Optimale Schaling van Interactie-effecten in een Model voor "Tellingen"

S. den Daas - 329225

1 juli 2013

Abstract

Interactie-effecten tussen categorische variabelen komen in allerlei toepassingen voor. Zodra er veel predictorvariabelen in een model opgenomen worden, is het moeilijk om alle interactie-effecten correct te schatten en te interpreteren. Een model voor de eenvoudige representatie van interactie-effecten en het terugbrengen van dimensionaliteit is het Optimaal Schalen van Interacties (OSI)-model. Het doel van het OSI-model is gemakkelijk inzicht te krijgen in de grootte en de vorm van de interactie-effecten. Aan de hand van een dataset omtrent delicten wordt aangetoond dat het OSI-model zinvol in de praktijk kan worden toegepast in een model voor "tellingen".

Trefwoorden: *interactie-effecten, optimaal schalen, gegeneraliseerde lineaire modellen, Poisson verdeling, tellingen, OSI-model, parameterisatie*

Inhoudsopgave

1 Inleiding

In veel toepassingen worden interactie-effecten verwacht tussen de verklarende variabelen. Zo kan een marketingmix, bijvoorbeeld een mix van prijsverlaging en display, gezamenlijk een grotere impact hebben dan enkel prijsverlaging of display op het aantal verkopen. Om het aantal verkopen goed in te schatten, is het noodzakelijk deze interactie-effecten ook op te nemen in het model.

Als er echter veel variabelen zijn loopt het aantal interacties al snel op. Hierdoor wordt het aantal te schatten parameters erg hoog, maar ook de interpretatie wordt onnodig ingewikkeld. Een model met minder parameters en betere interpretatiemogelijkheden beperkt het aantal interpretatiefouten en zorgt ervoor dat er minder tijd nodig is om de uitkomsten te interpreteren.

Van Rosmalen, Koning & Groenen (2009) hebben een model ontwikkeld die voorgaand probleem aanpakt. Hierbij maken zij gebruik van optimale schaling om de categorische variabelen te transformeren naar continue variabelen. Het doel van het 'Optimale Schaling van Interactie'-model (OSI) is het reduceren van de dimensionaliteit om het interpreteren van de data te vergemakkelijken en grafisch weer te geven. Eerdere onderzoeken gaan in op het bepalen van interactie-effecten, maar houden geen rekening met het exponentieel groeiend aantal parameters. Zo zijn Goodman's (1981) RC(M)-associatiemodellen een handige methode om interactie-effecten te analyseren. Goodman (1981) laat zien hoe de modellen gebruikt kunnen worden om de geschatte componenten voor interactie-effecten te verkrijgen. Echter, wordt geen rekening gehouden met de hoeveelheid parameters. Het OSI-model daarentegen houdt hier wel rekening mee. Dit model bestaat uit hoofdeffecten, kwantificaties en schaalfactoren. Hierbij zijn de hoofdeffecten het op zichzelf staande effect van de predictorvariabelen op de responsvariabele. De kwantificaties zijn verkregen met behulp van optimale schaling en kunnen gebruikt worden om de vorm van de interacties tussen predictorvariabelen te bepalen. Optimaal verwijst naar het feit dat deze kwantificaties gekozen zijn zodanig dat ze helpen om de categorieën van categorische variabelen een optimaal numerieke waarde toe te kennen. Hierbij wordt enkel ingegaan op de tweewegs interactie tussen de variabelen. Tot slot bepalen de schaalfactoren de grootte van de interacties tussen de predictorvariabelen. Deze orde van grootte wordt mede bepaald door de orde van grootte van de betreffende responsvariabele. Het OSI-model wordt toegepast in gegeneraliseerde lineaire modellen en kan gebruikt worden om interactie-effecten in een Poisson model te schatten. Het OSI-model zal het uitgangspunt vormen van dit onderzoek.

Zoals hierboven beschreven wordt het OSI-model toegepast op gegeneraliseerde lineaire modellen, nader GLM te noemen. Nelder & Wedderburn (1972) omschrijven in hun artikel een generalisatie van variantie-analyse, gebruikmakend van de log-likelihoods. Zij passen dit toe op vier verdelingen: de normale, binomiale, Poisson en gamma verdeling. In dit onderzoek wordt met name ingegaan op het GLM op de Poisson verdeling. Een lineair model bestaat uit systematische en willekeurige componenten. GLM combineert deze aspecten door allereerst een afhankelijk variabele, bijbehorende predictorvariabelen en een voorspelling met enkel systematiek te definiëren. Tot slot wordt een linkfunctie geïntroduceerd om de afhankelijke variabele en de voorspelling te koppelen. De parameters in de modellen schatten Nelder & Wedderburn aan de hand van het maximaliseren van de log-likelihood. Deze methode wordt verder toegelicht in sectie 2.3.

Interactie-effecten zijn ook terug te vinden bij conjunctanalyse. Een dergelijke analyse meet de preferenties van consumenten. Deze worden echter zelden meegenomen, aldus Van Rosmalen, Koning & Groenen (2009). Een reden hiervoor is dat het modelleren van alle interactie-effecten ook veel parameters vereist en er veel keuzesets per respondent bestaan. Daarom introduceren Van

Rosmalen et al. een nieuwe aanpak voor het modelleren van de tweeweginteracties gebaseerd op het hierboven besproken OSI-model. Deze aanpak is toe te passen op categorische variabelen, maar ook op continue producteigenschappen. Het doel is om de tweeweginteractie-effecten zo goed mogelijk te schatten door het aantal te schatten parameters zoveel mogelijk te beperken. In het onderzoek laten zij zien hoe conjunctanalyses geconstrueerd kunnen worden op basis van het speciaal voor conjuctanalyse ontworpen Conjoint-OSI-model.

Ook Anderson & Vermunt (2000) presenteren in hun artikel een manier om interactie-effecten te modelleren in een log-lineaire analyse met meer dan drie variabelen. Dit kunnen ze modelleren door aan te nemen dat er latente variabelen aanwezig zijn. Volgens de onderzoekers is het een krachtige en flexibele benadering voor de relaties tusssen nominale en/of ordinale variabelen in termen van latente continue variabelen. De geobserveerde variabelen zijn namelijk indicatoren van verschillende latente variabelen. Zodra de latente variabelen gecorreleerd zijn, wordt het moeilijk om de identificatie beperkingen te bepalen. Het model dat zij presenteren verschilt van factoranalyse in termen van de marginale verdeling. Bij factoranalyse wordt deze namelijk gegeven door multivariate normale verdeling, terwijl in het model van Anderson & Vermunt deze gegeven wordt door een mix van multivariate normale verdelingen. Dit model is een alternatief voor een traditioneel factoranalyse model. Echter verschaft het model van Anderson & Vermunt (2000) de mogelijkheid om, naast het reduceren van het aantal te schatten parameters, een multidimensionale schaling toe te passen op de verklarende variabelen. Welke methode het best toegepast kan worden, hangt af van de applicatie. Zo vereist het gebruik van het OSI-model geen aanname over de latente variabele, maar maakt daarentegen gebruik van optimale schaling van de kwantificaties.

In dit onderzoek wordt onderzocht hoe een dergelijke model opgezet kan worden in een model voor “tellingen”: Hoe kunnen interacties tussen categorische variabelen worden opgenomen in een model voor “tellingen”zonder extreem veel parameters te introduceren?

Om dit te onderzoeken zal gebruik gemaakt worden van een OSI-model voor een eenvoudige representatie van interactie-effecten. In dit onderzoek zal deze methode worden toegepast op een Poisson model; een model voor “tellingen”van gebeurtenissen.

2 Methoden

In deze sectie zal naar voren komen welke dataset wordt gebruikt in dit onderzoek. Daarnaast zal in worden gegaan op de Poisson verdeling, gegeneraliseerde modellen en optimale schaling om tot een methode te komen voor een OSI-model voor tellingen.

2.1 Data

Aan de hand van enquêtedata over criminaliteit zal dit onderzoek moeten gaan uitwijzen dat het OSI-model zinvol toegepast kan worden bij data met tellingen, ook wel ‘count data’ genoemd. In dit onderzoek tellen we hoe vaak individuen het slachtoffer zijn van criminaliteit. Dit aantal gaat verklaard worden aan de hand van hoofd- en interactie-effecten van categorische variabelen, zoals leeftijd, opleiding, stedelijkheid van de omgeving en risicofactoren.

In Tabel 1 zijn de kenmerken van de omgeving te vinden. Daarnaast hebben de respondenten aangegeven wat hun risicofactor is om slachtoffer te worden van een delict. Deze gegevens zijn in

Tabel 1: Kenmerken Omgeving

(a) Stedelijkheid

	Aantal	%
Niet stedelijk	474	16,1
Redelijk stedelijk	738	25,0
Stedelijk	543	18,4
Sterk stedelijk	665	22,5
Zeer stedelijk	531	18,0

(b) Staat van huizen

	Aantal	%
Zeer goede staat	800	27,2
Goede staat	1640	55,7
Neutraal	276	9,4
Slecht staat	170	5,8
Zeer slechte staat	56	1,9
Geen antwoord	9	

(c) Mate van lastiggevallen

	Aantal	%
Zeer frequent	29	1,0
Frequent	176	6,1
Neutraal	367	12,7
Niet Frequent	1652	57,2
Vrijwel nooit	662	22,9
Geen antwoord	65	

(d) Mate van criminaliteit

	Aantal	%
Zeer frequent	82	2,8
Frequent	338	11,6
Neutraal	676	23,2
Niet Frequent	1368	47,0
Vrijwel nooit	444	15,3
Geen antwoord	43	

(e) Mate van sociale interactie

	Aantal	%
Zeer frequent	443	15,1
Frequent	1365	46,6
Neutraal	652	22,3
Niet Frequent	407	13,9
Vrijwel nooit	62	2,1
Geen antwoord	22	

(f) Mate van drugsgebruik en -verhandeling

	Aantal	%
Zeer frequent	158	21,9
Frequent	502	39,1
Neutraal	389	14,5
Niet Frequent	1051	18,7
Vrijwel nooit	588	5,9
Geen antwoord	263	

Tabel 2: Delicten

(a) Frequenties

Frequentie	Seksueel misbruik	%	Zedendelicten	%	Aanrandingen	%
Nooit	2458	83.2	2502	84.7	2395	81.1
1	43	1.5	252	8.5	363	12.3
2	321	10.9	72	2.4	72	2.4
3	82	2.8	36	1.2	22	0.7
4	24	0.8	7	0.2	15	0.5
5	17	0.6	10	0.3	14	0.5
> 5	8	0.3	74	2.5	72	2.4

(b) Inschatting Risico

Risicofactor (%)	Seksueel Misbruik
Zeer klein	65,3
Klein	24,0
Neutraal	9,8
Groot	0,8
Zeer groot	0,1

Tabel 3: Kenmerken Respondenten

(a) Leeftijd

	Aantal	%
15-25	844	28,6
25-35	814	27,6
35-45	419	14,2
45-55	323	10,9
55-65	251	8,5
65-75	197	6,7
> 75	103	3,5

(b) Geslacht

	Aantal	%
Man	1315	44,6
Vrouw	1636	55,4

(c) Thuissituatie

	Aantal	%
Met partner	1587	66,0
Alleenstaand	75	3,1
Thuiswonend	705	29,3
Anders	37	1,5
Ontbreekt	547	

(d) Hoogst genoten opleiding

	Aantal	%
Basisonderwijs	230	7,8
Lager onderwijs	864	29,2
Middelbaar onderwijs	912	30,9
Hoger onderwijs	681	23,1
Universiteit	264	8,9

(e) Werkstatus

	Aantal	%
Betaalde baan	1389	47,3
Arbeidsongeschikt	68	2,3
Gepensioneerd	201	6,8
Werkloos	100	3,4
Huishouden	561	19,0
Scholier/student	573	19,4
Anders	57	2,0

(f) Inkomen

guldens pm	Aantal	%
<1000	112	4,7
1.001-2.000	400	16,6
2.001-3.000	614	25,6
3.001-4.000	506	21,0
4.001-5.000	419	17,4
>5.000	350	14,6
Geen antwoord	550	

Tabel 2b te vinden. Verder is in Tabel 2a te vinden hoe het aantal delicten zich spreidt onder de respondenten. Tot slot zijn in Tabel 3 eigenschappen van de respondenten te vinden.

2.2 Poisson Verdeling

Een Poisson verdeling is een discrete kansverdeling welke de kans uitdrukt op een gegeven aantal gebeurtenissen (y) in een vast tijdsinterval: $p_y(\lambda) = (e^{-\lambda})\lambda^y/y!$. Hierbij is het belangrijk dat het gemiddeld aantal gebeurtenissen (λ) in dit tijdsinterval bekend is en het aantal gebeurtenissen onafhankelijk is van het aantal gebeurtenissen in het vorige tijdsinterval. De kans op precies y successen in een binomiale kansexperiment met een kans r op succes, welke n maal wordt uitgevoerd, convergeert naar een Poisson verdeling zodra n voldoende groot is en de kans r voldoende klein (Poisson,1837; Haight, 1967). Zodra n concergeert naar oneindig wordt de kans r gekozen zodanig dat $n \cdot r$ convergeert naar λ. Dit betekent dat een Poisson verdeling gebruikt kan worden bij een grote steekproef, waarbij de kans op een gebeurtenis klein is. Zo kan deze verdeling gebruikt worden om bijvoorbeeld het aantal ongelukken, natuurrampen of aankomsten in een bepaald tijdsinterval te modelleren. De variantie van de Poisson verdeling is gelijk aan het gemiddelde en wordt dus gegeven door λ. Daarnaast kunnen we respectievelijk de likelihood en de log-likelihood definiëren als:

$$Ł(\lambda; y_1, ..., y_n) = \Pi_{i=1}^{n}(e^{-\lambda}\frac{\lambda^{y_i}}{y_i!}) = \frac{e^{-n\lambda}\lambda^{y_1+...+y_n}}{y_1!\ ...y_n!}. \tag{1}$$

$$LogL(\lambda; y_1, ..., y_n) = -n\lambda + \sum_i (y_i \log(\lambda) - \log(y_1!\ ...y_n!) \tag{2}$$

Deze $LogL$ zal worden gebruikt bij het ontwikkelen van een OSI-model voor count data. Sectie 2.5 zal hier verder op ingaan.

2.3 GLM

GLM maakt het mogelijk om complexe systematische lineaire modellen samen te stellen. De methode geeft een consistente manier om de systematische en willekeurig componenten te linken. Zo valt de discussie weg hoe ver data getransformeerd kan worden, omdat het model twee verschillende transformaties toelaat. De eenvoudigheid van het lineaire model blijft nog steeds bestaan (Nelder & Wedderburn, 1972). De lineairiteit van de systematische componenten wordt namelijk eerst toegelaten in het GLM. Vervolgens kan de gewenste verdeling van de waarnemingen worden toegelaten in hetzelfde GLM. De karaktereigenschappen van een dergelijk GLM zijn (Nelder & Wedderburn, 1972):

(i) Het model bevat een afhankelijke variabele y waarvan de verdeling met canonische parameter θ enkel het willekeurige component bevat. De dichtheidsfunctie van y wordt gegeven door:

$$\pi(y; \theta, \phi) = exp[\alpha(\phi)\{y\theta - g(\theta) + h(y)\} + \beta(\phi, y)] \tag{3}$$

Waarbij $\alpha(\phi) > 0$ de schaalfactor, zodanig dat voor een vaste ϕ een exponentiële familie ontstaat. ϕ is een dispersieparameter, waarmee de spreiding in de waarnemingen gemodelleerd kan worden.

(ii) Daarnaast bevat het model een set van predictorvariabelen $\mathbf{x}_1, ..., \mathbf{x}_m$ en een lineaire predictor $\eta = \sum_j \beta_j \mathbf{x}_j$, waarbij enkel het systematische effect uit het lineaire model is meegenomen.

Het is ook mogelijk om naast hoofdeffecten interactie-effecten mee te nemen in de lineaire predictor, $\eta = \sum_j \beta_j \mathbf{x}_j + \sum_j \sum_{k<j} \beta_{jk} \mathbf{x}_j \mathbf{x}_k$. Een model met alleen hoofdeffecten wordt een eerste orde GLM-model genoemd, zodra er interactie-effecten opgenomen worden is er sprake van een tweede orde GLM-model.

(iii) Tot slot is er een functie nodig om deze twee componenten te linken. De linkfunctie $\theta = f(\eta)$ verbindt de parameter θ van de verdeling van y met de lineaire predictor η.

Vervolgens wordt het logaritme van de likelihood gemaximaliseerd. Nelder & Wedderburn (1972) gebruiken de likelihood zodra θ en η overeenkomen. Hierdoor kunnen we het logaritme van de likelihood van de aselecte steekproef $y_1, \ ..., y_n$ schrijven als:

$$LogL = y\eta - g(\eta) + h(y) \tag{4}$$

Nelder & Wedderburn (1972) bewijzen dat het maximum van deze likelihoodfunctie gelijk is aan de oplossing van gewogen kleinste kwadraten.

2.4 OSI-model op GLM

Van Rosmalen, Koning & Groenen (2009) bewijzen dat een OSI-model toepassen op een GLM de dimensionaliteit reduceert, het interpreteren van de data vergemakkelijkt en het mogelijk maakt de bevindingen grafisch weer te geven. Zoals eerder vermeldt, bestaat het OSI-model uit hoofdeffecten, kwantificaties en schaalfactoren.

2.4.1 Modelomschrijving

Er wordt aangenomen dat de observaties $y_i, i = 1, \ ..., n$ onafhankelijk verdeeld zijn met $E(y_i) = \mu_i$. De verdeling wordt gegeven door (Van Rosmalen, Koning & Groenen, 2009):

$$f(y_i; \theta, \phi) = exp\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi), \tag{5}$$

Waarbij a(·), b(·), c(·) gegeven functies, θ de zogeheten natuurlijke parameter en ϕ de schaalparameter. Een linkfunctie h(·) relateert de lineaire predictor (η_i) aan de responsvariabele volgens $\eta_i = h(\mu_i)$. Veronderstel tot slot dat de continue predictorvariabelen ($\mathbf{x}_j, j = 1, \ ..., m$) bekend zijn. Dan kunnen de hoofd- en tweewegs interactie-effecten gegeven worden door (Van Rosmalen, Koning & Groenen,2009):

$$\eta_i = c + \sum_{j=1}^{m}(b_j x_{ij}) + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m}(w_{jl} s_{jl} x_{ij} x_{il}), \tag{6}$$

waar c de constante term, b_j het hoofdeffect van variabele $\mathbf{x}_j$ en s_{jl} de grootte van het interactie-effect tussen variabelen $\mathbf{x}_j$ en $\mathbf{x}_l$, ofwel de schaalfactor. Daarbij is w_{jl} gelijk aan 1 als het interactie-effect tussen de predictorvariabele j en l wordt meegenomen en 0 anders. Zodra men te maken heeft met categorische variabelen dienen deze eerst optimaal geschaald te worden. In het OSI-model worden de variabelen van de hoofd- en interactie-effecten apart van elkaar optimaal geschaald, zodanig dat (Van Rosmalen, Koning & Groenen,2009):

$$\eta_i = c + \sum_{j=1}^{m}(b_j r_{ij}) + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m}(w_{jl} s_{jl} q_{ij} q_{il}), \tag{7}$$

Hierbij is r_j de optimaal geschaalde variabele voor het hoofdeffect van predictorvariabele j en q_j is de gebruikte optimaal geschaalde variabele voor de interactie-effecten van variabele j. De waarde van deze continue optimaal geschaalde variabelen zijn niet bekend en moeten geschat worden. Zodra we r_j definiëren als $G_j a_j$, waarbij a_j een $K_j \times 1$ parameter vector die de kwantificaties bevat voor de hoofdeffecten van categorie j, en stellen dat $q_j = G_j y_j$, waarbij y_j een $K_j \times 1$ parameter vector die de kwantificaties bevat voor de interactie-effecten van variabele j, kunnen we in plaats van vergelijking (7) de volgende vergelijking schrijven (Van Rosmalen, Koning & Groenen,2009):

$$\eta_i = c + \sum_{j=1}^{m}(b_j g'_{ij} a_j) + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m}(w_{jl} s_{jl} g'_{ij} y_j y'_l g_{il}), \tag{8}$$

Zodra de kwantificaties a_j en y_j geschat zijn, kunnen de optimaal geschaalde variabelen behandeld worden zijnde continue variabelen. De overige parameters kunnen vervolgens worden uitgerekend volgens GLM.

2.4.2 Restricties

Echter zijn er wel beperkingen nodig op de parameters om het model te identificeren. Van Rosmalen, Koning & Groenen (2009) gebruiken beperkingen die volgen uit de 'optimale schaling'-methodologie. Zij leggen op dat de optimaal geschaalde variabelen r_j en q_j een gemiddelde van nul hebben en een variantie van één. Hierdoor ontstaan de volgende restricties (Van Rosmalen, Koning & Groenen, 2009):

$$1'q_j = \sum_{i=1}^{n} g'_{ij} y_j = 0 \qquad [locatie] \tag{9}$$

$$q'_j q_j = \sum_{i=1}^{n} g'_{ij} y_j y'_j g_{ij} = n \qquad [schaal] \tag{10}$$

$$1'r_j = \sum_{i=1}^{n} g'_{ij} a_j = 0 \tag{11}$$

$$r'_j r_j = \sum_{i=1}^{n} g'_{ij} a_j a'_j g_{ij} = n \tag{12}$$

Daarnaast is het niet mogelijk om $s_j l$ te schatten zodra het interactie-effect tussen variabelen j en l niet wordt opgenomen ($w_{jl} = 0$). Dus is het noodzakelijk de volgende restrictie toe te voegen:

$$s_{jl} = 0 \; als \; w_{jl} = 0 \tag{13}$$

Daarnaast kan het mogelijk zijn dat er nog extra beperkingen opgelegd dienen te worden, bijvoorbeeld als er weinig observaties beschikbaar zijn of de interactie-effecten voor weinig predictorvariabelen meegenomen worden. Of een dergelijke toevoeging van beperkingen nodig is, kan gecontroleerd worden door te kijken of de oplossing een unieke combinatie is die het logaritme van de likelihood maximaliseert door het algoritme te doorlopen met willekeurig gekozen startwaarden (Van Rosmalen, Koning & Groenen,2009).

2.4.3 Model Reparameterisatie

In deze paragraaf wordt een methode besproken waarbij de kwantificaties van de optimaal geschaalde variabelen in een één-dimensionaal OSI-model geparameteriseerd worden zonder bindende parameter restricties. Deze representatie is gebaseerd op het vervangen van de kwantificaties door functies van een nieuwe functie zodanig dat de resulterende optimaal geschaalde variabelen automatisch gemiddelde nul en variantie één hebben. De beperkingen die deze waarden opleggen, kunnen gegeven worden door (Van Rosmalen, Koning & Groenen, 2009):

$$\sum_{c=1}^{K_j}(f_c v_c) = 0 \qquad [gemiddelde = 0] \tag{14}$$

$$\sum_{c=1}^{K_j}(f_c v_c^2) = 1 \qquad [variantie = 1], \tag{15}$$

waarin v_c de kwantificaties van de optie c in categorie j en f_c de relatieve frequenties in categorie j. Hierbij wordt aangenomen dat de relatieve frequenties tot één optellen over de categorie, zodat $\sum_{k=1}^{K_j} f_k = 1$ met K_j het aantal categorieën in predictorvariabele j. Van Rosmalen et al. construeert een reparameterisatie $v_1,\ ..., v_{K_j}$ voor de parameters $x_1,\ ..., x_{K_j}$ gebaseerd op bovenstaande beperkingen, gebruikmakend van de kwantificatie parameters $\theta_1,\ ..., \theta_{K_{j-2}}$ in recursieve formules.

Als gevolg van dit recurrente verband split de parameterruimte in twee delen. Bij het meenemen van meerdere kwantificaties worden dit exponentieel veel delen. Daarom hebben Cleaver (2013) et al. gekozen voor een geometrische benadering om reparametrisatie waar te maken. Dit leidt ertoe dat de coördinatieparameters (ζ) uitgedrukt kunnen worden in poolcoördinaten van een bol. We kunnen de 'hoek'-parameters, ϕ, kiezen zodanig dat voldaan wordt aan:

$$\begin{aligned}
\zeta_1 &= \rho cos(\phi_1) \\
\zeta_2 &= \rho sin(\phi_1)cos(\phi_2) \\
&\vdots \qquad\qquad \vdots \\
\zeta_c &= \rho(\prod_{i=1}^{c-1} sin(\phi_i))cos(\phi_c) \\
&\vdots \qquad\qquad \vdots \\
\zeta_K - 2 &= \rho(\prod_{i=1}^{K-3} sin(\phi_i))cos(\phi_{K-2}) \\
\zeta_K - 1 &= \rho(\prod_{i=1}^{K-2} sin(\phi_i)),
\end{aligned} \tag{16}$$

hierbij wordt ϕ gegeven door (Cleaver, G., Donkers, B., Groenen, P.J.F., Koning, A.J. & Van Rosmalen, J., 2013):

$$\phi_c = \begin{cases} \pi\frac{e^{\theta_c}-1}{e^{\theta_c}+1} & als\ c = 1 \\ \frac{1}{2}\pi\frac{e^{\theta_c}-1}{e^{\theta_c}+1} & Anders \end{cases} \tag{17}$$

Door deze parameters op deze manier te kiezen kunnen we zeggen dat de som van kwadraten voor ζ gelijk zijn aan de gekwadrateerde straal ($\zeta'\zeta = \rho^2$).

2.4.4 Implementatie

Om het bovenstaande model te implementeren is een iteratief algoritme nodig om de optimale kwantificaties voor iedere categorie j te berekenen (Cleaver, G., Donkers, B., Groenen, P.J.F., Koning, A.J. & Van Rosmalen, J., 2010; 2013). Allereerst kunnen met behulp van de coördinatieparameters de herschaalde kwantificatieparameters uitgerekend worden. Vervolgens volgen hieruit de gewenste kwantificaties. Om de kwantificaties $v_1, ..., v_{K_j}$ voor gegeven kwantificatie parameters $\theta-1, ..., \theta_{K_j}$ van één van de categorische variabelen te berekenen, kan het volgende stappenplan gevolgd worden (Cleaver et al, 2010; 2013):

1. Initialiseer de frequenties (f_c) en het aantal opties in de categorische variabele (K_j)
2. Bereken S_c: $\sum_{l=1}^{c} f_l$ voor elke optie c in categorie j
3. Bereken ϕ volgens vergelijking (17)
4. Reken de hieruit volgende coördinatieparameters uit (ζ). Maak hierbij gebruik van (16)
5. Introduceer de expantiematrix $\mathbf{A}_{\ (K_j-1)\times(K_j-1)}$ met schalingsfactoren. Deze matrix wordt als volgt gedefinieerd:

$$\mathbf{A} = \begin{bmatrix} H_1 & H_2 & H_3 & \dots & H_{K_j-1} \\ -H_1S_1f_2 & H_2 & H_3 & \dots & H_{K_j-1} \\ 0 & -H_2S_2f_3 & H_3 & \dots & H_{K_j-1} \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & -H_{K_j-2}S_{K_j-2}f_{K_j-1} & H_{K_j-1} \\ 0 & \dots & 0 & 0 & -H_{K_j-1}S_{K_j-1}f_{K_j} \end{bmatrix} \tag{18}$$

   waarbij $H_c = \sqrt{\frac{f_{c+1}}{S_cS_{c+1}}}$
6. De herschaalde kwantificatieparameters volgen uit: $\xi = \mathbf{A}\zeta$. Deze hebben een gewogen som van nul en de gekwadrateerde lengte is gelijk aan de gekwadrateerde straal. Hierdoor is er een unieke manier om ζ naar ξ te transformeren, waarbij het gemiddelde nul is en de variantie één (Cleaver et. al)
7. Tot slot kunnen de kwantificaties voor elke optie c worden uitgerekend door de kwantificatieparameter voor c te delen door de wortel uit de frequentie van optie c: $v = \frac{\xi}{\sqrt{f}}$.

Dit stappenplan wordt gebruikt om de log-likelihood (4) te maximaliseren, gegeven de verbanden tussen deze kwantificaties, de kwantificatieparameters en deze log-likelihood.

2.5 OSI-model op Poisson verdeling

Vaak wordt gebruik gemaakt van een logaritmische linkfunctie als de standaardfouten Poisson verdeeld zijn (Van Rosmalen, Koning & Groenen, 2009). Het OSI-model volgt de volgende kansverdeling (19) (Van Rosmalen et al, 2009):

$$\log(f(y;\lambda)) = y\log(\lambda) - \lambda - \log(y!). \tag{19}$$

Hieruit volgt dat de natuurlijke parameter θ gelijk is aan $\log(\lambda)$, waardoor het logaritme van de kans op y met een bepaalde θ geschreven kan worden als:

$$\log(f(y;\theta)) = y\theta - e^{\theta} - \log(y!). \tag{20}$$

Zodra we kijken naar de verdeling van y voor individu i en de lineaire voorspeller η_i uit het GLM-model, kan deze als volgt geschreven worden:

$$\log(f(y_i;\eta_i)) = y_i\eta_i - e^{\eta_i} - \log(y_i!) \tag{21}$$

waarbij η_i gegeven is door:

$$\eta_i = c + \sum_j (b_j v_{ij}) + \sum_j \sum_{l>j} (b_{jl} v_{ij} v_{il}), \tag{22}$$

met c de constante, b_j het hoofdeffect en b_{jl} de grootte van het interactie-effect tussen de optimaal geschaalde variabelen $\mathbf{v}_j$ en $\mathbf{v}_l$. Hierbij is gekozen voor een lineaire predictor zonder kwadratische term. Doordat de log-likelihood van het Poisson model wordt gebruikt voor het vinden van optimale parameters in een model met tellingen, is het overbodig om een kromming op te nemen in de lineaire predictor. Vervolgens kan de log-likelihood, vergelijk met functie (2), gegeven worden door:

$$LogL(f(y_i;\eta_i)) = \sum_{i=1}^{n} \log(f(y_i;\eta_i)) = \sum_i (y_i\eta_i - e^{\eta_i} - \log(y_i)) \tag{23}$$

Allereerst kunnen de parameters optimaal geschaald worden aan de hand van de verhoudingen tussen de kwantificaties en de kwantificatieparameters beschreven in sectie 2.4.4. Hierbij nemen we $\rho = 1$. Uit dit algoritme worden de categorische variabelen vervangen door de optimaal geschaalde hoofdvariabelen (v_j voor $j = 1,\ ..., m$) en de interactie-variabelen ($v_j v_l$ voor $j = 1,\ ..., m,\ l = j+1,\ ..., m$) verkregen. Door deze te gebruiken in bovenstaande formule (21) kan de optimale λ_i voor elke individu i uit dit Poisson model berekend worden met behulp van het maximaliseren van de log-likelihood van deze verdeling (23). De hoofd- en respectievelijk de interactie-effecten tussen de categorische variabelen op het aantal delicten worden gegeven door $b_j v_j$ respectievelijk $b_{jk} v_j v_k$ uit de lineaire predictie vergelijking. Ook hier is het noodzakelijk om te kijken of er een uniek maximum ontstaat door verschillende startwaarden van het algoritme te vergelijken.

Om de verhouding van fit en complexiteit van het het model te bepalen wordt gebruik gemaakt van het deviance informatie criterium. Dit criterium wordt in een Poisson model gegeven door (Nelder & Wedderburn, 1972):

$$DIC = 2\{\sum(y\ ln(\frac{y}{\hat{\lambda}})) - \sum(y - \hat{\lambda})\} \tag{24}$$

Modellen met een kleinere DIC-waarde worden, hebben een hogere fit. Zo kan goed gekeken worden of de complexiteit van het toevoegen van een variabele afweegt tegen een betere fit.

3 Resultaten

In deze sectie zullen de resultaten besproken worden van het GLM-model en OSI-model op een Poisson verdeling. De resultaten zijn gebaseerd op het aantal aanrandingen en de categorische variabelen leeftijd, geslacht, thuissituatie en inkomen. Tevens zullen er voorbeeld interpretaties gegeven worden van enkele interactie-effecten. Tot slot worden de resultaten van de verschillende modellen vergeleken.

3.1 GLM op een Poisson verdeling

De invloed van het opnemen van interactie-effecten in een GLM-model kan het best onderzocht worden aan de hand van een vergelijking tussen een eerste en tweede orde GLM-model. Allereerst worden de resultaten van een eerste orde GLM-Model op een Poisson verdeling toegelicht. Vervolgens zullen de interactie-effecten in het tweede orde GLM-Model op een Poisson verdeling worden verduidelijkt.

3.1.1 Eerste orde GLM Model

Zodra we enkel de hoofdeffecten meenemen om de predictorvariabele η_i te schatten, resulteren hieruit de hoofdeffecten in Tabel 4 met bijbehorende standaardafwijkingen tussen haakjes gegeven.

Tabel 4: Hoofdeffecten - Eerste orde GLM

Constante												
-0,68	(0,084)											
Leeftijd												
15-25		25-35		35-45		45-55		55-65		65-75		Ouder dan 75	
$0,148^*$	(0,003)	$0,059$	(0,042)	$0,448^*$	(0,019)	$0,191^*$	(0,007)	$0,21^*$	(0,005)	$0,02^*$	(0,008)	$-0,207^*$	(0,025)
Geslacht													
Man		Vrouw											
$-0,781^*$	(0,064)	$0,461^*$	(0,011)										
Thuissituatie													
Met partner		Alleenstaand		Thuiswonend		Anders		Ontbreekt					
$-0,634^*$	(0,046)	$-0,568^*$	(0,047)	$0,057^*$	(0,022)	$0,435^*$	(0,014)	$0,083^*$	(0,107)				
Inkomen													
Minder dan 1000		1001-2000		2001-3000		3001-4000		4001-5000		Meer dan 5000			
$1,246^*$	(0,045)	$0,448^*$	(0,023)	$-0,086^*$	(0,009)	$-0,258^*$	(0,007)	$0,07^*$	(0,007)	$-0,148^*$	(0,028)		

**significant op 5%*

Uit deze tabel met hoofdeffecten is bijvoorbeeld af te lezen dat het zijn van een man respectievelijk een vrouw een significante negatieve respectievelijk positieve invloed heeft op het aantal aanrandingen. De kans om slachtoffer te worden van aanranding is dus kleiner voor mannen. Zo is ook te zien dat deze kans hoger is voor personen met een lagere leeftijd zijn en personen met een lager inkomen. De maximum log-likelihood van dit model is $370,398$. Daarnaast heeft dit model een DIC-waarde van $33079,654$.

3.1.2 Tweede orde GLM Model

Er is sprake van een unieke oplossing bij het maximaliseren van de log-likelihood in een GLM waar, naast de hoofdeffecten, ook interactie-effecten opgenomen worden. Door de grote hoeveelheid variabelen valt te betwijfelen hoe betrouwbaar de geschatte coëfficiënten zijn. Mede doordat de standaardafwijkingen van de coëfficiënten van de interactie-effecten niet gegeven kunnen worden. Door de extreme hoeveelheid te schatten parameters is het namelijk niet mogelijk om via het opgestelde GLM-algoritme de standaardafwijkingen te berekenen. Hierdoor is niet te achterhalen welke coëfficiënten significant afwijken van nul. Wat ons beperkt in het vergelijken van parameterwaardes

tussen een OSI-model en dit model.

Het tweede orde GLM model is in dit onderzoek door meerdere schattingsmethoden benaderd. Allereerst is met behulp van een standaard minimalisatie functie uit de 'MATLAB Toolbox' de negatieve log-likelihood geminimaliseerd. Door het optimaliseren van de coëfficiënten in de lineare predictor η. Hierbij wordt deze η voor elk individu i berekend door de juiste hoofd- en interactie-effecten te kiezen, overeenkomstig met de eigenschappen van dit individu. Zo worden voor elk individu slechts de relevante effecten meegenomen voor het schatten van de parameters. Ondanks dat voor elk effect een optimaal en positieve coëfficiënt gevonden wordt, is de standaardafwijking van de interactiecoëfficiënten niet bekend door de singulariteit tussen de opgenomen interactievariabelen. Ten tweede is de log-likelihood gemaximaliseerd aan de hand van het optimaliseren van η. Deze η geeft, samen met aangepaste predictorvariabelen voor elk effect, via de kleinste kwadraten methode de coëfficiënten voor de hoofd- en interactie-effecten. Om van elk effect de parameters te schatten is voor elk individu bepaald welke effecten van toepassing zijn. Op deze plaatsen in de aangepaste matrix van predictorvariabelen wordt de waarde van de eigenschap toegekend, op de andere plaatsen staat een nul. Als een individu bijvoorbeeld een vrouw tussen de 15-25 jaar is, zal het hoofdeffect op de plaats van de eigenschap 'vrouw' een 2 komen, op de plaats van '15-25 jaar' een 1 en van het interactie-effect tussen de eigenschappen 'vrouw' en '15-25' jaar een 2 (2×1). Door de singulaire matrix die hieruit onstaat worden de interactie-parameters op nul geschat en is het berekenen van de standaardafwijkingen wederom niet mogelijk. Daarnaast zijn beiden methoden ook uitgevoerd met dummy variabelen ter vervanging van de categorische variabelen en met verschillende startwaardes. Ook hier is singulariteit ontstaan. Kortom: geen van de schattingsmethoden heeft betrouwbare coëfficiënten van de interactie-effecten kunnen schatten.

Om toch een beeld te geven van de interpretatie van interactie-effecten in een tweede orde GLM-model zijn de coëfficiënten om de maximum likelihood ($2,72E+17$) te bereiken gegeven door de hoofdeffecten in Tabel 5 en interactie-effecten in Tabel 6. Dit model heeft een DIC-waarde van $5,43E+22$. Deze waarde is veel hoger dan de DIC-waarde van het eerste orde GLM-model. Het toevoegen van de interactie-effecten lijkt dus te zorgen voor een onnodig complex model; de grotere waarschijnlijkheid van het model weegt niet af tegen de grote hoeveelheid extra parameters.

Tabel 5: Hoofdeffecten - tweede orde GLM

Constante												
0,806	(1,648)											
Leeftijd												
15-25		25-35		35-45		45-55		55-65		65-75		Ouder dan 75	
$1,447^*$	(0,569)	$-0,105^*$	(0,050)	$-0,914^*$	(0,000)	$-2,204^*$	(0,000)	0,378	(1,048)	-1,645	(9,173)	$-2,869^*$	(0,595)
Geslacht												
Man		Vrouw										
$1,372^*$	(0,138)	-1,673	(3,284)									
Thuissituatie												
Met partner		Alleenstaand		Thuiswonend		Anders		Ontbreekt				
-0,375	(1,530)	$1,033^*$	(0,455)	$-0,845^*$	(0,001)	1,418	(1,098)	-1,362	(2,494)				
Inkomen												
< 1000		1001-2000		2001-3000		3001-4000		4001-5000		> 5000		Geen Antwoord	
$1,100^*$	(0,000)	0,994	(3,060)	-0,488	(1,645)	$-3,016^*$	(0,000)	$-1,015^*$	(0,015)	-0,454	(1,646)	$-1,561^*$	(0,001)

**significant op 5%*

Minder hoofdeffecten zijn significant van invloed op de afhankelijke variabele dan in het eerste orde GML-model. Daarnaast zijn de verschillen in coëfficiënten tussen het eerste en tweede orde GLM-model groot. Dit is te verklaren doordat de hoofdeffecten in het eerste orde GLM-model indirect ook de interactie-effecten meenemen. Wel kunnen we in de hoofdeffecten terugzien dat

Tabel 6: Interactie-effecten - tweede orde GLM

(a) Interactie tussen leeftijd en geslacht, thuissituatie en inkomen

	Geslacht		Thuissituatie					Inkomen						
Leeftijd	Man	Vrouw	Met partner	Alleenstaand	Thuiswonend	Anders	Ontbreekt	< 1000	1001-2000	2001-3000	3001-4000	4001-5000	> 5000	Geen Antwoord
15-25	-1,818	1,851	1,002	0,087	1,954	1,072	0,363	0,448	0,488	2,567	-1,744	-0,063	0,811	-2,256
25-35	1,705	1,53	-0,374	1,743	1,74	-0,029	1,452	2,071	0,516	0,481	0,314	-1,845	-1,309	-0,058
35-45	-0,42	-0,736	1,964	0,193	-1,089	-1,862	1,18	1,207	0,253	-2,898	0,779	-2,444	0,232	2,118
45-55	2,222	-1,704	1,396	-2,071	0,204	-0,03	0,497	0,504	2,345	-2,275	-2,808	-1,767	1,171	1,514
55-65	-0,653	1,466	-1,996	-2,078	-2,46	-2,114	1,892	-1,576	2,611	1,761	0,688	-2,121	-0,005	-1,378
65-75	1,615	2,261	2,155	-0,558	-2,33	-2,67	1,272	0,822	1,397	-0,762	-1,627	-1,866	0,215	-1,749
Ouder dan 75	-0,75	-1,044	2,939	1,221	-2,182	2,104	2,933	-2,499	0,489	1,662	-2,703	-2,744	-0,329	0,39

(b) Interactie tussen geslacht en thuissituatie en inkomen

	Thuissituatie					Inkomen						
Geslacht	Met partner	Alleenstaand	Thuiswonend	Anders	Ontbreekt	Minder dan 1000	1001-2000	2001-3000	3001-4000	4001-5000	Meer dan 5000	Geen Antwoord
Man	-2,997	0,675	0,166	1,808	-0,011	0,842	-1,764	-2,508	-2,148	0,726	-2,688	1,372
Vrouw	2,193	2,94	-0,123	-1,633	2,405	-0,498	2,688	-2,366	-2,001	0,442	2,587	1,427

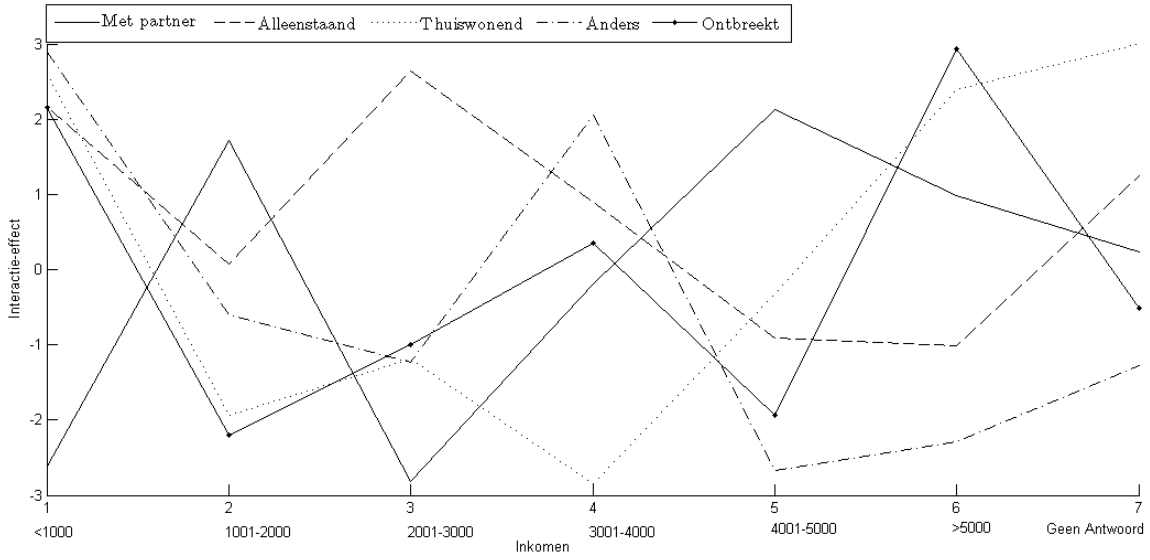
(c) Interactie thuissituatie en inkomen

	Thuissituatie				
Inkomen	Met partner	Alleenstaand	Thuiswonend	Anders	Ontbreekt
Minder dan 1000	-2,62	2,163	2,606	2,906	2,154
1001-2000	1,713	0,08	-1,934	-0,608	-2,196
2001-3000	-2,815	2,635	-1,192	-1,227	-1,002
3001-4000	-0,198	0,889	-2,849	2,053	0,354
4001-5000	2,125	-0,913	-0,324	-2,675	-1,937
Meer dan 5000	0,977	-1,015	2,391	-2,291	2,931
Geen Antwoord	0,24	1,242	2,997	-1,273	-0,513

personen met een lager inkomen een grotere kans hebben om slachtoffer te worden van een seksueel delict.

In Figuur 1 zijn de interactie-effecten tussen inkomen en thuissituatie grafisch weergegeven. In dit figuur is te zien dat de effecten van de thuissituatie en inkomen op het aantal delicten niet hetzelfde verloop vertonen. Dit betekent dat er sprake is van een interactie-effect tussen inkomen en thuissituatie, dus de invloed van het inkomen op het aantal delicten is afhankelijk van de thuissituatie. Het hebben van een partner heeft met name een ander interactie-effect met de inkomens van individuen met een andere thuissituatie. Zo is het interactie-effect tussen 'Met partner' en 'Minder dan 1000' negatief, terwijl het effect van dit lage inkomen met de andere thuissituaties positief is op het aantal delicten. Een dergelijke weergave is ook te maken voor de overige interactie-effecten.

Figuur 1: Interactie-effecten tussen inkomen en thuissituatie



3.2 OSI-model op Poisson verdeling

Het OSI-model schat 82, 3% minder parameters dan een tweede order GLM-model. Ondanks dat het OSI-model naast de coëfficiënten van de effecten ook de natuurlijke parameter θ schat. Door het schatten van minder parameters kan een grote hoeveelheid tijd aan het schatten, modelleren en interpretatie bespaard worden. De maximale log-likelihood van dit OSI-model is gelijk aan $1644, 463$. Het OSI-model heeft een hogere likelihood en daardoor een betere fit dan de eerste orde GLM-model. Echter is de fit van het OSI-model lager dan de fit van het tweede orde GLM-model. Daarnaast heeft het OSI-model een DIC-waarde van $280588, 569$. Deze waarde is hoger dan de DIC-waarde van het eerste orde GLM-model en lager dan deze waarde in het tweede orde GLM-model. Dit betekent dat de complexiteit van het OSI-model niet opweegt tegen de extra fit ten opzicht van het GLM model zonder interactie-effecten. Zodra er wel interactie-effecten opgenomen dienen te worden, weegt in het OSI-model voor een Poisson verdeling het verlies aan fit op tegen de enorme complexiteitsvermindering ten opzichte van een GLM-model van tweede orde. Tevens claimt een OSI-model makkelijker te schatten te zijn dan een tweede orde GLM-model. De bijbehorende optimaal geschaalde categorische variabelen zijn in Figuur 2 te vinden. Deze kwantificaties zijn ook in Tabel 7 terug te vinden samen met de hoofdeffecten. Tot slot zijn de geschatte coëfficiënten van een OSI-model op een Poissonmodel terug te vinden in Tabel 8 met bijbehorende standaarddeviaties.

Tabel 7: Hoofdeffecten en optimale kwantificaties

(a) Leeftijd

	Aantal	Hoofdeffect	Opt. kw.
15-25	844	0,386	-0,795
25-35	814	0,198	-0,408
35-45	419	-0,707	1,454
45-55	323	-0,888	1,828
55-65	251	0,540	-1,111
65-75	197	0,014	0,029
> 75	103	-0,241	0,496

(b) Geslacht

	Aantal	Hoofdeffect	Opt. kw.
Man	1315	-0,026	1,161
Vrouw	1636	0,019	0,861

(c) Thuissituatie

	Aantal	Hoofdeffect	Opt. kw.
Met partner	1587	-0,335	-0.743
Alleenstaand	75	-0,162	-0.36
Thuiswonend	705	1,690	3,747
Anders	37	0,519	1,151
Ontbreekt	547	-1,249	-2,769

(d) Inkomen

guldens pm	Aantal	Hoofdeffect	Opt. kw.
<1000	112	-7.23	-2,389
1.001-2.000	400	-4,507	1,400
2.001-3.000	614	1,920	0,634
3.001-4.000	506	4,013	1,325
4.001-5.000	419	-2,211	-0,73
>5.000	350	-0,018	0,006
Geen antwoord	550	0,709	0,234

Tabel 8: Geschatte coëfficienten OSI model voor count data

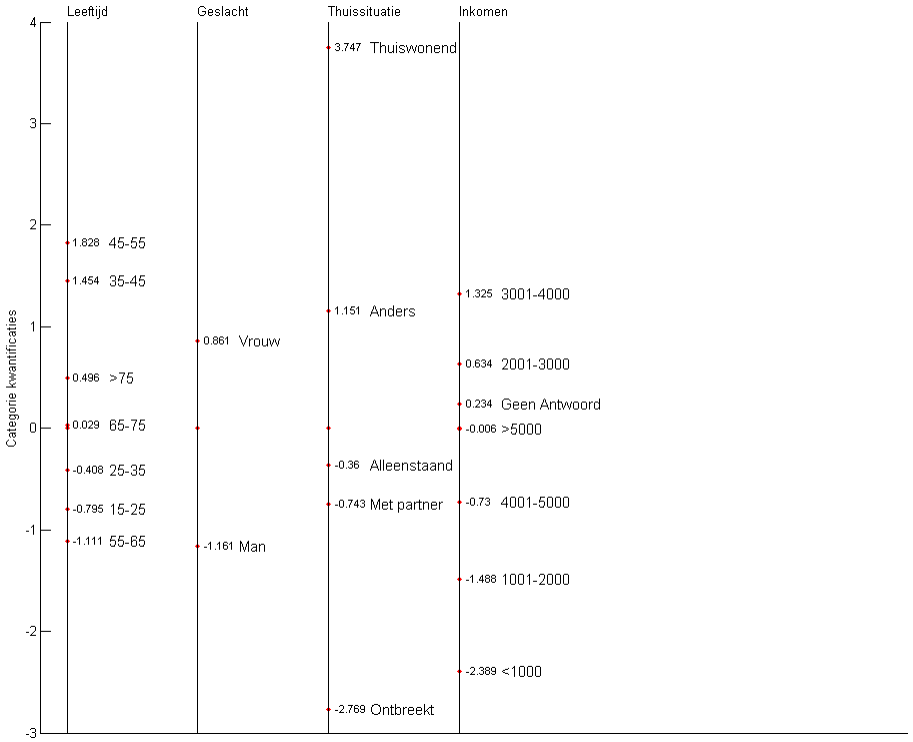
	Leeftijd		Geslacht		Thuissituatie		Inkomen	
Leeftijd	-		0,611	(0,633)	1,143	(0,862)	$1,674^*$	(0,423)
Geslacht	-		-		-0,376	(0,873)	$1,619^*$	(0,404)
Thuissituatie	-	-	-	-	-		$-1,25^*$	(0,625)
Inkomen		-	-	-	-	-	-	
Constante	-1,017	(0,638)						

**significant op 5%*

Het hoofdeffect van het hebben van een inkomen minder dan 1000 gulden is -7,234 ($3,029 \times -2,389$). Daarnaast is dit effect van het hebben van een modaal inkomen (tussen 3000 en 4000) gelijk aan 4,012 ($3,029 \times 1,325$). De grootte van deze effecten hangt af van de grootte van de af-

hankelijke variabelen en hebben geen directe betekenis. De hoofdeffecten zijn te interpreteren door deze met de andere hoofdeffecten te vergelijken. In dit onderzoek gaan we hier niet verder op in.

Figuur 2: Kwantificaties OSI-model in een Poisson verdeling



Door Figuur 2 te combineren met Tabel 8 is af te leiden welke interactie-effecten relatief veel invloed hebben op het aantal seksuele delicten. Grote interactie-effecten treden op als in Figuur 3 twee extreme groepen van verschillende gekwantificeerde variabelen gekozen worden. Om een interactie-effect tussen leeftijd en geslacht te bekijken, bijvoorbeeld het effect op het zijn van een man tussen de 25 en 35 jaar, kan de coëfficiënt van leeftijd en geslacht vermenigvuldigd worden met de kwantificaties in het OSI-model: 0.2894 $(0,611 \times -0,408 \times -1,161)$. Een man tussen de 45 en 55 jaar uit heeft een bijbehorend interactie-effect van -1.2967 $(0,611 \times 1,828 \times -1,161)$. Ook is het mogelijk om te kijken naar bijvoorbeeld het effect tussen inkomen en de thuissituatie. Het interactie-effect van een thuiswonend persoon met weinig tot geen inkomen respectievelijk een thuiswonend persoon met een modaal inkomen (3001-4000 gulden) is 11,190 $(-1,25 \times 3,747 \times -2,389)$ respectievelijk -6.3231 $(-1,25 \times 3,747 \times 1,325)$. Zodra een thuiswonend persoon weinig tot geen inkomen heeft is de kans aanzienlijk groter om slachtoffer te worden van aanranding.

De interactie-effecten van het OSI-model op een Poisson verdeling zijn groter dan deze effecten in het tweede orde GLM-model. De hoofdeffecten uit het OSI-model vertonen over het algemeen wel dezelfde richting als de significante hoofdeffecten in het eerste orde GLM-model. Een verschil in grootte is met name terug te vinden bij de eigenschappen thuissituatie en inkomen.

4 Discussie

In dit artikel is gekeken naar het opnemen van interactie-effecten in een model voor count data. Hierbij zijn GLM-modellen en een OSI-model op een Poisson verdeling naar voren gekomen. In dit onderzoek is gebleken dat minder hoofdeffecten significant van invloed zijn op het aantal delicten zodra er interactie-effecten worden toegevoegd. Ook zijn de verschillen in coëfficiënten van de hoofdeffecten tussen het eerste en tweede orde GLM-model groot. Daarbij is de fit van het tweede orde GLM-model aanzienlijk beter dan de fit van het eerste orde GLM-model. Echter weegt de verbeterde fit niet op tegen de complexiteit van het tweede orde GLM-model. Verder claimt een OSI-model makkelijker te schatten te zijn dan een tweede orde GLM-model. Het OSI-model heeft een betere fit dan het eerste orde GLM-model, maar is meer complex door het toevoegen van de interactie-effecten.

Om de conclusie omtrent de bruikbaarheid van het OSI-model voor een Poisson verdeling aan te scherpen, is een toekomstig onderzoek naar de standaardafwijkingen van de parameters in het tweede orde GLM-model noodzakelijk. Door een dergelijk onderzoek kunnen naast de fit en complexiteit, ook de schattingen van de modellen vergeleken worden. Uit een dergelijk onderzoek volgt of het OSI-model ook een goede schatting geeft van de richting en grootte van de interactie-effecten.

Wel is uit dit onderzoek gebleken dat het OSI-model bij deze data $82,3\%$ minder parameters schat dan een tweede orde GLM-model. Zodra er interactie-effecten opgenomen dienen te worden, geeft het OSI-model voor een Poisson verdeling een betere afweging tussen fit en complexiteit voor het werkelijk aantal delicten dan een GLM-model met interactie-effecten. Het schatten van extreem veel variabelen in een GLM-model resulteert dus in een onnodige complex model. Interacties tussen categorische variabelen in een Poisson verdeling kunnen het beste worden opgenomen in een OSI-model om het aantal predictorvariabelen in te perken, zonder teveel aan fit te verliezen. Mits de parameters in het OSI-model ook overeenkomen met de richting en grootte van parameters in een GLM-model van orde twee.

Tot slot is het aan te raden om in toekomstig onderzoek aandacht te besteden aan het correct meegeven van de W-matrix. Deze matrix is reeds naar voren gekomen in sectie 2.4.1. Deze kan gebruikt worden om een selectie te maken in de mee te nemen interactie-effecten. In dit onderzoek is elk effect toegevoegd en zijn alle elementen van W gelijk aan 1. Maar een model dat alleen significante effecten meeneemt, levert wellicht een model met een betere fit en minder complexiteit. Bij het uitvoeren van toekomstig onderzoek kan worden gekeken naar een algoritme om een W-matrix desdanig te construeren dat in het OSI-model enkel significante interacties meegenomen worden. Deze correct geschatte W-matrix kan dan gebruikt worden om deze betere fit te bereiken.

Referenties

[1] Van Rosmalen, J., Koning, A.J. & Groenen, P.J.F. (2009). Jaarboek MarktOnderzoek Association H11, *Optimaal schalen van interactie-effecten*, 177-193.

[2] Van Rosmalen, J., Koning, A.J. & Groenen, P.J.F. (2009). Multivariate Behavioral Research Models, *Optimal Scaling of Interaction Effects in Generalized Linear*, 44(1), 59-81.

[3] Van Rosmalen, J., Koning, A.J. & Groenen, P.J.F. (2009). ERIM PhD Series Vol. 165, *Segmentation and Dimension Reduction: Exploratory and Model-Based Approaches*, 9-30, 67-76.

[4] Cleaver, G., Donkers, B., Groenen, P.J.F., Koning, A.J. & Van Rosmalen, J. (2010). Werkdocument, Versie 0.12 November, *Interactions in conjoint analysis*, Work in Progress.

[5] Cleaver, G., Donkers, B., Groenen, P.J.F., Koning, A.J. & Van Rosmalen, J. (2013). Werkdocument, Versie 0.15 June, *Interactions in conjoint analysis*, Work in Progress.

[6] Koning, A.J. (2013). R, Package 'Choices', *Using choice data in conjoint analyis.*

[7] Koning, A.J. & Groenen, P.J.F. (2006). Multiple correspondence analysis and related methods, In: M. Greenache & J. Blasius (Eds.), *A new model for visualizing interactions in analysis of variance*, Chapman and Hall, 487-502.

[8] Haight, F.A. (1967). Publications in operations research No. 11, *Handbook of the poisson distribution*, United States of America: John Wiley & Sons, Inc., 1-99.

[9] Nelder, J.A. & Wedderburn, R.W.M. (1972). Journal of the Royal Statistical Society Series A, No. 135, part III, *Generalized Linear Models*, 370-384.

[10] Anderrson, C.J. & Vermunt, J.K. (2000). Sociological Methodology, Vol. 30, *Log-Multiplicative Association Models as Latent Variable Models for Nomonal and/or Ordinal Data*, 81-121.