

SMOOTHING EN INTERPOLATIE MET GEGENERALISEERDE LINEAIRE MODELLEN

P.H.C. Eilers*

Samenvatting. Het gegeneraliseerde lineaire model van Nelder en Wedderburn wordt gecombineerd met discrete smoothing, waarin eindige differenties als maat voor het gladde verloop van een reeks dienen. Zo ontstaat een gemakkelijk te hanteren en snel algoritme voor het smoothen en interpoleren van reeksen niet-normaal verdeelde variabelen. Toepassing wordt beschreven op het smoothen van reeksen tellingen en frekwentieverdelingen, alsmede op het interpoleren, met smoothing, van binaire dosis-response curven. Ook wordt toepassing op het gelijktijdig smoothen van schaal en locatie beschreven.

Dienst Centraal Milieubeheer Rijnmond
's Gravelandseweg 565
3119XT Schiedam
010-4273217

1 Inleiding

Bij een groot aantal statistische problemen, zoals de analyse van tijdreeksen, heeft men behoefte aan smoothing, interpolatie, of een combinatie van beide. Er is geen gebrek aan algoritmen voor dit doel. Het merendeel daarvan is gebaseerd op het kleinste kwadraten criterium, als maat voor de overeenkomst tussen de gegevens en een afgeleide reeks. Voor veel soorten gegevens is dit geen optimaal criterium, met name voor tellingen, binaire waarnemingen en scheef verdeelde variabelen. Dit probleem speelt niet alleen bij smoothing, maar bij tal van toepassingen waar men niet-normaal verdeelde variabelen in hun afhankelijkheid van covariabelen onderzoekt.

Door de introductie van het gegeneraliseerde lineaire model (GLM) gaven Nelder en Wedderburn (1972) een zeer algemeen toepasbare oplossing. De theorie van het GLM is elegant en betrekkelijk eenvoudig. Het schatten van parameters kan plaatsvinden d.m.v. iteratieve gewogen lineaire regressie. Dat vergemakkelijkt de implementatie voor praktische toepassingen. Het door Baker en Nelder (1978) ontwikkelde programma GLIM maakt het mogelijk met geringe programmeerinspanningen complexe modellen interactief te onderzoeken.

In de loop der jaren is het GLM op diverse manieren uitgebreid en nog algemener gemaakt; een goed overzicht geven McCullagh en Nelder (1983) en de verwijzingen daarin. Het merendeel van de uitbreidingen en toepassingen betreft modellen waarin de volgorde der waarnemingen niet van belang is. Recentelijk is het GLM gecombineerd met het Kalman-filter (West, Harrison en Migon, 1985) en met splines (Green en Yandell, 1985; O'Sullivan, Yandell en Raynor, 1986).

In dit artikel wordt het GLM gecombineerd met een algoritme voor smoothing en interpolatie dat gebruik maakt van eindige differenties (Eilers, 1987). Zo ontstaat een flexibel gereedschap — GLS: generalized linear smoothing — voor het analyseren van reeksen niet-normaal verdeelde waarnemingen.

Na een korte beschrijving, in §2 van de basis van het GLM, het smoothing-algoritme en de combinatie van beide, beschrijft §3 een aantal toepassingen: het smoothen van empirische frekwentieverdelingen en van reeksen tellingen, alsmede smoothing en interpolatie van binaire dosis-response gegevens. Tot slot wordt beschreven hoe smoothing van residuen van een regressiemodel plaats kan vinden en hoe daaruit een algoritme voor gelijktijdig smoothen van locatie en schaal volgt.

2 De theoretische bouwstenen

De theoretische bouwstenen zijn de familie van exponentiële verdelingen, (penalized) maximum likelihood schattingen en discrete smoothing. Deze worden hieronder kort beschreven en gecombineerd.

2.1 Exponentiële verdelingen

De familie van exponentiële verdelingen vormt de basis van het door Nelder en Wedderburn (1977) geïntroduceerde gegeneraliseerde lineaire model (GLM). De verdeling van variabele

y_i wordt exponentieel genoemd als ze geschreven kan worden als

$$f(y_i) = \exp[\{y_i\theta_i - b(\theta_i)\}/a(\phi_i) + h(y_i, \phi_i)]. \quad (1)$$

Hierin heeft θ_i de rol van de locatieparameter en ϕ_i die van schaalparameter. In het vervolg nemen we aan dat ϕ_i bekend, dan wel constant is. Om de notatie te vereenvoudigen schrijven we verder a_i voor $a(\phi_i)$, b_i voor $b(\theta_i)$ en h_i voor $h(y_i, \theta_i)$.

Er geldt:

$$E(y_i) = \mu_i = \partial b_i / \partial \theta_i; \quad (2)$$

$$E(y_i - \mu_i)^2 = \sigma_i^2 = a_i \partial^2 b_i / \partial \theta_i^2. \quad (3)$$

Als verklarende variabelen x_{ij} , $j = 1 \dots n$, bekend zijn, dan stelt het GLM een "link"-functie $t(\cdot)$ en een parameter-vector c voor, zó dat

$$\theta_i = t\left(\sum_{j=1}^n x_{ij}c_j\right). \quad (4)$$

De functie $t(\cdot)$ kan men in principe willekeurig kiezen, maar meestal gebruikt men de "natuurlijke link", die neerkomt op

$$\theta_i = \sum_{j=1}^n x_{ij}c_j. \quad (5)$$

Die keuze leidt niet alleen tot vereenvoudiging van de schattingsvergelijkingen en de berekeningen die daaruit volgen, maar garandeert ook dat de likelihood een convexe functie van c is (McCullagh en Nelder, 1983). In de afleidingen die volgen wordt de algemene vorm van $t(\cdot)$ gehanteerd, maar in de toepassingen wordt alleen de natuurlijke link gebruikt.

2.2 Penalized maximum likelihood schattingen

Voor het schatten van de parameters van een GLM wordt de maximum likelihood (ML) methode gebruikt. Voor een steekproef met m waarnemingen is de log-likelihood

$$L = \sum_{i=1}^m (y_i\theta_i - b_i)/a_i + \sum_{i=1}^m h_i = \sum_{i=1}^m L_i. \quad (6)$$

De parameter-vector c die L maximaal maakt is de ML-schatter. Nelder en Wedderburn (1972) tonen aan dat ML-schattingen voor een GLM berekend kunnen worden met iteratief gewogen lineaire regressie.

Stel nu dat we evenveel parameters, aan te duiden met z_i , introduceren als er waarnemingen zijn:

$$\theta_i = t(z_i). \quad (7)$$

In het algemeen is dit niet zo zinvol. Maximaliseren van L maakt dan μ_i gelijk aan de waargenomen y_i . De situatie verandert als we een "penalty" $P(z)$ introduceren, die beperkingen oplegt aan de waarden die de z_i 's aan kunnen nemen.

Een "penalized maximum likelihood" (PML) schatter maakt

$$L^* = L - \beta P(z) \quad (8)$$

minimaal. Met de waarde van de parameter β kan de gebruiker de invloed van de beperking op z instellen. Als de waarnemingen op gelijke onderlinge afstanden staan bestaat er een eenvoudige en zeer bruikbare vorm voor $P(z)$. Deze wordt hieronder besproken.

2.3 Discrete smoothing

Als een reeks met waarden z_i , $i = 1 \dots m$ gegeven is, dan is de tweede differentie (Fröberg, 1969) gegeven door

$$\Delta^2 z_i = z_i - 2z_{i-1} + z_{i-2}. \quad (9)$$

Hoe sterker de reeks fluctueert, des te groter de waarden die $\Delta^2 z_i$ aanneemt. Als penalty op het fluctueren van de reeks is derhalve

$$P(z) = \sum_{j=3}^m (\Delta^2 z_i)^2 \quad (10)$$

geschikt. Dit is al onderkend door Whittaker (1923), die onder andere voorstelde het smoothen van een reeks waarnemingen y te formuleren als het minimaliseren van

$$S = \sum_{i=1}^m (y_i - z_i)^2 + \beta \sum_{j=3}^m (\Delta^2 z_i)^2 \quad (11)$$

(Whittaker's voorstel was algemener: hij beschouwde ook hogere differenties dan de tweede). Hoe groter β , des te sterker de smoothing.

Men kan (11) algemener maken door in beide termen gewichten te introduceren:

$$S = \sum_{i=1}^m w_i (y_i - z_i)^2 + \beta \sum_{j=3}^m v_i (\Delta^2 z_i)^2. \quad (12)$$

Ontbrekende waarnemingen — onder andere bij interpolatie — laten zich door $w_i = 0$ te stellen eenvoudig verwerken. Knikken en sprongen in de reeks z kan men toelaten door plaatselijk $v_i = 0$ te stellen. Voor details raadplege men (Eilers, 1987). In dit artikel wordt alleen $v_i = 1$ voor $i = 3 \dots m$ en $v_i = 0$ voor alle andere waarden van i gebruikt.

2.4 Gegeneraliseerd lineair smoothen

Door het GLM te combineren met discrete smoothing ontstaat een flexibele techniek voor het smoothen en interpoleren van reeksen waarnemingen uit exponentiële verdelingen, die we gegeneraliseerd lineair smoothen (GLS) dopen. De te maximaliseren PML doelfunctie is

$$L^* = 2 \sum_{i=1}^m w_i L_i - \beta \sum_{j=3}^m (\Delta^2 z_i)^2. \quad (13)$$

Voor de partiële afgeleiden van L^* volgt nu:

$$c''_i z_{i-2} + c'_i z_{i-1} + c_i z_i + c'_{i+1} z_{i+1} + c''_{i+2} z_{i+2} = w_i t'(z_i)(y_i - \mu_i)/\beta, \quad (14)$$

waarin

$$c''_i = v_i; \quad c'_i = -2v_i - 2v_{i+1}; \quad c_i = v_i + 4v_{i+1} + v_{i+2}. \quad (15)$$

Hier is μ_i de verwachting van y_i als $\theta_i = t(z_i)$. Om de notatie te vereenvoudigen schrijven we

$$D_i^4 z_i = c''_i z_{i-2} + c'_i z_{i-1} + c_0 z_i + c'_{i+1} z_{i+1} + c''_{i+2} z_{i+2}. \quad (16)$$

2.5 Oplossing van de vergelijkingen met Newton iteraties

In het algemeen zullen de rechterzijden van de vergelijkingen in (14) niet lineair in z zijn. Evenals bij het GLM zoeken we een oplossing door middel van Newton iteraties. Stel dat een benadering $\hat{z}_i$ voor elke z_i gevonden is, die de verwachtingswaarde $\hat{\mu}_i$ oplevert. Stel verder

$$q_i = t'(z_i)(y_i - \mu_i)/a_i. \quad (17)$$

Dan volgt dat

$$q'_i = \partial q_i / \partial z_i = \{t''(z_i)(y_i - \mu_i) - t'(z_i)\mu'_i\}/a_i. \quad (18)$$

Het is gebruikelijk, zie Nelder en Wedderburn (1977), een vergelijking als (18) te vereenvoudigen door er de verwachting van te nemen. Daardoor verdwijnt de factor $(y_i - \mu_i)$, met als resultaat:

$$q'_i \approx -t'(z_i)\mu'_i = -t'(z_i)\sigma_i^2 \quad (19)$$

Als we nu schrijven

$$q_i(z_i) \approx q_i(\hat{z}_i) + q'(\hat{z}_i)(z_i - \hat{z}_i), \quad (20)$$

dan ontstaat het stelsel

$$\beta D_i^4 z_i + z_i \hat{\sigma}_i^2 t'(\hat{z}_i) = w_i t'(\hat{z}_i)(y_i - \hat{\mu}_i + \hat{\sigma}_i^2 \hat{z}_i). \quad (21)$$

Vergelijking met discrete smoothing leert dat het oplossen van (21) neerkomt op het smoothen van de reeks $u_i = y_i - \hat{\mu}_i + \hat{\sigma}_i^2 \hat{z}_i$ met effectieve gewichten $\hat{w}_i = w_i a_i t'(\hat{z}_i)$. De vergelijkingen zijn lineair en ze hebben een bandvorm; ze zijn daardoor efficiënt op te lossen.

Bij de bespreking van de toepassingen zal de specifieke vorm die (21) aanneemt voor de Poisson-, de binomiale en de chi-kwadraat-verdeling gegeven worden. Daar zal ook besproken worden wat het effect is van zeer sterke smoothing bij ieder van deze verdelingen. Bij zeer sterke smoothing is nl. de tweede term in (13) allesoverheersend. Dat betekent dat de tweede differentie van de reeks z overal vrijwel nul zal zijn, zodat die reeks een rechte lijn vormt.

2.6 Behoud van momenten

Door de coëfficiënten-matrix van $D_i^4 z_i$ uit te schrijven kan men controleren dat geldt:

$$\sum_{i=1}^m D_i^4 z_i = 0; \quad \sum_{i=1}^m i D_i^4 z_i = 0. \quad (22)$$

Als we dit combineren met (21) volgt:

$$\sum_{i=1}^m t'(z_i)(y_i - \mu_i)/a_i = 0; \quad \sum_{i=1}^m i t'(z_i)(y_i - \mu_i)/a_i = 0. \quad (23)$$

Voor het geval dat $t(z_i) = z_i$ en a_i constant volgt uit (23) dat de som van de waarnemingen gelijk is aan de som van de gesmoothede reeks. Ook volgt dat de eerste momenten van beide reeksen gelijk zijn. Deze eigenschappen zijn vooral van belang bij het smoothen van frekwentieverdelingen.

2.7 Quasi-likelihood

Uit (21) volgt dat de informatie over de exponentiële verdeling wordt samengevat in de grootheden μ_i en σ_i^2 . Daar volgt uit dat men met een eenvoudiger specificatie kan volstaan, namelijk één die μ_i en σ_i^2 als functies van de parameter θ_i geeft. Dit is in zijn algemeenheid al door Wedderburn (1974) vastgesteld. Hij introduceerde de benaming "quasi-likelihood" voor deze benadering.

3 Toepassingen

3.1 Het Poisson-model voor een tijdreeks van tellingen

De Poisson-verdeling laat zich schrijven als

$$f(y_i) = \exp(y_i \ln \mu_i - \mu_i - \ln y_i!). \quad (24)$$

Met $t(z_i) = z_i$ volgt daaruit:

$$a_i = 1; \quad \theta_i = \ln \mu_i; \quad \mu_i = \exp z_i; \quad \mu_i' = \sigma_i^2 = \mu_i. \quad (25)$$

De lineair te smoothen reeks u en de bijbehorende effectieve gewichten $\hat{w}$ worden dan:

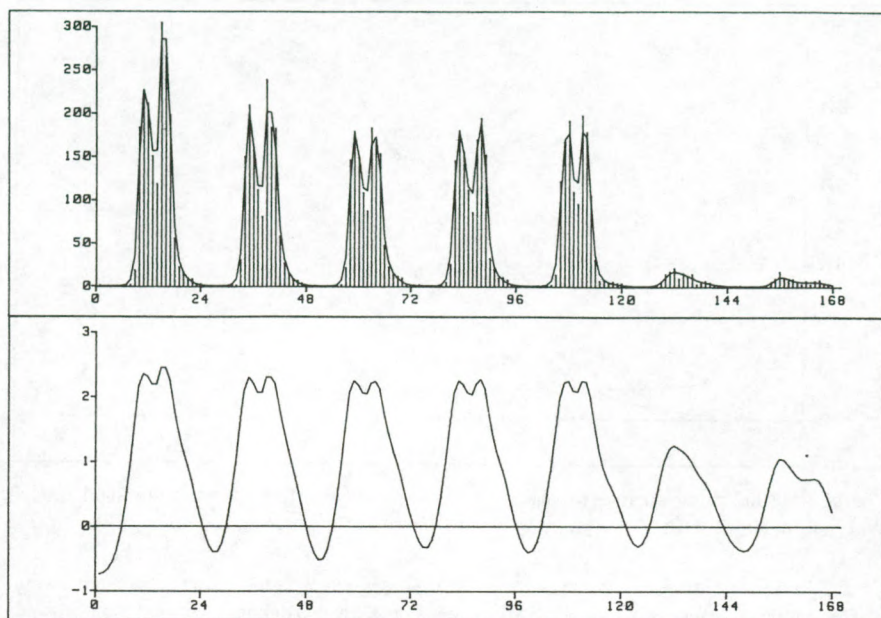
$$u_i = y_i - e^{\hat{z}_i} + \hat{z}_i e^{\hat{z}_i}; \quad \hat{w}_i = e^{\hat{z}_i}. \quad (26)$$

Een toepassing laat figuur 1 zien. De cijfers zijn ontleend aan West (1985); ze beschrijven het aantal binnenkomende gesprekken in een bedrijfstelefooncentrale, als uurtotalen gedurende één week. De ruwe gegevens zijn als vertikale lijnen in het bovenste deel van de figuur getekend. De gesmoothede reeks e^z voor $\beta = 1$ is daar als een vloeiende lijn doorheen getrokken. In het onderste deel van de figuur is de reeks z zelf getekend. Als startwaarden voor de iteraties diende $\hat{z}_i = \ln(y_i + 1)$.

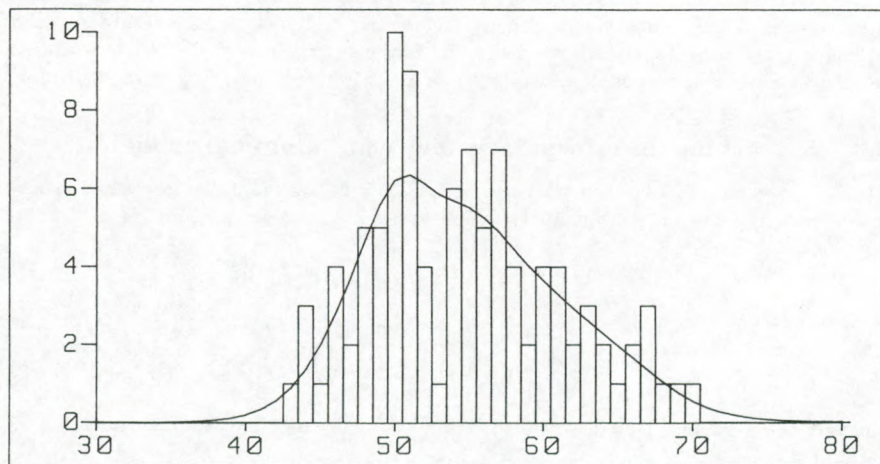
3.2 Smoothing van een frekwentieverdeling

Smoothing van een frekwentieverdeling gaat op dezelfde manier als smoothing van een tijdreeks van tellingen. De data worden aangeboden als tellingen y_i in intervallen van gelijke breedte, die met i geïndexeerd zijn. Een voorbeeld geeft figuur 2. De gegevens zijn breedten van kelkbladeren van *Iris setosa* en *Iris versicolor* (Kendall en Stuart, 1966); het aantal waarnemingen is 100 en de intervallen zijn 0.1 breed. De waarde van β is 100. De startwaarden waren $z_i = \ln(y_i + 1)$.

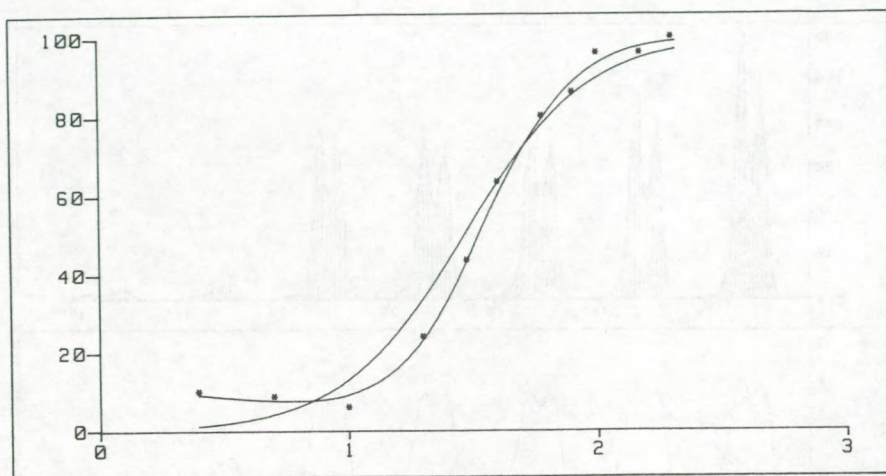
Een voordeel van deze wijze van smoothen is, dat gegarandeerd positieve waarden berekend worden voor de gladgestreken verdeling. Bij directe lineaire smoothing kan men daar niet zeker van zijn. Good en Gaskins (1980) werken met de wortel uit de tellingen en de gladde verdeling. Dat leidt zeker tot positieve gesmoothede waarden, maar er kunnen ongewenste knikken ontstaan waar de wortel van teken verandert. Het kwadraat van een glad verlopende reeks die zowel positieve als negatieve waarden bevat verloopt nl. niet glad.



Figuur 1: Boven: een tijdreeks van tellingen (vertikale lijnen) en de gesmoothede reeks ($\beta = 10$). Onder: de reeks z .



Figuur 2: Smoothing van een frekwentieverdeling met een Poisson-model ($\beta = 100$).



Figuur 3: Smoothing en interpolatie van dosis-response data met een binomiaal model. Waarnemingen en gefitte curven voor $\beta = 10^4$ en $\beta = 10^{12}$.

Bij smoothing op de beschreven wijze geldt dat de som van de waarnemingen en het eerste moment gelijk zijn aan die van de gesmoothede verdeling. Dit betekent dat de verwachtingswaarde van beide verdelingen gelijk is.

Bij zeer sterke smoothing volgen de z_i 's een rechte lijn: $z_i \approx c_0 + ic_1$ dus $\mu_i \approx e^{c_0 + ic_1}$. De gesmoothede verdeling is dan dus vrijwel gelijk aan een discrete exponentiële verdeling, met verwachtingswaarde gelijk aan het gemiddelde van de gegeven verdeling. Men kan aantonen dat bij smoothing met derde differenties de limiet bij hoge waarden van β en kleine intervallen de normale verdeling is. Dan geldt tevens dat de variantie van de gesmoothede verdeling, voor elke waarde van β , gelijk is aan die van de gegeven verdeling.

3.3 Smoothing en interpolatie met een binomiaal model

Als steekproeven met aantallen waarnemingen $n_i, i = 1 \dots m$, aantallen successen y_i vertonen, en we postulieren een binomiale verdeling met kans op succes p_i , dan geldt:

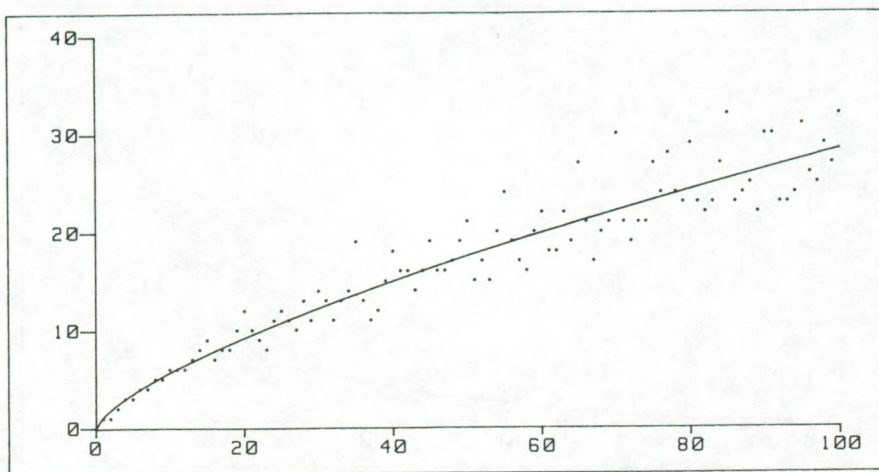
$$f(y_i) = \exp\{y_i \ln p_i + (n_i - y_i) \ln(1 - p_i) + \ln \binom{n_i}{y_i}\}. \quad (27)$$

Door te introduceren

$$\theta_i = \ln\left(\frac{p_i}{1 - p_i}\right), \quad b(\theta_i) = b_i = n_i \ln(1 - p_i), \quad (28)$$

ontstaat een verdeling met de vorm (1). Met $z_i = \theta_i$ volgt dan voor verwachting en variantie:

$$\mu_i = n_i / (1 + e^{-z_i}); \quad \sigma_i^2 = n_i p_i (1 - p_i). \quad (29)$$



Figuur 4: Goldbach-getallen $G(6i)$. De vloeiende lijn geeft de trend $y_i = 1.13i^{0.7}$.

Uit de laatste vergelijking volgt dat bij zeer sterke smoothing, als de reeks z een rechte lijn vormt, de logistische curve ontstaat:

$$\mu_i = n_i / (1 + e^{c_0 + c_1 i}). \quad (30)$$

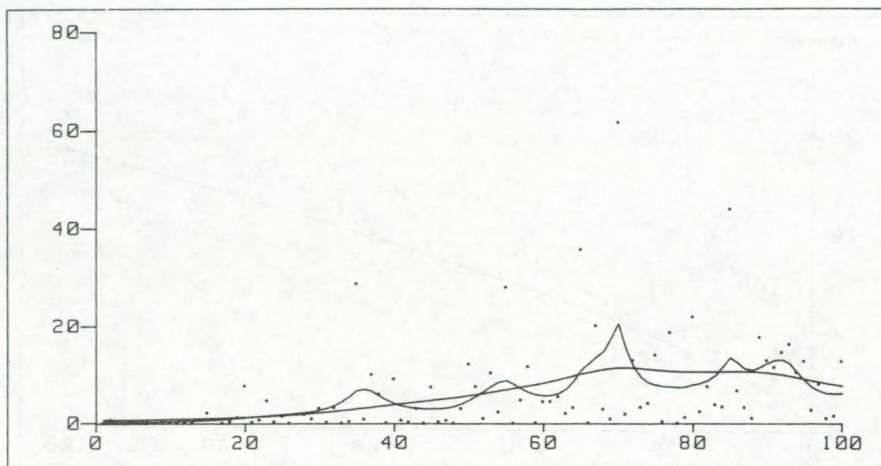
Figuur 3 geeft een voorbeeld. De data zijn ontleend aan Hewlett en Plackett (1979). De horizontale as geeft de logaritme van de toegepaste dosis van een bestrijdingsmiddel, de verticale as geeft het percentage gedood ongedierte in steekproeven van ca. 50 groot. De data zijn als punten uitgezet. Voor het toepassen van GLS is de x -as verdeeld in 50 intervallen. Waar een dosis in een interval valt worden n_i , de grootte van de steekproef, en y_i , het aantal gedode proefdieren, voor dat interval genoteerd. Voor intervallen waar geen dosis in valt, geldt $n_i = y_i = 0$. Verder zijn geen speciale maatregelen nodig, omdat het GLS-algoritme automatisch interpoleert. In figuur 3 zijn de resultaten voor $\beta = 10^4$ en $\beta = 10^{12}$ getekend. Het laatste geval komt effectief overeen met de logistische curve; deze mist het plateau bij lage doses, dat bij $\beta = 10^4$ wel gevolgd wordt.

3.4 Smoothing van residuen met de chi-kwadraat verdeling

De som s van de kwadraten van n normaal verdeelde onafhankelijke variabelen, met verwachting 0 en variantie 1, heeft een χ^2 -verdeling. Met $n = 2k$ wordt deze gegeven door:

$$f(s) = (s/2)^{k-1} e^{-s/2} / \Gamma(k). \quad (31)$$

Stel nu dat een reeks y gegeven is, waarvan aangenomen wordt dat y_i de som is van de kwadraten van $n_i = 2k_i$ normaal verdeelde variabelen met verwachting 0 en variantie τ_i . We zoeken een glad verlopende reeks die de trend van τ weergeeft.



Figuur 5: Kwadraten van de residuen (stippen) en twee trendlijnen die volgen uit smoothing met de χ^2 -verdeling ($\beta = 10^3$ en $\beta = 10^5$).

De verdeling van y_i is gegeven door:

$$f(y_i) = \left(\frac{y_i}{2\tau_i}\right)^{k_i-1} \exp\left(-\frac{y_i}{2\tau_i}\right) / 2\Gamma(k_i). \quad (32)$$

Deze uitdrukking is in de vorm (1) te schrijven als:

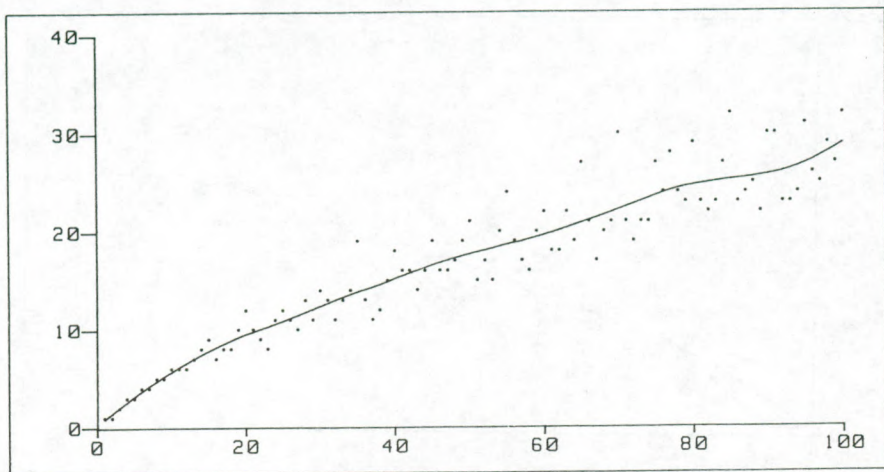
$$\theta_i = -1/2\tau_i; \quad b(\theta_i) = b_i = -k_i \ln(-\theta_i). \quad (33)$$

Hieruit volgt dan, met $\theta_i = z_i$:

$$\mu_i = -k_i/z_i; \quad \mu'_i = k_i/z_i^2. \quad (34)$$

De iteratie-vergelijkingen volgen hier direkt uit; voor de praktisch implementatie is het iets handiger te kiezen voor $\tau_i = n_i/z_i$, als $n_i = 2k_i$ het aantal vrijheidsgraden van y_i is.

Een voor de hand liggende toepassing van smoothing met een χ^2 -verdeling is het onderzoeken van heteroscedasticiteit van residuen van een regressiemodel. Dat wordt hieronder geïllustreerd met de zgn. Goldbach-getallen (Mosteller, 1972). Het Goldbach-getal $G(n)$ is het aantal manieren waarop een even getal n geschreven kan worden als de som van twee priemgetallen. Strikt genomen zijn gehele getallen niet de juiste bron voor de simulatie van normaal verdeelde residuen, maar als illustratie zijn ze goed te gebruiken; bovendien zijn de Goldbach-getallen eenvoudig te reproduceren. Figuur 4 geeft de reeks $G(6i)$. Presentatie op logaritmische x - en y -assen (niet afgebeeld) suggereert een trend die bij benadering evenredig is aan $i^{0.7}$. In figuur 4 is de trendlijn $1.13i^{0.7}$ getekend; deze is met ongewogen regressie bepaald. De kwadraten van de residuen t.o.v. deze trendlijn zijn in figuur 5 getekend.



Figuur 6: Gelijktijdige smoothing van locatie en schaal voor de Goldbach-getallen $G(6i)$ ($\beta_1 = 30$ en $\beta_2 = 10^5$).

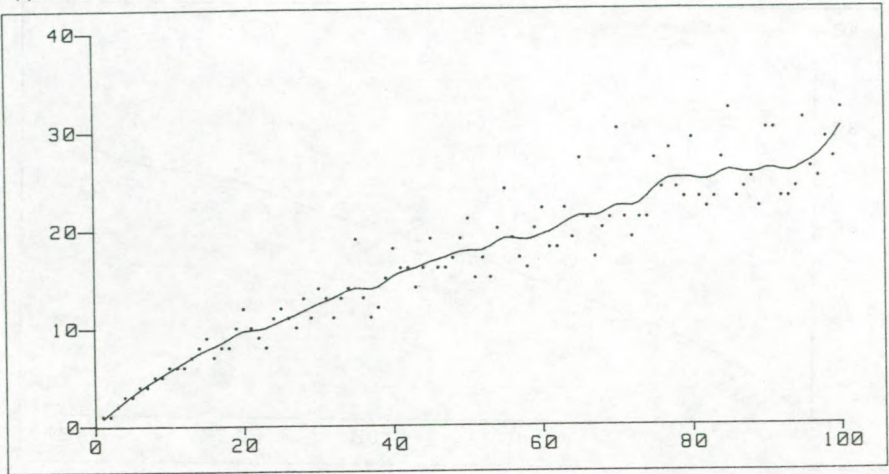
In dit geval is er voor ieder i een waarneming y_i en geldt $n_i \equiv 1$. Twee trendlijnen voor het verloop van τ zijn getekend, voor $\beta = 10^3$ en $\beta = 10^5$. De eerste verloopt duidelijk te grillig, maar illustreert de mogelijkheden van deze wijze van smoothen.

Bij veel toepassingen zullen residuen niet een reeks op gelijke afstanden vormen, maar min of meer willekeurig verdeeld over de x -as optreden. In dat geval past men dezelve techniek toe als hierboven bij de dosis-response relatie. De x -as wordt in m intervallen verdeeld en voor ieder interval vormt men y_i als de som van de kwadraten van de residuen en n_i als het aantal vrijheidsgraden. Waarschijnlijk zullen bepaalde intervallen $n_i = 0$ opleveren, omdat er geen residuen in vallen. Dat is geen probleem: er treedt automatisch interpolatie op, analoog aan de smoothing van de dosis-response curve.

3.5 Gelijktijdige smoothing van locatie en schaal

Als uit het verloop van de trend van τ blijkt dat de varianties van residuen niet constant zijn, dan is het efficiënter gewogen regressie toe te passen. Een mogelijkheid is, de gevonden trend te beschrijven met een parametrisch model. Maar het is ook mogelijk de door smoothing gevonden non-parametrische trend zelf te gebruiken. Als de residuen t.o.v. een non-parametrische trendlijn berekend zijn, dan komt men zo tot een techniek voor dubbele smoothing. Enerzijds wordt de locatie ("het gemiddelde") door smoothing bepaald, anderzijds de schaal (de variantie van de residuen).

Voor een nauwkeurige beschrijving moet de notatie iets aangepast worden, omdat nu twee reeksen tegelijk aan de orde zijn. Stel dat een reeks y gegeven is. De trend van y



Figuur 7: Smoothing van de locatie voor Goldbach-getallen $G(6i)$ met vaste gewichten ($\beta = 10.1$)

wordt geschat met de reeks z . De doelfunctie is:

$$S = \sum_{i=1}^m w_i (y_i - z_i)^2 + \beta_1 \sum_{j=3}^m (\Delta^2 z_i)^2. \quad (35)$$

De gewichten w_i zijn nog nader te bepalen uit de smoothing van de residuen. De kwadraten van de residuen $r_i = y_i - z_i$ worden gesmoothed met de χ^2 -verdeling zoals boven beschreven, met parameter β_2 ; het aantal vrijheidsgraden van r_i^2 is 1. De reeks t wordt gebruikt voor GLS van de reeks r^2 , dus de variantie van r_i wordt gegeven door $\tau_i = 1/t_i$ (de boven aangegeven praktische aanpassing wordt gebruikt).

Gewogen regressie is optimaal als de gewichten omgekeerd evenredig zijn aan de varianties van verstoringen. In het onderhavige geval zijn die varianties geschat als τ_i . Bijgevolg zijn de optimale gewichten evenredig aan t_i . De procedure wordt dus als volgt: bereken z , vervolgens r^2 en daaruit t , begin een nieuwe ronde met $w_i = t_i$. In de praktijk is gebleken dat één iteratie van de smoothing met χ^2 per ronde voldoende is.

Ter illustratie geeft figuur 6 nogmaals de Goldbach getallen. De trend is nu bepaald met dubbele smoothing; $\beta_1 = 30$ en $\beta_2 = 10^5$. Na 7 iteraties veranderden de z_i 's niet meer dan 10^{-6} . De startwaarden voor w waren $w_i = 100$.

Om het verschil met ongewogen smoothing duidelijk te maken, is in figuur 7 het resultaat daarvan getekend. Daar is $w_i = 1$ en $\beta = 10.1$, zijnde het gemiddelde van βw_i in figuur 6. Het is duidelijk te zien dat in figuur 7 bij hogere waarden van i slingeren op gaan treden, die ontbreken in figuur 6.

4 Discussie

Smoothing met gegeneraliseerde lineaire modellen is eerder voorgesteld. Green en Yandell (1985) en O'Sullivan, Yandell en Raynor (1986) gebruiken kubische splines; de eersten noemen eindige differenties als alternatief, maar passen die niet toe. Hun model is uitgebreider, doordat het ook extra verklarende variabelen inbouwt. In het hier voorgestelde model is dat op analoge wijze te verwezenlijken. Hastie en Tibshirani (1986) gebruiken een "nearest-neighbour smoother", in feite een lokale gewogen lineaire regressielijn. Voor het smoothen van frekwentieverdelingen heeft Simonoff (1983) een voorstel gedaan dat in feite neerkomt op het gebruiken van de kwadraten van eerste differenties van de reeks z als penalty.

De resultaten van deze verschillende modellen zullen elkaar bij smoothing niet veel ontlopen. Een voordeel van splines met een beperkt aantal knooppunten is, dat men de knooppunten en de parameters van de splines als een compacte samenvatting van de gesmoothede reeks kan geven. Voor interpolatie zullen de verschillen tussen splines en eindige differenties gering zijn; de "nearest-neighbour smoother" lijkt voor dit doel ongeschikt.

De voorbeelden van GLS die hier boven besproken werden, geven een indruk van het brede toepassingsterrein. Zo is GLS ook toegepast op het smoothen van Fourier-spectra van tijdreeksen met de χ^2 -verdeling en van tijdreeksen met Gamma-verdeelde waarden. Een ander voorbeeld is een reeks van tellingen die te sterke variaties vertoonden om aan een Poisson-verdeling te voldoen. Daarvoor bleek GLS met een negatief-binomiale verdeling te voldoen.

Voor praktische toepassingen is het van belang dat men de benodigde berekeningen snel en gemakkelijk uit kan voeren. Dit gaat voor GLS zeker op. Een programma in een hogere taal vraagt niet meer dan enkele tientallen regels en het aantal benodigde iteraties is meestal kleiner dan 10.

Implementatie binnen GLIM is minder aantrekkelijk, omdat men daarbij niet kan profiteren van een efficiënt algoritme voor het opslaan en oplossen van bandvormige vergelijkingen. Als expliciet de m bij m matrix opgeslagen en geïnverteerd moet worden, geeft dat mogelijk problemen met geheugenbeslag en rekentijd.

5 Programmatuur

De voorbeelden zijn berekend met programma's in BASIC voor een Hewlett-Packard HP9816 computer. Voor het smoothen van een frekwentieverdeling en smoothing en interpolatie van dosis-response gegevens zijn voorbeeldprogramma's afgedrukt in (Eilers, 1987). Voor afdrukken van deze en de andere programma's kan men zich wenden tot de auteur.

Referenties

- Baker R. J., Nelder J. A. (1978) *The GLIM System, Release 3, Generalized Linear Interactive Modeling*. Numerical Algorithms Group.
- Eilers P. H. C. (1987) Een uniform algoritme voor discrete smoothing en interpolatie. *Kwantitatieve Methoden* **24**, 115-126

- Eilers P. H. C. (1987) Het Digitale Strijkijzer, in: *Symposium Statistische Software 1987*, Rekencentrum der Rijksuniversiteit Groningen.
- Fröberg C. E. (1969) *Introduction to Numerical Analysis*. Addison-Wesley.
- Good I. J., Gaskins R. A. (1980) Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data. *J. Am. Stat. Ass.* **75**, 42-56.
- Green P. J. en B. S. Yandell (1985) Semi-parametric Generalized Linear Models, in: *Generalized Linear Models*, Proceedings Lancaster 1985, R. Gilchrist, B. Francis en J. Whittaker (eds.), Springer.
- Hastie T. en R. Tibshirani (1986) Generalized Additive Models. *Statistical Science* **1**, 297-318.
- Hewlett P. S. en R. L. Plackett (1979) *The Interpretation of Quantal Responses in Biology* Edward Arnold.
- Kendall M. G. en M. M. Stuart (1984) *The Advanced Theory of Statistics Vol. 3*, Griffin.
- McCullagh P. en J. A. Nelder (1983) *Generalized Linear Models*, Chapman and Hall.
- Mosteller F. (1972) A Data-Analytic Look at Goldbach Counts. *Statistica Neerlandica* **26**, 227-242.
- Nelder J. A. en R. W. M. Wedderburn (1972) Generalized Linear Models, *J. R. Stat. Soc. Series A* **135**, 370-384.
- O'Sullivan F., B. S. Yandell en W. J. Raynor Jr. (1986) Automatic Smoothing of Regression Functions in Generalized Linear Models, *J. Am. Stat. Ass.* **81**, 96-103.
- Simonoff J. S. (1983) A Penalty Function Approach to Smoothing Large Sparse Contingency Tables, *The Annals of Statistics* **11**, 208-218.
- Wedderburn R. W. M. (1974) Quasi-likelihood Functions, Generalized Linear Models and the Gauss-Newton Method, *Biometrika* **61**, 439-447.
- West M. (1985) Generalized Linear Models, Scale Parameters, Outlier Accomodation and Prior Distributions, in: *Bayesian Statistics 2*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley en A. F. M. Smith (eds.), North-Holland.
- West M., P. J. Harrison en H. S. Migon (1985) Dynamic Generalized Linear Models and Bayesian Forecasting (+ discussion), *J. Am. Stat. Ass.* **80**, 73-97.
- Whittaker E. T. (1923) On a New Method of Graduation, *Proceedings of the Edinburgh Mathematical Society* **41**, 63-75.

Ontvangen: 14-09-1987

Geaccepteerd: 06-04-1988