Enkele opmerkingen over "Adaptieve gewichten, een exploratieve techniek voor uitbijters en mengsels van regressiemodellen", door P.H.C. Eilers.

J.J. de Gruijter*)

In Kwantitatieve Methoden 23, pag. 63-83, verscheen onlangs een lezenswaardig artikel van P.H.C. Eilers over het gebruik van adaptieve gewichten bij regressie-analyse. Graag wil ik hierop enkele aanvullingen geven.

Eilers stelt voor de invloed van uitbijters bij regressie-analyse te reduceren door deze in de te minimaliseren som van gewogen kwadraten een klein gewicht te geven, en de gewichten adaptief te berekenen via een uitbreiding van de doelfunctie. Voor het meest algemene geval, een mengsel van q niet-lineaire deelmodellen, luidt de doelfunctie van Eilers (formule 4.7):

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

waarin de adaptieve gewichten w_{ik} de graduele toewijzingen van de individuen tot de deelmodellen voorstellen. De uitbreiding met de tweede term heeft tot gevolg dat de som van de gewichten per individu niet meer gelijk hoeft te zijn aan 1 en dat een uitbijter, d.w.z. een waarneming die bij geen enkel deelmodel goed past, een klein totaal gewicht krijgt.

Het lijkt mij nuttig hieraan toe te voegen dat de methode met adaptieve gewichten ook goed te formuleren is in termen van vage verzamelingen-theorie (zie b.v. Bezdek, 1981). Een gewicht w_{ik} wordt dan geinterpreteerd als het lidmaatschap van het i-de individu met betrekking tot een vage verzameling van individuen welke wordt gerepresenteerd door het k-de deelmodel. De bovengenoemde uitbreiding met de tweede term in (1) is dan te zien als het introduceren van een vage verzameling van uitbijters.
Dit idee is eerder voorgesteld en uitgewerkt door Ohashi (1984), met name in de context van fuzzy cluster-analyse, waarbij de rol van de residuen in

*) Instituut voor Toegepaste Informatica TNO, Postbus 100, 6700 AC Wageningen

de kwadraatsom wordt overgenomen door afstanden tussen individuen en clustercentra. Het minimalisatie-algorithme van Eilers is in essentie hetzelfde als dat van Ohashi. Ook Ohashi legt het verband met M-schatters; hij stelt bovendien een methode voor om p te kiezen.
In de formulering van Ohashi komt verder een in de fuzzy cluster-analyse bekende generalisatie voor, die mogelijk ook bij regressie-analyse interessant kan zijn. Dit houdt in dat de gewichten worden voorzien van een door de gebruiker te kiezen exponent ϕ. Eilers' doelfunctie wordt dan:

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

De exponent ϕ bepaalt, gegeven de dataset, de mate van vaagheid van de verzamelingen, zowel die van de verzamelingen gerelateerd aan de deelmodellen als die van de uitbijter-verzameling. Hoe groter ϕ, hoe vager de verzamelingen. Aangetoond kan worden (zie b.v. Bezdek, 1981), dat de oplossing van (2) bij de keuze $\phi=1$ leidt tot niet-vage verzamelingen, d.w.z. elk individu wordt dan geheel toegewezen tot juist één deelverzameling.
Het minimalisatie-algorithme kan, althans bij $\phi>1$, onder deze generalisatie ongewijzigd blijven, zij het dat de gewichten nu moeten worden herberekend volgens:

$$w_{ik} = \frac{r_{ik}^{-2/(\phi-1)}}{1 + \sum_{h=1}^{q} r_{ih}^{-2/(\phi-1)}}, \qquad (3)$$

met:

$$r_{ik} = (y_i - \hat{y}_{ik}(\underline{a}_k))/p \qquad (4)$$

Dit volgt uit nul stellen van de partiële afgeleiden van S naar w_{ik}.
(Vergelijk met Eilers' 4.4.)

In het geval $\phi=1$ moeten de w_{ik} in verband met de limietovergang in (3) anders worden herberekend, n.l.:

$$\begin{aligned} w_{ik} &= 1 \quad \text{als} \quad r_{ik}^2 = \min\{r_{i1}^2, \ldots, r_{iq}^2\} \quad \text{en} \quad r_{ik}^2 \leq 1 \\ &\quad 0 \quad \text{als} \quad r_{ik}^2 \neq \min\{r_{i1}^2, \ldots, r_{iq}^2\} \quad \text{of} \quad r_{ik}^2 > 1 \end{aligned} \qquad (5)$$

Dit komt neer op iteratieve re-allocatie van elk individu tot het deelmode waarvoor zijn residu in absolute waarde het kleinst is, tenzij die waarde groter is dan p. In dat geval wordt het betreffende individu in de lopendel iteratie-cyclus als uitbijter behandeld.
Ook indien een eenduidige toewijzing van de individuen tot de deelverzamelingen gewenst is (en derhalve ϕ=1 wordt gekozen), zou het wellicht de moeite lonen het iteratie-proces te beginnen met een waarde voor ϕ groter dan 1, en die dan stapsgewijs te verminderen. Omdat hierdoor de doelfunctie tijdelijk wordt afgevlakt, is het niet onaannemelijk dat zo de kans op stranden in een lokaal minimum is te verkleinen.

Een tweede generalisatie die zich vanuit de fuzzy cluster-analyse aandient betreft de varianties van de deelmodellen. Deze worden bij Eilers' methode impliciet gelijk verondersteld. Echter, zoals bij cluster-analyse de afstanden tot clustercentra kunnen worden gestandaardiseerd met een eigen covariantie-matrix voor elk van de clusters (Gustafson en Kessel, 1979), zo zouden bij regressie-analyse de residuen kunnen worden gestandaardiseerd met verschillende iteratief te berekenen varianties per deelmodel. Voor regressie-analyse is deze generalisatie overigens ook aangegeven door Aitkin en Tunnicliff Wilson (1980).

Referenties

Aitkin M., Tunnicliff Wilson G. (1980) Mixture models, outliers and the EM-algorithm, Technometrics 22, 325-331.

Bezdek, J.C. (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York.

Gustafson, D.E., Kessel W. (1979) Fuzzy clustering with a fuzzy covariance matrix. In: Proc. IEEE-CDC, Vol. 2 (K.S. Fu, ed.), 761-766, IEEE Press, Piscataway, New Jersey.

Ohashi, Y. (1984) Fuzzy clustering and robust estimation. 9th meeting SAS Users Group International, Hollywood Beach, Florida.

Ontvangen: 26-03-1987
Geaccepteerd: 03-05-1987