Trade-regressie

TUE-RC 82618

Eindhoven University of Technology
Computer Centre Note 46

TRADE-regressie

Jan B. Dijkstra

Samengesteld voor de Statistische Dag
9 april 1990.

TRADE-regressie

Jan B. Dijkstra

Samenvatting

TRADE-regressie is een tweestaps herwogen kleinste kwadraten methode. De naam is een afkorting van TRim And DElete. De eerste stap met (mogelijk asymmetrische) trimming van de residuen maakt de methode robuust tegen uitschieters. En de tweede stap (waarin alleen zeer verdachte punten worden genegeerd) herstelt de statistische efficientie voor normale verdelingen.

De laatste jaren krijgt high-breakdown regressie de volle aandacht van methode-ontwikkelaars. Denk hierbij aan Repeated Median van Siegel (1982), Least Median of Squares van Rousseeuw (1984), S-schatters van Rousseeuw en Yohai (1984), MM-schatters van Yohai (1987) en Tau-schatters van Yohai en Zamar (1988). Deze zeer robuuste methoden zijn helaas door hun grote bewerkelijkheid alleen geschikt voor modellen met een beperkt aantal variabelen. Wellicht mede daarom zijn ze (nog) niet opgenomen in de bekende statistische pakketten.

TRADE-regressie is een compromis. De methode is ontwikkeld met de volgende randvoorwaarden: (1) Ze moet toepasbaar zijn met de middelen die reeds in pakketten als SAS, SPSS en BMDP aanwezig zijn. (2) Ook voor grote data-sets moet de methode haalbaar blijven en (3) De robuustheid moet niet te zeer ten koste gaan van de efficientie.

Masking en Swamping

De titel van deze paragraaf bevat drie woorden waarvan twee anglicismen. Als er goede nederlandse alternatieven voor bestaan, dan hoort de auteur dezes dat graag.

Veronderstel dat U een aantal punten heeft in een p-dimensionale ruimte en dat hieraan een regressie-model kan worden aangepast dat alle denkbare soorten kritiek kan weerstaan. U laat er alle diagnostische hulpmiddelen van de laatste twintig jaar op los en vindt niets verdachts.

Deze ideale situatie gaan we nu bederven. In de prediktor-ruimte kiezen we een paar punten die buiten het kleinste convexe omhulsel van de overige punten liggen. Aan deze punten kennen we y-waarden toe die niet aan het regressie-model voldoen. Deze hefboom-punten (leverage points) situeren we relatief dicht bij elkaar. Vervolgens wordt het model opnieuw aangepast.

Omdat het wegnemen van een enkel punt hiervan de aanpassing nauwelijks aantast (de overige hefboom-punten nemen de rol gewoon over) maakt het groepje de invloed van een enkel lid onzichtbaar. Dit effect heet masking.

Doordat de toegevoegde hefboom-punten het aangepaste hypervlak doen kantelen nemen de residuen van sommige punten met extreme prediktor-waarden aanzienlijk toe (in absolute waarde). Deze op zich goed passende punten lijken daardoor uitschieters. Dat effect heet swamping.

Hawkins, Bradu en Kass (1984) schreven over deze problematiek een lezenswaard artikel en verzonnen een data-set ter illustratie van enkele eigenschappen. In de bijlage is deze data-set gegeven op pagina 1 en 2. Er zijn 75 punten in een 4-dimensionale ruimte die door hun index kunnen worden geidentificeerd. Ten behoeve van een regressie-model zijn er drie prediktoren x_1, x_2 en x_3 en er is een response-variabele y.

Met de punten 15 tot 75 is niets bijzonders aan de hand. De punten 1 tot 10 zijn kunstmatige hefboom-punten die op een kluitje liggen en flink aan het hypervlak trekken. En de punten 11 tot 14 hebben weliswaar extreme posities in de prediktor-ruimte, maar passen verder prima bij de punten 15 tot 75. Sommige auteurs gebruiken hiervoor de aanvechtbare aanduiding van goede hefboom-punten.

Klassieke aanpassing

Op pagina 3 en 4 van de bijlage zijn de resultaten te zien van een gewone regressie (met het pakket SAS op een VAX-VMS). Het ziet er zo op het eerste gezicht aardig uit. De hypothese dat y niet door de prediktoren wordt voorspeld levert een overschrijdingskans op van 0.0001. Dat is in overeenstemming met de determinatiecoefficient R^2 van 0.6018. Van de prediktoren is x_2 significant en x_3 zelfs zeer significant (overschrijdingskansen van 0.0343 respectievelijk 0.0040).

In de bijlage staan de gestudentiseerde residuen op pagina 5 en 6. Het is gebruikelijk om een punt als verdacht aan te duiden als de absolute waarde van het bijbehorende gestudentiseerde residu groter is dat 2.5. En dat gebeurt hier alleen bij de extreme (maar goed bij het hypervlak door punt 15 tot 75 passende) punten 11 tot 14. Een behoorlijk misleidend resultaat dus.

Naast de gestudentiseerde residuen worden de Cook (1977) statistics gegeven. Een waarde groter dan 1 betekent dat weglating van het bijbehorende punt de vector van regressiecoefficienten buiten het simultane 50 procent betrouwbaarheidsgebied zou plaatsen dat hoort bij alle waarnemingen. Deze maat is dus zeer geschikt om individuele hefboompunten op te sporen. Maar door hun verstrengeling worden de eerste tien punten niet ontmaskerd. Het enige verdachte punt (volgens de Cook-statistic) is punt 14 met een waarde van 2.114. Dit is echter misleidend; deze hoge waarde komt door de invloed van de eerste tien punten. Zonder deze zou punt 14 uitstekend passen bij het aangepaste model door de overige punten.

Formularium

Het beschouwde regressie-model is van de vorm $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$. De matrix X bestaat uit de elementen x_{ij}, voorafgegaan door een kolom eenen. Hierbij loopt i van 1 tot n en j van 1 tot k. In matrix-notatie wordt het regressie-model dus $y = X\beta + \epsilon$. De vector β heeft als lengte $p = k + 1$ en zijn elementen zijn genummerd van 0 tot k.

Laat b de schatter voor β zijn. Deze wordt vastgelegd door de normaalvergelijkingen $X^T Xb = X^T y$ zodat $b = (X^T X)^{-1} X^T y$. Voor de aangepaste waarden geldt $\hat{y} = Xb$. Substitutie levert vervolgens $\hat{y} = X(X^T X)^{-1} X^T y$. De matrix $X(X^T X)^{-1} X^T$ wordt nu H genoemd (de hat matrix die een hoedje op de y zet) zodat $\hat{y} = Hy$. De residuen worden nu gegeven door $e = y - \hat{y}$ ofwel $e = (I - H)y$.

De restvariantie σ^2 wordt geschat als de restkwadratensom gedeeld door het bijbehorende aantal vrijheidsgraden $s^2 = MSE = SSE/(n-p)$. Hierbij geldt $SSE = e^T e$ maar voor de berekening wordt doorgaans van $SSE = y^T y - b^T X^T y$ gebruik gemaakt.

Voor de variantie van de residuen geldt $V(e) = \sigma^2(I - H)$ zodat een natuurlijke definitie voor gestudentiseerde residuen als volgt luidt:

$$f_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

Hierbij stelt h_{ii} een diagonaal-element van de matrix H voor. De Cook-statistic kan nu met de voorgaande elementen worden opgebouwd:

$$D_i = \frac{f_i^2 h_{ii}}{p(1-h_{ii})}$$

Helaas bleken de in deze paragraaf behandelde klassieke middelen niet in staat om goed met de data van Hawkins, Bradu en Kass om te gaan. Maar misschien valt de tekortkoming nog te herstellen. Naast vele artikelen is er immers een lijvig handboek van Belsley, Kuh en Welsch (1980) over regressie-diagnostiek. En de hierin gepresenteerde hulpmiddelen zijn inmiddels opgenomen in de bekende statistische pakketten.

Invloedsmaten

De eerste maat van Belsley, Kuh en Welsch (dit trio wordt verder als BKW aangeduid) is RSTUDENT (voor de waarden zie pagina 7 tot 10 van de bijlage). Dit is een gestudentiseerd residu waarbij de Mean Square Error berekend wordt op basis van alle punten behalve het punt waarop dit studentized deleted residu betrekking heeft. Voor individuele hefboom-punten is dit een zeer goede maat, maar tegen verstrengeling is hij niet bestand. Voor de punten 11 tot en met 14 wordt de grens van 2.5 (in absolute waarde) overschreden. En dat resultaat is misleidend.

De diagonaal-elementen van de hat-matrix duiden de excentriciteit in de prediktor-ruimte aan. BKW bevelen als drempelwaarde $\frac{2p}{n}$ aan en dat is hier 0.1067. Voor de punten 12 tot en met 14 wordt deze grens overschreden. Dit is in overeenstemming met de kunstmatige wijze waarop de data geconstrueerd zijn.

COVRATIO is de verandering in de determinant van de covariantie-matrix van de regressie-coefficienten door weglating van steeds een enkel punt. De door BKW aanbevolen drempelwaarde is vastgelegd door:

$$|COVRATIO - 1| \geq \frac{3p}{n}$$

In dit geval betekent dat dat punten verdacht zijn als COVRATIO kleiner is dan 0.84 of groter dan 1.16. Weer zijn de punten 11 tot en met 14 de enige verdachte en dus kunnen we constateren dat ook deze maat door verstrengeling voor de gek gehouden kan worden.

DFFITS meet het geschaalde verschil tussen $\hat{y}_i$ en $\hat{y}_i(i)$. Bij de laatstgenoemde aangepaste waarde is het hypervlak berekend op basis van alle waarnemingen behalve de i-de. Hier is de grens $2\sqrt{p/n}$ en dat is 0.4619. Dit criterium markeert de punten 2, 7, 8 en 11 tot en met 14. Dus van de tien slechte hefboom-punten worden er drie ontmaskerd en van de vier goede hefboom-punten wordt gesuggereerd dat ze slecht zijn. Kortom: niet al te best.

DFBETAS meet het geschaalde verschil tussen b_i en $b_i(i)$ met een definitie analoog aan het bovenstaande. Omdat er vier parameters zijn levert dit voor iedere observatie vier waarden op. De door BKW aanbevolen drempel is $2/\sqrt{n}$ en dat is hier 0.2309. Wederom worden (naast het slechte hefboom-punt nummer 10) de punten 11 tot en met 14 verdacht bevonden en wel als volgt:

DF β_0: punten 13 en 14

DF β_1: punten 11, 13 en 14

DF β_2: punten 13 en 14

DF β_3: punten 10 tot 14

Zoals hierboven beschreven kan men met deze maten werken in SPSS, SAS en BMDP. Waarschijnlijk in nog wat meer pakketten, maar daar weet de auteur dezes niet veel van. STATGRAPHICS (versie 3.0 of lager) is in deze een tamelijk droevig geval: hier wordt alleen gereageerd op grote positieve waarde van DFFITS (de gebruiker zou de analyse moeten herhalen met -y in plaats van y).

Met routine-bibliotheken als NAG, IMSL en de op het Rekencentrum van de TUE ontwikkelde PP4 is nog een andere aanpak mogelijk die hier succesievelijke eliminatie zal worden genoemd: Zoek de observatie met de meest extreme invloedswaarde. Als deze boven de grens ligt, verwijder dan het punt en herhaal dit proces. Geen van de invloedsmaten van BKW leidt bij dit voorbeeld tot een bevredigend resultaat. Ter besparing van papier zal een verslag hiervan niet worden opgenomen.

Grafische diagnostiek

Op pagina 11 van de bijlage zien we het resultaat van de Kolmogorov toets op normaliteit. De toetsingsgrootheid (voor de residuen) is 0.180946 en dat levert een overschrijdingskans op van minder dan 0.01. De normaliteit is dus twijfelachtig en een grafisch onderzoek naar de aard van de afwijkingen gewenst.

Op pagina 12 van de bijlage zijn de resultaten grafisch weergegeven. Langs de horizontale as ligt de index en langs de vertikale het residu. De punten zijn aangegeven door letters. De

kunstmatige slechte leverage points (1 tot 10) met een b, de goede leverage points met een g (11 tot 14) en de overige (normale) punten met een n (15 tot 75). Rousseeuw en Leroy (1987) hebben deze data ook onderzocht en kwamen tot hetzelfde puntenpatroon (zij gebruikten geen codes om de drie categorien te onderscheiden). Door de figuur legden zij horizontale lijnen op de hoogtes $2.5\hat{\sigma}$ en $-2.5\hat{\sigma}$. Hierbij is σ geschat als $\sqrt{MSE}$. De bovenste grens werd door geen enkel residu overschreden, maar de onderste door de residuen 11, 12 en 13 ofwel drie van de vier goede hefboom-punten.

Een plotje van de residuen tegen de indices heeft in het algemeen alleen zin bij een tijdreeks en als bovendien de data in volgorde van waarneming zijn aangeboden. In dit geval ligt een plotje van de residuen tegen de voorspelde waarden meer voor de hand. Immers geldt onder de regressie-vooronderstellingen dat de residuen en de aangepaste waarden ongecorreleerd zijn. Een dergelijk plotje moet dus een ongestruktureerde chaos opleveren. Daarvan is echter o pagina 13 van de bijlage in het geheel geen sprake. Dit plaatje geeft zo duidelijk weer wat er aan de hand is dat een verbale explicatie weinig zinvol is.

TRADE regressie

Op pagina 14 tot 21 van de bijlage wordt TRim And DElete (TRADE) regressie gedemonsteerd. De eerste stap is een vorm van trimming. De variable ARES1 bevat de absolute waarden van de residuen. De mediaan hiervan is 0.725867. Alle waarnemingen waarvoor ARES1$\leqslant$0.725867 krijgen nu gewicht 1 en de overige gewicht 0. Dus de best passende helft wordt bewaard en de rest voorlopig terzijde gelegd.

Bewust is niet gekozen voor het toekennen van gewicht 1 aan de punten met residuen tussen het eerste en het derde kwartiel en gewicht 0 aan de buiten liggende punten. Dat zou namelijk symmetrische contaminatie veronderstellen. En bij het soort robuustheid dat TRADE regressie probeert na te streven wordt symmetrie in aantal bij de eventuele uitschieters niet verondersteld.

De fractie genegeerde waarnemingen is op 0.5 gesteld omdat (1) gekozen is voor zo groot mogelijke robuustheid in deze eerste stap en (2) bij een grotere fractie uitschieters deze toch niet meer van de goede waarnemingen onderscheidbaar zijn.

Het is goed nu reeds in te zien dat niet gegarandeerd kan worden dat hefboom-punten altijd gewicht 0 zullen krijgen. Hierop komen we verderop nog terug.

Met de gevormde gewichten wordt nu de aanpassing van het regressie-model herhaald. Teneinde de restvariantie schatbaar te houden moet het aantal waarnemingen met gewicht 1 strikt groter zijn dan het aantal aan te passen parameters. Daarom vergt TRADE regressie ruim tweemaal zoveel waarnemingen als er parameters zijn. Gelukkig is daar in de praktijk meestal wel aan voldaan.

Op bladzijde 16 en 17 van de bijlage is te zien dat de residuen van de leverage points zich verder verwijderd hebben van de bulk van de data. De residuen van de goede hefboom-punten zijn juist kleiner geworden. Kortom: een stap in de goede richting. Maar we zijn er nog niet, want op pagina 17 loopt de as van de puntenwolk door de normale waarnemingen en de goede hefboom-punten nog niet horizontaal. En dat zou toch wel moeten, willen de residuen en de aangepaste waarden onafhankelijk zijn.

Na deze TRIM stap komt nu de DELETE-stap die de efficientie voor normale verdelingen moet herstellen. Op pagina 18 van de bijlage wordt de mediaan berekend van de absolute waarden van de residuen na de TRIM-stap. De residuen zijn opgeslagen in ARES2 en de mediaan is 0.633128. Hiermee valt de schaalparameter σ als volgt robuust te schatten:

$$\hat{\sigma} = 1.4826 med_i ARES2_i$$

De factor 1.4826 maakt deze schatter bij benadering consistent voor normaal verdeelde fouten. Met $\hat{\sigma}$ vormen we nu gestandaardiseerde residuen $d_i = RES2_i / \hat{\sigma}$. De vector RES2 bevat hierbij de residuen na de TRIM-stap. Nu krijgen alle waarnemingen gewicht 1 behalve die waarvoor de absolute waarde van d_i groter is dat 2.5. Met deze gewichten wordt het model opnieuw aangepast. Op pagina 19 van de bijlage zien we de aanpassing en op pagina 20 en 21 de residuen. De leverage points zijn nu volkomen geisoleerd en de as door het gros van de punten ligt volkomen horizontaal. Prima dus. Verderop zullen we nagaan of dat toeval was, of dat deze strategie werkelijk kwaliteiten heeft.

In de meeste statistische pakketten kan TRADE regressie zonder veel moeilijkheden worden toegepast. Op het Rekencentrum van de TUE is dit gedaan met SAS, SPSS, BMDP en PP4. De noodzakelijke hulpmiddelen zijn dermate elementair dat ook bij andere pakketten geen problemen te verwachten zijn. Vaak zal het zelfs mogelijk zijn om een macro te schrijven dat iedere keer met hetzelfde gemak kan worden toegepast als een gewone regressie. In dat opzicht onderscheidt TRADE regressie zich in gunstige zin van enige verderop te behandelen robuuste alternatieven.

Least Median of Squares

Bij de data van Hawkins, Bradu en Kass ligt het gebruik van een high-breakdown methode meer voor de hand dan klassieke regressie (al zal de gebruiker zich dat op grond van de data-set alleen niet realiseren). Mogelijkheden zijn er genoeg: Repeated Medians van Siegel (1982), Least Median of Squares van Rousseeuw (1984), S-schatters van Rousseeuw en Yohai (1984), MM-schatters van Yohai (1987) en tenslotte de zeer efficiente τ-schatters van Yohai en Zamar (1988).

Het breakdown punt een schatter is in 1982 door Donoho en Huber als volgt gedefinieerd. Beschouw een schatter op basis van een steekproef met n elementen. Vervang nu hiervan m elementen door andere waarden (eventueel zo ongunstig mogelijk gekozen). Het breakdown punt ϵ^* is nu gedefinieerd als de kleinste waarde van $\frac{m}{n}$ waarvoor het niet mogelijk is de vervangende punten zo ongunstig te kiezen dat de schatter het oneindige ingetrokken wordt. Meestal wordt met ϵ^* de asymptotische waarde bedoeld; hierbij gaat n naar oneindig. De hoogst haalbare waarde voor ϵ^* is 0.5 (anders is de contaminatie niet meer van de goede data onderscheidbaar) en deze waarde wordt voor bovengenoemde high breakdown schatters gehaald.

Nu zal het resultaat van TRADE regressie vergeleken worden met dat van een high-breakdown methode. Gekozen is voor Least Median of Squares (voortaan aan te duiden als LMS) omdat de auteur dezes daarvoor een programma beschikbaar heeft, PROGRESS genaamd (auteurs: Rousseeuw en Leroy). Klassieke regressie heeft als doelfunctie:

$$\min_\beta \sum_{i=1}^{n} (y_i - x_i \beta)^2$$

Hierbij stelt x_i de i-de rij uit de designmatrix X voor. De doelfunctie van LMS is:

$$\min_\beta med_i (y_i - x_i \beta)^2$$

Dit doel wordt in benadering bereikt door alle p-tallen uit n te nemen en daarvoor steeds de exacte fit te bepalen en de waarde van de doelfunctie te bepalen. Bewaard wordt alleen de aanpassing waarvoor de doelfunctie zijn minimum bereikt. PROGRESS bevat ook nog een snellere variant, maar daarvan kan niet worden gegarandeerd dat dit een high-breakdown methode is en daarom is deze hier niet gebruikt.

De schaalparameter bij LMS wordt als volgt geschat:

$$s^0 = 1.4826[\min_\beta med_i (y_i - x_i \beta)^2]^{1/2}(1 + \frac{5}{n-p})$$

Met deze schaalparameter kan men vervolgens de gestandaardiseerde residuen berekenen. Ter verhoging van de efficientie is vervolgens een Reweighted Least Squares mogelijk: een klassieke aanpassing met gewicht 0 als het gestandaardiseerde residu in absolute waarde groter is dan 2.5 en anders gewicht 1.

Bij de data van Hawkins, Bradu en Kass worden de punten 1 tot met 10 op deze wijze verwijderd, alsmede punt 53. TRADE regressie verwijdert de punten 1 tot en met 12. In beide gevallen worden de tien slechte leverage points als zodanig herkend. LMS gooit er dan nog een ten onrechte weg en TRADE regressie twee. Maar de uitendelijke plaatjes van de residuen tegen de aangepaste waarden zijn niet met het blote oog onderscheidbaar.

Alle tot nu toe gevonden high breakdown methoden zijn zeer traag en voor grote data-sets onbruikbaar. Bovendien zitten ze niet in de gebruikelijke statistische pakketten en het is bepaald geen sinecure om ze er zelf aan toe te voegen. TRADE regressie daarentegen is eenvoudig te coderen en slechts driemaal zo bewerkelijk als gewone regressie.

Helaas is het breakdown punt van TRADE regressie 0. Atkinson (1987) noemt een kunstmatige data-set van Huber waarmee van allerlei robuuste schatters eenvoudig vastgesteld kan worden of hun breakdown punt bij een steekproef van 6 elementen de slechtst denkbare waarde van $\frac{1}{6}$ aanneemt. Dit zijn de data:

Nummer	x	y
1	-4	2.48
2	-3	0.73
3	-2	-0.04
4	-1	-1.44
5	0	-1.32
6	10	0.00

Het wordt aan de lezer over gelaten om vast te stellen wat hier aan de hand is. Wellicht bekruipt hem dan tijdens dit experiment hetzelfde gevoel als waarmee de auteur dezes al

gerume tijd kampt: een hoog breakdown punt is mathematisch interessant, maar voor data uit de praktijk niet zinvol. Meer hierover in de laatste paragraaf.

Karakterisering van het voorbeeld

Het voorbeeld bevat tien slechte hefboom-punten. Deze zijn kunstmatig aangebracht en betreffen de eerste tien pseudo-waarnemingen. Om na te kunnen gaan wat Hawkins, Bradu en Kass nu precies gedaan hebben, is het volgende experiment uitgevoerd. p is een hulp-variabele die de waarde 1 heeft voor observaties 11 tot 75 en de waarde 0 voor de eerste 10. Een dummy-variabele dus die de kunstmatige hefboom-punten identificeert.

Alle correlatie-coefficienten tussen p, x_1, x_2, x_3 en y zijn significant met een overschrijdingskans kleiner dan 0.0001. En een regressie met p als afhankelijke variabele en x_1, x_2 en x_3 als prediktoren geeft een significante x_2 en een zeer significante x_3. De determinatie-coefficient is bij dit model zelfs 0.6178 en dat is meer dan 0.6018 wat bereikt werd met y als response-variabele.

Nadat de hefboom-punten verwijderd zijn, daalt de determinatie-coefficient tot 0.0424. Deze waarde suggereert dat de prediktoren in de opgeschoonde data-set geen enkele voorspellende waarde ten aanzien van y hebben. Dat is in overeenstemming met het feit dat de oorspronkelijke significantie van x_2 en x_3 na verwijdering van de hefboom-punten totaal verdwenen is (zie pagina 19 van de bijlage). Dat er ook nog twee andere punten verwijderd zijn, speelt hier hoegenaamd geen rol.

Het beeld is nu duidelijk. Eerst leek er sprake te zijn van een uitstekende aanpassing van een regressie-model, waarbij men nog de verwijdering van de niet-significante x_1 zou kunnen overwegen. Maar na identificatie van de uitschieters blijkt er iets heel anders met de data-set aan de hand te zijn: de indicator-variabele voorspelt y heel wat beter dan de oorspronkelijke prediktoren. Een aanpassing met alleen p levert een determinatie-coefficient op van 0.9766 en dat is ordes beter dan wat met x_1, x_2 en x_3 bereikt kon worden. Kortom: een robuuste methode geeft soms meer inzicht in de struktuur van data dan een klassieke aanpak.

Evaluatie

Bij de Hawkins, Bradu en Kass data gedroeg TRADE regressie zich zeer bevredigend. Om na te gaan of dit goede gedrag toevallig was, is deze methode op nog een aantal andere data-sets losgelaten.

Tweedejaars studenten Wiskunde aan de TUE worden in het praktikum bij het college Regressie-analyse geconfronteerd met uitschieters en hefboom-punten. TRADE regressie is toegepast op de hierop betrekking hebbende opgaven. Alle kunstmatig opgenomen ellende werd moeiteloos herkend.

Op de jaarvergadering 1989 van de Sectie Statistische Programmatuur van de VVS werd een lastige data-set van Jansen (1988) met acht pakketten en wisselend succes onderzocht. Zie hiervoor Debets, Dijkstra, Eilers, van der Knaap, Otten, Raatgever, van der Sluis, Stemerdink, Verbeek en van Zomeren (1989). TRADE regressie had (evenals LMS) geen enkele moeite met de data. Dit in schrijnende tegenstelling tot klassieke methoden met

bijbehorende diagnostische hulpmiddelen.

Bij het programma PROGRESS wordt een aantal data-sets meegeleverd waarmee de gebruiker zelf overtuigd kan raken van de strukturele deficienties van klassieke Least Squares en de superieure robuustheid van Least Median of Squares. Hiervan zijn er 22 losgelaten op TRADE regressie en LMS. Daartoe behoren de Hawkins, Bradu en Kass data.

Bij beide methoden werden gestandaardiseerde residuen berekend en data verwijderd met een drempelwaarde van 2.5 voor de absolute waarde van deze gestandaardiseerde residuen. De laatste stap was in beide gevallen klassieke Least Squares. Dus als dezelfde data als uitschieters herkend worden, ontstaan dezelfde regresie-vergelijkingen. Van de 21 nog te bespreken gevallen (de Hawkins, Bradu en Kass dataset was immers reeds besproken) waren deze vergelijkingen in 15 gevallen volkomen identiek.

Van de zes overige was het verschil in drie gevallen uiterst klein. Het ging steeds om meerdere uitschieters, waarbij LMS er precies 1 meer vond dan TRADE regressie. En die ene die niet door TRADE gevonden werd, behoorde steeds bij het kleinste gestandaardiseerde residu (in absolute waarde) van LMS. De overige drie gevallen zijn dermate individueel dat ze nu apart besproken worden.

1. LMS heeft een aardige eigenschap: de exact fit property. Als de meerderheid van de data een lineair verband exact volgt, dan zou een robuuste regressie de hierbij behorende vergelijking moeten opleveren. Ter illustratie hiervan hebben Rousseeuw en Leroy een kunstmatige data-set geconstrueerd. Hierover struikelde TRADE regressie op overtuigende wijze. Het lijkt echter niet zo waarschijnlijk dat dit soort datasets in de natuur voorkomen.

2. Rousseeuw en Leroy hebben in een echte data-set vier kunstmatige uitschieters aangebracht. Deze uitschieters waren zo gekozen dat de criteria van Belsley, Kuh en Welsch er geen vat op hadden. Het is niet verwonderlijk dat ook TRADE regressie (die immers met klassieke least squares begint) hier ook niet succesvol in was.

3. Nu het laatste voorbeeld. De voorgaande voorbeelden suggereerden de volgende conclusie: Een reken-intensieve high-breakdown methode is alleen beter dan de snelle TRADE regressie voor kunstmatige data-sets. Bij data uit de natuur gedragen de methoden zich identiek of vrijwel identiek. Het nu te behandelen voorbeeld spreekt deze bewering tegen. Dit betreft het Hertzsprung-Russell diagram van het sterrecluster CYG OB1. De laatste pagina van deze notitie bevat de tweedimensionale dataset met klassieke regressie (de vrijwel horizontale lijn) en Least Median of Squares (door het gros van de data). TRADE is hier niet onderscheidbaar van klassieke regressie en levert pas een aanvaardbaar resultaat op zodra de aanpassing kwadratisch wordt. Dat is dan overigens dezelfde kwadratische aanpassing als die middels Least Median of Squares verkregen kan worden. Na bestudering van het plaatje is het de mening van de auteur dezes dat een lineaire aanpassing hier gewoon niet deugt en dat de data meer behoefte hebben aan een kwadratisch model dan aan een robuuste methode.

Samenvattend kan gesteld worden dat TRADE regressie zich in de praktijk zeer behoorlijk gedraagt. In afwachting van snelle high-breakdown methoden (die misschien wel nooit

zullen bestaan) is het een redelijk alternatief. Bijzonder aantrekkelijk is het gemak waarmee deze methode in de taal van de meeste statistische pakketten kan worden uitgedrukt.

Litteratuur

Siegel, A.F. (1982) Robust regression using repeated medians
Biometrika (69) 242-244

Rousseeuw, P.J. (1984) Least median of squares regression
Journal of the American Statistical Association (79) 871-880

Rousseeuw, P.J. and V.J. Yohai (1984) Robust regression by means of S-estimators
In: Robust and nonlinear time series (J. Franke, W. Hardle and D. Martin eds.) Lecture Notes in Statistics (26) 256-272 Springer, New York

Yohai, V.J. (1987) High breakdown-point and high efficency robust estimates for regression
The Annals of Statistics (Vol.15 No.20) 642-656

Yohai, V.J. and R.H. Zamar (1988) High breakdown-point estimates of regression by means of the minimization of an efficient scale
Journal of the American Statistical Association (Vol.83 No.402) 406-413

Hawkins, D.M. D. Bradu and G.V. Kass (1984) Location of several outliers in multiple regression data using elemental sets
Technometrics (Vol.26 No.3) 197-208

Cook, R.D. (1977) Detection of influential observations in linear regression
Technometrics (19) 15-18

Belsley, D.A. E. Kuh and R.E. Welsch (1980) Regression diagnostics: identifying influential data and sources of collinearity
Wiley, New York

Rousseeuw, P.J. and A.M. Leroy (1987) Robust regression and outlier detection
Wiley, New York

Donoho, D.L. and P.J. Huber (1983) The notion of breakdown point
In: A Festschrift for Erich Lehmann (Edited by P. Bickel K. Doksum and J.L. Hodges Jr.) Wadsworth, Belmont

Atkinson, A.C. (1985) Plots, transformations and regression
Oxford Statistical Science Series 1

Dijkstra, J.B. (1984) Oefeningen regressie-analyse voor tweedejaars studenten wiskunde
Technische Universiteit Eindhoven (Rekencentrum)

Jansen, F.J. (1988) Afstudeerverslag: Robuuste regressie-analyse
Faculteit der Wiskunde en Informatica (TUE)

Debets, P. en 9 andere auteurs (1989) Regressie-diagnostiek in verschillende pakketten
Kwantitatieve Methoden (32) 95-138

_1DUA23:[RCSAS]HAWKINS.LIS;3 **page 1**

hawkins, bradu and kass (1984) 1
9:54 TUESDAY, DECEMBER 19, 1989

OBS	INDEX	X1	X2	X3	Y
1	1	10.1	19.6	28.3	9.7
2	2	9.5	20.5	28.9	10.1
3	3	10.7	20.2	31.0	10.3
4	4	9.9	21.5	31.7	9.5
5	5	10.3	21.1	31.1	10.0
6	6	10.8	20.4	29.2	10.0
7	7	10.5	20.9	29.1	10.8
8	8	9.9	19.6	28.8	10.3
9	9	9.7	20.7	31.0	9.6
10	10	9.3	19.7	30.3	9.9
11	11	11.0	24.0	35.0	-0.2
12	12	12.0	23.0	37.0	-0.4
13	13	12.0	26.0	34.0	0.7
14	14	11.0	34.0	34.0	0.1
15	15	3.4	2.9	2.1	-0.4
16	16	3.1	2.2	0.3	0.6
17	17	0.0	1.6	0.2	-0.2
18	18	2.3	1.6	2.0	0.0
19	19	0.8	2.9	1.6	0.1
20	20	3.1	3.4	2.2	0.4
21	21	2.6	2.2	1.9	0.9
22	22	0.4	3.2	1.9	0.3
23	23	2.0	2.3	0.8	-0.8
24	24	1.3	2.3	0.5	0.7
25	25	1.0	0.0	0.4	-0.3
26	26	0.9	3.3	2.5	-0.8
27	27	3.3	2.5	2.9	-0.7
28	28	1.8	0.8	2.0	0.3
29	29	1.2	0.9	0.8	0.3
30	30	1.2	0.7	3.4	-0.3
31	31	3.1	1.4	1.0	0.0
32	32	0.5	2.4	0.3	-0.4
33	33	1.5	3.1	1.5	-0.6
34	34	0.4	0.0	0.7	-0.7
35	35	3.1	2.4	3.0	0.3
36	36	1.1	2.2	2.7	-1.0
37	37	0.1	3.0	2.6	-0.6
38	38	1.5	1.2	0.2	0.9
39	39	2.1	0.0	1.2	-0.7
40	40	0.5	2.0	1.2	-0.5
41	41	3.4	1.6	2.9	-0.1
42	42	0.3	1.0	2.7	-0.7
43	43	0.1	3.3	0.9	0.6
44	44	1.8	0.5	3.2	-0.7
45	45	1.9	0.1	0.6	-0.5
46	46	1.8	0.5	3.0	-0.4
47	47	3.0	0.1	0.8	-0.9
48	48	3.1	1.6	3.0	0.1
49	49	3.1	2.5	1.9	0.9
50	50	2.1	2.8	2.9	-0.4
51	51	2.3	1.5	0.4	0.7
52	52	3.3	0.6	1.2	-0.5
53	53	0.3	0.4	3.3	0.7
54	54	1.1	3.0	0.3	0.7

hawkins, bradu and kass (1984) 2
9:54 TUESDAY, DECEMBER 19, 1989

OBS	INDEX	X1	X2	X3	Y
55	55	0.5	2.4	0.9	0.0
56	56	1.8	3.2	0.9	0.1
57	57	1.8	0.7	0.7	0.7
58	58	2.4	3.4	1.5	-0.1
59	59	1.6	2.1	3.0	-0.3
60	60	0.3	1.5	3.3	-0.9
61	61	0.4	3.4	3.0	-0.3
62	62	0.9	0.1	0.3	0.6
63	63	1.1	2.7	0.2	-0.3
64	64	2.8	3.0	2.9	-0.5
65	65	2.0	0.7	2.7	0.6
66	66	0.2	1.8	0.8	-0.9
67	67	1.6	2.0	1.2	-0.7
68	68	0.1	0.0	1.1	0.6
69	69	2.0	0.6	0.3	0.2
70	70	1.0	2.2	2.9	0.7
71	71	2.2	2.5	2.3	0.2
72	72	0.6	2.0	1.5	-0.2
73	73	0.3	1.7	2.2	0.4
74	74	0.0	2.2	1.6	-0.9
75	75	0.3	0.4	2.6	0.2

hawkins, bradu and kass (1984) 3
9:54 TUESDAY, DECEMBER 19, 1989

DEP VARIABLE: Y

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	543.30014093	181.10004698	35.768	0.0001
ERROR	71	359.48572574	5.06317924		
C TOTAL	74	902.78586667			

ROOT MSE	2.250151	R-SQUARE	0.6018
DEP MEAN	1.278667	ADJ R-SQ	0.5850
C.V.	175.9764		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	-0.38755	0.41649961	-0.930	0.3553
X1	1	0.23918479	0.26245835	0.911	0.3652
X2	1	-0.334548	0.15505689	-2.158	0.0343
X3	1	0.38334082	0.12882514	2.976	0.0040

OBS	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL
1	9.7000	6.3196	0.5647	3.3804	2.1781
2	10.1000	6.1050	0.5506	3.9950	2.1818
3	10.3000	7.2974	0.6586	3.0026	2.1516
4	9.5000	6.9395	0.6385	2.5605	2.1577
5	10.0000	6.9390	0.6077	3.0610	2.1665
6	10.0000	6.5644	0.6186	3.4356	2.1635
7	10.8000	6.2870	0.5866	4.5130	2.1723
8	10.3000	6.4634	0.5653	3.8366	2.1780
9	9.6000	6.8910	0.6365	2.7090	2.1583
10	9.9000	6.8615	0.6634	3.0385	2.1501
11	-0.2	7.6312	0.6907	-7.8312	2.1415
12	-0.4	8.9717	0.8536	-9.3717	2.0820
13	0.7000	6.8180	0.7433	-6.118	2.1238
14	0.1000	3.9024	1.6894	-3.8024	1.4863
15	-0.4	0.2605	0.5413	-.660504	2.1841
16	0.6000	-.267081	0.6198	0.8671	2.1631
17	-0.2	-.846159	0.4458	0.6462	2.2055
18	0	0.3940	0.3417	-0.39398	2.2241
19	0.1000	-.553047	0.3973	0.6530	2.2148
20	0.4000	.0598083	0.4907	0.3402	2.1960
21	0.9000	0.2267	0.3857	0.6733	2.2168

hawkins, bradu and kass (1984) 4
9:54 TUESDAY, DECEMBER 19, 1989

OBS	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL
22	0.3000	-.634083	0.4813	0.9341	2.1981
23	-0.8	-.371969	0.3849	-.428031	2.2170
24	0.7000	-0.6544	0.3634	1.3544	2.2206
25	-0.3	.0049716	0.3335	-.304972	2.2253
26	-0.8	-.317941	0.4011	-.482059	2.2141
27	-0.7	0.6771	0.4597	-1.3771	2.2027
28	0.3000	0.5420	0.3450	-.242026	2.2235
29	0.3000	-.094949	0.3004	0.3949	2.2300
30	-0.3	0.9686	0.4857	-1.2686	2.1971
31	0	0.2689	0.5466	-.268896	2.1828
32	-0.4	-.955871	0.4295	0.5559	2.2088
33	-0.6	-.490861	0.3654	-.109139	2.2203
34	-0.7	-.023537	0.4025	-.676463	2.2139
35	0.3000	0.7010	0.4163	-.401029	2.2113
36	-1	0.1746	0.3421	-1.1746	2.2240
37	-0.6	-0.37059	0.5452	-0.22941	2.1831
38	0.9000	-.353562	0.3259	1.2536	2.2264
39	-0.7	0.5747	0.4207	-1.2747	2.2105
40	-0.5	-.477045	0.3900	-.022955	2.2161
41	-0.1	1.0021	0.5150	-1.1021	2.1904
42	-0.7	0.3847	0.5298	-1.0847	2.1869
43	0.6000	-1.1226	0.5539	1.7226	2.1809
44	-0.7	1.1024	0.4534	-1.8024	2.2040
45	-0.5	0.2635	0.3830	-.763451	2.2173
46	-0.4	1.0257	0.4372	-1.4257	2.2073
47	-0.9	0.6032	0.5762	-1.5032	2.1751
48	0.1000	0.9687	0.4542	-.868668	2.2038
49	0.9000	0.2459	0.4859	0.6541	2.1971
50	-0.4	0.2897	0.2825	-.689691	2.2323
51	0.7000	-.185911	0.4285	0.8859	2.2090
52	-0.5	0.6610	0.6020	-1.161	2.1681
53	0.7000	0.8154	0.6339	-.115411	2.1590
54	0.7000	-1.0131	0.4521	1.7131	2.2043
55	0	-.725867	0.4136	0.7259	2.2118
56	0.1000	-.682565	0.4344	0.7826	2.2078
57	0.7000	.0771377	0.3391	0.6229	2.2245
58	-0.1	-0.37596	0.4500	0.2760	2.2047
59	-0.3	0.4426	0.3024	-.742617	2.2297
60	-0.9	0.4474	0.5584	-1.3474	2.1798
61	-0.3	-.279318	0.5094	-.020682	2.1917
62	0.6000	-.090736	0.3271	0.6907	2.2263
63	-0.3	-.951059	0.4263	0.6511	2.2094
64	-0.5	0.3902	0.3639	-.890211	2.2205
65	0.6000	0.8917	0.3971	-.291656	2.2148
66	-0.9	-.635227	0.4273	-.264773	2.2092
67	-0.7	-.213942	0.3077	-.486058	2.2290
68	0.6000	.0580438	0.4815	0.5420	2.1980

hawkins, bradu and kass (1984) 5
9:54 TUESDAY, DECEMBER 19, 1989

OBS	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL
69	0.2000	.0050932	0.3820	0.1949	2.2175
70	0.7000	0.2273	0.3683	0.4727	2.2198
71	0.2000	0.1840	0.3095	.0160303	2.2288
72	-0.2	-.338124	0.3787	0.1381	2.2181
73	0.4000	-.041177	0.4645	0.4412	2.2017
74	-0.9	-.510211	0.5030	-.389789	2.1932
75	0.2000	0.5471	0.5606	-.347073	2.1792

OBS	STUDENT RESIDUAL	-2-1-0 1 2	COOK'S D
1	1.5520	|***|	0.040
2	1.8311	|***|	0.053
3	1.3955	|**|	0.046
4	1.1867	|**|	0.031
5	1.4129	|**|	0.039
6	1.5880	|***|	0.052
7	2.0775	|****|	0.079
8	1.7615	|***|	0.052
9	1.2552	|**|	0.034
10	1.4132	|**|	0.048
11	-3.6569	|******|	0.348
12	-4.5013	|******|	0.851
13	-2.8806	| *****|	0.254
14	-2.5582	| *****|	2.114
15	-.302419	| |	0.001
16	0.4009	| |	0.003
17	0.2930	| |	0.001
18	-.177144	| |	0.000
19	0.2949	| |	0.001
20	0.1549	| |	0.000
21	0.3037	| |	0.001
22	0.4250	| |	0.002
23	-.193068	| |	0.000
24	0.6099	|* |	0.002
25	-.137047	| |	0.000
26	-.217722	| |	0.000
27	-.625177	| *|	0.004
28	-.108847	| |	0.000
29	0.1771	| |	0.000
30	-.577418	| *|	0.004
31	-.123191	| |	0.000
32	0.2517	| |	0.001
33	-.049155	| |	0.000
34	-.305557	| |	0.001
35	-.181355	| |	0.000
36	-.528135	| *|	0.002

hawkins, bradu and kass (1984) 6
9:54 TUESDAY, DECEMBER 19, 1989

OBS	STUDENT RESIDUAL	-2-1-0 1 2	COOK'S D
37	-.105084	| | |	0.000
38	0.5630	| |* |	0.002
39	-.576686	| *| |	0.003
40	-.010358	| | |	0.000
41	-.503138	| *| |	0.003
42	-0.49599	| | |	0.004
43	0.7899	| |* |	0.010
44	-.817785	| *| |	0.007
45	-.344314	| | |	0.001
46	-.645925	| *| |	0.004
47	-0.6911	| *| |	0.008
48	-.394162	| | |	0.002
49	0.2977	| | |	0.001
50	-.308953	| | |	0.000
51	0.4011	| | |	0.002
52	-.535504	| *| |	0.006
53	-.053455	| | |	0.000
54	0.7772	| |* |	0.006
55	0.3282	| | |	0.001
56	0.3545	| | |	0.001
57	0.2800	| | |	0.000
58	0.1252	| | |	0.000
59	-0.33305	| | |	0.001
60	-.618147	| *| |	0.006
61	-.009436	| | |	0.000
62	0.3103	| | |	0.001
63	0.2947	| | |	0.001
64	-.400899	| | |	0.001
65	-.131683	| | |	0.000
66	-.119849	| | |	0.000
67	-.218059	| | |	0.000
68	0.2466	| | |	0.001
69	.0878953	| | |	0.000
70	0.2129	| | |	0.000
71	.0071925	| | |	0.000
72	.0622729	| | |	0.000
73	0.2004	| | |	0.000
74	-.177725	| | |	0.000
75	-.159267	| | |	0.000

SUM OF RESIDUALS	6.34215E-15
SUM OF SQUARED RESIDUALS	359.4857
PREDICTED RESID SS (PRESS)	498.5369

hawkins, bradu and kass (1984) 7
9:54 TUESDAY, DECEMBER 19, 1989

OBS	RESIDUAL	RSTUDENT	HAT DIAG H	COV RATIO	DFFITS	INTERCEP DFBETAS	X1 DFBETAS
1	3.3804	1.5678	0.0630	0.9839	0.4065	-0.0756	0.1156
2	3.9950	1.8627	0.0599	0.9277	0.4701	0.0057	-0.0430
3	3.0026	1.4051	0.0857	1.0357	0.4301	-0.0382	0.0800
4	2.5605	1.1902	0.0805	1.0624	0.3522	0.0450	-0.0845
5	3.0610	1.4230	0.0729	1.0186	0.3991	-0.0070	-0.0008
6	3.4356	1.6056	0.0756	0.9907	0.4591	-0.1409	0.2013
7	4.5130	2.1285	0.0680	0.8836	0.5747	-0.1564	0.1891
8	3.8366	1.7886	0.0631	0.9448	0.4642	-0.0338	0.0603
9	2.7090	1.2604	0.0800	1.0517	0.3717	0.0554	-0.0856
10	3.0385	1.4233	0.0869	1.0341	0.4392	0.0975	-0.1256
11	-7.8312	-4.0303	0.0942	0.5071	-1.2999	-0.0522	0.2384
12	-9.3717	-5.2872	0.1439	0.3224	-2.1676	-0.0240	-0.0246
13	-6.118	-3.0437	0.1091	0.7226	-1.0652	0.3666	-0.2568
14	-3.8024	-2.6660	0.5637	1.6475	-3.0302	0.5585	0.3368
15	-.660504	-0.3005	0.0579	1.1176	-0.0745	0.0146	-0.0609
16	0.8671	0.3985	0.0759	1.1349	0.1142	-0.0268	0.0918
17	0.6462	0.2911	0.0393	1.0963	0.0588	0.0507	-0.0396
18	-0.39398	-0.1759	0.0231	1.0814	-0.0270	-0.0086	-0.0141
19	0.6530	0.2930	0.0312	1.0871	0.0525	0.0395	-0.0271
20	0.3402	0.1538	0.0476	1.1097	0.0344	-0.0045	0.0246
21	0.6733	0.3018	0.0294	1.0848	0.0525	0.0048	0.0348
22	0.9341	0.4225	0.0457	1.0979	0.0925	0.0722	-0.0643
23	-.428031	-0.1918	0.0293	1.0880	-0.0333	-0.0057	-0.0150
24	1.3544	0.6072	0.0261	1.0641	0.0994	0.0452	0.0030
25	-.304972	-0.1361	0.0220	1.0810	-0.0204	-0.0167	0.0019
26	-.482059	-0.2163	0.0318	1.0902	-0.0392	-0.0318	0.0241
27	-1.3771	-0.6225	0.0417	1.0803	-0.1299	0.0084	-0.1035
28	-.242026	-0.1081	0.0235	1.0831	-0.0168	-0.0104	-0.0028
29	0.3949	0.1759	0.0178	1.0757	0.0237	0.0189	-0.0014
30	-1.2686	-0.5747	0.0466	1.0894	-0.1270	-0.1052	0.0513
31	-.268896	-0.1223	0.0590	1.1238	-0.0306	0.0045	-0.0258
32	0.5559	0.2500	0.0364	1.0945	0.0486	0.0322	-0.0211
33	-.109139	-0.0488	0.0264	1.0869	-0.0080	-0.0037	0.0004
34	-.676463	-0.3036	0.0320	1.0876	-0.0552	-0.0538	0.0287
35	-.401029	-0.1801	0.0342	1.0939	-0.0339	-0.0010	-0.0251
36	-1.1746	-0.5254	0.0231	1.0665	-0.0808	-0.0791	0.0457
37	-0.22941	-0.1043	0.0587	1.1237	-0.0261	-0.0226	0.0220
38	1.2536	0.5603	0.0210	1.0619	0.0820	0.0363	0.0248
39	-1.2747	-0.5740	0.0350	1.0763	-0.1092	-0.0358	-0.0525
40	-.022955	-0.0103	0.0300	1.0912	-0.0018	-0.0016	0.0011
41	-1.1021	-0.5005	0.0524	1.1010	-0.1177	0.0049	-0.0924
42	-1.0847	-0.4933	0.0554	1.1050	-0.1195	-0.1132	0.0875
43	1.7226	0.7878	0.0606	1.0876	0.2001	0.1309	-0.1246
44	-1.8024	-0.8159	0.0406	1.0622	-0.1678	-0.1093	0.0082
45	-.763451	-0.3422	0.0290	1.0827	-0.0591	-0.0204	-0.0293
46	-1.4257	-0.6433	0.0377	1.0743	-0.1274	-0.0825	0.0024

hawkins, bradu and kass (1984) 8
9:54 TUESDAY, DECEMBER 19, 1989

OBS	RESIDUAL	RSTUDENT	HAT DIAG H	COV RATIO	DFFITS	INTERCEP DFBETAS	X1 DFBETAS
47	-1.5032	-0.6885	0.0656	1.1025	-0.1824	0.0126	-0.1472
48	-.868668	-0.3918	0.0407	1.0937	-0.0807	-0.0062	-0.0566
49	0.6541	0.2958	0.0466	1.1046	0.0654	-0.0076	0.0516
50	-.689691	-0.3070	0.0158	1.0696	-0.0388	-0.0224	-0.0065
51	0.8859	0.3987	0.0363	1.0883	0.0773	0.0025	0.0525
52	-1.161	-0.5328	0.0716	1.1217	-0.1479	0.0196	-0.1252
53	-.115411	-0.0531	0.0794	1.1494	-0.0156	-0.0136	0.0104
54	1.7131	0.7750	0.0404	1.0658	0.1589	0.0587	-0.0130
55	0.7259	0.3261	0.0338	1.0888	0.0610	0.0473	-0.0337
56	0.7826	0.3523	0.0373	1.0916	0.0693	0.0129	0.0164
57	0.6229	0.2782	0.0227	1.0782	0.0424	0.0173	0.0186
58	0.2760	0.1243	0.0400	1.1015	0.0254	0.0006	0.0122
59	-.742617	-0.3310	0.0181	1.0711	-0.0449	-0.0404	0.0121
60	-1.3474	-0.6154	0.0616	1.1038	-0.1577	-0.1481	0.1231
61	-.020682	-0.0094	0.0512	1.1155	-0.0022	-0.0019	0.0018
62	0.6907	0.3083	0.0211	1.0754	0.0453	0.0390	-0.0071
63	0.6511	0.2928	0.0359	1.0924	0.0565	0.0224	-0.0034
64	-.890211	-0.3985	0.0261	1.0770	-0.0653	-0.0072	-0.0410
65	-.291656	-0.1308	0.0311	1.0913	-0.0234	-0.0133	-0.0037
66	-.264773	-0.1190	0.0361	1.0971	-0.0230	-0.0205	0.0157
67	-.486058	-0.2166	0.0187	1.0757	-0.0299	-0.0161	-0.0044
68	0.5420	0.2449	0.0458	1.1054	0.0536	0.0520	-0.0352
69	0.1949	0.0873	0.0288	1.0893	0.0150	0.0033	0.0091
70	0.4727	0.2115	0.0268	1.0847	0.0351	0.0346	-0.0222
71	.0160303	0.0071	0.0189	1.0788	0.0010	0.0004	0.0004
72	0.1381	0.0618	0.0283	1.0890	0.0106	0.0098	-0.0066
73	0.4412	0.1990	0.0426	1.1030	0.0420	0.0409	-0.0327
74	-.389789	-0.1765	0.0500	1.1121	-0.0405	-0.0364	0.0326
75	-.347073	-0.1582	0.0621	1.1268	-0.0407	-0.0372	0.0271

OBS	X2 DFBETAS	X3 DFBETAS
1	-0.0616	0.0389
2	0.0063	0.0923
3	-0.1799	0.1598
4	-0.0696	0.1610
5	-0.0900	0.1353
6	-0.0496	-0.0189
7	0.0267	-0.0548
8	-0.1076	0.1205
9	-0.1027	0.1905
10	-0.1669	0.2734
11	0.1739	-0.4905
12	1.1918	-1.2615

hawkins, bradu and kass (1984) 9
9:54 TUESDAY, DECEMBER 19, 1989

OBS	X2 DFBETAS	X3 DFBETAS
13	-0.4235	0.3588
14	-2.7946	1.9199
15	-0.0124	0.0501
16	0.0309	-0.0865
17	0.0202	0.0039
18	0.0061	0.0051
19	0.0261	-0.0079
20	0.0115	-0.0258
21	0.0026	-0.0262
22	0.0428	-0.0001
23	-0.0136	0.0226
24	0.0555	-0.0548
25	0.0084	-0.0065
26	-0.0163	0.0006
27	0.0141	0.0562
28	0.0093	-0.0050
29	-0.0033	0.0016
30	0.0869	-0.0993
31	0.0015	0.0159
32	0.0290	-0.0141
33	-0.0049	0.0043
34	0.0211	-0.0315
35	0.0058	0.0119
36	0.0065	-0.0291
37	-0.0061	-0.0075
38	0.0146	-0.0340
39	0.0626	-0.0137
40	-0.0005	-0.0002
41	0.0438	0.0239
42	0.0442	-0.0859
43	0.1245	-0.0358
44	0.1320	-0.1084
45	0.0249	0.0013
46	0.0981	-0.0778
47	0.0651	0.0438
48	0.0347	0.0088
49	0.0073	-0.0403
50	-0.0032	0.0089
51	0.0121	-0.0467
52	0.0429	0.0465
53	0.0086	-0.0130
54	0.1182	-0.0992
55	0.0290	-0.0070
56	0.0465	-0.0523
57	-0.0095	-0.0067
58	0.0141	-0.0203

hawkins, bradu and kass (1984) 10
9:54 TUESDAY, DECEMBER 19, 1989

OBS	X2 DFBETAS	X3 DFBETAS
59	0.0125	-0.0153
60	0.0504	-0.1124
61	-0.0006	-0.0005
62	-0.0151	0.0131
63	0.0397	-0.0343
64	-0.0070	0.0337
65	0.0165	-0.0104
66	-0.0065	-0.0029
67	-0.0078	0.0113
68	-0.0208	0.0360
69	-0.0022	-0.0048
70	-0.0043	0.0156
71	0.0001	-0.0004
72	0.0020	0.0018
73	-0.0034	0.0210
74	-0.0086	-0.0111
75	0.0203	-0.0319

hawkins, bradu and kass (1984) 11
9:54 TUESDAY, DECEMBER 19, 1989

UNIVARIATE

VARIABLE=RES1 RESIDUALS

MOMENTS

N	75	SUM WGTS	75
MEAN	8.456E-17	SUM	6.342E-15
STD DEV	2.20407	VARIANCE	4.85792
SKEWNESS	-1.5381	KURTOSIS	5.97878
USS	359.486	CSS	359.486
CV	99999	STD MEAN	0.254504
T:MEAN=0	3.323E-16	PROB>|T|	1
SGN RANK	58	PROB>|S|	0.761408
NUM ^= 0	75		
D:NORMAL	0.180946	PROB>D	<.01

QUANTILES(DEF=4)

100% MAX	4.51295	99%	4.51295
75% Q3	0.725867	95%	3.51578
50% MED	-0.0229549	90%	3.01696
25% Q1	-0.742617	10%	-1.3965
0% MIN	-9.3717	5%	-4.2655
		1%	-9.3717
RANGE	13.8846		
Q3-Q1	1.46848		
MODE	-9.3717		

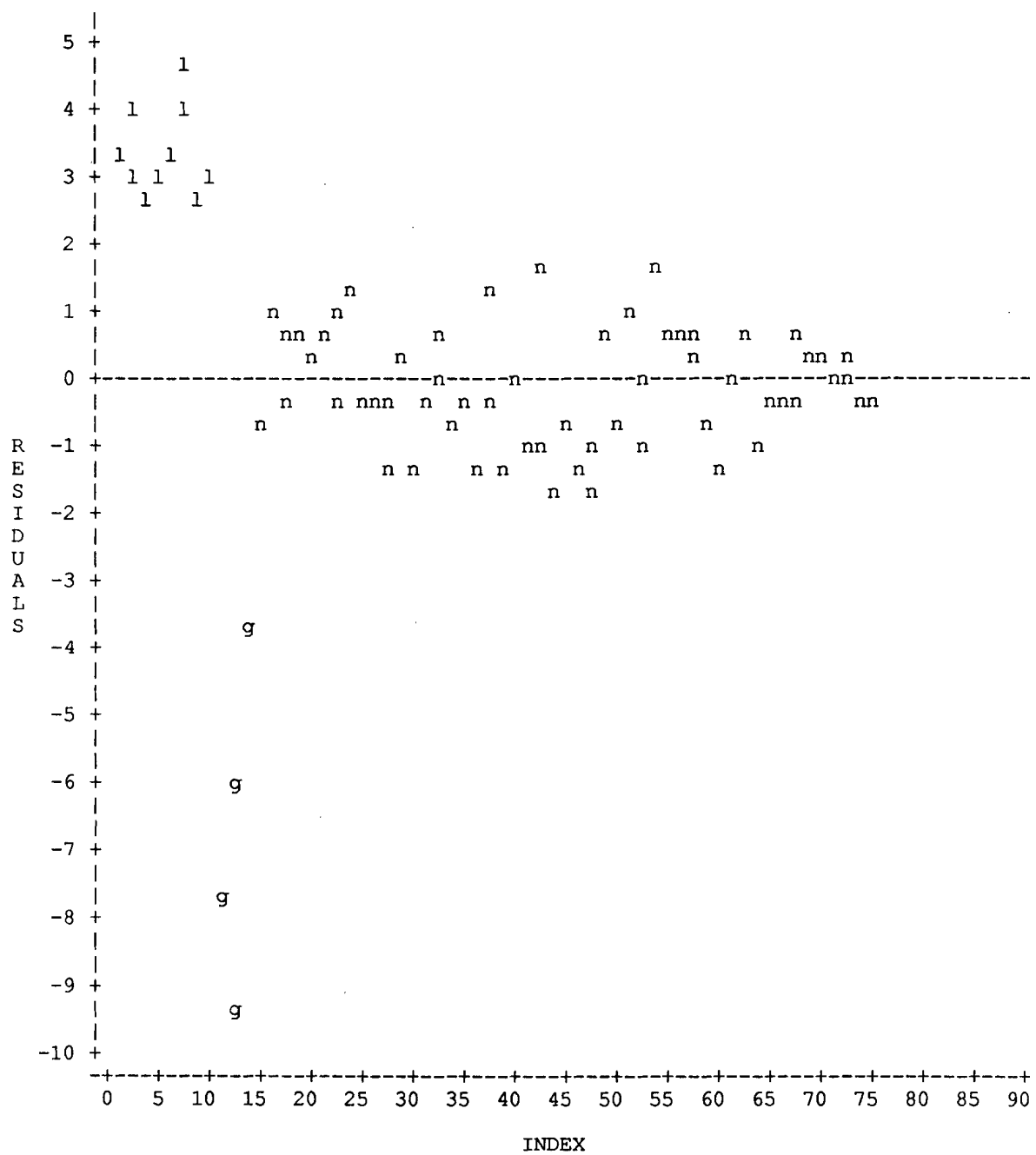
EXTREMES

LOWEST	HIGHEST
-9.3717	3.38039
-7.8312	3.43559
-6.118	3.83655
-3.8024	3.99499
-1.8024	4.51295

_1DUA23:[RCSAS]HAWKINS.LIS;3 **page 12**

hawkins, bradu and kass (1984) 12
9:54 TUESDAY, DECEMBER 19, 1989

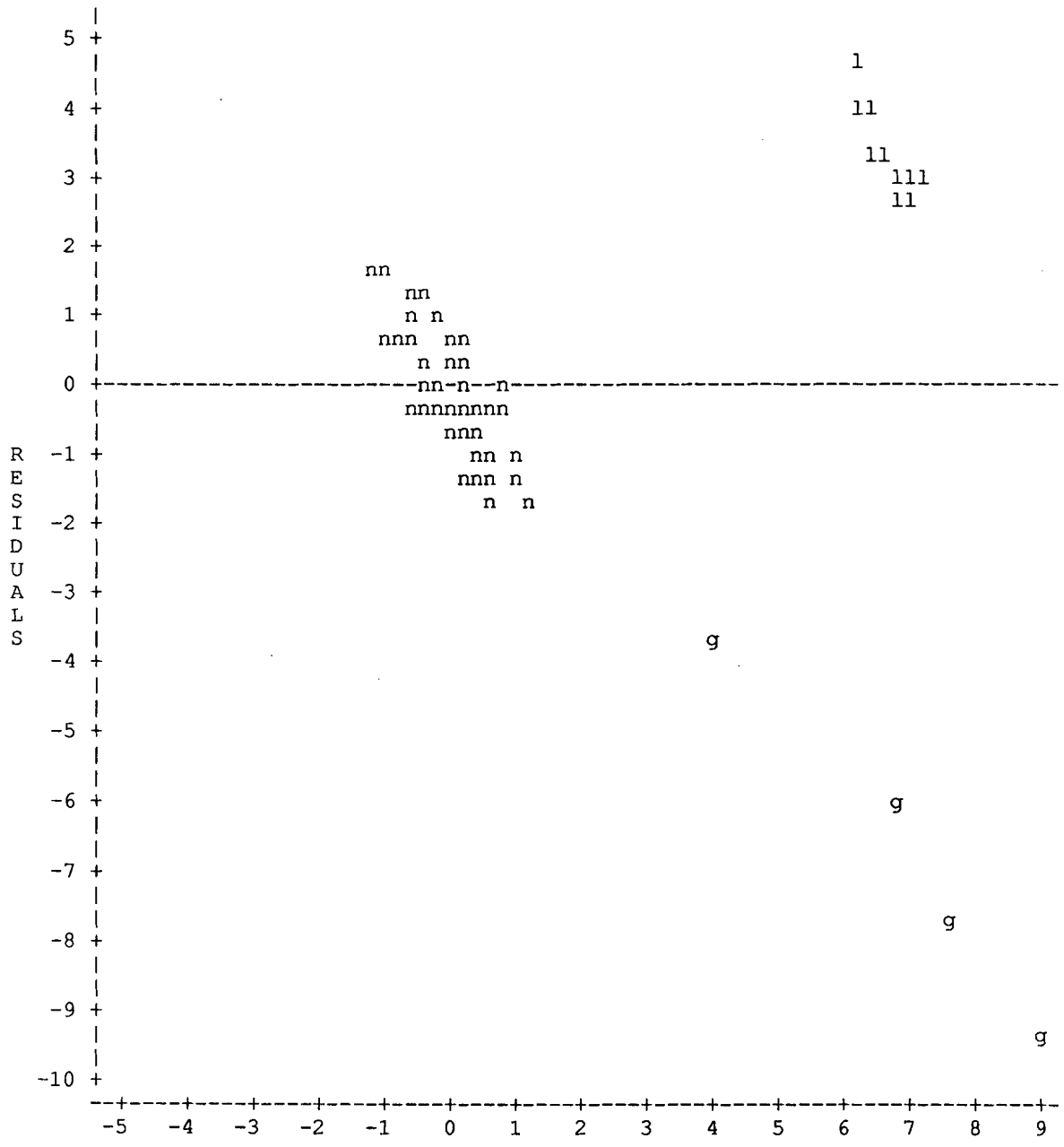
PLOT OF RES1*INDEX SYMBOL IS VALUE OF S



NOTE: 1 OBS HIDDEN

hawkins, bradu and kass (1984) 13
9:54 TUESDAY, DECEMBER 19, 1989

PLOT OF RES1*FIT1 SYMBOL IS VALUE OF S



NOTE: 23 OBS HIDDEN

hawkins, bradu and kass (1984) 14
9:54 TUESDAY, DECEMBER 19, 1989

UNIVARIATE

VARIABLE=ARES1

MOMENTS

N	75	SUM WGTS	75
MEAN	1.37959	SUM	103.469
STD DEV	1.71141	VARIANCE	2.92894
SKEWNESS	2.64883	KURTOSIS	8.21202
USS	359.486	CSS	216.741
CV	124.053	STD MEAN	0.197617
T:MEAN=0	6.98111	PROB>|T|	0.0001
SGN RANK	1425	PROB>|S|	0.0001
NUM ^= 0	75		

QUANTILES(DEF=4)

100% MAX	9.37166	99%	9.37166
75% Q3	1.42573	95%	4.83396
50% MED	0.725867	90%	3.58232
25% Q1	0.394949	10%	0.215608
0% MIN	0.0160303	5%	0.0919018
		1%	0.0160303
RANGE	9.35563		
Q3-Q1	1.03078		
MODE	0.0160303		

EXTREMES

LOWEST	HIGHEST
0.0160303	3.99499
0.020682	4.51295
0.0229549	6.118
0.109139	7.83125
0.115411	9.37166

hawkins, bradu and kass (1984) 15
9:54 TUESDAY, DECEMBER 19, 1989

DEP VARIABLE: Y

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	3.81927046	1.27309015	6.427	0.0014
ERROR	34	6.73441375	0.19807099		
C TOTAL	37	10.55368421			

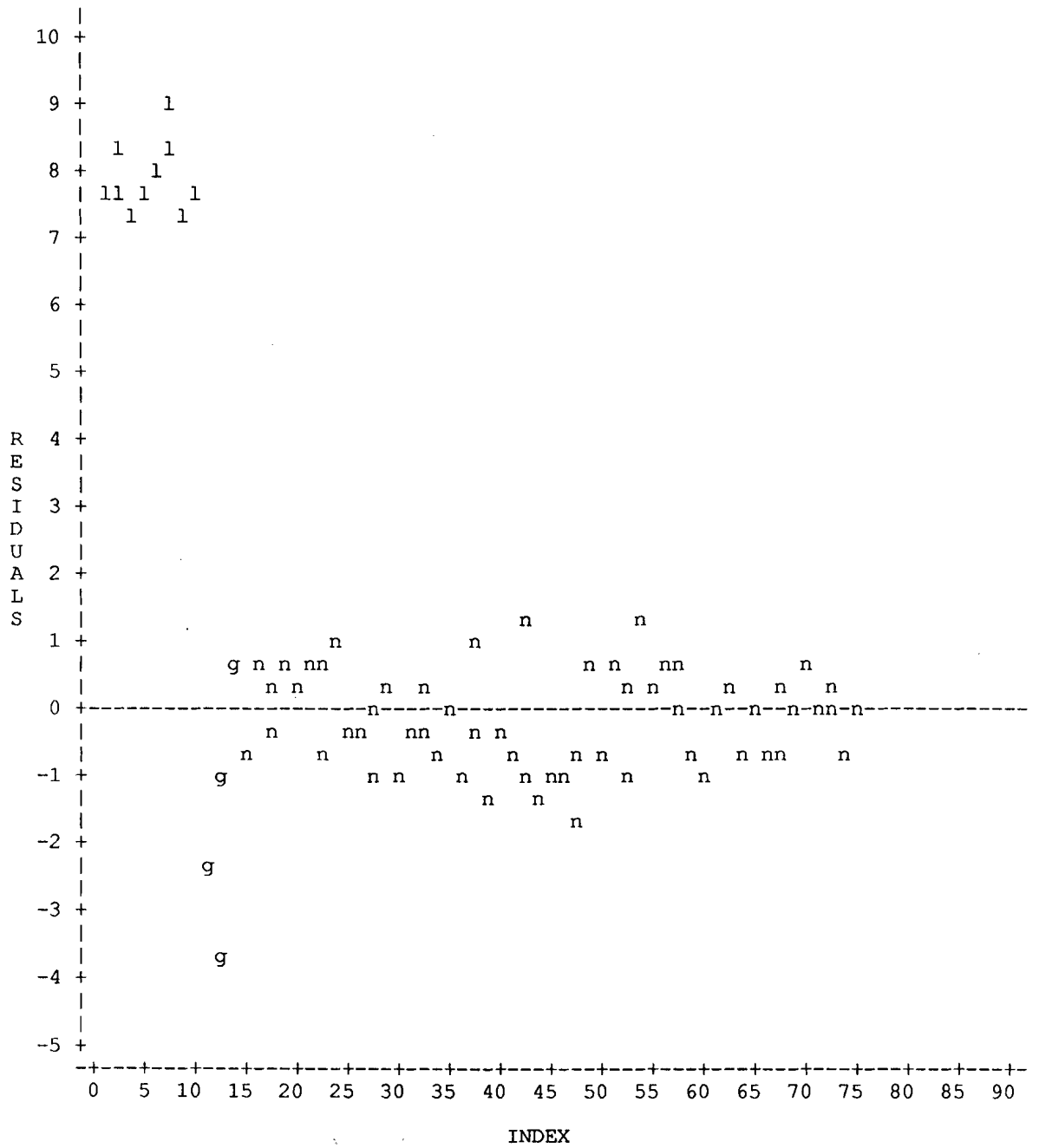
ROOT MSE	0.4450517	R-SQUARE	0.3619
DEP MEAN	-0.0263158	ADJ R-SQ	0.3056
C.V.	-1691.2		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	-0.0915959	0.18091906	-0.506	0.6159
X1	1	0.18900109	0.07305906	2.587	0.0141
X2	1	-0.26599	0.07360079	-3.614	0.0010
X3	1	0.19125724	0.0839313	2.279	0.0291

hawkins, bradu and kass (1984) 16
9:54 TUESDAY, DECEMBER 19, 1989

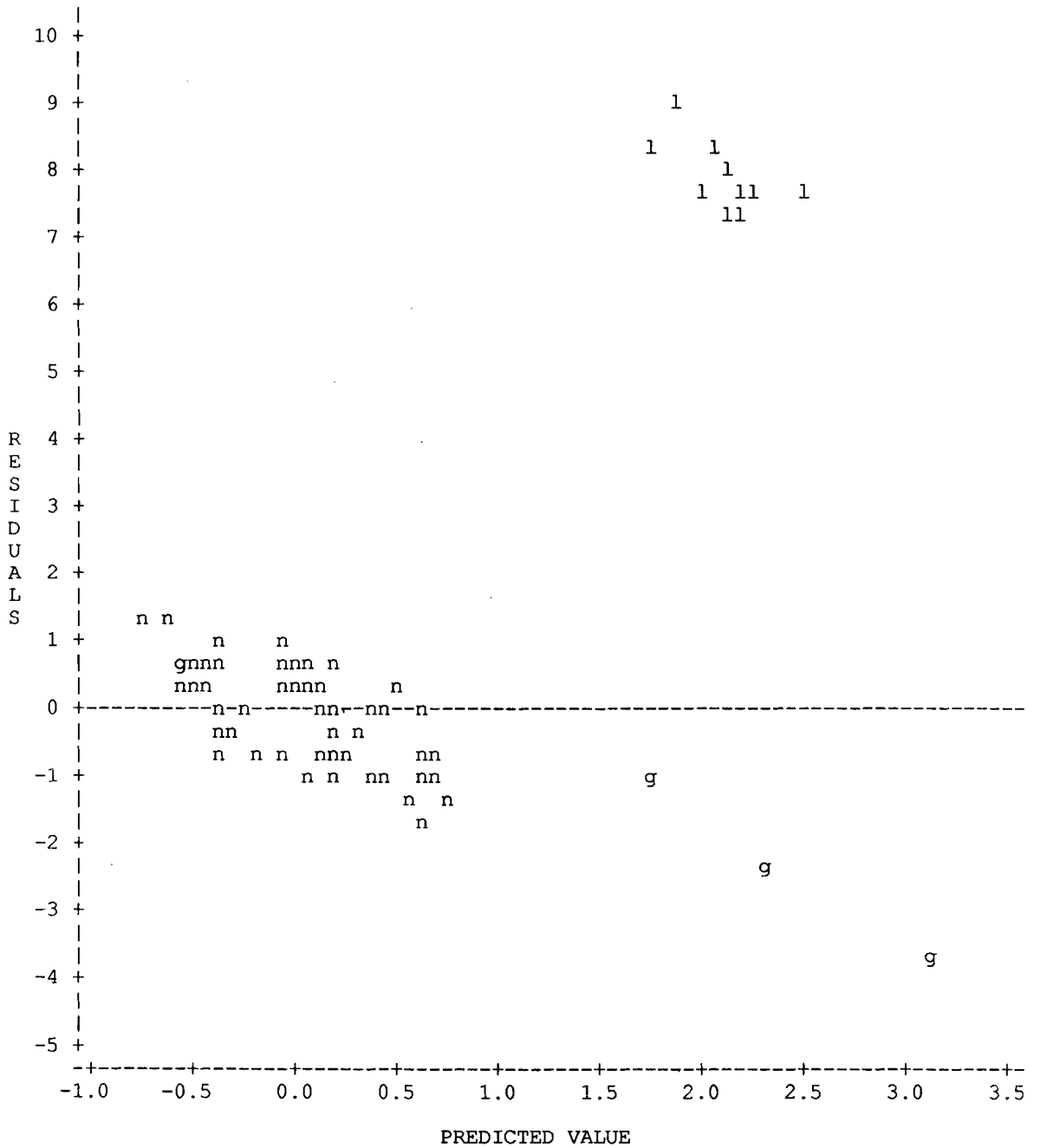
PLOT OF RES2*INDEX SYMBOL IS VALUE OF S



NOTE: 1 OBS HIDDEN

hawkins, bradu and kass (1984) 17
9:54 TUESDAY, DECEMBER 19, 1989

PLOT OF RES2*FIT2 SYMBOL IS VALUE OF S



NOTE: 14 OBS HIDDEN

hawkins, bradu and kass (1984) 18
9:54 TUESDAY, DECEMBER 19, 1989

UNIVARIATE

VARIABLE=ARES2

MOMENTS

N	75	SUM WGTS	75
MEAN	1.63811	SUM	122.858
STD DEV	2.53964	VARIANCE	6.44979
SKEWNESS	2.0491	KURTOSIS	2.55263
USS	678.539	CSS	477.285
CV	155.035	STD MEAN	0.293253
T:MEAN=0	5.58599	PROB>|T|	0.0001
SGN RANK	1425	PROB>|S|	0.0001
NUM ^= 0	75		

QUANTILES(DEF=4)

100% MAX	8.9007	99%	8.9007
75% Q3	1.10202	95%	7.95864
50% MED	0.633128	90%	7.68069
25% Q1	0.313178	10%	0.145166
0% MIN	0.0158108	5%	0.0419312
		1%	0.0158108
RANGE	8.88489		
Q3-Q1	0.788844		
MODE	0.0158108		

EXTREMES

LOWEST	HIGHEST
0.0158108	7.81332
0.0166075	7.89188
0.0232903	8.22569
0.0465914	8.32156
0.100878	8.9007

hawkins, bradu and kass (1984) 19
9:55 TUESDAY, DECEMBER 19, 1989

DEP VARIABLE: Y

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	0.83401801	0.27800600	0.872	0.4632
ERROR	59	18.81868040	0.31896068		
C TOTAL	62	19.65269841			

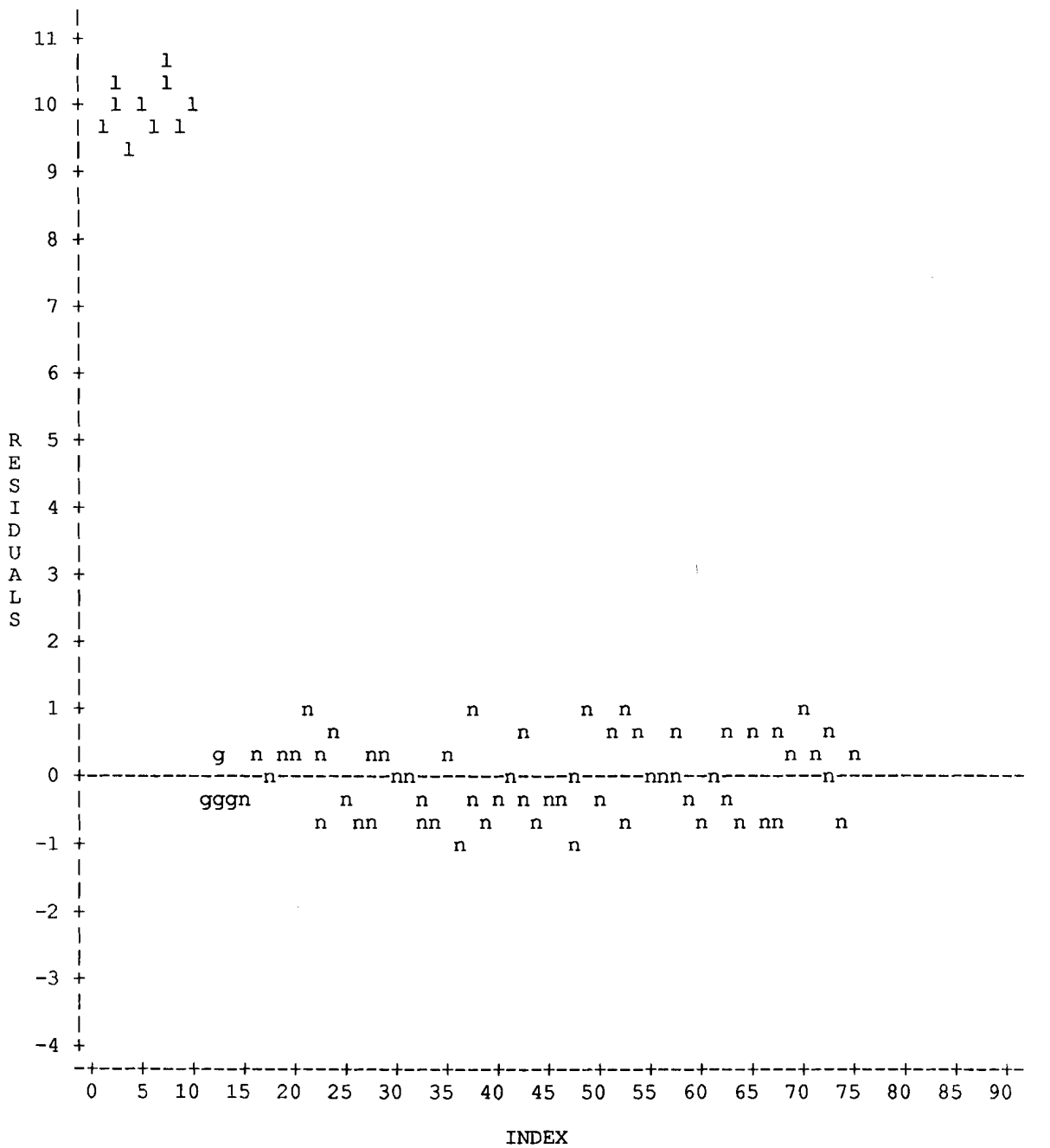
ROOT MSE	0.564766	R-SQUARE	0.0424
DEP MEAN	-0.0587302	ADJ R-SQ	-0.0063
C.V.	-961.629		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	-0.182467	0.10591323	-1.723	0.0902
X1	1	0.08189092	0.06764193	1.211	0.2309
X2	1	0.0236391	0.04887718	0.484	0.6304
X3	1	-0.0336746	0.04659436	-0.723	0.4727

hawkins, bradu and kass (1984) 20
9:55 TUESDAY, DECEMBER 19, 1989

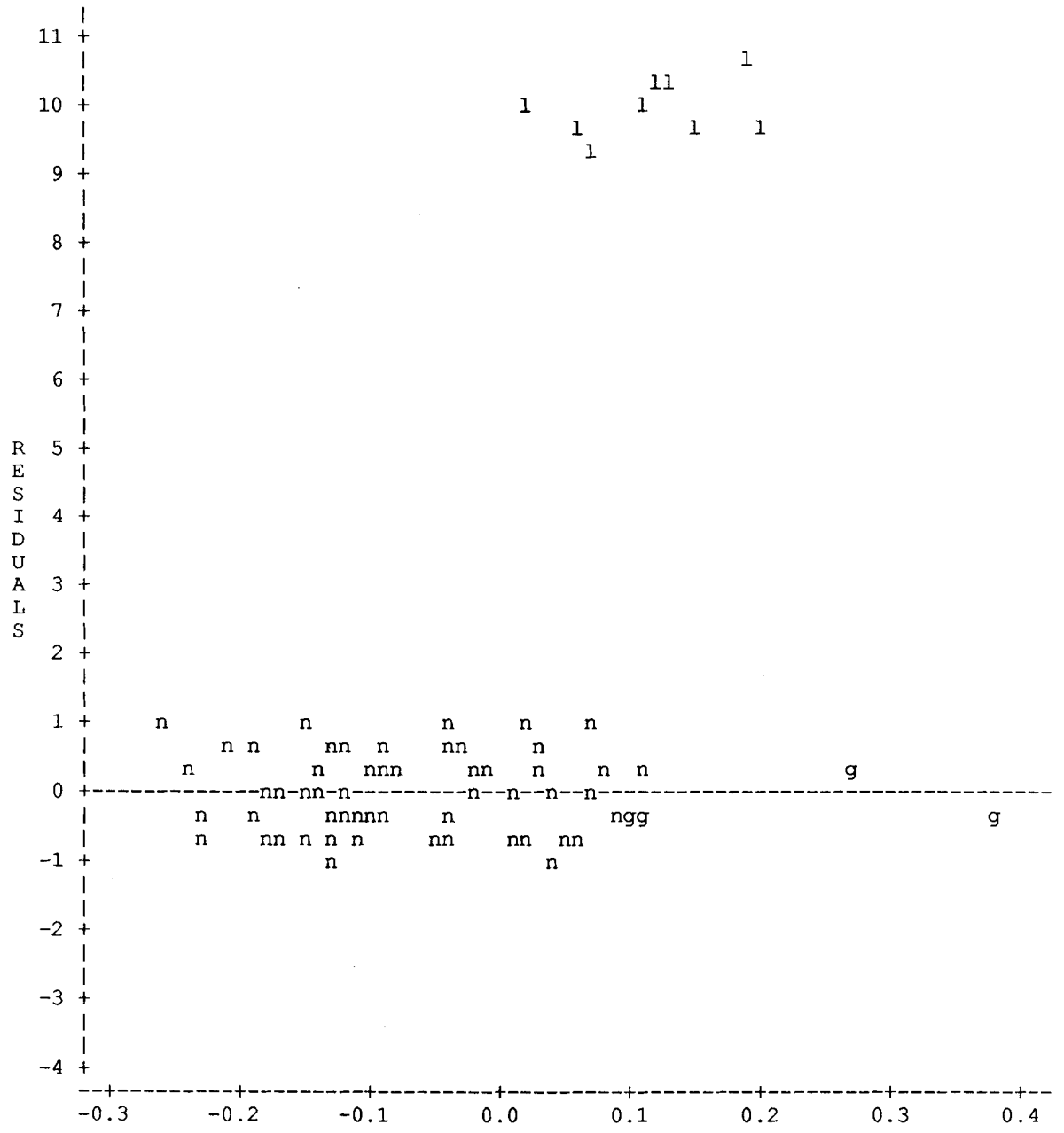
PLOT OF RES3*INDEX SYMBOL IS VALUE OF S



NOTE: 1 OBS HIDDEN

hawkins, bradu and kass (1984) 21
9:55 TUESDAY, DECEMBER 19, 1989

PLOT OF RES3*FIT3 SYMBOL IS VALUE OF S



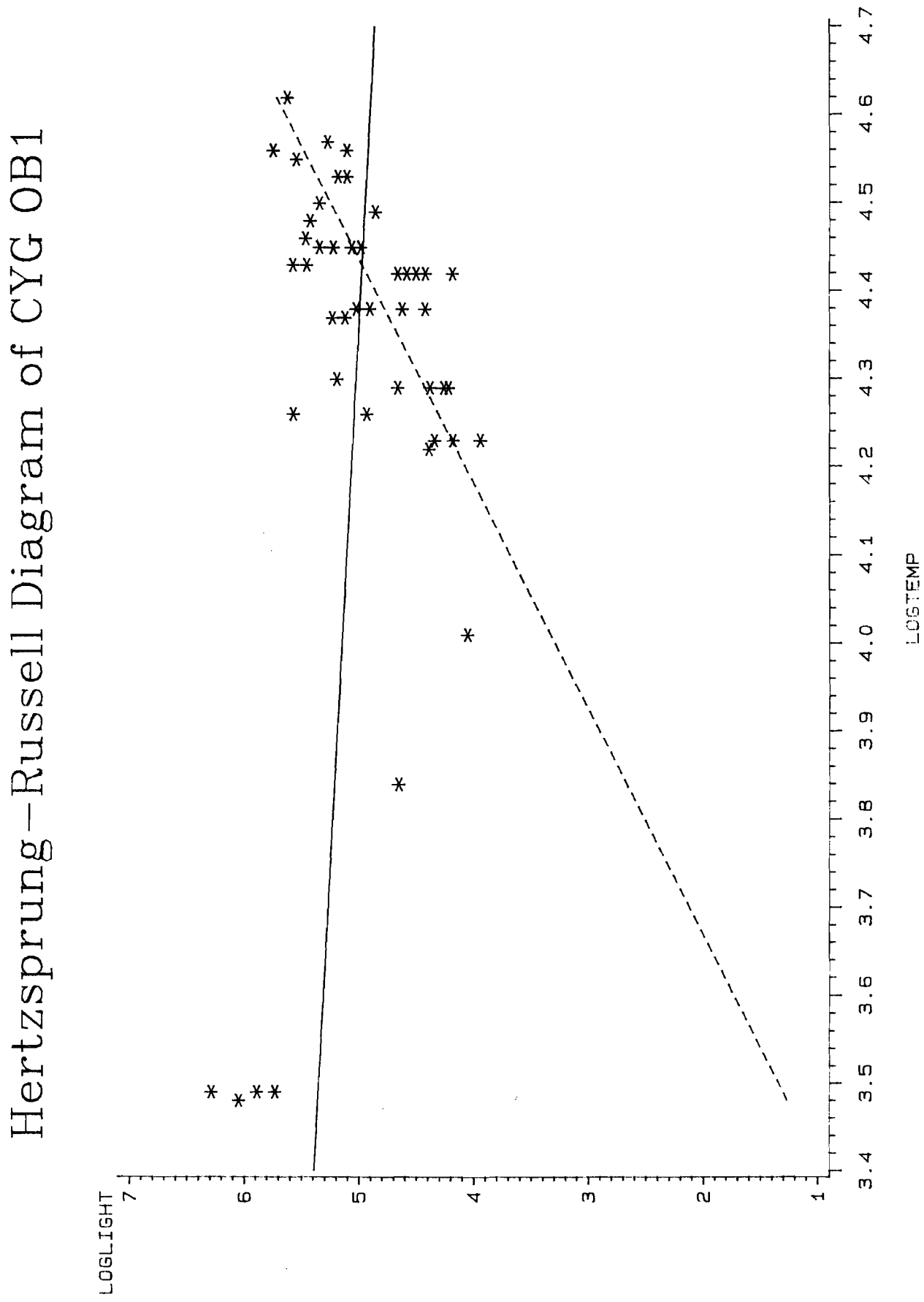
NOTE: 7 OBS HIDDEN

hawkins, bradu and kass (1984) 22
9:55 TUESDAY, DECEMBER 19, 1989

OBS	RES1	RES2	RES3	FIT1	FIT2	FIT3	S	W1	W2	D
1	3.3804	7.6835	9.5450	6.3196	2.01648	0.15497	l	0	0	8.1855
2	3.9950	8.3216	9.9931	6.1050	1.77844	0.10690	l	0	0	8.8652
3	3.0026	7.8133	10.1726	7.2974	2.48668	0.12736	l	0	0	8.3238
4	2.5605	7.3764	9.4310	6.9395	2.12357	0.06901	l	0	0	7.8583
5	3.0610	7.8092	9.8875	6.9390	2.19082	0.11251	l	0	0	8.3194
6	3.4356	7.8919	9.7991	6.5644	2.10812	0.20089	l	0	0	8.4075
7	4.5130	8.9007	10.6085	6.2870	1.89930	0.19151	l	0	0	9.4822
8	3.8366	8.2257	10.1782	6.4634	2.07431	0.12175	l	0	0	8.7631
9	2.7090	7.4353	9.5427	6.8910	2.16469	0.05729	l	0	0	7.9211
10	3.0385	7.6788	9.8755	6.8615	2.22120	0.02447	l	0	0	8.1805
11	-7.8312	-2.4976	-0.3071	7.6312	2.29765	0.10706	g	0	0	-2.6608
12	-9.3717	-3.5352	-0.4980	8.9717	3.13515	0.09796	g	0	0	-3.7661
13	-6.1180	-1.0634	0.4301	6.8180	1.76341	0.26990	g	0	1	-1.1329
14	-3.8024	0.6535	-0.2771	3.9024	-0.55351	0.37713	g	0	1	0.6962
15	-0.6605	-0.5813	-0.4938	0.2605	0.18128	0.09380	n	1	1	-0.6193
16	0.8671	0.6335	0.4867	-0.2671	-0.03349	0.11330	n	0	1	0.6749
17	0.6462	0.2789	-0.0486	-0.8462	-0.47893	-0.15138	n	1	1	0.2972
18	-0.3940	-0.3000	0.0236	0.3940	0.30004	-0.02364	n	1	1	-0.3196
19	0.6530	0.5058	0.2023	-0.5530	-0.40576	-0.10228	n	1	1	0.5388
20	0.3402	0.3893	0.3223	0.0598	0.01071	0.07768	n	1	1	0.4147
21	0.6733	0.7220	0.8815	0.2267	0.17802	0.01847	n	1	1	0.7692
22	0.9341	0.8038	0.4380	-0.6341	-0.50378	-0.13805	n	0	1	0.8563
23	-0.4280	-0.6276	-0.8087	-0.3720	-0.17237	0.00875	n	1	1	-0.6686
24	1.3544	1.0620	0.7385	-0.6544	-0.36204	-0.03848	n	0	1	1.1314
25	-0.3050	-0.4739	-0.1860	0.0050	0.17391	-0.11405	n	1	1	-0.5049
26	-0.4821	-0.4789	-0.6851	-0.3179	-0.32112	-0.11494	n	1	1	-0.5102
27	-1.3771	-1.1218	-0.7492	0.6771	0.42178	0.04921	n	0	1	-1.1951
28	-0.2420	-0.1183	0.3835	0.5420	0.41833	-0.08350	n	1	1	-0.1261
29	0.3949	0.2512	0.3899	-0.0949	0.04882	-0.08986	n	1	1	0.2676
30	-1.2686	-0.8993	-0.1179	0.9686	0.59929	-0.18214	n	0	1	-0.9580
31	-0.2689	-0.3132	-0.0708	0.2689	0.31318	0.07081	n	1	1	-0.3336
32	0.5559	0.1781	-0.3051	-0.9559	-0.57810	-0.09489	n	1	1	0.1897
33	-0.1091	-0.2542	-0.5631	-0.4909	-0.34578	-0.03686	n	1	1	-0.2708
34	-0.6765	-0.8179	-0.5267	-0.0235	0.11788	-0.17328	n	1	1	-0.8713
35	-0.4010	-0.1297	0.2729	0.7010	0.42970	0.02710	n	1	1	-0.1382
36	-1.1746	-1.0475	-0.8687	0.1746	0.04752	-0.13130	n	0	1	-1.1160
37	-0.2294	-0.2266	-0.4091	-0.3706	-0.37340	-0.19091	n	1	1	-0.2414
38	1.2536	0.9890	0.9380	-0.3536	-0.08903	-0.03800	n	0	1	1.0536
39	-1.2747	-1.2348	-0.6491	0.5747	0.53482	-0.05091	n	0	1	-1.3155
40	-0.0230	-0.2004	-0.3653	-0.4770	-0.29957	-0.13465	n	1	1	-0.2135
41	-1.1021	-0.7801	-0.1361	1.0021	0.68007	0.03613	n	0	1	-0.8310
42	-1.0847	-0.9155	-0.4748	0.3847	0.21551	-0.22518	n	0	1	-0.9753
43	1.7226	1.3783	0.7266	-1.1226	-0.77833	-0.12658	n	0	1	1.4684
44	-1.8024	-1.4276	-0.5690	1.1024	0.72763	-0.13100	n	0	1	-1.5209
45	-0.7635	-0.8557	-0.4553	0.2635	0.35566	-0.04472	n	0	1	-0.9116
46	-1.4257	-1.0894	-0.2757	1.0257	0.68938	-0.12427	n	0	1	-1.1606
47	-1.5032	-1.5018	-0.9386	0.6032	0.60181	0.03863	n	0	1	-1.5999
48	-0.8687	-0.5425	0.0918	0.9687	0.64249	0.00819	n	0	1	-0.5779
49	0.6541	0.7073	0.8335	0.2459	0.19272	0.06651	n	1	1	0.7535
50	-0.6897	-0.5152	-0.3580	0.2897	0.11518	-0.04196	n	1	1	-0.5488
51	0.8859	0.6794	0.6721	-0.1859	0.02062	0.02787	n	0	1	0.7238
52	-1.1610	-1.1020	-0.5615	0.6610	0.60202	0.06155	n	0	1	-1.1740
53	-0.1154	0.2101	0.9596	0.8154	0.48986	-0.25957	n	1	1	0.2239
54	1.7131	1.3243	0.7316	-1.0131	-0.62429	-0.03157	n	0	1	1.4108

hawkins, bradu and kass (1984) 23
9:55 TUESDAY, DECEMBER 19, 1989

OBS	RES1	RES2	RES3	FIT1	FIT2	FIT3	S	W1	W2	D
55	0.7259	0.4633	0.11509	-0.72587	-0.46334	-0.11509	n	1	1	0.4936
56	0.7826	0.5304	0.08973	-0.68257	-0.43043	0.01027	n	0	1	0.5651
57	0.6229	0.5037	0.74209	0.07714	0.19629	-0.04209	n	1	1	0.5366
58	0.2760	0.1555	-0.14393	-0.37596	-0.25548	0.04393	n	1	1	0.1656
59	-0.7426	-0.5260	-0.19718	0.44262	0.22600	-0.10282	n	0	1	-0.5604
60	-1.3474	-1.0973	-0.66643	0.44741	0.19727	-0.23357	n	0	1	-1.1690
61	-0.0207	0.0466	-0.12964	-0.27932	-0.34659	-0.17036	n	1	1	0.0496
62	0.6907	0.4907	0.71650	-0.09074	0.10928	-0.11650	n	1	1	0.5228
63	0.6511	0.2636	-0.26470	-0.95106	-0.56362	-0.03530	n	1	1	0.2808
64	-0.8902	-0.6943	-0.52009	0.39021	0.19428	0.02009	n	0	1	-0.7396
65	-0.2917	-0.0166	0.69306	0.89166	0.61661	-0.09306	n	1	1	-0.0177
66	-0.2648	-0.5204	-0.74952	-0.63523	-0.37957	-0.15048	n	1	1	-0.5544
67	-0.4861	-0.6083	-0.65543	-0.21394	-0.09167	-0.04457	n	1	1	-0.6481
68	0.5420	0.4623	0.81132	0.05804	0.13769	-0.21132	n	1	1	0.4925
69	0.1949	0.0158	0.21460	0.00509	0.18419	-0.01460	n	1	1	0.0168
70	0.4727	0.6331	0.84623	0.22732	0.06687	-0.14623	n	1	1	0.6745
71	0.0160	0.1009	0.22066	0.18397	0.09912	-0.02066	n	1	1	0.1075
72	0.1381	0.0233	-0.06343	-0.33812	-0.22329	-0.13657	n	1	1	0.0248
73	0.4412	0.4663	0.59180	-0.04118	-0.06631	-0.19180	n	1	1	0.4968
74	-0.3898	-0.5292	-0.71566	-0.51021	-0.37076	-0.18434	n	1	1	-0.5638
75	-0.3471	-0.1560	0.43600	0.54707	0.35598	-0.23600	n	1	1	-0.1662


Hertzsprung-Russell Diagram of CYG OB1
LOGLIGHT
7
6
5
4
3
2
1
3.4
3.5
3.6
3.7
3.8
3.9
4.0
4.1
4.2
4.3
4.4
4.5
4.6
4.7
LOGTEMP