

High-breakdown regressie en SAS

Citation for published version (APA):

Dijkstra, J. B. (1989). *High-breakdown regressie en SAS*. (Computing centre note; Vol. 45). Technische Universiteit Eindhoven.

Document status and date:

Gepubliceerd: 01/01/1989

Document Version:

Uitgevers PDF, ook bekend als Version of Record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Eindhoven University of Technology
Computing Centre Note 45

High-breakdown Regressie en SAS

Jan B. Dijkstra

Samengesteld voor de bijeenkomst van de Benelux Users Group SAS (BUGS)
op 31 oktober en 1 november 1989 in Scheveningen.

augustus 1989.

High-breakdown Regressie en SAS

Jan B. Dijkstra

Samenvatting

De regressie-mogelijkheden van SAS zijn zeer uitgebreid en met name de INFLUENCE-optie bij REG sluit vrijwel uit dat één geïsoleerde uitschieter onopgemerkt blijft. Helaas kunnen twee of meer uitschieters (met name in de prediktor-ruimte) de diagnostische hulpmiddelen van SAS behoorlijk voor de gek houden.

De laatste tien jaar is veel gepubliceerd over high-breakdown methoden. Hiermee aangepaste modellen volgen het gros van de data en identificeren de uitschieters. Helaas bevat SAS zulke technieken niet. Met NLIN kan men robuuste methoden coderen die resistent zijn tegen uitschieters in de y-richting, maar bijzondere punten in de prediktor-ruimte blijven een probleem.

Er zal een quick-and-dirty methode voor robuuste regressie gepresenteerd worden die niet dezelfde goede eigenschappen heeft als moderne high-breakdown methoden, maar die zich wel gemakkelijk in SAS laat coderen. Zelfs tegen uitschieters in de x-richting blijkt deze techniek redelijk bestand.

1. Inleiding

De hier te behandelen regressie-technieken zullen worden geïllustreerd aan de hand van een echte data-set (zie pag. 2). Deze is ontleend aan Rousseeuw en Leroy (1987) die zich op hun beurt weer hebben gebaseerd op Weisberg (1980) en Jerison (1973). De gegevens betreffen het lichaamsgewicht x (in kilogrammen) en het gewicht van de hersenen y (in grammen) van 28 dieren. Omdat de dieren variëren van een muis (0.023 kg) tot een uitgestorven brachiosaurus (87000 kg) ligt een logaritmische transformatie voor beide variabelen voor de hand. In de bijlage (pag. 11) zien we dat de waarnemingen hierdoor in dezelfde orde van grootte komen te liggen. Een enkelvoudige regressie die $\log(y)$ uit $\log(x)$ moet voorspellen levert de volgende aanpassing op:

$$\log(\hat{y}) = 2.554 + 0.4959 \log(x)$$

De determinatiecoëfficiënt is 0.6076 en dat is bemoedigend. Zowel de intercept als $\log(x)$ zijn significant. De t-waarden zijn hier respectievelijk 6.184 en 6.345 met 26 vrijheidsgraden. De procedure REG is aangeroepen met de opties R en INFLUENCE (zie de SAS-Statistics Manual uit 1987). Op de uitvoer van INFLUENCE zal in de derde paragraaf verder worden ingegaan. Met R verkrijgt men de gestudentiseerde residuen. Dit zijn de

Index	Soort	Lichaam	Hersenen
1	Bever	1.35	8.1
2	Koe	465	423
3	Wolf	36.33	119.5
4	Geit	27.66	115
5	Marmot	1.04	5.5
6	Diplodocus	11700	50
7	Olifant (Azie)	2547	4603
8	Ezel	187.1	419
9	Paard	521	655
10	Aap (Potar)	10	115
11	Kat	3.3	25.6
12	Giraffe	529	680
13	Gorilla	207	406
14	Mens	62	1320
15	Olifant (Afrika)	6654	5712
16	Triceratops	9400	70
17	Aap (Rhesus)	6.8	179
18	Kangaroo	35	56
19	Hamster	0.12	1
20	Muis	0.023	0.4
21	Konijn	2.5	12.1
22	Schaap	55.5	175
23	Jaguar	100	157
24	Chimpansee	52.16	440
25	Brachiosaurus	87000	154.5
26	Rat	0.28	1.9
27	Mol	0.122	3
28	Varken	192	180

residuen gedeeld door de individueel geschatte standaarddeviaties. Hiermee is niet alleen rekening gehouden met de stochastiek van de response-variabele, maar ook met de variatie van de aangepaste waarden. Gestudentiseerde residuen die in absolute waarde groter zijn dan 2.5 suggereren de aanwezigheid van uitschieters. In dit voorbeeld treedt dat niet op.

De optie R drukt tevens de Cook (1977) statistic af. Een waarde groter dan 1 voor een zeker punt betekent dat weglating van dit punt de vector van regressie-coëfficiënten buiten het 50% simultane betrouwbaarheidsgebied zou plaatsen dat hoort bij de complete data-set. Dit soort punten hebben dus grote invloed op het uiteindelijke aangepaste model. Daarom worden ze hefboom-punten genoemd: ze trekken immers krachtig aan het aangepaste hypervlak (in dit tweedimensionale voorbeeld een lijn). De Cook statistic neemt in dit voorbeeld nergens waarden groter dan 1 aan (zie pag. 13). Er is dus geen grond om de aanwezigheid van hefboom-punten te vrezen.

2. Least Median of Squares

Hoewel de gestudentiseerde residuen en de Cook-statistics dat niet suggereren, is er met deze data-set wel degelijk iets bijzonders aan de hand. Een plotje van $\log(\hat{y})$ tegen $\log(x)$ toont een drietal uitschieters die bij grote waarden van $\log(x)$ vlak bij elkaar duidelijk onder de denkbeeldige lijn door de overige punten liggen. Dit zijn de drie dinosaurussen (6, 16 en 25) die voor hun lichaamsgewicht verrassend kleine hersenen hebben. In dit tweedimensionale voorbeeld geeft een plaatje duidelijk aan wat er aan de hand is, maar bij meervoudige regressie gaat dat niet meer op. Daar ontstaat de behoefte aan robuuste methoden om de uitschieters op te sporen.

Hampel (1971) introduceerde het breakdown-punt voor schatters. Dit is de maximale fractie uitschieters die een steekproef kan hebben voordat de schatter het oneindige ingetrokken wordt. De klassieke least-squares schatter voor regressie heeft breakdown-punt 0. Immers: een enkele waarneming y die men loodrecht op de prediktor-ruimte verplaatst, blijft het aangepaste hypervlak zonder beperkingen met zich meeslepen. Het hoogst haalbare breakdown-punt is 0.5. Bij een hogere fractie zijn de goede waarnemingen en de uitschieters namelijk niet meer onderscheidbaar. Least Median of Squares is een regressiemethode van Rousseeuw (1984) die dit begerenswaardige breakdown-punt heeft.

Least Median of Squares minimaliseert niet de kwadratensom van de residuen, maar de mediaan van de kwadraten van de residuen. Dit resulteert in een weinig efficiënte maar zeer robuuste schatter. Een terreinwagen onder de regressie-methoden, zoals de originator het zelf eens uitdrukte. Least Median of Squares is helaas niet in SAS opgenomen, maar er is wel een public-domain PC-programma voor met de naam PROGRESS. De manual van Rousseeuw en Leroy (1987) behorende bij dit kosteloze programma is een duur boek.

Least Median of Squares is zeer reken-intensief. Daarom bevat PROGRESS een variant die onder de vooronderstelling van 50% uitschieters met een kans van 95% een aanpassing oplevert waarop geen enkele uitschieter vat heeft gehad. Bij minder dan 50% uitschieters nadert deze kans zeer snel naar 1. Dit proces is geïnspireerd op ideeën die Stahel (1981) voor een andere methode ontwikkeld had.

De efficiëntie van Least Median of Squares kan als volgt verhoogd worden. Bepaal de residuen r_i . Laat n het aantal waarnemingen zijn en p het aantal aangepaste parameters. Dan is s^0 zoals hieronder gedefinieerd een robuuste schatter voor de standaarddeviatie σ :

$$s^0 = 1.4826 \sqrt{\min_{\beta} \text{med}_i r_i^2 \left(1 + \frac{5}{n-p}\right)}$$

β is de vector van regressie-coëfficiënten. De factor 1.4826 maakt deze schatter bij benadering consistent voor normaal verdeelde fouten. De term $1 + \frac{5}{n-p}$ bevat een door simulatie gevonden correctie voor kleine steekproeven. Deze is specifiek voor Least Median of Squares nodig en dient bij andere regressie-methoden te worden weggelaten.

Vervolgens worden de gestandaardiseerde residuen $d_i = \frac{r_i}{s^0}$ berekend. Dan wordt een klassieke gewogen regressie uitgevoerd met gewicht 1 voor alle waarnemingen behalve die waarvoor $\text{abs}(d_i) > 2.5$. Deze kennelijke uitschieters krijgen gewicht 0. Bij normaal

verdeelde fouten impliceert dit dat ongeveer 2% van de waarnemingen wordt genegeerd. En daar valt mee te leven. In dit voorbeeld levert Least Median of Squares de volgende aanpassing op:

$$\log(\hat{y}) = 1.858 + 0.7380 \log(x)$$

De drie dinosaurussen krijgen nu extreme gestandaardiseerde residuen: Diplodocus (-9.00), Triceratops (-8.08) en Brachiosaurus (-9.66). Er zijn nog twee andere gestandaardiseerde residuen die de grens van 2.5 aan de positieve zijde overschrijden: Mens (4.23) en Rhesusaap (3.55). Totaal 5 observaties krijgen dus het gewicht 0 mee en de overige 23 het gewicht 1. Een klassieke regressie resulteert met deze gewichten in de volgende aanpassing:

$$\log(\hat{y}) = 2.001 + 0.7509 \log(x)$$

De t-waarden voor intercept en $\log(x)$ zijn gestegen tot 23.60 en 14.06 (was 6.184 en 6.345) met 21 vrijheidsgraden. Ook de determinatiecoëfficiënt is er beter op geworden. Deze steeg van 0.6076 tot 0.9637.

Least Median of Squares is niet de enige high-breakdown methode. Siegel (1982) introduceerde de Repeated Median. Rousseeuw en Yohai (1984) suggereerden de S-schatters en Yohai (1987) liet deze opvolgen door de efficiënte MM-schatters. Bovenstaand voorbeeld geeft duidelijk het nut van dit soort methoden aan. Het is dan ook jammer dat SAS hierin niets te bieden heeft.

In de volgende paragraaf zal het voorbeeld worden onderzocht met de REG-optie INFLUENCE. Daarna zal een weinig ambitieuze maar wel effectieve regressie-methode worden behandeld die (1) redelijk bestand is tegen uitschieters in alle richtingen en (2) tamelijk eenvoudig in SAS kan worden uitgedrukt.

3. De REG-optie INFLUENCE

INFLUENCE levert de volgende grootheden: RSTUDENT, HATDIAG, COVRATIO, DFFITS en voor iedere parameter (met inbegrip van de intercept) DFBETAS. Deze zullen nu elk afzonderlijk besproken worden (zie pag. 14).

RSTUDENT is een variant op het gestudentiseerde residu. Voor iedere waarneming wordt bij de berekening van de aangepaste waarde uitgegaan van een regressie door de overige punten. De invloed van dit punt komt daardoor duidelijker tot zijn recht. De dinosaurussen nummer 6 en 25 leveren waarden van respectievelijk -2.5074 en -2.5055. Dit overschrijdt in absolute waarde net de grens van 2.5. SAS staat toe dat de waarde van RSTUDENT in een vector wordt opgeslagen en daarom is het eenvoudig om een herwogen least-squares aanpassing te berekenen waarbij deze twee dinosaurussen het gewicht 0 krijgen en de overige subjecten het gewicht 1. Dit resulteert in een aanpassing met een determinatiecoëfficiënt van 0.7973 (dus een stuk minder dan de in de vorige paragraaf bereikte 0.9637). Opmerkelijk is wel dat de derde dinosaurus nu zeer invloedrijk blijkt te zijn: Voor nummer 16 wordt een waarde van -6.0454 bereikt.

Vanwege dit grote gestudentiseerde (deleted) residu zou men nu het proces kunnen herhalen zonder punt 16. Maar deze strategie maakt het aantal stappen onvoorspelbaar. Een betere (maar nog minder praktische) strategie zou de als volgt omschreven

succesievelijke eliminatie kunnen zijn zijn: (1) Bepaal het maximale gestudentiseerde (deleted) residu. (2) Indien dit in absolute waarde groter is dan 2.5 verwijder dan de bijbehorende observatie en pas het model opnieuw aan. (3) Herhaal (1) en (2) totdat de maximale waarde 2.5 niet meer overschrijdt. Dit soort strategieën zijn aardig voor routinebibliotheken als NAG en IMSL. Maar voor SAS zijn ze niet geschikt. Aan het einde van deze paragraaf wordt hierop teruggekomen.

Belsley, Kuh en Welsch (1980) bevelen voor de diagonaal van de hat-matrix de drempelwaarde van $\frac{2p}{n}$ aan. Deze grens wordt gepasseerd door de punten 20 en 25. Hiervan kan slechts gezegd worden dat ze in de prediktor-ruimte nogal excentrisch liggen. Weglating heeft dan ook weinig effect: de determinatiecoëfficiënt wordt 0.6254.

COVRATIO meet de verandering van de determinant van de covariantie-matrix van de schatters door de *i*-de observatie weg te laten. Als de waarde meer dan $\frac{3p}{n}$ van 1 afwijkt moet worden ingegrepen volgens het bovengenoemde trio. Dat gebeurt voor de nummers 6 en 27. Weglating heeft weinig effect: de determinatiecoëfficiënt wordt 0.6609.

DFFITS is een gevoeliger variant op de Cook-statistic. De aanbevolen drempel is hier $2\sqrt{p/n}$ en deze wordt overschreden door de nummers 6, 16, 20 en 25. Merkwaardig genoeg betreft dit de drie dinosaurussen en de muis. Weglating doet de determinatiecoëfficiënt stijgen tot 0.9036 en dat is in vergelijking met de 0.9637 uit de vorige paragraaf nog steeds wat teleurstellend.

DFBETAS meet de geschaalde verandering in de geschatte regressie-coëfficiënten door punt *i* weg te laten. De drempel is hier $\frac{2}{\sqrt{n}}$ en die wordt gepasseerd door de nummers 6, 16, 19, 20 en 25. Dus naast de drie dinosaurussen nu de muis en de hamster. Weglating van de kleine knaagdieren werkt hier kennelijk averechts, want de determinatiecoëfficiënt neemt ten opzichte van DFFITS nog af tot 0.8868.

Samenvattend kan gesteld worden dat met de informatie die INFLUENCE geeft zinvol kan worden verder gewerkt. Maar toch met duidelijk minder goede resultaten dan met een high-breakdown regressie-methode bereikt kunnen worden. Dit verschijnsel laat zich goed verklaren. De drie dinosaurussen liggen dicht bij elkaar. Weglating van steeds de *i*-de observatie (waarop de methoden van INFLUENCE berusten) heeft dus niet al te veel effect. Voor elke dinosaurus blijven er steeds twee over. Bij de dinosaurussen horen grote waarden van $\log(x)$ en bij de hamster en de muis juist kleine waarden. Samen trekken de dinosaurussen de lijn aan de rechterzijde naar beneden. Links wipt hij dus omhoog en dat geeft misleidende resultaten voor de kleine knaagdieren. De diagnostische hulpmiddelen van Belsley, Kuh en Welsch zijn uitstekend voor geïsoleerde uitschieters of hefboom-punten. Maar tegen verstrengeling zijn ze (in tegenstelling tot high-breakdown methoden) niet bestand.

In de hierboven besproken toepassingen werden de vermoedelijke uitschieters steeds gelijktijdig verwijderd. Bij succesievelijke eliminatie wordt per stap alleen de meest verdachte observatie verwijderd. Hierbij kan voor de gestudentiseerde (deleted) residuen als stopcriterium 2.5 gebruikt worden en voor de overige maten de suggesties van Belsley, Kuh

en Welsch. Hoewel dit in SAS zeer bewerkelijk is, is het hier besproken voorbeeld ook op deze wijze doorgerekend. Hierbij werd na 8 eliminaties gestopt omdat de auteur dezes er genoeg van begon te krijgen. Van de 5 criteria waren er 3 toen nog niet klaar met het eliminatiesproces.

RSTUDENT verwijdt achtereenvolgens: Diplodocus, Brachiosaurus, Triceratops, Mens, Rhesus aap, Chimpansee en Potar aap. Hierna wordt het stopcriterium 2.5 bereikt. Kortom: nog niet zo'n slecht resultaat.

HATDIAG verwijdt achtereenvolgens: Brachiosaurus, Muis, Mol, Hamster, Rat, Diplodocus, Triceratops en Afrikaanse olifant. Het stopcriterium is dan nog niet bereikt. Dit criterium houdt alleen rekening met de predictor-ruimte. Afwisselend worden zeer grote of zeer kleine dieren verwijderd, ongeacht de vraag of ze goed bij het model passen.

COVRATIO verwijdt Mol en Bever en stopt vervolgens. Een niet erg zinvol resultaat.

Veel beter begint DFFITS. Achtereenvolgens verdwijnen hier: Brachiosaurus, Diplodocus, Triceratops, Mens, Rhesus aap, Mol, Chimpansee en Potar aap. Het proces is dan nog niet voltooid, maar zou blijken de resultaten dat eigenlijk wel moeten zijn. Een stop na de Rhesus aap zou duidelijk beter zijn. Misschien valt er nog wat aan het criterium van Belsey, Kuh en Welsch te verbeteren.

De resultaten van DFBETAS zijn vrijwel identiek: Brachiosaurus, Diplodocus, Triceratops, Rhesus aap, Mol, Mens, Potar aap en Chimpansee. Ook hier was het stopcriterium nog niet bereikt.

4. Een quick-and-dirty methode

Pagina 15 van de bijlage behoort nog steeds bij de klassieke aanpassing op de volledige data-set. De residuen zijn hier uitgezet tegen de voorspelde waarden. Duidelijk is te zien hoe rechts-onder drie punten de as door de overige punten behoorlijk laten hellen. Theoretisch zouden de residuen en de aangepaste waarden onafhankelijk van elkaar moeten zijn onder de gebruikelijke regressie-vooronderstellingen. Hier is duidelijk niet aan voldaan.

Nu volgt een tamelijk grove en theoretisch nogal bedenkelijke stap. SAS laat toe dat de residuen door REG in een data-set worden geschreven. Hieraan wordt de naam REG1 toegekend (zie de log-file na het voorbeeld; FIT1 bevat de aangepaste waarden). In een data-step wordt de absolute waarde van REG1 toegekend aan AREG1. Met UNIVARIATE wordt hiervan de mediaan berekend (zie pag. 16). Nu krijgt een observatie gewicht 0 als AREG1 deze mediaan overschrijdt en anders de waarde 1. Dus alleen de best passende helft van de observaties doet in een volgende stap mee.

De determinatiecoëfficiënt krijgt bij een op bovenstaande wijze herwogen least-squares aanpassing de geflatteerde waarde van 0.9646 (niet verwonderlijk omdat de restvariantie structureel wordt onderschat). Op pagina 18 staat het hierbij behorende plaatje van de residuen RES2 tegen de aangepaste waarden FIT2. De as van de puntenwolk is lang niet meer zo scheef als in het vorige plaatje. En de residuen bij de dinosaurussen worden veel markanter.

Voor het berekenen van de schaalparameter kiezen we de volgende uitschieter-resistente methode:

$$s^1 = 1.4826 \text{med}_i \text{abs}(r_i)$$

waarbij r_i het i -de residu voorstelt. De mediaan van de absolute waarden van de residuen bij deze stap blijkt 0.4971 te zijn (berekend middels UNIVARIATE). De schaalparameter s^1 wordt dus 0.7370. De residuen in RES2 worden nu hierdoor gedeeld. Dat geeft gestandaardiseerde residuen d_i . Vervolgens wordt een tweede gewogen regressie aangeroepen met gewicht 0 als de absolute waarde van d_i de grens 2.5 overschrijdt en anders gewicht 1. De dinosaurussen 6, 16 en 25 hebben gestandaardiseerde residuen van respectievelijk -6.0268, -5.3759 en -6.2783. Ook de mens overschrijdt de grens; hier is het gestandaardiseerde residu 3.0688 (positief - een bemoedigend resultaat).

Deze regressie resulteert in een determinatiecoëfficiënt van 0.9440. Dit doet nauwelijks onder voor het resultaat van Least Median of Squares. Ook het plotje van de residuen RES3 tegen de aangepaste waarden FIT3 ziet er prima uit (zie pag. 21). De as van de puntenwolk ligt volkomen horizontaal en de dinosaurussen zijn nog verder dan in de vorige plaatjes van de rest verwijderd.

5. Evaluatie

De quick-and-dirty methode van de vorige paragraaf heeft geen overtuigend theoretisch fundament. Niettemin blijkt het uitstekend te werken bij het hier behandelde voorbeeld. 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 Technische Universiteit Eindhoven worden in een praktikum geconfronteerd met uitschieters en hefboom-punten (zie Dijkstra 1984). Bij de opgaven blijkt de quick-and-dirty methode voortreffelijk te werken. Niettemin zal de auteur dezes die methode angstvallig buiten het wiskunde-onderwijs houden.

Op de jaarvergadering 1989 van de Sectie Statistische Programmatuur van de Vereniging voor Statistiek (VVS) zijn 9 voordrachten gehouden over een lastige data-set van Jansen (1988). De tekst van deze voordrachten zal binnenkort in *Kwantitatieve Methoden* verschijnen. Zie hiervoor Debets, Dijkstra, Eilers, van der Knaap, Otten, Raatgever, van der Sluis, Stemerdink, Verbeek en van Zomeren (1989). Ook met dit lastige voorbeeld had de quick-and-dirty methode geen enkele moeite.

Bij PROGRESS van Rousseeuw en Leroy (1987) wordt een aantal data-sets meegeleverd die geselecteerd of geconstrueerd zijn om de noodzaak van high-beakdown methoden te illustreren. Hiervan zijn 9 gevallen voor enkelvoudige regressie geprobeerd en 11 voor meervoudige regressie. Steeds zijn Least Median of Squares (LMS) en de quick-and-dirty (QD) methode toegepast, in beide gevallen afgesloten met herwogen least-squares. Hierbij werden uitschieters met gestandaardiseerde residuen groter dan 2.5 (in absolute waarde) buiten de aanpassing gehouden.

In 15 van de 20 gevallen kwamen de resultaten van LMS en QD precies met elkaar overeen. In 3 gevallen trad het effect op dat reeds zichtbaar was bij het voorbeeld van de hersenen. Het Rhesus aapje werd door LMS wel als uitschieter herkend en niet door QD. Wanneer dit effect voorkwam betrof het steeds een enkel punt waarbij het residu van LMS het kleinste was dat de grens van 2.5 passeerde.

Voor dit effect zijn verscheidene verklaringen denkbaar. In sommige gevallen onderschat LMS de restvariantie (Dijkstra, 1987). Misschien is dat hier gebeurd. Een andere mogelijke verklaring is dat bij de berekening van s^1 de residuen gemakshalve ten opzichte van 0 genomen zijn in plaats van ten opzichte van hun mediaan. De correcte aanpak zou zijn:

$$s^1 = 1.4826 \text{med}_j \text{abs}[r_i - \text{med}_j(r_j)]$$

Dit valt echter in SAS nauwelijks uit te drukken en bovendien zal $\text{med}_j(r_j)$ in het algemeen niet veel van 0 afwijken. Niettemin kan hier sprake zijn van een systematische overschatting van de restvariantie en dus een onderschatting van de gestandaardiseerde residuen. Een derde mogelijke verklaring is dat de uitschieters in de eerste stap te veel aan het model trokken. In 2 van de 3 gevallen uit deze categorie kregen alle uitschieters initieel het gewicht 0 mee. Maar in 1 geval kreeg de geringste door LMS gedetecteerde uitschieter bij de eerste stap van de QD methode het gewicht 1. Het is dan ook begrijpelijk dat dit punt in de tweede stap niet meer werd ontmaskerd.

Van de 20 data-sets bij PROGRESS moeten er nu nog 2 besproken worden. 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 QD op overtuigende wijze. Het lijkt echter niet erg waarschijnlijk dat dit soort data-sets in de natuur voorkomen.

Tenslotte was er nog een echte data-set waarin Rousseeuw en Leroy 4 kunstmatige uitschieters aanbrachten. Deze uitschieters waren zo gekozen dat de criteria van Belsley, Kuh en Welsch er geen vat op hadden. Het is niet verwonderlijk dat ook QD (die immers met een klassieke least-squares begint) hier ook niet succesvol in was.

Samenvattend blijkt de gebrekkigheid van QD nogal me te vallen. Bij geen enkele echte data-set faalde de methode. En de twee gevallen waarin QD tekort schoot waren kunstmatig en speciaal geconstrueerd om varianten van klassieke least-squares te kunnen ontmaskeren.

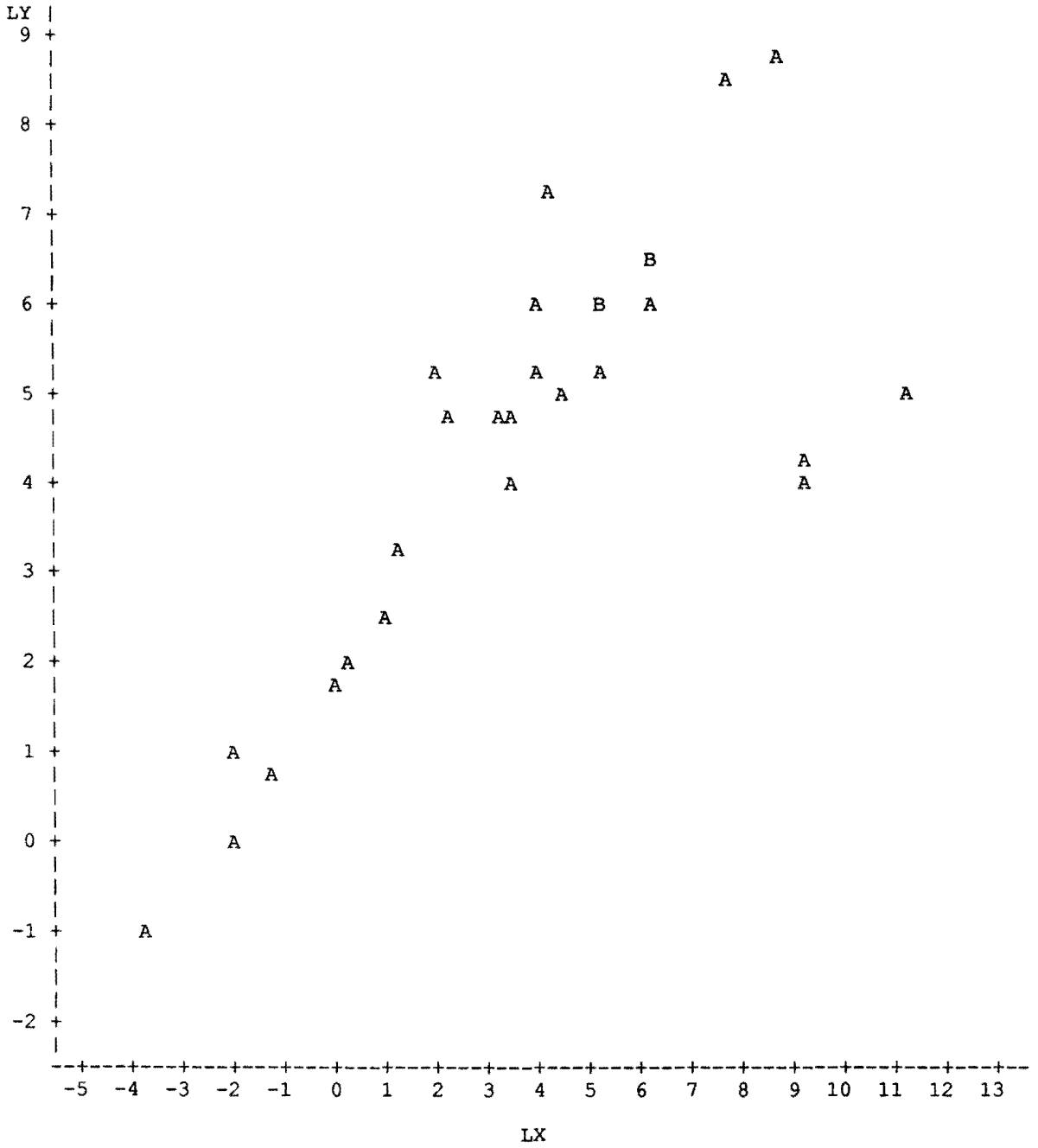
6. Litteratuur

- Rousseeuw, P.J. and A.M. Leroy (1987) Robust regression and outlier detection
John Wiley & Sons, New York
- Weisberg, S. (1980) Applied linear regression
John Wiley & Sons, New York
- Jerison, H.J. (1973) Evolution of the brain and intelligence
Academic Press, New York
- SAS/STAT Guide for personal computers (1987) Version 6 edition
SAS Institute Inc.
- Cook, R.D. (1977) Detection of influential observations in linear regression
Technometrics (19) 15-18
- Hampel F.R. (1971) A general qualitative definition of robustness
Annals of Mathematical Statistics (42) 1887-1896

- Rousseeuw, P.J. (1984) Least median of squares regression
Journal of the American Statistical Association (79) 871-880
- Siegel, A.F. (1982) Robust regression using repeated medians
Biometrika (69) 242-244
- Rousseeuw, P.J. and V.J. Yohai (1984) Robust regression by means of S-estimators
Lecture Notes in Statistics (26) 256-272
- Yohai, V.J. (1987) High breakdown-point and high efficiency robust estimates for regression
The Annals of Statistics (20) 642-656
- Belsley, D.A., E. Kuh and R.E. Welsch (1980) Regression diagnostics
John Wiley & Sons, New York
- Dijkstra, J.B. (1984) Opgavenbundel regressie-analyse voor tweedejaars wiskunde-
studenten
Technische Universiteit Eindhoven
- Jansen, F.J. (1988) Afstudeerverslag: Robuuste regressie-analyse
Technische Universiteit Eindhoven
- Debets, P., J.B. Dijkstra, G. Eilers, H. van der Knaap, A. Otten, J. Raatgever, D. van der
Sluis, G. Stemerding, A. Verbeek en B. van Zomeren (1989) Regressie-diagnostiek in
verschillende pakketten
Verschijnt binnenkort in Kwantitatieve Methoden
- Dijkstra, J.B. (1980) Analysis of means in some non-standard situations
CWI Tract (47) Centrum voor Wiskunde en Informatica

body and brain weight for 28 animals 1
9:10 MONDAY, AUGUST 7, 1989

PLOT OF LY*LX LEGEND: A = 1 OBS, B = 2 OBS, ETC.



body and brain weight for 28 animals

1
14:47 FRIDAY, JULY 28, 1989

OBS	X	Y	LX	LY
1	1.4	8.1	0.3001	2.09186
2	465.0	423.0	6.1420	6.04737
3	36.3	119.5	3.5926	4.78332
4	27.7	115.0	3.3200	4.74493
5	1.0	5.5	0.0392	1.70475
6	11700.0	50.0	9.3673	3.91202
7	2547.0	4603.0	7.8427	8.43446
8	187.1	419.0	5.2316	6.03787
9	521.0	655.0	6.2558	6.48464
10	10.0	115.0	2.3026	4.74493
11	3.3	25.6	1.1939	3.24259
12	529.0	680.0	6.2710	6.52209
13	207.0	406.0	5.3327	6.00635
14	62.0	1320.0	4.1271	7.18539
15	6654.0	5712.0	8.8030	8.65032
16	9400.0	70.0	9.1485	4.24850
17	6.8	179.0	1.9169	5.18739
18	35.0	56.0	3.5553	4.02535
19	0.1	1.0	-2.1203	0.00000
20	0.0	0.4	-3.7723	-0.91629
21	2.5	12.1	0.9163	2.49321
22	55.5	175.0	4.0164	5.16479
23	100.0	157.0	4.6052	5.05625
24	52.2	440.0	3.9543	6.08677
25	87000.0	154.5	11.3737	5.04019
26	0.3	1.9	-1.2730	0.64185
27	0.1	3.0	-2.1037	1.09861
28	192.0	180.0	5.2575	5.19296

body and brain weight for 28 animals 2
 14:47 FRIDAY, JULY 28, 1989

DEP VARIABLE: LY

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	94.43902399	94.43902399	40.261	0.0001
ERROR	26	60.98798786	2.34569184		
C TOTAL	27	155.42701185			
ROOT MSE		1.531565	R-SQUARE	0.6076	
DEP MEAN		4.425446	ADJ R-SQ	0.5925	
C.V.		34.60816			

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > T
INTERCEP	1	2.55489813	0.41313696	6.184	0.0001
LX	1	0.49599469	0.0781694	6.345	0.0001

OBS	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL
1	2.0919	2.7037	0.3967	-.611884	1.4793
2	6.0474	5.6013	0.3437	0.4461	1.4925
3	4.7833	4.3368	0.2898	0.4465	1.5039
4	4.7449	4.2016	0.2916	0.5433	1.5036
5	1.7047	2.5744	0.4110	-.869603	1.4754
6	3.9120	7.2011	0.5245	-3.289	1.4389
7	8.4345	6.4448	0.4302	1.9896	1.4699
8	6.0379	5.1498	0.3111	0.8881	1.4996
9	6.4846	5.6577	0.3486	0.8269	1.4914
10	4.7449	3.6970	0.3114	1.0480	1.4996
11	3.2426	3.1471	0.3527	0.095515	1.4904
12	6.5221	5.6653	0.3492	0.8568	1.4912
13	6.0064	5.1999	0.3141	0.8065	1.4990
14	7.1854	4.6019	0.2908	2.5835	1.5037
15	8.6503	6.9211	0.4883	1.7292	1.4516
16	4.2485	7.0925	0.5103	-2.844	1.4440
17	5.1874	3.5057	0.3237	1.6817	1.4970
18	4.0254	4.3183	0.2899	-0.29298	1.5039
19	0	1.5033	0.5439	-1.5033	1.4317
20	-.916291	0.6839	0.6569	-1.6002	1.3835
21	2.4932	3.0094	0.3655	-.516168	1.4873
22	5.1648	4.5470	0.2901	0.6178	1.5038
23	5.0562	4.8390	0.2967	0.2172	1.5026

body and brain weight for 28 animals 3
 14:47 FRIDAY, JULY 28, 1989

OBS	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL
24	6.0868	4.5162	0.2898	1.5706	1.5039
25	5.0402	8.1962	0.6610	-3.156	1.3816
26	0.6419	1.9235	0.4891	-1.2817	1.4514
27	1.0986	1.5115	0.5428	-.412845	1.4321
28	5.1930	5.1626	0.3119	.0303689	1.4995

OBS	STUDENT RESIDUAL	-2-1-0 1 2	COOK'S D
1	-.413635		0.006
2	0.2989		0.002
3	0.2969		0.002
4	0.3614		0.002
5	-.589401	*	0.013
6	-2.2857	****	0.347
7	1.3536	**	0.078
8	0.5922	*	0.008
9	0.5545	*	0.008
10	0.6988	*	0.011
11	.0640863		0.000
12	0.5746	*	0.009
13	0.5380	*	0.006
14	1.7181	***	0.055
15	1.1912	**	0.080
16	-1.9695	***	0.242
17	1.1234	**	0.030
18	-.194817		0.001
19	-1.05	**	0.080
20	-1.1566	**	0.151
21	-.347047		0.004
22	0.4108		0.003
23	0.1446		0.000
24	1.0443	**	0.020
25	-2.2843	****	0.597
26	-.883077	*	0.044
27	-.288272		0.006
28	.0202531		0.000

SUM OF RESIDUALS -4.05231E-15
 SUM OF SQUARED RESIDUALS 60.98799
 PREDICTED RESID SS (PRESS) 76.79557

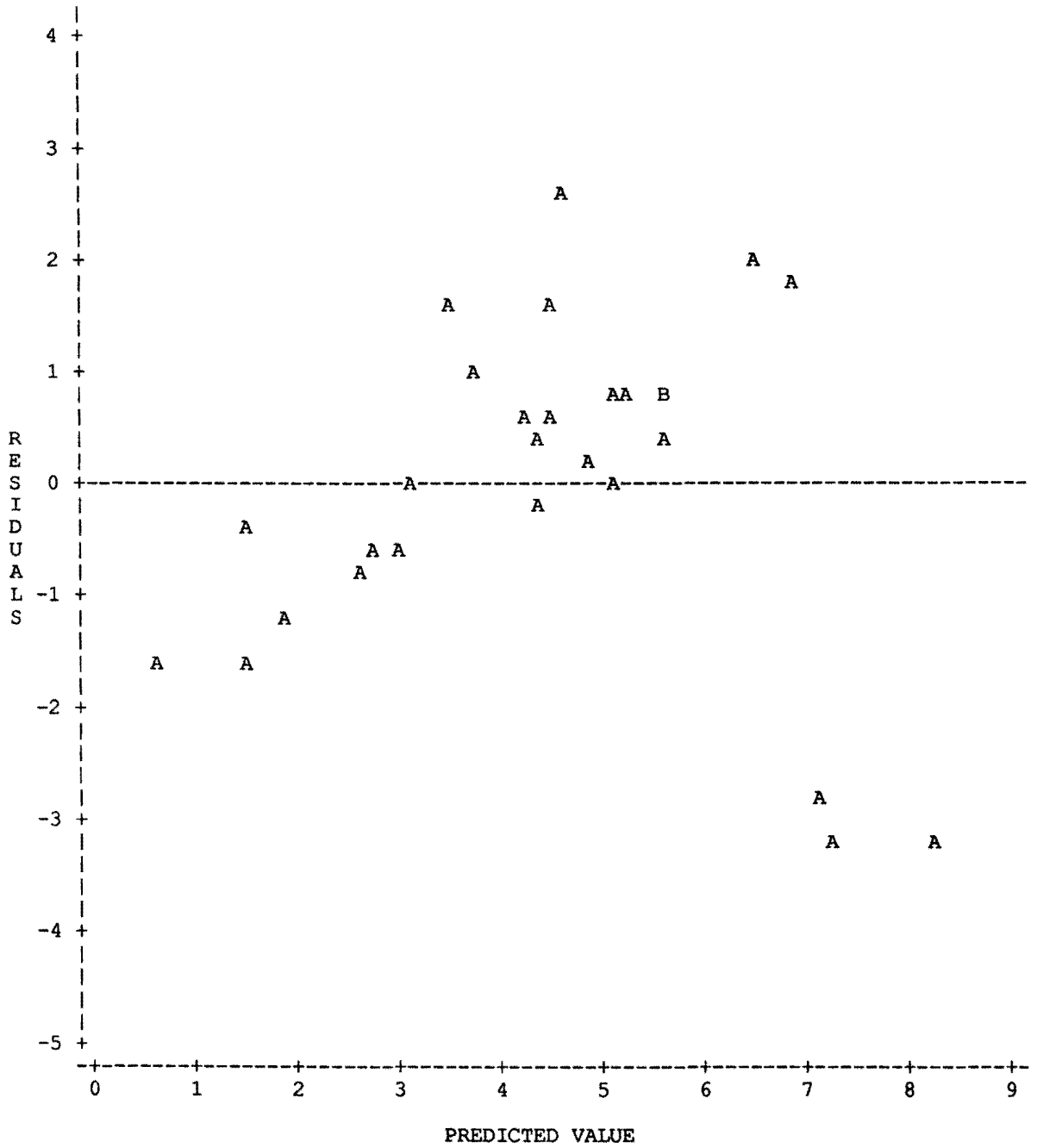
body and brain weight for 28 animals 4
14:47 FRIDAY, JULY 28, 1989

OBS	RESIDUAL	RSTUDENT	HAT DIAG H	COV RATIO	DFFITs	INTERCEP DFBETAS	LX DFBETAS
1	-.611884	-0.4069	0.0671	1.1442	-0.1091	-0.1090	0.0746
2	0.4461	0.2936	0.0504	1.1311	0.0676	0.0139	0.0365
3	0.4465	0.2916	0.0358	1.1142	0.0562	0.0413	-0.0027
4	0.5433	0.3552	0.0362	1.1110	0.0689	0.0539	-0.0083
5	-.869603	-0.5819	0.0720	1.1346	-0.1621	-0.1621	0.1151
6	-3.289	-2.5074	0.1173	0.7824	-0.9140	0.1906	-0.7622
7	1.9896	1.3767	0.0789	1.0146	0.4029	-0.0228	0.2981
8	0.8881	0.5847	0.0413	1.0979	0.1213	0.0473	0.0445
9	0.8269	0.5469	0.0518	1.1139	0.1278	0.0235	0.0712
10	1.0480	0.6918	0.0413	1.0862	0.1436	0.1313	-0.0530
11	0.095515	0.0628	0.0530	1.1418	0.0149	0.0146	-0.0085
12	0.8568	0.5670	0.0520	1.1121	0.1328	0.0241	0.0743
13	0.8065	0.5305	0.0421	1.1041	0.1112	0.0409	0.0432
14	2.5835	1.7893	0.0360	0.8817	0.3460	0.2177	0.0331
15	1.7292	1.2013	0.1017	1.0762	0.4041	-0.0645	0.3255
16	-2.844	-2.0937	0.1110	0.8807	-0.7399	0.1409	-0.6094
17	1.6817	1.1293	0.0447	1.0249	0.2442	0.2310	-0.1094
18	-0.29298	-0.1912	0.0358	1.1185	-0.0369	-0.0273	0.0021
19	-1.5033	-1.0521	0.1261	1.1350	-0.3997	-0.3905	0.3384
20	-1.6002	-1.1645	0.1840	1.1925	-0.5529	-0.5248	0.4963
21	-.516168	-0.3411	0.0569	1.1363	-0.0838	-0.0830	0.0512
22	0.6178	0.4041	0.0359	1.1073	0.0780	0.0508	0.0051
23	0.2172	0.1418	0.0375	1.1220	0.0280	0.0147	0.0062
24	1.5706	1.0462	0.0358	1.0296	0.2016	0.1340	0.0100
25	-3.156	-2.5055	0.1863	0.8492	-1.1987	0.4013	-1.0777
26	-1.2817	-0.8792	0.1020	1.1333	-0.2963	-0.2933	0.2389
27	-.412845	-0.2831	0.1256	1.2291	-0.1073	-0.1049	0.0908
28	.0303689	0.0199	0.0415	1.1284	0.0041	0.0016	0.0015

body and brain weight for 28 animals

5
14:47 FRIDAY, JULY 28, 1989

PLOT OF RES1*FIT1 LEGEND: A = 1 OBS, B = 2 OBS, ETC.



body and brain weight for 28 animals 6
14:47 FRIDAY, JULY 28, 1989

UNIVARIATE

VARIABLE=ARES1

MOMENTS

N	28	SUM WGTS	28
MEAN	1.16983	SUM	32.7551
STD DEV	0.916316	VARIANCE	0.839636
SKEWNESS	1.00707	KURTOSIS	0.189976
USS	60.988	CSS	22.6702
CV	78.3293	STD MEAN	0.173168
T:MEAN=0	6.75546	PROB> T	0.0001
SGN RANK	203	PROB> S	0.0001
NUM ^= 0	28		

QUANTILES (DEF=4)

100% MAX	3.28903	99%	3.28903
75% Q3	1.66132	95%	3.22916
50% MED	0.86321	90%	2.87519
25% Q1	0.463906	10%	0.205038
0% MIN	0.0303689	5%	0.0596847
		1%	0.0303689
RANGE	3.25866		
Q3-Q1	1.19741		
MODE	0.0303689		

EXTREMES

LOWEST	HIGHEST
0.0303689	1.98964
0.095515	2.58345
0.217208	2.84399
0.29298	3.15598
0.412845	3.28903

body and brain weight for 28 animals

14:47 FRIDAY, JULY 28, 1989 ⁷

DEP VARIABLE: LY

ANALYSIS OF VARIANCE

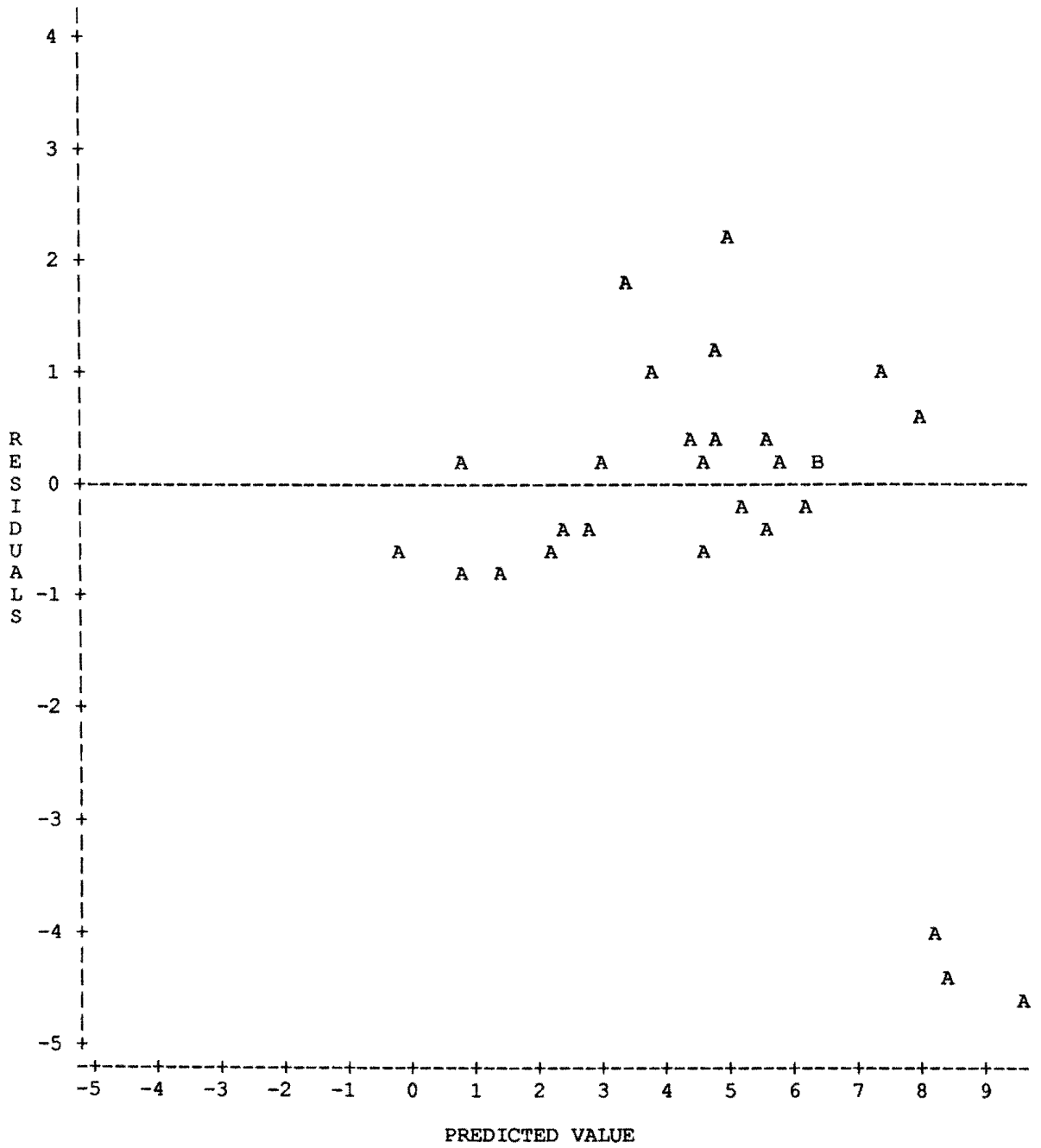
SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	35.94520302	35.94520302	326.851	0.0001
ERROR	12	1.31969004	0.10997417		
C TOTAL	13	37.26489306			
ROOT MSE		0.3316235	R-SQUARE	0.9646	
DEP MEAN		4.496737	ADJ R-SQ	0.9616	
C.V.		7.37476			

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > T
INTERCEP	1	2.22139656	0.15393127	14.431	0.0001
LX	1	0.65470548	0.03621353	18.079	0.0001

body and brain weight for 28 animals 8
14:47 FRIDAY, JULY 28, 1989

PLOT OF RES2*FIT2 LEGEND: A = 1 OBS, B = 2 OBS, ETC.



body and brain weight for 28 animals 9
14:47 FRIDAY, JULY 28, 1989

UNIVARIATE

VARIABLE=ARES2

MOMENTS

N	28	SUM WGTS	28
MEAN	1.00965	SUM	28.2703
STD DEV	1.27775	VARIANCE	1.63264
SKEWNESS	2.101	KURTOSIS	3.46263
USS	72.6246	CSS	44.0814
CV	126.553	STD MEAN	0.241472
T:MEAN=0	4.18124	PROB> T	.000273522
SGN RANK	203	PROB> S	0.0001
NUM ^= 0	28		

QUANTILES (DEF=4)

100% MAX	4.6276	99%	4.6276
75% Q3	1.06283	95%	4.54418
50% MED	0.497151	90%	4.01043
25% Q1	0.264306	10%	0.193559
0% MIN	0.167565	5%	0.173242
		1%	0.167565
RANGE	4.46004		
Q3-Q1	0.79852		
MODE	0.167565		

EXTREMES

LOWEST	HIGHEST
0.167565	1.71097
0.180181	2.26193
0.195046	3.96245
0.19525	4.44223
0.209796	4.6276

body and brain weight for 28 animals 10
14:47 FRIDAY, JULY 28, 1989

DEP VARIABLE: LY

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	138.62005664	138.62005664	371.155	0.0001
ERROR	22	8.21661753	0.37348261		
C TOTAL	23	146.83667416			
ROOT MSE		0.6111322	R-SQUARE	0.9440	
DEP MEAN		4.313599	ADJ R-SQ	0.9415	
C.V.		14.16757			

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > T
INTERCEP	1	2.09625860	0.16973072	12.350	0.0001
LX	1	0.74345084	0.03858997	19.265	0.0001

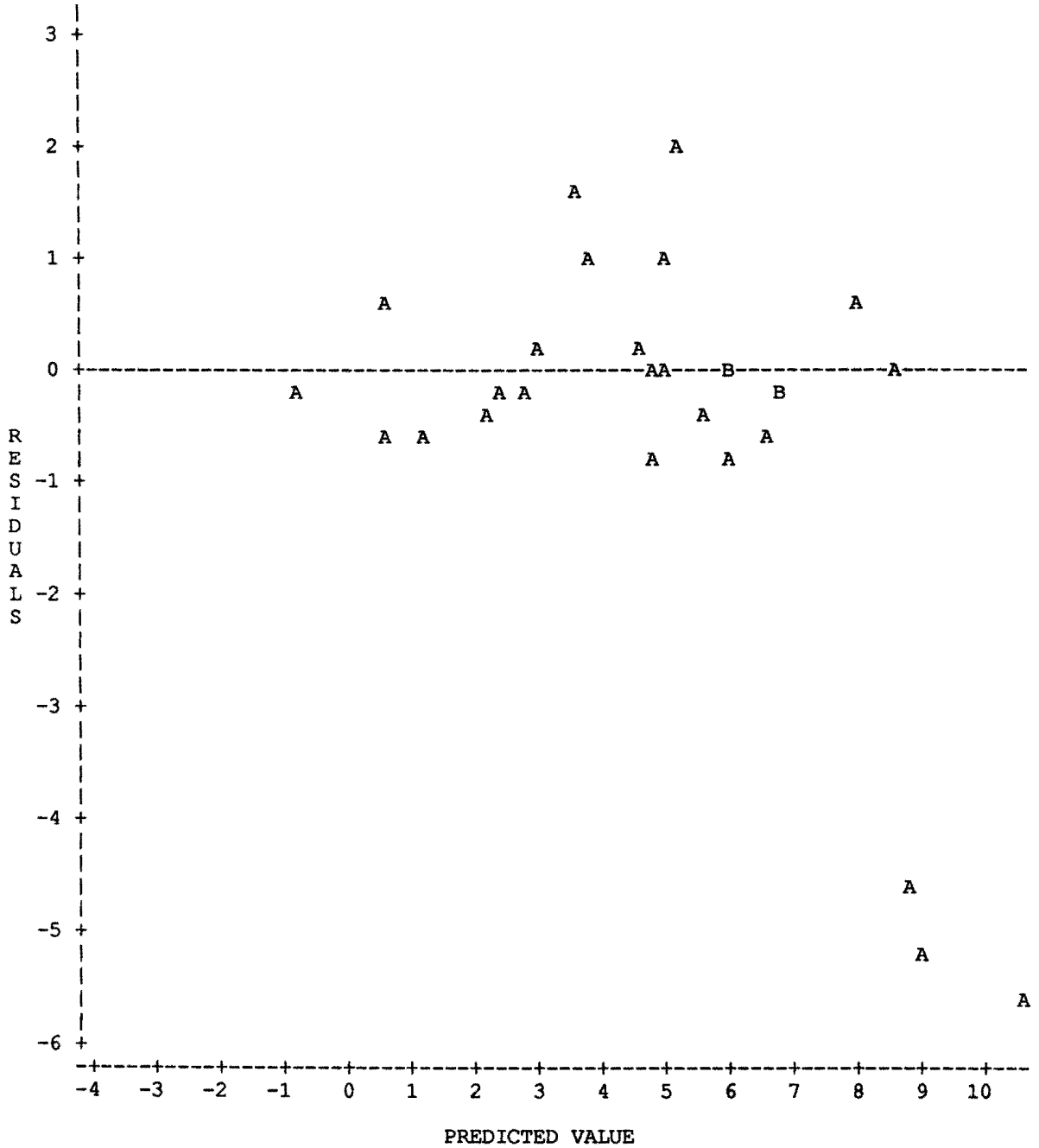
body and brain weight for 28 animals

11

14:47 FRIDAY, JULY 28, 1989

PLOT OF RES3*FIT3

LEGEND: A = 1 OBS, B = 2 OBS, ETC.



body and brain weight for 28 animals 12
14:47 FRIDAY, JULY 28, 1989

OBS	FIT1	FIT2	FIT3	RES1	RES2	RES3	W1	W2	D
1	2.70375	2.41788	2.3194	-0.6119	-0.3260	-0.2275	1	1	-0.4423
2	5.60132	6.24262	6.6626	0.4461	-0.1952	-0.6152	1	1	-0.2649
3	4.33683	4.57352	4.7672	0.4465	0.2098	0.0161	1	1	0.2846
4	4.20159	4.39501	4.5645	0.5433	0.3499	0.1804	1	1	0.4747
5	2.57435	2.24707	2.1254	-0.8696	-0.5423	-0.4207	0	1	-0.7358
6	7.20105	8.35425	9.0604	-3.2890	-4.4422	-5.1484	0	0	-6.0268
7	6.44482	7.35604	7.9269	1.9896	1.0784	0.5076	0	1	1.4631
8	5.14977	5.64658	5.9857	0.8881	0.3913	0.0521	0	1	0.5309
9	5.65772	6.31707	6.7471	0.8269	0.1676	-0.2625	1	1	0.2273
10	3.69697	3.72891	3.8081	1.0480	1.0160	0.9368	0	1	1.3784
11	3.14708	3.00306	2.9839	0.0955	0.2395	0.2587	1	1	0.3250
12	5.66528	6.32705	6.7584	0.8568	0.1950	-0.2363	1	1	0.2646
13	5.19990	5.71276	6.0609	0.8065	0.2936	-0.0545	1	1	0.3983
14	4.60193	4.92345	5.1646	2.5835	2.2619	2.0208	0	0	3.0688
15	6.92113	7.98475	8.6408	1.7292	0.6656	0.0095	0	1	0.9030
16	7.09249	8.21095	8.8977	-2.8440	-3.9625	-4.6492	0	0	-5.3759
17	3.50568	3.47642	3.5214	1.6817	1.7110	1.6660	0	1	2.3213
18	4.31833	4.54910	4.7395	-0.2930	-0.5238	-0.7141	1	1	-0.7106
19	1.50326	0.83325	0.5199	-1.5033	-0.8332	-0.5199	0	1	-1.1305
20	0.68388	-0.24832	-0.7082	-1.6002	-0.6680	-0.2081	0	1	-0.9062
21	3.00937	2.82130	2.7775	-0.5162	-0.3281	-0.2843	1	1	-0.4451
22	4.54700	4.85094	5.0822	0.6178	0.3138	0.0825	1	1	0.4258
23	4.83904	5.23643	5.5200	0.2172	-0.1802	-0.4637	1	1	-0.2445
24	4.51622	4.81031	5.0361	1.5706	1.2765	1.0507	0	1	1.7318
25	8.19617	9.66780	10.5520	-3.1560	-4.6276	-5.5118	0	0	-6.2783
26	1.92351	1.38798	1.1499	-1.2817	-0.7461	-0.5080	0	1	-1.0123
27	1.51146	0.84407	0.5322	-0.4128	0.2545	0.5664	1	1	0.3453
28	5.16259	5.66351	6.0049	0.0304	-0.4706	-0.8120	1	1	-0.6384

Copyright (c) 1986 SAS Institute Inc., Cary, N. C. 27512, U. S. A.
NOTE: VMS SAS Production Release 5.16 at EINDHOVEN UNIVERSITY OF TECHNOLOGY (04717001).
NOTE: LICENSED CPUID MODEL = VAX(CLASS H), SERIAL = 06FF5322.

```
1 options pagesize=60 linesize=80;
2 data een;
3 title 'body and brain weight for 28 animals';
4 infile brain;
5 input x y;
NOTE: INFILE BRAIN IS FILE USER4:[RCSAS]BRAIN.DAT.
NOTE: 28 LINES WERE READ FROM INFILE BRAIN.
      THE MINIMUM LINE LENGTH IS 5.
      THE MAXIMUM LINE LENGTH IS 11.
NOTE: THE DATA SET WORK.EEN HAS 28 OBSERVATIONS AND 2 VARIABLES.
NOTE: THE DATA STEP USED 00:00:00.39 CPU SECONDS, 355 PAGEFAULTS.
6 data twee;
7 set een;
8 lx=log(x);
9 ly=log(y);
NOTE: THE DATA SET WORK.TWEE HAS 28 OBSERVATIONS AND 4 VARIABLES.
NOTE: THE DATA STEP USED 00:00:00.33 CPU SECONDS, 82 PAGEFAULTS.
10 proc print;
NOTE: THE PROCEDURE PRINT USED 00:00:00.24 CPU SECONDS, 94 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 1.
11 proc reg;
12 model ly=lx/r influence;
13 output out=drie p=fit1 r=res1;
NOTE: THE DATA SET WORK.DRIE HAS 28 OBSERVATIONS AND 6 VARIABLES.
NOTE: THE PROCEDURE REG USED 00:00:00.53 CPU SECONDS, 63 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGES 2 THROUGH 4.
14 proc plot;
15 plot res1*fit1/vref=0;
NOTE: THE PROCEDURE PLOT USED 00:00:00.24 CPU SECONDS, 15 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 5.
16 data vier;
17 set drie;
18 ares1=abs(res1);
NOTE: THE DATA SET WORK.VIER HAS 28 OBSERVATIONS AND 7 VARIABLES.
NOTE: THE DATA STEP USED 00:00:00.28 CPU SECONDS, 1 PAGEFAULTS.
19 proc univariate;
20 var ares1;
NOTE: THE PROCEDURE UNIVARIATE USED 00:00:00.21 CPU SECONDS, 7 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 6.
21 data vijf;
22 set vier;
23 w1=1;
24 if ares1>.86321 then w1=0;
NOTE: THE DATA SET WORK.VIJF HAS 28 OBSERVATIONS AND 8 VARIABLES.
NOTE: THE DATA STEP USED 00:00:00.30 CPU SECONDS, 14 PAGEFAULTS.
25 proc reg;
26 model ly=lx;
27 weight w1;
28 output out=zes p=fit2 r=res2;
NOTE: THE DATA SET WORK.ZES HAS 28 OBSERVATIONS AND 10 VARIABLES.
NOTE: THE PROCEDURE REG USED 00:00:00.22 CPU SECONDS, 5 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 7.
29 proc plot;
30 plot res2*fit2/vref=0;
```

S A S L O G VMS SAS 5.16 2
14:47 FRIDAY, JULY 28, 1989

NOTE: THE PROCEDURE PLOT USED 00:00:00.15 CPU SECONDS, 7 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 8.
31 data zeven;
32 set zes;
33 ares2=abs(res2);
NOTE: THE DATA SET WORK.ZEVEN HAS 28 OBSERVATIONS AND 11 VARIABLES.
NOTE: THE DATA STEP USED 00:00:00.26 CPU SECONDS, 0 PAGEFAULTS.
34 proc univariate;
35 var ares2;
NOTE: THE PROCEDURE UNIVARIATE USED 00:00:00.13 CPU SECONDS, 0 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 9.
36 data acht;
37 set zeven;
38 s0=1.4826*.497151;
39 d=res2/s0;
40 w2=1;
41 if abs(d)>2.5 then w2=0;
NOTE: THE DATA SET WORK.ACHT HAS 28 OBSERVATIONS AND 14 VARIABLES.
NOTE: THE DATA STEP USED 00:00:00.39 CPU SECONDS, 1 PAGEFAULTS.
42 proc reg;
43 model ly=lx;
44 weight w2;
45 output out=negen p=fit3 r=res3;
NOTE: THE DATA SET WORK.NEGEN HAS 28 OBSERVATIONS AND 16 VARIABLES.
NOTE: THE PROCEDURE REG USED 00:00:00.24 CPU SECONDS, 3 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 10.
46 proc plot;
47 plot res3*fit3/vref=0;
NOTE: THE PROCEDURE PLOT USED 00:00:00.18 CPU SECONDS, 0 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 11.
48 proc print;
49 var fit1 fit2 fit3 res1 res2 res3 w1 w2 d;
NOTE: THE PROCEDURE PRINT USED 00:00:00.22 CPU SECONDS, 4 PAGEFAULTS.
NOTE: THE PROCEDURE PRINTED PAGE 12.
NOTE: SAS INSTITUTE INC., SAS CIRCLE, BOX 8000, CARY, N. C., 27512-8000.