KM 32(1989)
pag 95 - 138

Regressie-diagnostiek in verschillende pakketten

Inhoud

1 Probleemstelling

1.1 Samenvatting

De auteur dezes had onlangs het genoegen een afstudeerder aan de TUE te begeleiden die een aantal robuuste regressie-methoden onderling vergeleek. Middels een aantal lastige datasets (Jansen, 1988) werd het kaf van het koren gescheiden en manifesteerde zich duidelijk de noodzaak om in combinatie met klassieke kleinste kwadraten regressie behoorlijke diagnostische hulpmiddelen te gebruiken. Hieronder volgt een dataset waar zo op het eerste gezicht niets mee aan de hand is. De bedrieglijke schijn zal worden toegelicht en een aantal invloedsmaten zal worden behandeld en getoetst aan dit voorbeeld. Vervolgens worden de diagnostische hulpmiddelen van verschillende pakketten vergeleken.

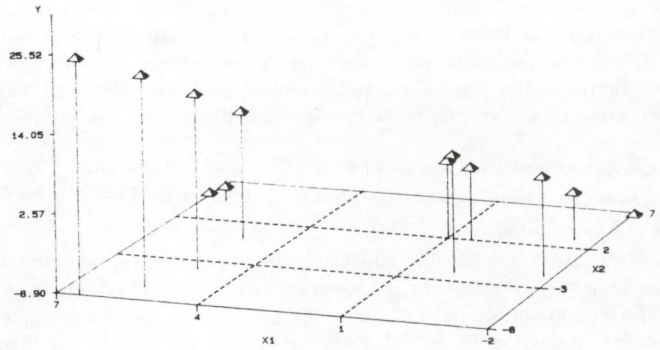
1.2 Inleiding

Het model bij regressie-analyse luidt $y = \beta_0 + \beta_1 x_1 + \ \ldots \ + \beta_k x_k + \varepsilon$. Van de afwijkingen ε_i veronderstellen we dat deze onafhankelijk normaal verdeeld zijn met verwachting 0 en constante variantie σ^2. De instellingen van de voorspellende variabelen zijn x_{ij}. Hierbij identificeert i de observaties en deze index loopt van 1 tot n. De prediktoren worden door j geïdentificeerd en hiervoor geldt $j = 1, \ \ldots \ , k$. Vanwege de intercept β_0 wordt het aantal parameters p gegeven door $p = k + 1$. De matrix X wordt gevormd door de elementen x_{ij}, voorafgegaan door een kolom enen. Hiermee wordt de hat-matrix $H = X(X^T X)^{-1} X^T$ gevormd waarmee de aangepaste waarden $\hat{y} = Hy$ berekend kunnen worden. Dit levert dan weer de residuen $e_i = y_i - \hat{y}_i$. De kwadratensom is $SSE = \sum_{i=1}^{n} e_i^2$. Delen door het aantal vrijheidsgraden $n - p$ levert de mean square error MSE en dat is een schatter voor σ^2. De standaard afwijking s wordt berekend middels $s = \sqrt{MSE}$. Hiermee kunnen de residuen worden gestandaardiseerd tot $d_i = e_i / s$. Dit is een eerste diagnostisch hulpmiddel: als een zekere d_i in absolute waarde groter is dan 2, dan rechtvaardigt dat een nadere inspectie van de bijbehorende waarneming.

1.3 Voorbeeld

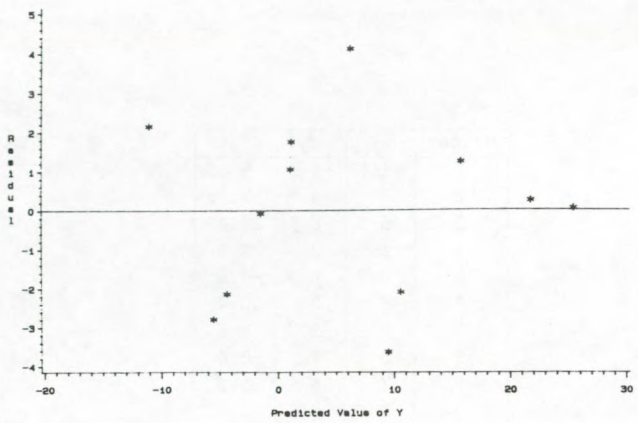
Zo op het eerste gezicht zien de data in deze tabel er niet bijzonder uit. Maar bestudering van figuur 1.1 brengt de onderzoeker tot andere gedachten: 10 punten liggen ongeveer in één vlak en 2 vallen daar duidelijk buiten. Dit zijn de punten met rangnummers 5 en 9. Problematisch is dat ze vlak bij elkaar liggen (weglating van de ene laat de invloed van de andere gewoon wat toenemen). Zo mogelijk nog problematischer is dat ze in het prediktorvlak opgespannen door x_1 en x_2 maar net buiten het kleinste convexe omhulsel voor de overige punten vallen (zodat de gebruikelijke afstandsmaten geen verontrustende resultaten zullen geven). We doen nu eerst even of onze neuzen collectief bloeden en passen het klassieke regressiemodel aan. Dat resulteert in $\hat{y} = 5.815 + 0.0862 x_1 - 2.387 x_2$ met een determinatiecoëfficiënt $R^2 = 0.9602$. De prediktor x_1 is niet significant; de overschrijdingskans is 0.7114, zodat de argeloze gebruiker het model wellicht opnieuw zou aanpassen

Nummer	x_1	x_2	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



figuur 1.1

figuur 1.1



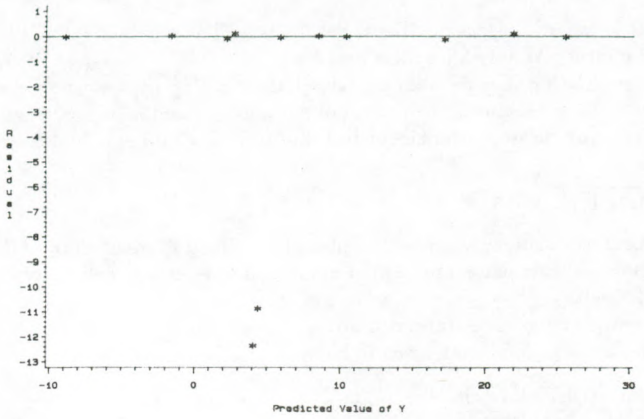
figuur 1.2

zonder x_1. De andere prediktor x_2 is wel significant: de overschrijdingskans is hier 0.0001. Onder de regressie-vooronderstellingen zijn e_i en $\hat{y}_i$ ongecorreleerd, zodat een plaatje waarin e_i tegen $\hat{y}_i$ wordt uitgezet een ongestructureerde zwerm punten zou moeten opleveren. In figuur 1.2 is hieraan redelijk voldaan. De normaliteit van de residuen kan onder andere worden onderzocht met de toets van Shapiro en Wilk (1965) zoals die in SAS is opgenomen. Dit levert een overschrijdingskans van 0.8249, zodat ook hier geen reden tot ongerustheid lijkt te zijn. Tenslotte zijn nog de gestandaardiseerde residuen berekend. Voor punt 6 wordt de grootste waarde van 1.64 bereikt en dat is veilig minder dan de kritieke waarde van 2.

Omdat het driedimensionale plaatje de punten 5 en 9 als verdacht aftekende is het model opnieuw aangepast na weglating van deze punten. Dit resulteerde in

$$\hat{y} = 5.093 + 1.062x_1 - 1.693x_2$$

met $R^2 = 0.9999$. Beide prediktoren zijn nu significant op een niveau van 0.0001 en de residuen tegen de voorspelde waarden geven een schitterend beeld zoals uit figuur 1.3 blijkt. De punten 5 en 9 geven nu bijzonder grote residuen, maar dat was te verwachten. Dit zijn immers bijzondere punten en in een fatsoenlijk onderzoek zouden ze ontmaskerd moeten worden. In dit driedimensionale geval lukt dat met een plaatje. Maar met meerdere prediktoren is zoiets niet meer zichtbaar te maken en dan ontstaat de behoefte aan formelere maten. Dit soort datasets komen in de praktijk werkelijk voor. Denk maar aan (1) Metingen aan wolven waarvan enkele van een afwijkend ras, (2) Scheepvaart door de Sont gedurende deze eeuw met enkele oorlogsjaren en (3) Oogmetingen waarbij sommige proefpersonen prenataal geconfronteerd zijn met Rode Hond.



figuur 1.3

1.4 Invloedsmaten

Bij standaardisering worden de residuen gedeeld door een schatter voor σ. Hierbij is rekening gehouden met de variantie van y_i, maar niet met de variantie van $\hat{y}_i$. Bij de gestudentiseerde residuen gebeurt dat wel. $\mathrm{Var}(\hat{y}_i) = \sigma^2 h_{ii}$, zodat $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$, waarbij h_{ii} een diagonaalelement van H voorstelt. De **gestudentiseerde residuen** worden nu gegeven door:

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

Indien de absolute waarde hiervan groter is dan 2, dan is nader onderzoek gewenst. De grootste waarde van 1.755 wordt hier bereikt voor punt 6 en deze waarde is niet alarmerend. De volgende maat is ontleend aan **Cook** (1977) en wordt doorgaans aangeduid als D_i. Indien voor zekere i deze maat groter is dan 1, dan zou weglating van punt i de vector van schatters voor de regressie-coëfficiënten buiten het simultane 50% betrouwbaarheidsgebied van de complete dataset plaatsen. Een punt met deze eigenschap is dus zeer invloedrijk. De formule voor de Cook-statistic luidt:

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

In dit voorbeeld is de grootste waarde 0.533 en deze wordt bereikt voor punt 9. Maar de waarde is te klein om verdere aktie te rechtvaardigen.
Met de diagonaal van de **hat-matrix** H kunnen punten worden opgespoord die in de prediktorruimte sterk excentrisch liggen. Er geldt namelijk $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$ en $|h_{ij}| \leq h_{ii}$ voor alle j. Dus h_{ii} is de mate waarin $\hat{y}_i$ bepaald wordt door y_i zelf. Grote h_{ii} wijst op (te) veel invloed van een enkele waarneming op de schatting. Belsley, Kuh en Welsch (1980)

bevelen als kritieke waarde $2p/n$ aan. Hier is dat 0.5 terwijl de grootste optredende waarde 0.4298 is (voor punt 9). Weinig bijzonders dus.
Een alternatieve schatting voor de standaardafwijking σ is $s(i)$, gebaseerd op alle waarnemingen behalve i. Deze notatie zal verderop ook bij andere grootheden gebruikt worden.
Een **alternatief voor de gestudentiseerde residuen** wordt nu gevormd door:

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

Ook hierbij wordt als kritieke waarde 2 aanbevolen. En die grens wordt hier zowaar gepasseerd. Helaas gebeurt dat echter bij het oninteressante punt 6 met de waarde 2.041. Een misleidend resultaat.
CR_i meet de **verandering in de determinant** van de covariantie-matrix van de regressiecoëfficiënten die ontstaat door punt i weg te laten.

$$CR_i = \frac{\det(s^2(i)[X(i)^T X(i)]^{-1})}{\det[s^2(X^T X)^{-1}]}$$

Ingrijpen is gewenst als de absolute waarde van $CR_i - 1$ groter is dan $3p/n$. In dit geval is de grens dus 0.75 en deze wordt overschreden voor punt 3 ($CR_3 = 1.962$), punt 11 ($CR_{11} = 2.314$) en punt 12 ($CR_{12} = 1.804$). Dit help ons geen stap verder omdat met deze punten niets aan de hand is. Een door Belsley, Kuh en Welsch voorgestelde maat is $DFFITS$. Deze is gedefinieerd als:

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

Dit is het **gestandaardiseerde verschil tussen** $\hat{y}_i$ en $\hat{y}_i(i)$. Als kritieke waarde wordt $2\sqrt{p/n}$ aanbevolen (in dit geval dus 1) voor de absolute waarde van $DFFITS$. Hier wordt voor de eerste observatie -1.218 bereikt en voor de negende -1.364. De overigen zijn in absolute waarde kleiner dan 1. Het eerste punt wordt hier ten onrechte als problematisch aangeduid. Bij punt 9 is de alarmering terecht, maar punt 5 is hierbij niet herkend als vrijwel net zo invloedrijk.
De laatste hier te behandelen maat heet $DFBETAS$. Deze meet de **verandering in de regressiecoëfficiënten** ten gevolge van het weglaten van punt i:

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

De kritieke waarde is hierbij $2/\sqrt{n}$ en dat is hier 0.5773. Voor observatie nummer 1 wordt deze grens twee keer overschreden: $\Delta\beta_{10} = -1.144$ en $\Delta\beta_{11} = 0.9924$. Voor observatie nummer 5 vinden we $\Delta\beta_{51} = -0.5893$. En voor observatie nummer 9 worden er tenslotte twee gevonden: $\Delta\beta_{91} = -1.034$ en $\Delta\beta_{92} = -0.9567$. De punten 5 en 9 worden hier terecht als problemen geïdentificeerd. Bij punt 1 is de alarmering misleidend.

1.5 Statistische software

Bij de hierboven beschreven dataset is het lastig om met eenvoudige diagnostische hulpmiddelen de juiste aanpassing te vinden. Met een plaatje in de driedimensionale ruimte valt het zo te zien; de statisticus die deze figuur inspecteert zal zijn aanpassing op tien punten baseren en aan de onderzoeker melden dat met twee punten wellicht iets bijzonders aan de hand is. Helaas hebben veel praktijkproblemen meer dan twee prediktoren zodat grafische representatie niet haalbaar is. Mede daarom is het interessant om na te gaan welke diagnostische hulpmiddelen in de gebruikelijke statistische software zijn geïmplementeerd. In de volgende paragrafen wordt beschreven hoe deze dataset met acht verschillende pakketten kan worden onderzocht. In een volgend nummer van KM volgt dan nog een vergelijkend overzicht met betrekking tot het diagnostisch gereedschap dat deze pakketten bij hun regressie-modules bieden.

1.6 Literatuur

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

Shapiro. SS em M.B. Wilk (1969) Approximations to the null-distribution
of the W-statistic
Technometrics, Volume 10, pag. 861-866

SAS/STAT Guide for personal computers - Version 6 edition
SAS Institute Inc. Cary (USA) 1987

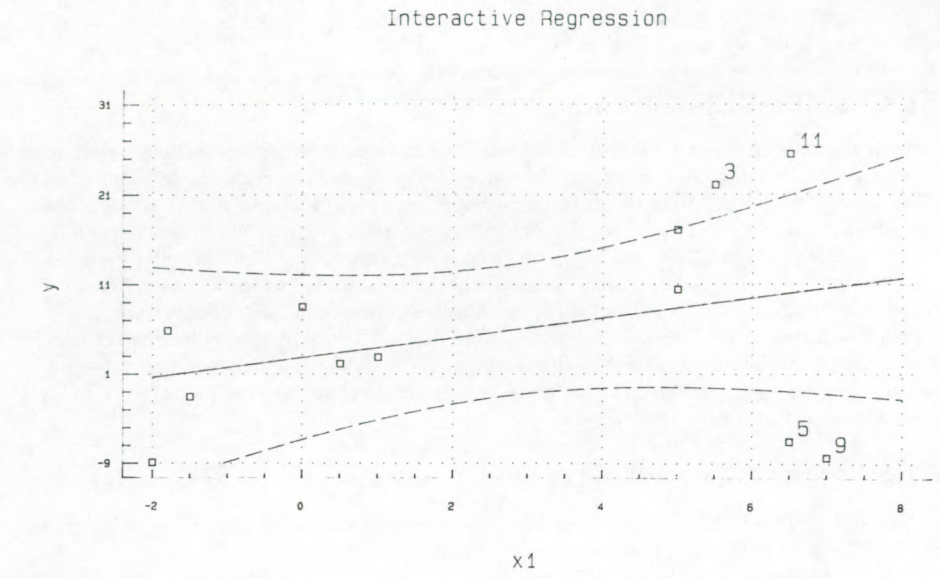
Cook, R.D. (1977) Detection of influential observations in linear regression
Technometrics, Volume 19

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

2 STATGRAPHICS 2.6

2.1 Inleiding

Door velen wordt STATGRAPHICS beschouwd als een compleet pakket, gebruiksvriendelijk en gemakkelijk te leren. De kracht van het pakket zit in de vele grafische mogelijkheden ter ondersteuning van de statistische analyses. Plots kunnen interaktief worden bijgewerkt. De statistische mogelijkheden zijn echter niet echt diepgaand. Voor een meer uitgebreide beschrijving en evaluatie van de mogelijkheden van STATGRAPHICS verwijzen we naar Snijkers & Debets (1989).
STATGRAPHICS heeft verschillende mogelijkheden voor regressie analyse: enkelvoudige regressie, interactive outlier analyse, multiple regressie, stapsgewijze regressie, ridge regressie en niet-lineaire regressie analyse.



figuur 2.1

2.2 Analyse van de dataset

Met behulp van de interactieve outlier analyse kan men voor elke x-variabele afzonderlijk, uitschieters in de data opsporen. In de plot van y tegen x (zie figuur 2.1) zijn aangegeven de regressielijn en het 95% betrouwbaarheidsinterval voor het gemiddelde van y, gegeven een bepaalde x-waarde. Bij een plot van y tegen x1 vallen de punten 3, 11, 5 en 9 buiten het interval. De fit in termen van de correlatie tussen y en x1 bedraagt 0.343. Weglating van de punten 3 en 11 heeft een verlaging van de fit ten gevolge ($r = 0.029$) terwijl weglating van de punten 5 en 9 de fit enorm verbetert ($r = 0.887$). In de plot van y tegen x2 vallen de punten 1, 6 en 7 buiten het interval maar levert weglating van een der punten nauwelijks nog een verbetering van de oorspronkelijke correlatie (-0.980).
De uitvoer van de multipele regressie bestaat in eerste instantie uit de basisgegevens: coefficiënten, standaard errors, multipele correlatie, enz. Van de andere resultaten zijn t.b.v. regressiediagnostiek van belang:

- plots van de residuen tegen de voorspelde waarde y, de index van de observaties of een andere variabele. Het plaatje van de residuen tegen voorspelde waarde geeft zoals eerder gezien geen reden tot ongerustheid.

```
                    Residual Summary
-----------------------------------------------------------------
Number of observations = 12    (0 missing values excluded)
Residual average  = -2.34997E-15
Residual variance = 6.41199
Residual standard error = 2.53219
Coeff.  of skewness = 0.0217282     standardized value = 0.0307283
Coeff.  of kurtosis = -0.559969     standardized value = -0.395958

Durbin-Watson statistic = 3.10967
```

figuur 2.2

- informatie over de verdeling van de residuen (zie figuur 2.2).
  Naast gemiddelde, variantie en standaard error krijgt men de kurtosis en de skewness en de gestandaardiseerde waarden hiervan als een soort van toets op afwijking van normaliteit.
  De Durbin-Watson statistic is hier een toets op de auto-correlatie tussen de residuen.

- de plot van de predicted versus de observed y waarden geeft mogelijkerwijs aanwijzingen voor het fitten van een ander (niet-lineair) model (hier niet van toepassing).

- de normal probability plot van de residuen als een aanvulling op de normaliteitstoetsen. In deze plot lijken er overeenkomstig eerdere opmerkingen geen serieuze problemen in deze dataset aanwezig te zijn. Als bijkomstigheid wordt hier opgemerkt dat STATGRAPHICS de punten foutief identificeert.

- en uiteraard de influence measures. Hier krijgt men een overzicht van de observaties die als invloedrijk zijn gedefinieerd met behulp van een van de criteria
  leverage $> 2p/n$
  DFITS $> 2\sqrt{p/n}$.

STATGRAPHICS geeft hier de melding:
"Number of flagged observations (high leverage or DFITS) = 0".
Zoals eerder gezien is er volgens de leverage maat niets aan de hand, maar zouden volgens het DFITS criterium de punten 1 en 9 als problematisch aangeduid moeten worden. STATGRAPHICS ontdekt beide punten niet. Helaas krijgt men dan ook niet de maten voor de afzonderlijke punten te zien. Dit moet men via een omweg doen door de resultaten van de regressieanalyse te bewaren. De voorspelde waarde, het residu, het gestandaardiseerde residu, de leverage waarde, de DFITS waarde en de Mahalanobis afstand voor alle observaties kunnen als variabelen bewaard worden. Uiteraard kan men ze dan ook printen (zie figuur 2.3). Het blijkt dan dat de DFITS waarden van de punten 1 en 9 inderdaad groter zijn dan het genoemde criterium. Het probleem hier is dat STATGRAPHICS enkel toetst op positieve afwijkingen, immers als men de analyse uitvoert op de variabele -y dan ontdekt STATGRAPHICS deze punten wel. Dit probleem kwam ook al naar voren in versie

Print van de bewaarde resultaten

row	OBSERVED	FITTED	RESIDS	SRESIDS	LEVERAGE	DFITS	MAHALD
1	5.85	9.478	-3.628	-1.934	0.284	-1.218	3.059
2	17.05	15.793	1.257	0.522	0.168	0.235	1.111
3	22.05	21.802	0.248	0.108	0.277	0.067	2.931
4	2.90	1.128	1.772	0.721	0.109	0.252	0.315
5	-6.50	-4.365	-2.135	-1.060	0.358	-0.792	4.673
6	10.40	6.246	4.154	2.041	0.127	0.777	0.539
7	8.50	10.588	-2.088	-0.902	0.182	-0.425	1.313
8	-8.90	-11.064	2.164	1.067	0.349	0.781	4.448
9	-8.30	-5.515	-2.785	-1.571	0.430	-1.364	6.629
10	2.15	1.085	1.065	0.428	0.120	0.158	0.459
11	25.52	25.469	0.051	0.024	0.385	0.019	5.343
12	-1.55	-1.474	-0.076	-0.032	0.211	-0.016	1.767

figuur 2.3

2.1 (Dijkstra, 1988).
De Mahalanobis distance is de afstand van een observatie ten opzichte van het gemiddelde voor de onafhankelijke variabelen van alle observaties en geeft aldus een indicatie of een case een uitschieter is. De waarde kan vergeleken worden met de waarden van de χ^2-verdeling met $p-1$ vrijheidsgraden. De kritieke grens voor de MD bedraagt in dit geval 5.99 volgens welk criterium punt 9 als verdacht aangewezen wordt.
Deze maat hanterend wordt in een volgende analyse waarbij punt 9 buiten beschouwing is gelaten punt 5 als verdacht aangewezen. Wat betreft de gestandaardiseerde residuen moet hier nog opgemerkt worden dat hier sprake is van de gestudentiseerde residuen waarbij met de waarneming zelf geen rekening is gehouden. Volgens de kritieke waarde van 2 wordt punt 6 aangewezen. De gestandaardiseerde residuen kan men wel in een regressiemodule opvragen samen met de geobserveerde en voorspelde waarde. Bij deze uitvoer hanteert STATGRAPHICS echter 3 als kritieke waarde.

2.3 Conclusie

Een interessante optie in STATGRAPHICS is de interactieve outlier analyse. Hoewel in eerste instantie bedoeld voor enkelvoudige regressie kan men met behulp hiervan op het spoor komen van bijzondere observaties zoals in dit geval de punten 5 en 9. Men moet er dan wel naar op zoek zijn.
Want ondanks de vele plaatjes en de implementatie van een aantal regressie diagnostiek maten ontdekt men de bewuste punten niet met behulp van de uitvoer van het multipele regressiemodule. Het feit dat het DFITS criterium alleen op positieve waarden werkt en de omweg die men moet maken om die maten te zien vergemakkelijkt e.e.a. beslist niet. Met behulp van de Mahalanobis distance had men immers de bewuste punten wel kunnen

opsporen; deze maat krijgt men in de uitvoer van het regressiemodule echter alleen als een punt al als uitschieter is aangewezen.
Bovendien hanteert men in afwijking van de literatuur 3 als kritieke grens voor de gestudentiseerde residuen.
Als eindkonklusie zou ik willen geven:
In STATGRAPHICS zijn de juiste middelen aanwezig voor regressiediagnostiek. De criteria worden echter niet goed gehanteerd. Een goede uitleg en keuze en definitie van de criteria overlaten aan de gebruiker zou beter geweest zijn.

2.4 Literatuur

Snijkers, G.J.M.E. & Debets, P.
STATGRAPHICS 2.6, Statistisch of grafisch pakket?
Kwantitatieve Methoden 30 (1989), pag. 129-143

3 Analyse met SAS

3.1 Inleiding

SAS kan worden beschouwd als het kanon onder de statistische pakketten. Het pakket biedt uitgebreide mogelijkheden voor elementaire (SAS/BASE), grafische (SAS/GRAPH) en statistische (SAS/STAT) analyses op mainframes (IBM), mini's (VAX), workstations (SUN en HP) em micro's (IBM compatibles). Een groot voordeel daarbij is dat de stuurtaal voor al deze soorten computers gelijk is. In deze bijdrage wordt de betreffende dataset geanalyseerd met de standaardtools welke SAS biedt. Omdat SAS een programmeertaal is, met mogelijkheden voor matrixbewerkingen (SAS/IML), zou een veel uitgebreider analyse zoals in GLIM mogelijk zijn. In de dagelijkse praktijk is er echter geen tijd voor dit soort uitgebreide analyses.

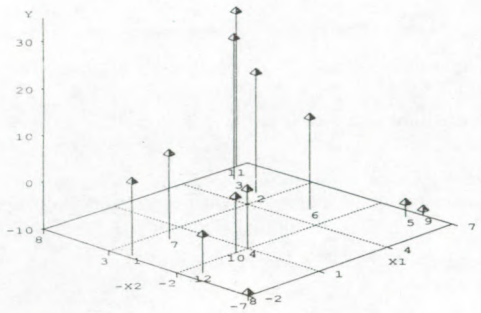
De 'standaard-analyse ' verloopt in drie stappen:

1. Het XYZ-plaatje vanuit een goede hoek bekeken
2. Zoeken in de SAS/STAT manual wat SAS te bieden heeft wat betreft invloedsmaten
3. Uitvoeren van deze analyse

3.2 Plotten met SAS/GRAPH

Voor het maken van een driedimensionaal scatterdiagram zijn de volgende SAS-statements voldoende

```
goptions device=beeldscherm, plotter of grafische printer;
proc g3d;
 scatter x*y=z;
run;
```



figuur 3.1

Het is nog niet mogelijk om in SAS interaktief de puntenwolk te bekijken, doch SAS heeft inmiddels de Canadese firma Neo-visuals Inc. opgekocht om deze technieken in huis te halen. Voor een geroutineerd SAS-gebruiker is het echter niet al te moeilijk om het plaatje te roteren en er labels in te hangen. Met een eenvoudig programma is de gezichtshoek veranderd en worden de labels van de punten in het grondvlak weergegeven (zie figuur 3.1).

3.3 De SAS documentatie

Uitschieters in de zin van ver weg liggende punten zijn in een regressieprobleem meestal niet met residuen-analyse op te sporen, omdat het residu van een ver verwijderd punt gewoonlijk klein is. Met residuen-analyse wordt dit punt alleen als uitschieter opgemerkt door een systematisch verloop in de residuen van de overige punten. Daarom is alleen gekeken naar de eerder besproken invloedsmaten. De SAS/STAT manuals geven onder de procedure PROC REG een vrij volledige en up-to-date opsomming van invloedsmaten, inclusief richtlijnen voor de grootte van de overschrijdingsgrenzen en eventueel gecorrigeerd voor het aantal punten dat in de regressie is meegenomen. Hier valt weinig aan toe te voegen en SAS verdient een pluim op dit punt. Opvallend is dat Cook's D niet als standaarduitvoer verschijnt, maar wel kan worden opgevraagd en weggeschreven naar een nieuwe dataset.

3.4 Uitvoeren van de analyse met SAS

Hieronder wordt de uitvoer gegeven bij de analyse met invloedsmaten van de gegeven dataset (zie figuur 3.2, 3.3 en 3.4). Allereerst wordt de standaard-analyse wat betreft invloedsmaten uitgevoerd (via de parameter `influence` in de procedure PROC REG). In de uitvoer zijn de getallen welke de 95% betrouwbaarheidsgrenzen (gecorrigeerd voor het aantal punten) overschrijden met een tekstverwerker vetgemaakt en onderstreept. Alleen de invloedsmaat DFFITS merkt punt 9 op als de grootste uitschieter. Indien we dit punt in de volgende stap weglaten, wijzen alle invloedsmaten vervolgens punt 5 als uitschieter

SAS 15:34 Wednesday, January 18, 1989 3

Obs	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	X1 Dfbetas	X2 Dfbetas
1	-3.6284	-1.9343	0.2841	0.6290	-1.2185	-1.1454	0.9924	0.5626
2	1.2573	0.5219	0.1681	1.5479	0.2346	0.0873	0.0679	-0.1217
3	0.2475	0.1085	0.2774	1.9618	0.0672	0.0208	0.0138	-0.0471
4	1.7723	0.7215	0.1090	1.3227	0.2524	0.2191	-0.0799	0.0619
5	-2.1351	-1.0597	0.3583	1.4961	-0.7918	0.0844	-0.5839	-0.5441
6	4.1541	2.0407	0.1265	0.4637	0.7766	0.1981	0.4534	0.1320
7	-2.0883	-0.9023	0.1818	1.3012	-0.4254	-0.4041	0.2818	0.2206
8	2.1643	1.0674	0.3488	1.4666	0.7812	0.5021	-0.3449	0.4435
9	-2.7849	-1.5707	0.4298	1.1149	-1.3638	0.2078	-1.0340	-0.9567
10	1.0654	0.4277	0.1203	1.5124	0.1582	0.1437	-0.0680	0.0303
11	0.0513	0.0243	0.3847	2.3135	0.0192	0.0040	0.0051	-0.0137
12	-0.0755	-0.0317	0.2111	1.8041	-0.0164	-0.0147	0.0109	-0.0027

Sum of Residuals	6.661338E-16
Sum of Squared Residuals	57.7079
Predicted Resid SS (Press)	108.6320

figuur 3.2

aan. Na het weglaten van punt 5 is het opmerkelijk dat de invloedsmaat COVRATIO onder de overgebleven 10 punten 6 uitschieters opmerkt. Deze maat is waarschijnlijk volstrekt ongeschikt om uitschieters op te sporen. Kortom: er is één weg welke leidt naar Rome, namelijk via de invloedsmaat DFFITS. Dit is op zich niet zo verwonderlijk, aangezien zich onder de invloedsmaten geen enkele maat bevindt die geschikt is voor het opsporen van simultane uitschieters, zoals bij dit probleem het geval is.

3.5 Conclusie

+ goede grafische analyse-mogelijkheden
+ alle standaard-invloedsmaten zitten in SAS
+ goede documentatie wat betreft invloedsmaten, met up-to-date referenties en richtlijnen voor de grootte van overschrijdingsgrenzen, eventueel gecorrigeerd voor het aantal punten
+ zowel op de PC als het mainframe te draaien met dezelfde stuurkaarten

- Cook's D niet standaard als uitvoer bij influence-parameter (wel op te geven als uitvoer)
- geen invloedsmaten voor simultane uitschieters (zodat voor dit probleem slechts de maat DFFITS tot de juiste oplossing leidt).

Totaal: SAS doet het prima voor een algemeen statistisch pakket.

SAS 15:34 Wednesday, January 18, 1989 4

Model: MODEL2 (11 punten, punt 9 eruit)

SAS 15:34 Wednesday, January 18, 1989 5

Obs	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	X1 Dfbetas	X2 Dfbetas
1	-2.8108	-1.6035	0.3332	0.8757	-1.1336	-1.0325	0.9514	0.6392
2	0.9714	0.4314	0.1741	1.6705	0.1981	0.0669	0.0673	-0.0671
3	0.2368	0.1111	0.2774	2.0550	0.0688	0.0210	0.0120	-0.0410
4	1.4317	0.6238	0.1176	1.4382	0.2277	0.1807	-0.0242	0.0778
5	-4.0496	-73.8137	0.6277	0.0000	-95.8455	15.8909	-79.1294	-75.1590
6	3.2545	1.7115	0.1860	0.6425	0.8181	0.1099	0.5834	0.3387
7	-1.5989	-0.7392	0.1994	1.4882	-0.3689	-0.3462	0.2552	0.2131
8	1.7061	0.9005	0.3643	1.6901	0.6816	0.4066	-0.1688	0.3964
9	0.8330	0.3579	0.1243	1.6144	0.1348	0.1163	-0.0344	0.0342
10	0.0542	0.0275	0.3847	2.4252	0.0218	0.0045	0.0048	-0.0133
11	-0.0284	-0.0128	0.2112	1.8924	-0.0066	-0.0059	0.0038	-0.0008

Sum of Residuals	-4.66294E-15
Sum of Squared Residuals	44.1058
Predicted Resid SS (Press)	168.3017

figuur 3.3

SAS 15:34 Wednesday, January 18, 1989 6

Model: MODEL3 (10 punten, punten 5 en 9 eruit)

SAS 15:34 Wednesday, January 18, 1989 7

Obs	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	X1 Dfbetas	X2 Dfbetas
1	-0.0400	-0.6048	0.5075	2.6999	-0.6139	-0.5187	0.5479	0.4532
2	-0.1248	-1.7756	0.2014	0.5601	-0.8916	-0.2048	-0.4321	-0.0374
3	0.1118	1.6251	0.2778	0.7361	1.0079	0.2933	0.1458	-0.3944
4	0.1311	1.8369	0.1560	0.4933	0.7896	0.4494	0.2367	0.4437
5	-0.00284	-0.0387	0.4269	2.7686	-0.0334	0.0024	-0.0291	-0.0243
6	0.0212	0.2548	0.2590	2.0750	0.1507	0.1365	-0.1152	-0.1051
7	-0.0180	-0.2466	0.4318	2.7112	-0.2149	-0.0973	-0.0288	-0.1411
8	-0.0879	-1.0664	0.1436	1.1016	-0.4366	-0.3089	-0.0463	-0.1864
9	-0.0197	-0.2602	0.3848	2.4959	-0.2058	-0.0408	-0.0335	0.0853
10	0.0291	0.3407	0.2113	1.9010	0.1764	0.1549	-0.0710	0.0134

Sum of Residuals	-7.54952E-15
Sum of Squared Residuals	0.0566
Predicted Resid SS (Press)	0.0939

figuur 3.4

4 Genstat

4.1 Inleiding

Voor wie slechts af en toe iets van Genstat onder ogen krijgt: de *syntax* van de commandotaal is met ingang van versie 5 zeer drastisch gewijzigd. Tegelijk zijn de *mogelijkheden* hier en daar wat uitgebreid. Zie R.W. Payne e.a. (1987). *Genstat 5 Reference Manual.* Clarendon Press, Oxford.
Met de opdracht FIT kunnen zowel lineaire als gegeneraliseerde lineaire modellen (niet te verwarren met GLM van SAS!) worden aangepast. FIT drukt op verzoek alle $\hat{y}_i$ en f_i af (notatie inleiding). FIT drukt los daarvan automatisch een waarschuwing af voor elke waarneming met grote f_i of h_{ii}. De drempels zijn c voor f_i en cp/n voor h_{ii}. De waarde van c is afhankelijk van het aantal vrijheidsgraden $n-p$, op zo'n manier dat het verwachte aantal meldingen bij fatsoenlijke gegevens ongeveer 1 is:

$$c = \begin{cases} 2.0, & n-p > 20 \\ 4.0, & n-p \text{ zeer groot} \\ -\Phi^{-1}(0.5/(n-p)), & \text{andere gevallen.} \end{cases}$$

De manual verwijst naar R.D. Cook and S. Weisberg(1982). *Residuals and influence in regression.* Chapman and Hall, New York.
Met de opdracht RKEEP kunnen de $\hat{y}_i$, f_i en h_{ii} worden bewaard in nieuwe variabelen. Daarna kan er mee gerekend worden en kunnen grafieken worden geproduceerd met de opdracht GRAPH.
Na het een en ander te hebben uitgeprobeerd hebben we alle benodigde opdrachten ondergebracht in een doorlopende programmafile die in de uitvoer is afgedrukt op de doorlopend genummerde regels.

4.2 Eerste stap

Fitten van het model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, en beoordeling van de f_i (door Genstat *standardized* residuals genoemd) en de h_{ii} (*leverage*). Geen waarschuwingen of andere bijzonderheden.

```
Genstat 5 Release 1.2 (VAX/VMS4)
Copyright 1987, Lawes Agricultural Trust (Rothamsted Experimental Station)
  1 JOB [OUTPRINT=*] regression_diagnostics
  2 UNITS [NVALUES=12]
  3 FACTOR [LEVELS=12; VALUES=1...12] veld
  4 SCALAR n,p; 12,3
  5 OPEN CHANNEL=2; NAME='kreng.geg'; FILETYPE=input; ACCES=readonly
  6 READ [CHANNEL=2; PRINT=*] x1,x2,y
  7 MODEL Y=y;
  8 FIT [PRINT=m,s,e,f] x1+x2
```

```
** Summary of analysis ***

             d.f.     s.s      m.s.
Regression      2  1392.18  696.089
Residual        9    57.71    6.412
Total          11  1449.89  131.808

Percentage variance accounted for 95.1

** Estimates of regression coefficients ***

          estimate   s.e.      t
Constant     5.815  0.948   6.14
x1           0.086  0.226   0.38
x2          -2.387  0.173 -13.80

** Fitted value and residuals ***

                                        Standardized
      Unit  Response  Fitted value      residual     Leverage
         1      5.85          9.48         -1.69         0.28
         2     17.05         15.79          0.54         0.17
         3     22.05         21.80          0.11         0.28
         4      2.90          1.13          0.74         0.11
         5     -6.50         -4.36         -1.05         0.36
         6     10.40          6.25          1.76         0.13
         7      8.50         10.59         -0.91         0.18
         8     -8.90        -11.06          1.06         0.35
         9     -8.30         -5.52         -1.46         0.43
        10      2.15          1.08          0.45         0.12
        11     25.52         25.47          0.03         0.38
        12     -1.55         -1.47         -0.03         0.21
Mean            5.76          5.76         -0.04         0.25
```

De h_{ii} kunnen slechts via RKEEP opgevraagd worden.

```
9  RKEEP RESID=f; FITTED=pred; LEVERAGE=h; DEVIANCE=sse; DF=dfe
```

```
 10 PRINT veld,x1,x2,y,pred,f,h; F=9; D=0,1,1,2,3,3,3
```

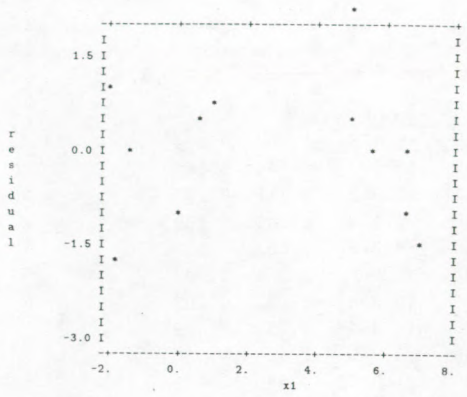
veld	x1	x2	y	pred	f	h
1	-1.8	-1.6	5.85	9.478	-1.694	0.284
2	5.0	-4.0	17.05	15.793	0.544	0.168
3	5.5	-6.5	22.05	21.802	0.115	0.277
4	1.0	2.0	2.90	1.128	0.741	0.109
5	6.5	4.5	-6.50	-4.365	-1.053	0.358
6	5.0	0.0	10.40	6.246	1.755	0.127
7	0.0	-2.0	8.50	10.588	-0.912	0.182
8	-2.0	7.0	-8.90	-11.064	1.059	0.349
9	7.0	5.0	-8.30	-5.515	-1.456	0.430
10	0.5	2.0	2.15	1.085	0.449	0.120
11	6.5	-8.0	25.52	25.469	0.026	0.385
12	-1.5	3.0	-1.55	-1.474	-0.034	0.211

Genstat leent zich bijzonder goed voor vector- en matrix-rekenwerk. Ter illustratie worden enkele grootheden uitgerekend onder de daarin gebruikte benaming. De laatste CALCULATE berekent de grootheden CR_i met een formule die equivalent is aan die van de inleiding. We zullen de uitvoer, als zijnde niet-standaard, verder niet gebruiken.

```
 11 CALCULATE resid=y-pred                "gewoon residu"
 12  & zresid = resid/sqrt(sse/dfe)       "gestandaardiseerd"
 13  & sresid = zresid/sqrt(1-h)          "gestudentiseerd (identiek aan f)"
 14  & lever  = h - 1/12                  "constante doet niet mee in SPSSX"
 15  & mahal  = 11 * lever                "afstand**2 tot centrum in x1-x2-vlak"

 16  & cook_d = (sresid**2/3)*h/(1-h)  "COOK's distance"
 17  & cr     = ((n-p-f**2)/(n-p-1))**p/(1-h) "cov ratio, zie inleiding"
 18 PRINT veld,pred,resid,zresid,sresid,lever,mahal,cook_d,cr; F=8; D=0,8(3)
```

veld	pred	resid	zresid	sresid	lever	mahal	cook_d	cr
1	9.478	-3.628	-1.433	-1.694	0.201	2.208	0.379	0.629
2	15.793	1.257	0.497	0.544	0.085	0.932	0.020	1.548
3	21.802	0.248	0.098	0.115	0.194	2.135	0.002	1.962
4	1.128	1.772	0.700	0.741	0.026	0.283	0.022	1.323
5	-4.365	-2.135	-0.843	-1.053	0.275	3.024	0.206	1.496
6	6.246	4.154	1.641	1.755	0.043	0.475	0.149	0.464
7	10.588	-2.088	-0.825	-0.912	0.099	1.084	0.062	1.301
8	-11.064	2.164	0.855	1.059	0.266	2.921	0.200	1.467
9	-5.515	-2.785	-1.100	-1.456	0.346	3.811	0.533	1.115
10	1.085	1.065	0.421	0.449	0.037	0.407	0.009	1.512
11	25.469	0.051	0.020	0.026	0.301	3.315	0.000	2.314
12	-1.474	-0.076	-0.030	-0.034	0.128	1.405	0.000	1.804



figuur 4.1

4.3 Tweede stap

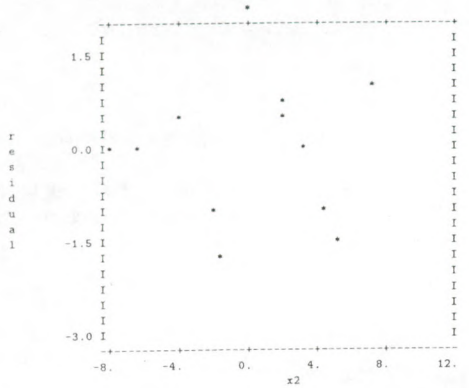
Grafische beoordeling van de gestudentiseerde residuen.
Residu tegen x_1. Er zit een opvallend gat in de ingestelde x_1-waarden. Punt 6 ligt qua residu nogal geïsoleerd. In de residuen lijkt mede daardoor een verloop te zitten: eerst stijgend, dan dalend. Residu tegen x_2. Vrijwel hetzelfde beeld als residu tegen aangepaste waarde (figuur inleiding), wat niet zo verwonderlijk is omdat x_1 niet duidelijk (althans niet significant) mee doet. Opvallend is de lange reeks monotoon dalende punten 6-4-10-12-5-9 (zie figuur 4.2). Deze grafieken veroorzaken het gevoel dat er iets niet goed zit, maar wijzen niet in de richting van 9 en 5 als boosdoeners. Overwogen kunnen worden het splitsen van de data in twee groepen naar x_1-waarde, of het opnemen van een tweede-graads effect van x_1.

```
19 GRAPH [YTITLE='residual'; XTITLE='x1'; NROWS=20; NCOLUMNS=50] f; x1
```

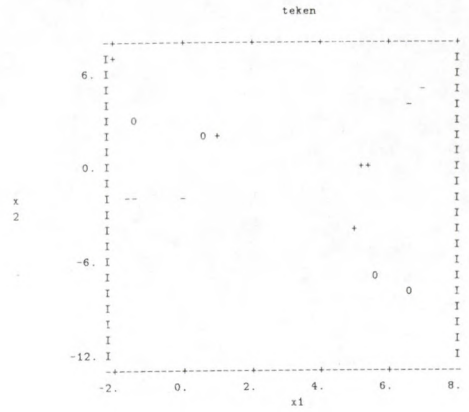
```
20 GRAPH [YTITLE='residual'; XTITLE='x2'; NROWS=20; NCOLUMNS=50] f; x2
```

Teken van het residu in het x_1-x_2-vlak (zie figuur 4.3). Nuttig voor het opsporen van interactie. De plotsymbolen `--,-,0,+` en `++` horen bij de arbitraire indeling van de residuen in klassen met grenzen `-1.5,-0.5, 0.5` en `1.5`. Het gat in de x_1-instellingen is weer prominent aanwezig; veel opvallender dan een wat geïsoleerd liggen van 9 en 5. In de tekens lijkt een golvend verloop te zitten van linksonder naar rechtsboven: achtereenvolgens `--,-,0,+` en `++` en vervolgens weer `-`. Dit lijkt op een tweedegraads component in genoemde richting en was voor ons aanleiding om het model uit te breiden tot een tweedegraads responsie-oppervlak. *Achteraf* gezien was hier ook de 'gouden' constatering mogelijk geweest dat 5 en 9 de lineaire trend `--,-,0,+,++` verstoren door niet `+++` te zijn maar `-`. Zo zorgvuldig worden grafiekjes niet altijd bestudeerd.

```
21 TEXT [NVALUES=5; VALUES='--','-','0','+','++'] teken
```



figuur 4.2



figuur 4.3

```
22 FACTOR [LABELS=teken] klasse
23 SORT [GROUPS=klasse; LIMITS=!(-1.5,-0.5,0.5,1.5)] f
24 GRAPH [T='teken'; YT=,x2'; XT='x1'; NR=20; NC=50] x2; x1; SYMBOL=klasse
```

4.4 Derde stap

Dan maar eens een tweedegraads responsie-oppervlak proberen (natuurlijk qua aantal parameters wel erg ruim). De gecorrigeerde R_2 stijgt hierdoor van 95.1 naar 99.4, en MSE daalt van 6.41 naar 0.76. Genstat geeft nu een waarschuwing bij punt 6: groot residu (2.22). Na het schrappen van x_1^2 (niet significant) verandert er weinig; nu blijkt punt 12 een groot residu (–2.06) te vertonen.

```
25 CALCULATE x11=x1**2 & x22=x2**2 & x12=x1*x2
26 FIT x1+x2+x11+x22+x12

             d.f.      s.s       m.s.
 Regression     5  1445.313   289.0626
 Residual       6     4.572     0.7621
 Total         11  1449.885   131.8078

Percentage variance accounted for 99.4

MESSAGE: The following units have large residuals:
                    6          2.22

           estimate   s.e.      t
 Constant     5.799  0.411   14.11
 x1           0.977  0.288    3.39
 x2          -1.514  0.134  -11.27
 x11         -0.0636 0.0683  -0.93
 x22         -0.1128 0.0356  -3.17
 x12         -0.2157 0.0372  -5.79

  27 FIT x1+x2+x22+x12

             d.f.      s.s       m.s.
 Regression     4  1444.652   361.1630
 Residual       7     5.234     0.7477
 Total         11  1449.885   131.8078

Percentage variance accounted for 99.4

MESSAGE: The following units have large residuals:
                   12         -2.06
```

```
          estimate    s.e.      t
Constant    5.673    0.384    14.76
x1          0.730    0.112     6.52
x2         -1.470    0.124   -11.81
x22        -0.1389   0.0218   -6.38
x12        -0.2378   0.0284   -8.37

  28 STOP
```

4.5 Conclusie

De standaard uitvoer bevat geen aanwijzing dat er met 9 en 5 iets mis is. De grafiek van tekens van residuen wel, maar pas bij ongebruikelijk zorgvuldige beschouwing. Als de onderzoeker niet met meer informatie komt of tegensputtert dan blijft het bij de constatering dat een tweedegraadsoppervlak de gegevens aardig beschrijft en dat een gewoon vlak door de punten te slecht past.

5 Analyse met SYSTAT

5.1 Besturing van SYSTAT

SYSTAT bestaat uit een aantal modules die min of meer los van elkaar staan. Om van de een naar de ander te gaan moest je in oudere versies eerst met het commando QUIT terug naar het operating system en dan de volgende module oproepen. Bij de nieuwe versie 4.0 kan je nu ook binnen SYSTAT van de ene naar de andere module overstappen, althans volgens de manual; bij mij (op een Olivetti M280) lukte het niet om binnen SYSTAT naar de module MGLH te gaan.
Om de regressie te draaien moeten we eerst naar de DATA module om de data in te voeren:

```
SYSTAT
DATA
SAVE VOORB
INPUT OBS$ Y X1 X2
```

de data

~ (tilde karakter)

```
RUN
QUIT
QUIT
```

Na deze commando's is een SYSTAT-file aangemaakt met de naam VOORB. We zijn nu weer in MS-DOS en draaien vervolgens de module MGLH (Multivariate General Linear Hypothesis).

```
MGLH
USE VOORB
```

```
MODEL Y=CONSTANT+X1+X2
SAVE MATEN/MODEL
ESTIMATE
QUIT
```

De residuals en de verschillende maten zijn nu tesamen met de variabelen gesaved in een nieuwe SYSTAT-file met de naam MATEN. Met behulp van de module DATA kunnen ze nu geprint worden:

```
DATA
USE MATEN
CASELIST
RUN
```

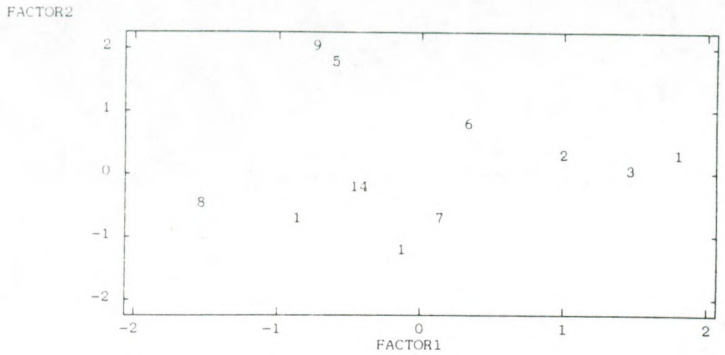
5.2 Invloedsmaten

De uitvoer van SYSTAT omvat de volgende maten:
LEVERAGE (Belsley, Kuh en Welsch), COOK (Cook's D), STUDENT, de studentized residuals en SEPRED, de standard error of predicted values. De maten LEVERAGE en COOK zijn dezelfde als in het SAS pakket. De uitvoer wat betreft de verschillende maten is te zien in tabel 1.

Tabel 1. Invloedsmaten volgens de SYSTAT-uitvoer.

		LEVERAGE	COOK	STUDENT	SEPRED
CASE	1	0.284	0.397	-1.934	1.350
CASE	2	0.168	0.020	0.522	1.038
CASE	3	0.277	0.002	0.108	1.334
CASE	4	0.109	0.022	0.721	0.836
CASE	5	0.358	0.206	-1.060	1.516
CASE	6	0.127	0.149	2.041	0.901
CASE	7	0.182	0.062	-0.902	1.080
CASE	8	0.349	0.200	1.067	1.496
CASE	9	0.430	0.533	-1.571	1.660
CASE	10	0.120	0.009	0.428	0.878
CASE	11	0.385	0.000	0.024	1.571
CASE	12	0.211	0.000	-0.032	1.163

In de SYSTAT-manual wordt vermeld dat COOK vergeleken kan worden met een F-verdeling met p en $n-p$ vrijheidsgraden. Dit levert een grenswaarde op van 3.86 bij een rechter overschrijdingskans $p = 0.05$. De maat STUDENT gedraagt zich onder de gebruikelijke voorwaarden voor regressie als een t-verdeling met $n-p-2$ vrijheidsgraden. Dus de grenswaarde bij $p = 0.05$ hiervoor is 2.36. SEPRED vertoont sterke gelijkenis met de maat LEVERAGE. In de SYSTAT-uitvoer is er geen enkele maat die alarm slaat voor de punten 5 en 9.



figuur 5.1

5.3 Principale Componenten Analyse

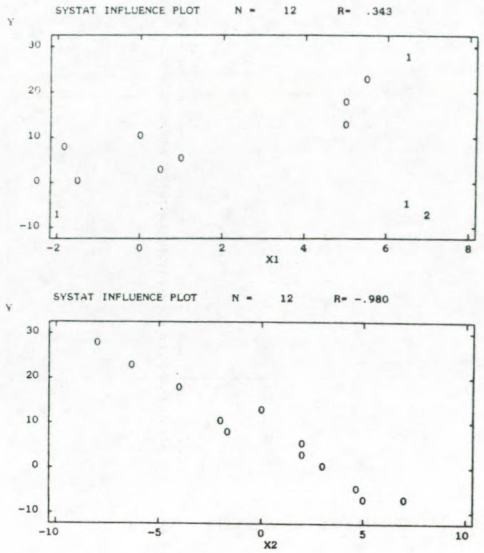
Bij sommige statistiek afdelingen is het gewoonte om op nieuwe data eerst Principale Componenten Analyse te doen om uitbijters te vinden (tip van H. van der Knaap, Unilever). Als we dit uitvoeren met SYSTAT op de drie variabelen y, x_1 en x_2 zien we dat er eigenlijk maar twee factoren zijn, waarvan de eerste factor het belangrijkste is. De eerste factor correleert hoog met y en met $-x_2$ en minder met x_1, de tweede correleert alleen vrij hoog met x_1. De percentages verklaarde varianties voor drie factoren en eigenwaarden zijn te zien in de volgende tabel:

	% verklaarde var.	eigenwaarde
Factor 1	72.2	2.170
Factor 2	27.1	0.809
Factor 3	0.7	0.020

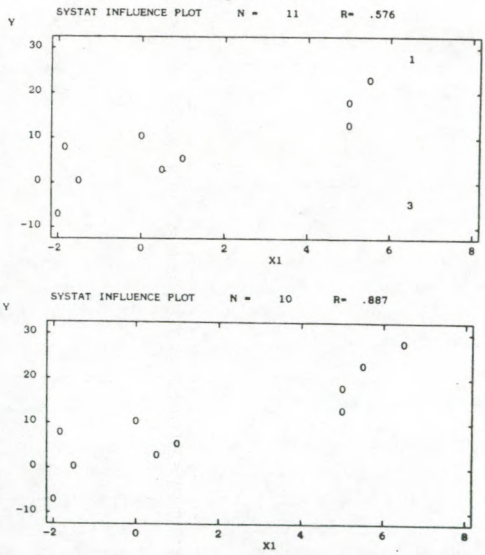
Figuur 5.1 geeft een plaatje van de factorscores van factor 1 tegen factor 2. We zien dat de eerste factor de hoofdrichting van alle punten weergeeft, terwijl de tweede factor bepaald wordt door de punten 5 en 9 en in mindere mate door punt 1. Bij Principale Componenten Analyse zou dus zeker de aandacht op de punten 5 en 9 vallen.

5.4 Influence plots

In de GRAPH-module van SYSTAT zit de mogelijkheid om zogenaamde influence plots te maken. Dit zijn X-Y plotjes met alle punten weergegeven door nullen terwijl punten die een grote invloed hebben op de correlatiecoëfficiënt door een ander cijfer worden weergegeven. Het cijfer 1 betekent dat de correlatiecoëfficiënt met een waarde tussen de 0.1 en 0.2 zou veranderen als het punt weggelaten zou worden, bij een 2 zou de correlatiecoëfficiënt veranderen tussen 0.2 en 0.3 enzovoort. Figuur 5.2 geeft de influence plots voor y tegen x_1 en y tegen x_2. Het plotje van y tegen x_2 geeft geen probleemgevallen aan: alle punten



figuur 5.2



figuur 5.3

worden weergegeven door een nul. De plot van y tegen x_1 geeft vier punten met een grote invloed op de correlatiecoëfficiënt: de punten 8, 9 en 11 worden weergegeven door een 1 terwijl punt 9 als een 2 afgedrukt is. Het ligt voor de hand om punt 9 weg te laten. Weglating van punt 9 verandert de influence plot drastisch, zie figuur 5.3. Nu wordt punt 5 duidelijk herkend als een punt met een grote influence namelijk een 3. Weglaten van punt 5 geeft een influence plot die geheel bestaat uit nullen.
Er kunnen tegen deze methode enige bezwaren opgeworpen worden. Allereerst wordt de enkelvoudige correlatie beschouwd tussen y en x_1 en tussen y en x_2 en niet de multipele correlatie. Meestal zullen uitbijters echter ook wel op deze manier ontmaskerd kunnen worden. Een tweede bezwaar is dat bij grotere n de invloed van een of twee punten op de correlatiecoëfficiënt steeds kleiner wordt. Mogelijk worden dan ver weg liggende punten dan niet meer herkend. Dus deze methode is alleen geschikt voor kleine aantallen. Voor het onderhavige probleem werkt de methode goed.

6 Verwerking met SPSS

Verwerking regressieprobleem van J. Dijkstra met SPSSX 2.0 op een CYBER 170/760 onder bedrijfssysteem NOS van CDC.

SPSS (Statistical package for the Social Sciences) is een algemeen statistisch pakket, ontwikkeld op de Stanford University in de USA. Later heeft SPSS INC de ontwikkeling en verspreiding overgenomen. Het pakket groeide en de naam dekte al lang niet meer de lading, zodat een andere naam bedacht is, namelijk Superior Performing Software Systems.

SPSS is een gebruikersvriendelijk pakket waarbij met behulp van een keyword-taaltje (commands en subcommands) tamelijk gemakkelijk een heleboel dingen gedaan kunnen worden. Het pakket draait op bijna alle mainframe computers, veel mini's, workstations en IBM compatible pc's, hetgeen erg plezierig is voor de uitwisseling.

De invoer van een standaard regressieprobleem met drie variabelen ziet er met SPSSX als volgt uit:

```
REGRESSION     VARIABLES=X1 X2 Y
             / DEPENDENT = Y
             / ENTER
```

Deze drie subcommands moeten altijd opgegeven worden.
De standaard uitvoer is dan:

```
* * * *  M U L T I P L E   R E G R E S S I O N  * * * *

VARIABLE LIST NUMBER 1   LISTWISE DELETION OF MISSING DATA
EQUATION NUMBER 1    DEPENDENT VARIABLE..   Y
BEGINNING BLOCK NUMBER  1.   METHOD:  ENTER
 VARIABLE(S) ENTERED ON STEP NUMBER   1..     X2
                                      2..     X1
```

```
MULIPLE R          .97990  ANALYSIS    OF     VARIANCE
R SQUARE           .96020                DF  SUM OF SQUARES  MEAN SQUARE
ADJ. R SQUARE      .95135  REGRESSION    2     1392.17757     696.08878
STANDARD ERROR    2.53219  RESIDUAL      9       57.70792       6.41199
                             F =      108.56047       SIGNIF F =  .0000

------------------- VARIABLES IN THE EQUATION -----------------

VARIABLE         B        SE B     BETA       T      SIG T
X2           -2.38669    .17291  -.97082  -13.803    .0000
X1             .08621    .22573   .02686     .382    .7114
(CONSTANT)    5.81489    .94772             6.136    .0002

FOR BLOCK NUMBER  1   ALL REQUESTED VARIABLES ENTERED.
```

De variabele X1 is niet significant (overschrijdingskans = 0.7114), maar variabele X2 wel (overschrijdingskans=0.0000).
De regressievergelijking wordt dan:

```
Y=5.8149 + 0.0862*X1 - 2.3867*X2
```

met een RQUARE van 0.9602.
SPSSX biedt evenwel de mogelijkheid om meer gegevens af te drukken met behulp van de subcommands DESCRIPTIVES en STATISTICS, bijvoorbeeld:

```
/  DESCRIPTIVES=ALL
/  STATISTICS=ALL
```

De parameter ALL is in principe verwerpelijk, maar in sommige gevallen toch wel handig.

De extra uitvoer van DESCRIPTIVES boven hetgeen hierboven afgedrukt is, ziet er als volgt uit:

```
         MEAN   STD DEV  VARIANCE  LABEL
 X1     2.642    3.577    12.795
 X2      .117    4.670    21.809
 Y      5.764   11.481   131.808

N OF CASES =    12

CORRELATION, COVARIANCE, SIGNIFICANCE, CROSS-PRODUCT:
```

	X1	X2	Y
X1	1.000	-.326	.343
	12.795	-5.438	14.082
	.999	.302	.275
	140.749	-59.818	154.901
X2	-.326	1.000	-.980
	-5.438	21.809	-52.520
	.302	.999	.000
	-59.818	239.897	-577.715
Y	.343	-.980	1.000
	14.082	-52.520	131.808
	.275	.000	.999
	154.901	-577.715	1449.885

De extra uitvoer van STATISTICS boven hetgeen hierboven afgedrukt is, ziet er als volgt uit:

```
R SQUARE CHANGE     .96020
F CHANGE          108.56047
SIGNIF F CHANGE     .0000

CONDITION NUMBER BOUNDS:        1.119,      4.474

VAR-COVAR MATRIX OF REGRESSION COEFFICIENTS (B)
BELOW DIAGONAL:  COVARIANCE    ABOVE:  CORRELATION

            X2      X1
 X2      .02990  .32554
 X1      .01271  .05096
```

XTX MATRIX	X2	X1	Y
X2	1.11854	.36412	.97082
X1	.36412	1.11854	-.02686
Y	-.97082	.02686	.03980

```
EQUATION NUMBER 1    DEPENDENT VARIABLE..   Y

---------------------- VARIABLES IN THE EQUATION ----------------------
```

VARIABLE	95 CONFDNCE INTRVL BB		CORREL	PART COR	PARTIAL	TOL.
X2	-2.77783	-1.99555	-.97957	-.91794	-.97719	.89403
X1	-.42444	.59685	.34290	.02540	.12628	.89403
(CONSTANT)	3.67101	7.95877				

SUMMARY TABLE

```
 STEP  MULTR   RSQ   ADJRSQ  F(EQU)  SIGF  RSQCH   FCH    SIGCH
  1
  2   .9799  .9602   .9514  108.560  .000  .9602  108.560  .000

     VARIABLE  BETAIN  CORREL
IN:     X2     -.9796  -.9796
IN:     X1      .0269   .3429
```

SPSSX biedt ook nog de mogelijkheid om invloedsmaten af te drukken met behulp van het subcommand CASEWISE, bijvoorbeeld:

```
CASEWISE=DEFAULTS ALL MAHAL COOK SRESID SDRESID LEVER
```

De extra uitvoer ziet er als volgt uit:

```
CASEWISE PLOT OF STANDARDIZED RESIDUAL

*:  SELECTED   M: MIS

                -3.0       0.0       3.0
 CASE #  IDENT  0:.........:.........:0      Y      *PRED
     1    1.00  .      *   .         .    5.85     9.4784
     2    2.00  .          .*        .   17.05    15.7927
     3    3.00  .          *         .   22.05    21.8025
     4    4.00  .          .  *      .    2.90     1.1277
     5    5.00  .       *  .         .   -6.50    -4.3649
     6    6.00  .          .    *    .   10.40     6.2459
     7    7.00  .        * .         .    8.50    10.5883
     8    8.00  .          .   *     .   -8.90   -11.0643
     9    9.00  .       *  .         .   -8.30    -5.5151
    10   10.00  .          .*        .    2.15     1.0846
    11   11.00  .          *         .   25.52    25.4687
    12   12.00  .          *         .   -1.55    -1.4745
 CASE #  IDENT  0:.........:.........:0      Y      *PRED
                -3.0       0.0       3.0
```

figuur 6.1

CASE #	*RESID	*SRESID	*SDRESID	*MAHAL	*C
1	-3.6284	-1.6935	-1.9343	2.2082	.3793
2	1.2573	.5444	.5219	.9321	.0200
3	.2475	.1150	.1085	2.1351	.0017
4	1.7723	.7415	.7215	.2827	.0224
5	-2.1351	-1.0526	-1.0597	3.0241	.2062
6	4.1541	1.7553	2.0407	.4749	.1487
7	-2.0883	-.9117	-.9023	1.0836	.0616
8	2.1643	1.0592	1.0674	2.9206	.2003
9	-2.7849	-1.4565	-1.5707	3.8114	.5331
10	1.0654	.4486	.4277	.4070	.0092
11	.0513	.0258	.0243	3.3150	.0001
12	-.0755	-.0336	-.0317	1.4053	.0001
CASE #	*RESID	*SRESID	*SDRESID	*MAHAL	*C

RESIDUAL STATISTICS:

	MIN	MAX	MEAN	STD DEV	N
*PRED	-11.0643	25.4687	5.7642	11.2500	12
*ZPRED	-1.4959	1.7515	.0000	1.0000	12
*SEPRED	.8361	1.6601	1.2351	.2906	12
*ADJPRED	-12.2238	25.4367	5.9900	11.1898	12
*MAHAL	.2827	3.8114	1.8333	1.2351	12
*COOK D	.0001	.5331	.1319	.1729	12
*RESID	-3.6284	4.1541	-.0000	2.2905	12
*ZRESID	-1.4329	1.6405	-.0000	.9045	12
*SRESID	-1.6935	1.7553	-.0382	1.0521	12
*DRESID	-5.0681	4.7557	-.2259	3.1337	12
*SDRESID	-1.9343	2.0407	-.0489	1.1446	12
*LEVER	.0257	.3465	.1667	.1123	12

TOTAL CASES = 12

waarbij:

```
PRED        - Unstandardized predicted values
RESID       - Unstandardized residuals
DRESID      - Deleted residuals
ADJPRED     - Adjusted predicted values
ZPRED       - Standardized predicted values
ZRESID      - Standardized residuals
SRESID      - Studentized residuals
SDRESID     - Studentized deleted residuals (Hoaglin & Welsh, 1978)
SEPRED      - Standard errors of the predicted values
MAHAL       - Mahalanobis' distances
COOK        - Cook's distances (Cook, 1977)
LEVER       - Leverage values (Velleman & Welsh, 1981)
```

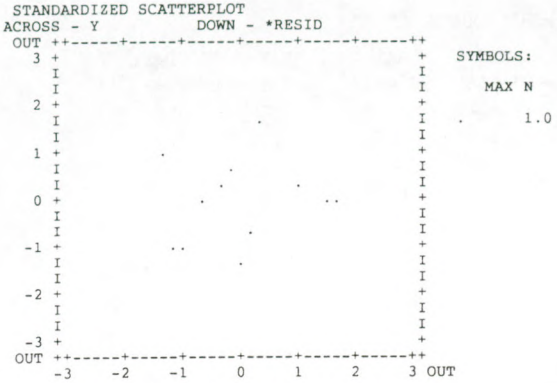
SPSSX biedt ook nog de mogelijkheid om residuen af te drukken met behulp van het subcommand RESID, bijvoorbeeld:

```
    / RESID=DEFAULT SIZE(SMALL) id(IDENT)
```

De extra uitvoer ziet er dan als volgt uit:

```
OUTLIERS - STANDARDIZED RESIDUAL
   CASE #      IDENT      *ZRESID
        6       6.00       1.64051
        1       1.00      -1.43292
        9       9.00      -1.09979
        8       8.00        .85473
        5       5.00       -.84320
        7       7.00       -.82469
        4       4.00        .69990
        2       2.00        .49654
       10      10.00        .42074
        3       3.00        .09775

HISTOGRAM
STANDARDIZED RESIDUAL
NEXP N      (* = 1 CASES,     .  :  = NORMAL CURVE)
0  .01   OUT
0  .02  3.00
0  .05  2.67
0  .11  2.33
0  .22  2.00
1  .40  1.67 *
0  .66  1.33 .
1  .97  1.00 :
1 1.27   .67 :
```



figuur 6.2

```
2 1.50   .33 *:
3 1.59     0 *:*
0 1.50  -.33  .
1 1.27  -.67 :
2  .97 -1.00 :*
1  .66 -1.33 :
0  .40 -1.67
0  .22 -2.00
0  .11 -2.33
0  .05 -2.67
0  .02 -3.00
0  .01   OUT
```

Bovendien kunnen een aantal plotjes afgedrukt worden met het subcommand SCATTERPLOT, bijvoorbeeld:

```
/ SCATTERPLOT (*RES,*PRE)
```

De extra uitvoer ziet er dan als volgt uit: De uitvoer geeft geen aanleiding tot verdere stappen, maar gegeven is dat er met deze punten iets aan de hand is. Alleen de invloedsmaten kunnen een indruk geven dat er iets aan de hand is. De invloedsmaat *MAHAL met een waarde 3.8114 is bijna significant voor case 9. Bovendien blijkt dat LEVER met een waarde van .3465 significant is voor case 9.

De overweging is dan om hetzelfde nog eens te doen zonder case 9.
De uitkomsten zijn dan:

```
Y=5.6323 + 0.3026*X1 - 2.2333*X2
```

met een RSQUARE van 0.9643.

De bijbehorende RESIDUALS STATISTICS zijn:

	MIN	MAX	MEAN	STD DEV	N
*PRED	-10.6061	25.4658	7.0427	10.9087	11
*ZPRED	-1.6179	1.6888	-.0000	1.0000	11
*SEPRED	.8051	1.8603	1.1890	.3144	11
*ADJPRED	-11.5837	25.4319	7.6135	10.6967	11
*MAHAL	.2666	5.3680	1.8182	1.4937	11
*COOK D	.0000	4.4904	.4797	1.3349	11
*RESID	-4.0496	3.2545	.0000	2.1001	11
*ZRESID	-1.7247	1.3861	.0000	.8944	11
*SRESID	-2.8266	1.5363	-.0898	1.2086	11
*DRESID	-10.8774	3.9982	-.5708	4.0585	11
*SDRESID	-73.8137	1.7115	-6.5459	22.3264	11
*LEVER	.0267	.5368	.1818	.1494	11

Uit de invloedsmaten blijkt dan dat er wat aan de hand is met case 5. De volgende stap is dan om case 5 weg te laten. Met de weglating van case 5 en 9 ziet de regressielijn er dan als volgt uit:

```
Y=5.0928 + 1.0620*X1 - 1.6930*X2
```

met een RSQUARE van 0.9999.

De bijbehorende RESIDUALS STATISTICS zijn:

	MIN	MAX	MEAN	STD DEV	N
*PRED	-8.8820	25.5397	8.3970	10.7098	10
*ZPRED	-1.6134	1.6007	-.0000	1.0000	10
*SEPRED	.0341	.0641	.0482	.0106	10
*ADJPRED	-8.8684	25.5520	8.4003	10.7067	10
*MAHAL	.3920	3.6676	1.8000	1.1568	10
*COOK D	.0004	.2743	.0888	.0974	10
*RESID	-.1248	.1311	.0000	.0793	10
*ZRESID	-1.3877	1.4583	.0000	.8819	10
*SRESID	-1.5528	1.5873	-.0141	.9969	10
*DRESID	-.1562	.1554	-.0033	.1021	10
*SDRESID	-1.7756	1.8369	.0065	1.1046	10
*LEVER	.0436	.4075	.2000	.1285	10

De uitkomsten zijn nu prima, maar als je weer naar de invloedsmaten kijkt, dan blijkt dat case 1 een significante MAHAL en LEVER invloedsmaat heeft. In principe zouden we dus verder moeten gaan door nu case 1, 5 en 9 weg te laten.

Enige informatie betreffende SPSSX:

- de instructie set en uitkomsten zijn gelijk aan die van SPSS/PC V2.0.

- SPSSX werkt veel met default waarden, maar REGRESSION kent geen default model, hetgeen een sterk punt is.
- volgens het manual SPSSX, 2nd edition moet de keuze van het model voorafgegaan worden door het subcommand MODEL. In de batch versie is dat juist fout, het subcommand moet achterwege gelaten worden.
- met het subcommand CASEWISE kunnen 12 tijdelijke variabelen zoals *PRED, *RESID, ..., *LEVER aangemaakt en afgedrukt worden, maar helaas niet in één keer (run). Per keer kunnen er maar maximaal zes tijdelijke variabelen afgedrukt worden. Op de CYBER is het mij niet gelukt om de variabele LEVER afgedrukt te krijgen, daarentegen ging dat met SPSS/PC prima.
- het significantie niveau voor LEVER is $2 * p/n$, waarbij met p in SPSS het aantal onafhankelijke variabelen wordt bedoeld in plaats van alle variabelen.
- in de beschrijving van REGRESSION zitten wat slordigheden, bijvoorbeeld met BCOV wordt de correlatiematrix de ene keer boven de diagonaal en de andere keer weer onder de diagonaal afgedrukt.
- P-P plot met residuen levert niet voldoende informatie op. Beter is het om te kijken naar de 10 extrema van de outliers.

7 Het BMDP pakket

7.1 BMDP in relatie tot SAS, SPSS etc.

BMDP heeft een pricipiëel andere opzet dan pakketten als SAS en SPSS. In deze laatsten zijn alle technieken in één programma verenigd, in BMDP is voor iedere techniek een apart programma voorhanden.
Misschien kan dit nog het best geïllustreerd worden door SAS en SPSS te vergelijken met de bekende 'Swiss Army knives'. In een allesomvattend gereedschap vindt men behalve messen ook een schaar, een rasp, een pincet etc. Wanneer men een 'full feature' mes aanschaft heeft men zelfs een compleet opvouwbaar bestek op zak. BMDP daarentegen lijkt meer op een verzameling afzonderlijke gereedschappen. Met SAS en SPSS heeft men vrijwel àlle benodigde instrumenten op zak, met BMDP moet men, afhankelijk van het probleem, overschakelen naar een ander programma.

7.2 Standaardanalyse van het standaardprobleem in BMDP

Bij de analyse van het standaardprobleem is uitgegaan van de positie van de 'argeloze' gebruiker. Er is niet bij voorbaat gebruik gemaakt van de wetenschap die het bestuderen van het plotje van de data levert. De doorsnee BMDP gebruiker zal in eerste instantie het programma 1R kiezen. Wel zal een BMDP gebruiker doorgaans iets beter statistisch

onderlegd zijn en daarom eerder geneigd de volledige set 'regression-diagnostics' te vragen.

In dit geval zijn plotjes gevraagd van:
— residuen vs. voorspelde waarden
— kwadraten van residuen vs. voorspelde waarden
— geobserveerde en voorspelde waarden vs. X_1, resp. X_2.
— residuen vs. X_1, resp. X_2.
— 'gecorrigeerde' residuen vs. X_1, resp. X_2.
— 'detrended' en gewone normal probability plots van de residuen.

Voorts zijn de gebruikelijke resultaten als regressiecoëfficiënten, met standaardfout etc., multiple R^2 en een variantenanalyse tabel geproduceerd.

Helaas, zoals de inleiding van Dijkstra al vermeldt, niets levert ook maar enig vermoeden op dat er iets met de data aan de hand zou zijn. Wel is een duidelijke conclusie dat de plotjes niet erg geschikt zijn om de feitelijke boosdoeners te identificeren, mochten deze bestaan. Het is vrijwel onmogelijk aan de hand van de plotjes uit te maken welk punt bij welke set gegevens hoort. Aan de hand van de tabel met X_1, X_2, Y en de voorspelde en residuele waarden is er wel uit te komen, maar bij meer reële problemen met enkele honderden waarnemingen zal ook dit problematisch worden.

7.3 Seriële correlatie van de residuen

Aandachtige bestudering van de standaardanalyse leverde plotseling een vermoeden van een interessant verschijnsel. De seriële correlatie van de residuen bleek maar liefst `-0.7615` te zijn. Daarom werd de analyse nog eens gedraaid, eerst met alle cases gesorteerd volgens de waarden van X_1, daarna volgens die van X_2, resp. Y. In alle drie gevallen nam de correlatie af tot bijna 0. De toevallige volgorde waarin de waarnemingen werden aangeboden zal dus wel te maken hebben met het constructieproces van de gegevens.

7.4 Op zoek naar meer diagnostics

Uitgaande van de wetenschap dat Cook's distance in ons geval de boosdoeners zou kunnen aanwijzen, werd in het manual gevonden dat `9R`, 'All Possible Subsets Regression' die kan geven. De '87 update informatie leerde dat vanaf versie 87 ook met `2R`, 'Stepwise Regression', deze maat berekend kan worden, tesamen met nog een groot aantal toetsingsgrootheden.

`2R` levert in totaal 21 grootheden, die in 3 groepen verdeeld kunnen worden, de *Residual Measures* welke aanwijzingen kunnen geven voor uitbijters in de Y-variabele, de *Leverage Measures* welke extreme waarden in de predictoren kunnen opsporen en de *Influence Measures* die cases aangeven welke een ongewoon grote invloed op de schatters van de regressiecoëfficiënten hebben. In deze laatste groep vinden we Cook, Modified Cook, Belsley, Kuh & Welsch en Andrews-Pregibon.

Besloten werd, zoals een doorsnee gebruiker zou doen, eerst het programma 9R te gebruiken om daarna tot de ontdekking te komen dat 2R ook het gezochte kan leveren.

7.5 De resulaten van 9R

Hier openbaarde zich helaas een van die kleine onverwachte problemen die ieder pakket van tijd tot tijd vertoont. 9R doet zeer veel, het weigert echter de best-possible subset uit een serie van slechts 2 predictoren te berekenen. Zoals zo vaak is ook dit te omzeilen; in dit geval door een derde predictor toe te voegen. Door vervolgens de analyse tot X_1 en X_2 te beperken krijgen we exact wat we willen, maar het is een vreemde omweg.

Naast de plotjes die al met 1R waren verkregen werden nu ook plotjes gevraagd met de Cook's distance vs X_1, resp. X_2. Tevens werd de tabel met geobserveerde, voorspelde etc. waarden uitgebreid met de Mahalanobis en Cook distances.

BMDP 9R vermeldt dat case 9 de grootste Mahalanobis distance van 3.81 vertoont. De grootste Cook's distance behoort ook bij case 9, namelijk 0.53. Tevens wordt aangegeven dat verwijdering van case 9 tot gevolg zal hebben dat alle regressiecoëfficiënten zich zullen verplaatsen tot binnen de 39.58% 'confidence ellipsoid'.

Opvolgen van deze raad en opnieuw draaien, maar dan zonder case 9, levert een herhaling van de procedure. Nu wordt geadviseerd ook case 5 te verwijderen. Dit zal tot gevolg hebben dat de coëfficiënten tot binnen de 95.07% ellipsoid opschuiven, hetgeen zeer bevredigend mag heten. De regressievergelijking is dezelfde als door Dijkstra in de probleemstelling gegeven.

7.6 De resultaten van 2R

Ook bij 2R was er even een initiëel probleem. Het programma voegt stapsgewijs variabelen toe aan de regressievergelijking en verwijdert eventueel stapsgewijs variabelen. Hierbij worden criteria als de F-waarde behorend bij de op te nemen (of te verwijderen) variabele gehanteerd. Default kent 2R daarvoor grenzen en variabele X_1 voldoet hier niet aan, zodat bij de eerste run de regressievergelijking beperkt bleef tot X_2. Men kan echter op eenvoudige wijze de opname van X_1 forceren en zodoende toch de hele vergelijking oplossen.

Uit de tabel met allerlei toetsingsgrootheden blijkt ook hier case 9 de grootste storing te veroorzaken. Verwijderen en opnieuw draaien wijst net als bij 9R vervolgens case 5 aan als grootste afwijker.

Helaas ontbreekt in 2R de rapportage over de 'confidence ellipsoid' die in 9R wel wordt gegeven als bijprodukt van de Cook's distance.

7.7 Conclusies

BMDP rapporteert de cases 9 en 5 als vreemde eenden in de bijt. Dit gebeurt echter pas als toetsen als Cook en Mahanalobis worden berekend. Het verdient aanbeveling deze

grootheden standaard in alle BMDP regressieprogramma's op te nemen.

Het zou bijzonder prettig zijn als er een grote mate van standaardisatie werd nagestreefd. 1R, 2R en 9R hebben kleinere en grotere verschillen in de mogelijke toetsingsgrootheden, de manier waarop de plotjes worden gespecificeerd, etc. Een uniforme set mogelijkheden zal het gebruikersgemak zéér ten goede komen.

Vanaf versie 87 is 2R het meest geëigende programma voor standaard regressieanalyses. Daarnaast biedt het zeer uitgebreide mogelijkheden voor stepwise regressie en vooral voor residuen-analyse.

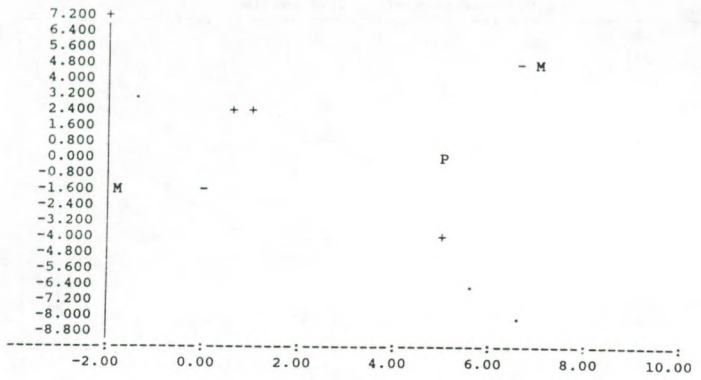
8 Analyse met behulp van GLIM

Het aardige van GLIM is dat men makkelijk verder kan rekenen met gefitte waarden, residuen de diagonaal van de 'hat-matrix', etc. Bovendien kan men macro's schrijven voor bepaalde analysetechnieken, en tenslotte kan men in GLIM heel makkelijk plotjes maken (met de resolutie van een regeldrukker). Deze voordelen zijn niet uniek voor GLIM. Verschillende modulen voor niet-lineaire regressie bieden ongeveer dezelfde mogelijkheden plus nog andere, en hetzelfde geldt voor verschillende talen voor matrix-manipulatie, zoals bijvoorbeeld APL, MATLAB en SAS-IML.

Hier is een voorbeeld van een plot zoals men dat makkelijk met GLIM maakt. Langs de assen staan (de) twee predictoren. Door middel van 'M','-','.','+' en 'P' is aangegeven of de residuen sterk negatief zijn, negatief, ongeveer nul, positief of sterk positief. Voor het discretiseren van de residuen heeft GLIM $ASsign en $MAP ter beschikking.

```
$CAL r=(%YV-%FV) $SORT rr r $PRI rr      $C %YV=Y, %FV=Y-hat, r=residu:
-3.63  -2.78  -2.14  -2.09  -0.076  0.051  0.248  1.07  1.26  1.77  2.16
4.15
$ASS b=-2.4, -.8, .8, 2.4 :  v=1,2,3,4,5 $C Breekpunten en nieuwe waarden
$MAP r5=r values v interv *b* $FAC r5 5 $C Hercodeer r in 5 categorieen
$PLOT x2 x1 'M-.+P' r5                   $C Grootte van res in x1-x2 plot
```

Dit plaatje suggereert een kwadratische kromming van het regressievlak in de richting van de diagonaal $X_1 = X_2$. Op de diagonaal $X_1 = -X_2$ in het midden positieve residuen, naar buiten toe sterk negatieve. (Dit is echter misleidend; de ware oorzaak is dat waarnemingen 5 en 9, rechts bovenin, ver beneden het regressievlak door de overige punten liggen, waardoor het KK-vlak door alle 12 punten gekanteld is ten opzichte van het KK-vlak door de overige 10 punten. Het verband is niet kwadratisch maar lineair-met-twee-uitzonderingen). Dit zou aanleiding kunnen zijn om maar eens een paar modellen met kwadratische termen aan te passen. In een ijverige bui heb ik ze maar alle 11 geprobeerd. (11 Modellen waaronder een model met 6 parameters aanpassen bij 12 waarnemingen is natuurlijk niet echt aan te bevelen. Maar ik doe het hier om te demonstreren wat men met GLIM kan doen.) Hieronder volgt het 'deviance-df'-plot (zie figuur 8.2) van deze 11 modellen, gemaakt met twee eigen macro's (#save en #dvdf). Vertikaal de deviance (=residuele kwadraatsom),



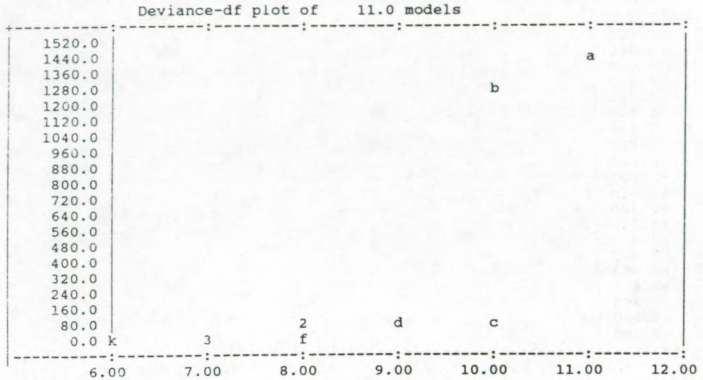
figuur 8.1

horizontaal de vrijheidsgraden. De goede modellen zitten dus langs de rechter-onderrand, hier: a (=model 1), c, f, (één van) de 3 modellen bij 7 df, en k (=model 11). Links-onder de goedpassende, complexe modellen; rechtsboven de spaarzame, iets minder goedpassende modellen. Hoe men spaarzaamheid tegen goodness-of-fit moet afwegen, is mijns inziens meer een inhoudelijke keuze dan een statistische; maar het is in het algemeen verstandig om een model alleen te accepteren als het significant beter past dan alle spaarzamer modellen. Daarmee sluiten we dus de modellen op het meest linkse deel van de rand uit.

```
$FIT $ #save $FIT x1 #save ...  $FIT x1+x2+x12+x22+x1x2 #save #dvdf
```

	model	deviance	df	dev/df	R2	R2adj
	11.00	4.57	6.00	0.762	0.997	0.9942
	10.00	5.23	7.00	0.748	0.996	0.9943
	8.00	12.22	7.00	1.745	0.992	0.9868
	9.00	30.12	7.00	4.303	0.979	0.9674
	6.00	35.69	8.00	4.461	0.975	0.9662
...	5.00	42.16	8.00	5.269	0.971	0.9600
	7.00	57.61	8.00	7.201	0.960	0.9454
	4.00	57.71	9.00	6.412	0.960	0.9514
	3.00	58.64	10.00	5.864	0.960	0.9555
	2.00	1279.41	10.00	127.941	0.118	0.0293
	1.00	1449.89	11.00	131.808	0.000	0.0000

```
Here R2 is defined as 1-dev/dev0,
and R2adj is defined as 1-(dev/df)/(dev0/df0),
where dev0 and df0 refer to model    1.00.
```



figuur 8.2

Model 10, d.w.z. $1 + X_1 + X_2 + X_1^2 + X_2^2$, lijkt erg aantrekkelijk. Significant beter dan spaarzamer modellen, $R^2 = 0.996$. En wat biedt GLIM nu aan regressiediagnostiek? Zonder macro's komt men met GLIM voor dit probleem niet ver. En ook de standaard macrobibliotheek MACLIB.GLM, Release 1.0 van januari 1985 biedt hier te weinig: een plot van de 'leverages' en een q-q-plot van de gestandaardiseerde residuen. Maar mijn eigen macrobibliotheek bevat een aantal macro's die bij deze data (en algemener) goede diensten bewijzen. Mijn standaard macro voor regressiediagnostiek levert (zonder parameters, heel gebruiksvriendelijk) voor het laatst aangepaste model het volgende op.

- Een q-q-plot van de gestudentiseerde residuen, inclusief een regressielijn door de middelste 50% van de residuen.
- Een plot van deze residuen tegen hun 'leverage'.
- Voor elk van de volgende grootheden, de vijf waarnemingen met de meest extreme waarden, inclusief natuurlijk het nummer van de waarneming, en de kritieke grens van de grootheid: gestudentiseerde residuen $\mathtt{f_i(i)}$, 'leverage' $\mathtt{h_{ii}}$, Cooks afstand $\mathtt{D_i}$ en de verandering in de determinant $\mathtt{CR_i}$. Ook de DFFITS zou makkelijk te implementeren zijn, en ook de $\mathtt{DFBETAS}_{ij}$ zou wel kunnen, maar het is omvangrijker en bewerkelijker.

De moraal van het verhaal is: wat andere pakketten in huis hebben als opties kan men in GLIM vrij makkelijk implementeren, en als men van alle pakketten de beste ideeën jat, en daar eigen voorkeuren en vondsten aan toevoegt, komt men gemakkelijk boven het niveau van de standaardpakketten uit. (Maar, hoe nuttig zijn dit soort diagnostieken? Zeker niet onnuttig, maar zijn echt robuuste schattingsmethoden niet veel beter? Later meer daarover.)

Voor het model $1 + X_1 + X_2 + X_1^2 + X_2^2$ levert mijn macro nogal verontrustende informatie; er gebeuren rare dingen, maar steeds in andere punten.

Maar zoals we al in de inleiding gezien hebben, kom je bij het analyseren van dit probleem niet ver met diagnostische maten gebaseerd op het weglaten van telkens één punt. Het model achter deze data is een populatie waarbij een meerderheid nauwkeurig aan één regressiemodel voldoet, terwijl een (latente) minderheid zich mogelijk heel anders gedraagt. Alleen met hierop gebaseerde, dus robuuste schattingsmethoden kan men het geheim van deze data ontsluieren. Nu kan men met GLIM ook heel makkelijk de robuuste regressietechniek IRLS toepassen: Iteratively Reweighted Least Squares, zie Holland & Welsch (1977) of ROSEPACK. In elke iteratieslag baseert men de gewichten w_i voor de volgende fit op de residuen res_i van de vorige fit. Waarnemingen met gewicht nul hebben natuurlijk wel gewoon residuen. Als men bijvoorbeeld na een fit het gewicht w_i gelijk maakt aan $1/abs(res_i)$ en men minimaliseert vervolgens $\sum_i w_i * res_i^2$, en die procedure convergeert, dan heeft men de L_1-schatter te pakken, die de som van de absolute waarden van de residuen minimaliseert. (Maar dat levert hier niets op, integendeel. Zie de inleiding.)

Interessanter is het om de grootste x% van de residuen een heel klein gewicht te geven. Ik heb even gauw een macro gemaakt, die heel ruw de in absolute waarde meest extreme 50% van de residuen bij de volgende iteratieslag gewicht nul geeft (maar ze leveren daar wel weer residuen op, natuurlijk). Dat is nogal bruut. De discontinuïteit kan problemen met convergentie geven en doet dat ook. GLIM meldt divergentie, en de deviance gaat inderdaad op en neer. Maar de schattingen blijken (hier) wel stabiel. Bovendien kan bij deze methode de keus van een startwaarde kritisch zijn. Immers, stel men heeft een punt dat in de predictorenruimte een extreme uitbijter is (-999 op alle predictoren doet het vaak aardig). Zo'n punt heeft een enorme 'leverage' (en is daar ook aan te herkennen), trekt het regressievlak naar zich toe, en heeft een heel klein residu, en houdt dat bij iteratie. Het voornaamste verschil met Least Median of Squared Residuals lijkt me dat daar een veel betere, maar ook veel duurdere startwaarde gekozen wordt (zie de bijdrage van Bert van Zomeren). Ik heb hier simpelweg voor kleinste kwadraten als startwaarde gekozen. Eventueel kan men het residu schatten op basis van het model waarin dat ene punt zelf niet meedoet, en men moet alsnog uitkijken voor punten met veel 'leverage'.

Ondanks de ruwe implementatie, die zeker voor verbetering vatbaar is, levert deze methode hier feilloos het model op waarbij (o.a) waarneming 5 en 9 niet meedoen, en hun residuen nu onmiddellijk opvallen (z-score ruim -70). Elke methode die niet-meegewogen waarnemingen wel kan analyseren zou dit nu laten zien. Merkwaardig is dat op de CR grootheid (zie inleiding) veel punten extreem scoren, maar dat 5 en 9 niet tot de vijf meest extreme behoren. Wellicht ligt dat aan de steekproefverdeling van CR, waarvoor bij kleine steekproeven (althans bij mij) erg weinig bekend is.

IRLS is een gebruik van GLIM buiten het kader van de exponentiële familie, terwijl het manual ten onrechte suggereert dat GLIM alleen binnen dat kader toepasbaar is. Binnen een exponentiële functie is de variantie een functie van de verwachting. Deze 'variance function' karakteriseert de verdelingen. Een gevolg daarvan is, dat normale verdelingen de enige

translatiefamilie is, en de enige exponentiële familie met alleen symmetrische verdelingen. Binnen de exponentiële familie kun je dus L_1-regressie niet modelleren, ondanks het feit dat L_1-regressie juist de ML-schatting is bij een Laplace verdeling van de storingsterm. (Ook met 'quasi-likelihood' gaat dit niet.) In GLIM mag de 'variance function' echter wel degelijk ook van de waarnemingen afhangen, en niet alleen maar van de gefitte waarden. Juist deze enigszins ongedocumenteerde vrijheid is zeer fundamenteel, en maakt bijvoorbeeld IRLS mogelijk.

9 Verwerking met PROGRESS

9.1 Abstract

Een regressie dataset die slechts twee hefboompunten bevat wordt geanalyseerd met behulp van programmatuur voor robuuste regressie en lokatie schatting. Het blijkt dat een plot van robuuste residuen tegen robuuste afstanden in de predictorruimte een krachtig diagnostisch hulpmiddel is. Een routinematige bewerking wijst beide hefboompunten zonder moeite als afwijkend aan.

9.2 Inleiding

Vanuit het uitgangspunt van robuustheid kan het probleem der regressiediagnostiek als volgt worden bezien. Waarom heeft men eigenlijk diagnostics nodig? Onder andere omdat men inmiddels weet dat een regressie schatting, uitgerekend met de methode der kleinste kwadraten, sterk beïnvloed kan zijn door eventuele slechte waarnemingen. Dit kan tot gevolg hebben dat *goede* waarnemingen ver van de (foute) oplossing komen te liggen. Waarom niet direct een betrouwbare, in dit geval dus een *robuuste* methode gehanteerd? Residuen ten opzichte van een robuuste regressie oplossing zijn, zoals zal blijken, *wel* als diagnostische hulpmiddelen te gebruiken.

Wanneer men vervolgens wil onderzoeken of er in de geanalyseerde dataset waarnemingen aanwezig zijn die een meer dan gemiddelde invloed op de oplossing hebben, (de zogenaamde *hefboompunten*), dan verdient het aanbeveling om een robuuste afstandsmaat te gebruiken en dus niet de klassieke (Mahalanobis) afstand (of de daar direct mee samenhangende hat matrix).

Door kombinatie van robuust residu en robuuste afstandsmaat komt men tot een zeer krachtig diagnostisch hulpmiddel: een grafiek waarin het robuuste residu is uitgezet tegen de robuuste afstand. In zo'n plaatje kan men direct vier groepen punten onderscheiden: 1. uitbijters in de y-richting, 2. goede hefboompunten, 3. slechte hefboompunten, en 4. alle andere punten.

9.3 Robuuste Regressie

Het asymptotisch breekpunt van de kleinste kwadraten regressie schatter heeft de waarde 0. Dit wil zeggen dat het mogelijk is om de schatter een willekeurige waarde te laten krijgen door het veranderen van slechts één punt. In kringen van "robustniks" wordt dit als een nadelige eigenschap ervaren. Een regressie schatter met het grootst mogelijke breekpunt van 50% werd in 1984 geïntroduceerd door P J Rousseeuw [1]. Deze inmiddels vrij bekende "Least Median of Squares" methode hanteert als verliesfunctie

$$\underset{\beta}{\text{mediaan}}\ (y_i - \beta^t \mathbf{x}_i)^2, \tag{1}$$

de mediaan der gekwadrateerde residuen dus. De LMS breekt zelfs niet indien er slechte hefboompunten in het data materiaal aanwezig zijn, zoals dat in de voorbeeld dataset het geval is.

De verliesfunctie (1) is niet eenvoudig te minimaliseren vanwege de aanwezigheid van vele lokale minima. Men kan een benaderende oplossing vinden door het toepassen van een zogenaamde "subsampling" strategie. Dit houdt in dat men random subsamples uit de data trekt ter grootte p, waarin p het aantal onbekende regressie parameters is. Voor elk der subsamples berekent men regressie oplossing en bijbehorende verliesfunctie. De oplossing met de kleinste verliesfunctie is de LMS oplossing.

Voor het standaardiseren der residuen wordt bij deze techniek uiteraard een robuuste spreidingsmaat gebruikt die gebaseerd is op de mediaan der gekwadrateerde residuen. Van de op robuuste wijze gestandaardiseerde residuen wordt aangenomen dat ze buiten het interval `[-2.5,2.5]` vallen indien het regressie-uitbijters betreft.

9.4 Robuuste afstanden

Of een waarneming een scharnierpunt (leverage point) is, hangt alleen af van de ingestelde waarden der x-variabelen. Doorgaans berekent men voor elke observatie i de klassieke, gekwadrateerde Mahalanobis afstand

$$MD_i^2 = (\mathbf{x}_i - T(X))^t C(X)^{-1} (\mathbf{x}_i - T(X)), \tag{2}$$

waarin x_i de kolomvector der x-variabelen is, $T(X)$ de lokatie van de waarnemingen in de x-ruimte is (doorgaans het rekenkundig gemiddelde) en $C(X)$ de klassieke variantie-covariantie matrix. De schatters $T(X)$ en $C(X)$ zijn niet robuust. De hefboompunten, eigenlijk uitbijters in de x-ruimte, zullen zowel $T(X)$ als $C(X)$ beïnvloeden, waardoor bijvoorbeeld de 95% tolerantie ellipsen die op vergelijking 2 gebaseerd zijn opgeblazen zullen worden. Vooral groepjes van scharnierpunten zullen op die manier worden *gemaskeerd*. Omdat hat matrix en Mahalanobis afstand samenhangen volgens

$$MD_i^2 = (n-1)(h_{ii} - 1/n),$$

kunnen we van de diagonaalelementen van de hat matrix niet verwachten dat er ook groepjes invloedrijke punten mee gevonden kunnen worden.

Wanneer men echter robuuste afstandsmaten konstrueert door in vergelijking (2) voor

T en C robuuste schatters in te vullen bereikt men een veel beter resultaat. Een multivariate lokatie schatter met het hoogst mogelijke breekpunt van 50% is de Minimum Volume Ellipsoide Estimator (MVE) van Rousseeuw [2,3], gedefinieerd als het middelpunt van de ellipsoide met het kleinste volume die ten minste $[n/2] + 1$ observaties bevat. De bijbehorende covariantie schatter is dan gebaseerd op die ellipsoide zelf.

Wanneer de onderliggende verdeling multivariaat normaal is zijn de robuuste kwadratische afstanden, gebaseerd op de MVE, bij benadering verdeeld volgens $\chi^2(p)$. Om ver verwijderde punten te vinden dient men de afstanden te vergelijken met bijvoorbeeld het 97.5 percentiel van deze verdeling.

De MVE is moeilijk te berekenen, maar laat zich op dezelfde manier benaderen als de LMS, namelijk door middel van subsampling. Het blijkt zelfs dat beide algorithmen zich zeer goed laten kombineren.

9.5 Software

In het programma PROGRESS van Rousseeuw en Leroy (zie bijv. [2]) wordt de LMS regressie schatter berekend. Tevens wordt een tabel met op robuuste wijze gestandaardiseerde residuen geproduceerd. PROGRESS is beslist geen statistisch pakket in de zin die men er tegenwoordig aan geeft. Men kan eerder spreken van een prototype programma dat een relatief nieuwe techniek uitvoert. Het is geschreven voor PC's in standaard Fortran.

Op de TU Delft werd het programma MINVOL ontwikkeld waarmee de bovenvermelde robuuste afstanden berekend worden. Dit eveneens in standaard Fortran 77 geschreven programma is al evenmin een pakket, maar loopt wel op bijvoorbeeld de CONVEX minisuper van de TU Delft. Dit is van groot belang in verband met het rekenintensieve karakter van algorithmen gebaseerd op resampling.

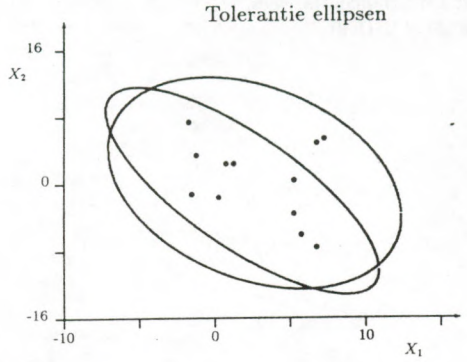
Recentelijk werd in Delft door BC van Zomeren en E Maryniak de ontwikkeling gestart van een pakket voor RObuste Multivariate Analyse, ROMA genaamd. Hierin worden LMS en MVE op efficiënte wijze gekombineerd. De "pilot" versie van ROMA is reeds draaiend beschikbaar op minisuper en PC. Het programma berekent LMS-residuen, op de MVE gebaseerde afstanden in de x-ruimte, maar kan ook voor multivariate lokatie- en schaalschatting alleen worden gebruikt. Het produceert naast het reeds besproken plotje van robuuste residuen tegen robuuste afstanden tevens een plotje van een puntenwolk in het platte vlak, met daarin getekend de klassieke en robuuste tolerantie-ellipsen.

9.6 Resultaten

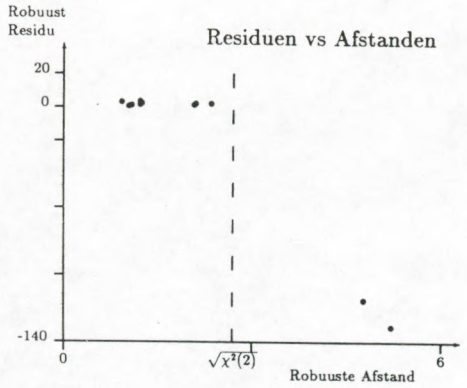
De hierboven genoemde analyse technieken hebben geen enkele moeite met het analyseren van de voorbeeld data set en het diagnostiseren van de punten die de moeilijkheden veroorzaken. Figuur 1 toont de observaties als punten in het (x_1, x_2)-vlak. De grootste ellips is de klassieke tolerantie ellips gebaseerd op rekenkundig gemiddelde en covariantiematrix. Alle punten liggen *binnen* deze ellips, hetgeen verklaart waarom bijvoorbeeld de hat matrix geen leverage punten aanwijst. De smallere ellips is gebaseerd op de MVE. De twee punten die buiten deze ellips vallen zijn inderdaad de gezochte punten 5 en 9.

In figuur 2 zien we de residuen geplot tegen de robuuste afstanden. Er zijn duidelijk twee punten die ver van de andere af liggen en dus leverage punten zijn. Deze punten hebben ook een overdreven groot residu. Uiteraard betreft het hier de gezochte boosdoeners.

Tot slot dient vermeld te worden dat PROGRESS en ROMA standaard ook een zogenaamde "reweighted" LS oplossing berekenen. dat is een gewogen kleinste kwadraten oplossing waarbij de gewichten gebaseerd zijn op de LMS. Dit heeft het voordeel dat men weer kan beschikken over bijvoorbeeld confidentie intervallen en lineaire hypothesen. In deze oplossing is er absoluut geen sprake van het niet significant zijn van de variabele x_1.



figuur 9.1



figuur 9.2

9.7 Literatuur

[1] Peter J Rousseeuw, 1984,
Least Median of Squares Regression,
Journal of the American Statistical Association, 79, 871-880

[2] Peter J Rousseeuw and Annick M Leroy, 1987,
Robust Regression & Outlier Detection,
Wiley: New York.

[3] Peter J Rousseeuw and Bert C van Zomeren, 1988,
Unmasking Multivariate Outliers And Leverage Points by means of Robust Covariance Matrices,
Research Report 88-29, TU Delft