# HET MULTIPROPORTIONEEL POISSON-MODEL

Een reactie op het artikel "Analyse van ongevallen in verkeerssituaties met een multiproportioneel Poisson-model" van prof. dr. ir. R. Hamerslag & ir. J.P. Roos in Verkeerskunde 31 (1980) 11: 567 t/m 571

Artikel Verkeerskunde 32 (1981) 3: 124 en 125

R-81-3

Drs. S. Oppe

Voorburg, 1981

Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV

In november 1980 is in Verkeerskunde een artikel gepubliceerd "Analyse van ongevallen in verkeerssituaties met een multiproportioneel Poisson-model" van prof. dr. ir. R. Hamerslag en ir. J.P. Roos.
Gezien de nauwe relatie die bestaat tussen dit model en het onderwerp van mijn eigen bijdragen op dit gebied die in juli en december 1980 in Verkeerskunde verschenen, voel ik mij genoodzaakt een aantal opmerkingen te maken bij het bovenomschreven artikel.

1. De inleiding

Allereerst wil ik ingaan op een aantal onjuistheden in de inleiding. Hierin wordt gesteld:
"De ongevallen die incidenteel plaatsvinden zijn met "black spot" analyse moeilijk te analyseren. Beter is dit mogelijk met wiskundige modellen. Veelal gebruikt men regressie-analyse ...."
Op het gebruik hiervan wordt dan verder ingegaan. Gesteld wordt:
"Het gebruik van meervoudige lineaire regressie houdt in, dat impliciet verondersteld wordt dat de frequentie van de waarnemingsuitkomsten normaal verdeeld is. Gegeven het probleem, het gaat namelijk om de analyse van verkeerssituaties waar weinig ongevallen gebeuren, is deze veronderstelling in hoge mate onwaarschijnlijk."
Dit is zo algemeen gesteld natuurlijk niet juist.
Bij het schatten van de regressiecoëfficiënten met behulp van een kleinste-kwadratenmethode worden de best bij de gegevens passende coëfficiënten berekend. Hierbij worden in het geheel geen verdelingsassumpties gebruikt.
Pas bij het toetsen van uitspraken omtrent deze earameters wordt vaak (en inderdaad vaak al te gemakkelijk) aangenomen dat de gegevens normaal verdeeld zijn.
Het probleem bij regressie-analyse is dan ook niet zozeer deze normaal-verdelingsassumptie, dan wel de veel hachelijker aanname dat de onafhankelijke variabelen lineair samenhangen met de afhankelijke variabele. In gevallen waar het in deze context over gaat zou variantie-analyse meer voor de hand liggen.

Echter ook dan zijn er nog problemen. Allereerst is het de vraag of een kleinste-kwadratenoplossing welke bij regressie-analyse en variantie-analyse wordt toegepast bij dit soort gegevens wel zo efficiënt is, maar vooral of een daarbij steeds verondersteld additief verband wel aannemelijk is. Oppe (1979) laat zien dat het multiplicatieve model niet alleen theoretisch gezien aannemelijker is, maar ook de gegevens beter beschrijft.
Over de aantrekkelijkheid van het multiplicatieve model bestaat dus geen verschil van mening met de auteurs. We constateren wel dat in het multiproportionele Poisson-model alleen ten aanzien van het schatten van de modelparameters van de Poisson-verdeling wordt uitgegaan, maar dat bij het toetsen van hypothesen over gevonden parameters toch ook hier weer een beroep wordt gedaan op het normaal verdeeld zijn van de aantallen ongevallen. Dus bij te kleine aantallen ongevallen zal men ook bij deze analysemethode geen toetsbare hypothesen kunnen formuleren.
De schrijvers vervolgen met:
"Nog sterker doet het bezwaar van een onjuiste veronderstelling omtrent de waarnemingsdistributie zich gevoelen bij het door een logaritmische transformatie gelineariseerde multiplicatieve model. De logaritme van nul is niet gedefinieerd en een nulwaarneming kan dus niet bij het onderzoek worden betrokken."
Het lijkt hier te gaan om het log-lineaire analysemodel, maar dat is onjuist. Verondersteld wordt namelijk dat het voor de berekening van de parameters noodzakelijk is de logaritme van de geobserveerde waarden te nemen.
In het algemeen geldt zowel bij log-lineaire analyses als bij het multiproportionele Poisson-model en de gewone Chi-kwadraatanalyse van kruistabellen dat, als wordt uitgegaan van maximale aannemelijkheidsschatters, er geen verwachte waarden, c.q. parameters, gelijk mogen zijn aan nul.
Zou bijvoorbeeld bij het multiproportionele Poisson-model in formule (3.5a) $Y_k^+$ gelijk zijn aan nul voor één k, dan is daarmee ook $\hat{a}_k$ gelijk aan nul en de oplossing ongedetermineerd. De categorie k dient dan te worden verwijderd.
In principe wordt nl. in de drie genoemde gevallen dezelfde functie

gemaximaliseerd. Dit probleem is niet direct afhankelijk van de geobserveerde waarden, maar vooral van de modelspecificaties. We komen op dit punt nog terug.
Bij de variant van het log-lineaire model waarbij gemodificeerde minimum Chi-kwadraatschatters worden gebruikt in plaats van maximale aannemelijkheidsschatters (het model dat de kleinste Chi-kwadraatwaarde oplevert wordt gekozen, in plaats van het model met de grootste kans op de gevonden aantallen) wordt wel gewerkt met de logaritme van de geobserveerde aantallen. Ook deze schatters echter zijn efficiënt en asymptotisch gelijk aan de maximale aannemelijkheidsschatters.
Bewezen kan worden (vgl. o.a. De Leeuw & Oppe, 1976, Bijlage 1) dat bij deze schattingsprocedure een correctie voor bias gewenst is. Deze komt erop neer dat bij alle observaties 1/2 wordt opgeteld.
Het betreft hier niet een noodsprong van een in het nauw gebrachte onderzoeker die zijn heil zoekt in een "kunstgreep" om aan het log-probleem te ontkomen, maar een correctie voor bias, waarvan bewezen is dat hiermee betere schatters worden verkregen dan wanneer geen correctie wordt toegepast. Dat hiermee en passant het log-nul probleem verdwijnt is natuurlijk meer dan een gelukkig toeval.
De inleiding eindigt met: "Het is beter om de analysemethode te richten op het geringe aantal ongevallen". We zullen nagaan of dit met het multiproportionele Poisson-model gebeurt.

## 2. Het multiproportionele Poisson-model

We lezen hier: "Het hier geïntroduceerde multiplicatieve model is een logisch vervolg op de analyse van kruistabellen met één of twee kenmerken, waarbij de gedetailleerde informatie die aanwezig is, in zijn geheel kan worden geanalyseerd." We zullen nagaan of dit zo is. Maar eerst enige opmerkingen over de weging van kruistabellen. Aandacht hiervoor is inderdaad vrij recent, al zijn er heel wat voorbeelden te noemen zoals De Leeuw (1975), De Leeuw & Oppe (1976), Andersen (1977), Thomsen (1980), alle binnen de

context van het (verkeers-)veiligheidsonderzoek. Ook ten aanzien van deze weging is er geen verschil van mening. Deze mogelijkheid is zeer gemakkelijk. Voor de rest van het betoog is de weging echter niet relevant. We zullen er verder vanuit gaan dat alle observaties een gewicht één krijgen en $L_{klmn}$ voortaan uit de formules weg laten. Dit kan zonder verlies van informatie. De L-waarden zijn gegeven constanten en geen te schatten parameters. We zouden ze dus eigenlijk beter naar het linker lid van de vergelijking kunnen brengen, bijv. in formule (2.1). Zoals gezegd laten we ze voor het gemak maar even weg.
Formule (2.1) wordt dan $\mu_{klmn} = a_k . b_l . c_m . d_n . \ldots .$

Dit model zegt in feite dat $\mu_{klmn}$ (bijv. de onveiligheid op een bepaalde locatie) het produkt is van een aantal onafhankelijk van elkaar te beschouwen factoren. Verondersteld wordt dus dat er tussen de factoren geen enkele vorm van interactie bestaat. Er wordt alleen terloops opgemerkt dat de parameters ook kunnen worden gebruikt voor combinaties van kenmerken. Laten we aan de hand van een eenvoudig voorbeeld kijken naar de consequenties hiervan voor de "gedetailleerde informatie" die in de tabel aanwezig kan zijn. Stel we hebben de volgende drie-wegtabel:

| A1 | | | |
|---|---|---|---|
| | B1 | B2 | |
| C1 | 25 | 15 | 40 |
| C2 | 15 | 25 | 40 |
| | 40 | 40 | 80 |

| A2 | | | |
|---|---|---|---|
| | B1 | B2 | |
| C1 | 15 | 25 | 40 |
| C2 | 25 | 15 | 40 |
| | 40 | 40 | 80 |

De constructie is duidelijk: er bestaat een sterke samenhang tussen B en C. Deze is omgekeerd voor A1 en A2. Het model (2.1) wordt nu:

$$\mu_{klm} = a_k . b_l . c_m$$

Uit de formules (3.5) en (3.6) blijkt dat voor de schatting van deze parameters wordt uitgegaan van de marginale verdelingen van A, B en C.

In dit geconstrueerde geval geldt dat

$$Y^+_{1..} = Y^+_{2..} = Y^+_{.1.} = Y^+_{.2.} = Y^+_{..1} = Y^+_{..2} = 80$$

Voor alle parameters $a_1$, $a_2$, $b_1$, $b_2$, $c_1$, $c_2$, wordt gevonden dat deze gelijk zijn aan elkaar en dus $\sqrt[3]{20}$ bedragen. Voor elke $\mu_{klm}$ vinden we dan de waarde 20.
Kortom alle parameters zijn aan elkaar gelijk, de tabel levert geen enkele informatie!

Bij oppervlakkige bestudering lijkt het model behalve multiproportioneel ook multivariaat. Dit voorbeeld, en bestudering van de formules (3.5) en (3.6) laat zien, dat er slechts sprake is van een aaneengeschakelde serie enkelvoudige analyses. De parameters voor kenmerk A zijn evenredig met de marginale aantallen van A, alleen de evenredigheidsconstante hangt af van verdere uitsplitsing naar andere factoren. Dit is al bekend uit de normale Chi-kwadraattoets voor een twee-wegtabel. Om het model $\mu_{kl} = a_k . b_1$ te toetsen gebruiken we als schatters de marginalen van de tabel en als gezamelijke constante voor het produkt $a_k . b_1$ kiezen we het totale aantal observaties n, ofwel:

$$\hat{\mu}_{kl} = Y^+_{k.} \cdot Y^+_{.1} \Big/ n$$

Bij de Chi-kwadraatanalyse wordt getoetst of het model $\mu_{kl} = a_k . b_1$ de tabel voldoende beschrijft. Het model wordt verworpen als er sprake is van interactie, dus als de Chi-kwadraatwaarde, gebaseerd op de verschillen $Y_{kl} - \hat{\mu}_{kl}$, significant is. Pas als dit niet zo is wordt aangenomen dat het model de individuele gegevens voldoende beschrijft.
In het multiproportionele Poisson-model wordt echter een dergelijke toets (de $G^2$-toets) alleen gebruikt als hulpmiddel bij het kiezen van de kenmerken, met andere woorden, om te zien of het weglaten van bijvoorbeeld de a-parameters het model significant slechter maakt. Er wordt niets gezegd over de houdbaarheid van het uitgangsmodel zelf, terwijl de mogelijkheid van toetsing van dit model juist

een van de grote voordelen is van dit soort modellen t.o.v. regressie-analyse.
We hebben gezien welke consequenties dit kan hebben voor de parameterschatting en de eruit voortvloeiende schattingen voor de individuele celwaarden.
Voor alternatieve aanpakken bij grote aantallen kenmerken verwijs ik naar mijn bijdrage in Verkeerskunde juli 1980. In de daar genoemde technieken wordt gezocht naar een beschrijving van de individuele observaties in termen van de relaties tussen de diverse relevant geachte kenmerken, hetgeen juist het uitgangspunt van multivariate analyse is!

## 3. Log-lineaire analyse

Bij log-lineaire analyse gaat men uit van precies dezelfde basisaannamen als bij het multiproportionele Poisson-model, maar met twee essentiële verschillen:
1. het model heeft niet alleen parameters voor de geïsoleerde kenmerken, maar ook voor de interactie-effecten;
2. het model wordt gepresenteerd in de logaritmevorm.
Bij een log-lineaire analyse van bijv. een drie-wegtabel luidt het meest complete (verzadigde) model:

$$\mu_{klm} = 1.1^{A}_{k}.1^{B}_{l}.1^{C}_{m}.1^{AB}_{kl}.1^{AC}_{km}.1^{BC}_{lm}.1^{ABC}_{klm} \qquad (a)$$

Hierin is 1 een algemene parameter (de evenredigheidsconstante), $1^{A}_{k}$, $1^{B}_{l}$ en $1^{C}_{m}$ komen (op zo'n evenredigheidsconstante na) overeen met de parameters $a_k$, $b_l$ en $c_m$ van formule (2.1). De overige parameters zijn interactieparameters: $1^{AB}_{kl}$ voor de interactie tussen kenmerk A en B, etc., $1^{ABC}_{klm}$ voor het "unieke" van elke cel-waarde. Dit model beschrijft de celwaarden volledig (Chi-kwadraat is nul). In het algemeen wordt nu eerst getoetst of de parameters $1^{ABC}_{klm}$ in het model kunnen worden gemist (nulhypothese: de celwaarden hebben niets unieks). Met name bij black-spotanalyse lijkt dit een zeer zinnige hypothese! Op een dergelijke wijze kan worden getoetst of elke groep interactieparameters mag worden weggelaten uit het

model. Dit zal in de praktijk zelden gebeuren. Is dit echter het geval en mag het model worden beschreven als $\mu_{klm} = 1.1^{A}_{k}.1^{B}_{l}.1^{C}_{m}$, dan is beschrijving van de tabel met behulp van het log-lineaire model identiek aan die met behulp van het multiproportionele Poisson-model. De parameters en de schattingsprocedure zijn equivalent. Essentieel verschil is echter dat in het eerste geval getoetst is of het model toelaatbaar is.
Bij een log-lineaire analyse wordt het model (a) eerst herschreven tot een lineair model:

$$\ln(\mu_{klm}) = \lambda + \lambda^{A}_{k} + \ldots\ldots\ldots\ldots + \lambda^{ABC}_{klm} \qquad (b)$$

Hierin is $\lambda = \ln(1)$, $\lambda^{A}_{k} = \ln(1^{A}_{k})$, etc.
In feite komen deze parameters ook al voor in formule (3.3).
We zullen hier niet ingaan op de formules voor het schatten van deze nieuwe parameters. Zijn deze $\lambda$-parameters echter gevonden, dan zijn ze direct vertaalbaar in termen van de eerdere 1-parameters.
De herschrijving vindt zijn oorsprong in de statistische eenvoud van het lineaire model. Definieren we een lineaire vectorruimte met de $\lambda$'s als uitgangspunt voor de basis, dan is elke mogelijke tabel van uitkomsten op te vatten als een vector in deze ruimte. Het probleem is nu gereduceerd tot het zoeken van een zo klein mogelijke lineaire deelruimte waar een vector van observaties nog redelijk inpast. Deze deelruimte is direct te interpreteren in termen van hoofdeffecten en interactie-effecten.
Een model kan in matrixnotatie als volgt worden omschreven:
$\ln(\underline{\mu}) = V\underline{\lambda}$,
waarin $\ln(\underline{\mu})$ de vector van verwachte log-observaties betreft en $\underline{\lambda}$ de vector van alle parameters. De "design matrix" V bepaalt de interpretatie van de parameters. Deze matrix V kan nu zo worden gekozen dat bepaalde parameters en daarmee de veronderstelde aanwezigheid van interacties uit het model verdwijnen. Het voordeel van de presentatie met behulp van het lineaire model is verder dat eenvoudig uit de betrouwbaarheid van de observaties valt af te leiden wat de betrouwbaarheid is van de geschatte parameters, zo-

dat deze of groepen ervan kunnen worden getoetst (vgl. toetsing van interactie in een twee-wegtabel met behulp van een $X^2$-toets). Bij deze toetsing zijn veel bruikbare varianten te definieren, door een zorgvuldige keuze van de design matrix. In het algemeen geldt dat parameters die overbodig zijn, gelijk zijn aan nul. Voor de parameters kan worden nagegaan of deze significant van nul verschillen (bijv. de parameter voor klasse 1 van kenmerk A is significant hoger dan die voor de andere klassen van kenmerk A). Maken we de gebruikelijke normaal-verdelingsassumpties die gelden voor een gewone Chi-kwadraatanalyse, dan kan ook voor de $\lambda$'s worden afgeleid dat deze normaal verdeeld zijn. Het toetsen van individuele parameters of groepen van parameters is direct mogelijk. Deze aanpak van de toetsingsproblemen die in het beschouwde artikel worden beschreven, maar daar niet bevredigend worden opgelost, maakt het log-lineaire model juist zo aantrekkelijk. De scheefheid van de ratio's van parameters rond de waarde 1 vormen geen probleem meer. Monte-Carlo studies zijn daardoor overbodig (vgl. par. 5: Simulatie).

Nog wel blijft het probleem van de kleine aantallen observaties bestaan. Hoe specifieker een model is, met andere woorden, hoe hoger het interactieniveau dat verondersteld wordt in de gegevens aanwezig te zijn, hoe meer er uitgesplitst moet worden. Bij de toetsingen zoals hier genoemd, dus ook bij het multiproportionele Poisson-model, wordt geen specifiek gebruik gemaakt van de sterke aannamen van Poisson verdeelde of multinomiaal verdeelde observaties. Er wordt steeds van uitgegaan dat de aantallen zo groot zijn, dat (met behulp van de centrale-limietstelling) mag worden aangenomen dat deze aantallen observaties (of sommen ervan) normaal verdeeld zijn. Wil men hieraan echt iets verbeteren, dan zou men moeten zoeken naar generaliseringen van bijv. Fisher's exacte toets voor 2x2-tabellen. Pas dan wordt de pretentie echt waar gemaakt dat het beter is "om de analysemethode te richten op het geringe aantal ongevallen", zoals aan het slot van de inleiding wordt gesteld.

LITERATUUR

Andersen, E.B. (1977). Multiplicative Poisson models with unequal cell rates. Scand. J. Statist. 4.

De Leeuw, J. (1975). Maximum likelihood estimation for weighted Poisson models. RN005-75. Rijksuniversiteit Leiden, Afd. Datatheorie, Leiden, 1975.

De Leeuw, J. & Oppe, S. (1976). Analyse van kruistabellen: Loglineaire Poisson-modellen voor gewogen aantallen. R-76-8. SWOV, Voorburg, 1976.

Oppe, S. (1978). The use of multiplicative models for analysis of road safety data. R-78-18. SWOV, Voorburg, 1978. Ook in: Accid. Anal. & Prev. 11 (1979) 2 (June) 101-115.

Thomsen, L.K. (1980). Statistik analyse of faerdselulykker. IMSOR, Lyngby, 1980.