

ADAPTIEVE GEWICHTEN, EEN EXPLORATIEVE TECHNIK VOOR

UITBIJTERS EN MENGSELS VAN REGRESSIEMODELLEN.

P.H.C. Eilers (*)

SAMENVATTING

Aan de som van gewogen kwadraten in een lineair regressiemodel wordt een term toegevoegd die de inhomogeniteit van de gewichten meet. Zo kan de invloed van uitbijters gereduceerd worden. Een uitbreiding maakt het mogelijk op eenvoudige wijze mengsels van lineaire regressiemodellen te schatten, ook als uitbijters aanwezig zijn.

Het minimalisatieprobleem laat zich oplossen door herhaald toepassen van gewogen regressie, voor elk van de deelmodellen in het mengsel apart. Als toepassing worden (mengsels) van regressielijnen, smoothing van tijdreeksen en mengsels van periodieke tijdreeksen gepresenteerd.

Voor enkelvoudige regressiemodellen wordt de equivalente M-schatter afgeleid.

(*) Dienst Centraal Milieubeheer Rijnmond

's-Gravelandseweg 565

3119 XT Schiedam

tel. (010) 4 273 217

1. Inleiding

In de toegepaste statistiek neemt de kleinste kwadratenmethode voor het schatten van parameters in (lineaire) regressiemodellen een zeer prominente plaats in. Deze problemen zijn analytisch uitstekend te behandelen en met de algemene beschikbaarheid van computers is de numerieke oplossing nauwelijks een probleem meer, behalve in zeer omvangrijke en/of slecht geconditioneerde stelsels van vergelijkingen. Als verstoringen normaal verdeeld en onafhankelijk zijn, zijn kleinste kwadratenschatters optimaal en zijn uitvoerig getabelleerde afgeleide verdelingen (zoals t - en F -verdeling) bekend.

Helaas zijn kleinste kwadratenmethoden lang niet altijd bruikbaar. Twee regelmatig terugkomende problemen zijn het onderwerp van wat volgt: uitbijters en mengsels van modellen. Uitbijters zijn min of meer geïsoleerde waarnemingen, die buiten het algemeen patroon van de waarnemingen vallen. In problemen met lage dimensie, zoals bivariate regressie, zijn uitbijters meestal gemakkelijk aan te wijzen op een plaatje van de gegevens. Afhankelijk van hun positie kunnen ze een grote en ongewenste invloed hebben op de geschatte parameters van statistische modellen.

Er is sprake van mengsels van verdelingen als de waarnemingen meer dan één patroon volgen. Bij bivariate regressie uit zich dat bijvoorbeeld als één of meer knikken of sprongen in de regressielijn. Bij univariate waarnemingen vormen zich twee of meer clusters met mogelijk ongelijke spreidingen en eventuele overlappingsen.

Voor de behandeling van uitbijters kan men twee belangrijke hoofdrichtingen onderscheiden. De eerste is gebaseerd op het min of meer formeel toetsen van verdachte waarnemingen op het uitbijterschap, om ze na positieve uitkomst te elimineren. De tweede richting is die der robuuste schattingsmethoden. Deze zijn zo geformuleerd dat extreme waarnemingen geen of weinig invloed hebben op geschatte waarden van parameters.

Vooraf onder de noemer van Exploratory Data Analysis (EDA) kregen de robuuste schattingsmethoden de laatste decennia een groeiende populariteit. Bekende inleidende boeken zijn: Tukey, 1977; Mosteller en Tukey, 1977; Velleman en Hoaglin, 1981; Hoaglin, Mosteller en Tukey, 1983.

Een minder bekende aanpak van uitbijters is voorgesteld door Aitkin en Tunnicliffe Wilson (1980). Zij beschouwen de waarnemingen als afkomstig uit een mengsel van twee of meer normale verdelingen. Eén daarvan geldt voor het merendeel der waarnemingen en de beschrijving van het patroon dat men onderzoekt, de andere verdelingen beschrijven de uitbijters. Aitkin en Tunnicliffe Wilson geven aan hoe hun techniek in principe voor iedere regressiesituatie toepasbaar is.

Het modelleren van waarnemingen als afkomstig van een mengsel van twee of meer verdelingen kent een omvangrijke geschiedenis; het boek van Everitt en Hand (1981) geeft een uitgebreide introductie. Voor continu verdeelde variabelen is een normale verdeling - na een eventuele transformatie - het meest toegepaste model. De nadruk ligt meestal op univariate of multivariate verdelingen zonder covariabelen. Een uitzondering daarop vormen technieken voor geknikte of gebroken regressielijnen ("switching regressions"), die meestal gebaseerd zijn op het model van de normale verdeling.

De meeste methoden voor het modelleren van mengsels doen expliciete aannamen over onderliggende verdelingen.

In het onderstaande wordt een alternatief gepresenteerd dat gebaseerd is op gewogen kleinste kwadraten, waarbij aan de doelfunctie een extra term toegevoegd wordt, die de inhomogeniteit van de gewichten meet. Het minimaliseren van de doelfunctie komt neer op iteratief gewogen regressies. Dat maakt deze methode eenvoudig te implementeren en toepasbaar op zeer uiteenlopende situaties waarin uitbijters, mengsels van regressiemodellen of beide optreden. Ter illustratie zal ze toegepast worden op gebroken regressielijnen, smoothing van tijdreeksen, en mengsels van periodieke tijdreeksen. De nadruk ligt hierbij sterk op exploratieve aspecten. Er blijkt een sterke overeenkomst te bestaan met bepaalde soorten robuuste schatters, de zogenaamde M-schatters. Die wordt kort uitgewerkt.

2. Een gewogen vorm van kleinste kwadraten

Stel dat waarnemingen y_i , $i = 1..m$, gegeven zijn. Wij willen een "gemiddelde" van deze waarnemingen vormen, dat enerzijds een goede aanduiding geeft van de ligging van de waarnemingen op de y-as, anderzijds echter niet gevoelig is voor een enkele waarneming die ver weg ligt.

Het rekenkundig gemiddelde, $\bar{y}$, is de oplossing van:

$$\text{minimaliseer } S_0 = \sum_{i=1}^m (y_i - \bar{y})^2 \quad (2.1)$$

Een uitbreiding hiervan is een gewogen vorm van kwadraten:

$$\text{minimaliseer } S_1 = \sum_{i=1}^m w_i^2 (y_i - \bar{y})^2 \quad (2.2)$$

Door geschikte waarden aan de w_i 's te geven, kan men de invloed van verdachte waarnemingen reduceren of elimineren. Dat vraagt echter kennis over ongelijke varianties voor het kiezen van de gewichten, zoals bij "officiële" gewogen kleinste kwadraten, dan wel om een min of meer subjectieve beoordeling van de gegevens en keuze van de gewichten. Het is echter ook mogelijk (2.2) zo uit te breiden dat de gewichten uit de gegevens zelf afgeleid worden.

Als alle w_i 's 1 zijn, dan zijn (2.1) en (2.2) identiek. Als de gewichten dusdanig genormaliseerd zijn dat de hoogste waarde 1 is, dan is $1 - w_i$ een indicatie van de afwijking van ongewogen kleinste kwadraten. Een globale vorm daarvan is:

$$S_2 = \sum_{i=1}^m (1 - w_i)^2 \quad (2.3)$$

Hoe kleiner S_2 , des te geringer de afwijking van ongewogen kleinste kwadraten.

In het onderstaande zal blijken dat een uit gewogen optelling van S_1 en S_2 gevormde doelfunctie tot zeer bruikbare resultaten leidt; die doelfunctie is:

$$\text{minimaliseer } S = \sum_{i=1}^m w_i^2 (y_i - \bar{y})^2 + p^2 \sum_{i=1}^m (1 - w_i)^2 \quad (2.4)$$

Een meer algemene vorm gaat uit van n onafhankelijke variabelen, gegeven door x_{ij} , $i = 1..m$, $j = 1..n$. De te minimaliseren doelfunctie wordt dan:

$$S = \sum_{i=1}^m w_i^2 (y_i - \sum_{j=1}^n a_j x_{ij})^2 + p^2 \sum_{i=1}^m (1 - w_i)^2 \quad (2.5)$$

Er is geen beperking tot lineaire schatters. In zijn meest algemene vorm wordt (2.5):

$$S = \sum_{i=1}^m w_i^2 (y_i - \hat{y}_i)^2 + p^2 \sum_{i=1}^m (1 - w_i)^2 \quad (2.6)$$

Waarin $\hat{y}_i$ een functie is van onafhankelijke variabelen en te schatten parameters.

Uit de partiële afgeleiden van S in (2.4) volgt:

$$\bar{y} = \frac{\sum_{i=1}^m w_i^2 y_i}{\sum_{i=1}^m w_i^2} \quad (2.7)$$

$$w_i = \frac{p^2}{p^2 + (y_i - \bar{y})^2} = \frac{1}{1 + (y_i - \bar{y})^2 / p^2} \quad (2.8)$$

Uit deze vergelijking volgt dat $\bar{y}$ een gewogen gemiddelde van de waarnemingen is en dat de gewichten bepaald worden uit de residuen $y_i - \bar{y}$, na schaling met p . Hoe groter p des te kleiner de invloed van de residuen; als p erg groot is, zullen alle gewichten nauwelijks van 1 verschillen en is de uitkomst dezelfde als voor ongewogen middelen. Merk op dat voor de gewichten volgens (2.8) altijd geldt dat $0 < w_i < 1$.

De partiële afgeleiden van (2.5) geven analoge resultaten:

$$\sum_{j=1}^n \left(\sum_{i=1}^m w_i^2 x_{ij} x_{ik} \right) a_j = \sum_{j=1}^m w_i^2 x_{ik} y_i \quad k = 1..n \quad (2.9)$$

$$\hat{y}_i = \sum_{j=1}^n a_j x_{ij} \quad (2.10)$$

$$w_i = \frac{1}{1 + (y_i - \hat{y}_i)^2 / p^2} \quad (2.11)$$

De vergelijkingen (2.9) zijn de normaal-vergelijkingen voor gewogen lineaire regressie. Uit (2.11) blijkt weer dat de gewichten berekend worden uit de geschaalde residuen.

De gekoppelde vergelijkingen (2.7) en (2.8), respectievelijk (2.9) en (2.11), zijn niet lineair en hebben geen expliciete oplossing. Ze suggereren echter een simpele iteratieve rekenwijze, waarbij men om en om (2.7) en (2.8), respectievelijk (2.9) en (2.11), oplost. Voor de oplossing van (2.9) kan men gebruik maken van de mogelijkheden voor gewogen regressie die veel standaard programmatuur zoals SPSS en GLIM bieden; het vormen van de gewichten via (2.11) is erg eenvoudig.

De convergentie blijkt wisselend van snelheid te zijn. De uitkomsten zijn niet altijd uniek, omdat de doelfunctie lokale minima kan bezitten, zie ook § 3. Het is daarom van belang de iteraties te starten met zinvolle schattingen van de gewichten of parameterwaarden. Een bruikbare methode is, alle gewichten 1 te maken, behalve die van vermoedelijke uitbijters, die gewicht nul krijgen; daarbij kan het geen kwaad conservatief te zijn door ruimschoots gewichten nul te maken.

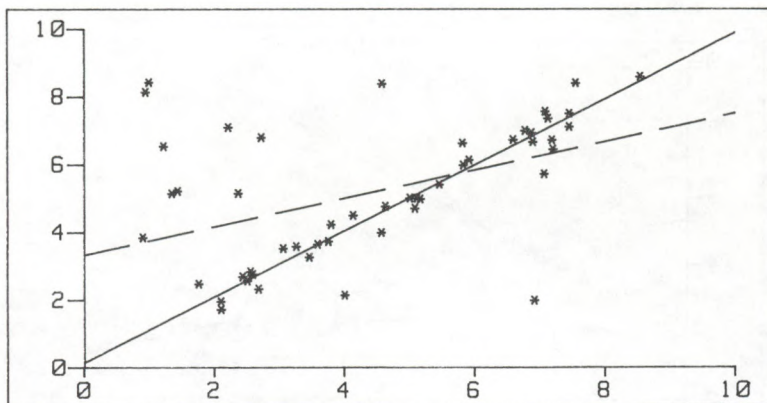
De keuze van p eist enig inzicht in het patroon van de gegevens. Ruwweg kan men stellen dat $3p$ de grens is waar de invloed van het verschil tussen een waarneming en de gefitte waarde tot 1% van de ongewogen invloed gereduceerd is. Als $y_i - \hat{y}_i = 3p$, dan is $w_i = 0.1$ en $w_i^2 = 0.01$. In bivariate situaties is een geschikte waarde van p meestal goed aan te geven aan de hand van een grafische weergave van de data. Subtieler is een adaptieve keuze van p , bijvoorbeeld de (gewogen) standaarddeviatie van de residuen $y_i - \hat{y}_i$. Daar is echter nog geen ervaring mee.

3. Toepassing op uitbijters

In deze paragraaf onderzoeken we de toepassing van adaptieve weging op enkele situaties waarin uitbijters optreden. Hoewel het gemiddelde van een univariate steekproef als drager van de afleiding gebruikt werd, is het als toepassingsvoorbeeld vrij triviaal. Interessanter zijn lineaire regressie en tijdreeksen.

3.1. Bivariate regressie

Figuur 3.1 toont gesimuleerde data, die een puntenwolk vormen waarvan een deel duidelijk wijst op een lineaire regressielijn, terwijl het andere deel uit verspreide uitbijters bestaat. De streeplijn is het resultaat van ongewogen regressie, de volle lijn van adaptief gewogen regressie met $p = 0.5$. Deze waarde van p werd gekozen na inspectie van de figuur.



Figuur 3.1: Simulatie van regressielijn met uitbijters.

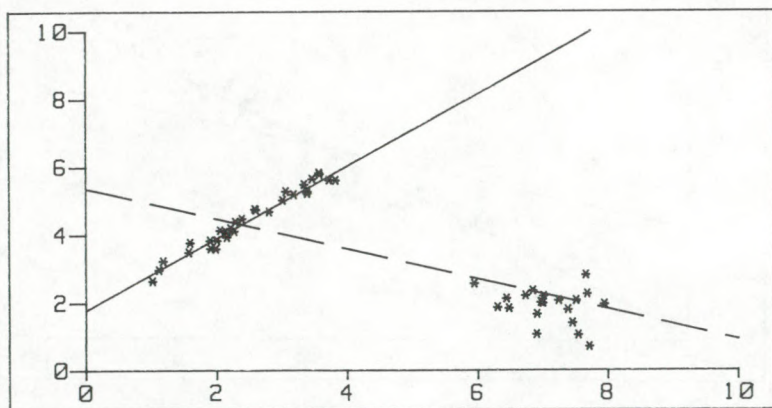
Binnen zekere grenzen maakt de waarde van p niet veel uit, zie tabel 3.1. In deze tabel geeft It het aantal iteraties dat nodig was om de veranderingen in a en b kleiner dan 10^{-4} te maken.

Tabel 3.1: Invloed van p op a en b in $\hat{y} = ax + b$ en aantal iteraties.

P	1000	100	10	1	0.5	0.2	0.1
a	0.417	0.418	0.478	0.973	0.983	0.981	0.964
b	3.351	3.348	3.031	0.170	0.139	0.204	0.327
It	2	3	7	8	14	39	55

Een tweede voorbeeld geeft figuur 3.2. Hier vormen de uitbijters een cluster. De gegevens zijn gesimuleerd op basis van parameterwaarden die Rousseeuw (1984) geeft. De streeplijn geeft de ongewogen regressielijn aan.

In dat artikel laat hij zien dat bekende robuuste schatters geen goede resultaten geven, in tegenstelling tot zijn eigen voorstel, de kleinste mediaan van kwadraten. Zoals uit de figuur blijkt, opereren adaptief gewogen kleinste kwadraten ($p = 0.5$) hier met succes. Rousseeuw geeft aan dat zijn methode kostbaar in rekentijd is en een specifiek programma vereist.



Figuur 3.2: Toepassing op Rousseeuw's voorbeeld.

Dit voorbeeld toont aan dat goede beginschattingen van essentieel belang zijn. Gebruiken we de parameters van de ongewogen regressielijn als startwaarden, dan resulteert, ongeacht de waarde van p , een uiteindelijke lijn die hier weinig van af wijkt. Voor het resultaat in figuur 3.2 werd gestart met $a = 1$ en $b = 0$.

In § 5 wordt het sterke verband tussen adaptief gewogen regressie en M-schatters uitgewerkt. Dat verband houdt in dat de door Rousseeuw onderzochte M-schatters weinig andere resultaten zullen geven dan adaptief gewogen regressie. De conclusie is derhalve dat zijn voorbeeld een lokaal minimum van de doelfunctie van M-schatters laat zien. Dit lokale minimum is het gevolg van de ongelukkige beginschattingen die de kleinste kwadratenlijn geeft.

3.2. Smoothing van tijdreeksen

Meer dan een halve eeuw geleden stelden Whittaker (1923) en Henderson (1924) voor, de smoothing van een reeks waarnemingen y_i te formuleren als het minimaliseren van:

$$T = \sum_{i=1}^m (y_i - z_i)^2 + \beta \sum_{i=1}^m (\nabla^k z_i)^2 \quad (3.1)$$

Hier is z_i de gesmoothde reeks en ∇^k de operator van de k-de centrale differentie. De som loopt over alle i waar voor deze gedefinieerd is; β is een parameter die de smoothing beïnvloedt: hoe groter β , des te gladder de reeks z .

De ervaring heeft uitgewezen dat $k = 2$ goed voldoet:

$$\nabla^2 z_i = z_{i-1} - 2 z_i + z_{i+1} \quad (3.2)$$

De doelfunctie wordt dus:

$$T = \sum_{i=1}^m v_i (y_i - z_i)^2 + \sum_{i=2}^{m-1} (z_{i-1} - 2 z_i + z_{i+1})^2 \quad (3.3)$$

Toepassingen van deze vorm van smoothing (in combinatie met seizoen-correctie) vindt men onder andere bij Akaike (1981), Schlicht (1980) en Cleveland (1983). Het aantrekkelijke van deze smoothingmethode is dat de vergelijkingen gemakkelijk zijn op te lossen. Ook zijn voor ontbrekende waarnemingen voorzieningen getroffen, Akaike en Ishiguro (1980), door de a-priori gewichten v_i ; deze zijn 0 voor ontbrekende waarden, anders 1.

Figuur 3.3 toont een gesimuleerde tijdreeks gevormd door:

$$y_i = \sin(0.08 i) + e_i \quad (3.4)$$

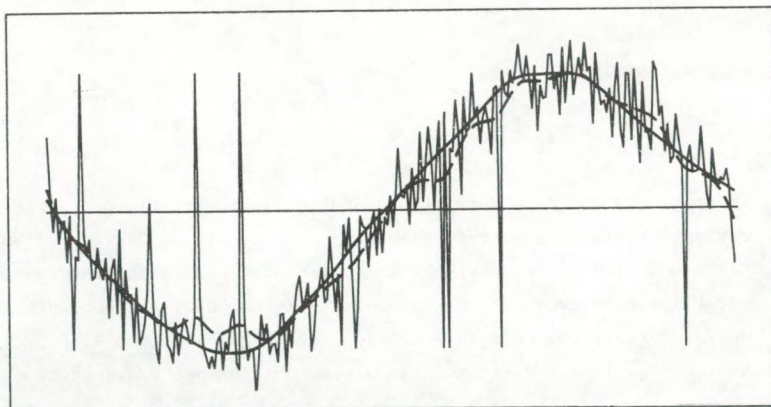
Waarin e_i normaal verdeeld is met gemiddelde 0 en standaarddeviatie 0.1. Verder zijn met gulle hand uitbijters met waarden -1 of 1 toegevoegd. De gesmoothde reeks ($\beta = 1000$, de streeplijn in de figuur) blijkt bij de uitbijters duidelijke slingeren te vertonen. Met adaptieve weging valt dit effect te bestrijden. Daartoe wordt de te minimaliseren doelfunctie geformuleerd als:

$$S = \sum_{i=1}^m v_i w_i^2 (y_i - z_i)^2 + \beta \sum_{i=2}^{m-1} (z_{i-1} - 2z_i + z_{i+1})^2 + p^2 \sum_{i=1}^m (1 - w_i)^2 \quad (3.5)$$

De details van de vergelijkingen die uit (3.5) volgen, blijven achterwege. De structuur is echter identiek aan wat we eerder tegenkwamen: de afzonderlijke gewichten volgen uit de corresponderende, met p geschaalde, residuen en in de smoothing-vergelijkingen worden gewichten w_i^2 geïntroduceerd, die het oplossen nauwelijks compliceren. De zelfde iteratieve rekenwijze als voorheen wordt toegepast.

In figuur 3.3 staat het resultaat als dikke lijn getekend voor $\beta = 1000$ en $p = 0.1$.

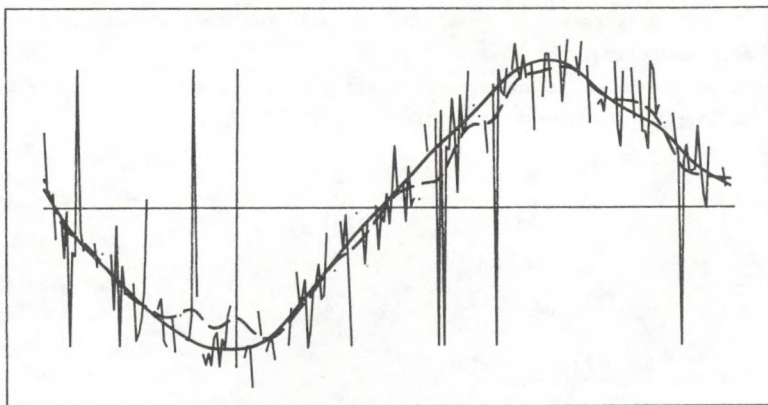
De waarde van p werd gekozen op basis van inspectie van de grafiek. Na 10 iteraties veranderden de z_i 's minder dan 0,1%.



Figuur 3.3: Smoothing van reeks met uitbijters, met en zonder adaptieve gewichten.

Wie weleens in een boek als dat van Tukey (1977) of Velleman en Hoaglin (1981) de hoofdstukken over smoothing van tijdreeksen heeft gelezen, zal getroffen zijn door de wel haast ambachtelijke benadering die daar gevolgd wordt. Het smoothing-proces wordt opgebouwd uit bouwstenen met klinkende namen als "running medians", "hanning", "twicing", "reroughing", "resmoothing", die in talloze variaties opgestapeld kunnen worden. Daarbij vragen de uiteinden van de reeks en mogelijke ontbrekende waarnemingen bijzondere ingrepen. In vergelijking daarmee is de hier voorgestelde methode bijzonder simpel en effectief. De uiteinden van de reeks vragen geen aparte ingrepen en ontbrekende waarnemingen vormen door de a-priori gewichten v_i geen probleem.

Een illustratie van een reeks met ontbrekende waarnemingen geeft figuur 3.4. Van de reeks uit de figuur 3.3 is 30% van de waarnemingen (exclusief de uitbijters) aselect verwijderd. De gesmoothde reeks ($\beta = 1000$ en $p = 0.1$) verschilt weinig met die in figuur 3.3.



Figuur 3.4: Smoothing met adaptieve gewichten van reeks met ontbrekende waarnemingen.

4. Uitbreiding naar mengsels

4.1. Mengsels van regressielijnen

De uitbreiding van adaptieve weging naar mengsels van modellen laat zich eenvoudig illustreren aan de hand van een bivariate gebroken regressielijn, zoals getekend in figuur 4.1. De doelfunctie wordt geformuleerd als:

$$S = \sum_{i=1}^m w_{i1}^2 (y_i - a_1 x_i - b_1)^2 + \sum_{i=1}^m w_{i2}^2 (y_i - a_2 x_i - b_2)^2 + p^2 \sum_{i=1}^m (1 - w_{i1} - w_{i2})^2 \quad (4.1)$$

Er zijn nu twee termen met gewogen kwadraten van residuen, één voor elk van de beide delen van de gebroken regressielijn, elk met zijn eigen gewichten. Merk op dat de som in beide delen over alle waarnemingen loopt. De laatste term beschrijft in welke mate de som van de gewichten, per waarneming, van 1 verschilt.

De gewichten en parameterwaarden die S minimaal maken volgen uit het nul stellen van de partiële afgeleiden. Het resultaat is:

$$\sum_{i=1}^m w_{ik}^2 x_i^2 a_k + \sum_{i=1}^m w_{ik}^2 x_i b_k = \sum_{i=1}^m w_{ik}^2 x_i y_i \quad k = 1, 2 \quad (4.2)$$

$$\sum_{i=1}^m w_{ik}^2 x_i a_k + \sum_{i=1}^m w_{ik}^2 b_k = \sum_{i=1}^m w_{ik}^2 y_i \quad k = 1, 2 \quad (4.3)$$

$$w_{ik} = \frac{1/r_{ik}^2}{1 + 1/r_{i1}^2 + 1/r_{i2}^2} \quad k = 1, 2 \quad (4.4)$$

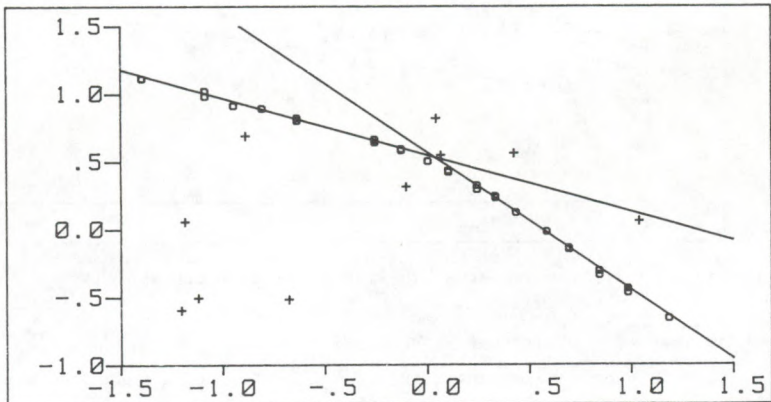
met:

$$r_{ik} = (y_i - a_k x_i - b_k) / p \quad k = 1, 2 \quad (4.5)$$

De vergelijkingen (4.2) en (4.3) zijn normaalvergelijkingen voor gewogen regressie. Er zijn aparte stelsels voor a_1 en b_1 respectievelijk a_2 en b_2 . De gewichten w_{ik} worden berekend uit de met p geschaalde residuen ten opzichte van de beide regressielijnen. De numerieke oplossingswijze is weer iteratief, waarbij om en om de gewichten en de parameters van de lijnen berekend worden.

Goede startwaarden zijn ook hier van belang, omdat S lokale minima kan hebben. Een effectieve methode is het stellen van $w_{i1} = 1$ en $w_{i2} = 0$ voor een aantal punten dat duidelijk bij één van de regressielijnen hoort en $w_{i1} = 0$ en $w_{i2} = 1$ voor een aantal punten dat duidelijk bij de andere lijn hoort; alle andere punten krijgen $w_{i1} = 0$ en $w_{i2} = 0$. Met deze keuzen van de gewichten start men dan de iteratief gewogen regressies.

Een illustratie geeft figuur 4.1. De 29 door cirkels gemerkte punten zijn ontleend aan Bacon en Watts (1971); zij geven het verband tussen (de logaritme van) de stroomsnelheid en (de logaritme van) de hoogte van de stagnante oppervlaktelaag in een hellend kanaal. De 10 met kruisjes gemerkte punten zijn aselekt getrokken (x uniform tussen -1.2 en 1.2 , y uniform tussen -0.6 en 1.1). Voor twee punten op de "linker tak" werd $w_{i1} = 1$ gesteld en voor twee punten op de "rechttertak" werd $w_{i2} = 1$ gesteld; alle andere gewichten kregen nul als startwaarde. De waarde van p was 0.05 .

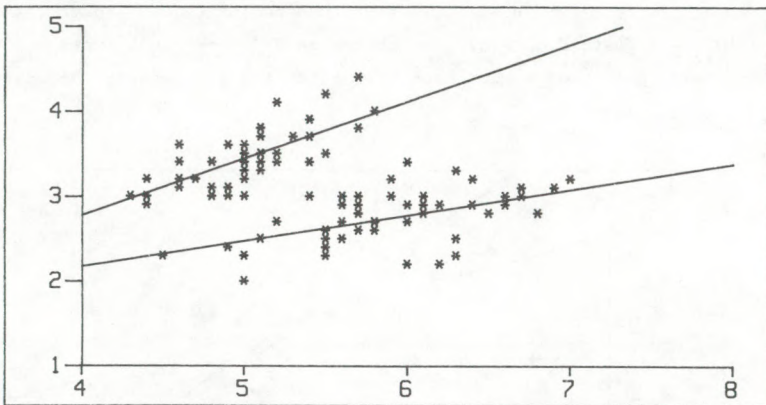


Figuur 4.1: Voorbeeld van gebroken regressielijn.

De situatie in figuur 4.1 is "klassiek" in de zin dat de geldigheid van één van beide lijnen afhangt van het feit of x kleiner of groter is dan een zekere waarde.

De positie van het snijpunt van beide lijnen geeft een schatting van deze waarde. Pollard (1977); geeft een voorbeeld van een methode die neerkomt op het kiezen van a-priori gewichten (met waarden 0 of 1), op basis van de x coördinaat, waarbij het scheidingspunt systematisch gevarieerd wordt om de som van de kwadraten van de residuen minimaal te maken.

Het gebruik van adaptieve gewichten is niet afhankelijk van een dergelijke gunstige splitsingsmogelijkheid. Figuur 4.2 toont het verband tussen lengte (horizontaal) en breedte van kelkbladen van irissen (iris setosa en iris versicolor). De gegevens zijn ontleend aan Fisher. Men kan ze onder andere vinden bij Kendall en Stuart (1966) en Chen (1973).



Figuur 4.2: Iris-gegevens van Fisher met twee regressielijnen.

De iteraties werden gestart door het hierboven aangegeven kiezen van gewichten voor enkele punten in elk van de "wolken". Omdat er geen indicaties van uitbijters zijn, is een hoge waarde van p ($p=10$) gekozen. Als de regressielijnen gevonden zijn, kan een klassificatie van de punten naar een van de regressielijnen eenvoudig plaatsvinden op basis van de gewichten volgens (4.4). Bij een hoge waarde van p zal $w_{i1} + w_{i2}$ weinig van 1 afwijken. Ligt w_{ij} dicht bij 1, dan "hoort punt i waarschijnlijk bij regressielijn j " en weegt zwaar mee in de berekening van die lijn. Liggen beide gewichten dichtbij 0.5, dan komen beide lijnen in aanmerking.

Zijn uitbijters aanwezig en is de waarde van p zodanig dat hun invloed effectief gereduceerd wordt, dan zullen van de uitbijters zowel w_{i1} als w_{i2} laag zijn. De uitbijters kunnen derhalve op basis van lage waarden van $w_{i1} + w_{i2}$ gedetecteerd worden.

Het is eenvoudig (4.1) uit te breiden naar meer algemene modellen en meer mengsels:

$$S = \sum_{i=1}^m \sum_{k=1}^q w_{ik}^2 (y_i - \sum_{j=1}^n x_{ij} a_{jk})^2 + p^2 \sum_{i=1}^m (1 - \sum_{k=1}^q w_{ik})^2 \quad (4.6)$$

Hier is sprake van een mengsel van q modellen, met elk een parameter-vector met elementen a_{jk} . De basisfuncties x_{ij} zijn voor elk model gelijk. Dat is echter geen eis. Een zeer algemene formulering is:

$$S = \sum_{i=1}^m \sum_{k=1}^q w_{ik}^2 (y_i - \hat{y}_{ik}(\underline{a}_k))^2 + p^2 \sum_{i=1}^m (1 - \sum_{k=1}^q w_{ik})^2 \quad (4.7)$$

Met $\hat{y}_{ik}(\underline{a}_k)$ is hier de geschatte waarde van y_i volgens model k aangegeven. De dimensies van de verschillende parametervectoren hoeven niet gelijk te zijn. Het is ook mogelijk dat één of meer parameters in meer deelmodellen tegelijk voorkomen.

4.2. Mengsels van periodieke tijdreeksen

Tot nu toe waren de voorbeelden van mengsels niet sterk afwijkend van wat men in de literatuur tegenkomt. Er zijn echter tal van situaties waarin het idee van een mengsel, gecombineerd met de hierboven gepresenteerde schattingsmethode, vruchtbaar kan zijn. Dat zal nu met enkele voorbeelden aangetoond worden.

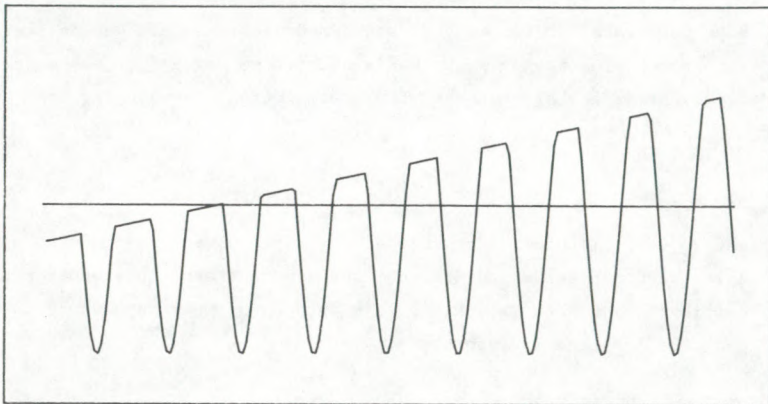
Figuur 4.3 toont een sinusvormig signaal dat afgekapt is door een rechte lijn. In figuur 4.4 is hetzelfde signaal getekend, voorzien van enige ruis. Ook zijn de twee gefitte modellen van het mengsel getekend. We nemen aan dat de hoekfrequentie f van het sinussignaal bekend is. Als mengselmodel stellen we:

$$\hat{y}_{i1} = a_1 i + b_1 \quad ; \quad y_{i2} = a_2 \sin(fi) + b_2 \cos(fi) \quad (4.8)$$

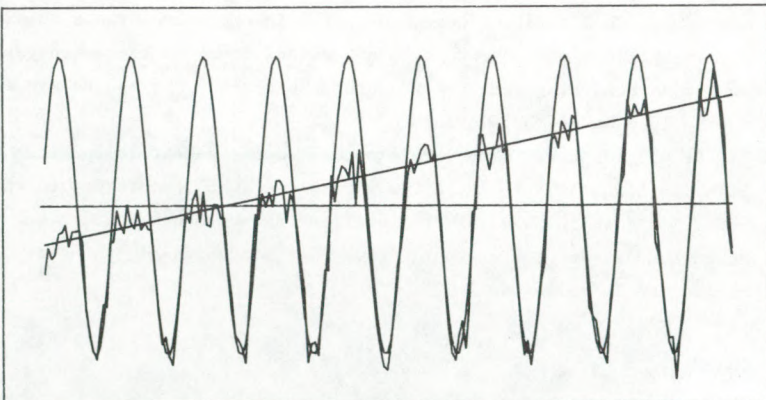
De regressors in de deelmodellen zijn hier dus totaal verschillend. Merk op dat zowel de rechte lijn als de sinus voor alle waarden van i gelden; de gewichten w_{ik} bepalen echter in welke mate een van beiden past bij y_i .

Op dit thema zijn vele variaties mogelijk. Zo kan men bijvoorbeeld een "gelijkgerichte" sinus ($y_i = |\sin(fi)|$), plus ruis, modelleren als een mengsel van twee sinussignalen:

$$\hat{y}_{i1} = a_1 \sin(fi) + b_1 \cos(fi); \hat{y}_{i2} = a_2 \sin(fi) + b_2 \cos(fi) \quad (4.9)$$



Figuur 4.3: Een afgeknot sinusvormig signaal.



Figuur 4.4: Fit van mengsel op afgeknotten sinus met ruis.

5. Een equivalente M-schatter

Een kenmerk van kleinste kwadraten als schattingsmethode is, dat extreme waarnemingen een zeer grote invloed op parameterschattingen kunnen hebben. In de doelfunctie:

$$S = \sum_{i=1}^m (y_i - y)^2 = \sum_{i=1}^m u_i^2 \quad (5.1)$$

voor het schatten van een gemiddelde, kan één zeer hoge of zeer lage y_i , S volledig gaan overheersen. Dit komt doordat $(y_i - y)^2$ onbeperkt groot kan worden, ongeacht de waarde van y , door y_i te laten toe- of afnemen. Onder de robuuste schattingsmethoden nemen de M-schatters een prominente plaats in, zie bijvoorbeeld Goodall (1983). Bij bepaalde M-schatters wordt u_i^2 vervangen door een functie $\rho(u_i)$ die voor kleine waarden van $|u_i| = |y_i - y|$ weinig verschilt van u_i^2 , maar voor grotere waarden begrensd is, of minder sterk stijgt dan u_i^2 . De doelfunctie wordt dan:

$$S = \sum_{i=1}^m \rho(u_i) = \sum_{i=1}^m \rho(y_i - y) \quad (5.2)$$

De afgeleide van $\rho(u)$, $\psi(u)$ is evenredig aan de zogenaamde invloedsfunctie. Minimaliseren van S komt overeen met het oplossen van:

$$0 = \sum_{i=1}^m \psi(y_i - y) = \sum_{i=1}^m \psi(u_i) \quad (5.3)$$

De gevoeligheid van de oplossing van (5.3), $\bar{y}$, voor kleine verschuivingen in y_i wordt gegeven door de afgeleide functie $\psi'(y_i - y)$. Details vindt men bij Goodall (1983).

Voor het schatten van een gemiddelde met adaptieve weging is de doelfunctie:

$$S = \sum_{i=1}^m w_i^2 (y_i - y)^2 + p^2 \sum_{i=1}^m (1 - w_i)^2 \quad (5.4)$$

In § 2 vinden we dat in het optimum de gewichten voldoen aan:

$$w_i = \frac{p^2}{p^2 + (y_i - \bar{y})^2} \quad (5.5)$$

Deze uitdrukking is ook te beschouwen als een functie die w_i geeft voor elke waarde van y . Vullen we die in (5.4) in, dan ontstaat:

$$S = \sum_{i=1}^m \frac{p^2 (y_i - y)^2}{p^2 + (y_i - y)^2} = \sum_{i=1}^m \frac{p^2 u_i^2}{p^2 + u_i^2} \quad (5.6)$$

Een minimum van (5.6) is tevens een minimum van (5.4).

Het is duidelijk dat (5.6) de vorm (5.2) heeft, met:

$$\psi(u) = \frac{p^2 u^2}{p^2 + u^2} \quad (5.7)$$

Voor kleine waarden van u komt $\psi(u)$ overeen met u^2 , voor hoge waarden van $|u|$ nadert $\psi(u)$ asymptotisch naar p^2 . Hoe hoger p , des te hoger dit plafond en des te groter het gebied waar geen noemenswaardige afwijking van ongewogen kleinste kwadraten optreedt.

Goodall (1983) beschrijft hoe voor M-schatters equivalente schatters afgeleid kunnen worden die gebruik maken van adaptieve weging.

In het bovenstaande is het omgekeerde gebeurd: voor een vorm van adaptieve weging is de equivalente M-schatters afgeleid. Verder onderzoek lijkt hier de moeite waard: mogelijk zijn uit de uitbreiding naar mengsel veralgemeniseerde M-schatters af te leiden.

De lokale minima van de doelfunctie, die al enkele malen gesignaleerd zijn, zullen in het algemeen bij M-schatters op kunnen treden. De literatuur is hier nogal zwijgzaam over. Meestal worden de normale kleinste kwadratenschatters als startwaarden voor iteraties gebruikt. Dat is geen garantie van het vinden van een globaal minimum, zoals het voorbeeld in figuur 3.2 laat zien.

6. Discussie

Zoals in het voorafgaande is gebleken, biedt adaptieve weging interessante en gemakkelijk te implementeren technieken voor het bestrijden van de invloed van uitbijters en het modelleren van mengsels. De bijbehorende theorie is (nog) eenvoudig en het bestaan van een equivalente M-schatter geeft een sterk aanknopingspunt aan de omvangrijke literatuur over robuuste schattingsmethoden.

Iteratieve weging is een veel toegepaste techniek en voor M-schatters heeft men equivalente algoritmen afgeleid die daar op gebaseerd zijn. Voor zover mij bekend is de hier voorgestelde toevoeging van een term in de gewichten, aan de som van gewogen kwadraten, nog niet eerder voorgesteld. Ook de afleiding van een M-schatter vanuit de weging is ongebruikelijk. Deze M-schatter is afwijkend van de gebruikelijke, doordat zijn invloedsfunctie asymptotisch naar nul gaat, in plaats van tamelijk abrupt. Moberg, Ramberg en Randles (1980) pasten een rationale functie toe als onderdeel van een systeem van invloedsfuncties.

De uitbreiding naar mengsels van regressiemodellen is zeer eenvoudig en leidt tot simpele rekentechnieken. Met name is waardevol dat de normaal-vergelijkingen alleen via de gewichten gekoppeld zijn. Door de iteratieve weging kan men de normaal-vergelijken per model apart oplossen. Vergeleken met veel bestaande modellen voor mengsels is het ontbreken van een aanname over de verdelingen van de deelpopulaties prettig. Het geeft de gebruiker veel vrijheid, die stimuleert tot het verruimen van het toepassingsgebied van mengsels, zoals aan de hand van enkele voorbeelden werd aangetoond.

Op theoretisch gebied lijken interessante mogelijkheden aanwezig te zijn voor het veralgemeniseren van M-schatters naar mengsels.

Een probleem is het bestaan van lokale extremen in de doelfunctie, zeker gezien het feit dat beginschattingen die afkomstig zijn van ongewogen kleinste kwadraten niet altijd geschikt blijken. Voorzichtigheid is dus geboden. In bivariate problemen, die zich goed grafisch laten presenteren, is een visuele controle mogelijk. Moderne rekenapparatuur biedt in toenemende mate interactieve grafische mogelijkheden, die het aan-geven van beginschattingen voor parameters en/of gewichten vergemakkelijken.

In meerdimensionale problemen is een systematisch minimalisatie algoritme gewenst. Dat vereist echter complexe programmatuur, waardoor de charme van de eenvoud verdwijnt. Daarom lijkt de hier voorgestelde iteratieve weging in eerste instantie vooral geschikt voor exploratieve toepassingen. Nader onderzoek naar de eigenschappen van de schatters lijkt de moeite waard.

De keuze van de parameter p vraagt meer theoretische en experimentele aandacht. Zo kan p mogelijk adaptief berekend worden, zoals bij sommige M-schatters.

7. Referenties

- Aitkin M., Tunnicliffe Wilson G. (1980), Mixture models, outliers and the EM-algorithm, *Technometrics* 22, 325-331
- Akaike H. (1981), Seasonal adjustment by a Bayesian modelling, *Journal of Time Series Analysis* 1, 1-13
- Akaike H., Ishiguro M. (1980), Trend estimation with missing observations, *Ann. Inst. Statist. Math.* 32 part B, 481-488
- Bacon D.W., Watts D.G. (1971), Estimating the transition between intersecting straight lines, *Biometrika* 58, 525-534
- Chen C.H. (1973), *Statistical Pattern Recognition*, Hayden
- Cleveland W.S. (1983), Seasonal and calendar adjustment, *Handbook of Statistics*, Vol. 3, D.R. Brillinger and P.R. Krisniah (eds.), Elsevier
- Everitt B.S., Hand D.J. (1981), *Finite Mixture Distributions*, Chapman and Hall
- Goodall C. (1983), M-estimators of location: an outline of the theory, Hoofdstuk 11 in Hoaglin e.a. (1983)
- Henderson R. (1924), A new method of graduation, *Transactions of the Actuarial Society of America* 25, 29-40
- Hoaglin D.C., Mosteller F., Tukey J.W. (1983), *Understanding Robust and Exploratory Data Analysis*, Wiley
- Kendall M.G., Stuart A. (1966), *The Advanced Theory of Statistics* Vol. 3, Hafner
- Moberg Th. F., Ramberg J.S., Randles H. (1980), An adaptive multiple regression procedure based on M-estimators, *Technometrics* 22, 213-224
- Mosteller F., Tukey J.W. (1977), *Data Analysis and Regression*, Addison-Wesley

- Pollard J.H., (1977), A Handbook of Numerical and Statistical Techniques, Cambridge University Press
- Quandt R.E., Ramsey J.B. (1978), Estimating mixtures of normal distributions and switching regressions, J. Am. Statist. Ass. 73, 730-738
- Rousseeuw P.J. (1984), Least median of squares regression, J. Am. Stat. Ass. 79, 875-880
- Schlicht E. (1981), A seasonal adjustment principle and seasonal adjustment, method derived from this principle, J. Am. Stat. Ass. 76, 373-378
- Tukey J.W. (1977), Exploratory Data Analysis, Addison-Wesley
- Velleman P.F., Hoaglin D.C. (1981), Applications, Basics and Computing of Exploratory Data Analysis, Duxbury Press
- Whittaker E.T. (1923), On a new method of graduation, Proceeding of the Edinburgh Mathematical Society 41, 63-75

Ontvangen: 16-04-86
Geaccepteerd: 21-10-86