Advances in Survival Analysis and Optimal Scaling Methods

Willems, S.J.W.

Samenvatting

Dit proefschrift bestaat uit vijf wetenschappelijke artikelen over verschillende statistische onderwerpen, namelijk overlevingsanalyse, optimale-schalingtransformaties en statistiekcommunicatie, of een combinatie van deze onderwerpen. Elk onderzoek is gemotiveerd door een specifieke dataset of door een interesse in een bepaalde eigenschap van data.

Hoofdstuk 1

De motivatie voor het onderzoeksproject dat wordt beschreven in het eerste hoofdstuk is een vraag van de afdeling Psychiatrie van het Leids Universitair Medisch Centrum. Doctoren op deze afdeling verzamelden data over patiënten met stemmings- en angststoornissen om het beloop van hun suïcidale gedachten te onderzoeken. In het bijzonder wilden zij op basis van de eigenschappen van een patiënt voorspellen hoeveel tijd er zit tussen de diagnose en de remissie van suïcidale gedachten. Het doel van het schatten van deze hersteltijd was om te achterhalen welke patiënten een verhoogde kans hebben op aanhoudende suïcidale gedachten.

De statistische methoden die gebruikt worden om dit soort vragen te beantwoorden worden *overlevingsanalysemethoden* genoemd. Deze methoden worden gebruikt om de tijd te modelleren tussen een specifiek startpunt en het moment waarop de gebeurtenis waar de interesse naar uitgaat plaatsvindt. Bij het schatten van de hersteltijd is het startpunt dus het moment waarop de patiënt de diagnose krijgt en de interesse gaat uit naar het moment van remissie van suïcidale gedachten. Overlevingsanalyse kan naast hersteltijd ook gebruikt worden om andere tijdsintervallen te modelleren, bijvoorbeeld de overlevingsduur van patiënten (tijd vanaf diagnose tot overlijden) of werkloosheidsduur (tijd tussen het verliezen van een baan het starten van een nieuwe).

Om de hersteltijd van patiënten te bepalen moeten deze patiënten over lange tijd geobserveerd worden, namelijk tot het moment dat de remissie wordt geconstateerd. Het kost dus tijd om alle benodigde data te verzamelen en dit verhoogt het risico dat de remissie voor sommige personen niet geobserveerd zal worden. Als een patiënt bijvoorbeeld gedurende het onderzoek verhuist naar

een andere stad en daarom overstapt naar een ander ziekenhuis, dan zullen de onderzoekers in het eerste ziekenhuis het herstel van deze patiënt niet waarnemen. Dit fenomeen wordt *censurering* genoemd omdat het herstel van deze persoon niet geobserveerd wordt en dus *gecensureerd* is. Hoewel de data van een gecensureerd persoon incompleet zijn, is de beschikbare informatie wel waardevol. Van de verhuisde patiënt is het bijvoorbeeld wel bekend dat hij of zij nog niet hersteld was op het moment van het laatste ziekenhuisbezoek. Dit geeft aan dat herstel op zijn vroegst *ná* dit bezoek plaats zal vinden.

Aangezien censurering regelmatig voorkomt wanneer data worden verzameld om tijdschattingen te maken, zijn de overlevingsanalysemethoden zo ontworpen dat zij ook dit soort incomplete maar waardevolle data kunnen meetellen. Om deze modellen simpel te houden zijn ze ontwikkeld onder de aanname dat de censurering willekeurig is. Dit betekent dat iedereen dezelfde kans moet hebben om gecensureerd te worden. Die kans mag daarom niet afhangen van de eigenschappen van een persoon of het moment waarop zijn of haar gebeurtenis plaatsvindt. Als een patiënt bijvoorbeeld van ziekenhuis verandert na een verhuizing dan kunnen we ervan uitgaan dat dit net zo goed bij een andere patiënt had kunnen gebeuren. In andere woorden, de oorzaak van censurering is *onafhankelijk* van de gezondheid van de patiënt. Deze onafhankelijkheid is belangrijk voor een correcte analyse.

De doctoren wie de hersteltijd probeerden te schatten vermoedden echter dat juist voornamelijk de patiënten wie zich beter begonnen te voelen niet meer naar de kliniek kwamen. Dus, het vermoeden was dat de kans op censurering gerelateerd is aan de gezondheid van een patiënt en dat deze dus niet voor alle patiënten gelijk is. Daarom zou de aanname van onafhankelijke censurering incorrect kunnen zijn en daardoor kunnen de standaard overlevingsanalyse-methoden onnauwkeurige resultaten geven. De data zouden namelijk partijdig zijn omdat deze meer informatie bevatten over patiënten met een lange hersteltijd en minder informatie over degenen met een kort herstel. Hierdoor kan de hersteltijd overschat worden. Om te voorkomen dat de door afhankelijke censurering veroorzaakte partijdigheid ook in de resultaten doordringt kunnen er speciale correcties worden toegepast in de overlevingsanalyse.

Dit eerste hoofdstuk richt zich op een specifieke correctiemethode genaamd *Inverse Probability Censoring Weighting Estimator*. Wanneer een observatie gecensureerd is compenseert deze methode voor het verlies aan informatie door in de analyse extra gewicht te geven aan de individuen wie het meest lijken op de gecensureerde persoon en nog wel geobserveerd worden.

Als onderdeel van dit proefschrift is een gebruiksvriendelijke implementatie van deze correctiemethode ontwikkeld voor het veelgebruikte statistische software-programma *R*. Deze implementatie maakt de methode makkelijker beschikbaar

voor onderzoekers wie zelf weinig programmeerervaring hebben.

De methode is daarna toegepast op de data van patiënten met stemmings- en angststoornissen om te corrigeren voor de vermoedelijke afhankelijke censurering. De resultaten lieten zien dat de hersteltijd inderdaad overschat werd zonder deze correctie, maar het verschil was erg klein. Dit geeft aan dat de afhankelijke censurering ofwel niet problematisch was ofwel dat belangrijke informatie over de patiënten mistte waardoor de gewichten niet correct verdeeld konden worden over de patiënten die nog wel geobserveerd werden.

De methode is daarna getest op data met afhankelijke censurering door middel van een simulatiestudie. Deze simulatiestudie toonde aan dat de correctiemethode inderdaad kan corrigeren voor afhankelijke censurering. Echter, zoals te verwachten is, presteert deze methode het beste als voor iedereen de kans op censurering correct kan worden geschat. Om de hersteltijd goed te kunnen schatten moet daarom data over veel personen worden verzameld en moeten ook alle eigenschappen die geassocieerd zijn met censurering geregistreerd worden.

Hoofdstuk 2

In het tweede hoofdstuk ligt de focus op een toepassing van overlevingsanalyse en optimale-schalingtechnieken om mogelijke voorspellers van werkloosheidsduur te identificeren. De dataset die gebruikt is komt van het Nederlandse Uitvoeringsinstituut Werknemersverzekeringen (UWV). In Nederland kunnen werkzoekenden wie recent werkloos zijn geworden een werkloosheidsuitkering aanvragen bij het UWV en ook begeleiding en advies krijgen bij het zoeken naar een nieuwe baan. Omdat het budget beperkt is moeten de kosten van deze adviesservice gereduceerd worden. Een besparingsstrategie is het verminderen van één-op-één persoonlijke begeleiding. Het UWV heeft daarom een online vragenlijst ontwikkeld om werklozen te selecteren wie moeite hebben om zelf een baan te vinden en daarom baat zullen hebben bij persoonlijke begeleiding. Het UWV zal veel kosten besparen wanneer de persoonlijke begeleiding alleen aangeboden wordt aan de personen wie dit echt nodig hebben.

Via de online vragenlijst verzamelt het UWV data over de werkzoekenden op 17 verschillende factoren, zoals *werkzoekgedrag*, *werkzoekattitude*, *acceptatiebereidheid* en *zelf-effectiviteit m.b.t. sollicitatievoorbereiding*. Deze factoren worden gemeten met 1–5 vragen in de vragenlijst. De laatstgenoemde factor wordt bijvoorbeeld gemeten met behulp van de drie volgende uitspraken:

– *Ik kan goed informatie vinden over vacatures.*
– *Ik kan mijn sterke en zwakke punten voor een baan goed uitleggen.*
– *Ik kan een goede (digitale) sollicitatie schrijven.*

Werkzoekenden worden gevraagd om aan te geven in hoeverre zij het eens zijn

met elk van deze stellingen en de gegeven antwoordopties zijn *zeer mee oneens*, *mee oneens*, *niet eens/niet oneens*, *mee eens* en *zeer mee eens*.

De gegeven antwoorden zijn meestal sterk aan elkaar gerelateerd. Als een persoon bijvoorbeeld in staat is om zelfstandig een baan te zoeken en sollicitaties te versturen, dan zullen zijn of haar antwoorden op de bovenstaande drie uitspraken allemaal *mee eens* of *zeer mee eens* zijn, en vice versa. In andere woorden, de antwoorden op de vragen binnen een factor zijn veelal hetzelfde. Sommige factoren zijn ook sterk aan elkaar gerelateerd. Als een persoon bijvoorbeeld positief staat tegenover het zoeken van een baan, dan zal hij of zij waarschijnlijk ook actief op zoek gaan. Daarom zal deze persoon het ook eens zijn met de uitspraken over zowel de factor *werkzoekattitude* als de factor *werkzoekgedrag*.

Voor statistische analyses is het beter als sterk gerelateerde variabelen niet individueel in een model worden opgenomen, maar eerst worden samengevat in een klein aantal samenvattingsscores. Een simpele manier om deze scores te krijgen is door het (gewogen) gemiddelde te berekenen over de vragen binnen een factor. Deze aanpak met gewogen gemiddelden is gebruikt om elke factor samen te vatten en *factor analyse* is toegepast om de optimale gewichten te berekenen voor elke vraag. Omdat het onmogelijk is om berekeningen te doen met namen van categorieën zoals *zeer mee eens*, werden de vijf antwoordopties vervangen door de gehele getallen 1–5 zodat deze waarden gebruikt konden worden in de berekeningen.

Daarna werden de samenvattingsscores en een aantal aanvullende variabelen, zoals leeftijd, gender en opleiding, gebruikt om de kans dat een werkloze een baan vindt te modelleren. Voor deze analyse werd *logistische regressie* gebruikt welke de werkzoekende kan classificeren in één van twee groepen. De eerste groep bevat de personen wie waarschijnlijk binnen een jaar een baan zullen vinden en de tweede groep bevat de personen wie dat waarschijnlijk niet lukt. Naast de werkstatus na één jaar was het UWV ook geïnteresseerd in deze status na 6 en 9 maanden. Daarom werden er drie analyses gedaan, namelijk om de kansen te berekenen op het vinden van een nieuwe baan binnen 6, 9 en 12 maanden.

Ook al geeft deze analyse al een goed inzicht in welke factoren de kans op het vinden van een nieuwe baan beïnvloeden, zijn er ook andere methodes die gebruikt kunnen worden om de data voor te bereiden en te analyseren. In het tweede hoofdstuk van dit proefschrift wordt voor deze beide stappen een alternatieve methode voorgesteld en toegepast op de dataset van het UWV.

Voor de datavoorbereiding wordt de methode *optimale-schaling principale componentenanalyse (OSPCA)* voorgesteld. Zoals bij factor analyse heeft deze methode als doel om de vele variabelen samen te vatten in een kleinere verzameling van samenvattingsscores. Echter worden deze scores in OSPCA niet individueel per factor berekend, maar zijn het gewogen gemiddelden over *alle*

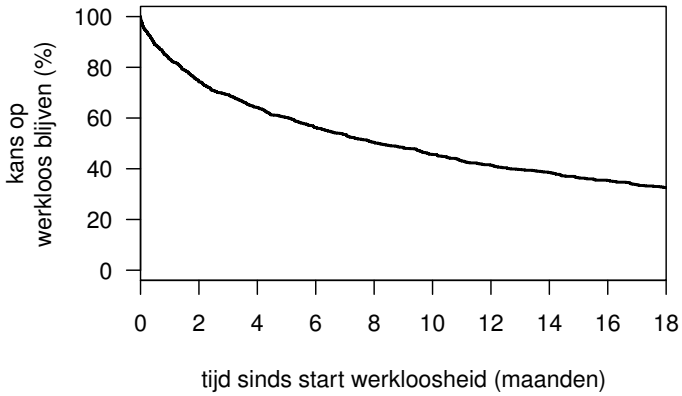
oorspronkelijke variabelen. De gewichten worden op zo een manier gekozen dat de samenvattingsscores samen zo veel mogelijk van de variatie tussen de personen in de dataset beschrijven.

Het grootste voordeel van dit model is dat het ook optimale waarden vindt voor de categorieniveaus voordat het gewogen gemiddelde berekend wordt. Het zou bijvoorbeeld kunnen blijken dat de waarden 1,5, 2, 3, 4 en 4,5 betere vervangers zijn voor de categorieën dan de gehele getallen 1, 2, 3, 4 en 5, omdat de stap van het niet hebben van een mening (*niet eens/niet oneens*) naar het (on)eens zijn met een stelling groter is dan de stap tussen het (on)eens zijn en zeer (on)eens zijn met een stelling. De stappen naar de extremere mening zijn kleiner en worden daarom in dit voorbeeld weergegeven door een kleinere stapgrootte. Op deze manier wordt de volgorde van de categorieniveaus behouden door de vervangende waarden zonder dat er gelijke afstanden tussen de opeenvolgende niveaus worden geforceerd zoals dat werd gedaan bij de vervanging door gehele getallen.

Verder kan OSPCA ook inzicht geven in de relaties tussen de verschillende factoren. Deze methode zal de data van het UWV namelijk samenvatten door gewogen gemiddelden te berekenen over de antwoorden op alle stellingen in plaats van dit per factor te doen. Op deze manier geven de samenvattingsscores ook inzicht in de onderliggende relatie tussen de 17 factoren. Dit kan bijvoorbeeld gebruikt worden om de verwachte relatie tussen *werkzoekattitude* en *werkzoekgedrag* te verifiëren. Dit vermogen om ook de onderlinge relaties tussen factoren te bestuderen is een ander voordeel van OSPCA. Het kan echter ook een nadeel zijn aangezien de samenvattingsscores lastiger te interpreteren zijn als zij berekend worden over alle variabelen in plaast van per factor.

De voorgestelde alternatieve methode voor de uiteindelijke analyse is overlevingsanalyse. Deze methode (welke ook in het eerste hoofstuk is gebruikt) kan de kans om werkloos te blijven schatten over een heel tijdsinterval in plaats van op een specifiek tijdstip zoals één jaar. Het verloop van deze kans over het tijdsinterval wordt dan gevisualiseerd met een *overlevingsfunctie*. Een voorbeeld van een overlevingsfunctie is gegeven in Figuur 1; deze toont aan dat ongeveer 60% van de personen binnen 12 maanden een baan heeft gevonden.

In dit hoofdstuk worden de verschillen tussen de twee combinaties van methoden beschreven en voor een aantal situaties wordt besproken welke combinatie het meest geschikt is.



Figuur 1: *Voorbeeld van een overlevingsfunctie welke de geschatte kans om tot een bepaald moment werkloos te blijven aangeeft. Deze kans is bijvoorbeeld 40% na 12 maanden wat aangeeft dat 60% van de werkzoekende een baan vindt binnen die tijd.*

Hoofdstuk 3

De motivatie achter de eerste twee hoofdstukken van dit proefschrift waren vragen van onderzoekers en de geleverde datasets bevatte verschillende typen variabelen. Sommige variabelen, zoals leeftijd, zijn *numerieke* variabelen gemeten op een continue schaal. Andere variabelen zijn *categorisch* en duiden een categorieniveau aan. Dit type data kan verder worden uitgesplitst naar *nominale* categorische data waarvan de categorieniveaus geen rangorde hebben, zoals gender, of *ordinale* categorische data welke wel een rangorde hebben, zoals opleidingsniveau of de mate waarin een persoon instemt met een stelling (*zeer mee oneens*, *mee oneens*, *niet eens/niet oneens*, *mee eens* en *zeer mee eens*).

Voordat er enige berekeningen kunnen worden gedaan op categorische data moeten deze worden omgezet in numerieke waarden. In andere woorden, elke van de categorieniveaus moet worden *gekwantificeerd*. De resulterende numerieke waarden worden ook wel *kwantificaties* genoemd. In hoofdstuk 2 werden de categorieniveaus bijvoorbeeld vervangen door de gehele getallen 1–5 zodat het gewogen gemiddelde berekend kon worden. Echter, gehele getallen zijn mogelijk geen goede vervangers. Het optimaal kwantificeren van categorische data is geen eenvoudig proces en er zijn meerdere kwantificatiestrategieën beschikbaar.

In de overlevingsanalyse worden twee verschillende kwantificatiestrategieën toegepast. De eerste optie is om elk categorieniveau te vervangen door een nieuwe variabele die de waarde 0 of 1 aanneemt en de tweede optie is om de niveaus te vervangen door gehele getallen, zoals werd gedaan in hoofdstuk 2.

Wanneer de eerste methode wordt toegepast krijgt iedere persoon in de dataset een waarde voor elke nieuwe variabele die een categorieniveau vertegenwoordigt. Deze waarde is 1 als deze persoon onder dat specifieke niveau valt, en zo niet dan is de waarde 0. Als alle categorieniveaus op deze manier worden gecodeerd krijgt iedere persoon de waarde 0 voor elke nieuwe variabele behalve degene die zijn of haar categorie vertegenwoordigt. Met deze strategie worden dus vele extra variabelen aangemaakt. Het effect van elk van deze variabelen op de overlevingskans wordt daarna individueel berekend zonder daarbij rekening te houden met de andere categorieniveaus. Wanneer deze effecten op deze manier individueel worden berekend, dan kan het zijn dat een eventuele rangorde van de categorieën verloren gaat. Dit kan een probleem zijn wanneer de rangorde van ordinale categorieën behouden moet blijven. Als bijvoorbeeld de depressie van een patiënt gemeten wordt op een schaal van *laag*, *medium* naar *hoog* dan is het waarschijnlijk dat de patiënten in de laagste groep ook de kortste hersteltijd hebben, en vice versa. Als de effecten individueel worden berekend, dan kan het gebeuren dat deze rangorde niet teruggevonden wordt.

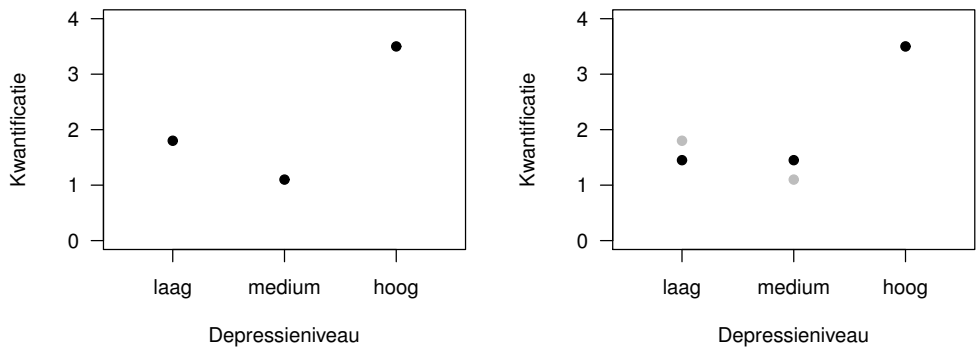
De tweede optie, waarbij gehele getallen worden gebruikt voor de kwantificering, wordt vaak toegepast op ordinale data omdat het de rangorde van de categorieën behoudt. De correcte rangorde wordt namelijk afgedwongen door de niveaus te vervangen voor oplopende gehele getallen (*laag* = 1, *medium* = 2, *hoog* = 3). Echter introduceren deze kwantificaties ook gelijke afstanden tussen de opeenvolgende niveaus wat mogelijk niet geschikt is voor de data.

Een andere kwantificatiestrategie is *optimale-schaling* welke als doel heeft om optimale kwantificaties te vinden voor elk categorieniveau en tegelijkertijd het effect van elke variabele op de uitkomst schat. Deze strategie is voorheen al in een aantal statistische methoden geïmplementeerd, zoals in optimale-schaling principale componentenanalyse (hoofdstuk 2).

De optimale-schalingsmethode begint met het vinden van kwantificaties voor elk categorieniveau zonder daarbij rekening te houden met de andere niveaus. Dit kan dus resulteren in kwantificaties die een andere rangorde hebben dan de originele categorieniveaus. Dit wordt het *nominale-schalingniveau* genoemd. In Figuur 2a wordt een voorbeeld van de uitkomst van deze stap weergegeven voor de depressieschaal. In dit voorbeeld is de rangorde onjuist omdat patiënten met een laag depressieniveau een hogere kwantificatie hebben gekregen dan patiënten met een medium niveau.

Als de rangorde van de categorieniveaus behouden moet blijven dan gaat het algoritme verder met de volgende stap waarin kwantificaties worden berekend voor het *ordinale-schalingniveau*. In deze stap worden de nominale kwantificaties die in een dalende volgorde liggen vervangen voor waarden die niet dalen. Het resultaat voor de depressieschaal is weergegeven in Figuur 2b. De eerste twee categorieën zijn in de verkeerde volgorde en worden vervangen door hun gemiddelde. Zij

krijgen dezelfde kwantificatie waardoor alle kwantificaties in een niet-dalende volgorde liggen, zodat de originele rangorde behouden blijft.



(a) *Nominale-schalingniveau* **(b)** *Ordinale-scalingniveau*

Figuur 2: *Voorbeelden van* nominale *en* ordinale *kwantificaties voor de depressieschaal van* laag, medium *naar* hoog. *De nominale kwantificaties uit figuur (a) zijn als grijze stippen weergegeven in figuur (b).*

In het derde hoofdstuk wordt de optimale-schalingmethode geïmplementeerd in overlevingsanalyse en dit algoritme wordt in detail uitgelegd. Verder wordt er een simulatiestudie gedaan om de prestaties van deze methode te vergelijken met de twee andere kwantificatiestrategieën. Hiervoor werd overlevingsdata gesimuleerd met een ordinale categorische variabele.

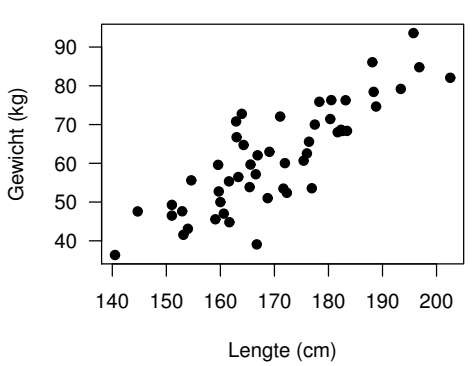
Uit deze simulatiestudie werd geconcludeerd dat de eerste strategie waarin de effecten individueel worden geschat goed presteert voor ordinale data, tenzij de correcte rangorde van de categorieniveaus lastig is op te sporen. Als de rangorde overduidelijk is of makkelijk te vinden is door de grote hoeveelheid data dan kan deze methode de juiste volgorde vinden. In andere situaties kan het gunstig zijn om de juiste rangorde af te dwingen met de optimale-schalingmethode.

Terwijl de eerste strategie dus prima presteert in een aantal situaties, blijkt uit de simulatiestudie dat het gebruik van gehele getallen in veel gevallen een slechte keuze is. Omdat het gelijke afstanden creëert tussen opeenvolgende niveaus wordt het effect van alle categorieniveaus namelijk gemiddeld. Daardoor zal het resultaat eventuele verschillen in effect tussen de niveaus verbergen.

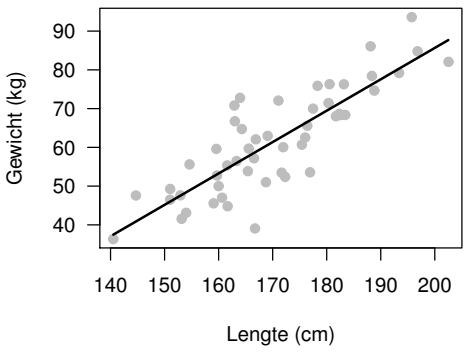
Dus het gebruik van gehele getallen lijkt een slechte strategie te zijn en ofwel nieuwe variabelen kunnen worden geïntroduceerd voor elk categorieniveau ofwel optimale-schaling kan worden toegepast om een beter inzicht te krijgen in de verschillende effecten van categorieniveaus op de overlevingskans.

Hoofdstuk 4

De eerste drie hoofdstukken van dit proefschrift waren gericht op modellen om overlevingsdata te analyseren. Er zijn echter veel meer soorten modellen om data te analyseren. Een voorbeeld is het veelgebruikte *lineaire regressiemodel*. Dit model beschrijft de relatie tussen een groep variabelen en een numerieke uitkomst. Er kan bijvoorbeeld data zijn verzameld over de lengte en het gewicht van tienermeisjes om te onderzoeken hoe het gewicht van een meisje kan worden voorspeld op basis van haar lengte. Dit datavoorbeeld is weergegeven in Figuur 3a waarin elke punt de combinatie van lengte en gewicht van één meisje weergeeft. Het langste meisje in de dataset is bijvoorbeeld iets langer dan twee meter en weegt ongeveer 80 kilogram. Het lineaire regressiemodel zal de relatie tussen de lengte en het gewicht van meisjes samenvatten door een *lineaire* (rechte) lijn te vinden die de geobserveerde data zo goed mogelijk beschrijft, zoals weergegeven in Figuur 3b.



(a) *Geobserveerde combinaties van gewicht en lengte van tienermeisjes*

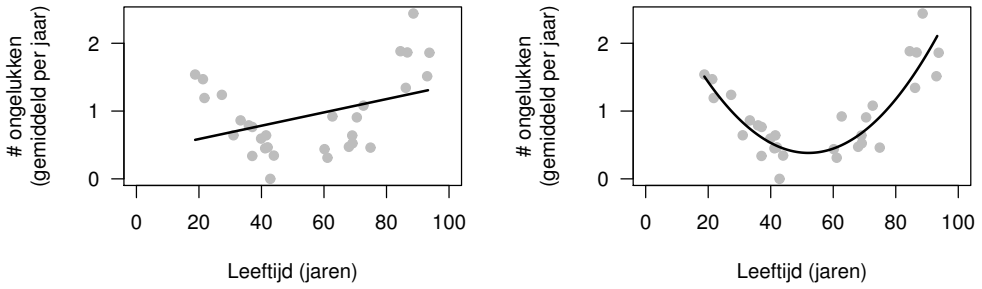


(b) *Geschatte rechte lijn die de geobserveerde data (grijze stippen) het beste beschrijft*

Figuur 3: *Voorbeeld van lineaire regressie op data.*

Ook al geeft deze lijn een goed beeld van de relatie tussen lengte en gewicht, is een rechte lijn niet geschikt om alle mogelijke relaties te beschrijven. Een typisch voorbeeld hiervan is de relatie tussen de leeftijd van een bestuurder en het aantal ongelukken dat hij of zij veroorzaakt. In het algemeen veroorzaken jongeren en ouderen meer ongelukken dan mensen in de leeftijdsgroep daartussen. Hierdoor heeft deze relatie een u-vorm welke niet beschreven kan worden met standaard lineaire regressie. De rechte lijn die het beste bij deze data zou passen zou namelijk het aantal ongelukken veroorzaakt door jonge en oude bestuurders onderschatten en het aantal voor de leeftijdscategorie daartussen overschatten

(zie Figuur 4a). Een meer flexibel model is daarom nodig voor deze *niet-lineaire* relatie zodat de u-vorm correct beschreven wordt (zie Figuur 4b).



(a) *Geschatte lineaire relatie* **(b)** *Geschatte niet-lineaire relatie*

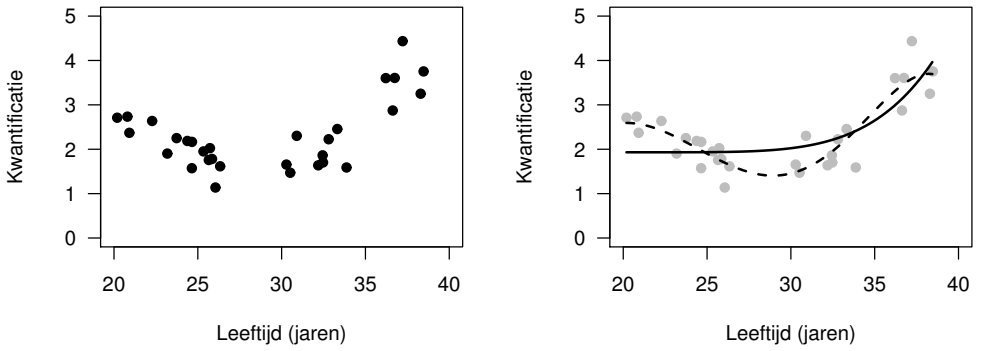
Figuur 4: *Voorbeelden van lineaire regressie en niet-lineaire regressie op data over de leeftijd van een bestuurder en het aantal ongelukken dat hij of zij veroorzaakt (grijze stippen).*

De *optimale-schalingmethode* beschreven in hoofdstuk 3 is uitgebreid zodat het ook numerieke data kan kwantificeren. Met deze uitbreiding kan de methode gebruikt worden om complexere relaties tussen variabelen te modelleren. In het algoritme wordt elke geobserveerde unieke numerieke waarde gezien als een aparte categorie. Zoals bij categorische data start de methode met het individueel schatten van het effect van elke van deze categorieën. Zo worden in deze stap kwantificaties berekend voor iedere numerieke waarde in de originele dataset. Als bijvoorbeeld de leeftijd van patiënten is verzameld en de eerste stap van de optimale-schalingmethode wordt toegepast op de geobserveerde leeftijden, dan kan dit de kwantificaties opleveren die in Figuur 5a worden weergegeven.

Aangezien niet alle mogelijke leeftijden tussen 20 en 40 jaar zijn geobserveerd zijn er alleen kwantificaties berekend voor een beperkt aantal leeftijden in dit interval. Er zijn bijvoorbeeld geen schattingen voor 27 tot 30 jaar. Om dit gat te dichten zal de optimale-schalingmethode een gladde continue functie tekenen door de nominale kwantificaties. Deze functie kan ofwel stijgend zijn of zowel stijgend als dalend, zoals weergegeven door de ononderbroken lijn en de stippellijn in Figuur 5b. Deze functie geeft dan ook kwantificaties voor de leeftijden die niet waren geobserveerd.

De optimale-schalingmethode om zowel numerieke als categorische data optimaal te kwantificeren is al toegepast in het lineaire regressiemodel om niet-lineaire relaties te modelleren. Dit lineaire regressiemodel kan echter alleen gebruikt worden voor numerieke uitkomsten zoals lichaamsgewicht en aantal ongelukken. Voor andere soorten uitkomsten worden vaak gegeneraliseerde versies van het lineaire model gebruikt, genaamd *gegeneraliseerde lineaire modellen*. Het logisti-

sche model dat in hoofdstuk 2 werd gebruikt om de kans te berekenen dat iemand na één jaar een baan had gevonden en is een voorbeeld van een gegeneraliseerd lineair model. Voor dit model zijn namelijk maar twee uitkomstopties, want iemand vindt binnen één jaar een nieuwe baan of niet.



(a) *Nominale-schalingniveau* **(b)** *Continue-schalingniveaus*

Figuur 5: *Voorbeelden van gladde continue kwantificatiefuncties, ofwel stijgend (ononderbroken lijn) ofwel zowel stijgend als dalend (stippellijn), op numerieke data.*

In het vierde hoofdstuk van dit proefschrift wordt beschreven hoe de optimale-schalingsmethode kan worden geïmplementeerd in gegeneraliseerde lineaire modellen. Speciale aandacht wordt gegeven aan de implementatie in het logistische regressiemodel en deze combinatie wordt toegepast op drie verschillende datasets om te laten zien hoe de methode gebruikt kan worden en welke voordelen het heeft.

Hoofdstuk 5

Nu er steeds meer data worden verzameld worden statistische resultaten ook steeds vaker gebruikt in besluitvorming. Een belangrijke taak voor statistici is daardoor om hun bevindingen begrijpelijk te communiceren. Als de resultaten verkeerd worden begrepen dan is al het werk gestoken in het verzamelen van de data, het onwikkelen van de analysemethoden en het toepassen daarvan voor niets geweest.

Omdat statistische modellen vaak worden gebruikt om voorspellingen te maken is het duidelijk communiceren van de geschatte kansen belangrijk. Eerder onderzoek heeft aangetoond dat personen wie deze kansen communiceren dat liever verbaal doen met kanswoorden zoals *onwaarschijnlijk*, *meestal* en *misschien*,

omdat deze uitspraken ook een zekere mate van onzekerheid overbrengen. Deze voorkeur geeft aan dat er een vertaalstap nodig is van de geschatte numerieke kans naar een passende uitdrukking. In sommige gevallen worden kansschalen gebruikt om deze vertaalstap te standaardiseren. Een kansschaal geeft bijvoorbeeld aan dat de uitdrukking *zeer waarschijnlijk* gebruikt moet worden voor kansen tussen 90% en 95% en *extreem waarschijnlijk* voor kansen tussen 95% en 99%. Dit soort schalen zijn meestal ook symmetrisch. Dus, als *zeer waarschijnlijk* 90–95% uitdrukt, dan duidt *zeer onwaarschijnlijk* op 5–10%.

Er is al uitgebreid onderzoek gedaan naar de interpretatie van Engelse verbale kansuitdrukkingen. Dit onderzoek toonde aan dat er tussen mensen grote verschillen zijn in hun interpretatie van deze uitdrukkingen en dat deze vaak asymmetrisch is. Bijvoorbeeld, de geïnterpreterde kansen van gespiegelde kanswoorden zoals *waarschijnlijk* en *onwaarschijnlijk* tellen gemiddeld niet op tot 100%. Verder wordt de interpretatie ook beïnvloed door de aanvankelijke verwachting over een uitspraak waarin een kansuitdrukking geplaatst wordt. De numerieke interpretatie van het woord *waarschijnlijk* in de uitspraak *"Het is waarschijnlijk dat het in juni zal regenen in Manchester, Engeland"* is bijvoorbeeld meestal hoger dan in de zin *"Het is waarschijnlijk dat het in juni zal regenen in Barcelona, Spanje"*.

Deze onderzoekresultaten geven aan dat het onmogelijk is om verbale kansuitdrukkingen samen te vatten in een (symmetrische) kansschaal op een manier zodat iedereen het eens is met deze schaal. Toch gebruiken veel organisaties dit soort schalen.

Ook al is er uitgebreid onderzoek gedaan naar Engelse uitspraken, is het belangrijk om deze ook in andere talen te bestuderen. Veel internationale organisaties publiceren documenten in meer dan één taal en de betekenis van de verbale kansuitdrukkingen zou verloren kunnen gaan bij de vertaling.

In het vijfde en laatste hoofdstuk van dit proefschrift word een onderzoeksproject naar de interpretatie van Nederlandse uitdrukkingen beschreven. In de studie werden zowel kansuitdrukkingen zoals *waarschijnlijk* en *misschien* bestudeerd als uitdrukkingen over frequencies zoals *soms* en *doorgaans*.

De onderzoeksresultaten zijn erg vergelijkbaar met de conclusies uit de Engelse studies. Er werden namelijk ook grote verschillen gevonden in de interpretatie van de kanswoorden en ook de asymmetrie bij gespiegelde uitspraken werd waargenomen. Dit toont aan dat ook Nederlandstalige kansschalen niet door iedereen zouden worden ondersteund.

In dit onderzoek werden ook de interpretaties van leken vergeleken met die van statistici en er werden geen structurele verschillen gevonden tussen die groepen. Zelfs de personen wie vaak kansen schatten en communiceren zijn het onderling niet eens over de interpretatie van de kansuitdrukkingen. Uit de resultaten

van deze studie werd geconcludeerd dat het gebruik van kansuitdrukkingen om geschatte kansen over te brengen een risico op miscommuncatie geeft.