ANALYSE VAN KRUISTABELLEN: LOG-LINEAIRE POISSON MODELLEN VOOR GEWOGEN AANTALLEN

R-76-8
J. de Leeuw (R.U. Leiden)
S. Oppe (SWOV)
Voorburg, 1976
Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV

INHOUD

1. KRUISTABELLEN

Kruistabellen (cross-tables, contingency tables) zijn tabellen waarin elementen van populaties of steekproeven (van allerlei aard) zijn geklassificeerd t.a.v. een of meer kenmerken. Bv. de klassifikatie van dodelijke slachtoffers over het jaar 1974 naar de kenmerken leeftijd en wijze van verkeersdeelname. Indien sprake is van slechts één kenmerk, dan wordt vaak gesproken van marginale tabellen. Maar ook wel bij tabellen die ontstaan indien over een of meer variabelen van een kruistabel wordt opgeteld en over een of meer andere niet. Omdat er geen wezenlijk verschil is tussen een marginale tabel en een kruistabel (enkel een funktioneel verschil), zullen we het voortaan slechts hebben over kruistabellen.
De engelse term 'contingency table' (tabel met toevallige gebeurtenissen) zou in feite beter zijn, omdat in deze term iets wordt uitgedrukt van de assumpties die worden gemaakt bij de analyse van kruistabellen t.a.v. de toevalsfaktoren die geacht worden een rol te spelen bij het tot stand komen van de tabel. Dit aspect is met name bij steekproeven essentieel.
M.b.v. een steekproef proberen we enerzijds een beschrijving te geven van de populatie waaruit de steekproef is getrokken, anderzijds proberen we uitspraken over die populatie te toetsen. In de hieronder beschreven analysemodellen wordt er vanuit gegaan dat een steekproef een beeld geeft van de populatie dat, enkel afhankelijk van toevalsfluktuaties, meer of minder juist is.
Aannamen over de wijze waarop het toeval een rol speelt vormen de basis van het analysemodel. Daarbinnen zijn weer verschillende specifikaties van het model mogelijk.
Bij analyse van kruistabellen wordt meestal niet verondersteld dat er specifieke relaties (zoals orde-relaties of zelfs metrische relaties) tussen de klassen van een kenmerk aanwezig zijn.
Deze extra veronderstellingen zijn echter binnen specifieke modellen wel mogelijk. Bv. t.a.v. een variabele als leeftijd.
De laatste jaren is er een nieuwe ontwikkeling te konstateren in de manier waarop kruistabellen worden geanalyseerd. Was het vroeger gebruikelijk om (meestal d.m.v. chi-kwadraat toetsing) overall-

hypotheses te toetsen over een tabel met één of twee kenmerken, nu ligt bij de analyse de nadruk steeds meer op de gedetailleerde informatie welke in de tabel aanwezig is. Verder is het mogelijk om ook tabellen van hogere orde (uitsplitsing naar meerdere kenmerken) te analyseren, zodat ook meer gekompliceerde samenhangen, dus relaties tussen meer dan twee kenmerken tegelijk, kunnen worden onderzocht.

2. MODEL

2.1. Basisaanname

De basisaanname is dat de aantallen doden in de cellen van de kruistabel onafhankelijke random variabelen zijn die een Poissonverdeling hebben, waarbij de parameters van de Poissonverdeling kunnen verschillen. Strikt genomen kan deze assumptie slechts gelden voor de aantallen dodelijke ongevallen, maar er wordt aangenomen dat de (om praktische redenen gemaakte) keuze voor de aantallen doden slechts een gering vertekenend effekt kan hebben. Om het enigszins konkreet te houden: stel dat we te maken hebben met een twee-weg kruistabel, met r rijen en k kolommen dan zouden we de Poissonaanname voor iedere cel als volgt kunnen schrijven: er zijn getallen $\lambda_{ij} > 0$ $(i=1, .., r; j=1, .., k)$ zodanig dat

$$\text{prob}\left[\underset{\sim}{X}_{ij} = x_{ij}\right] = e^{-\lambda_{ij}} \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!}$$

Hierbij is $\underset{\sim}{X}_{ij}$ de stochastische veranderlijke in cel (i,j) die als waarden de natuurlijke getallen $x_{ij} = 0,1,2,\ldots$ kan aannemen. Een kortere schrijfwijze voor deze aanname is

$$\underset{\sim}{X}_{ij} \sim \mathcal{P}(\lambda_{ij}),$$

wat we kunnen lezen als: X_{ij} heeft een Poissonverdeling met parameter λ_{ij}.

2.2. Onafhankelijkheidsaannamen betreffende kenmerken in modellen

Hoewel we aannemen dat de $\underset{\sim}{X}_{ij}$ onafhankelijk zijn is het natuurlijk wel mogelijk dat er tussen de parameters λ_{ij} relaties bestaan. Door de relaties tussen deze parameters te onderzoeken kunnen we nagaan of ook de kenmerken die de variabelen bezitten onafhankelijk van elkaar zijn.

Wat bedoelen we als we zeggen dat de rijen en kolommen van een rxk kruistabel (met onafhankelijke Poissonvariabelen $\underset{\sim}{X}_{ij}$) korresponderen met onafhankelijke rij- en kolomvariabelen? Stel $\underset{\sim}{X}_{i.}$ en $\underset{\sim}{X}_{.j}$ zijn de marginale verdelingen, d.w.z.

$$\underset{\sim}{X}_{i.} = \sum_{j=1}^{k} \underset{\sim}{X}_{ij}$$

en

$$\underset{\sim}{X}_{.j} = \sum_{i=1}^{r} \underset{\sim}{X}_{ij}$$

De eis dat de rij- en kolomvariabelen onafhankelijk zijn komt erop neer dat de kansen van de r konditionele verdelingen binnen rijen

$$\text{prob}\left[(\underset{\sim}{X}_{i1} = x_{i1}) \wedge (\underset{\sim}{X}_{i2} = x_{i2}) \wedge \ldots \wedge (\underset{\sim}{X}_{ik} = x_{ik}) \,\middle|\, \underset{\sim}{X}_{i.} = x_{i.}\right]$$

hetzelfde zijn voor alle i=1, ..,r, en dat de k konditionele verdelingen binnen kolommen

$$\text{prob}\left[(\underset{\sim}{X}_{1j} = x_{1j}) \wedge (\underset{\sim}{X}_{2j} = x_{2j}) \wedge \ldots \wedge (\underset{\sim}{X}_{rj} = x_{rj}) \,\middle|\, \underset{\sim}{X}_{.j} = x_{.j}\right]$$

hetzelfde zijn voor alle j=1, ..,k. Gebruik makend van de onafhankelijkheid van de $\underset{\sim}{X}_{ij}$ en de Poisson aanname kunnen we afleiden dat de konditionele verdelingen binnen rijen gelijk zijn aan de multinomiaal verdelingen

$$\left(\frac{x_{i.}!}{\prod_{j=1}^{k} x_{ij}}\right) \prod_{j=1}^{k} \left(\frac{\lambda_{ij}}{\lambda_{i.}}\right)^{x_{ij}}$$

terwijl de konditionele verdelingen binnen kolommen gelijk zijn aan de multinomiaal verdelingen

$$\left(\frac{x_{.j}!}{\prod_{i=1}^{r} x_{ij}}\right) \prod_{i=1}^{r} \left(\frac{\lambda_{ij}}{\lambda_{j}}\right)^{x_{ij}}$$

De rij- en kolomvariabelen zijn dus onafhankelijk wanneer $\left(\frac{\lambda_{ij}}{\lambda_{i.}}\right)$ hetzelfde is voor alle j, en $\left(\frac{\lambda_{ij}}{\lambda_{.j}}\right)$ hetzelfde is voor alle i.

Noodzakelijk en voldoende hiervoor is dat er getallen $\alpha_i \geq 0$ $(i=1,..,r)$, $\beta_j \geq 0$ $(j=1,..,k)$, en $\mu \geq 0$ zijn, zodanig dat $\lambda_{ij} = \mu \alpha_i \beta_j$ voor alle i,j.

Dit multiplikatieve model wordt meestal herleid tot een lineair model door het nemen van de logarithme:

$$\ln \lambda_{ij} = \mu + \alpha_i + \beta_j \qquad (1)$$

waarbij $\alpha_i = \ln \alpha_i$ enz.

Vandaar dat dergelijke modellen dan ook wel log-lineaire modellen worden genoemd

Het log-lineaire model is dus equivalent met de eis van onafhankelijkheid van de rij- en kolomvariabelen.

2.3. Verzadigde en onverzadigde modellen

Zoals genoemd is er naast het toetsen van hypothesen omtrent tabellen soms ook belangstelling voor beschrijving van de tabellen. In het geval dat de kenmerken niet onafhankelijk zijn en het bovengenoemde model (1) dus niet opgaat, is het model uit te breiden met specifieke parameters voor de cellen. In dat geval geldt dus het volgende model:

$$\ln \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (2)$$

In dit geval is het altijd mogelijk zodanige parameters μ, α_i, β_j en γ_{ij} te vinden dat er een volledige overeenstemming is tussen de tabel die men wil beschrijven en het model dat hierbij gebruikt wordt. De betekenis van de beschrijving is nu dat de variatie in de aantallen observaties van de cellen van de tabel wordt weergegeven in relatie tot de struktuur van de tabel: men kan bijvoorbeeld zien in hoeverre de variatie toe te schrijven is aan een rij-effekt, een kolom-effekt of een interaktie-effekt. Hoewel men hier evenveel parameters heeft als cellen en er dus geen reduktie van informatie is, is er wel een ordening van informatie. Merk verder op dat model (1) een bijzonder geval is van model (2): het is gelijk op de restriktie na, dat $\gamma_{ij} = 0$ voor alle i, j. Er zijn nog andere restrikties mogelijk, bv. dat de α_i's onderling

een lineaire relatie vormen of bv. gelijk aan nul zijn. In al deze gevallen spreken we van onverzadigde modellen. Als we te maken hebben met een steekproef dan kunnen we deze niet-verzadigde modellen zien als toetsbare hypothesen omtrent de populatie waaruit de steekproef afkomstig is. Bij een verzadigd model is deze toetsing niet mogelijk omdat het model de gegevens volledig beschrijft.
T.a.v. de keuze van het analysemodel is er een grote overeenkomst met lineaire modellen zoals gebruikt bij variantie-analyse. Ook hier kunnen we spreken van een afbraak van de tabel in komponenten: hoe groot is de rij-bijdrage, de kolombijdrage, de unieke celbijdrage voor iedere cel? Voor een willekeurige tabel is dit na te gaan door schatting van de parameters van het model.
Deze systematische afbraak geeft dus een efficient overzicht van de informatie die in de tabel aanwezig is. Verder is het mogelijk om betrouwbaarheidsgrenzen van de schatters voor de parameters te geven zodat ook toetsing van individuele schatters mogelijk is.

Een goede weergave van de relatie tussen variantie-analysemodellen en log-lineaire modellen vindt men bij Nelder & Wedderburn [6].

2.4. Gewogen Poisson-modellen

Tot nu toe hebben we enkel gesproken over aantallen doden als funktie van een aantal kenmerken. Soms zijn we echter geïnteresseerd in de analyse van dodencijfers die genormeerd zijn op een bepaalde expositiegrootheid zoals inwonertal, lengte van wegen, etc. Indien we de aantallen doden in de tabel aanvullen met een expositiemaat per cel, die mag verschillen van cel tot cel, dan kunnen we een algemener Poisson-model hanteren. De fundamentele aanname wordt nu

$$\underset{\sim}{X}_{ij} \sim \mathcal{P}\ (e_{ij}\ \lambda_{ij}),$$

waarbij de e_{ij} de gegeven expositiegrootheden zijn, en waarbij voor de λ_{ij} weer een log-lineair model aangenomen wordt.

3. DE DESIGN-MATRIX

3.1. Algemeen

In matrix-notatie is de algemene vorm van een log-lineair model voor n Poisson-variabelen $\underset{\sim}{X}_1 \sim P(\lambda_1)$ te schrijven als

$$\eta = V\theta,$$

waarbij η een vektor is van waarden $\eta_1 = \ln \lambda_1$. V is een gegeven matrix van de orde n x p (de zogenaamde design-matrix), en θ een vector van p onbekende parameters. Zijn de X_1 geordend in een twee-weg tabel en vervangen we de index 1 door de rij- en kolomindexen i en j, dan kunnen we het model

$$\eta_{ij} = \ln \lambda_{ij} = \mu + \alpha_i + \beta_j$$

in het geval waarin r = k = 2 bijvoorbeeld herschrijven als

$$\begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Merk op dat in dit geval de design-matrix V van de orde 4 x 5 en van de rang 3 is. Dit wordt duidelijk als we het model herschrijven in de equivalente vorm

$$\begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

met

$$\theta_1 = \mu + \bar{\alpha} + \bar{\beta}$$
$$\theta_2 = (\alpha_1 - \bar{\alpha}) = -(\alpha_2 - \bar{\alpha})$$
$$\theta_3 = (\beta_1 - \bar{\beta}) = -(\beta_2 - \bar{\beta})$$

waarbij met $\bar{\alpha}$ en $\bar{\beta}$ respectievelijk het gemiddelde van de α's en β's wordt bedoeld.

In het algemeen is het altijd mogelijk (en wenselijk) om de design-matrix zo te kiezen dat zijn rang gelijk is aan zijn aantal kolommen. Dit voorkomt dat extra restrikties aan de parameters moeten worden opgelegd om een unieke oplossing te vinden. In het geval dat we een rechtstreekse oplossing voor de α's en β's zouden zoeken zouden deze restrikties zijn: $\alpha_1 + \alpha_2 = 0$ en $\beta_1 + \beta_2 = 0$. Het gelijk zijn van de rang aan de matrix aan het aantal kolommen is bijvoorbeeld altijd het geval als we V zo kiezen dat V'V diagonaal is, waarbij met $\mathbf{V}'$ de getransponeerde van matrix V wordt aangeduid (V heet dan kolomsgewijs orthogonaal) of zo dat V'V gelijk is aan de eenheidsmatrix (V heet dan kolomsgewijs orthonormaal).

3.2. Drie bruikbare vormen van design-matrixen

3.2.1. Helmert-matrices

We bekijken eerst het geval waarin we een enkelvoudige klassifikatie hebben. Voorbeeld: i=1,...,n korrespondeert met n leeftijdskategorieën, $\underset{\sim}{X}_i$ is het aantal ongevallen in ieder van die kategorieën. Een eerste type design-matrix dat dikwijls gebruikt wordt is de Helmert-matrix. Een komplete Helmert-matrix voor n = 4 ziet er als volgt uit

1 -1 -1 -1
1 1 -1 -1
1 0 2 -1
1 0 0 3

Merk op dat deze V kolomsgewijs orthogonaal is. Het model $\eta = V\theta$ is dus verzadigd. Een perfekte aanpassing is mogelijk als we $\hat{\theta} = (V'V)^{-1}V'\eta$ kiezen. Onverzadigde modellen zijn mogelijk door kolommen van V weg te laten, wat overeenkomt met de hypothese dat sommige van de elementen van θ in het verzadigde model gelijk zijn aan nul. De interpretatie van Helmert-effekten wordt duidelijk uit de volgende equivalenties:

$$\theta_1 = 0 \Longleftrightarrow \Sigma\eta_1 = 0 \Longleftrightarrow \sqrt[n]{\prod_{i=1}^{n}\eta_i}\lambda_{i=1}$$

$$\theta_2 = 0 \Longleftrightarrow \eta_2 = \eta_1 \Leftrightarrow \lambda_2 = \sqrt[1]{\lambda_1}$$

$$\theta_3 = 0 \Leftrightarrow 2\eta_3 = \eta_1 + \eta_2 \Leftrightarrow \lambda_3 = \sqrt[2]{\lambda_1 \lambda_2}$$

$$\theta_4 = 0 \Leftrightarrow 3\eta_4 = \eta_1 + \eta_2 + \eta_3 \Leftrightarrow \lambda_4 = \sqrt[3]{\lambda_1 \lambda_2 \lambda_3}$$

Hieruit kunnen we bijvoorbeeld afleiden

$$\theta_3 = \theta_4 = 0 \Leftrightarrow \lambda_3 = \lambda_4 = \sqrt[2]{\lambda_1 \lambda_2},$$

enzovoort. Helmert-effekten vergelijken dus iedere λ_i afzonderlijk met het geometrisch gemiddelde (g.g.) van de voorafgaande λ_i. Op zo'n manier kunnen we uitvinden of er een trend in onze data zit, of misschien een plotselinge sprong.

3.2.2. Orthogonale polynomen

Stel dat de leeftijdskategorieën in ons voorbeeld intervallen zijn met gelijke lengte. We zouden ons dan kunnen interesseren voor het funktionele verband tussen leeftijd en aantal ongevallen. We kunnen dit funktionele verband beschrijven als een polynoom, dat wil zeggen als een lineaire kombinatie van orthogonale polynomen, voor n = 3 levert dit bv. de volgende (kolomsgewijs orthogonale) design matrix op:

1 -1 1
1 0 -2
1 1 1

Iedere konstante funktie op (1,2,3) is vanzelfsprekend een veelvoud van de eerste kolom van V, iedere lineaire funktie op (1,2,3) is een lineaire kombinatie van de eerste twee kolommen, en iedere tweedegraadsfunktie is een lineaire kombinatie van de eerste drie kolommen. Iedere funktie op (1,2,3) kan opgevat worden als een tweedegraadsfunktie: dit is alleen maar een andere manier om te zeggen dat het model gedefinieerd door V verzadigd is. Onverzadigde modellen zijn over het algemeen van de vorm $\theta_3 = 0$ of $\theta_2 = \theta_3 = 0$. De hypothese $\theta_3 = 0$ zegt dat de drie punten $(1,\eta_1)$, $(2,\eta_2)$, en $(3,\eta_3)$ op een rechte lijn liggen, de hypothese $\theta_2 = \theta_3 = 0$ zegt dat $\eta_1 = \eta_2 = \eta_3$. In het algemeen kan de hypothese dat $(\eta_1,\ldots,\eta_n)$ een q-de graadspolynoom is van $(1,2,\ldots,n)$ geschreven worden als $\eta_i = \pi_q(i)$. Uit onze diskussie volgt

$$\eta_i = \pi_q(i) \Leftrightarrow \theta_{q+1} = \ldots = \theta_n = 0.$$

De interpretatie van polynoomeffekten in log-lineaire modellen wordt bemoeilijkt door het gebruik van de log-transformatie. Immers

$$\eta_i = \pi_q(i) \Leftrightarrow \lambda_i = \exp(\pi_q(i)) = \exp(\alpha_o + \alpha_1 i + \ldots + \alpha_q i^q) =$$

$$= [\exp(i^0)]^{\alpha 0} \; [\exp(i^1)]^{\alpha 1} \ldots [\exp(i^q)]^{\alpha q}$$

Deze laatste funktie is wat minder simpel en vertrouwd als een polynoom.

3.2.3. Tussen-Binnen kontrasten

In veel gevallen vallen kategorieën van onze klassifikatie op natuurlijke wijze uiteen in verschillende groepen. Leeftijd kan bv. gegroepeerd worden in twee groepen beneden en boven de veertig. Deze indeling kunnen we in verzadigde design-matrix vorm weergegeven als

00-20	1	-1	-1	0
20-40	1	-1	+1	0
40-60	1	1	0	-1
60-80	1	1	0	+1

In dit geval zijn de metingen zelf dus in vier kategorieën, en gaan we als het ware na of een indeling in minder kategorieën mogelijk is zonder al te veel verlies van informatie. De eerste kolom van V korrespondeert zoals gewoonlijk met het totaalgemiddelde, de tweede kolom kontrasteert de twee groepen (het effekt tussen groepen), en de derde en vierde kolom bekijken de effekten binnen de groepen afzonderlijk. Als er K groepen zijn met n_k elementen ($\sum_{k=1}^{K} n_k = n$), dan zijn er in het algemeen K-1 tussen-groepeffekten, en $\sum_{k=1}^{K} (n_k - 1) = n - K$ binnen-groepeffekten. De meest voorkomende onverzadigde modellen stellen dat alle θ-waarden korresponderend met tussen-groepeffekten nul zijn. Dit komt overeen met de hypothese dat de arithmetische gemiddelden van de η_i gelijk zijn voor iedere groep, wat equivalent is met het feit dat de geometrische gemiddelden van de λ_i hetzelfde zijn voor iedere groep.

3.2.4. Kombinatie van design-matrices

We bekijken nu een tweevoudige klassifikatie met bijvoorbeeld twee klassen in het eerste kenmerk (Noord-Brabant tegen de rest van Nederland), en vier klassen in het tweede kenmerk (bv. de vier leeftijdskategorieën uit de vorige paragraaf). We kiezen eerst twee design matrices V_1 en V_2 voor de kenmerken afzonderlijk.

Bijvoorbeeld

$$V_1 = \begin{matrix} +1 & -1 \\ +1 & +1 \end{matrix}$$

$$V_2 = \begin{matrix} +1 & -1 & -1 & 0 \\ +1 & -1 & +1 & 0 \\ +1 & +1 & 0 & -1 \\ +1 & +1 & 0 & +1 \end{matrix}$$

We vormen vervolgens van alle 2 x 4 = 8 kombinaties van kolommen van V_1 en V_2 het uitwendig produkt (het uitwendig produkt van een n-vector x en een m-vector y is een n x m matrix met als elementen $x_i y_j$). Dit geeft de volgende acht matrices

V_1	V_2	Produkt				V_1	V_2	Produkt			
1	1	+1	+1	+1	+1	2	1	-1	-1	-1	-1
		+1	+1	+1	+1			+1	+1	+1	+1
1	2	-1	-1	+1	+1	2	2	+1	+1	-1	-1
		-1	-1	+1	+1			-1	-1	+1	+1
1	3	-1	+1	0	0	2	3	+1	-1	0	0
		-1	+1	0	0			-1	+1	0	0
1	4	0	0	-1	+1	2	4	0	0	+1	-1
		0	0	-1	+1			0	0	-1	+1

We kunnen deze acht matrices opvatten als acht vektoren van acht elementen, en zo een design-matrix V_{12} vormen met deze vektoren als kolommen. Dus:

Design-matrix	behorend bij vektor
+1 -1 -1 0 -1 +1 +1 0	η_{11}
+1 -1 +1 0 -1 +1 -1 0	η_{12}
+1 +1 0 -1 -1 -1 0 +1	η_{13}
+1 +1 0 +1 -1 -1 0 -1	η_{14}
+1 -1 -1 0 +1 -1 -1 0	η_{21}
+1 -1 +1 0 +1 -1 +1 0	η_{22}
+1 +1 0 -1 +1 +1 0 -1	η_{23}
+1 +1 0 +1 +1 +1 0 +1	η_{24}

De zo gevormde matrix V_{12} is weer kolomsgewijs orthogonaal, en definieert een verzadigd model. We kunnen zeggen dat V_{12} gevormd is via uitwendige produkten. Bij gebruik van een design-matrix die op deze manier opgebouwd is willen we gewoonlijk een bepaald soort onverzadigde modellen onderzoeken. We zullen deze onverzadigde modellen voor ons voorbeeld bekijken. We kiezen eerst de kolom die korrespondeert met de eerste kolom van V_1 en de eerste kolom van V_2. Dit is de eerste kolom van V_{12}. De hypothese $\theta_1 = 0$ is equivalent met de hypothese dat het arithmetrisch gemiddelde van de η_{ij} $(i=1,2;j=1,2,3,4)$ nul is, d.w.z. dat het geometrisch gemiddelde van de λ_{ij} één is.

We kiezen vervolgens de groep van kolommen van V_{12} die samengesteld zijn uit de eerste kolom van V_1 en kolom twee, drie of vier van V_2. Dit zijn kolommen 2, 3, 4 van V_{12}. De hypothese $\theta_2 = \theta_3 = \theta_4 = 0$ is equivalent met de hypothese dat de kolomgemiddelden van de η_{ij} identiek zijn, ofwel

$$\eta_{.1} = \eta_{.2} = \eta_{.3} = \eta_{.4}$$

Dit is equivalent met

$$\lambda_{11}\lambda_{21} = \lambda_{12}\lambda_{22} = \lambda_{13}\lambda_{23} = \lambda_{14}\lambda_{24}.$$

Op dezelfde manier kunnen we de groep kolommen kiezen die samengesteld zijn uit de eerste kolom van V_2 en een niet-eerste kolom van V_1. Deze groep bestaat uit de vijfde kolom van V_{12}. De hypothese $\theta_5 = 0$ is

$$\lambda_{11}\lambda_{12}\lambda_{13}\lambda_{14} = \lambda_{21}\lambda_{22}\lambda_{23}\lambda_{24}.$$

Tenslotte is er de groep van kolommen 6,7,8 die korrespondeert met een niet-eerste kolom van V_1 en niet-eerste kolom van V_2. De hypothese $\theta_6 = \theta_7 = \theta_8 = 0$ korrespondeert met

$$\eta_{ij} = \frac{\eta_{i.}}{k} + \frac{\eta_{.j}}{r} - \frac{\eta_{..}}{kxr}$$

dat wil zeggen met het ontbreken van additieve interaktie in de η_{ij} (vergelijk model (1) op blz. 5), wat weer hetzelfde is als het ontbreken van multiplikatieve interaktie in de λ_{ij} (voor een vergelijking van deze twee vormen van interaktie verwijzen we naar Darroch 1974, Lancaster 1973, 1975). Het is duidelijk dat deze vorm van analyse via uitwendige produkten gegeneraliseerd kan worden naar tabellen met meer dan twee klassifikaties. We beginnen steeds met design matrices voor ieder van de kenmerken, vormen uitwendige produkten, en groeperen de kolommen van de uiteindelijke design matrix door na te gaan welke eerste kolommen erin voorkomen. Zo vormen we groepen effekten die overeenkomen met de additieve interakties van de η's (die bekend zijn uit gewone variantie-analyse), en met multiplikatieve interakties van de λ's (die geïnterpreteerd kunnen worden op de manier van paragraaf 2.2. als onafhankelijkheidsmodellen). Het is van belang om in te zien dat een <u>interaktie-hypothese</u> van de vorm $\theta_6 = \theta_7 = \theta_8 = 0$ uit bovenstaand voorbeeld waar is of niet waar is, <u>onafhankelijk van de keuze van de oorspronkelijke</u> V_1, V_2 ..... De keuze van de design matrix voor een bepaald kenmerk is dus alleen van belang om de individuele θ's beter te kunnen interpreteren, maar voor het beschrijven van de tabel naar de bijdragen van de kenmerken of de interakties tussen de kenmerken is de keuze van geen belang.

4. PARAMETERSCHATTING EN HYPOTHESETOETSING

4.1. Inleiding

We vatten voor het gemak nog even samen wat de fundamentele aannamen zijn van de klasse-modellen waarin we geïnteresseerd zijn.

A_1: $\underset{\sim}{X}_i \sim \mathcal{P}(\rho_i \lambda_i^o)$

A_2: $\underset{\sim}{X}_i$ zijn onafhankelijk

A_3: $\eta^o = V\theta^o$.

In A1 is ρ dus een bekende vektor van gewichten (of ekspositiematen), in A3 is $\eta^o = \ln \lambda^o$, en is V een bekende n x p design-matrix, waarvan we zullen aannemen dat hij kolomsgewijs orthonormaal is. Het superscript 'o' bij θ, η, en λ dient om de 'werkelijke' waarde van deze parameters aan te duiden, en om ze te onderscheiden van schatters en variabelen in bepaalde funkties. Wat ons interesseert is in de eerste plaats het schatten van de p onbekende parameters, en in de tweede plaats het toetsen of het model A1,A2,A3 juist is. Het is hierbij van belang om A3 ook nog op andere (equivalente) manieren te formuleren. Als V een n x p kolomsgewijs orthonormale matrix is, dan bestaat er een n x (n - p) kolomsgwijs orthonormale matrix V_c zodanig dat $V'V_c = 0$. Het is duidelijk dat A3 ook geschreven kan worden als

A3: $V_c' \eta^o = 0$.

Een derde formulering is mogelijk als we de p-dimensionale lineaire ruimte $\mathcal{V}$ definiëren als

$$\mathcal{V} = \{\eta \mid V_c' \eta = 0\} = \{\eta \mid \eta = V\theta\}.$$

dan

A3: $\eta^o \in \mathcal{V}$.

Over het algemeen is het ondoenlijk schatters en testprocedures te gebruiken die optimaal zijn voor alle mogelijke steekproefgroottes. We zullen daarom asymptotische argumenten gebruiken, en schatters en toetsen afleiden die optimale eigenschappen hebben als bepaalde grootheden naar oneindig gaan. Voor dit doel herformuleren we A1 als

A1: $\underset{\sim}{X}_i \sim P(m\rho_i \lambda_i^o)$.

De grootheid m geeft aan hoe groot onze gewichten ρ_i en parameters λ_i^o gemiddeld zijn. Als we doorgaan met waarnemen dan zullen de $\underset{\sim}{X}_i$ vanzelfsprekend naar oneindig gaan. De aanname A1 zegt nu in feite dat alle $\underset{\sim}{X}_i$ even snel naar oneindig gaan: als m oneindig groot wordt dan geldt dat de waarden $\underset{\sim}{X}_i/m$ naar de vaste grootheden $\rho_i \lambda_i^o$ konvergeren (in waarschijnlijkheid).
Voor onze analyses is het in het algemeen niet nodig om de waarde van m te kennen, we moeten alleen bereid zijn deze aanname te maken.
Uit de algemene theorie van de asymptotische statistische analyse zijn de volgende feiten bekend. In de eerste plaats zullen we geïnteresseerd zijn in schatters die <u>konsistent</u> zijn, dat wil zeggen dat als $m \to \infty$ dan $\hat{\theta}(m) \xrightarrow{p} \theta^o$. In de tweede plaats zijn we geïnteresseerd in schatters die <u>asymptotisch normaal</u> zijn, wat wil zeggen dat hun verdeling steeds meer op een multinormale verdeling gaat lijken als m naar oneindig gaat. Voor schatters met deze twee eigenschappen, die we samen kunnen vatten als

T1: $m^{\frac{1}{2}} (\hat{\theta}(m) - \theta^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$

geldt dat de asymptotische dispersie matrix Σ voldoet aan de ongelijkheid

$$\Sigma \geq (V'M^oV)^{-1}$$

Waarbij M^o de diagonale matrix is met op de diagonaal de waarden $\rho_i \lambda_i^o$. Schatters uit deze klasse waarvoor de genoemde ongelijkheid een gelijkheid is, en die dus in zekere zin zo precies mogelijk zijn, noemen we <u>efficient</u>. Hoewel vrijwel alle voor de hand liggende schatters voldoen aan T1, voldoen ze niet noodzakelijkerswijs aan de strengere eis

T2: $m^{\frac{1}{2}} (\hat{\theta}(m) - \theta^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (V'M^oV)^{-1})$.

Omdat efficientie een wenselijke eigenschap is, zullen wij ons tot efficiente schatters (dat wil zeggen tot schatters die voldoen aan T2) beperken. Bovendien is het van belang op te merken dat betrouwbaarheidsintervallen van schatters en toetsen van hypothesen over deze schatters over het algemeen asymptotisch optimaal zijn als de schatters efficient zijn.
Het is bekend dat efficiënte schatters gevonden kunnen worden door het maximaliseren van de aannemelijkheidsfunktie die de aannemelijkheid van de observaties als funktie van de parameters geeft, en dat

een asymptotisch optimale test van A3 binnen A1-A2 mogelijk is door het berekenen van de aannemelijkheidsverhouding tussen de best passende schatter(s) en de hypothetische waarde van de parameter(s). De schattings- en toetsingstheorie gebaseerd op deze "maximum likelihood" is voor log-lineaire Poisson modellen uiteengezet in Haberman (1974). De theorie is aangepast voor gewogen Poisson modellen in De Leeuw (1975). Omdat de berekeningen gebaseerd op aannemelijkheid over het algemeen niet erg eenvoudig zijn bekijken we hier een andere klasse van schatters en toetsen (ook optimaal en efficient), gebaseerd op de gemodificeerde minimum chi-kwadraat methode van Neyman (1949).

4.2. Gemodificeerde minimum chi-kwadraat methoden

We beginnen deze paragraaf met een bekende grenswaarde stelling voor Poisson variabelen die, op A1 toegepast, zegt dat voor $m \to \infty$

$$\frac{\underset{\sim}{X}_i - m\varrho_i\lambda_i^o}{(m\varrho_i\lambda_i^o)^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Als we definiëren

$$\underset{\sim}{Y}_i = \underset{\sim}{X}_i / (m\varrho_i)$$

dan kunnen we dit herschrijven in de wat handiger vorm

$$m^{\frac{1}{2}}(\underset{\sim}{Y}_i - \lambda_i^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{\lambda_i^o}{\varrho_i}).$$

Als we tenslotte definiëren

$$\underset{\sim}{Z}_i = \ln \underset{\sim}{Y}_i$$

volgt hieruit

$$m^{\frac{1}{2}}(\underset{\sim}{Z}_i - \mathcal{N}_i^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (\varrho_i\lambda_i^o)^{-1})$$

De gemodificeerde minimum chi-kwadraat methode welke we hieronder zullen bespreken heeft een eenvoudige geometrische interpretatie. We definiëren de afstandsmaat

$$\delta(\eta_1, \eta_2) = (\eta_1 - \eta_2)'\underset{\sim}{X}(\eta_1 - \eta_2).$$

De matrix $\underset{\sim}{X}$ is diagonaal, en op de diagonaal staan de $\underset{\sim}{X}_i$. Merk op dat we tot nu toe al aangetoond hebben dat

$$\delta(\underset{\sim}{Z}, \eta^o) \xrightarrow{\mathcal{L}} \chi^2_n$$

als $m \to \infty$ (dit volgt uit de limietverdeling van Z, en uit $\underset{\sim}{X}_i/m \xrightarrow{P} \varrho_i \lambda^o_i$). Voor schattingen bekijken we de afstand tussen de vektor $\underset{\sim}{Z}$ van waarnemingen, en de verzameling van toegestane schatters η. Voor het berekenen van de gemodificeerde minimum chi-kwadraat schatters moeten we $\hat{\eta}$ zodanig kiezen dat

$$\delta(\underset{\sim}{Z}, \hat{\eta}) = \min_{\eta \in \mathcal{U}} \delta(\underset{\sim}{Z}, \eta).$$

Dit levert een schatting $\hat{\eta}$ op voor η^o. De korresponderende schatter voor θ^o is $V'\hat{\eta}$, en de statistiek die gebruikt wordt om A3 te toetsen is $\delta(Z, \hat{\eta})$. In de volgende paragraaf bestuderen we verdelingen van schatters en toetsgrootheden.

4.3. Berekeningen en limietverdelingen

Het probleem

$$\min_{\eta \in \mathcal{U}} \delta(\underset{\sim}{Z}, \eta)$$

kan op twee verschillende manieren geformuleerd worden. De eerste formulering is

$$\min_{\theta} \delta(\underset{\sim}{Z}, V\theta).$$

Dit levert op schatters $\hat{\theta}_I$, en vervolgens $\hat{\eta}_I = V\hat{\theta}_I$. Formulering II gebruikt Lagrange vermenigvuldigers en kan geschreven worden als

$$\min_{\eta} \max_{\omega} \delta(\underset{\sim}{Z}, \eta) + 2\omega' V_c' \eta.$$

Dit levert schatters $\hat{\eta}_{II}$ en $\hat{\omega}$ op, en vervolgens $\hat{\theta}_{II} = V'\hat{\eta}_{II}$.

De (n-p)-element vector ω is een vector van onbepaalde vermenigvuldigers. Het model kan nu ook worden geschreven als:

A3 : $\omega = o$

Omdat de oplossing van het oorspronkelijke probleem uniek is, geldt nu vanzelfsprekend $\hat{\theta}_I = \hat{\theta}_{II} = \hat{\theta}$ en

$$\hat{\eta}_I = \hat{\eta}_{II} = \hat{\eta}.$$

Uit formulering I volgt dat $\hat{\theta}$ gegeven is door

$$\hat{\theta} = (V'\underset{\sim}{X}V)^{-1}V'\underset{\sim}{X}\underset{\sim}{Z},$$

en dus

$$\hat{\eta} = V(V'\underset{\sim}{X}V)^{-1}V'\underset{\sim}{X}\underset{\sim}{Z}.$$

Er volgt verder uit dat zowel $\hat{\theta}$ als $\hat{\eta}$ efficiente schatters zijn, m.a.w.

$$m^{\frac{1}{2}} \ (\hat{\theta} - \theta^o) \xrightarrow{\mathcal{L}} \mathcal{N} \ (0, \ (V'M^oV)^{-1} \ )$$

$$m^{\frac{1}{2}} \ (\hat{\eta} - \eta^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \ V(V'M^oV)^{-1}V')$$

De asymptotische dispersiematrixen kunnen geschat worden door

$$S(\hat{\theta}) = (V'\underset{\sim}{X}V)^{-1}$$

$$S(\hat{\eta}) = V(V'\underset{\sim}{X}V)^{-1}V'$$

Bovendien volgt uit de gegeven resultaten

$$\delta(z, \hat{\eta}) \xrightarrow{\mathcal{L}} \chi^2_{n-p}.$$

Formulering II geeft andere nuttige informatie. We vinden

$$\hat{\omega} = (V_c'\underset{\sim}{X}^{-1}V_c)^{-1}V_c'\underset{\sim}{Z}$$

$$\hat{\eta} = \underset{\sim}{Z} - \underset{\sim}{X}^{-1}V_c(V_c'\underset{\sim}{X}^{-1}V_c)^{-1}V_c'\underset{\sim}{Z}.$$

De vektoren $\hat{\omega}$ en $\hat{\eta}$ zijn asymptotisch onafhankelijk, en

$$m^{\frac{1}{2}}\hat{\omega} \xrightarrow{\mathcal{L}} \eta(0, \ (V_c' \ M_o^{-1}V_c)^{-1})$$

Uit vergelijking van I en II volgt ook dat we $\delta(\underset{\sim}{Z}, \hat{\eta})$ kunnen schrijven in drie verschillende vormen.

$$\begin{aligned}\delta(\underset{\sim}{Z}, \hat{\eta}) &= \underset{\sim}{Z}'\left[\underset{\sim}{X} - \underset{\sim}{X}V(V'\underset{\sim}{X}V)^{-1}V'\underset{\sim}{X}\right]\underset{\sim}{Z} \\ &= \hat{\omega}'V_c'\underset{\sim}{X}^{-1}V_c\,\hat{\omega} \\ &= \underset{\sim}{Z}'V_c(V_c'\underset{\sim}{X}^{-1}V_c)^{-1}V_c'\underset{\sim}{Z}.\end{aligned}$$

De statistiek $\delta(Z, \hat{\eta})$ wordt dus ook gevonden als we A3 toetsen in de vorm $V_c'\eta = 0$ of $\omega = 0$ door gebruik te maken van de asymptotische verdeling van $V_c'\underset{\sim}{Z}$ en $\hat{\omega}$. Deze tests worden respektievelijk de <u>Wald test</u> en de <u>Lagrange vermenigvuldiger test</u> genoemd, in deze kontekst zijn ze dus equievalent aan de Neyman methode).
Met name V_c een matrix van lage rang is, zal de Wald-test de voorkeur verdienen.

LITERATUUR

1. J.N. Darroch: 'Multiplicative and additive interaction in contingency tables', Biometrika, 1974, p. 207.

2. L.A. Goodman: 'The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications', J.A.S.A., 1970, p. 226.

3. L.A. Goodman: 'Guided and Unguided Methods for the selection of Models for a set of T Multidimensional Contingency Tables', J.A.S.A., 1973, p. 165.

4. S.J. Haberman: 'The Analysis of Frequency Data', Univ. of Chicago press, 1974.

5. H.O. Lancaster: 'The multiplicative definition of interaction', Austral. J. Statist., 1971, p. 36.

6. J. de Leeuw: 'Maximum Likelihood Estimation for Weighted Poisson Models', RN005-75, afd. Datatheorie der R.U. Leiden.

7. J.A. Nelder and R.W.M. Wedderburn: 'Generalized Linear Models', J.R. Statist. Soc. A, 1972, p. 370.

8. J. Neyman: 'Contributions to the theory of the X^2-test', Proc. of the Berkeley Symp. on Math., Statist. and Probability, 1949, p. 239.

9. R.L. Plackett: 'The Analysis of Categorical Data', Griffin, London, 1974.

BIJLAGE 1: KORREKTIE VOOR BIAS

We kunnen een korrektie voor bias toepassen door alvorens de Z_i waarden te berekenen eerst $\frac{1}{2}$ op te tellen bij de X-waarden, zodat Z_i nu gedefinieerd wordt als

$$Z_i = \ln \frac{X_i + \frac{1}{2}}{m \rho_i}$$

Waarom $\frac{1}{2}$? Welnu stel we definiëren

$$\underset{\sim}{Z} = \ln \frac{\underset{\sim}{X} + a}{m \rho}$$

(We laten in het vervolg voor het gemak even de i en de kleine superscript o weg, we definiëren ook $\mu = \rho\lambda$). Stel

$$\underset{\sim}{U} = \frac{(\underset{\sim}{X} - m\mu) + a}{m\mu} .$$

dan

$$\underset{\sim}{Z} = \eta + \underset{\sim}{U} - \tfrac{1}{2}\underset{\sim}{U}^2 + \tfrac{1}{3}\underset{\sim}{U}^3 - \tfrac{1}{4}\underset{\sim}{U}^4 + \ldots\ldots$$

Hieruit volgt

$$E(\underset{\sim}{Z}) = \eta + \frac{1}{m}\left(\frac{2a - 1}{\mu}\right) - \left(\frac{1}{m}\right)^2 \left(\frac{6a^2 - 12a + 5}{12\mu^2}\right) + o(m^{-2}).$$

Deze korrektie heeft verder als prettig neveneffekt dat Z nu ook gedefinieerd is voor X = o.

BIJLAGE 2: KOMPUTERPROGRAMMA

In het komputerprogramma is het nodig per variabele een design matrix van orthogonale kolomvektoren in te lezen waarbij de eerste kolomvektor wordt gegenereerd. De definitieve design matrix wordt in het programma gekonstrueerd m.b.v. de uitwendig produkt methode en daarna omgezet in een orthonormale matrix. Indien men dus niet in afzonderlijk effekten geïnteresseerd is, is het de meest eenvoudige methode om Helmert-effekten in te voeren. De θ's van het verzadigde model worden berekend m.b.v. de formule

$$\theta = (V' \underset{\sim}{X} V)^{-1} V' \underset{\sim\sim}{XZ}$$

Deze formule reduceert in het geval van een verzadigd model voor de orthonormale V-matrix tot

$$\theta = V' X^{-1} V V' \underset{\sim\sim}{XZ} = V' \underset{\sim}{Z}$$

de bijbehorende varianties, op grond waarvan de standaardscores zijn berekend, staan op de diagonaal van de matrix $(V'XV)^{-1}$ welke matrix voor het verzadigde geval wordt berekend als $V' X^{-1} V$ zodat inverteren niet nodig is.

Voor het toetsen van hypotheses waarbij (telkens beperkte) groepen van θ's op nul worden gesteld wordt formulering II van blz. 16 gebruikt omdat in dit geval slechts een matrix van beperkte orde dient te worden geinverteerd om

$$\delta(\underset{\sim}{Z}, \hat{\zeta}) = \underset{\sim}{Z}' V_c (V_c' \underset{\sim}{X}^{-1} V_c)^{-1} V_c' Z \text{ te krijgen}$$

De matrix $V_c' \underset{\sim}{X}^{-1} V_c$ is gegeven als deelmatrix van de al berekende matrix $V' X^{-1} V$.

BIJLAGE 3: VOORBEELD VAN EEN ANALYSE

Ter illustratie volgt hier een uitgewerkt voorbeeld.
Dit voorbeeld is gekozen vanwege de eenvoud van de tabel.
Met name vanwege het gevaar van verkeerde interpretatie van de variabele 'alkohol-gebruik' zal men deze tabel niet in het rapport terugvinden.

Gekozen is voor een drie-weg tabel waarbij de variabelen zijn:
A: Noord-Brabant tegen de Rest van Nederland
B: Alkoholgebruik gekonstateerd versus niet gekonstateerd
C: Plaats op de weg (kruispunt, rechteweg, hoek/bocht).
In de cellen van de tabel staan aantallen doden over de jaren 1971-1973 (CBS-gegevens), binnen de bebouwde kom.

		C (kr.p.)	C (r.w.)	C (h/b)
A_1 (N-Br.)	B_1 (alk)	22	48	14
	B_2 (geen alk)	243	272	48
A_2 (Rest N)	B_1	97	202	68
	B_2	1206	1442	189

Deze gegevens zijn in de analyse gewogen naar het aantal inwoners voor Noord-Brabant met faktor 18.80 en voor de Rest van Nederland met een faktor 115.08.
Bij de analyse is gebruik gemaakt van de volgende design matrix, die was opgebouwd uit Helmert-effekten:

matrix:		effekt:
$V' =$	1 1 1 1 1 1 1 1 1 1 1 1	T : totaal
	1 1 1 1 1 1-1-1-1-1-1-1	A : N-Brabant tegen Rest v. Nederland
	1 1 1-1-1-1 1 1 1-1-1-1	B : alc.gebruik wel tegen niet gekonstateerd
	1-1 0 1-1 0 1-1 0 1-1 0	C_1: kruispunt tegen rechte weg
	1 1-2 1 1-2 1 1-2 1 1-2	C_2: kr.p.+rechte w. tegen hoek/bocht
	1 1 1-1-1-1-1-1-1 1 1 1	A x B
	1-1 0 1-1 0-1 1 0-1 1 0	A x C_1
	1 1-2 1 1-2-1-1 2-1-1 2	A x C_2
	1-1 0-1 1 0 1-1 0-1 1 0	B x C_1
	1 1-2-1-1 2 1 1-2-1-1 2	B x C_2
	1-1 0-1 1 0-1 1 0 1-1 0	A x B x C_1
	1 1-2-1-1 2-1-1 2 1 1-2	A x B x C_2

Hieronder volgen de resultaten van de schattingen voor het verzadigde model.
De (2x2x3 = 12) schatters komen overeen met een totaal-effekt, de hoofdeffekten, eerste-orde interaktie-effekten en twee-orde interaktie-effekten.

	enkelvoudige scores		Chi-kwadraat-waarden	dfr (vrijheidsgr
	grenswaarde ± 1.96		95% grenswaarde	
			3.84	1
			5.99	2
totaaleffekt	+ 27.59	**	761.28	1
hoofdeffekten:				
A-effekt:	+4.02	**	16.16	1
B-effekt:	-24.24	**	587.35	1
C-effekten:	- 5.98	**	265.27	2
	+14.19	**		
eerste-orde interaktie-effekten:				
A x B effekt:	+ .41	N.S.	0.17	1
A x C effekten:	+ .10,	N.S.	0.23	2
	-.46	N.S.		
B x C effekten:	-4.04,	**	43.26	2
	-5.70	**		
tweede-orde interaktie-effekten:				
A x B x C effekten:	- .35,	N.S.	1.31	2
	+1.03	N.S.		

Scores staan in standaard-vorm. De hypothese dat de cellen identieke Poissonparameters hebben is dus zeer onwaarschinlijk: $\hat{0}_t$ (= 27.59) $\gg$ 1.96 (= de grenswaarde bij toetsen op 5%-niveau, tweezijdig). Op deze wijze is al na te gaan welke effekten signifikant zijn op bv. 5%-niveau. Deze zijn gemerkt met **. Dit toetsen kan ook plaatsvinden m.b.v. een X^2-toets. We vinden dan voor elk effekt één X^2-waarde (zie kolom 2) met een bijbehorende aantal vrijheidsgraden (zie kolom 3). Merk op dat de X^2 waarden bij dfr=1 gelijk zijn aan het kwadraat van de enkelvoudige scores.
Bij de X^2-toetsen wordt dan telkens verondersteld dat alle schatters die met het effekt overeenkomen gelijk zijn aan nul, en geeft de X^2-waarde aan hoe groot de diskrepantie tussen het zo verkregen model en de data is.
Waar we te maken hebben met één vrijheidsgraad per effekt is de signifikantie van beide toetsen per definitie identiek. Bij deze analyse geldt ook voor de andere X^2-waarden dat ze hetzelfde resultaat opleveren als de toets van de enkelvoudige scores.
Dit behoeft niet altijd zo te zijn. De enkel-voudige scores kunnen bijvoorbeeld allen (net) niet signifikant zijn, maar gezamenlijk wel een signifkante X^2-waarde opleveren. Ook is het mogelijk dat maar één enkelvoudige score signifikant is waardoor de totale X^2-waarde niet signifikant behoeft te zijn.
In dergelijke gevallen leveren de X^2-waarde en de enkelvoudige scores dus additieve informatie.

Interpretatie van de gegevens

In het algemeen zijn de hoofdeffekten en het totale effekt op zich niet zo veelzeggend bij een interpretatie van de gegevens. Hier echter, waar een korrektie voor het aantal inwoners is toegepast, is over het A-effekt op te merken dat er per inwoner minder ongevallen plaatsvinden in de Rest van Nederland dan in Noord-Brabant (de richting van het effekt blijkt uit het teken!).

Om een interpretatie te geven aan dit verschijnsel zouden we iets moeten weten over bv. de urbanisatiegraad in Noord-Brabant en de Rest van Nederland en verder op zijn minst iets over de aantallen reizigers-/voertuigkm's. Voor een interpretatie van het B x C-effekt is het van belang om zich te realiseren dat het hier gaat om gekonstateerd alkoholgebruik. Het lijkt interessant ook bebouwing hierbij te betrekken.
Uit dit alles moet duidelijk zijn dat de interpretatie van de effekten een aktiviteit is die los staat van de analyse zelf.