KM 39 (1992)
pg.39-52

LAV-REGRESSIE

Eric Schulte Nordholt

Samenvatting

In dit artikel bekijkt de auteur[1] hoe LAV (=Least Absolute Values)-regressie en OLS (=Ordinary Least Squares)-regressie verschillen wanneer deze beide methoden op een zelfde dataset worden toegepast. Nadelen van de OLS-regressie zijn de gevoeligheid voor uitschieters in de data en de slechte resultaten als we te maken krijgen met dikstaartige foutverdelingen. LAV-regressie is in deze gevallen robuuster. Uit een simulatie-experiment blijkt dat bij een normale en een bimodale verdeling de standaardfouten van de schattingen van de OLS-methode kleiner zijn dan die van de LAV-methode. Daarentegen is het bij de Cauchy en Laplace verdeling net andersom. Omdat de foutverdeling bij praktijkdata onbekend is lijkt het verstandig beide methoden naast elkaar te gebruiken en de schattingen te vergelijken alvorens een conclusie te trekken.

[1] van der Zaenlaan 17 1215 SG Hilversum tel. 035-47267

Inleiding In het kader van mijn afstuderen in de Wiskunde aan de Rijksuniversiteit te Utrecht (RUU) liep ik in de periode september-december 1990 stage bij AKZO Research Laboratories Arnhem (ARLA). Ik deed onderzoek naar de LAV (=Least Absolute Values)-regressie. Dit artikel is een bewerkte versie van een deel van het verslag van mijn onderzoek.

LAV-schatters zijn schatters die sommen van absolute waarden minimaliseren. Een bekend voorbeeld is het volgende. Stel we hebben n waarnemingen $y_1,...,y_n$ en een te schatten locatieparameter b. Als de som

$$\sum_{i=1}^{n} |y_i - b|$$

moet worden geminimaliseerd door het schatten van b resulteert de mediaan m_y van de waarnemingen $y_1,...,y_n$ als schatter voor b. De mediaan m_y is nu de LAV-schatter. Het "break down point" (zie Hampel, Ronchetti, Rousseeuw en Stahel, 1986) van de mediaan is 50 %. Het lijkt mij aannemelijk dat dit ook voor de andere LAV-schatters geldt. Een bewijs voor dit vermoeden heb ik helaas niet kunnen vinden. De LAV-schatter wordt ook wel L_1-schatter genoemd omdat hij overeenstemt met een L_p-schatter met $p = 1$. De L_p-schatter zou in het gegeven voorbeeld de som

$$\sum_{i=1}^{n} |y_i - b|^p$$

minimaliseren. Als we $p = 2$ kiezen hebben we de OLS (=Ordinary Least Squares)-schatter. De schatter b in het voorbeeld is dan gelijk aan het gemiddelde van de y_i's.

De LAV-schatter is minder gevoelig voor uitschieters in de data en dus een robuustere schatter dan de OLS-schatter. Dit komt doordat bij de LAV-methode een afwijking van de regressielijn minder zwaar weegt dan bij de OLS-methode: bij de LAV-methode worden de absolute waarden van de afwijkingen van de regressielijn geminimaliseerd terwijl bij de OLS-methode de kwadraten van deze afwijkingen worden geminimaliseerd. Hierdoor hebben uitschieters in de data niet die enorme invloed op een LAV-schatter die ze op een OLS-schatter hebben. Ook voor dikstaartige verdelingen van de residuen (bijvoorbeeld de Cauchy verdeling) is dit van belang en is de dan meer efficiënte LAV-schatter te verkiezen boven de alom gebruikte OLS-schatter. In dit verslag wordt een schatter efficiënt genoemd als de bijbehorende standaardafwijking klein is. Uitschieters in de data opsporen is niet altijd realiseerbaar. Zo kunnen in een meer-dimensionaal regressiemodel uitschieters niet eenvoudig worden opgespoord om vervolgens uit de dataset te worden verwijderd. In een twee-dimensionaal regressiemodel is het door het tekenen van een plaatje meestal wel mogelijk een idee te krijgen wat de uitschieters zouden kunnen zijn. In het geval van meer dan twee dimensies is dit niet meer mogelijk.

LAV-schatters zijn niet nieuw: zo gebruikte Galileo in 1632 de som van de absolute waarden als criterium in zijn sterrenkundig onderzoek (zie Hald, 1986). Boscovich had in 1757 het idee het minimaliseren van de som van de absolute waarden van de residuen onder de voorwaarde dat de residuen gemiddeld 0 zijn als criterium te gebruiken (zie Koenker en Bassett, 1985). Laplace noemde in 1793 het bepalen van de LAV-schatter in een eenvoudig model de 'method of situations' (zie Dodge, 1987). Tenslotte kwam Edgeworth in 1887 met een lineaire regressie methode die gebruik maakte van LAV-schatters (zie Dodge, 1987).

In deze eeuw was aanvankelijk toch de OLS-schatter de dominante schatter. Een reden hiervoor is dat de OLS-schatter in de praktijk eenvoudiger is te berekenen terwijl LAV-schatters met behulp van ingewikkelde iteratieve procedures worden berekend. Met de komst van snelle computers heeft dit argument zijn waarde verloren. Veel onderzoek is de laatste jaren verricht naar hoe deze iteratieve procedures kunnen worden verfijnd.

De resulterende schatters hoeven niet uniek te zijn. De som van de absolute waarden van de residuen neemt voor de verschillende LAV-schattingen echter wel altijd dezelfde minimale waarde aan. In praktijkgevallen zal niet-uniciteit soms voorkomen, maar dit is meestal geen probleem omdat de verschillende minima dan dicht bij elkaar liggen.

Een andere reden voor de geringe populariteit van de LAV-schatter was het ontbreken van kleine steekproef theorie. Tot voor kort was ook maar weinig bekend over hoe de standaardafwijkingen van de LAV-schatters zouden kunnen worden geschat en konden dus ook geen betrouwbaarheidsintervallen worden bepaald.

Natuurlijk zal het niet meevallen mensen die jarenlang gewoon zijn geweest OLS-schatters te gebruiken te overtuigen van het nut ook naar andere schatters te kijken. De laatste jaren is er een stroom van artikelen in wetenschappelijke tijdschriften gepubliceerd over LAV-schatters. Veel tot dusverre ontbrekende informatie over de LAV-schatter is hierin te vinden. Hopelijk zal deze stroom artikelen helpen om mensen te overtuigen van het nut van LAV-schatters.

In de paragraaf Theoretische Beschouwingen zijn beknopt wat achtergronden van de LAV-methode weergegeven. Ik heb op een IBM gewerkt met het TSO systeem. De berekeningen zijn uitgevoerd met behulp van een SAS-programma. In dit programma worden de SAS-procedures LAV en REG gebruikt. In de SAS-procedure LAV worden de schattingsresultaten bepaald door het minimalisatieprobleem eerst te formuleren als een Lineair Programmerings-probleem en dit vervolgens met behulp van een verbeterde versie van de methode van Barrodale en Roberts (1974) op te lossen. Het schatten van standaardafwijkingen voor de LAV-methode ontbreekt in de SAS-procedure LAV en heb ik zelf geprogrammeerd met behulp van SAS/IML. Data van Organon zijn gebruikt om te bekijken hoe in de praktijk de resultaten van de LAV- en OLS-methode verschillen. Het farmaceutische

bedrijf Organon te Oss maakt deel uit van het AKZO-concern. Een probleem waar men zich bij Organon mee bezighoudt is het verbeteren van tabletteerprocessen. Deze processen kunnen worden geanalyseerd met behulp van statistische methoden door verschillende grootheden in deze processen te variëren.

In het onderzoek bleek dat het kan voorkomen dat door uitschieters in de data de resultaten van de OLS-methode van weinig waarde meer zijn. Na weglating van deze uitschieters bleken er vaak nog meer uitschieters in de data te zitten. Na herhaald weglaten van uitschieters was het resultaat van de OLS-methode vergelijkbaar met dat van de LAV-methode toegepast op de gehele dataset inclusief de uitschieters. Het grote voordeel van de LAV-methode is dus dat de omslachtige, tijdrovende en dubieuze procedure van herhaald weglaten van uitschieters achterwege kan blijven.

Er was in het onderzoek een dataset waar in het geheel geen uitschieters in zaten. Als criterium voor een uitschieter is genomen dat de absolute waarde van het gestandaardiseerde residu (=het residu gedeeld door de bijbehorende standaardafwijking) van de OLS-methode niet groter dan een bepaalde tabelwaarde mag zijn. Deze tabelwaarden zijn afhankelijk van het aantal waarnemingen. De datasets van Organon bestonden uit 53 waarnemingen. De tabelwaarde is ook afhankelijk van het aantal te schatten parameters in het model (inclusief de intercept). Bij 53 waarnemingen ligt de tabelwaarde bij een onbetrouwbaarheid van 5 % tussen de 3.0 en 3.2. De precieze tabelwaarden zijn te vinden in het boek *Outliers in Statistical Data* van Barnett en Lewis (1978). In genoemde dataset zonder uitschieters leverden de LAV- en OLS-methode vergelijkbare resultaten op. In een andere dataset moesten tot drie keer toe uitschieters worden weggelaten. De dataset telde toen nog maar 49 waarnemingen. Pas na dit herhaalde weglaten leverden de LAV- en OLS-methode vergelijkbare resultaten op. Bij de gehele dataset waren in dit geval de resultaten van de LAV-methode ook zonder het weglaten van de uitschieters al goed. De geschatte standaardafwijkingen van de parameters in het model bij toepassing van de OLS-methode op de gehele dataset waren hier enorm groot. De complete onderzoeksresultaten zijn bij de auteur te verkrijgen.

Het is interessant te bekijken in welke gevallen de LAV-methode beter is dan de OLS-methode en in welke gevallen het net andersom is. Er moet wel worden gedefinieerd wanneer een methode als beter dan een andere methode wordt beschouwd. Een criterium is te bekijken welke methode de efficiëntste is. Om dit nader te bekijken is geëxperimenteerd met gesimuleerde data. De resultaten van dit Simulatie-experiment worden in de gelijknamige paragraaf gegeven. De conclusies van het onderzoek en ideeën waar in de toekomst de aandacht zich op zou kunnen richten zijn te vinden in de paragraaf Conclusies en Aanbevelingen voor Toekomstig Onderzoek. Veel literatuur is inmiddels beschikbaar over LAV-

schatters. Vaak overlappen verschillende artikelen elkaar. Een greep van zo min mogelijk overlappende literatuur is te vinden in de literatuurlijst.

Veel dank ben ik verschuldigd aan mijn begeleiders prof.dr. R.D. Gill (RUU), drs. P. Kuiper (ARLA) en dr. A. Sieders (ARLA), die ondanks hun volle agenda's toch telkens weer tijd voor mij vrij wilden maken, en ook aan de andere medewerkers van ARLA die mij met het onderzoek hebben geholpen. Alle goede tips waren zeer welkom en aan de opbouwende kritiek heb ik veel gehad. Zonder genoemde steun had ik het onderzoek nooit op deze wijze kunnen realiseren! Aan prof.dr. R.D. Gill is het bovendien te danken dat dit artikel kon worden gerealiseerd.

Theoretische Beschouwingen Zoals in de vorige paragraaf is vermeld is de OLS-schatter niet zo robuust aangezien uitschieters de schatter sterk kunnen beïnvloeden. De LAV-schatter is hier minder gevoelig voor en zou dus de voorkeur kunnen verdienen. Waarom dit in het nabije verleden niet het geval is geweest is in de vorige paragraaf duidelijk gemaakt.

Alvorens het regressiemodel te beschouwen zal eerst het begrip M-schatter worden gedefinieerd en zal worden gekeken naar locatiemodellen. Na deze opstap zal het regressiemodel zelf nader worden bekeken.

Beide schatters behoren tot de klasse van M-schatters, de zogenaamde gegeneraliseerde maximale aannemelijkheidsschatters. We geven nu de definitie van deze klasse van M-schatters (zie ook Hampel, Ronchetti, Rousseeuw en Stahel, 1986).

Definitie:
Zij $x_1,...,x_n$ identiek en onderling onafhankelijk verdeelde stochastische variabelen met kansdichtheidsfunctie f. De M-schatter $T_n = T_n(x_1, ..., x_n)$ is een van deze stochastische variabelen afhankelijke schatter die de som

$$\sum_{i=1}^{n} \rho(x_i, T_n)$$

minimaliseert met ρ een functie.

Opmerking:
Als ρ een differentieerbare functie is en

$$\psi(x_i, \theta) = \frac{\partial \rho(x_i, \theta)}{\partial \theta}$$

dan is de M-schatter T_n een oplossing van de vergelijking

$$\sum_{i=1}^{n} \psi(x_i, T_n) = 0.$$

We zijn in het bijzonder geïnteresseerd in locatiemodellen en dus in M-schatters T_n die

$$\sum_{i=1}^{n} \rho(x_i - T_n)$$

minimaliseren. Ter illustratie volgen nu twee voorbeelden van zulke M-schatters.

Voorbeeld 1:
Kies

$$f(x_1 - \theta, ..., x_n - \theta) = c^n \exp(-\sum_{i=1}^{n} \rho(x_i - \theta)),$$

$$c = \frac{1}{2} \text{ en } \rho(x_i - \theta) = \frac{|x_i - \theta|}{\sigma}$$

(de x_i's zijn nu verdeeld volgens de Laplace verdeling). De som

$$\sum_{i=1}^{n} \rho(x_i - \theta)$$

is nu minimaal voor de L_1-schatter $\widehat{\theta} = \text{mediaan}(x_1, ..., x_n)$. We vinden dus de maximale aannemelijkheidsschatter voor een steekproef uit de Laplace verdeling.

Voorbeeld 2:
Kies f als in voorbeeld 1, maar kies nu

$$c = \frac{1}{\sigma\sqrt{2\pi}} \text{ en } \rho(x_i - \theta) = \frac{(x_i - \theta)^2}{2\sigma^2}$$

(de x_i's zijn nu normaal verdeeld). Teneinde de M-schatter te bepalen minimaliseren we nu

$$\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}.$$

Er geldt dat

$$\psi(x_i - \theta) = \frac{\partial \rho(x_i - \theta)}{\partial \theta} = -\frac{x_i - \theta}{\sigma^2}$$

en het oplossen van de vergelijking

$$\sum_{i=1}^{n} \psi(x_i - \theta) = \sum_{i=1}^{n} -\frac{x_i - \theta}{\sigma^2} = 0$$

levert de L_2-schatter

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

We vinden dus de maximale aannemelijkheidsschatter voor een steekproef uit de normale verdeling. Geconcludeerd kan worden dat M-schatters een natuurlijke generalisatie zijn van de 'gewone' maximale aaannemelijkheidsschatters omdat in bekende gevallen zoals in de bovenstaande twee voorbeelden de M-schatter samenvalt met de 'gewone' maximale aannemelijkheidsschatter.

Er zijn nog veel meer mogelijkheden om f te kiezen, bijvoorbeeld de Cauchy verdeling of, zoals Huber in zijn boek *Robust Statistics* (1981) voorstelt, een combinatie van de voorbeelden 1 en 2. Hoewel deze mogelijkheden zeker interessant zijn en niet altijd zoals in de twee voorbeelden bekende maximale aannemelijkheidsschatters opleveren beperkt dit onderzoek zich tot een vergelijking tussen LAV-schatters als in voorbeeld 1 en OLS-schatters als in voorbeeld 2. Er is dus niet gekeken naar L_p-schatters met $1 < p < 2$ of naar combinaties van LAV- en OLS-schatters. De twee methoden worden naast elkaar gebruikt om het verschil in schattingsresultaten in een zelfde situatie te illustreren. Bij een groot verschil hebben we waarschijnlijk niet te maken met bij benadering normaal of Laplace verdeelde data. Het is aan te raden de dan meer robuuste LAV-schatter te gebruiken.

We gaan uit van het lineaire regressiemodel $y = X\beta + e$, met e_i, $i = 1,...,n$, de storingstermen, die worden verondersteld identiek en onafhankelijk verdeeld te zijn met kansdichtheidsfunctie f, β_j, $j = 1,...,p$, de te schatten parameters, x_{ij}, $i = 1,...,n$, $j = 1,...,p$, een verzameling van n waarnemingen van p variabelen en y_i, $i = 1,...,n$, de bijbehorende metingen van de afhankelijke variabele. $x_{i1} = 1$ voor $i = 1,...,n$ want we beschouwen een model met intercept. De LAV-schatter bepaalt een vector $\hat{\beta}$ zodanig dat de som

$$\sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} x_{ij}\beta_j|$$

wordt geminimaliseerd.

Voor het bepalen van de standaardafwijkingen van de geschatte parameters van de LAV-schatting is gebruik gemaakt van het resultaat van Bassett en Koenker (1978) dat $\hat{\beta}$ asymptotisch normaal verdeeld is met verwachting β en covariantiematrix

$$\tau^2(X'X)^{-1} = \frac{1}{4f^2(m)}(X'X)^{-1}.$$

In de laatste formule stelt m de mediaan voor. Meer informatie over dit resultaat is te vinden in Bloomfield en Steiger (1983) en in werk van Pollard dat in 1990 is verschenen.

Deze laatste formule kan aannemelijk worden gemaakt door het locatiemodel te beschouwen. De matrix X is dan een kolom enen en er volgt:

$$\tau^2(n)^{-1} = \frac{1}{4f^2(m)}(n)^{-1} = \frac{1}{4nf^2(m)}.$$

Dat dit de asymptotische variantie van de steekproefmediaan met n waarnemingen is valt als volgt in te zien. Laat F_n de verdelingsfunctie van de data in de steekproef zijn, F de werkelijke verdeling en M_n de steekproefmediaan. Dan geldt $F_n(M_n) = F(m) = \frac{1}{2}$ en dus volgt $0 = \sqrt{n}(F_n(M_n) - F(m)) = \sqrt{n}(F_n(M_n) - F(M_n)) + \sqrt{n}(F(M_n) - F(m))$. De eerste term van deze laatste uitdrukking is bij benadering een gestandaardiseerde binomiale verdeling en dus bij benadering $\mathrm{N}(0, \frac{1}{4})$ verdeeld, terwijl met behulp van een Taylor ontwikkeling naar M_n is in te zien dat de tweede term bij benadering gelijk is aan $\sqrt{n}(M_n - m)f(m)$. We mogen concluderen dat $\sqrt{n}(M_n - m)f(m)$ bij benadering $\mathrm{N}(0, \frac{1}{4})$ is verdeeld, waarmee aannemelijk is geworden dat M_n asymptotisch $\mathrm{N}(m, 1/4nf^2(m))$ is verdeeld.

Dit resultaat impliceert dat voor elke verdeling van de ϵ_i's, waarvoor de steekproefmediaan asymptotisch een kleinere standaardafwijking heeft dan het steekproefgemiddelde, de LAV-methode efficiënter is dan de OLS-methode. Het is in het geval van dikstaartige verdelingen en in het geval dat er uitschieters in de data zitten dus efficiënter de mediaan in plaats van het gemiddelde als schatter te gebruiken. In deze gevallen is de LAV-methode dus efficiënter dan de OLS-methode.

Omdat in de praktijk $f(m)$ onbekend is zal deze moeten worden geschat om standaardafwijkingen te kunnen schatten. In het boek *Density Estimation for Statistics and Data Analysis* van Silverman (1986) zijn verschillende methoden beschreven om $f(m)$ te schatten. In dit onderzoek is er voor gekozen om $f(m)$ te schatten met $\hat{f}(m)$ waarbij

$$\hat{f}(m) = \frac{2}{\sqrt{n}(r_{hoog} - r_{laag})}.$$

Hierin stelt r_{hoog} het grootste van de $\sqrt{n}$ kleinste positieve residuen voor en analoog stelt r_{laag} het kleinste van de $\sqrt{n}$ grootste negatieve residuen voor.

Om te toetsen of de vector $\hat{\beta}$ significant van 0 verschillend is wordt de door McKean en Sievers (1987) gesuggereerde toetsgrootheid D gebruikt met

$$D = (\sum_{i=1}^{n} |y_i - \hat{\alpha}_0| - \sum_{i=1}^{n} |y_i - \hat{\alpha} - \hat{\beta}x_i|) \times 4f(m).$$

De eerste som is de gecorrigeerde som van absolute waarden ($\hat{\alpha}_0$ is de mediaan van de y_i's) en de tweede som is de residuele som van absolute waarden (=de som van de absolute waarden van de residuen). D heeft asymptotisch een $\chi^2(q)$ verdeling onder de nulhypothese

$H_0 : \beta = 0$. Hieruit volgt dat D/q asymptotisch een $F(q, n-q)$ verdeling heeft onder dezelfde H_0. Telkens stelt q het aantal elementen in de vector β voor.

Het gebruikelijke OLS-analogon van de toetsgrootheid D is de toetsgrootheid X^2 met

$$X^2 = (\sum_{i=1}^{n}(y_i - \hat{\alpha}_1)^2 - \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2) \times \frac{1}{\hat{\sigma}^2}.$$

$\hat{\alpha}_1$ stelt hierin het gemiddelde van de y_i's voor en $\hat{\sigma}^2$ de geschatte variantie. Ook X^2 heeft asymptotisch een $\chi^2(q)$ verdeling onder H_0 en X^2/q heeft net als D/q asymptotisch een $F(q, n-q)$ verdeling onder H_0.

De toetsen D en X^2 zijn te generaliseren naar situaties waarin wordt getoetst of slechts een aantal elementen van de vector β gelijk aan 0 zijn. Als we ω als de gerestricteerde ruimte voor de vector β en Ω als de ongerestricteerde ruimte voor de vector β definiëren en als SAE de Sum of Absolute Errors voorstelt en SSE de Sum of Squared Errors zijn D en X^2 als volgt te schrijven:

$$D = \frac{SAE(\omega) - SAE(\Omega)}{\hat{\tau}/2} = (SAE(\omega) - SAE(\Omega)) \times 4f(m)$$

en

$$X^2 = \frac{SSE(\omega) - SSE(\Omega)}{\hat{\sigma}^2}.$$

Simulatie-experiment Zoals in de inleiding reeds is vermeld is het interessant te bekijken in welke gevallen de LAV-methode efficiënter is dan de OLS-methode en in welke gevallen het net andersom is. In deze paragraaf worden de LAV- en OLS-methode vergeleken voor gesimuleerde data. We gaan uit van het lineaire regressiemodel $y = X\beta + e$ dat in de paragraaf Theoretische Beschouwingen is beschreven. X wordt vast gekozen door een eenvoudig model met intercept en x_1 en x_2 als verklarende variabelen te beschouwen. x_1 en x_2 zijn de belangrijkste verklarende variabelen in de dataset van Organon. De achterliggende gedachte om deze variabelen in dit model op te nemen is dat op deze wijze de storingstermen relatief klein blijven. Het model is nu te schrijven als $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. De vector β wordt geïnitialiseerd als de nulvector. Omdat we 53 waarnemingen hebben worden eerst 53 randomgetallen gegenereerd. Deze randomgetallen worden in de vector e gezet. Daar β was geïnitialiseerd als de nulvector betekent dit dat de randomgetallen in de vector y terecht komen. Nu we X en y hebben worden met behulp van zowel de LAV- als de OLS-methode de parameters in β geschat. Ook de bijbehorende standaardafwijkingen worden op de bekende manier geschat.

Voor het verkrijgen van randomgetallen van de uniforme, normale en Cauchy verdeling is gebruik gemaakt van de SAS-functies ranuni, rannor en rancau. De randomgetallen van

de uniforme verdeling zijn vervolgens getransformeerd tot randomgetallen van de Laplace verdeling. Tenslotte is nog een bimodale verdeling bekeken. Randomgetallen van de $N(0,1)$ verdeling zijn getransformeerd tot randomgetallen van de $N(-1,1)$ en $N(1,1)$ verdelingen. De bimodale verdeling is verkregen door uit beide laatst genoemde verdelingen een gelijk aantal randomgetallen te nemen. Het SAS-programma, dat bij de auteur is te verkrijgen, is enigszins aangepast om de experimenten met de randomgetallen uit te kunnen voeren.

Voor de normale, Cauchy, Laplace en bimodale verdeling is het experiment 100 keer gedaan met elke keer andere randomgetallen uit de genoemde verdelingen. De gemiddelde lengten van de betrouwbaarheidsintervallen van de LAV- en OLS-methode voor de parameterschattingen van β_0, β_1 en β_2 zijn bepaald. De resultaten zijn te vinden in de tabel simulatie-experiment 100 experimenten. Boven twee kolommen in deze tabel staat gemiddelde lengte betr.int., hetgeen duidt op de gemiddelde lengte van het 95 % betrouwbaarheidsinterval van de betreffende parameterschattingen. Namen in de tabel beginnen met een N, C, L of B en de bijbehorende waarden zijn dan bepaald met behulp van randomgetallen van respectievelijk de normale, de Cauchy, de Laplace of de bimodale verdeling. Vervolgens wordt vermeld of het resultaten van de LAV- of de OLS-methode zijn en tenslotte volgt nog β_0, β_1 of β_2.

	gemiddelde lengte betr.int.	# sig.		gemiddelde lengte betr.int.	# sig.
NLAVβ_0	0.64	7	LLAVβ_0	0.48	9
NLAVβ_1	0.79	6	LLAVβ_1	0.59	9
NLAVβ_2	0.98	12	LLAVβ_2	0.73	14
NOLSβ_0	0.54	6	LOLSβ_0	0.54	5
NOLSβ_1	0.67	4	LOLSβ_1	0.67	6
NOLSβ_2	0.83	5	LOLSβ_2	0.83	5
CLAVβ_0	0.98	5	BLAVβ_0	1.15	3
CLAVβ_1	1.21	9	BLAVβ_1	1.43	2
CLAVβ_2	1.50	7	BLAVβ_2	1.76	3
COLSβ_0	9.92	3	BOLSβ_0	0.79	1
COLSβ_1	12.25	6	BOLSβ_1	0.97	0
COLSβ_2	15.17	10	BOLSβ_2	1.20	1

tabel simulatie-experiment 100 experimenten

Zoals te verwachten was is bij de randomgetallen van de normale verdeling de OLS-methode efficiënter dan de LAV-methode en is het bij de randomgetallen van de Laplace verdeling net andersom. Bij de randomgetallen van de normale verdeling was de OLS-methode 78 van de 100 keer efficiënter dan de LAV-methode en bij de randomgetallen van de Laplace verdeling was de LAV-methode 68 van de 100 keer efficiënter dan de OLS-methode. Uit de tabel valt ook op te maken dat de OLS-methode het bij de randomgetallen van de Cauchy verdeling dramatisch slecht doet. Hier was de LAV-methode maar liefst 99 van de 100 keer efficiënter dan de OLS-methode. Bij de randomgetallen van de bimodale verdeling tenslotte doet de LAV-methode het slecht. Van de 100 keer was de OLS-methode 95 keer efficiënter dan de LAV-methode.

In de tabel wordt behalve de gemiddelde lengten van de betrouwbaarheidsintervallen ook het aantal keren dat een bepaalde parameter met een onbetrouwbaarheid van 5 % significant is weergegeven. Dit aantal keren wordt met # sig. aangeduid. De aantallen die zijn bepaald met behulp van de OLS-methode liggen rond de 5, terwijl de aantallen die zijn bepaald met behulp van de LAV-methode wat hoger lijken uit te komen. Opvallend is dat de aantallen bij de randomgetallen van de bimodale verdeling lager zijn dan de aantallen bij de randomgetallen van de andere verdelingen.

Uit dit simulatie-experiment kunnen we concluderen dat, als we zo efficiënt mogelijke schatters willen hebben, we bij de randomgetallen van de normale en bimodale verdeling de voorkeur moeten geven aan de OLS-methode, terwijl bij de randomgetallen van de Laplace en Cauchy verdeling de LAV-methode de voorkeur moet krijgen. In de praktijk is het niet bekend volgens welke verdeling de data bij benadering zijn verdeeld en is het verstandig de twee methoden naast elkaar te gebruiken om het verschil in schattingsresultaten te kunnen bekijken. Als slechts één van beide methoden wordt toegepast kan informatie verborgen blijven. De robuustheid van de LAV-methode wordt fraai geïllustreerd bij de randomgetallen van de Cauchy verdeling: de gemiddelde lengten van de betrouwbaarheidsintervallen van de OLS-methode zijn hier zeer veel groter dan die van de LAV-methode.

Conclusies en Aanbevelingen voor Toekomstig Onderzoek In dit artikel zijn de LAV- en OLS-methode voor het schatten van modellen vergeleken. In het simulatie-experiment werd geïllustreerd hoe weinig efficiënt de OLS-methode bij de randomgetallen van de dikstaartige Cauchy verdeling is. Uit oogpunt van efficiëntie verdient de LAV-methode in het geval van dikstaartige verdelingen de voorkeur. Zoals te verwachten was bleek dat bij de randomgetallen van de normale verdeling de OLS-methode efficiënter was dan de LAV-methode en dat het bij de randomgetallen van de Laplace verdeling net andersom was. In het simulatie-experiment werd ook geïllustreerd dat bij de randomgetallen

van de bimodale verdeling de OLS-methode efficiënter is dan de LAV-methode. Het is duidelijk dat er geen algemene uitspraak valt te doen welke methode de efficiëntste is. Wel is duidelijk dat de LAV-methode robuuster is dan de OLS-methode. Het verdient de voorkeur beide methoden naast elkaar te gebruiken en de resultaten van de twee methoden te vergelijken alvorens conclusies te trekken over de parameterschattingen en bijbehorende standaardafwijkingen.

Het berekenen van LAV-schatters is geen probleem nu men de beschikking heeft over snelle computers. Standaardafwijkingen van de geschatte parameters kunnen ook worden bepaald. Een probleem dat blijft bestaan is het ontbreken van kleine steekproef theorie. Als er in verhouding tot het aantal te schatten parameters weinig waarnemingen zijn is het aantal vrijheidsgraden gering. Een gevolg is dan dat de schatting voor $f(m)$ niet erg stabiel is. Omdat deze schatting wordt gebruikt voor het bepalen van de standaardafwijkingen van de geschatte parameters zijn deze standaardafwijkingen niet zo betrouwbaar. Wellicht is een oplossing gebruik te maken van de bootstrapmethode, zodat dan ook voor kleine steekproeven de LAV-methode resultaten oplevert, waarin we meer vertrouwen kunnen hebben. Recent zijn een aantal artikelen verschenen over het gebruik van de bootstrapmethode. In Stangenhaus (1987) wordt de bootstrapmethode gebruikt om standaardafwijkingen en betrouwbaarheidsintervallen voor LAV-schatters te bepalen. Overigens is het de vraag hoeveel vertrouwen we mogen hebben in de geschatte standaardafwijkingen van de OLS-methode bij een gering aantal vrijheidsgraden. Als de veronderstelde normaliteit van de storingstermen geen redelijke veronderstelling is, dan is het een illusie te denken dat we nog vertrouwen mogen hebben in de met de OLS-methode bereikte resultaten.

Een andere aanbeveling voor toekomstig onderzoek is te bekijken of de toetsgrootheid D van McKean en Sievers (1987) kan worden verbeterd door deze te vermenigvuldigen met $(n-q)/n$. Ook de eigenschappen van de toetsgrootheid D/q en andere toetsen met een F verdeling zouden kunnen worden bestudeerd. Over het introduceren van de correctieterm $(n-q)/n$ en verschillende toetsen met een F verdeling is meer informatie te vinden in McKean en Schrader (1987) en Schrader en McKean (1987).

Tenslotte is het ook interessant een vergelijking te maken tussen de LAV-methode en de Least Median of Squares methode van Rousseeuw (1984). Beide methoden zijn robuuster dan de OLS-methode. Welke methode is echter het meest "robuust"?

Literatuurlijst

Barnett, Vic, en Lewis, Toby, *Outliers in Statistical Data*, (Wiley, Chichester), 1978

Barrodale, I., en Roberts, F.D.K., Algorithm 478: Solution of an Over-Determined System of Equations in the l_1 Norm, *Communications of the Association of Computing Machinery*, 1974, Volume 17, Number 6, blz. 319-320

Bassett, Gilbert Jr., en Koenker, Roger, Asymptotic Theory of Least Absolute Error Regression, *Journal of the American Statistical Association*, September 1978, Volume 73, Number 363, blz. 618-622

Bloomfield, Peter, en Steiger, William L., *Least Absolute Deviations Theory, Applications and Algorithms*, (Birkhäuser, Boston), 1983

Dielman, Terry E., en Pfaffenberger, Roger C., Bootstrapping in Least Absolute Value Regression: an Application to Hypothesis Testing, *Communications in Statistics B, Simulation and Computation*, 1988, Volume 17, Number 3, blz. 843-856

Dodge, Yadolah, An Introduction to L_1-norm based Statistical Data Analysis, *Computational Statistics & Data Analysis*, 1987, Volume 5, blz. 239-253

Draper, N.R., en Smith, H., *Applied Regression Analysis*, (Wiley, New York), 1966

Hald, A., Galileo's Statistical Analysis of Astronomical Observations, *International Statistical Review*, 1986, Volume 54, Number 2, blz. 211-220

Hampel, Frank R., Ronchetti, Elvezio M., Rousseeuw, Peter J., en Stahel, Werner A., *Robust Statistics, The Approach Based on Influence Functions*, (Wiley, New York), 1986

Huber, Peter J., *Robust Statistics*, (Wiley, New York), 1981

Koenker, Roger, en Bassett, Gilbert, On Boscovich's Estimator, *The Annals of Statistics*, 1985, Volume 13, Number 4, blz. 1625-1628

McKean, Joseph W., en Schrader, Ronald M., Least Absolute Errors Analysis of Variance, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, 1987, blz. 297-305

McKean, Joseph W., en Sievers, Gerald L., Coefficients of Determination for Least Abso-

lute Deviation Analysis, *Statistics & Probability Letters*, 1987, Volume 5, blz. 49-54

Pollard, David, *Empirical Processes: Theory and Applications*, (Institute of Mathematical Statistics, Hayward, California), 1990

Pollard, David, Asymptotics for Least Absolute Deviation Regression Estimators, *Econometric Theory*, 1990

Rousseeuw, Peter J., Least Median of Squares Regression, *Journal of the American Statistical Association*, December 1984, Volume 79, Number 388, blz. 871-880

Schrader, Ronald M., en McKean, Joseph W., Small Sample Properties of Least Absolute Errors Analysis of Variance, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, 1987, blz. 307-321

Schwarz, Gideon, Least-Absolute-Value Prediction Lines in Closed Form, *Journal of the American Statistical Association*, December 1987, Volume 82, Number 400, blz. 1150-1152

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, (Chapman and Hall, London), 1986

Sposito, V.A., en Tveite, Michael D., On the Estimation of the Variance of the Median used in L_1 Linear Inference Procedures, *Communications in Statistics A, Theory and Methods*, 1986, Volume 15, Number 4, blz. 1367-1375

Stangenhaus, Gabriela, Bootstrap and Inference Procedures for L_1 Regression, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, 1987, blz. 323-332

ontvangen 13- 2 -1991
geaccepteerd 22- 10-1991