_x op leeftijd x aan. De sterftetafels zijn in het algemeen niet gebaseerd op waarnemingen van een werkelijke groep nuljarigen, maar op een fictieve groep nuljarigen. Deze fictieve groep van nuljarigen ($=l_0$) wordt ook wel de radix van de sterftetafel genoemd. De radix van een sterftetafel wordt willekeurig gekozen en kan dus verschillen van andere sterftetafels. Meestal is de radix een veelvoud van een macht van tien, bijvoorbeeld 100000. De fictieve groep nuljarigen wordt vervolgens gevolgd tot en met de laatste levende. Het aantal waargenomen levenden van leeftijd x vermenigvuldigd met het overlevingsquotiënt van die leeftijd levert het aantal personen dat één jaar later nog in leven is, ofwel

$$ l_{x+n} = {}_n\hat{p}_x \cdot l_x \quad \text{voor} \quad x = 0,1,\ldots,\omega-1. \tag{3.3.11} $$

Hiermee kan het aantal personen ${}_n d_x$ dat komt te overlijden op het leeftijdsinterval $[x, x+n)$ berekend worden volgens

$$ {}_n d_x = {}_n\hat{q}_x \cdot l_x \quad \text{voor} \quad x = 0,1,\ldots,\omega-1. \tag{3.3.12} $$

Dit kan ook in termen van het overlevingsquotiënt worden uitgedrukt. Als het sterftequotiënt in bovenstaande vergelijking wordt vervangen door het overlevingsquotiënt dan leidt dit tot de volgende relatie

$$ {}_n d_x = l_x(1-{}_n\hat{p}_x) = l_x - {}_n\hat{p}_x \cdot l_x = l_x - l_{x+n}. \tag{3.3.13} $$

Hieruit volgt dat het aantal overlijdensgevallen in de groep x-jarigen gelijk is aan het verschil tussen het aantal levenden van leeftijd x en het aantal levenden n jaren later. De variabelen l_x en ${}_n d_x$ zijn afhankelijk van de radix en komen dan ook niet overeen met de werkelijke geobserveerde data. Ze geven enkel het aantal levenden en het aantal overledenen voor de sterftetafel met radix l_0 weer. In de vijfde kolom staat de fractie ${}_n \mathrm{a}_x$. Dit is de fractie van het leeftijdsinterval $[x, x+n)$ die de overleden personen ${}_n d_x$ gemiddeld hebben geleefd. Alle ${}_n d_x$ personen zijn tijdens het interval $[x, x+n)$ komen te overlijden en hebben dus x complete jaren geleefd plus een fractie van het het interval $[x, x+n)$. Voor iedere leeftijd wordt het gemiddelde van deze fracties weergegeven door ${}_n \mathrm{a}_x$.

De sterfte-intensiteit ${}_n m_x$ wordt bepaald door het quotiënt te nemen van de waargenomen sterfte ${}_n d_x$ binnen het leeftijdsinterval $[x, x+n)$ en het aantal waargenomen levensjaren ${}_n L_x$ in de cohort in het interval. Dit leidt tot onderstaande vergelijking.

$$ {}_n m_x = \frac{{}_n d_x}{{}_n L_x} \quad \text{voor} \quad x = 0,1,\ldots,\omega. \tag{3.3.14} $$

Het aantal geleefde jaren ${}_n L_x$ binnen het leeftijdsinterval $[x, x+n)$ is in kolom zeven opgenomen. Elk persoon die aan het eind van het interval $[x, x+n)$ nog in leven is, levert een bijdrage van n jaar aan ${}_n L_x$, terwijl elk persoon die komt te overlijden binnen dit interval gemiddeld een fractie ${}_n \mathrm{a}_x$ bijdraagt. Hieruit volgt voor $x = 0,1,\ldots,\omega - 1$:

$$ {}_n L_x = n \cdot (l_x - {}_n d_x) + n \cdot {}_n \mathrm{a}_x \cdot {}_n d_x \tag{3.3.15} $$

De eerste term aan de rechterkant van de bovenstaande vergelijking geeft het totale aantal jaren dat wordt geleefd door de levenden, namelijk $n \cdot (l_x - {}_n d_x)$. De tweede term geeft het aantal jaar weer dat werd geleefd door de ${}_n d_x$ overledenen. De hoogstvoorkomende leeftijd ω in de sterftetafel is een open interval.Voor dit interval wordt ${}_\infty L_x$ berekend op basis van de sterfte-intensiteit voor de personen vanaf leeftijd ω. De berekening van ${}_\infty L_x$ geschiedt volgens:

$$ {}_\infty L_x = \frac{l_\omega}{{}_\infty m_x}. \tag{3.3.16} $$

Voor de hoogst voorkomende leeftijd ω in de sterftetafels is het aantal geleefde jaren binnen de leeftijdsinterval gelijk aan het quotiënt van het aantal levenden op

leeftijd ω en de sterfte-intensiteit op leeftijd ω. Met L_x kan het totale aantal jaren T_x, geleefd door de personen vanaf leeftijd x, in kolom acht weergegeven. Deze grootheid is belangrijk voor de berekening van de levensverwachting. T_x is gelijk aan de som van het aantal geleefde jaren in elke leeftijdsinterval vanaf leeftijd x:

$$T_x = L_x + L_{x+1} + \ldots + \mathrm{L}_\omega \quad \text{voor} \quad x = 0,1,\ldots,\omega. \tag{3.3.17}$$

Uitgedrukt in termen van T_{x+1}:

$$T_x = L_x + T_{x+1} \quad \text{voor} \quad x = 0,1,\ldots,\omega\text{-}1. \tag{3.3.18}$$

In de laatste kolom wordt de *resterende* levensverwachting $\hat{e}_x$ voor leeftijd x vermeld. Dit getal geeft het gemiddeld aantal jaren weer dat een persoon met leeftijd x zal leven. De levensverwachting wordt als volgt berekend:

$$\hat{e}_x = \frac{T_x}{l_x} \quad \text{voor} \quad x = 0,1,\ldots,\omega. \tag{3.3.19}$$

Als de leeftijd x hierbij wordt opgeteld levert dit de levensverwachting vanaf geboorte op.

3.4 Voorbeeld

Hieronder is een overlevingstafel te zien zoals deze op het CBS wordt gepubliceerd. In deze tabel zijn niet alle kolommen volgens de methode van Chiang vermeld. De eerste kolom geeft leeftijd (x) weer. In de tweede kolom is de sterftekans ($_n\hat{q}_x$) opgenomen. De derde en vierde kolommen vermelden respectievelijk het aantal levenden (l_x) en overledenen ($_n d_x$). Tenslotte geeft de laatste kolom de resterende levensverwachting weer ($\hat{e}_x$).

Overlevingstafels; geslacht en leeftijd

Vrouwen
2000

Leeftijd (jaar)	Sterftekans	Levenden (aantal)	Overledenen (aantal)	Levensverwachting (jaar)
0	0.00418	100000	418	80.58
0,5	0.00076	99582	76	80.42
1,5	0.0003	99506	30	79.48
2,5	0.00023	99476	23	78.5
3,5	0.00013	99453	13	77.52
4,5	0.0002	99440	20	76.53
5,5	0.00006	99420	6	75.55
6,5	0.00009	99414	9	74.55
7,5	0.00013	99405	13	73.56
8,5	0.00013	99392	13	72.57
.	.	.	.	.
.	.	.	.	.
96,5	0.30894	4539	1402	2.45
97,5	0.30977	3137	972	2.33
98,5	0.378	2165	818	2.15

3.5 Kwaliteitsindicatoren

Schattingen voor de variantie van de standaardschatters $\hat{q}_j$, $\hat{p}_j$ en $\hat{P}_j$ worden gegeven door de formules die zijn afgeleid door Greenwood (1926). In dit geval geldt:

$$\hat{var}(\hat{p}_j) = \frac{\hat{p}_j \hat{q}_j}{N'_j} \tag{3.5.1}$$

en

$$\hat{var}(\hat{q}_j) = \frac{\hat{q}_j - \hat{q}_j^2}{N'_j} \tag{3.5.2}$$

Hieruit volgt dat

$$\hat{var}(\hat{P}_j) = \hat{P}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i - \hat{q}_i^2}{(1-\hat{q}_i)N'_i} = \hat{P}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{N'_i \hat{p}_i}. \tag{3.5.3}$$

In het speciale geval waarbij $W_j = 0$, kan bovenstaande gesimplificeerd worden tot

$$\hat{var}(\hat{P}_j) = \frac{\hat{P}_j(1-\hat{P}_j)}{n} \tag{3.5.4}$$

Betrouwbaarheidsintervallen voor $\hat{P}_j = \hat{S}(a_j)$ volgen dan direct uit de standaard theorie voor de binomiale verdeling. De betrouwbaarheidsintervallen voor de geschatte sterftequotiënten volgens de methode van Chiang zijn symmetrisch en gebaseerd op de aanname dat leeftijdsspecifieke mortaliteit binomiaal verdeeld is (Chiang, 1984). De variantie van $\hat{q}_x$ is dan asymptotisch normaal verdeeld en kan geschat worden met

$$\hat{\text{var}}(\hat{q}_x) = \frac{\hat{q}_x^2(1-\hat{q}_x)}{D_x}. \tag{3.5.5}$$

Het $100(1-\alpha)\%$ betrouwbaarheidsinterval zou dus geschat kunnen worden met

$$\hat{q}_x \pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}(\hat{q}_x)}, \tag{3.5.6}$$

waarbij $z_{1-\alpha/2}$ is het $1-\alpha/2$ de percentiel van de standaard normale verdeling.

4. Kaplan-Meier-schatter voor de survivalfunctie

4.1 Korte beschrijving

De Kaplan-Meier-schatter geeft een niet-parametrische schatting van de survivalfunctie. Deze kan vervolgens gebruikt worden om bijvoorbeeld de mediane of gemiddelde duur uit te rekenen.

4.2 Toepasbaarheid

De Kaplan-Meier-schatter is te gebruiken voor het schatten van de survivalfunctie voor continue duren. Schattingen van bijvoorbeeld de gemiddelde of mediane duur kunnen gebaseerd worden op de Kaplan-Meier-schatter.

Aangezien de Kaplan-Meier-schatter niet parametrisch is, worden er geen aannames gedaan over het verloop van de duurverdeling. Het nadeel hiervan is dat het beperkt mogelijk is om de invloed van bepaalde achtergrondkenmerken op de duur te bestuderen. Het is natuurlijk wel mogelijk om voor verschillende groepen de survivalfunctie te schatten en deze met elkaar te vergelijken. Dit is het onderwerp van hoofdstuk 6. De schatter is dus vooral beschrijvend te gebruiken.

De Kaplan-Meier-schatter is gebaseerd op de aanname van niet-informatieve censurering, wat inhoudt dat kennis van het censureringstijdstip geen extra informatie geeft over de overlevingskansen op een tijdstip na censurering aannemend dat de persoon niet gecensureerd is. Aan deze aanname wordt bijvoorbeeld niet voldaan als personen die in een slechte gezondheidstoestand verkeren vaker worden gecensureerd dan personen die gezond zijn. Als niet aan de voorwaarden is voldaan zijn de schatters niet zuiver.

4.3 Uitgebreide beschrijving

Stel we hebben n individuen en er zijn k $(k \leq n)$ verschillende tijdstippen $t_1 < t_2 < \ldots < t_k$ waarop gebeurtenissen plaatsvinden. Er kunnen meerdere gebeurtenissen plaatsvinden op een tijdstip t_j; d_j is het aantal gebeurtenissen dat plaatsvindt op t_j. De Kaplan-Meier-schatter is gedefinieerd als:

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j}, \tag{4.3.1}$$

waarbij n_j het aantal personen is dat risico loopt op t_j. Dit is het aantal personen dat een duur heeft langer of gelijk aan t_j en nog niet gecensureerd is voor t_j. Censurering wordt dus verwerkt in n_j: tot het censureringsmoment telt een persoon mee in n_j daarna niet meer. Op het moment van censurering daalt n_j.

Als een persoon gecensureerd is op een tijdstip t_j en er vindt ook een gebeurtenis plaats op tijdstip t_j dan wordt aangenomen dat de censurering net iets na t_j

plaatsvindt. De persoon wordt dan dus tot de risicogroep gerekend op t_j. In feite wordt dus aangenomen dat de duur van de persoon langer is dan t_j, wat wel zeer waarschijnlijk is voor iemand die op t_j wordt gecensureerd. Als de langste waargenomen duur gecensureerd is, dan is de schatter gedefinieerd tot dit censureringsmoment.

De schatter kan gezien worden als gebaseerd op vergelijking (2.3.12), waar de survivalfunctie ook wordt gegeven als een product van één min de hazard. Hier wordt de hazard geschat met d_j/n_j, het aantal gebeurtenissen gedeeld door het aantal personen dat een gebeurtenis zou kunnen ondergaan.

De Kaplan-Meier-schatter is onder vrij algemene voorwaarden een consistente schatter van de survivalfunctie (de schatting nadert de werkelijke survivalfunctie als $n \to \infty$).

Door gebruik te maken van vergelijking (2.3.6) kan de schatter ook gebruikt worden om een schatting te verkrijgen van de cumulatieve hazard:

$$\hat{H}_{KM}(t) = -\log\left(\hat{S}(t)\right). \tag{4.3.2}$$

Het subscript 'KM' wordt gebruikt om een onderscheid te maken met de in het volgende hoofdstuk besproken Nelson-Aalen-schatter voor de cumulatieve hazard. Het is echter niet goed mogelijk om de Kaplan-Meier-schatter te gebruiken voor een schatting van de hazard. Het aantal gebeurtenissen d_j op een tijdstip t_j is vaak klein, waardoor de schatting nogal sterk fluctueert. Sommige pakketten (waaronder STATA) kunnen wel een soort 'gladgestreken' schatting geven die alleen gebruikt kan worden om kwalitatief naar de data te kijken.

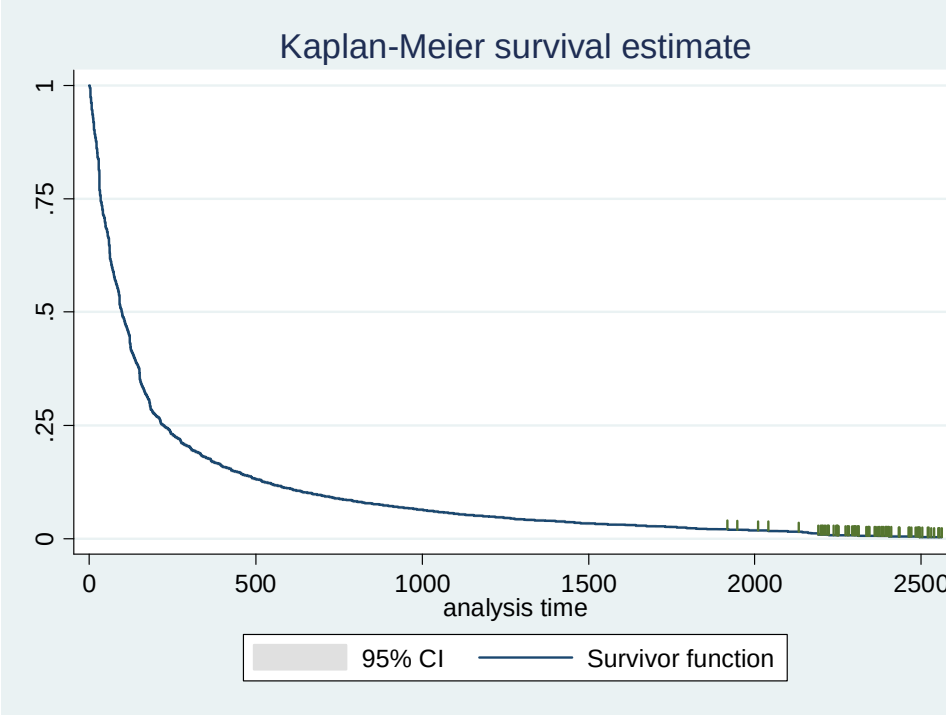
De mediaan is gedefinieerd als de kleinste waarde van *t* waarvoor *S(t)* kleiner of gelijk is dan 0.5 is en kan dus makkelijk uit de Kaplan-Meier-schatter geschat worden.

$$\hat{\xi}_{0.5} = \inf\left\{t : \hat{S}(t) \le 0.5\right\}. \tag{4.3.3}$$

Dit is dus de duur waarop $\hat{S}(t)$ van een waarde groter dan 0.5 naar een waarde kleiner dan 0.5 'springt'. Het gemiddelde wordt geschat door de oppervlakte onder de geschatte survivalfunctie.

$$\hat{\mu} = \int_0^T \hat{S}(t)\mathrm{d}t\,, \tag{4.3.4}$$

waarin T de langste duur is die is waargenomen. Als de langste duur een gebeurtenis betreft, dan bereikt $\hat{S}(t)$ daar nul en is $\hat{\mu}$ een zuivere schatter. Als de langste duur een censurering betreft dan bereikt $\hat{S}(t)$ nul niet en zal $\hat{\mu}$ een onderschatting geven van het werkelijke gemiddelde.



Figuur 4.4.1 Kaplan-Meier-schatter van de survivalfunctie voor werkloosheids-uitkeringsduren in dagen. Het betrouwbaarheidsinterval is in grijs weergegeven (door de grote hoeveelheid data is het betrouwbaarheidsinterval dermate klein dat het slecht te zien is). De groene streepjes geven censureringen aan.

4.4 Voorbeeld

Figuur 4.4.1 toont de Kaplan-Meier-schatter voor werkloosheidsuitkeringen. Duidelijk is te zien dat het merendeel van de personen een duur heeft kleiner dan een jaar. De mediaan is 97 dagen.

Aangezien dit een grote dataset betreft, lijkt de schatter continue. Echter, de survivalfunctie is, zoals uit vergelijking (4.3.1) blijkt, een (rechtscontinue en linksgelimiteerde) stapfunctie, die iedere keer als er een gebeurtenis optreedt een stap naar beneden gaat.

4.5 Kwaliteitsindicatoren

De variantie van de Kaplan-Meier-schatter wordt gegeven door de formule van Greenwood (Lawless, 1982)

$$\hat{\text{var}}\left\{\hat{S}(t)\right\} = \hat{S}(t)^2 \sum_{j:t_j<t} \frac{d_j}{n_j\left(n_j - d_j\right)}. \qquad (4.5.1)$$

De schatting $\hat{S}(t)$ is asymptotisch normaal verdeeld. Het $100(1-\alpha)\%$ betrouwbaarheidsinterval zou dus geschat kunnen worden met

$$\hat{S}(t) \pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}\left\{\hat{S}(t)\right\}}, \tag{4.5.2}$$

waarin $z_{1-\alpha/2}$ het $1-\alpha/2^{\text{de}}$ percentiel is van de standaard normale verdeling. Echter om te zorgen dat het betrouwbaarheidsinterval binnen het interval [0,1] blijft wordt meestal een transformatie toegepast.

$$\hat{S}(t)^{\exp\left(\pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}\left\{\hat{S}(t)\right\}}/\left(\hat{S}(t)\log\hat{S}(t)\right)\right)}. \tag{4.5.3}$$

Voor het bepalen van een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor $\xi_{0.5}$ wordt gebruik gemaakt van ideeën van Brookmeyer en Crowley (1982). Zij ontwikkelden een betrouwbaarheidsinterval op basis van hypothesetoetsen. Als interval neemt men alle waardes van $\xi_{0.5}^0$ die niet verworpen zullen worden als men de nulhypotheses $\xi_{0.5} = \xi_{0.5}^0$ test tegen de alternatieve hypothese $\xi_{0.5} \neq \xi_{0.5}^0$ met een onbetrouwbaarheidsdrempel van α. Meer formeel bestaat het betrouwbaarheidsinterval uit alle waardes van $\xi_{0.5}^0$ die voldoen aan

$$\frac{\left|g(\hat{S}(\xi_{0.5}^0)) - g(0.5)\right|}{\left|g'(\hat{S}(\xi_{0.5}^0))\right|\hat{S}(\xi_{0.5}^0)\hat{\tau}(\xi_{0.5}^0)} \leq z_{1-\alpha/2}. \tag{4.5.4}$$

waarbij *g* een nader te bepalen functie is, zoals *g(x) = x* of *g(x) = log(-log(x))* wat meestal gebruikt wordt. De ondergrens van het betrouwbaarheidsinterval wordt bepaald door de kleinste waarde voor *t* te kiezen waarvoor geldt dat

$$\hat{S}(t)^{\exp\left(-z_{1-\alpha/2}\sqrt{\hat{\text{var}}\left\{\hat{S}(t)\right\}}/\left(\hat{S}(t)\log\hat{S}(t)\right)\right)} \leq 0.5. \tag{4.5.5}$$

Voor de bovengrens van het betrouwbaarheidsinterval kiest men de kleinste waarde *t* waarvoor geldt dat

$$\hat{S}(t)^{\exp\left(+z_{1-\alpha/2}\sqrt{\hat{\text{var}}\left\{\hat{S}(t)\right\}}/\left(\hat{S}(t)\log\hat{S}(t)\right)\right)} \leq 0.5. \tag{4.5.6}$$

De variantie voor het gemiddelde wordt gegeven door Klein en Moeschberger 2003, paragraaf 4.5,

$$\text{var}(\hat{\mu}) = \sum_{i=1}^{k}\left(\int_{t_i}^{T}\hat{S}(t)\text{d}t\right)^2 \frac{d_i}{n_i(n_i - d_i)}, \tag{4.5.7}$$

waaruit het $100(1-\alpha)\%$ betrouwbaarheidsinterval geschat kan worden door

$$\hat{\mu} \pm z_{1-\alpha/2}\sqrt{\text{var}(\hat{\mu})}. \tag{4.5.8}$$

5. Nelson-Aalen-schatter voor de cumulatieve hazard

5.1 Korte beschrijving

De Nelson-Aalen-schatter geeft een niet-paramatrische schatting van de cumulatieve hazard. Deze kan vervolgens gebruikt worden om bijvoorbeeld de mediane of gemiddelde duur uit te rekenen.

5.2 Toepasbaarheid

De Nelson-Aalen-schatter is te gebruiken voor het schatten van de cumulatieve hazard voor continue duren. Schattingen van bijvoorbeeld de gemiddelde of mediane duur kunnen gebaseerd worden op de Nelson-Aalen-schatter.

Aangezien de Nelson-Aalen-schatter niet parametrisch is, worden er geen aannames gedaan over het verloop van de duurverdeling. Het nadeel hiervan is dat het beperkt mogelijk is om de invloed van bepaalde achtergrondkenmerken op de duur te bestuderen. Het is natuurlijk wel mogelijk om voor verschillende groepen de survivalfunctie te schatten en deze met elkaar te vergelijken. Het vergelijken van survivalfuncties wordt besproken in hoofdstuk 6. De schatter is dus vooral beschrijvend te gebruiken.

De Nelson-Aalen-schatter is gebaseerd op de aanname van niet-informatieve censurering, wat inhoudt dat kennis van het censureringstijdstip geen extra informatie geeft over de overlevingskansen op een tijdstip na censurering aannemend dat de persoon niet gecensureerd is. Aan deze aanname wordt bijvoorbeeld niet voldaan als personen die in een slechte gezondheidstoestand verkeren vaker worden gecensureerd dan personen die gezond zijn. Als niet aan de voorwaarden is voldaan zijn de schatters niet zuiver.

5.3 Uitgebreide beschrijving

Stel we hebben n individuen en er zijn k $(k \leq n)$ verschillende tijdstippen $t_1 < t_2 < \ldots < t_k$ waarop gebeurtenissen plaatsvinden. Er kunnen meerdere gebeurtenissen plaatsvinden op een tijdstip t_j; d_j is het aantal gebeurtenissen dat plaatsvindt op t_j. De Nelson-Aalen-schatter is gedefinieerd als:

$$\hat{H}(t) = \sum_{j:t_t<t} \frac{d_j}{n_j}. \qquad (5.3.1)$$

waarbij n_j het aantal personen is dat risico loopt op t_j. Dit is het aantal personen dat een duur heeft langer of gelijk aan t_j en nog niet gecensureerd is voor t_j. Censurering wordt dus verwerkt in n_j: tot het censureringsmoment telt een persoon mee in n_j daarna niet meer. Op het moment van censurering daalt n_j.

Als een persoon gecensureerd is op een tijdstip t_j en er vindt ook een gebeurtenis plaats op tijdstip t_j dan wordt aangenomen dat de censurering net iets na t_j plaatsvindt. De persoon wordt dan dus tot de risicogroep gerekend op t_j. In feite wordt dus aangenomen dat de duur van de persoon langer is dan t_j, wat wel zeer waarschijnlijk is voor iemand die op t_j wordt gecensureerd. Als de langste waargenomen duur gecensureerd is, dan is de schatter gedefinieerd tot dit censureringsmoment.

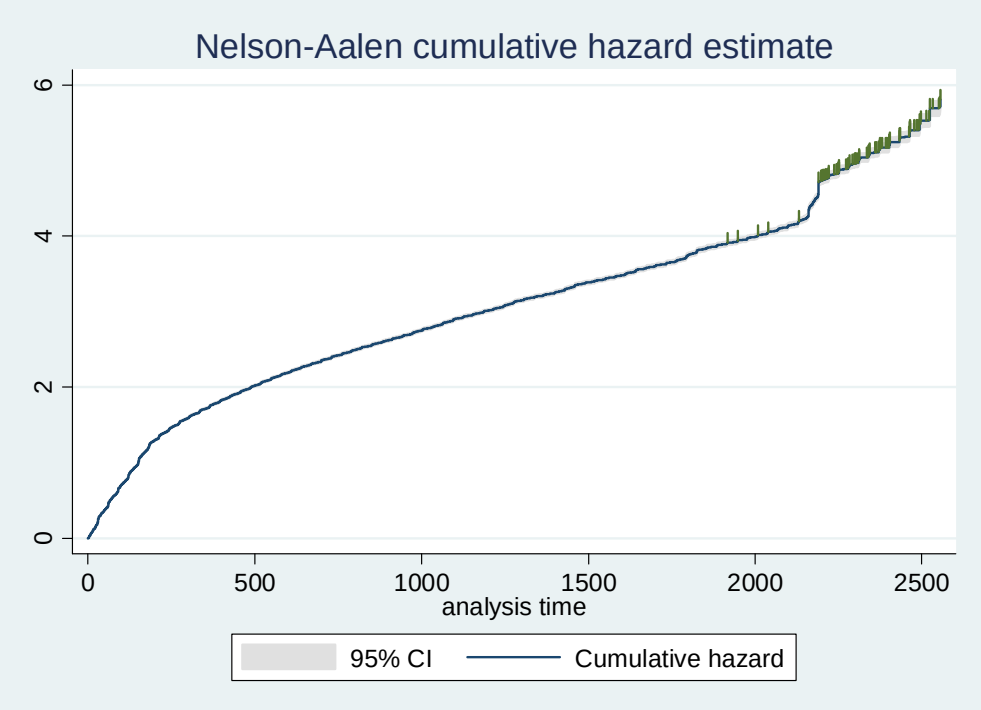
Door gebruik te maken van vergelijking (2.3.6) kan de schatter ook gebruikt worden om een schatting te verkrijgen van de survival functie:

$$\hat{S}_{FH}(t) = \exp\left(-\hat{H}(t)\right). \tag{5.3.2}$$

Deze schatter voor de survivalfunctie wordt ook wel de Flemming-Harrington-schatter genoemd. Vandaar dat het subscript 'FH' gebruikt wordt. Voor grote datasets ($n \to \infty$) is de Flemming-Harrington-schatter gelijk aan de Kaplan-Meier-schatter.

Het verloop van de hazard kan gemakkelijker gezien worden met de cumulatieve hazard dan met de survivalfunctie. Een constante hazard geeft een lineaire cumulatieve hazard; een toenemende hazard een convexe cumulatieve hazard en een afnemende hazard een concave cumulatieve hazard.

5.4 Voorbeeld



Figuur 5.4.1 Nelson-Aalen-schatter van de cumulatieve hazard voor werkloosheids-uitkeringsduren in dagen. Het betrouwbaarheidsinterval is in grijs weergegeven (door de grote hoeveelheid data is het betrouwbaarheidsinterval dermate klein dat het slecht te zien is). De groene streepjes geven censureringen aan.

Figuur 5.4.1 toont de Nelson-Aalen-schatter voor werkloosheidsuitkeringen. De cumulatieve hazard is duidelijk concaaf: naarmate personen langer een werkloosheidsuitkering hebben neemt de kans om uit de werkloosheidsuitkering te stromen af. De kans op uitstromen neemt na 6 jaar (ongeveer 2200 dagen) sterk toe. Dit wordt veroorzaakt doordat de lengte van de duur wettelijk beperkt is.

5.5 Eigenschappen

Op theoretische gronden kan er niet echt een voorkeur aan de Nelson-Aalen-schatter of Kaplan-Meier-schatter gegeven worden. Voor grote datasets geven beide nagenoeg dezelfde uitkomsten.

5.6 Kwaliteitsindicatoren

De variantie van de Nelson-Aalen-schatter wordt gegeven door (Klein en Moeschberger, 2003)

$$\hat{\text{var}}\left(\hat{H}(t)\right) = \sum_{j:t_j<t} \frac{d_j}{n_j^2}. \qquad (5.6.1)$$

Voor varianties en betrouwbaarheidsintervallen voor overige statistieken gebaseerd op de Nelson-Aalen-schatter zie paragraaf 4.5.

6. Het vergelijken van duurverdelingen

6.1 Korte beschrijving

In dit hoofdstuk worden verschillende toetsen besproken die gebruikt kunnen worden om duurverdelingen te vergelijken.

6.2 Toepasbaarheid

Deze methoden kunnen gebruikt worden om te toetsen of groepen met verschillende achtergrondkenmerken significant verschillende duurverdelingen hebben. Zo kan men bijvoorbeeld toetsen of personen met een opleiding een kortere uitkeringsduur hebben dan personen zonder opleiding.

6.3 Uitgebreide beschrijving

Er zijn verscheidene toetsen die gebruikt kunnen worden. We zullen toetsen bespreken die het meest in de praktijk gebruikt worden. Dit zijn de volgende toetsen: logrank-toets, Wilcoxon-toets en de Tarone-Ware-toets. Andere meer specifieke toetsen kan men vinden in Lawless (2003).

In het algemeen is het wenselijk om de duurgegevens aan een beschrijvend onderzoek te onderwerpen voordat men gaat toetsen. Het is verstandig om een grafiek te maken van de survivalfuncties van de verschillende groepen die men wenst te vergelijken. Tevens is het aan te raden om op zijn minst de schatting van de mediaan en het gemiddelde met hun variantie en betrouwbaarheidsinterval op te vragen. Bovendien is het belangrijk om te bekijken wat het censureringspatroon is van de verschillende groepen die vergeleken dienen te worden. In ieder geval moet het percentage censureringen per groep bekend zijn.

Stel dat n_j het aantal personen is dat risico loopt op t_j. De riskset wordt aangegeven door n. Stel verder dat er k ($k \leq n$) verschillende tijdstippen $t_1 < t_2 < \ldots < t_k$ worden waargenomen waarop gebeurtenissen plaatsvinden en dat we r groepen met elkaar willen vergelijken.

De toetsingsgrootheden van de verschillende toetsen kunnen in een zelfde vorm worden geschreven. De toetsen verschillen alleen maar in het soort gewicht dat gekozen wordt. De toetsingsgrootheid is

$$\mathbf{u}'\mathbf{V}^{-1}\mathbf{u} \sim \chi^2_{r-1}. \tag{6.3.1}$$

waarbij

$$\mathbf{u}' = \sum_{j=1}^{k} W(t_j)(d_{1j} - E_{1j}, \ldots, d_{rj} - E_{rj}). \tag{6.3.2}$$

met d_{rj} als het aantal gebeurtenissen in groep r op tijdstip t_j, E_{rj} het verwachte aantal gebeurtenissen in groep r op tijdstip t en $W(t_j)$ het gewicht dat bij een specifieke toets hoort op tijdstip t_j.

Verder is V een r x r covariantie matrix met elementen

$$V_{il} = \frac{\sum_{j=1}^{k} W^2(t_j) n_{ij} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(d_{il} - \frac{n_{ij}}{n_j} \right). \qquad (6.3.3)$$

met $i, l = 1, \ldots, r$ en $d_{il} = I[i = l]$.

Een erg populaire toets is de logrank-test. In de meeste software paketten is dit de standaard toets die wordt berekend. Bij de logrank-test is $W(t_j)=1$. Opgemerkt moet worden dat deze toets beter niet gebruikt kan worden als de survivalfuncties kruizen. Een andere toets is de Wilcoxon-toets, hierbij is $W(t_j)=n_j$. Door dit gewicht wordt er meer nadruk gelegd bij vroege duren. Helaas is deze toets redelijk onbetrouwbaar als de censureringspatronen over de groepen verschillen. Tenslotte is er de Tarone-Ware toets. Deze is minder gevoelig dan de Wilcoxon-toets voor verschillende censureringspatronen bij groepen. Het gewicht bij de Tarone-Ware-toets is $W(t_j) = \sqrt{n_j}$. Tenzij men in het bijzonder geïnteresseerd is in korte duren zal de Tarone-Ware-toets in het algemeen betere resultaten laten zien.

7. Parametrisch model voor de hazard

7.1 Korte beschrijving

Bij deze methode wordt een parametrisch model geschat voor de hazard. De hazard hangt daarbij af van achtergrondkenmerken van de individuen. Zo zou bijvoorbeeld de uitkeringsduur kunnen afhangen van geslacht en leeftijd. Doordat duren vaak erg scheef verdeeld zijn en doordat er censurering optreedt is het niet mogelijk om 'normale' regressie toe te passen op de waargenomen duren.

Het model dat het meest gebruikt wordt is het proportional-hazards-model:

$$h_i(t) = r_i h_0(t) . \qquad (7.1.1)$$

Hierin wordt aangenomen dat de hazard voor iedereen dezelfde vorm heeft, namelijk $h_0(t)$, de baselinehazard. Individuele personen wijken van deze baseline hazard af met de risicofactor r_i, welke afhangt van de achtergrondkenmerken van het individu. Hoe hoger de risicofactor des te groter is de kans dat de gebeurtenis optreedt.

De baselinehazard modelleert de duurafhankelijkheid van de hazard. Voor $h_0(t)$ kunnen zeer veel functies gekozen worden. Zo zijn het exponentiële model en het Weibullmodel veel gebruikte modellen, maar ook polynomen en splines worden wel gebruikt. In de beschrijving zullen we ons beperken tot het exponentiële model en het Weibullmodel.

Naast het proportional-hazards-model bestaat er ook het accelerated-failure-time-model, waarbij personen dezelfde verdelingsfunctie voor de duren hebben, maar waarbij de tijd sneller of langzamer loopt voor personen afhankelijk van de achtergrondkenmerken van de personen. Deze modellen zullen niet besproken worden. Hiervoor wordt naar de literatuur verwezen (bijvoorbeeld Lawless, 1982; Klein en Moeschberger, 2003).

7.2 Toepasbaarheid

Met deze methode kan een parametrische beschrijving verkregen worden van de duurverdeling. De methodes zijn toepasbaar voor continue duren. De invloed van bepaalde achtergrondkenmerken op de duur kan hiermee bestudeerd worden.

Het is echter wel noodzakelijk om een parametrisch model aan te nemen voor de duurverdeling. Het kan lastig zijn om een model te vinden dat de waargenomen verdeling goed beschrijft. Voor een correcte schatting van de parameters is het echter noodzakelijk om een correct model te gebruiken. Als men alleen geïnteresseerd is in de relatieve invloed van achtergrondkenmerken op de hazard en niet in het modelleren van de gehele verdeling kan het in hoofdstuk 8 besproken Cox-model gebruikt worden, dat geen aannames doet voor de baselinehazard $h_0(t)$.

7.3 Uitgebreide beschrijving

Zoals gezegd wordt de hazard gemodelleerd als een baselinehazard vermenigvuldigd met de risicoscore. Nu moeten op een of andere manier de *k* achtergrondkenmerken van individu *i*, $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$, in deze risicoscore verwerkt worden. Hiervoor wordt meestal een exponent gebruikt:

$$h_i(t) = r_i h_0(t) = \exp(\mathbf{X}_i \beta) h_0(t) = \exp(\beta_1 x_{i1} + \ldots + \beta_k x_{ik}) h_0(t). \qquad (7.3.1)$$

Dit zorgt ervoor dat de risicoscore altijd positief is; de hazard mag/kan namelijk niet negatief zijn. De risicoscore geeft aan hoeveel meer risico een persoon loopt op het optreden van een gebeurtenis ten opzichte van de baselinehazard.

Om het model te kunnen schatten is het verder nodig om voor $h_0(t)$ een functie te kiezen. Hiervoor zijn zeer veel keuzes mogelijk. Hier worden alleen het exponentiële model en het Weibullmodel besproken. Andere keuzes zijn het lognormale, log-logistische en Gompertz model.

Bij het exponentiële model, wordt aangenomen dat de hazard constant is voor iedereen. Alleen het niveau verschilt tussen individuen. De baselinehazard wordt dus gegeven door

$$h_0(t) = \lambda \quad \text{met } \lambda > 0. \qquad (7.3.2)$$

Een constante hazard vertaald zich in een lineair toenemende cumulatieve hazard aangezien volgens vergelijking (2.3.5)

$$H(t) = \int_0^t h(u)\mathrm{d}u = \int_0^t \lambda \mathrm{d}u = \lambda t. \qquad (7.3.3)$$

De Nelson-Aalen-schatter zou dus gebruikt kunnen worden om na te gaan of het exponentiële model een redelijk model zou kunnen zijn.

In de praktijk zal dit model niet vaak voorkomen, omdat de hazard in het algemeen wel afhangt van de duur. Zo neemt het risico op overlijden toe met leeftijd (na een daling in de eerste levensjaren) en neemt het risico op het vinden van een baan in het algemeen af met de werkloosheidsduur. Een model dat wel met een stijgende of dalende hazard kan omgaan is het Weibullmodel, waarvan de baselinehazard gegeven wordt door

$$h_0(t) = \lambda p t^{p-1} \quad \text{met } \lambda > 0 \text{ en } p > 0. \qquad (7.3.4)$$

Als *p* gelijk is aan één dan is dit model gelijk aan het exponentiële model. Voor *p>1* neemt de hazard toe met de duur en voor *p<1* neemt de hazard af. Een toe- of afnemende hazard vertaald zich in een convexe of concave hazard.

7.4 Kwaliteitsindicatoren

Voor de beoordeling van de kwaliteit van duurmodellen zijn veel verschillende residuen beschikbaar. Dit komt voor een groot gedeelte doordat geen van de

residuen op zichzelf voldoende is om de kwaliteit te beoordelen. Voor een goede beoordeling van de kwaliteit is het dus nodig om alle residuen in ogenschouw te nemen. De verschillende residuen worden in de volgende paragrafen besproken.

Naast de residuen kan een eenvoudige plot van de door het model voorspelde survivalfunctie voor verschillende groepen tegen de met de Kaplan-Meier-schatter geschatte survivalfunctie voor de groepen al een goede indicatie geven van de fit van het model.

7.4.1 Martingale residuen

Stel N_i is het aantal gebeurtenissen dat een individu *i* heeft ondergaan. In de situatie zoals deze hier besproken wordt is dit één als de duur is afgesloten en er dus een gebeurtenis is waargenomen en nul als de duur is gecensureerd. De martingale residu m_i is dan gedefinieerd als het verschil tussen het aantal waargenomen gebeurtenissen en het aantal verwachte gebeurtenissen:

$$m_i = N_i - \mathrm{E}[N_i]. \tag{7.4.1}$$

Het aantal verwachte gebeurtenissen is altijd groter of gelijk aan nul en kan ook groter zijn dan één. Iemand kan bijvoorbeeld zo ongezond leven en toch dermate oud worden dat hij eigenlijk al ‘meerdere keren had moeten overlijden’. Als een individu gecensureerd is dan is $m_i \leq 0$. Als een individu een gebeurtenis ondergaat dan:

- $0 < m_i \leq 1$: de gebeurtenis was ‘te vroeg’;
- $m_i < 0$ de gebeurtenis was ‘te laat’.

Bij het beoordelen van de residuen moet rekening gehouden worden met het feit dat de residuen scheef verdeeld zijn: ze zijn altijd kleiner of gelijk aan één. De residuen zijn op twee manieren te gebruiken:

- Een plot van de residuen tegen een covariaat die niet is meegenomen in het model, geeft het functionele verband aan. Een lineair verband tussen de residuen en de covariaat geeft aan dat de invloed van de covariaat lineair in het model kan worden meegenomen (zoals hierboven is besproken).
- Een plot van de residuen tegen een covariaat die is meegenomen in het model zou geen verband moeten laten zien.

Om het verband beter te kunnen zien kan bijvoorbeeld een Lowess-smooth door de residuen geplot worden.

7.4.2 Devianceresiduen

Deviance residuen zijn een transformatie van martingale residuen:

$$d_i \approx \frac{N_i - \mathrm{E}[N_i]}{\mathrm{E}[N_i]}. \tag{7.4.2}$$

De residuen zijn symmetrisch verdeeld rond nul. Bij kleine hoeveelheden censurering (ongeveer <25%) zijn de residuen bij benadering normaal verdeeld. Bij grotere hoeveelheden censurering zijn de residuen nog wel symmetrisch, maar is er wel een groter aandeel kleine waardes. Een waarde kleiner dan nul geeft aan dat de gebeurtenis later plaats vond dan verwacht en een waarde groter dan nul dat de gebeurtenis eerder plaatsvond dan verwacht.

De residuen kunnen gebruikt worden om duren op te sporen die niet goed door het model voorspeld worden. Waardes die bijvoorbeeld groter/kleiner zijn dan ±2.5/3 zijn verdacht. Een plot van de residuen tegen de risicoscores kan gebruikt worden om na te gaan of bepaalde risicoscores slecht voorspeld worden.

7.4.3 Cox-Snellresiduen

De Cox-Snellresiduen volgen uit de baseline cumulatieve hazard gewogen met de riskscore van de individuen:

$$cs_i = \hat{H}_0(t_i)\exp\left(\mathbf{X}_i\hat{\boldsymbol{\beta}}\right). \tag{7.4.3}$$

Als het model goed bij de data past, dan volgen zij een exponentiële verdeling met $\lambda = 1$. Zoals volgt uit vergelijking (7.3.3), moet de cumulatieve hazard van de residuen een rechte lijn door de oorsprong met helling 1 zijn. Om dit na te gaan kan de Nelson-Aalen-schatter van de residuen bepaald worden. Deze kan dan naast een lijn met helling 1 geplot worden om na te gaan hoe goed het model de waarnemingen beschrijft. Doordat het aantal lange duren in het algemeen klein is, zal er bij lange duren altijd enige afwijking zijn.

7.4.4 Schoenfeldresiduen

De Schoenfeldresiduen, ook wel partial residuals genoemd, zijn gedefinieerd als (Schoenfeld, 1982)

$$\mathbf{s}_i = \mathbf{X}_i - \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}, t_i)\,, \tag{7.4.4}$$

waarbij $\mathbf{X}_i$ de covariatenvector van indiviu *i* is en $\bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}, t_i)$ de uit het model volgende gemiddelde covariatenvector van alle individuen op het tijdstip t_i waarop individu *i* zijn gebeurtenis onderging. De residuen geven dus aan in welke mate de individu *i* afwijkt in achtergrondkenmerken van de overige individuen die op dat moment in leven zijn. De residuen zijn dus vectoren met een waarde voor ieder in het model meegenomen covariaat.

De residuen kunnen gebruikt worden om na te gaan of het effect dat bepaalde covariaten hebben op de hazard afhangt van de duur. Zo kan het bijvoorbeeld zijn dat de eerste tijd na werkloos worden er weinig verschil is tussen mannen en vrouwen wat betreft de kans op het vinden van een baan, maar dat naarmate de werkloosheid langer duurt er verschillen beginnen op te treden. In dit geval geldt dus niet meer dat de hazard proportional is. Een plot van de Schoenfeldresiduen van geslacht tegen de duur zou in dit geval een verband laten zien. Het is mogelijk om

deze tijdsafhankelijke invloed mee te nemen in het model door in feite de β te laten variëren met de duur. Deze uitbreiding op het model wordt hier verder niet besproken. Meer informatie hierover is te vinden in (Singer en Willet, 2003; Klein en Moeschberger, 2003).

7.4.5 Geschaalde Schoenfeldresiduen

De geschaalde Schoenfeldresiduen $\mathbf{s}_i^*$ zijn gedefinieerd als (Grambsch en Therneau, 1994):

$$\mathbf{s}_i^* = \mathbf{V}^{-1}\left(\hat{\boldsymbol{\beta}}, t_i\right)\mathbf{s}_i\,, \tag{7.4.5}$$

waar $\mathbf{V}^{-1}\left(\hat{\boldsymbol{\beta}}, t_i\right)$ gezien kan worden als de covariantiematrix van de Schoenfeldresiduen. Het voordeel van deze schaling is ten eerste dat de correlatie tussen de residuen vermindert. Bij de ongeschaalde Schoenfeldresiduen kan het voorkomen dat een tijdsafhankelijkheid in één covariaat zichtbaar wordt in de Schoenfeldresiduen van een andere covariaat. Dit is bij de geschaalde Schoenfeldresiduen minder. Het tweede voordeel is dat het model dat gebruikt moet worden voor de tijdsafhankelijkheid, makkelijker af te lezen is uit de residuen doordat bij benadering geldt

$$\beta_j(t_i) \approx \mathrm{E}\left[s_{ij}^*\right] + \hat{\beta}\,. \tag{7.4.6}$$

De geschaalde Schoenfeld residuen zijn dus makkelijker in gebruik dan de ongeschaalde residuen. In het algemeen zal men de residuen van een parameter plotten tegen de duur in een scatterplot. Een Lowess smooth door de plot kan helpen bij het interpreteren van de residuen. Om te kijken of er een tijdsafhankelijkheid is, kan ook de correlatiecoëfficient tussen de residuen en de duur berekend worden. Een significante correlatie geeft aan dat een tijdsafhankelijkheid aanwezig is.

8. Cox-model / proportional-hazard-model

8.1 Korte beschrijving

Als men geïnteresseerd is in de relatieve invloed van achtergrondkenmerken op de hazard en niet in een modelleren van de gehele verdeling dan kan men kiezen voor een Cox model (Cox, 1972). Het cox model is als volgt gedefinieerd:

$$h(t) = h_0(t) r_i = h_0(t) \exp(\mathbf{X}'\beta). \tag{8.1.1}$$

Merkt op dat ook hier weer gekozen is voor het exponent. Dit zorgt ervoor dat de risicoscore altijd positief is. De hazard mag namelijk nooit negatief zijn.

8.2 Toepasbaarheid

Het Cox model kan men gebruiken als de interesse uitgaat naar de relatieve invloed van achterkenmerken op de hazard en niet in het modelleren van de gehele duurverdeling. Bij het coxmodel worden geen aannames gedaan over precieze vorm van de baselinehazard $h_0(t)$. Het model is daarom bijzonder flexibel en er wordt geen parametrisch model aangenomen. Het model is goed geschikt om het effect van achtergrondkenmerken op de hazard te onderzoeken.

8.3 Uitgebreide beschrijving

Bij het cox model is het mogelijk om het effect van achtergrondkenmerken op de hazard te onderzoeken zonder dat er aannames gedaan worden over de baselinehazard $h_0(t)$. Voor de volledigheid het coxmodel is als volgt gedefineerd:

$$h(t) = h_0(t) r_i = h_0(t) \exp(\mathbf{X}'\beta). \tag{8.3.1}$$

Een coxmodel maakt hiervoor gebruik van de proportional-hazards-assumptie. Dat betekent dat er wordt aangenomen dat iedereen dezelfde baseline hazard heeft en dat alleen de schaling tussen individuen met verschillende achtergrondkenmerken verschilt. Volgens de proportional-hazard-assumptie moet voor een individu *i* en individu *j* het volgende gelden:

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i) \cdot h_0(t)}{\exp(\boldsymbol{\beta}'\mathbf{X}_j) \cdot h_0(t)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}{\exp(\boldsymbol{\beta}'\mathbf{X}_j)} = \exp(\boldsymbol{\beta}'(\mathbf{X}_i - \mathbf{X}_j)). \tag{8.3.2}$$

Wat direct opvalt aan bovenstaande vergelijking is het wegvallen van de baselinehazard voor beide individuen. Vanwege dit feit hoeven er geen aannames gedaan over de precieze vorm van de baselinehazard.

Net als bij het parametrische model wordt de hazard gemodelleerd als een baselinehazard vermenigvuldigd met een risicoscore. Net als bij het parametrische model wordt hierbij aangenomen dat de hazard voor iedereen dezelfde vorm heeft

gegeven door $h_0(t)$, de baselinehazard. Individuele personen wijken hier dan vanaf met de risicofactor r_i, welke afhangt van de achtergrondkenmerken van de individu. De risicoscore geeft aan hoeveel meer risico een persoon loopt op het optreden van een gebeurtenis ten opzichte van de baselinehazard. Ook hier geldt weer: hoe hoger de risicofactor des te groter is de kans dat de gebeurtenis optreedt. Verder moet opgemerkt worden dat er geen intercept is opgenomen in $\mathbf{X}'\beta$. De intercept is namelijk opgenomen in de baselinehazard $h_0(t)$.

Ook in de likelihood van een Cox model komt er geen baselinehazard voor. De likelihood die gebruikt wordt om de coefficienten β te schatten is een partiële likelihood. Strikt genomen is deze partiële likelihood geen normale likelihood, maar onder milde veronderstellingen kan deze toch opgevat worden als een normale likelihood. Men kan deze likelihood dan gebruiken in likelihoodratio tests en voor hypothese-toetsen en betrouwbaarheidsintervallen. Dat dit gerechtvaardigd is wordt beschreven in Lawless (2003, par. 7.1.3).

Het model kan niet zonder meer geschat worden als er duren van gelijke lengte (in het Engels ties genoemd) voorkomen in de steekproef. De partiële likelihood kan in eerste instantie niet omgaan met duren van gelijke lengte. Gelukkig zijn er aanpassingen beschikbaar. Twee populaire aanpassingen zijn exact partial likelihood en Efrons likelihood. Bij de exact partial likelihood wordt elke mogelijke volgorde van gebeurtenissen afgegaan en deze is daarom erg langzaam als er veel duren van gelijke lengte optreden. Efrons likelihood benadert de likelihood, maar is veel sneller dan de exact partial likelihood. Het advies is standaard de exact partial likelihood te gebruiken. Als er echter veel duren van gelijke lengte zijn of een grote steekproef dan dient men uit te wijken naar Efrons likelihood. Voor meer informatie over beide aanpassingen zie Lawless (2003, par. 7.1.4).

8.4 Kwaliteitsindicatoren

Men kan hiervoor dezelfde kwaliteitsindicatoren gebruiken als bij parametrische modellen. Er wordt dan ook verwezen naar paragraaf 7.4.

9. Logistisch model voor discrete duren

9.1 Korte beschrijving

Zoals besproken in hoofdstuk 1, is de hazard de kans dat een gebeurtenis optreedt in een zeker tijdsinterval gegeven dat deze aan het begin van het interval nog niet is opgetreden. Het logistische regressiemodel wordt vaak gebruikt om kansen te modelleren. Het ligt daarom voor de hand om een model voor de hazard van discrete duren ook op het logistische model te baseren.

Het blijkt dat door de duurdata anders te organiseren (dan bij de meeste andere methodes gebruikelijk is), de logistische-regressieroutines zoals deze in bijna alle statistische pakketten voorkomen te gebruiken zijn om discrete duren te modelleren.

9.2 Toepasbaarheid

Het is met deze methode mogelijk om de discrete verdeling van duren te beschrijven afhankelijk van bepaalde achtergrondkenmerken. Er wordt daarbij dus aangenomen dat de duren discreet zijn. Indien de duren niet discreet zijn, maar slechts gediscretiseerd zijn (bijvoorbeeld werkloosheidsduur gemeten in jaren), wordt aangenomen dat gecensureerde personen het hele interval risico liepen. Daarnaast zal er in dit geval ook een verlies in efficiëntie optreden omdat door de discretisatie informatie verloren gaat.

Het voordeel van deze methode ten opzichte van bijvoorbeeld parametrische regressie en Cox-regressie, is dat de baselinehazard automatisch meegeschat wordt. Daarbij kunnen aannames gedaan worden, zoals een constante hazard. Het model met deze aannames kan vervolgens getoetst wordt tegen het model zonder aannames.

9.3 Uitgebreide beschrijving

In het geval van discrete duren, is de hazard p_{ij} de kans dat de duur van individu i gelijk is aan t_j gegeven dat duur gelijk of langer is dan t_j:

$$p_{ij} = \Pr\left(T_i = t_j \mid T_i \geq t_j\right). \tag{9.3.1}$$

We willen deze hazard p_{ij} modelleren aan de hand van achtergrondkenmerken $\mathbf{X_{ij}}$. Zoals de index j al aangeeft, is er een vector met achtergrondkenmerken voor iedere tijdsperiode. Het is dus mogelijk dat de covariaten variëren met de tijd (zoals inkomen). Ze kunnen natuurlijk ook constant zijn in de tijd (zoals geslacht). Aangezien p_{ij} een kans is, ligt het voor de hand om dezelfde transformatie toe te passen als bij logistische regressie:

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j + \mathbf{X}'_{ij}\boldsymbol{\beta}, \tag{9.3.2}$$

waarin α_j de baselinehazard is en $\mathbf{X}'_{ij}\boldsymbol{\beta}$ het effect van de achtergrondkenmerken op de baselinehazard modelleert. Uit de geschatte waarden van α_j en $\boldsymbol{\beta}$ kan vervolgens de hazard teruggevonden worden met

$$p_{ij} = \frac{\exp(\alpha_j + \mathbf{X}'_{ij}\boldsymbol{\beta})}{1+\exp(\alpha_j + \mathbf{X}'_{ij}\boldsymbol{\beta})}. \tag{9.3.3}$$

De survivalfunctie en cumulatieve hazard kunnen teruggevonden worden met de formules uit hoofdstuk 1.

Om het model te kunnen schatten in de meeste statistische programma's, is het nodig om de data anders te rangschikken dan gewoonlijk is bij duuranalyses. In figuur 9.3.1 is links het formaat weergegeven zoals dat normaal gebruikt wordt: voor iedere persoon is er één record met daarin de duur en daarbij of deze wel of niet is afgesloten en eventuele covariaten. Om het logistische duurmodel te kunnen schatten is het nodig om per persoon en per periode (waarin de persoon 'in leven is') een record te hebben: het persoon-periode-formaat. Voor iedere periode wordt aangegeven of er een gebeurtenis is opgetreden of niet. Dit formaat is rechts weergegeven in figuur 9.3.1. In SPSS kan hiervoor de routine VARSTOCASES gebruikt worden; STATA heeft een speciale routine voor duurdata STSPLIT.

persoon-formaat

ID	Duur	Gebeurtenis	Geslacht	Werk1	Werk2	Werk3
1	2	1	m	0	0	x
2	3	0	v	0	1	1
3	1	1	m	0	x	x
4	5	1	m	1	1	1
5	2	1	m	1	0	x

persoon-periode-formaat

ID	Periode	Gebeurtenis	Geslacht	Werk
1	1	0	m	0
1	2	1	m	0
2	1	0	v	0
2	2	0	v	1
2	3	0	v	1
3	1	1	m	0
4	1	0	m	1
4	2	0	m	1
4	3	0	m	1

Figuur 9.3.1. Het persoon-formaat en het persoon-periode-formaat voor het opslaan van duurdata

Het blijkt dat de likelihood van het logistische duurmodel overeenkomt met de likelihood van een logistische regressie op het wel of niet optreden van een gebeurtenis (variabele 'gebeurtenis' in figuur 9.3.1). De normale logistische regressieroutines kunnen dus gebruikt wordt om het model te schatten.

Een paar voorbeelden van mogelijke modellen:

Constante hazard (geen duurafhankelijkheid):

$$\text{logit}(p_{ij}) = \alpha + \beta_1 X_1 + \beta_2 X_{2j}. \tag{9.3.4}$$

In dit model wordt aangenomen dat de baselinehazard niet duurafhankelijk is. De hoogte van de hazard wordt verklaard aan de hand van de covariaten X_1 en X_{2j}, waarin de laatste verandert in de tijd. In de situatie in figuur 9.3.1 zouden X_1 en X_{2j} respectievelijk ‘geslacht’ en ‘werk’ kunnen zijn.

Duurafhankelijke hazard:

$$\text{logit}(p_{ij}) = \alpha_j + \beta_1 X_1 + \beta_2 X_{2j}. \tag{9.3.5}$$

Door de duur als categoriale variabele mee te nemen in het model, zodat er voor iedere duurperiode een dummy variabele gecreëerd wordt, kan een baseline hazard worden meegenomen die in iedere tijdsperiode een andere waarde kan aannemen. Er worden dus geen aannames gedaan over het verloop van de baseline hazard.

Hazard lineair afhankelijk van de duur:

$$\text{logit}(p_{ij}) = \alpha + \beta_0 t_j + \beta_1 X_1 + \beta_2 X_{2j}. \tag{9.3.6}$$

Door duur niet als categoriale variabele maar als continue variabele mee te nemen kan een baseline hazard die lineair van de duur afhangt worden aangenomen in het model.

9.4 Eigenschappen

Hoewel de standaard logistische-regressieroutines gebruikt worden bij het schatten, is de situatie niet gelijk aan die bij logistische regressie. Alleen de likelihood komt overeen. Zaken die bij normale logistische regressie gebruikt worden om bijvoorbeeld de kwaliteit te beoordelen zijn daarom niet per se ook toepasbaar voor logistische duurregressie. Alleen zaken gebaseerd op de likelihood, zoals AIC, likelihoodratiotests, zijn wel bruikbaar.

9.5 Kwaliteitsindicatoren

Geneste modellen zoals bijvoorbeeld in vergelijkingen (9.3.4) tot (9.3.6) kunnen vergeleken worden met behulp van een likelihoodratiotest.

10. Literatuur

Brookmeyer, R. en Crowley, J.J. (1982), A confidence interval for the median survival time. *Biometrics* 38, 29-41.

Berkson, J. en Gage, R.P. (1950), Calculation of survival rates for cancer. *Proceedings of Staff Meetings of the Mayo Clinic* 25, 270-286.

Chiang, C.L. (1984), *The life table and its applications*. Robert E. Krieger Publishing Company, Malabar.

Chiang, C.L. (1968), *Introduction to stochastic processes in biostatistics*. Wiley, New York.

Chiang, C.L. (1960a), A stochastic study of life table and its applications: I. Probability distributions of the biometric functions. *Biometrics* 16, 618-635.

Chiang, C.L. (1960b), A stochastic study of life table and its applications: II. Sample variance of the observed expectation of life and other biometric functions. *Human Biology* 32, 221-238.

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

Gehan, E.A. (1969), Estimating survival functions from the life table. *Journal of Chronic Disease* 21, 629-644.

Cutler, S.J. and Ederer, F. (1958), Maximum utilization of the life table method in analyzing survival. *Journal of Chronic Disease* 8, 699-712.

Grambsch, P.M. en Therneau, T.M. (1994), Proportional hazard tests and diagnostics based on weighted residuals. *Biometrika* 81, 515–526.

Greenwood, M. (1926), The natural duration of cancer. *Reports of Public Health and Medical Subjects* 33, Her Majesty's stationery office, London.

Klein, J.P. en Moeschberger, M.L. (2003), *Survival analysis: techniques for censored and truncated data*. Springer, New York.

Lawless, J.F. (2003), *Statistical models and methods for lifetime data. 2nd edition*, Wiley, New York.

Meulen, A. van der (2009), *Overlevingstafels en Longitudinale analyse: deelthema Overlevingstafels*. Rapport Methodenreeks, CBS, Den Haag.

Meulen, A. van der en Jansen, F. (2007), *Achtergronden en berekeningswijzen van CBS-overleveringstafels*. Bevolkingstrends, 3e kwartaal 2007.

Schoenfeld, D. (1982), Partial residuals for the proportional hazards regression model. *Biometrika* 69, 239–241.

Singer, J. D. en Willet, J.B. (2003), *Applied longitudinal data analysis: modeling change and event occurrence*. Oxford University Press.

Wolthuis H. en Bruning, R. (1996), *Levensverzekeringswiskunde*. Ceuterick, Leuven.