Tekortkomingen van de Cook-statistic

Eindhoven University of Technology
Computing Centre Note 44

Tekortkomingen van de Cook-statistic

Jan B. Dijkstra

Samengesteld voor de Statistische Dag
op 20 maart 1989 in Utrecht.

januari 1989.

Tekortkomingen van de Cook-statistic

Jan B. Dijkstra

Samenvatting

Gebruikers van regressie-analyse inspecteren doorgaans de residuen. Omdat bij de toetsing verondersteld wordt dat de residuen onafhankelijk normaal verdeeld zijn met verwachting 0 en constante variantie wordt de inspectie er meestal op gericht om afwijkingen van deze vooronderstellingen op te sporen. Voor de hand liggen dan een normaliteitstoets, een plaatje waarin de residuen worden uitgezet tegen de aangepaste waarden en een tabel met gestandaardiseerde residuen. Als de normaliteit niet verworpen wordt, het plaatje een ongestruktureerde zwerm punten voorstelt en geen enkel gestandaardiseerd residu in absolute waarde groter is dan 2, dan worden de regressie-resultaten met vertrouwen geinterpreteerd. Uitzonderlijk scrupuleuze onderzoekers gaan soms nog een stapje verder. Zij weten dat zeer invloedrijke punten soms niet herkenbaar zijn door grote waarden voor de gestandaardiseerde residuen en inspecteren dus ook nog een invloedsmaat. De laatste 10 jaar geniet in dit opzicht de Cook-statistic een toenemende populariteit en deze is dan ook opgenomen in statistische pakketten als SAS, SPSS en BMDP. Helaas is deze maat echter niet fool-proof, zoals in het nu volgende verhaal zal worden geillustreerd.

Inleiding

Het model bij regressie-analyse is $y = \beta_0+\beta_1 x_1+ \cdots +\beta_k x_k +\epsilon$. De waarnemingen y_i worden gedaan bij instellingen x_{ij}. Hierbij geldt $i = 1, \cdots, n$ en $j = 1, \cdots, k$. Er zijn dus n waarnemingen en k voorspellende variabelen. Met betrekking tot de notatie geldt verder de volgende conventie: Een hoofdletter representeert een matrix, een kleine letter zonder index een vector en een kleine letter met index een scalar. De designmatrix X bestaat uit de de elementen x_{ij}, voorafgegaan door een kolom van n eenen. Van de afwijkingen ϵ_i wordt verondersteld dat ze onafhankelijk normaal verdeeld zijn met verwachting 0 en constante variantie σ^2. In dat geval wordt een maximum-likelihood schatter voor β verkregen door $\sum_{i=1}^{n} \epsilon_i^2$ te minimaliseren als functie van β. Dit leidt tot de normaal-vergelijkingen $X^T Xb = X^T y$, waarin b de oplossing van het stelsel representeert. Hiermee worden dan weer de aangepaste waarden $\hat{y} = Xb$ gevormd. De uiteindelijke residuen zijn $e_i = y_i - \hat{y}_i$. De restkwadratensom $SSE = \sum_{i=1}^{n} e_i^2$ gedeeld door het aantal

Rekencentrum Technische Universiteit, Postbus 513, 5600 MB Eindhoven, telefoon 040-474535/474089

vrijheidsgraden $n-p$ levert het gemiddelde kwadraat *MSE* op. Het aantal parameters p is hierbij gelijk aan $k+1$. Omdat *MSE* een schatter is voor de variantie σ^2 geldt dat $\hat{\sigma} = \sqrt{MSE}$. Hiermee worden de gestandaardiseerde residuen $d_i = e_i/\hat{\sigma}$ gevormd.

Een voorbeeld

Gegeven zijn een aantal waarnemingen y bij ingestelde waarden voor x.

X	Y
1.5	4.3
1.6	3.7
2.8	3.5
2.8	2.4
3.9	2.0
4.5	1.0
13.8	5.2

Het aangepaste model is $\hat{y} = 2.539+0.1399x$ met als determinatie-coefficient de waarde 0.173. Omdat dit de fractie door het model verklaarde variantie voorstelt is het een nogal pover resultaat. Een onderzoek naar de significantie van x levert een overschrijdingskans van 0.3534 op. De normaliteit van de residuen wordt onderzocht met de toets van Shapiro en Wilk (1965). De overschrijdingskans is 0.627, zodat er geen aanleiding bestaat om aan de normaliteit te twijfelen. Ook de gestandaardiseerde residuen leveren niets verontrustends op: de grootste absolute waarde is 1.511 en deze wordt bereikt voor de zesde waarneming.

De residuen e_i en de aangepaste waarden $\hat{y}_i$ zijn ongecorreleerd. Een grafische weergave van het verband tussen deze twee variabelen moet dus bij voorkeur een chaotisch beeld opleveren. Bijlage 1 is dit echter allerminst. Heel duidelijk zien we dat er een punt is dat de lijn sterk naar zich toe trekt. Dit hefboom-punt is hier zo eenvoudig te ontmaskeren omdat er slechts een enkele voorspellende variabele is. In bijlage 2 zijn de punten en de aangepaste lijn getekend, en ook de lijn die verkregen wordt door bij de berekening het hefboom-punt buiten beschouwing te laten. Duidelijk is te zien dat het meest invloedrijke punt bij de eerste lijn niet aanleiding geeft tot het grootste residu, maar bij de tweede lijn wel. In bijlage 3 zijn de residuen weergegeven die behoren bij de tweede lijn. Het hefboom-punt is hier op overtuigende wijze geisoleerd.

De Cook-statistic

Het is niet zo eenvoudig om hefboom-punten op te sporen als er meerdere prediktors zijn. Een plotje van de residuen tegen de aangepaste waarden levert dan niet altijd de nodige informatie en dus is er behoefte aan een analytisch hulpmiddel. Daartoe gebruiken we de hat-matrix $H = X(X^TX)^{-1}X^T$. De diagonaal-elementen hiervan zijn h_{ii}. Er geldt $Var(e_i) = \sigma^2(1-h_{ii})$. Dit is geen constante omdat er behalve de variabiliteit van de waarnemingen ook nog de stochastiek van het aangepaste model in is verwerkt. Een alternatief voor de gestandaardiseerde residuen worden nu gevormd door de gestudentiseerde residuen

$f_i = e_i/\sqrt{MSE(1-h_{ii})}$. Hiermee heeft Cook (1977) zijn invloedsmaat D_i geconstrueerd:

$$D_i = \frac{f_i^2 Var(\hat{y}_i)}{pVar(e_i)} = \frac{f_i^2 h_{ii}}{p(1-h_{ii})}$$

Als aan de regressie-vooronderstellingen voldaan is, dan volgt D_i bij benadering de F_{n-p}^{p} verdeling. Zie bijvoorbeeld Montgomery en Peck (1982). Ongeacht de waarden van a en b levert een onbetrouwbaarheid $\alpha = 0.5$ een kritieke waarde voor F_b^a op van 1. Daarom is het interessant om punten te zoeken waarvoor $D_i > 1$. Als dit optreedt geldt: Weglating van punt i zou de vector b buiten het 50% simultane betrouwbaarheidsgebied voor β plaatsen dat behoort bij alle waarnemingen. Een dergelijk punt moet dus beschouwd worden als zeer invloedrijk.

Voortzetting van het voorbeeld

Hieronder wordt in een tabel de waarde gegeven van de residuen, de gestandaardiseerde residuen, de gestudentiseerde residuen en de Cook-statistics. Eerst betreft dit de complete dataset en in een tweede tabel worden de resultaten gegeven voor de dataset waarin het hefboom-punt buiten de berekening van de regressiecoefficienten is gehouden.

Complete dataset			
Residu	Standaard	Student	Cook
1.55	1.08	1.223	0.211
0.9368	0.6525	0.763	0.074
0.5688	0.3962	0.434	0.019
-0.5312	-0.37	-0.405	0.016
-1.085	-0.7559	-0.818	0.057
-2.169	-1.511	-1.632	0.222
0.7292	0.508	2.142	38.52

Na weglating van het hefboom-punt			
Residu	Standaard	Student	Cook
0.1782	0.3877	0.509	0.093
-0.3251	-0.7072	-0.9	0.252
0.635	1.381	1.513	0.23
-0.465	-1.011	-1.108	0.123
0.1984	0.4316	0.523	0.064
-0.2216	-0.4819	-0.714	0.304
12.96	28.21		

In de eerste tabel valt op dat bij het hefboom-punt (de laatste observatie) het gestudentiseerde residu zoveel groter is dan het gestandaardiseerde residu. Dit vergt enige toelichting. Er geldt $e_i = y_i - \hat{y}_i$. Dit herschrijven we tot $y_i = e_i + \hat{y}_i$. Voor de varianties wordt dit dus $Var(y_i) = Var(e_i) + Var(\hat{y}_i) + 2Cov(e_i, \hat{y}_i)$ Omdat e_i en $\hat{y}_i$ ongecorreleerd zijn valt de

laatste term weg. Er geldt dus $Var(e_i) = Var(y_i) - Var(\hat{y}_i)$. De variantie $Var(y_i) = \sigma^2$ van de waarnemingen zelf wordt geschat door MSE. De variantie $Var(\hat{y}_i) = h_{ii}\sigma^2$ van de aangepaste waarde hangt af van de plaats. In het centrum van de prediktor-waarden is deze variantie het kleinst, maar naar mate men zich hier verder van verwijdert wordt de aanpassing onzekerder. Daarmee neemt de variantie van de residuen dus af, zodat de gestudentiseerde residuen de gestandaardiseerde flink kunnen overheersen. Door deze twee maten te vergelijken kunnen excentrische punten in de prediktor-ruimte worden opgespoord. De Cook-statistic voor het hefboom-punt is 38.52. Dit vergelijken we met een F-verdeling met 2 en 5 vrijheidsgraden. Dit geeft een overschrijdingskans van 0.0009169 zodat weglating van punt 7 een aangepast model oplevert waarvan de parameters buiten het 99.91 procent simultane betrouwbaarheidsgebied vallen dat geldt voor de parameters bij de volledige dataset.

Verstrengelde hefboom-punten

Binnen het kader van een door de auteur dezes begeleid afstudeerproject heeft Jansen (1988) een pathologisch lastige dataset geconstrueerd. Het betreft kunstmatige waarnemingen y bij ingestelde waarden voor de prediktoren x_1 en x_2 die in de volgende tabel zijn weergegeven.

Nummer	X1	X2	Y
1	-1.8	-1.6	5.85
2	5	-4	17.05
3	5.5	-6.5	22.05
4	1	2	2.9
5	6.5	4.5	-6.5
6	5	0	10.4
7	0	-2	8.5
8	-2	7	-8.9
9	7	5	-8.3
10	0.5	2	2.15
11	6.5	-8	25.52
12	-1.5	3	-1.55

Bijlage 4 laat zien wat hier aan de hand is. Op twee na liggen alle punten ongeveer op een lijn in de driedimensionale ruimte. De punten met rangnummer 5 en 9 vallen daarbuiten. Niet omdat hun y-waarde zo uitzonderlijk is, maar omdat de instellingen voor x_1 en x_2 buiten het kleinste convexe omhulsel voor de overige instellingen in de prediktor-ruimte vallen. Bovendien liggen deze bijzondere punten vlak bij elkaar waardoor weglating van punt 5 of punt 9 de aanpassing niet erg sterk zal beinvloeden zolang het andere punt in de berekeningen nog is meegenomen.

Laten we beginnen met de standaard-analyse. Het aangepaste model voor alle waarnemingen is

$$\hat{y} = 5.814+0.0862x_1-2.386x_2$$

Hierbij hoort een determinatiecoefficient van 0.9602 zodat het model bijzonder goed lijkt te passen. De significantie van x_1 en x_2 is respectievelijk 0.7114 en 0.0001 zodat de argeloze gebruiker x_1 zonder problemen uit het model zou verwijderen. Het grootste gestudentiseerde residu is 1.755 en die waarde wordt bereikt voor de zesde observatie. De grootste Cook-statistic is 0.533 voor waarneming nummer 9. Een toets op de normaliteit van de residuen levert als overschrijdingskans 0.8249 zodat ook daar geen reden tot bezorgdheid ligt. Bijlage 5 tenslotte toont de grafiek van de residuen tegen de aangepaste waarden en ook daar lijkt niets mee aan de hand. De hefboompunten worden hier dus niet ontmaskerd. Niet middels de Cook-statistic omdat ze met elkaar verstengeld zijn en niet middels de grafiek van de residuen omdat de dimensie van de prediktor-ruimte groter is dan 2.

Hefboom-punten met meerdere prediktoren

Om het probleem van de verstrengeling te elimineren wordt nu punt 5 buiten de berekening gehouden en het model opnieuw aangepast. Het resultaat is

$$\hat{y} = 5.735+0.2171x_1-2.293x_2$$

met een determinatiecoefficient van 0.9606. De overschrijdigskansen voor x_1 en x_2 zijn respectievelijk 0.421 en 0.0001 zodat er wat dat betreft weinig veranderd is. Punt 9 heeft nu in absolute waarde het grootste gestudentiseerde residu van -2.827 en dat is verdacht hoog. Bij dit punt hoort ook de hoogste Cook-statistic van 5.389 en door deze waarde te vergelijken met een F-verdeling met 3 en 8 vijheidsgraden komen we tot een overschrijdingskans van 0.03294 zodat weglating van punt 9 de geschatte coefficienten buiten het simultane 96.71 procent betrouwbaarheidsgebied voor de coefficienten behorende bij de overige waarnemingen zou plaatsen. Punt 5 is hierbij uiteraard buiten beschouwing gelaten. In dit geval is de Cook-statistic dus de juiste weg om een hefboom-punt op te sporen. Maar zou dit punt nu ook met primitievere middelen gevonden kunnen worden?

Het grootste gestandaardiseerde residu (in absolute waarde) is -1.625. Deze waarde wordt weliswaar voor punt 9 bereikt, maar is nog dermate klein dat hierin geen aanleiding voor verder onderzoek kan worden gevonden. In bijlage 6 zijn de residuen geplot tegen de aangepaste waarden. De verschillen met bijlage 5 zijn maar klein en ook hierin valt niets verontrustends te zien. Een toets op de normaliteit van de residuen levert vervolgens een overschrijdingskans van 0.5422 op en ook daar lijkt dus niets mee aan de hand. De conclusie van deze sectie is duidelijk: Als er meerdere prediktoren zijn, vindt de Cook-statistic hefboom-punten die niet op primitievere manieren zichtbaar gemaakt kunnen worden. De voorwaarde is hierbij echter wel dat er geen sprake is van verstrengeling.

Robuuste aanpassing

Bijlage 7 geeft de residuen nadat het model is aangepast zonder met observaties nummer 5 en 9 rekening te houden. De aanpassing is dan

$$\hat{y} = 5.0928+1.062x_1-1.692x_2$$

met een determinatiecoefficient van 0.9999. De significantie van beide prediktoren is 0.0001. Kortom: een ideale aanpassing en de vraag dringt zich op of dit resultaat niet ook op een andere manier bereikt had kunnen worden.

Van een procedure die de som van de absolute waarde van de residuen minimaliseert valt niet veel te verwachten. Dit is een ML-schatter voor dubbelexponentieel verdeelde fouten en dat betekent dat met grotere afwijkingen in de y-richting beter wordt omgegaan. De afwijking zit in dit voorbeeld echter in de prediktor-ruimte. De weinig bevredigende aanpassing

$$\hat{y} = 7.448 - 0.372x_1 - 2.562x_2$$

is dus niet erg verrassend. De kwaliteit van de aanpassing wordt hier en verderop in deze sectie bepaald door een vergelijking met de hierboven gegeven ideale fit waarbij de punten 5 en 9 buiten beschouwing zijn gebleven.

In Holland en Welsch (1977) wordt een methode voor iteratief herwogen kleinste kwadraten beschreven. Dit proces is hier ook gebruikt met de gewichtsfunctie van Huber (1973). Bij de klassieke regressie-methode is de invloed van de residuen kwadratisch. Bij Huber worden de residuen eerst gestandaardiseerd middels een robuuste schatter voor σ. Vervolgens is hun invloed kwadratisch voor $|d_i| \leq H$ en lineair als $|d_i| > H$. Voor H is hier de waarde 1.345 gekozen wat resulteert in 95% efficiency voor normale fouten. Dit iteratieve proces heeft een beginschatting nodig. Hiervoor is de routine gekozen die de som van de absolute waarde van de residuen minimaliseert. De uiteindelijke aanpassing is

$$\hat{y} = 5.808 + 0.083x_1 - 2.387x_2$$

en dat is een matig resultaat. Ook Huber's methode is uitsluitend geschikt om uitschieters in de y-richting te dempen.

Een heel andere methode heet Least Median of Squares en is afkomstig van Rousseeuw (1984). Hier wordt de mediaan van de gekwadrateerde residuen geminimaliseerd. In de tweedimensionale ruimte valt deze methode als volgt qua effect te schetsen. Knip een strook karton uit en probeer die over 50% van de data-punten te leggen. Lukt dat, maak de strook dan wat smaller en probeer het weer. Ga met kleine stapjes zo door tot het niet meer lukt. Leg dan de vorige strook terug en de middellijn hiervan is de oplossing. Heuristische algorithmes voor deze methode werken heel anders, maar de hier gegeven beschrijving illustreert duidelijk tegen wat voor soort ellende Least Median of Squares is opgewassen. Tussen uitschieters in de y-richting en in de prediktor-ruimte wordt nauwelijks onderscheid gemaakt. Het is dan ook niet verwonderlijk dat hier een goede aanpassing werd gevonden:

$$\hat{y} = 5.059 + 1.067x_1 - 1.69x_2$$

Least Median of Squares is weliswaar een zeer robuuste methode, maar de efficiency laat veel te wensen over. Daarom lijkt het aantrekkelijk de variantie van de schatters te verminderen door een iteratief herwogen kleinste kwadraten proces met deze methode als beginschatter. Het lijkt niet verstandig om de gestandaardiseerde residuen zo invloedrijk te laten als bij Huber. Uiteindelijk trekt de aanpassing dan toch naar de hefboom-punten toe

en ontstaat het model

$$\hat{y} = 5.662 + 0.285x_1 - 2.239x_2$$

dat veel te wensen overlaat. In Hampel, Rousseeuw, Ronchetti en Stahel (1986) wordt een invloedsfunktie gebaseerd op de tangens hyperbolicus beschreven. Hierbij hebben residuen met een grote absolute waarde in het geheel geen invloed meer, zodat de hefboom-punten geheel buiten het rekenproces blijven. Omdat bij deze methode de grootste efficiency-winst doorgaans al in de eerste stap bereikt wordt, is er daarna niet verder door geitereerd. Het aangepaste model wordt dan

$$\hat{y} = 5.095 + 1.062x_1 - 1.693x_2$$

en dat is een zeer bevredigend resultaat.

Een vergelijking van de residuen

Hieronder worden de residuen gegeven van de hierboven beschreven aanpassingen.

REF	LS	LAV	LAVHU	LMS	LMSTH	LMSHU
-0.04	-3.62	-6.36	-3.62	0.00	-0.04	-2.88
-0.12	1.25	1.21	1.27	-0.10	-0.12	1.00
0.11	0.24	-0.00	0.27	0.13	0.10	0.26
0.13	1.77	0.94	1.78	0.15	0.12	1.43
-10.87	-2.13	-0.00	-2.10	-10.88	-10.87	-3.93
-0.00	4.15	4.81	4.17	0.00	-0.00	3.31
0.02	-2.08	-4.07	-2.08	0.06	0.01	-1.64
-0.01	2.16	0.84	2.16	0.00	-0.02	1.68
-12.36	-2.78	-0.33	-2.75	-12.37	-12.36	-4.76
-0.08	1.06	0.01	1.07	-0.06	-0.09	0.82
0.07	0.05	-0.00	0.07	0.00	-0.02	0.09
0.02	-0.07	-1.87	-0.07	0.06	0.02	-0.06

In de tabel zijn de volgende afkortingen gebruikt: REF is de referentie-aanpassing. Dat is het resultaat van klassieke regressie na weglating van punt 5 en 9. LS staat voor Least Squares. Dit is klassieke regressie voor de volledige data-set. LAV is Least Absolute Values en staat voor een aanpassing die de som van de absolute waarden van de residuen minimaliseert. In LAVHU wordt dit proces gevolgd door een na-iteratie volgens Huber. LMS staat voor Least Median of Squares. In LMSTH volgt daar nog een variantie-verminderende stap op middels de op de tangens hyperbolicus gebaseerde invloedsfunktie. LMSHU staat voor Least Median of Squares gevolgd door na-iteratie volgens Huber.

REF, LMS en LMSTH leveren bevredigende resultaten. De overige methoden niet. Het is duidelijk dat de hefboom-punten buiten de aanpassing gehouden moeten worden, maar door de verstrengeling was de Cook-statistic niet de juiste weg om ze op te sporen.

De moderne robuuste statistiek biedt genoeg methoden om met een bescheiden aantal uitschieters in de y-richting om te gaan. Maar voor uitschieters in de prediktor-ruimte is er

nog niet veel. De Cook-statistic kan alleen maar een enkelvoudig geisoleerd punt opsporen.

Andere diagnostics

Belsley, Kuh en Welsch (1980) bieden een aantal alternatieve regressie-diagnostics. De eerste hiervan is een alternatief voor de gestudentiseerde residuen. Hierbi wordt de grootheid $s(i)$ gebruikt. Dat is een schatting voor σ die gebaseerd is op een aangepast model waarbij punt i buiten beschouwing is gelaten. We krijgen hier dus de formule:

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

De aanbevolen drempel in absolute waarde hierbij is 2. Deze grens wordt maar net gepasseerd bij het oninteressante punt nummer 6 met $s_6(6) = 2.041$.

Een andere maat is *DFFITS*. Deze is gedefinieerd als:

$$DFFITS_i = \frac{e_i\sqrt{h_{ii}}}{s(i)(1-h_{ii})}$$

In feite is dit het gestandaardiseerde verschil tussen $\hat{y}_i$ en $\hat{y}_i(i)$. Analoog aan de berekening van $s(i)$ wordt bij de laatstgenoemde aangepaste waarde de waarneming nummer i buiten beschouwing gelaten. Als kritieke waarde wordt $2\sqrt{p/n}$ aanbevolen. In dit geval is dat de waarde 1. De grens heeft betrekking op de absolute waarde en deze wordt overschreden voor de punten met nummer 1 en 9. *DFFITS* is hier -1.218 en -1.364 respectievelijk. Hier wordt dus een probleem-punt ontmaskerd en een ander genegeerd. Tevens wordt gewaarschuwd voor een punt waar niets mee aan de hand is.

Tenslotte is er nog de maat *DFBETAS*. Deze meet de verandering in de regressiecoefficienten ten gevolge van het weglaten van het punt i:

$$DFBETAS_{ij} = \frac{b_j - b_j(i)}{s(i)\sqrt{(X^TX)^{-1}_{jj}}}$$

De kritieke waarde is hierbij $2/\sqrt{n}$ en dat is hier 0.5773. Voor observatie nummer 1 wordt deze grens twee keer overschreden: $\delta\beta_0 = -1.144$ en $\delta\beta_1 = 0.9924$. Voor observatie nummer 5 vinden we $\delta\beta_1 = -0.5893$. En voor observatie nummer 9 worden er tenslotte twee gevonden: $\delta\beta_1 = -1.034$ en $\delta\beta_2 = -0.9567$. De punten 5 en 9 worden hier terecht als problemen geidentificeerd. Bij punt 1 is de alarmering misleidend.

De Cook-statistic kan uitgebreid worden om meerdere invoedrijke punten simultaan op te sporen. Zie hiervoor Cook en Weisberg (1982). Voor iedere tweetal, drietal, viertal en meer in het algemeen m-tal punten kan worden nagegaan wat de invloed is van het weglaten van deze punten op de schatting voor de regressie-coefficienten. De rekentijd die deze techniek vergt maakt hem echter voor praktijkproblemen al gauw onbruikbaar.

Slotopmerking

SAS, SPSS en BMDP bieden nu de Cook-statistic. De gebruikers zouden er meer mee gediend zijn als Least Median of Squares opgenomen zou worden, eventueel met een variantie-reducerende na-iteratie die grote gestandardiseerde residuen (in absolute waarde)

buiten beschouwing laat. Achteraf kan middels een plotje van de residuen tegen de aangepaste waarden of middels speciaal voor dit doel ontwikkelde statistics (zie Rousseeuw en Leroy 1987) worden onderzocht of het gebruik van deze robuuste methode noodzakelijk was. Zo nee, dan kan de gebruiker met een gerust hart de resultaten van de klassieke regressie interpreteren.

Literatuur

SAS/STAT Guide for Personal Computers - Version 6 Edition
SAS Institute Inc. Cary (USA) 1987

SPSSX User's Guide - A complete guide to SPSSX language and operations
SPSS Inc. and McGraw-Hill Book Company, New York 1983

BMDP Statistical Software Manual - 1985 Reprinting
University of California Press, Berkeley

Shapiro, SS and M.B. Wilk (1969) Approximations to the null-distribution of the W-statistic
Technometrics, Volume 10

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

Montgomery, D.C. and E.A. Peck (1982) Introduction to linear regression analysis
John Wiley & Sons, Inc.

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

Holland, P.W. and R.E. Welsch (1977) Robust regression using iteratively reweighted least-squares
Communications in Statistics (A 6-9) 813-827

Huber, P.J. (1973) Robust regression: asymptotics, conjectures and Monte Carlo
Annals of Statistics (1) 799-821

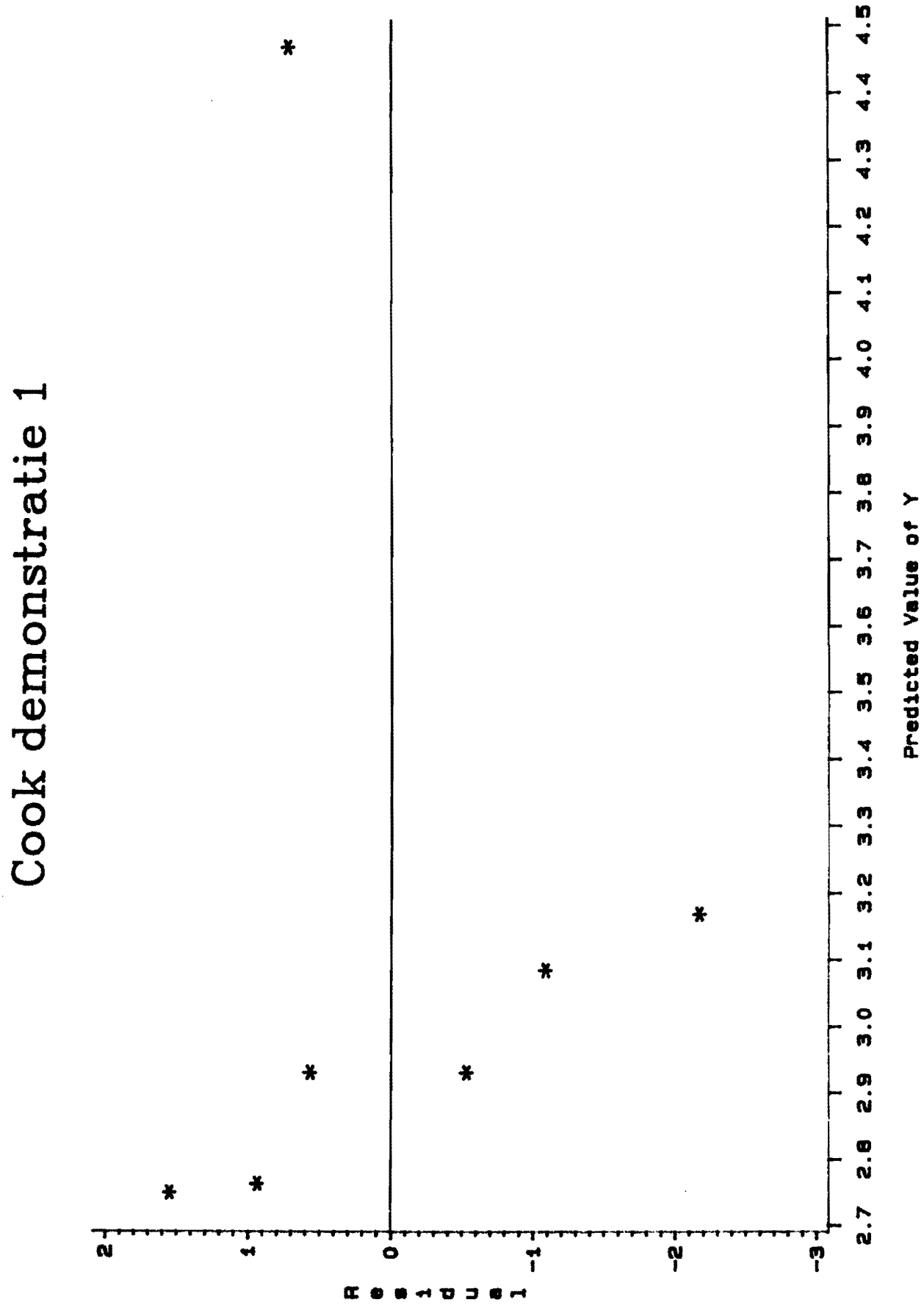
Rousseeuw, P.J. (1984) Least median of squares regression
Journal of the American Statistical Association (79) 871-880

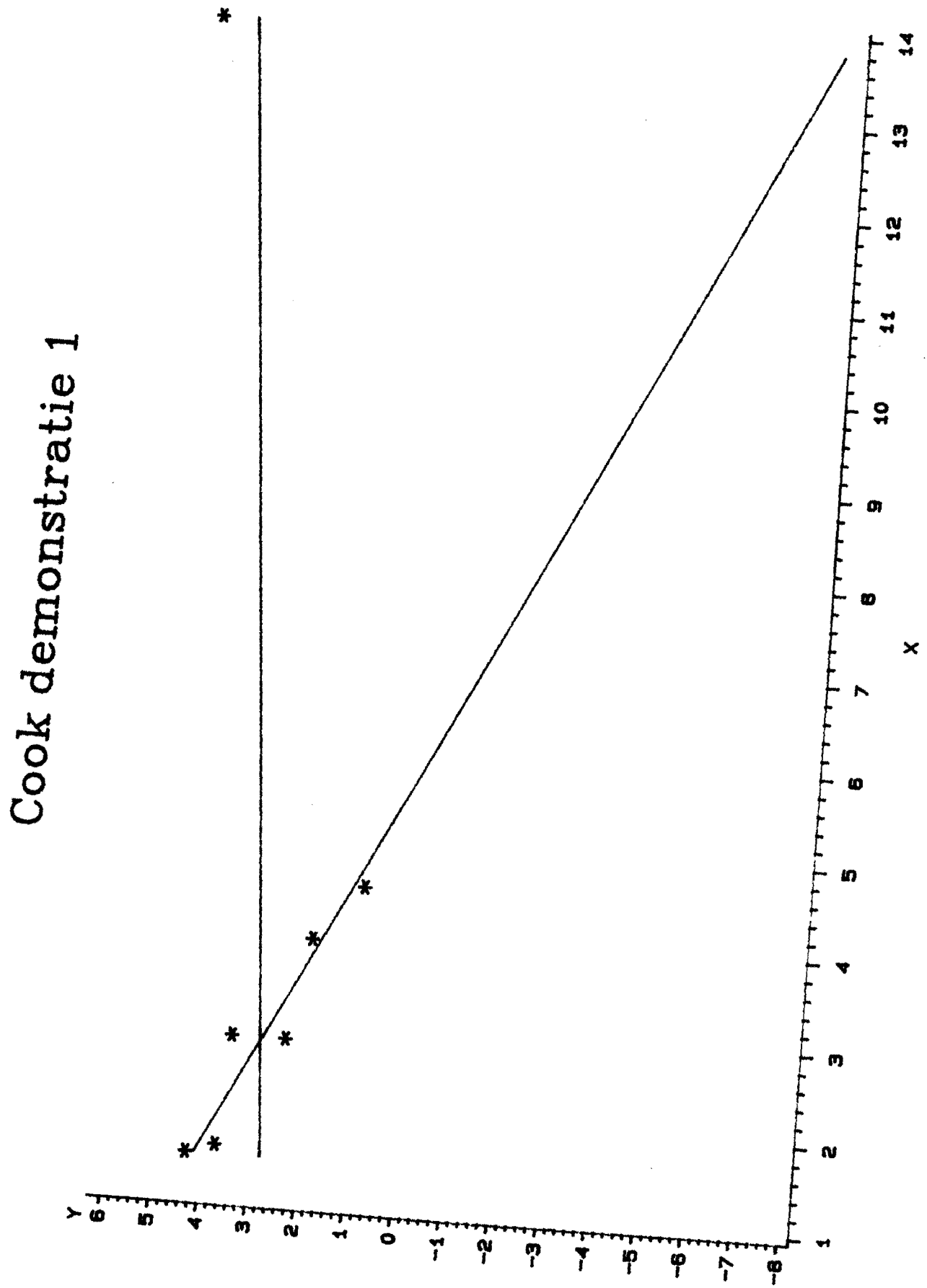
Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel (1986) Robust statistics: The approach based on influence functions
John Wiley & Sons, Inc. (New York)

Belsley, D.A., E. Kuh and R.E. Welsch (1980) Regression diagnostics
John Wiley & Sons, Inc. (New York)

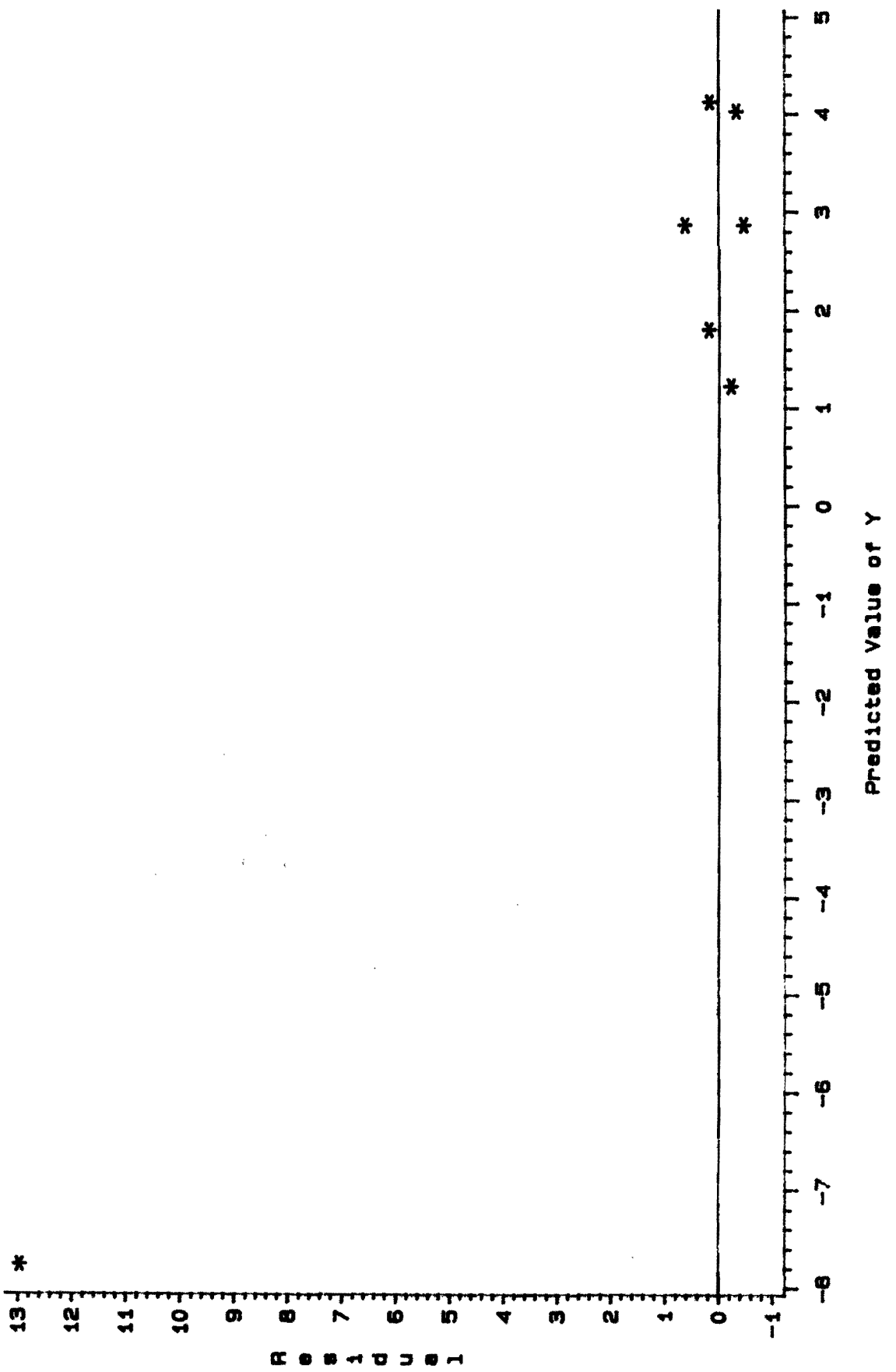
Cook, R.D. and S. Weisberg (1982) Residuals and influence in regression
Chapman & Hall, London

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

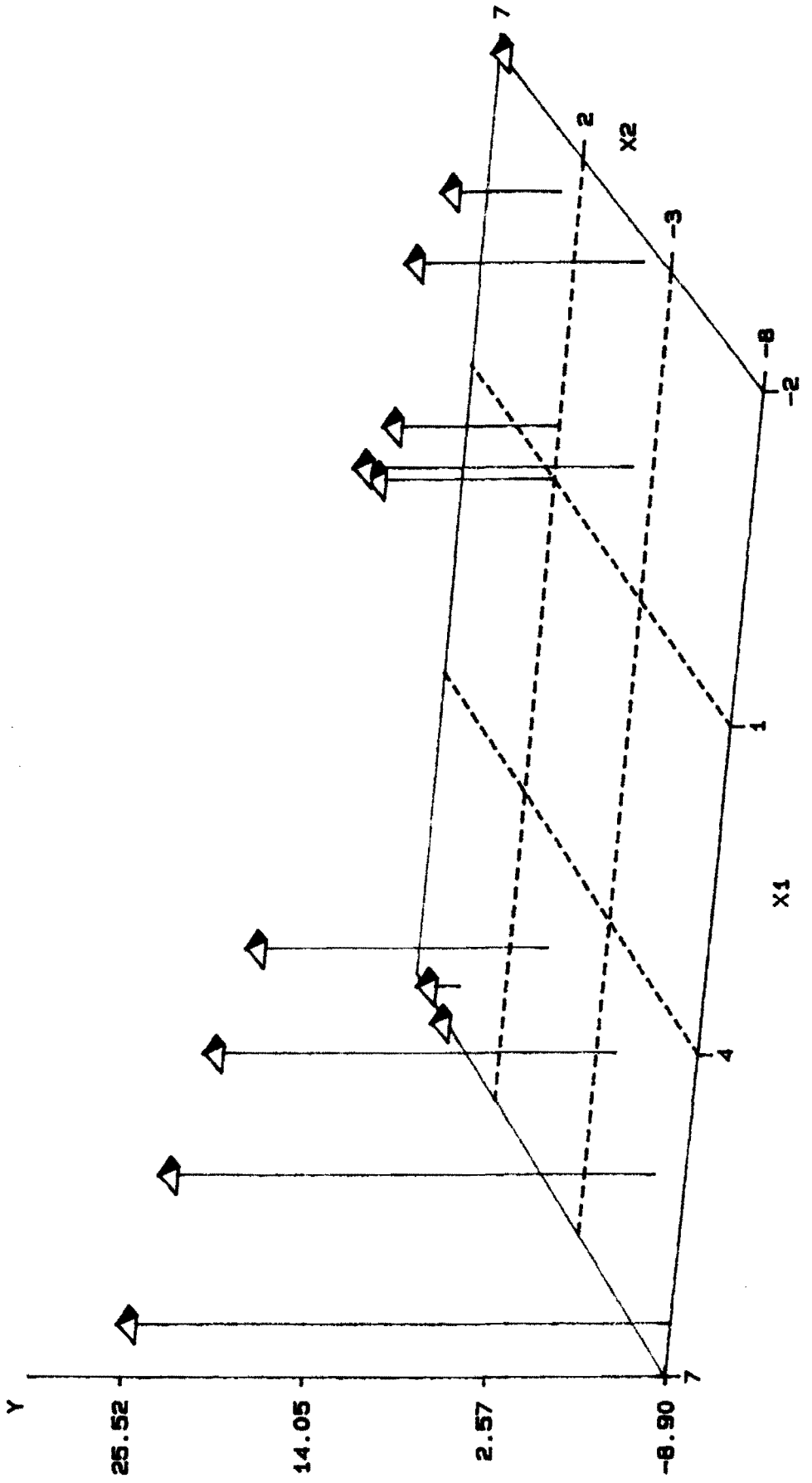

Cook demonstratie 1
Residuen
2
1
0
-1
-2
-3
2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4 4.5
Predicted Value of Y

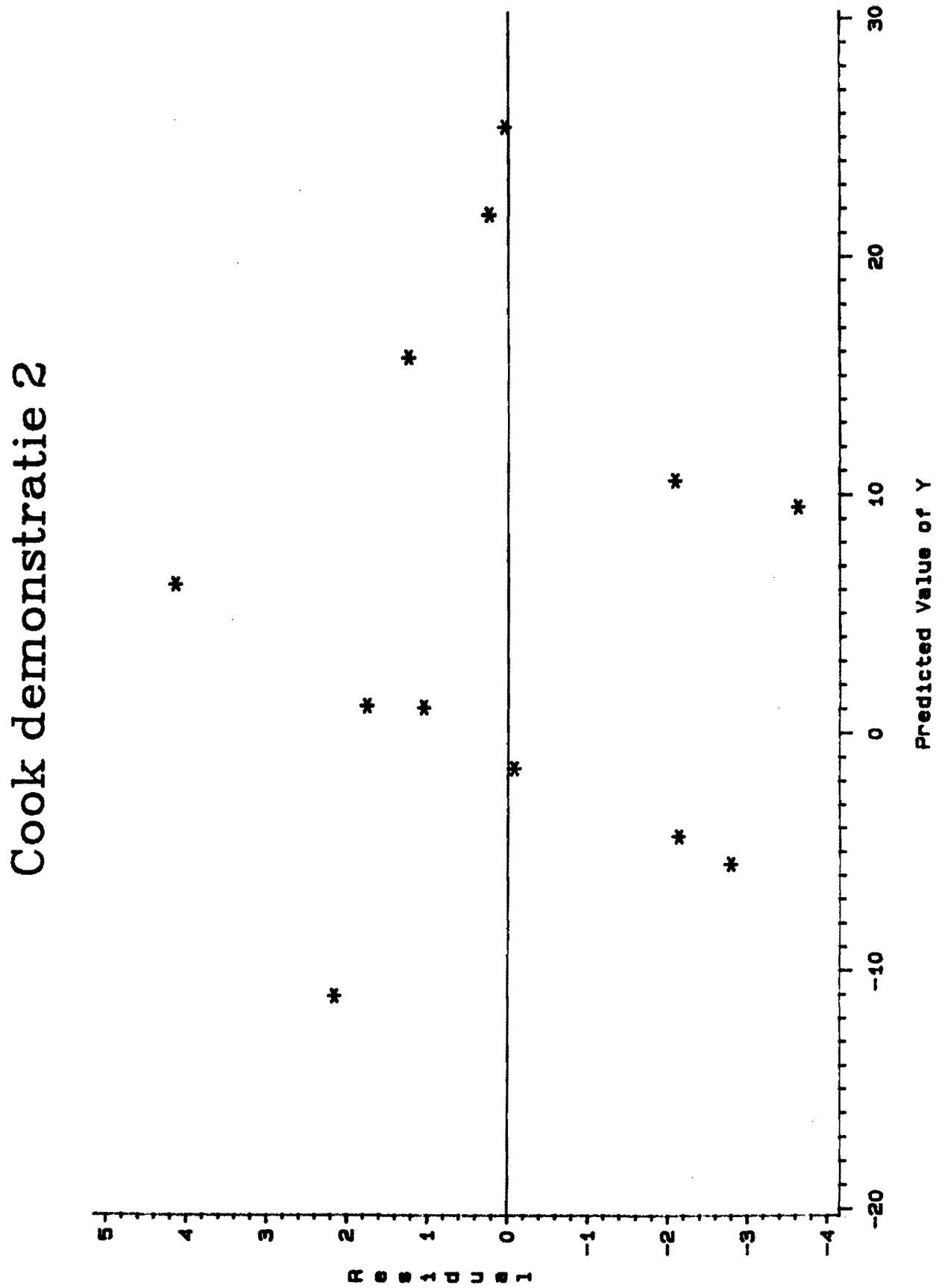

Cook demonstratie 1
Y
X
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
1
2
3
4
5
6
7
8
9
10
11
12
13
14

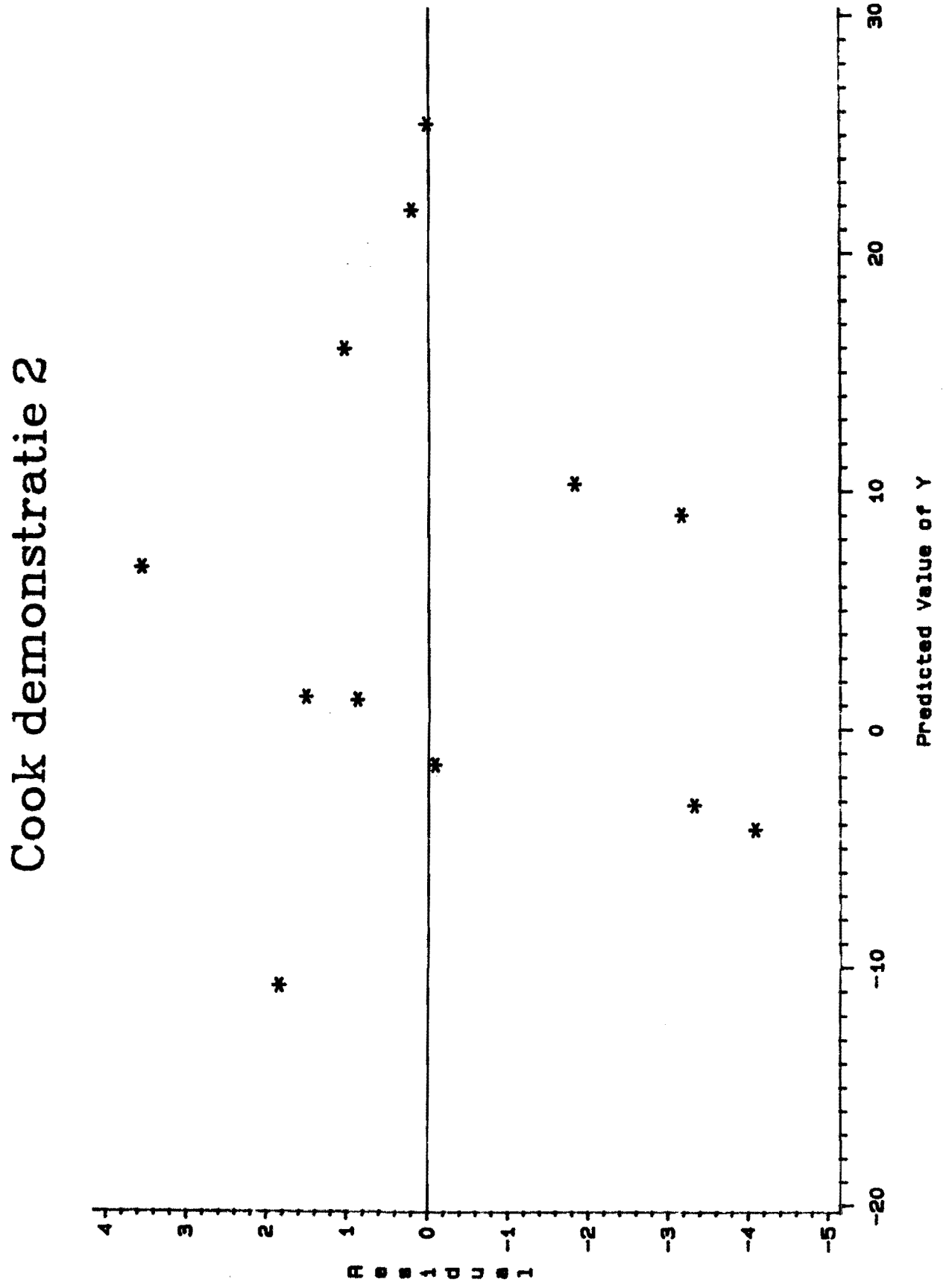
Cook demonstratie 1

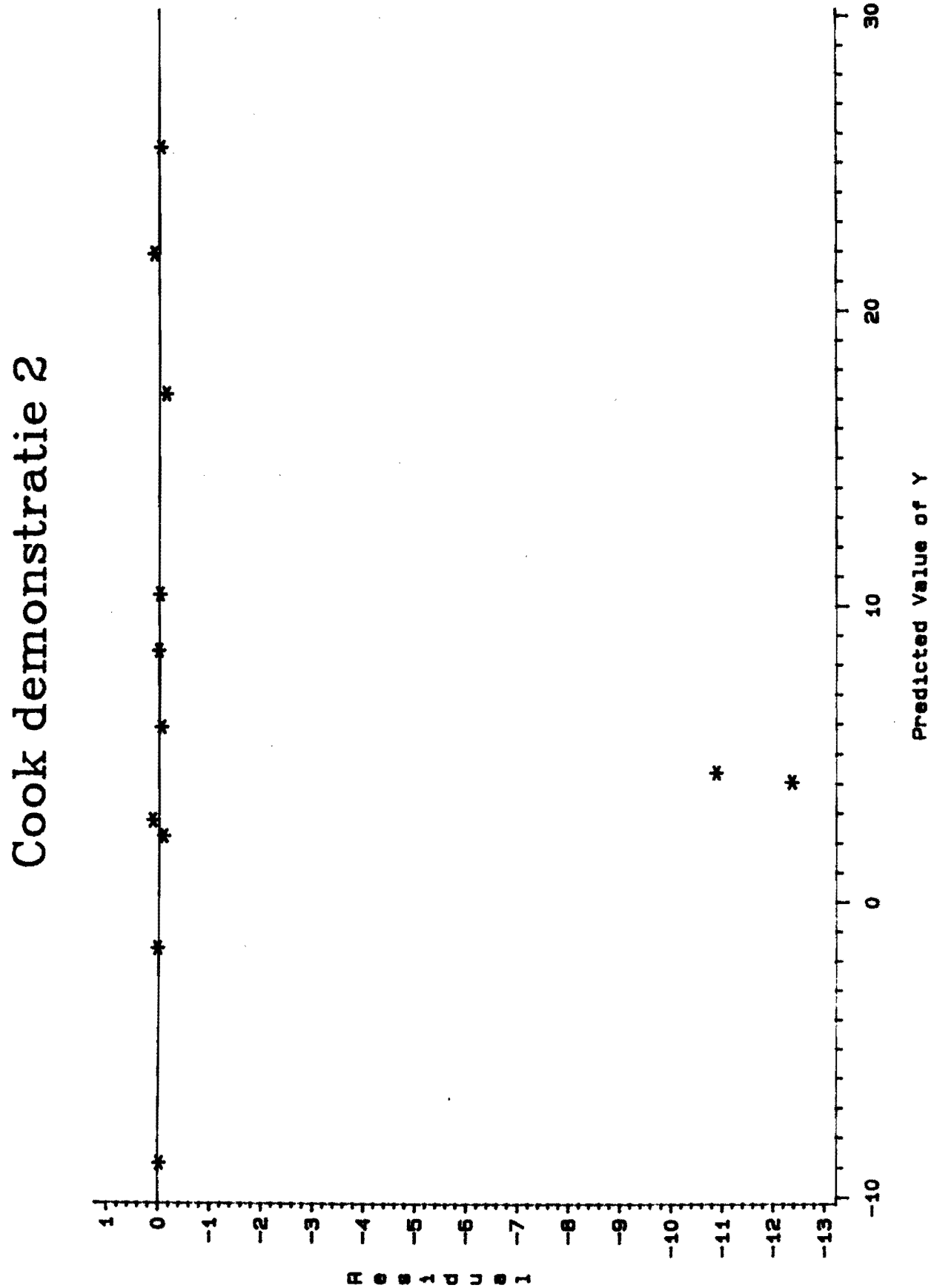


Cook demonstratie 2




Cook demonstratie 2
Residual
5
4
3
2
1
0
-1
-2
-3
-4
-20
-10
0
10
20
30
Predicted Value of Y


Cook demonstratie 2
Residuen
Predicted Value of Y
4
3
2
1
0
-1
-2
-3
-4
-5
-20
-10
0
10
20
30


Cook demonstratie 2
Residual
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-10
0
10
20
30
Predicted Value of Y